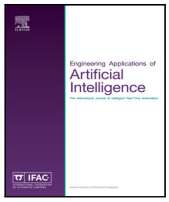
Engineering Applications of Artificial Intelligence 136 (2024) 108910

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Data mining process to detect suicidal behaviour in out-of-hospital emergency departments

José del Campo-Ávila [a,*], Javier Ramos-Martín [b], Carlos Gómez-Sánchez-Lafuente [c,d], Johanna García-Pedrosa [a], Saúl García-Martín [a], Ana I. Martínez-García [e], José Guzmán-Parra [c,d], Rafael Morales-Bueno [a], Berta Moreno-Küstner [b,d]

[a] Universidad de Málaga, Andalucía Tech, Departamento de Lenguajes y Ciencias de la Computación, Campus de Teatinos, 29071 Málaga, Spain
[b] Universidad de Málaga, Andalucía Tech, Departamento de Personalidad, Evaluación y Tratamiento Psicológico, Ampliación del Campus de Teatinos, 29071 Málaga, Spain
[c] Department of Mental Health, University General Hospital of Malaga, Spain
[d] Institute of Biomedical Research in Malaga (IBIMA), Spain
[e] Unidad de Gestión Clínica del Dispositivo de Cuidados Críticos y Urgencias del Distrito Sanitario Málaga-Coín-Guadalhorce, Malaga, Spain

ARTICLE INFO

Keywords:
Emergency calls
Suicidal behaviour detection
Mental health disorders
Class-imbalanced data
Supervised learning
Support clinical decision-making

ABSTRACT

Out-of-hospital emergency departments receive multiple types of requests daily. Their management requires a balance to be found between available resources and the actual needs of the requesting party. Those regarding suicidal behaviour, which are resource heavy, are few in number in terms of the bulk of requests, and detecting them correctly is therefore important. Previous research, using machine learning algorithms to analyse suicide, has typically focused on discovering insights to be used by medical personnel. This proposal extends its use in two directions: knowledge that can be used by non-exclusively medical staff, such as telephone operators, and the models that have been incorporated into a software prototype to help in the decision-making of an emergency department. In addition, previous research has often included a range of information from different sources that are not available when processing an emergency call request, for example, data that is only obtained at the end of the intervention. A full-scale data mining process has been performed using data from the out-of-hospital emergency service in Malaga (Spain). Sensitivity has been the primary goal to avoid missing cases requiring special attention, but this objective has been pursued without overlooking a good trade-off with specificity. The best models can offer such a compromise between sensitivity and specificity, and show more than 80% in both metrics simultaneously. The experts validate that the modelling phase showed that the algorithms have automatically identified already known situations. This lays the groundwork for further iterations with a promising outlook.

1. Introduction

Suicide is one of the leading causes of unnatural death worldwide, which is a public health problem. There are estimated to be 20 earlier attempts for every suicide death (World Health Organization, 2014). The history of earlier suicide attempts is a strong predictor of the occurrence of suicidal behaviour in the future (Parra-Uribe et al., 2017). Identifying and registering these episodes of suicidal behaviour is important to improve future detection. The WHO predicts that the desired reduction in suicides will not be achieved by 2030 (World Health Organization, 2019), and it proposes four interventions, one of which is directly related to this work: *early identification, assessment, management and follow-up of people affected by suicidal behaviour*.

Traditionally, research on suicidal behaviour has used conventional statistical techniques to identify, characterise and predict this behaviour (Franklin et al., 2017). These techniques, based primarily on a manual study of the data, are limited to working with datasets consisting of not so many instances, nor described by too many attributes. Most research in this context uses conventional null hypothesis testing statistical methods, where the researchers manually infer the hypothesis.

* Corresponding author.
E-mail addresses: jcampo@uma.es (J. del Campo-Ávila), javierramos@uma.es (J. Ramos-Martín), carlos.gomez.s.sspa@juntadeandalucia.es (C. Gómez-Sánchez-Lafuente), johannagp9@uma.es (J. García-Pedrosa), saugarmar@uma.es (S. García-Martín), anai.martinez.sspa@juntadeandalucia.es (A.I. Martínez-García), jose.guzman.parra.sspa@juntadeandalucia.es (J. Guzmán-Parra), rmorales@uma.es (R. Morales-Bueno), bertamk@uma.es (B. Moreno-Küstner).

https://doi.org/10.1016/j.engappai.2024.108910
Received 22 January 2024; Received in revised form 27 April 2024; Accepted 28 June 2024
Available online 13 July 2024

For some years now, research in psychology and psychiatry has started to use a second approach to improve the modelling task. In particular, machine learning is complementing the traditional statistical study in this phase of the data mining process. The main reason is to increase the possibility of working with larger datasets and to analyse many patterns, which would be unmanageable manually. Thus, there is a better chance of being able to classify suicidal ideation and other related characteristics (Nordin et al., 2022).

Different approaches have long been developed to advance behavioural identification and they vary widely in terms of the size of the datasets or the way they are constructed. Even the most recent work continues to show this diversity, ranging from small and manually collected datasets that study close interactions between patients and clinicians (Venek et al., 2017), to big datasets that collect publicly available data from social networks and analyse the annotations by experts (Dhelim et al., 2023a; Kodati and Dasari, 2024).

It is common to determine whether longitudinal historical data can be used to predict patients' future risk of suicidal behaviour (Barak-Corren et al., 2017). These *longitudinal studies* use Electronic Health Records (EHRs) over several years and work with millions of instances (records) defined by hundreds and even thousands of attributes. Using that great volume of data enables further variants, such as *case-cohort studies* (Gradus et al., 2020). In this type of study, a manual and directed sampling process selects all the positive cases (minority class) and a small percentage of negative cases (majority class) to build a dataset where the skew between minority and majority classes is softened.

These types of studies can be conducted in some cases, but other constraints apply in others. Accessing every patient's clinical history is sometimes impossible or the access cannot be in real-time. This is the case of out-of-hospital emergency services, where the variety and volume of data for every person requesting help is limited. A second kind of study can be conducted, the *cross-sectional study*. It uses the instances recorded during a short period (a few years) as referred to the same time. This type of study has also been used to characterise and classify suicidal behaviour (Ramos-Martín et al., 2022) and is an alternative when longitudinal historical data are not available.

As already mentioned, machine learning has been previously used in the context of suicide risk from several perspectives and domains (Nordin et al., 2022). For example, Lin et al. (2020) studied suicide ideation in military personnel and Su et al. (2020) focused on children and adolescents. The study sometimes seeks to measure the impact of certain attributes on the models, as Gradus et al. (2020) did when studying the sex-specific suicide risk and two different datasets were generated (one for men and another for women). However, sex information can also be included in the dataset and the algorithms allowed to determine its importance.

What is not so common is the study of suicidal behaviour in the context of out-of-hospital (or pre-hospital) emergency services. Few studies have been conducted and they all use traditional statistical approaches. They study different issues ranging from the more detailed to the more generic. There are papers on one specific cause, such as substance-related issues (Kabadayi and Usul, 2023), and others focused on several separate sets of causes, including cutting self-harm and deliberate self-poisoning (Norotte et al., 2023). There are also two studies that analyse the number and characteristics of suicide attempts according to the available sociodemographic, temporal, and health care variables (Mejías-Martín et al., 2018; Moreno-Küstner et al., 2019). The importance of obtaining a good insight into suicidal behaviour can be applied to different tasks, including: assigning emergency and transport services (Kabadayi and Usul, 2023), evaluating and improving the telephone assessment (Tilley et al., 2024), or proposing sufficient training and clinical support for telephone operators (Doan et al., 2024).

Motivations and key contributions

The data mining process carried out here seeks to predict when a request to an out-of-hospital emergency service involves suicidal behaviour. The results obtained by different models are statistically compared before selecting the best candidates. After validating the usefulness of the new knowledge, the experts have proposed its inclusion in a prototype system that can help in the decision making in such an emergency service.

Therefore, the most significant contributions are summarised as follows:

- Use of data mining process to learn from large datasets in terms of number of instances and attributes in the context of suicidal behaviour at emergency services.
- Learning only considers the information available at the time a request call is processed, and ignoring other information that is available when the case is closed.
- Conducting a data mining process that overcomes the difficulties inherent to an imbalanced problem, because suicidal behaviour is only related to 1% of the emergency service calls.
- Obtaining relevant patterns in terms of model sensitivity and comprehensibility, which can be validated by experts in the field.
- Incorporating discovered knowledge in a software prototype that could assess the telephone operators.

The rest of the paper is organised as follows. Section 2 presents a literature review, with background information, for the scientific positioning of this work. Some preliminaries, relevant to understand the work developed, are introduced in Section 3. CRISP-DM methodology is summarised in Section 4 and the implementation details developed for every phase in this methodology are described in detail. Finally, Section 5 concludes the paper and includes some future lines of research.

2. Related work

For some years now, it has been evident that machine learning techniques are becoming relevant in the study of suicidal behaviour prediction (Nordin et al., 2022). Almost two-thirds of research has been conducted using longitudinal studies where the number of instances is usually high (collected from Electronic Health Records – EHR – for many years). The remaining third has used cross-sectional studies where datasets are usually smaller.

Data sources can be varied and come from EHR systems (Walsh et al., 2017), self-reported questionnaires (Lin et al., 2020), social media (Kodati and Dasari, 2024) or even audio-visual cues (Dhelim et al., 2023b). Depending on the type of data source and the filtering performed to prepare the experimental design, the imbalance between users with suicidal and non-suicidal behaviour can vary. For example, when the study is specific and tries to determine if a particular mental disorder influences the result, the dataset is usually small and balanced (Barros et al., 2017). However, when the dataset is not filtered, the prevalence of suicide behaviour is minimal (near 1%). This is an unaddressed problem (Nordin et al., 2022) with few exceptions that use oversampling techniques to rebalance data (Oh et al., 2020). Moreover, the oversampling solution seems insufficient because it increases the likelihood of overfitting.

A systematic review of research has identified eight types of machine learning algorithms for suicidal behaviour prediction (Nordin et al., 2022): Bayesian-based, instance-based, artificial neural network, regularisation, decision tree, support vector machine, regression, and ensemble techniques. Section 3.3 provides a more detailed description of this algorithm, but the following paragraphs present its use to detect suicidal behaviour.

Bayesian-based models can be considered a naïve approach, but they can perform well in some contexts. For example, Barak-Corren

et al. (2017) predicted suicidal behaviour in healthcare centres, with 80% accuracy, but with highly imbalanced values for sensitivity and specificity. Oh et al. (2020) studied suicidal ideation in the Korean population and obtained more accurate results while keeping a balance between sensitivity and specificity. While Bayesian-based models can obtain accurate models in some cases, instance-based models (like kNN) or artificial neural networks (ANN, including deep neural networks) usually show low or moderately accurate results. The regularisation technique also usually obtains moderate results with an imbalance between sensitivity and specificity far from 80%. Only using regularisation over large longitudinal studies, with millions of instances and hundreds of attributes, reaches balanced sensitivity and specificity near to 80% (Chen et al., 2020).

Decision trees (DT) are a type of algorithm widely used in this context. This is mainly because they can be used to classify and predict while maintaining robustness even with very complex and non-parametric data. Prediction ability is common in machine learning algorithms, but explainability of the discovered knowledge is rare. For example, Edgcomb et al. (2021) applied this method in the specific context of detecting suicidal behaviour after the hospitalisation of adults with serious mental illness. They conducted a longitudinal study with structured Electronic Health Records (EHR) data that included many insights about the patients. Sensitivity and specificity results are balanced and close to 80%. The explainability of the model, with rules up to 9 depth level, was as important as predictive power. Support Vector Machine (SVM) is another algorithm whose use is as widespread as that of decision trees. It usually obtains moderately accurate results, but it is revealed as the best option with a higher frequency (Nordin et al., 2021).

The most extended methods to learn about suicidal behaviour, reaching the most diverse domains in this context, are ensemble methods (such as bagging or boosting) (Nordin et al., 2022). The ensemble methods train multiple learners (known as base learners) and combine them to create a new learner. Random forest (RF) is an extension of bagging, and it is used frequently as it usually avoids overfitting problems, which sometimes appear when using isolated decision trees. For example, Cho et al. (2021) obtains high accuracy values, although sensitivity and specificity need to be more balanced. A minor disadvantage of Random forest models is their explainability because combining different rules from different trees is less direct than consulting a unique decision tree.

Obtaining reliable models offers an excellent opportunity to share such knowledge with experts in the field. Incorporating such models in software that could assess the decision-making task is a crucial advantage that could complete the previously stored information. Computerised clinical decision support systems (CDSS) have been used for many years (Garg et al., 2005). However, their implementation in the suicidal behaviour domain is only testimonial. Kurian et al. (2012) implements patient-self reports to be completed before seeing their clinician. This information is available to the clinician in order to be validated and for a better treatment visit. Etter et al. (2018) proposed a closer case of CDDS to screen suicide behaviour where a similar kind of report (called previsit screener form) is completed by the patient and by the nurse prior to the provider (or clinician) encounter. Other screening tools are used by non-mental health professionals for suicide assessment (Joe and Bryant, 2007). The closest idea to something similar to a system providing some kind of alert is proposed as future avenues for research in a very recent review (Barua et al., 2024). However, to the best of our knowledge, there is no software that assists in supporting decision making in the suicidal behaviour domain, let alone in the context of out-of-hospital emergency departments.

Once the literature mentioned in this section has been examined, we can position this work. The availability and processing of data differs in an out-of-hospital emergency department from other departments, such as primary or secondary care. Although data may be related to each requested call, not all data can be used while handling that call. Two leading causes are limiting data availability during call handling: anonymity, which limits access to clinical data from unknown callers, or limited availability, as certain data are only known at the end of the call handling. The address is a piece of information that is known for practically every call. A more complete dataset can be created with that information and a socioeconomic database storing aggregated indicators for each territorial unit (with 1000 and 2500 inhabitants).

The techniques and algorithms used in the data mining process are well-known, but they must be appropriately organised to overcome several challenges. The imbalance of the data due to the paucity of suicidal behaviour in this cross-sectional study is a significant problem. The learning problem can worsen when the dataset is larger than those traditionally used in cross-sectional studies. The need to obtain balanced results on sensitivity and specificity metrics while prioritising explainable Artificial Intelligence (XAI) models also calls for a specific process set-up. The latter is essential because experts in the field must validate and understand the results. Including such knowledge in a computerised clinical decision support system (CDSS) is then desirable.

3. Preliminaries

This section describes the sources of information and the background methods used to conduct the data mining process proposed in Section 4.

3.1. Raw datasets

Two sources of information have been used in this study. Although a more detailed description will be given in Phase 2 of the CRISP-DM methodology (see Sections 4.3 and 4.4), an initial approximation is made here.

The first source is the Emergency Coordinating Centre (ECC) database serving the capital of Malaga, Spain. This database, in its original version, comprises a total of 83 946 instances collected over three years (2018, 2019 and 2020). Every instance registers the basic information received by an operator processing a telephone service request: information on the requesting party (id, sex, age), place, date and time of the request, or comments on the diagnosis. It is an update of the database used in an earlier paper where the classification of out-of-hospital emergency service (OES) requests related to suicidal behaviour is validated (Ramos-Martín et al., 2022).

The second source is a database that aggregates socioeconomic data from the vicinity of the point where the request is made. These data are collected and made available by the National Statistics Institute (INE) (Instituto Nacional de Estadística, 2021) and provide indicators on the distribution of household income. The database uses territorial units, known as census sections, that subdivide the municipality into zones with a population size between 1000 and 2500 inhabitants. The health district that includes Malaga city and its immediate area of influence is composed by 470 census sections.

3.2. Resampling methods for dealing with imbalanced datasets

Most standard algorithms assume problems where classes are approximately equally represented or have equal misclassification costs, but their performance decreases when that is not the case (He and Garcia, 2009). Imbalanced datasets, where classes have strong distribution skews, are present in real-world domains. Although the context of suicidal behaviours is a clear case in which the issue of imbalance appears, there is hardly any research that has considered this problem (Nordin et al., 2022). Assigning distinct costs to training examples or resampling the original dataset (by undersampling the majority class or oversampling the minority class) (Chawla et al., 2002) are the two main alternatives are used to address the issue of class imbalance.

In the first case, we can talk about cost-sensitive learning, where the aim is to minimise costs related to the dataset instead of the

error itself. Even though there are different types of costs, we consider the misclassification costs (Petrides and Verbeke, 2022). When the problem includes a binary class (with positive or negative labels), four situations can appear depending on the real and predicted labels. Thus, True Positive (TP) and True Negative (TN) denote instances correctly classified as positive or negative, respectively; and False Positive (FP) and False Negative (FN) are the equivalents for misclassified instances. The impact of a mistake varies according to the class for which the error occurs. The matrix known as *cost matrix* (CM) summarises the importance:

$$CM = \begin{bmatrix} C_{TP} & C_{FN} \\ C_{FP} & C_{TN} \end{bmatrix}$$

It is common to consider costs only for errors (hits are usually costless) and the cost changes depending on the type of error (FN or FP). Thus, when predicting a positive instance as negative is worse than predicting a negative instance as positive, the cost of FN (C_{FN}) is larger than the cost of FP (C_{FP}). This circumstance can occur when it is important to identify sick persons (instead of considering them as healthy), to detect fraud cases (instead of considering them as harmless transactions) or to identify suicidal behaviour (instead of classifying it as a less severe circumstance).

In the second case, to facilitate the classifying task to the algorithms, the aim is to reduce the class imbalance. Multiple methods grouped into two approaches can be used and He and Garcia (2009) which usually are combined:

(a) Undersampling the majority class. This approach can be executed using several techniques, such as random undersampling (removing a set of instances of majority class), informed undersampling (that tries to reduce information loss) and Tomek links (that clean up unwanted overlapping between classes).
(b) Oversampling the minority class. There is a variety of methods to perform this oversampling, such as random oversampling (duplicating instances of minority class), synthetic sampling with data generation (SMOTE Chawla et al., 2002) and adaptive synthetic sampling (ADASYN Haibo He et al., 2008).

3.3. Machine learning algorithms

A wide variety of algorithms can be used to induce models that represent the knowledge that can be extracted from data. They are mainly separated into supervised and unsupervised methods. Supervised methods try to classify the instances into as many classes as labels are defined for the class attribute (or dependent attribute). Unsupervised methods do not know which those classes are and group the instances in different clusters that maintain coherence inside the group and search for separation from other groups.

Some examples of supervised algorithms are:

- Naive Bayes (NB) classifier optimally performs when attributes are independent given the class, but it also shows good performance even when dependencies exist between attributes (Domingos and Pazzani, 1997). A variant, the complement Naive Bayes classifier (Rennie et al., 2003), was designed to correct assumptions made by the standard Multinomial Naive Bayes classifier and is particularly suited for imbalanced datasets.
- K Nearest Neighbours (kNN) is a simple classifier mainly used for predictive purposes because of its poor performance in descriptive tasks. It employs the local information close to the instance to predict the class and, as occurs with other classical classifiers, its performance is usually unsatisfactory with imbalanced data (Sun and Chen, 2021). Resampling strategies could improve its performance and much research has been conducted to improve kNN alternatives, although more work is still required.
- Artificial Neural Networks (ANN) are interconnected neurons organised into several layers (input, hidden and output) (Murtagh, 1991). The connections (configured as weights) define the influence between inputs and outputs, and adjust the input values to generate a desired output. ANNs often use the backpropagation technique to adjust the network weights so that the desired output matches the input variables of the labelled examples. The behaviour of the network and its learning process can be adapted with the help of different activation functions.
- Decision trees (DT) are algorithms that can be used for a predictive and descriptive task because they model the knowledge as trees where rules can be extracted. Each branch in the tree represents a rule expressed as a conditional statement that can be understood without expert knowledge. CART (Breiman et al., 1984) and C4.5 (Quinlan and Ross, 1993) are widely used decision tree algorithms.
- Support Vector Machines (SVM) are a type of classification algorithm that utilise linear classifiers to identify the hyperplane that divides data into different categories (Cortes et al., 1995). Support vectors refer to the subset of the instances that identify the location of the separating hyperplane with maximum symmetric margin on both sides that best captures the data trend. Classical SVM algorithms are designed for binary classification problems, although transformations can be used to extend their use.
- Random Forest (RFs) is a decision tree-based technique that involves generating multiple decision trees using different subsets of attributes (Breiman, 2001). It is a learning method in which multiple weak models, trained with different sets of randomly selected observations and different subsets of attributes from the dataset, are combined to get a prediction. The use of different subsets of attributes for different trees is a key aspect that has shown great success (Wyner et al., 2017).
- XGBoost (XGB), or eXtreme Gradient Boosting, is an algorithm that uses the boosting principle to generate multiple weak prediction models sequentially (Chen and Guestrin, 2016). Each model builds on the results of the previous model to create a new model, so it hinders the comprehensibility of the global model, but it achieves improved predictive power and greater stability in its results. An optimisation algorithm such as Gradient Descent combines these weak models into a more robust model.

The above algorithms may include partial or full resampling methods to deal with imbalanced datasets (see Section 3.2). Resampling methods (such as undersampling and oversampling) can be used at the pre-processing stage, so that all algorithms can take advantage of this technique, since the pre-processed dataset is the same for all. The inclusion of a class weight related to misclassification costs provides additional information with which not all algorithms can work directly. The original NB, kNN or ANN algorithms do not use such information and only use the resampling approach.

3.4. Metrics for assessment

Metrics are needed to find suitable algorithms that model the underlying knowledge. Accuracy is the most commonly used evaluation metric for classification, but it may not be a good choice for imbalanced scenarios (Haixiang et al., 2016), because bias towards the majority class masks what happens to the minority class. Other frequently used metrics are Sensitivity, Specificity, Precision, F-Measure or AUC/ROC. There are doubts about the suitability of AUC/ROC as it is based on TP and FP rate, and therefore do not depend on class distribution, what is important in imbalanced scenarios (Fawcett, 2006). Other questions are likewise open related to the use of optimal thresholds in AUC/ROC (Hand, 2009). The four measures to be used in this study are described below:

$$sensitivity = \frac{TP}{TP + FN} \qquad specificity = \frac{TN}{TN + FP}$$

$$precision = \frac{TP}{TP + FP} \qquad F_1 = \frac{2 \times TP}{2 \times TP + FP + FN}$$

These metrics should not be measured arbitrarily. A few guidelines need to be followed to avoid overfitting and to obtain statistical support.

Overfitting occurs when a model includes more terms or uses more complicated approaches than necessary (Hawkins, 2004). This occurs when the algorithm learns the training data set too well, but loses generalisation ability and responds poorly to cases that were not used for training. Strategies such as cross validation can be used to obtain a correct approximation on the performance of the final models. In particular, a good setup of this method suggests using a stratified 10-fold cross-validation (Kohavi, 1995).

After inducing classifiers, considering statistical tests to obtain statistical support for their performance is essential. For example, the Wilcoxon signed-rank test can be used when comparing two classifiers pairwise. Friedman's test with corresponding post-hoc tests may also be suitable when comparing more classifiers over multiple datasets (Demšar, 2006). Multiple results obtained from different runs (reordering, sampling, or cross-validating the original data set) are needed to perform the tests.

3.5. Technologies for implementation

Two different stages can be distinguished in the implementation process: one related to pre-processing and modelling phases and the other related to the deployment phase.

Currently, `R` and `python` are the most widely used languages for data mining, machine learning or artificial intelligence processes. Python (Van Rossum and Drake, 2009) has been selected for this research as it has the necessary libraries to perform all necessary pre-processing and modelling actions (`numpy`, `pandas`, `matplotlib`, `imblearn` or `sklearn`). In general, unless otherwise indicated (see Table 2), the parameters used are the default ones. Full implementation details can be found in the source code available in the public repository.[1]

The development of the software prototype has been carried out with a more engineering process. The web application is based on a client/server structure. The implementation allows the interactive part to be deployed on the user side (using the React library). The server has been created with an API (FastAPI based on python) that provides documentation for each of the endpoints when deployed.

The server needs to use the socio-economic data associated with the patient's address. A database available in the cloud (MongoDB Atlas) has been used to facilitate access to this information.

3.6. Ethical considerations

This research met the ethical research criteria and was approved by the Ethics and research Committee of north-east Málaga (session of May 28, 2020). The dataset after pre-processing phases does not contain identifiable information on patients. All members in the research groups signed a specific confidentiality commitment.

4. Application of data mining process

This section describes briefly a methodology for data mining and presents the results achieved during the deployment of such process.

[1] https://github.com/jcampoavila/ECC-DataMining.

4.1. Methodology for data mining process: CRISP-DM

Conducting data mining processes in many different fields is common nowadays due to the abundance of data and the well-known advantages of its application. Several methodologies have been proposed to guide such a process. One such is CRISP-DM (Chapman et al., 2000), which has become the "*de facto* standard for developing data mining projects" (Marbán et al., 2009) and has been applied to a wide variety of domains since its definition, twenty years ago (Martinez-Plumed et al., 2021).

The upper part of Fig. 1 summarises the six phases in which the CRISP-DM methodology is structured. The first three phases (1, 2 and 3) are equivalent to what is known as pre-processing in other methodologies and they are in charge of constructing a final dataset from which to learn. The modelling phase (4) is responsible for applying statistics and machine learning algorithms to discover new patterns in the data. Finally, the two last phases (5 and 6), also known as the post-processing phase, provide new information to the experts, who can use it to improve their previous knowledge and even include it in software applications.

The CRISP-DM methodology has been applied to carry out the process described in this work and will also be used to structure its presentation in the following subsections. The lower part of Fig. 1 outlines the most relevant information obtained in each phase.

4.2. Business understanding (Phase 1)

It is important to stress that this phase is crucial. Experts in the field here determine the objectives, what will condition all subsequent phases.

Experts in this field include psychologist researchers specialised in suicidal behaviour and professionals at the Emergency Coordinating Centres (ECC). Processing one telephone requests at the ECC involves two activities: the initial classification made by one telephone operator at the ECC and the assessment provided by one health professional that finally attended the emergency. In some cases, the participation *in situ* of the professional is not needed as the request is minor and only requires the first activity to be implemented. Therefore, several roles participate in defining the objectives: psychologists, telephone operators, doctors coordinating the final emergency response, and healthcare professionals who attend patients *in situ*.

The main goals for this research are to:

(a) Analyse the impact of data mining processes when trying to discover novel and useful knowledge in the field, and prioritising the sensitivity metric to avoid false negative cases. This is an usual goal in this area (Gradus et al., 2020; Ramos-Martín et al., 2022). This objective will be achieved if findings, that are already known to the experts, are automatically identified. In addition, it is desirable for the process to include statistical information on the results.
(b) Obtain models (black or white models) that can estimate, in a reliable way, the probability of a case of suicidal behaviour. This goal will be accomplished if any induced model simultaneously obtains sensitivity and specificity values higher than 80%.
(c) Create one software prototype that could be tested at the ECC to assess whether the telephone operators correct label every incoming call. This objective will be achieved if the prototype is implemented in a web-accessible way, so that the emergency service's own software does not need to be modified.

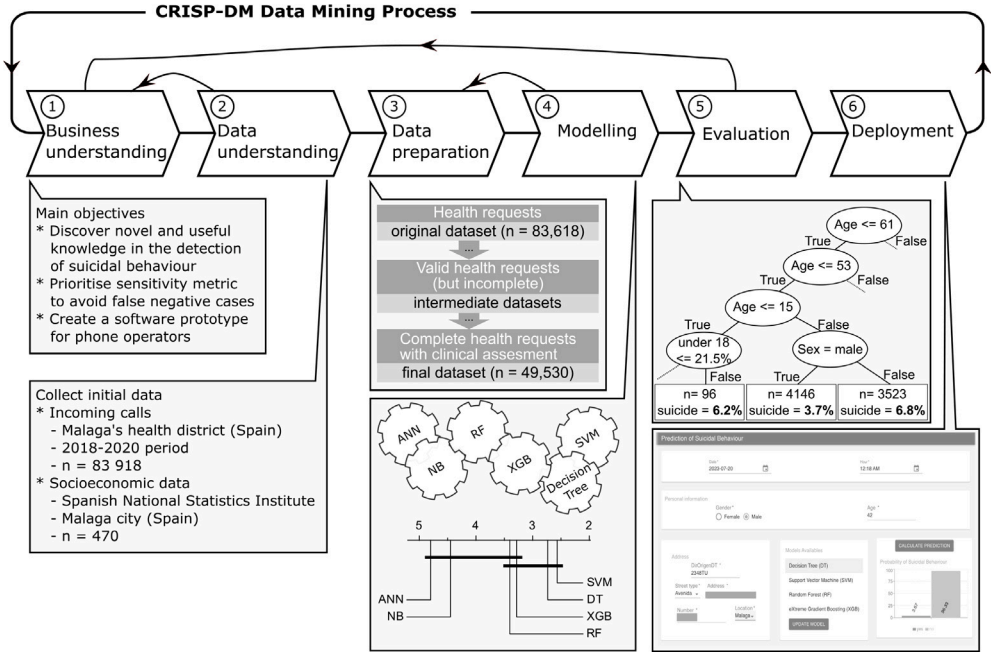


Fig. 1. Diagram of the CRISP-DM methodology (upper part) showing the most relevant results of each phase (lower part).

4.3. Data understanding (Phase 2)

The Emergency Coordinating Centre (ECC) dataset stores data from 83 918 incoming calls (or requests) recorded for three years (2018–2020) in the health district that includes Malaga city (Spain) and its immediate area of influence. This dataset collects information related to the activity that corresponds to an ECC. This dataset is defined by 16 raw attributes that code information on the caller (id, sex or age), place, date and time of the request. Comments regarding the a priori and a posteriori diagnosis, included during the call by the telephone operator or afterwards by the professional deployed to the site, are also stored for every request.

The essential task of defining what is considered suicidal behaviour can be performed using the above information. In our case, such suicidal behaviour label is the dependent variable – the class attribute – about which the learning process revolves. At this point, the clinical assessment made by physicians and researchers is fundamental. The first key to labelling a request as suicidal behaviour is the code given by the professional deployed to the site. If that clinical assessment, according to the International Classification of Disease ninth revision (ICD-9) (World Health Organization, 2011), is identified as V62.84 code (*suicide ideation*) or E950-E959 (*suicide and self-inflicted injuries*), it is positively labelled as suicidal behaviour. Moreover, in order not to miss any possible case of suicidal behaviour, the second key relates to drug poisoning, according to the selection made in the study by Mejías-Martín et al. (2018). When clinical assessment identifies the request with ICD-9 codes 305.4, 305.9 or 969 and, simultaneously, the telephone operator presumes a suicide threat, the request is positively labelled as suicidal behaviour too.

The dataset from the Spanish National Statistics Institute (INE) that includes socioeconomic data for subdivisions of the municipality is used to add new information to every request. It is defined by 30 variables that aggregate indicators such as income (depending on its source), percentages of the population over and under income-based thresholds, percentages of the population in different age ranges or indexes about economic inequality.

4.4. Data preparation (Phase 3)

Several transformations have been applied to the data in order to assure the quality of the dataset that will be used in the modelling phase. In summary, there is a filtering process to obtain valid requests for this domain, a combination of two datasets (telephone requests information and socioeconomic data) and a transformation in the attributes dimension (calculating new variables and deleting some).

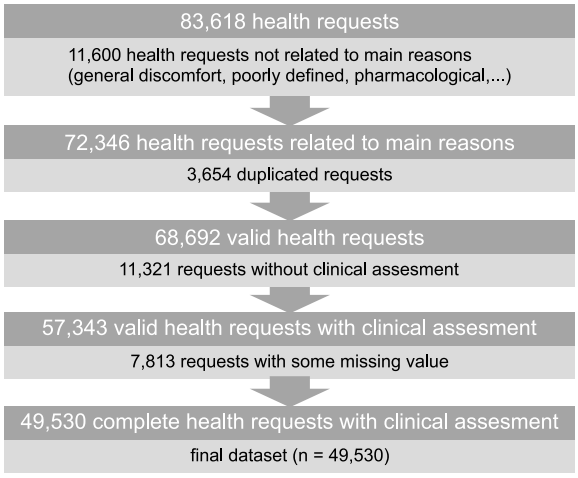


Fig. 2. Flow chart of the filtering process to discard invalid requests in ECC dataset. The result is a dataset without missing values that include potentially relevant instances.

First, a filtering process reduced the size of the ECC original dataset from 83 618 requests to 49 530. Basically, the deleted instances are unrelated to the problem, duplicated or incomplete. Details on filtering are given in Fig. 2. After that filtering, a descriptive statistical analysis was conducted to know the main characteristics of the dataset. The dataset is not perfectly balanced attending to the sex, but the difference is slight: 57.9% for women and 42.1% for men. Considering the age, analysed in 15-year brackets, the brackets with the most requests is shown to be over 75 years old (56.9%), and the second is between 61 and 75 years old (21.1%). The tendency is maintained for the rest of the brackets; the number of requests falls the younger the people. The youngest age bracket, under 15 years of age, is very rare (0.85%).

Calculated attributes are added to the dataset, but the most critical one is the attribute that codes in a binary label the presence or absence

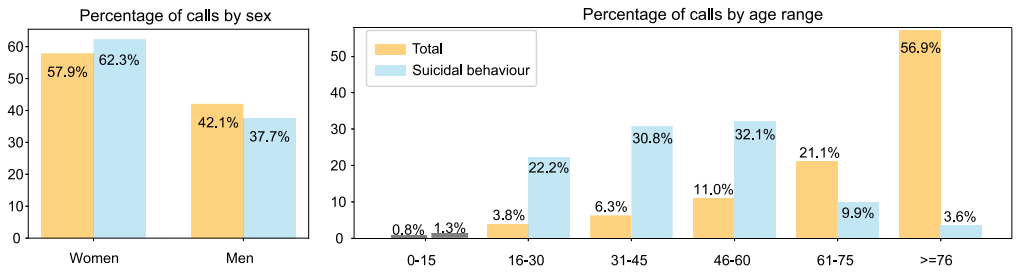


Fig. 3. Percentage of calls in the complete dataset and in the subset with suicidal behaviour depending on the sex and the age range.

Table 1
Characteristics of most of the attributes considered in call requests (full dataset and subset with suicidal behaviour). Nominal values calculated from the date of the call (such as day of the week or month) are not included due to the excess of values for each attribute. There are no missing attributes in the dataset.

Attributes	Total *n = 49 530*	Suicidal *n = 523*
Sex – *n (%)*		
Men	*20 841 (42.0%)*	*197 (37.7%)*
Women	*28 689 (58.0%)*	*326 (62.3%)*
Age – *mean (sd)*	*72.55 (18.91)*	*43.62 (15.97)*
Socioeconomic – *mean (sd)*		
% population with income below 50% of the median	*20.62 (10.26)*	*22.10 (10.64)*
% population with income below 5000 euros	*11.10 (6.51)*	*12.00 (6.80)*
% of single-person households	*26.58 (7.36)*	*26.98 (8.10)*
% of population over 65 years old	*18.81 (6.36)*	*18.10 (6.37)*
% of population under 18 years old	*17.71 (4.38)*	*17.90 (4.51)*
Average net income per person (in thousands)	*10.56 (3.20)*	*10.25 (3.15)*
Average household size	*2.64 (0.28)*	*2.63 (0.29)*
Number of requests made in the last month – *mean (sd)*	*0.96 (3.96)*	*0.95 (4.32)*

of suicidal behaviour. It is defined based on expert knowledge (Phase 2). The number of requests that present suicidal behaviour is 523, which is 1.1%. This percentage coincides with percentages observed in other research (Moreno-Küstner et al., 2019; Ramos-Martín et al., 2022). Descriptive statistics of the population that present a suicidal behaviour does not match with those of population that made an emergency request. For example, the difference between sex increases (62.3% of women versus 37.7% of men) and the most frequent age brackets are then those between 46 and 60 (32.1%), between 31 and 45 (30.8%) and between 16 and 30 (22.2%). Fig. 3 shows the distributions of the total number of requests and number of requests presenting a suicidal behaviour, according to the sex and the age. It is easy to make a visual comparison between the different distributions.

The original set of attributes is enriched with other variables to study the importance of the date, time or location of the request. However, the most relevant information addition comes from the integration of the two datasets defined in Phase 2. The quality of socioeconomic data, provided by the Spanish National Statistics Institute (INE), is very high. Knowing the address from where the request is made allows a correspondence between the aggregated measures and the individual involved in the request to be established. However, the associated information cannot be exact because the measures are defined for the census section where that person lives, so that information corresponds to the person's neighbourhood, not to the specific person.

On the other hand, some of the original attributes are deleted in order to only consider relevant attributes. For example, unique keys or attributes that are not available at the beginning of the request (such as those related to diagnosis, priority or resolution) are discarded. The number of attributes that are finally used is 19 (including the class attribute). They describe characteristics of the person (sex and age), information on the time of the request (day of the week, day of the month, month, quarter, week of the month, week of the year, whether it is a working day, day or night), socioeconomic parameters referring to the person's neighbourhood (percentage of population over 65 years old or under 18 years old, percentage of single-person households, average household size, average net income per person, population with income below 5000 euros or below 50% of the median), use of the ECC (number of requests made by that person in the last month) and presumed suicidal behaviour (class attribute). Table 1 summarises most of the attributes considered in the call request. A dataset with demonstration data is available.[2] Filling this dataset with real data allows the data mining process to be reproduced.

4.5. Modelling (Phase 4)

The standard setting for modelling must be defined prior to detailing the learning process for individual algorithms. Experiments have been conducted following a 3×10-fold stratified cross validation. Therefore, multiple results (30) obtained by slight variations of the dataset can be combined to get statistically supported conclusions. The experiments measure four dimensions relevant to the experts in this field. Details of the metrics for assessment can be found in Section 3.4.

The dataset ready to be used in the modelling phase is defined by 19 attributes and consists of 49 530 instances (requests received at the ECC). The first issue before modelling is the imbalance of the dataset, as only 523 instances are labelled as "suicidal behaviour", representing 1.1% of the entire dataset. To overcome this problem, resampling methods and cost-sensitive learning, described in Section 3.2, have been used.

As every learning process in a 10-fold cross validation uses 90% of instances to train and 10% to test, resampling methods were applied before the training phase, with the test subset (4953 instances) remaining unaltered. Specifically, the oversampling SMOTE parameter was defined as 0.1, and the undersampling rate was defined as 0.2. Therefore, the initial 471 positive instances come to 4410, and the initial 44 106 negative instances to 22 050. The rate of positive instances passes from 1% to 16%. The costs associated to FN and FP are 0.2 and 0.8,

[2] `BDsocioeconomic_dummy.csv` dataset in `INPUT_dataset` folder from repository https://github.com/jcampoavila/ECC-DataMining.

Table 2
Hyperparameter tuning. The best configurations are underlined.

Algorithm	Hyperparameters	Analysed values
NB	Type	{*gaussian*, *complement*}
kNN	Nearest neighbours	{5, 7, 9, 11, 13, 15, 17, 19}
	Weight function	{*uniform*, *distance*}
ANN	Neurons in hidden layer	{100, 200, 300}
	Iterations	{100, 200, 300}
	Regularisation term	{0.001, 0.0001, 0.00001}
DT	Max depth	{4, 7, 10, 13, 16}
	Min samples per leaf	{4, 7, 10, 13, 16}
SVM	Regularisation param	{0.04, 0.2, 1, 5, 25}
	Kernel type	{*rbf*, *linear*, *poly*, *sigm*}
RF	Estimators	{10, 20, 30}
	Max features	{3, 4, 5, 6}
	Max depth	{5, 10, 15}
	Min samples per leaf	{5, 10, 15}
XGB	Balance pos/neg weight	{2, 3, 4, 5, 7}
	Learning rate	{0.1, 0.5, 1.0}
	Max depth	{2, 5, 10}

respectively. The reason is to encourage that cases of suspected suicidal behaviour are not overlooked.

Seven different algorithms have been used to learn from the dataset: Naive Bayes (NB), k Nearest Neighbours (kNN), Artificial Neural Networks (ANN), Decision Trees (DT), Support Vector Machines (SVM), Random Forests (RF), and XGBoost (XGB), all of which have already been introduced in Section 3.3. A limitation imposed by the algorithms themselves or by their implementation does not allow the use of costs for every algorithm. NB, kNN and ANN do not take this alternative into account. Furthermore, XGB cannot include such costs directly, but offers an option to integrate costs.

Since all these algorithms have different *hyperparameters* that control their execution (and therefore can impact their performance), the optimal configuration for each needs to be determined. Table 2 shows the values selected for the parameters of every algorithm. A grid search has been conducted to obtain the configuration that would provide the best results for each algorithm (Bischl et al., 2023). This configuration is marked in Table 2 with the corresponding values underlined.

Following the criteria defined by the experts in Phase 1, sensitivity is one of the most critical metrics to decide the quality of the model. At the same time, high values of sensitivity must not compromise low values of specificity (and accuracy). Taking this into consideration, the objective is to find models with sensitivity of over80% of sensitivity, while the specificity remains over 80% as well. The highest sensitivity value is preferred if the above criterion is not satisfied. If the criterion is satisfied, the highest F_1 score is preferred. These criteria have been used to select the best hyperparameter combination for each algorithm (see Table 2).

The results achieved by those algorithms and configurations are detailed in Table 3. Some ideas can be extracted from these results. One of the most notable ideas is related to the importance of cost-sensitive learning: algorithms that cannot use a customised cost for different types of errors do not reach the desired 80% threshold. The rest of the algorithms do and perform similarly. SVM presents the best sensitivity result.

A statistical validation has been conducted using the Wilcoxon test (with a minimum p-value of 0.05) to detect significant differences concerning the SVM result. The results that are worse and statistically proven are depicted in Table 3 with the symbol ⊖. One algorithm, DT, does not exhibit such a difference (indicated with ⊙), which will be relevant for the next phase. This is so because we can consider both results as equivalent, and DT is more understandable than SVM. Thus, we can keep the best of each option: the resulting decision tree rules can be discussed with the experts in the field, while the SVM model can be used as a black box in a decision support tool.

Table 3
Performance of the machine learning algorithms. SVM shows the best result for sensitivity, although the decision tree does not show significant differences.

Algorithm	Sensitivity (%)	Specificity (%)	Precision (%)	F_1 score (%)
NB	76,7 ± 7,5 ⊖	80,5 ± 0,6	4,0 ± 0,4	7,7 ± 0,7
kNN	27,4 ± 6,6 ⊖	93,2 ± 0,4	4,1 ± 0,9	7,2 ± 1,6
ANN	54,7 ± 32,7 ⊖	82,9 ± 20,8	4,9 ± 2,2	7,7 ± 2,8
DT	81,8 ± 7,5 ⊙	80,1 ± 1,7	4,2 ± 0,3	8,0 ± 0,6
SVM	***82,2*** ± ***4,8***	81,4 ± 0,5	4,5 ± 0,3	8,5 ± 0,5
RF	80,2 ± 5,7 ⊖	81,1 ± 1,0	4,3 ± 0,3	8,2 ± 0,5
XGB	80,6 ± 6,6 ⊖	81,8 ± 1,3	4,5 ± 0,4	8,6 ± 0,8

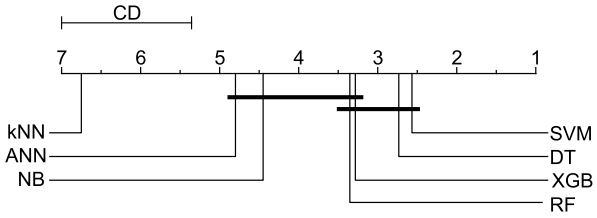


Fig. 4. Diagram to compare performance of different algorithms using the Critical Distance (CD).

Similar conclusions can be observed when the alternative option checking for statistical differences for multiple algorithms is considered simultaneously, although separation requirements are relaxed and there are not as many differences. Fig. 4 describes the results calculated with the autorank tool (Demšar, 2006; Herbold, 2020). Two groups can be identified, one consisting of algorithms that exceed the 80% threshold (those that can use cost-sensitive learning) and the other of algorithms below the threshold.

4.6. Evaluation (Phase 5)

The experts present throughout the data mining process are once again taking on a leading role. They evaluate the results achieved and validate their suitability. Although the results from this study should not be interpreted as causal effects, the patterns discovered can help. We have found four models that meet the requirements of the experts, two of which are of particular interest, SVM and DT.

Age and sex, in order of importance, are the attributes for these two models that contribute most to learning. This is in line with what has been observed in studies conducted by experts in the field because a higher risk of suicide attempts is related to age and sex (Nordin et al., 2022). The third most important attribute for SVM is the number of call requests made in the last month while that third attribute for the DT is the percentage of people under 18. This last attribute is available in the socioeconomic database included in this data mining process, which reveals the usefulness of having included it.

Considering results obtained by the ensemble methods (RF or XGB) is relevant to note that the attributes from socioeconomic database are often among the most informative attributes, apart from age and sex.

The model created by SVM can be used in the prototype designed to help the Emergency Coordinating Centre (ECC). Experts perceive that the final decision can be improved by incorporating a new prediction system, as it better addresses possible suicidal requests and better exploits resources.

The induced decision tree that best fulfils the experts' expectations is small enough to understand easily. It only presents 12 decision rules with a maximum of 4 attributes per rule. What interests the experts are the patterns where the highest number of cases of suicidal behaviour are concentrated. The global rate in the dataset is known to be 1%, so rules that double or triple such percentage are candidates to be useful. In this case, three rules show that increase and the details are given

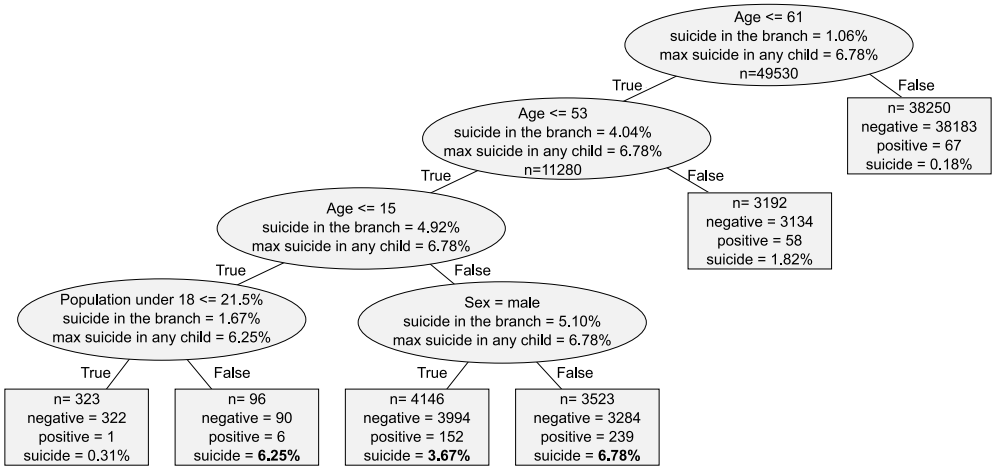


Fig. 5. Decision tree induced with the best configuration and statistics from the original dataset. The "suicide in the branch" label on an internal node means the rate of requests with suicidal behaviour that match the decisions set from the root node to that internal node. It is similar to the "suicide" label at the leaves, but referring to internal nodes. The "max suicide in any child" label on an internal node means the maximum rate of suicidal behaviour that will be observed in the leaves reachable from that internal node. Experts have become interested in branches where these metrics are high. Three leaves are highlighted due to a great growth in the rate of suicide behaviour, far higher than 3% which is three times or more than the global rate.

in Fig. 5. Age seems to be the most discriminating attribute, although sex and the significant presence of the young population can sometimes influence it.

The two most informative rules include 75% of positively labelled suicidal behaviours, while only covering 15% of requests from the whole dataset. Age is common to both rules and reveals itself as the most discriminatory attribute, affecting people between the ages of 16 and 53. The two rules differ in a second attribute, sex. There is a different distribution in the leaves of the tree according to sex. Men show a lower rate than women, rising from a 3.67% to 6.78% respectively.

A third branch (with two rules), that affects a small subset of requests related to people under 16 years old (age ≤ 15), is also revealed as of interest. It considers 419 instances (323 + 96) and there are only 7 (1 + 6) cases identified as suicidal behaviour in that age range. However, the interesting point is how the rate of suicidal behaviour varies in terms of the percentage of young people (under 18) in the neighbourhood of the requester. In the full dataset, 83% of the requests are registered in neighbourhoods where the youth population (under 18) accounts for less than 21%. Nevertheless, most cases in this rule, labelled as positive suicidal behaviour, live in neighbourhoods where the young population is greater than 21%, that is, there is a greater presence of young people in the neighbourhood. This rule will require a more detailed study in a second iteration of the data mining process.

It would be desirable to compare these results to previous research, but that is not possible. To the best of our knowledge, this is the first research of its kind in the out-of-hospital emergency department setting. The most similar studies, which share some similarities, are two cross-sectional studies. They use large datasets, which are not so common in cross-sectional studies, and the class is imbalanced. However, the imbalance is about 10%, which is not as imbalanced as the 1% of our study. Jung et al. (2019) obtains the best result using the XGB method with a sensitivity of 78.5% and a specificity of 79.4%. Something similar happens in the research by Oh et al. (2020), as the LogitBoost ensemble method obtains the best result with a sensitivity of 81.0% and a specificity of 78.7%.

4.7. Deployment (Phase 6)

The last phase is designed to use and disseminate the acquired knowledge. The methodology, the novel knowledge and its potential lines of improvement are disseminated, inter alia, with the publication of this paper. Furthermore, the code that can be distributed is freely available at https://github.com/jcampoavila/ECC-DataMining. The dataset is not distributed due to personal data protection measures, although a demonstration version is available.

The use of such novel knowledge, coded as a predictive model, likewise aims to create a tool that can be used by Emergency Coordinating Centre (ECC) staff when making decisions and creating a diagnosis of possible suicidal behaviour when receiving a call. Fig. 6 shows a screenshot of such a prototype, a web application, where data can be input and where the prediction is made.

The use of this prototype requires the user to consult an additional application to the one normally used, which entails a slight overload. The integration of the prototype into the existing application is not expected for the moment, but it would be feasible when the emergency service considers it appropriate to undertake this improvement.

5. Conclusions

This paper applies a data mining methodology to discover new knowledge for detecting suicidal behaviour in the context of an Emergency Coordinating Centre (ECC) that receives telephone requests. As a final result, once this knowledge had been validated, a software solution has been implemented in the form of a prototype.

In view of the results, one of the strengths of this work is that it has achieved the established objectives. Several techniques have been used in the preprocessing phases, while four machine learning algorithms have been revealed to help achieve the minimum requirements established by experts in this field. A minimum of 80% sensitivity and specificity rates have been achieved simultaneously for this imbalanced dataset. Once those results had been shown to be statistically similar and validated by experts, they have been incorporated into a software prototype that is expected to be helpful at the ECC.

According to the experts' observations after testing the potential of data mining, they would like more detailed rules to be extracted. This would allow them to identify more specific features, even if the group discovered that such a rule was smaller. For this purpose, a second iteration of the CRISP-DM methodology is the next natural step after this first iteration. Due to the good results, future lines appear promising, such incorporating a new dataset with more instances (from more years and regions) and more attributes (meteorological, medical). Testing new preprocessing techniques and various algorithms and configurations are presumed to find more accurate models. Even a new approach from the unsupervised learning perspective seems favourable.

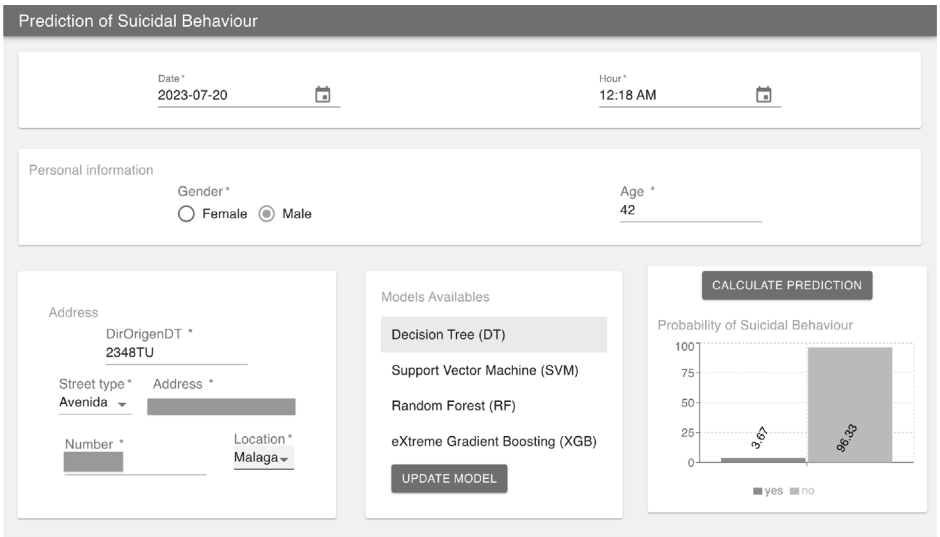


Fig. 6. Screenshot with the prototype software designed for the Emergency Coordinating Centre (ECC).

CRediT authorship contribution statement

José del Campo-Ávila: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Data curation, Conceptualization. **Javier Ramos-Martín:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation, Conceptualization. **Carlos Gómez-Sánchez-Lafuente:** Writing – review & editing, Writing – original draft, Validation. **Johanna García-Pedrosa:** Software, Methodology, Data curation. **Saúl García-Martín:** Software, Methodology, Data curation. **Ana I. Martínez-García:** Validation, Supervision, Resources. **José Guzmán-Parra:** Writing – review & editing, Writing – original draft, Validation, Methodology. **Rafael Morales-Bueno:** Writing – review & editing, Writing – original draft, Validation, Methodology. **Berta Moreno-Küstner:** Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential, but there is a dummy dataset in the public repository to test the code.

Acknowledgements

This study was funded by the Fundación Progreso y Salud (Junta de Andalucía, Spain). Number: AP-0226-2019. Funding for open access charge: Universidad de Málaga/CBUA, Spain.

References

Barak-Corren, Y., Castro, V.M., Javitt, S., Hoffnagle, A.G., Dai, Y., Perlis, R.H., Nock, M.K., Smoller, J.W., Reis, B.Y., 2017. Predicting suicidal behavior from longitudinal electronic health records. Am. J. Psychiatry 174 (2), 154–162. http://dx.doi.org/10.1176/appi.ajp.2016.16010077.

Barros, J., Morales, S., Echávarri, O., García, A., Ortega, J., Asahi, T., Moya, C., Fischman, R., Maino, M.P., Núñez, C., 2017. Suicide detection in Chile: proposing a predictive model for suicide risk in a clinical sample of patients with mood disorders. Braz. J. Psychiatry 39, 1. http://dx.doi.org/10.1590/1516-4446-2015-1877.

Barua, P.D., Vicnesh, J., Lih, O.S., Palmer, E.E., Yamakawa, T., Kobayashi, M., Acharya, U.R., 2024. Artificial intelligence assisted tools for the detection of anxiety and depression leading to suicidal ideation in adolescents: a review. Cogn. Neurodyn. 18, 1–22. http://dx.doi.org/10.1007/S11571-022-09904-0.

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., Lindauer, M., 2023. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. WIREs Data Min. Knowl. Discov. e1484. http://dx.doi.org/10.1002/widm.1484.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees. Routledge, Monterey, CA, http://dx.doi.org/10.1201/9781315139470.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., 2000. CRISP-DM 1.0. pp. 1–76.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic minority over-sampling technique. J. Artificial Intelligence Res. 16, 321–357. http://dx.doi.org/10.1613/jair.953.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 785–794. http://dx.doi.org/10.1145/2939672.2939785.

Chen, Q., Zhang-James, Y., Barnett, E.J., Lichtenstein, P., Jokinen, J., D'Onofrio, B.M., Faraone, S.V., Larsson, H., Fazel, S., 2020. Predicting suicide attempt or suicide death following a visit to psychiatric specialty care: A machine learning study using Swedish national registry data. PLoS Med. 17, e1003416. http://dx.doi.org/10.1371/journal.pmed.1003416.

Cho, S.E., Geem, Z.W., Na, K.S., 2021. Development of a suicide prediction model for the elderly using health screening data. Int. J. Environ. Res. Public Health 18, http://dx.doi.org/10.3390/IJERPH181910150.

Cortes, C., Vapnik, V., Saitta, L., 1995. Support-vector networks. Mach. Learn. 20 (3), 273–297. http://dx.doi.org/10.1007/BF00994018, 1995 20:3.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7 (Jan), 1–30.

Dhelim, S., Chen, L., Das, S.K., Ning, H., Nugent, C., Leavey, G., Pesch, D., Bantry-White, E., Burns, D., 2023a. Detecting mental distresses using social behavior analysis in the context of COVID-19: A survey. ACM Comput. Surv. 55, http://dx.doi.org/10.1145/3589784.

Dhelim, S., Chen, L., Ning, H., Nugent, C., 2023b. Artificial intelligence for suicide assessment using audiovisual cues: a review. Artif. Intell. Rev. 56, 5591–5618. http://dx.doi.org/10.1007/S10462-022-10290-6.

Doan, T.N., Rashford, S., Sims, L., Wilson, K., Garner, S., Bosley, E., 2024. Suicide-related out-of-hospital cardiac arrests in Queensland, Australia: Temporal trends of characteristics and outcomes over 14 years. Prehospital Emerg. Care 28, 431–437. http://dx.doi.org/10.1080/10903127.2023.2230595.

Domingos, P., Pazzani, M., 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Mach. Learn. 29, http://dx.doi.org/10.1023/A:1007413511361.

Edgcomb, J.B., Shaddox, T., Hellemann, G., Brooks, J.O., 2021. Predicting suicidal behavior and self-harm after general hospitalization of adults with serious mental illness. J. Psychiatr. Res. 136, 515–521. http://dx.doi.org/10.1016/J.JPSYCHIRES.2020.10.024.

Etter, D.J., McCord, A., Ouyang, F., Gilbert, A.L., Williams, R.L., Hall, J.A., Tu, W., Downs, S.M., Aalsma, M.C., 2018. Suicide screening in primary care: Use of an electronic screener to assess suicidality and improve provider follow-up for adolescents. J. Adolesc. Health 62, 191–197. http://dx.doi.org/10.1016/J.JADOHEALTH.2017.08.026.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874. http://dx.doi.org/10.1016/j.patrec.2005.10.010.

Franklin, J.C., Ribeiro, J.D., Fox, K.R., Bentley, K.H., Kleiman, E.M., Huang, X., Musacchio, K.M., Jaroszewski, A.C., Chang, B.P., Nock, M.K., 2017. Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. Psychol. Bull. 143, 187–232. http://dx.doi.org/10.1037/BUL0000084.

Garg, A.X., Adhikari, N.K., McDonald, H., Rosas-Arellano, M.P., Devereaux, P.J., Beyene, J., Sam, J., Haynes, R.B., 2005. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: A systematic review. JAMA 293, 1223–1238. http://dx.doi.org/10.1001/JAMA.293.10.1223.

Gradus, J.L., Rosellini, A.J., Horváth-Puhó, E., Street, A.E., Galatzer-Levy, I., Jiang, T., Lash, T.L., Sørensen, H.T., 2020. Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. JAMA Psychiatry 77, http://dx.doi.org/10.1001/jamapsychiatry.2019.2905.

Haibo He, Yang Bai, Garcia, E.A., Shutao Li, 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). IEEE, pp. 1322–1328. http://dx.doi.org/10.1109/IJCNN.2008.4633969.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Gong, B., 2016. Learning from class-imbalanced data: Review of methods and applications. Expert Syst. Appl. 73, http://dx.doi.org/10.1016/j.eswa.2016.12.035.

Hand, D.J., 2009. Measuring classifier performance: A coherent alternative to the area under the ROC curve. Mach. Learn. 77 (1), 103–123. http://dx.doi.org/10.1007/S10994-009-5119-5.

Hawkins, D.M., 2004. The problem of overfitting. J. Chem. Inf. Comput. Sci. 44 (1), 1–12. http://dx.doi.org/10.1021/ci0342472.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21 (9), 1263–1284. http://dx.doi.org/10.1109/TKDE.2008.239.

Herbold, S., 2020. Autorank: A python package for automated ranking of classifiers. J. Open Source Softw. 5, 2173. http://dx.doi.org/10.21105/JOSS.02173.

Instituto Nacional de Estadística, 2021. Atlas de Distribución de Renta de los Hogares. Technical Report, Instituto Nacional de Estadística, URL: https://www.ine.es/experimental/atlas/experimental_atlas.htm.

Joe, S., Bryant, H., 2007. Evidence-based suicide prevention screening in schools. Child. Sch. 29, 219. http://dx.doi.org/10.1093/CS/29.4.219.

Jung, J.S., Park, S.J., Kim, E.Y., Na, K.S., Kim, Y.J., Kim, K.G., 2019. Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. PLoS ONE 14, e0217639. http://dx.doi.org/10.1371/JOURNAL.PONE.0217639.

Kabadayi, E., Usul, E., 2023. Prehospital emergency service use for substance-related issues before and during COVID-19. Emerg. Med. Int. 2023, 1–6. http://dx.doi.org/10.1155/2023/8886832.

Kodati, D., Dasari, C.M., 2024. Negative emotion detection on social media during the peak time of COVID-19 through deep learning with an auto-regressive transformer. Eng. Appl. Artif. Intell. 127, 107361. http://dx.doi.org/10.1016/J.ENGAPPAI.2023.107361.

Kohavi, R., 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. pp. 1137–1143.

Kurian, B.T., Grannemann, B., Trivedi, M.H., 2012. Feasible evidence-based strategies to manage depression in primary care. Curr. Psychiatry Rep. 14, 370–375. http://dx.doi.org/10.1007/S11920-012-0290-Y.

Lin, G.M., Nagamine, M., Yang, S.N., Tai, Y.M., Lin, C., Sato, H., 2020. Machine learning based suicide ideation prediction for military personnel. IEEE J. Biomed. Health Inf. 24 (7), 1907–1916. http://dx.doi.org/10.1109/JBHI.2020.2988393.

Marbán, O., Segovia, J., Menasalvas, E., Fernández-Baizán, C., 2009. Toward data mining engineering: A software engineering approach. Inf. Syst. 34 (1), 87–107. http://dx.doi.org/10.1016/j.is.2008.04.003.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M.J., Flach, P., 2021. CRISP-DM twenty years later: From data mining processes to data science trajectories. IEEE Trans. Knowl. Data Eng. 33 (8), 3048–3061. http://dx.doi.org/10.1109/TKDE.2019.2962680.

Mejías-Martín, Y., Martí-García, C., Rodríguez-Mejías, C., Valencia-Quintero, J.P., García-Caro, M.P., Luna, J., 2018. Suicide attempts in Spain according to prehospital healthcare emergency records. PLoS ONE 13 (4), e0195370. http://dx.doi.org/10.1371/journal.pone.0195370.

Moreno-Küstner, B., del Campo-Ávila, J., Ruíz-Ibáñez, A., Martínez-García, A.I., Castro-Zamudio, S., Ramos-Jiménez, G., Guzmán-Parra, J., 2019. Epidemiology of suicidal behavior in malaga (Spain): An approach from the prehospital emergency service. Front. Psychiatry 10, http://dx.doi.org/10.3389/fpsyt.2019.00111.

Murtagh, F., 1991. Multilayer perceptrons for classification and regression. Neurocomputing 2, 183–197. http://dx.doi.org/10.1016/0925-2312(91)90023-5.

Nordin, N., Zainol, Z., Mohd Noor, M.H., Chan, L.F., 2022. Suicidal behaviour prediction models using machine learning techniques: A systematic review. Artif. Intell. Med. 132, 102395. http://dx.doi.org/10.1016/j.artmed.2022.102395.

Nordin, N., Zainol, Z., Noor, M.H.M., Fong, C.L., 2021. A comparative study of machine learning techniques for suicide attempts predictive model. Health Inf. J. 27, http://dx.doi.org/10.1177/1460458221989395.

Norotte, C., Zeltner, L., Gross, J., Delord, M., Richard, C., Bembaron, M.-C., Caussanel, J.-M., Herbillon, A., Rousseau, C., Chiquet, C., Ehly, C., Pain, A., Vadillo, F., Morisset, L., Roux, P., Passerieux, C., Lambert, Y., Koukabi-Fradelizi, M., Younes, N., Richard, O., 2023. Telephone assessment of suicidal risk at prehospital emergency medical services: A direct comparison with face-to-face evaluation at psychiatric emergency service. Arch. Suicide Res. 1–15. http://dx.doi.org/10.1080/13811118.2023.2265432.

Oh, B., Yun, J.Y., Yeo, E.C., Kim, D.H., Kim, J., Cho, B.J., 2020. Prediction of suicidal ideation among Korean adults using machine learning: A cross-sectional study. Psychiatry Investig. 17, 331–340. http://dx.doi.org/10.30773/PI.2019.0270.

Parra-Uribe, I., Blasco-Fontecilla, H., Garcia-Parés, G., Martínez-Naval, L., Valero-Coppin, O., Cebrià-Meca, A., Oquendo, M.A., Palao-Vidal, D., 2017. Risk of re-attempts and suicide death after a suicide attempt: A survival analysis. BMC Psychiatry 17 (1), 163. http://dx.doi.org/10.1186/s12888-017-1317-z.

Petrides, G., Verbeke, W., 2022. Cost-sensitive ensemble learning: a unifying framework. Data Min. Knowl. Discov. 36, 1–28. http://dx.doi.org/10.1007/s10618-021-00790-4.

Quinlan, J.R.J.R., Ross, J., 1993. C4.5 : Programs for Machine Learning. Morgan Kaufmann Publishers, p. 302.

Ramos-Martín, J., Rueda-Artero, É.L., del Campo-Ávila, J., Martínez-García, A.I., Castillo-Jiménez, P., Moreno-Küstner, B., 2022. Validity of the classification of emergency service requests related to suicidal behavior. Salud Mental 45 (2), 53–59. http://dx.doi.org/10.17711/SM.0185-3325.2022.008.

Rennie, J.D.M., Shih, L., Teevan, J., Karger, D.R., 2003. Tackling the poor assumptions of naive Bayes text classifiers. In: Proceedings of the Twentieth International Conference on International Conference on Machine Learning. AAAI Press, pp. 616–623.

Su, C., Aseltine, R., Doshi, R., Chen, K., Rogers, S.C., Wang, F., 2020. Machine learning for suicide risk prediction in children and adolescents with electronic health records. Transl. Psychiatry 10, http://dx.doi.org/10.1038/s41398-020-01100-0.

Sun, B., Chen, H., 2021. A survey of k nearest neighbor algorithms for solving the class imbalanced problem. Wirel. Commun. Mob. Comput. 2021, 1–12. http://dx.doi.org/10.1155/2021/5520990.

Tilley, D., Christopher, L.D., Farrar, T., Naidoo, N., 2024. Emergency medical service responses as latent social capital toward deliberate self-harm, suicidality and suicide. Psychol. Health Med. 29, 743–753. http://dx.doi.org/10.1080/13548506.2023.2214867.

Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Venek, V., Scherer, S., Morency, L.-P., Rizzo, A.S., Pestian, J., 2017. Adolescent suicidal risk assessment in clinician-patient interaction. IEEE Trans. Affect. Comput. 8, 204–215. http://dx.doi.org/10.1109/TAFFC.2016.2518665.

Walsh, C.G., Ribeiro, J.D., Franklin, J.C., 2017. Predicting risk of suicide attempts over time through machine learning. Clin. Psychol. Sci. 5 (3), 457–469. http://dx.doi.org/10.1177/2167702617691560.

World Health Organization, 2011. International classification of diseases, 9th revision, clinical modification.

World Health Organization, 2014. Preventing Suicide: A Global Imperative. World Health Organization, p. 89.

World Health Organization, 2019. Suicide Worldwide in 2019: Global Health Estimates. World Health Organization.

Wyner, A.J., Olson, M., Bleich, J., Mease, D., 2017. Explaining the success of adaboost and random forests as interpolating classifiers. J. Mach. Learn. Res. 18.