International Journal on Artificial Intelligence Tools
Vol. XX, No. X (2022) 1–38

World Scientific
www.worldscientific.com

PREPROCESSSING AND ARTIFICIAL INTELLIGENCE FOR INCREASING EXPLAINABILITY IN MENTAL HEALTH

Xavier Angerri*

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya—BarcelonaTech, 08034 Barcelona, Spain

xavier.angerri@upc.edu

Karina Gibert

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya—BarcelonaTech, 08034 Barcelona, Spain

karina.gibert@upc.edu

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

This paper shows the added value of using the existing specific domain knowledge to generate new derivated variables to complement a target dataset and the benefits of including these new variables into further data analysis methods. The main contribution of the paper is to propose a methodology to generate these new variables as a part of preprocessing, under a double approach: creating 2nd generation knowledge-driven variables, catching the experts criteria used for reasoning on the field or 3rd generation data-driven indicators, these created by clustering original variables. And Data Mining and Artificial Intelligence techniques like Clustering or Traffic light Panels help to obtain successful results. Some results of the project INSESS-COVID19 are presented, Basic descriptive analysis gives simple results that eventhough they are useful to support basic policy-making, especially in health, a much richer global perspective is acquired after including derivated variables. When 2nd generation variables are available and can be introduced in the method for creating 3rd generation data, added value is obtained from both basic analysis and building new data-driven indicators.

Keywords: data science, intelligent decision support. Health, COVID19, Mental Health, Traffic Light Panels, Preprocessing, Explainable AI, Intelligent decision support

1. Introduction

In March 2020 the World Health Organization (WHO) declared the disease caused by SARS-COV2 as a Pandemics and a long lockdown started all over the world. Consequently, the Centre for Cooperation in Development (CCD) of the Universitat Politècnica de Catalunya opened a Special Call for projects with the aim of financing research that could help to advance in overcoming the COVID19 crisis. One of the projects receiving funds from this call was INSESS-COVID19 (see section 2), that used Data

* IDEAI. Nexus II building, groundfloor. Campus Nord. Universitat Politècnica de Catalunya, C. Jordi Girona 29, Barcelona 08034, Spain.

Mining and Artificial Intelligence techniques to quickly obtain new knowledge from citizens from a given target phenomenon.

The main aim of the INSESS-COVID19 project was to characterize the impact of the lockdown in the vulnerability of citizens. The outcomes were intended to help Catalan Government and BASS (Basic Areas of Social Services) to improve their policies and services, which means improving people welfare and health. The project provided a powerfull AI-based technology to quickly and automatically analyze data coming from a questionnaire and automatically generate the associated report.

In the current work our main goal is to go further and propose a data engineering methodology allowing to augment the information extracted from data as much as possible. The paper shows how introducing a preprocessing step to create new relevant derived variables that enlarge the original dataset increases the added value extracted from data in further analysis and allows a further understanding of the analysed target phenomenon. The creation of new derived variables is done from a data engineering perspective using two successive strategies, expert-based and data-driven. On the one hand, specific domain knowledge provided by experts is injected in the process and used to create what we call 2nd generation variables. On the other hand, data science techniques like clustering are used to created new data-driven variables. Moreover, when these new variables are introduced in further analysis under correct conditions, the final results are richer, provide relevant additional information from data and contribute to get a wider perspective of the target phenomenon.

In addition, the paper shows the benefit of adding the new 2nd generation variables (expert-based) to the original dataset and to consider them as well in the derivation of the 3rd generation variables (data-driven), this leading to a significant improvement in the final analysis.

These goals are directly related with knowledge representation issues and have the interest that new derived variables might catch nonlinear relationships of original ones, so enlarging the scope of the analysis and the possibility of getting additional added value from data.

Automatic interpretation oriented tools like Traffic Lights Panel (TLP) are proposed to elicit a conceptual interpretation of the 3rd generation data-driven variables, created with clustering techniques, so that the resulting class variable is converted into a new ordinary qualitative variable, with modalities representing labels (concepts). This facilitates the use of these new variables in further predictive models.

Finally, from the wide scope on vulnerability addressed in INSESS-COVID19 project, the paper focus on the specific field of Mental Health and shows under this use-case, how the

introduction of these knowledge-based and data-driven new indicators can be automatically processed and when used in further analysis they contribute to get a deeper understanding of the target phenomenon.
In synthesis, the main methodological contributions of the paper regards the introduction of a preprocessing step to create new derived variables into a target dataset. Two main mechanisms (knowledge-based or data- driven) are proposed here and tested on real data. The impact of using these new variables in further data analysis is also analyzed for the specific case of profiling problems using clustering techniques. The added value extracted from data when the new variables are considered is seen in the real application to Mental Health data coming from the INSESS-COVID19 database and additional knowledge obtained from data is analysed in detail.

The overall methodology proposed in the paper is a contribution to the field of Intelligent Decision Support Systems as well, since the additional knowledge extracted from the dataset including original, 2nd and 3rd generation data is deeper and useful to support more complex decisions compared with those supported with original data by itself. Also the results better approaches the usual reasoning of the expert, thus, increasing the understanding of the experts on the target phenomenon.

This is the structure of this paper. In section 2 the framework project is presented. In section 3 material and methods used is presented and the methodological proposal described in section 4. Section 5 shows the results of the application of the proposed methodology to the mental health data from INSESS-COVID19 database. Section 6 shows the conclusions and future work.

2. The framework project INSESS-COVID19

2.1. *The framework project INSESS-COVID19*

The INSESS-COVID19 project (Identification of Emerging Social Needs as a consequence of COVID19 and effect on the Social Services of the territory [1]), led by Professor Karina Gibert is one of the projects launched under the special call of the UPC for COVID projects at the beginning of the pandemics in Catalonia. By March 2020 most of the research was focusing on modelling the propagation of the pandemics like in [2], [3] or [4]. Nonetheless, at that time, an overflow of the Social Services was already expected to occur after the overflow observed on the Healthcare system. Given this intuition, there was an interest from the Social Services System to get soon information that could help them to understand the nature of the new needs of the vulnerable population raised by the pandemics and to be prepared to attend those needs ontime. The INSESS-COVID19 project is aligned with these goals and introduces the potential of Digital Transformation and Artificial Intelligence to provide elements to support decision-making and policy-making at the 107 Basic Areas of Social Services (BASS) of Catalonia and the Social Services Department from the Catalan

government. A prospective study to identify the social vulnerabilities of the Catalan population was performed.
The project started in April 2020, the data collection started in July 2020 and closed by December 6, 2020. The final results of project were described in a general report [5] published in the project web [1] and presented to the Catalan Government by December 15th 2020, only 9 days after data collection closed, thanks to the powerful AI technology developed not only to automatically clean data and perform the analysis, but including automatic reporting as well. The report was delivered to the Government by 2 general directorates (Social Services and Equality and Equity), and distributed to the 107 BASS (basic areas of social services) all over Catalonia.

The field of automatic reporting is more linked to the real practice than to academic and scientific publications. We can find some surveys on the added value of automatic reporting in the fields of medical image processing [6] [7], or in very specific and concrete applications [8][9]. But there are few foundational and formal references to these respect. Anyway, authors believe (and demonstrated in the INSESS-COVID19 project) that the time between the end of data collection and the data-driven decision-making is sensibly reduced when automatic reporting procedures are inserted in between the end of the data analytics process and the beginning of consumption of these results from the decision-makers side.

The knowledge presented in the report was obtained from a specifically designed instrument (the INSESS-COVID19 questionnaire, see section 2.2.) answered by 971 people belonging to 68 BASS. The participants called to answer this questionnaire were proposed by BASS and fulfilled one or more target subpopulations pre-defined by social services experts. Because the project aimed to discover the impact of the pandemics on vulnerable population, the inclusion and exclusion criteria to determine who would participate into the study focused on segments of population with different socio-economic vulnerabilities or risks of vulnerabilities which were defined and prioritized by a team of social services experts coming from the Catalan government and the local Social Services spread out on the Catalan territory. The study followed 20 target vulnerable subpopulations among which people with mental health disorders were considered as well. One of the target profiles regarded intellectual disabled persons who received assistance from Social Services to respond the questionnaire. The followed profiles were defined by the Social Services experts according to standard methodology and they are extensively described and justified in [10].

The INSESS-COVID19 report had information from all over Catalonia and helped the Catalan Government to make global decisions, soon in 2020, while the COVID19 was still enacting. Nevertheless, taking into account that each BASS has its own idiosyncrasy, additional local reports were sent to each one of the local BASS distributed around Catalonia. In the final report, basic descriptive information was provided. It was

automatically generated by using data science techniques and some Artificial Intelligence so that a final report was automatically build in a final layout, ready to provide support to executive meetings. Details on the technology specifically developed in the project to process INSESS-COVID19 database are described in [10].

The information provided by the INSESS-COVID19 was quickly transferred to Social Services authorities and provided results that allowed the Social Services System to modify some protocols to attend the gender violence victims during 2nd and subsequent lockdowns, arised warnings about the need of attending the mental health of the people and public call centers attended by psychologist were put into place and so on. However, the analysis was done on basic original data and regarding future follow-ups of the Social Services System, more sophisticated analysis that finds more complex patterns an needs of the population are suitable. This paper elaborates on the previous data engineering science that can be introduced for a more powerful preprocessing of the original data that can explore a wider perspective from data and extract more useful knowledge.

The INSESS-COVID19 project was a finalist of the European Social Services Awards 2021.

2.2. *The questionnaire INSESS-COVID19*

The questionnaire INSESS COVID19 was available to the selected participants through the web of the project under authentication protocols and constituted the mean to collect data directly from citizens, and to get them available for analysis in real time. Being the COVID19 a disrupting process, no available questionnaires existed to obtain the required relevant information to improve social services for new emergent vulnerabilities provoked by COVID19 crisis.

For this reason a new questionnaire had to be designed, specifically oriented to evaluate the impact of lockdown and COVID19 into the vulnerability of the citizens. To build the questionnaire, a state of the art was done, by identifying other related surveys launched by other institutions (see [10] for the state of the art). The INSESS-COVID19 questionnaire was specifically designed by the occasion by adapting the Self Sufficiency Matrix (SSM)(presented and used in [12],[13],[14])] a standardized and validated questionnaire devoted to evaluate vulnerability, well-known internationally and used in the Catalan Social Services System (SSM.cat) [15]. The SSM.cat is at the kernel of the digital transformation of Social Services described in the Catalan Strategy for Social Services[16] and represents a concept of vulnerability very much aligned with the current structure of primary care Social Services in Catalonia.

The adaptation of SSM.cat into the INSESS-COVID19 questionnaire was developed by the Social Services experts of the Catalan government and the local Social Services in Catalonia participating in the project, together with the knowledge engineers. The resulting

questionnaire was further reviewed by the Advisory Board of the project and a social service expert, and the texts of some questions were improved and disambiguated. The questionnaire was validated in a pilot with real citizens from two BASS early in July 2020. Minor changes were implemented as a result of the pilots and the survey was launched by mid July 2020. The INSESS-COVID19 questionnaire [10] had 195 questions distributed in 20 different blocks, with different number of variables each, related with basic social aspects in life, like participation in the society, Digital Gap, labour situation and health. Specially, health was targeted in one of the principal blocks, providing information on Mental Health (with special attention on addictions), health situation related with COVID-19 and information related with disability and dependency. The questionnaire contains 15 questions regarding health aspects, representing 21 columns in the dataset. The entire questionnaire and the domain values can be found in [1]. The type of data obtained from the questionnaire includes complex innovative types of advanced data which are modelled and described in detail in [10].

3. Antecedents

In this section the required additional information to provide the complete context of the proposal is provided

3.1. *Descriptive methodology*

The work [10] establishes the methodology to be used for analyzing data coming from the INSESS-COVID19 instrument. In this paper the descriptive analysis methodology used is the one proposed in [10]. The descriptive techniques used in the report [5] included Extended 5-Number Summary, Extended Frequency Table, Marginal Bar Plot, Multivalued Frequency Table, Trajectory Graphs, Trajectory Frequency Tables, Grid of Pies, Multiple Stacked Bar Plot [10] and Traffic Lights Panels [17], some of them being advanced descriptive statistics techniques, that produced relevant new knowledge about the pandemics and 1st wave in Catalonia.

3.2. *Clustering process*

Clustering is applied when the interest is to discover groups among data [18]. Data is grouped into classes or clusters, so that data from the same cluster is quite similar and different from those in other clusters. When clusters are found, characteristic patterns interrelating certain attributes describe the discovered clusters.

There are many families of clustering techniques, all of them distance-based methods, so that distances between two objects are used to build the groups under a variety of criteria that define the different algorithms. There are a large number of distance measurements and clustering algorithms. In this project Gibert Mixed Metric is used [19] since the INSESS-COVID19 instrument combines numerical and qualitative data together. Also, the Ward method will be used for the clustering, since it is based on the relationship between

the within clusters inertia and between clusters inertia and being inertias associated with the quantity of information [11] the Ward's method provides clusters with easier conceptual interpretation than other methods.

3.3. *Profiling classes methodology*

3.3.1. *Class Panel Graph (CPG)*

A CPG is a graphical representation of the conditional distributions of variables versus a set of clusters in a single panel, proposed in [17]. The clusters can come from a previous clustering algorithm. The details on applying CPG for automatic interpretation of clusters can be found in [20], directly linked with the explainable AI field [21] [22]. Many authors provide surveys giving a wide perspective of XAI from different approaches, but most of them refer to the supervised machine learning field [23] [24][25][26] or specific applications like recommender systems[27]. In [28] the importance of visualization for explainability is highlighted and in [29] the need to go further, analyzing causability as well, specially in the medical field, is addressed. In our work, we go in depth on the use of visual tools to help to the automatic interpretation of clusters, the result of a non-supervised machine learning family of methods, in order to facilitate the concepts induction associate to the clusters, so contributing to explain what clusters mean, represent, and how they were formed.

The CPG visualizes the joint behaviour of many variables with respect to the classes together in a compact way and provides to the expert a quick understanding of the particularities of each class, facilitating the process of class labelling, i. e., identifying the concept represented by the class, the *interpretation*. The CPG is in fact facilitating the inductive learning to the experts and helps to conceptualize the clusters. However, this resource resulted to be still complex for those stakeholders without technical skills and it was later evolving to TLP to bridge the gap for non-technical stakeholder.

3.3.2. *Traffic Light Panel (TLP)*

The TLP technique as a method for interpreting classes [20] was introduced by Dr. Karina Gibert in as a symbolic abstraction of the CPG [17] which has been extensively proved as a useful interpretation-oriented tool in previous works [17] [20] [30].
The TLP is based on identifying the dominant levels of each variable in each class. To make a TLP, the analyst must carefully read the CPG, mark the central trend for each class (High, Central or Low values) of each variable, and assign to the qualitative levels of the variables the colours of the traffic light, in accordance with the interpretative codes of the expert. The context and meaning of each colour must be related to some latent domain concept that allows the association between the variable polarity and the idea of improvement or worsening. In [20] propose two basic ways to assign colours to the scale of the variable.

(1) Direct colour coding (red-yellow-green) associated with low-medium-high values.
(2) Inverse colour coding (green-yellow-red) associated with low-medium-high values.

A very important property is that when the classes are well constructed, they must be distinguishable and must represent different profiles. Therefore, there should not be two rows of TLPs with the same colour scheme. TLP can be projected over a CPG by transferring the colours of the cells of the TLP to the cells of the CPG.
The TLP culminates the journey from data collection to new knowledge on the target phenomenon, providing a more symbolic support to the interpretation of clusters, with the main focus on areas such as statistics, data mining, and more recently data science.

3.3.3. *Annotated Traffic Lights Panel*

In [30] a further improvement of the TLP is presented, namely *annotated Traffic Light Panel (aTLP)*, which serve to manage the intrinsic uncertainty associated to the interpretation of the prototypes resulting from a clustering. aTLPs associate nominative traffic light colours (those determined in the TLP) with two colour dimensions (hue and saturation) that are used to represent the central trend and purity of the cells of the TLP. The purity is based on a Generalized Variation Coefficient (valid for numerical and qualitative variables) and a model of uncertainty developed in [11]. In this sense, pure colours represent little or no variability in the central trend of the variables in a certain class, while darker colours represent an increase in heterogeneity and consequently a loss of reliability in decisions based on the darker cells. In [11] a colour-based model for the automatic calculation of the saturation of each colour and an algorithm for the association of a hue to each cell in the aTLP are presented. The darkness of the colours is associated to higher variance. As with TLP, the aTLP can be projected over a CPG.

4. Methodological proposal

In [31] a complete preprocessing methodology with several steps is presented. This paper contributes to the step related to the creation of new variables and its effect on increasing the knowledge extracted from data when these new variables are included in the further analytics processes. In addition, the contribution of this paper is strongly related with the main idea developed by Fayyad in [32], who is pointing the importance of using the best data mining techniques to extract as much knowledge as possible from data.

4.1. *The added value of building derived variables*

Data-driven models are normally build with the variables originally obtained from the data set, sometimes arranged during the data cleaning process on basic transformations.

However, when new relevant variables can be defined as combination of original variables, added value appears on further data mining processes. The new variables could be generated by different methods. In this paper, 2 mechanisms are proposed:

- Knowledge-based creation of new 2nd generation variables: they add new concepts to the data base approaching the experts reasoning parameters
- Data-driven creation of 3rd generation new indicators: they synthesize blocks of thematic variables in single indicators by using data-driven modelling techniques.

Thus, new dimensions (usually corresponding to nonlinear combinations of original variables) could be interpreted by the expert using its conceptual universe. Once, several variables are created, this process ends and the further modelling process provides results easier to interpret and providing a deeper perspective on the analyzed data, as it will be seen in the results section (5).

4.2. *Creation of new knowledge-based 2nd generation variables*

The idea is to use the specific domain knowledge provided by experts to build new variables as a combination (often nonlinear) of the original variables available in the dataset. The knowledge-based new variables are approaching the reasoning made by the expert, and as a consequence, the further modelling including these variables results in easier interpretation by the expert since the models are expressed in terms he/she uses in his natural reasoning. The idea of creating new knowledge-based variables, materializing the know-how of the expert was originally presented in [31]. Nevertheless, techniques described in [31] are used in several works like [33] [34]. The idea that data alone is not enough to disclose complexity of real phenomena is known and several authors gathered attention to this issue, in [35] the proposal is to use specific domain knowledge to combine various extrapolation and prediction methods to understand the collected data for many application fields like health or finances. In [37] authors claim the importance of using domain knowledge all along the analytics process, from the very early stage of what questions to ask, to each of the steps of the preprocessing to build more precise and robust predictive models and thus obtain better insights. However in the preprocessing block it does not tackle the important step of creating new variables, what we address in this work. In [38] it is also stated that domain knowledge is essential in data science processes. Although the paper focus on marketing applications in this paper we explore the importance of a domain specific knowledge-based guidance to create new variables before the analysis from a transversal perspective non specific for a particular application field.

Indeed, often experts think in terms of complex transformations of original data into some new variables representing concepts they use to reason or to evaluate a certain scenario (for example evaluating if a person debuted on a mental health disorder based on an integral evaluation of a set of binary variables indicating if the person suffers or not from a particular disorder before the lockdown and after it, or reasoning over the total number of mental disorders suffered by a person instead of reasoning over the list of binary variables indicating the existence or not of each separated mental disorder). This new generation of variables consumes the knowledge that experts have about data, and can create many

different variables referring new concepts not present in original data, useful for further analysis.

In this paper, a reduced set of mechanisms to create this kind of 2nd generation variables is considered, all of them based on using the specific domain knowledge from the experts to combining several original variables in specific ways indicated by them with the aim of representing more complex concepts that can improve further data mining models. We consider here three mechanisms:

a. *Indicators*: A variable X with modalities $D_x=\{m_1, m_2, m_3, \ldots, m_X\}$is converted into a set of binary variables X'_m $\forall m \epsilon D_x$, such that $D_{xm}=D$, with $D=\{m'_1=YES, m'_2=NO\}$.

   The transformation rule being: if $X=m$, $\forall m \epsilon D_x$ then $X'_m=YES$; otherwise $X'_m=NO$.

b. *Multivariable conditions*: Given a set of variables $X_1 \ldots X_K$ with m_X modalities in D_x a new variable is created based on the evaluation of a Boolean function f built over the original variables.

   The transformation generates a new variable X' according to the following transformation rule: if $f(X_{1\ldots..} X_K)$ = TRUE then X'=YES; else X'=No.

c. *Counts:* When a pack of variables refer the same concept (i. e. several symptoms of a disease or several mental disorders), a new aggregated variable can be created for example by counting the number of positive values in the entire pack. Given a set of binary variables $(X_{1\ldots..} X_K)$ variables with the same modalities, $D_X=D$, with $D=\{m_1, m_2, m_3, \ldots, m\}$, a new numerical variable X' is created as a count of the variables in $(X_{1\ldots..} X_K)$ pointing to a certain subset of reference values $A \subset D$.

   The transformation rule is $X' = card\{X_m : X_m \epsilon A\}_{m \epsilon \{1:K\}}$.

   This are only three basic mechanisms to be used for generation of new expert-based derived variable, but there are many more that can be used in real projects.

4.3. *Creation of new data-driven 3rd generation new indicators*

In this case, new variables are built by using auxiliary multivariate analysis techniques. This is usefully to synthesize results of a subset of variables in a new single variable. These variables could be obtained using different data-driven models like PCA, clustering, etc. In this paper, clustering is used. The variables used for these models should correspond to a same theme (or to a same block of a questionnaire) and the new variables synthesize the relevant information of the entire pack in a single (or a reduced number of) new variables.

In our case we are using clustering to synthesize the information of a certain block of variables into a single one because the resulting clusters identify profiles (in terms of patterns of a certain combination of values for a certain subset of variables followed by the individuals belonging to the profile) that can be associated to individuals. Thus, the

resulting class variable matches with the structure required for a new column of a dataset. In [36] criteria to choose the proper data mining method to solve a certain problem are provided and this work was our reference to choose the use of clustering here. Indeed the resulting class variable is a new qualitative variable, namely a new indicator, that has the values properly labelled and can be easily added to the original dataset. Transforming the class variable into a new qualitative variable of the dataset means that the interpretation of the classes and further labelling is critical to get a meaningful D set of variable modalities. The proposed process is:

1. Determine a target topic, among the themes represented by the original dataset (eventually, the enlarged dataset including the 2nd generation variables can be used as well, given that often will represent nonlinear combinations of original variables and clustering do not requires independency)
2. Variable selection: Select a subset of variables regarding the target topic. They will become the *components* of the new indicator (i.e. a block of health variables, or a block of economic variables)
3. Clustering of topic variables: Using Ward's Method with Gibert Mixed Metrics selected variables are clustered and a new class variable results. In its original form (where classes received a numerical identifier), the variable is not interpretable by itself and a post-processing is required to interpret the clusters, achieving representative labels for each cluster, so that the class variable becomes a new qualitative variable with meaningful modalities. This is addressed in the following steps.
4. Creating CPG: Build the CPG from all the components of the new indicator identified in step 2 versus the new class variable obtained in step 3.
5. Creating TLP: Built the TLP (o the aTLP) from the same structure of the CPG in previous step. This abstracts the central trend of each component in each cluster by using a colour code as indicated in section 3.3.2 or 3.3.3.
6. Label modalities of new indicator: With the projection of the TLP/aTLP over the CPG, the particularities of the different variables in each class emerge, so the experts can induce proper labels to all the clusters, according to their main characteristics and summarizing the main concept behind each class. There is a bijective relationship between the class variable from step 3 and the set of labels created here. Thanks to the TLP, the class variable becomes an interpretable new qualitative variable with modalities having semantical meaning.
7. Label the new qualitative variable: Associate a label to the variable (often the name of the topic) and a description of the variable itself and each of its modalities in order to fix interpretation.
8. Adding the data-driven indicator to general database. The new 3rd generation qualitative variable becomes a new column of the dataset indicating a labelled cluster for each individual.

4.4. *Validation methodology*

From the methodological point of view, we aimed to compare the nature of the knowledge extracted when data is consumed in further data science models in the two scenarios of using only original variables, or including 2nd and 3rd generation variables as well.

To this purpose two groups of experts have been collaborating in the research:

- The first group was exposed to the analysis of original variables and the kind of knowledge on mental health provided by a global clustering of the original data and the further interpretation of resulting clusters into meaningful profiles.
- The second group worked with the knowledge engineers to build 2nd and 3rd generation variables. Then, the results of the clustering of the enriched dataset including the new created variables where shown and they participated to the interpretation of the resulting profiles.

The nature of the new knowledge obtained in the two scenarios was then put in common and discussed with all the experts.

5. Case Study and Results

Being this work oriented to health and specifically mental health, an analysis of INSESS-COVID19 database focused on this topic is done. From the original INSESS-COVID19 questionnaire (presented in section 2.1), a total of 15 questions spread out among several blocks, regard the impact of the first wave or COVID19 in health and mental health. The involved blocks are the following:

- ***Block VIII: Dependency*** (contains 3 questions about dependency)
- ***Block XIV: Teleworking and Teletraining*** (contains 1 questions to identify if the teleworking or teletraining impacted mental health)
- ***Block XVI: Health COVID19*** (2 questions to see if the person had COVID19)
- ***Block XVII: COVID19 (***2 specific questions for those that passed the disease)
- ***Block XVIII: Health*** (5 questions on mental health, abuse of substances and disability).
- ***Block XIX: Disability evolution*** (2questions for impact of disease on disability)

Among this questions, 2 of them are of type TQQ (Temporal Qualified Variable, see [10]), so that each question generates 4 variables. This means that from 15 questions, 21 variables are derived into the working database. Briefly, a TQQ Variable is in fact a triplet (X, T, Q) where X is a qualitative variable replicated T times and Q is a Likert variable to qualify the modalities of X at each timestamp from T. In this case, it inspects a target issue in the baseline (January 2020), generating one variable, and after the first wave (in July 2020), but generating three additional variables to indicate if the person feels better in July 2020 with regards to January or not, if he/she feels the same as in January or he/she feels worse.

5.1. *Health related variables in INSESS-COVID19 questionnaire*

The original INSESS-COVID19 variables regarding health are in Table 1. The 4th column indicates the type of the variable according to the typology established in [10].

Table 1. Variables in questionnaire

Block	Question code	Variable	Modalities	Type of variable
B8	D1	Do you have any dependency degree?	1. No; 2. Grade I (moderate dependence); 3. Grade II (severe dependence); 4. Grade III (high dependency); 5. No Answer	Ordinal
B8	D2	Do you think that if they valued you now you would have a variation in the degree of dependence?	1. I have improved; 2. The same; 3. I got worse; 4. No Answer	Ordinal
B8	D3	Do you attribute this variation to COVID-19?	1. Yes; 2. No; 3. I have not varied; 4. No Answer	Nominal
B16	SC1	Are you a member of any COVID-sensitive risk group19?	1. Yes; 2. No; 3. No Answer	Nominal
B16	SC2	Have you passed COVID-19?	1. Yes, I have been to the ICU; 2. Yes, I was admitted but not to the ICU; 3. Yes, diagnosed with symptoms and telephone medical care at home; 4. Yes, at home and with telephone medical care; 5. I have had symptoms but it is not known; 6. Yes, diagnosed but asymptomatic; 7. I had no trouble; 8. No answer	Ordinal
B17	Cov1	When did they detect you?	1. Between 1 and 15 March 2020; 2. Between 16 and 31 March 2020; 3. Between 1 and 15 April 2020; 4. Between 16 and 30 April 2020; 5. Between 1 and 15 May 2020; 6. Between 16 and 31 May 2020; 7. Between 1 and 15 June 2020; 8. Between 16 and 30 June 2020; 9. Between 1 and 15 July 2020; 10. Between 16 and 30 July 2020; 11. Between 1 and 15 August 2020; 12. Between 16 and 31 August 2020; 99. No answer	Temporal
B17	Cov2	Do your health conditions after passing COVID-19 allow you to resume your usual activity?	1. Yes, no problem; 2. With difficulty; 3. I won't be able to do it until September; 4. I will not be able to do this until January 2021; 5. With sequelae that prevent the recovery of normal activity; 6. No answer	Ordinal
B18	S9	Do you have a recognized disability?	1. Physics; 2. Sensory; 3. Intellectual; 4None 5. No answer	Nominal
B19	ED1	Has your degree of disability worsened between January 2020 and July 2020?	1. Yes, it has improved; 2. It has not changed; 3. Yes, it has gotten worse; 4. No answer	Ordinal
B19	ED2	If you feel worse, do you attribute this to the situation generated by COVID19?	1. Yes; 2. No; 3. I haven't gotten worse; 4. No Answer	Nominal
B14	TT2.TT.TFsuportG20	Did teleworking / teletraining require emotional support? *Until the moment, Jan2021*	1.Professional Network; 2.PersonalNetwork; 3.No; 4.No Answer	Temporal basic variable
B18	S3	Did the COVID-19 situation require emotional support?	1.Professional Network; 2.PersonalNetwork; 3.No; 4.No Answer	Temporal basic variable
B18	S4	Do you have a mental health problem diagnosed and how did you evolve during the COVID crisis19?*Jan 2020 and July 2020 I feel better, same or worse*	1. Severe Mental Disorder; 2. Border line personality disorder; 3. Post-traumatic stress disorder; 4. Anxiety 5. Depression; 6. Other; 7. None	Temporal Qualified Qualitative
B18	S5	Are you getting medication?	1. Yes; 2. No	Binary

B18	S6//S7//S8	Indicate your consumption of the substances or behaviors indicated: Tobacco; Drugs; Alcohol; Addictive behaviors (gambling addiction, cyber addictions). *in January 2020, July 2020 and January 2021*	1. No; 2. Use; 3. Abuse; 4. Addiction	Temporal Qualified Qualitative

These variables produce some direct information related with health and mental health. With the INSESS-COVID19 analytics technology developed in [10] first generation results are obtained, but more complex analysis can be done when these original variables are combined into new derivate variables either knowledge-based or data-driven, according to the proposed methodology.

5.2. *Focusing on mental health*

In this section, the improvement in the understanding of the health situation of the respondents is shown in a progressive sequence where:

(1) Only original variables are used for basic statistics
(2) Knowledge-based 2nd generation variables are introduced and basically described
(3) The data-driven 3rd generation indicators are introduced based only on original variables
The 3rd generation indicators are build including original and second generation variables together

5.3. *Using only original variables*

From the variables regarding health identified in INSESS-COVID19 project, we will focus on mental health topic. In Table 1 there are 4 variables regarding mental health (coded with TT2.TT.TF, S3, S4, S5). From them, 2 are temporal basic variables, 1 is temporal qualified variable (TQQ, S4) and 1 nominal. These are specific types of variables described in [10] for the first time. Among the more complex types is the TQQ variable (see above).

The question S4: *Do you have a mental health problem diagnosed and how did you evolve during the COVID19 crisis?* is a TQQ, which is one of the complex types of variables introduced in [10], where X is a qualitative variable for diagnosed mental disorders with S=8 modalities in D_{S4} = {Severe mental disorder (SMD), Personality limit disorder (PLD), Post-traumatic Stress Disorder (PSTD), Depression, Anxiety, Others, Nothing}.

X generates a first qualitative baseline vector X_{t1}= Do you have a mental health problem diagnosed in Jan 2020

Q is a Likert set of values with the common set of possible values that each mental disorder can take in each temporal timestamp. For S4 the Q has 3 possible values m_s, Q={Better, Equal, Worse}. So that for each mental disorder we can see if the person improves or worsens his mental health in between two consecutive timestamps.

Apart from X, S4 generates other 3 qualitative variables representing in which mental disorders the person feels better than in X by July 2020, in which he/she feels same, in which she/he feels worse.

S4 has a multivalued nature since a single person can suffer simultaneously from several mental disorders. So that the values of D_{S4} are not disjunct among them and an individual can simultaneously improve his anxiety and depression in July 2020 and be worse on Post-Traumatic Stress and Severe Mental Disorder. This confers a complex structure to the entire variable. On top of that, since Q is qualifying evolution, the first timestamp needs a different codification to indicate a baseline situation of having or not the mental disorder in January 2020.

As it is known, multivalued questions provide a lot of information, but sometimes they are too complex to use in the further data modelling.

However, taking advantage of the expertise of the stakeholders, we can build some interesting indicators on top of this original complex structure that helps in the analysis and the understanding of the target phenomenon.

The implemented digital questionnaire produced a structure of 4 variables to represent S4 (see Table 2) where the code of the question, the text of the question are concatenated with the "modality" represented in each column. In Table 2 the cells are multivalued, since in each cell several mental disorders might appear simultaneously:

ID	**S4. Do you have a mental health problem diagnosed and how did you evolve during the COVID crisis19? [1. January 2020]**	**S4. Do you have a mental health problem diagnosed and how did you evolve during the COVID crisis19? [2. July 2020 I'm better]**	**S4. Do you have a mental health problem diagnosed and how did you evolve during the COVID crisis19? [3. July 2020 I'm the same]**	**S4. Do you have a mental health problem diagnosed and how did you evolve during the COVID crisis19? [4. July 2020 I'm worse]**	
ID1	Depression	None	Depression	None	…
ID2	None	None	None	Post-Traumatic Stress	…
ID3	Post-traumatic Stress Disorder;Depression;Anxiety	Anxiety	Post-traumatic Stress Disorder	Depression	…
:	:	:	:	:	:

Table. 2. S4 database representation example

This information shows which mental disorders improved or got worse during the first wave of lockdown.

And the basic automatic descriptive analysis of these variables is shown in Figure 1, consisting of basic Paretto diagrams for the baseline situation of the persons, and the improvements or worsening by July 2019:

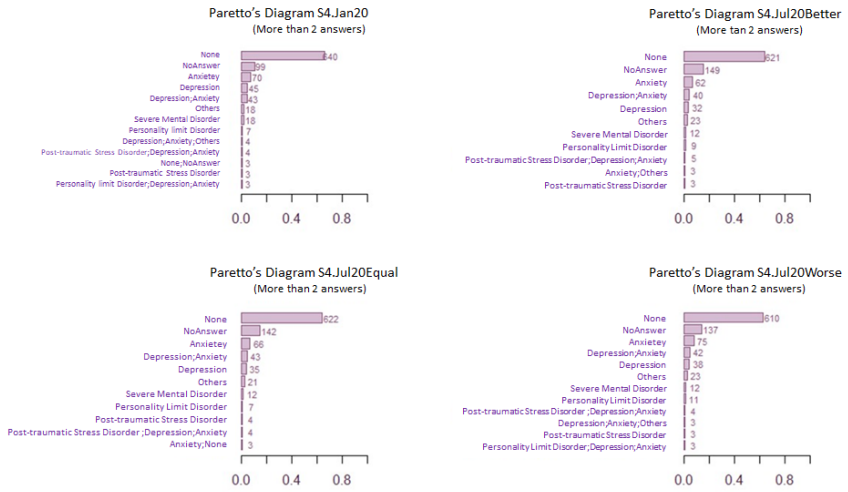

Fig. 1. Basic descriptive analysis of original data (origin of graphs, Klass software)

In [10] the techniques used to describe each type of variable are described. From the basic description of these variables it can be seen that:

- Most of the people report no mental health disorder
- 23.58% of participants reported some mental health problem in January 2020
- The most prevalent observed psychiatric profile is suffering from Anxiety, followed by Depression in January 2020. Then, the third group suffered from a combination of Anxiety and Depression simultaneously. After the first wave having these two becomes more prevalent than Depression alone
- There is people suffering from many mental disorders together but they are few people in global

Analysing the other variables from mental health we could see that

- 38% of participants required emotional support until July 2020, but their prevision is that the emotional support decreases for January 2021 to 33,47%.
- 22% of participants are getting medication.
- 54% with Intellectual Disability declare to feel worse by July 2020.

Mental health was among the most relevant impacts produced by COVID19 crisis affecting all kind of people.

It is clear that with the mental health original variables described above we cannot answer interesting questions directly like how many people debuted with a mental health disorder during the pandemics, or how many people improved his mental health or got worse due to pandemics or how the mental disorder evolved. Creating derived variables can help in this matter.

5.4. *Adding knowledge-based 2nd generation derived variables into the analysis*

The main idea is to introduce new variables representing the expert reasoning criteria. The methodological principles announced in section 4.2 are used. In this research, according to the experts, the following knowledge-based variables were created.

(1) **Creating basic dummies**
From S4 a set of six dummy variables are created indicating if the person suffers from each mental health disorder (d) considered in the modalities of original S4 variable, at a certain timestamp (t) .

This corresponds to the case of creating new indicators indicated in 4.2

The new resulting variables are in table 3

$$X_{df} = \begin{cases} 1 \; if \text{ the person has mental disorder d at timestamp t} \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \tag{1}$$

Table 3. Basic dummies created

d (the mental disorder)	**January 2020**
SMD (Severe Mental Disorder)	S4.SMDJan20
PLD (Personality Limit Disorder)	S4.PLDJan20
PSTD (Posttraumatic Stress Disorder)	S4.PSTDJan20
Depression	S4.DepressionJan20
Anxiety	S4.AnxietyJan20
Others	S4.OthersJan20

(2) **Dynamic dummies** on the evolution of mental health diagnoses. These dummies show if the person improved or get worse on each mental health diagnoses during first wave of the pandemics. These are of new variables of type "Condition" from

those described in section 4.2. A total of 18 binary variables are created based on the evolution of each mental health diagnoses between January 2020 and July 2020. See them in table 4

Table 4. Dynamic dummies created

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-

(3) **Debut on mental health disorder:** Also, a new variable to know if the person has debuted in mental health during the first wave. The new variable is called *S4. Mental Problem Debut during pandemics 1st wave (S4. MentDebut):* it has been created with a combination of the variables created in the previous step. It is considered that a person debuts with a Mental Health Problem if he/she had no that diagnose in January 2020 and he/she has it by July 2020 feeling worse than before. This is a new variable of type Condition and it is created as follows:

$$S4.MentDebut = \begin{cases} YES\ if\ \forall d, X_{dJan20} = \{NO\}\ and\ \exists d, X_{dJul20} = \{Pijtor\} \\ NO \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \quad (2)$$

(4) **Mental disorder in January 2020 (s4.MentalDisorderG20):** This is a binary variable D={YES, NO} showing if the person have any mental disorder diagnosed. This are variables of type Indicator.

$$S4.MentalDisorderG20 = \begin{cases} YES\ \ if\ \exists d, X_{dJan20} = \{YES\} \\ NO\ \ if\ \forall d, d, X_{dJan20} = \{NO\} \end{cases} \quad (3)$$

(5) **Mental disorder in January 2020 getting medication (S4S5.MedMental):** This is a binary variable D={YES, NO} showing if the persons with some mental disorder diagnosed is getting medication, which is an indicator of severity. These are of type condition.

$$S4S5.MedMental = \begin{cases} YES\ \ if\ S4.MentalDisorderG20 = \{YES\}\ and\ S5 = \{Yes\} \\ NO \qquad\qquad\qquad\qquad\qquad\qquad otherwise \end{cases} \quad (4)$$

(6) **Most impacting mental disorders in January 2020** getting medication *(S4S5.MedMentalSevere*): Given that COVID19 and the lockdown is seriously

impacting in certain mental disorders, this is a binary variable D={YES, NO} showing if a person suffering from the most serious mental disorders (SMD, PLD or PSTD is getting medication. This indicates that in January this person had already a quite severe mental health impairment. These is of type "Condition".

$$S4S5.MedMentalSevere = \begin{cases} YES \;\; if \;\exists d \;\; in \{[SVM, PLD, PSD]\} such\ that\ X_{dJan20} = \{YES\}\ and\ S5 = \{YES\} \\ NO \quad otherwise \end{cases} \tag{5}$$

(7) **Under 30 (U30)**: In sociodemographic block there is the question *P3.Age* of participant. A new binary variable U30={YES, NO} is showing if the participant is younger than 30.

$$U30 = \begin{cases} YES\ if\ P3.age < 30 \\ YES\ if\ P3.age \geq 30 \end{cases} \tag{6}$$

(8) **Mental Health Comorbidity:** is of type COUNT and indicates how many mental health diagnoses have the person before the lockdown (if a person has a diagnoses of anxiety and depression simultaneously, this variable takes value 2, and so on)

With these new variables, we have better capacity to learn on mental health situation of the respondents. Analysing results (shown in Table 5) from second-generation variables the conclusions are richer and add value to the analysis. With these new created variables, it is easy to see some new pieces of knowledge that were not evident over simple analysis of original data:

From the set of binary variables created above it can be seen the baseline situation before lockdown

Table 5. Basic dummies results

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

- A 24% of the participants declared to have a diagnosis on mental health. From those, 46.28% declares depression. A 58.51% is affected by anxiety. The 10% suffer from severe mental health disorder and a 6.1% from a Limit Personality disorder. A 4.8% is due to Post Traumatic disorder and 11.4% other mental disorders.
- From those, it can be also seen that a 96.94% of them are getting medication.

In Figure 2 more results are shown. The big circle is a percentage from all participants. The percentage of the small circles is relative data of the subpopulation it is referring to. Thus,

- A 73,78% of participants with Mental Health disorders feel worse in July 2020 (Fig. 2)
- A 5.12% of participants developed new mental issues during the COVID19 first lockdown.
- A 68.86% with depression feel worse in July
- 72.38% of participants that suffered from anxiety feel worse.

In addition, from the count new variable created, one can learn that 17% of the participants suffer from a diagnoses in mental health, a 5% has at least two mental disorders simultaneously (like post-traumatic stress and depression) and 1% registers even three different mental pathologies.

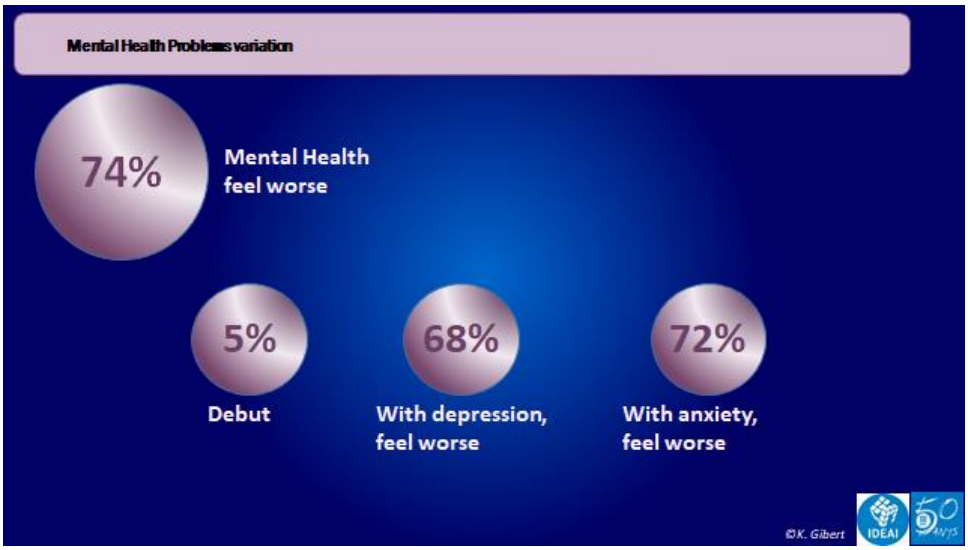

Fig. 2. Mental health data-driven results.

When new variable U30 is used to filter original data and the analysis is repeated for young people (bellow 30) different results are observed. In Figure 3, the results are shown and the main results, obtained from the original data and the pre-processed variables.

- 17% of young participants declared to have a diagnoses on mental health,
- 55% of them are getting medication (in contrast with the 96,94% found for global population)
- 36% declares depression
- 51% declares anxiety
- 12% severe mental health disorder
- 9% Limit Personality disorder

- 6% other mental disorders.
- 38% of young people required emotional support.

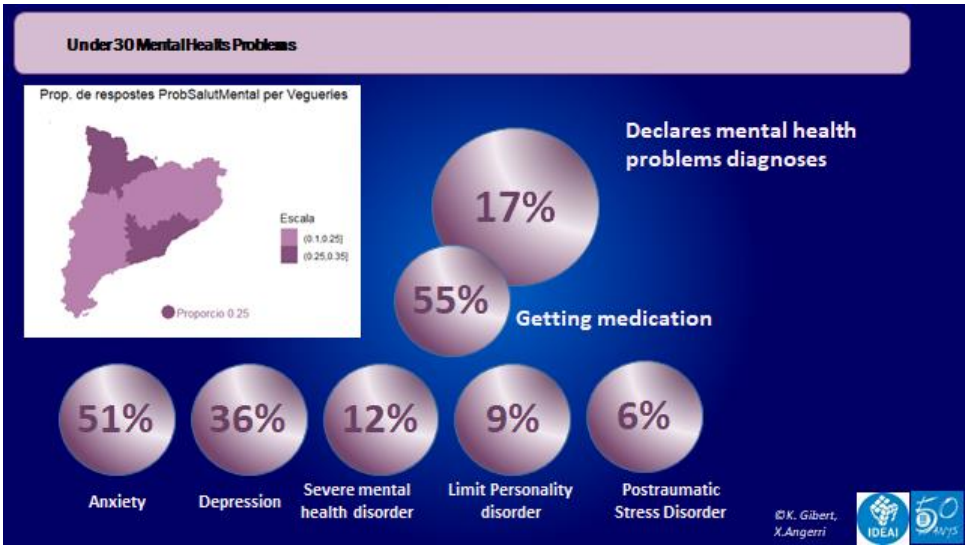

Fig. 3.Under 30 mental health data-driven results.

5.5. *Adding 3rd generation data-driven indicators*

To know the health global situation using all health variables information, 3 new data-driven indicators had been created using the new methodology explained in section 4.3:
(1) Health
(2) Mental Health
(3) Substance abuse

As said in 5.1, there are several blocks containing Health questions. With the experts, the health variables were distributed in three views such that one clustering process per view will provide the corresponding indicator. The components of the three new indicators are shown in Table 6:

Table 6. Components of the data-driven indicators

Data-driven Indicator	**Modalities**	**Components (labels according to Table 1)**
Health	Variables related with COVID disease, dependency degree and disability of the person.	D1, D2, D3, SC1, SC2, Cov1, Cov2, S9, ED1, ED2
Mental Health	Those variables related with emotional support and mental health disorders	TT2.TT.TF, S3, S4, S5
Substance Abuse	are variables related specific with Substance abuse	S6, S7, S8

In the questionnaire, it is possible to see several variables that could be related with health and mental health, but they are not included there because experts considered they do not provide much relevant information. Each group of components is clustered and clusters are interpreted using thermometers and TLPs, together with the CPG and a description of the general situation is available.

The results are shown in an aTLP (see Fig.4), which helps to view how the mental health behaves on the sample data. The TLP method is described in 3.3.2

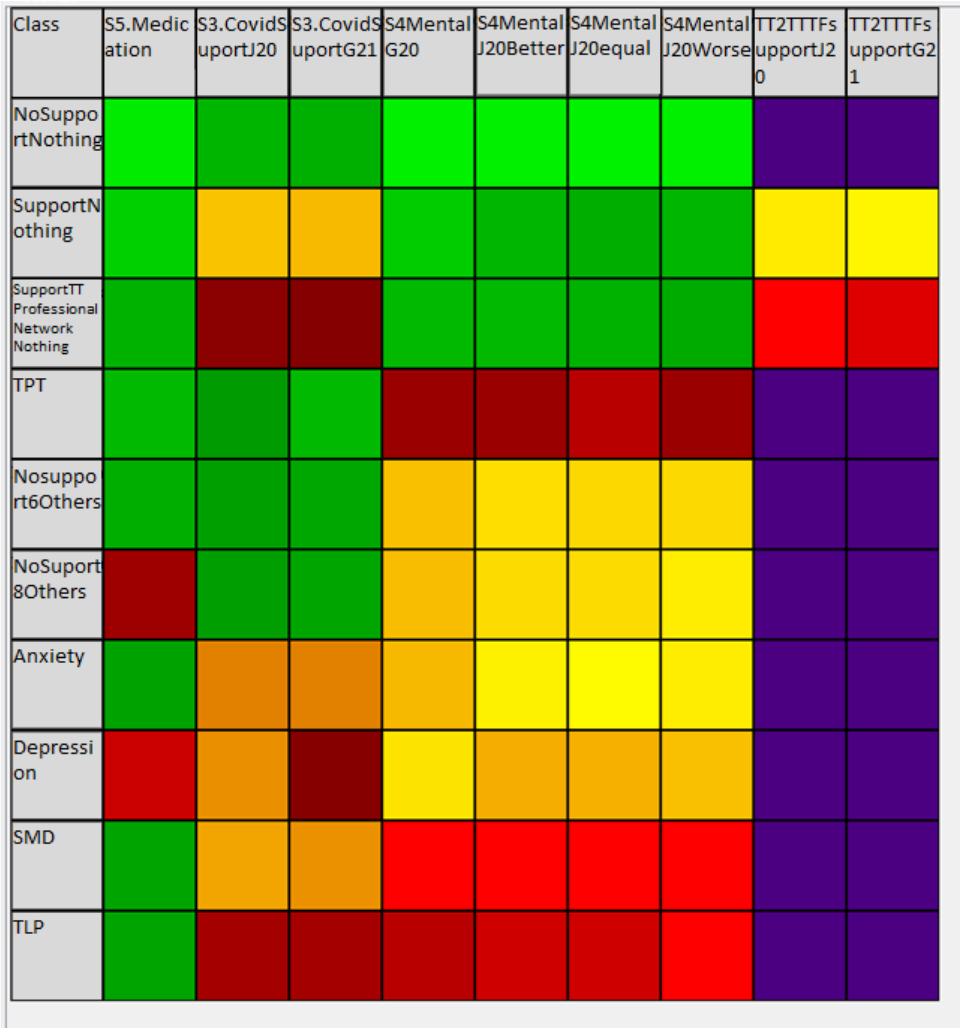

Fig. 4. Mental Health TLP using only original variables (graph: KLASS).

According to the dendrogram inspection, 10 classes are appearing when we are clustering all variables. Using CPGs and TLPs they can be described as follows:

(1) **NoSupportNothing:** They did not telework and they did not have any mental problem. They did not require emotional support and they are not get medication.

(2) **SupportNothing:** They did teleworking and they required emotional support from their personal network. They did not have any mental problem. They did not require professional emotional support and they are not getting medication.

(3) **SupportTTProfessionalNetworkNothing:** They did teleworking and they required emotional support from a professional. They did not have mental problems. They did not require emotional support and they are not getting medication.

(4) **TPT:** They have post-traumatic stress disorder diagnosed. They did not telework. The majority did not need emotional support in July 2020 with the same expectation for January 2021. The majority are not getting medication.

(5) **NoSupport6Others:** They did not do telework and they have other mental problem that are not specified. They did not require emotional support and they are not getting medication.

(6) **NoSupport8Others:** They did not do telework and they have other mental problem that are not specified. They did not require emotional support and they are getting medication.

(7) **Anxiety:** They have anxiety diagnosed. They did not do teleworking. Some of them required support from personal network and others from a professional until July 2020. The majority are not getting medication.

(8) **Depression:** They have Depression diagnosed. They did not do teleworking. Some of them required support from personal network and others from a professional until July 2020. In January 2021 the majority of this group expect to need professional support. The majority are getting medication.

(9) **SMD:** They have severe mental disorder diagnosed. They did not do teleworking. Some of them required support from personal network and others from a professional until July 2020 with the same expectation for January 2021. The majority are not getting medication.

(10) **TLP:** They have borderline disorder diagnosed. They did not do teleworking. They required emotional support from a professional in July 2020 with the same expectation for January 2021. The majority are not getting medication.

5.5. Creating 3rd generation data-driven indicators including 2nd generation knowledge-based variables

Once the knowledge-based variables have been created, they behave as ordinary variables from the technical point of view. So, there is no limitations to use these variables to be included in any further analysis, in particular the creation of 3rd generation data driven

variables, as it is mentioned. 3.2.3. Of course, one have to be careful to use these variables in a correct way according to the technical requirements of the data model to be used (for example, in classical multivariate regression models, independency of regressors is required and this will rise an incompatibility between original and derived variables to coexist in a regression model; each model has its own requirements).

So, the new mental health variables derived from original ones were included in a new clustering process where they are mixed with the original variables. The process is repeated and a new TLP is created (see Figure 5).

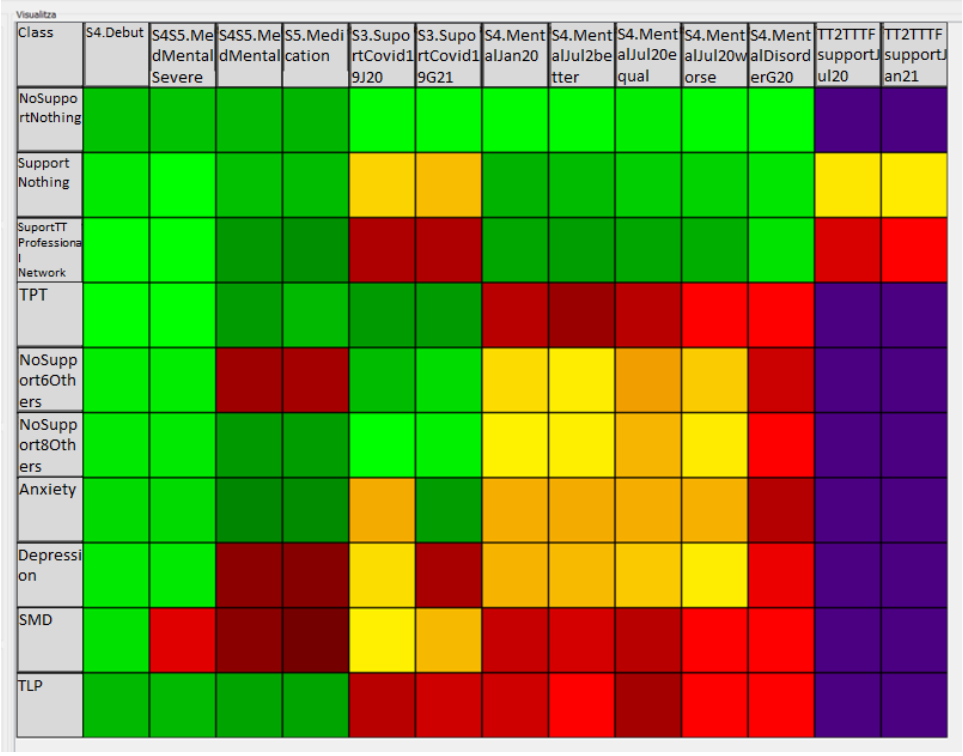

Fig.5. Mental Health with Knowledge-based derived variables TLP (graph: KLASS).

5.5.1. The added value of knowledge-based derived variables when creating data-driven indicators

Thanks to the use of 2nd generation variables in clustering, we are obtaining a lot of additional information.

On the one hand, the colours of the original variables are more brilliant, meaning that the clusters have lower degrees of heterogeneity.

Also, the conclusions that emerge from data where knowledge-based derived variables are included in the analysis are:

- The new variable S4.MentalDisorderG20 reports on having some mental disorder before the pandemics. With its introduction on the analysis, the TLP (Fig 5) clearly shows that 3 of the clusters discovered are mainly composed by people without any mental health issue and this information was not that evident in the TLP with original variables (Fig 4).
- Building variables that indicate whereas the person with certain profile of mental disorders are intaking medication or not, allows to visualize in the TLP that those groups under medication are, in fact, intaking medication for their mental problem.
- Experts were interested in focusing on 3 specific mental disorders as well (SMD, PLS, PSTD). The group of people mainly affected by a SMD (Severe Mental Disorder) are getting medication.
- The groups with higher proportion of people debuting in Mental Health disorders during first wave are concentrated in two main profiles: The one of people suffering Anxiety, the one of people suffering other non-specified diseases and that did not require additional emotional support. There is a third profile with slightly impact on debuts in mental health, the one regarding people that were teleworking during the first wave and those required emotional support either from their personal network or professionals. This means that the people is basically debuting in non-severe mental disorders at the first stage.

For this reason, the preferred TLP is the one in Figure 6.

5.5.2 The extended interpretation of mental health data-driven profiles

The new data-driven indicator for Mental Health is reinterpreted in the next section by including all the new information provided by the knowledge-based derived variables.

(1) **NoSupportNothing:** No one from this group debuted with a mental health issue during the 1rst wave of the pandemics. They did not do teleworking and they did not have any mental problem. They did not require emotional support and the majority are not getting medication. Nevertheless, if those people who is getting medications is not because of any mental health disorder.

(2) **SupportNothing:** The majority did not debuted from mental health problem. They did teleworking and they required emotional support from their personal network, they also say that this emotional support required is due teleworking. They did not have any mental problem. They did not required emotional support. the majority are not getting medication. Nevertheless, if those people who is getting medications is not because of any mental health disorder.

(3) **SupportTTProfessionalNetworkNothing:** The majority did not debuted from mental health problem Due to teleworking they required professional emotional support. They did not have any mental disorder diagnosed. They are not getting medication.

(4) **TPT:** No one from this group debuted from mental health problem. The majority are not getting medication, but those who are getting medication is have mental problem diagnosed. They have post-traumatic stress disorder diagnosed. They did not do teleworking. The majority did not require emotional support in July 2020 with the same expectation for January 2021.

(5) **NoSupport6Others:** The majority did not debuted from mental health problem**. T**hey did not do teleworking and they have other mental problem that are not specified. They did not require emotional support and they are getting medication due to their mental problem.

(6) **NoSupport8Others:** No one from this group debuted from mental health problem. Those who are getting medication is due to mental health disorder. They did not do teleworking and they have other mental problem, which is not specified. They did not require emotional support and they are getting medication.

(7) **Anxiety:** Some people in this group debuted with anxiety during pandemics. They did not teleworking. Some of them required emotional support from personal network and others from a professional until July 2020. The majority are not getting medication, some of the people who is getting medication is due to his mental disorder.

(8) **Depression:** The majority have Depression diagnosed before the pandemics. They did not teleworking. Some of them required emotional support from personal network and others from a professional until July 2020. In January 2021 the majority of this group expect to need professional support. They are getting medication due to their mental problem.

(9) **SMD:** They have severe mental disorder diagnosed. They are getting medication due to this severe mental health problem. They did not telework. Some of them required emotional support from personal network and others from a professional until July 2020 with the same expectation for January 2021.

(10) **TLP:** They have borderline disorder diagnosed. They did not teleworker. They required emotional support from a professional in July 2020 with the same expectation for January 2021. Those who are getting medication is due to their mental health problem.

5.6. *Basic descriptive statistics of original Health variables*

- 12.25% of participants had suffered COVID-19 between March 2020 and December 2020.
- 32% of participants belong a COVID19 risk group.
- 15,75% of participants suffer some disability
- In terms of disability, 15.75% of participants have a disability, and the majority are physical disability (81.04% of people with disabilities). Of these, 57.93% say they have worsened in July 2020 compared to January and of these 58.33% attribute it to COVID19.

- In case of dependency degree 14.9% have a dependency degree declared and 14.2% of people declared that their grade of dependency is worse

5.7. *Data-driven indicators in Health*

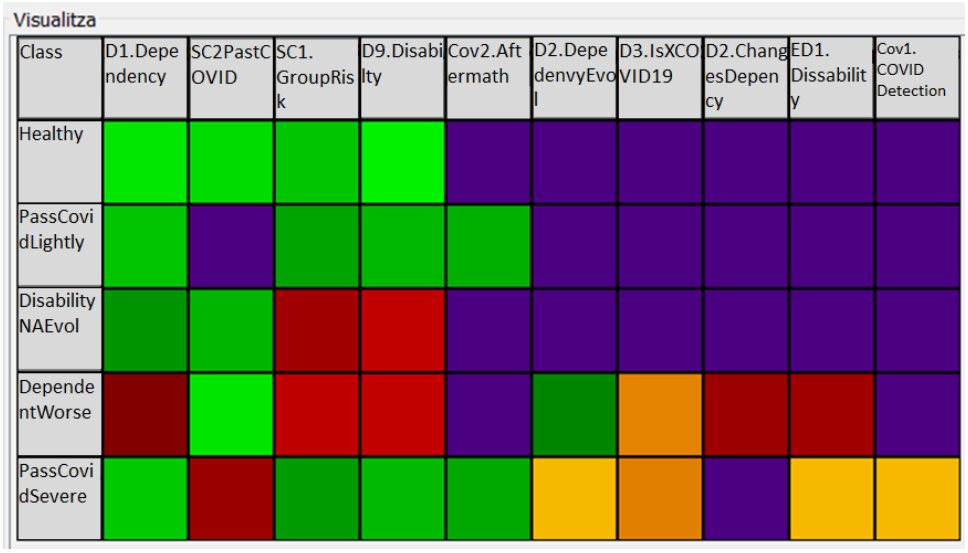

Fig.6. Health TLP (graph: KLASS)

Some of the variables were used to generate a new health indicator (see Fig. 6) and identified the following scenarios

(1) **Healthy:** People in this group does not have health problems. They did not suffer from COVID19, they are not belonging to a COVID Risk group and they are not disabled people.
(2) **PassMildCovid:** They are not disability people neither disabled people. They did suffer from COVID without difficulties and they were able to do normal life after COVID. They are not belonging a COVID Risk group.
(3) **DisabilityNAEvol:** They are disabled people, which means they are belonging a COVID risk group. Nevertheless, they did not suffer from COVID.
(4) **DependentWorse:** They are disabled and disability people, which means they are belonging a COVID risk group. Nevertheless, they did not suffer from COVID. Their dependency is getting worse due to COVID.

(5) **PassCovidSevere:** They are not disabled people. They did suffer from severe COVID and they were able to do normal life after COVID. They are not belonging a COVID Risk group.

5.8. *Second Generation Substance Abuse variables*

The blocks S6, S7 and S8 implement the question about substance abuse. From the structural point of view, it is a TQQ variable and the basic descriptive analysis of original data provides information about the combination of substances that people follows, but more interesting is the information we can obtain by transforming original data into new indicators as described in section 3.2. New variables were created, like those reporting the level of use or the different substances in the three target timestamps. Fig 7 visualizes this information.

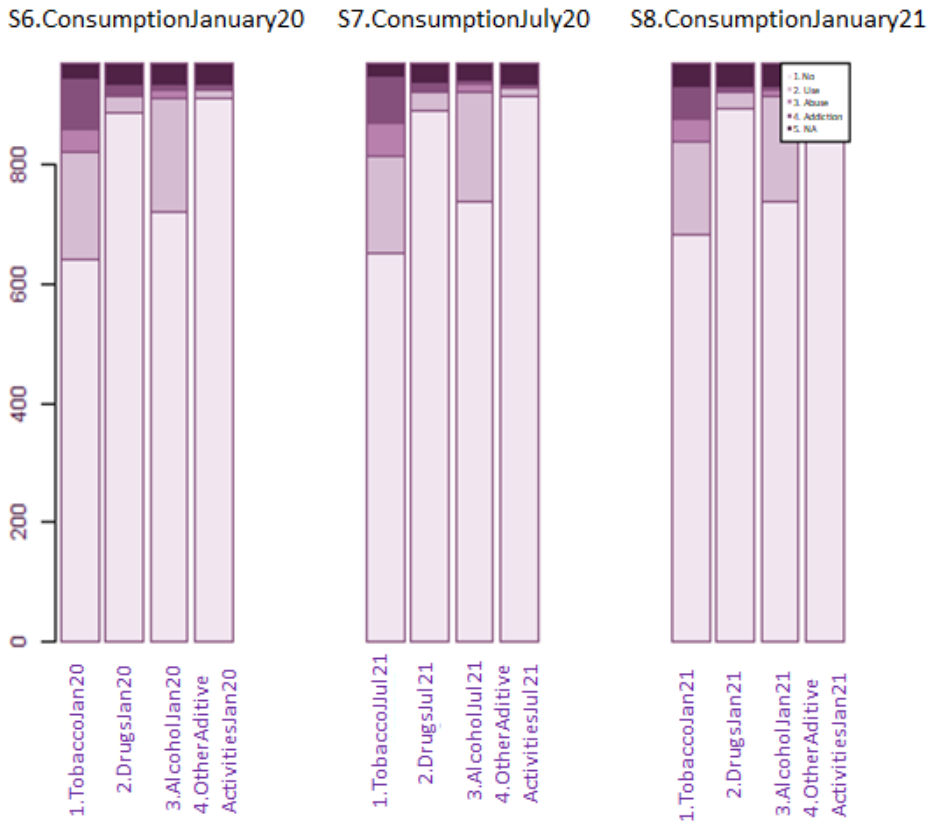

Fig. 7 Multiple barplot of substance abuse.

In this case, an alternative visualization of these type of data is used to better understand the evolution of substances consumption along time. Figure 8 visualizes how consumption of substances change along time.

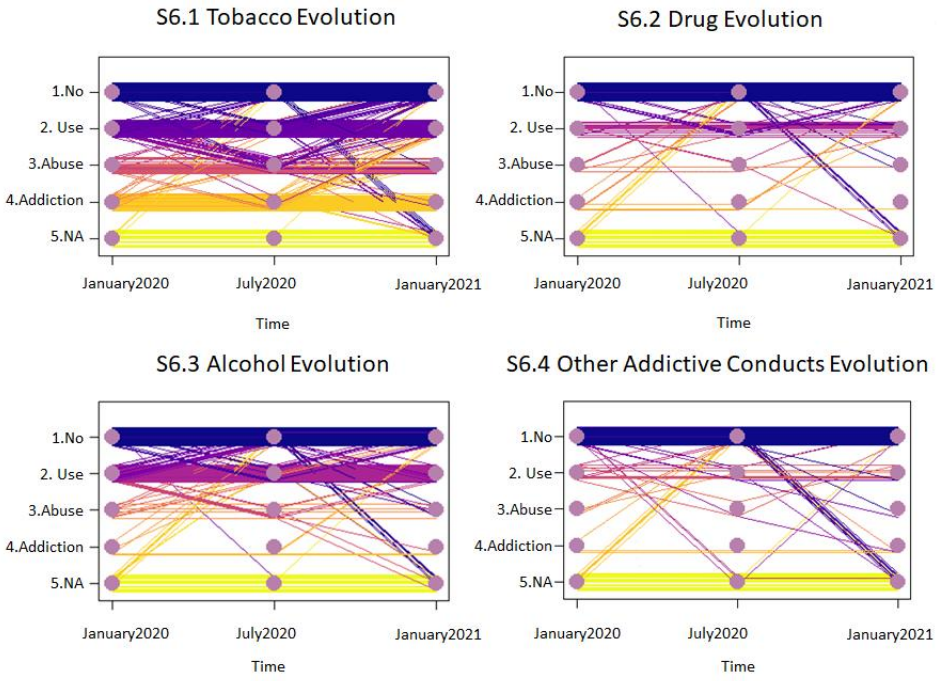

Fig. 8 Trajectory map of substance abuse.

In this block we can see (Fig 7) that Tobacco is the most consumed substance and the number of persons consuming it slightly decreased along the period studied. Indeed 30.8% of the people were smokers in January 2020, and decreased to 25,5% by January 2021. We can also see that:

- Tobacco is the substance with more addicted people (1.8% in January 2020)
- The majority of people are not using substances.

Additionally, Fig 8 shows that:

- In all substances, the majority of people are not moving from their state of addiction (85% for tobacco, 92% for drugs, 88% for alcohol, 92% for other substances)
- In all variables, there is some people who did not answer in January 2020 and January 2021 but they answered "No" in July, that is why in all trajectory maps this ∧ pattern appears.

5.8.4. *Data-driven Substance Abuse indicator*

For those variables composing the indicator of substance abuse, after clustering process (see Fig. 9) the new indicator will have 10 modalities corresponding to the following meanings:

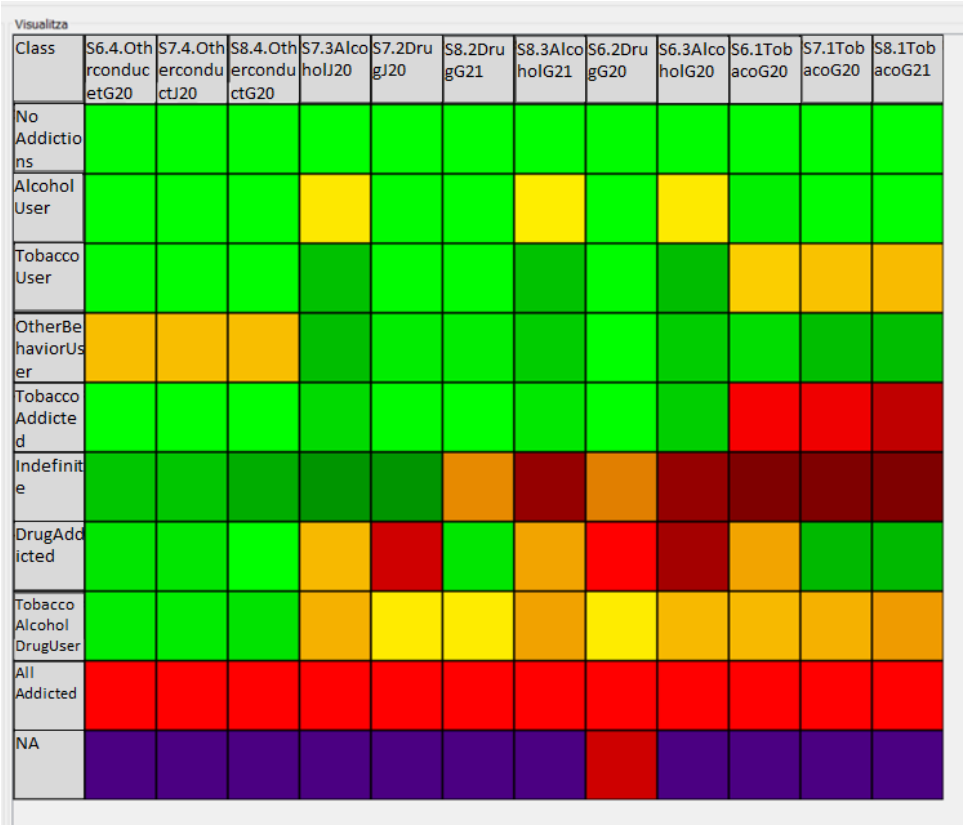

Fig. 9. Substance abuse Traffic Lights panel.

(1) **NoAddictions:** They did not have any addition.
(2) AlcoholUser: They consume alcohol but not other substances.
(3) TobaccoUser: They consume tobacco but not other substances.
(4) OtherBehaviourUser: They follow other conducts like gambling but they do not have other addictions.
(5) TobaccoAddicted: They are tobacco addicts but they do not have other addictions.
(6) Indefinite: This is a mixed group without a clear description. Nevertheless, we could say that they have problems with tobacco and they do not have problems with other addictions. In July 2020 they are not having problems with drugs and alcohol.

(7) DrugAddicted: They were drugs addicted until July 2020. They expect not to be addict by January 2021. They also have problems with tobacco and alcohol
(8) TobaccoAlcoholConsumer: They are drugs, tobacco and alcohol consumers.
(9) AllAddicted: They are addicted to all substances along the entire time..
(10) NA: They did not answer these questions.

6. Conclusions

This work is a contribution to the fundamental data science research and addresses the importance of engineering original (clean) data collections by devoting special attention to the step of deriving new features in the pre-processing process. The proposal focus on two basic mechanisms of addressing this task of very different nature, which provide complementary added value to data for further analytics.

The knowledge-based 2nd generation variables relies on the interaction with domain experts to derive either quantitative or qualitative criteria representing expert's criteria, whereas the 3rd generation data-driven indicators use clustering for further derivation of other qualitative criteria representing new synthetic indicators.

Combining both quantitative and qualitative data with the human in the loop also shows the importance of developing hybrid research methodologies where using DM and AI models increases the quality of observations thanks to complementary data engineering tasks oriented to enrich the dataset with additional information that, on the one hand approach the elements of analysis to the conceptual framework of the experts and their natural reasonings, so that results gain understandability and, on the other hand opens the possibility to explore more complex interactions between the variables. The paper shows that introducing these intermediate steps based on the creation of derived variables of different natures can improve the further analysis and allow a better understanding of a specific target phenomenon. It also shows that AI and DM allow to derive this kind of added-value knowledge and, as a consequence, stakeholders including policy makers can take better decisions.

The methodological proposal is illustrated in a real case study using qualitative research methods (i.e., questionnaires applied to a target population), about health data in the medical arena (i.e., mental health study in the COVID-19 pandemic context). Using the real data on mental health from INSESS-COVID19 database, profiles of vulnerability are sought using clustering techniques, in order to provide support to the policy-making or definition of new social services or updating of current ones to better attend the needs of vulnerable population as a consequence of COVID19 crises.

The paper shows the impact on the on extracted knowledge from data when the 2nd and 3rd generation variables are introduced into the clustering process.

The INSESS COVID19 report, include several conclusions, the majority of which come from a basic descriptive analysis described in previous works [1][10]. Among them, we can highlight that:

- The impact of the COVID-19 crisis on the social services of the territory has two dimensions interacting among them: the impact on people with social care needs, and the impact on practice of professional social services teams.
- The INSESS-COVID19 study shows that COVID-19 crisis is resulting in social realities completely new and different.
- Among the most harmed people are women and elderly.

These conclusions were relevant enough to support some of the first political decisions after the first lockdown.

From the particular Mental Health data, the conclusions obtained are also simple:

- The impact of COVID19 crisis is high in Mental Health.
- 41% of participants needed emotional support during first lockdown.
- 7% of participants suffer only from anxiety (Fig1)
- 5% of participants suffer only from depression (Fig1)
- 5% of participants suffer from depression and anxiety simultaneously in January 2020. (Fig1)
- In January 640 participants didn`t have any mental health problem diagnosed (Fig1).

The basic descriptive analysis on the original variables are easy and fast to use. Nevertheless, the learning does not go much over than simple percentages.

However, data contains much more information that can be extracted with the proper knowledge engineering and data science techniques. The way why original data is combined and transformed into new derived variables provide new layers of knowledge representation that opens the door to richer analyses and better learning from data.

The paper provides a methodology to enrich original data sets with an additional step of pre-processing devoted to creation of new derived variables. Although this step is already mentioned in [31] there were not yet a well-established methodology to address the particular step of creating new derived variables. In this paper, two possible ways to enrich the original dataset with new variables are provided:

- One based on expert knowledge (2nd generation variables),
- The other based on using clustering to create a new variable that synthesizes a complete dimension of data, followed by the use interpretation oriented tools to

label properly the modalities of the new data-driven variable (3rd generation variables).

Thus, the analysis is improved with 2nd and 3rd generation data, created with data mining and Artificial Intelligence methods. This methodology, proposed in this paper, is implemented providing new derived variables based on expert-knowledge and clustering. Therefore, the main conclusion of the paper is that using knowledge engineering techniques to build 2nd and 3rd generation knowledge-based variables and analysing the new enlarged dataset more knowledge is extracted from data.

To go into details, the interpretation of question S4 is really difficult referring the situation in July 2020.

In July20Worse question, 610 participants declare not having mental health disorders diagnosed. There is a loss of 30 people regarding the 640 without mental health disorders in January 2020. These 30 participants choose other modalities in the question July20Worse, meaning that they moved from no mental health problem in January to some mental diagnosis in July. In this case this means the person debuted with some mental diagnoses and in the particular diagnosis from which he/she feels worse. Similar thing happens with the 18 people lost from None category of the question S4Jul20 Equal, and here, interpretation is somehow more confusing, because it does not have much sense that a person with no mental problem in January moves to "equal in depression" or "equal in Anxiety" in July.

Thus, from this basic analysis of Paretto independent graphs, it is difficult to understand in detail which mental health disorders increased prevalence, or severity and which improved after the lockdown.

This is because in the first basic analysis of original data the variables are analysed independently with simple descriptive tools and the interactions between variables are ignored. For the particular case of S4, there is a baseline variable (January 2020) and the other three represent evolutions with respect to the baseline that cannot be caught with simple descriptive tools.

Much added value can be extracted from data when derived variables are created by using the knowledge of experts on the target domain or building data-driven new indicators on data. Applying this new methodology additional knowledge is obtained about mental health issues after lockdown:

- 5% of participants debuted with a Mental disorder during lockdown (this can be quantified upon the transformation of original dataset into a set of dummies). However none of the new mental problems acquired during the lockdown are a Severe Mental Disorder. People debuts with depression, anxiety and post-

traumatic stress disorder mainly.*73,8% from those suffering some mental disorder in January feel worse in July 2020 (this became evident with derived variables created).*

- Specifically, a 68.86% of participants with depression feel worse in July
- A 72.38% of participants that suffered from anxiety feel worse in July.
- An 86,95% of participants that suffered from SMD feel worse in July.

Creating a binary variable (U30) and combining it with other variables the results for young people were obtained.

- 17% of young participants declared to have a diagnoses on mental health (55% of them are getting medication, in contrast with the 96,94% found for global population)
    - 36% declares depression
    - 51% declares anxiety
    - 12% severe mental disorder
    - 9% Limit Personality disorder
    - 6% other mental disorders.

38% of young people required emotional support. The 3rd generation variables can be obtained from clustering, and the resulting qualitative variable interpreted and labelled based on the TLP. By creating 3rd generation variables, it becomes possible to say that there are 10 groups of people related with mental health problems. Three (3) of the 10 emerged clusters discovered are people without any mental health issue. This are the main results on mental Health:

- People who gets medication is due to their mental health problems.
- Participants affected by a SMD are getting medication.
- People who debut are not debuting in severe mental disorders at the first stage.
- People who did telework has needed emotional support.
- People who are suffering an specific Mental Disorder have similar features between them
- The group of people suffering Anxiety, the one of people suffering other non-specified diseases and that did not require additional emotional support. There is a third profile with slightly impact on debuts in mental health, the one regarding people that were teleworking during the first wave and those required emotional support either from their personal network or professionals

As it is seen, the amount and the quality of information increases adding 2nd and 3rd generation variables.

This is also been proved with Health block of questions. The basic descriptive gives the following knowledge:

- 12.25% of participants had suffered COVID-19 during first lockdown.
- 32% of participants belong to a COVID19 risk group.
- 15.75% of participants suffer some disability.
- 14.9% of participants have a dependency degree.

Adding 3rd generation data-driven variables, 5 groups appear.

- People with a Severe COVID episode, tends to be people with disability or dependency
- Participants with dependency, who belong a group Risk their dependency has changed.

In substance abuse block, which is strongly related with mental health block, the basic results obtained after working with 2nd generation variables are the following.

- Tobacco is the substance with more addicted people
- The majority of people are not using substances.
- The majority of people are not moving from their state of addiction after the lockdown.

Adding 3rd generation data-driven variables, 9 groups appear:

- There are people who do not use addictive substances.
- There are people who are addicted to all substances
- Tobacco is the substance that more groups are in contact.

As it is seen, developing more variables thanks to pre-processing techniques and clustering interpreted with the TLP technique, gives more information.

In all cases, results obtained with raw data had been shown to experts. After, results coming from enriched data had been also shown to the mental-health experts, who evaluated very positively the quantity and complexity of knowledge obtained with the enriched datasets, this validating the added value of the 2nd and 3rd generation variables into the further data-driven models. Also, we shown evidences that introducing the new variables into further analysis helps. The paper reports on the clustering of original variables in Mental Health and the impact on extracted knowledge from data when the 2nd generation variables are introduced into the clustering process. The results of a first clustering without second-generation variables, show that the people who are suffering a specific Mental Disorder have similar features between them, because they are in the same group. Moreover, shows that who had done teleworking need emotional support. However, with the clustering using both original and second-generation variables, the number and quality of conclusions increases. For example, it allows a detailed analysis of specific mental disorders and provides information about debuts along pandemics. It is also seen that those groups intaking medication, get specifically a mental health medication.

So as it is seen, in each level, some conclusions from the results are obtained. Nevertheless, if data mining and Artificial Intelligence techniques are being used, more comprehensible are the conclusions.

As a future research lines, the new indicators (either knowledge-based or data-driven) added to the original database become available for more complex analysis providing a more perspective of COVID19 impact. The main limitation of clustering approaches is related with the unsupervised nature of the method, so that performances cannot be directly computed and new mechanisms to quantify the quality of the results are being searched, in order to complete the current validation mechanisms based on interpretability of the results and health experts. In this context it make sense to prepare an experimental setting with synthetic datasets where the real clusters are known and can be used as ground truth, so that quantitative performances and comparisons can be properly addressed. However, this is not a trivial task that can be performed easily. The handicap is to find a synthetic dataset suitable for knowledge-based generation of 2nd generation variables and for concept induction from clusters that allow the real interpretation of profiles and covers the cognitive part of the proposed methodology. As the current extension of the paper do not allow to include this part, we will address this experimental issues in the near future. Finally, although the use of the TLP to support interpretation has been extensively proved in the past, its current main limitation is that in real settings often appears a clusters representing the missing data and this requires in fact a fourth colour (like violet) which is not yet implemented in KLASS, although work is in progress.

Acknowledgement

This research has been partially funded by Development Cooperation Centre (CCD) of the UPC under a special COVID-19 call (CCD-COVID-L019) and by Generalitat de Catalunya with the predoc grant ref:2021-FISDU-00409. Authors want also to thank the partner iSocial and specially Toni Codina for his contribution to the project.

7. References

1. K. Gibert (2020), https://insess-covid19.upc.edu/
2. La Gatta, Valerio, et al. "An epidemiological neural network exploiting dynamic graph structured data applied to the covid-19 outbreak." IEEE Transactions on Big Data 7.1 (2020): 45-55.
3. 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.
4. Vespignani, Alessandro, et al. "Modelling covid-19." Nature Reviews Physics 2.6 (2020): 279-281.
5. K. Gibert, T. Codina and X. Angerri, Informe INSESS-COVID19: identificació de necessitats socials emergents com a conseqüència de la Covid19 i efecte sobre els serveis socials del territori., *Web del project INSESS-COVID19* (2020), http://www-eio.upc.edu/~karina/INSESS/InformeINSESS-COVID19.pdf
6. Messina, Pablo, et al. "A survey on deep learning and explainability for automatic report generation from medical images." ACM Computing Surveys (CSUR) (2022).
7. Kaur, Navdeep, Ajay Mittal, and Gurprem Singh. "Methods for automatic generation of radiological reports of chest radiographs: a comprehensive survey." Multimedia Tools and Applications (2021): 1-31.
8. 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.

9. 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 9.9 (2020): 496.
10. Gibert, Karina, and 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 11.7 (2021): 3110
11. Benzécri, J. P. (1973). L'analyse des données (Vol. 2, p. l). Paris: Dunod.
12. Lauriks, S., Buster, M. C. A., de Wit, M. A. S., van de Weerd, S., Tigchelaar, G., & Fassaert, T. (2012). The Dutch version of the self-sufficiency matrix (SSM-D).
13. The Self-Sufficiency Standard. Available online: https://depts.washington.edu/selfsuff/standard.html (accessed on 13 September 2022)
14. 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.
15. "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.
16. Generalitat de Catalunya, Departament de Treball, and Afers Socials i Famílies. "PLA ESTRATÈGIC DE SERVEIS SOCIALS 2021-2024."
17. Gibert, Karina, Alejandro García-Rudolph, and Gustavo Rodríguez-Silva. "The role of KDD Support-Interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation." Acta Informatica Medica 16.4 (2008): 178.
18. Gibert, Karina. "The use of symbolic information in automation of statistical treatment for ill-structured domains." AI Communications 9.1 (1996): 36-37.
19. Gibert, K., & Cortés García, C. U. (1997). Weighting quantitative and qualitative variables in clustering methods. Mathware & soft computing. 1997 Vol. 4 Núm. 3.
20. 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.
21. Royal Society. Explainable AI. Available online: https://royalsociety.org/topics-policy/projects/explainable-ai/ (accessed on 13 September 2021).
22. Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." Artificial intelligence 267 (2019): 1-38.
23. Burkart, Nadia, and Marco F. Huber. "A survey on the explainability of supervised machine learning." Journal of Artificial Intelligence Research 70 (2021): 245-317.
24. Vilone, Giulia, and Luca Longo. "Notions of explainability and evaluation approaches for explainable artificial intelligence." Information Fusion 76 (2021): 89-106.
25. 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
26. 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
27. 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
28. 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
29. Holzinger, Andreas, et al. "Causability and explainability of artificial intelligence in medicine." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 9.4 (2019): e1312.

30. Gibert, Karina, and Dante Conti. "aTLP: A colour-based model of uncertainty to evaluate the risk of decisions based on prototypes." AI Communications 28.1 (2015): 113-126.
31. Gibert, Karina, Miquel Sànchez–Marrè, and Joaquín Izquierdo. "A survey on pre-processing techniques: Relevant issues in the context of environmental data mining." AI Communications 29.6 (2016): 627-663.
32. Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data." Communications of the ACM 39.11 (1996): 27-34.
33. 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.
34. Vergara, Camila, et al. "Learning on the relationships between respiratory desease and the use of traditional stoves in Bangladesh households." (2016).
35. Hartmann, Thomas, et al. "Model-driven analytics: Connecting data, domain knowledge, and learning." arXiv preprint arXiv:1704.01320 (2017).
36. Gibert, K., 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.
37. 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.
38. Ahlemeyer-Stubbe, Andrea, and Agnes Müller. "Why domain knowledge is essential for data scientists in marketing." Applied Marketing Analytics 7.4 (2022): 362-373.