International Journal of Combinatorial Optimization Problems and Informatics, 12(1), Jan-April 2021, 43-57. ISSN: 2007-1558.

www.editada.org

Identification of possible suicide cases using a Bayesian Classifier with the database the Emergency Service 911 of Aguascalientes

Carlos Manuel Ramírez López, Martín Montes Rivera, Alberto Ochoa Zezzatti, Julio César Ponce Gallegos, José Eder Guzmán Mendoza

Universidad Politécnica de Aguascalientes, México.
mc180007@alumnos.upa.edu.mx, martin.montes@upa.edu.mx, alberto.ochoa@uacj.mx, julk.ponce@gmail.com, jose.guzman@upa.edu.mx

Abstract. Nowadays, data storage has become a routine task in the information systems of companies or public institutions. Within the daily work of these organizations, considerable volumes of information are generated, which processed through Data Mining techniques allow decision making. These accumulations of data can be processed and analyzed so that they become the raw material to obtain a new product, knowledge. This study processes the Database of the Emergency Service 911 of the state of Aguascalientes through the application of different techniques of Data Mining for determining the main attributes to consider in a Bayesian classifier that allows prediction of potential suicide cases. The purpose of this processing is to find patterns in consummated suicides, which can help with the identification of potential suicide cases in Aguascalientes.

Keywords: Bayesian Classifier, Data Mining.

Article Info
Received: April 07, 2020.
Accepted: Dec 16, 2020.

1 Introduction

Currently, data storage has become a routine task in the information systems of companies or public institutions [1].

The 911 service daily work generates considerable volumes of information in Aguascalientes, and its processing with different Data Mining (DM) techniques, could allow finding an association, combination, and prediction of data, so that an intelligible model can facilitate the decision-making process related with suicide cases within institutions or companies.

One factor to consider in the DM information pre-processing stage is the selection of attributes for testing in classification. DM allows improving performance, by reducing the dimensionality of data [2], discarding those with trivial or unimportant information, giving an inherent advantage to the computational cost of classification.

The selection of attributes permits finding a minimum subset of the original set of attributes to increase the quality of the classification, facilitating the visualization, understanding, the storage requirements, and reducing the time execution. The selection methods are classified into "Filter," "Layer," and "Hybrid Method" [2].

Another remarkable mechanism in the DM pre-processing stage is the statistical interpolation methods (EBK, Empirical Bayesian Kriging), which generate a prediction surface from a set of scattered points [3]. EBK is useful to know the geospatial information to predict future cases of potential suicide.

EBK results allow a geographical representation of the data distribution, which generates a prediction surface. This interpolation method also presupposes a distance or direction between the points that show a spatial correlation, which in turn can be used to explain the variation of the surface [3].

The pre-processing of the information allows having a data set free of irrelevant information, which is later analyzed with a Bayesian Classification algorithm, allowing the identification of suicide patterns in the database of the Emergency Service 911 of the state of Aguascalientes.

A common theme at a national level and one that has increased considerably is the phenomenon of suicide since it is a social problem that violates the tranquility of the citizens of Aguascalientes. According to the Emergency Service 911 statistics, the behavior of this phenomenon has been on the rise, based on its behavior in the last three years and until October 2019 (the date limit used in the dataset for this study).

The information in the dataset bases on citizen reports to the 911 emergency service, where is expected to detect patterns in cases of completed suicides, allowing early detection of cases that could end in a suicide.

2 Backgrounds

Exist a great diversity of works related to the selection of attributes, application of the EBK method, and probabilistic classification methods. Below, we present the literature related to this study and considered in the development of this research.

Table 1. Backgrounds.

	Title	**Year**	**Type**	**Results**
Feature Selection	Feature selection Based on Class-Dependent Densities for High-Dimensional Binary Data	2012	Filter	BER=6.38%, FD=70.8%
	Attribute selection method by class	2009	Filter	Percentage o error MLP 15.19%
	Heuristic Search over a Ranking for Feature Selection	2016	Filter	ACC=95.54 – ATT=10
Suicide	Acute mental distress associated with suicidal behavior in a clinical sample of patients with affective disorders: Determination of critical variables using artificial intelligence tools	2017	Decision Trees	58.85%
	Comparative study of artificial neural networks applied to the identification of school violence in educational institutions	2018	RNR	96.17%
	Women's violence by their partners according to the Demographic and Family Health Survey, ENDES.	2018	ANN's	87.7%
Empirical Bayesian Kriging	Intelligent system for predicting motorcycle accident by Reaching into a smart city using a kriging model to achieve its reduction and the reduction of deaths in the medium term	2019	Kriging	Z* (x0, y0) = 3-39 WNCP of influenza-like illness (WNCP) Weekly Number of Cases per Practitioner
	Empirical Bayesian Kriging, Implemented in ArcGIS Geoestatistical Analyst	2009	EBK	Prediction 14.72 and 3.52 Error Standard 2.19 and 0.52
	Use of Empirical Bayesian Kriging for Revealing Heterogeneities in the Distribution of organic Carbon on	2017	EBK and Kriging	Error Standard Ordinary Error Standard EBK

	Agricultural Lands			Kriging 0.00005 0.38291	0.00594 0.36979
Bayes	Bayesian System for Diabetes Prediction	2017	Bayes	Matches between 87.6% y 96.9%	
	Design and Construction of a Predictive Model of Depression in Older Adults in Chile	2019	Bayes	Accuracy 0.89 Precision 0.63	

3 Overview of the problem

Suicide in the state of Aguascalientes has become a public health problem that has reached alarming figures. Figure 1 shows the information taken from the database of the Emergency Service 911, with the statistics of suicide in the state of Aguascalientes in the period 2016 - 2019 and where the behavior of this phenomenon can be observed for each month of the year.

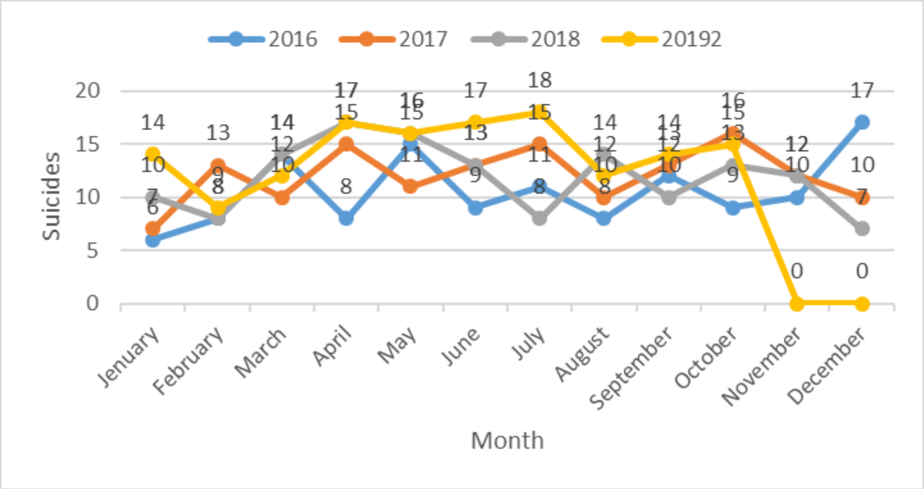

Fig. 1. Behavior of suicide in Aguascalientes during period 2016 - 2019.

As shown in the previous graph, the behavior of this phenomenon is presented for each year and month, where it is clear that the months in which it manifests itself in greater or lesser quantity can be identified, obtaining Table 1 with the comparative statistics:

Table 2. Total completed suicides during period 2016-2019 in Aguascalientes.

Year	Suicides
2016	127
2017	144
2018	142
Until October 2019	144

The analysis of completed suicide cases allows deriving of a characteristics pattern observed in most cases, directly related to the issue of violence in some of its variants (violence against women, violence against partners and family violence). According to statistics from the Emergency Service 911 of the state of Aguascalientes, alarming data related to cases of violence are observed, as shown in Table 2.

Table 3. Reports of family violence received in 2017 and 2018 in the 911 emergency service.

Year	Family violence reports received at the 911 Emergency Service
2017	More than 19 thousand 529
2018	More than 28 thousand 300

This problem is on the rise in the state, as can be seen in the figures of the Executive Secretariat of the National Public Security System (SESNSP), in its report on the incidence of crime in the Common Jurisdiction 2018, which shows this trend in the register of investigation files for the period January-May 2018 and 2019 [12].

Table 4. Records of research folders received in 2018 and 2019 for the period January-May according to the SESNSP in Aguascalientes.

Year	Period	Research folder log
2018	January-May	658
2019	January-May	959

As can be seen in Table 3, the figures are lower than those of the 911 Emergency Service, since not all reports reach the Public Prosecutor's Offices.

In addition to violence, the pattern of characteristics observed in cases of completed suicide showed behaviors related to:

- Suicide attempts.
- Alcohol consumption.
- Drug use.
- Psychological support for mental disorders.

Therefore, we propose the application of a Bayesian Classification algorithm that allows the early detection of possible suicidal behaviors.

4 Theoretical Framework.

Supervised Attribute Selection

The process of reducing the dimensions of the raw information in the database is an essential step in pattern recognition, as is reducing dimensions, which is also essential in exploratory data analysis. There are many potential benefits of selecting variables and characteristics, some of which are: facilitating the visualization and understanding of data, reducing measurement and storage requirements, reducing training and utilization times, and challenging the curse of dimensionality to improve predictive performance. [13].

The methods for selecting attributes can then be classified into three categories [2]:

- Filter: Selects the subset of characteristics based on the intrinsic characteristics of the data.
- Wrapper: Requires an algorithm to determine the best subset of characteristics, guarantees good results but has a high computational cost for large data sets.
- Hybrid Method: Apply the Filter Method initially, and then the Wrapper Method.

Weak and robust relevance

Before starting with methods for the selection of attributes it is indispensable to know in what the selection is based on, since an attribute can be relevant or not for the class, in other words, two degrees of relevance are required "weak and robust". We will take as an example an optimal Bayes' classifier for a given problem where an attribute (characteristic) X is said to be powerfully relevant, since if it is removed, it will cause a deterioration in the performance of the classifier, on the other hand, an attribute (characteristic) X, is weakly relevant if it is not firmly relevant. There is also a subset of characteristics, S, so that the performance of a classifier is worse than the performance of S U {X}. Therefore, an attribute (characteristic) is irrelevant if it is not actively or weakly relevant [14].

Variable ranking

A wide variety of algorithms for the selection of variables includes "Variable Ranking", as the primary auxiliary in the selection mechanism, because of simplicity, scalability and good empirical results; basically, the criterion of Variable Ranking defines individual variables and independent of others [13].

The selection of attributes by filtering has the main disadvantage, that it totally ignores the effects of the selected subset of characteristics on the performance of the induction algorithm [14], on the contrary algorithms based on filtering use the sequential selection of the characteristics that have the best criteria value are computationally more efficient than the Wrapper methods [15].

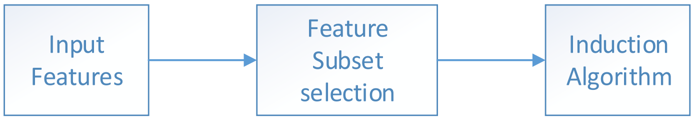

Fig. 2. Approach to feature filtering [14].

When rating Feature Ranking characteristics, weights are typically assigned to the characteristics by evaluating each characteristic individually according to some criteria such as the degree of relevance to the class, the characteristics are ranked according to their weight in descending order [4]

The algorithms for the selection of attributes used for database processing in the Weka tool [16] are shown below:

1. **ChiSquaredAttributeEval,** it evaluates the value of an attribute by calculating the value of the chi-square statistic with respect to the class.
2. **CVAttributeEval Variation Clustering,** seeks a useful subset of attributes to improve the accuracy of supervised learning technique classification in classification problems with a large number of attributes involved. It first creates a ranking of attributes based on the Variation value, then divides it into two groups, the last using the Verification method to select the best group. It evaluates the value of an attribute by calculating the value of the CV with respect to the class. Evaluates the value of an attribute by calculating the value of the CV value with respect to the class.
3. **FilteredAttributeEval,** Class for running an arbitrary subset evaluator on data that has passed through an arbitrary filter (note: filters that alter the order or number of attributes are not allowed). Like the evaluator, the structure of the filter is based exclusively on the data from the training.
4. **GainRatioAttributeEval,** it evaluates the value of an attribute by measuring the ratio of profit to class. The information gain ratio is the relationship between information gain and essential information. It was proposed to reduce bias towards multivariate attributes by taking into account the number and size of branches when choosing an attribute.
5. **InfoGainAttributeEval,** it evaluates the value of an attribute by measuring the information gain with respect to the class. Measures how each characteristic contributes to decreasing the overall entropy.
6. **OneRAttributeEval,** Evaluates the value of an attribute using the OneR classifier, uses the minimum error attribute for prediction, discretizing the numerical attributes.
7. **SignificanceAttributeEval,** it evaluates the value of an attribute by calculating probabilistic significance as a two-way function. (attribute-class and attribute-class association).
8. **SymmetricalUncertAttributeEval,** it assesses the value of an attribute by measuring the symmetric uncertainty with respect to the class.

Analysis of the geospatial situation

The analysis of the geospatial situation is of great interest since it allows us to visualize the geolocalized points and through spatial prediction techniques to obtain the spatial dependence of the data.

Interpolation

The concept of interpolation is based on the fact that spatially distributed objects are spatially correlated; therefore, close things tend to have similar characteristics, in other words, it follows that close points are more likely to be similar than those further away [3].

Kriging Method

The Kriging method, a technique currently used, assume that a geographic data set can be modeled according to a stochastic process [17].

It is a probabilistic predictor, and as such establishes a statistical model of the data, this method has the standard error parameter, which quantifies the associated uncertainty of the predicted values, uses a semivariogram, which consists of a distance and direction function separating two locations to quantify the spatial dependence of the data [18].

The method is based on statistical models that include autocorrelation; in other words, statistical relations between the midpoints. These techniques not only do they provide a prediction surface, but they also give certainty to the predictions.

An important feature is that this method adjusts a mathematical function to a specified number of points to determine the output value for each location. The Kriging method is composed of several steps, which are mentioned below [3]:

1. Exploratory statistical analysis of the data.
2. Variogram modeling.
3. Creation of the surface.

$$\hat{z}(s_0) = \sum_{i=1}^{N} \lambda_i z(s_i) \tag{1}$$

Where: Z(s_i) = the measured value at location *i*. λ_i = an unknown weighting for the measured value at location *i*.
s_0 = the location of the prediction. *N* = number of measured values.

Another concept that is important to know is "Variography", also called model fit or spatial modeling, which consists of the modeling of the midpoints calculated with the following equation.

$$Semivariogram(Distance\ h) = 0.5 * Average\ ((Value\ i - Value\ j)^2) \tag{2}$$

Kriging uses a semivariogram, a function of distance and direction that separates two locations, to quantify the spatial dependence in the data. A semivariogram is constructed by calculating half the mean square difference of the values of all pairs of meters at locations separated by a given distance h [18].

Method EBK

The Empirical Bayesian Kriging method is a geographic interpolation method that automates the most complex aspects of creating a valid Kriging model. Other Kriging methods require manual adjustment of parameters to obtain accurate results, and EBK automatically calculates these parameters through a process of creating subsets and simulations [3].

Another difference of EBK related to other classical Kriging methods is that it considers the error introduced by the semivariogram model estimation, determined with several models instead of a single one.

The method can predict values where observations are not available; this interpolation method consists of obtaining a value for a variable of interest in a place where there are no data using data from places where the data have been collected [18], then we observe the spectrum of a semivariogram produced by the EBK method.

The process of estimating the semivariograms is as follows [19]:

1. A semivariogram is estimated from the subset data.
2. The new data is simulated with the previous semivariogram.
3. A new semivariogram is estimated from the estimated data.
4. Steps 2 and 3 are repeated a specified number of times.

The EBK method, differs from other versions of Kriging by its operation principle that for interpolates any point within the mapping, which is independent of the mapped data, a restricted neighborhood is formed, a variogram is estimated and the predicted value of the point is calculated with the data of its neighborhood, in other words, the interpolation at any point of the mapping is done using only its sub-population of available observations, this makes the method independent of trends [20].

The importance of highlighting this methodology is focused on some of the advantages it represents, such as [19]:

- Standard errors of prediction are more accurate than in other Kriging methods.
- It allows for accurate predictions of moderately non-stationary data.
- It is more accurate than other Kriging methods for small datasets.

Similarly, it is vital to highlight the disadvantages of using this method presented below [19].

- Processing time increases rapidly as the number of entry points increases.
- The Empirical Logarithmic transformation is particularly sensitive to outliers.

Bayesian theory

As a basic concept, it is necessary to know the meaning of Bayesian Probability. To better understand the theory, an experimental case is presented, where it is assumed that there are two classes w1 and w2 where the patterns belong, and there are two initial probabilities for each class. [21].

$P(w1), P(w2)$

If by any chance these probabilities are not known, they can be obtained in the following way:

$$P(w1) \approx \frac{N2}{N} \tag{3}$$

Where:

N = This is the total number of training guidelines available.

$N2$ = Class membership.

Another statistical amount that can be assumed to be known is the conditional probability density of the class $P(x|wi), i = 1,2,$ which describes the distribution of the characteristics vector in each of the classes. [21].

Now you have all the elements needed to obtain the conditional probability, Eq.4. Shows the Bayes' Rule.

$$P(wi|x) = \frac{P(x|wi)P(x)}{P(x)} \tag{4}$$

Where:

$P(\omega_i|x)$ Probability that a class ω_i is presented given a characteristic x.

$p(x|\omega_i)$ Probability of a characteristic x given a class ω_i.

$P(\omega_i)$ Probability of a class ω_i.

$P(x)$ Probability that characteristic x is present.

$$p(x) = \sum_{i=1}^{2} p(x|w_i)P(w_i) \tag{5}$$

Therefore, the Bayesian classification can be declared as follows:

Si $P(w1|x) > P(w2|x), x$ is classified as w1

Si $P(w1|x) < P(w2|x), x$ is classified as w2

The following equation was used to obtain the probability of any characteristic being present:

$$p(x|w_i) = \left(\frac{1}{\sqrt{2\pi s}}\right) e^{-\frac{(x-m_i)^2}{2x^2}} \tag{6}$$

Where:

σ (Sigma) is the variance or standard deviation.

μ_i (Miu) is the average.

$$\sigma = \sqrt{(\mu - \mu_i)^2} \tag{7}$$

5 Methodology

The stage of the methodology comprises three processes, the initial one corresponds to the selection of attributes, which helps to determine the attributes that will be during the pre-processing of the information, then there is the process of Geostatistical Analysis, which provides essential information in the prediction of new cases of suicide, and finally the Bayesian Classifier, which by means of Bayes' probabilistic method classifies the set of data in order to determine whether or not a specific case is a candidate.

Attribute selection with Weka®

The pre-processing of the dataset is necessary to consider the data entry of Weka® Attribute Selection algorithm. In this stage, the database is converted to ARFF, which is an ASCII text file with the database adapted to Weka® [22]. The 911 database is in a SQL Server, so it is necessary to execute a query to extract a subset of attributes of the database.

The result of the query is exported in Excel in CSV format, where it is possible to visualize the types of data (numerical, text or null), and adapt it appropriately to the algorithms for the Selection of Attributes of Weka®.

Weka® identifies different types of data, each of which is listed below [22]:

- Numerical, they can be real or whole numbers.
- Nominal, values with predefined labels.
- String, arbitrary text strings.
- Date, data with date format.
- Relational, other relations.

Once the format of the database has been changed, it is evaluated by the Weka® Attribute Selection algorithms, so we proceed to open the Explorer tool, where the database loads the .arff format and where the characteristics of the attributes to be selected can be observed. It also allows editing the database if it is necessary.

In the following list, we show the 25 attributes evaluated by each attribute selection algorithm, all of them belong to the class SUICIDE.

14.- Id_Community
15.- Municipality_Code
16.- Full Name
17.- Id_Sex
18.- Age
19.- Age1
20.- Id_Rol_Person
21.- Diagnostic
22.- Id_Mode
23.- Telephone
24.- Inc_No_Ext
25.- Suicide

In the same way, the algorithms for the selection of attributes used in Weka® are shown below.

1. ChiSquaredAttributeEval.
2. CVAttributeEval.
3. FilteredAttributeEval.
4. GainRatioAttributeEval.
5. InfoGainAttributeEval.
6. OneRAttributeEval.
7. SignificanceAttributeEval.
8. SymmetricalUncertAttributeEval.

Geostatistical Analysis with Empirical Bayesian Kriging in ArcMap®

The first step in the EBK method is filtering the database corresponding to completed suicides, for allowing its georeference, as shown in Fig.4. The filtering obtains a total of 537 completed suicides within the period of 2014-2018.

The data was georeferenced on the Google Maps® platform, which allowed the location to be obtained in the WGS84 (Decimal Degrees) coordinate system (Table 5).

Table 5. Location in WGS84 coordinate system (Decimal Degrees).

Country	Position
México	21.88833, -102.27196
México	22.23094, -102.32273
México	21.86333, -102.3105
México	21.90667, -102.28137
México	21.89201, -102.29914
México	21.87234, -102.26647

The geostatistical analysis implies the "Convert Coordinate Notation" tool of the ArcMap® software since it does not adequately process the WGS84 (Decimal Degrees) coordinate system. Once the conversion of the coordinate system is done, it is necessary to add the points to the ArcMap® base map.

For this topic of study the base map "Open Street Maps" was selected, which allows the visualization of the streets.

The dataset has 25 characteristics which are mentioned below:

1. Id_Mode
2. Id_Rol_Person
3. Full Name
4. Date
5. Hour
6. Age
7. Inc_No_Int
8. Id_Community
9. Telephone
10. Id_Street
11. Domestic Violence
12. Alcoholic
13. Age1
14. Droug Consumption
15. Id_Colony
16. Inc_No_Ext
17. Domestic Violence_2
18. Municipal_Code
19. Psychological support
20. Dating Violence
21. Diagnostic
22. Suicide threat
23. Id_Sex
24. Violence against women
25. Suicide

As it can be observed for the case study only some characteristics are of interest, as it is the case of:

- Id_Street
- Inc_No_Int
- Inc_No_Ext
- Id_Colony
- Id_Community
- Municipal_Code

These allow using the Google Maps® tool, to generate the geolocation of each one of the 537 suicides, to obtain the coordinates X and Y, of each eventuality, to enter them to the ArcMap® software later. Below, the geolocated points are shown in both tools.

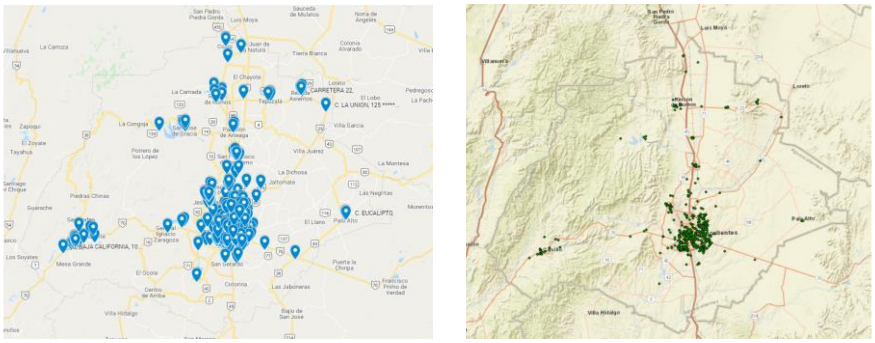

Fig. 3. Geolocation of suicides - Google Maps® (Left).
Geolocation of suicides - ArcMap®(Right).

The next step is to select the data source you want to work with EBK algorithm, which is the previously geolocalized data. Within the "General Properties," it is necessary to set the size of the data subset to 100 for showing more detailed information.

Bayesian classifier

The Bayesian Classifier was developed in Matlab®, where for the application of this classifier it was necessary to modify the content of the attributes shown below: Full_Name, Inc_No_Int, Inc_No_Ext, and Diagnostic.

The procedure followed for the development of the Bayesian classification algorithm corresponds to the following flowchart:

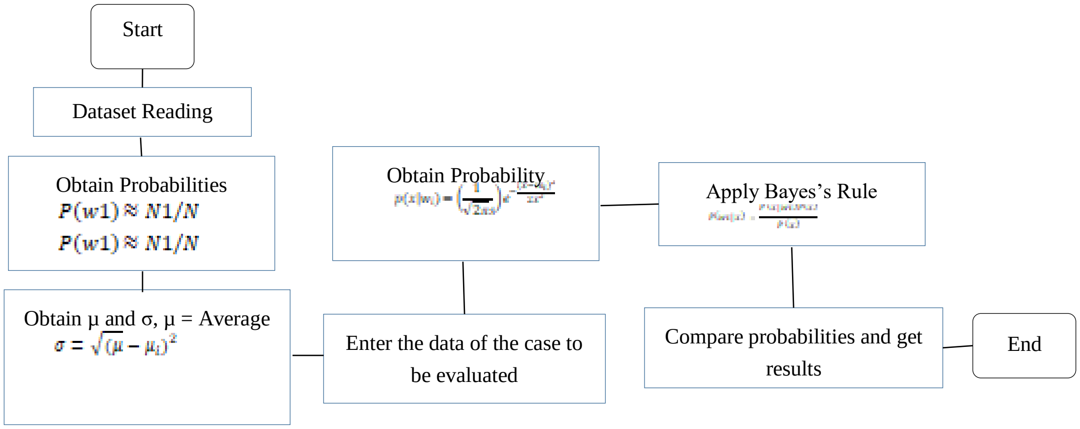
Start
Dataset Reading
Obtain Probabilities
$P(w1) \approx N1/N$
$P(w1) \approx N1/N$
Obtain μ and σ, μ = Average
$\sigma = \sqrt{(\mu - \mu_i)^2}$
Obtain Probability
$p(x|w_i) = \left(\frac{1}{\sqrt{2\pi\sigma}}\right)e^{-\frac{(x-\mu_i)^2}{2\sigma^2}}$
Enter the data of the case to be evaluated
Apply Bayes's Rule
$P(w_i|x) = \frac{P(x|w_i)P(w_i)}{P(x)}$
Compare probabilities and get results
End

Fig. 4. Flowchart for the development of the Bayesian classification algorithm.

6 Results

This section shows the results obtained in each of the processes, and the different techniques used in its development.

Results obtained in the selection of attributes

Below are the results obtained when evaluating the subset of 24 attributes and 1 million 48 thousand 490 instances of the database, with each of the eight algorithms mentioned above. The result of the eight algorithms, allows the calculation of the average for each attribute, to rearrange them in an ascending way and visualize them ordered according to its importance.

Table 6. Comparison of results in algorithms for the selection of attributes.

No	Feature		Pos		Pos		Pos		Pos		Pos		Pos		Pos		Pos		Avrg
1	Id_Mode	ChiSquaredAttributeEval	4	CVAttributeEval	6	FilteredAttributeEval	10	GainRatioAttributeEval	4	InfoGainAttributeEval	10	OneRAttributeEval	1	SignificanceAttributeEval	7	SymmetricalUncertAttributeEval	4	Miu	5.75
2	Id_Rol_Person		5		9		6		2		6		13		4		2		5.875
3	Full Name		1		14		1		5		1		24		5		5		7
4	Date		8		20		4		13		4		3		2		13		8.375
5	Hour		3		21		2		9		2		23		1		9		8.75
6	Age		13		8		11		3		11		15		9		3		9.125
7	Inc_No_Int		7		1		7		11		7		22		10		11		9.5
8	Id_Community		12		4		12		8		12		8		17		8		10.125
9	Telelephone		2		22		3		12		3		21		8		12		10.375
10	Id_Street		6		19		5		15		5		18		3		14		10.625
11	Domestic Violence		17		17		14		7		14		5		12		7		11.625
12	Alcoholic		16		18		13		6		13		12		11		6		11.875
13	Age1		11		23		17		1		17		16		16		1		12.75
14	Droug Consumption		18		15		15		10		15		6		13		10		12.75
15	Id_Colony		10		13		8		20		8		17		6		20		12.75
16	Inc_No_Ext		9		7		9		22		9		19		14		21		13.75
17	Domestic Violence_2		20		11		19		14		19		2		19		15		14.875
18	Municipal Code		15		12		18		19		18		9		18		18		15.875
19	Psychological support		21		10		21		16		21		4		20		16		16.125
20	Dating Violence		22		5		22		17		22		7		22		17		16.75
21	Diagnostic		14		2		16		24		16		20		21		24		17.125
22	Suicide threat		23		3		23		18		23		10		23		19		17.75
23	Id_Sex		19		16		20		23		20		11		15		22		18.25
24	Violence against women		24		24		24		21		24		14		24		23		22.25

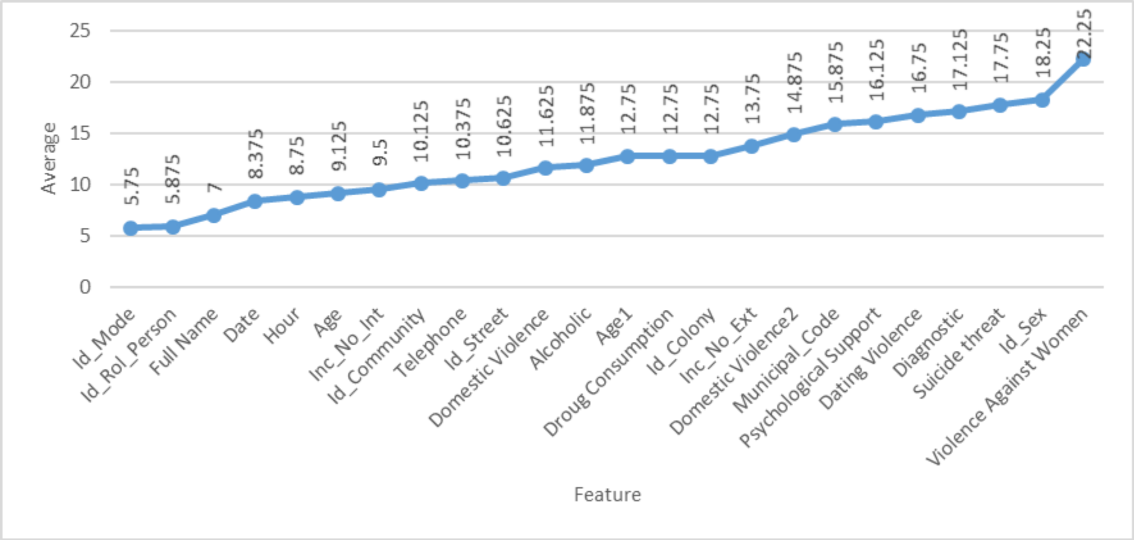

Fig. 5. The average of each attribute ordered according to their power of discrimination.

Below is the list of attributes in the database, ordered by their discrimination power, where we can see that attribute No.1 "Id_Mode" has a higher discrimination power for the suicide class.

1. Id_Mode
2. Id_Rol_Person
3. Full Name
4. Date
5. Hour
6. Age
7. Inc_No_Int
8. Id_Community
9. Telephone
10. Id_Street
11. Domestic Violence
12. Alcoholic
13. Age1
14. Droug Consumption
15. Id_Colony
16. Inc_No_Ext
17. Domestic Violence_2
18. Municipal Code
19. Psychological Support
20. Dating violence
21. Diagnostic
22. Suicide threat
23. Id_Sex
24. Violence Against Women

Results of the application of the EBK statistical interpolation method

Next, it is shown the prediction made by the EBK method, where it is possible to observe the places where the EBK method predicts that another case of suicide could appear. For this case, 100 simulations with linear semivariograms are used.

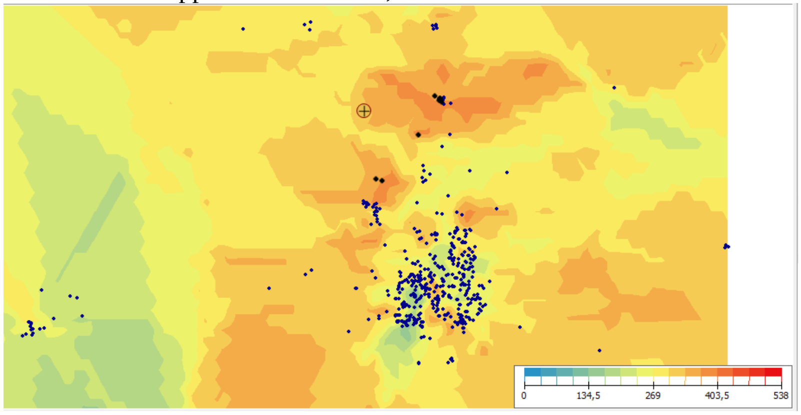

Fig.4. Prediction semivariogram using the EBK method.

Next, the figure corresponding to the probability is shown, where it is possible to observe those areas where the probability of another suicide occurring is higher, close to 0.9922, as well as the areas where the probability of a suicide occurring is very low, 0.007766.

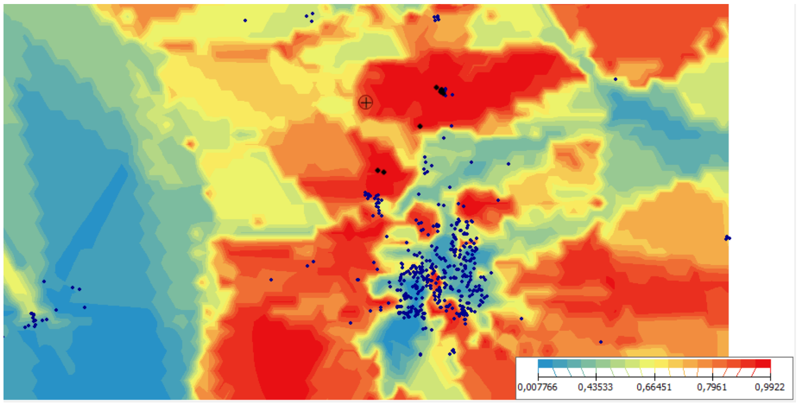

Fig. 6. Probability semivariogram using the EBK method.

Finally, the figure above shows the prediction of the standard error where the areas where the error is most present due to the absence of suicides and those areas where the error decreases due to the history of suicide can be identified.

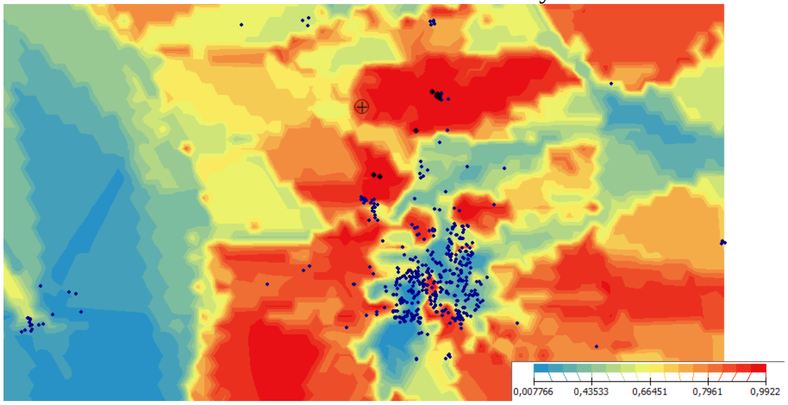

Fig. 7. Semivariogram of the standard error prediction using the EBK method.

Results Bayesian Classifier

For the classification stage, the classifier was carried out with two characteristic vectors. For the first of these, the example of a healthy person was taken, and in most characteristics, the result was negative.

e 1 (Non suicidal):

1. Id_Mode: 0
2. Id_Rol_Person: 2
3. Full Name: 1
4. Date: 41924
5. Hour: 870221678240341
6. Edad: 22
7. Inc_No_Int: 1
8. Id_Community: 1
9. Telephone: OMITIDO
10. Id_Street: 1967
11. Domestic Violence: 0
12. Alcoholic: 1
13. Age1: 0

e 2 (Suicidal):

1. Id_Mode: 1
2. Id_Rol_Person: 2
3. Full Name: 2
4. Date: 38703
5. Hour: 750299768515106
6. Edad: 35
7. Inc_No_Int: 340
8. Id_Community: 1
9. Telephone: OMITIDO
10. Id_Street: 334
11. Domestic Violence: 0
12. Alcoholic: 0
13. Age 1: 3

14. Droug Consumption: 0	14. Droug Consumption: 1
15. Id_Colony: 54	15. Id_Colony: 226
16. Inc_No_Ext: 0	16. Inc_No_Ext: 1007
17. Domestic Violence_2: 0	17. Domestic Violence_2: 0
18. Municipal Code: 8	18. Municipal Code: 3
19. Psychological Support: 0	19. Psychological Support: 0
20. Dating Violence: 0	20. Dating Violence: 0
21. Diagnóstic: 1	21. Diagnóstic: 2
22. Suicide threat: 0	22. Suicide threat: 0
23. Id_Sex: 1	23. Id_Sex: 0

Below are the results obtained after running the Bayesian Classifier in Matlab.

Example 1:

Test =[0,2,1,41924,880221678240341,22,1,1,OMITIDO,1967,0,1,0,0,54,0,0,8,0,0,1,0,1];

Result = '**NON SUICIDAL'**

Example 2:

Test =[1,2,2,38703,750299768515106,35,340,1,OMITIDO,334,0,0,3,1,226,1007,0,3,0,0,2,0,0];

Result = '**SUICIDAL**'

7 Conclusions

The conclusions, as the results section, is divided in order of execution. Initially, the database is pre-processed, which allows the elimination of trivial or unimportant information; after that, the geospatial analysis of the information is undertaken, and the classification stage is completed.

Attribute Selection

Each one of the algorithms of selection attributes contributes essential information that allows making an analysis when obtaining the average of each attribute, this way it is possible to visualize those attributes that have a higher power of discrimination; therefore, a subset of characteristics actively is generated for the class of "SUICIDE".

The use of different algorithms, provides an inherent advantage, since it addresses from different methods the selection of attributes, ensuring a subset of characteristics with the best quality of pre-processing, eliminating redundant attributes that do not affect the objective of the subject of study.

The clear visualization of the attributes with the most significant power of discrimination allows the determination of the number of attributes and on the other hand, those that should be ignored. Therefore, there is the possibility of giving continuity to the work of identifying patterns for potential cases of suicide.

Application of the EBK

The application of the EBK, for the analysis of each case with geospatial orientation, allows having a more exact approximation to the expected values since unlike other variants of the Kriging method, this one calculates multiple semivariograms, which evaluate the different zones differently. Because it does through subsets allowing the presence of spatial autocorrelation between the points, in other words, those points that are near, have similar characteristics and at the same time have different characteristics from those points that are far from them.

Then the suicidal cases, are translated into the characteristics of the environment for those eventualities that arise in different places of the state of Aguascalientes, which means that the problem that causes suicide, in the east of the city where the socioeconomic level of life is medium-low unlike the north of the city where the standard of living is medium-high, this situation makes the EBK method, the most suitable for this geospatial analysis.

Bayesian Classifier

The "Bayesian Classifier" is a very efficient and useful tool to carry out the identification of patterns in large databases and, in turn, perform a useful classification.

In the previous analysis, the effectiveness of the Bayesian Classifier is observed when performing the corresponding calculations with low computational cost obtaining and satisfactory results. Since they show a satisfactory classification by efficiently identifying potential suicide cases from those that are not, this method also demonstrates its effectiveness by being a small sample of suicides compared to the universe of data.

References

1. T. Aluja, “La Minería de Datos, entre la estadística y la Inteligencia Artificial,” vol. 25, pp. 479–498, 2001.
2. K. Sutha and J. . Tamilselvi, “A review of feature selection algorithms for data mining techniques,” Int. J. Comput. Sci. Eng., vol. 7, no. 6, pp. 63–67, 2015, doi: 10.1111/j.1532-849X.2011.00718.x.
3. ArcGIS ESRI, “Cómo funciona Krigin,” 2016. http://desktop.arcgis.com/es/arcmap/10.3/tools/3d-analyst-toolbox/how-kriging-works.htm#ESRI_SECTION1_E112B7FAED26453D8DA4B9AEC3E4E9BF (accessed Dec. 07, 2019).
4. K. Javed, H. A. Babri, and M. Saeed, “Feature selection based on class-dependent densities for high-dimensional binary data,” IEEE Trans. Knowl. Data Eng., vol. 24, no. 3, pp. 465–477, 2012, doi: 10.1109/TKDE.2010.263.
5. R. Ruiz, J. Riquelme, and J. Aguilar-Ruíz, “Heuristic Search over a Ranking for Feature Selection,” Dep. Comput. Sci. Univ. Seville, Spain, no. May, p. 9, 2016, doi: 10.1007/11494669.
6. W. Tang and K. Z. Mao, “Feature selection algorithm for mixed data with both nominal and continuous features,” Pattern Recognit. Lett., vol. 28, no. 5, pp. 563–571, 2007, doi: 10.1016/j.patrec.2006.10.008.
7. S. Morales et al., “Acute mental discomfort associated with suicide behavior in a clinical sample of patients with affective disorders: Ascertaining critical variables using artificial intelligence tools,” Front. Psychiatry, vol. 8, no. FEB, pp. 1–16, 2017, doi: 10.3389/fpsyt.2017.00007.
8. L. E. C. Soncco, “Estudio comparativo de redes neuronales artificiales aplicadas a la identificación de violencia escolar en las instituciones educativas.,” p. 148, 2018.
9. L. F. M. Quiñones and G. T. C. Tarazona, “Aplicación de redes neuronales artificiales sobres la violencia de la mujer por su pareja según la encuenta demográfica y de salud familiar, endes 2016,” Universidad Nacional de Ancash “Santiago Antúnez De Mayolo,” 2018.
10. O. D. Castrillón, W. Sarache, and E. Castaño, “Sistema bayesiano para la predicción de la diabetes,” Inf. Tecnol., vol. 28, no. 6, pp. 161–168, 2017, doi: 10.4067/S0718-07642017000600017.
11. M. A. G. Gómez, “Diseño y Construcción de un Modelo de Predicción de Depresión en Adultos Mayores en Chile,” Universidad de Chile, 2019.
12. Secretariado Ejecutivo del Sistema Nacional de Seguridad Pública, “Incidencia Delictiva del Fuero Federal,” Gob.Mx, 2019, [Online]. Available: https://www.gob.mx/sesnsp/acciones-y-programas/incidencia-delictiva-del-fuero-federal.
13. K. P. Singh, N. Basant, and S. Gupta, “Support vector machines in water quality management,” Anal. Chim. Acta, vol. 703, no. 2, pp. 152–162, 2011, doi: 10.1016/j.aca.2011.07.027.
14. K. K. Seth, “Charmonium - In and out of the nucleus,” Nucl. Phys. A, vol. 629, no. 1–2, pp. 358–365, 1998, doi: 10.1016/S0375-9474(97)00711-2.
15. K. Torkkola, “Feature extraction by non-parametric mutual information maximization,” J. Mach. Learn. Res., vol. 3, pp. 1415–1438, 2003.
16. B. B. P. Bautista, “Método de Selección de Atributos por Clase,” p. 110, 2009.
17. F. Carrat and A. J. Valleron, “Epidemiologic mapping using the ‘kriging’ method: Application to an influenza-like epidemic in France,” Am. J. Epidemiol., vol. 135, no. 11, pp. 1293–1300, 1992, doi: 10.1093/oxfordjournals.aje.a116236.
18. K. Krivoruchko, “Empirical Bayesian Kriging Implemented in ArcGIS Geostatistical Analyst,” Shanghai Jiaotong Daxue Xuebao/Journal Shanghai Jiaotong Univ., vol. 43, no. 11, pp. 1813–1817, 2009.
19. ArcGis Desktop, “¿Qué es un Kriging bayesiano empírico?,” 2016. https://desktop.arcgis.com/es/arcmap/10.4/extensions/geostatistical-analyst/what-is-empirical-bayesian-kriging-.htm (accessed Jan. 23, 2020).
20. V. P. Samsonova, Y. N. Blagoveshchenskii, and Y. L. Meshalkina, “Use of empirical Bayesian kriging for revealing heterogeneities in the distribution of organic carbon on agricultural lands,” Eurasian Soil Sci., vol. 50, no. 3, pp. 305–311, 2017, doi: 10.1134/S1064229317030103.
21. W. Tae Kim and Y. J. Park, “Batalin-Tyutin quantization of the (2+1) dimensional nonabelian Chern-Simons field theory,” Phys. Lett. B, vol. 336, no. 3–4, pp. 376–380, 1994, doi: 10.1016/0370-2693(94)90548-7.
22. R. R. Bouckaert, E. Frank, R. Kirkby, P. Reutemann, A. Seewald, and D. Scuse, “WEKA Manual for Version 3-7-2,” p. 343, 2018, [Online]. Available: http://netcologne.dl.sourceforge.net/project/weka/documentation/3.7.x/WekaManual-3-7-12.pdf.