Metodologia integral de perfilat ràpid intel·ligent i interpretable pel suport a la presa de decisions complexes (MIPRI2D). Aplicacions a Sostenibilitat

Xavier Angerri Torredeflot

Departament d'Estadística i Investigació Operativa

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya

Programa de Doctorat en Sostenibilitat
Universitat Politècnica de Catalunya
Directora de la tesi: Karina Gibert Oliveras
Barcelona, Juliol de 2023

Xavier Angerri Torredeflot
Tesi doctoral

DEAI-UPC 2023

LA VIDA

En la vida hi ha pedres i sots,

Cadascú trepitja com vol o com pot

Viure és la manera d'arribar-hi

I l'esperança és la llum amagada que ens guia.

Xavier Angerri Texidó (L'avi Angerri)

AGRAIMENTS

Començar a escriure aquesta pàgina em fa molta il·lusió, doncs és el moment d'agrair a totes les persones que han fet possible aquesta tesi.

En primer lloc, agrair la Dra. Karina Gibert i Oliveras, catedràtica en Intel·ligència Artificial i directora de la tesi. Gràcies per la teva immensa generositat, la teva paciència i totes les estones compartides al llarg d'aquests anys de tesi doctoral. Cada minut compartit amb tu ha estat un aprenentatge en molts aspectes de la vida, especialment acadèmics. El teu coneixement ha estat transcendent per a la realització d'aquesta tesi, així com la teva contribució al món de la ciència amb infinitat d'articles molt valuosos per a la meva recerca.

A la Sra. Montse Guardia, Presidenta del Consell Social de la UPC, pel seu optimisme i els ànims incansables que m'ha enviat durant tants mesos, donant suport en tot moment amb un somriure al seu rostre.

Al Sr. Toni Codina, director de la Fundació iSocial, per tots els aprenentatges apresos en el camí del projecte INSESS-COVID19.

A tot l'equip d'IDEAI-UPC, sempre a punt per donar un cop de mà. Al Víctor Garcia, senzill i humil de cor que escolta sempre amb el cor obert; a l'Esther Parra que amb la seva expertesa i bonança sempre té bons consells per donar-te i moments per compartir; a la Sonia Garcia persona de pau interior, qui sempre m'ha aportat pau i bones vibres; a la Berta Serracanta, detallista i atenta als altres qui va ser la meva primera companya de camí i va portar alegria a IDEA; al Dr. Dante Conti, gran professor, per la infinitat de consells que m'ha donat en aquest camí, pel seu suport i el seu coneixement en l'àmbit de recerca. Les seves contribucions prèvies han estat importants per al desenvolupament d'aquesta tesi; al Sr. Sergi Ramírez, gran professor i company que m'ha ajudat sempre que l'he necessitat amb els seus coneixements posant sempre humor en el dia a dia; al Miquel Umbert i Hasnain Hussain, grans estudiants amb qui ha estat un plaer compartir estones de treball; al Kevin Sánchez, persona somrient i treballadora, que des de la distància sempre ha aportat bon rollo allà on ens hem trobat.

A la tribuKLASS, desenvolupadors de KLASS, el programa clau per a aquesta tesi. Per les seves hores de programació i desenvolupament davant el codi creant un programa "user-friendly" capaç de generar càlculs i gràfics de gran valor afegit i que han constituït un excel.lent punt de partida per aquesta tesi.

A la Colla Castellera Jove de Barcelona i les diferents juntes directives, per acceptar que les enquestes anuals 2021 i 2022 es facin seguint la metodologia INSESS i per entendre sempre la meva posició.

A la beca 2021 FISDU 00409 concedida per la Generalitat de Catalunya que ha permès el desenvolupament d'aquesta tesi, així com el projecte INSESS-COVID19 (2020-L019) finançat pel Centre de Cooperació per al Desenvolupament de la.

A totes les persones que des d'iTNOW per tot el suport per tal que jo pugues tirar endavant aquesta tesi. Al Damiàn Fernando Moreno, qui amb el seu bon rollo em va ajudar a continuar aquesta tesi en un punt crític, a la Sílvia Bosch qui sempre ha confiat en mi, al Lucinio Gutiérrez, qui sempre té frases positives. A moltes persones que estan i han passat per l'equip de Gestió de la Configuració qui en el seu moment em van donar eines per veure la llum. Ells i elles són L'Ana Sanz, el Guifré Rovira, el Carles Hernan, el Miguel Crespo, el Carlos Sebastiani, la Merche Claudin, la Cinthya Colchado, la Cristina Martinez i moltes altres persones amb qui he compartit camí. Una menció especial el Francesc González, que sempre que ho he necessitat ha estat un suport i ha compartit camí amb mi, una persona que escolta i que treballa sense descans per assolir tots els objectius.

A l'equip de Stefanini Spain, en especial a la Beatriz Mendez i a l'Eli Marín, per donar-me el seu suport i animar-me en tot moment.

En l'àmbit personal, també penso que és necessari anomenar a totes les persones que han donat suport en aquest procés i que han aportat el seu granet de sorra.

Al Sr. Artur Torredeflot, el meu tiet que va ser qui, de camí a Boston, en una interessant i potent conversa em va interpel·lar per iniciar aquest intens camí. Gràcies per posar la primera pedra d'aquesta tesi.

Al Dr. Sergi Grau, qui en la primera crisi d'aquest camí em va escriure una frase que ha estat llum, «Qui resisteix, venç».

A la Dra. Júlia Melià, una gran amiga, exemple a seguir en qui sempre he trobat suport i empatia. Excel·lent científica que sempre està disposada a donar un cop de mà.

A la Dra. Maria Huerta, pels seus ànims incondicionals, que sempre i en tot moment ha cregut en aquest projecte. Les seves paraules sempre han estat forces per continuar amb la recerca.

Al doctorand Antoni Ginot, qui podrà llegir la seva tesi amb èxit. Per parlar-me de la beca i acompanyar-me en aquest procés. Per la empatia i les experiències de recerca compartides.

A la doctoranda Ana Matres, gran amiga. Per l'empatia i la força que demostra en cada moment. El seu somriure i el seu carinyo són força per continuar endavant.

A la Laura Madurell, algú qui ajuda sempre amb un somriure d'orella a orella. Per dir sempre les paraules més adequades, per ser-hi malgrat la distància física. Per ser un suport permanent i acompanyar de la manera que ho fa, sempre des del positivisme i l'amor.

A l'Ignacio Maestre, humil i gran amic. Per acompanyar-me en les nits d'estrès mostrant el camí de l'esport com quelcom imprescindible per a fer una bona recerca.

A la família Batlle Juanola, un model de família a seguir qui acull sempre amb els braços oberts. Al Simon, per animar-me sempre i escoltar-me quan ho necessitava. A la Neus, per fer sempre les preguntes adequades quan han estat necessàries que m'han ajudat a triar els camins d'aquest procés. A l'Enric pels seus sentits brindis, al Simon pel seu humor i a la Gemma per la seva alegria.

A l'Alex Tobella, un referent per a moltes persones. I autor de dos llibres. Per tots els esmorzars i sopars compartits. Per escoltar-me i aconsellar-me en els moments més complicats de la necessaris i animar-me a escriure sempre. Per donar-me l'oportunitat de publicar un pròleg al seu llibre.

Al Javi Jordan, mestre de mestres i gran acompanyant. Per tots els esmorzars on posava la veu de la saviesa en els moments més de dubte. Per creure sempre amb mi.

Al Luis Lacanal, un exemple incondicional d'entrega al que fa. Pel seu exemple de lluita i constància en l'estudi i per tots els moments compartits entorn la taula. Persones com ell donen forces per continuar construint un món millor.

Al Joan Carles Bailach, el millor company de pis i millor amic, una persona constant, treballadora i tenaç. Per totes les nits de conversa, per totes les paraules d'ànim en els moments més adients, per la seva comprensió i empatia. Sempre has posat la llum en els moments més foscos i t'has alegrat dels meus èxits.

A la Isa Flaquer, la millor companya de pis i millor amiga, una persona exigent que ajuda a treure el millor de qui té aprop. Per totes les sobretaules compartides, per totes les paraules d'ànim i per ser-hi. Per la seva comprensió amb les meves coses als espais comuns.

A l'avi Angerri, que en pau descansi, que va ser present durant l'elaboració d'[Angerri & Gibert, 2023]. Avi, gràcies per dir-me que jo havia de picar el més amunt possible i confiar sempre en mi.

I com no, a la meva família, la millor família del món que sempre hi és i té allò que sempre necessites abans que ho demanis. Al meu germà Joan, el millor germà del món, qui sempre sap allò que vols abans que obris la boca, un exemple d'esforç i treball, un germà que regala abraçades i alegries. Gràcies, Joan per ser-hi. A la Núria, la millor germana del món, qui sap treure el millor de mi dient-me sempre tot allò que pensa. Per recordar-me sempre el meu potencial i les meves fortaleses i no fer-me desviar del meu àmbit de recerca. Al Marc, una persona humil, que des del silenci fa gran la seva presència. Per les seves paraules, sempre plenes de carinyo i estima que sempre calmen l'ànima i ajuden a seguir endavant, traient el més positiu de totes les situacions. Al papà, un excel·lent mestre i un exemple a seguir, per les abraçades en els moments més necessaris i per la companyia. Pels seus escrits, que ajuden a tirar endavant. Per les seves paraules d'admiració. A la mamà, per totes les converses mantingudes on sempre m'ha donat la llibertat de triar, qui sempre m'ha recordat els meus talents i m'ha ajudat a aixecar-me sempre. Per saber allò que vull abans que ningú, per ensenyar-me el valor de la cultura de l'esforç i del treball, tant transcendents en el transcurs d'aquesta recerca.

I per acabar, dono gràcies a Déu, per totes les senyals que m'ha enviat en els moments més necessaris.

RESUM

Vivint en l'era del big data i la intel·ligència artificial, aquesta tesi contribueix, per una banda, amb una nova metodologia de suport a la decisió basada en dades per decidir ràpidament després de tancar la recollida de dades i, per una altra,amb una infraestructura tecnològica per a processos participatius de recollida de dades. L'objectiu principal d'aquesta tesi és "Definir una metodologia ràpida de diagnòstic d'un domini (territorial o no) basat en tècniques de perfilat, que incorpori l'ús del clustering i el TLP basat en termòmetres com a peces clau d'una nova metodologia d'Intel·ligència Artificial explicable i orientada al suport a la presa de decisions complexes i estratègiques." Així, s'ha desenvolupat MIPRI2D integrant diverses contribucions. La consulta INSESS és l'instrument principal; es creen qüestionaris atemporals per obtenir informació d'alguns moments concrets sense tenir en compte quan respon el ciutadà. La metodologia permet manegar bases de dades heterogènies que impliquen molts tipus de variables, incloent-hi variables de resposta múltiple o temporals i nous tipus, que hem proposat, de més complexes i expressives com Variables de quadrícula, Variables Qualificades Temporals (TQQ) o Variables Bàsiques Temporals (TBV). Es proposa un model conceptual de metainformació (MdM), per a que MIPRI2D pugui tractar qualsevol tipus de qüestionari, sempre que les variables estiguin correctament definides en el model de metainformació. Això permet un processament automàtic del preprocessament de les dades i dels mètodes de mineria de dades que transformen les dades en informació de valor afegit. L'anàlisi descriptiva que proposa MIPRI2D va més enllà de l'estat de l'art amb noves eines per descriure els nous tipus de variables considerats.

El procés d'adquisició de dades proposat i la generació automàtica d'un informe final amb l' anàlisi descriptiva són eines potents per donar suport al desenvolupament depolítiques davant de situacions disruptives. Una altra contribució, és la proposta de càlcul de l'error estadístic de mostreig, per tal que el secret estadístic es mantingui preservat i es redueixi el risc de reidentificació. Es proposen també uns nous tipus de variables derivades de 2a i 3a generació que enriqueixen el conjunt de dades i milloren els resultats del clustering i perfilat de les classes. La tesi també aborda el repte d'obtenir clústers coherents quan les dades tenen una component territorial i presenta una metodologia per identificar les millors variables representatives (TFSM) entre un conjunt d'àmbits, de manera que tant la coherència com la interpretabilitat dels resultats del clustering es conserven des d'un punt de vista geogràfic. La proposta inclou la introducció del termòmetre, unaeina d'adquisició de coneixement que permet introduir la semàntica de les variables en els quadres de semàfor i millorala interpretabilitat dels resultats. La tesi contribueix a trobar clusters a partir de dades que descriguin diversos àmbits, proporcionant resultats interpretables i amb consistència geogràficament. Això

permet obtenir resultats interpretables del clustering de dades multitopic i amb desequilibrades i amb estructura territorial. A la tesi hi ha resultats teòrics i aplicats. El principal resultat teòric es troba en les diverses passes de la metodologia MIPRI2D proposada, una metodologia genèrica adequada per a qualsevol tema i domini d'aplicació. Els resultats pràctics són l'aplicació de MIRPI2D a 4 casos d'us reals amb diferents tipus de consultes píublic/privades que mostren la flexibilitat i versatilitat de la proposta. La metodologia s'ha testejat i validat en el context del projecte INSESS-COVID19, on es van descobrir, interpretar i documentar grups territorials de població vulnerable en menys de 15 dies després del tancament de la recollida de dades. A més, s'han realitzat també aplicacions relatives a la Transformació Digital en entitats de tercer sector, una ONG i dades de consum elèctric.

ABSTRACT

Living in the era of big data and artificiall intelligence this thesis wants to contribute with a new methodology to support data-driven decisions rapidly after data collection, and a technological infrastructure for participatory processess in the data collection step.Thus the main goal is "Define a rapid methodology for making diagnoses in complex domains (territorial or not) based on profiling techniques, incorporating the use of clustering and TLP based on thermometers as key parts of a new explainable Artificial Intelligence methodology and oriented to support complex and strategic decisions". To achieve it, MIPRI2D has been developed thanks to several contributions. INSESS consultation is the main instrument in MIPRI2D. In it, non-temporal questionnaires are created to obtain information from some timestamps disregarding when the citizen answer the questionnaire. The methodology can deal with heterogeneous databases involving many types of variables, including multiple response, compositional or temporal variables and new variable types more complex and expressive proposed in this thesis likeGrid variables, Temporal Qualified Variables (TQQ) or Temporal Basic Variables (TBV). A Conceptual model for metainformation (MdM) is proposed, so that MIPRI2D can deal with any kind of questionnaire, provided that the variables are properly defined in the metainformation model. This allows an automatic processing of preprocessing and data mining methods that transforms data in valuable added information. The descriptive analysis proposed in MIPRI2D goes beyond the state of the art by proposing new tools to describe the new type of variables introduced in the thesis..

The data acquisition process proposed and the automatic generation of a finalist descriptive analisys is a powerfull tool to support basic policy-making in front of disruptive situations where systemic data is non available. A contribution on the calculation of the estimation statistical error is proposed, so that statistical secrecy keeps preserved and risk of redidentification reduced. New types of derived variables named 2^{nd} and 3^{rd} generation variables enriches the dataset and improves clustering and profiling results. This thesis also tackles the challenge of getting coherent clusters when data has a territorial component and presents a new methodology to identify the better representative variables (TFSM) among a set of topics, such that both coherence and interpretability of clustering results is preserved from a geographical point of view. The proposal includes the introduction of the Thermometer, a new knowledge acquisition tool used to introduce the semantics of the variables in the the Traffic Light Panels (TLP) improving the interpretability of the clustering results. The proposal contributes to find clusters on datasets describing several topics, by providing interpretable results with geographical consistency. The present proposal opens the door to get interpretable results of clustering multitopic unbalanced data linked to a territorial structure. In this thesis there are 2 types of results, theorical ones and applied. The main theoretical result

are encompassed in the several steps of MIPRI2D contribution. MIPRI2D is a generic methodology suitable for any topic and application domain. The practical results are the aplication of the thesis proposal to 4 real use cases where it was used for different kinds of private/public consultations proving the flexibility and versatility of the proposal. The methodology is tested and validated in the context of the INSESS-COVID19 project, where territorial groups of vulnerable population were discovered, interpreted and reported in less than 15 days after data collection closure. Also, successfull applications to Digital Transformation in Third Sector entities, NGO and electric consumption data have been obtained as well.

Index

1. Introducció

Vivim en un món ple de dades. Estem envoltats per molts emissors i receptors d'informació, que van des dels propis ciutadans amb dispositius mòbils a dades institucionals, passant per sensors de meteorologia, qualitat de l'aire o mobilitat.

La idea inicial de la tesi era centrar-se en la proposta d'un mètode innovador de ciència de les dades i Intel·ligència Artificial per a la interpretació de perfils automàtics que permetin la comprensió i diagnòstic d'un problema/ecosistema/sistema complex.

En aquest moment, encara lluny de poder construir models predictius, on encara no es coneix bé quines són les variables, quines d'elles són inputs o outputs, quines són causa i quines són efecte, les tècniques de clustering resulten útils per descobrir i caracteritzar l'estructura del domini. I acompanyar-les d'eines que permeten interpretar automàticament els resultats del procés de cluster permet generar resultats que vagin molt més enllà de les simples agrupacions d'objectes que el clustering troba, sinó que facin emergir el significat de les classes trobades de forma automàtica, explicar la raó d'esser d'aquestes classes i aportar una representació simbòlica d'aquests dominis complexos que puguin sustentar processos de comprensió per part de l'expert, suficientment potents com per activar accions o decisions associades als perfils identificats.

Aquestes tècniques, que serien la continuació d'una línia de recerca iniciada als anys 90 per la directora de la tesi, i que utilitzarien el TLP com a eina bàsica, representen excel.lents eines de suport a la conceptualització de les classes i als processos d'aprenentatge inductiu que deriven de l'anàlisi intel.ligent de dades i actualment es troben perfectament aliniades amb el que es coneix com XAI (explainable AI) [Royal Society, 2019] o IA-explicable una nova branca de recerca en IA que està a l'agenda més actual d'aquesta disciplina.

Així, inicialment es va pensar en formalitzar la construcció automàtica del quadre semàfor (Traffic Light Pannel (TLP)), una eina d'interpretació de classes útil en les fases primerenques de comprensió d'un problema complex, quan l'expert, l'analista o el decisor s'enfronten per primer cop amb el problema en qüestió i encara no coneixen bé els mecanismes que regeixen aquella realitat. Fins al moment, l'ús del TLP ha demostrat utilitat en aplicacions prèvies, tot i que s'ha anat utilitzant de forma no automàtica. En aquesta tesi es desenvolupen mecanismes d'intel·ligència artificial que permeten automatitzar la construcció del quadre semàfor de manera objectivable i generar conceptualitzacions de perfils fàcils d'entendre per usuaris no experts en tecnologia. El TLP automàtic s'utilitzarà de forma combinada amb l'aplicació de mètodes de clustering per identificar els patrons bàsics subjacents a un domini complex. Però es voldrà anar una passa enllà i que la construcció automàtica del TLP pugui aprofitar models de representació de coneixement sobre el domini en estudi que permetin ampliar l'automatització a una abstracció simbòlica que descrigui l'estructura del domini tot incorporant un primer nivell de semàntica que tingui en compte el significat concepual de les variables de la base de dades, més enllà de poder estudiar únicament les relacions algebraiques entre les mateixes.

Afegir semàntica als models de dades permet estrenyer el pont entre la mineria de dades i els processos de presa de decisions de les organitzacions i contribueix a insertar bé la dada en aquests processos de decisions, perquè bàsicament facilita la comprensió dels resultats per part dels experts, degut a que venen expressats representant la semàntica de les coses i els experts els poden interpretar de forma autònoma. En aquest sentit la tesi explorarà les possibilitats d'introduir una eina de representació de la semàntica de les variables en el procés d'interpetació de les classes i integrar-lo amb l'ús del TLP.

D'altra banda, el que es planteja a la tesi és la interpretació de classificacions realitzades automàticament a partir de dades. Quan aquestes dades tenen una component territorial esdevé especialment important assegurar que el resultat final serà precisament consistent amb l'estructura territorial. De fet, la presa de decisions globals per un únic territori, podria no ser una bona opció quan el territori manifesta comportaments heterogenis en les distintes zones. Molt sovint, diverses parts de la població es comporten de manera diferent a les altres. La presa de decisions sostenibles, sovint significa prendre decisions diferents associades als diferents escenaris que succeeixen en les diferents àrees del territori, i associar aquestes decisions a zones més petites on els individus són més similars. No obstant això, l'establiment de polítiques per a unitats territorials locals podria no ser convenient, perquè el secret estadístic podria estar en perill quan els grups resultants siguin massa petit.

Trobar mecanismes per identificar grups d'unitats territorials similars que puguin compartir les mateixes polítiques és una bona opció per donar suport a una formulació

de polítiques a un nivell intermedi de granularitat territorial que preservi el secret estadístic i vagi més enllà del territori global més general i les decisions imprecisses. Per trobar una agrupació adequada d'unitats territorials, és interessant explorar tècniques avençades de la família dels mètodes de clustering, que puguin garantitzar la consistència des del punt de vista territorial. Com és sabut, el mètodes de clustering estan orientats a identificar blocs d'individus similars susceptibles de rebre un mateix tractament. Els grups resultants contindran grups d'unitats territorials que comparteixen un únic tractament base o una única decisió per grup (i que eventualment es podrà personalitzar per l'individu concret, però la base de la decisió/tractament serà inicialment comú).

L'aplicació dels mètodes de clustering directament a la base de dades d'individus no garanteix que els resultats siguin coherents des d'un punt de vista territorial. Això significa que els clústers podrien formar-se amb individus similars procedents de llocs molt diferents en el territori, i això farà impossible dissenyar polítiques sota una perspectiva territorial.

Així mateix, les circumstàncies actuals han generat l'oportunitat d'ampliar les mires d'aquesta tesi a un context més ampli, i construir una metodologia nova i global per diagnosticar un territori i expressar els resultats de forma interpretada. Així que la tesi proposa el MIPRI2D (Metodologia Integral de Perfilat ràpid i Intel·ligent i interpretable pel suport a la presa de decisions complexes), una metodologia ràpida de diagnòstic d'un territori basat en tècniques de perfilat i que incorpora l'ús del clustering i el TLP com a peces clau de la metodologia, lligant-la amb processos de presa de decisions estratègiques d'una forma molt més visible, propocionant context al problema de tesi que inicialment ens plantejavem i alhora aportant la possibilitat de realitzar proves de concepte a escala real i sostinguda en el temps, que permetran validar les propostes de tesi amb major contundència.

Així doncs, l'objectiu definitiu de la tesi és definir una metodologia ràpida de diagnòstic d'un domini (territorial o no) basat en tècniques de perfilat, que incorpori l'ús del clustering i el TLP basat en termòmetres com a peces clau d'una nova metodologia d'Intel·ligència Artificial explicable i orientada al suport a la presa de decisions complexes i estratègiques.

Tanmateix, aquesta tesi s'ha desenvolupat majoritàriament en un moment històric, una pandèmia mundial causada pel virus de la SARS-COV2 i que va afectar tots els àmbits de la societat, on les conseqüències han estat devastadores des del punt de vista sanitari, però estem també veient com també ho són des del punt de vista econòmic i del benestar social. La crisi de la Covid-19 va generar una situació mai vista abans i en el moment d'encarar el projecte de tesi, el Centre de Cooperació al Desenvolupament de

la UPC va obrir la convocatòria especial Covid-19 per tal de finançar projectes relacionats amb el coronavirus.

A diferència de molta de la recerca que s'ha fet i es fa en l'àmbit de la Covid-19, que se centra en la predicció de les infeccions, la supervivència, la propagació de la malaltia o el diagnòstic, l'equip de la Dra. Karina Gibert conjuntament amb la ONG Fundació iSocial va presentar a la convocatòria el projecte INSESS-COVID19, el qual va néixer amb la voluntat de posar l'accent en els Serveis Socials (SS), grans oblidats de l'equació en la gestió de la pandèmia i on hi ha fortes necessitats d'aportar-hi la dada com un actiu per a la gestió i la millora del propi sistema i dels serveis als ciutadans.

La proposta va ser una de les 21 guanyadores de la convocatòria. El projecte INSESS-COVID19 (Identificació de Necessitats Socials Emergents a conseqüència de la Covid-19 i efecte sobre els SS del territori, insess-covid19.upc.edu), és un estudi prospectiu per conèixer les vulnerabilitats de la població catalana i aportar elements de decisió a les 107 Àrees Bàsiques de SS de Catalunya que hauran de fer-hi front.

La tesi recull els desenvolupaments teòrics realitzats durant el projecte INSESS-COVID19 com a nucli de la proposta, però generalitza la metodologia proposada a un context més general on l'àmbit d'interès poden ser flors, cotxes o edificis. MIPRI2D desenvolupa un plantejament innovador basat en uns mecanismes d'obtenció ràpida de dades a partir de processos participatius que involucren experts en SS i també ciutadania; una metodologia mixta que combina tècniques de ciència de les dades, gestió del coneixement i intel·ligència artificial, la qual ha permès aportar elements de suport a l'elaboració de polítiques. L'eina tecnològica desenvolupada a INSESS-COVID19, constitueix el nucli de la primera part de la tesi, i l'anomenarem consultaINSESS i constata la viabilitat de poder tenir diagnòstics ràpids de territori sempre que sigui necessari, no només en Serveis Socials, sinó en qualsevol àrea de govern, i permet superar les limitacions dels sistemes d'informació més clàssics en relació amb el suport a la presa de decisions en situacions inesperades.

D'acord amb els objectius inicials de la tesi, on es dibuixa l'interès en aplicacions a problemes reals relacionats amb els 17 objectius de desenvolupament sostenible, els objectius del projecte INSESS-COVID19, centrats a estudiar les vulnerabilitats de la població catalana en els mesos posteriors a la pandèmia i aportar elements de decisió per permetre a les 107 Àrees Bàsiques de Serveis Socials de Catalunya de fer-hi front, estan directament relacionats amb l'acompliment dels següents ODS:

- Objectiu 1: Erradicar la pobresa a tot el món i en totes les seves formes
- Objectiu 3: Garantir una vida sana i promoure el benestar per a totes les persones a totes les edats.
- Objectiu 10: Reduir la desigualtat en i entre països.

- Objectiu 16: Promoure societats pacífiques i inclusives per tal d'aconseguir un desenvolupament sostenible, proporcionar accés a la justícia per totes persones i desenvolupar institucions eficaces, responsables i inclusives a tots els nivells.

A més el projecte incorpora una mirada de gènere, doncs ha obtingut, a més del suport de la Direcció General de SS de la Generalitat de Catalunya, el suport de la Direcció General d'Igualtat, per tant, també contribueix al cinquè objectiu, aconseguir la igualtat de gènere i empoderar totes les dones i nenes.

Així doncs, i atenent que el projecte INSESS-COVDI19 era un projecte d'emergència social, aquesta tesi ha desenvolupat la tecnologia INSESS per un diagnòstic ràpid de territori basat en informació directa de ciutadaniaà, que també treballa la cerca de perfils de ciutadania que responen a determinats patrons de vulnerabilitat social fruit de la crisi de la COVID-19, que s'analitzen aplicant tècniques de clustering i d'interpretació automàtica d'aquests clusters, integrant tots els objectius de la tesi en un projecte marc, real, de gran impacte.

Els resultats inicials del projecte INSESS-COVID19 es van presentar en sessió pública al Palau Robert el passat 15 de desembre de 2020, davant les directores generals de SS i Igualtat i representants de l'Associació Catalana de Municipis, la Federació de Municipis de Catalunya, l'Àrea Metropolitana de Barcelona i la Diputació de Barcelona, amb un altíssim impacte mediàtic (15 notícies a diaris i premsa local i nacional, presència a dues agències de notícies, entre elles EFE, presència al telenotícies migdia i nit TV3, i a diferents televisions locals i ràdios).

La tesi ha abordat finalment el problema des d'una perspectiva més local, on canvis de temàtica de les dades i del qüestionari de referènia puguin ser viables i ha construit una metodologia paraigua (MIPRI2D) que conté totes les altres com a filles.

L'estructura d'aquest document presenta en primer lloc un estat de l'art complert seguit de la introducció a diferents conceptes que són necessaris per a la comprensió de les contribucions d'aquesta tesi, que es troben en el capítol 5. El capítol 5 és el que cetralitza les contribucion de la tesi des del punt de vista metodològic i que presenta la formalització de totes les contribucions. Inicia presentant la visió general de la metodologia MIRPI2D, que s'estructura en 5 fases, començant per les consultes INSESS, el preprocessament de les dades i l'anàlisi descriptiva de les dades. Posteriorment, es defineix el quadre semàfor i la seva aplicació en la construcció automàtica dels semàfors basats en la informàció semàntica que aporta el termòmetre. S'han formalitzat mètodes de generació de noves variables derivades i un nou mètode de selecció de variables. El capítol 6 presenta les validacions d'aquesta metodologia seguides per un capítol 7, que

presenta les 4 aplicacions de la metodologia, donant un especial èmfasi al projecte INSSES-COVID19.

2. Estat de l'art

2.1. Disseny de qüestionaris

Vivim en l'era de les dades, les dades ens envolten, moltes persones prenen decisions basades en les dades i algunes polítiques públiques es basen en enquestes com el Baròmetre Europeu [UE, 2003], on la Comissió Europea mesura l'opinió pública sobre diverses qüestions relacionades amb la UE, com l'economia, les qüestions socials i el medi ambient.

[Alcañiz & Planas, 2011] defineix l'enquesta com una tècnica que utilitza un conjunt de procediments estandarditzats de recerca que es recullen i analitzen diverses dades. Es tria una mostra per representar la població d'estudi que es pretén explorar, descriure, explicar o predir una sèrie de característiques.

Per recollir dades dels participants es desenvolupa un qüestionari a causa de la seva utilitat en estudis a gran escala que impliquen un nombre significatiu de participants que proporcionen informació valuosa.

Per tant, hi ha una gran importància en el pas de disseny del qüestionari. Un qüestionari dolent podria implicar una enquesta inútil perquè s'obtenen conclusions incorrectes. I el que és pitjor, es podrien prendre decisions deficients i afectar la població en la direcció equivocada. En conseqüència, els principals autors han escrit algunes directrius sobre el tema. (Mildred, 2017) havia escrit un llibre amb diverses instruccions, pistes i exemples per dissenyar les preguntes que apareixen al qüestionari. [Alcañiz & Planas, 2011] havia escrit un llibre molt comprensible amb diversos exemples i les principals pistes en tots els passos del qüestionari.

2.2. Eines d'elaboració de formularis digitals

Actualment, a la xarxa es poden trobar multiplicitat d'eines per a la realització de formularis en línia. Les prestacions i capacitats de les diferents opcions és variada. Per tal de fer la tria de l'eina més adient, s'ha fet una anàlisi de les característiques principals d'algunes de les eines disponibles, per tal de triar la que millor s'ajusti a les necessitats del projecte. A [Marra & Bogue, 2006] hi ha una anàlisi comparativa d'eines per fer qüestionaris en línia, però és tan

antiga (2006!) que hem realitzat una recerca específica en aquest àmbit. Les que s'han analitzat, han estat:

- Google Forms: Aplicació de Google que genera formularis. Aquesta aplicació permet un nombre il·limitat de preguntes, facilitant l'opció per a fer filtres de les preguntes
- SuveyMonkey: https://es.surveymonkey.com/
- Encuestafacil: https://www.encuestafacil.com/
- Wufo: https://www.wufoo.com/
- Jot Form: https://www.lancetalent.com/blog/herramientas-crear-formularios-online-mh/
- Typeform: https://www.typeform.com/
- ZohoForms: https://www.zoho.com/forms/
- Formdesk: https://en.formdesk.com/
- CognitoForms: https://www.cognitoforms.com/
- Formsite: https://www.formsite.com/
- Formstack: https://www.formstack.com/
- Arengu : https://marketing4ecommerce.net/herramientas-de-formularios-online/
- Formidable forms: https://formidableforms.com/?utm_source=wprepo&utm_medium=link&utm_campaign=liteversion
- Zerion: https://www.zerionsoftware.com/iformbuilder
- FormBuldier: https://formbuilder.online/
- MicrosoftForms: https://www.microsoft.com/en-us/microsoft-365/online-surveys-polls-quizzes
- Online encuesta: https://www.onlineencuesta.com/
- Survio: https://www.survio.com/es/
- QuestionPro https://www.questionpro.com/es/
- Eval&Go: https://www.evalandgo.com/es/
-

Per tal d'assegurar que es copçava bé la funcionalitat real d'aquestes eines es va dissenyar una petita enquesta de prova que contenia 15 preguntes de diferents tipologies i estructures que permetien testejar les prestacions de les diferents eines. De les 15 preguntes, es van considerar inicialment les següents tipologies

- 2 binàries
- 3 Likert
- 1 oberta
- 1 pregunta tipus filtre binari
- 2 numèrica discreta
- 1 ordinal
- 1 numèrica contínua
- 1 interval
- 1 multivaluada
- 2 nominal

En una primera revisió es va veure que Wufo era només gratuït per 3 formularis i fins a 100 entrades al mes. O des de $ 14.95 al mes permetia fins a 10 formularis i 500 entrades. Jot Form admet 5 formularis i fins 100 entrades al mes en versió lliure i per $ 19 al mes s'accedeix al Plan Bronce que permet 25 formularis I fins 1,000 entrades per mes. Typeform en versió lliure només admetia 10 camps per formulari i 100 respostes per mes. Per $ 35 al mes el Plan Pro admet camps il.limitats I respostes amb salts de lògica. La revisió s'exten per a totes les eines identificades i la principal conclusió és que la majoria limiten o bé el número de preguntes o bé el número de respostes en versió lliure, o bé l'obtenció del fitxer que conté totes les dades recollides. En algunes eines només es pot baixar la descriptiva univariant de les preguntes i l'accés a les dades també és de pagament. Algunes altres no admetien preguntes multivaluades i d'altres no admetien posar preguntes de filtre que determinessin preguntes adicionals concretes per subpoblacions específiques. A les webs

- https://www.lancetalent.com/blog/herramientas-crear-formularios-online-mh/
- https://www.genbeta.com/herramientas/11-herramientas-para-crear-el-formulario-online-perfecto
- https://marketing4ecommerce.net/herramientas-de-formularios-online/
- https://zapier.com/learn/forms-surveys/best-online-form-builder-software/

Es troben recomancions més actualitzades que ens han servit per identificar les eines a revisar. Donat que nosaltres esperavem més de 1000 respostes, no volíem limitacions en el número de preguntes de l'enquesta, ni limitacions temporals en el procediment de recollida de dades. Sabíem que l'enquesta requeriria preguntes addicionals per subgrups de respondents de certs perfils. I volíem poder-nos baixar les dades per fer la nostra pròpia anàlisi multivariant amb les nostres pròpies tècniques. I això d'entrada limitava molt el rang d'eines apropiades per al nostre estudi.

Es va implementar el qüestionari en Google Forms, SurveyMonkey i EncuestaFacil i el resultat va ser que amb EncuestaFacil no es podien descarregar les dades de forma gratuïta,

SurveyMonkey només permet 10 preguntes gratuïtes. Finalment, aprofitant el compte corporatiu de la UPC amb google, que aportava la possibilitat de fer servir Google Forms sense limitacions, es va decidir que l'eina de suport del qüestionari digital seria Google Forms.

No obstant això, hem trobat algunes limitacions que han tingut impacte rellevant en el projecte. La principal és que per les preguntes de filtre l'enquesta no dona cap possibilitat de decidir com omplir el camp que se salta per un cert perfil d'usuari, i per construcció els deixa en blanc. La gestió de blancs és complicada quan es vol fer una anàlisi en R i ha calgut invertir moltes hores en la codificació d'scripts per preprocessar correctament la gestió dels blancs. L'altra inconvenient és que totes les preguntes multivaluades les entrega en format multivaluat, per tant amb totes les respostes separades per comes en una mateixa casella del fitxer de dades, la qual cosa també ha requerit esforços extra de transformació a variables dummy apropiades pel tipus d'anàlisis que necessitàvem. Finalment totes les preguntes que inclouen un camp "altres" amb espai per especificar, les entrega incorporant directament les opcions noves que l'usuari descriu a la variable, sense cap revisió de la qualitat de les respostes, i ha calgut manipular el fitxer de dades revertint aquesta passa per poder concentrar l'anàlisi a les opcions inicialment planificades de les preguntes i realitzar el preprocessament de les opcions noves d'usuari en una segona ronda depurant els continguts i recodificant-los adequadament per tal que representessin opcions noves però correctes de les preguntes.

2.3. Seguretat i Secret estadístic

Actualment totes les lleis que regulen les operacions estadístiques (RGPD, etc), estableixen que cal preservar l'anonimitat del ciutadà que respon, tot preservant la confidencialitat de les dades, garantint que sigui impossible la seva identificació un cop fet públics els resultats de l'anàlisi. A més, cal tenir en compte que quan les dades són sensibles és molt important que siguin secretes i, per tant, és necessari vetllar per la privacitat des del primer moment; des de la recollida de dades fins a la publicació de resultats fins a l'anàlisi seguint el següent:

- Recollida de dades: En aquest pas és molt important vigilar com es realitzen les preguntes, d'una forma que no resulti violenta ni agresiva, i sempre, que al participant li quedi clar que no es violarà en cap cas la seva privacitat i al llarg de l'estudi es compliran tots els protocols de seguretat.
- Emmagatzemament de dades: Desar les dades de forma anònima és transcendent per a la confidencialitat de les dades, així com és molt important que es garanteixi la seguretat del servidor on s'emmagatzemen les dades, garantint que ningú aliè a l'estudi hi pot accedir per fer-ne un us fraudulent.
- Anàlisi de dades: En el moment que aparguin publicats els resultats de l'anàlisi cal garantir que que sigui impossible la reidentificació dels participant.

Tanmateix en cap ocasió es concreta la metodologia que cal seguir per tal de garantir la privacitat I el secret estadístic de l'enquestat. [Damgard, Pedersen & Pfitzmann, 1998] presenten i comparen dues definicions del que anomenen protocols de preservació estadística per acabar concloent cal seguir investigant sobre com protegir el secret estadístic en cas que es vulguin mostrar combinacions de variables on es mostrin diferents individus.

La pràctica clàssica de no publicar resultats sobre subpoblacions massa petites no és una solució en el context d'aquest projecte, ja que les minories vulnerables (fins i tot quan no són estadísticament significatives) requereixen atenció i no poden desaparèixer del panorama (pensem en les dones víctimes de violència domèstica, mai no són massa, afortunadament, i això no és un motiu per amagar en l'anàlisi què passa amb aquest segment de població, per petit que sigui; en aquest cas, és especialment delicat assegurar que cap dels resultats que es publiquin vulnerarà l'anonimat i la total protecció de la identitat que requereixen aquestes persones).

2.4. Mostreig

La mostra, els participants en una enquesta que representen la població no és fàcil de trobar. Per obtenir un mostreig representatiu hi ha algunes tècniques per obtenir-lo.

Els mètodes probabilístics i no-probabilístics es defineixen a [Taherdoost, 2016].

- **Mètodes probabilístics:** Mètodes on tots els individus pertanyents a la població són coneguts i tenen possibilitats de ser escollits com a part del mostreig.
  - Mostreig aleatori senzill: Cada cas de la població té la mateixa probabilitat d'inclusió en la mostra i els participants són seleccionats un per un. La loteria és el millor exemple d'això.
  - Mostreig sistemàtic: Cada cas de la població té la mateixa probabilitat d'inclusió en la mostra, encara que només es fa un dibuix. Tots els participants estan en una llista i el començament està seleccionat aleatòriament.
  - Mostreig aleatori estratificat: El mostreig estratificat és on la població es divideix en estrats (o subgrups) i una mostra aleatòria es pren de cada subgrup. Els estrats són heterogenis, formats per persones amb moltes diferències entre ells.
  - Mostreig de clústers: mostreig de clústers és on tota la població es divideix en grups o grups. Posteriorment, es pren una mostra aleatòria d'aquests grups. Els cúmuls són homogenis, formats per persones sense moltes diferències entre ells.
  - Mostreig en múltiples etapes: El mostreig en múltiples etapes és un procés de moure's d'una mostra ampla a una estreta, utilitzant un procés pas a pas.

- **Mètodes no probabilístics:** Normalment no es coneix la llista d'individus que pertanyen a la població.
    - **Mostreig de quota:** Els participants es trien sobre la base de característiques predeterminades de manera que la mostra tingui la mateixa distribució de característiques que la població més àmplia
    - **Mostreig de bola de neu:** El mostreig de bola de neu és un mètode de mostreig no aleatori que utilitza alguns casos per ajudar a animar altres casos a participar en l'estudi, augmentant així la grandària de la mostra. S'utilitza en poblacions inaccessibles.
    - Mostreig de conveniència: Els participants se seleccionen perquè sovint estan disponibles fàcilment i fàcilment.
    - **Mostreig de propòsit o judici:** Es seleccionen deliberadament persones o esdeveniments particulars per tal de proporcionar informació important que no es pot obtenir d'altres opcions.

A [Blair, Czaja & Blair, 2014] hi ha algunes explicacions relacionades amb la grandària de la mostra i el seu disseny. S'ha de determinar la grandària de la mostra. Seguint [Lopèz & Fachelli, 2015] i [Krejcie & Morgan, 1970] la mida de la mostra depèn dels paràmetres següents:

- N: Grandària de la població
- σ^2: Variància de població de la variable aleatòria a estimar (variables numèriques)
- p: Probalitat de succes (variables qualitatives). q=1-p
- 1-α: Nivell de confiança
- z: α /2 Percentil a Z~Normal(0,1) e: marge d'error acceptat en les estimacions obtingudes.

Quan es fa la inferència per obtenir informació d'una variable numèrica, el paràmetre de raonament és normalment la mitjana de la població. Si la variable és qualitativa, això està directament relacionat amb la taxa d'èxit que és el paràmetre de raonament.

També és necessari distingir si estem treballant amb poblacions finites o infinites. Les expressions que determinen la grandària necessària de n per a una mostra que permet trobar diferències significatives de grau de a (1-α)% de confiança són les següents:

- Si el paràmetre estudiat és la mitjana per a una població infinita:

$$n = \frac{z^2 \cdot \sigma^2}{e^2} \qquad (1)$$

- Si el paràmetre estudiat és la mitjana per a una població finita:

$$n = \frac{z^2 \cdot \sigma^2 \cdot N}{(N-1) \cdot e^2 + z^2 \cdot \sigma^2} \qquad (2)$$

- Si el paràmetre estudiat és la proporció per a una població infinita:

$$n = \frac{z^2 \cdot p \cdot q}{e^2} \quad (3)$$

- Si el paràmetre estudiat és la proporció per a una població finita:

$$n = \frac{z^2 \cdot \sigma^2}{e^2} \quad (4)$$

2.5. Estimació de l'error estadístic

Els resultats de totes les estimacions sobre les dades dels qüestionaris tenen associats errors de mostreig. S'ha consultat a les principals oficines d'estadística en el nostre context i s'utilitzen dos mètodes diferents per a calcular-les.

2.5.1. Institut Estadístic de Catalunya (IDESCAT)

IDESCAT és l'oficina estadística de Catalunya i utilitza el Coeficient Variància (CV) de l'estimació $\hat{\theta}$ com a estimació de l'error de mostreig relatiu per a l'estimació $\hat{\theta}$. CV es publica a les taules d'error de mostreig. El CV estimat permet obtenir un interval de confiança en el 95% de la característica estimada (θ):

$$[\hat{\theta} \pm 1.96\, \widehat{CV} \times \hat{\theta}] \quad (5)$$

Per a les variables numèriques $\hat{\theta}$ és la mitjana observada i per a les qualitatives és la proporció obsegada. El

$$\widehat{CV}(\hat{\theta}) = \sqrt{\widehat{V(\hat{\theta})}}/\hat{\theta} \quad (6)$$

com sempre. Per tant, la part més important en el nostre cas és estimar $V(\hat{\theta})$. Per a variables numèriques, s'estima com el quadrat de la mostra de desviació quasi-estàndard. Per a les variables qualitatives, cada modalitat es considera que segueix una distribució de Bernoulli, de manera que , representa la proporció d'aquesta modalitat, mentre que

$$V(\hat{\theta}) = \frac{\hat{\theta}(1-\hat{\theta})}{n} \quad (7)$$

A més, la confiança de la qüestió qualitativa en el seu conjunt es proporciona mitjançant la desviació estàndard conjunta de totes les modalitats.

Al seu torn, per al càlcul del $\widehat{CV}$. segueix les recomanacions d'Eurostat i el grup de treball Net-SILC2 [Di Meglio et al. 2013], de manera que s'utilitza l'agrupació d'errors i l'enfocament de clúster definitiu. D'acord amb aquesta metodologia, per al càlcul de la variància de l'error de mostreig, només es té en compte la variació entre els totals de les unitats de mostreig primari (els tractes censals). Això podria ser paral·lel al paper de ABSS en el nostre cas.

2.5.2. Institut Nacional Estadístic (INE, Instituto Nacional de Estadistica)

Els errors de mostreig de les estimacions d'algunes de les principals característiques investigades es calculen trimestralment. S'utilitza un mètode de remostreig per obtenir els errors de mostreig. L'INE utilitza el mètode de les semimostres reiterades [EPA, 2005], [EPA, 2022], en la majoria dels seus panells importants, entre ells l'EPA(Enquesta Població Activa) [EPA, 2021].

Aquest procediment consisteix en l'obtenció de r semi mostres a partir de dades (sent una semi mostra una submostra de grandàia n/2, amb n la grandària de la mostra original). A partir de cada semimostra s, es calcula l'estimació ($\widehat{\theta_s}$) . del paràmetre de destinació θ. Un cop s'han calculat totes les estimacions, així com l'estimació de la mostra sencera $\hat{\theta}$, l'estimador de variància ve donat per:

$$\widehat{V(\hat{\theta})} = \frac{1}{r}\sum_{s=1}^{r}\left(\widehat{\theta_s} - \hat{\theta}\right)^2 \quad (8)$$

on r és el nombre de submostres considerades, ($\widehat{\theta_s}$) . és l'estimació de θ. obtinguda amb la semisample s (una tècnica de reponderació s'aplica utilitzant el programari CALMAR) i $\hat{\theta}$és l'estimació global del paràmetre objectiu, basat en la mostra completa.

En el cas de l'EPA, el nombre de iteracions utilitzades és de 40, format fent parelles amb les seccions de cada estrat, assegurant que les dues seccions de cada parell pertanyen al mateix desplaçament de rotació. Cada iteració està constituïda per un nombre de seccions equivalent al 50% de la mostra (semimostra) i cada secció apareix a la meitat de les iteracions. L'enquesta publica l'error de mostreig relatiu com a percentatge (coeficient de variació):

$$\widehat{CV}(\hat{\theta}) = \frac{\sqrt{\widehat{V(\hat{\theta})}} \times 100}{\hat{\theta}} \quad (9)$$

2.6. Eines de preprocessament

El preprocessament és un dels passos en el descobriment de coneixement a partir del procés de dades, com es diu en diversos articles com (Garcia, Ramírez-Gallego, Luengo, Benítez, . Herrera, 2016) que s'aproxima a les tècniques de mineria de dades des d'una perspectiva Big Data. A (Mishra, Alessandra, Regar, Marini, & Rutledge, 2020) es troba una revisió amb diverses aplicacions de cada pas. No obstant això, a (Gibert, Sachez-Marre, & Izquierdo, 2016) s'expliquen les principals tècniques i passos en el preprocessament, mostrats a la Figura 1.

- **Detecció i tractament d'errors:** algunes dades estan malmeses i el valor que indica és fals. Per identificar dades corruptes és necessari interactuar amb un expert a causa del seu coneixement de domini específic. La correcció i la manca de producció són les solucions per resoldre les dades corruptes.
- **Detecció i tractament de dades que falten:** Seguint la literatura principal sobre dades que falten com (Alison, 2002) i (Little & Donald B, 2019) les dades que falten podrien ser aleatòries (produïdes aleatòriament i no segueixen cap patró particular) o no aleatòries (produïdes per causes identificables). La prova de Little és una prova estadística basada en la comprovació de diferències en la distribució multivariant de dades entre el conjunt d'observacions amb un determinat patró de dades que manca i les altres. Les causes de la manca de dades no a l'atzar són: Dades ocultes deliberadament, dades no proporcionades explícitament, dades corresponents a un valor especial, dades no és possible obtenir, dades perdudes, mancades són estructurals. Les dades que manquen es poden representar utilitzant diferents valors (“NA”,”?”,””, 99,...). Un cop detectades les dades que falten s'han de tractar. La millor opció és imputar-los. Com es diu a (Gibert, Sachez-Marre, . Izquierdo, 2016), la imputació és un procés complex per convertir les dades que falten en dades utilitzant tècniques d'estimació, com el mètode MIMMI (Mixed intel·ligent-multivariate missing imputation) presentat a (Gibert, 2013), una metodologia que utilitza tècniques d'agrupació amb un subconjunt de variables sense valors que manquen per imputar els valors que manquen de les variables incompletes amb els mitjans condicionals dels cúmuls resultants. Una altra metodologia d'imputació és MICE (Multiple Imputation by Chained Equations), presentada a (Azur, Stuart, Frangakis, . Leaf, 2011) basada en un model de regressió construït amb les variables completes per imputar les variables incompletes.

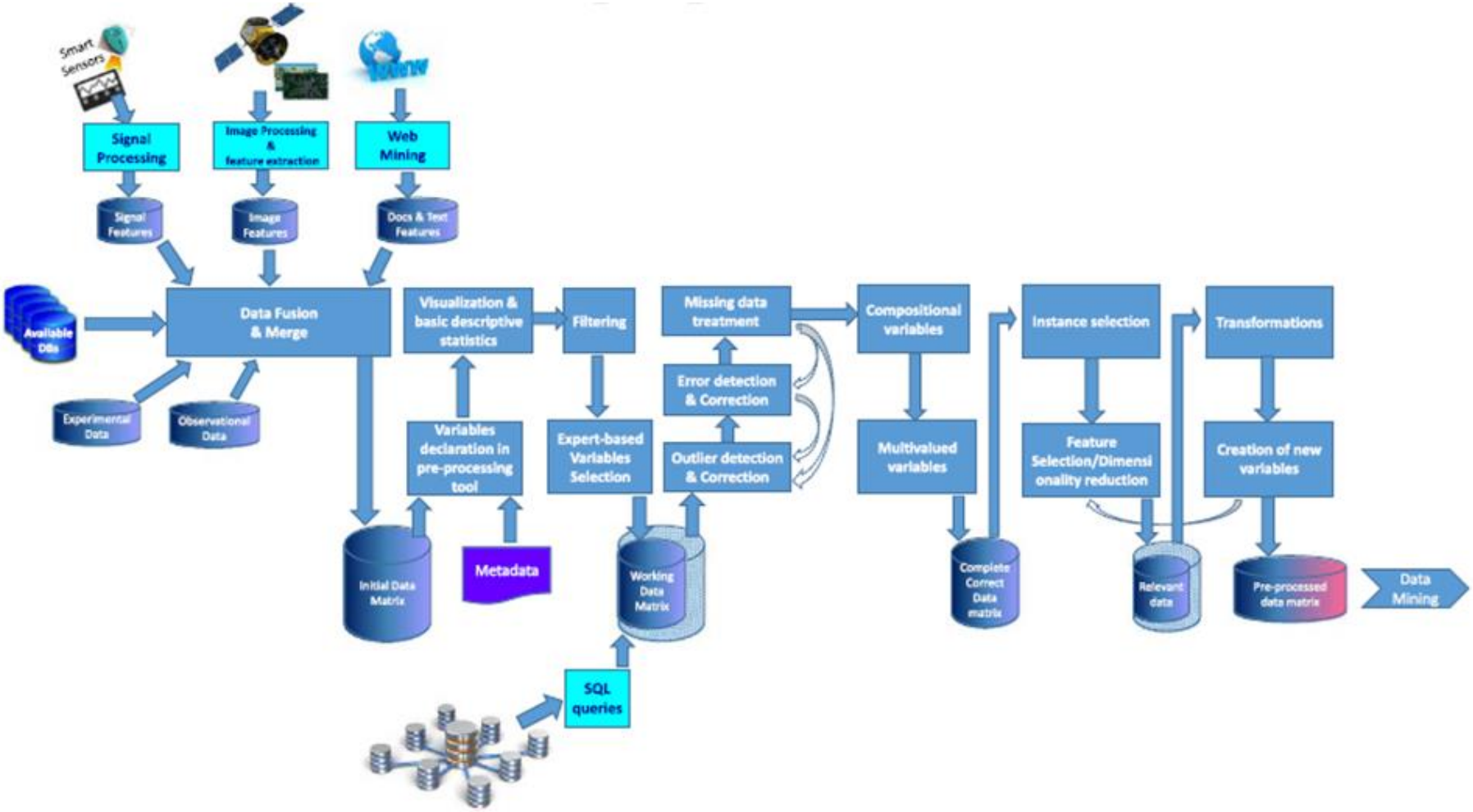

Figura 1 Mètodes de preprocessament

Aquest procés es repeteix fins que no hi hagi més valors perduts. Altres autors com (Farhangfar, Kurgan, . Dy, 2008) o (Luengo, García, . Herrera, 2012) mostren que el millor mètode d'imputació és en funció del mètode de mineria de dades escollit per aplicar a l'anàlisi de dades. A més, els mètodes a distància es poden aplicar com a (Batista & Monard, 2002) i (Rahman . Islam, 2014). Per tractar les dades que manquen és necessari l'expert en el camp, que té els coneixements necessaris per guiar l'estadística per fer una estimació plausible, especialment quan falta no és aleatori

- **Detecció de rellevància o redundància i reducció de la dimensionalitat:** Un cop totes les dades no tenen valors perduts, a vegades hi ha diverses columnes o files que no són rellevants per a l'anàlisi des del punt de vista tècnic perquè no proporcionen cap informació addicional. Reduir el nombre d'exemples del conjunt de dades amb una pèrdua mínima d'informació és l'objectiu principal de les tècniques de selecció d'exemple presentades a (Jankowski & Grochowski, 2004), (Liu). Motoda, 2001) i (Reinartz, 2002). Hi ha dos grups de mètodes per aplicar la selecció d'instàncies: filtra els mètodes de selecció d'instàncies i els mètodes de selecció d'instàncies d'embolcall. En el cas de la rellevància de les variables i la reducció de la dimensionalitat, en cas que no hi hagi experts disponibles 2 tipus de metodologies disponibles. Mètodes de ponderació de característiques equiparen la rellevància si les variables, proporcionant una classificació (pes) de les variables segons el seu grau de rellevància. La selecció de característiques, sent una especialització de la ponderació de característiques on tots els pesos obtenen variables binàries 0 o 1, identifica el subconjunt de les variables rellevants. La següent sessió dona més detalls sobre la selecció de característiques.
- **Creació de noves variables:** El preprocessament és un pas important en el procés de mineria de dades perquè permet obtenir més quantitat i qualitat d'informació utilitzant mètodes precisos. Tècniques de preprocessament com transformacions o creacions de noves variables es presenten a [Gibert, Sanchez-Marre & Izquierdo, 2019] Una d'elles és la idea de crear noves variables que corresponguin als paràmetres de raonament de l'expert. A [Angerri & Gibert, 2023] es presenten diversos mètodes per crear noves variables. Els indicadors i recomptes són exemples d'això. Com es diu, les bases de dades ambientals contenen diverses variables que corresponen al mateix tema, anomenat en aquest document com a bloc. En aquest article es presenta la manera de crear un nou indicador nou de tercera generació basades en dades. Aquesta metodologia es millorarà en aquest document i diversos documents mostraran la manera de crear noves variables. La construcció de nous indicadors augmenta el nombre de variables per bloc.

2.6.1. Gestió de la metainformació

La investigació sobre l'estructura de metainformació és un problema obert. La quantitat de dades augmenta cada dia i gestionar-la és complex. Així, la investigació en el camp de metadades es basa en la creació de programari i tecnologies per emmagatzemar tota la metainformació d'una gran base de dades, com en [Dai et al., 2022] o [Gribova, 2016], on es descriu una metodologia per implementar una eina extensible per editar la metainformació per l'expert en el domini.

Actualment en el context del BigData la metainformació ha agafat molta rellevància. Disposar d'un model de metainformació machine-readable és clau per la interoperabilitat de les dades i els futurs entorns de dades federades. A [Zhao et al. 2017]es treballa la inducció de metainformació a través de models de topic modelling. A [Kim et al. 2023] es proposa un model de metainformació general per dataspaces i orientat a l'etiquetat de corpus. A [Maccioni et al, 2018] introdueixen un framework que permet optimitzar pipelines de preprocessing en datalakes No obstant això, en l'actualitat enara l'accent es posa en fer les dades interoperables i consumibles per un usuari aliè al que les produeix o comparteix, però més a nivell de formats que de poder fer després una interpretació prou acurada de la dada per a poder analitzar els resultats de la mineria de dades.

Els autors no coneixen altres investigacions sobre la relació entre el model de metainformació i la presentació automàtica d'informes de resultats.

2.7. Elaboració automàtica d'informes (Automatic Reporting)

En general, la metodologia d'informes es presenta com a [Boynton & Greenhalgh, 2004] no mostra cap eina d'informes automàtics perquè el camp de l'informe automàtic està més vinculat a la pràctica real que a les publicacions acadèmiques i científiques. No obstant això [Messina, 2022] i [Kaur, Mittal & Singh, 2022] van publicar la seva investigació sobre el valor afegit de la presentació automàtica d'informes en els camps del processament d'imatges mèdiques i [Mathew, 2005] [Lei, 2020] en aplicacions molt específiques i concretes.

2.8. Selecció de variables

Per fer una bona selecció de característiques cal prendre variables amb diferències significatives entre les modalitats. Per abordar aquesta qüestió hi ha diverses metodologies per fer-ho, totes diferents i adients per a cada una de les situacions. Tanmateix, en les bases de dades de sostenibilitat,solen existir variables que fan referencia al territori.. Per saber si una variable es comporta d'una manera diferent en un territori d'un altre és pot utilitzar el Test-Value proposat a [Lebart, Morineau & Fénelon, 1990], per tal que ens indiqui les modalitats significatives de les variables en aquella modalitat. Aquesta nova metodologia es basa en [Gibert, Sevilla-Villanueva & Sànchez-Marrè, 2016]

Quines són les millors variables a usar en l'agrupació final? Aquesta és la qüestió principal a resoldre abans de les variables d'agrupació. Aquest problema s'ha investigat diverses vegades i hi ha diversos documents que resumeixen els mètodes proposats.

[Li et al, 2017] estan proporcionant una visió general exhaustiva i estructurada dels avenços recents en la investigació de selecció de característiques. També, [Li et al, 2017] estan revisant la investigació de la selecció de característiques des d'una perspectiva de dades i revisant els algoritmes de selecció de característiques representatives per a dades convencionals, dades estructurades, dades heterogènies i dades de transmissió. Per a cada tipus de dades, es presenten diversos mètodes. Mètodes basats en similitud, Mètodes basats en la informació-teòrica, Mètodes basats en Sparse-Learning, Mètodes basats en estadístiques es defineixen per a dades convencionals. Per a característiques estructurades hi ha alguns mètodes per a estructures de característiques de grup, estructures de característiques d'arbre i estructures de característiques de grafs. Mètodes per a algorismes de selecció de característiques amb dades enllaçades, selecció de característiques de múltiples fonts i algorismes de selecció de característiques amb dades multivista existeixen en cas de dades heterogènies. Per a la transmissió de dades existeixen algorismes amb característiques de flux i algorismes amb Estructura de dades amb fluxos de dades.

De tots els mètodes de [Li et al, 2017] els mètodes amb més impacte són els següents:

La puntuació laplaciana utilitzada en[He, Cai & Niyogi, 2005] i ReliefF utilitzada en [Robnik-Šikonja & Kononenko, 2003] són mètodes basats en similitud per a dades convencionals.

Mutual Information Feature Selection used in [Battiti, 1994], minimum Redundancy Maximum Relevance used in [Peng, Long & Ding, 2005], Conditional Mutual Information Maximization used in [Fleuret et al., 2004] and Fast Correlation-Based Filter used in [Yu & Liu, 2003] are Information-Theoretical-Based Methods for Conventional Data.

La selecció de característiques amb lp -Norm Regularizer utilitzat en [Tibshirani, 1996] és un mètode basat en Sparse-Learning per a dades convencionals

T-Score utilitzat en [Davis & Sampson, 1986] és un mètode basat en estadístiques per a dades convencionals.

Grup Lasso utilitzat en [Yuan & Lin, 2006]i Superposició Grup dispers Lasso utilitzat en [Jacob, Obozinski & Vert, 2009] són mètodes per a Estructures de característiques de grup per Característiques estructurades.

També trobem literatura per treballs ambit no supervisat estan emmarcades en un model de case-based reasoning com [Nuñez & Sanchez-Marre, 2004] i [Nuñez & Sanchez-Marre, 2005]

2.9. Tècniques de clusterització (Clustering)

La agrupament s'aplica quan no se sap a quin grup pertanyen les dades i és necessari trobar-los. A [Xu et al, 2015] hi ha una revisió interessant sobre l'àrea de la clusterització automàtica. Com es va dir en [Gibert, 1996] clustering és un procés d'agrupament de dades en classes o cúmuls, de manera que les dades del mateix cúmul són bastant similars i diferents d'un grup a un altre. No obstant això, quan s'aplica l'agrupació, s'han de decidir diversos elements. Es podrien aplicar mètodes jeràrquics o de partició. En aquest cas, es treballa a la tesi amb el mètode ascendent jeràrquic amb criteris d'agregació Ward presentats en [Ward, 1963] i utilitzats en diversos estudis.

Per tal de poder realitzar aquest agrupament de dades, cal considerar una mesura de la distància entre dos objectes. Hi ha un gran nombre de mesures de distància. En aquest treball, s'utilitza la distància mixta de Gibert [Gibert & Cortés, 1997] ja que es treballa el cas general en que els qüestionaris a analitzar combinin dades numèriques i qualitatives.

A més, el el clustering podria ser condicionat o no. En aquest document, es presentarà una proposta de metodologia que utilitza el clustering condicionat quan els individus amb una certa característica haurien d'estar en el mateix clúster. En els projectes on l'objecte principal és crear grups territorials, el clustering condicionat és una bona opció. [Lefkovitch, 1980] presenta diversos exemples de Clustering Condicionat,.

2.9.1. Clusterització multivista (Multiview clustering)

El Clustering Multiview tracta l'alta dimensionalitat de les dades [Sevilla-Villanueva, Gibert & Sanchez-Marre, 2015], [Sevilla-Villanueva, Gibert & Sanchez-Marre, 2017]. En aquest enfocament, les variables es divideixen en diversos grups segons diferents temes o temes referits al conjunt de dades (com situació de treball, biomarcadors, opinions, etc.). Cada grup o vista s'analitza independentment de l'altre, encara que algunes vistes es poden agrupar en un grup més gran i analitzar-les junts; això ajuda quan els temes de diverses vistes estan relacionats o contenen poques variables i diversos blocs que tenen sentit s'agrupen. Llavors, els objectes s'agrupen sota cada grup de variables o visió.

Una vegada que totes les vistes estan agrupades i analitzades, apareix una nova variable qualitativa. En l'agrupació multivista, per fer un cúmul general, s'utilitzen les variables de classe resultants de cada vista. Les tècniques de clúster multivista s'utilitzen a [Bickel & Scheffer, 2004] on les variables es divideixen en dos grups independents pel seu significat.

2.10. Estat de l'art sobre els Serveis Socials a Catalunya

Donat que el projecte INSESS-COVID19 ha estat un dels motors més rellevants per al desenvolupament d'aquesta tesi i que l'aplicació de la metodologia proposada a les dades de vulnerabilitat ha representat l'aplicació més gran de la tesi, s'ha considerat dedicar un apartat

específic de l'estat de l'art a l'àmbit d'aplicació dels Serveis Socials a Catalunya. En aquest apartat es tractarà l'estat actual de l'estructura dels Serveis Socials de Catalunya, el servei on més ha impactat aquesta tesi. En primer lloc, es pot observar la revisió realitzada, l'estructura actual de la cartera de Serveis Socials i una de les mostres, així com una revisió dels qüestionaris i informes existents en el moment d'iniciar la recerca i realitzar el qüestionari. Aquest apartat és transcendent per tal de poder realitzar correctament el qüestionari INSSES-COVID19

2.10.1. Revisió del marc conceptual dels Serveis Socials

Consisteix en la recerca, lectura i anàlisi d'articles científics que tracten sobre els aspectes estadístics a treballar, així com la revisió de fonts d'informació i elements existents sobre les eines a utilitzar o bé per tal de conèixer informació de context del camp d'aplicació.

En el cas específic de la revisió del marc conceptual dins el projecte INSESS-COVID19, es refereix a la fase de documentació i revisió de la mateixa per entendre l'estructura orgànica i la de les dades de SS a Catalunya, la revisió de les principals enquestes i estadístiques oficials amb informació rellevant per SS, així com en l'aprofundiment del coneixement de l'estructura de SS a Catalunya, a nivell de departament i d'Ajuntaments i Consells Comarcals. També la revisió d'informes oficials amb les xifres de població disponibles per poder determinar la grandària de la mostra esperada i encaminar el disseny mostral.

En aquesta fase de la tesi s'ha completat la revisió del marc conceptual dels SS de Catalunya i s'ha realitzat una primera revisió dels aspectes estadístics més rellevants per la realització de l'informe INSESS-COVID19, com el càlcul de la variància dels estimadors. En el segon any de tesi s'aprofundirà en la revisió crítica de la literatura en referència a les contribucions metodològiques de la tesi.

2.10.2. Estructura dels SS a Catalunya

L'article 1 de la Constitució Espanyola conté un mandat per tal que els poders públics exerceixin un paper promocional del benestar social: «España se constituye en un Estado social y democrático de Derecho, que propugna como valores superiores de su ordenamiento jurídico la libertad, la justicia, la igualdad y el pluralismo político» [Alemán, 1993]. La Ley de Bases de Régimen Local (1985), en l'article 25.2.k) diu que el municipi exercirà competències d'acord amb la legislació estatal i autonòmica en matèria de prestacions de SS i de promoció i reinserció social. A l'article 26.1.c) diu: «Los municipios con población superior a 20.000 habitantes deberán prestar en todo caso Servicios Sociales». I l'article 36 diu «son competencias propias de la Diputación la prestación de servicios públicos de carácter supramunicipal, y en su caso supracomarcal».

Així, el 23 de juny de 1980 el Govern de la Generalitat nomena, en sessió parlamentària, el primer Director General de SS de Catalunya [Vila, 2004]. Entre l'estiu de 1980 i del 1981 es

transfereixen les competències en SS de l'Estat i de la Seguretat Social (IMSERSO) a Catalunya, i es comença a desplegar l'estructura de SS. La primera llei de SS de Catalunya s'aprova el 27 de desembre de 1985 (Llei 26/1985, de 27 de desembre).

L'estructura que actualment tenen els SS a Catalunya, que bàsicament es divideix en dos grans subsistemes, d'acord amb el marc legal estatal d'origen:

- Els SS d'atenció primària
- Els SS especialitzats

L'atenció primària es realitza al món local, en una estructura de 107 Àrees Bàsiques de SS que són de dos tipus:

- ABSS municipal: Tots els municipis de Catalunya que tenen més de 20.000 habitants tenen àrea d'atenció primària en SS a l'Ajuntament
- ABSS comarcal: Tots els municipis de Catalunya de menys de 20.000 habitants depenen d'un servei comú d'atenció primària en SS que s'articula a través dels Consells Comarcals, que agrupen el servei per tots els municipis «petits» d'aquella comarca.

Els serveis especialitzats depenen de la direcció general de SS del govern de la Generalitat, que actualment depèn del departament de Treball, Afers Socials i Famílies. La Figura 2 representa la cartera de SS especialitzats i d'atenció primària que actualment existeixen. La Llei 12/2007, d'11 d'octubre estableix l'actual marc de gestió dels SS a Catalunya, els quals es defineixen per períodes de 4 anys entre els ens locals (les ABSS, siguin municipals o comarcals) i el Departament, a través dels contractes programa, on es detalla la cartera de serveis d'atenció primària que el Departament cofinança per cada territori.

El 10 de desembre de 2020 el Parlament de Catalunya ha aprovat el nou Pla Estratègic de SS que revisa la cartera de serveis a la llum de la problemàtica emergida de la crisi de la COVID-19. El CG del 29 de desembre de 2020 aprova definitivament aquest nou Pla.

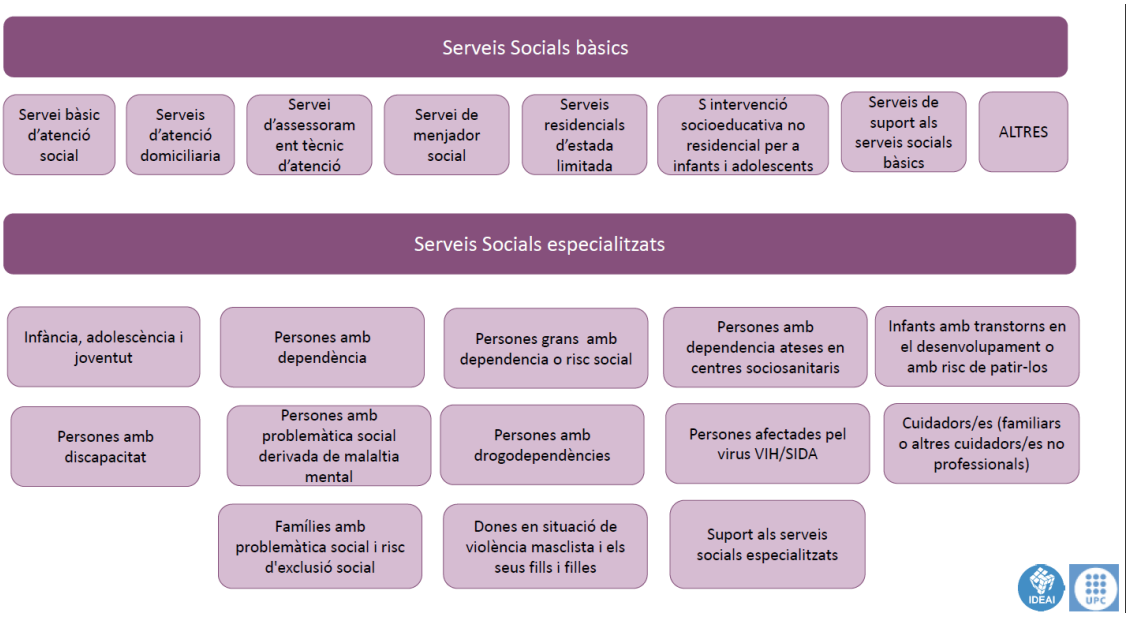

Figura 2 Estructura dels SS de Catalunya I cartera de serveis

2.11. Estructura dels qüestionaris sobre vulnerabilitat social i informes de referència existents

Com a part de la revisió del marc conceptual dels SS de Catalunya, s'ha fet una recerca sobre els principals qüestionaris de referència al territori en matèria d'estadística oficial i informes de referència en SS, així com d'enquestes específiques sobre la crisi de la COVID-19 que s'havien posat en marxa fins al moment, per tal d'observar el tipus d'informació que es demana, i la forma com es formulen les diferents preguntes. S'ha elaborat una anàlisi que queda resumit en la taula següent:

Tipologia	Títol	Entitat promotora	Link
Enquesta especial COVID19	Survey on the impact of COVID-19. 2020	CEO (Centre d'Estudis d'Opinió)	http://ceo.gencat.cat/ca/estudis/registre-estudis-dopinio/estudis-dopinio-ceo/societat/detall/index.html?id=7588
	Survey on time uses in lockdown. 2020	CEO	http://ceo.gencat.cat/ca/estudis/registre-estudis-dopinio/estudis-dopinio-ceo/societat/detall/index.html?id=7608
	Special Barometer May 2020	CIS (Centro de Investigaciones Sociológicas)	http://www.cis.es/cis/opencms/ES/NoticiasNovedades/InfoCIS/2020/Documentacion_3281.html
	Covid 19 Impact Survey	Dr. Nuria Oliver, commissioned for AI and COVID-19. Generalitat Valenciana	https://covid19impactsurvey.org/
	Gestioemocional.cat	Health department, Generalitat de Catalunya.	https://gestioemocional.catsalut.cat/
	Social Service Survey	ACM	https://docs.google.com/forms/d/e/1FAIpQLSe7MBgTSeA4NtfWIzWM3yDtdsVUXHX118r-FvwHXLimgvKVCA/viewform
Enquesta oficial de referència	Encuesta Condiciones de Vida	INE (National Statistics Institute)	https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176807&menu=ultiDatos&idp=1254735976608
Informes de referència	Anàlisi de les necessitats socials de dones i homes	Fundació "la Caixa"	https://observatoriosociallacaixa.org/ca/informe-necesidades-sociales-mujeres-y-hombres?utm_source=newsletter&utm_medium=email&utm_campaign=ObservatorioSocial.Newsletter%20junio1&utm_content=CAT%20Tema%20de%20portada%20leer%20m%C3%A1s&utm_term=General
	Condicions de vida de les treballadores de la llar i les cures centreamericanes a Barcelona	Centre d'Informació per a Treballadors Estrangers (CITE)	https://www.ccoo.cat/pdf_documents/2020/estudi_CITE_llar_2020.pdf
	La Matriu d'autosuficència SSM-CAT (I ES)	Generalitat de Catalunya. Departament de Treball, Afers Socials i Famílies	https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwisl8Sb0Z7uAhUVjhQKHW8qBd0QFjACegQIAhAC&url=https%3A%2F%2Fwww.fadq.org%2Fwp-content%2Fuploads%2F2020%2F06%2F01_02.-2020_05_19_Eina-SSM-CAT_M_BALLESTER.pdf&usg=AOvVaw34TEc80iGsf0ccdm2M6ujn

Taula 1: Revisió d'enquestes I informes de referència

Es van analitzar també els diferents temes sobre els quals tracten els qüestionaris especials COVID-19. La taula 1 recull el contingut d'aquests qüestionaris.

Titol	Percepció situació actual	Salut	Impacte economic	Impacte en la societat (de valors)	Condicions del confinament	Valoració política	Dades Sociodemo-gràfiques	Situació respecte al treball en el confinament	Tasques domèstiques al confimante	Violencia masclista	Situació serveis socials	Actuacions realitzades
Enquesta sobre l'impacte de la COVID-19. 2020	x	x	x	x	x	x	x					
Enquesta sobre els usos del temps i el confinament. 2020		x			x		x	x	x	x		
BARÓMETRO ESPECIAL DE MAYO 2020	x	x	x		x	x	x					
Covid 19 Impact Survey		x	x		x	x						
GestioEmocional.cat		x										
Impacte social de la COVID-19 a Tarragona	x	x	x				x	x				
Impacte social de la COVID-19 a Tarragona	x	x	x				x					
Condicions de vida												
SSM-cat												
Enquesta Serveis Socials		x	x								x	x
Encuesta Condiciones de Vida		x	x				x	x				

Taula 2:Àmbits continguts a les enquestes de referència

2.11.1. El model Self Sufficiency Matrix

D'altra banda, el govern Català adopta el model SSM.cat [DGSS,2019] com a instrument per a avaluar la vulnerabilitat social de les famílies, com a eina nuclear del nou Sistema digital de SS de Catalunya (s-Social), en el marc de la transformació digital dels SS anunciada en el Pla Estratègic de SS de Catalunya [PESS 2020] aprovat el 29 de desembre de 2020.

El model SSM.cat està inspirat al seu torn en la versió holandesa del Model de Matriu d'Autosuficiència, desenvolupat per Diana Pearce per les Wider Opportunities for Women, com a part del projecte estatal Family Economic Self-Sufficiency [Pearce, 1996] [Brooks, Pearce 2000]. Aquest model defineix un concepte de vulnerabilitat social que s'expressa sobre 11 àrees de la vida, com es pot veure a la figura 3.:

Economia	Abús de substàncies, i conductes addictive
Feina i Formació	Activitats de la vida diària
Ús del temps	Activitats instrumentals de la vida diària
Allotjament	Relacions socials
Relacions convivencials	Participació en la comunitat
Salut mental	Aspectes judicials i d'ordre públic
Salut física	

Figura 3: Àmbits compresos al model SSM.cat

Aquesta revisió és la base per la construcció del questionari INSESS-COVID19 que més endavant es detallarà.

3. Antecedents i conceptes bàsics

3.1. Introducció

En aquesta tesi, com s'ha vist en el capítol anterior està emmarcada en l'àmbit de l'estadística avençada i la ciència de dades tot incorporant eines d'Intel·ligència Artificial, amb una forta component aplicada i inclou aplicacions en l'àmbit dels Serveis Socials, el Tercer Sector, les associacions sense ànim de lucre i educatiu. Tanmateix, les contribucions d'aquesta tesi pivotaran sobre conceptes de naturalesa estadística, concretament en tècniques d'anàlisi multivariant, intel·ligència artificial, en les seves vessants d'aprenentatge no supervisat, interpretació automàtica de resultats, i generació automàtica d'informes, i la ciència de dades, des d'on s'integren diferents disciplines intensives en dades per a extreure'n valor afegit per la presa de decisions estratègiques usades l'àmbit de la mineria de dades. Aquestes tècniques són fonamentals per a l'extracció de conclusions que serveixin per a prendre decisions, el gran objectiu d'aquesta tesi.

Això requereix utilitzar una sèrie d'antecedents sobre els quals se sustentaran les contribucions de la tesi, que es descriuen tot seguit per a poder tenir en aquest document la informació necessària per a la comprensió del capítol 4 on es presenta la proposta de tesi i totes les contribucions metodològiques.

3.2. Eines per a la interpretació automàtica de classes i perfilat

A continuació s'introdueixen diferents eines orientades a la interpretació de dades en contextos de clustering, que s'han desenvolupat en els darrers 20 anys en el grup de recerca de la Dra. Gibert. Aquestes eines jugaran un paper fonamental en el desenvolupament de tesi que es proposa tot seguit.

3.2.1. Panell de classes (Class Panel Graph, CPG)

Un CPG és una representació gràfica de les distribucions condicionals de variables versus un conjunt de clústers en un sol panell, proposat en [Gibert, Garcia-Rudolph & Rodriguez-Silva, 2008]. Els clústers poden provenir d'un algorisme d'agrupament anterior. Els detalls sobre l'aplicació de CPG per a la interpretació automàtica de clústers es poden trobar a [Gibert, Conti & Vrecko, 2012], directament relacionats amb el camp AI explicable [Royal Society, 2019] [Miller, 2019]. Molts autors proporcionen enquestes que donen una àmplia perspectiva de XAI des de diferents enfocaments, però la majoria es refereixen al camp d'aprenentatge automàtic supervisat com es pot veure a [Burkart & Hubert, 2021] [Vilone & Luca, 2021] [Guidotti et al. 2018] [Alonso, Castiello & Mencar, 2018] o a aplicacions específiques com sistemes de recomanació [Tintarev & Masthoff, 2007]. [Vellido, Martin-Guerrero & Lisboa 2012] es destaca la importància de la visualització per a l'explicabilitat i a [Holzinger et al. 2019] la necessitat d'anar més enllà, analitzant també la causalitat, especialment en el camp mèdic. En el nostre treball, aprofundim en l'ús d'eines visuals per ajudar a la interpretació automàtica dels clústers, el resultat d'una família de mètodes d'aprenentatge automàtic no supervisat, per tal de facilitar la inducció de conceptes associats als clústers, contribuint així a explicar què signifiquen, representen i com es van formar.

El CPG visualitza el comportament conjunt de moltes variables respecte a les classes de manera compacta i proporciona a l'expert una comprensió ràpida de les particularitats de cada classe, facilitant el procés d'etiquetatge de classe, és a dir, identificant el concepte representat per la classe, la interpretació. El CPG, de fet, facilita l'aprenentatge inductiu als experts i ajuda a conceptualitzar els clústers. No obstant això, aquest recurs va resultar ser encara complex per a aquells interessats sense habilitats tècniques i més tard es va evolucionar a TLP per salvar la bretxa per a les parts interessades no tècniques.

3.2.2. El quadre semàfor (Traffic Lights Panel TLP)

Un dels eixos centrals de la Tesi doctoral es basarà en l'automatització de la interpretació de les diferents classes obtingudes després d'una classificació i haver obtingut el panell de classes. La tècnica emprada per a la interpretació de les classes s'anomena Traffic Light Pannels (TLP), la qual podria ser traduïda al Català com a tècnica del quadre semàfor.

La tècnica del TLP com a mètode per interpretar classes [Gibert, Conti & Vrecko, 2012] va ser introduïda per la Dra. Karina Gibert després de realitzar histogrames o boxplots i representar-los en un panell de classes [Gibert, Garcia-Rudolph & Rodriguez-Silva, 2008]. El CPG va suposar en els seus orígens una eina interessant de suport a la interpretació perquè permetia representar conjuntament el comportament de moltes variables respecte de les classes de forma compacta. I una comprensió ràpida de les particularitats de cada classe per part de l'analista. Aquest recurs resultava complex encara per aquells experts en camps d'aplicació que no tenien competències tècniques. El TLP suposa una abstracció simbòlica de la

informació del CPG més interpretable, basada en identificar els nivells dominants de cada variable en cada classe.

Per fer un TLP, l'analista ha de llegir atentament el CPG, marcar la tendència central per a cada classe de cada variable i assignar als nivells qualitatius els colors del quadre semàfor, en correspondència amb els codis interpretatius de l'expert. El context i significat de cada color ha d'estar relacionat amb algun concepte latent del domini que permet l'associació entre la polaritat variable i la idea de millora o empitjorament.

A [Gibert, Conti & Vrecko, 2012] es proposen dues maneres bàsiques d'assignar colors a l'escala de la variable.

a. Codis de color directes (vermell-groc-verd) associats als valors baixos-mitjans-alts.
b. Codis de colors inversos (verd-groc-vermell) associats als valors baixos-mitjans-alts.

Una propietat molt important és que si les classes estan ben construïdes, han de ser distingibles i han de representar diferents perfils. Per tant, no hi hauria d'haver dues files del TLP amb la mateixa combinació de colors.

El TLP culmina el camí que es recorre des de la recollida de dades fins a l'obtenció de coneixement nou sobre un cert fenomen, objectiu principal d'àrees com l'estadística, la mineria de dades i recentment el data science. El TLP aporta doncs el suport per la interpretació de les classes.

A [Angerri, 2015] es fa una primera proposta d'automatització del TLP que es basa sobre la utilització del mètode d'inducció de regles basada en boxplots proposat a [Gibert & Perez, 2006] i consisteix a induir una discretització de la variable numèrica a partir del condicionament per la variable de classe per més tard estudiar com es creua aquesta variable discretitzada amb les mateixes classes i detectar les tendències centrals de cada classe de forma automàtica a partir de la taula creuada corresponent. El treball de fi de grau esmentat planteja una proposta que en aquesta tesi s'explorarà amb major profunditat.

3.2.3. Annotated Traffic Ligth Panel (Quadre semàfor anotat)

A [Gibert & Conti, 2015] es presenta una millora per als TLP presentats anteriorment, l'*annotated Traffic Ligth Panel*(aTLP), els quals serveixen per a gestionar la incertesa intrínseca que es produeix quan s'interpreten els prototips resultants d'una classificació. El CPG es limitava a representar la tendència central de les variables en les classes, però l'anàlisi de les variables associades s'havia de fer amb unes taules d'estadístiques bàsiques condicionades a les classes amb el propi KLASS que presenta entre altres les desviacions típiques de cada variable en cada classe. Els aTLP associen als colors del quadre semàfor amb dues dimensions del color (to i saturació) que serveixen per a mesurar la tendència central i la puresa dels

prototips basant-se en els Coeficients de Variació (CV) i un model d'incertesa. En aquest sentit, els colors purs representen baixa o nul·la variabilitat en la tendència central de les variables, mentre que els colors més enfosquits representen un increment en l'heterogeneïtat i per conseqüència una pèrdua de fiabilitat en la presa de decisions basades en les caselles més fosques del TLP. A [Gibert & Conti, 2015] es presenta el model de càlcul automàtic de la saturació de cada color calculat sobre la base dels coeficients de variació o el factor d'incertesa de cada variable en cada classe. La figura 4 mostra el model de gradació o pèrdua de puresa dels colors per a coeficients de variació en un interval entre [0,1].

RED Scale

x	R	G	B	Color
0,00	255	0	0	
0,05	244	0	0	
0,10	233	0	0	
0,15	222	0	0	
0,20	212	0	0	
0,25	202	0	0	
0,30	192	0	0	
0,35	182	0	0	
0,40	173	0	0	
0,45	164	0	0	
0,50	155	0	0	
0,55	146	0	0	
0,60	138	0	0	
0,65	130	0	0	
0,70	122	0	0	
0,75	114	0	0	
0,80	107	0	0	
0,85	100	0	0	
0,90	93	0	0	
0,95	86	0	0	
1,00	80	0	0	

GREEN Scale

x	R	G	B	Color
0,00	0	255	0	
0,05	0	244	0	
0,10	0	233	0	
0,15	0	222	0	
0,20	0	212	0	
0,25	0	202	0	
0,30	0	192	0	
0,35	0	182	0	
0,40	0	173	0	
0,45	0	164	0	
0,50	0	155	0	
0,55	0	146	0	
0,60	0	138	0	
0,65	0	130	0	
0,70	0	122	0	
0,75	0	114	0	
0,80	0	107	0	
0,85	0	100	0	
0,90	0	93	0	
0,95	0	86	0	
1,00	0	80	0	

Proposed YELLOW Scale

x	R	G	B	Color
0,00	255	255	0	
0,05	255	244	0	
0,10	254	233	0	
0,15	253	222	0	
0,20	252	212	0	
0,25	251	202	0	
0,30	249	192	0	
0,35	247	182	0	
0,40	245	173	0	
0,45	242	164	0	
0,50	239	155	0	
0,55	236	146	0	
0,60	232	138	0	
0,65	227	130	0	
0,70	222	122	0	
0,75	217	114	0	
0,80	211	107	0	
0,85	204	100	0	
0,90	197	93	0	
0,95	189	86	0	
1,00	180	80	0	

Figura 4: Model de gradació de o pèrdia de puresa de colors

El model de color es basa en el model RGB que representa els colors com un vector de 3 components, descomponent els colors en quantitats de vermell, verd i groc, cada component pren valors entre 0 i 255 on 0 representa l'absència del color i 255 la presència del color amb la màxima saturació. El grau de saturació d'un color d'acord amb el coeficient de variació x:

- Per als colors vermell i verd (primaris): $S(x) = 80 + 125(1-x)^2$
- Per al color groc (color compost): $S'(x) = 180 + 180(1-x)-143(1-x)2+38(1-x)^3$

Per a determinar el valor de x s'usa un identificador de la variació interna de classe com s'indica en [Gibert & Conti, 2015]. Si X_k és numèrica x = CV|C sent C la classe de la qual es vol calcular el coeficient de variació. Si X_k és qualitativa, es correspon a la proporció de valors diferents a la freqüència dominant de la classe. És a dir, per a una classe identificada majoritàriament com a dones, es correspondria a la proporció d'homes que conté la classe.

3.3. KLASS

Java-KLASS és un software estadístic creat per la Dra. Karina Gibert a la dècada dels 90, amb l'objectiu de poder automatitzar la classificació de dominis poc estructurats. És un programa

que evoluciona constantment, al qual es van afegint cada dia noves funcions. A continuació, es presenta el programari més detalladament.

3.3.1. Antecedents

Fruit del treball i estudi sobre la classificació de dominis poc estructurats va sorgir la primera versió de KLASS, un sistema orientat a la classificació automàtica de dominis poc estructurats. Aquest paquet ha anat evolucionant de forma continuada des de l'aparició de la primera versió que formava part de la tesina [Gibert, 1991] i també de la tesi doctoral [Gibert, 1994] de Karina Gibert i ha estat objecte dels antics projectes finals de carrera, tant de la diplomatura d'estadística com de les Enginyeries en Informàtica de la UPC i de la UIB (Universitat Illes Balears) i d'alguns treballs finals de grau o treballs finals de màster dels plans nous. L'eina està integrada per un conjunt d'eines per gestionar i ajudar experts en els processos de mineria de dades.

El sistema en la seva primera versió va ser desenvolupat en el llenguatge de programació LISP i la seva execució es realitzava sobre UNIX. Quan la UPC va deixar de mantenir les llicències de LISP, va ser reimplementat en JAVA, pels avantatges que aquest llenguatge de programació presentava, com l'eliminació de costos de llicència, la portabilitat, una aplicació desenvolupada en JAVA pot executar-se en qualsevol sistema operatiu i la possibilitat de distribuir un programari executable independent del seu codi font, entre altres.

Actualment, KLASS és un paquet que s'està utilitzant en diferents entorns i per aquest motiu un dels objectius principals que es van marcar era que les noves versions havien de mantenir sempre totes les funcionalitats de les anteriors com a mínim.

Actualment Java-KLASS [Gibert & Nonell 2008] i [Gibert & Nonell 2005a] inclou 5 MB de codi Java i 160.000 línies de codi i ofereix funcionalitats relacionades amb la identificació i interpretació de perfils i tècniques de gestió de dades, coneixement i modelat relacionades.

3.3.2. Funcionalitats de Java-KLASS

Es presenta el conjunt de funcionalitats que dona JAVA-KLASS en la seva versió més recent:

- Representació de matrius de dades.
- Representació i gestió de la metainformació associada a les dades.
- Selecció de variables i individus basada en criteris per generar submatrius basades en mostreig aleatori.
- Recodificació o discretització de variables i generació de variables noves.
- Preprocessament de dades incloent mètodes diversos de tractament de dades mancants, entre ells el MIMMI [Gibert 2013]

- Gestió de Bases de coneixement.
- Gestió d'Ontologies[Gibert, Valls, Batet, 2014]
- Gestió de termòmetres
- Estadística descriptiva extensa de dades.
- Visualització 3D
- Càlcul de distàncies simples, mixtes o semàntiques
- Classificació automàtica amb mètodes jeràrquics clàssics, basats en regles, en ontologies i amb mètodes basats en densitats com BDSCAN o OPTICS.
- Visualització i gestió d'arbres jeràrquics (dendrogrames) de clustering, reachability plots, tall de l'arbre.
- Raonament automàtic
- Interoperabilitat de mètodes
- Anàlisi dinàmica [Gibert 2010].
- Interpretació automàtica de les classes amb CPG [Gibert, Garcia-Rudolph & Rodriguez-Silva, 2008], amb TLPs [Gibert, Conti & Vrecko, 2012] [Gibert, 2013], amb aTLPs [Gibert & Conti, 2015] i conceptualment [Gibert, 2014].
- Gestió de Sistemes heterogenis que incloguin informació numèrica, qualitativa, semàntica i ontologies.

3.3.3. Cronologia

A continuació, es resumeix breument el desenvolupament del sistema Java-KLASS fins a la darrera versió

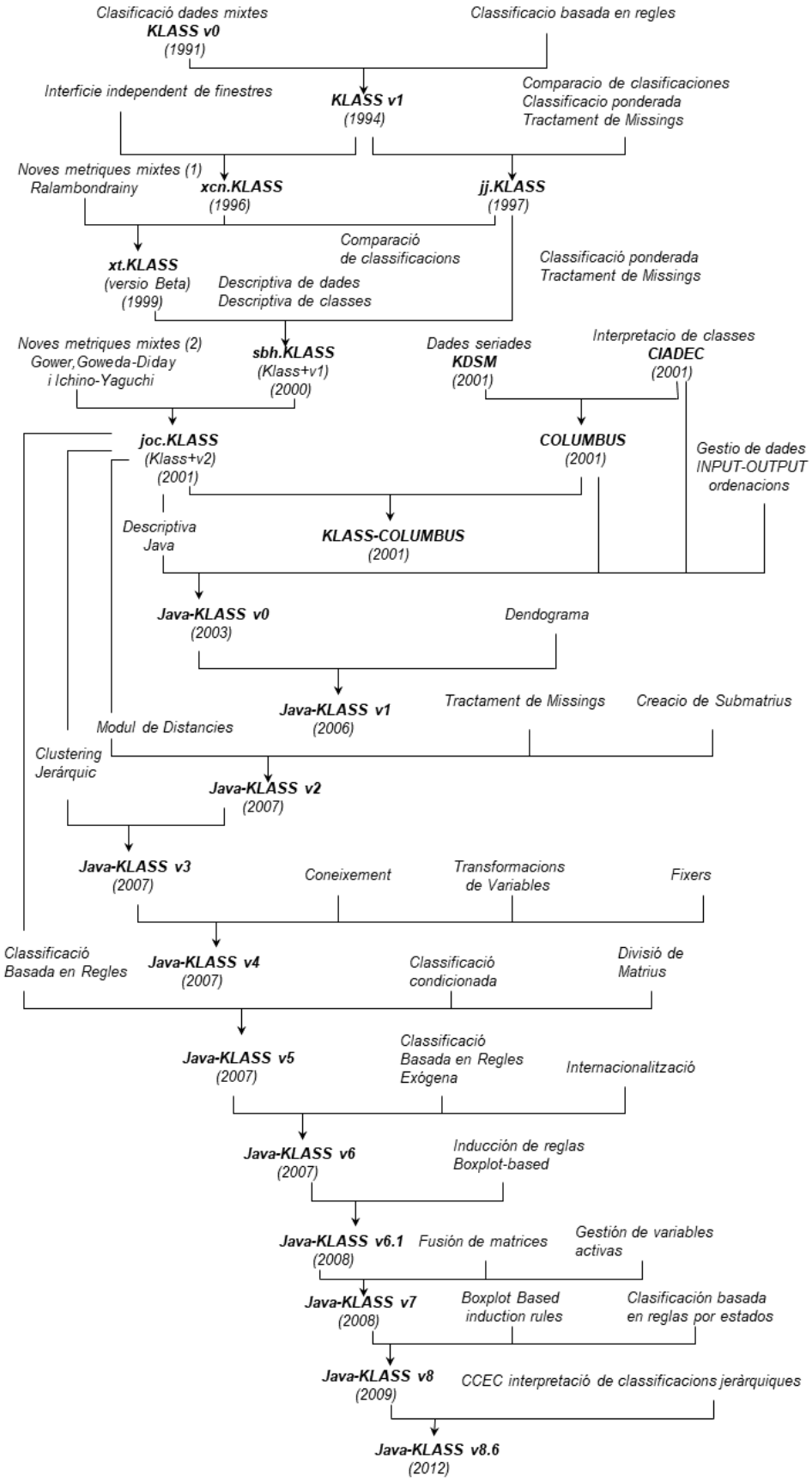

Figura 5: Cronologia de KLASS

4. Objectius de la tesi i contribucions

4.1. Objectiu general

L'objectiu principal d'aquesta tesi és:

Definir una metodologia ràpida de diagnòstic d'un domini (territorial o no) basat en tècniques de perfilat, que incorpori l'ús del clustering i el TLP basat en termòmetres com a peces clau d'una nova metodologia d'Intel·ligència Artificial explicable i orientada al suport a la presa de decisions complexes i estratègiques.

La tesi doctoral consta de 3 parts principals, la primera que es troba al capítol 3 de la tesi i es dona per conclosa. Aquesta inclou l'estat de l'art, realitzat gràcies a una intensa recerca bibliogràfica seguida del capítol 3 d'aquesta tesi, on es mostren les tècniques de mineria de dades emprades per al desenvolupament de la tesi. La 2a part consta en la presentació de la metodologia MIPRI2D, una conceptualització dels mètodes i innovacions aplicats en el projecte INSESS-COVID19, la gran aplicació d'aquesta tesi que es troba exposada en una tercera part de forma més amplia, on es detallen els resultats obtinguts de la recerca aplicada, juntament amb 4 altres aplicacions.

4.2. Objectius específics

Per a fer-ho cal

- Dissenyar l'esquema d'infraestructura tecnològica que suporta el procés
- Dissenyar els mecanismes de recollida de dades i qüestionaris associats

- Dissenyar els mètodes d'anàlisi automàtica de les dades i de quantificació de l'error estadístic amb tècniques de mineria de dades l'intel·ligència artificial.
- Dissenyar els mecanismes de preservació del secret estadístic associats a tot el procés multivariant.
- Dissenyar un mètode automàtic d'interpretació conceptual de perfils basat en l'automatització del ***TLP basat en termòmetres*** que porti els resultats de l'anàlisi de dades a conceptes directament comprensibles per part de l'usuari final.
- Formalització de la proposta metodològica complerta
- Implementació de la proposta en un sistema de suport al diagnòstic de territori basat en dades
- Generació automàtica de conceptualitzacions de perfils fàcils d'entendre per usuaris no experts en tecnologia.
- Aplicació de la proposta de tesi a diversos casos d'estudi reals de l'àmbit de la sostenibilitat
- Validació de la proposta metodològica i consolidació dels resultats de les aplicacions.

5. Proposta Metodològica

5.1. La metodologia MIPRI2D

Aquesta secció presenta una visió global de la metodologia proposada en la tesi, que hem anomenat *Metodologia Integral de Perfilat Ràpid, Intel·ligent i Interpretable de suport a la presa de Decisions complexes (MIPRI2D)* que reuneix una sèrie de passes, algunes de les quals han requerit recerca específica i la resolució de problemes oberts fins al moment.

MIPRI2D és una metodologia innovadora que permet assolir els objectius de tesi plantejats al cap 4.

La figura 6 presenta un esquema general de la proposta MIPRI2D. Els passos principals de la proposta es detallen a continuació. En les següents subseccions, es proporcionen detalls sobre cada passa tot destacant la part innovadora.

La metodologia MIPRI2D s'estructura en 5 Fases principals:

- Fase I: Anàlisi del fenomen i disseny de les eines d'observació
- Fase II: Tallers i adquisició de dades
- Fase III: Anàlisi Intel·ligent de dades
- Fase IV: Perfilat intel·ligent de les classes
- Fase V: Interpretació de resultats i elaboració de diagnòstic i recomanacions finals

Les fases de PÍPRID estan orientades a cobrir la cadena de valor sencera entre el disseny de l'experiència, potser col·laborativa, la recollida de dades i l'elaboració d'informes finals, de valor afegit. A continuació es detalla breument cadascuna d'aquestes fases per més endavant donar més detalls de la seva conceptualització

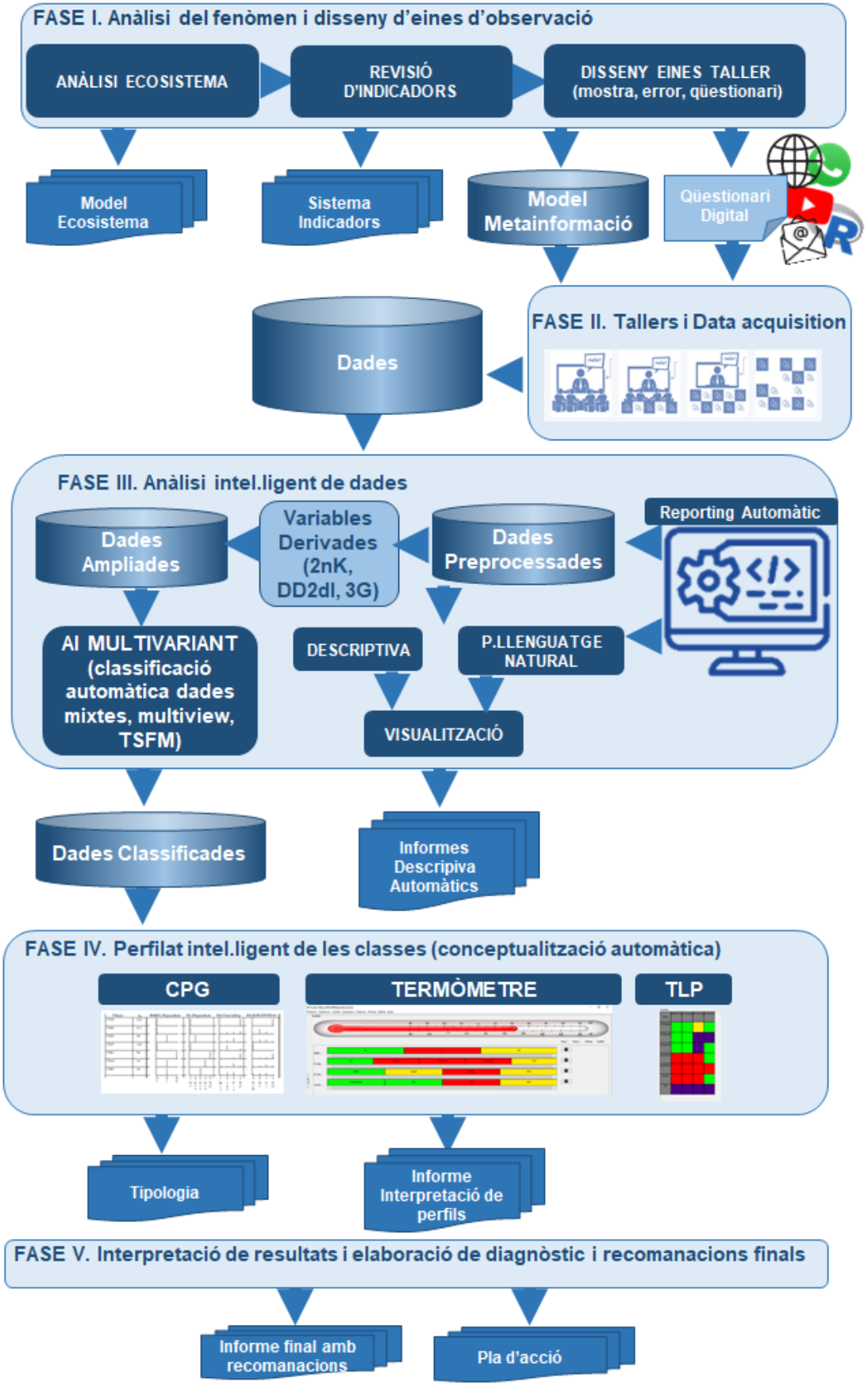

Figura 6: La metodologia MIPRI2D

FASE I. Anàlisi del fenomen i disseny de les eines d'observació El primer pas de la metodologia MIPRI2D consisteix en la preparació de totes les eines necessàries per a poder realitzar una consulta INSESS. Així, aquesta fase es basa en l'anàlisi del fenomen i disseny de les eines necessàries per a dur a terme la consulta i es desenvolupa en tres grans blocs que descrivim a continuació:

Abans de començar amb la part tècnica de la metodologia basada en l'obtenció i gestió de dades, i la seva anàlisi, el mètode proposat suggereix començar per la comprensió de l'estructura de l'ecosistema objectiu. A partir d'aquesta anàlisi, apareixerà una idea clara sobre el disseny de la mostra, d'una banda, i el tipus de preguntes requerides també pels participants. A més, la manera en què es recolliran les dades requereix atenció.

1. Anàlisi de l'ecosistema de destinació
2. Identificació de població en estudi
3. Revisió d'indicadors
4. Disseny d'eines per als tallers
    4.1. Augment de l'expressivitat dels qüestionaris. Variables més expressives. Com es veurà al llarg del document, la tesi incorpora l'ús de tipus de variables amb estructures complexes, amb més potència expressiva i que permeten poden analitzar fenòmens més complexos. Algunes d'elles expressen al llarg de diverses columnes o no a la base de dades. Per a fer front a aquesta situació és necessari desenvolupar alguns nous components metodològics que es detallen al llarg del document.
    4.2. Definició de la tipologia de variables
    4.3. Creació del model de metainformació
5. Construcció de l'instrument del qüestionari
6. Disseny de la infraestructura tecnològica
    6.1. Protocols d'alta, vistes, emmagatzematge de dades, tutorials, mails, plantilles de whatsapp....
    6.2. Disseny de la seguretat i preservació del secret estadístic (privadesa, reidentificació...)
7. Informació territorial

El que s'obté d'aquesta fase és:

a. El model d'ecosistema
b. El Sistema d'indicadors
c. El model de Metainformació
d. Qüestionari digital
e. Eines tecnològiques del taller: La Web, vídeos, tutorials, material pels tallers, codi RStudio i Scripts, repositori de dades, campanyes de captació de participants, etc.

FASE II Tallers i Data acquisition: La informació és recollida arreu mitjançant tallers; eventualment, i segons l'aplicació, en tot el territori. Els tallers permeten explicar l'objectiu de l'estudi i presentar el qüestionari, explicant l'orientació de les preguntes i es resolen els dubtes que puguin sorgir. La realització dels tallers és un tret fonamental d'aquesta metodologia, ja que evita biaixos per la mal interpretació de les preguntes per part dels participants. Els tallers poden ser de 4 modalitats que seran presentats a més endavant.

8. Criteris d'inclusió i exclusió
9. Determinació de la grandària de la mostra i disseny mostral
10. Tipologia de tallers

El que s'obté d'aquesta fase és:

f. La composició de la mostra d'observació
g. El resultat dels tallers (bases de dades, principalment)

FASE III Anàlisi intel·ligent de dades. Basada en l'ús de tècniques de ciència de les dades, reporting automàtic i intel·ligència artificial per a extreure valor estratègic de les dades, bàsicament a través d'eines visuals i estadístiques descriptives, models d'aprenentatge no supervisat, etc.

11. Preprocessament de dades.
    - 11.1. Tècniques bàsiques d'homogeneïtzació
    - 11.2. Imputació de dades mancants
    - 11.3. Recodificacions de preguntes
    - 11.4. Modalitats i supressió de duplicitats.
12. Anàlisi descriptiva i territorial (numèrica i gràfica i mapes estadístics)
    - 12.1. Eines gràfiques innovadores: Diagrama de teler
    - 12.2. Taula de freqüències ampliada
    - 12.3. Diagrama de barres, pastís o taula de freqüències marginals
    - 12.4. Taula de freqüències multivaluada
    - 12.5. Taula de freqüències de trajectôria
    - 12.6. Diagrama de barres múltiples
    - 12.7. Quadrícula de diagrames de pastís
    - 12.8. Taules de transició
    - 12.9. Diagrama de barres apilades múltiple
13. Estructura de l'anàlisi descriptiva per tipus de variables
    - 13.1. Variables numèriques
    - 13.2. Variables categòriques
    - 13.3. Variables multivaluades

13.4. Variables de quadrícula
13.5. Variables bàsiques temporals
13.6. Variables TQQ
13.7. Anàlisi de preguntes Obertes mitjançant PLN.
13.8. Estudis específics
13.9. Reporting automàtic

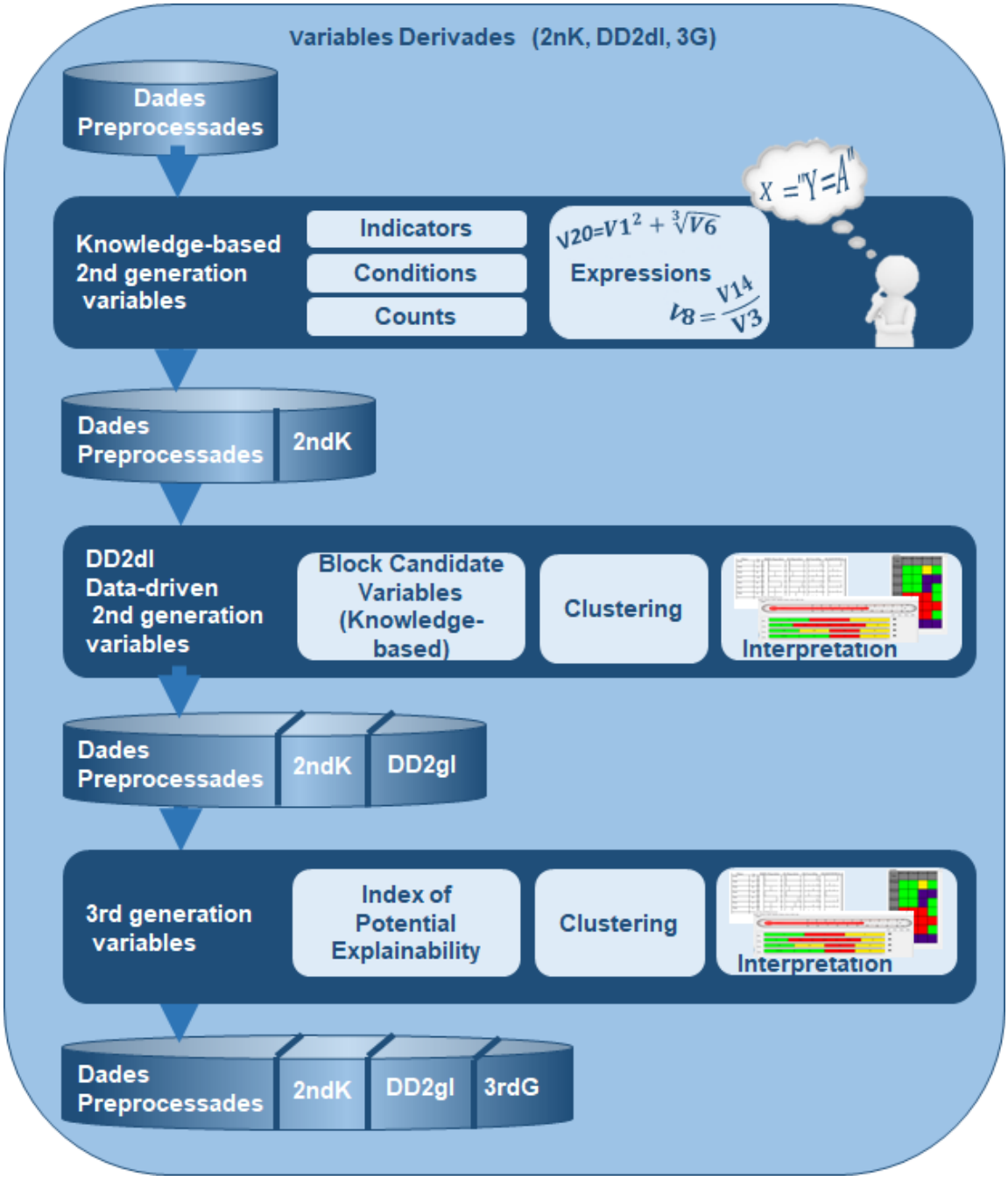

Figura 7:Variables derivades

14. Eines de suport a la interpretació de classes i noves variables
    14.1. TERMÒMETRE
    14.2. TLP BASAT EN TERMÒMETRE
15. Ampliació del preprocessament amb tècniques de creació de noves variables
    15.1. Variables de 2a generació basades en coneixement expert
    15.2. Variables de 2a generació basades en dades (DD2gI)
    15.3. Variables de 3a generació
16. Anàlisi Multivariant:
    16.1. Selecció de variables basades en criteris territorials (TFSM

D'aquesta fase s'obtenen:

h. Les dades preprocessades, ampliades i classificades
i. Informes d'anàlisi descriptiva automàtics

FASE IV Perfilat intel.ligent de les classes: El darrer pas de la fase anterior consta la realització d'un clustering multivista del qual s'obtenen grups per als quals cal fer una tasca d'interpretació i de conceptualització contextualitzada a l'àmbit concret d'aplicació. Les passes són:

17. Descripció i identificació els patrons de comportament de cada grup mitjançant eines de suport a la interpretació de classes
18. Representació territorial dels grups obtinguts

D'aquesta fase s'obtenen:

19. Dades classificades

j. Mapes estadístics
k. Informes d'interpretació de perfils automàtics
l. Tipologia

FASE V Interpretació dels resultats, elaboració del diagnòstic i recomanacions finals La darrera conclouen l'estudi contextualitzant els resultats i emetent diagnòstic i les recomanacions adients.

En aquesta darrera fase pren importància la realització de reunions amb les persones expertes i amb aquelles que tenen el poder de prendre decisions per tal de presentar-los les conclusions a les quals s'ha arribat.

D'aquesta fase es pot obtenir:

m. Un informe amb recomanacions finals
n. Un pla d'acció

5.2. FASE I Anàlisi del fenomen i disseny de les eines d'observació

5.2.1. Anàlisi de l'ecosistema diana

Com s'ha dit la primera passa abans de plantejar una consulta INSESS és analitzar l'ecosistema objecte d'estudi. La proposta consisteix a incloure els següents aspectes en aquesta anàlisi:

1. **Comprensió de l'estructura del domini de destinació.** Com s'organitza el domini, quines administracions públiques tenen competències en els diferents tipus de serveis, el conjunt de serveis disponibles, etc. El conjunt d'actors que hi intervenen, els rols i relacions entre ells, i el coneixement que tenen les persones que han de participar en el taller sobre el tema en qüestió, etc.
2. **Revisió de les fonts actuals d'estadístiques oficials sobre el domini** d'aplicació que es pot utilitzar com a referència per a l'anàlisi. La revisió literària és útil. No en el sentit acadèmic, sinó en la cerca d'informes oficials que descriguin el domini objectiu (en aquest cas, les estadístiques oficials i les enquestes seguides durant la pandèmia).

   Per al cas INSESS-COVID19, cal entendre l'estructura dels Serveis Socials a Catalunya. Pel cas de DIMCARE cal entendre quines dades hi ha disponibles sobre l'ús de les eines digitals avui en dia a les entitats del tercer sector i quin és el coneixement per part dels seus treballadors. Pel que fa a la colla castellera, en el moment de fer la redacció cal conèixer i entendre els termes castellers i les activitats que es desenvolupen en una colla castellera. En relació amb el cas del consum energètic d'una llar, al ser un qüestionari dirigit als infants és necessari conèixer quins són els electrodomèstics que hi pot haver a casa d'una família amb canalla, així com els coneixements pel que fa a tecnologia que poden tenir aquests infants per tal que puguin entendre les preguntes que es formularan al qüestionari.

Aquesta anàlisi, realitzada juntament amb els experts en matèria de dominis, donarà lloc a una identificació clara de:

- Tipus de participants requerit
- Tipus d'informació rellevant que es voldrà obtenir els participants
- Aportacions per a les decisions preses en els pròxims passos.

5.2.2. Identificació de població en estudi

Un cop s'ha analitzat l'estat de l'ecosistema diana, es coneixen els actors que intervenen en aquest i qui pot ser coneixedor de les dades que és necessari per a l'estudi. Un cop fet l'anàlisi en profunditat sobre la informació ens pot aportar cada part implicada, es definirà la població diana, és a dir, els perfils dels individus d'interès. Aquest fet permetrà a l'investigador

dissenyar els perfils dels individus d'interès, així com establir els criteris d'inclusió i exclusió dels associats.

Establerta la població diana, caldrà establir quin dels mètodes de mostreig s'estabiliran per a la selecció de la mostra i quins seran els mecanismes de recollida de dades associats per garantir la representativitat.

Per tal de poder realitzar de forma correcta aquesta fase, es duran a terme diferents reunions amb experts.

En el cas d'INSESS-COVID19, després d'una profunda comprensió de la llista de serveis socials disponibles oferts en el sistema d'atenció social primària, es va definir una llista de 20 perfils objectiu i els criteris d'inclusió corresponents juntament amb els professionals dels Serveis Socials, tant dels governs, ajuntaments com dels consells regionals (consells comarcals). Els perfils proposats assenyalen segments de població a priori que s'espera que siguin significativament danyats per la pandèmia. A L'apartat xxxxx es donen detalls d'aquest disseny.

En el cas de DIMCARE per exemple, les unitats d'estudi no són persones, sinó entitats del tercer sector de 3 regions europees participants al projecte. A L'apartat xjkfhldkjhdkfhgd es pot observar el context del projecte i la seva aplicació.

El cas de l'associació sense ànim de lucre la població diana són els associats i associades a l'entitat, ja que són totes les persones de les quals és necessari conèixer l'opinió.

5.2.3. Revisió d'indicadors

Aquesta part consisteix en la recerca d'informes existents en relació a la temàtica concreta a tractar. Per dur-lo a terme, cal cercar informes publicats que donin dades relatives als conceptes. Aquests informes poden contenir els resultats d'una enquesta o bé conclusions extretes a partir de l'anàlisi de bases de dades, ja formin part de l'estadística oficial o no. Hi ha temàtiques sobre les quals fer la recerca d'indicadors al respecte és poc factible degut a la manca de dades del sector, fet que justifica la realització d'una consulta INSESS per a la realització de la mateixa.

5.3. FASE I. Disseny d'eines per als tallers

5.3.1. Augment de l'expressivitat dels qüestionaris

Els fenòmens que la tecnologia INSESS vol englobar són d'un nivell de complexitat que no en té prou amb qüestionaris de preguntes de resposta simple i múltiple que obliguen a simplificar enormement el tipus de relacions que es poden obtenir de l'anàlisi de dades d'enquesta.

En aquesta tesi es fa un pas endavant en aquesta direcció i es defineixen estructures de dades noves, basades sobre blocs de columnes de la matriu de dades i que representen variables compostes de major complexitat a les utilitzades habitualment i que incorporen major potència expressiva, permetent de poder indagar sobre aspectes més complexos dels fenòmens en estudi.

Més endavant s'explicarà quina infraestructura tecnològica acompanya la introducció d'aquest nou tipus de variables i la metodologia per poder-les analitzar en tota la seva complexitat, i per poder-ho fer de forma automàtica.

5.3.2. Definició de la tipologia de variables

En aquesta secció s'introdueixen uns tipus nous de variables d'estructura complexa, que les introduïm perquè permeten augmentar la potència expressiva dels qüestionaris. I això ens permet estudiar fenòmens més complexos i extreure'n informació més rica. Quan parlem de variables complexes no ens referim a les que s'expressen en l'espai dels números imaginaris, sinó en aquelles que representen característiques de fenòmens d'estructura complicada, amb interaccions d'ordre superior entre les variables, amb relacions causa-efecte difícils de modelar..., A [Gibert & Angerri, 2021] ja es van introduir alguns tipus de variables que es recullen a continuació (vegeu taula 6):

Type	Form of question	DB structure	Example
Numerica	Integer field	One numerical column	Age
Likert or Ordinal	Simple choice response	One alphanumeric column	Pending processes
Nominal	Simple choice response	One alphanumeric column	Gender
Nominal Multivalued	Multiple choice response	One column with lists of values separated by ";"	Impacted area (by process)
OrdinalXtime	Simple choice grid	One alphanumeric column per timeStamp	Convivential Unit
OrdinalXordinalxTime	Several Packs of several Ordinal	Several Packs of several columns	Kinds in charge
Open question	Open textual window to be edited by the respondant	Textual column with complete text	Why didn't you receive the payments by July 10th?

Taula 3: Tipologia de variables de INSESS-COVID19

En les properes seccions es formalitzen i defineixen aquests nous tipus

5.3.3. Variables multivaluades

Una variable multivaluada és una variable qualitativa X amb S possibles modalitats D = {m_1, m_2,...,m_S} de manera que un individu pot prendre simultàniament diversos valors de D. Les variables multivaluades permeten tractar variables on l'individu té més d'una modalitat simultàniament. Seria el cas de símptomes de malalties, gènere de les pel.lícules (drama+ historic).

Aquest és el cas, per exemple, de la variable X="Qui genera violència", que pren valors:

D={1. No soc objecte de violència, 2. Un superior o ascendent (pare, tiet, responsable de feina, prof…) 3. Un igual (germà, company, amic, veí…), 4. Un subaltern o descendent (fills, empleats, …)}.

Els valors de X, $x_i \in P(D)$, de manera que un individu pugui patir violència simultàniament per un superior i un igual, o un igual i un subaltern. El concepte de variable multivaluada no és nou, tanmateix en aquesta tesi s'introdueixen eines descriptives específiques per extreure millor informació d'aquest tipus de variables com es podrà veure a la secció 5.12.3

R4. En cas que siguis objecte de violència, qui exerceix aquesta violència?
1. No soc objecte de violència
1. No soc objecte de violència
1. No soc objecte de violència, 2. Un superior o ascendent (pare, tiet, responsable de feina, professor…), 3. Un igual (germà, company, amic, veí…), 4. Un subaltern o descendent (fills, empleats, …), 5. No contesta

Figura 8: Visualització de les variables multivaluades a la base de dades original

5.3.4. Variable de quadrícula

Variable de graella (X, Q) és una variable qualitativa X amb S modalitats en D. Per a cada modalitat $m_s \in D$, es dona un valor (de Q) que indica la qualificació de m_s (Q és un conjunt Likert o ordinal de valors). Aquesta variable permet representar les variables qualitatives on les seves modalitats estan avaluades per una likert o similar

Aquest és el cas, per exemple, de X="Quines llengües parles", D={1. Català, 2. Castellà, 3. Anglès, 4. Francès, 5. Romanès, 6. Àrab, 7. Altres} i Q={Bé, Regular, amb dificultats, No la parlo, No contesta}

P5. Quines llengües parles? [1. Català]	P5. Quines llengües parles? [2. Castellà]	P5. Quines llengües parles? [3. Anglès]	P5. Quines llengües parles? [4. Francès]	P5. Quines llengües parles? [5. Romanès]	P5. Quines llengües parles? [6. Àrab]	P5. Quines llengües parles? [7. Altres]
1. Bé	2. Regular	3. Amb dificultats	1. Bé	2. Regular	3. Amb dificultats	3. Amb dificultats
No contesta	No la parlo	No la parlo	Amb dificultats	Amb dificultats	Regular	Regular
Bé	Bé	Bé	Amb dificultats	No la parlo	No la parlo	No la parlo
Bé	Bé	Bé	Regular	Bé	Regular	Bé
Regular	Regular	Amb dificultats	Amb dificultats	Bé	No la parlo	No la parlo

Figura 9: Visualització de les variables de quadricula a la base de dades original

5.3.5. Variables bàsiques temporals (Temporal Basic Variable TBV)

La variable temporal (X, T) defineix X com una variable qualitativa (nominal, ordinal, binària o Likert) amb les modalitats S en D que es mesura n_T vegades al llarg del temps, proporcionant columnes nT en el conjunt de dades com a rèpliques temporals de X. Podem denotar aquestes rèpliques com X_{t1}, X_{t2}, ..., X_T. Aquest tipus permet analitzar conjuntament sèries temporals d'una variable qualitativa en diferents moments del temps.

Aquest és el cas per exemple de X="Esquema de convivència",

amb D={Sol/a, MM-MP, Nucli, Reagrupades, Extensa, No-Fami i No Contesta} amb T=3. (Gener 2020, Juliol 2020 i Gener 2021)

F2. Marca el teu esquema de convivència en aquests tres moments. [Gener 2020]	F2. Marca el teu esquema de convivència en aquests tres moments. [Juliol 2020]	F2. Marca el teu esquema de convivència en aquests tres moments. [Gener 2021]
3. Unitat familiar de pare(s) i mare(s) i fills propis	2. Família monomarental o monoparental	2. Família monomarental o monoparental
No-Fami	Extensa	Reagrupades
Reagrupades	Reagrupades	Reagrupades

Figura 10: Visualització de les variables TBV a la base de dades original

5.3.6. Variable de quadrícula multivaluada

(X, Q) és una variable multivaluada qualitativa X amb modalitats S en D, replicada a cada valor de Q (Q també pot ser diverses marques de temps). Per a cada modalitat $m_s \in D$, i el valor de Q una columna binària m_{sq} indica la presència/absència de m_s en el grup q.

Aquest és el cas de X="Ajuts demanats" amb D={1. Ajuts per alimentació, 2. Ajut per tenir cura dels infants, 3. Accés a material informàtic, 4. Serveis de repartiment a domicili de productes de primera necessitat, 5. Ajuts per pagar el lloguer de la vivenda, 7. Ajuts per pagament de taxes i tributs, 8. Renda Mínima d'inserció, 9. Ingrés Mínim Vital, 10. Serveis d'acollida, 11. Assistència psicològica, 12. Teleassistència, 13. Iniciar un procés d'ERTO al meu negoci, 14. Altres} i Q={No ho he necessitat, Ajuntament i/o Consell Comarcal, Generalitat de Catalunya, Gobierno de Espanya, Entitats d'Acció Social, Altres, No contesta}

E2. Has necessitat acollir-te a algun dels ajut o prestació incloent els especials que s'han posat en marxa per mitigar la problemàtica per la COVID-19? [2. Ajut per tenir cura dels infants]	E2. Has necessitat acollir-te a algun dels ajut o prestació incloent els especials que s'han posat en marxa per mitigar la problemàtica per la COVID-19? [3. Accés a material informàtic]	acollir-te a algun dels ajut o prestació incloent els especials que s'han posat en marxa per mitigar la problemàtica per la COVID-19? [4. Serveis de repartiment a domicili de productes de primera necessitat]	E2. Has necessitat acollir-te a algun dels ajut o prestació incloent els especials que s'han posat en marxa per mitigar la problemàtica per la COVID-19? [5. Ajuts per pagar lloguer de la vivenda]	acollir-te a algun dels ajut o prestació incloent els especials que s'han posat en marxa per mitigar la problemàtica per la COVID-19? [6. Ajuts per pagar subministraments basics (llum, aigua, gas....)]	E2. Has necessitat acollir-te a algun dels ajut o prestació incloent els especials que s'han posat en marxa per mitigar la problemàtica per la COVID-19? [7. Ajuts per pagament de taxes i tributs]
Entitats d'Acció Social	Gobierno de España	Generalitat de Catalun	Entitats d'Acció Social	Entitats d'Acció Social	Gobierno de España
Gobierno de España	Generalitat de Catalun	Ajuntament i/o Consell	Generalitat de Catalun	Gobierno de España	Gobierno de España
No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat
Generalitat de Catalun	Ajuntament i/o Consell	Ajuntament i/o Consell	Ajuntament i/o Consell	Generalitat de Catalun	Generalitat de Catalu
No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat
Generalitat de Catalun	Ajuntament i/o Consell	Gobierno de España	Entitats d'Acció Social	Altres	Ajuntament i/o Conse
No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat
No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat
No contesta	No contesta	No contesta	No contesta	No contesta	No contesta
No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat	No ho he necessitat

Figura 11: Visualització de les variables de quadricula multivaluades a la base de dades original

5.3.7. Variable TQQ: Temporal Qualified Qualitative

(X, T, Q) és una variable qualitativa X amb modalitats S en D, replicat n_T vegades. Per a cada modalitat m_s, un valor (de Q) indica la qualificació de m_s (Q és Likert o conjunt ordinal).

Com a exemple, la variable X= participació en la societat, està prenent quatre modalitats

D={Associacions, Xarxes, Volontarietat, Altres}, el que indica el tipus d'accions participatives que la persona segueix.

La variable es replica T=3 vegades (t_1 = gener2020, t_2 = juliol 2020, t_3 = gener 2021).

Cada X_t és, al seu torn, un conjunt de Likerts, tal que per a cada modalitat de X en t1 tenim un Likert qualificant el grau de participació de la persona en ell. Q={Molt (molt, Una mica (alguns), Gens (cap), NC (resposta perduda). Al nivell de representació interna, cada variable TQQ proporciona columnes de card(D)xT Likert cadascuna prenent valors en Q que requereixen una descripció conjunta.

Soc1. La teva participació en activitats de la comunitat al gener del 2020 era: [1. Associacions culturals o esportives]	Soc1. La teva participació en activitats de la comunitat al gener del 2020 era: [2. Xarxes de veins, AFAs, etc...]	Soc1. La teva participació en activitats de la comunitat al gener del 2020 era: [3. Voluntariat]	Soc1. La teva participació en activitats de la comunitat al gener del 2020 era: [4. Altres]
Lleugerament implicat	Lleugerament implicat	Lleugerament implicat	Gens implicat
Lleugerament implicat	Lleugerament implicat	Lleugerament implicat	Lleugerament implicat
Gens implicat	Gens implicat	Molt implicat	Lleugerament implicat
Lleugerament implicat	Gens implicat	Lleugerament implicat	No contesta

Figura 12: Visualització de les variables TQQ a la base de dades original

5.4. FASE 1. Creació del Model de metainformació

Continuem a la Fase 1 de MIPRI2D, però en un ordre de coses ben diferent. Per crear un bon fitxer de metadades, és molt important modelar i formalitzar adequadament els diversos tipus de variables presents a les dades.

Una vegada que s'han definit els diferents tipus de variables, i les eines estadístiques i gràfiques per analitzar cada tipus de variable són clares, es requereix un mecanisme per proporcionar intel·ligència als scripts que realitzen l'anàlisi descriptiva. Això es basa en la declaració de la variable i la implementació està dissenyada sobre la base d'un fitxer de metainformació que proporciona tota la informació conceptual requerida al sistema R per executar una anàlisi descriptiva adequada, capaç d'utilitzar procediments descriptius predefinits per a cada tipus de variable. El fitxer de metainformació ha de contenir tota la informació contextual de les dades. La proposta Out és utilitzar un fitxer de metainformació en forma de taula (implementable com a fitxer csv per exemple) amb l'estructura següent: Les files estan associades a variables. Alguns variables proporcionen metainformació a través de diverses files.

5.4.1. Model de metainformació

Per fer possible l'anàlisi automàtica de les dades, el model de metainformació conceptual ha estat dissenyat amb tot el coneixement necessari per generar l'informe automàtic i tenir la base de dades original en una estructura separada. El Model de metadades (MdM) és una matriu de dades que conté tantes files com sigui necessari, amb un mínim d'1 fila per pregunta per a preguntes numèriques i obertes. Per a variables qualitatives amb les modalitats S, el nombre de files és 1+S. Per a les variables TOT i quadrícula el nombre de files és: 1+Max(Q, S). Aquestes són les 16 columnes de la taula. Algunes de les files en Mm descriuen variables. Altres descriuen les modalitats de les variables qualitatives

- k (Col): Nombre de columna on la variable (X_k) es troba en el conjunt de dades original (I és un conjunt d'individus $\mathcal{I}= \{i_1 \dots i_n\}$ descrits per les variables K $\{X_1,X_2,...,X_k,...X_K\}$). Aquesta columna no es compleix en files on es descriu la modalitat d'una pregunta categòrica. Per a les variables TQQ k és la posició de les columnes en el conjunt de dades que conté les modalitats.
- $\tilde{k}$ (Eticol): Aquest paràmetre només s'omple en les files MdM referint les modalitats. Per a aquestes files $\tilde{k} = k$ Indica la posició de la variable qualitativa X_k.
- B (Bloc): En cas que el qüestionari estigui dissenyat per blocs temàtics, s'especifica el número de bloc. Cada bloc amb diverses variables del mateix tema.
- ℓ_B: Especifica el nom del bloc
- ℓ^*_B: Etiqueta curta per al bloc que s'utilitzarà a l'informe

- $\mathbb{Q}$(Pregunta): Text complet de la pregunta tal com apareix al qüestionari original. Anomenem r_k la fila del MdM on es descriu.
- A (respostes): Si X_k és qualitatiu, les files r∈ $\{r_k + 1, \ldots, r_k + S\}$ s'utilitzen per descriure les D_k les modalitats de X_k
- J: El valor per defecte és «inexistent». Només s'utilitza per a variables quadrícula o TQQ. Els valors de Q corresponents es descriuen en files r∈ $\{r_k + 1, \ldots, r_k + S\}$.
- A* (Rephrasing): Una expressió curta per a les preguntes i possibles respostes que s'utilitzaran en les taules estadístiques i gràfiques, ja que els textos llargs se superposaran i faran difícil la lectura.
- J* (Colcurt): Etiqueta curta per als valors Q en les variables TQQ i quadrícula. En cas que la pregunta no sigui un TQQ, cap de les dues xarxes no es compleix.
- O (Objecte): Indica el tipus d'informació representada a la fila: «pregunta», «modalitat», «nom del bloc» o separador (usat entre les variables)
- τ (Tipus de variable): Dona el tipus de variable representada en Q (només en files amb O ≠missing): «Numerical»; «Simple response» per a Likerts o qualitatius; «Multivalued variable» per a Multivalued (Nominal, Ordinal o Likert); «grid» per a TQQ, Temporal basic and Grid Variables;; «Open» per a respostes textuals;.
- τ_D (tipus descriptiu): procediment descriptiu associat (alguns utilitzats per a més d'un tipus de variable). Cada procediment descriptiu utilitza una combinació específica d'eines descriptives numèriques i gràfiques, algunes específicament dissenyades en aquest treball. τ_D només s'especifica quan es defineix τ.

Tipus	Tipus Descriptiu	Tipus	Tipus Descriptiu
Numerica	Numerical	TBV	NomxT, LikertxT
Resposta simple	Nominal, Ordinal, Likert	Multivalued Grid	NomMVxNom + TQQ
Variable Multivaluada	NomMV OrdMV, LikertMV	TQQ	NomxOrdxT
Quadricula	NomxLikert, OrdxLikert + TBV	Open	NomxOrdxT

Taula 4: Tipus de Kit descriptiu

- k_0 (PackIni): Per a la fila r on τ_r ∈{“Resposta múltiple”, ”Quadricula”, “TBV”, “Grid multivalorada”, ”TQQ”}, $k_0 = k_{r-1_f} + 1$ indica la columna de la base de dades on la informació sobre aquesta variable $\tilde{k}_r$ comença(bàsicament una columna després de la posició anterior final de la variable).
- k_f (PackFi): Per a la fila r on τ_r ∈{“Resposta múltiple”, ”Grid”, “TBV”, “Grid multivalorada”, ”TQQ”}, k.findica l'última columna de la base de dades que conté informació sobre la variable $\tilde{k}_r$ Vegeu Taula 5 per als valors de k_f per a diferents tipus de variables

- n_C (nCols): És el nombre de possibles modalitats a les cel·les de l'estructura de dades associada amb la variable, de vegades corresponen al nombre de modalitats de Q o T o X depenent del tipus de variables. Només disponible per a variables amb τ_r ∈{"Resposta múltiple", "Grid", "TBV", "Grid multivalorada", "TQQ"}.

Type	k_f	n_C
Variable multivaluada	$k_f = k_0 + S - 1$	2
Quadrícula	$k_f = k_0 + S - 1$	n_Q
TBV	$k_f = k_0 + n_T - 1$	S
Quadrícula multivaluada	$k_f = k_0 + S \cdot n_q - 1$	2
TQQ	$k_f = k_0 + S \cdot n_T - 1$	n_Q

Taula 5: Càlcul de k_f

- Reference: En cas que aquesta pregunta s'inspiri en una enquesta de referència (vegeu Taula 1), s'espera aquí trobar el enllaç a l'enquesta mare, per poder validar.
- Visualització: Per a la interpretació de patrons, dividim les variables en diferents grups, depenent del tema principal, i podem indicar quin kit de vistualització li correspon.

Aquest model permet processar qualsevol tipus de qüestionari, tingui la configuració de variables que tingui i, per tant, obre la porta a processar en cada moment informació provinent de diferents qüestionaris i, per tant, podrà adaptar-se a qualsevol població diana també.

5.5. FASE 1. Construcció de l'instrument del qüestionari

Un cop estudiat el marc conceptual i identificada la població d'estudi, es procedeix conceptualitzar els aspectes que es volen analitzar i es decideix el model de referència a seguir. Per cada aspecte es deriven les preguntes, focalitzades en els àmbits a estudiar, però dirigides a fer aflorar aspectes relacionats amb l'objecte de l'estudi. Un cop decidides les qüestions, i identificat de quina tipologia són, segons la taula presentada a l'apartat 4.3, es procedeix a redactar—les garantint que no admetin interpretacions variades ni ambigües i es defineixen les ordenacions dels grans blocs del qüestionari, així com el de les preguntes dins cada bloc.

Al cas INSESS-COVID19, del marc conceptual se'n va derivar una conceptualització de les àrees de la vida que es volen estudiar inspirades en el model de referència SSM.cat, referit a l'estat de l'art. Així, el qüestionari INSESS-COVID19 va focalitzar preguntes sobre els àmbits de SSM.cat, però dirigides a fer aflorar, no només la vulnerabilitat social, sinó també l'impacte de la COVID-19 en aquesta vulnerabilitat. El resultat és un qüestionari amb 21 preguntes que generen fins a 190 ítems interns, d'estructures variades, segons el tipus de preguntes. Des del punt de vista de la resposta, respostes numèriques o categòriques, algunes de resposta múltiple, i algunes quadrícules.

En general, l'estructura del qüestionari que volem contemplar en aquesta recerca és la d'aglutinar variables prou expressives sota blocs temàtics, que recullin informació sobre aspectes específics i permeti també paquets de preguntes específics d'un subgrup d'interès quan sigui necessari. La relació entre blocs de està esquematitzat a la Figura 13:

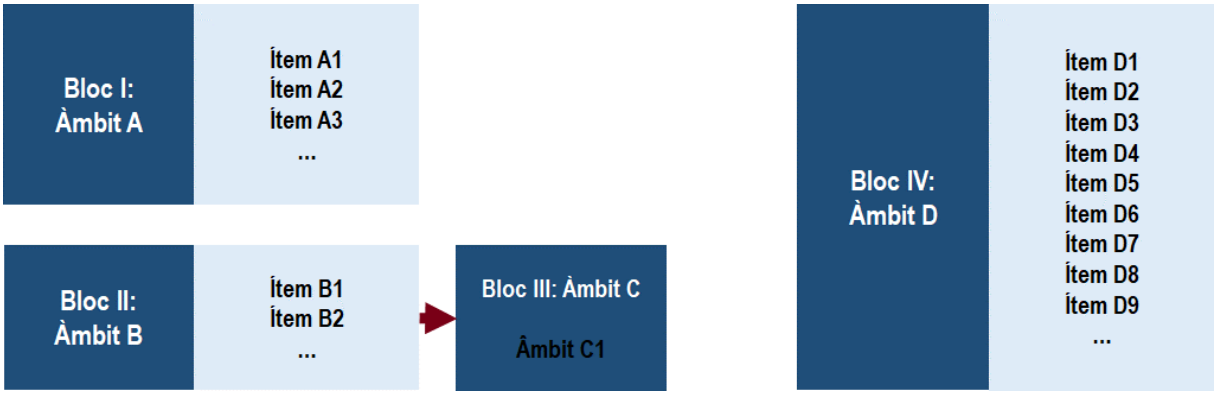

Figura 13: Disseny d'un qüestonari

5.5.1. Qüestionaris atemporals. Robustesa respecte del moment de la mostra

Un dels objectius del qüestionari era guanyar tota la robustesa possible respecte del moment que la persona respon. Per això es planteja el qüestionari com una reflexió en diversos moments ***FIXES*** del temps i es demana als respondents que responguin (el dia que sigui) des d'aquests moments establerts. Això permet, per exemple, que dades recollides al setembre es puguin tractar conjuntament amb les de juliol perquè les preguntes refereixen a gener, a juliol i a la previsió per desembre per exemple.

L'instrument INSESS introdueix, doncs, una estructura innovadora en el qüestionari, que permet un llarg període de recollida de dades preservant la comparabilitat de les dades recollides. Aquesta és una característica molt rellevant del qüestionari doncs un extens termini en la recollida de dades no introdueix cap mena de biaix en relació al moment que el participant participa en el taller, sense poder analitzar totes les dades de manera conjunta, sense haver de condicionar al moment de la resposta al qüestionari. Això proporciona un important avantatge davant de petites mostres, ja que proporcionar un període més llarg per a la recollida de dades la grandària de la mostra pot augmentar sense límit, sense alterar la la validesa de les dades recollides anteriorment.

Per tal de garantir el que s'ha comentat, la proposta de la metodologia presentada en aquesta tesi és que totes les preguntes del qüestionari es divideixin en dues categories:

- Estàtica: Característica que manté estàtica durant tot el període d'estudi (edat, sexe, lloc de naixement, etc.)

- Dinàmica: Característica que pot canviar al llarg del període d'estudi

La proposta és sol·licitar respostes en uns moments determinats al llarg del temps per a totes les preguntes dinàmiques del qüestionari

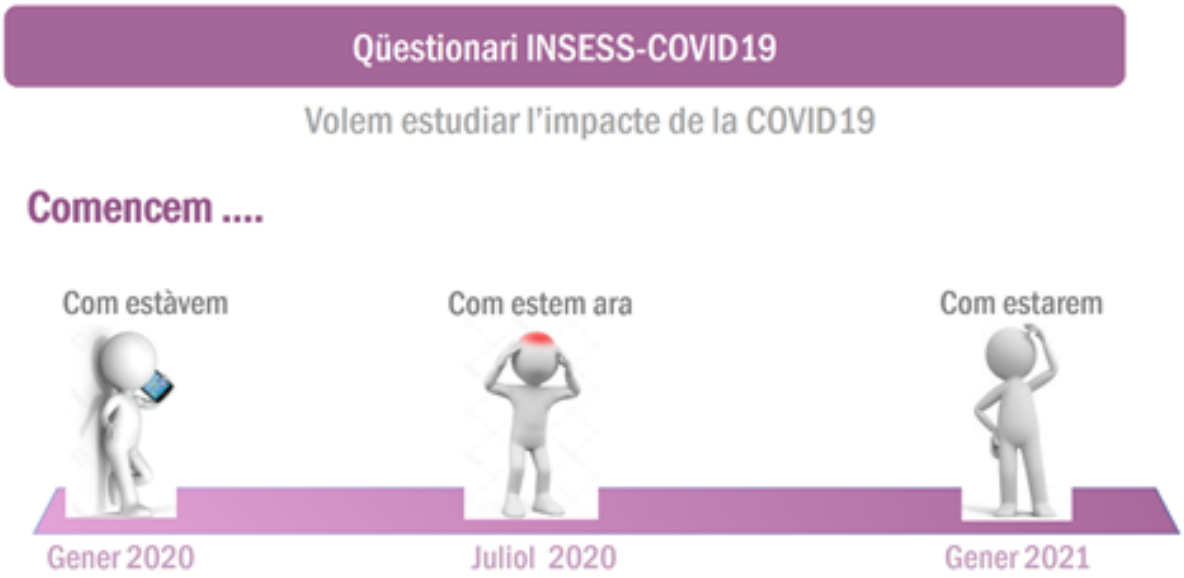

Figura 14: Les tres marques fices de temps del qüestionari INSESS-COVID19

La introducció d'aquest disseny en el qüestionari té la propietat que l'estudi guanya robustesa respecte a la data específica en què el ciutadà participa en el projecte. El qüestionari es pregunta sobre situacions/percepcions en aquests tres punts de temps fixos, de manera que el procés de recollida de dades es pot realitzar sempre que sigui necessari i les dades encara permeten l'anàlisi de la dinàmica del fenomen. Les respostes de les persones que participen al llarg del període de recollida de dades, sense una data concreta per respondre al qüestionari, encara proporcionen informació sobre la situació de la persona en uns punts d'observació especificats a l'estudi (per exemple, en el cas de INSESS-COVID19 al gener de 2020, juliol de 2020 i gener de 2021), de manera que les dades de tots els respondents es poden analitzar conjuntament, independentment del moment que s'ha entregat la resposta. Aquesta solució permet superar les limitacions de recollir les dades totes de cop en un període curt de temps i per exemple en el cas de INSESS-COVID19 va ser de gran utilitat perquè la consulta es va llençar en plena desescalada de la pandèmia i encara hi havia moltes ABSS tancades i les que no, estaven desbordades i no es podien entretenir a buscar ciutadans que participessin a l'estudi.

Aquest disseny permet avaluar l'impacte del qual s'està estudiant entre les dates que es pregunten al qüestionari. Això permet tractar el qüestionari com una anàlisi pre-post i poder estudiar les dinàmiques generades entre els diferents moments del temps, així com copsar la

percepció que els respondents tenen sobre cada moment del temps, sigui present, passat o futur. Si alguna de les dates és futura, llavors es fa possible desenvolupar estudis prospectius

La conseqüència és que aquest disseny introdueix paquets de variables al qüestionari, que ja no són independents, i es requeriran procediments específics per analitzar-les de la manera correcta. Aquests s'introdueixen més tard en aquesta tesi doctoral.

5.6. FASE 1: Disseny de la infraestructura tecnològica

Per tal que els individus puguin omplir l'enquesta fàcilment cal desenvolupar una infraestructura tecnològica adequada. A més, també és necessari desenvolupar les eines que permeten analitzar d'una forma molt ràpida els resultats obtinguts a les enquestes i preservin l'anonimat dels respondents.

La metodologia inclou (la Figura 15 mostra el resum):

- El disseny de la infraestructura tecnològica:
    - Un servidor al núvol que compleix amb tot el RGPD acull el qüestionari digital la UPC garanteix els nivells de seguretat del servidor que allotjarà les dades.
    - Un web d'accés a documentació i qüestionari
- Protocols d'emmagatzematge de dades al núvol: les dades recollides no inclouran dades personals. Si n'hi ha alguna (per evitar respondents que repeteixin el qüestionari per exemple), es xifrarà amb clau irreversible al mòbil abans de viatjar cap a servidor.
- Els protocols d'accés al web: que garanteixin l'accés fàcil i segur al qüestionari als ciutadans que participaran en l'estudi
    - La web ha de mantenir dues vistes separades
        - Una per les entitats que organitzen els tallers de la consulta taller i que conté documents de suport perquè faciliten l'organització dels tallers.
        - Una altra pels ciutadans participants amb tota la informació per saber com respondre el qüestionari.
        - Protocols d'alta de participants: via procés d'autenticació en el mateix web (amb contrasenyes diferents i genèrics per cada tipus d'usuari).
        - El web es podrà accedir utilitzant telèfon mòbil, tauleta o PC (ordinador personal).
- Un servei de mail, i fer-lo accessible des del web
- Implementar el qüestionari en un formulari digital per recollir les dades del qüestionari
- Tutorials: Cal desenvolupar la documentació que permetrà als usuaris interactuar amb l'eina de forma indefinida.

- El procés per generar els informes automàtics definitius basats en les dades recollides en aquests qüestionaris. Després de posar en marxa els primers tallers, tota l'artilleria d'anàlisi automàtica i intel·ligent de dades es va anar desenvolupant a l'espera que es tanqués la recollida de dades amb tot a punt per analitzar-les de forma automàtica i en poques hores. Bàsicament, es va combinar l'ús del paquet estadístic R amb l'ús del software KLASS (en la seva versió java). Les dades recollides al qüestionari es descarreguen (periòdicament) per ser processades automàticament a través de scripts R i KLASS [Gibert et al. 2015] i un informe de treball ben editat es genera automàticament amb Word, on els resultats de l'anàlisi es mostren i es formategen com un document final.
- Els protocols de contacte amb els respondents: En el cas de INSESS-COVID19, per exemple es va decidir que serien les ABSS de SS qui contactarien directament amb les persones que complien els diferents perfils per convidar-los a participar de manera que els analistes de dades ja no poden saber la identitat dels participants.

Després de la implementació i el desplegament, es va dur a terme la validació tècnica dels scripts, del comportament del servidor, la funcionalitat de web i disponibilitat de tots els materials necessaris.

El projecte s'ha desenvolupat amb 3 portàtils i un servidor per centralitzar la recollida de dades, permanentment disponible des del 15 de juny de 2020.

Formular els diferents esquemes de confidencialitat que hem fet servir en els diferents casos en teòric i després a cada cas ho repetim.

Figura 15: Infraestructura tecnologica del Sistema INSESS-COVID19

5.7. FASE 1. Disseny de la seguretat i preservació del secret estadístic

Un dels fonaments de la professió d'estadístic, correspon a la preservació del secret estadístic. És necessari prendre decisions envers el moment necessari de donar una informació o no per tal de mantenir l'anonimat de les respostes enfront del principi de donar la màxima informació possible.

En general en qualsevol base de dades de qüestionari, té ple sentit creuar un nombre gran de variables i estudiar subpoblacions que responguin a perfils molt específics. En el cas d'INSESS-COVID19, els 20 perfils de ciutadans objectiu del projecte s'enfoquen a algunes subpoblacions que representen minories presumptament afectades pel COVID-19. El procés de recollida de dades s'ha distribuït pel territori per tal de minimitzar els esforços requerits als professionals d'ABSS, ja col·lapsats per la gestió dels casos afectats per la pandèmia. Algunes de les ABSS proporcionaven més dels 20 ciutadans necessaris, però moltes només n'aportaven 20 o de vegades menys. Això significa que per a alguns perfils, una ABSS pot proporcionar un o dos individus només. Això planteja greus limitacions per publicar estadístiques clàssiques a nivell ABSS, ja que seria fàcil per als professionals de l'ABSS identificar la persona concreta i que es violés el secret estadístic. Aquest fenomen no només es produeix quan es presenten dades a nivell d'ABSS, sinó fins i tot quan s'estudien perfils minoritaris a nivell de tot Catalunya, creuant amb altres variables que poden revelar informació suficient per identificar les persones.

INSESS-COVID19 proposa i aplica algunes bones pràctiques que preserven el secret estadístic fins i tot davant de subpoblacions molt petites.

S'han tingut en compte totes les dades per al càlcul d'estadístiques globals

S'han amagat de l'informe públic totes les modalitats de variables qualitatives amb un nombre massa reduït de respostes (només s'han publicat aquelles amb un mínim de 10 respostes). Les modalitats amb algunes respostes, però que no són suficients per ser públiques es detallen a l'informe. Per tant, es pot saber que menys de deu persones han estat comptabilitzades a l'estudi per a aquestes modalitats, però no hi ha un nombre exacte.

Els perfils objectiu amb menys de 3 participants només apareixen a la mostra com a perfils actuals, però sense el nombre exacte d'enquestats. Això és particularment important quan es presenten els resultats a nivell de ABSS.

Els perfils de destinació no s'exclouen mútuament. Així, molts dels ciutadans que participen en l'estudi compleixen simultàniament diversos perfils: per exemple, dones monoparentals que també treballen en l'àmbit dels serveis essencials, o homes amb salaris molt baixos i en situació d'infrahabitatge, etc. és possible reduir el llindar publicable fins a 3, ja que no es pot saber si les persones d'aquest "perfil ocult" només tenen aquesta característica o algunes altres i es manté la identificació de la persona.

Es planteja a partir d'aquí estudiar mètodes que determinin els llindars de preservació (3, 10) a partir de les variables que es creuin.

5.7.1. Privadesa

Moltes de les preguntes contingudes en el qüestionari INSESS-COVID19 són sensibles (ser objecte de violència, estar en situació irregular al país, patir desordres mentals, etc.). Garantir la privacitat i l'anonimat dels enquestats és crucial per assegurar-los que poden respondre a totes les preguntes sense tenir por.

Aquesta és la raó per la qual el qüestionari és autònom i anònim, de manera que no es pot identificar al demandat i les seves respostes no es poden creuar amb cap altra base de dades a nivell individual. En particular, no es poden creuar amb els sistemes d'informació dels serveis socials. De manera que no podem esperar obtenir informació addicional sobre la persona del qüestionari. Algunes preguntes requereixen informació que els serveis socials ja tenen sobre les persones, però preferim preguntar-ho de nou i evitar sentiments de desconfiança que podrien limitar les respostes proporcionades pels enquestats.

Per garantir aquesta seguretat, els professionals de l'ABSS identifiquen les persones que participen en els tallers, però no comparteixen amb l'equip INSESS-COVID19 les seves identitats. Envien als participants els enllaços i contrasenyes per entrar al lloc web del projecte i al qüestionari, però utilitzant una contrasenya comuna el sistema no pot rastrejar les identitats dels enquestats, de manera que les respostes mantinguin l'anonimat i la seguretat. El servidor que allotja la base de dades del qüestionari també compleix el RGPD, i l'equip INSESS-COVID19 conserva les microdades sense compartir amb cap altra institució que no sigui les dades agregades.

No obstant això, totes aquestes bones pràctiques no són suficients per a garantir el secret estadístic dels enquestats.

5.7.2. Risc de reidentificació

Els perfils dels ciutadans que dirigeix el projecte INSESS-COVID19 se centren en algunes subpoblacions que representen minories presumptament afectades per la COVID. El procés de recollida de dades s'ha distribuït al llarg del territori per minimitzar els esforços necessaris per als professionals de l'ABSS, ja col·lapsats per la gestió dels casos afectats per la pandèmia. Alguns de les ABSS estaven proporcionant més que els 20 ciutadans requisats, però alguns d'ells van proporcionar al voltant de 20 o de vegades menys. Això significa que per a alguns perfils, un baix pot proporcionar una o dues persones soles. Això planteja serioses limitacions per a la publicació d'estadístiques descriptives clàssiques a nivell de baix, ja que seria fàcil per als professionals de baix revelar el secret estadístic identificant la persona. Aquest fenomen es produeix no només quan es presenten dades a nivell de baix, sinó fins i tot quan s'estudien

perfils minoritaris a nivell català, creuats amb altres variables que poden revelar informació suficient per identificar les persones.

La pràctica clàssica de no publicar resultats sobre subpoblacions massa petites no és una solució en el context d'aquest projecte, ja que les minories vulnerables (encara que no és estadísticament significatiu) requereixen atenció i no poden desaparèixer de la imatge (pensem en les dones víctimes de la violència domèstica, mai són massa, bus no és una raó per amagar en l'anàlisi el que succeeix amb aquest segment de població, oi?)

INSESS-COVID19 proposa i aplica algunes bones pràctiques que preserven el secret estadístic fins i tot davant de subpoblacions molt petites.

S'han tingut en compte totes les dades per al càlcul de les estadístiques globals.

Totes les modalitats de variables qualitatives amb un nombre massa petit de respostes s'han amagat de l'informe públic (només s'han publicat aquelles amb un mínim de 10 respostes). En l'informe s'enumeren les modalitats amb algunes respostes, però no suficients per a ser públiques. Per tant, es pot saber que menys de 10 persones han estat comptabilitzades en l'estudi per a aquestes modalitats, però el número exacte no està disponible.

Els perfils de destinació amb menys de tres participants només es llisten com a perfils actuals a la mostra, però sense el nombre exacte d'enquestats. Això és especialment important quan els resultats es comuniquen a nivell de ABSS.

Els perfils objectius no són mútuament excloents. Per tant, molts dels ciutadans que copien en l'estudi es troben simultàniament amb diversos perfils: per exemple, dones monoparentals que també treballen en l'àmbit dels serveis essencials, o homes amb salaris molt baixos i en una situació d'infrahabitatge, etc. Això fa possible reduir el llindar de publicable fins a tres, ja que no es pot saber si les persones en aquest "perfil amagat" només tenen aquesta característica o algunes altres i la identificació de la persona manté preservada.

5.8. FASE 1. Informació territorial

Com és habitual quan les dades es recullen en un territori, un mapa que visualitza la informació estadística és molt rellevant. A INSESS-COVID19, quatre nivells territorials eren adequats: Ciutats i pobles, BASS, Vegueries, i Províncies. Els 947 municipis catalans s'agrupen en un primer nivell administratiu a 42 Comarques. Cada comarca és un baix que gestiona tots els municipis de la comarca amb menys de 20.000 habitants. Els municipis amb més de 20.000 habitants també són un baix. Per tant, Catalunya té 107 ABSS al territori. Vegueries és un grup intermedi de comarques. Catalunya té vuit Vegueries i quatre províncies. La província és massa gran per ser considerada en l'estudi INSESS-COVID19, ja que l'heterogeneïtat dins d'una sola província és massa alta des del costat de la vulnerabilitat social. Per tant, BASS i Vegueries són els dos nivells territorials considerats per a la representació geogràfica.

Val la pena esmentar que les variables qualitatives no poden representar-se en els mapes en el seu conjunt, però cal seleccionar algunes modalitats específiques i les seves proporcions territorials representen una per una.

5.9. FASE II Tallers i Data acquisition

La proposta de tesi planteja tallers cara a cara amb la ciutadania i part d'ells està orientat a crear un espai comú on un grup de participants ompli el qüestionari de forma síncrona. Els principals avantatges d'aquest disseny són:

- Es redueix les limitacions produïdes per l'escletxa digital de la població vulnerable durant el taller i es garanteix la participació de forma correcta d'aquestes persones.
- El temps de recollida de dades es redueix. En les dues hores que dura el taller es recullen totes les respostes d'un territori determinat
- Les dades mancants es redueixen.
- Es redueix o s'elimina la interpretació errònia de les preguntes

5.9.1. Criteris d'inclusió i d'exclusió

En col·laboració amb els experts en el domini d'aplicació, i després d'analitzar el marc conceptual, es defineixen perfils d'interès per al projecte, amb la intenció d'aprofundir en aquells segments de població a priori més lligats als objectius de l'estudi.

Un cop clar els segments objectius, cal preparar un document distribuit a les entitats col·laboradores del projecte.

5.9.2. Determinació de la grandària de la mostra i disseny mostral

Per tal de definir quina és la grandària de la mostra, s'ha construit un simulador que permeti veure com evoluciona la grandària de la mostra necessària en funció de l'error mostral i la confiança, el qual permeti avaluar les dificultats en la realització dels tallers en relació amb el valor de l'evidència aportada per les dades recollides en el mateix.

La primera conclusió que s'extreu de la simulació és que es pot treballar amb la suposició de població infinita, donat que per una població de més d'un milió d'habitants els valors de les corresponents expressions finites han convergit ja als valors límit expressats sota hipòtesi de població infinita.

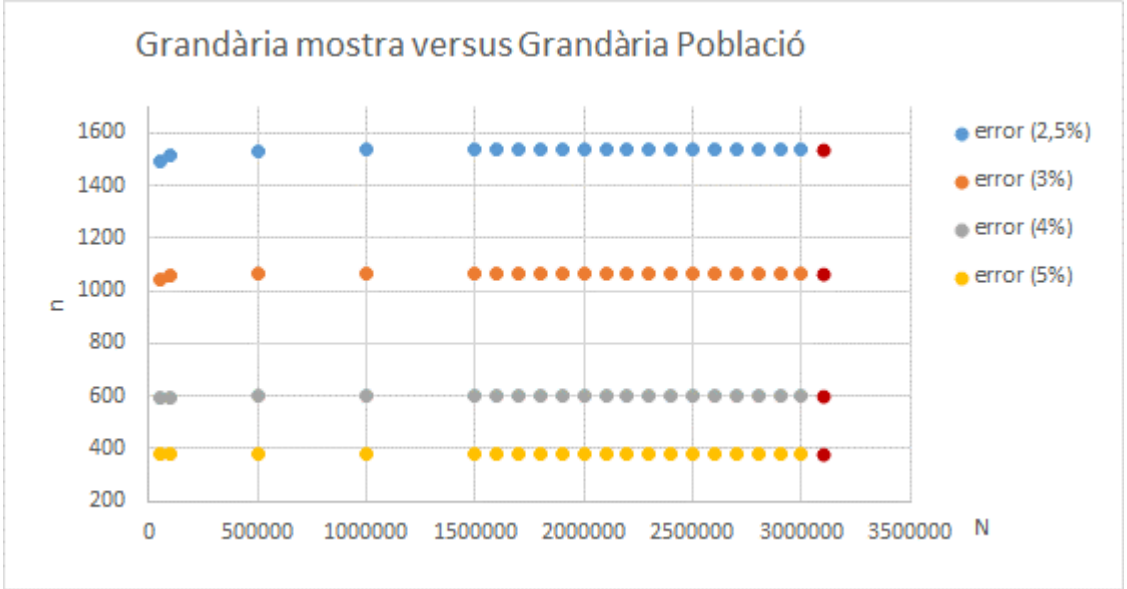

Figura 16: Grandària de la mostra vs Grandària de la població

En segon lloc, es mostra l'impacte de modificar la confiança o el marge d'error en la grandària de la mostra necessària per obtenir validesa estadística.

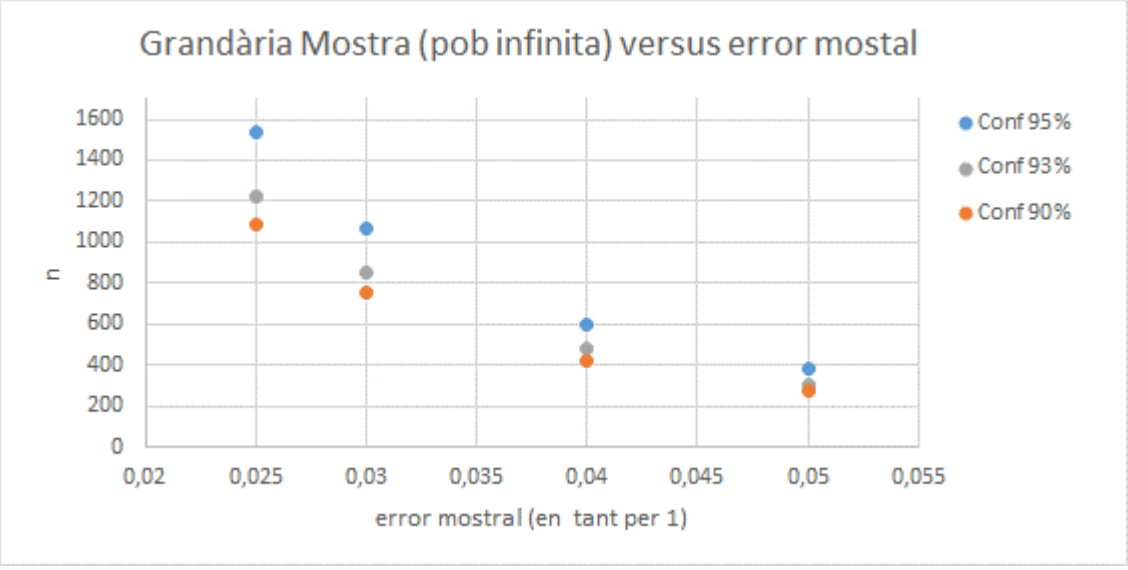

Figura 17: Grandàroa de la mostra vs error mostral

Si es dissenya una enquesta amb un mínim de 5 opcions per triar en totes les variables qualitatives (i s'assumeix resposta simple) l'exigència és menor encara.

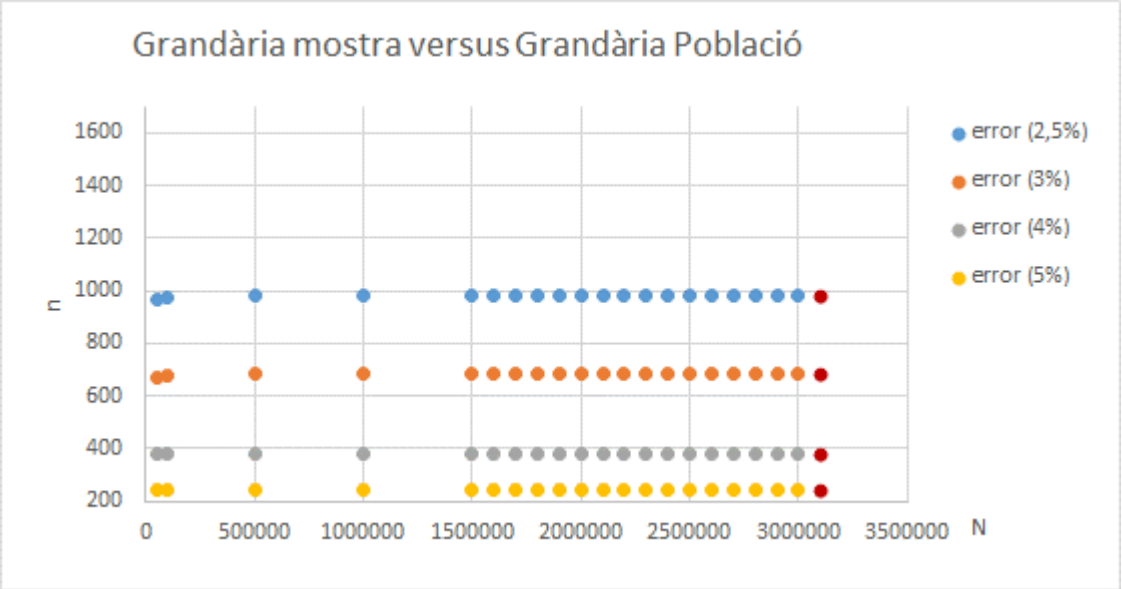

Figura 18: Grandària de la mostra vs Grandària de la població

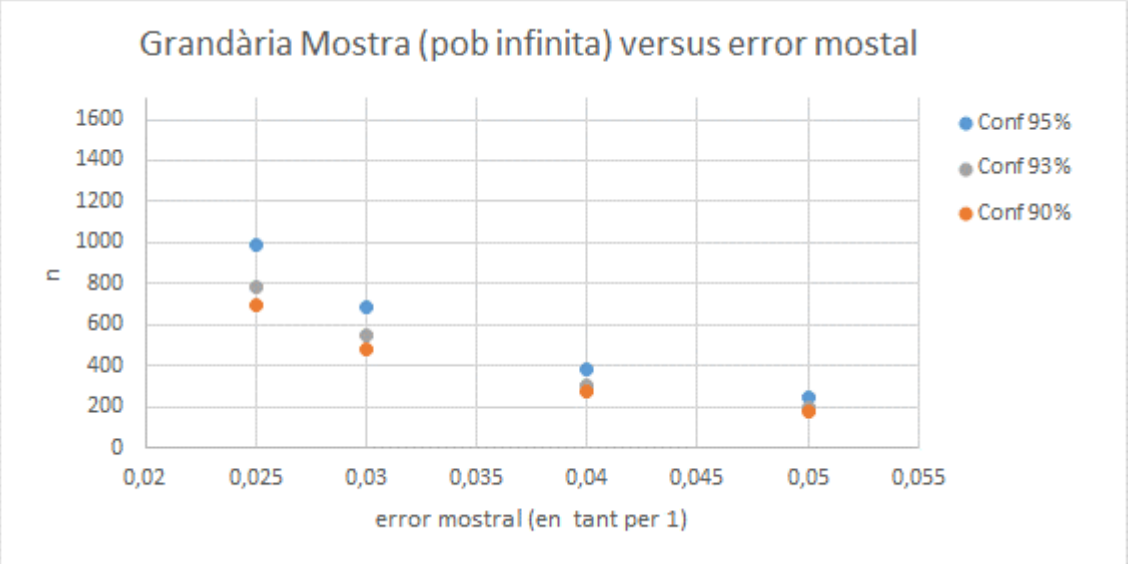

Figura 19: Grandària de la mostra vs error mostral

5.10. Tipologia de tallers

Es dissenya un mètode de recollida de dades mitjançant diferents tipus de tallers, així com la definició del principi de la metodologia del taller.

La infraestructura abans esmentada obeeix a l'objectiu de poder recollir dades directament de ciutadans i ciutadanes, corresponents als perfils d'interès, i que ens puguin omplir el qüestionari com a principal font d'alimentació de la base de dades sobre la que se sustentarà l'anàlisi intel·ligent de dades.

La combinació síncrona-remota dels tallers permet també distribuir l'àmbit de l'estudi i accedir a persones que viuen en zones perifèriques del territori.

La metodologia del taller es basa en els següents principis:

- L'estructura territorial de suport no hauria d'aportar cap dada; cal tenir en compte que la disrupció invalida la possibilitat d'entrenar el model utilitzant dades històriques.
- Cada unitat territorial hauria de veure's mínimament impactat per participar en el projecte, especialment en situacions disruptives on els professionals han hagut de fer un gran esforç per atendre el dia a dia i es buscava una participació que els requerís el menor esforç possible.
- Per això la unitat territorial només se li demanaria que seleccionés les participants en el projecte i un espai.
- Només s'utilitzaria la informació del qüestionari per poder assegurar l'anonimat i el secret estadístic de la participant, i per això el qüestionari seria autocontingut
- Es minimitzaria l'error d'interpretació de les participants per garantir la qualitat de la dada recollida i evitar que es malmetés l'esforç dels participants i les unitats involucrades.
- Es mantindria control sobre els temps de resposta de les participants per evitar el problema habitual en les enquestes clàssiques de no resposta o d'haver d'invertir molts esforços a aconseguir-les.

Per donar cobertura a aquestes premisses es van crear els tallers INSESS-COVID19

La unitat organitzadora buscaria el nombre de ciutadans que prèviament ciutadans dels perfils indicats i els citaria en un dia i hora concret per a realitzar un taller amb nosaltres consistent a:

- Donar el context del projecte a totes les participants que poguessin entendre la importància de la seva participació
- Crear una visió homogènia del qüestionari (això evitaria les mal interpretacions de preguntes)
- Crear totes les condicions perquè tothom pogués respondre el qüestionari dins el taller (això evitaria el ràtios de no resposta i el cost de seguiment de la resposta en el projecte).
- Respondre en directe tots els dubtes que es poguessin plantejar en el procés de resposta i donar suport a l'escletxa digital.

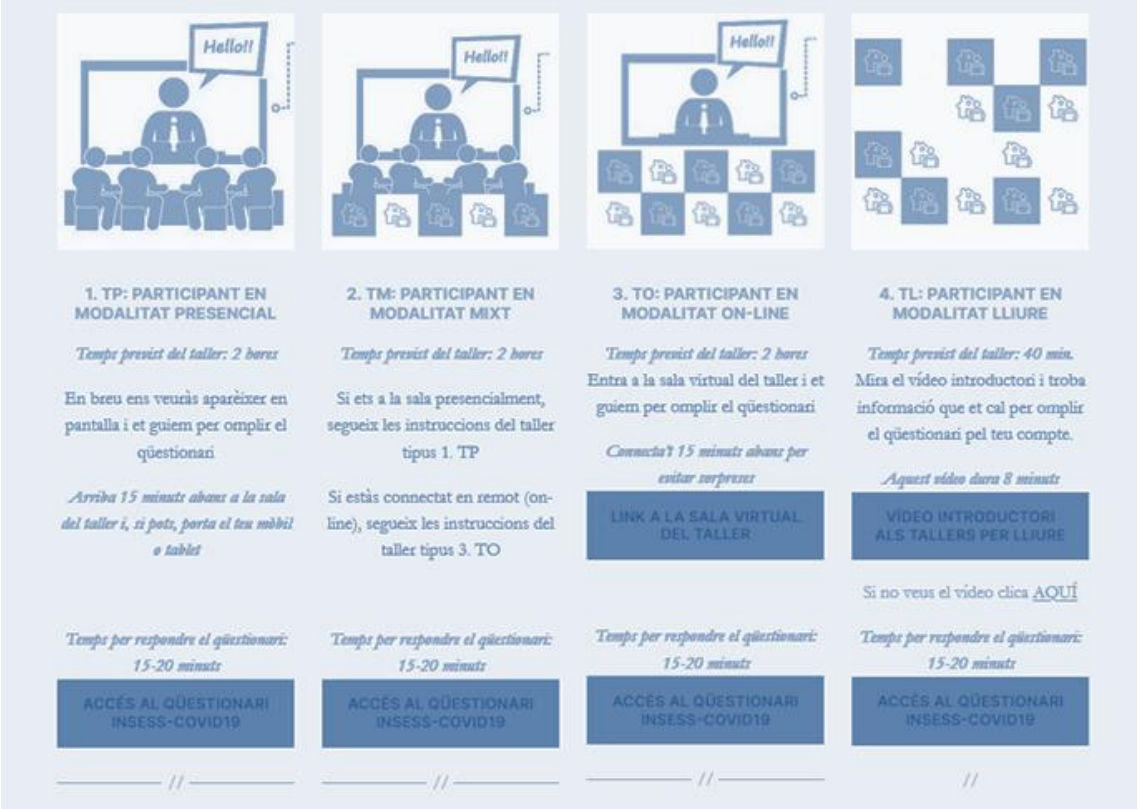

Figura 20: Tipologia de tallers

L'allargament de la pandèmia i els rebrots de juliol van posar en risc tota la concepció del projecte, inicialment basat en la covocatòria de tallers tradiconals i presentcials per tot el territori. Ens vam afanyar a evolucionar el concepte de ***Taller INSESS*** creant 4 modalitats noves de tallers sota diferents graus de presencialitat, mantenint però l'ABSS cost per als professionals de les unitats d'anàlisi que ens donen suport i la facilitat de respondre en uns 20 minuts per part del ciutadà. Aquesta adaptació va suposar un canvi radical del rol de la web en el projecte i dels seus continguts.

Es consideren, doncs, 4 tipus diferents de taller:

Figura 21: Taller presencial

- **Taller presencial**: Totes les persones que estan participant en el taller es reuneixen en un lloc proporcionat per l'organitzador L'equip INSESS s'uneix a través de

videoconferència per dirigir el taller. La principal limitació d'aquest disseny és requerir coincidència en el temps i l'espai entre participant i organitzador local i requereix logística i habitacions específiques ofertes per la unitat territorial per a la celebració del taller.

Figura 22: Taller mixt

- **Taller mixt** : Alguns participants es troben a la sala assignada per l'organitzador del taller. A més, alguns participants s'uneixen al taller a través de videoconferència. L'equip INSESS-COVID19 s'uneix a través de la videoconferència.

Figura 23: Taller online

- **Taller Online**: Tots els participants s'uneixen al taller a través de videoconferència. L'equip INSESS-COVID19 s'hi uneix mitjançant videoconferència per dirigir el taller.

Figura 24: Taller lliure

- **Taller Lliure**: El taller per lliure dona tota la llibertat al respondent de fer el taller quan vulgui des de casa i des del seu propi mòbil, amb un vídeo de suport que substitueix la tasca de contextualització inicialment dissenyada per als tallers presencials o semipresencials.

Amb les circumstàncies de la pandèmia prolongades en brots successius, es va activar una versió en línia, deslocalitzada en el temps i l'espai. La contextualització de

l'activitat es va pre-enregistrar en vídeos, es va penjar a la web, de manera que cada participant ha d'entrar a la web, seguir els vídeos (10 min) i respondre al qüestionari, tots disponibles en una àrea web privada. En aquesta modalitat, les propietats del taller són:

- No cal oferir una sala específica per celebrar el taller
- No cal fixar un dia i hora per celebrar el taller
- El temps per a la recopilació de dades ha de ser un període més llarg. Es requereix un seguiment addicional dels participants per garantir el lliurament dels qüestionaris a temps
- Les dades que falten poden augmentar

La malinterpretació de les preguntes encara es redueix a través dels vídeos, però no hi ha cap interacció disponible, de manera que podria no ser eliminada del tot

- Es requereix un suport humà específic a llarg termini per resoldre les diferències digitals de la població vulnerable. La unitat organitzadora ha d'oferir una persona a aquest propòsit.

- L'objectiu principal dels minivídeos és garantir que tots els participants tinguin la mateixa comprensió de les preguntes i coneguin els objectius principals del projecte, ajudant així a reduir les dues interpretacions errònies de les preguntes.

El projecte va considar quatre modalitats de taller (Figura 20)

5.11. FASE III Anàlisi intel·ligent de dades

5.11.1. Preprocessament de les dades:

La primera fase està formada pel que s'anomena preprocessing, que té per objectiu principal la depuració i preparació de les dades per al seu posterior tractament i manipulació amb els diferents softwares que s'empraran al llarg de l'anàlisi. La realització correcta d'aquest punt és molt important per tal de garantir el valor de l'anàlisi resultant.

Per al cas d'estudi, el disseny dels tallers ha ajudat al fet que les persones responguessin amb responsabilitat i no ha calgut eliminar respostes del qüestionari. Tampoc s'han detectat mal interpretacions en les preguntes.

La implementació de canvis de format en les dades per poder-les passar d'un software a un altre i per poder-les passar de format "dades de qüestionari" a matrius d'anàlisi són les operacions que han requerit més esforç d'implementació per aquesta fase.

Un cop s'ha creat l'enquesta juntament amb el fitxer de metadades corresponent e, es pot abordar el pas del preprocessament. En aquest treball, el pas del preprocessament és

implementat per R, on les tècniques especificades en [Gibert, Sànchez-Marrè & Izquierdo, 2019] s'apliquen com la imputació o recodificació que falta. El pas del preprocessament té cinc passos principals:

- **Detecció i correcció d'errors:** on els errors a la base de dades són detectats i tractats utilitzant la tècnica apropiada, com es mostra a [Gibert, Sanchez-Marre & Izquierdo, 2019].
- **Imputació de mancants:** aquelles preguntes que són contestades per tots els participants, contenen la modalitat "No Resposta" en preguntes qualitatives i es manté com una modalitat addicional. Per a les preguntes de filtre, les cel·les s'inicialitzen a «Filtre», de manera que no es generen manques per als subgrups que salten la pregunta.
- **Modalitats i supressió de duplicitats:** els formularis de Google generen una variable automàtica de temps de dades amb la marca horària de les respostes. En el cas de 2 files repetides en les mateixes marques de temps, s'elimina la redundància. En el cas de variables amb correlació 1, o redundants, una d'aquestes també s'elimina.
- **Recodificacions de preguntes:** La variable original i els noms de les modalitats poden ser massa llargs en moltes aplicacions reals per permetre un etiquetatge correcte dels títols, eixos, variables o etiquetes individuals de la trama. Utilitzant la informació de MdM donada en preguntes k ., Q, A* i J*, els noms de les variables i les modalitats es recodifiquen en formes curtes comprensibles i compatibles amb la representació gràfica. Per a les variables representades en més d'una columna del conjunt de dades, com preguntes TQQ o Variables de graella, s'utilitza un procés específic tenint en compte les columnes n_D implicades. El tipus de columna variable en el fitxer de metadades determina l'algoritme apropiat per a reformular cada pregunta.
- **Verificació del preprocessament:** Per a validar que la reformulació s'ha fet correctament, les taules de contingència es construeixen entre les variables originals i preprocessades. El resultat hauria de ser com una matriu diagonal, validant que totes les modalitats de la variable original corresponen al seu nou nom curt.

Un cop recollides les dades és necessari procedir a l'anàlisi mitjançant la infraestructura desenvolupada amb anterioritat. La metodologia que cal emprar per procedir a l'anàlisi és el següent:

5.11.2. Anàlisi descriptiva i territorial (mapes estadístics):

Variable a variable, es mostren els estadístics bàsics que ens permeten fer-nos una idea de la distribució de la distribució de la variable. L'anàlisi descriptiva es realitzada automàticament i forma part de la documentació generada en el report automàtic.

Per al cas d'estudi, l'anàlisi descriptiva més bàsica ha permès disposar d'informació sobre la participació en tot moment, de com es representaven els perfils objectiu a les diferents àrees territorials i de com es distribuïen les respostes a les diferents preguntes.

S'ha dissenyat una component de realització de mapes automàtics que funciona per diferents nivells d'agregació territorial (municipis, vegueries, províncies)

A continuació es presenten detalladamentles eines estadístiques més innovadores que s'utilitzen per analitzar variables temporals o multivaluades, i els altres tipus complexos definits

De fet, algunes de les eines utilitzades són molt bàsiques, però altres s'han desenvolupat ex professo en aquest projecte i obert la porta per ampliar el coneixement proporcionat en la primera anàlisi descriptiva de qualsevol base de dades, atès que el tipus de variables són adequadament conceptualitzades abans de la pròpia anàlisi. A continuació, es proposa la descripció de les noves eines descriptives avançades.

Cadascuna d'aquestes eines s'ha validat correctament abans d'incloure en els procediments utilitzats per analitzar les dades del projecte. En primer lloc, la proposta es va validar amb les parts interessades de l'informe per a veure si apreciaven la informació útil facilitada per l'eina. Després es va realitzar la validació tècnica dels scripts que els implementaven. Finalment, la interpretabilitat dels resultats es va utilitzar com a criteris de validació final quan es va presentar tot l'informe del projecte a les parts interessades finals.

El programa desenvolupat en la fase de desenvolupament de la infraestructura tecnològica descansa sobre un software que genera documentació de forma automàtica en un programa editable [Gibert & Nonell 2005].

En el cas del projecte, la major part de les anàlisis elaborades han descansat sobre tecnologies que composaven automàticament documents de Word editables amb els resultats de l'anàlisi, per tal d'escurçar els temps entre la tancada de la recollida de dades i l'obtenció dels informes de resultats.

			Graphical tools											Numerical tools									
			Univariant					Multivalued		Bivariant		Trivariant	4-variant	Univariant				Multivalued		Bivariant			
			Pie chart	Barplot	Histogram	Boxplot	WordCloud	Marginal Pie chart	Marginal Barplot	Multiple barplot	Grid of pies	Trajectory map	Multiple Stacked Barplots	Frequency Table	Extended S-Number Summary	Standard error	95% CI error	Multivalued Freqs table	Trajectory freqs table	Cross table (counts)	Cross table (proportions)	Transition table	
Type	**Form of question**	**DB structure**																					**Example**
Numerica	Integer field	One numerical column			x	x									x	x	x						Age
Likert or Ordinal	Simple choice response	One alphanumeric column	x	x										x		x	x						Pending processes
Nominal	Simple choice response	One alphanumeric column	x	x										x		x	x						Gender
Nominal Multivalued	Multiple choice response	One column with lists of values separated by ";"						x	x					x				x					Impacted area (by process)
OrdinalXtime	Simple choice grid	One alphanumeric column per timeStamp								x	x	x							x	x	x	x	Convivential Unit
OrdinalXordinalxTime	Several Packs of several Ordinal	Several Packs of several columns								several	several		x							several	several		Kinds in charge
Open question	Open textual window to be edited by the respondant	Textual column with complete text					x																Why didn't you receive the payments by July 10th?

Figura 25 Tipologia de preguntes, variables, estructures de dades i eines d'anàlisi

5.11.3. Eines gràfiques innovadores. Diagrama de teler

Originalment introduït en [Gibert et al. 2009] i a [Gibert, Rodriguez S & Rodriguez R, 2010], consisteix en un gràfic bidimensional amb les modalitats de la variable qualitativa de destinació (ordenada o depenent si és nominal, o Likert o ordinal). El temps es representa en l'eix X i és discret. Aquesta eina s'utilitza per representar totes les variables bàsiques temporals. Les corresponents a les característiques dinàmiques i mesurades en diferents moments de temps. Per a cada individu, els nodes que representen les seves opcions al llarg del temps estan enllaçats amb una aresta. Vores del mateix color representen la mateixa trajectòria dels individus. El gruix de les trajectòries representa la proporció dels enquestats que segueixen aquest patró. Els gràfics de trajectòria representen en un únic paquet d'eines de 3 columnes diferents en el fitxer de dades corresponents a la mateixa variable X mesurada en 3 marques de temps X_{T1}, X_{T2}, X_{T3}; on cada X_{Ti} és una rèplica de X que mostra el valor al llarg del temps. Els gràfics de trajectòria ensenyen quins individus evolucionen de manera similar. Donen l'oportunitat d'identificar els patrons temporals i trobar quines variables els distingeixen. Aquesta anàlisi interpretativa genera hipòtesis sobre quins factors estan associats a evolucions negatives o perjudicials per als individus. L'eina és transversal, i s'ha utilitzat en [Gibert et al. 2009] per identificar causes de deteriorament funcional en pacients neurològics amb lesió de la medul·la espinal durant el procés d'inclusió social després de la descàrrega. En [Gibert, Rodriguez S & Rodriguez R, 2010] es va utilitzar per entendre els patrons d'evolució del mode d'operació de les plantes de tractament d'aigües residuals diàriament. Aquí s'aplica descobrir les principals tendències de l'evolució temporal de les principals variables del qüestionari INSESS-COVID19 un per un.

L'ús de gràfics de teler en R és una contribució d'aquesta tesi, sent aquesta la primera vegada que s'implementa en R per ser representat automàticament en informes automàtics. La figura 9 mostra la gràfica de trajectòria de les relacions convivencials.

La qualitat variable de les relacions convivencials és ordinal i pot prendre 10 modalitats diferents (des de 01. Satif (Satisfactoria) a 9. Inexistent i 10. NC (falta). Aquesta variable ha estat mesurada per tres marques de temps al qüestionari. Una línia del gràfic representa cada enquestat. Tots els enquestats que segueixen el mateix camí temporal es mostren amb el mateix color de línia.

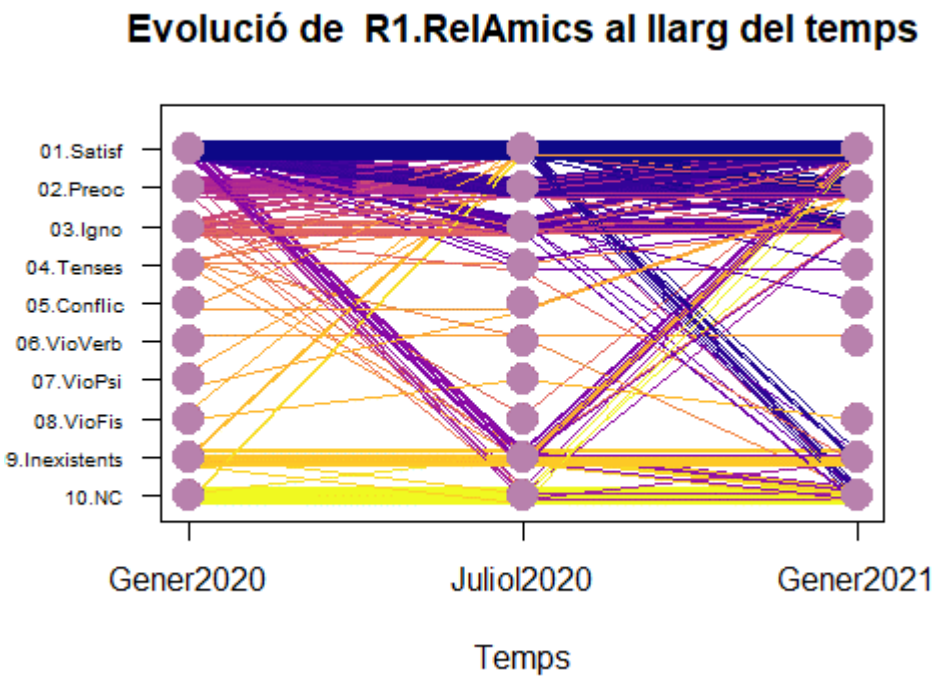

Figura 26: Diagrama de teler de la pregunta R1.

El patró “VΛ” és un patró especial identificat per primera vegada durant aquesta tesi. Correspon a una doble dinàmica en el mateix procés, on part dels individus segueixen un patró de "V" (entre els 2 primers moments del temps empitjoren la seva situació i es recuperen per al 3r moment del temps, tornant on eren el primer moment), mentre que un altre segment dels individus segueixen el patró oposat").”Λ”” (estaven en males condicions en el primer moment i ho tornaran a estar en el tercer moment del temps. Entre el primer i segon període, les persones participants han millorat en altres àmbits, això fa que en el 2n període de temps estigun millor per a la característica preguntada).

5.11.4. Eines innovadores. Taules de transicions

Taula creuada de la categorització de les successives taules de transició que quantifica quants canvis d'estat s'observen tant en la primera com en la segona transició. Les taules següents són exemples de canvis en la qualitat de les relacions en la Unitat Convivencial.

	Millor	Igual	Pitjor	NC	Sum
Millor (better)	0.009	0.042	0.010	0.004	0.066
Igual (Same)	0.027	0.690	0.016	0.009	0.743
Pitjor (Worse)	0.056	0.040	0.004	0.009	0.109
NC	0.001	0.005	0.000	0.076	0.082
Sum	0.093	0.778	0.031	0.099	1.000

Taula 6:Els canvis reportats al llarg del temps (relatiu)

	Freq	Prop
Improve	76	0.078
V pattern	4	0.004
Balance	670	0.690
Λ pattern	10	0.010
Enworse	59	0.061

Taula 7:Canvia els patrons (unitat convivent).

	Freq	Prop
Improve	76	0.078
V pattern	5	0.005
Balance	642	0.661
Λ pattern	13	0.013
Enworse	83	0.085

Taula 8:Figura 14: (a) Canvia els patrons en les relacions familiars amb la persona que viu fora de casa

	Freq	Prop
Improve	65	0.067
V pattern	7	0.007
Balance	662	0.682
Λ pattern	10	0.010
Enworse	80	0.082

Taula 9:Canvia els patrons en les relacions amb els veïns

5.11.5. Resum ampliat de 5 nombres

Sent X una variable numèrica (x_1,...-x_n), el resum de 5 nombres [Moore et al. 1993] és un conjunt de 5 estadístiques prou robustes utilitzades per descriure variables numèriques. Està composat pel mínim, el primer quartil (Q1), la medianaitjana, el tercer quartil (Q3) i i elmàxim. En la nostra versió, l'ampliem afegint la mitjana, la quasi desviació estàndar i el coeficient de variació de manera que la informació sobre la simetria de la variable i la rellevància de la variància també es pot avaluar.

Min	Q1	Mediana	Mitjana	Q3	Max	Des	CV
min(X)	$x\ tq\ card(X \leq x) = 0.25n$	$x\ tq\ card(X \leq x) = 0.5n$	$\frac{\sum_{i=1}^{n} x_i}{n}$	$x\ tq\ card(X \leq x) = 0.75n$	Max(X)	$\frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$	$\frac{s}{\bar{X}}$

Taula 10: Resum ampliat de 5 nmbres

5.11.6. Eines innovadores: Taula de freqüències ampliada

Com que X és una variable nominal, la taula de freqüències exteses (Taula 11) estén la tradicional amb l'error estàndard, calculat segons les expressions descrites en aquest article i la desviació estàndard agrupada de totes les modalitats juntes com un indicador de bondat de la qüestió en el seu conjunt. Per a variables qualitatives nominals, les modalitats es presenten en ordre descendent, en un estil Paretto, de manera que les modalitats més freqüents apareixen a la part superior de la taula. Per a les variables de Likert, es presenta l'ordre original de les modalitats.

P2.Gènere	Freq.	Prop.	Std. Err
2.Dona	655	0.675	0.0152
1.Home	307	0.316	0.0148
3.NoBinary	5	0.005	0.0032
4.No Contesta	4	0.004	0.0000

95% CI error: ± 5x10-4e-04 Std. Error of the question: 0.0107

Taula 11Teula de frequencies ampliada de gènere

5.11.7. Diagrama de barres, sectors o taula de freqüències marginals

Les variables multivaluades proporcionen respostes multivaluades compostes per subconjunts de modalitats. Aquest és el cas, per exemple, dels dispositius digitals utilitzats per una persona (poden ser múltiples, Telèfon mòbil, tauleta, pc, portàtil...). Sent X una variable multivaluada (x_1,...,x_n), on xi és una llista de modalitats separades per «;». Les freqüències de cada modalitat de la variable no estan disponibles per anàlisi directa.

El diagrama de barres marginals, com a la figura 27, aparentment sembla el diagrama de barres clàssic, però es construeix sobre una variable multivaluada. Això significa que un individu pot estar representat en diverses barressimultàniament. En conseqüència, la columna de proporcions corresponent supera el 100%. De manera que la taula de freqüències marginal té un aspecte similar a la taula de freqüències però representa proporcions que sumen més del 100%. El mateix passa amb el diragrama de pastís. Tots ells representen els recomptes marginals o proporcions dels dummies (eventualment) que representen cadascuna de les modalitats de la variable, independentment de com aquesta variable es representa internament en la base de dades (com una sola columna de llistes de valors en les cel·les, o com un conjunt de dummies, un per modalitat). La figura27 mostra l'àmbit de la vida impactat per processos judicials no resolts. La mateixa persona pot tenir diverses àrees afectades simultàniament, com l'estat civil (procés de divorci, per exemple) i l'economia i la família.

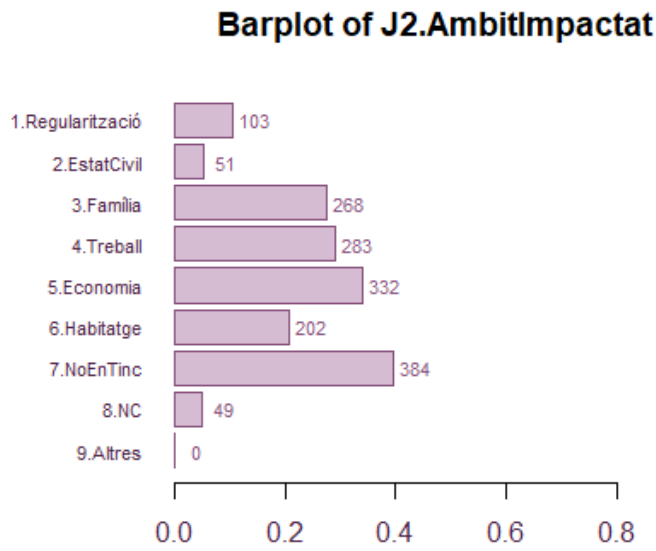

Figura 27: Gràfic de barres marginals de la pregunta J2.

5.11.8. Taula de freqüències multivaluada

Com que les variables nominals multivaluades estan representades per columnes amb llistes de modalitats a les cel·les, proposem la taula de freqüències multivaluada per analitzar les bosses de modalitats seleccionades pels enquestats. A la taula de freqüències multivaluada, tots els subconjunts de modalitats proporcionats com a respostes es mostren amb els seus corresponents recomptes i freqüències. De fet, això representa un subconjunt de la distribució de probabilitat de la variable. Per preservar el secret estadístic les combinacions es publiquen només per freqüències superiors a tres. El nombre de combinacions ocultes també s'informa al final, així com les mètriques d'incertesa (uncertainity methods). Aquestes variables s'implementen a través de preguntes d'elecció múltiple al qüestionari. Quan es col·lapsa en bosses de modalitats, el seu pes en l'anàlisi es manté com una variable. Quan es representen com a variables fictícies, com en la manera tradicional, poden biaixar l'anàlisi a mesura que augmenten la dimensionalitat del conjunt de dades innecessàriament.

5.11.9. Taula de freqüències de trajectòria

Per a variables bàsiques temporals: Aparentment sembla una taula de freqüències multivaluada. La diferència principal és que s'ha construït a partir d'un conjunt de diverses variables qualitatives (una per marca de temps), cadascuna d'elles són d'elecció simple i es representa en una columna diferent en el conjunt de dades. Quantifica la informació que es mostra al diagrama de teler. Vegeu a la taula 12 la taula de freqüències de trajectòria corresponent a la R1. Variable RelUConv presentada més tard a la secció Resultats.

R1.RelUConv	Frequencies
01.Satisf+01.Satisf+01.Satisf	596
10.NC+10.NC+10.NC	64
02.Preoc+02.Preoc+02.Preoc	33
09.Inexistents+09.Inexistents+09.Inexistents	25
01.Satisf+02.Preoc+02.Preoc	24
01.Satisf+02.Preoc+01.Satisf	23
02.Preoc+01.Satisf+01.Satisf	14
02.Preoc+02.Preoc+01.Satisf	12
01.Satisf+01.Satisf+02.Preoc	9
01.Satisf+01.Satisf+10.NC	7
04.Tenses+04.Tenses+04.Tenses	7
09.Inexistents+01.Satisf+01.Satisf	7
01.Satisf+04.Tenses+01.Satisf	6
02.Preoc+01.Satisf+02.Preoc	6
05.Conflic+05.Conflic+05.Conflic	6
03.Igno+01.Satisf+01.Satisf	5
04.Tenses+04.Tenses+01.Satisf	5
01.Satisf+03.Igno+01.Satisf	4
01.Satisf+09.Inexistents+10.NC	4
02.Preoc+04.Tenses+04.Tenses	4
04.Tenses+05.Conflic+01.Satisf	4
05.Conflic+01.Satisf+01.Satisf	4
09.Inexistents+03.Igno+10.NC	4
10.NC+01.Satisf+01.Satisf	4
01.Satisf+03.Igno+03.Igno	3
01.Satisf+04.Tenses+04.Tenses	3
03.Igno+03.Igno+03.Igno	3
04.Tenses+01.Satisf+01.Satisf	3
04.Tenses+04.Tenses+03.Igno	3
05.Conflic+04.Tenses+01.Satisf	3

Taula 12: Taula de trajectòries

5.11.10. Diagrama de barres múltiples

Com és habitual, representa la distribució de probabilitat conjunta de 2 variables qualitatives. En aquest cas, una és el temps. L'altre és una variable nominal, ordinal o Likert. Per a variables bàsiques temporals. Vegeu un exemple a la Figura 28

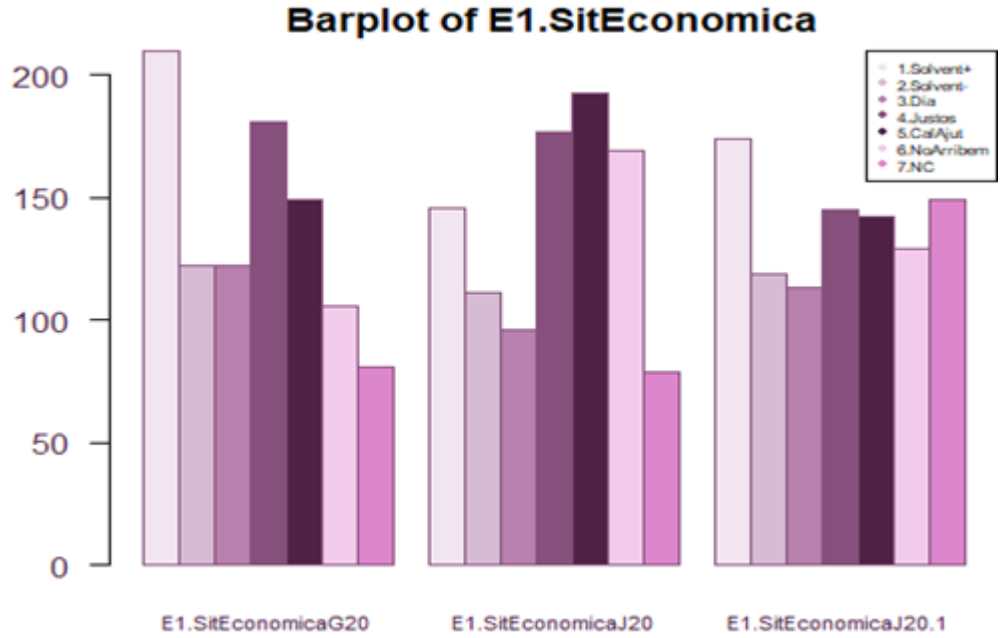

Figura 28. Diagrama de barres múltiples de situació econòmica.

5.11.11. Graella de diagrames de pastis

Per a les variables bàsiques temporals, les columnes T que representen el temps es poden analitzar independentment com si fossin variables qualitatives ordinàries. Es pot fer una representació gràfica de pastís per a cada marca horària i es presenten en una xarxa Vegeu un exemple en la figura 29 per a la situació econòmica.

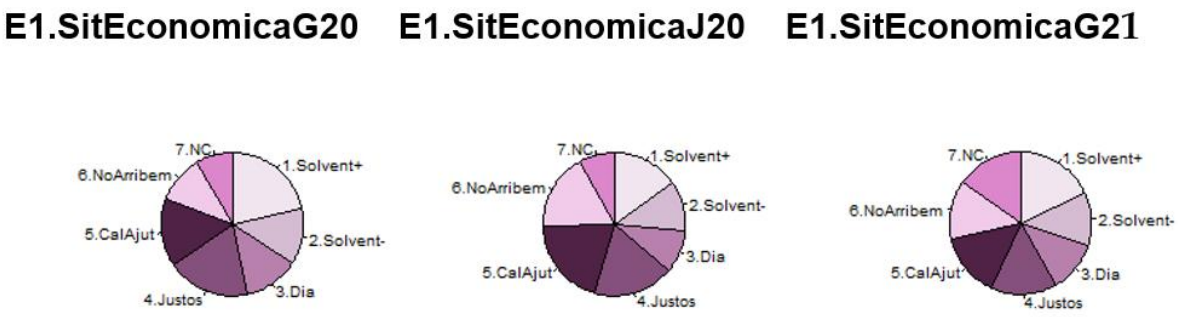

Figura 29: Grella dels diagrames de pastis de la pregunta E1. Situació econòmica.

5.11.12. Eines innovadores: Taules de transició

Taules que quantifiquen les transicions entre dues marques de temps consecutives, en recomptes o proporcions. Donada una variable bàsica temporal (X,T), és la taula creuada entre X_t i X_t+1, t = {1:T-1}. Vegeu un exemple a la taula 13 els canvis en la qualitat de les relacions en la unitat convivencial entre gener de 2020 i juliol de 2020.

	01. Satisf	02. Preoc	03. Igno	04. Tenses	05. Conflic	06. VioVerb	07. VioPsi	08. VioFis	09.Inexist ents	10. NC

01.Satisf	616	50	9	12	5	0	1	0	6	4
02.Preoc	21	48	0	8	2	1	0	0	1	0
03.Igno	5	0	3	0	1	0	0	0	0	0
04.Tenses	3	1	0	16	4	1	1	0	1	0
05.Conflic	4	1	3	4	7	0	1	0	0	1
06.VioVerb	1	3	2	0	0	0	0	0	0	0
07.VioPsi	1	1	0	0	0	0	2	1	0	0
08.VioFis	0	0	0	0	0	0	1	1	1	0
09.Inexistents	9	0	4	0	0	0	0	0	28	3
10.NC	4	0	0	1	0	0	0	0	1	66

Taula 13 Canvis entre gener i juliol de 2020 (variable gener 2020—Juliol 2020)

5.11.13. Diagrama de barres apilades múltiple

Aquesta és una proposta de representació gràfica per proporcionar una vista compacta d'una variable de tipus TQQ. En aquest cas, les tres parcel·les de barres apilades representen la participació en la societat a través del temps. Per a cada marca horària, un gràfic de barres bivariants apilat representa la relació entre la Q Likert (en barres) i les modalitats de X, en aquest cas, indicant si la persona participa més o menys en activitats socials (com xarxes de veïnatge, associacions o moviments voluntaris). Els canvis al llarg del temps també es poden analitzar. Vegeu Figura 30.

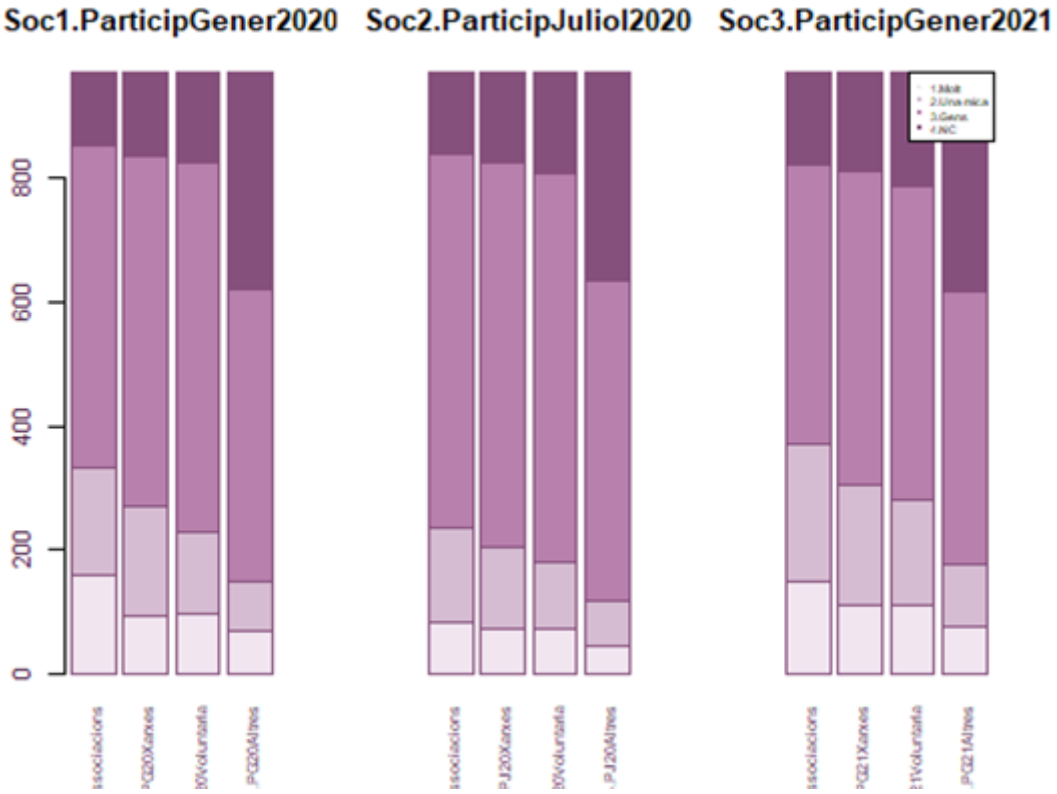

Figura 30 Gràfic apilat múltiple o Pregunta Soc1-2-3.

5.12. Estructura de l'anàlisi descriptiva per tipus de variables

5.12.1. Variables numèriques

Quants anys fa que vius a Catalunya? Aquesta és una pregunta que genera una variable numèrica. El participant respon amb un número a l'enquesta i a la base de dades generada per Google es mostra de la següent manera:

P3. Edat
53
41
83
59

Figura 31: Visualtizacio d'una variable numèrica a la base de dades generada per Google Forms

Així mateix, Google forms mostra el resultat en el següent gràfic:

Figura 32: Visualització de les variables numèriques a Google Forms

Com es pot observar, aquesta descripció no és suficient per extreure tota la informació que aquesta variable pot aportar. Així doncs, per al cas de les variables numèriques la metodologia MIPRI2D proposa que les variables numèriques es descriguin amb les següents tècniques:

- 5-number summary (resum en 5 números) ampliat, d'acord amb el que s'ha exposat a XXX
- Un histograma com es pot veure a la figura 33
- Un diagrama de caixes com es pot veure a la figura 34
- L'error estàndard de la pregunta calculat amb la fórmula expressat a XX

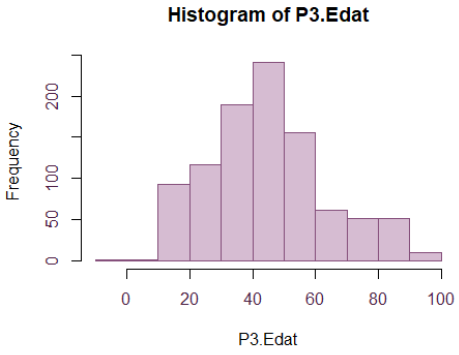

Figura 33: Histograma

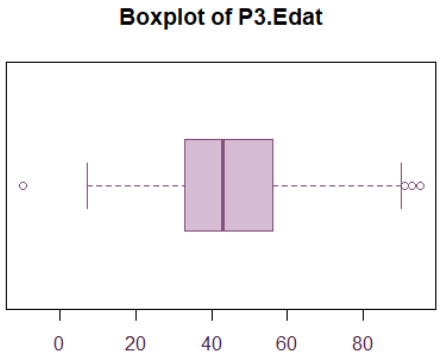

Figura 34: Boxplot

5.12.2. Variables categòriques

Tens una situació regularitzada? Aquesta és una pregunta que genera una variable categòrica X. El participant respon amb una de les modalitats a l'enquesta i a la base de dades generada per Google es mostra de la següent manera:

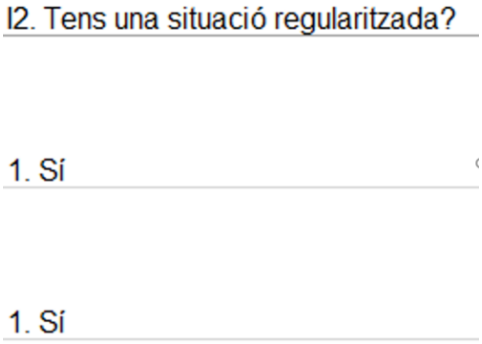

Figura 35: Visualtizacio d'una variable categòrica a la base de dades generada per Google Forms

Així mateix, Google forms mostra el resultat en el següent gràfic:

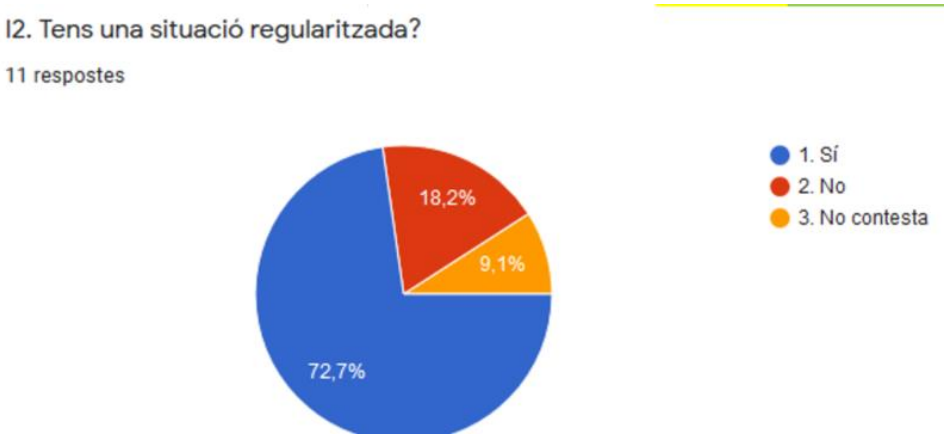

Figura 36 Visualització de les variables categòriques a Google Forms

Com es pot observar, aquesta descripció no és suficient per extreure tota la informació que aquesta variable pot aportar. Així doncs, per al cas de les variables categòriques la metodologia MIPRI2D proposa que les variables categòriques es descriguin amb les següents tècniques:

- Diagrama de sectors com es pot veure a la figura 38
- Un diagrama de barres com es pot veure a la figura 37
- Una taula de freqüències ampliada,

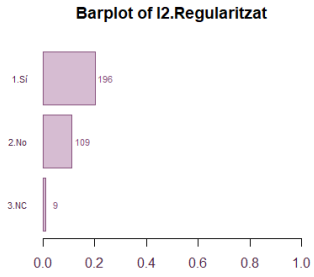

Figura 37: Diagrama de barres

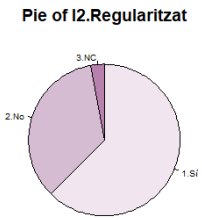

Figura 38: Diagrama de pastís

5.12.3. Variables multivaluades

R4. En cas que siguis objecte de violència, qui exerceix aquesta violència? Aquesta és una pregunta que genera una variable multivaluada com les descrites a la secció 5.3.3. El participant respon amb una o més modalitats a l'enquesta i a la base de dades generada per Google es mostra de la següent manera:

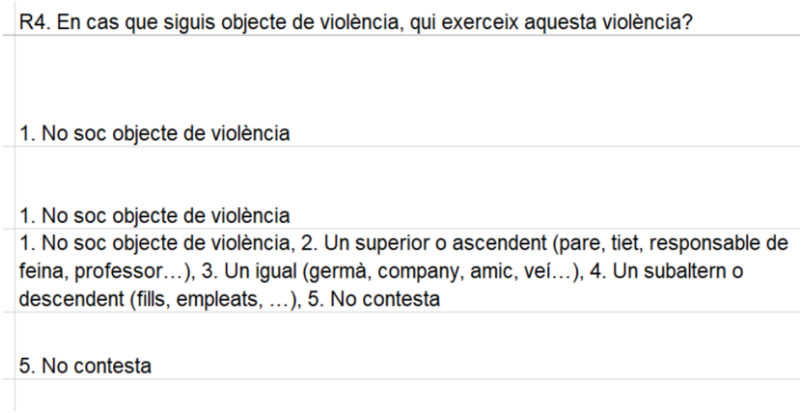

Figura 39: Visualtizacio d'una variable multivaluada a la base de dades generada per Google Forms

Així mateix, Google forms mostra el resultat en el següent gràfic:

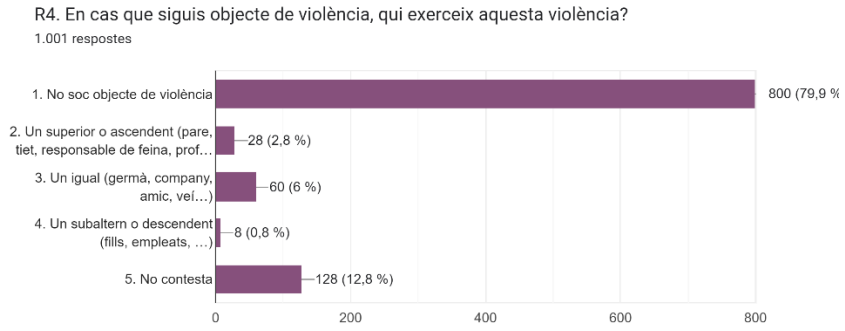

Figura 40: Visualització de les variables categòriques a Google Forms

Com es pot observar, aquesta descripció no és suficient per extreure tota la informació que aquesta variable pot aportar. Així doncs, per al cas de les variables numèriques la metodologia MIPRI2D proposa que les variables categòriques es descriguin amb les següents tècniques:

- Diagrama de sectors marginal
- Diagrama de barres marginals
- Taula de freqüències marginals
- Taula de freqüències multivaluada

5.12.4. Variables de quadrícula

P5. Quines llengües parles? Aquesta és una pregunta que genera una variable de quadrícula com les descrites a la secció 5.3.4. El participant respon amb el nivell que té de cada una de les llengües a l'enquesta i a la base de dades generada per Google es mostra de la següent manera:

P5. Quines llengües parles? [1. Català]	P5. Quines llengües parles? [2. Castellà]	P5. Quines llengües parles? [3. Anglès]	P5. Quines llengües parles? [4. Francès]	P5. Quines llengües parles? [5. Romanès]	P5. Quines llengües parles? [6. Àrab]	P5. Quines llengües parles? [7. Altres]
1. Bé	2. Regular	3. Amb dificultats	1. Bé	2. Regular	3. Amb dificultats	3. Amb dificultats
No contesta	No la parlo	No la parlo	Amb dificultats	Amb dificultats	Regular	Regular
Bé	Bé	Bé	Amb dificultats	No la parlo	No la parlo	No la parlo
Bé	Bé	Bé	Regular	Bé	Regular	Bé

Figura 41: Visualtizacio d'una variable de quarícula a la base de dades generada per Google Forms

Així mateix, Google forms mostra el resultat en el següent gràfic:

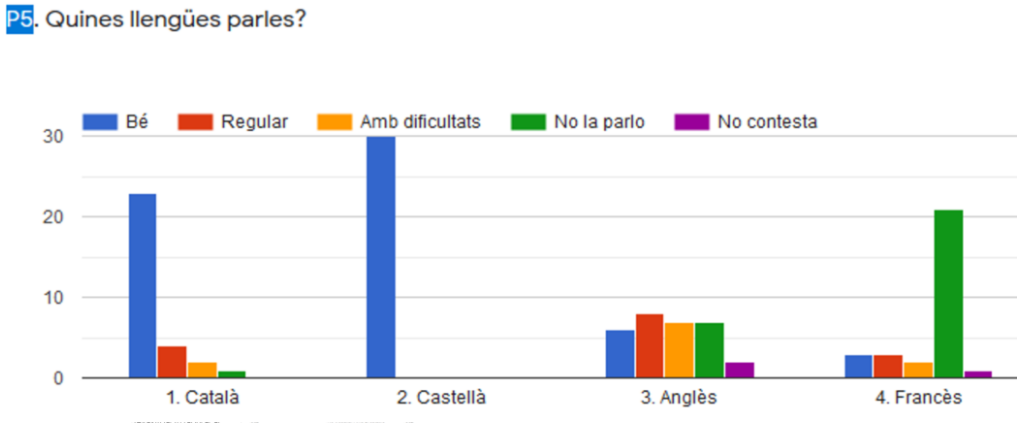

Figura 42: Visualització de les variables de quadrícula a Google Forms

Com es pot observar, aquesta descripció no és suficient per extreure tota la informació que aquesta variable pot aportar. Així doncs, per al cas de les variables de quadrícula la metodologia MIPRI2D proposa es descriguin amb les següents tècniques:

- Diagrama de barres múltiple
- Taules de contingència de freqüències absolutes, com es mostra a la taula 14
- Taules de contingència de freqüències relatives, com es mostra a la taula 15
- Una quadrícula de diagrames de sectors

	P5.1.Català	P5.2.Castellà	P5.3.Anglès	P5.4.Francès	P5.5.Romanès	P5.6.Àrab	P5.7.Altres
1.Be	598	889	102	61	11	111	90
2.Regular	166	68	182	101	3	10	10
3.Dificultat	107	12	191	154	62	55	62
4.NoParlo	95	2	434	584	794	709	553
5.NC	5	0	62	71	101	86	256

Taula 14: Taula de frequencies absolutes

	P5.1.Català	P5.2.Castellà	P5.3.Anglès	P5.4.Francès	P5.5.Romanès	P5.6.Àrab	P5.7.Altres
1.Be	0.616	0.916	0.105	0.063	0.011	0.114	0.093
2.Regular	0.171	0.070	0.187	0.104	0.003	0.010	0.010
3.Dificultat	0.110	0.012	0.197	0.159	0.064	0.057	0.064
4.NoParlo	0.098	0.002	0.447	0.601	0.818	0.730	0.570
5.NC	0.005	0.000	0.064	0.073	0.104	0.089	0.264

Taula 15: Taula de freqüències relatives

5.12.5. Variables bàsiques temporals

Marca el teu esquema de convivència en aquests tres moments. Aquesta és un enunciat del qüestionari que genera una variable bàsica temporal, com les descrites a la secció XXX. El participant respon l'esquema de convivència que presenta a cada moment del temps a l'enquesta i la base de dades generada per Google es mostra de la següent manera:

F2. Marca el teu esquema de convivència en aquests tres moments. [Gener 2020]	F2. Marca el teu esquema de convivència en aquests tres moments. [Juliol 2020]	F2. Marca el teu esquema de convivència en aquests tres moments. [Gener 2021]
3. Unitat familiar de pare(s) i mare(s) i fills propis	2. Família monomarental o monoparental	2. Família monomarental o monoparental
No-Fami	Extensa	Reagrupades
Reagrupades	Reagrupades	Reagrupades
Nucli	Nucli	Nucli

Figura 43: Visualtizacio d'una variable básica temporal a la base de dades generada per Google Forms

Així mateix, Google forms mostra el resultat en el següent gràfic:

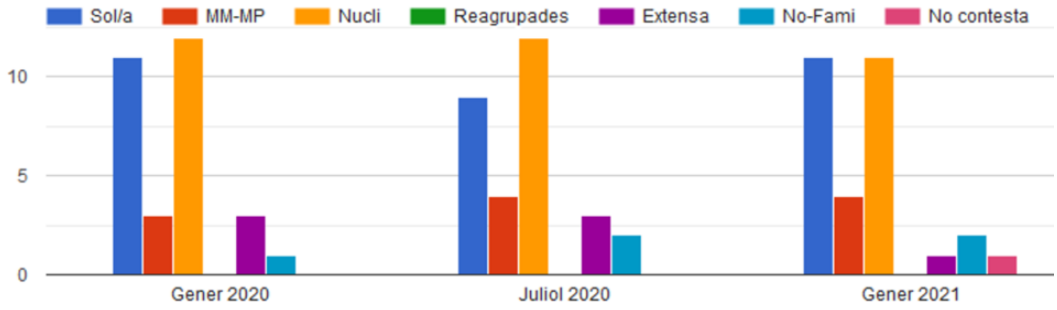

Figura 44: Visualització de les variables bàsiques temporals a Google Forms

Com es pot observar, aquesta descripció no és suficient per extreure tota la informació que aquesta variable pot aportar. Així doncs, per al cas de les variables bàsiques temporals la metodologia MIPRI2D proposa que les variables es descriguin amb les següents tècniques:

- Diagrama de teler

- Taules de freqüència de trajectòria
- Diagrama de barres múltiple.
- Taules de contingència de freqüències absolutes i relatives, com es mostra a la taula 14 i 15
- Una quadrícula de diagrames de sectors.
- Les taules de transicions de taules de transicions.

Per a aquest tipus de variables s'ha desenvolupat un diagrama de teler on cada individu

5.12.6. Variables TQQ

Soc1. La teva participació en activitats de la comunitat al gener del 2020 era: Soc2. Com ha variat aquesta participació en activitats de la comunitat al juliol del 2020? Soc3. Com creus que serà aquesta participació en activitats de la comunitat al gener del 2021? Aquesta és una pregunta del qüestionari que genera una variable de tipus TQQ, com les descrites a la secció 5.3.7. El participant respon l'esquema de convivencia que presenta a cada moment del temps a l'enquesta i la base de dades generada per Google es mostra de la següent manera:

Soc1. La teva participació en activitats de la comunitat al gener del 2020 era: [1. Associacions culturals o esportives]	Soc1. La teva participació en activitats de la comunitat al gener del 2020 era: [2. Xarxes de veïns, AFAs, etc...]	Soc1. La teva participació en activitats de la comunitat al gener del 2020 era: [3. Voluntariat]	Soc1. La teva participació en activitats de la comunitat al gener del 2020 era: [4. Altres]	Soc2. Com ha var aquesta participac activitats de la co al juliol del 2020? Associacions cult esportives]
Lleugerament implicat	Lleugerament implicat	Lleugerament implicat	Gens implicat	Lleugerament imp
Lleugerament implicat	Lleugerament implicat	Lleugerament implicat	Lleugerament implicat	Lleugerament imp
Gens implicat	Gens implicat	Molt implicat	Lleugerament implicat	Lleugerament imp
Lleugerament implicat	Gens implicat	Lleugerament implicat	No contesta	Lleugerament imp

Figura 45: Visualtizacio d'una variable TQQ a la base de dades generada per Google Forms

Així mateix, Google forms mostra el resultat en el següent gràfic:

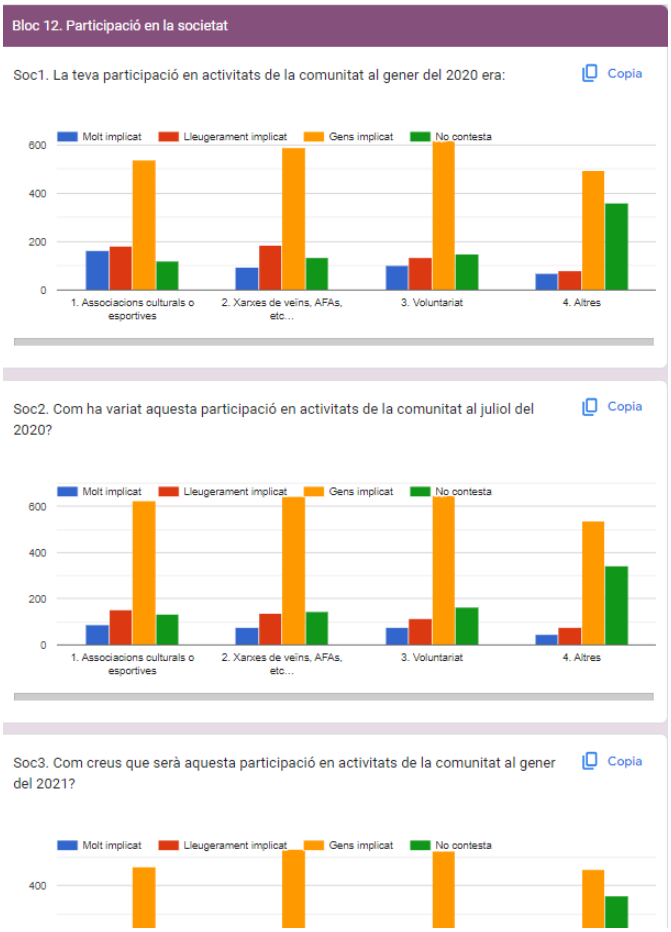

Figura 46: Visualització de les variables bàsiques temporals a Google Forms

Com es pot observar, aquesta descripció no és suficient per extreure tota la informació que aquesta variable pot aportar. Així doncs, per al cas de les TQQ la metodologia MIPRI2D proposa que les variables es descriguin amb les següents tècniques:

- Un diagrama de barres apilades múltiple

Aquest tipus de variables, alhora es poden entendre com un conjunt de variables bàsiques temporals, així doncs, per cada una de les variables s'apliquen els mètodes descrits per a les variables bàsiques temporals. Una quadrícula de diagrames de sectors

5.12.7. Anàlisi de preguntes obertes mitjançant PLN

L5. Perquè? Aquesta és una pregunta del qüestionari que genera una variable de oberta. El participant respon de forma oberta les possibilitats de trobar feina en un any. i la base de dades generada per Google es mostra de la següent manera:

L5. Per què?
RES
o
La meva formació i les meves capa
Impossibilitat d'horaris
tinc experiència suficient
RES
Res
Por la burocracia administrativa, si
Res

Figura 47: Visualtizacio d'una variable TQQ a la base de dades generada per Google Forms

Així mateix, Google forms mostra el resultat de la següent forma:

Figura 48 Visualització de les variables TQQ a Google Forms

Com es pot observar, aquesta descripció no és suficient per extreure tota la informació que aquesta variable pot aportar. Així doncs, per al cas de les preguntes obertes la metodologia MIPRI2D proposa que les variables es descriguin amb un núvol de paraules tal i com es pot veure a la figura 49

Figura 49: Núvol de punts

5.12.8. Estudis específics

En el cas que es desitgin realitzar estudis específics per a sectors de la població hi ha 3 possibles maneres de condicionar. Es tracta de generar un informe o diferents informes amb totes les variables descrites usant la descripció que correspon a cada tipus de variables seguint els mètodes anteriorment exposats.

- Un únic informe per a una subpoblació: Aquest informe estarà condicionat a què els individus compleixin una certa condició, per exemple, persones grans de la província de Girona.
- Comparativa entre les diferents modalitats: Aquest informe, però contindrà una comparativa per totes les preguntes en funció de les modalitats d'una variable categòrica determinada.
- Diferents informes independents en funció d'una variable: En aquest cas, es generen múltiples informes, un per cada una de les modalitats, però són informes que són independents un de l'altre. Per exemple, la realització d'informes per Vegueries. S'han generat 8 informes, un per cada Vegueria, que no depenen l'un de l'altre.

5.13. Reporting automàtic

La clau per a obtenir una resposta ràpida i, en conseqüència, un suport ràpid per a la presa de decisions és disposar de la infraestructura tecnològica preparada per a recopilar dades, així com analitzar les dades tan aviat com es tanqui el període de recollida.

Les dades arriben al qüestionari en línia automàticament tan aviat com els participants proporcionen les seves respostes sense intervenció addicional de l'equip de recerca, a part de garantir la disponibilitat permanent del servidor.

En qualsevol moment, les dades es poden descarregar des del qüestionari en línia en forma d'un fitxer CSV, de manera que es poden tractar diverses ones per formar un panell continu si es torna a demanar.

El contingut del fitxer csv representa les diverses preguntes del qüestionari que segueixen els formats descrits a la figura 25 d'acord amb el tipus de les variables que representen les diferents preguntes.

Donat un cert qüestionari, un fitxer de metadades es pot enllaçar amb ell, indicant quin tipus correspon a cada variable, i quines columnes contenen la informació relativa a aquesta variable en el csv.

Cada qüestionari necessita el seu propi fitxer de metadades. Canviar el qüestionari és relativament simple, de manera que les modificacions en el qüestionari digital corresponent es poden fer fàcilment, i el fitxer de metadades corresponent s'ha de modificar en conseqüència.

L'anàlisi de les dades recollides en el qüestionari es processa automàticament a través d'alguns scripts R i Rmarkdown, que introdueixen tant el conjunt de dades en format csv com el fitxer de metadades corresponent.

També s'implementa un component de coneixement, de manera que els procediments saben en cada moment quin tipus d'anàlisi s'apropia per a cada variable, segons el seu tipus. Això dóna la intel·ligència al sistema i és capaç de gestionar excepcions. A més, es pot modificar per afegir nous tipus de dades, incloent-hi altres eines d'anàlisi quan sigui necessari. Aquest component és el que inclou totes les directrius que garanteixen la preservació del secret estadístic davant de petites mostres esmentades en seccions anteriors.

A més, una part molt important del procediment és que Rmarkdown ha estat dissenyat per a la presentació automàtica d'informes de manera que produeixi un document de Word formatat amb els resultats. Així, el resultat de l'anàlisi és un fitxer de Word editable preparat per ser llegit, comentat i postprocessat de manera molt fàcil pel mateix responsable de la presa de decisions, només requerint una experiència de domini específica per seleccionar els resultats rellevants, per afegir explicacions complementàries per als resultats analítics, per sintetitzar els resultats en una breu visió general o per reordenar-los en un racional que tingui sentit per a la comunicació de resultats.

Quan l'anàlisi s'ha de repetir periòdicament (cada sis mesos, per exemple), el sistema també està preparat per afegir els criteris de reordenació i selecció a la part d'informes automàtics, produint així un document de resultats molt més proper al que l'expert necessita per comunicar els resultats.

Com s'ha dit abans, el qüestionari INSESS-COVID19 està generant un fitxer csv amb 195 columnes que representen 25 blocs d'informació. Algunes de les variables es divideixen en moltes columnes per representació interna, com s'ha explicat abans. El total de temps transcorregut que s'ha de descarregar el fitxer csv del qüestionari (situat al servidor) i obtenir el fitxer Word que conté els resultats de l'anàlisi mitjançant l'ús dels scripts dissenyats en el projecte és d'aproximadament 15 minuts de mitjana. I l'aspecte del document obtingut és molt pròxim a un informe final, com es pot veure en la Figura 50

Després d'aplicar aquesta metodologia, l'informe automàtic genera un document Word, amb una disposició completa i tota la informació numèrica i gràfica associada als tipus de dades simples i complexos descrits (fig1).

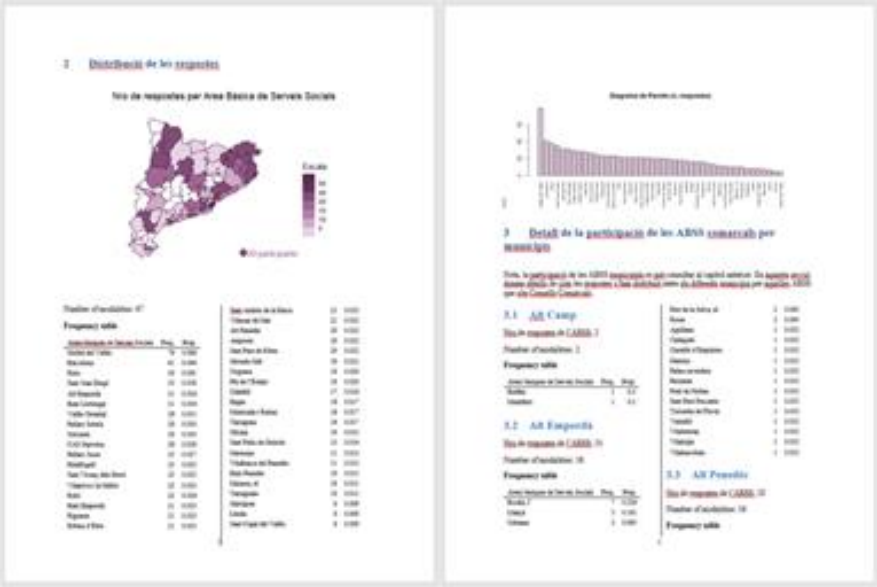

Figura 50. Informe automàtic generat

Com s'ha dit abans, les anàlisis s'adapten al tipus de variable i el procediment R troba informació rellevant en el conjunt de dades d'acord amb tota la informació representada en MdM i algunes regles per procedir en cada cas.

5.14. Eines de suport a la interpretació de clases i noves variables

5.14.1. Termòmetre

A [Canudes, 2016] s'introdueix l'ús del termòmetre com una eina d'adquisició de coneixement expert relacionat amb la interpretació de la polaritat semàntica de cada variable. Es proposa un model de termòmetres en el qual el rang de les variables quantitatives és dividit en 3 intervals de manera que el primer interval es correspon amb els valors més pròxims als mínims de la variable, el segon interval per als valors intermedis i el tercer interval amb els valors més pròxims als màxims de la variable. L'usuari transmet al sistema a través d'aquesta eina visual en quin sentit han d'interpretar-se els valors extrems de les variables numèriques. En línia amb els treballs d'automatització del TLP a [Gibert, Conti & Vrecko, 2012], [Gibert, Conti & Sànchez-Marrè, 2012], a [Canudes, 2016] es proposa un model per a transferir la semàntica de les variables expressada en un termòmetre al quadre semàfor amb la possibilitat d'assignar el color verd o vermell del quadre semàfor tant al primer com a l'últim interval, d'acord amb el criteri de l'expert. No obstant en aquesta tesi s'ha fet una contribució a nivell de formalització d'aquesta eina com es veurà més endavant.

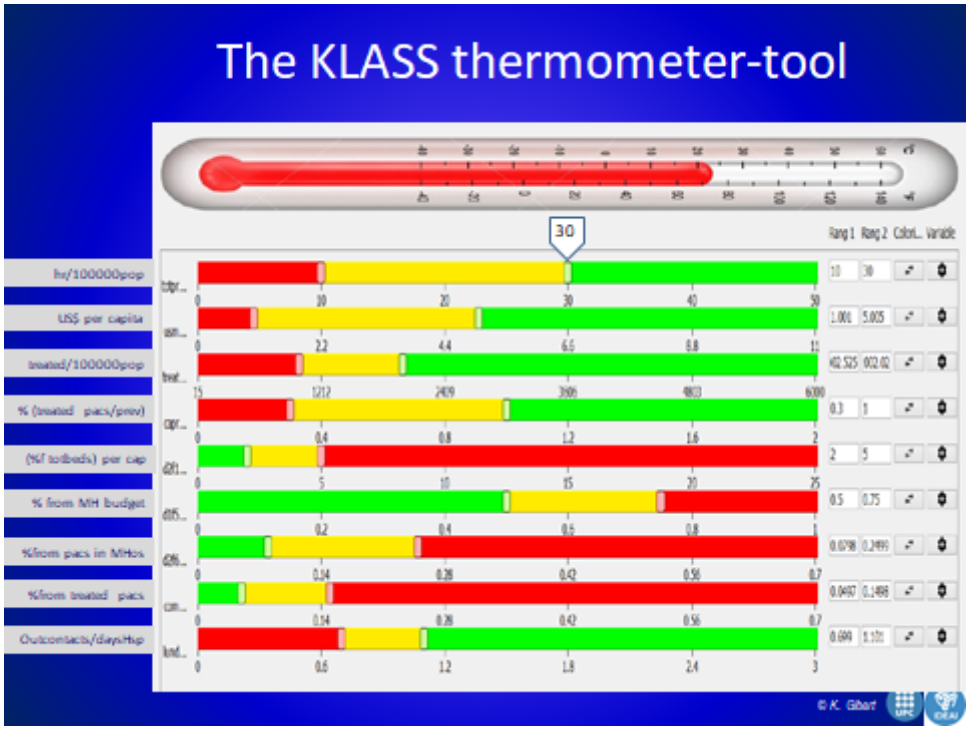

Figura 51: Exemple de termòmetre

A [Gibert, Conti & Sànchez-Marrè, 2012] Es presenten dues metodologies per construir automàticament el TLP. Aquests mètodes es basen en diferents tendències centrals Estadístiques de variables dins de la classe. En aquest apartat es mostra un nou mètode basat en el coneixement expert.

El termòmetre (T) és una eina d'adquisició de coneixement que permet representar la semàntica associada a una variable d'una manera formal, de manera que es pot injectar en altres mètodes d'anàlisi de dades. A T representa una abstracció simbòlica d'aquesta semàntica d'acord amb dos principis principals:

Hi ha un concepte de referència latent que pot guiar l'avaluació dels valors variables com a promoció o no dels individus pel que fa a aquest concepte latent (és a dir, qualitat de l'aigua, bondat d'un sistema d'atenció, availabitament de serveis...). Aquest concepte latent de referència està alineat amb els objectius de l'anàlisi

Un conjunt de colors del semàfor {vermell (r), groc (y), verd (g)} s'associarà amb la semàntica dels valors de les variables segons el concepte de referència latent de tot ell. Per exemple, les variables que indiquen aigua bruta s'associaran al vermell en problemes de qualitat de l'aigua i l'aigua neta associada al verd. A més, la viola s'utilitzarà per als valors que manquen.

La representació formal d'un termòmetre es descriu en el següent:

- $\mathcal{I}$ és un conjunt d'individus $\mathcal{I}= \{i_1 \dots i_n\}$ descrits per les variables K $\{X_1,X_2,...,X_j,...X_K\}$
- $D_k=\{m_1, m_2, m_3,..., m_k\}$ és el conjunt de modalitats per a una variable qualitativa X_k

- $T = \{t_1, t_2, t_3, .. ., t_K\}$ és el panell de termòmetre disponible, on t_k, $k \in \{1 : K\}$ és el termòmetre de la variable $X_k \in K$. Quan X_k és qualitatiu, $t_k=\{(m_1;q_1),(m_2,q_2),...,(m_k,q_k)\}$ on:
    - $m \in D_k$ és una modalitat de la variable X_k
    - q_k és el color assignat a m_k.
- Quan X_k és quantitatiu, $t_k = \{r_1, r_2, o\}$, on:
    - r_1 és un valor numèric per a X_k, tal que $\min(X_k) , r_1) \max(X_k)$
    - r_2 és un valor numèric per a X_k, tal que, $r_1 , r_2 \leq \max(X_k)$
    - o és la polaritat semàntica de la variable ($o \in$ {direct, inverse}). Representa una associació directa dels significats variables amb els colors dels quadre semàfors o l'invers (valors alts de variables numèriques poden enllaçar al vermell si mesuren contaminants d'aigua, o al verd si mesuren, per exemple, biodiversitat en problemes de qualitat d'aigua).

VARIABLES NUMÈRIQUES

El termòmetre d'una variable numèrica indica els punts de tall d'un edifici de tall tres intervals en el rang variable, que s'associaran a tres zones de color, d'acord amb els coneixements dels experts anteriors i els principis del termòmetre.

Per convenció, a causa de la forta relació que aquest model visual manté amb els quadre semàfors, només es permeten un màxim de tres zones amb els colors vermell, groc i verd i es mereix violeta dels valors que falten. Aquest model restringit en tres colors està orientat a aprofitar tots els codis d'interpretació implícits associats amb quadre semàfors, de manera que els conceptes d'inducció dels resultats de l'anàlisi estiguin autoritzats. Les zones de color estan disposades de manera que:

Primera zona de color: des de $\min(X_k)$ a r_1

Segona zona de color: de r_1 a r_2.

Tercera zona de color: de r_2 a $\max(X_k)$

La idea és que l'expert pot indicar en T tres àrees on la semàntica de la variable canvia de baixa, normal i alta, i els punts de tall r_1 i r_2 determinen els valors on la variable canvia de significat

En aquesta primera estructura, una segona capa semàntica de la variable es transfereix al sistema en un segon component: el color, que associa el verd amb la banda més benevolent en termes de la interpretació de la variable i el vermell amb la menys benevolent, sempre d'acord amb l'expert en el tema estudiat.

L'associació entre colors (verd, groc, vermell) i nivells qualitatius (baix, mitjà, alt) es pot fer de dues maneres:

- Directe: valors baixos vermells i alts verds
- Inversa: valors verds baixos i valors vermells alts

Determinar l'associació de color directe o inversa depèn de la semàntica de la variable i els seus valors i està completament estirada al costat expert. Mantenir l'expert en el bucle es torna crític per a la construcció del termòmetre.

DISSENY DEL TERMÒMETRE

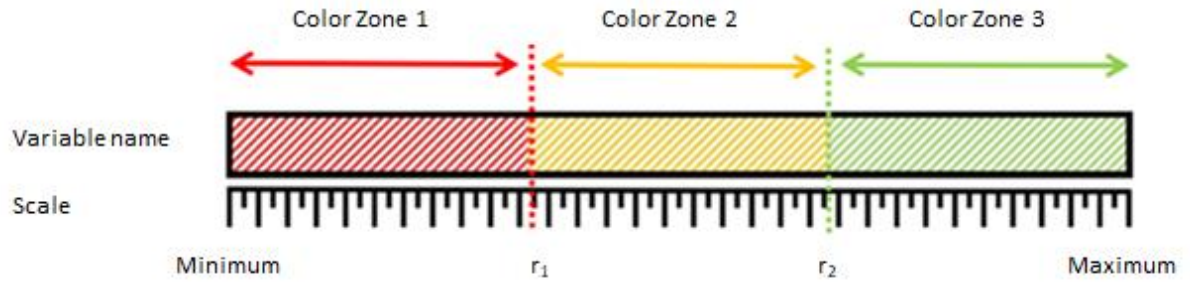

Figura 52:. Disseny de termòmetre variable numèric.

Figura 52 visualitza un t_k a partir d'una variable numèrica. Mostra els elements següents:

- Nom de la variable: nom de la variable quantitativa representada en el termòmetre.
- Mínim: valor mínim de X_k observat a la mostra
- Màxim: valor màxim de X_k observat a la mostra
- Escala: eix graduat amb els valors possibles de X_k. r_1: límit superior de la primera zona de color.
- r_2: límit inferior de la zona del tercer color

Quan o és directe, el verd és a la Tercera Zona, si és Invers és a la Primera Zona.

El resultat obtingut mitjançant aquesta tècnica és un model visual molt potent que proporciona molta informació al sistema i als agents profans en la naturalesa danyada de les dades, a través d'una representació molt visual, intuïtiva i intel·ligible.

A causa de la seva forma compacta, en un espai petit, els termòmetres de diverses variables es poden mostrar junts com files d'un sol panell que, a un cop d'ull molt ràpid, permet a l'expert entendre la conceptualització de les variables d'una manera molt intuïtiva. El termòmetre és una eina tècnica que recull la semàntica de la variable i facilita la creació d'una nova variable qualitativa que captura la semàntica associada als valors de les variables baixes, mitjanes o altes a través de colors simbòlics vermell (r), groc (y), verd (g) i violeta (v). Es pot

utilitzar molt sovint com una eina auxiliar per produir automàticament la interpretació dels resultats de l'anàlisi de dades.

VARIABLES QUALITATIVES

Quan la variable és qualitativa, el termòmetre determina una recodificació directa entre les modalitats originals de la variable i els colors simbòlics.

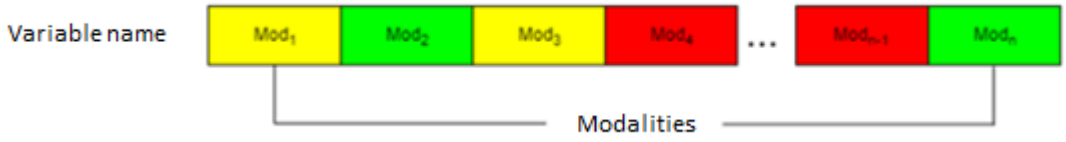

Figura 53. Disseny de termòmetre variable qualitativa.

Com que no hi ha zones de tall, els termòmetres per a les variables qualitatives s'han dissenyat de manera que totes les modalitats poden prendre qualsevol dels 3 colors, d'aquesta manera es garanteix que l'expert pot assignar una semàntica a cada modalitat per separat, i se li permet assigna el mateix color a diverses modalitats. Per a les variables qualitatives ordinals, les modalitats s'han de representar amb l'ordre correcte i els colors assignats seguint aquest ordre (Vegeu Figura 53).

5.14.2. Quadre semàfor basat en Termòmetre

A [Avila Montalvo, 2018] es proposa la integració del quadre semàfor i el termòmetre en un primer procediment innovador de construcció de quadre semàfors basats en la semàntica que els termòmetres aporten a les classes. El treball de [Avila Montalvo, 2018] és un primer intent de construir una eina on l'usuari transfereix al sistema la semàntica de les variables a través del termòmetre i a través d'un procediment intern es calcula com es distribueixen els nivells qualitatius induïts pel termòmetre i la semàntica que representa sobre els quadres semàfors en les classes. El nivell dominant per cada classe determinarà el color de la casella i la semàntica farà referència a la variable latent de referència que s'ha utilitzat per definir els colors del termòmetre. Els quadre semàfors basats en termòmetres admeten també el corresponent procés d'anotació i poden presentar-se sobre colors bàsics o amb la degradació del model de color descrit a l'apartat 2.7.2. En aquesta tesi es continua treballant en el perfeccionament d'aquesta eina com es veurà més endavant.

La metodologia original s'actualitza per incorporar la informació semàntica proporcionada al termòmetre. Sigui $\{X_1, X_2, X_3, ..., X_K\}$ el conjunt qualitatiu o quantitatiu de variables que es representaran en un TLP, T el termòmetre disponible per al mateix conjunt de variables, i P la variable de classe, que és la variable categòrica de destinació que s'explicarà en el TLP. La generació del TLP basat en T es compon del vector següent:

1. Fase de discretització / recodificació: Creeu Z_k una nova variable qualitativa resultant de la recodificació o discretització de X_K segons el seu tipus original i la informació del termòmetre:
   1.1. Si X_k és quantitatiu, crea Z_k=dis(X_k, t_k), discretitzant X_k segons els valors de tall indicats en el termòmetre i els colors associats.
   1.2. Si X_k és qualitatiu, creeu Z_k=rec(X_k, t_k) recodificant X_k segons els colors donats en el termòmetre a cada modalitat
   1.3. Si t_k ;T llavors assigna groc a tots els valors de X_k ; l'usuari pot editar manualment

   Discretització (X_k :quantitatiu):

Per a tot $i \in [1 : n]$, sent xi el valor de X_k de l'individu i el valor z_i de Z_k és:

- Si x_i és un valor vàlid de X_k
  - Si és $x_i \leq r1$
    - Si o=directe llavors z_i=”r”
    - Si o=invers llavors z_i=”g”
  - Si ($x_i > r_1$) ($x_i \leq r_2$) →z_i=”y”
  - Si és $x_i > r_2$
    - Si o=directe, llavors z_i=”g”
    - Si o=invers llavors z_i=”r”
  - Si falta el x_i, llavors z_n=”v”

  Recodificació (X_k :qualitatiu):

  Per a tot i, [1: n], sent xi el valor de Xk de l'individu i el valor z_i de Z_k és:

  Sent $t_k=\{m_m,q_m\}_{m=1:nk}$

- Si ($x_i=m_m$) llavors $z_i=q_m$
- Si falta el x_i, llavors z_i=”v”

2. Fase de creació de matrius creuades

$$M_k = P\ x\ Z = \begin{bmatrix} n_{11} & n_{12} & n_{13} & n_{14} \\ n_{21} & n_{22} & n_{23} & n_{24} \\ n_{31} & n_{32} & n_{33} & n_{34} \\ \vdots & & & \\ n_{c1} & \cdots & n_{cq} & \cdots \end{bmatrix} \qquad (10)$$

On D_Z ={r,g,y,v} són els colors associats a les columnes M_K i donats c∈P i q∈D_Z, l'element n_{cq} és el nombre d'individus de classe c amb Z_k = q .

$$n_c = \sum_{q=1}^{4} n_{cq} \, , N = \sum_{\forall c \in P} n_c \qquad (11)$$

3. Fase d'assignació de colors. Sigui $F_c \in M_k$ una fila de la matriu anterior. El color de la cèl·lula es denota S_C i s'expressa de la manera següent.
   - Variables qualitatives binàries
     - Si argmax(F_c) = 4 LLAVORS S_c=v
     - Si argmax(F_c) = 3 LLAVORS S_c=y
     - Si argmax(F_c) = 1
       - Si $n_{c1}/n_c \geq \gamma$ LLAVORS S_c=r
       - Si $n_{c1}/n_c < \gamma$ LLAVORS S_c=y
     - Si argmax(F_c) = 2
       - Si $n_{c2}/n_c \geq \gamma$ LLAVORS S_c=g
       - Si $n_{c2}/n_c < \gamma$ LLAVORS S_c=y

On $\gamma \in [0,1]$ i determina la proporció del llindar d'una modalitat que es considerarà no de color groc. Això és necessari perquè les variables binàries tenen només dues modalitats i representen una dicotomia bàsica i en una classe el nombre d'elements vermells o verds s'ha de transformar en un sol color de la cela. El valor per defecte per . hauria de ser 0.5, de manera que més del 50% dels elements d'una classe amb color verd determinarien una cel·la verda del TLP. El paràmetre γ permet més flexibilitat per tractar falsos negatius positius i falsos i proporciona la possibilitat de mantenir l'assignació de groc a una cel·la fins a una proporció més alta de verd (per exemple, γ =0,7, l'algoritme no assignarà verd si la classe té menys del 70% dels elements verds)

- Altres variables
  - Si card[argmax(F_c)] = 1 LLAVORS $q_{argmax(F_c)}$
  - Si card[argmax(F_c)]>1Λ (argmax(F_c)=3) LLAVORS S_C=y
  - Si $n_{c*}/n_c > \gamma$ LLAVORS S_C=v

On $q_{argmax(F_c)}$ és el color assignat segons la posició de q a Dz={r,g,y,v}

5.15. Ampliació del preprocessament amb tècniques de Generació de variables derivades

Els models basats en dades es construeixen normalment amb les variables originalment obtingudes del conjunt de dades, de vegades arranjats durant el procés de neteja de dades en transformacions bàsiques.

No obstant això, quan noves variables rellevants es poden definir com una combinació de variables originals, el valor afegit apareix en més processos de mineria de dades. Les noves variables podrien ser generades per diferents mètodes. En aquest document es proposen 2 mecanismes:

Creació basada en el coneixement de noves variables de segona generació: afegeixen nous conceptes a la base de dades que s'acosten als paràmetres de raonament experts

Creació de variables basades en dades de 3a generació d'indicadors nous: sintetitzen blocs de variables temàtiques en indicadors individuals mitjançant l'ús de tècniques de modelatge basades en dades.

Així, noves dimensions (normalment corresponents a combinacions no lineals de variables originals) podrien ser interpretades per l'expert utilitzant el seu univers conceptual. Una vegada que es creen diverses variables, aquest procés acaba i el procés de modelització proporciona resultats més fàcils d'interpretar i proporcionar una perspectiva més profunda de les dades analitzades, com es veurà a la secció de resultats (5).

5.15.1. Variables de segona generació basades en el coneixement

La idea és utilitzar el coneixement de domini específic proporcionat per experts per construir noves variables com una combinació (sovint no lineal) de les variables originals disponibles al conjunt de dades. Les noves variables basades en el coneixement s'aproximen al raonament fet per l'expert, i com a conseqüència, el modelatge addicional que inclou aquestes variables resulta en una interpretació més fàcil per l'expert, ja que els models s'expressen en termes que ell/ella utilitza en el seu raonament natural. La idea de crear noves variables basades en el coneixement, materialitzant els coneixements de l'expert es va presentar originalment a [Gibert, Sanchez-Marre & Izquierdo, 2019]. No obstant això, les tècniques descrites a [Gibert, Sanchez-Marre & Izquierdo, 2019] s'utilitzen en diverses obres com [Torres, Hernan & Janeth 2009] [Vergara et al. 2016]. Es coneix la idea que les dades per si soles no són suficients per revelar la complexitat dels fenòmens reals i diversos autors van cridar l'atenció sobre aquest tema, a [1] la proposta és utilitzar el coneixement de domini específic per combinar diversos mètodes d'extrapolació i predicció per entendre les dades recollides per a molts camps d'aplicació com la salut o les finances. En [Ahlemeyer-Stubbe & Agnes, 2021] els autors reclamen la importància d'utilitzar el coneixement del domini al llarg del procés analític, des de la primera etapa de quines preguntes fer, fins a cadascun dels passos del preprocessament per construir models predictius més precisos i robusts i així obtenir millors coneixements. No obstant això, en el bloc de preprocessament no aborda l'important pas de crear noves variables, la qual cosa abordem en aquest treball. En [Ahlemeyer-Stubbe & Agnes, 2022] també s'afirma que el coneixement de domini és essencial en processos científics de dades.

Tot i que el paper se centra en les aplicacions de màrqueting en aquest document, explorem la importància d'una guia específica basada en el coneixement per crear noves variables abans de l'anàlisi des d'una perspectiva transversal no específica per a un camp d'aplicació particular.

De fet, sovint els experts pensen en termes de transformacions complexes de dades originals en algunes noves variables que representen conceptes que utilitzen per raonar o per avaluar un escenari determinat (per exemple, avaluar si una persona va debutar en un trastorn de salut mental basat en una avaluació integral d'un conjunt de variables binàries que indiquen si la persona pateix o no un trastorn particular abans del confinament i després d'ell, o raonar sobre el nombre total de trastorns mentals patits per una persona en lloc de raonar sobre la llista de variables binàries que indiquen l'existència o no de cada trastorn mental separat). Aquesta nova generació de variables consumeix el coneixement que els experts tenen sobre les dades, i pot crear moltes variables diferents fent referència a conceptes nous que no estan presents en les dades originals, útil per a una anàlisi més detallada.

En aquest document es considera un conjunt reduït de mecanismes per crear aquest tipus de variables de segona generació, totes basades en l'ús del coneixement de domini específic dels experts per combinar diverses variables originals de maneres específiques indicades per ells amb l'objectiu de representar conceptes més complexos que puguin millorar els models de mineria de dades. Considerem tres mecanismes:

- **Indicadors:** Una variable X amb modalitats $Dx=\{m_1, m_2, m_3, ..., m_X\}$ es converteix en un conjunt de variables binàries X'_m ,$m \in D_x$, tal que $D_{xm}=D$, amb $D=\{m'_1$SI, m'_2=NO}. La regla de transformació és: si X=m, ,$m \in D_x$ llavors X'_m=YES; en cas contrari X'_m=NO.
- **Condicions multivariable**s: Donades un conjunt de variables X_1 ... X_K amb modalitats m_X en D_x es crea una nova variable basada en l'avaluació d'una funció booleana f construïda sobre les variables originals. La transformació genera una nova variable X' segons la següent regla de transformació: si $f(X_1..... X_K)$ = CERT després X'=SÍ; en cas contrari X'=No.
- **Comptadors**: Quan un paquet de variables es refereix al mateix concepte (és a dir, diversos símptomes d'una malaltia o diversos trastorns mentals), es pot crear una nova variable agregada, per exemple, comptant el nombre de valors positius en tot el paquet. Donat un conjunt de variables binàries ($X_1..... X_K$) variables amb les mateixes modalitats, $D_X=D$, amb $D=\{m_1, m_2, m_3, ..., m\}$, es crea una nova variable numèrica X' com a recompte de les variables en ($X_1..... X_K$) apuntant a un cert subconjunt de valors de referència $A \subset D$.La regla de transformació és $X' = card\{Xm: X_m \in A\}_{m \in \{1:K\}}$.

Aquests són només tres mecanismes bàsics per a la generació de noves variables derivades basades en experts, però hi ha molts més que es poden utilitzar en projectes reals.

5.15.2. Variables de segona generació basats en dades (DD2gI)*

A [Angerri & Gibert, 2023] es presenten diverses maneres de construir nous indicadors amb les dades originals. Aquests mètodes utilitzaven fórmules matemàtiques per a ser construïdes. En [Angerri & Gibert, 2023] la metodologia es va basar en l'ús de coneixement de domini específic proporcionat pels experts per construir noves variables com una combinació de les variables originals. En aquest document, s'introdueix una metodologia addicional per construir una variable de segona generació basades en dades, mentre que la metodologia original presentada en treballs anteriors es coneixerà a partir d'ara com a variable de segona generació basada en el coneixement.

En diverses bases de dades, les variables temporals són presents i en general es representen com diverses columnes del conjunt de dades, cadascuna amb la rèplica de la variable de destinació en diferents marques de temps. En altres ocasions, algunes variables estan fortament relacionades entre si perquè responen al mateix concepte (com un conjunt de variables per a diferents tipus de subsidis socials representats en un conjunt de variables binàries). Si aquestes variables han de ser entrades d'un procés de clustering, el risc de biaix dels resultats és alt, ja que el mateix "concepte" es representa amb més dimensions en el conjunt de dades i el pes d'aquest concepte en la formació de cúmuls augmentaria. Per evitar aquest tipus de biaix, es crea una variable de segona generació basades en dades:

1. Seleccioneu les variables compents que s'han de sintetitzar en una única variable basada en dades: els experts haurien d'identificar els subconjunts de variables en aquesta situació i determinar quines variables seràn components de la nova variable.
4. Clusterització de les variables de components seleccionats: Utilitzant el mètode de Ward amb Mètriques Mixtes Gibert les variables seleccionades són agrupades i una nova variable de classe obtinguda P. En la seva forma original (on les classes reben un identificador numèric), la variable no és interpretable per ella mateixa i es requereix un postprocessament per interpretar els cúmuls, aconseguint etiquetes representatives per a modalitats més comprensibles. Això s'adreça en els següents passos.
5. Creació de CPG: construeix un CPG a partir de tots els components escollits en 1 vs. P identificats en 2.
6. Creació de termòmetres per a les variables de component seleccionades en 1: Juntament amb l'expert en el camp, d'acord amb \S 2.6.1.
7. Creació de TLP basat en termòmetre: utilitzant la metodologia descrita en \S 2.6.2.
8. Crea el nou indicador interpretat $\mathcal{P}$: mitjançant l'etiquetatge de les modalitats de P d'acord amb la informació proporcionada per TLP/aTLP i la representació conjunta de TLP/aTLP sobre CPG. Aquestes eines mostren les particularitats de les diferents variables

en cada classe, de manera que els experts del domini poden induir etiquetes adequades a totes les classes de P, segons les seves característiques principals i resumint el concepte principal darrere de cada classe. Hi ha una relació bijectiva entre P i $\mathcal{P}$. Gràcies a la TLP i el termòmetre, la variable de classe es converteix en una nova variable qualitativa interpretable amb modalitats amb significat semàntic.

9. Nom $\mathcal{P}$: Associeu una etiqueta a la variable (sovint el nom del concepte) i una descripció de la mateixa variable i cadascuna de les seves modalitats per corregir la interpretació.
10. Afegir $\mathcal{P}$ a la base de dades general: ampliar el conjunt de variables amb aquesta nova variable. $\mathcal{P}$ P esdevé una nova columna del conjunt de dades que indica un cúmul etiquetat per a cada individu amb l'estructura d'una variable qualitativa ordinària.

5.15.3. Variables de tercera generació basades en dades

En aquest cas, noves variables es construeixen utilitzant tècniques auxiliars d'anàlisi multivariant. Això és útil per sintetitzar resultats d'un subconjunt de variables en una nova variable. Aquestes variables es podrien obtenir usant diferents models basats en dades com PCA, clustering, etc. En aquest document s'utilitza l'agrupació. Les variables usades per a aquests models haurien de correspondre a un mateix tema (o a un mateix bloc d'un qüestionari) i les noves variables sintetitzen la informació rellevant de tot el paquet en un únic (o un nombre reduït de) noves variables.

En el nostre cas, estem utilitzant el clustering per sintetitzar la informació d'un cert bloc de variables en un únic perquè els cúmuls resultants identifiquen perfils (en termes de patrons d'una certa combinació de valors per a un determinat subconjunt de variables seguits pels individus que pertanyen al perfil) que poden ser associats a individus. Per tant, la variable de classe resultant coincideix amb l'estructura requerida per a una nova columna d'un conjunt de dades. En [36] es proporcionen criteris per escollir el mètode de mineria de dades adequat per resoldre un problema determinat i aquest treball era la nostra referència per escollir l'ús de clustering aquí. De fet, la variable de classe resultant és una nova variable qualitativa, és a dir, un nou indicador, que té els valors degudament etiquetats i es pot afegir fàcilment al conjunt de dades original. Transformar la variable de classe en una nova variable qualitativa del conjunt de dades significa que la interpretació de les classes i un etiquetatge addicional és fonamental per obtenir un conjunt D significatiu de modalitats variables. El procés proposat és:

Determinar un tema objectiu, entre els temes representats pel conjunt de dades original (eventualment, també es pot utilitzar el conjunt de dades ampliat, incloses les variables de segona generació, atès que sovint representaran combinacions no lineals de variables originals i l'agrupació no requereix independència)

- Selecció de variables: Seleccioneu un subconjunt de variables pel que fa al tema de destinació. Es convertiran en els components del nou indicador (és a dir, un bloc de variables sanitàries o un bloc de variables econòmiques)

- Clistering de variables: Utilitzant el mètode de Ward amb Mètriques Mixtes Gibert, les variables seleccionades s'agrupen i es produeixen noves variables de classe. En la seva forma original (on les classes rebien un identificador numèric), la variable no és interpretable per si mateixa i es requereix un postprocessament per interpretar els clústers, aconseguint etiquetes representatives per a cada clúster, de manera que la variable de classe esdevé una nova variable qualitativa amb modalitats significatives. Això s'adreça en els següents passos.

- Creació de CPG: Creeu el CPG a partir de tots els components del nou indicador identificat en el pas 2 versus la nova variable de classe obtinguda en el pas 3.

- Creació del termòmetre: Creació de termòmetres per a les variables de component seleccionades en 1: Juntament amb l'expert en el camp, d'acord

- Crear TLP: Construït el TLP (o l'atLP) a partir de la mateixa estructura del CPG en el pas anterior. Això resumeix la tendència central de cada component en cada grup utilitzant un codi de color com s'indica a la secció 5.14.2

- Modalitats d'etiquetatge del nou indicador: Amb la projecció del TLP/aTLP sobre el CPG, emergeixen les particularitats de les diferents variables en cada classe, de manera que els experts poden induir etiquetes adequades a tots els cúmuls, d'acord amb les seves característiques principals i resumint el concepte principal darrere de cada classe. Hi ha una relació bijectiva entre la variable de classe del pas 3 i el conjunt d'etiquetes creades aquí. Gràcies a la TLP, la variable de classe es converteix en una nova variable qualitativa interpretable amb modalitats amb significat semàntic.

- Etiqueta la nova variable qualitativa: Associa una etiqueta a la variable (sovint el nom del tema) i una descripció de la mateixa variable i cadascuna de les seves modalitats per tal de corregir la interpretació.

Afegir l'indicador basat en dades a la base de dades general. La nova variable qualitativa de tercera generació es converteix en una nova columna del conjunt de dades que indica un cúmul etiquetat per a cada individu.

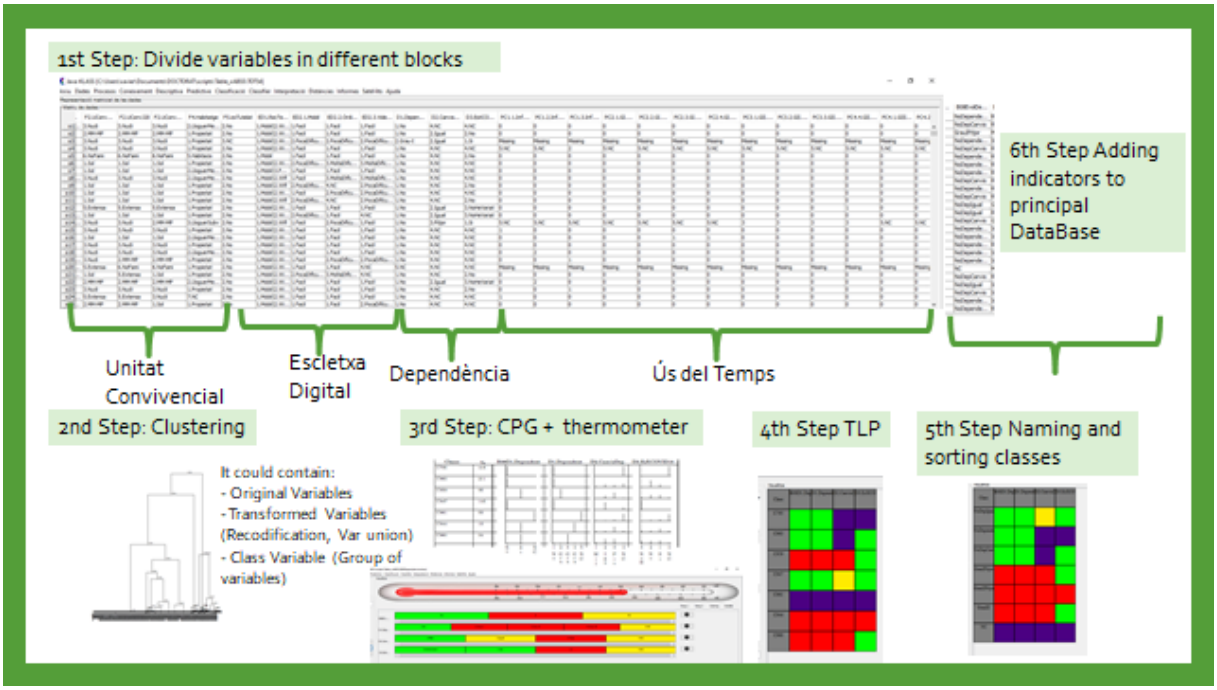

Figura 54 Nou contorn d'indicadors nous basats en dades de 3a generació

5.16. Anàlisi Multivariant.

Un cop preprocessades les dades, és necessari aplicar tècniques de classificació automàtica per estudiar patrons multivariats. Cal analitzar quina serà la tècnica de classificació millor, així com establir la mètrica necessària. Per a la interpretació de patrons es faran servir eines de suport a la interpretació automàtica de les classes resultants del clustering, donat l'objectiu de suport a la presa de decisions del projecte.

Per al projecte en concret, s'ha utilitzat la tècnica de classificació jeràrquica de Ward amb una mètrica mixta (mètrica de Gibert) que permet treballar amb variables numèriques i qualitatives quan és el cas, i amb la distància χ^2 quan només hi ha qualitatives.

Donat que ens enfrontem a qüestionaris de gran complexitat amb un nombre important de variables, algunes d'elles relacionades entre si i agrupades per blocs temàtics, s'ha plantejat una primera proposta de reducció de la dimensionalitat basada en coneixement i que permeti concentrar el valor de les dades en un únic indicador per bloc temàtic. Aquesta proposta requereix més recerca per a que es consolidi definitivament, però s'esbossa en les següents passes:

Per cada bloc temàtic:

1. Realitzar un clustering utilitzant només les variables del bloc temàtic
2. Construir el termòmetre de les variables implicades
3. Realitzar un TLP basat en el termòmetre
4. Interpretar els perfils a partir del TLP

5. Conceptualitzar els perfils i etiquetar les classes

Això col·lapsa totes les variables del bloc temàtic en un únic indicador (que s'anomena com el bloc temàtic) i que recull els diferents perfils que es donen dins el bloc. Al final del procés la Base de dades s'ha reduït a tantes variables com blocs temàtics té el qüestionari (21 en lloc de 190) més les variables del bloc anagràfic (sexe, estat civil, etc)

5.16.1. Mètode de selecció de característiques territorials (TFSM)

Després de trobar els cúmuls d'una visió i interpretar-los, la selecció de característiques ha de fer grups territorials. El focus és trobar quina variable és la millor per representar la vista en el clúster final, de manera que es permetrà una variable per bloc per evitar la sobre representació o tergiversació d'un tema.

En aquest punt cada bloc té el seu propi indicador. És el moment de seleccionar les variables per crear un cúmul general, tenint una variable per bloc. La contribució metodològica per seleccionar la variable apropiada de cada bloc es basa en els valors de prova de Lebart, presentats en 2.1.2. L'objectiu principal d'aquest mètode és obtenir la variable seleccionada per bloc que s'utilitzarà en el clúster global final. Sent $\mathcal{I}$ un conjunt d'individus $\mathcal{I} = \{i_1 \dots i_n\}$descrits per les variables K $\{X_1,X_2,...,X_j,...X_K\}$

Els passos són els següents:

1. Seleccioneu la variable d'ubicació: En aquest cas, la variable territorial s'ha de seleccionar des de la base de dades original. Aquesta és una variable informativa en el conjunt de dades original, que no pertany a cap bloc. La variable territorial XLoc és una variable qualitativa on les modalitats són ubicacions (L) $Dloc=\{l_1, l_2,...,l,...l_L\}$, $Loc \in 1:K$
2. Seleccioneu el bloc: En el pas 1 de Creació de nous indicadors de tercera generació basats en dades s'han creat alguns blocs. Ara s'ha de seleccionar un bloc.
3. Seleccioneu les variables candidates: En aquest moment, el bloc pot contenir variables originals, variables de segona generació i l'indicador de tercera generació basat en dades. Només la variable de tercera generació i les seves variables de components se seleccionen com a variables candidates. Les variables candidates són $\chi=\{X_k \text{ tq } k \in 1:K\}$ amb modalitats $D_k=\{m_1,m_2,...,m,...m_{nk}\}$, m, D_k
4. Calcula un rànquing de variables segons la seva capacitat per explicar la distribució territorial: Avalua les variables candidates per determinar la selecció. Repetiu els passos següents per a cada variable candidata preseleccionada en el pas 3.
   4.1. Calcula els valors de prova de Lebart: utilitzant la metodologia de 2,1.2 tots els valors de prova de Lebart es calculen per a cada variable candidata qualitativa contra *Loc*. Per a cada qualitat de X_k, la sortida d'aquest procés és una taula V_k amb $l \in Dloc$ en files i columnes de n_k, anomenades $M_{m\ (m=1:\ nk)}$, sent v_{lm}, són els valors p del valor de prova de Lebart de $M_m\ versus\ l \in\ D_{Loc}$
   4.2. Crea una nova variable S_k amb $l \in D_{Loc}$ en files

4.3. Calcula la relació d'ubicacions significatives per a la variable X_k utilitzant la informació proporcionada per S_k

$$R_k = \frac{card\{s \in S_k : s > 0\}}{L} \quad (12)$$

Aquest indicador avalua el percentatge d'ubicacions que es poden caracteritzar per algunes modalitats de X_k. Estarem interessats en les variables X_k que proporcionen modalitats significatives a tantes ubicacions com sigui possible, el que significa que la mateixa variable pot explicar una part més gran del territori.

4.4. Donat X_k, i el seu corresponent V_k, creeu la taula Π_k amb l ∈ D_{Loc} en files i columnes de n_k, anomenada $M_{m\ (m=1:\ nk)}$, on π_{lm}, correspon a la probabilitat empírica de X_k condicionada a una ubicació donada per a les cel·les significants:

$$\pi_{lm} = \begin{cases} 0 & v_{lm} > 0{,}05 \\ \dfrac{card_{i\in\mathcal{I}}\{x_{ik} = M_m \ and \ x_{iloc} = l\}}{card_{i\in\mathcal{I}}\{x_{iloc} = l\}} & v_{lm} \leq 0{,}05 \end{cases} \quad (13)$$

Això ajuda a entendre, mentre que les modalitats significatives d'una determinada ubicació cobreixen una gran porció de la població de la ubicació o, al contrari, un grup limitat d'individus. De fet, tenir una modalitat significativa que cobreixi una minoria no és suficient per explicar un clúster (o una ubicació).

4.5. Crea una variable nova $\Pi_{Loc\ *k}$amb *l* ∈ D_{Loc} en files i:

$$\Pi_{l*,k} = \sum_{m=1}^{n_k} \pi_{lm} \quad (14)$$

$\Pi_{Loc*,k} = (\Pi_{1*,k}, \ldots, \Pi_{L*,k})$ indica la proporció d'individus involucrats en ubicacions amb modalitats significatives de X_k segons el valor de prova de Lebart.

4.6. Calcula la mitjana de $\Pi_{l*,k}$ per a totes les ubicacions, de manera que obtenim una estimació de la contribució mitjana de les modalitats significatives de X_k per a la població de la ubicació.

$$\widetilde{\Pi}_{Lock} = \frac{\sum_{l=1}^{L} \Pi_{l*,k}}{card_{l=1:L}\{\Pi_{l*,k} > 0\ \}} \quad (15)$$

Aquest indicador és una estimació de la cobertura mitjana de les modalitats significants de $X_k$al llarg del territori. Per a valors baixos de $\widetilde{\Pi}_{Lock}$, les modalitats significatives de X_k corresponen a una minoria de les diferents ubicacions i la variable no és tan informativa com sigui necessari.

4.7. Calcula l'índex de potencial explicabilitat de la variable X al territori com:

$$E_k = R_k \cdot \widetilde{\Pi}_{Loc,k} \quad (16)$$

L'índex E_k pondera la proporció mitjana d'individus implicats en territoris significatius per a una variable donada per la proporció d'ubicacions significatives en tot el territori. Estem interessats en variables amb capacitat per caracteritzar tantes ubicacions com sigui possible, amb la major cobertura possible. La rellevància de la variable disminueix en el cas que sigui significativa en molts llocs amb poca cobertura o en pocs llocs, tot i que implica un gran nombre d'individus. Aquesta és una correcció important per equilibrar l'impacte en l'anàlisi de les grans ciutats que concentren una gran part de la població com és Barcelona al territori català. E_k es defineix per obtenir robustesa pel que fa a la distribució desequilibrada de la població en el territori o les modalitats que assenyalen a grups excepcionals d'individus amb baixa presència. Això evita que les grans ciutats dominin tota l'anàlisi i també les minories excepcionals. En calcular el valor E_kd'per a totes les variables candidates, es pot construir un rànquing pel que fa al potencial de X_k per explicar la distribució territorial.

5. Selecció de variables: per a cada bloc, la variable amb el màxim E_k està seleccionada per representar el bloc en una anàlisi posterior

5.17. Fase VI Perfilat intel·ligent de les classes

5.17.1. Descripció i identificació de patrons de comportament de cada grup mitjançant eines de suport a la interpretació de classes

1. Clustering de les variables seleccionades: El conjunt de variables seleccionades en el pas 5 tenen una variable representativa per a cada bloc en el conjunt de dades. Aquest conjunt de variables és l'entrada a un mètode de Clusterització Condicional utilitzant un algorisme jeràrquic basat en el mètode de Ward, i la métrica imita de Gibert i X_{Loc}

com a variable condicional. Una nova variable de classe $\mathcal{P}$ resultant del procés. En la seva forma original (on les classes descobertes van rebre un identificador numèric), la variable no és interpretable per si mateixa i es requereix un postprocessament per interpretar els clústers, aconseguint etiquetes representatives per a cada clúster, de manera que la variable de classe es converteix en una nova variable qualitativa amb modalitats significatives. Això s'adreça en els següents passos.

2. Interpretació i etiquetatge de les classes obtingudes: Repetiu els passos 3 a 8 de la creació de variables de segona generació
3. Perfil de classes: Analitza la importància de les variables d'entrada en les classes i les distribucions condicionals per identificar les característiques rellevants de cada classe, de manera que es pugui fer una breu descripció de cada classe característica.

5.18. Fase V. Interpretació de resultats, elaboració del diagnòstic i recomanacions finals

La última fase es desenvolupa conjuntament amb els experts en el domini d'aplicació (que en realitat han d'acompanyar tot el procés) i consisteix a generar les interpretacions en termes de l'aplicació necessàries per convertir els resultats de l'anàlisi de les dades en elements de suport efectiu a la presa de decisions. Convé en aquesta fase abordar qüestions com:

- Contextualització dels resultats obtinguts a les fases prèvies, en l'aplicació real concreta
- Diagnosticar la situació de l'ecosistema de manera objectiva.
- Identificació de punts forts i febles
- Elaborar unes recomanacions de suport a la presa de decisions per posar solucions als punts febles i reforçar els punts forts
- Elaboració dels informes amb el que s'ha generat en tot el procés.

Això significa que la metodologia desenvolupada pel projecte INSESS-COVID19 afavoreix una infraestructura tecnològica que permet obtenir informació directa i fresca dels ciutadans, col·lectius professionals específics o actors rellevants implicats en una decisió sobre el cer-tain mitjançant eines de participació directa, on:

El responsable de la presa de decisions pot decidir què fer, fins i tot si el seu sistema d'informació no recull aquesta informació (la modificació del qüestionari requereix menys de 2 h)

El responsable de la presa de decisions pot decidir qui ha de rebre el qüestionari i quan (el disseny de mostra i la representativitat dels enquestats són crucials)

El responsable de la presa de decisions pot decidir si la resposta al qüestionari és voluntària o obligatòria i els terminis de resposta

En funció del cas, la crida als enquestats pot ser immediata si hi ha correus personals disponibles, o pot requerir més temps, si les institucions intermèdies han de trobar-los i trucar. Tanmateix, això es troba fora de la part tecnològica de la metodologia proposada.

Una vegada convocats els participants i activat un nou qüestionari, 20 min serien suficients per respondre a un qüestionari d'extensió similar a la construcció d'INSESS-COVID19, i 15 min proporcionarien al document de treball els resultats de l'anàlisi per al diagnòstic i la interpretació, constituint així una eina molt potent per al diagnòstic ràpid de situacions rellevants per a la presa de decisions, i per a la implementació d'estratègies de participació directa en una nova forma de formulació de polítiques.

Per descomptat, les eines proposades no són restrictives per a l'elaboració de polítiques, però el seu ús pot ampliar-se per a controlar qualsevol mena de procés industrial o empresarial a través de l'emmagatzematge de dades, simplement modificant el qüestionari o les dades d'entrada dels scripts corresponents.

6. Disseny de les validacions i experimentació

6.1. Validació de l'enquesta i els perfils

El qüestionari i el disseny de mostres de la primera fase de la metodologia es validen àmpliament a través de diverses rondes d'experts.

1. Dos experts del consell assessor del projecte especialitzat en innovació per als serveis socials van analitzar tant la llista de preguntes com el conjunt de perfils objectiu definits i van proporcionar comentaris positius i alguns suggeriments per millorar l'escriptura per reduir ambigüitats
2. Les versions actualitzades dels perfils objectiu i del qüestionari es van presentar a la Comissió de Serveis Socials de la Federació Catalana de Municipis i es va celebrar un taller per avaluar la proposta que van acceptar plenament els experts
3. El personal tècnic del Departament de Serveis Socials de la Generalitat de Catalunya també ha revisat els materials amb comentaris amb èxit
4. Els professionals dels serveis socials van comprovar els materials amb una retroalimentació positiva i petites esmenes sobre l'ambigüitat dels escrits

Cap d'ells va detectar cap perfil que manqués en el disseny de la mostra o la pregunta en el qüestionari i alguns van destacar l'interès d'alguns perfils o preguntes aparegudes a conseqüència de la revisió sistemàtica proposada en el document que no s'hauria inclòs des d'un enfocament més tradicional basat en experts (com focus grups).

6.2. Validació de la mostra

Després de la recollida de dades, s'hauria de requerir una nova validació de la representabilitat de la mostra. A més a més, això es pot perseguir fent comparacions de proporcions proves estadístiques i proves d'homogeneïtat per comprovar, mentre que la distribució de la mostra és homogènia a la distribució de la població. No obstant això, aquesta és la primera vegada a Catalunya (i probablement a Espanya) que un estudi s'adreça a 20 perfils vulnerables, independentment si són usuaris actuals del Sistema de Serveis Socials o no. A la condició publicitària, no hi ha dades de població disponibles per fer aquesta validació. De fet, totes les estadístiques oficials de referència o informes consultats com a Estat de l'Art tenen algunes similituds amb el nostre estudi, però les poblacions objectiu no són directament comparables, impedint així la possibilitat de provar aquesta part. Sent la primera vegada que s'analitza una

població d'aquest tipus, aquest treball es convertirà en la referència per provar altres estudis en el futur.

Malgrat aquesta limitació, intentem anar més lluny i inspeccionar algunes de les estadístiques oficials de referència i informes per a veure si podíem obtenir algunes pistes i indicacions que la nostra mostra representa efectivament bé a la població de referència.

6.3. Metodologia de validació 2nd generation

Des del punt de vista metodològic, es va intentar comparar la naturalesa del coneixement extret quan les dades es consumeixen en altres models de ciència de dades en els dos escenaris d'utilitzar només variables originals, o incloent-hi també variables de segona i tercera generació.

A aquest efecte, dos grups d'experts han estat col·laborant en la recerca:

El primer grup va ser exposat a l'anàlisi de variables originals i el tipus de coneixement sobre salut mental proporcionat per un agrupament global de les dades originals i la interpretació posterior dels cúmuls resultants en perfils significatius.

El segon grup va treballar amb els enginyers de coneixement per construir variables de segona i tercera generació. Després, els resultats del clustering del conjunt de dades enriquit incloent les noves variables creades on es mostren i van participar en la interpretació dels perfils resultants.

La naturalesa dels nous coneixements obtinguts en els dos escenaris es va posar en comú i es va debatre amb tots els experts.

6.4. Metodologia de validació del Termòmetre .

En aquest apartat es presenten els protocols de validació de les diferents contribucions metodològiques del document:

Per validar la introducció del termòmetre en la generació de TLPs automàtics: s'utilitza termòmetre per millorar el TLP. Per tant, la metodologia de validació proposada es basa en la comparació de dos TLP. Una construcció tradicional, on el color d'una cèl·la és decidit per un humà basat en l'anàlisi de CPG; la segona, el TLP basat en el termòmetre, on el color de les celes és decidit pel mètode proposat en la secció 5.14.1. Tant els TLP com la seva interpretació es mostren a un grup d'experts en el camp, que discuteixen quin dels dos és més creïble o quin d'ells proporciona una millor comprensió del domini de destinació

6.5. Metodologia de validació del TFSM

Per a validar el mètode de selecció de característiques territorials (TFSM): es validen dos aspectes:

1. **La pertinència de l'índex d'explicabilitat potencial proposat:** Els resultats de l'ús de la metodologia proposada TFSM es comparen amb l'estat de l'art d'utilitzar la prova de χ^2 de la variable territorial (X_{Loc}) enfront cada variable X_k com a eina per seleccionar la variable més discriminant per als passos de clustering posteriors. No obstant això, tot X_k proporciona un valor p de prova significatiu χ^2 independent i no és útil per reduir el nombre de variables que s'utilitzaran per representar els blocs en el procés de clustering, mentre que l'índex proposat permet un rànquing de variables de més discriminant a menys i permet el procés de selecció de variables.
2. **Classificació final obtinguda:** Es comparen dues classificacions. Un resultat de l'aplicació de TFSM,anomenat $\wp$ Un altre resultat d'agrupar el conjunt de les variables de tercera generació creades per a cada bloc, és a dir $\wp'$ Hi ha dues maneres de comparar i validar si la obtinguda a través de TFSM és millor, una basada en eines gràfiques i una altra basada en el numèric.
   2.1. **Validació numèrica:** Calcula els valors de prova de Lebart: utilitzant la metodologia de 2.1.2 tots els valors de prova de Lebart s'obtenen per a cada c $\in \wp$ versus X_{Loc} i per a cada c' $\in \wp'$. Es calculen taules $V_{\wp}$ i $V_{\wp'}$ similars a les taules V_k presentades en S2.6.3. Aquí, el clustering ha estat condicionada a les ubicacions, tots els individus d'una ubicació s'agrupen en una sola classe. De manera que $s_{lk} = 1$ $\forall\ l = 1{:}L, k = \wp, \wp'$ i $R_{\wp} = R_{\wp'} = 1$ per construcció, de manera que $E_{\wp} = E_{\wp'} = 1$ i l'índex d'explicabilitat potencial proposat no es poden utilitzar en mètodes de clustering condicional. Per aquesta raó, simplement calcularem la proporció global de cel·les significatives en $V_{\wp}$ $and\ V_{\wp'}$ per veure quina de les dues particions pot explicar una part més gran del territori. Donada una partició P amb n_P classes, (P $\in\{\wp, \wp'\}$), l'índex és

$$S_P = \frac{card_{ln|\ l=1:L, c=1:n_P}\{v_{lc} \leq 0.05\}}{L * n_P} \qquad (17)$$

   Aquest índex explica la proporció de cel·les significatives en un V_k table. Com més gran és el valor de S_P, millor P distribueix al llarg d'Ubicacions. Proposem comparar $V_{\wp}\ i\ V_{\wp'}$ i consegüentment $S_{\wp}\ i\ S_{\wp'}$,per tal de veure quina de les dues particions es distribueix d'una manera més significativa al territori.

   2.2. **Validació gràfica:** s'ha de dibuixar un mapa per classificació. El mapa està pintat en funció de la classe que pertany a la ubicació. Es considerarà la cohesió territorial de les classes per a l'avaluació

7. Aplicacions a casos reals i resultats

7.1. Introducció

En aquesta tesi s'ha treballat principalment en un gran cas real d'aplicació, que és el projecte INSESS-COVID19, però a més s'ha utilitzat fins a 4 aplicacions reals diverses que ha permès mostrar la robustesa i versatilitat de la metodologia proposada.

L'estructura d'aquest capítol és:

1. INSESS-COVID19
2. Consum energètic de les famílies
3. Enquestes de valoració a una associació sense ànim de lucre
4. Projecte DIMCARE

7.2. INSESS-COVID19

Descripció del projecte INSESS-COVID19

El projecte INSESS-COVID19 (Identificació de necessitats socials emergents a conseqüència de la COVID-19 i efecte en els Serveis Socials del territori [Gibert 2020] es va centrar en la comprensió i l'anticipació del desbordament dels Serveis Socials esperats després del desbordament del sistema sanitari a causa del confinament plantejat per la crisi SARS-COV2 (març 2020). El Sistema de Serveis Socials de Catalunya estava interessat a obtenir aviat informació sobre les noves necessitats possibles de les persones vulnerables. El projecte INSESS-COVID19 intenta donar respostes i introdueix el potencial de la clusterització multivariant a través del TLP basat en el termòmetre per realitzar un estudi prospectiu que permeti identificar les vulnerabilitats socials de la població catalana d'una manera prou comprensible perquè els resultats proporcionin elements de suport a la presa de decisions i la presa de polítiques a les 107 Àrees Bàsiques de Serveis Socials (ABSS) de Catalunya i al Departament de Serveis Socials de la Generalitat de Catalunya.

Va començar l'abril de 2020, la recopilació de dades es va tancar el 6 de desembre de 2020 i els resultats finals del projecte es van escriure en un informe general [Gibert, Codina & Angerri, 2020]. Va ser publicat al web del projecte i presentat a la Generalitat de Catalunya el

15 de desembre de 2020, només nou dies després del tancament de la recollida de dades. Es va presentar un informe general al Govern i es va distribuir al 107 ABSS (àmbits bàsics dels serveis socials) de tot Catalunya, que conté informació de tots els ABSS junts. No obstant això, les idees d'aquest informe no van proporcionar elements per a donar suport a la presa de decisions i la presa de polítiques per a cada ubicació.

Tots els detalls metodològics s'expliquen clarament en [Gibert, Codina & Angerri, 2020], que també conté els resultats globals del projecte. Es descriuen amb diverses taules, grafs i mapes. A més, en [Gibert & Angerri, 2021] es defineixen les noves tècniques a aplicar i la feina anterior relacionada amb INSESS-COVID. Aquesta informació va ser comunicada a la Generalitat de Catalunya que va poder prendre decisions globals ràpidament el 2020, mentre la COVID-19 encara s'estava promulgant.

Aquest projecte ha estat finalista dels Premis de Serveis Socials Europeus 2021 i ha contribuït al premi Muncunill 2021 de l'Ajuntament de Terrassa, al qual havia guanyat la Universitat Politècnica de Catalunya.

Base de dades INSESS-COVID19

La font de dades és el qüestionari INSESS-COVID19. Un dels objectius principals del projecte era obtenir informació de diversos àmbits de la vida, cosa que significa que el qüestionari hauria de preguntar sobre ells. Els temes objectiu apareixen en el model conceptual de referència anomenat model SSM.cat [DIXIT CDSS, 2021], un instrument per avaluar la vulnerabilitat social adoptada pel Govern de Catalunya per formar part del nou sistema de serveis socials és (e-Social), planificat com a nucli de la transformació digital dels serveis socials objectiu en el Pla Estratègic [PESS, 2020] i molt alineat amb l'estructura actual dels serveis socials d'atenció primària a Catalunya.

Els temes anteriors es van convertir en l'etiqueta principal dels blocs del qüestionari. No obstant això, tots els blocs no estan compostos pel mateix nombre de variables. Això també pot ser diferent Això significa que no es poden analitzar utilitzant les mateixes tècniques, a més en el moment de la selecció de característiques.

Els primers blocs contenen preguntes personals, com l'edat i el lloc de residència. Aquesta última es convertirà en una variable important en les següents seccions. Els blocs següents contenen informació relacionada amb els temes del model SSM.cat. La base de dades original conté un total de 195 variables. Gràcies a les noves variables creades utilitzant nous mètodes mostrats en [Angerri & Gibert, 2023] i en aquest document la quantitat de variables augmenta a 258.

Les respostes del qüestionari es van obtenir a través d'un taller especial desenvolupat en el projecte INSESS-COVID19 i descrit en [Gibert & Angerri, 2021]. L'esquema de la proposta és la següent:

Figura 55: Disseny MIPRI2D per a INSESS-COVID19

7.2.1. FASE I Anàlisi del fenòmen i disseny d'eines d'observació

Anàlisi de l'ecosistema

Analitzar l'ecosistema va consistir en la comprensió del sistema de serveis socials de Catalunya. Per aquest fi, es van consultar tota una sèrie d'informes i enquestes realitzades al sector que es poden veure a la secció 2.10 d'aquest document.

Identificació de les subpoblacions i perfils de destinació Identificació de la població en estudi*

Com s'ha dit abans, després d'una profunda comprensió de la llista de serveis socials disponibles oferts en el sistema d'atenció social primària, es va definir una llista de 20 perfils objectiu i els criteris d'inclusió corresponents juntament amb els professionals dels Serveis Socials, tant dels governs, ajuntaments com dels consells regionals (consells comarcals). Els perfils proposats assenyalen segments de població a priori que s'espera que siguin significativament danyats per la pandèmia:

1. Famílies monoparentals
2. Persones joves desocupades
3. Persones desocupades majors de 50 anys
4. Persones nouvingudes en situació irregular

5. Menors no acompnyats/des i joves extutelats/des
6. Treballadors/es pobres (salaris molt baixos)
7. Treballadors/es pobres (temporals i fixos-discontinus)
8. Treballadors/es pobres (economia submergida)
9. Persones afectades per un ERTO o un acomiadament
10. Treballadors autònoms i petits empresaris en fallida
11. Persones grans amb dependència
12. Persones grans que viuen soles
13. Persones amb discapacitat (física, sensorial o intel·lectual)
14. Cuidadores no-professionals
15. Persones amb malaltia o trastorn mentals
16. Persones de la comunitat LGTBI en situació de vulnerabilitat
17. Persones amb alcohòliques i drogodependents
18. Dones víctimes de violència masclista
19. Persones sense llar o en situació d'infrahabitatge.
20. Professionals de serveis essencials socials i de salut

7.2.2. Fase I Disseny del qüestionari INSESS-COVID19

Al cas INSESS-COVID19, del marc conceptual se'n va derivar una conceptualització de les àrees de la vida que es volen estudiar inspirades en el model de referència SMM.cat, referit a l'estat de l'art.

Així, el qüestionari INSESS-COVID19 focalitzant preguntes sobre els àmbits de SSM.cat, però dirigides a fer aflorar, no només la vulnerabilitat social, sinó també l'impacte de la COVID-19 en aquesta vulnerabilitat. El resultat és un qüestionari amb 21 preguntes que generen fins a 190 ítems interns, d'estructures variades, segons el tipus de preguntes.

Des del punt de vista de la resposta, respostes numèriques o categòriques, algunes de resposta múltiple, i algunes quadrícules.

L'estructura del qüestionari, obre blocs específics per alguns aspectes concrets i està esquematitzat a la 56:

Després d'una extensa anàlisi del marc conceptual, es va acordar amb els experts una conceptualització de les àrees objectiu de la vida a estudiar. Entre totes les estructures, enquestes i informes analitzats, el model conceptual de referència va ser el model SSM.cat [DIXIT CDSS, 2021], un instrument per calcular la vulnerabilitat social adoptada pel Govern de Catalunya per formar part del nou sistema de serveis socials (e-Social), planificat com a nucli de la transformació digital dels serveis socials objectiu del Pla Estratègic de Serveis Socials de Catalunya [PESS, 2020] i molt alineat amb l'estructura actual dels serveis socials d'atenció primària a Catalunya. El procés pel qual es va seleccionar aquest model de referència és nou,

ja que es basa en una revisió sistemàtica de l'Estat de l'Art, incloent-hi l'elaboració d'una taxonomia d'indicadors, agrupats per temes, i la descripció de les enquestes revisades en termes del nombre de variables (i de dalt a dalt) relacionades amb cada tema, l'avaluació experta de la utilitat d'aquestes qüestions sobre els objectius de l'estudi, i el disseny dels blocs temàtics i la seqüència segons això.

El model SSM.cat es va inspirar en la versió neerlandesa del model d'Autosuficiència Matrix, desenvolupat per la Universitat d'Amsterdam [Lauricks et al., 2012], que al seu torn és una adaptació de la matriu d'autosuficiència original desenvolupada per Diana Pearce per a Oportunitats més àmplies per a les dones com a part del Projecte Estatal d'Organització de l'Autosuficiència econòmica familiar [SSS, 2002] [Brooks & Perarce, 2000].

Inspirat en SSM.cat, INSESS-COVID19 avalua la vulnerabilitat social de les 11 àrees de la vida diària que disminueixen:

- Ingressos
- Activitats diàries
- Casa
- Relacions internes
- Salut mental
- Salut física
- Abús de substàncies
- Habilitats d'activitats diàries
- Xarxa social
- Participació comunitària
- Marc jurídic

El qüestionari INSESS-COVID19 s'ha desenvolupat centrant les preguntes en aquestes àrees. Cada àrea pot contenir un nombre diferent de preguntes, principalment orientades a posar en relleu no només la vulnerabilitat social, sinó també l'impacte de la COVID19 en aquesta vulnerabilitat. El resultat és un qüestionari amb 21 blocs que generen fins a 195 elements, de diferents estructures, segons el tipus de preguntes. La figura 56 mostra l'estructura global de l'enquesta.

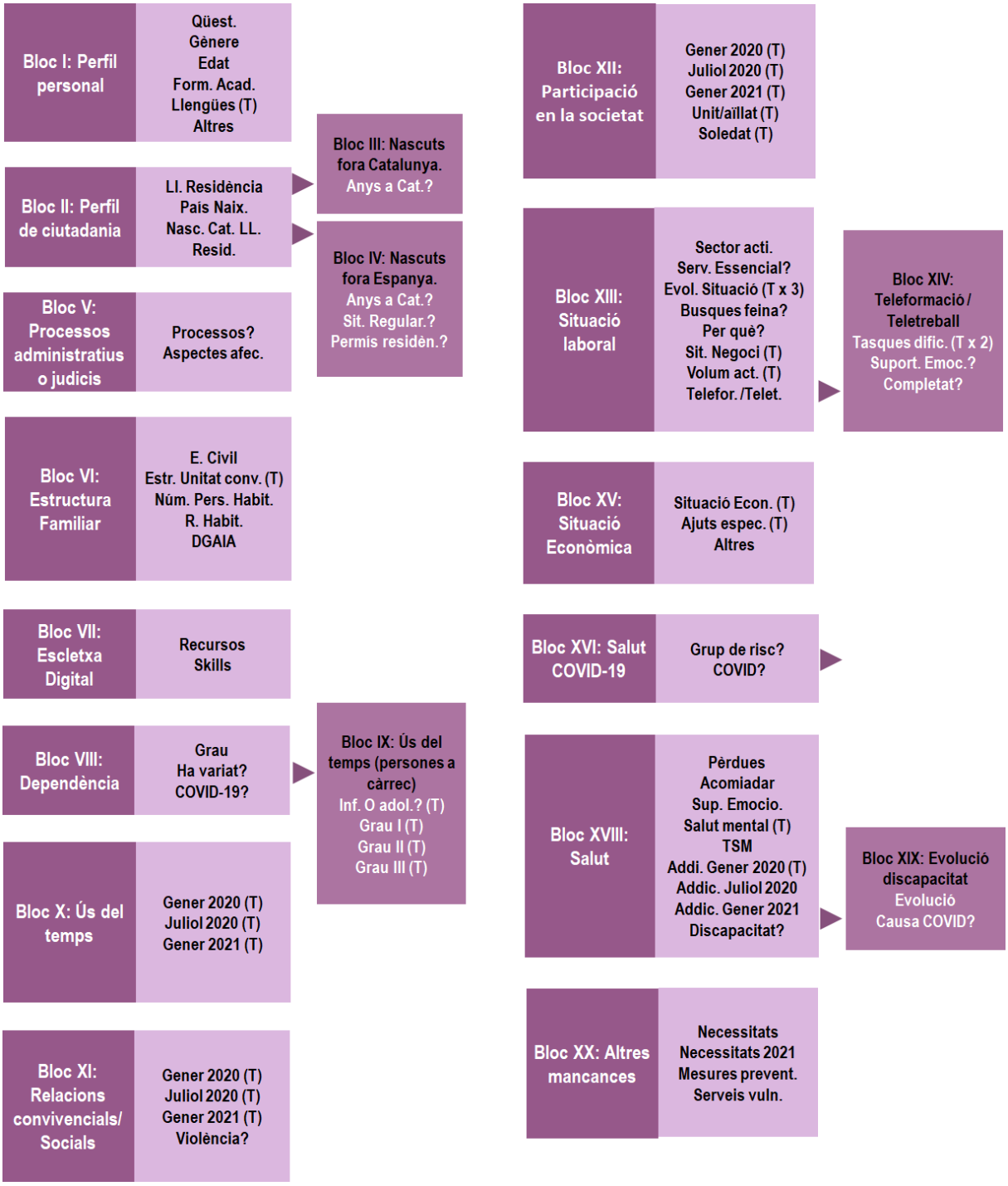

Figura 56: Disseny del qüestionari INSESS-COVID19

A tall d'exemple, a continuació es mostraran les variables específiques del bloc relatiu a la salut i específicament a la salut mental, ja que a [Angerri & Gibert 2023] es presenta una anàlisi de la base de dades INSESS-COVID19 enfocada en aquest tema. Del qüestionari original INSESS-COVID19, un total de 15 preguntes repartides entre diversos blocs, consideren l'impacte de la primera onada o COVID19 en salut mental i salut. Els blocs implicats són els següents:

Bloc VIII: Dependència (conté 3 preguntes sobre dependència)

Bloc XIV: Teletreball i teleformació (conté 1 pregunta per identificar si el teletreball o el teleformació van afectar la salut mental)

Bloc XVI: COVID-19 de salut (2 preguntes per veure si la persona tenia COVID-19)

Bloc XVII: COVID-19 (2 preguntes específiques per a aquells que van passar la malaltia)

Bloc XVIII: Salut (5 preguntes sobre salut mental, abús de substàncies i discapacitat).

Bloc XIX: Evolució de la discapacitat (2 preguntes per a l'impacte de la malaltia sobre la discapacitat)

Entre aquestes preguntes, 2 d'elles són del tipus TQQ (Temporal Qualified Variable, vegeu [Gibert & Angerri, 2021]), de manera que cada pregunta genera 4 variables. Això significa que a partir de 15 preguntes, 21 variables es deriven a la base de dades de treball. Breument, una variable TQQ és de fet un triplet (X, T, Q) on X és una variable qualitativa replicada T vegades i Q és una variable Likert per qualificar les modalitats de X a cada marca horària de T. En aquest cas, inspecciona un problema objectiu a la línia de base (gener de 2020), generant una variable, i després de la primera onada (juliol de 2020), però generant tres variables addicionals per indicar si la persona se sent millor al juliol de 2020 respecte a gener o no, si ell/ella sent el mateix que al gener o ell/ella sent pitjor.

Les variables originals INSESS-COVID19 pel que fa a la salut es troben a la Taula 1. La quarta columna indica el tipus de la variable segons la tipologia establerta a [Gibert & Angerri, 2021].

Bloc	Codi pregunta	Variable	Modalitats	Tipus variable
B8	D1	Tens algun grau de dependència?	1. No; 2. Grau I (dependència moderada); 3. Grau II (dependència severa); 4. Grau III (gran dependència); 5. No Contesta	Ordinal
B8	D2	Creus que si et valoressin ara tindries una variació en el grau de dependència?	1. He millorat; 2. El mateix; 3. He empitjorat; 4. No Contesta	Ordinal
B8	D3	Atribueixes aquesta variació a la COVID-19?	1. Sí; 2. No; 3. No he variat; 4. No Contesta	Nominal
B16	SC1	Ets persona d'algun grup de risc sensible a la COVID-19?	1. Si; 2. No; 3. No Contesta	Nominal
B16	SC2	Has Passat la COVID-19?	1. Sí, he estat a l'UCI; 2. Sí, m'han ingressat però no a l'UCI; 3. Sí,	Ordinal

			diagnosticat amb símptomes i atenció mèdica telefònica a casa; 4. Sí, a casa i amb atenció mèdica telefònica; 5. He tingut símptomes, però no se sap; 6. Sí, diagnosticat però asimptomàtic; 7. No he tingut cap molèstia; 8. No contesta	
B17	Cov1	Quan te'l van detectar?	1. Entre l'1 i el 15 de març; 2. Entre el 16 i 31 de març; 3. Entre l'1 i el 15 d'abril; 4. Entre 16 i 30 d'abril; 5. Entre l'1 i 15 de maig; 6. Entre el 16 i 31 de maig; 7. Entre l'1 i 15 de juny; 8. Entre el 16 i 30 de juny; 9. Entre l'1 i 15 de juliol; 10. Entre el 16 i 31 de juliol; 11. Entre l'1 i el 15 d'agost; 12. Entre el 16 i 31 d'agost; 99. No contesta	Temporal
B17	Cov2	Les teves condicions de salut després de passar la COVID-19 et permeten reprendre la teva activitat habitual?	1. Sí, sense problema; 2. Amb dificultat; 3. No ho podré fer abans del setembre; 4. No ho podré fer fins al gener 2021; 5. Amb seqüeles que impedeixen recuperar l'activitat habitual; 6. No contesta	Ordinal
B18	S9	S9. Tens alguna discapacitat reconeguda?	1. Física; 2. Sensorial; 3. Intel·lectual; 4. Cap; 5. No contesta	Nominal
B19	ED1	Creus que el teu grau de discapacitat ha variat respecte del gener de 2020?	1. Sí, ha millorat; 2. No ha variat; 3. Si, ha empitjorat; 4. No contesta	Ordinal
B19	ED2	Si has empitjorat, ho atribueixes a la situació generada per la COVID-19?	1. Sí; 2. No; 3. No he empitjorat; 4. No contesta	Nominal
B14	TT2.TT.TFsuportG20	El teletreball / teleformació t'ha requerit suport emocional? *Fins al moment, Com et veus a gener 2021*	De la teva xarxa de suport personal; Professional; No; No contesta	Temporal basic variable
B18	S3	Has requerit suport emocional degut al COVID-19?	1.Xarxa Professional; 2.Xarxa Personal; 3.No; 4.No Answer	Temporal basic variable
B18	S4	Tens algun problema de salut mental diagnosticat i com has evolucionat durant la crisi de la COVID19?	Trastorn mental greu; Trastorn límit de la personalitat; Trastorn d'Estrès Post-traumàtic; Depressió; Ansietat; Altres; Cap; No contesta	Temporal Qualified

		Gener 2020; Juliol 2020 estic millor; Juliol 2020 estic igual; Juliol 2020 estic pitjor		Qualitative
B18	S5	Estàs rebent tractament farmacològic?	1. Sí; 2. No	Binary
B18	S6// S7// S8	Indica el teu consum de substàncies o conductes indicades: Tabac; Drogues; Alcohol; Conductes additives (ludopatia, ciberadiccions….) *Al Gener 2020, al Juliol 2020 I al Gener 2021*	1. No; 2. Ús; 3. Abús; 4. Addicció	Temporal Qualified Qualitative

Taula 16. Variables al qüestionari

Aquestes variables produeixen informació directa relacionada amb la salut i la salut mental. Amb la tecnologia analítica INSESS-COVID19 desenvolupada en la primera generació s'obtenen resultats, però es pot fer una anàlisi més complexa quan aquestes variables originals es combinen en noves variables derivades, siguin basades en el coneixement o basades en dades, segons la metodologia proposada.

Disseny d'eines per als tallers

Considerant la situació crítica de les Àrees Bàsiques durant la 1a onada de la pandèmia, els tallers es van dissenyar tenint en compte que la majoria d'Àrees Bàsiques del territori no podien dedicar temps a l'estudi al juny-juliol, i sense introduir el tipius de disseny de tallers per lliure al disseny, el projecte no era viable.

L'impacte de la 1a onada de la pandèmia es mesura a través de les diferències entre juliol i gener de 2020.

Amb aquest exercici es fa possible tractar el qüestionari com una anàlisi prepost i poder estudiar les dinàmiques provocades per la pandèmia, així com copsar la percepció que els participants tenen sobre el seu futur (a mitjà termini pels que responien al mes de juliol, a curt pel que han respost al novembre).

Aquest disseny va possibilitar allargar la recollida de dades, inicialment planificada pel mes de juliol, fins molt més tard, donant així la màxima oportunitat a participar de l'estudi a totes les ABSS que ho han volgut fer, esperant el moment que les seves pròpies circumstàncies ho han permès, i ates el nivell de sobresaturació que els SS viuen des del març de 2020.

7.2.3. FASE II Tallers i data aquisition

Segons les estadístiques oficials de l'últim Baròmetre del Tercer Sector [BTSS, 2017], la població vulnerable de Catalunya és d'1.584.000 persones. La grandària de la mostra pot ser minada per deterció sota l'aproximació de població infinita, ja que l'asímptota de l'error de la mostra sota l'aproximació de població finita s'aconsegueix al voltant d'1.000.000 de població. Segons expressions clàssiques [Krejcie & Morgan, 1970], una mostra de 1067 ciutadans que participen en el projecte proporcionaria un error de mostra de 0,03 a un nivell de confiança de 0,95.

Tenint en compte que els ABSS es trobaven en una crisi de desbordament a causa de la pandèmia, assumim que al voltant d'un 20% d'ells no podrien participar en el projecte, així que vam determinar que es demanaria a la xarxa de 107 ABSS de tot el territori que trobés 20 ciutadans cadascun, seguint un mínim de 10 dels perfils objectiu. Els professionals serveis socials per a cada baix estaven seleccionant 20 ciutadans d'un subconjunt de perfils que representaven adequadament els principals problemes que ocorren en les seves àrees geogràfiques. Els equips de serveis socials estaven en joc en aquest pas en un senyal de mostra de dues etapes, en una combinació entre metodologies de cocreació col·laboratives i estratègies clàssiques de mostreig en múltiples etapes.

Els ciutadans seleccionats van ser convidats a participar en el projecte seguint els tallers INSESS-COVID19 en qualsevol de les seves formes. El qüestionari INSESS-COVID19 es va obrir a partir del 17 de juliol de 2020 i ha estat recollint dades contínuament fins al 6 de desembre de 2020. El 7 de desembre de 2020, es van recollir 971 respostes en una base de dades que contenia 195 variables i es van descarregar per a l'anàlisi automàtica segons allò que s'ha explicat a la seccio

Altres modalitats: Al llarg del procés de recollida de dades, algunes ABSS van utilitzar mecanismes creatius per implicar la ciutadania en l'estudi:

- Rubí va decidir organitzar un taller quasi-cara-a-cara pel seu compte, utilitzant els materials electrònics disponibles per al Taller Lliure al lloc web del projecte.
- Mollet del Vallès, va convocar el departament de Cultura per celebrar un taller quasi cara a cara en un dia obert per tal d'involucrar més ciutadans i proporcionar informació a 80 ciutadans.
- Reus estava utilitzant una estratègia de xarxa distribuïda, de manera que cadascun dels especialistes havia de trobar només dos o tres participants i es va seguir una entrevista telefònica per omplir el qüestionari
- Cervemakers i l'Institut de Cervelló, van proposar la participació en IN-SESS-COVDI19 com a activitat voluntària per a estudiants de la 4a ESO i també han col·laborat com a agents del projecte mitjançant el seguiment de la participació ciutadana.

Validació del disseny del taller i de les infraestructures tecnològiques

El 2 de juliol es van dur a terme dos pilots:

1. El ABSS Castell-Platja d'Aro va trobar 20 persones que van conèixer alguns dels perfils sol·licitats i van demanar el taller en un lloc proporcionat pels Serveis Socials. A causa de la pandèmia, l'equip INSESS-COVID 19 es va unir remotament a la reunió, a través de la videoconferència, i va presentar el projecte, va donar el context i totes les instruccions i va respondre tots els dubtes sobre el qüestionari als participants. Després del taller de 2 h, les 20 respostes ja es van pujar al servidor INSESS-COVID19. Cap de les preguntes va ser mal interpretada i es van donar totes les respostes.

2. El segon pilot va tenir lloc al ABSS la Noguera. En aquest cas, intentant salvar la bretxa digital, els professionals dels serveis socials van seleccionar els 20 participants, i van donar una trucada telefònica per aprovar el qüestionari; el professional estava transcrivint les respostes citi-zen al servidor INSESS-COVID19. El temps necessari per recollir respostes dels 20 participants va durar més de 2 mesos. Cap dels enquestats va malinterpretar cap pregunta.

Validació de la mostra

Les estadístiques oficials de l'INE o de l'IDESCAT, com el cens o el padró, proporcionen dades sobre la proporció de persones amb discapacitat a Catalunya, per exemple, i com que totes les persones amb discapacitat obtenen una certificació dels Serveis Socials, succeeix que si la mostra INSESS-COVID19 és vàlida, la proporció de mostra de persones amb discapacitat hauria de ser igual a la proporció real reportada a l'IDESCAT. El mateix passa amb l'habitatge assignat; totes les famílies que han obtingut el dret a tenir una casa de gratuïtes s'han vinculat al sistema de Serveis Socials per gestionar-la i l'IDESCAT a l'Anuari Estadístic de Catalunya 2019 informa de la proporció de la població catalana en aquesta situació que és comparable a la que apareix a la mostra INSESS-COVID19. La mateixa situació es produeix amb les persones vídues, que es reporta oficialment en el cens de l'INE i totes elles processen la seva pensió a través del sistema de serveis socials. No obstant això, la proporció de persones casades no seria comparable. De fet, ja que el cens es fa per a tota la població i és vulnerable o no afecta directament a la capacitat de casar-se (que és un indicador d'estabilitat), les estadístiques oficials del cens sobre les persones casades no es poden comparar directament amb les de la nostra mostra, on només s'adreça la població vulnerable.

L'informe oficial dels Serveis Socials a Catalunya (informe Rudel) no es pot utilitzar per a la comparació, ja que només es refereix als Serveis Socials Bàsics, i també incloem en el nostre estudi altres segments de poblacions com els pacients de salut mental que són usuaris de Serveis Socials Especialitzats i el mateix succeeix amb altres perfils inclosos a la mostra. A més, el tercer baròmetre sectorial proporciona informació interessant, però només pel que fa als usuaris del tercer sector, com s'esperava, i en la nostra mostra, incloem persones que mai

abans s'havien vinculat al sistema de serveis socials ni a altres entitats del tercer sector. Per exemple, els empresaris que havien fet fallida s'inclouen en la causa de mostra INSESS-COVID19, que són un grup vulnerable que mereix atenció i que podria convertir-se en usuaris del sistema de serveis socials en la propera boira, però aquestes persones mai han format part de cap de les estadístiques proporcionades per l'informe del tercer sector Barometer o Rudel. A més, els treballadors dels serveis essencials es van produir en la mostra INSESS-COVID19 provinents del sistema sanitari, del sistema de serveis socials i del sector hostaler. Cap d'elles estava estructuralment relacionada amb els serveis socials abans. A més, les estadístiques oficials sobre la grandària d'aquests sectors professionals també són inusuals, ja que inclouen a persones no vulnerables, que no estan dirigides en el disseny de mostres IN-SESS-COVID19.

Segment de població	Proporció mostral	Proporció poblacional	p-Val	Diferencia Significativa	Font
Persones discapacitades	15,8	14,8	0,82	No	IDESCAT Enquesta econòmica
Cuidadores	0,033	0,036	0,64	No	IDESCAT Anuari Estadístic de Catalunya
Viudes	0,084	0,075	0,27	No	INE cens

Taula 17 Validació. Relació de mostra contra la proporció de població

En síntesi, per a aquells indicadors on hi ha estadístiques oficials externes disponibles i comparables amb la configuració de la mostra INSESS-COVID19, la mostra sembla representativa, però la validació global no és adequada, sent INSESS-COVID19 un estudi pioneer en la seva categoria.

Finalment, l'error estadístic global de la mostra és del 3%, que és prou petit com per proporcionar resultats significatius.

7.2.4. Fase III Anàlisi intel·ligent de dades

*Base de dades INSESS-COVID19**

La font de dades és el qüestionari INSESS-COVID19. Un dels objectius principals del projecte era obtenir informació de diversos àmbits de la vida, cosa que significa que el qüestionari hauria de preguntar sobre ells. Els temes objectiu apareixen en el model conceptual de referència anomenat model SSM.cat [DIXIT CDSS, 2021], un instrument per avaluar la vulnerabilitat social adoptada pel Govern de Catalunya per formar part del nou sistema de serveis socials és (e-Social), planificat com a nucli de la transformació digital dels serveis socials objectiu en el Pla Estratègic [PESS, 2020] i molt alineat amb l'estructura actual dels serveis socials d'atenció primària a Catalunya.

Els temes anteriors es van convertir en l'etiqueta principal dels blocs del qüestionari. No obstant això, tots els blocs no estan compostos pel mateix nombre de variables. Això també pot ser diferent Això significa que no es poden analitzar utilitzant les mateixes tècniques, a més en el moment de la selecció de característiques.

Els primers blocs contenen preguntes personals, com l'edat i el lloc de residència. Aquesta última es convertirà en una variable important en les següents seccions. Els blocs següents contenen informació relacionada amb els temes del model SSM.cat. La base de dades original conté un total de 195 variables. Gràcies a les noves variables creades utilitzant nous mètodes mostrats en [Angerri & Gibert, 2023] i en aquest document la quantitat de variables augmenta a 258.

Les respostes del qüestionari es van obtenir a través d'un taller especial desenvolupat en el projecte INSESS-COVID19 i descrit en [Gibert & Angerri, 2021]

7.2.5. Fase III Preprocessament de les dades

En el projecte INSESS-COVID19 s'han realitzat les següents passes del preprocessing:

- Duplicitats: Es va detectar la presència a la base de dades que hi havia variables que es duplicaven i feien referència a una mateixa pregunta, el que va generar que es fusionessin les dues preguntes. En preguntes nominals de resposta semitancada, es van detectar diferents valors a la modalitat de resposta lliure que es referien al mateix concepte, aspecte que va obligar a substituir un dels 2 per tal d'unificar-ho amb la mateixa etiqueta
- Rephrasing: Les preguntes del qüestionari són majoritàriament oracions interrogatives, fet que genera uns texts molt llargs impossibles de processar en el moment de generar gràfics i taules de forma automàtica. Això genera que s'hi hagin reduït els texts, canviant frases de més de paraules per 1 o 2 mots que representen el concepte al qual fa referència la pregunta. Així mateix, les modalitats existents en els diferents tipus de variables s'han escurçat per tal que la visualització dels gràfics i taules es vegin correctament.
- Verificació del preprocessament: Per tal de validar que les modalitats són substituïdes adientment s'han generat taules de contingència, enfrontant la variable original i la mateixa preprocessades.

7.2.6. Fase III: Analisi descriptiva i territorial

En els següents resultats principals del qüestionari, presentat a la Generalitat el passat 15 de desembre de 2020, es sintetitzen de manera que s'il·lustren les diferents eines utilitzades en l'anàlisi i es discuteixen els resultats globals. La cobertura territorial dels enquestats és raonable, encara que algunes zones de la província de Tarragona no van participar en el projecte INSESS-COVID19 a conseqüència del desbordament dels serveis socials ja esmentats.

Aquí el nombre de respostes es presenten de manera agregada. Més tard, les ABSS amb menys de cinc enquestats es conserven dels resultats públics, i només s'usen per a l'anàlisi interna i per a la construcció dels resultats globals finals.

La Figura 57 visualitza la participació del ABSS proporcionant alguna resposta al qüestionari. Blanc correspon a ABSS que no va participar en el projecte. La figura 58 mostra el diagrama de Paretto. Es pot veure que algunes ABSS específics provenen més que els 20 participants requerits. La figura 59 proporciona participació a nivell de Vegueria.

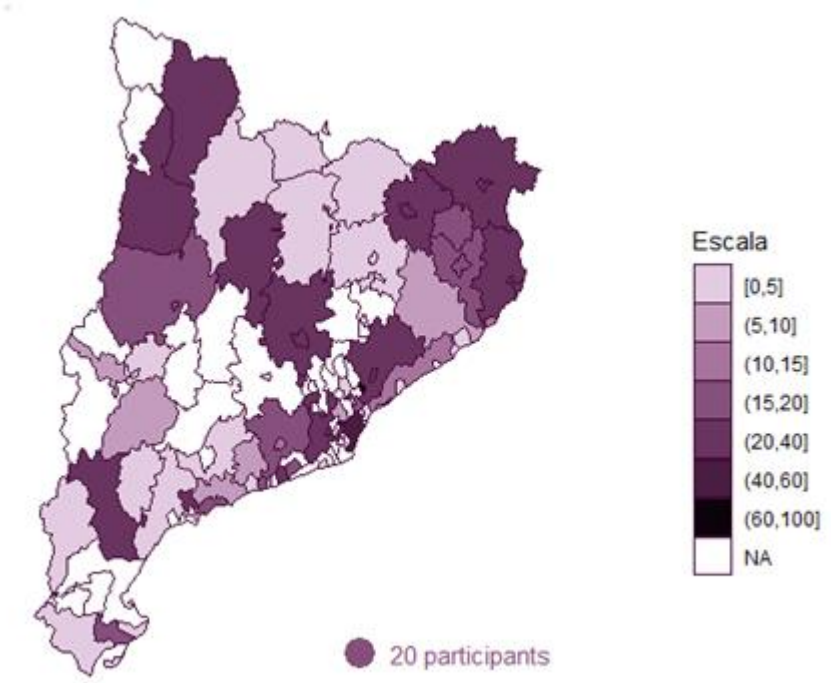

Figura 57. Nombre de respostes per ABSS.

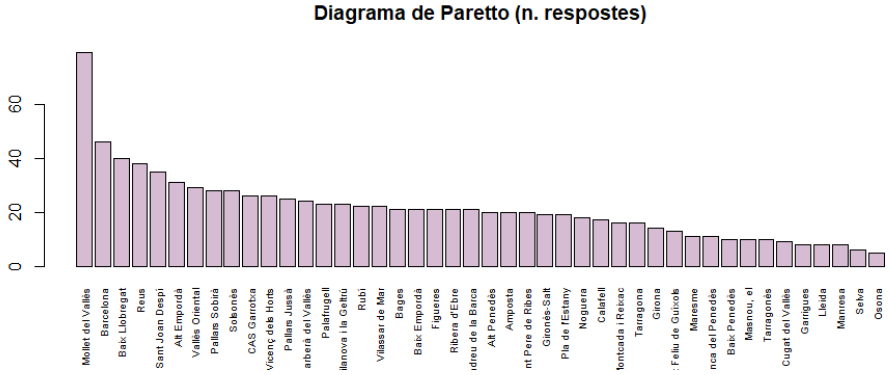

Figura 58 Nombre de respostes per ABSS.

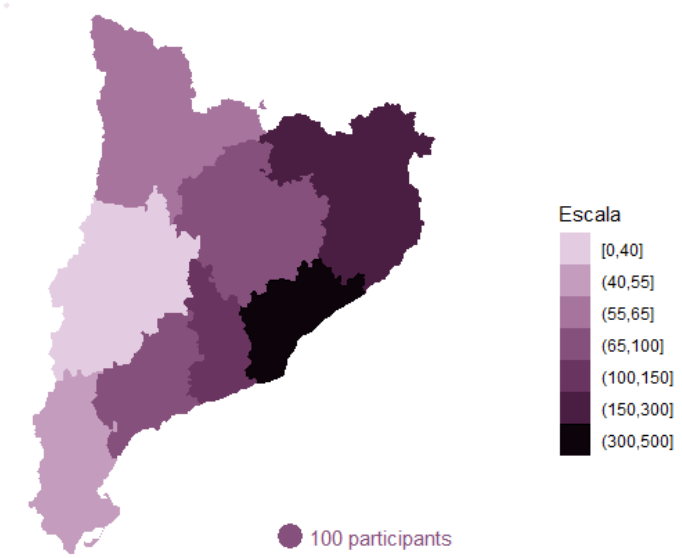

Figura 59 Nombre de respostes per Vegueria.

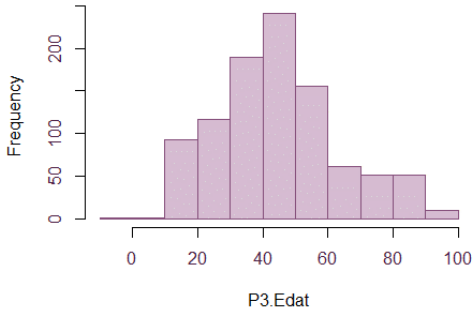

Figura 60: Histograma de l'edat

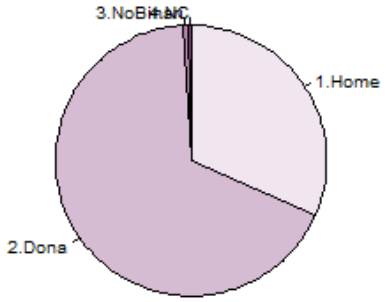

Figura 61: Diagrama de pastís de gènere

Min	Q1	Mediana	Mitjana	Q3	Max	Desviació Estàndard	CV
-10	33	43	45.39	56	95	18.316	0.404

Taula 18: Resum ampliat de 5 xifres.

Salut

De les variables relatives a la salut identificades en el projecte INSESS-COVID19, ens centrarem en el tema de la salut mental. A la Taula 19 hi ha 4 variables relacionades amb la salut mental (codificades amb TT2.TT.TF, S3, S4, S5). D'ells, 2 són variables bàsiques temporals, 1 és variable qualificada temporal (TQQ, S4) i 1 nominal. Aquests són tipus específics de variables descrites en [Gibert & Angerri, 2021] per primera vegada. Entre els tipus més complexos hi ha la variable TQQ (vegeu més amunt).

La pregunta S4: Tens un problema de salut mental diagnosticat i com vas evolucionar durant la crisi de COVID19? és un TQQ, que és un dels tipus complexos de variables introduïts a [10], on X és una variable qualitativa per a trastorns mentals diagnosticats amb S=8 modalitats en DS4 = {trastorn mental greu (SMD), trastorn límit de personalitat (PLD), trastorn de tensió posttraumàtic (PSTD), depressió, anxietat, altres, res}.

X genera un primer vector qualitatiu de base X_{t1}= Tens un problema de salut mental diagnosticat el gener de 2020

Q és un conjunt de valors Likert amb el conjunt comú de valors possibles que cada trastorn mental pot prendre en cada marca temporal. Per S4 la Q té 3 valors possibles ms, Q={Millor, Igual, Pitjor}. Perquè per cada trastorn mental puguem veure si la persona millora o empitjora la seva salut mental entre dues marques de temps consecutives.

A part de X, S4 genera altres 3 variables qualitatives que representen en quins trastorns mentals la persona se sent millor que a X per a juliol de 2020, en què ell/ella sent el mateix, en què ella/ella se sent pitjor.

S4 té una naturalesa multivaluada, ja que una sola persona pot patir simultàniament diversos trastorns mentals. Perquè els valors de DS4 no siguin disjunts entre ells i un individu pugui millorar simultàniament la seva ansietat i depressió al juliol de 2020 i ser pitjor en l'estrès posttraumàtic i el trastorn mental greu. Això confereix una estructura complexa a tota la variable. A més d'això, ja que Q és una evolució qualificadora, la primera marca horària necessita una codificació diferent per indicar una situació de base de tenir o no el trastorn mental al gener de 2020.

Com se sap, les preguntes multivaluades proporcionen molta informació, però de vegades són massa complexes per utilitzar-les en la modelització de dades.

No obstant això, aprofitant l'experiència de les parts interessades, podem construir alguns indicadors interessants a més d'aquesta estructura complexa original que ajuda en l'anàlisi i la comprensió del fenomen objectiu.

El qüestionari digital implementat va produir una estructura de 4 variables per representar S4 (vegeu Taula 19) on el codi de la pregunta, el text de la pregunta es concatenen amb la "modalitat" representada en cada columna. A la taula 19, les celes són multivaluades, ja que a cada cèl·lula poden aparèixer diversos trastorns mentals simultàniament:

ID	S4. Tens algun problema de salut mental	S4. Tens algun problema de salut mental	S4. Tens algun problema de salut mental	S4. Tens algun problema de salut mental

	diagnosticat i com has evolucionat durant la crisi de la COVID19 [1. Gener 2020]	diagnosticat i com has evolucionat durant la crisi de la COVID19? [2. Juliol 2020 estic millor]	diagnosticat i com has evolucionat durant la crisi de la COVID19? [3. Juliol 2020 estic igual]	diagnosticat i com has evolucionat durant la crisi de la COVID19? [4. Juliol 2020 estic pitjor]	
ID1	Depressió	Cap	Depressió	None	…
ID2	Cap	Cap	Cap	Estres PostTraumatic	…
ID3	Estres PostTraumatic; Depressió;Ansietat	Ansietat	Estres PostTraumatic	Depressió	…
:	:	:	:	:	:

Taula 19: Exemple de representació de la base de dades S4

Aquesta informació mostra quins trastorns mentals van millorar o van empitjorar durant la primera onada de confinament.

I l'anàlisi descriptiva automàtica bàsica d'aquestes variables es mostra a la Figura 62, que consisteix en diagrames bàsics de Paretto per a la situació bàsica de les persones, i les millores o empitjorament per al juliol de 2019:

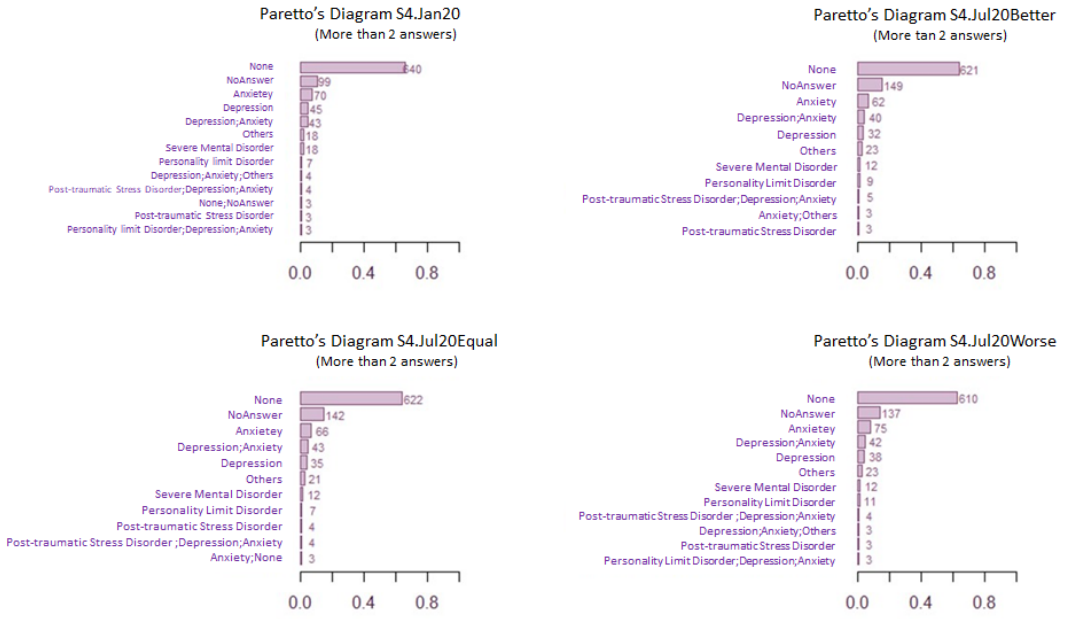

Figura 62:Diagrames de paretoper a S4

Impacte econòmic i de treball

Pregunta L3.1.: Indica la vostra categoria de treball personal al gener de 2020, juliol de 2020 i les vostres previsions per al gener de 2021 («Indica la teva categoria laboral a gener i juliol de 2020 i quina creus que és la teva categoria laboral al gener de 2021»)

Resposta per tota la mostra. Algunes conclusions són visibles a la taula 20, les freqüències de la categoria de treball

- El nombre de persones que no treballen i no reben cap benefici augmenta un 50%
- El nombre de persones que no treballen i reben algun benefici augmenta un 17,6%
- El nombre de persones que no tenen ocupació o ocupació augmenta un 11%

	L3.1.CategLaboralG20	L3.1.CategLaboralJ20	L3.1.CategLaboralG21
1.Estudiant	84	53	64
2.1aFeina	35	46	83
3.NoTreballaNprest	86	130	73
4.NotreballaSprest	125	147	86
5.Mcasa	39	42	37
6.Jubilat	123	126	129
7.Cap	312	275	287
8.NC	167	152	212

Taula 20: L3.1 Freqüències de Categoria de Treball.

Pregunta L3.2. i L3.3.: Indica la vostra situació laboral personal al gener de 2020, juliol de 2020 i les vostres previsions per al gener de 2021 (“Indica la teva situació laboral a gener i juliol de 2020 i quina creus que és la teva situació laboral al gener de 2021”)Vegeu Figures laborals a la figura 23

	L3.3.SituacioLaboralG20	L3.3.SituacioLaboralJ20	L3.3.SituacioLaboralG21
1.Ampliacio	61	44	112
2.Reduccio	55	75	59
3.Acomiadat	36	57	16
4.Plegat	5	16	6
5.Autònom	35	27	27
6.Empressari	<5	<5	<5
7.Cap	575	548	493
8.NC	203	203	257

Taula 21 L3.3. Freqüències de situació laboral.

Aquestes dues preguntes proporcionen diferents modalitats per a la situació laboral:

- El nombre de persones que han estat llicenciades (Acomiadat) o plegades (Plegat) augmenta un 78.04%.

- El nombre de persones que van reduir la seva jornada laboral (Reducció) augmenta un 36,36%.

De taules similars fetes a la pregunta L3.2. (1. Cindefinit (contracte permanent), 2. CtempActiu (contracte de termini fix), 3. TreballPerCTemp (contractes temporals intermitents), 4. TeballNregul (activitat de treball irregular), 5. ERTO (procés de regulació temporal), 6. RecentCtemp (contracte temporal recentment iniciat), 7. TrobaFeinaFixa (fix treball trobat) es va trobar que:

- El nombre de persones que tenien un negoci i deixaven de treballar durant el confinament o entraven en fallida va augmentar un 110%
- El nombre de persones amb condicions de treball no precàries (contractes permanents o de durada determinada) va disminuir un 41,62%.
- El 51,25% de les persones sense treball tenen por de no treballar abans de gener de 2021 (els motius més esmentats són que moltes empreses van tancar a causa de la COVID-19, després d'una certa edat, les possibilitats de contreure's de nou disminueixen, per a determinats sectors, la gent té por d'estar infectada per l'empleat i prefereix no contractar nous treballadors).

Situació econòmica:

Pregunta E1: Situació econòmica al gener de 2020 i juliol, i previsió per al gener de 2021 ("Situació econòmica a gener i juliol de 2020 i previsió per gener de 2021")

Pregunta E2.: Cal presentar alguns dels suports especials per rebre fons per mitigar el posat en marxa per mico la problemàtica per la COVID-19?")problema creat per la COVID-19? ("Ha necessitat acollir-te a algun dels ajust especials que s'han

	E1.SitEconomicaG20	E1.SitEconomicaJ20	E1.SitEconomicaG21
1.Solvent+	0.216	0.150	0.179
2.Solvent-	0.126	0.114	0.123
3.Dia	0.126	0.099	0.116
4.Justos	0.186	0.182	0.149
5.CalAjut	0.153	0.199	0.146
6.NoArribem	0.109	0.174	0.133
7.NC	0.083	0.081	0.153

Taula 22 Taula de proporcions temporals d'E1.SitEconomica per nivells. Cada columna representa la distribució observada de la variable E1 per cada moment del temps

- El nombre de persones amb problemes econòmics augmenta un 23,34% (això explica als que tenen dificultats per resistir-se a tot el mes, als que tenen nous deutes a finals de mes i als que requereixen ajuda econòmica externa per seguir endavant)
- Un 42,8% d'ells està convençut que tindran escassetat econòmica per a gener de 2021

- Un 62,20% dels enquestats tenia algunes necessitats de serveis socials
- Un 46,1% necessitava suport alimentari (d'ells un 64,51% el buscava al ABSS)
- Un 25,00% necessitava suport per pagar el lloguer de la casa (i el 51,85% d'ells la buscava al ABSS)
- Un 11,8% va demanar Renda mínima garantida i un 51,3% d'ells la va buscar al govern català
- Un 15,7% necessitava suport psicològic, i el 51,3% d'ells el va buscar al ABSS
- Un 51,8% necessitava un suport que implicava algun benefici econòmic. Lamentablement, un 70,37% d'ells no va rebre el pagament abans de l'1 de juliol de 2020. Alguns d'ells no podien completar la presentació electrònica per falta d'habilitats digitals, alguns (14,41%) estaven fora dels criteris d'elegibilitat restrictius. Vegeu a la

	E2. 1	**E2. 2.**	**E2. 3.**	**E2. 4**	**E2. 5**	**E2. 6**	**E2. 7**	**E2. 8**	**E2. 9**	**E2. 10.**	**E2. 11**	**E2. 12.**	**E2. 13.**	**E2. 14.**
1.NoNecessita	423	568	591	570	491	487	572	519	517	569	521	543	557	463
2.Ajuntament	254	65	30	70	109	162	43	18	23	31	68	50	14	15
3.Gencat	21	20	14	8	40	10	5	51	23	3	18	7	1	4
4.Gobierno	6	1	1	1	5	4	3	8	32	3	4	1	12	6
5.Entitats	73	14	7	29	18	26	9	4	3	9	14	7	3	5
6.Altres	17	16	22	21	29	26	23	14	15	15	27	21	12	25
7.NC	132	253	288	254	253	236	303	345	343	333	303	331	365	431

Taula 23: Cross taula d'E2.AjutsCOVID19 per nivells.

Una atenció especial requereix les dificultats en les condicions de vida, suavitzades per l'estat d'alarma, ja que tots els processos de desnonament van ser interromputs. No obstant això, tornaran a sorgir en els pròxims mesos:

- Un 27,18% viu en cases socials o comparteix una habitació en un pis (Pregunta F4. Via de vida (Habitatge) del qüestionari)
- Un 27,1% necessitava suport per pagar les factures d'electricitat o gas (i 67.30% d'ells buscaven ajuda al ABSS)
- Un 10,8% d'ajuda necessària per pagar impostos

Impacte social

- Un 15,24% són persones dependents (Pregunta D1: Teniu alguna dependència de grau (deu grau de dependència) del qüestionari?) Vegeu el diagrama de barres a la figura 63 (a)

- D'ells, un 55,40% es refereix a un procés de dependència cada vegada pitjor a partir de gener de 2020 (i un 53,65% d'ells atribueix un empitjorament directe a COVID-19) (Pregunta D2: Creus que el teu nivell de dependència seria diferent si ara et reavaluessin? Creus que si et valoressin ara tindries una variació en el grau de dependència?) Vegeu Taula creuada de T1.1 per nivells a la taula 24

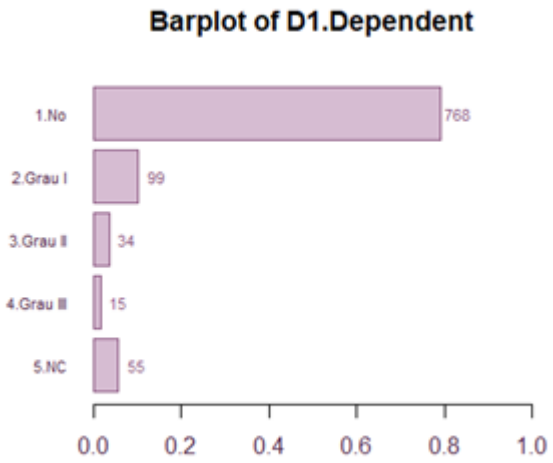

Figura 63: Barplot de D1.Dependent

	T1.1.CuraDepG20	T1.1.CuraDepJ20	T1.1.CuraDepG21
1.<10%	0.344	0.632	0.379
2.10%-30%	0.058	0.058	0.046
3.30%-50%	0.044	0.042	0.047
4.50%-70%	0.024	0.045	0.029
5.>70%	0.044	0.217	0.053
9.NC	0.486	0.005	0.446

Taula 24 T1.1 per nivells

El qüestionari també rep informació de l'altra banda de la dependència. El costat dels cuidadors informals:

- Un 16,99% dels enquestats tenien persones dependents a càrrec al gener de 2020
- El nombre de persones amb dependents a càrrec ha augmentat un 40,43%

Pregunta PC2.1: Quantes persones dependents et tenen al càrrec, segons l'edat? (Quantes persones en Grau I de dependència tens a càrrec en les diferents franges d'edat? (0–11) anys)

Aquesta variable té un nivell de complexitat més, perquè la dependència es classifica en tres grups de severitat creixent introduint una quarta variable a l'anàlisi. Per analitzar aquest element, les tres variables considerades són:

1. Grau dependència: Variable Ordinal amb tres modalitats: Grau I (desamortització menor), Grau II i Grau III (desamortització més alta)

2. Grup d'edat: Variable Ordinal amb 4 modalitats determinades per experts: Nens: (1-11) anys, Adolescents: (12–17)anys, Adults: (18–69)anys, Majors: més de 70 anys
3. Nombre de persones dependents a càrrec: variable discreta: (0,1,2....)

A més, el qüestionari inclou un bloc sencer dedicat a l'ús del temps, des del qual es pot veure:Vegeu el diagrama de barres apilat múltiple de les preguntes PC2, PC3 i PC4 a la Figura 64.

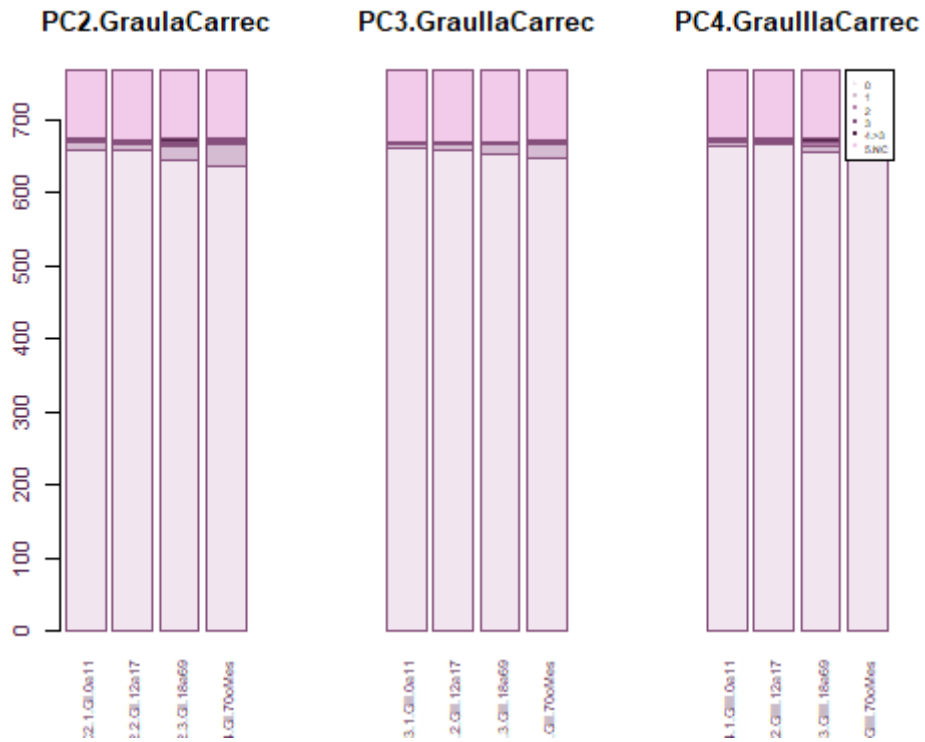

Figura 64 Diagrama de barres múltiples apilades de la pregunta PC2-3-4. Dependents persones a càrrec.

Per a cada grau de severitat, l'anàlisi interna replica l'estructura del gràfic de barres múltiple anterior al qüestionari PC2 amb el nombre de persones dependents de grau I a càrrec del grup d'edat Figura 65

Grau de dependència I:

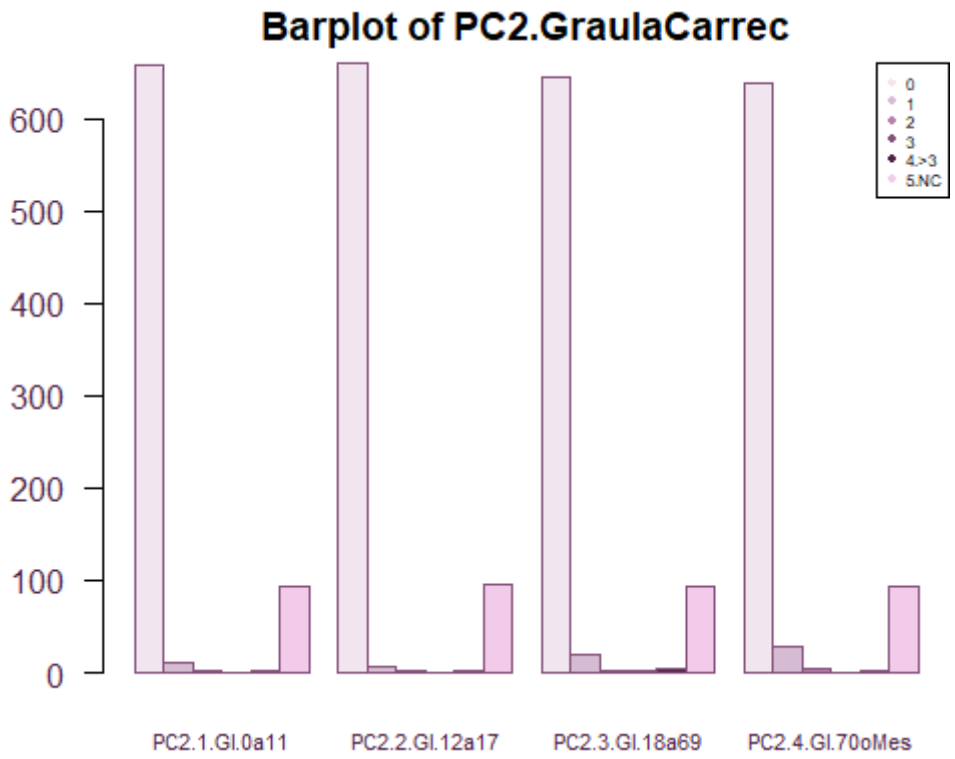

Figura 65. Múltiples bars o persones dependents del grau I que s'encarreguen per grup d'edat.

Vegeu Taula creuada de la pregunta PC2 amb el nombre de persones dependents del grau I que s'encarreguen per grup d'edat taula 25, Taula de proporcions temporals a Figura 30 graella de diagrames de sectors a Figura 66

	PC2.1.GI.0a11	PC2.2.GI.12a17	PC2.3.GI.18a69	PC2.4.GI.70oMes
0	658	660	645	638
1	12	6	20	29
2	3	2	2	4
3	0	1	2	1
4.>3	2	3	5	2
5.NC	93	96	94	94

Taula 25 Taula creuada del PC2.GraulaCarrec per nivells.

	PC2.1.GI.a0a11	PC2.2.GI.12a17	PC2.3.GI.18a69	PC2.4.GI.70oMes
0	0.857	0.859	0.840	0.831
1	0.016	0.008	0.026	0.038
2	0.004	0.003	0.003	0.005
3	0.000	0.001	0.003	0.001
4.>3	0.003	0.004	0.007	0.003
5.NC	0.121	0.125	0.122	0.122

Taula 26 Relació temporal del PC2.GraulaCarrec per nivells.

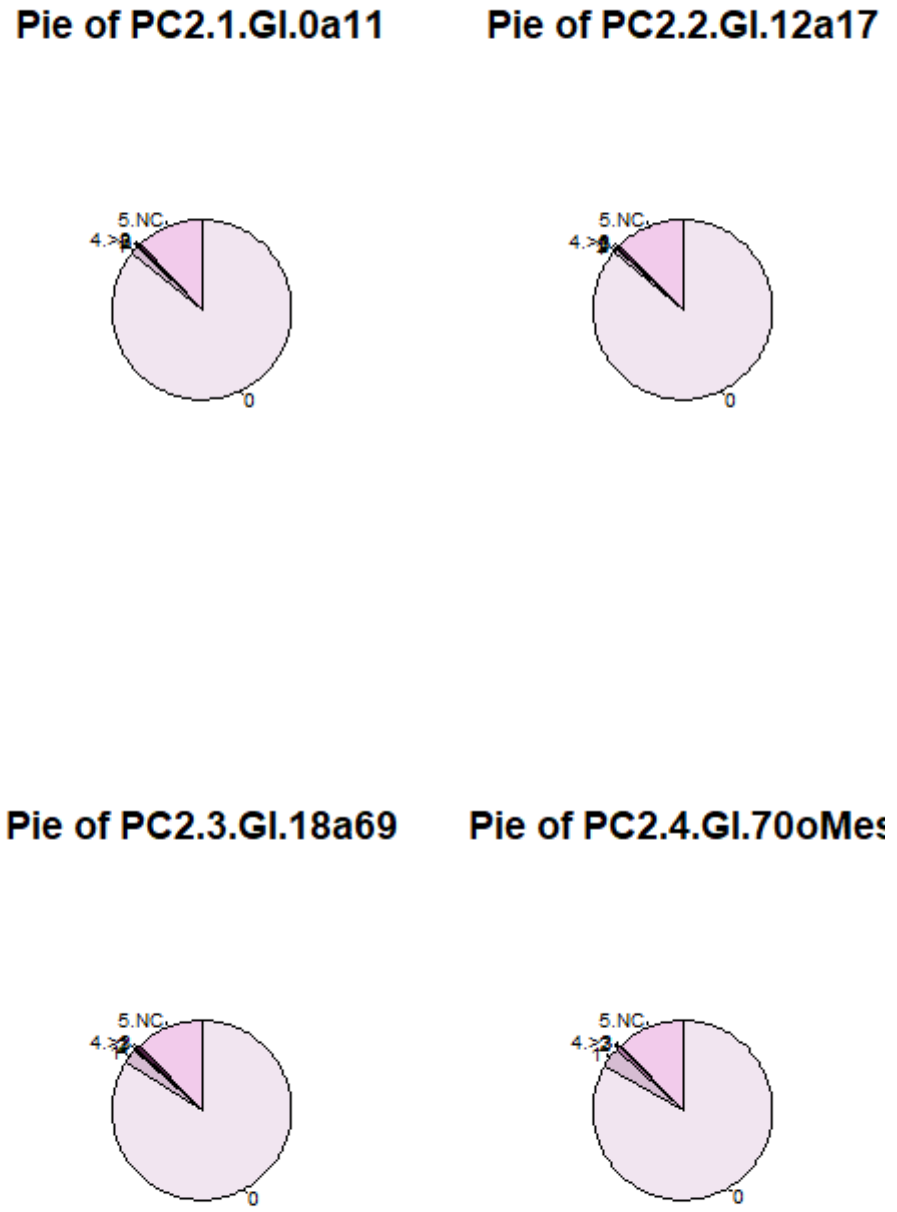

Figura 66 Graella de diagrames de pastís de la pregunta PC2.1-PC25.4.

A més, el qüestionari inclou tot un bloc dedicat a l'ús del temps, des del qual podem veure que:

- Un 29,69% dels cuidadors requereixen ara més temps per a tenir cura de les seves persones dependents a càrrec
- Alguns cuidadors van augmentar la dedicació requerida fins a 5 vegades més que abans de la pandèmia.

Pregunta R1. RelUConv: Com eren de mitjana les relacions en els següents entorns (Com eren majoritàriament les relacions que mantenies amb les persones en els diferents sectors?).

Aquest és un paquet de preguntes que demanen unitat convivencional (Unitat convivencial), Família, Veïns, Amics, Treballadors i altres. En tots ells, el patró "VΛ" s'observa més o menys intensament.

S'observen un total de 93 patrons dels quals 30 es poden llistar, ja que els altres tenen una freqüència massa petita per ser publicats sota la garantia de preservar la secreció estadística.

Vegeu la taula de freqüències de trajectòria a la pregunta R: RelConv a la taula 27:

R1.RelUConv	Freq.	R1.RelUConv	Freq.
01.Satisf+01.Satisf+01.Satisf	596	03.Igno+01.Satisf+01.Satisf	5
10.NC+10.NC+10.NC	64	04.Tenses+04.Tenses+01.Satisf	5
02.Preoc+02.Preoc+02.Preoc	33	01.Satisf+03.Igno+01.Satisf	4
09.Inexistents+09.Inexistents+09.Inexistents	25	01.Satisf+09.Inexistents+10.NC	4
01.Satisf+02.Preoc+02.Preoc	24	02.Preoc+04.Tenses+04.Tenses	4
01.Satisf+02.Preoc+01.Satisf	23	04.Tenses+05.Conflic+01.Satisf	4
02.Preoc+01.Satisf+01.Satisf	14	05.Conflic+01.Satisf+01.Satisf	4
02.Preoc+02.Preoc+01.Satisf	12	09.Inexistents+03.Igno+10.NC	4
01.Satisf+01.Satisf+02.Preoc	9	10.NC+01.Satisf+01.Satisf	4
01.Satisf+01.Satisf+10.NC	7	01.Satisf+03.Igno+03.Igno	3
04.Tenses+04.Tenses+04.Tenses	7	01.Satisf+04.Tenses+04.Tenses	3
09.Inexistents+01.Satisf+01.Satisf	7	03.Igno+03.Igno+03.Igno	3
01.Satisf+04.Tenses+01.Satisf	6	04.Tenses+01.Satisf+01.Satisf	3
02.Preoc+01.Satisf+02.Preoc	6	04.Tenses+04.Tenses+03.Igno	3
05.Conflic+05.Conflic+05.Conflic	6	05.Conflic+04.Tenses+01.Satisf	3

Taula 27+Taula de freqüència de trajectòria per a la pregunta R1: Rel UConv.

L'informe automàtic proporciona el gràfic múltiple bivariant i la taula de freqüències i la graella de diagrames de sectors també. Aquí es mostra la taula de proporcions. Vegeu Taula 28.

	R1.RelUConvG20	R2.RelUConvJ20	R3.RelUConvG21
01.Satisf	0.724	0.684	0.734
02.Preoc	0.083	0.107	0.087
03.Igno	0.009	0.022	0.020
04.Tenses	0.028	0.042	0.020
05.Conflic	0.022	0.020	0.009
06.VioVerb	0.006	0.002	0.001
07.VioPsi	0.005	0.006	0.004
08.VioFis	0.003	0.002	0.000
09.Inexistents	0.045	0.039	0.033
10.NC	0.074	0.076	0.093

Taula 28 Proporcions de R1.RelUConv per hora.

La taula 29 mostra la taula de transició entre juliol de 2020 i gener de 2021, i es pot veure quins canvis en la qualitat de les relacions són més freqüents. Durant el confinament, un 7,92% dels participants van passar de relacions satisfactòries amb persones que vivien en la mateixa llar a situacions pitjors (la majoria d'ells a relacions preocupants o tibants), mentre que un 4,53% va millorar les seves relacions inicials a satisfactòries

	01. Satisf	02. Preoc	03. Igno	04. Tenses	05. Conflic	06. VioVerb	07. VioPsi	08.VioFis	09.Inexistents	10.NC
01.Satisf	635	15	2	2	0	0	0	3	7	635
02.Preoc	37	60	3	2	0	0	1	0	1	37
03.Igno	7	1	8	0	1	0	0	0	4	7
04.Tenses	16	5	4	14	0	0	0	0	2	16
05.Conflic	8	2	0	0	7	0	1	0	1	8
06.VioVerb	0	0	0	0	0	1	0	0	1	0
07.VioPsi	2	0	0	1	1	0	2	0	0	2
08.VioFis	2	0	0	0	0	0	0	0	0	2
09.Inexistents	2	1	1	0	0	0	0	28	6	2
10.NC	4	0	1	0	0	0	0	1	68	4

Taula 29. Canvis esperats juliol 2020—gener 2021.

Vegeu els canvis reportats al llarg del temps en la pregunta R1,RelUConv per temps en la Figura 67.

Vegeu el diagrama de teler a la Figura 67 i canvieu els patrons a lataula 30 sobre la relació amb els amics. Vegeu els patrons de canvi en les relacions laborals a la taula 31

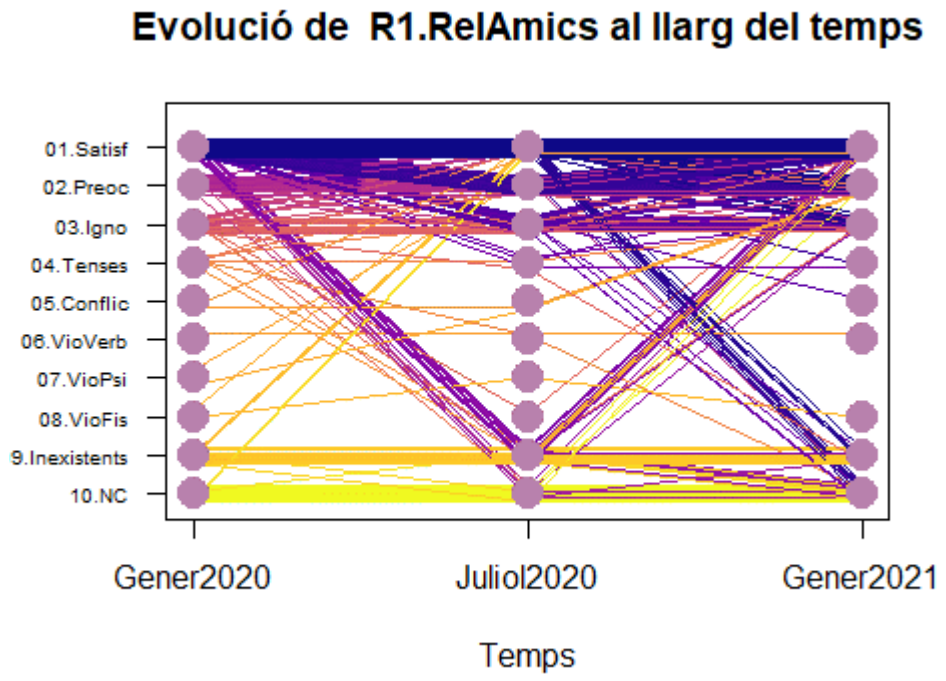

Figura 67Diagrama de telerde les relacions amb els amics.

	Freq	Prop
Improve	44	0.045
V pattern	7	0.007
Balance	677	0.697
Λ pattern	5	0.005
Enworse	84	0.087

Taula 30: Canvia els patrons en les relacions amb els amics.

	Freq	Prop
Improve	60	0.062
V pattern	4	0.004
Balance	523	0.539
Λ pattern	12	0.012
Enworse	64	0.066

Taula 31Canvia els patrons en les relacions laborals

El patró "VΛ" apareix de nou aquí, amb una certa proporció de persones que es comporta més amb altres persones durant la pandèmia, i les que se senten més aïllades

Pregunta Soc4.: La pandèmia han creat vincles amb altres persones (família, amics, veïns, etc.)? La pandèmia: T'ha creat vincles d'unió amb altres persones (família, amistats, veïnatge, etc.) Vegeu el diagrama de barplot de sentiments d'aïllament durant la pandèmia a la Figura68 i el gràfic de barres d'intensificació d'enllaços a causa de la pandèmia a la Figura 69. A la taula 32 es mostra la taula de freqüències dels sentiments d'aïllament i a la taula 33 de freqüències d'intensificació en enllaços

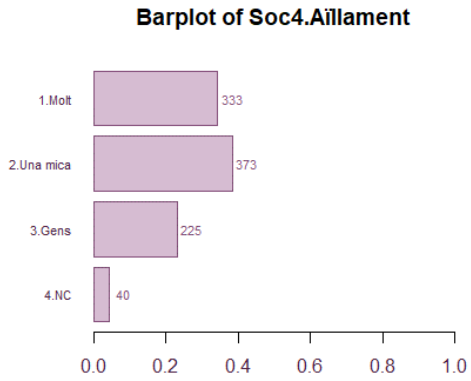

Figura 68: Barplot de sentiments d'aïllament durant la pandèmia

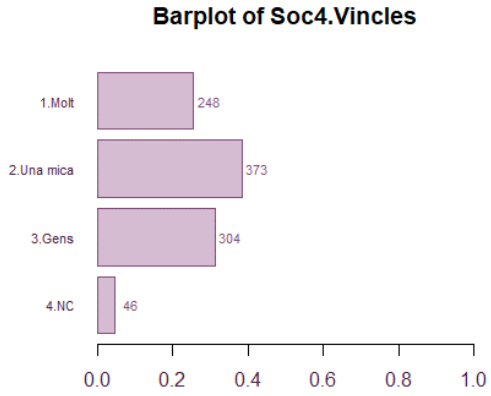

Figura 69: Barplot de intensificació d'enllaços a causa de la pandèmia

Soc4.Aïllament	Freq.	Prop.	Std. Err
2.Una mica	373	0.384	0.0155
1.Molt	333	0.343	0.0152
3.Gens	225	0.232	0.0134
4.NC	40	0.041	0.0063

Taula 32: Taula de freqüència dels sentiments d'aïllament durant la pandèmia

Soc4.Vincles	Freq.	Prop.	Std. Err
2.Una mica	373	0.384	0.0155
3.Gens	304	0.313	0.0148
1.Molt	248	0.255	0.0141
4.NC	46	0.047	0.0071

Taula 33: Taula de freqüència de la identificació dels enllaços a causa de la pandèmia

Pregunta: T'has sentit o et sents sol?

Els resultats es mostren en diferents figures. Vegeu el mapa de trajectòria dels sentiments solitaris a la Figura 70, la taula de friccions de les Trajectories a la taula 34, múltiples barplot a la Figura 71 i proporcions per nivell a la taula 35. Vegeu Graella de diagrames de sectors a la Figura 72. Els canvis gener 2020–juliol 2020 es mostren a la taula 36 i veuen els canvis previstos al juliol 2020–gener 2021 a la taula 37

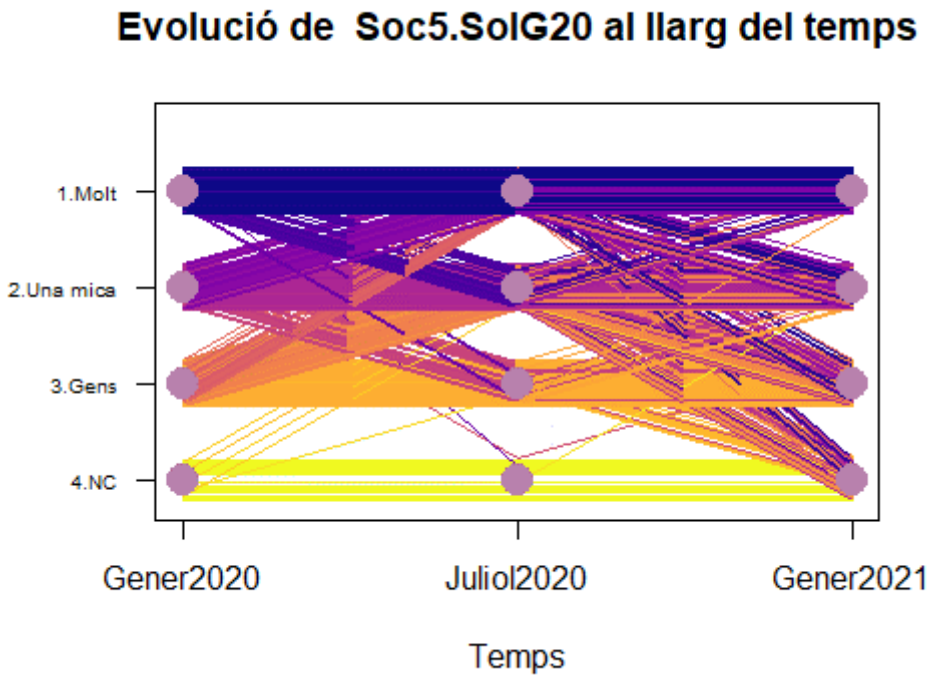

Figura 70: Diagrama de teler dels sentimentde solitud

Es relacionen tot seguits les combinacions amb freqüència mínima de 3

Soc5.SolG20	Freqüències	Soc5.SolG20	Freqüències
3.Gens+3.Gens+3.Gens	277	3.Gens+1.Molt+1.Molt	13
2.Una mica+2.Una mica+2.Una mica	142	3.Gens+1.Molt+2.Una mica	12
1.Molt+1.Molt+1.Molt	80	2.Una mica+2.Una mica+1.Molt	9
3.Gens+2.Una mica+2.Una mica	59	1.Molt+1.Molt+4.NC	8
3.Gens+2.Una mica+3.Gens	45	3.Gens+3.Gens+2.Una mica	8
2.Una mica+1.Molt+1.Molt	37	1.Molt+1.Molt+3.Gens	7
2.Una mica+1.Molt+2.Una mica	32	3.Gens+1.Molt+3.Gens	7
4.NC+4.NC+4.NC	31	3.Gens+2.Una mica+1.Molt	7
2.Una mica+2.Una mica+3.Gens	26	1.Molt+2.Una mica+1.Molt	6
2.Una mica+3.Gens+3.Gens	24	1.Molt+2.Una mica+3.Gens	6
3.Gens+2.Una mica+4.NC	21	1.Molt+2.Una mica+4.NC	5
2.Una mica+2.Una mica+4.NC	19	2.Una mica+1.Molt+4.NC	5
3.Gens+3.Gens+4.NC	17	3.Gens+1.Molt+4.NC	5
1.Molt+1.Molt+2.Una mica	16	1.Molt+3.Gens+3.Gens	4
1.Molt+2.Una mica+2.Una mica	15	2.Una mica+3.Gens+2.Una mica	3
2.Una mica+1.Molt+3.Gens	13	2.Una mica+3.Gens+4.NC	3

Taula 34: Taula de freqüències de les trajectòries

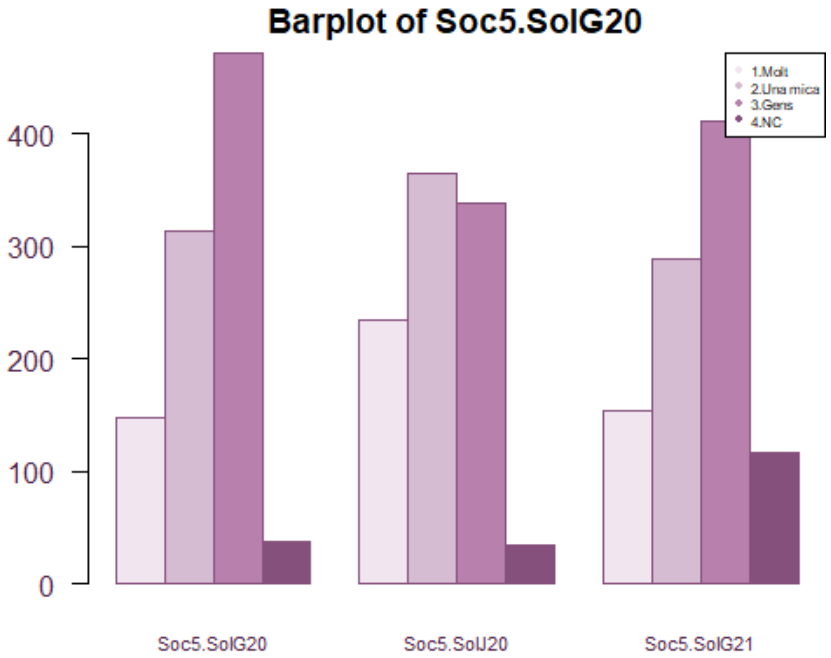

Figura 71: Diagrama múltiple de barres de sentiments de solitud

(a)

	Soc5.SolG20	Soc5.Sol	Soc5.SolG21
1.Molt	0.152	0.242	0.158
2.Una mica	0.323	0.375	0.298
3.Gens	0.486	0.348	0.424
4.NC	0.038	0.035	0.120

Taula 35: Proporcions de Soc5.SolG20 per nivells

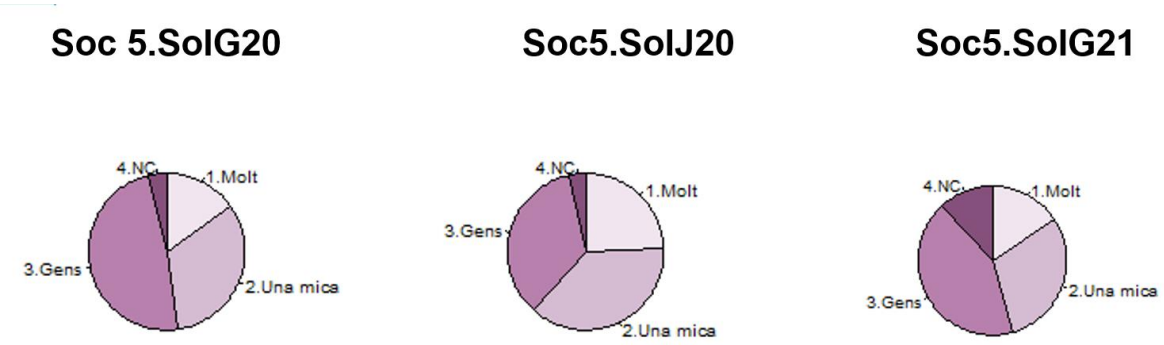

Figura 72: *Graella de gràfics de pastís per a sentiments de solitud.*

	1.Molt	2.Una mica	3.Gens	4.NC
1.Molt	111	32	4	1
2.Una mica	87	196	30	1
3.Gens	37	132	303	0
4.NC	0	4	1	32

Taula 36 Canvis gener 2020–juliol 2020

	1.Molt	2.Una mica	3.Gens	4.NC
1.Molt	130	60	27	18
2.Una mica	22	217	78	47
3.Gens	1	11	306	20
4.NC	0	1	1	32

Taula 37: Canvis planificats juliol 2020–gener 2021

- Un 72,7% dels enquestats se sentia més aïllat
- El sentiment de solitud augmenta un 29,3% i aquest és un patró seguit principalment per dones (70%) de més de 60 anys (de mitjana), vivint soles, amb alguna manca d'habilitats digitals i un 52,18% d'elles que requereixen suport emocional durant la pandèmia.
- Un 41,45% dels participants requeria suport psicològic a causa de la COVID-19 i d'ells un 30,95% requeria suport emocional.
- Un 23,58% dels participants van fer referència a algun trastorn mental al gener de 2020. D'ells: un 96,94% va rebre tractament farmacològic. Entre les persones amb trastorns mentals, el 46,28% pateixen depressió i el 58,51% pateixen trastorn d'ansietat.
-
- El 73,78% de les persones amb trastorns mentals declarats com a pitjors al juliol de 2020
- Un 68,86% de persones amb depressió declara sentir-se pitjor al juliol de 2020
- Un 72,3% de les persones amb ansietat declaren sentir-se pitjors per a juliol de 2020
- Un 5,12% de les persones sense trastorns mentals al gener de 2020 declaren sentir-se pitjor al juliol de 2020
- Un 57,93% de les persones amb discapacitat se senten pitjors i un 58,33% d'elles atribueix una deterioració de la COVID-19

A més, un 49,64% dels participants s'han teletreballat o han seguit formació en línia durant la pandèmia, mentre que només un 5,1% d'ells ja va fer teleformacio al gener de 2020. El juliol de 2020, un 21,84% de les persones involucrades en les teleactivitats (treball o educació) van patir l'impacte en les activitats assistencials (relatives, ancians, nens...). Un 54,4% d'ells requeria suport emocional.

Vegeu la taula de freqüències dels trastorns mentals a la taula 38 i al diagrama de barres Marginal dels trastorns mentals a la figura 73

Salut Mental	Freq.	Prop.	95% CI error
S4.SM2G20Ansietat	134	0.585	0.0326
S4.SMG20Depressio	106	0.463	0.0330
S4.SMG20Altres	26	0.114	0.0210
S4.SMG20TMG	23	0.100	0.0197
S4.SMG20TLP	14	0.061	0.0158
S4.SMG20TPT	11	0.048	0.0141

95% CI error: ± 0.0021; SE: 0.0239

Taula 38: Taula de freqüència de trastorns mentals.

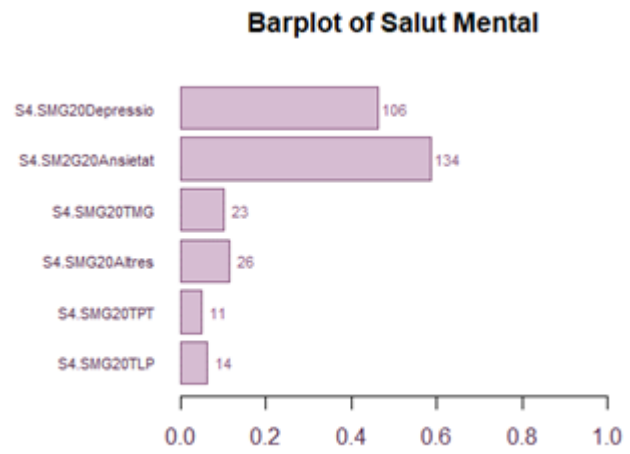

Figura 73 Diagrama de barres marginals de trastorns mentals

Violència

Entre les opcions per escollir la qualitat de les relacions en diferents entorns (preguntes R1 a R9 del qüestionari), opcions particulars que es pregunten si la persona està sent objecte de violència, sigui físicament emocional o psíquica. En total, un 6,38% dels enquestats es declara víctima d'alguna forma de violència. D'ells, el 72,58% són dones. L'estatus civil, la professió i el nivell acadèmic són transversals entre aquests grups (17,4% tenen estudis universitaris). Una característica comuna d'aquestes persones és que el 90,33% de les persones tenen

precariat laboral (sense contracte estable o temporal, però altres formes irregulars de treball o desocupació). El qüestionari planteja dues preguntes addicionals per obtenir més detalls sobre el patró de l'agressor i les forces equilibrin amb la víctima.

Pregunta R4. Agressor: Si heu estat objecte de violència, qui la realitza? ("Si té indicat ser objecte de violència, qui exerceix aquesta violència?") Vegeu la taula de freqüències de tipus d'agressor a la taula 39 i el diagrama de barres que mostra qui és l'agressor a figura 74 Vegeu la taula de freqüències de R4.Agressor a la taula 40

R4.Agressor	Freq.	Prop.	Std. Err
1.NoViolencia	775	0.798	0.0130
5.NC	125	0.129	0.0110
3.Igual	58	0.060	0.0077
2.Superior	27	0.028	0.0055
4.Subaltern	8	0.008	0.0032

Taula 39: Taula de freqüència de tipus agressor

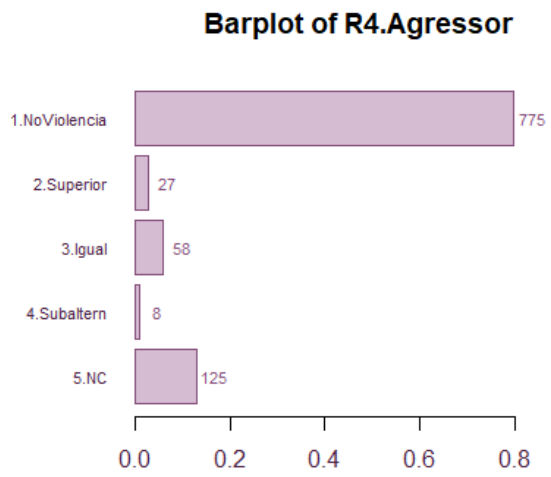

Figura 74 Qui és l'agressor?

R4.Agressor	Freq.	Prop.	Std. Err
1.NoViolencia	764	0.787	0.0130
5.NC	117	0.120	0.0105
3. Un igual (germà, company, amic, veí…)	48	0.049	0.0071
2. Un superior o ascendent (pare, tiet, responsable de feina, professor…)	20	0.021	0.0045
1.NoViolencia;5.NC	7	0.007	0.0032
4. Un subaltern o descendent (fills, empleats, …)	5	0.005	0.0032

2. Un superior o ascendent (pare, tiet, responsable de feina, professor…);3. Un igual (germà, company, amic, veí…)	4	0.004	0.0000

Taula 40:. Taula de freqüència.

Consum de substàncies

Els blocs S6, S7 i S8 apliquen la qüestió de l'abús de substàncies. Des del punt de vista estructural, és una variable TQQ i l'anàlisi descriptiva bàsica de les dades originals proporciona informació sobre la combinació de substàncies que la gent segueix, però més interessant és la informació que podem obtenir transformant les dades originals en nous indicadors, tal com es descriu a l'apartat 3.2. Es van crear noves variables, com les que informen del nivell d'ús o les diferents substàncies en els tres segells de temps objectiu. Fig 7 visualitza aquesta informació.

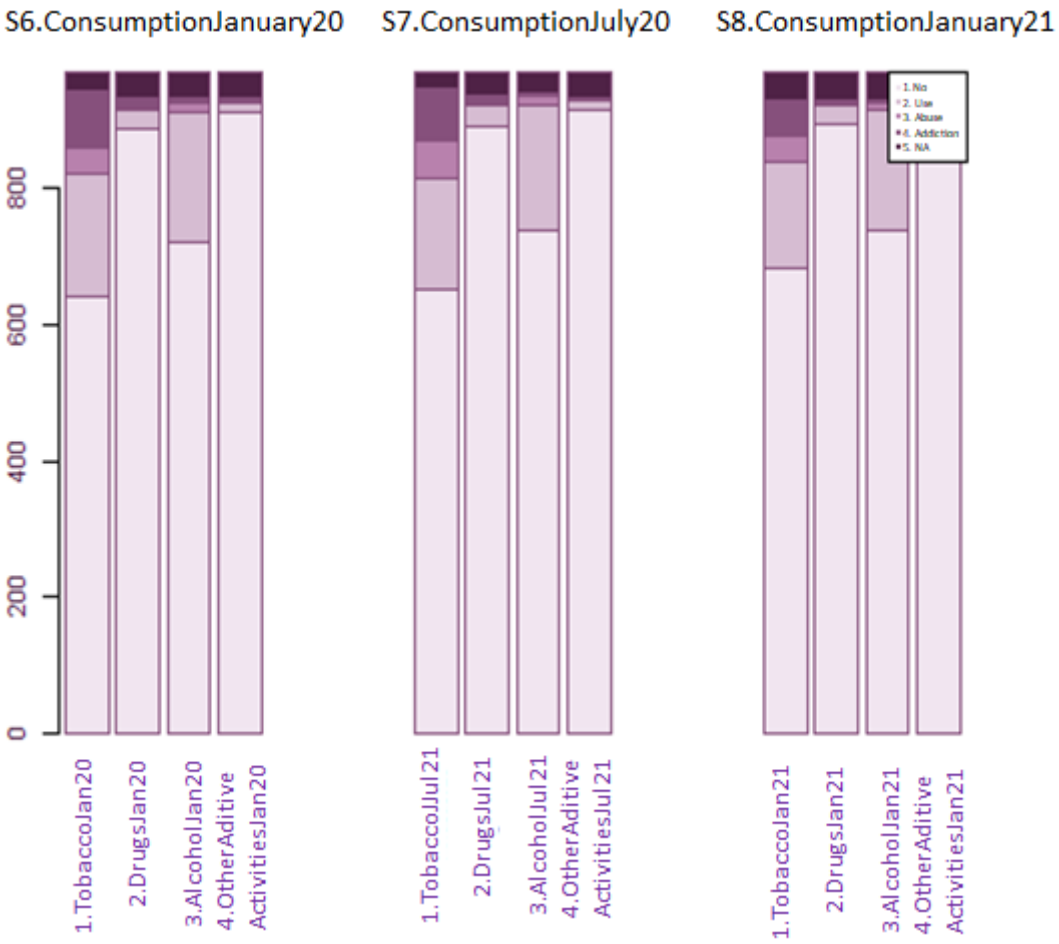

Figura 75: Diagrama de barres múltiple d'abús de substàncies.

En aquest cas, una visualització alternativa d'aquest tipus de dades s'utilitza per entendre millor l'evolució del consum de substàncies al llarg del temps. La Figura 76 mostra com el consum de substàncies canvia al llarg del temps.

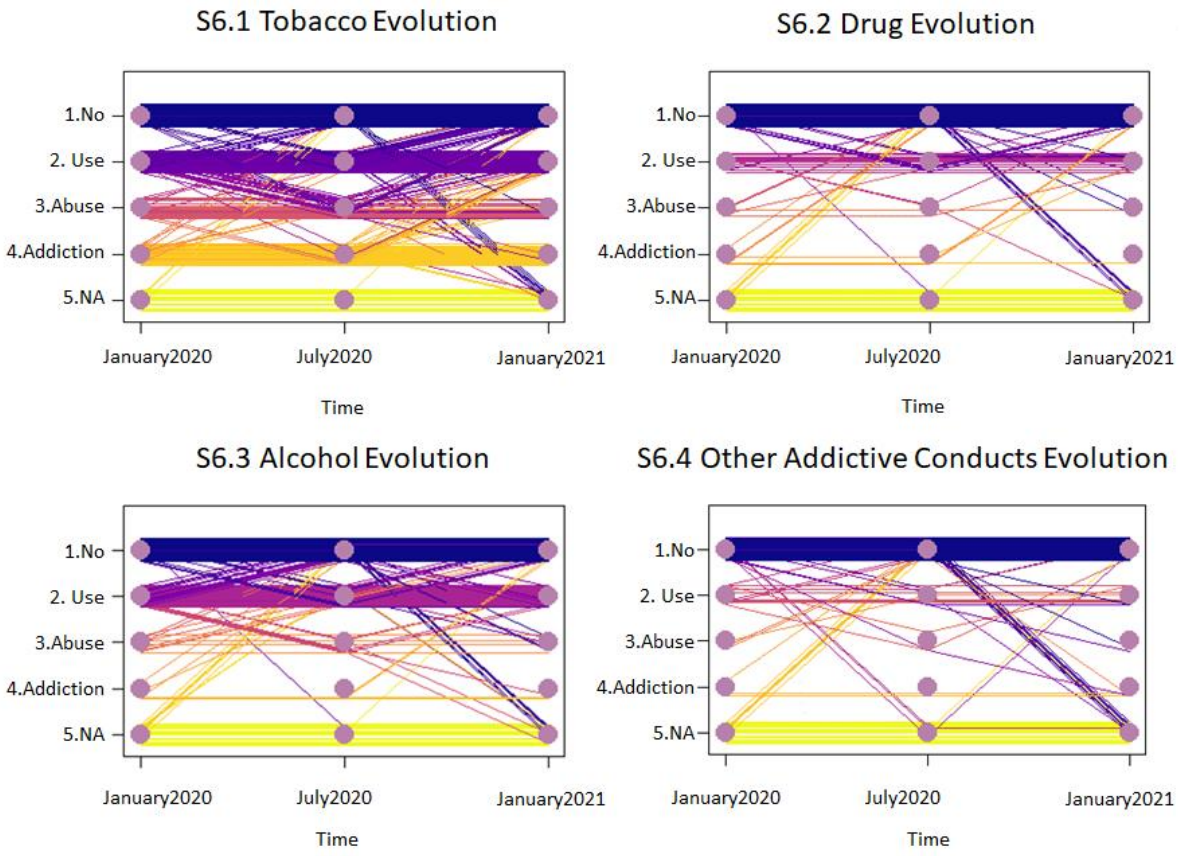

Figura 76:Diagrama de sectors d'abús de substàncies.

En aquest bloc podem veure (Figura 75) que el tabac és la substància més consumida i el nombre de persones que el consumeixen lleugerament va disminuir al llarg del període estudiat. De fet, el 30,8% de les persones eren fumadors al gener de 2020, i va disminuir al 25,5% al gener de 2021. També podem veure que:

- El tabac és la substància amb més addictes (1,8% al gener de 2020)
- La majoria de les persones no utilitzen substàncies.

A més, Fig 76 mostra que:

- En totes les substàncies, la majoria de les persones no varien el seu estat d'addicció (85% per al tabac, 92% per a les drogues, 88% per a l'alcohol, 92% per a altres substàncies)
- En totes les variables, hi ha algunes persones que no van respondre al gener de 2020 i gener de 2021 però van respondre "no" al juliol, per això en tots els diagremes de teler apareix aquest patró ..

7.2.7. Fase III: Síntesi dels resultats

A continuació, sintetitzem els resultats d'aplicar els scripts intel·ligents automàtics a tot el conjunt de dades.

IMPACTE ECONÒMIC I LABORAL

- El nombre de persones que no treballen i no reben cap benefici augmenta un 50%
- El nombre de persones que no treballen i reben algun benefici augmenta un 17,6%.
- El nombre de persones que no tenen ocupació o ocupació augmenta un 11%
- El nombre de persones que han estat llicenciades o plegades augmenta un 78.04%
- El nombre de persones que han reduït la seva jornada laboral augmenta un 36,36%.
- El nombre de persones que tenien el seu propi negoci i deixaven de treballar durant el confinament o entraven en fallida va augmentar un 110%
- El nombre de persones amb condicions de treball no precàries va disminuir un 41,62%.
- El 51,25% de les persones sense treball tenen por de no treballar encara al gener de 2021 (els motius més esmentats són que moltes empreses van tancar a causa de la COVID-19, després que disminueixin de nou les possibilitats d'edat per a contreure una altra vegada, per als certs sectors, la gent té por d'estar infectada per l'empleat i prefereix no contractar nous treballadors).
- El 51,25% de les persones sense treball tenen por de no treballar encara al gener de 2021
- El nombre de persones amb problemes econòmics augmenta un 23,34%.
- Un 42,8% d'ells està convençut que tindran escassetat econòmica per a gener de 2021
- Un 62,20% dels enquestats tenia algunes necessitats de serveis socials
- Un 46,1% necessitava suport alimentari (d'ells un 64,51% el buscava a l'ABSS)
- Un 25,00% necessitava suport per pagar el lloguer de la casa (i el 51,85% d'ells la buscava al ABSS)
- Un 11,8% va demanar Renda mínima garantida i un 51,3% d'ells la va buscar al govern català
- Un 15,7% necessitava suport psicològic, i el 51,3% d'ells el va buscar a l'ABSS
- Un 51,8% necessitava un suport que implicava algun benefici econòmic. Per desgràcia, un 70,37% d'ells no van rebre el pagament abans de l'1 de juliol de 2020. Alguns d'ells no podien completar la tramesa electrònica per falta d'habilitats digitals, alguns (14,41%) estaven fora dels criteris d'elegibilitat restrictius

- Un 27,18% viu en cases socials o comparteix una habitació en un pis
- Un 27,1% necessitava suport per pagar les factures d'electricitat o gas (i 67.30% d'ells buscaven ajuda a l'ABSS)
- Un 10,8% necessitava ajuda institucional per pagar impostos i tributs
- Un 51,8% dels participants van presentar sol·licituds de suport econòmic a les administracions durant la primera onada
- El 70,37% dels sol·licitants d'ajuda econòmica no van rebre ni un cèntim abans de l'1 de juliol de 2020 (esmenten diverses raons entre les quals podem destacar el retard en les resolucions, les dificultats per a presentar la proposta, l'impacte de la bretxa digital de fer l'aplicació digital, els criteris restrictius d'elegibilitat que van deixar exclosos un 14,41% de les persones que declaren necessitar el suport).

IMPACTE SOCIAL

- Un 67,5% dels participants al projecte INSESS-COVID19 són dones.
- Un 15,24% dels participants són persones dependents
- A partir d'ells un 55,40% es refereix a un procés de dependència cada vegada pitjor a partir de gener de 2020 (i un 53,65% d'ells atribueix un empitjorament directe a COVID-19)
- Un 16,99% dels enquestats tenien persones dependents a càrrec al gener de 2020
- El nombre de persones amb dependents a càrrec va augmentar un 40,43% al juliol
- Un 18,02% dels enquestats no van declarar la dedicació a persones dependents al gener i van declarar dedicar més del 70% del seu temps diari a aquest assumpte al juliol
- Un 72,7% dels enquestats se sentia més aïllat
- El sentiment de solitud augmenta un 29,3% i aquest és un patró seguit principalment per dones (70%) de més de 60 anys (de mitjana), vivint soles, amb alguna manca d'habilitats digitals i un 52,18% d'elles que requereixen suport emocional durant la pandèmia.
- Un 41,45% dels participants requeria suport psicològic a causa de la COVID-19 i d'ells un 30,95% requeria suport emocional.
- El 73,78% de les persones amb trastorns mentals declarats com a pitjors al juliol de 2020
- Un 68,86% de persones amb depressió declara sentir-se pitjor al juliol de 2020
- Un 72,3% de les persones amb ansietat declaren sentir-se pitjors per a juliol de 2020
- Un 5,12% de les persones sense trastorns mentals al gener de 2020 declaren sentir-se pitjor al juliol de 2020
- Un 57,93% de les persones amb discapacitat se senten pitjors i un 58,33% d'elles atribueix una deterioració de la COVID-19

- Un 54,4% de les persones que feien teletreball o teleeducació durant la pandèmia requerien suport emocional

VIOLÈNCIA

- En total, un 6,38% dels enquestats declaren ser víctimes d'alguna forma de violència.
- D'ells, un 72,58% són dones.
- L'estat civil, la professió i el nivell acadèmic són transversals entre aquests grups (17,4% tenen estudis universitaris).
- Un 90,33% de les persones es troben en una escassetat laboral.
- Sovint la persona és víctima de més d'un agressor simultàniament: 14,51% reb la violència de dos perfils dels agressors simultàniament; 27,42% de tres
- En un 93,54% dels casos, l'agressor comparteix una relació d'igualtat amb la víctima, sent veí, amic, germà...
- En un 45,55% dels casos l'agressor té una relació de poder amb la víctima, sent pare, cap, professor, etc.

SALUT

- El 12,25% dels participants havien patit COVID-19 entre març de 2020 i desembre de 2020.
- 32% dels participants pertanyen a un grup de risc COVID-19.
- 15,75% dels participants pateixen alguna discapacitat
- En termes de discapacitat, el 15,75% dels participants tenen una discapacitat, i la majoria són discapacitats físiques (81,04% de les persones amb discapacitat). D'aquests, el 57,93% diuen que han empitjorat al juliol de 2020 en comparació amb gener i d'aquests 58,33% l'atribueixen a COVID-19.
- En cas de grau de dependència, el 14,9% té un grau de dependència declarat i el 14,2% de les persones declara que el seu grau de dependència és pitjor

SALUT MENTAL

La majoria de les persones no informen de cap trastorn de salut mental

El 23,58% dels participants van informar d'algun problema de salut mental el gener de 2020

- El perfil psiquiàtric observat més prevalent és el de l'ansietat, seguit de la depressió al gener de 2020. Després, el tercer grup va patir una combinació d'ansietat i depressió simultàniament. Després que la primera onada tingui aquests dos esdevingui més prevalent que la depressió sola
- Hi ha persones que pateixen molts trastorns mentals junts, però són poques persones en el món

Analitzant les altres variables de la salut mental podem veure que:

- El 38% dels participants van necessitar suport emocional fins a juliol de 2020, però la seva pretensió és que el suport emocional disminueix per a gener de 2021 a 33,47%.
- El 22% dels participants reben medicaments.
- El 54% de la discapacitat intel·lectual declara que se sent pitjor per a juliol de 2020.

La salut mental va ser un dels impactes més rellevants produïts per la crisi de COVID19 que afectava a tota mena de persones.

7.2.8. Fase III Ampliació del preprocessament amb generanció de variables derivades. Variables de segona generació basades en el coneixement

La idea principal és introduir noves variables que representin els criteris de raonament expert. S'utilitzen els principis metodològics anunciats a la secció 5.15.1. En aquesta recerca, segons els experts, es van crear les següents variables basades en el coneixement.

Creació de la variable d'ubicació

Es va demanar als participants el municipi on viuen. No obstant això, aquesta variable no és útil per a aconseguir el nostre objectiu. A Catalunya hi ha 947 municipis. Cada ciutat pertany a un ABSS, de manera que es crea una nova variable a la base de dades anomenada ABSS. Els experts ens van donar la llista de municipis pertanyents a cada ABSS i vam crear la nova variable anomenada ABSS, que indica el ABSS de residència del participant.

Indicadors

DUMMIES BÀSIQUES

A partir de S4 es crea un conjunt de sis variables fictícies que indiquen si la persona pateix de cada trastorn de salut mental (d) considerat en les modalitats de la variable S4 original, en una marca horària (t) .

Això correspon al cas de la creació de nous indicadors indicats en el 4.2

Les noves variables resultants són a la taula 41

$$X_{df} = \begin{cases} 1 \text{ si la persona té el trastorn de salut mental d al moment t} \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad altrament \end{cases} \quad (18)$$

d (Trastorn Salut Mental)	January 2020
SMD (Severe Mental Disorder)	S4.SMDJan20
PLD (Transtorn Límit de la Personalitat)	S4.PLDJan20
PSTD (Estrès Post Traumàtic)	S4.PSTDJan20
Depressió	S4.DepressionJan20
Ansietat	S4.AnxietyJan20
Altres	S4.OthersJan20

Taula 41 Variables resultants

DUMMIES DINÀMIQUES

Dummies dinàmiques sobre l'evolució dels diagnòstics de salut mental. Aquests dummies mostren si la persona va millorar o empitjorar en cada diagnòstic de salut mental durant la primera onada de la pandèmia. Aquestes són les noves variables de tipus "Condició" de les descrites a la secció 5.15.1. Es creen un total de 18 variables binàries basades en l'evolució de cada diagnòstic de salut mental entre gener de 2020 i juliol de 2020. Veure'ls a la taula 42

D	Better in July 2020	Equal in July 2020	Worse in July 2020
SMD (Severe Mental Disorder)	S4.SMDJul20+	S4.SMDJuly20=	S4.SMDJuly20-
PLD (Personality Limit Disorder)	S4.PLD Jul20+	S4.PLD Jul20=	S4.PLD Jul20-
PSTD (Posttraumatic Stress Disorder)	S4.PSTDJul20+	S4.PSTDJul20=	S4.PSTDJul20-
Depression	S4.DepressionJul20+	S4.DepressionJul20=	S4.DepressionJul20-
Anxiety	S4.AnxietyJul20+	S4.AnxietyJul20=	S4.AnxietyJul20-
Others	S4.OthersJul20+	S4.OthersJul20=	S4.OthersJul20-

Taula 42 Dummies dinàmiques creades

DEBUT EN SALUT MENTAL

Debut en el trastorn de salut mental: També, una nova variable per saber si la persona ha debutat en salut mental durant la primera onada. La nova variable s'anomena S4. Problema mental Debut durant la pandèmia 1a onada (S4. MentDebut): s'ha creat amb una combinació de les variables creades en el pas anterior. Es considera que una persona debuta amb un problema de salut mental si no tenia el diagnòstic al gener de 2020 i que el té al juliol de 2020 sentint-se pitjor que abans. Aquesta és una nova variable de tipus Condicio i es crea de la manera següent:

$$S4.MentDebut = \begin{cases} SÍ & \text{SI } \forall d, X_{dJan20} = \{NO\} \; and \; \exists d, X_{dJul20} = \{Pijtor\} \\ NO & altrament \end{cases} \quad (19)$$

TRANSTORN MENTAL AL GENER DE 2020

Transtorn Mental al gener de 2020 (s4.MentalDisorderG20): Aquesta és una variable binària D={SI, NO} que mostra si la persona té algun trastorn mental diagnosticat. Aquestes són variables de tipus Indicador.

$$S4.MentalDisorderG20 = \begin{cases} YES \quad si\ \exists d, X_{dJan20} = \{SI\} \\ NO \quad si\ \forall d, d, X_{dJan20} = \{NO\} \end{cases} \quad (20)$$

Condicions multivariables

TRANSTORN MENTAL AL GENER DE 2020

Trastorn mental al gener de 2020 rebent medicació (S4S5.MedMental): Aquesta és una variable binària D={YES, NO} que mostra si les persones amb algun trastorn mental diagnosticat estan rebent medicació, que és un indicador de gravetat. Són de tipus de condició.

$$S4S5.MedMental = \begin{cases} YES \ \ if\ S4.MentalDisorderG20 = \{SI\}\ and\ S5 = \{\text{SI}\} \\ NO \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad altrament \end{cases} \quad (21)$$

TRANSTORN MENTAL GREU

La majoria dels trastorns mentals impactants al gener de 2020 reben medicació (S4S5.MedMentalSevere: Atès que la COVID-19 i el confinament tenen un impacte greu en certs trastorns mentals, aquesta és una variable binària D={SI, NO} que mostra si una persona que pateix els trastorns mentals més greus (SMD, PLD o PSTD està rebent medicació. Això indica que al gener aquesta persona ja tenia un deteriorament de salut mental bastant greu. És del tipus "Condició".

$$S4S5.MedMentalSevere = \begin{cases} SI \text{ si } \exists d \quad in\ \{[SVM, PLD, PSD]\} such\ that\ X_{dJan20} = \{SI\}\ and\ S5 = \{SI\} \\ NO \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad altrament \end{cases} \quad (22)$$

MENORS DE 30

Menors de 30 (U30): En bloc sociodemogràfic hi ha la pregunta P3.Edat del participant. Una nova variable binària U30={SÍ, NO} mostra si el participant és menor de 30 anys.

$$U30 = \begin{cases} SI \text{ si } P3.age < 30 \\ SI \text{ si } P3.age \geq 30 \end{cases} \quad (23)$$

Comptadors

Comorbiditat de salut mental: és de tipus Comptador i indica quants diagnòstics de salut mental tenen la persona abans del confinament (si una persona té un diagnòstic d'ansietat i depressió simultàniament, aquesta variable pren valor 2, i així successivament)

Amb aquestes noves variables, tenim una millor capacitat d'aprendre sobre la situació de salut mental dels enquestats. Analitzant els resultats (mostrats a la Taula 43) de les variables de segona generació les conclusions són més riques i afegeixen valor a l'anàlisi. Amb aquestes noves variables creades, és fàcil veure algunes peces noves de coneixement que no eren evidents sobre una anàlisi simple de les dades originals:

A partir del conjunt de variables binàries creades més amunt es pot veure la situació de la línia de base abans del confinament

Mental Health	Freq.	Prop	95% CI error
S4.AnxietyJan20	134	0.585	0.0326
S4.DepresionJan20	106	0.463	0.0330
S4.OthersJan20	26	0.114	0.0210
S4.SMDJan20	23	0.100	0.0197
S4.PLDJan20	14	0.061	0.0158
S4.PSTDJan20	11	0.048	0.0141

Taula 43 Resultats bàsics de dummies

- Un 24% dels participants declaren estar diagnosticat sobre salut mental. D'aquests, el 46,28% declara depressió. Un 58,51% es veu afectat per l'ansietat. El 10% pateix un trastorn de salut mental greu i un 6,1% d'un trastorn de personalitat límit. Un 4,8% es deu al trastorn posttraumàtic i un 11,4% a altres trastorns mentals.
- D'aquests, també es pot veure que un 96,94% d'ells estan rebent medicació.
- En la figura 77 es mostren més resultats. El gran cercle és un percentatge de tots els participants. El percentatge dels cercles petits és dades relatives de la subpoblació a la qual es refereix. Així,
- Un 73,78% dels participants amb trastorns mentals se senten pitjors al juliol de 2020 (Fig. 77)
- Un 5,12% dels participants van desenvolupar nous problemes mentals durant el primer confinament de la COVID-19.
- Un 68,86% amb depressió empitjoren al juliol
- El 72,38% dels participants que patien ansietat se senten pitjors.
- A més, a partir del recompte de nova variable creada, es pot aprendre que el 17% dels participants pateixen un diagnòstic en salut mental, un 5% té almenys dos trastorns mentals simultàniament (com l'estrès posttraumàtic i la depressió) i un 1% registra fins i tot tres patologies mentals diferents.

Figura 77: Resultats de salut mental.

- Quan la nova variable U30 s'utilitza per filtrar les dades originals i l'anàlisi es repeteix per als joves (per sota de 30) s'observen diferents resultats. En la figura 78, es mostren els resultats i els resultats principals, obtinguts a partir de les dades originals i les variables preprocessades.
- El 17% dels joves participants declarats diagnòstics sobre salut mental,
- El 55% d'ells estan rebent medicació (en contrast amb el 96,94% trobat per a la població global)
- El 36% declara depressió
- El 51% declara ansietat
- 12% de trastorn de salut mental greu
- 9% Limita el trastorn de personalitat
- Un 6% més trastorns mentals.
- El 38% dels joves necessitaven suport emocional.

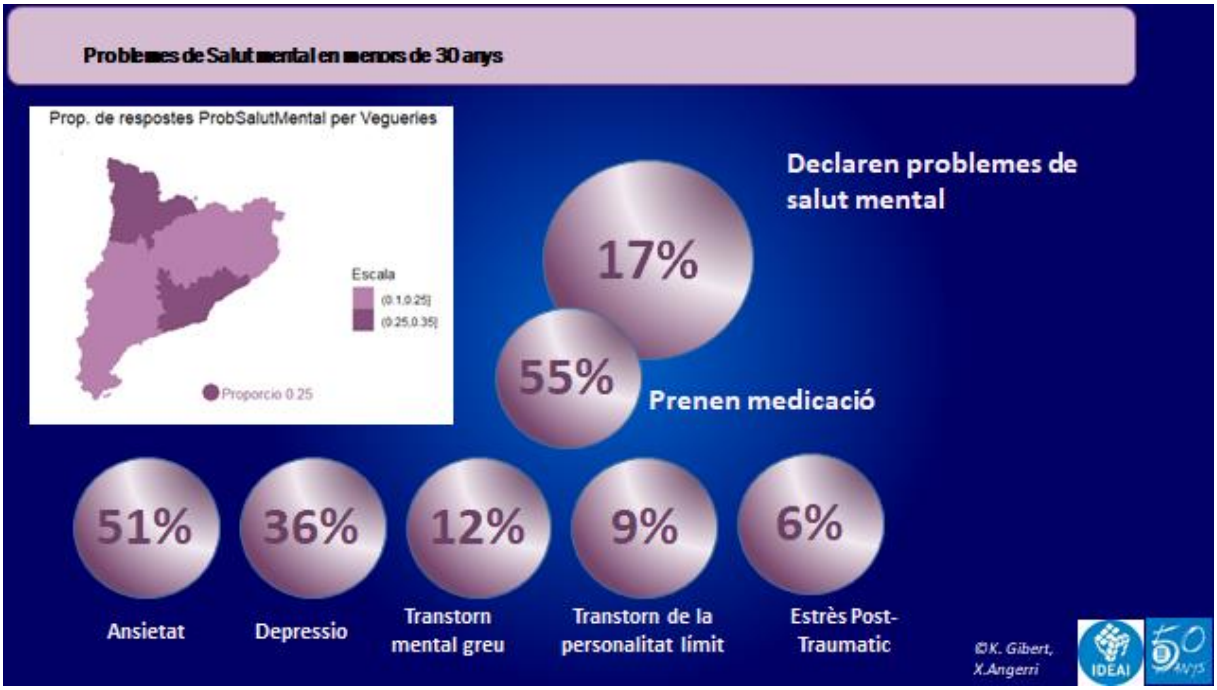

Figura 78: Problemes de salut mental en menors de 30 anys

7.2.9. Fase III Variables de segona generació basats en dades

A l'apartat 2.6.2 es presenta una nova metodologia per crear indicadors de segona generació basats en dades que resumeix la idea principal de crear noves variables.

Ara es mostrarà una demostració de com es podria construir la variable de segona generació utilitzant la nova metodologia presentada.

1. Seleccioneu les variables del component a sintetitzar:

Al qüestionari hi ha diverses preguntes que proporcionen informació sobre la situació de la persona al gener de 2020, juliol de 2020 i gener de 2021, de manera que es poden analitzar junts.

En aquest cas, s'ha preguntat als participants sobre la seva unitat convivencial. Cada participant va respondre a la pregunta: “amb qui vius”. Les opcions possibles són: «1. Sol: visc sol.;2. MM-MP: família monoparental o monoparental.;3. Nucleus: Família de pare i mare i fills propis (si hi ha fills).;4. Reagrupada: Famílies reagrupades (fills de diverses parelles).;5: Extensió: Familiar ampliada (pares, mares, fills, avis, ties, etc.).;6: No-família: visc amb persones que no són família.”; 7.NoAnswer

S'ha respost 3 vegades, creant 3 variables diferents. Per analitzar-la, aquestes variables seran els components de la nova variable.

2. Clustering de les variables de component seleccionades

Utilitzant el Mètode de Ward amb Mètriques Mixtes Gibert, les variables seleccionades es clusteriten i es produeixen noves variables de classe. La figura 79 mostra el dendrograma resultant.

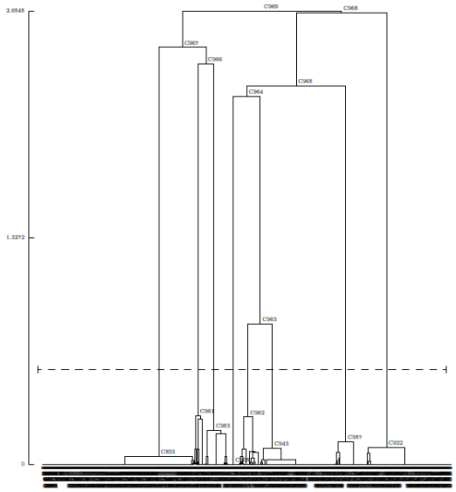

Figura 79: Dendrograma de la unitat de convivencial

Utilitzant els criteris de [Calinski & Harabasz, 1974], la base de dades es divideix en 8 clústers.

3. Creació del CPG

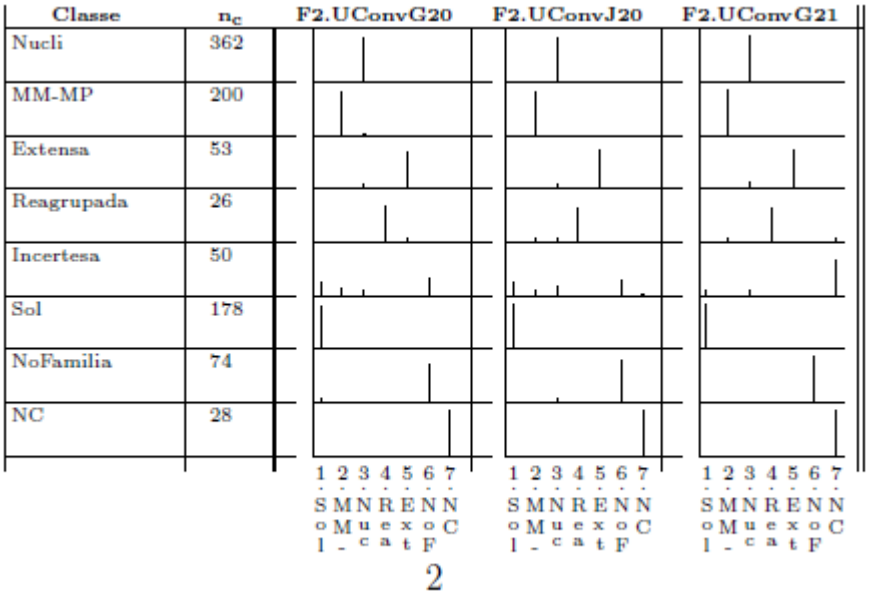

Figura 80 CPG de la unitat convivencial

4. CREACIÓ DEL TERMÒMETRE

Seguint la metodologia presentada a la secció 5.14.1. S'ha creat un termòmetre (vegeu la figura 81) juntament amb l'expert de camp, s'ha assignat un color per a cada modalitat.

Figura 81:Termòmetre de la unitat convivencial

El verd s'assigna a «2.MM-MP» i «3.Nucleus" modalitats perquè l'expert considera que viure amb la teva família del nucli i la gent que t'estima és positiu. Teniu gent que us pot ajudar en cas de necessitat i les dificultats són fàcils de superar.

El groc s'assigna a «4. Reagrupat» i «5.Diferents" modalitats perquè l'expert considera que viure amb persones que pertanyin al vostre grup familiar és positiu. D'altra banda, aquest tipus de famílies solen estar compostes per molts membres i és possible l'aparició de conflictes entre ells. Per tant, com que és negatiu, està marcat de groc .

El vermell s'assigna a «1.Sol" i "6.NoFamily". Viure amb persones que no són de la teva família i, per tant, no hi ha cap inclinació emocional és com viure sol, perquè tot s'ha de fer per tu mateix. És més difícil trobar ajuda. A més, si vius amb persones que no són la teva família, és més fàcil que apareguin conflictes.

5. Creació de TLP basat en el termòmetre:

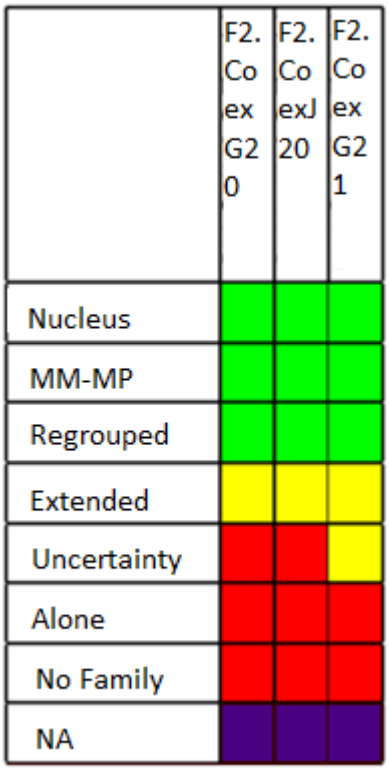

Figura 82: TLP de la unitat convivencial

- Nucleus: Les persones d'aquest grup van respondre durant els 3 preíodes de temps “3.Nueclus”
- MM-MP: Les persones d'aquest grup van respondre durant els 3 preíodes de temps «2.MM-MP»
- Reagrupats: les persones d'aquest grup van respondre durant els 3 preíodes de temps “4.Reagrupat”
- Estesa: Les persones d'aquest grup van respondre durant els 3 preíodes de temps “5.Estesa”
- Incertesa: La gent no sap què fer al gener de 2021. (el qüestionari es va respondre al 2020, durant la primera onada de COVID-19). Durant el 2020 tenen situacions diferents.
- Sols: les persones d'aquest grup van respondre durant 3 períodes de temps “1.Sol”
- Sense família: les persones d'aquest grup van respondre durant 3 períodes de temps “6.No-família”
- NA: Les persones d'aquest grup van respondre durant 3 períodes de temps “7.Sense resposta”

En resum, hi ha 8 classes. 7 són per a persones sense canvis. Persones que van respondre a la mateixa opció el gener de 2020, juliol de 2020 i gener de 2021. En l'últim grup, la gent va canviar d'opció.

6. Creació de l'indicador $\mathcal{P}$ interpretat i el nom $\mathcal{P}$:

Com que aquesta nova variable ve de la variable F2.CoexMYY, on M correspon al mes i YY a l'any, l'etiqueta és F2.CoexLab, on el laboratori prové de l'etiqueta i Uconv de la unitat convivencial.

7. Afegir $\mathcal{P}$ a la base de dades general:

Un cop fets tots els passos P s'afegeix a la base de dades general.

Durant el procés aquest procediment s'ha repetit diverses vegades, s'han creat diversos DD2gI. També s'han afegit a la base de dades indicadors nous de tercera generació presentats a [Angerri & Gibert, 2023]. Les xifres 8, 9 i 10 representen totes les variables i indicadors de la base de dades.

7.2.10. Fase III Variables de tercera generacio basades en dades

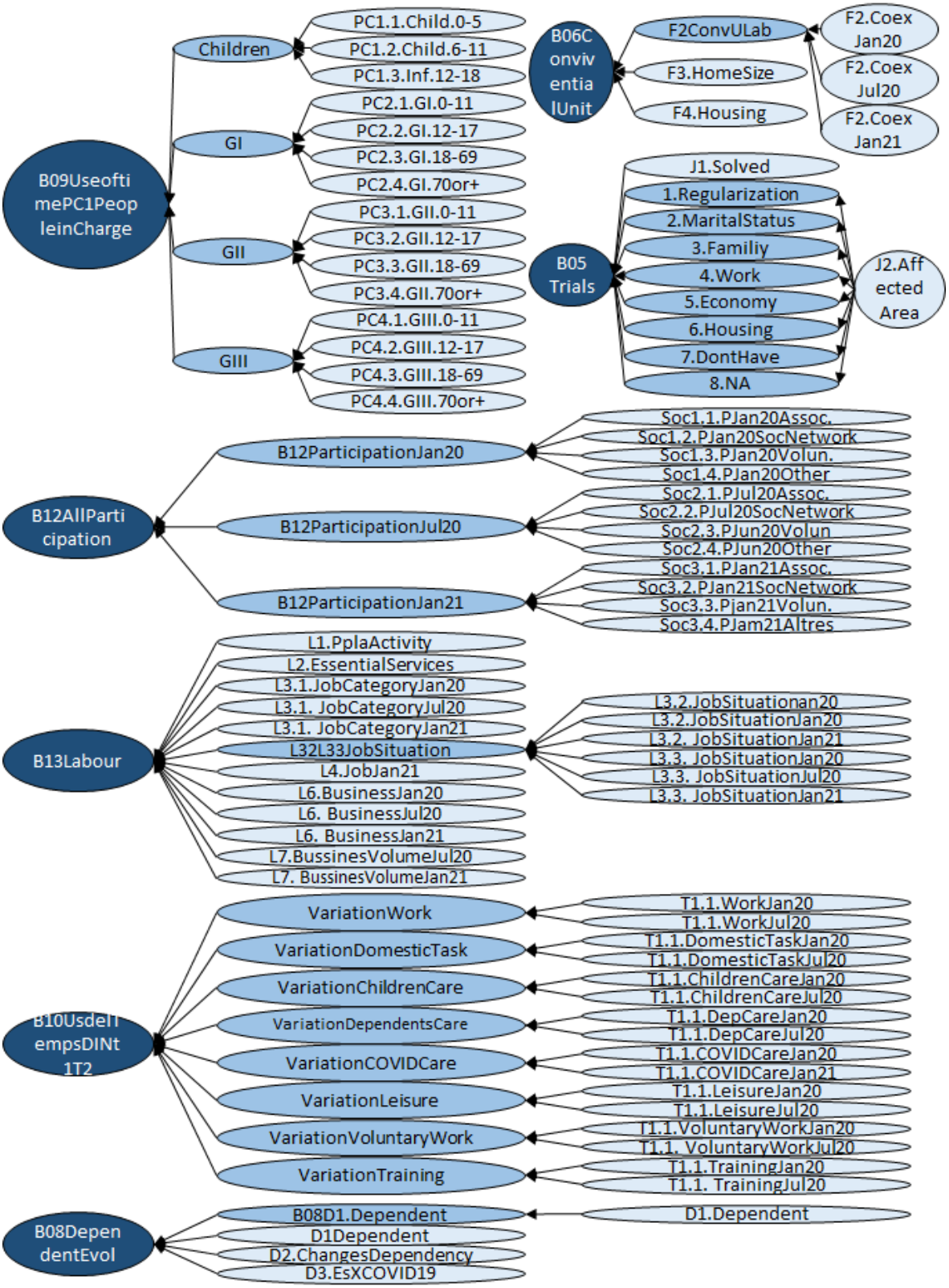

Figura 83: Components de les variables de tercera generació

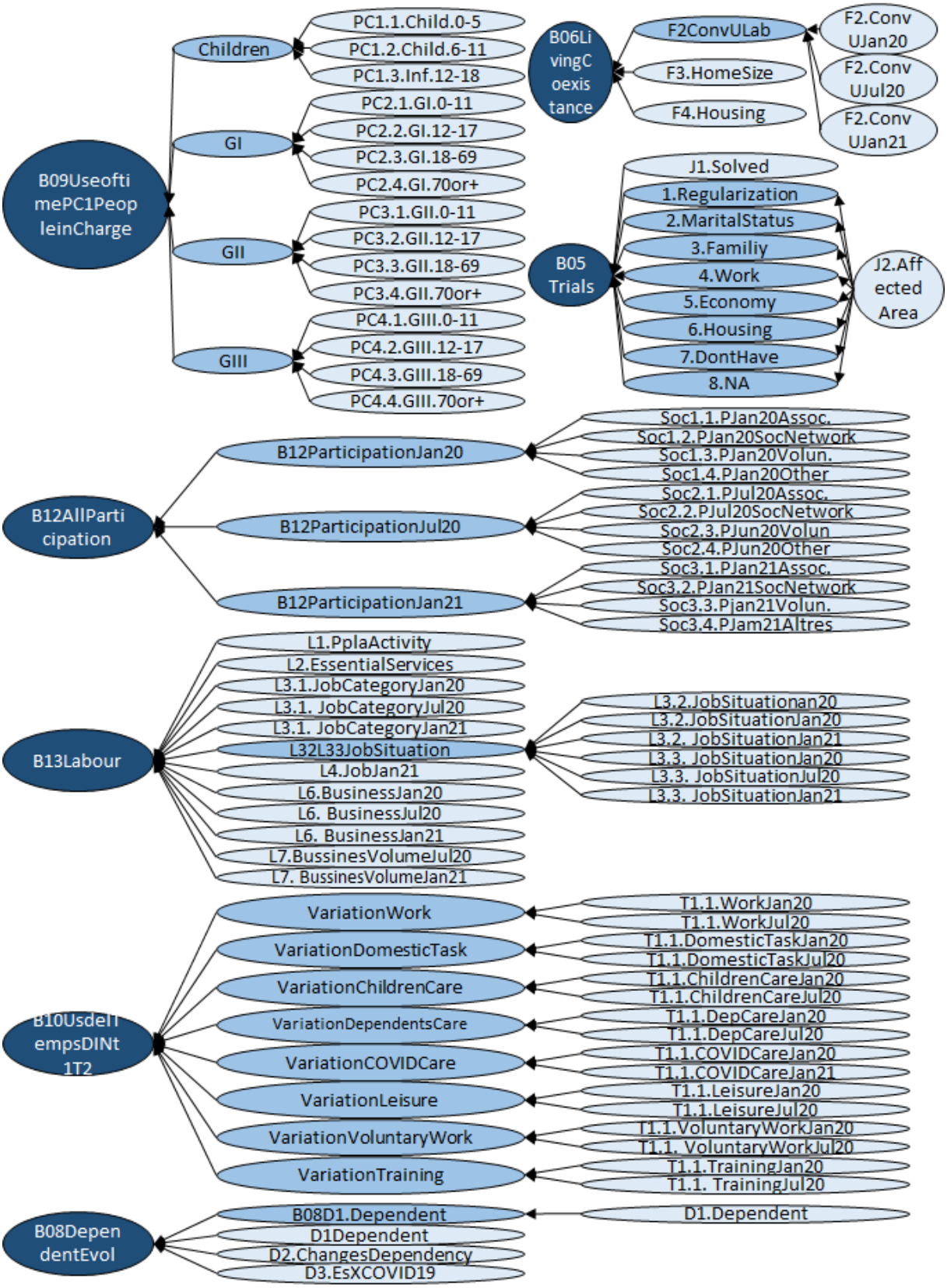

Figura 84: Components de les variables de tercera generació

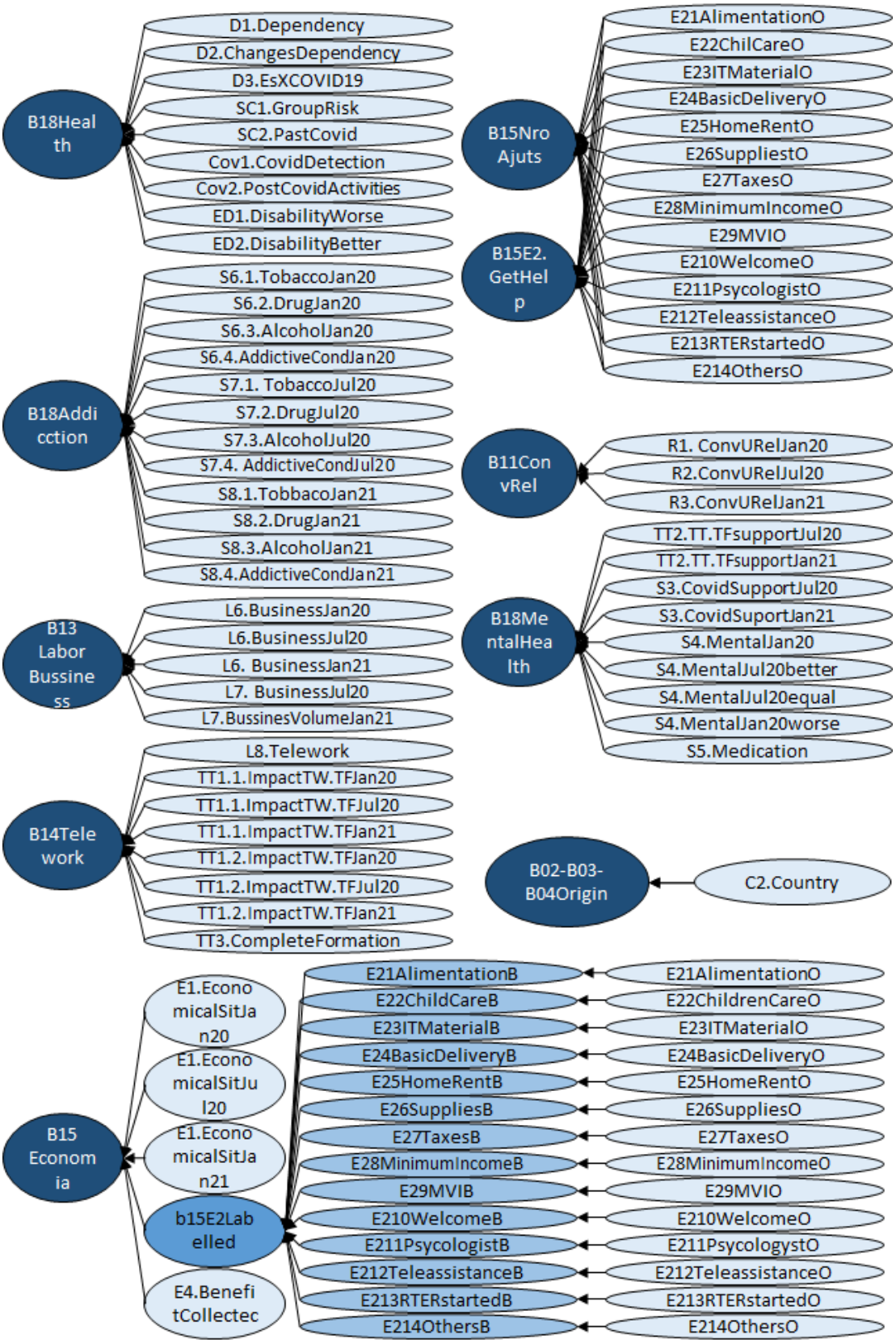

Figura 85: Components de les variables de 3a generació

Per conèixer la situació global de la salut utilitzant tota la informació de les variables de salut, s'han creat 3 nous indicadors basats en dades utilitzant la nova metodologia explicada a la secció 5.13.3:

- Salut
- Salut mental
- Abús de la substància

Hi ha diversos blocs que contenen qüestions de salut. Amb els experts, les variables de salut es van distribuir en tres punts de vista, de manera que un procés de clustering per visió proporcionarà l'indicador corresponent. Els components dels tres nous indicadors es mostren a la taula 44:

Indicador Data-driven	**Modalitats**	**Components (etiquetes en relació amb la taula)**
Salut	Variables relacionades amb la malaltia de la COVID, grau de dependència i discapacitat de la persona	D1, D2, D3, SC1, SC2, Cov1, Cov2, S9, ED1, ED2
Salut Mental	Aquelles variables relacionades amb el support emociona i els transtorns de salut mental	TT2.TT.TF, S3, S4, S5
Abús de substàncies	Variables relacionades amb l'abús de substàncies	S6, S7, S8

Taula 44: Components dels indicadors

En el qüestionari, és possible veure diverses variables que podrien estar relacionades amb la salut i la salut mental, però no s'hi inclouen perquè els experts van considerar que no proporcionen molta informació rellevant. Es realitza un procés de clustering i les classes obtingudes són interpretats utilitzant termòmetres i TLPs, juntament amb el CPG i una descripció de la situació general és disponible.

A continuació, es mostra el termòmetre que servirà per construir posteriorment el aTLP:

Figura 86: Termòmetre per a l'indicadot de salut mental

Els resultats es mostren en un aTLP (vegeu Fig.4), que ajuda a veure com es comporta la salut mental en les dades de mostra. El mètode TLP es descriu en 3.3.2

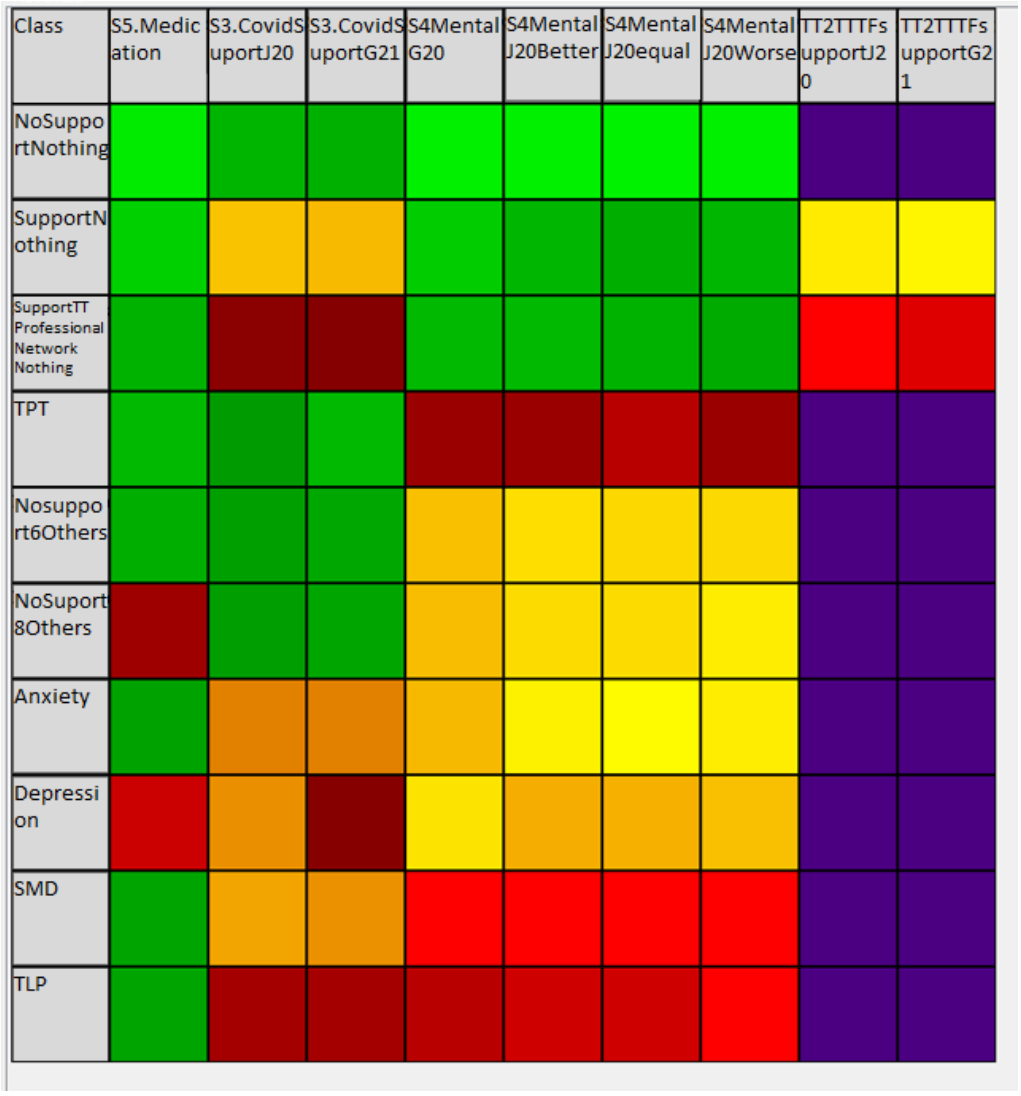

Figura 87 TLP de salut mental utilitzant només variables originals (graf: KLASS).

Segons la inspecció de dendrograma, apareixen 10 classes quan s'agrupen totes les variables. Utilitzant CPG i TLP es poden descriure de la manera següent:

- NoSupportNothing: No van teletreballar i no van tenir cap problema mental. No requerien suport emocional i no reben medicació.
- SupportNothing: van fer teletreball i van necessitar suport emocional de la seva xarxa personal. No tenien cap problema mental. No requerien suport emocional professional i no reben medicació.
- SupportTTProfessionalNetworkNothing: Van fer teletreball i van requerir el suport emocional d'un professional. No tenien problemes mentals. No requerien suport emocional i no estan rebent medicació.
- TPT: Tenen un trastorn d'estrès posttraumàtic diagnosticat. No van teletreballar. La majoria no va necessitar suport emocional al juliol de 2020 amb la mateixa expectativa per al gener de 2021. La majoria no reben medicaments.
- NoSupport6Altres: No van fer teletreball i tenen altres problemes mentals que no s'especifiquen. No requerien suport emocional i no estan rebent medicació.
- NoSupport8Altres: No van fer teletreball i tenen altres problemes mentals que no s'especifiquen. No requerien suport emocional i estan rebent medicació.
- Anxietat: tenen ansietat diagnosticada. No van fer teletreball. Alguns d'ells van requerir el suport de la xarxa personal i altres professionals fins al juliol de 2020. La majoria no reben medicaments.
- Depressió: Té diagnosticada Depressió. No van fer teletreball. Alguns d'ells van requerir el suport de la xarxa personal i altres professionals fins al juliol de 2020. Al gener de 2021, la majoria d'aquest grup espera necessitar suport professional. La majoria reben medicaments.
- SMD: Tenen un trastorn mental greu diagnosticat. No van fer teletreball. Alguns d'ells van requerir el suport de la xarxa personal i altres professionals fins a juliol de 2020 amb la mateixa expectativa per al gener de 2021. La majoria no reben medicaments.
- TLP: Tenen un trastorn límit diagnosticat. No van fer teletreball. Van requerir el suport emocional d'un professional el juliol de 2020 amb la mateixa expectativa per al gener de 2021. La majoria no reben medicaments.

Creació d'indicadors basats en dades de tercera generació, incloses les variables basades en el coneixement de la segona generació

Una vegada que s'han creat les variables basades en el coneixement, es comporten com a variables ordinàries des del punt de vista tècnic. Per tant, no hi ha limitacions per utilitzar aquestes variables per ser incloses en cap anàlisi addicional, en particular la creació de

variables de tercera generació basades en dades, com s'esmenta. 3.2.3. Per descomptat, cal anar amb compte d'utilitzar aquestes variables d'una manera correcta d'acord amb els requisits tècnics del model de dades que s'utilitzarà (per exemple, en els models de regressió multivariant clàssica, es requereix la independència dels regressors i això augmentarà la incompatibilitat entre les variables originals i derivades per coexistir en un model de regressió; cada model té els seus propis requisits).

Per tant, les noves variables de salut mental derivades de les originals es van incloure en un nou procés de clustering on es mesclen amb les variables originals. El procés es repeteix i es crea un nou TLP generat a partir de les variables de 2a que ha usat el termòmetre de la figura

Figura 88: Termòmetre de salut mental incloent les variables de 2a generacio

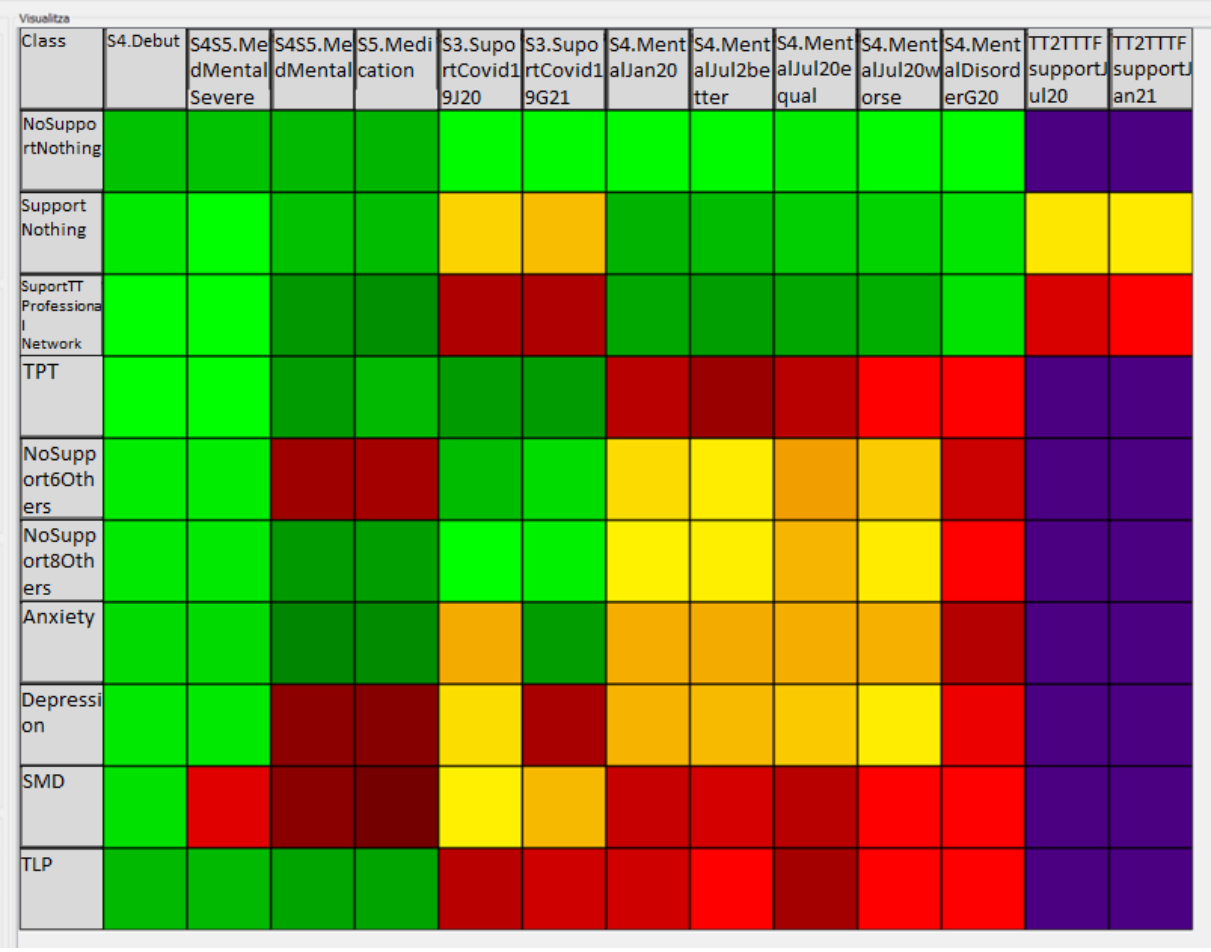

Figura 89. Salut mental amb variables derivades basades en el coneixement TLP (graf: KLASS).

El valor afegit de les variables derivades basades en el coneixement en crear indicadors basats en dades

Gràcies a l'ús de variables de segona generació en el procés de clustering, estem obtenint molta informació addicional.

D'una banda, els colors de les variables originals són més brillants, el que significa que els cúmuls tenen graus més ABSS d'heterogeneïtat.

A més, les conclusions que emergeixen de les dades on s'inclouen les variables derivades basades en el coneixement en l'anàlisi són:

La nova variable S4.MentalDisorderG20 informa sobre tenir algun trastorn mental abans de la pandèmia. Amb la seva introducció en l'anàlisi, el TLP (Fig 5) mostra clarament que 3 dels

cúmuls descoberts estan compostos principalment per persones sense problemes de salut mental i aquesta informació no era tan evident en el TLP amb variables originals (Fig 4).

La construcció de variables que indiquen que mentre que la persona amb un perfil determinat de trastorns mentals està prenent medicació o no, permet visualitzar en el TLP que els grups sota medicació, de fet, estan prenent medicació pel seu problema mental.

Els experts estaven interessats a centrar-se en 3 trastorns mentals específics (SMD, PLS, PSTD). El grup de persones principalment afectades per un trastorn mental greu estan rebent medicaments.

Els grups amb major proporció de persones que debuten en trastorns de salut mental durant la primera onada es concentren en dos perfils principals: El de les persones que pateixen ansietat, el de les persones que pateixen altres malalties no especificades i que no requereixen suport emocional addicional. Hi ha un tercer perfil amb un petit impacte en els debuts en salut mental, el que fa referència a les persones que estaven teletreballant durant la primera onada i aquells que requerien suport emocional de la seva xarxa personal o professionals. Això significa que la gent està debutant bàsicament en trastorns mentals no greus en la primera etapa.

El nou indicador de dades per a la salut mental es reinterpreta en la següent secció incloent tota la informació nova proporcionada per les variables derivades basades en el coneixement.

- NoSupportNothing: No van teletreballar i no van tenir cap problema mental. No requerien suport emocional i no reben medicació.
- Suport Res: Van fer teletreball i van requerir suport emocional de la seva xarxa personal. No tenien cap problema mental. No requerien suport emocional professional i no reben medicació.
- SupportTTProfessionalNetworkNothing: Van fer teletreball i van requerir el suport emocional d'un professional. No tenien problemes mentals. No requerien suport emocional i no estan rebent medicació.
- TPT: Tenen un trastorn d'estrès posttraumàtic diagnosticat. No van teletreballar. La majoria no va necessitar suport emocional al juliol de 2020 amb la mateixa expectativa per al gener de 2021. La majoria no reben medicaments.
- NoSupport6Altres: No feien teletreball i tenien altres problemes mentals que no s'especificaven. No requerien suport emocional i no estan rebent medicació.
- NoSupport8Altres: No feien teletreball i tenien altres problemes mentals que no s'especificaven. No requerien suport emocional i estan rebent medicació.
- Anxiety: Tenen ansietat diagnosticada. No van fer teletreball. Alguns d'ells van requerir el suport de la xarxa personal i altres professionals fins al juliol de 2020. La majoria no reben medicaments.

- Depressió: Té diagnosticada Depressió. No van fer teletreball. Alguns d'ells van requerir el suport de la xarxa personal i altres professionals fins al juliol de 2020. Al gener de 2021, la majoria d'aquest grup espera necessitar suport professional. La majoria reben medicaments.
- SMD: Tenen un trastorn mental greu diagnosticat. No van fer teletreball. Alguns d'ells van requerir el suport de la xarxa personal i altres professionals fins a juliol de 2020 amb la mateixa expectativa per al gener de 2021. La majoria no reben medicaments.
- TLP: Tenen un trastorn límit diagnosticat. No van fer teletreball. Van requerir el suport emocional d'un professional el juliol de 2020 amb la mateixa expectativa per al gener de 2021. La majoria no reben medicaments.

Indicadors basats en dades en salut

Figura 90: Termòmetre per a les variables de salut

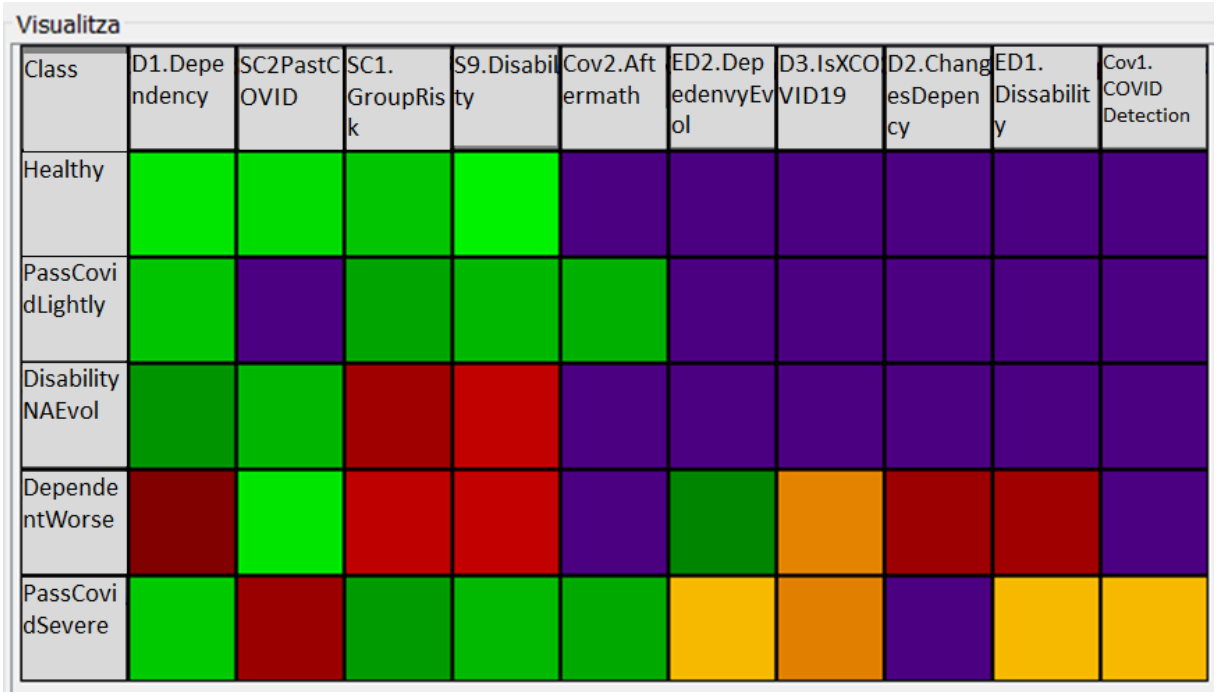

Figura 91 TLP de salut (graf: KLASS)

Algunes de les variables es van utilitzar per generar un nou indicador de salut (vegeu Fig. 6) i van identificar els escenaris següents

- Salut: Les persones d'aquest grup no tenen problemes de salut. No van patir COVID-19, no pertanyen a un grup de risc COVID i no són persones amb discapacitat.
- PassMildCovid: No són persones amb discapacitat ni persones amb discapacitat. Va patir COVID sense dificultats i va poder fer vida normal després de COVID. No pertanyen a un grup de risc COVID.
- DiscapacitatNevol: Són persones amb discapacitat, cosa que significa que pertanyen a un grup de risc de COVID. No obstant això, no va patir de COVID.
- DependentWorse: Són persones amb discapacitat i amb discapacitat, cosa que significa que pertanyen a un grup de risc de COVID. No obstant això, no va patir de COVID. La seva dependència està empitjorant a causa de la COVID.
- PassCovidSevere: No són persones discapacitades. Va patir COVID greu i va poder fer vida normal després de COVID. No pertanyen a un grup de risc COVID.

Indicador de consum de substàncies

Per a aquestes variables que componen l'indicador d'abús de substàncies, després del procés d'agrupació (vegeu Fig. 93) el nou indicador tindrà 10 modalitats corresponents als següents significats:

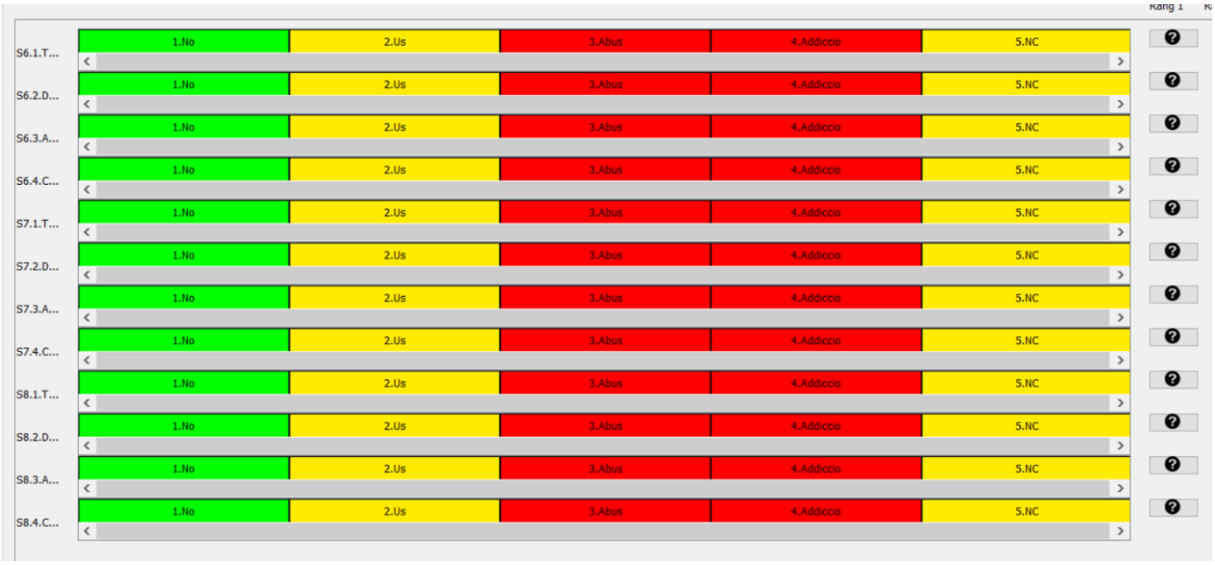

Figura 92: Termòmetre de salut mental

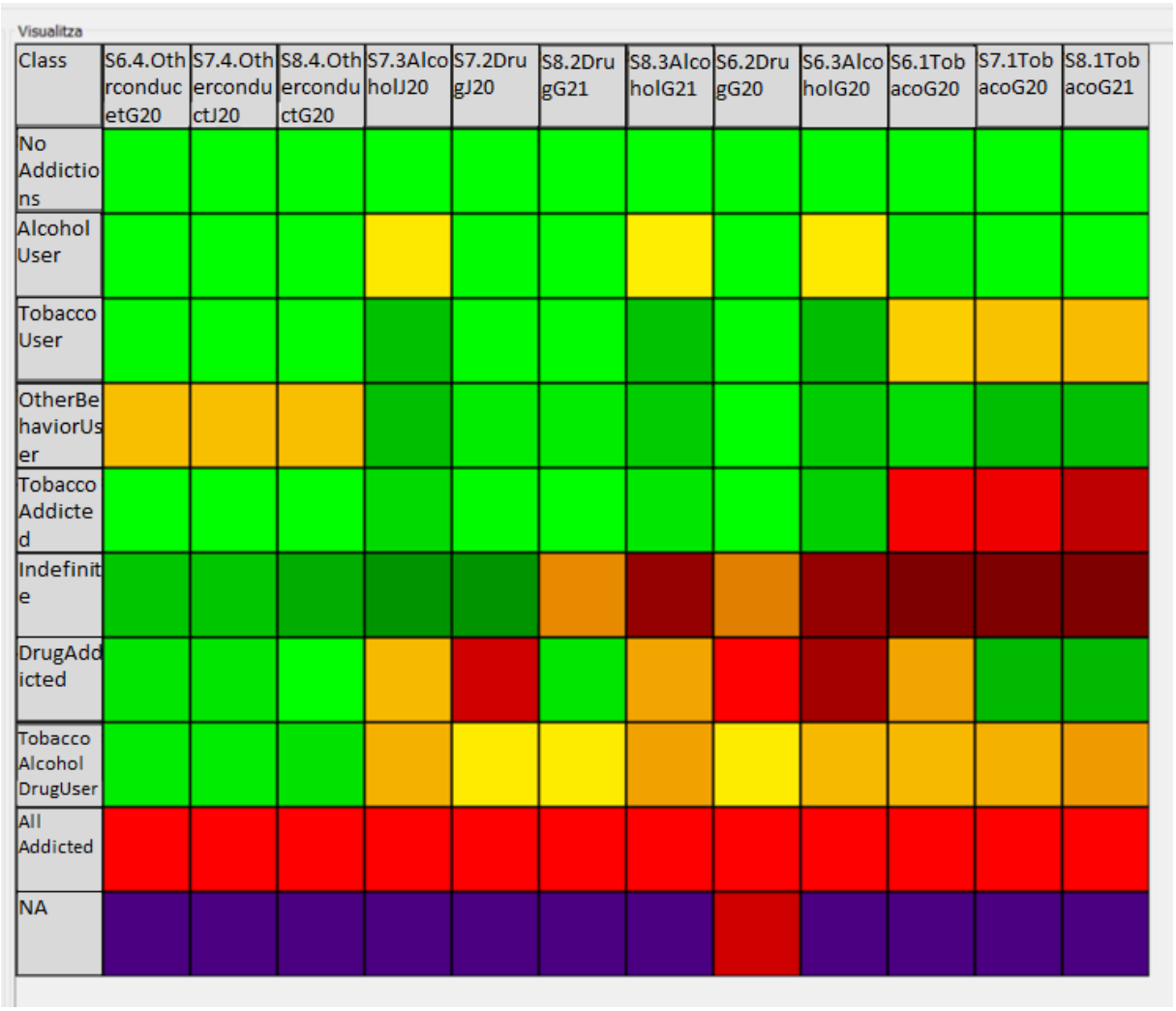

Figura 93:. TLPs d'abus de substàncies.

- NoAddiccions: No tenien cap afegit.
- AlcoholUser: consumeixen alcohol però no altres substàncies.
- El tabac: consumeixen tabac però no altres substàncies.
- Altres usuaris: Segueixen altres conductes com el joc, però no tenen altres addiccions.
- TobaccoAddicted: Són addictes al tabac però no tenen altres addiccions.
- Indefinit: Aquest és un grup mixt sense una descripció clara. No obstant això, podríem dir que tenen problemes amb el tabac i que no tenen problemes amb altres addiccions. Al juliol de 2020 no tenen problemes amb les drogues i l'alcohol.
- Drogoaddictes: Són addictes a les drogues fins al juliol de 2020. S'espera que no siguin addictes al gener de 2021. També tenen problemes amb el tabac i l'alcohol
- TobaccoAlcoholConsumer: Són consumidors de drogues, tabac i alcohol.
- AllAdictes: Són addictes a totes les substàncies al llarg de tot el temps.
- NA: No van respondre aquestes preguntes

7.2.11. Fase III Anàlisi Multivariant.

En les files següents, mostrarem com se seleccionen les variables que s'escolliran per al procés de clustering.Com a exemple, utilitzarem el Bloc XI: Us de l'indicador de temps (la gent que s'ocupa de) . Aquests indicadors estan compostos per 4 variables. Aquestes variables indiquen si el participant té fills a càrrec o persones de dependència de grau I, II o III a càrrec. Aquestes variables són les variables de segona generació construïdes utilitzant la metodologia proposada a [Angerri & Gibert, 2023]

1. Seleccioneu la variable d'ubicació:

La variable territorial serà l'ABSS. Aquesta variable s'ha creat a la secció 7.2.8. Aquesta és una variable que prové de la variable original Municipalitat de residència.

2. Seleccioneu el bloc

En aquest document es mostrarà el cas del Bloc XI: Us of time (caretakers), que s'ha construït més amunt.

3. Seleccioneu les variables candidates

Com es pot veure, 4 variables de segona generació són el component de l'indicador. Per tant, Fills, GI, GII i GIII són les variables candidates que s'han de seleccionar per estar en el grup final.

4. Calcular una classificació de variables segons la seva capacitat per explicar la distribució territorial:

Utilitzant la metodologia explicada en 5.16.1 s'estan calculant tots els resultats.

Indicador/ Component	Bloc XI: Us of time (Persones a càrrec)	Infants	GI	GII	GIII
R_k	0,7	0,5	0,5	0,43	0,45
$\widetilde{\Pi}_{Lock}$	0,47	0,56	0,58	0,54	0,57
E_k	0,329	0,28	0,29	0,23	0,26

Taula 45: $R_k, \widetilde{\Pi}_{Lock}, E_k$ valors per a cada component.

5. Selecció de variables:

En aquest cas, la variable que representa el bloc és la variable seleccionada serà el bloc XI: Ús del temps (Caretakers).

Un cop seleccionades totes les variables del model i s'ha fet la selecció de característiques, les variables seleccionades finals són R3 RelConvG21, B06UConvivencial, S3SuportCovid19J20, B11VioG21, B02-B03-B04Origen, B09UsTempsPC1FentCarrec, E210AcollidaO, 5Economia, B15E2Labelled, L32L33SituacioLab, B12ParticipacioTot, VariacioCuraCOVID, D3EsXCOVID19,L6NegociG20

7.2.12. Fase IV Perfilat intel·ligent de classes

6.Clustering de les variables seleccionades:

Una vegada seleccionades totes les variables del model i s'ha realitzat la selecció de característiques, les variables seleccionades finals són les següents:

Totes aquestes variables i el ABSS estan agrupats. Per fer el clustering s'ha utilitzat un clustering condicional per ABSS mitjançant els criteri de Ward i la distància de Gibert mixta. A la figura 94 es mostra el dendrograma resultant.

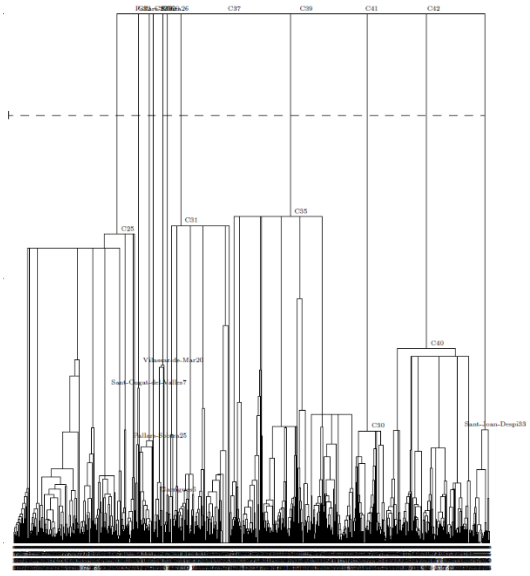

Figura 94: Dendrograma final

Utilitzant criteris de Zelinski-Harabasz [Calinski & Harabasz, 1974], la base de dades es divideix en 11 grups.

7. Creació del CPG

Després del dendrograma, es crea un CPG. El seu aspecte és com les xifres 5 i 10 mostren l'aspecte que té.

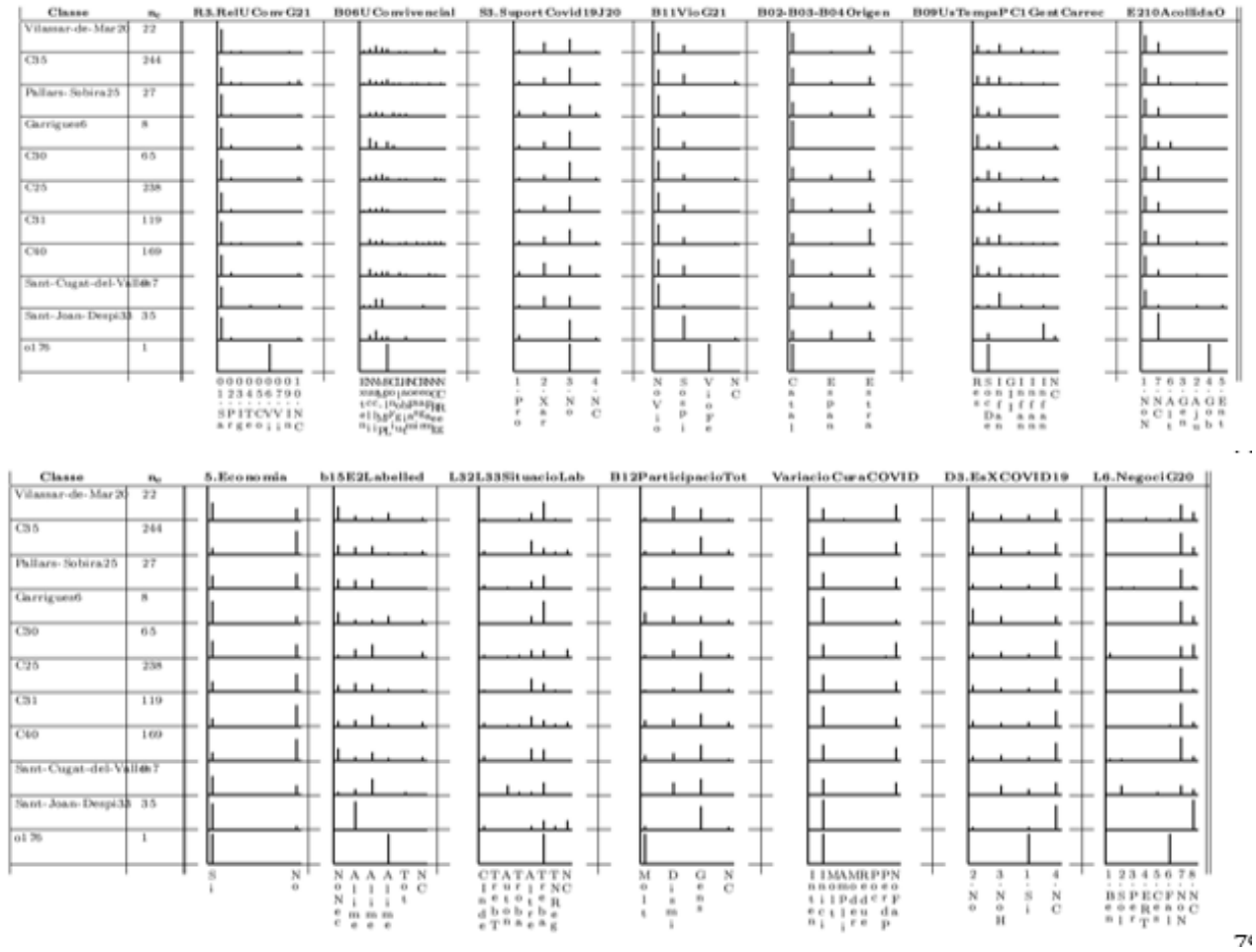

Figura 95 CPG Final

9. Creació del termòmetre

El termòmetre final està construït. Les variables que s'han seleccionat podrien ser indicadors o components d'indicadors. En cas que la variable seleccionada sigui un component d'indicadors, el termòmetre construït per aquest moment es reutilitza en aquest pas. En el cas que sigui un indicador, el termòmetre es construeix utilitzant el coneixement expert i el TLP resultant per construir l'indicador de la tercera generació de dades.

Figura 96: Termòmetre final

Cada variable té el seu propi termòmetre. A la Figura 96 es mostra el termòmetre per a totes les variables.

Crear TLP basat en el termòmetre

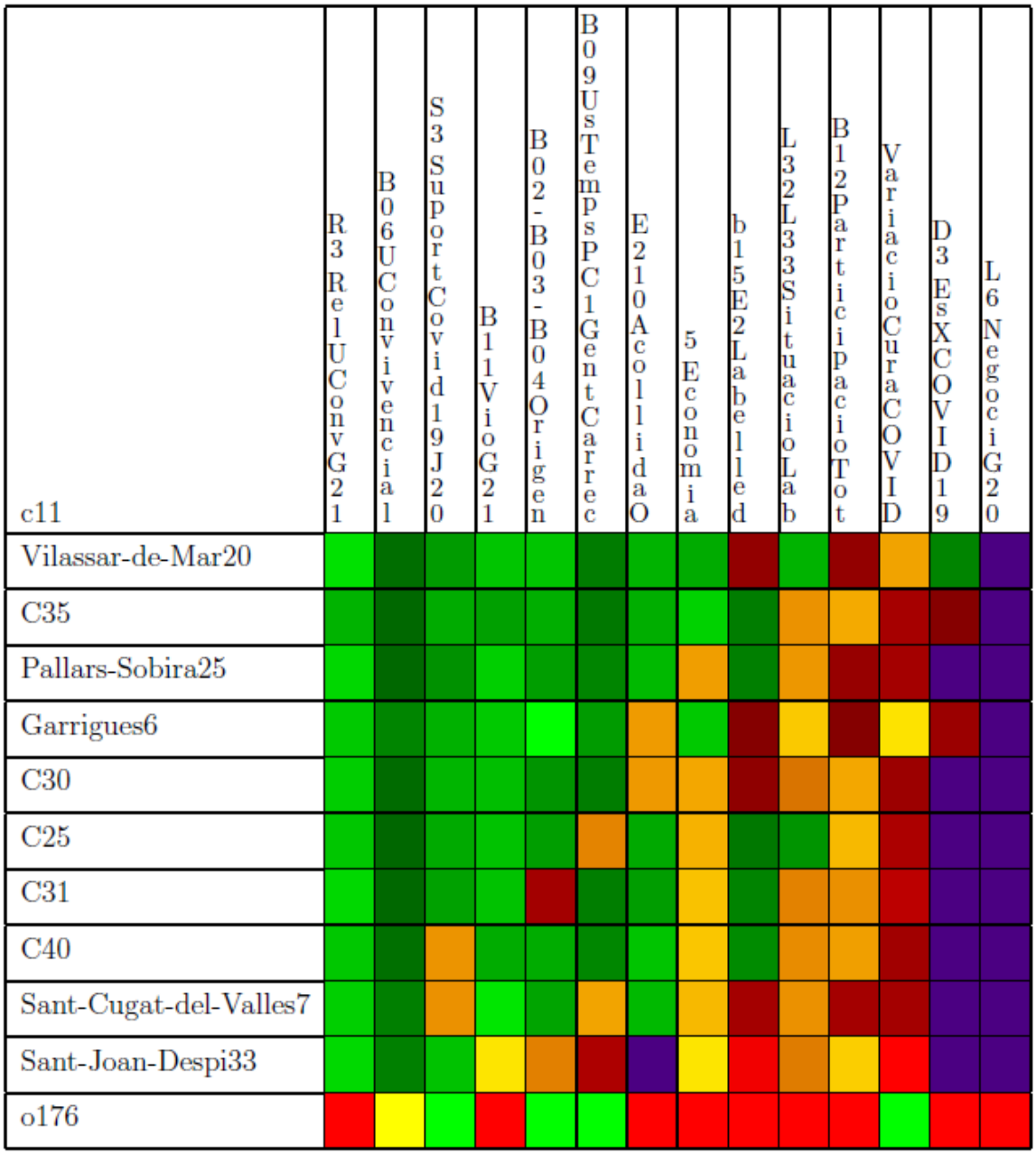

Figura 97 T-aTLP final

<u>Validació estructural de la classificació:</u>

Validació del pas 4:

Com es va dir a la secció 6.5 es va aplicar la prova per a tots els components i indicador contra l'ABSS. Els valors p singificatius (<0,05) indiquen que una variable discrimina el BASS. La idea de comprovar quines variables i l'ABSS depenen o no.

Per exemple, mostrarem els valors obtinguts a partir de 2 indicadors. El primer és el que s'ha seguit en aquest document, el BLOC IX. Per complementar-lo, també es demostra el bloc XI.

A les taules 46 i 47 es veuen els valors .2 per cada component variable.

Indicador/ Component	Block XI: Us of time (People taking care from)	Children	GI	GII	GIII
χ^2p-value	1,37E-11	4,27E-26	2,88E-26	1,83E-22	1,57E-22

Taula 46 p vaalors de χ^2 per ús dels components del block IX ús del temps

Indicador/ Component	B11Rel Conv	R1.RelU ConvG20	R2.RelU ConvJ20	R3.RelU ConvG21
χ^2p-value	1,37E-11	2,95E-06	1,58E-17	1,88E-10

Taula 47: p vaalors de χ^2 per ús dels components p vaalors de χ^2 per ús dels components de temps

Totes les variables són significatives, cosa que significa que no podem descartar cap variable. Seguint aquest criteri, totes les variables han de ser preses en precisió i no hem estat capaços de definir un criteri.

Després s'ha utilitzat la metodologia proposada en aquest document a la secció 6.5.

Com es veu a la Taula 47, B09Use of time (Gent a càrrec) l'indicador és la variable escollida perquè té la relació de contribució més alta (0,3290).

Indicator/ Component	B11Rel Conv	R1.RelU ConvG20	R2.RelU ConvJ20	R3.RelU ConvG21
R_k	0.61	0.66	0.68	0.66
$\widetilde{\Pi}_{Lock}$	0.38	0.33	0.34	0.39
E_k	0.23	0.22	0.23	0.26

Taula 48: . $R_k, \widetilde{\Pi}_{Lock}, E_k$per a cada component B11 RelConv

No obstant això, l'indicador B11RelConv és la variable seleccionada en aquest cas, R3.RelUconvG21 és la variable seleccionada, amb el percentatge de contribució més alt.

Aquests són 2 exemples del procés de selecció de característiques. Això es repeteix en tots els blocs i aquest fenomen s'està repetint. Gairebé tots els valors p obtinguts amb valors p són millors.

Després de fer clustering i aplicar el termòmetre a la TLP (vegeu la figura 18), el resultat són 11 grups, descrits en la secció següent:

- Vilassar de Mar: No presenten problemes de relació en la seva unitat de coexistència i aquesta és una família en cases de propietat o lloguer. No

necessitaven suport psicològic a causa de la COVID, la majoria són d'origen català. Les persones responsables no perden el temps cuidant de la unitat. No necessiten cap ajuda per COVID. Alguns tenen demandes pendents relacionades amb el camp econòmic. Alguns no necessitaven assistència social durant la COVID. Són treballadors i persones amb altres situacions laborals poc convencionals. La seva socialització amb el medi ambient ha disminuït o és inexistent.

- C35: No presenten problemes de relació amb la unitat de coexistència i tenen una unitat de convivència familiar (extensiva, tradicional o monoparental) en cases de propietat o lloguer. No necessitaven suport psicològic a causa de la COVID, la majoria són d'origen català. Les persones a càrrec no perden el temps cuidant de la unitat... No necessiten cap ajuda per COVID. No tenen decisions econòmiques pendents i algunes persones necessiten ajuda amb els aliments. Són persones amb altres situacions laborals poc convencionals i persones que no van treballar. No participen en el seu entorn.
- Pallars: No presenten problemes de relació amb la unitat de coexistència i tenen una unitat de coexistència familiar (extensiva, tradicional o monoparental) en cases de propietat o lloguer. No necessitaven suport psicològic a causa de la COVID, la majoria són d'origen català. Les persones a càrrec no perden el temps cuidant de la unitat... No necessiten cap ajuda per COVID. No tenen situacions econòmiques pendents i alguns d'ells necessiten ajuda amb aliments. Són persones que van treballar i amb altres situacions laborals poc convencionals. No participen en el seu entorn.
- Garrigues: No presenten problemes de relació amb la unitat de coexistència i tenen una unitat de coexistència familiar (extensiva, tradicional o monoparental) en cases de propietat o lloguer. No necessitaven suport psicològic a causa de la COVID, la majoria són d'origen català. Les persones responsables no perden el temps cuidant de la unitat. La seva situació de dependència no ha disminuït a causa de la COVID. Alguns d'ells necessiten ajuda dels serveis socials a causa de la COVID. Tenen pendents decisions econòmiques i algunes persones no han necessitat ajuda i altres han necessitat ajuda relacionada amb els aliments. Són persones que van treballar i amb altres situacions laborals poc convencionals. La participació en el seu entorn social és variada. Alguns participen molt, uns altres disminueixen i altres no van participar en absolut.
- C30: No presenten problemes de relació amb la unitat de coexistència i tenen una unitat de convivència familiar (ampliada, tradicional o monoparental) en cases de propietat o lloguer. No necessitaven suport psicològic a causa de la COVID, la majoria són d'origen català. No tenen cap persona al càrrec. - Alguns necessiten ajuda dels serveis socials a causa de la COVID. Tenen pendents decisions financeres i algunes persones necessiten assistència alimentària, així com altres ajudes

específiques que han sorgit durant la pandèmia. Són persones que van treballar i persones amb altres situacions laborals poc convencionals. No participen en el seu entorn ni s'han cuidat de persones amb COVID.

- C25: No presenten problemes de relació amb la unitat de coexistència i tenen una unitat de convivència familiar (ampliada, tradicional o monoparental) en cases de propietat o lloguer. No necessitaven suport psicològic a causa de la COVID, la majoria són d'origen català. La majoria d'ells tenen fills dependents que ocupen part del seu temps. No necessiten cap ajuda a causa de la COVID. No tenen decisions econòmiques pendents i algunes persones necessiten ajuda amb els aliments. Alguns són persones amb altres situacions laborals poc convencionals i altres són treballadors. No participen en el seu entorn ni s'han cuidat de persones amb COVID.
- C31: No presenten problemes de relació amb la unitat de coexistència i tenen una unitat de convivència familiar (extensiva, tradicional o monoparental) en cases de propietat o lloguer. No necessitaven suport psicològic a causa de la COVID, tenen un origen estranger. Les persones responsables no perden el temps cuidant de la unitat. No necessiten cap ajuda per COVID. . No tenen decisions econòmiques pendents i algunes persones necessiten ajuda amb els aliments. Alguns són persones amb altres situacions laborals poc convencionals i altres són treballadors. No participen en el seu entorn ni s'han cuidat de persones amb COVID.
- C40: No presenten problemes de relació amb la unitat de coexistència i tenen una unitat de convivència familiar (extensiva, tradicional o monoparental) en cases de propietat o lloguer. Algunes persones han necessitat suport psicològic a causa de la COVID des de la seva xarxa personal, majoritàriament d'origen català. Els responsables no perden el temps cuidant de la unitat que no tenen decisions econòmiques pendents i algunes persones necessiten ajuda amb els aliments. Alguns són persones amb altres situacions laborals poc convencionals i altres són treballadors. No participen en el seu entorn ni s'han cuidat de persones amb COVID.
- Sant Cugat del Vallès: No presenten problemes de relació amb la unitat de coexistència on tenen una bona unitat de coexistència, viuen principalment amb el seu nucli o són famílies monoparentals. Algunes persones han necessitat suport psicològic a causa de la COVID des de la seva xarxa personal, majoritàriament d'origen català. Cuiden de nens que consumeixen part del seu temps lluny d'ells. No necessiten anar als serveis de recepció durant la COVID. Tenen pendents decisions econòmiques i la majoria de la gent necessita ajuda amb els aliments. Algunes persones van treballar abans de la pandèmia i han perdut els seus llocs de treball i altres són autònoms. Algunes persones no participen en el seu entorn i en uns altres han disminuït.

- Sant Joan Despí: No presenten problemes de relació amb la unitat de coexistència i tenen una unitat de convivència familiar (extensiva, tradicional o monoparental) en cases de propietat o lloguer. No necessiten suport psicològic a causa de la COVID de la seva xarxa personal, són d'origen espanyol i se sospita que són víctimes de la violència. Tenen fills al càrrec, de manera que s'encarreguen de les persones dependents de grau III que es fan càrrec del seu temps. Tenen demandes pendents relacionades amb assumptes econòmics i totes les persones necessiten ajuda relacionada amb els aliments. Algunes persones van treballar abans de la pandèmia i van perdre els seus llocs de treball. No tothom està implicat en el seu entorn.
- o176: Una persona que patirà violència verbal per la seva unitat de cohabitació, però ara viu sola en un pis. No ha necessitat suport emocional i sofrirà violència física pel treball i la violència psicològica a la resta. És una persona dependent que viu a Catalunya i ha acceptat les mesures proposades pel govern espanyol per resoldre la COVID. Té una demanda pendent sobre qüestions econòmiques i ha rebut múltiples subvencions. Va treballar abans de la pandèmia. És una persona molt involucrada.

Creació de mapes per a visualitzar classes

En la figura 98 és fàcil veure com es distribueixen els grups. Aquestes són les distribucions:

- C25: Alt Empordà, Amposta, Baix Penedès, Calafell, Mollet del Vallès, Sant Vicenç dels Horts, Solsonès, Tarragona, Vilafranca del Penedès.
- C30: Lleida, Montcada i Reixac, Sant Andreu de la Barca, Sant Pere de Ribes
- C31: Barcelona, Manresa, Masnou, el, Rubí, Tarragonès
- C35: Alt Penedès, Bages, Baix Llobregat, CAS Garrotxa, Girona, Maresme, Osona, Pallars Jussà, Reus, Ribera d'Ebre, Vilanova i la Geltrú
- C40: Baix Empordà, Barberà del Vallès, Figueres, Gironès-Salt, Noguera, Pla de l'Estany, Sant Feliu de Guíxols, Selva, Vallès Oriental
- Garrigues: Garrigues
- Pallars-Sobira25: Pallars Sobirà
- Sant-Cugat-del-Valles7: Sant Cugat del Vallès
- Sant-Joan-Despi33: Sant Joan Despí
- Vilassar-de-Mar20: Vilassar de Mar

En aquest cas podem veure com s'agrupen les ABSS.

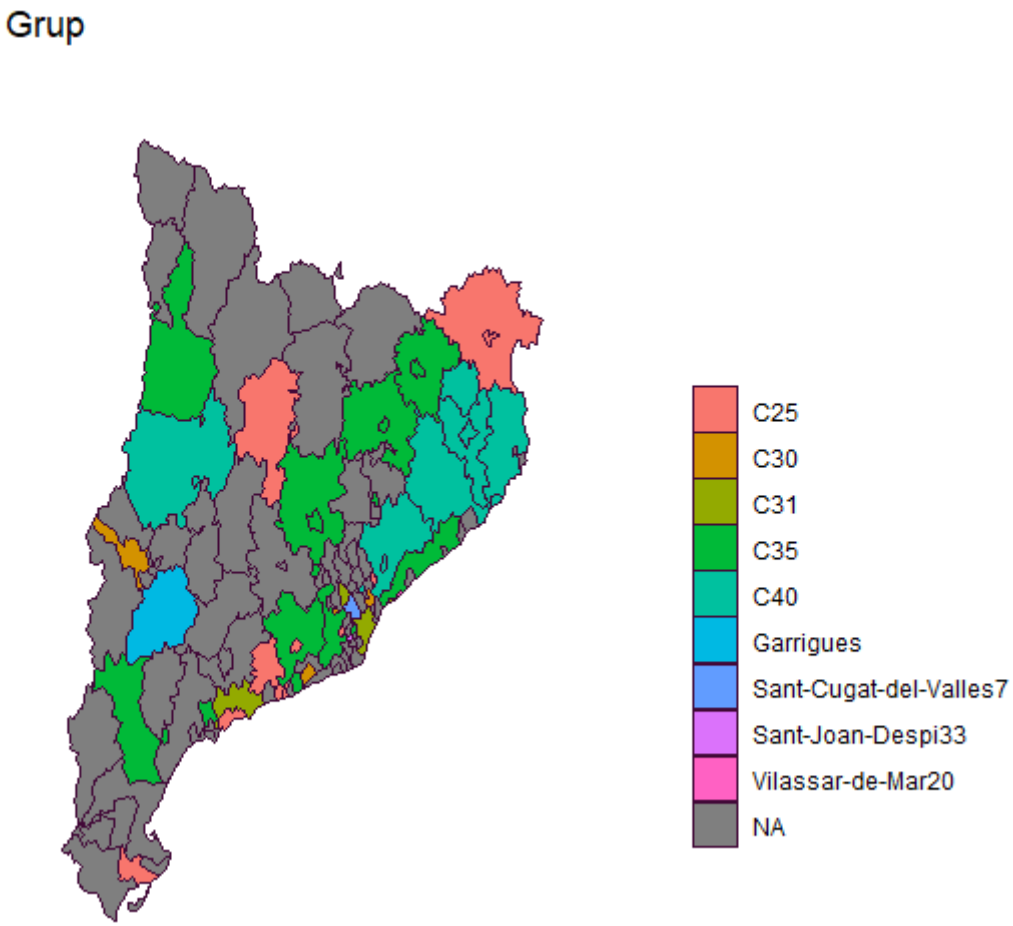

Figura 98: Representació territorial del clustering amb les variables seleccionades

Validació de la classificació estructural

Utilitzant la metodologia explicada a secció 6.5 es compararan 2 classificacions.

Validació numèrica

El test de Lebart amb valors P s'ha calculat utilitzant dues variables que són la variable ABSS i la variable de classe tal com s'explica a la secció 2.8

El S_P en la classificació utilitzant variables seleccionades en TFSM és 0.343, el qual és més alt el percentatge resultant sense aplicar el TFSM, on es veu que S_P=0,29.

Validació gràfica

La primera validació que s'ha fet és gràfica. A la Figura 99 es mostra el mapa quan es fa clustering amb tots els indicadors nous de tercera generació basats en dades.

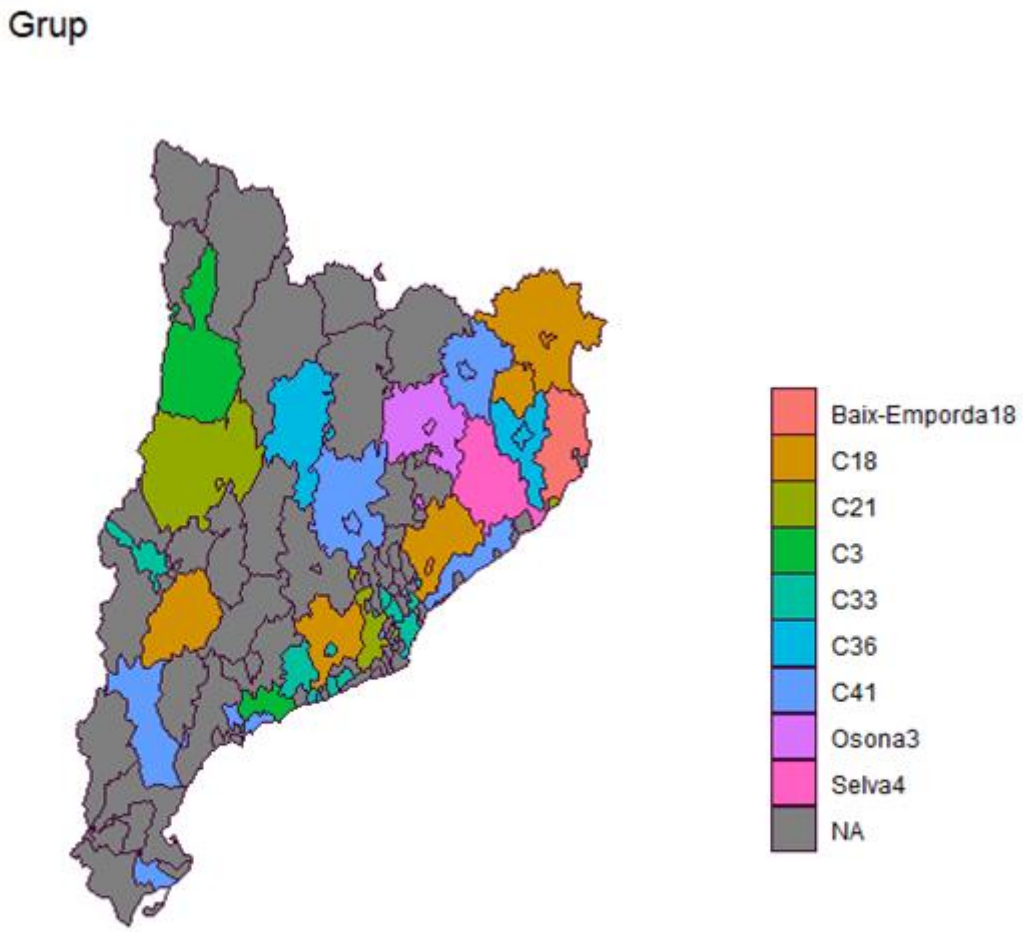

Figura 99: Representació territorial de la classificació obtinguda sense aplicar el TFSM

7.2.13. FASE V Interpretacio dels resultats, elaboració del diagnòstic i recomanacions finals

L'impacte de la crisi de la COVID-19 en els Serveis Socials al llarg del territori en envolta dues dimensions relacionades: l'impacte en les persones que necessiten assistència social i l'impacte en la praxi dels equips professionals de serveis socials. La crisi de la COVID-19 va aparèixer en un moment en què, com a societat, encara no ens havíem recuperat del problema econòmic iniciat el 2008, i ha afegit una càrrega addicional als serveis socials de tot el país que ja estan molt desbordats.

Aquesta pandèmia va arribar quan la nostra maduresa tecnològica va ser menys avançada del que a nosaltres, com a societat, ens hagués agradat. Encara no podem utilitzar les dades com un actiu immediat en la gestió de crisi.

El projecte INSESS-COVID19 mostra com la crisi de la COVID-19 està afectant, d'una banda, en els mateixos segments de població que ja estaven danyats per l'última crisi econòmica (2008) i, d'altra banda, afecta nous grups, creixents necessitats socials que també requereixen l'atenció dels serveis socials: dones i persones majors.

L'alentiment econòmic que va començar en 2008 va afectar el mercat laboral d'una manera especialment negativa i va generar un doble procés d'empobriment degut, d'una banda, a la caiguda dels ingressos i a la creixent desigualtat en la seva distribució de llocs de treball i, per un altre, a l'enfonsament dels ingressos més baixos. Aquesta situació limitava les oportunitats que els individuals i les famílies havien de resoldre les seves dificultats econòmiques i augmentava les diferències socials existents. L'augment de la desocupació, la desocupació prolongada, els salaris precaris, la discontinuïtat laboral o l'ABSS poder adquisitiu de les pensions de jubilació van afeblir les economies familiars, i això va augmentar els problemes i la complexitat de les situacions socials, i va accentuar els processos d'exclusió social de les persones i les famílies. Entre els perfils més afectats es trobaven els que havien perdut els seus llocs de treball, els joves desocupats que buscaven el seu primer treball, les famílies joves amb fills dependents, les dones solteres amb responsabilitats familiars, els homes solters sense llar, les dones majors amb pensions. Immigrants no contributius i irregulars.

L'anàlisi de les 971 respostes recollides de ciutadans pertanyents als 20 perfils i distribuïdes per tot Catalunya, i incloent també alguns individus que mai havien estat usuaris de Serveis Socials, però comencen a ser després de la primera ronda de la pandèmia.

L'informe INSESS COVID19 inclou diverses conclusions, la majoria de les quals provenen d'una anàlisi descriptiva bàsica descrita en obres anteriors [Gibert & Angerri, 2021] i [Angerri & Gibert, 2023]. Entre ells, podem destacar que:

- .L'impacte de la crisi COVID-19 en els serveis socials del territori té dues dimensions que interaccionen entre ells: l'impacte en les persones amb necessitats d'atenció social i l'impacte en la pràctica dels equips de serveis socials professionals.
- . L'estudi INSESS-COVID19 mostra que la crisi de la COVID-19 està provocant realitats socials completament noves i diferents.
- . Entre les persones més perjudicades es troben les dones i les persones majors.

Aquestes conclusions van ser prou rellevants per a donar suport a algunes de les primeres decisions polítiques després del primer confinament.

De les dades de salut mental en particular, les conclusions obtingudes també són senzilles:

- L'impacte de la crisi COVID-19 és alt en la salut mental.
- 41% dels participants necessitaven suport emocional durant el primer confinament.
- 7% dels participants pateixen només ansietat (Fig1)

- 5% dels participants pateixen només depressió (Fig1)
- El 5% dels participants pateixen depressió i ansietat simultàniament al gener de 2020. (Fig1)
- Al gener 640 participants no van tenir cap problema de salut mental diagnosticat (Fig1).

Per a entrar en els detalls, la interpretació de la pregunta S4 és realment difícil de referir-se a la situació al juliol de 2020.

A la pregunta de juliol de 20 de juliol, 610 participants declaren no tenir trastorns de salut mental diagnosticats. Hi ha una pèrdua de 30 persones pel que fa als 640 sense trastorns de salut mental al gener de 2020. Aquests 30 participants trien altres modalitats en la pregunta July20Worse, el que significa que van passar de cap problema de salut mental al gener a algun diagnòstic mental al juliol. En aquest cas això significa que la persona va debutar amb alguns diagnòstics mentals i en el diagnòstic particular del qual se sent pitjor. Una cosa similar passa amb les 18 persones perdudes de Cap categoria de la pregunta S4Jul20 Equal, i aquí, la interpretació és d'alguna manera més confusa, perquè no té molt sentit que una persona sense cap problema mental al gener es mogui a "igual en depressió" o "igual en ansietat" al juliol.

Això és degut al fet que en la primera anàlisi bàsica de les dades originals les variables s'analitzen independentment amb eines descriptives simples i les interaccions entre variables s'ignoren. Per al cas particular de S4, hi ha una variable de línia de base (gener de 2020) i els altres tres representen evolucions respecte a la línia de base que no es poden agafar amb simples eines descriptives.

Es pot extreure molt valor afegit de les dades quan es creen variables derivades mitjançant l'ús del coneixement d'experts en el domini objectiu o la construcció de nous indicadors basats en dades sobre les dades. Aplicant aquesta nova metodologia s'obté coneixement addicional sobre problemes de salut mental després del confinament:

El 5% dels participants van debutar amb un trastorn mental durant el confinament (això es pot quantificar amb la transformació del conjunt de dades original en un conjunt de dummies). No obstant això, cap dels nous problemes mentals adquirits durant el confinament són un trastorn mental greu. Les persones debuten amb depressió, ansietat i trastorn d'estrès posttraumàtic principalment, el 73,8% dels que pateixen algun trastorn mental al gener se senten pitjors al juliol de 2020 (això es va fer evident amb les variables derivades creades).

- Concretament, un 68,86% dels participants amb depressió se senten pitjors al juliol
- Un 72,38% dels participants que patien ansietat se senten pitjors al juliol.
- Un 86,95% dels participants que van patir febre aftosa se senten pitjors al juliol.

Es va crear una variable binària (U30) i es va combinar amb altres variables els resultats per als joves.

- 17% dels joves participants declarats que tenen diagnòstics sobre salut mental (55% d'ells reben medicació, en contrast amb el 96,94% trobat per la població global)
- 36% declara depressió
- 51% declara ansietat
- 12% de trastorn mental greu
- 9% Limita el trastorn de personalitat
- 6% altres trastorns mentals.

El 38% dels joves necessitaven suport emocional. Les variables de tercera generació es poden obtenir a partir del clustering, i la variable qualitativa resultant interpretada i etiquetada basada en el TLP. En crear variables de tercera generació, es pot dir que hi ha 10 grups de persones relacionades amb problemes de salut mental. Tres (3) dels 10 grups descoberts són persones sense problemes de salut mental. Aquests són els principals resultats en salut mental:

- Les persones que reben medicaments es deuen als seus problemes de salut mental.
- Els participants afectats per la malaltia reben medicació.
- Les persones que debuten no debuten en trastorns mentals greus en la primera etapa.
- La gent que feia teletreball necessitava suport emocional.
- Les persones que pateixen un trastorn mental específic tenen característiques similars entre elles

El grup de persones que pateixen ansietat, el de persones que pateixen altres malalties no especificades i que no requereixen suport emocional addicional. Hi ha un tercer perfil amb un petit impacte en els debuts en salut mental, el relatiu a les persones que estaven teletreballant durant la primera onada i aquells que requerien suport emocional de la seva xarxa personal o professionals

Com es veu, la quantitat i la qualitat de la informació augmenta afegint variables de segona i tercera generació.

Això també s'ha demostrat amb el bloc de preguntes sobre salut. La descripció bàsica dona els següents coneixements:

- El 12,25% dels participants havien patit COVID-19 durant el primer confinament.
- 32% dels participants pertanyen a un grup de risc COVID-19.
- El 15,75% dels participants pateixen alguna discapacitat.
- 14,9% dels participants tenen un grau de dependència.

Afegint variables derivades de tercera generació, apareixen 5 grups.

Persones amb COVID greu, tendeixen a ser persones amb discapacitat o dependència

Els participants amb dependència, que pertanyen a un grup, han canviat la seva dependència.

En el bloc d'abús de substàncies, que està fortament relacionat amb el bloc de salut mental, els resultats bàsics obtinguts després de treballar amb variables de segona generació són els següents.El tabac és la substància amb més addictes

- La majoria de les persones no utilitzen substàncies.
- La majoria de les persones no abandonen el seu estat d'addicció després del confinament.

Afegint variables de tercera generació basades en dades, apareixen 9 grups:

- Hi ha persones que no utilitzen substàncies addictives.
- Hi ha persones addictes a totes les substàncies
- El tabac és la substància amb la qual estan en contacte més grups.

Com es veu, el desenvolupament de més variables gràcies a les tècniques de preprocessament i clustering interpretats amb la tècnica TLP, dona més informació.

Una anàlisi addicional del conjunt de dades INSESS-COVID19, incloent-hi l'anàlisi condicional per territoris específics, fa evident que no es van atendre altres necessitats perquè es van centrar en una ubicació i es van ocultar involuntàriament sota els patrons globals trobats en l'anàlisi general. Això va crear la necessitat d'anar més enllà amb un enfocament sistemàtic per informar de la localitat a cada territori (idealment nivell de baix). No obstant això, la mostra no era prou gran per descendir a aquest nivell de granularitat sense un alt risc de violar el secret estadístic i la reidentificació dels participants individuals. Això va augmentar la necessitat de trobar un nou enfocament per permetre informes conjunts per a grups de baix similars, de manera que la grandaria de la mostra garanteixi els principis de privacitat necessaris, mentre que la cohesió territorial també es va considerar.

La base de dades de destinació prové de les respostes del qüestionari INSESS-COVID19 amb 195 preguntes de 19 temes diaris. Per millorar la quantitat i la qualitat de la informació obtinguda a partir d'ella, la base de dades preprocessada aixeca fins a 258 variables útils incloent algunes noves variables de segona i tercera generació que proporcionen valor afegit.

La selecció de les variables representatives per a cada bloc temàtic original de l'enquesta que s'ha d'introduir en el procés de clustering territorial requereix nous criteris que es desenvolupin. El mètode de selecció de característiques territorials proposat es mostra per proporcionar millor clustering de ABSS que les obres anteriors [Angerri & Gibert, 2023]. La comparació dels cúmuls resultants mostra com l'explicabilitat dels cúmuls obtinguts amb TFSM és més alta que l'obtinguda en [Angerri & Gibert, 2023] segons el criteri SP, i també, la

visualització dels cúmuls obtinguts sobre mapes (Figures 16 i 17) mostra grups més cohesionats en el clustering TFSM, on diverses ABSS similars s'agrupen junts en una classe. Les principals conclusions del clustering són les següents: Dels 11 grups obtinguts 5 representen un sol baix i dels altres 6 grups entre 4 i 11 ABSS en grups més grans amb característiques similars que van d'un sol baix (classe Sant Joan Despí) amb persones amb bona situació coexistent, però amb moltes necessitats i precarietat en economia, condicions de treball a la classe Vilassar amb millors condicions en coexistència, deixant, economia i dificultats concentrades a l'accés a les subvencions socials, participació en la comunitat, per una gradació de clústers intermedis on el suport emocional gradualment és menys necessari, el suport econòmic millora, o les condicions de treball milloren. Els patrons que s'han trobat en aquesta anàlisi són extremadament útils per entendre la situació especial dels immigrants (C31, distribuït a través de 5 ABSS amb problemes econòmics, necessitats de treball i dificultats per participar en la comunitat) o altres patrons que podrien estar connectats amb l'activació de nous protocols específics per atendre vulnerabilitats especials. I els resultats es van publicar en pocs dies després del tancament de la recollida de dades, gràcies a la potent metodologia científica de dades dissenyada en el sistema INSESS-COVID.

INSESS-COVID19 mostra com alguns dels impactes més rellevants entre gener i juliol de 2020 els indicadors econòmics enumerats a la Secció 7.2.5 Impacte Econòmic i de Treball Entre la classe treballadora, hi ha un pessimisme que mereix atenció. La gent està preocupada pel seu futur i una part important d'ells pensa que aquesta situació no millorarà ni a curt ni a mitjà termini.

La disminució ininterrompuda dels ingressos de moltes persones i famílies està accelerant l'increment del risc de pobresa en la societat catalana, tant per a la pov-erty moderada com per a l'extrema. Això ha plantejat les demandes de necessitat de serveis socials públics i entitats del tercer sector, així com les sol·licituds d'ajuda i suport econòmic.

Una atenció especial requereix les dificultats sobre les condicions de vida, suavitzades per l'estat d'alarma. El sobtat empobriment d'amplis segments de població tindrà un efecte retardat en la falta de capacitats per a pagar impostos, factures de serveis domèstics com el gas o l'electricitat, el lloguer de la casa o les cotitzacions bancàries per a préstecs i hipoteques. L'estat d'alarma declarat pel govern ha interromput tots els processos de desnonament. No obstant això, tornaran a sorgir en els pròxims mesos, tan aviat com s'abolí l'estat d'alarma, i aquests processos es reactivaran en un context molt pitjor que quan es trenquin.

De fet, l'ERTO (regulació temporal dels procediments d'ocupació), ajuda institucional als autònoms i altres fons econòmics proporcionats pels governs van contribuir d'alguna manera a suavitzar els efectes econòmics de la pandèmia, ineficiències i retards en la gestió i resolució de les sol·licituds va disminuir sensiblement l'impacte positiu que podrien haver tingut.

Fins aquí, la conclusió principal és que la COVID19 ha provocat una crisi que impacita en segments de població ja castigats per la crisi anterior del 2008, que encara no s'han recuperat, creant així un impacte amplificat cap a la pobresa i la vulnerabilitat social en moltes necessitats crítiques, com l'habitatge o el treball.

No obstant això, com s'ha dit abans, hi ha una cosa nova en la crisi de la COVID19, que no es va observar en crisis anteriors i que empitjora encara més les vulnerabilitats socials de les persones. Com s'ha vist en l'apartat 7.2.6, on es mostren els indicadors. La crisi de la COVID19 també és una crisi social i de relacions, i està afectant també a altres segments de població diferents dels afectats per la crisi de 2008, com a conseqüència de les noves vulnerabilitats socials sorgides de les mesures de desgovern social que requereix la gestió de la pandèmia: mobilitat restringida, confinament domiciliari, aïllament social, teletreball, transformació digital accelerada, interrupció i retard dels processos judicials i administratius, etc. Aquestes mesures van causar efectes seriosos per a les dones i les persones majors.

A més, això és molt diferent de les xifres comunes oficials. De fet, conforme a CCI2018, un 58,5% dels usuaris dels serveis socials són dones. La proporció de dones a INSESS-COVID és significativament més alta. A més, això assenyala un major impacte de la pandèmia en la vulnerabilitat de les dones, sempre que el gènere no fos un criteri utilitzat en cap dels 20 perfils objectiu definits per participar en el projecte. Així que, quan en BASS va trobar gent seguint aquests 20 perfils, va passar que la majoria d'ells estaven satisfets amb les dones. De fet, les dones van assumir una pesada càrrega en els pitjors períodes de la pandèmia, els cuidadors informals solen ser dones, famílies de famílies sense pare, dones a càrrec de persones dependents, nens o persones amb discapacitat o trastorns mentals. No obstant això, la majoria dels perfils professionals dels serveis socials i de salut que s'han destacat durant la pandèmia també solen ser obres femenines. Finalment, moltes vídues també són dones, de manera que les dones majors que marxen soles també es veuen greument afectades per l'aïllament, la soledat i els problemes de dependència durant la pandemia.

D'altra banda, les restriccions de mobilitat i l'alta vulnerabilitat a la COVID-19 de les persones majors van causar greus impactes relacionats amb el confinament en aquest segment de població: soledat, aïllament, depressió, bretxa digital, etc.

El qüestionari també rep informació de l'altra banda de la dependència. El costat dels cuidadors informals

Pel que fa a les relacions socials i la participació, en totes les àrees, treballant, familiar, amistats... el patró “VΛ” s'observa durant el període de confinament que implica la doble dinàmica de:

, Reforçar els vincles, augmentar la solidaritat i intensificar la participació, fins i tot en activitats de voluntariat

• Desconnectar i aïllar de parents, amics, veïns, col·legues

Per tant, molts segments de població requerien un suplement psicològic i emocional. Els sentiments solitaris, l'aïllament i les deficiències mentals van sorgir en moltes persones, especialment en persones d'edat avançada.

Finalment, la violència també ha estat present durant la pandèmia com a mostra els indicadors de violència a la secció 7.2.6.

El qüestionari també inclou informació sobre la bretxa digital i la interrupció/retard o processos judicials i administratius (divorcis, regularització, desnonaments, etc.). Els grups més afectats són les dones en diferents formes, i un dels patrons més impressionants és el de les dones víctimes de la violència, que van haver de passar el confinament a casa juntament amb l'agressor mentre s'interrompien els divorcis o les ordres de restricció en els tribunals.

Com a principal avantatge de la metodologia proposada, estem proporcionant una eina per a la participació directa que pot proporcionar accés als ciutadans de consulta (o professionals quan sigui necessari) en temps ràpids, i processar les dades recopilades molt ràpidament. Les preguntes es poden adaptar a cada experiència de l'aplicació, i l'anàlisi es mantindrà automàtica, sempre que el fitxer csv de Metainformació es proporcioni juntament amb el conjunt de dades. En l'addició, un disseny especial del qüestionari pot resoldre el petit nombre de respostes mitjançant la substitució dels enquestats que poden retardar el temps sense perdre la validesa del conjunt de dades durant molt de temps. Finalment, els resultats s'ofereixen en forma o document de treball en Word que no redueix radicalment la capacitat d'utilitzar aquests resultats en reunions estratègiques imme-diadament després de la descàrrega de dades del qüestionari digital.

Com tots els estudis basats en dades ciutadanes, els resultats depenen de la veritat de les respostes proporcionades pels participants.

El temps necessari per completar l'estudi depèn de la celeritat dels participants en la cerca de baix, i el temps de resposta dels participants. Els tallers presencials proposats a INSESS-COVID19 van ser dissenyats per mitigar el temps necessari per a la recollida de dades i els pilots van demostrar la seva eficàcia, fins i tot si la pandèmia es van restringir a treballar sota la modalitat "lliure". Les limitacions per celebrar els tallers tal com es van dissenyar originalment es van resoldre mitjançant el desenvolupament de noves modalitats per als tallers. Contrapartida, el taller lliure augmenta la cobertura, però perd el control del temps per donar la resposta. En qualsevol cas, la tecnologia desenvolupada i les noves eines estadístiques descriptives proposades per forma correcta i ràpida, i proporcionaran resultats molt valuosos quan la resposta al qüestionari sigui obligatòria (això depèn del tema de la consulta)

La metodologia proposada constitueix una eina poderosa per revelar els patrons subjacents de vulnerabilitat social al territori català, però encara no està proporcionant prediccions sobre les necessitats socials de la població en els mesos actuals. Una vegada que s'hagin descobert els patrons, el model predictiu per als patrons específics es podrà arribar amb el següent pas de l'anàlisi i les tècniques de classificació.

Les dades provenen de tot el territori català, però els participants seleccionats identificats pel BASS no estan obligats a respondre, de manera que, alguns d'ells poden saltar-se el pagament i generar dades pobres a partir d'algun BASS. A més a més, podrien aparèixer els territoris entrellaçats per mostres. No obstant això, aquesta heterogeneïtat s'associa amb la característica intrínseca del propi territori, per la qual cosa no és necessàriament errònia. No obstant això, això pot ser compensat mitjançant la crida a nous substituts amb els mateixos perfils i les seves respostes seran vàlides gràcies a la introducció de variables temporals incloses en el qüestionari, fins i tot si estava responent a les preguntes amb un important retard.

7.3. Conusm energètic de les famílies

7.3.1. Contextualització de la base de dades

En el marc de les accions de STEAM de IDEAI i del projecte aquí STEAM de la UPC Aquí STEAM UPC (iniciativa per atreure talent femení als estudis de tecnologia i enginyeria, adreçada específicament a noies d'entre 9 i 14 anys de Catalunya. El programa, que compta amb el suport de la Secretaria de Polítiques Digitals de la Generalitat de Catalunya, vol trencar els estereotips i rols de gènere establerts en la societat i fer visibles nous referents femenins d'una manera atractiva i propera per a les noies) [UPC STEAM,2023].

Es va dissenyar una activitat inspiracional dissenyada per augmentar l'interès en la mineria de dades a alumnes d'una escola de primària.. Inicialment, una enginyera informàtica va visitar una classe primària real i com a part del taller proposava una activitat. Va demanar als nens i nenesque durant 7 dies seguits recollissin algunes dades en relació als consums de casa. A més, es preguntaa de quins electrodomèstics es feien ús diàriament. (en aquest cas, consums elèctrics de les llars i el tipus d'electrodomèstics encesos). Aquestes dades eren introduïdes en un formulari dissenyat amb la metodologia MIPRI2D i analitzades automàticament amb la tecnologia desenvolupada en aquesta tesi. . Uns dies més tard, la mateixa científica torna al centre escolar per tal de compartir amb les criatures els resultats de l'anàlisi de les dades recollides. D'aquetsta manera els infants van poder aprendre quin és l'electrodoméstic que més electricitat consumeix a les seves llar i l'impacte que té el fet de deixar els llums encesos

a casa. A més, van descobrir el funcionament de la tecnologia MIPRI2D en primera persona sent els protagonistes de l'aplicació. El 2022 l'anàlisi es va fer amb la tecnologia INSESS i els nens podien aprendre sobre l'eina que consumeix més electricitat a les seves llars.

7.3.2. Estructura de la base de dades

L'activitat va engrescar tant als infants que la setmana del 28 de març al 4 d'abril de 2022 es van recollir un total de 175 respostes al qüestionari, ja que la majoria d'infants van recollir diàriament les dades de la seva llar. A la taula, es pot observar la composició de la base de dades, en funció del tipus de variables que conté.

Tipus de variable	
Variables Numèriques	12
Variables categòriques	15
Variables multivaluades	0
Variables de quadrícula	0
TBV	0
Quadrícules multivaluades	0
Variables TQQ	0

Taula 49: Estructura de la base de dades

7.3.3. Principals resultats obtinguts

D'aquesta aplicació en un àmbit molt local, els resultats més destacats que es poden extreure són els següents:

- La recollida de dades es va fer en llars que en la seva gran majoria estan habitades per 2 adults i 2 infants amb un o dos banys a la llar.
- Van sopar a casa una mitjana de 3.47 persones la setmana de la recollida de dades, mentre que només ho van fer 1,61 persones de mitjana.
- Més del 50% de les ocasions els infants van fer ús de la rentadora, la televisió, el forn o el microones
- Menys del 50% dels dies els infants assenyalen que no fan ús del rentaplats, l'aspiradora o la *Roomba,* la planxa, l'assecador de cabell,
- Un 31% de les ocasions els infants han respost que no tenen *Roomba* al seu habitatge, així com ho han marcat en un 25% de les ocasions per al rentaplats
- El 54% de les ocasions es deixen els electrodomèstics en StanBy enlloc d'apagats.

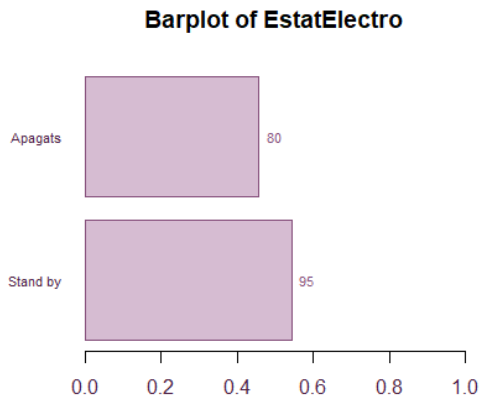

Figura 100: Diagrama de barres de Electrodomestics

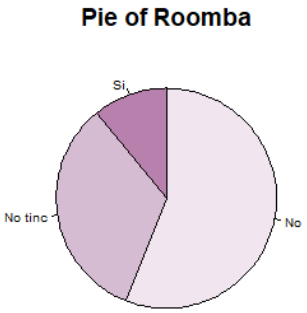

Figura 101: Diagrama de pastís

7.4. Enquestes de valoració a una associació de sense ànim de lucre

7.4.1. Contextualització de la base de dades

Colla Castellera Jove de Barcelona, és una assocació sense ànim de lucre que té com a finalitat la pràctica castellera, mitjançant castells. Segons [Sole 2021] els castells són una mostra folklòrica que ha format part de la cultura catalana des de l'any 1791, quan es va formar la primera colla castellera a la ciutat de Valls. Tenen l'origen en el Ball de Valencians, un ball arrelat a la Catalunya Nova durant una part dels segles XVII i XVIII [Cervelló, 2017]. Són construccions formades per un conjunt de persones que, enfilant-se ordenadament els uns sobre les espatlles dels altres, formen torres humanes de diversos pisos d'alçada. [Brotons, 1995]. amb

La colla castellera Jove de Barcelona realitza una enquesta de valoració anual entre els seus membres per tal que aquests puguin valorar l'activitat anual de l'entitat. L'activitat anual acaba al voltant del 20 de desembre i els resultats de l'enquesta són presentats a l'Assemblea General Ordinaria que té lloc al mes de Gener.al gener. L'Enquesta Anual d'Opinió 2021 i 2022 s'han analitzat a través de consultes INSESS i informes automàtics proposats en aquest document.

7.4.2. Esctructura de la base de dades

Any 2021

Tipus de variable

Variables Numèriques	0
Variables categòriques	6
Variables multidentades	0
Variables de quadrícula	5
TBV	2
Quadrícules multivaluades	0
Variables TQQ	0

Taula 50: Estructua de la base de dades per al 2021

Any 2022

Tipus de variable	
Variables Numèriques	0
Variables categòriques	8
Variables multivaluades	0
Variables de quadrícula	8
TBV	0
Quadrícules multivaluades	0
Variables TQQ	0

Taula 51: Estructura de la base de dades per al 2022

L'estructura del qüestionari es troba a la figura 102

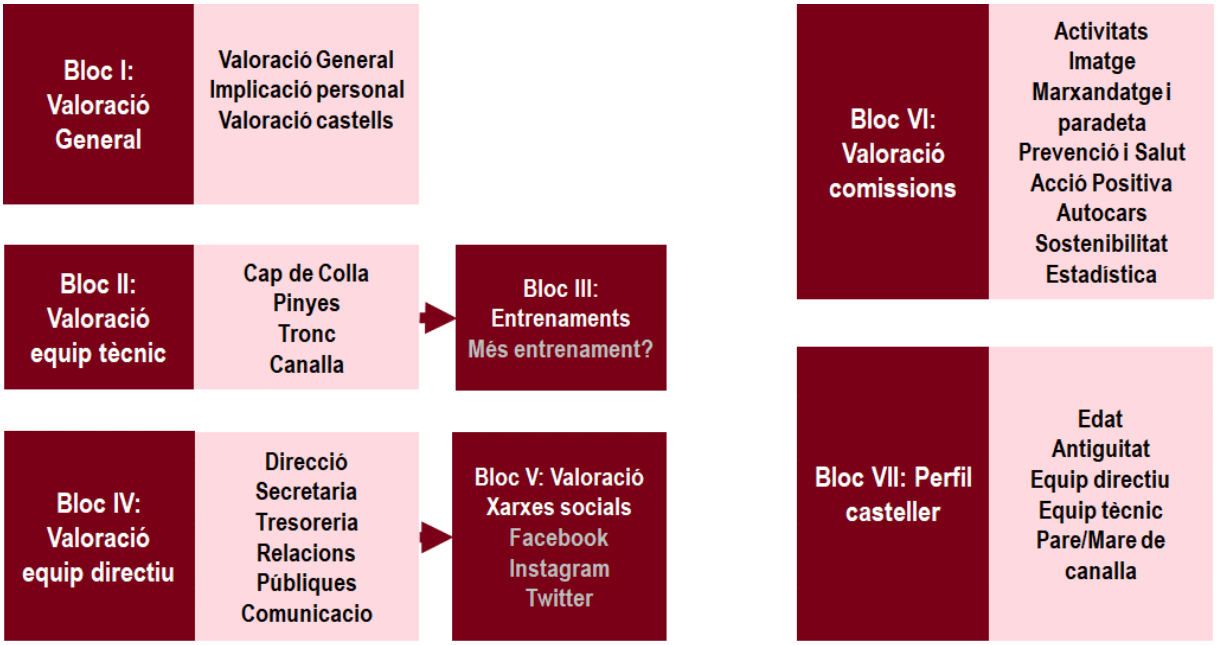

Figura 102: Estructura de la base de dades per a la

7.4.3. Resultats

Els resultats de les valoracions no són de domini públic, ja que aquestes són confidencials. Per tant, no poden ser publicades en aquesta tesi. Tanmateix, una de les preguntes és oberta I es dona l'oportunitat als participants d'opinar d'allò que considerin més oportú de manera totalment lliure. La metodologia MIPRI2D genera aquest núvol de punts per tal de mostrar els resultats de les preguntes obertes.

Figura 103: Núvol de punts per a les respostes obertes

Com es pot veure a la figura 103 hi ha molts conceptes, però la gran majoria esmenten la paraula colla, el concepte que motiva l'enquesta.Entre els comentaris que hi ha s'esmenten diferents comentaris l'equip tècnic, anomenat també com la técnica. Entre ells hi ha

felicitacions per a la temporada que s'ha fet. Es considera que en general s'ha fet una bona feina.

7.5. DIMCARE

7.5.1. Contextualització del projecte

El projecte europeu DIMCARE té per objectiu analitzar el nivell de digitalització en el teixit de l'economia social i solidària que es dedica a la cura de les persones, per tal d'identificar les seves necessitats en aquest àmbit. El projecte busca abordar la necessitat de digitalització en el sector de la cura a les persones dins de l'ecosistema de l'economia social i solidària. [DIMCARE, 2023] [Mataro, 2023]

Dintre del projecte hi ha un paquet de treball dedicat a l'anàlisi quantitatiu de la situació, per al qual es va dissenyar un qüestionari que va ser omplert per les entitats del tercer sector de les ciutats de Mataró (Espanya),Prato (Itàlia) I Varazdin (Croacia).

Estructura de la base de dades

La base de dades està composta de la següent manera:

Tipus de variable	
Variables Numèriques	0
Variables categòriques	2 3
Variables multivaluades	7
Variables de quadrícula	2
TBV	0
Quadrícules multivaluades	0
Variables TQQ	0

Taula 52: Estructura de la base de dades DIMCARE

Resultats

L'anàlisi de les dades es va realitzar amb la tecnologia desenvolupada per a la metodologia MIPRI2D. Tanmateix, aquests resultats no poden ser exposits en aquesta tesi degut a la confidencialitat de les dades.

Conclusions

L´única part de la metodologia MIPRI2D aplicada en aquest projecte és l'anàlisi de les dades. Aquest fet va obligar a fer el disseny del model de metainformació partint d'un qüestionari dissenyat prèviament, aspect que va generar que hi hagués preguntes que fossin impossibles d'analitzar a causa de l'heterogeneïtat de les modalitats que hi havia com a resposta.

A més, una de les preguntes no es va poder analitzar, ja que la interpretació de la pregunta havia estat diferent per a unes entitats i per a unes altres. Aquest fet és degut al fet que no es van dur a terme els tallers dissenyats en la tecnologia MIPRI2D.

8. Conclusions

Aquesta tesi aporta contribucions en l'àmbit de la ciència intel·ligent de dades per a poder realitzar informes automàtics sobre dades recollides directament d'individu objecte d'estudi, per medis tecnològics de forma molt ràpida que permetin tenir un diagnòstic ràpid i complert sobre un fenomen, i prendre decisions informades de forma efectiva i en situacions crítiques on les dades sistèmiques no contenen la informació rellevant per la decisió en qüestió. En aquesta tesi el diagnòstic s'ha materialitzat en forma de perfils interpretats automàticament i que descriuen com s'estructura el fenomen en estudi.

És molt rellevant destacar, que en contra del que darrerament s'ha fet més popular en l'àmbit de la ciència de dades i les tècniques d'intel·ligència artificial basades en dades (aprenentatge automàtic), en aquesta tesi no es construeixen models predictius, ja que ens situem en un moment anterior del procés cognitiu, que és la ***comprensió*** d'un fenomen complex, per entendre què passa, i no encara per poder-nos anticipar, si bé s'ha vist a la tesi que els resultats obtinguts sí que han permès prendre decisions per corregir o mitigar events no desitjats, i prevenir-los a futur, però no encara predir-los amb anterioritat. Les tècniques d'aprenentatge no supervisat són una branca importantíssima de la Intel.ligència Artificial que tenen per objecte ***descriure i caracteritzar*** les estructures subjacents en fenòmens de certa complexitat, i en aquest sentit estan en un procés anterior del procés cognitiu. S'ocupen de fer-nos ***conèixer*** què passa, i no de ***reconeixer*** quan es dona el que ja sabem que passa, que seria el que farien els mètodes d'aprenentatge supervisat, altrament dits models predictius d'intel.ligència artificial.

La proposta de tesi s'ha emmarcat en una metodologia paraigües, global, que aporta directrius per estructurar tot el procés des de la recollida de dades fins a l'elaboració de l'informe final de resultats, realitzat aquest de forma automàtica. El procés és molt complex, i ha requerit recerca per poder avançar en moltes de les seves passes intermèdies, el que ha donat lloc a diferents mètodes i tècniques per abordar algunes fases que es van presentant a la tesi amb nom propi i que finalment s'integren totes en

una única metodologia general. General perquè resol el problema global, però també perquè resol un problema genèric, que no està lligat a cap aplicació concreta, sino que s'adapta de forma molt versàtil a qualsevol domini d'aplicació on hi pugui haver dades disponibles o recol.lectables.

La metodologia MIPRI2D, la contribució principal d'aquesta tesi, és un conjunt de passes que permeten recollir ràpidament dades directament dels individus de la població diana i generen informes automàtics que descriuen la població en estudi a partir d'un conjunt de perfils distingibles i automàticament construïts i interpretats i que, essent territorialment consistents, permeten entendre punts forts i dèbils de cada perfil i encaminar decisions, estratègies i polítiques públiques (segons l'aplicació) a partir dels perfils. MIPRI2D s'ha estructurat en 5 fases que treballen diferents etapes del procés i que eventualment constitueixen contribucions independents de la tesi que es podríen fer servir aïlladament en altres contextos.

En la Fase I de MIPRI2D s'aborda tot el referent a la construcció de les eines necessàries per obtenir les dades a analitzar. Es proposa fer-ho a partir del que hem anomenat consultes INSESS, i que s'estructuren al voltant d'uns tallers que poden tenir diferents formats i on participen els individus objecte d'estudi. L'objectiu principal d'aquests tallers és centralitzar la recollida d'informació en moments concrets del temps. L'element clau dels tallers són els qüestionaris, que es dissenyen específicament per cada aplicació, o s'aprofiten de l'estat de l'art si el motiu de la consulta INSESS ho permet o es repeteixen d'una consulta INSESS prèvia si es vol fer seguiment d'un procés analitzat sota MIPRI2D prèviament. Els tallers accel.leren el procés de recollida de dades alhora que disminueixen els riscos d'haver d'eliminar preguntes dels qüestionaris recollits per mala interpretació o manca de respostes. El resultat dels tallers és una base de dades per ser analitzada amb tècniques també enriquides amb desenvolupaments propis d'aquesta tesi.

Una de les components més crítiques de la proposta de tesi és tota la infrastructura tecnològica i logística que s'ha dissenyat per garantir l'èxit i tempos curts de materialització d'una consulta INSESS, entre els que cal comptar, una web de suport, el qüestionari, les plantilles de mailing per captar els participants, els scripts d'anàlisi de dades, de preprocessament, etc etc. Entre aquestes eines podem destacar la construcció d'un qüestionari digital molt senzill d'ús, adequat per recollir informació fresca i directa dels individus objecte d'estudi gairebé immediata, preparada per ser analitzada amb processos genèrics i automatitzats que poden ser molt útils per fer front a situacions noves i inesperades, com la gestió d'emergències i situacions disruptives (com ho va ser COVID-19). És interessant tenir en compte que una eina així pot tenir un impacte molt significatiu en la forma de fer política perquè obre la porta a les consultes participades als col.lectius rellevants en cada cas i l'obtenció directa d'informació rellevant en decisions per les que potser no es disposa de dades sistèmiques.

La metodologia és flexible per treballar en tota mena de qüestionaris i poblacions amb modificacions mínimes. La metodologia MIPRI2D proporciona una nova eina per obtenir coneixement rellevant per a la presa de decisions de molts nivells diferents, des de la més operativa fins a la més estratègica, incloent-hi el suport a la presa de decisions polítiques i el suport a l'elaboració de polítiques públiques.

MIPRI2D representa doncs una metodologia innovadora per recollir, analitzar i reportar dels resultats d'aquestes dades als responsables que hagin de prendre decisions.

Pel que fa a les components de MIPRI2D anirem describint les contribucions realitzades a la tesi.

Dins la proposta metodològica que presentem, s'aporta un mecanisme especialment dissenyat a la tesi per construir qüestionaris atemporals, que permeten obtenir informació d'un determinat moment del temps, i si s'escau, veure l'evolució temporal d'una determinada característica, independentment del moment que es reculli la informació. Aquesta característica és clau per poder allargar els processos de recollida de dades en el temps fins a reunir un nombre suficient de respostes, però referides totes al mateix moment del fenòmen en estudi, sigui quan sigui que respon el participant. Pel cas concret a l'aplicació INSESS-COVID19 aquesta característica va ser realment rellevant. Es va estructurar el qüestionari de manera que es demanava informació per Gener 2020, Juliol 2020 i Gener 20201 (prospectiva) i mentre les ABSS estaven tancades per COVID, els respondents confinats a casa i els professionals de serveis socials també es va fer necessari que la recollida de dades que havia iniciat al juliol es pogués allargar fins al desembre, i es poguessin recopilar les respostes de fines a 971 ciutadans. Una grandària de mostra molt considerable si ens referim a dades de població vulnerable.

Pel que fa al disseny de qüestionaris, permetem que les preguntes presents en un qüestionari puguin incloure un seguit de tipus de variables nous, també dissenyats específicament en aquesta tesi amb l'objectiu d'enriquir l'expressivitat del qüestonari i representar qüestions de gran complexitat. A part de les variables més clàssiques com les númeriques o les categòriques de resposta simple o múltiple, els qüestionaris MIPRI2D admeten nous tipus de variables que han estat introduïts en aquesta tesi, com les variables bàiques temporals o les TQQ (Temporal Qualified Qualitative) que recullen preguntes d'estructura més complexa i amplien la capacitat del qüestionari.

A més, per tal de calcular correctament la grandària mostra, s'ha construït un petit simulador que permet analitzar els errors de mostreig associats a diferents preguntes davant algunes configuracions dades. Com era d'esperar, hem pogut provar que per a una població molt gran, el càlcul de la grandària de la mostra per poblacions infinites coincideix amb les de població finita i elaborar una proposta de càlcul de la grandària de la mostra que afita la precissió dels resultats i tindrà repercussions en la definició d'una

política de publicació de resultats que garantitzi la protecció de la privacitat i eviti el risc de reidentificació de l'individu o el que és el mateix, el risc de violació del secret estadístic.A més, s'aporten també criteris per calcular l'error de mostreig, sobre la base del que fan les estadístiques oficials a Catalunya i Europa.

Per tal de poder garantir universalitat en el preprocessament i anàlisi de les dades, s'ha formalitzat un model de metainformació MdM crucial per al correcte desenvolupament de la metodologia i la generació automàtica d'informes intel·ligents sobre qualsevol tipus de qüestionari que utilitzi els tipus de variables reconeguts en la metodologia MIPRI2D. Aquesta tesi contribueix amb una proposta d'estructura de metainformació oberta que es pot implementar en qualsevol eina de suport digital (un fitxer csv, una base de dades Oracle, un artefacte de dataspace...) capaç de llegir, gestionar i descriure bases de dades i de connectar amb el llenguatge de programació que realitzi l'anàlisi pròpiament dit. Disposar d'un model de metainformació machine-readable que descriu la base de dades objecte d'anàlisi, permet fer una passa ràpida i automàtica de preprocessament i construir un informe automàtic, on s'utilitzen diverses eines descriptives que depene del tipus de variable, també indicat al model de metainformació. D'altra banda, ja que el Model de Metainformació està dissenyat a priori, juntament amb el qüestionari, i la tecnologia proposada INSESS és flexible per a qualsevol combinació de base de dades+MdM, obtenir resultats finals es converteix en un procés accelerat, ja que el primer esborrany d'informe s'obté en molt pocs minuts després de tancar el procés de recollida de dades. Això contrasta fortament amb els tempos habituals d'un procés d'aquestes característiques quan parlem de dades sistèmiques. Sense anar més lluny, pel cas de Serveis Socials, la situació a l'inici del projecte INSESS-COVID19 era que a gener 2020 el departament disposava com a eina de decisió de les xifres oficials de l'informe RUDEL 2019, que recull els resultats de totes les dades de territori de 2018. Això dibuixa l'escenari esgarrifós que, el procés de centralització de dades de serveis socials des dels ajuntaments al departament de Serveis Socials de la Generalitat de Catalunya triga dos anys, entre que els Ajuntaments omplen els formularis anuals, el CTTI els rep, els depura, els analitza, filtra els resultats, els interpreta, realitza les consultes necessàries al món local per anotar correctament els resultats, elabora l'informe final i el remet al departament de Serveis Socials. Aquest és un procés relativament freqüent quan parlem d'elaboració d'informes sobre dada sistèmica o informes periòdics amb xifres i estadístiques oficials. Els períodes de temps poden ser més curt si la dada distribuïda al món local està harmonitzada o no, però en tot cas, estaríem parlant de dos anys en el cas de Serveis Socials i potser un en altres situacions, però no molt menys. Això significa que moltes de les polítiques públiques es sustenten amb dades de dos anys anteriors quan es dissenyen, i que, si el domini on impacta la política en qüestió té una dinàmica d'evolució més ràpida, la política corre el risc de ser obsoleta abans d'estrenar-se. Aquesta no és una situació específica de l'administració pública i en moltes organitzacions es donen situacions similars encara. Si a això li afegim que quan hem de decidir sobre una qüestió més nova, és freqüent que

la dada que necessitem no s'hagi recollit mai encara, es trigarà vora un any en modificar protocols per incloure-la i els dos addcionals fins a reportar-la des que es recull, la dada oficial necessària per la decisió arribaria TRES anys tard respecte del moment en que apareix la necessitat de tenir-la damunt la taula. És per aquest motiu que disposar d'una eina com les consultes INSESS que en escasos dos mesos pot obtenir informació directament de la població d'interès i disposar d'un primer informe final que descrigui la situació, obre la porta a incrementar el nombre de polítiques participades, però també a poder prendre decisions basades en dades davant situacions noves amb molta cel.leritat. Quan encara el nou fenomen és susceptible de ser influït per una decisió oficial. En el cas d'INSESS-COVID19, els dos anys de delay en les dades de Serveis Socials van quedar reduïts a 2 mesos de recollida de dades i 15 dies per tancar l'informe final, entre obtenir l'informe automàtic i afegir-hi les contextualitzacions, discussions i conclusions sobre els resultats oferts per les dades. I, de fet, el lapsus dedicat a la recollida de dades va venir fortament influenciat per la pandèmia, el confinament i les restriccions per fer els tallers. Però és un tempo que depèn completament de la reactivitat dels consultats per respondre el qüestionari i en cap cas depèn de la tecnologia de suport. Si els participants responen depressa (el qüestionari es responia en 15minuts). Si tothom hagués respost en una setmana per exemple, entre el llençament de la consulta i l'informe final haguéssin passat únicament 3 setmanes.

La transcendència d'aquesta contribució rau en el fet que permet veure el que està passant, permet entendre les principals tendències en diferents parts del territori i prendre les decisions necessàries molt ràpidament. Per exemple, en el cas d'INSSES-COVID19 es va veure que les dones maltractades s'havien quedat aïllades confinades en mans dels seus maltractadors a la primera onada, i per la segona es van poder habilitar espais públics on la gent que no es volia/podia confinar a casa, es podia dirigir.

Addicionalment, la consulta INSESS posa de manifest quins indicadors del qüestionari resulten més rellevants per a futurs estudis on fer seguiment o analitzar els paràmetres rellevants a afegir a les bases de dades sistèmiques, o a introduir com a variables d'entrada en la construcció demodels predictius que, ara si, ajudin els responsables a anticipar-se. A més, els resultats obtinguts del qüestionari, després d'un procés automàtic de preprocessament generat automàticament produeixen les dades principals que podríen fàcilment alimentar sistemes de simulació de suport a la presa de decisions.

Per tal de poder realitzar el ràpid preprocessament i anàlisi de les dades recollides amb l'instrument de la consulta INSESS, s'han desenvolupat algorismes de prepocessament automàtic de les dades que treballen no només amb les dades recollides, sinó amb la implementació del model de metainformació. En el cas d'aquesta tesi s'ha optat per un fitxer csv, donat que el nombre de variables no era molt gran (pocs centenars en el pitjor dels casos) i que la implementació s'ha fet en R en un sistema local que llegeix molt

fàcilment aquest tipus de fitxers. Tot i així MdM es pot implementar en tecnologies cloud i connectar a qualsevol altre tipus d'algorisme de processament de dades (preprocessament o anàlisi) connectat a dades que estiguin també en altres arquitectures. Aquesta contribució és interessant, perquè permet escurçar de forma considerable la fase del preprocessament de les dades, una de les etapes més costoses en el desenvolupament d'un projecte d'anàlisi de dades, i quantificat en els estàndards amb el 80% del temps del cicle de vida d'un projecte de dades. Amb la present proposta, aquest temps es redueix a uns pocs minuts.

En definitva, el temps entre la finalització de la recollida de dades (o el tancament del qüestionari) i la publicació dels resultats es torna extremadament més curt si s'aplica la metodologia MIPRI2D mitjançant consultes INSESS, convertint la proposta en una eina molt potent per donar suport a la presa de decisions basada en dades, amb un valor afegit especial per a aquelles decisions que requereixen informació no disponible en les bases de dades del moment i requereixen consulta directa a una població de referència (població vulnerable, associats d'una ONG, entitats del tercer sector, llars d'un determinat districte... entre d'altres) que anirà canviant en funció de la pròpia consulta.

A més, aquesta tesi contribueix a la ciència de dades amb noves eines de visualització de dades o descriptives per tal de descriure les variables d'estructura complexa introduïdes en aquesta tesi i que permeten estudiar aspectes més complexos dels fenòmens d'interès. Eines com el diagrama de teler o la taula de transicions permeten extreure patrons complexos de les dades que amb les eines clàssiques d'estadística descriptiva no es poden obtenir. El diagrama de teler és una eina molt visual que facilita la comprensió de variables temporals, permetent observar la tendència de la població en un aspecte concret al llarg del temps. Aquesta aportació és de gran valor afegit a la ciència de dades, facilitant a persones no expertes en mineria de dades la comprensió del que assenyalen les dades. Els diagrames de teler tenen sentit en aquelles variables que s'observen de forma replicada al llarg del temps. Per exemple, en el cas de INSESS-COVID19 que es van demanar respostes per gener 2020, juliol 2020 i gener 2021 de molts diferents aspectes, com per exemple la situació laboral, el diagrama de teler permet visualitzar de forma molt i molt intuïtiva com evoluciona l'estabilitat laboral de les persones al llarg de la pandèmia i quins patrons aparèixen. I permet estudiar quin perfil de persona hi ha darrera de cada perfil (dels que han perdut la feina durant el confinament, dels que han aconseguit contractes nous etc etc). Associat a aquest tipus de diagrama, s'han introduït altres eines de caràcter numèric, com les taules de transicions, que compten per exemple quàntes persones passen d'una situació a una altra en dos moments consecutius del temps i estudiar quan els camins d'evolució es bifurquen a partir d'un cert moment què hi ha de diferent entre les persones que segueixen un camí o l'altre. Les Variables tipus TQQ poden representar estructures més complexes com per exemple el nombre de membres en una família (numèrica) segregat per diferents graus de dependència (ordinal o Likert) i distribuïts per grups d'edat

(ordinal) entre d'altres. Sense entrar en tècniques d'estadística multivariant, aquesta tesi ofereix una evolució de les eines descriptives uni i bivariants que permeten ampliar els informes bàsics d'un qüestionari amb dades de gran riquesa expressiva i que permeten estudiar patrons de certa complexitat.

Adicionalment, en aquesta tesi s'específica una metodologia per a garantir que la publicació de resultats sobre l'anàlisi de dades preserva el secret estadístic per a les subpoblacions minoritàries, de manera que la informació pugui utilitzar-se per a les decisions sense risc de revelar la identitat dels participants. Aquesta contribució és transcendent per a la comunitat científica, perquè permet la difusió de resultat de manera pública i alhora de forma anònima, fent que la transparència augmenti.

Pel que fa a la generació d'informes automàtics, es preveu també la generació d'informes parcials, basats en tècniques l'anàlisi condicional per algun factor de condicionament. En el cas d'INSESS-COVID19 s'han realitzat estudis condicionats a territoris específics (vegueries), de gènere, de joves, de persones grans, que posin l'accent en aquelles necessitats específiques del col.lectiu o interès i que potser fan aflorar patrons locals que han quedat emmascarats en l'informe general

Pel que fa als informes de territori, la mostra no era prou gran per descendir a aquest nivell de granularitat sense un alt risc de violar el secret estadístic i la reidentificació dels participants individuals. Això va crear la necessitat de trobar un nou enfocament per permetre informes conjunts per a grups de territoris similars, de manera que la grandària de la mostra garanteixi els principis de privacitat necessaris, mentre que la cohesió territorial també es pugui garantir. Així, en la Fase IV de MIPRI2D es proposa la identificació de perfils territorials consistents, que han de premetre realitzar els informes locals amb garanties de privacitat.Aquest procés, es sustenta en tècniques d'introducció de noves variables derivades i tècniques de clustering multivista, de les que parlarem més avall, i eines de suport a la interpretació automàtica de perfils, de desenvolupament propi del grup de recerca dirigit per Karina Gibert, algunes de les quals es remonten a l'inici dels 2000. No obstant això, en aquesta tesi s'avença en aquesta direcció amb eines intel.ligents noves que permeten arribar més lluny en el procés d'interpretació automàtica dels perfils. Així doncs, s'ha introduït la formalització del termòmetre com una eina d'adquisició de coneixement que permet injectar la semàntica de les variables en els processos d'interpretació i que s'utilitzarà en el DD2gI que descriurem més avall. Si bé és cert que en treballs anteriors del grup de recerca s'havien fet algunes proves incipients amb aquesta eina, en aquesta tesi s'ha pogut formalitzar i delimitar com utilitzar-la de forma integrada en la metodologia proposada. A part d'utilitzar-la per conèixer la semàntica de les variables, s'ha integrat en el procés de construcció automàtica dels TLPS, tot enriquint-los amb la semàntica de les variables definida pels termòmetres. Els quadres semafor basats en termòmetres, determinen automàticament els colors de les caselles en funció de les indicacions recollides en el

termòmetre, la qual cosa augmenta significativament el potencial interpretatiu del quadre semàfor, que fins al moment s'havia construït mitjançant la inspecció visual de les distribucions condicionals de les variables i representades en els CPG combinades amb els resultats d'algunes proves d'hipòtesi. Amb el quadre semàfor basat en termòmetres s'obté un mecanisme objectivable per construir el quadre semàfor. La proposta ha mostrat resultats molt prometedors en aplicacions reals i s'ha validat pels experts d'INSESS-COVID19 per valoració comparada dels TLP basats en termòmetres o tradicionals.

Pel que fa a la introducció de noves variables derivades que enriqueixin l'expressivitat de la base de dades i es puguin incloure en el procés de clustering, la tesi planteja mecanismes per fer-ho de forma que el procés final d'identificació de perfils es pugui explicar en termes més propers a l'univers conceptual dels experts. De fet, la combinació de dades quantitatives i qualitatives amb coneixement humà a priori sobre el fenòmen d¡interès i específic del domini d'aplicació permet l'aplicació de tècniques de de ciància de dades complementàries orientades a enriquir la base de dades originalment recollit amb el qüestionari amb informació addicional. Així, d'una banda, s'aproxima la recerca al marc conceptual dels experts i als seus mecanismes de raonaments naturals, de manera que els resultats guanyen explicabilitat i, d'altra banda, obre la possibilitat d'explorar relacions i associacions més complexes entre diferents variables de la base de dades que puguin jugar un paper rellevant en la modelització. Aquesta tesi és una contribució fonamental a la ciència de dades, ja que aborda la importància de derivar noves variables a partir de les originals en processos de preprocessamnt de dades previs a l'anàlisi multivariant. La proposta se centra en tres mecanismes que tracten aquesta tasca des d'un enfoc molt diferent, però que proporcionen un valor afegit complementari a les dades originals enriqueixen l'anàlisi addicional i la obtenció d'informació complementaria, augmentant així la quantitat i la qualitat de coneixement obtingut amb l'anàlisi de dades, el qual millora, necessàriament, la qualitat de les decisions que es puguin prendre a partir dels resultats obtinguts amb aquesta metodologia. A aquest respecte, la tesi sistematitza les formes habituals de construir indicadors derivats sota el que hem anomenat variables derivades de 2a generació basades en coneixement que es construeixen a partir de la interacció amb experts de domini per obtenir criteris quantitatius o qualitatius que representen els criteris dels experts. Però introdueix dos mecanismes més consistents en la clusterització de conjunts parcials de variables per reduir la dimensionalitat de les dades i equilibrar el pes dels diferents blocs temàtics en un número reduït de variables potencialment més discriminants: els indicadors de 2a generació basats en dades (DD2gI) utilitzen mètodes de clustering per a la derivació d'altres criteris qualitatius que representen nous indicadors sintètics. A més, s'introdueix el concepte de variables de 3aa generació, que agrupen en una única variable qualitativa que representa el concepte d'un bloc determinat, obtenint al final tantes variables com blocs temàtics conté la base de dades. L'aplicació de quadres semàfors basats en dades a les variables resultants dels processos

de clustering permet la interpretació i etiquetat automàtic de les classes resultants i la generació de noves variables qualitatives convenientment interpretades que poden passar a formar part de nous models multivariants o predictius com una variable més.Amb aquesta nova base de dades enriquida amb variables originals, de 2a i 3a generació, s'introdueix valor afegit ar la tasca de trobar perfils distingibles d'individus que descriguin, per exemple, el territori. Aquesta tasca, no es pot abordar adequadament quan les dades s'organitzen per blocs temàtics amb un nombre de variables desigual a cada bloc i la proposta que presenta aquesta tesi fa ús de mètodes de clustring multivista per a l'anàlisi global. En aquest cas s'ha assajat l'interès de treballar el multivista amb els blocs desiguals o trobar la millor manera de determinar les variables representatives de cada bloc per utilitzar en el procés de clustering posterior.

A la secció 2.8 d'aquest document s'estudia l'estat de l'art pel que fa a mètodes de selecció de variables. La majoria de referències de la literatura estan orientades a la selecció de variables en el camp de l'aprenentatge automàtic supervisat i les que hem trobat d'aprenentatge no supervisat s'emmarquen dins l'àmbit dels models de case-based reasoning. El cas que ens ocupa, es centra en la realització d'un clustering, que és troba en el camp de l'aprenentatge automàtic no supervisat, però queda lluny dels supòsits dels mètodes d'aprenentatge no supervisat vistos a l'estat de l'art. A més, MIPRI2D presenta una restricció addicional que requereix atenció; les dades són territorials i la cohesió territorial dels clústers resultants també és un objectiu, en cas contrari, la presa de decisions podria incórrer en algunes inconsistències. Per aquest motiu s'aporta la metodologia TFSM que al seu torn descansa en un criteri innovador també de potencial explicabilitat d'una variable respecte d'una variable de classe, i que permet determinar un ranking de les variables i triar per cada bloc temàtic les que millor discriminen les classes. Aplicant aquesta metodologia es pot realitzar una classificació final basada en les variables representatives de cada bloc temàtic i interpretar-la amb les eines de suport aportades també a la tesi com el semàfor basat en termòmetres. Amb això es completaria l'objectiu principal de la tesi, que era d'identificar perfils distingibles i interpretables cohesionats territorialment.

Les contribucions d'aquesta tesi s'han il·lustrat amb 4 casos diferents, ja que la metodologia proposada és general i es pot aplicar a qualsevol altre tipus de conjunt de dades (sigui procedent d'una enquesta o no) sempre que les dades es relacionin amb diferents blocs temàtics (que actualment és una situació molt comuna).

Com s'ha vist a la tesi, la proposta metodològica s'ha aplicat principalment al projecte INSESS-COVID-19. Aquest fet ha permès contribuir amb el primer informe sobre l'impacte del confinament de primera onada COVID19 a la població vulnerable de Catalunya, posant xifres als problemes que van sorgir de la primera onada de la pandèmia. Aquest informe, presentat al Palau Robert de Catalunya el 15 de desembre

de 2020 davant les directores generals de Serveis Socials i Igualtat i feminismes, va tenir un impacte mediàtic molt important (amb més de 25 impactes en premsa i ràdio) i va ser de molt interès per al Govern de Catalunya, ja que encara s'estava gestionant la crisi de la COVID i va permetre prendre mesures que van afavorir la qualitat de vida d'algunes persones en les onades següents de la pandèmia.

La metodologia proposada s'ha aplicat en altres casos d'ús reals amb resultats molt satisfactoris. D'una banda s'ha fet una consulta INSESS per a la valoració d'una associació sense ànim de lucre a Catalunya (Colla Castellera Jove de Barcelona), essent l'entitat no gaire gran, es van donar 15 dies per recollir les respostes a partir dels mòbils i en un parell de dies després de tancar la metodologia es van podre entregar resultats

També s'ha aplicat la proposta d'aquesta tesi a les entitats del tercer sector de tres ciutats pilot diferents (Mataró (Espanya), Prato (Itàlia) i Varazdin (Croàcia)) dins del projecte europeu SMP-COSME-2021-RESILIENCE 101074115-DIMCARE aconseguint caracteritzar l'actitud davant la transformació digital dels membres d'aquestes entitats, i que no es poden reportar els resultats per estar limitats per la RGPD.

En el marc del projecte AquiSTEAM de la UPC, s'ha dissenyat també una activitat inspiradora dedicada als nens i nenes de 9 anys d'escoles primàries que recullen dades sobre el consum elèctric de les seves cases durant 15 dies i tots els dispositius elèctrics utilitzats (TV, calefacció, etc.), i s'analitzen aquestes dades amb metodologia MIPRI2D de manera que els nens poden entendre els dispositius més consumidors, portar algunes recomanacions a casa, i despertar interès pels camps de la mineria de dades i la Intel.ligència Artficial de cara a la seva progressió formativa. Actualment, algunes altres aplicacions reals són en el camp del turisme

La síntesi del que hem abordat en aquesta tesi ens permet constatar que la intel·ligència artificial i la mineria de dades permeten obtenir coneixement de valor afegit sobre fenomens complexos, en conseqüència, les parts interessades, inclosos els responsables polítics, poden prendre millors decisions. La proposta de tesi que avui presentem escurça molt significativament el temps transcorregut entre la recollida de dades i l'obtenció de l'informe d'anàlisi de dades, relegant l'interval més gran de temps del procés, que és la recollida de dades al que trigui el participant en involucrar-se, participant, responent al qüestionari,.

La tesi deixa moltes línies de recerca que permetran continuitat per poder avançar cap a sistemes intel.ligents de diagnòstic integrats i on totes les peces desenvolupades en aquesta tesi es puguin integrar en un únic procés.

No obstant això, abans de descriure les línies de treball futur, volem fer palesa les especials condicions de desenvolupament d'aquesta tesi, que es va iniciar amb l'inici del projecte INSESS-COVID19, aprovat l'abril de 2020, en ple confinament, i que es va desenvolupar amb la directora de tesi, cadascu confinat a casa seva, sota tota la pressió

d'estar simultàniament desenvolupant materials on-line per poder impartir les classes del període de confinament de la universitat. Aquesta circumstància, no només va dificultar enormement l'arrencada del projecte, sino també l'obtenció d'informació i les interaccions amb els experts, que, tractant-se de serveis socials, estaven completament desbordats gestionant l'emergència i teníen poca capacitat, sino nul·la, de poder parar el dia a dia per a reunir-se amb nosaltres. No obstant, tothom va fer l'esforç de trobar moments i nosaltres vam reduir el que necessitavem al que era estrictament imprescindible. D'altra banda, el disseny del procés de recollida de dades originalment basat en tallers presencials que haguessin pogut accel.lerar enormement el procés, va saltar completament pels aires quan les ABSS van tancar, i les restriccions COVID no permetien ni reunions ni grans ni peties. Aquesta circumstància ens va forçar a replantejar els tallers i d'aquí en va sortir una metodologia molt interessant sobre com fer els tallers des de diferents visions i perspectives, que s'ha presentat a la tesi.

Per a treballs futurs, s'està millorant la interpretació automàtica dels grups resultants: d'una banda, s'introduiran expressions regulars per crear les descripcions textuals de les classes d'acord amb els resultats del TLP. D'altra banda, s'està començant a treballar en l'ordenació automàtica de les files i columnes del TLP introduint processos intel·ligents que poden tenir en compte la relació semàntica entre blocs temàtics i puguin generar una representació simbòlica del domini de perfils més beneficiosos a menys, d'acord amb la semàntica introduïda en el termòmetre.

A més, de cara millorar la fase I de MIPRI2D, seria interessant poder implementar una interfície que permeti a usuaris no especialistes en tecnologia poder construir el fitxer de metainformació assitits per una interfície proactiva, i gestionar tota la part tecnològica i enllaçar les dades que resulten de la consulta amb el generador automàtic d'informes.

Finalment, seria bo poder convertir MIPRI2D en una eina de suport habitual en l'àmbit de les polítiques públiques, permetent escalar la solució a un sistema de suport a les consultes periòdiques on usuaris no experts puguin llençar les consultes i obtenir els informes de forma autònoma.

Ampliar l'abast de l'eina a poder fer consultes periòdiques sobre un tema concret el qual es vulgui tenir monitoritzat al llarg del temps requereix desenvolupar mòduls nous que més enllà de generar informes periòdics actualitzant les dades a les noves consultes, puguin analitzar transversament l'evolució dels indicadors al llarg del temps i eventualment efectuar prediccions a mig i llarg termini.

Això també es pot integrar en una eina de "what if analysis" que completi un sistema intel·ligent de suport a la presa de decisions que integri les prediccions MIPRI2Dbasades en la repetició de consultes al llarg del temps amb les actuals prestacions de la tecnologia.

9. LLista de contribucions de la tesi

Aquesta tesi realitza múltiples contribucions que es llisten a continuació

Aquesta tesi realitza múltiples contribucions que es llisten a continuació

- **Metodologia MIPRI2D:** metodologia ràpida d'identificació de perfils explicables d'un domini (territorial o no) a partir de dades d'individus diana, orientada al suport a la presa de decisions complexes i estratègiques
- **Metodologia de consultes INSESS:** La metodologia que estructura la recollida de dades mitjançant el disseny d'un instrument, la població diana, els tallers de recollida de dades i tota la infrastructura tecnològica necessària per acompanyar el procés. Qüestionaris atemporals: robustos a estudiar situacions en un moment del temps concret, però amb participació dels respondents en altres periodes posteriors
- **Càlcul de l'error de mostreig:** Adaptació del mètodeutilitza IDESCAT i l'INE es proposa un nou mètode de càlcul.
- **Augmentació de la capacitat expressiva dels qüestionaris a partir de la introducció de nous tipus de variables complexes:**
    - Variable de quadrícula
    - Variables bàsiques temporals (Temporal Basic Variable TBV)
    - Variable de quadrícula multivaluada
    - Variable TQQ: Temporal Qualified Qualitative
- **Model de metainformació (MdM):** Model de representació del coneixement per a la metainformació de qualsevol qüestionari que es vulgui utilitzar en una consulta INSESS. S'ha formalitzat per tal que qualsevol que construeixi dades per una consulta pugui acompanyar-les d'aquesta definició. És la clau pel tractament automàtic de les dades
- **Mètode de Pre-Processament automàtic de les dades:** Conjunt de mètodes desenvolupats per a la realització automàtica del preprocessament de les dades.
- **Eines descriptives innovadores**:
    - Diagrama de teler,
    - Taules de transicions
    - Taula de freqüències ampliada
    - Taula de transició

- **Política de no-reidentificació:** Establiment d'una política per a la no reidentificació del ciutadà individual a partir dels resultats publicats sobre la descriptiva de les dades.
- **Informe de descriptiva automàtica de les dades:** Creació d'un procediment automàtic que construeix un informe finalista i maquetat amb els resultats de l'anàlisi descriptiva de totes les variables descrites mitjançant les eines proposades en aquesta tesi.
- **Formalització del Termòmetre** com a eina d'adquisió de coneixement semàntics de les variables que permet modelar formalment la polatirat de la interpretació del significat d'una variable
- **Quadre semàfor basat en termometres:** Generalització d'una eina d'interpretació automàtica de les classes prèvia que era el quadre semàfor, mitjatçant la introducció del termòmetre per a la construcció automàtica del quadre semàfor
- **Metodologia de derivació de noves variables** per facilitar l'obtenció de perfils territorialment consistents.
    - Variables derivades de 2a generació basades en coneixement de l'expert, recollint informació específica de domini proporcionda pels experts i combinant variables originals d'acord amb les indicacions dels experts
    - Variables derivades de 2a generació basades en dades (DD2dl), a partir de processos automàtics de clustering i interpretació automàtica de classes que sintetitzin un conjunt de variables candidates indicat pels experts
- **Variables derivades de 3a generació**, utilitzen la mateixa metodologia que les de 2a generació basades en dades, però admeten les variables de 2a generació com a variables candidates Mètode de selecció de variables discriminants per territori (TSFM): Identifica quina variable original, de 2ª o 3ª generació representarà un bloc temàtic en el clustering multivista final que formarà els perfils territorials i construeix els perfils i la seva interpretació conceptual i automàtica
- **Índex de potencial explicabilitat d'una variable** : Criteri que permet ordenar quina variable té més potencial explicatiu per una variable de classe donada i que TSFM utilitza internament
- **Primer informe sobre l'impacte del confinament en la vulnerabilitat social a Catalunya de la primera onada COVID:** Informe entregat a la Generalitat de Catalunya el 15 de desembre de 2020, resultat del projece INSESS-COVID19
-

10. Publicacions sorgides d'aquesta tesi

Al llarg d'aquest primer any de tesi s'han realitzat 3 publicacions diferents:

1.Informe oficial sobre els resultats de la primera consulta INSESS-COVID19 que implementa la metodologia proposada a la tesi doctoral. Informe que es va entregar AL GOVERN en el seu moment com a document de referència per establir polítiques socials arran de les vulnerabilitats emergents conseqüència de la COVID19 i que respon a la següent referència

GIBERT, Karina; ANGERRI, Xavier; Codina, Toni (2021). Informe INSESS-COVID19: Identificació de necessitats Socials Emergents com a conseqüència de la Covid19 i efecte sobre els Serveis Socials del territori. Publica Projecte INSESS-COVID19.
http://www-eio.upc.edu/~karina/INSESS/InformeINSESS-COVID19.pdf

- Informes derivats específics, que no són de caràcter públic: 8 informes específics per cada Vegueria, un informe per l'Ajuntament de Girona enfocat a persones grans; un informe específic per la FEDAIA de salut mental en adolescents; un informe de gènere per la setmana de la dona de la UPC 2021.
- Un article en revista d'impacte que avença en la formalització de la proposta metodològica de la tesi:

  GIBERT, Karina; ANGERRI, Xavier. The INSESS-COVID19 Project. Evaluating the impact of the COVID19 in social vulnerability while preserving privacy of participants from minority subpopulations. Applied Sciences, 2021, vol. 11, no 7, p. 3110.
  - *IF: 2.32; QII: Engineering, multidisciplinariety*
- Un segon paper publicat en revista d'impacte (JCR) que formalitza la creació de noves variables derivades:

  ANGERRI, X.; GIBERT, K. Preprocessing and Artificial Intelligence for Increasing Explainability in Mental Health. International Journal on Artificial Intelligence Tools, 2022. IF: 1.059; QIV: Artificial Intelligence
- Un tercer paper en revista d'impacte (JCR) que està sotmès des de fa unes setmanes

  Xavier Angerri, Karina Gibert: Dimensionality reduction with automatic interpretation of clusters. Mathematics 2023 (submitted)
  IF: 2.592; QI: Mathematics
- Abstract i presentació de congrés internacional.

  Angerri Torredeflot, Xavier; Gibert, Karina (2021) Evaluating the impact of the COVID19 in mental health in people with social vulnerability through INSESS-COVID19 technology. Book of abstracts

at International Scientific Symposium on Biometrics p. 20. Croatian Biometrics Association. 2021

- Una presentació en taula rodona a Barcelona:

  Xavier Angerri (2021) L'atenció a la infància en l'era postpandèmia: balanç, tendències i propostes de futur. 25è FORUM FEDAIA (Federació d'Entitats d'Atenció a la Infància i a l'Adolescència)

- Un segon abstract en congrés internacional:

  X. Angerri, K. Gibert: Variables Selection for improving clustering multiview processes, in Procs int'l Environmental Modelling and Software Society (iEMSs) 2022. Brussels. Publisher: iEMSSs

- Un tercer article de congrés internacional amb actes a IOSPress que s'està revisant actualment i rebra feedback en poques setmanes

  X. Angerri, K. Gibert: A General model for the metainformation of complex questionnaires for automatic preprocessing and reporting under INSESS methodology. In Proc. International Conference of the Catalan Association for Artificial Intelligence. In Frontiers in Artificial Intelligence and Applications, IOSPress, The Netherlands. (Submitted)

- Report de treball Document monogràfic que descriu el mètode usat per al càlcul de la grandària en funció del nivell de confiança i de la grandària mostral, així com la calculadora interactiva dissenyada a tal efecte i utilitzada per determinar la grandària mostral al projecte INSESS-COVID19

  Angerri X, Gibert K (2020): Determinació de la grandària mostral"

Referències

[ACCWG,2005] Australasian Creatinine Consensus Working Group. "Chronic kidney disease and automatic reporting of estimated glomerular filtration rate: a position statement." Clinical Biochemist Reviews 26.3 (2005): 81.

[Ahlemeyer-Stubbe & Agnes, 2021] Ahlemeyer-Stubbe, Andrea, and Agnes Müller. "The importance of domain knowledge for successful and robust predictive modelling." Applied Marketing Analytics 6.4 (2021): 344-352.

[Ahlemeyer-Stubbe & Agnes, 2022] Ahlemeyer-Stubbe, Andrea, and Agnes Müller. "Why domain knowledge is essential for data scientists in marketing." Applied Marketing Analytics 7.4 (2022): 362-373.

[Alcañiz & Planas, 2011] Alcañiz, M., & Planas, D. (2011). Disseny d'enquestes per a la investigació social. Barcelona: Universitat de Barcelona. Retrieved from http://hdl.handle.net/2445/18302

[Alemán, 1993] Alemán Bracho, M. D. C. (1993). Una perspectiva de los servicios sociales en España. Alternativas. Cuadernos de Trabajo Social, N. 2 (octubre 1993); pp. 195-205.

[Alison 2002] Alison, P. (2002). Missing Data. Sage Publications.

[Alonso, Castiello & Mencar, 2018] Alonso Jose M., Castiello Ciro, Mencar Corrado A bibliometric analysis of the explainable artificial intelligence research field International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer, Cádiz, Spain (2018), pp. 3-15

[Angerri & Gibert, 2023] ANGERRI, X.; GIBERT, K. Preprocessing and Artificial Intelligence for Increasing Explainability in Mental Health. International Journal on Artificial Intelligence Tools, 2023., 32.2 doi: 10.1142/S0218213023400110

[Angerri, 2015] Angerri, X. (2015). Aplicació de tècniques avançades de mineria de dades a la identificació de patrons d'agitació en malalts mentals greus (TFG. Grau en Estadística). UPC-UB.

[Avila Montalvo, 2018] Avila Montalvo, J. J. (2018). Interpretación automática de clases a partir de la extensión del cuadro termómetro a variables cualitativas a través de KLASS (TFM. Master en Enginyeria Informática). UPC

[Azur et al., 2011] Azur, M., Stuart, E., Frangakis, C., & Leaf, P. (2011). Multiple imputation by chained equations: what is it and how does it work? International Journal of Methods in Psychiatric Researh, 20(1), 40-49. doi:10.1002/mpr.329

[Barnett & Toby, 1994] Barnett, V., & Toby, L. (1994). Outliers in statistical data (Vol. 3). New York: Wiley.

[Batista & Monrad, 2002] Batista, G., & Monard, M. (2002). A Study of K-Nearest Neighbour as an Imputation Method. His, 85(251-260), 48.

[Battiti, 1994] BATTITI, Roberto. Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on neural networks, 1994, 5.4: 537-550. doi: 10.1109/72.298224

[Benzécri, 1973] Benzécri, J. P. (1973). L'analyse des données (Vol. 2, p. l). Paris: Dunod.

[Bickel & Scheffer, 2004] BICKEL, Steffen; SCHEFFER, Tobias. Multi-view clustering. In: ICDM. 2004. p. 19-26.

[Blair, Czaja & Blair, 2014] Blair, J., Czaja, R., & Blair, E. (2014). Designing surveys. SAGE Publications.

[Brooks & Perarce, 2000] Brooks, Jennifer, and Diana Pearce. "Meeting needs, measuring outcomes: The self-sufficiency standard as a tool for policy-making, evaluation, and client counseling." Clearinghouse Rev. 34 (2000): 34.

[Brotons, 1995] Brotons, X. (1995). Castells i castellers: Guia completa del món casteller. Barcelona: Lynx Edicions

[BTSS, 2017] Barometre del Tercer Sector Social del 2017. Available online: http://www.tercersector.cat/el-sector-catalunya (accessed on 15 February 2021).

[Burkart & Hubert, 2021] Burkart, Nadia, and Marco F. Huber. "A survey on the explainability of supervised machine learning." Journal of Artificial Intelligence Research 70 (2021): 245-317.

[CAIXA, 2020] Observatori Social de la Caixa (2020) Anàlisi de les necessitats socials de dones i homes

[Calinski & Harabasz, 1974] CALIŃSKI, Tadeusz; HARABASZ, Jerzy. A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 1974, 3.1: 1-27.

[Canudes, 2016] Canudes Solans, D. (2016). Apropant el datamining a l'expert a través del pre i postprovessament de resultats: l'ús del termòmetre en la construcció automàtica dels quadres semàfors de klass. TFM, MEI,UPC.

[CATALUNYA, 2023] CATALUNYA.COM. Castells [online]. [Accessed 26 April 2023]. Available from: https://www.catalunya.com/castells-1-3-12567?language=en

[Cervelló 2017] Cervelló, A. (2017). Els orígens del fet fasteller: Del Ball De Valencians as Xiquets de Valls (del segle XVIII al 1849). Valls: Cossetània

[CITE, 2020] Centre d'Informació per a Treballadors Estrangers. (2020). Condicions de vida de les treballadores de la llar i les cures centreamericanes a Barcelona

[Dai et al., 2022] DAI, Hao, et al. The State of the Art of Metadata Managements in Large-Scale Distributed File Systems—Scalability, Performance and Availability. IEEE Transactions on Parallel and Distributed Systems, 2022, 33.12: 3850-3869. doi: 10.1109/TPDS.2022.3170574

[Damgard, Pedersen & Pfitzmann, 1998] Damgard, I. B., Pedersen, T. P., & Pfitzmann, B. (1998). Statistical secrecy and multibit commitments. IEEE Transactions on Information Theory, 44(3), 1143-1151.

[Davis & Sampson, 1986] DAVIS, John C.; SAMPSON, Robert J. Statistics and data analysis in geology. New York: Wiley, 1986.

[DGSS,2019] La Direcció General de Serveis Socials desplega un nou instrument per detectar i prevenir situacions de necessitat social . (2019). Recuperado 8 de diciembre de 2020, de DIXIT Centre de Documentació de Serveis Socials website:

https://dixit.gencat.cat/es/detalls/Noticies/direccio_general_serveis_socials_desplega_instrument_detectar_prevenir_situacions_necessitat_social

[Di Meglio et al. 2013] Di Meglio, E.; Osier, G.; Berger, Y.G.; DiFalco, E. Standard error estimation for EU-SILC target indicators–first results of the Net-SILC2 project. 2013.

[DIMCARE, 2023] DIMCARE - digital missions for Care Social Economy's resilience. Yunus Social Business Centre University of Florence [online]. 25 January 2023. [Accessed 26 April 2023]. Available from: http://sbflorence.org/en/dimcare-social-economy/

[DIXIT CDSS, 2021] DIXIT Centre de Documentació de Serveis Socials. Available online: https://dixit.gencat.cat/ca/detalls/Noticies/tsf_presenta_eina_cribratge_ajudar_identificar_gestionar_casos_socials_complexos.html (accessed on 15 February 2021).

[DIXIT CDSS, 2022] El Departament presenta una eina de cribatge per ajudar a identificar i gestionar els casos socials més complexos. DIXIT Centre de Documentació de Serveis Socials, dixit.gencat.cat/ca/detalls/Noticies/tsf_presenta_eina_cribratge_ajudar_identificar_gestionar_casos_socials_complexos.html. Accessed on 6 april 2022.

[EPA, 2002] Encuesta de población activa. Metodología. (2002). INE, 2002

[EPA, 2005] Encuesta de Población Activa, Metodologia 2005. Descripcion general de la Encuesta. Available online: https://www.ine.es/inebaseDYN/epa30308/docs/resumetepa.pdf (accessed on 24 March 2021).

[EPA, 2021] EPA. Available online: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736176918&menu=ultiDatos&idp=1254735976595 (accessed on 15 February 2021).

[EPA, 2022] Encuesta de Población Activa. Diseño de la Encuesta y Evaluación de la calidad de los datos. Informe Técnico. Available online: https://www.ine.es/inebaseDYN/epa30308/docs/epa05_disenc.pdf (accessed on 15 February 2021).

[EUR-LEX, 2021] EUR-LEX Available online: https://eur-lex.europa.eu/eli/reg/2016/679/2016-05-04 (accessed on 15 February 2021).

[Explainable AI, 2021] Explainable AI, Royal Society Available online https://royalsociety.org/topics-policy/projects/explainable-ai/ (accessed on 14 March 2021).

[Farhangfar et al., 2008] Farhangfar, A., Kurgan, L., & Dy, J. (2008, December). Impact of imputation of missing values on classification error for discrete data. Pattern Recognition, 41(12), 3692-3705. doi:10.1016/j.patcog.2008.05.019

[Fayyad et al., 1996] Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 1996, 39.11, 27-34

[FCRI, 2023] Web d'fcri. 100tifiques.cat [online]. [Accessed 26 April 2023]. Available from: https://100tifiques.cat/

[Fleuret et al., 2004] FLEURET, François. Fast binary feature selection with conditional mutual information. Journal of Machine learning re-search, 2004, 5.9.

[Fonseca, 2021] Fonseca i Casas, Pau, et al. "Sars-cov-2 spread forecast dynamic model validation through digital twin approach, catalonia case study." Mathematics 9.14 (2021): 1660.

[Garcia et al. 2016] Garcia, S., Ramirez-Gallego, S., Luengo, J., Benitez, J., & Herrera, F. (2016). Big data preprocessing: methods and prospects. Big Data Analytics. doi:10.1186/s41044-016-0014-0

[Gibert, 1991] Gibert, Karina; (1991). Klass. estudi d'un sistema d'ajuda al tractamentestadıstic de grans bases de dades. Master's thesis, Master's thesis, UPC

[Gibert, 1994] Gibert Karina; (1994), L'ús de la informació simbòlica en l'automatització del tractament estadístic de dominis poc estructurats. (Direcció Ulises Cortes). Tesi doctoral. Universitat Politècnica de Catalunya

[Gibert, 1996] Gibert, Karina. "The use of symbolic information in automation of statistical treatment for ill-structured domains." AI Communications 9.1 (1996): 36-37.

[Gibert, 2013] Gibert, Karina; (2013). Mixed intelligent-multivariate missing imputation. (T. a. Francis, Ed.) International Journal of Computer Mathematics, 91(1), 85-96. doi:10.1080/00207160.2013.783209

[Gibert, 2014] Gibert Karina; 2014 Automatic generation of classes interpretation as a bridge between clustering and decision making International Journal of Multicriteria Decision Making 4(2):154-182 Inderscience

[Gibert & Angerri, 2021] GIBERT, Karina; ANGERRI, Xavier. The INSESS-COVID19 Project. Evaluating the impact of the COVID19 in social vulnerability while preserving privacy of participants from minority subpopulations. Applied Sciences, 2021, 11.7: 3110. doi: 10.3390/app11073110

[Gibert & Conti, 2015] GIBERT, Karina; CONTI, Dante. aTLP: A color-based model of uncertainty to evaluate the risk of decisions based on proto-types. AI Communications, 2015, 28.1: 113-126. doi: 10.3233/AIC-140611

[Gibert & Cortes, 1997] GIBERT, Karina; CORTÉS GARCÍA, Claudio Ulises. Weighting quantitative and qualitative variables in clustering methods. Mathware & soft computing. 1997 Vol. 4 Núm. 3, 1997.

[Gibert & Nonell, 2005] Gibert Karina; Nonell, R. 2005 Descriptive statistics with KLASS.Supporting LaTeX documents ellaborat-ion. In Procs 3rd World Conf on Computational Statistics and Data Analysis pp 90 Limassol (Cyprus)

[Gibert & Nonell, 2008] Gibert, Karina; R. Nonell 2008 Pre and post-processing in KLASS. Proc. of the iEMSs IVth Int'l Congress of Environmental Modeling and Software (DM-TES'08 Workshop), vol III: 1965-1966

[Gibert & Perez, 2006] Gibert, Karina; Pérez-Bonilla, A. (2006). Revised boxplot based discretization as the kernel of automatic interpretation of classes using numerical variables. In Data Science and Classification (pp. 229-237). Springer, Berlin, Heidelberg.

[Gibert 2020] GIBERT, Karina. Covid19. Web del Project INSESS-COVID19 [online]. [Accessed 27 April 2023]. Available from: https://insess-covid19.upc.edu/

[Gibert et al. 2008] GIBERT, Karina, et al. Response to TBI-neurorehabilitation through an AI& Stats hybrid KDD methodology. Medical Archives, 2008, 62.3: 132-135.

[Gibert et al. 2009] Gibert, Karina; García Rudolph, A.; Curcoll, L.; Soler, D.; Pla, L.; Tormos, J.M. Knowledge discovery about quality of life changes of spinal cord injury patients: clustering based on rules by states. Stud. Health Technol. Inform. 2009, 150, 579–583.

[Gibert et al. 2015] Gibert, Karina; Nonell, R., Velarde, J. M., and Colillas, M. M. Knowledge Discovery with clustering: impact of metrics and report-ing phase by using KLASS. Neural Network World, 2015, 15(4), 319-326

[Gibert et al. 2018] Gibert, Karina;, Izquierdo, J., Sànchez-Marrè, M., Hamilton, S. H., Rodríguez-Roda, I., & Holmes, G. (2018). Which method to use? An assessment of data mining methods in Environmental Data Science. Environmental modelling & software, 110, 3-27.

[Gibert et al., 2018] Gibert, Karina;, Horsburgh, J. S., Athanasiadis, I. N., & Holmes, G. Environmental data science. Environmental Modelling & Soft-ware, 2018, 106, 4-12

[Gibert, Codina & Angerri, 2020] GIBERT, Karina; CODINA, Toni; ANGERRI TORREDEFLOT, Xavier. Informe INSESS-COVID19: identificació de necessitats socials emergents com a conseqüència de la Covid19 i efecte sobre els serveis socials del territori. 2020.

[Gibert, Conti & Sànchez-Marrè, 2012] GIBERT, Karina; CONTI, Dante; SÀNCHEZ-MARRÈ, Miquel. Decreasing uncertainty when interpreting profiles through the traffic lights panel. In: Advances in Computational Intelligence: 14th International Conference on Information Processing and Man-agement of Uncertainty in Knowledge-Based Systems, IPMU 2012, Catania, Italy, July 9-13, 2012. Proceedings, Part II 14. Springer Berlin Heidelberg, 2012. p. 137-148. doi: 10.1007/978-3-642-31715-6_16

[Gibert, Conti & Vrecko, 2012] Gibert, Karina, Dante Conti, and Darko Vrecko. "Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants." Environmental Engineering and Management Journal 11.5 (2012): 931-944.

[Gibert, Garcia-Rudolph & Rodriguez-Silva, 2008] Gibert Karina; A. Garc-Rudolph, G. Rodriguez-Silva 2008: The role of KDD Support-Interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation. Acta Informatica Medica 16(4) 178-182

[Gibert, Rodriguez S & Rodriguez R, 2010] Gibert, Karina; Rodriguez Silva, G.; Rodriguez Roda, I. Knowledge discovery with clustering based on rules by states: A water treatment application. Environ. Model. Softw. 2010, 25, 712–723.

[Gibert, Sanchez-Marre & Izquierdo, 2019] GIBERT, Karina; SÀNCHEZ–MARRÈ, Miquel; IZQUIERDO, Joaquín. A survey on pre-processing techniques: Relevant issues in the context of environmental data mining. AI Communications, 2016, 29.6: 627-663. doi: 10.3233/AIC-160710

[Gibert, Sevilla-Villanueva & Sànchez-Marrè, 2016] GIBERT, Karina; SEVILLA-VILLANUEVA, Beatriz; SÀNCHEZ-MARRÈ, Miquel. The role of significance tests in consistent interpretation of nested partitions. Journal of computational and applied mathematics, 2016, 292: 623-633. doi:10.1016/j.cam.2015.01.031

[Gribova, 2016] GRIBOVA, V. V., et al. Implementation of a model of a metainformation-controlled editor of information units with a complex structure. Automatic Documentation and Mathematical Linguistics, 2016, 50: 14-25. doi: 10.3103/S0005105516010052

[Guidotti et al. 2018] Guidotti Riccardo, Monreale Anna, Ruggieri Salvatore, Turini Franco, Giannotti Fosca, Pedreschi Dino A survey of methods for explaining black box models ACM Comput. Surv. (CSUR), 51 (5) (2018), pp. 93:1-93:42, 10.1145/3236009

[Hartmann, 2017] Hartmann, Thomas, et al. "Model-driven analytics: Connecting data, domain knowledge, and learning." arXiv preprint arXiv:1704.01320 (2017).

[He, Cai & Niyogi, 2005] HE, Xiaofei; CAI, Deng; NIYOGI, Partha. Laplacian score for feature selection. Advances in neural information processing sys-tems, 2005, 18.

[Holzinger et al. 2019] Holzinger, Andreas, et al. "Causability and explainability of artificial intelligence in medicine." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9.4 (2019): e1312.

[INE, 2016] INE (2016)Encuesta de Población Activa. Diseño de la Encuesta y Evaluación de la calidad de los datos. Informe Técnico Recuperat de: https://www.ine.es/inebaseDYN/epa30308/docs/epa05_disenc.pdf

[INE, 2020] INE (2020) Encuesta de Población Activa Informe Técnico

[Jacob, Obozinski & Vert, 2009] JACOB, Laurent; OBOZINSKI, Guillaume; VERT, Jean-Philippe. Group lasso with overlap and graph lasso. In: Proceedings of the 26th annual international conference on machine learning. 2009. p. 433-440. doi: 10.1145/1553374.1553431

[Kaur, Mittal & Singh, 2022] KAUR, Navdeep; MITTAL, Ajay; SINGH, Gurprem. Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey. Multimedia Tools and Applications, 2022, 81.10: 13409-13439. Doi: doi.org/10.1007/s11042-021-11272-6

[Kim et al., 2023]Kim, J., Kim, Y., de Langis, K., Shin, J., & Kang, D. (2023). infoVerse: A Universal Framework for Dataset Characterization with Multidimensional Meta-information. arXiv preprint arXiv:2305.19344.

[Krejcie & Morgan, 1970] Krejcie, R., & Morgan, D. (1970). Determining Sample Size for Research Activities. Educational and Psychological Measurement, 30(3), 607-610. doi:10.1177/001316447003000308

[La Gatta et al., 2020] La Gatta, V.; Moscato, V.; Postiglione, M.; Sperli, G. An Epidemiological Neural network exploiting Dynamic Graph Struc-tured Data applied to the COVID-19 outbreak. IEEE Trans. Big Data 2020,7,45–55.

[Lauricks et al., 2012] Lauriks, S.; Buster, M.C.A.; de Wit, M.A.S.; van de Weerd, S.; Tigchelaar, G.; Fassaert, T. The Dutch version of the self-sufficiency matrix (SSM-D). 2012

[Lebart, Morineau & Fénelon, 1990] LEBART, L.; MORINEAU, A.; FÉNELON, J. P. Traitement statistique des données. Dunod, Paris, 1990, 34.

[Lefkovitch, 1980] LEFKOVITCH, Leonard P. Conditional clustering. Biometrics, 1980, 43-58. doi: 10.2307/2530494

[Lei, 2020] LEI, Yujiao, et al. Research of Automatic Generation for Engineering Geological Survey Reports Based on a Four-Dimensional Dynamic Template. ISPRS International Journal of Geo-Information, 2020, 9.9: 496.doi: 10.3390/ijgi9090496

[Li et al, 2017] LI, Jundong, et al. Feature selection: A data perspective. ACM computing surveys (CSUR), 2017, 50.6: 1-45. doi: 10.1145/3136625

[Little & Donald, 2019] Little, R., & Donald B, R. (2019). Statistical analysis with missing data (Vol. 793). John Wiley & Sons.

[Lopèz & Fachelli, 2015] López-Roldán, P., & Fachelli, S. (2015). El diseño de la muestra. In P. López-Roldán, & S. Fachelli, Metodología de la investigación social cuantitativa (p. 64). Bellaterra: Universitat Autònoma de Barcelona. Retrieved from https://ddd.uab.cat/pub/caplli/2017/185163/metinvsoccua_cap2-4a2017.pdf

[Luengo, Garcia & Herrera, 2012] Luengo, J., García, S., & Herrera, F. (2012). On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowledge and Information Systems, 32, 77-108. doi:10.1007/s10115-011-0424-2

[Maccioni et al 2018]Maccioni, A., & Torlone, R. (2018). KAYAK: a framework for just-in-time data preparation in a data lake. In Advanced Information Systems Engineering: 30th International Conference, CAiSE 2018, Tallinn, Estonia, June 11-15, 2018, Proceedings 30 (pp. 474-489). Springer International Publishing.

[Marra & Bogue, 2006] Marra, R. M., & Bogue, B. (2006). A critical assessment of online survey tools. Women in Engineering ProActive Network.

[Mataro, 2023] Mataró participa en el projecte europeu DIMCARE en l'àmbit de l'Economia Social i Solidària. (2022). Ajuntament de Mataró. https://www.mataro.cat/ca/actualitat/noticies/2022/mataro-participa-en-el-projecte-europeu-dimcare-en-l2019ambit-de-l2019economia-social-i-solidaria

[Mathew, 2005] MATHEW, Timothy H. Chronic kidney disease and automatic reporting of estimated glomerular filtration rate: a position statement. The Medical Journal of Australia, 2005, 183.3: 138-141. doi: 10.5694/j.1326-5377.2005.tb06958.x

[Messina, 2022] MESSINA, Pablo, et al. A survey on deep learning and explainability for automatic report generation from medical images. ACM Computing Surveys (CSUR), 2022, 54.10s: 1-40. doi: 10.1145/3522747

[Miller, 2019] Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.

[Milred, 2017] Mildred, P. (2017). Questionnaire Research A practical guide. New York: Routledge.

[Mishra et al. 2020] Mishra, P., Alessandra, B., Regar, J., Marini, F., & Rutledge, D. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. Trends in Analytical Chemistry, 132(116045). doi:10.1016/j.trac.2020.116045

[Moore et al. 1993] Moore, D; McGabe, G P,, ¬CraigB. A., Introduction to the Practice of Statistics; H.Freeman, New York, USA, 1993

[Nuñez & Sanchez-Marre, 2004] Núñez, H., & Sànchez-Marrè, M. (2004, August). Instance-based learning techniques of unsupervised feature weighting do not perform so badly!. In *ECAI* (Vol. 16, p. 102).

[Nuñez & Sanchez-Marre, 2005] Núñez, H., & Sànchez-Marrè, M. (2005, May). A Case-Based Methodology for Feature Weighting Algorithm Recommendation. In *CCIA* (pp. 223-230).

[Pearce, 1996] Pearce, D. (1996). The self-sufficiency standard. See: http://www. sixstrategies. org.

[Peng, Long & Ding, 2005] PENG, Hanchuan; LONG, Fuhui; DING, Chris. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence, 2005, 27.8: 1226-1238. doi: 10.1109/TPAMI.2005.159

[Perez, 2018] Pérez Tamayo, L. D. (2018). Tractamiento de variables multivaluadas en ciència de datos a través de KLASS. (Tesis de Master de Ingenieria Informática, FIB, UPC

[PESS, 2020] Generalitat de Catalunya. Departament de Treball, Afers Socials i Famílies (2020). Pla Estratègic de Serveis Socials 2021-2024

[Rahman & Islam, 2014] Rahman, M., & Islam, M. (2014). FIMUS: A framework for imputing missing values using co-appearance, correlation and similarity analysis. Knowledge-Based Systems, 56, 311-327. doi:10.1016/j.knosys.2013.12.005

[Robnik-Šikonja & Kononenko, 2003] ROBNIK-ŠIKONJA, Marko; KONONENKO, Igor. Theoretical and empirical analysis of ReliefF and RReliefF. Machine learn-ing, 2003, 53: 23-69. doi: 10.1023/A:1025667309714

[Royal Society, 2019] Royal Society. Explainable AI. Available online: https://royalsociety.org/topics-policy/projects/explainable-ai/ (accessed on 13 September 2021).

[Sevilla-Villanueva, Gibert & Sanchez-Marre, 2015] Sevilla-Villanueva, B., Gibert, K., & Sànchez-Marrè, M. (2015). Identifying nutritional patterns through integrative multiview clustering. Artif. Intell. Res. Dev, 277, 185.

[Sevilla-Villanueva, Gibert & Sanchez-Marre, 2017] SEVILLA-VILLANUEVA, Beatriz; GIBERT, Karina; SÀNCHEZ-MARRÈ, Miquel. A methodology to discover and under-stand complex patterns: Interpreted Integrative Multiview Clustering (I2MC). Pattern Recognition Letters, 2017, 93: 85-94. doi: 10.1016/j.patrec.2017.02.008

[Singh & Upadhyaya, 2012] Singh, K., & Upadhyaya, S. (2012). Outlier Detection: Applications And Techniques. International Journal of Computer Science Issues, 9(1), 307 - 323.

Solé Amenós, M. (2021). CASTELLS I TRASTORNS DE LA CONDUCTA ALIMENTÀRIA: Estudi de les característiques del fet casteller que poden influir en el desenvolupament d'un TCA i anàlisi de la incidència de TCA i comportament associat en castelleres joves. [Treball Final de Master]. Universitat de Girona.

[SSM, 2020] Generalitat de Catalunya. Departament de Treball, Afers Socials i Famílies (2020). La matriu d'autosuficiència SSM-CAT.

[SSS, 2002] The Self-Sufficiency Standard. Available online: https://depts.washington.edu/selfsuff/standard.html (accessed on 13 September 2022)

[Taherdoost, 2016] Taherdoost, H. (2016). Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. International Journal of Academic Research in Management, 5, 18-27. doi:10.2139/ssrn.3205035

[Tibshirani, 1996] TIBSHIRANI, Robert. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Method-ological), 1996, 58.1: 267-288. doi: 10.1111/j.2517-6161.1996.tb02080.x

[Tintarev & Masthoff, 2007] Tintarev Nava, Masthoff Judith A survey of explanations in recommender systems IEEE 23rd International Conference on Data Engineering Workshop, IEEE, Istanbul, Turkey (2007), pp. 801-810, 10.1109/icdew.2007.4401070

[Torres, Hernan & Janeth 2009] Torres, Patricia, Camilo Hernán Cruz, and Paola Janeth Patiño. "Índices de calidad de agua en fuentes superficiales utilizadas en la producción de agua para consumo humano: Una revisión crítica." Revista Ingenierías Universidad de Medellín 8.15 (2009): 79-94.

[UE, 2003] Union, E. (2023). Eurobarometer. Retrieved May 5, 2023, from https://europa.eu/eurobarometer/screen/home

[UPC STEAM,2023] https://aquisteam.upc.edu/ca

[Vellido, Martin-Guerrero & Lisboa 2012] Vellido Alfredo, Martín-Guerrero José David, Lisboa Paulo J.G. Making machine learning models interpretable European Symposium on Artificial Neural Networks, ESANN, vol. 12, i6doc, Bruges, Belgium (2012), pp. 163-172

[Vergara et al. 2016] Vergara, Camila, et al. "Learning on the relationships between respiratory desease and the use of traditional stoves in Bangladesh households." (2016).

[Vespignani, 2020] Vespignani, Alessandro, et al. "Modelling covid-19." Nature Reviews Physics 2.6 (2020): 279-281.

[Vila, 2004] Vilà Mancebo, A. (2004). Els serveis socials a CataluAnya. Una visió històrica. Universitat de Girona.

[Vilone & Luca, 2021] Vilone, Giulia, and Luca Longo. "Notions of explainability and evaluation approaches for explainable artificial intelligence." Information Fusion 76 (2021): 89-106.

[Ward, 1963] WARD JR, Joe H. Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 1963, 58.301: 236-244.

[Weber et al. 2018] Weber, M., Coldewey-Egbers, M., Fioletov, V., Frith, S., Jeannette D., W., Burrows, J., . . . Loyola, D. (2018). Total ozone trends from 1979 to 2016 derived from five merged observational datasets – the emergence into ozone recovery. Atmospheric Chemistry and Physics, 18(3), 2097–2117. doi:10.5194/acp-18-2097-2018

[Xu et al, 2015]Xu, D, & Tian, Y (2015). A comprehensive survey of clustering algorithms. Annals of Data Science, 2:165-193

[Yu & Liu, 2003] YU, Lei; LIU, Huan. Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of the 20th international conference on machine learning (ICML-03). 2003. p. 856-863.

[Yuan & Lin, 2006] YUAN, Ming; LIN, Yi. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2006, 68.1: 49-67. doi: 10.1111/j.1467-9868.2005.00532.x

[Zhao et al. 2017]Zhao, H., Du, L., Buntine, W., & Liu, G. (2017, November). MetaLDA: A topic model that efficiently incorporates meta information. In 2017 IEEE International Conference on Data Mining (ICDM) (pp. 635-644). IEEE.