Suicide prediction in workers using Naive Bayes

Daniel Alejandro Barajas Aranda [1][0000-0002-8220-3877] and María Dolores Torres Soto [1][0000-0002-7245-1076] and Aurora Torres Soto[1][0000-0002-2930-824X]

[1] Universidad Autónoma de Aguascalientes, Aguascalientes, México
alengot@hotmail.com

Abstract. The suicidal tendency is a noticeably big problem in Aguascalientes, Mexico, especially in the young and working population, which constitutes more than 50% of completed suicides. This brings with it a great affectation to the economy of the state, not only when the workers manage to commit suicide, but also when it is only attempted, since their work responsibilities are stopped. In this study, a database compiled by the psychology department of the Autonomous University of Aguascalientes with factors associated with mood is analyzed, in which we can find features associated with sleep disturbance, self-esteem problems, and even affectations in the weight. This database contains information on people with suicidal tendencies, and a control group. To identify the key features that suicidal people present, the total set of typical testors was obtained and the informational weight of each feature was obtained. In the same way, a predictor was made using a classifier based on the naive bayes theory, analyzing its effectiveness with the total set of features, and using only the best features, with an informational weight greater than 40%.

Keywords: Suicide, Typical testors, Prediction.

1 Introduction

In Aguascalientes the suicide rate is very high. Since 2012 there were more than 100 suicides per year, and this situation is increasing as years go by [1]. In 2020, a historical value of 184 suicides was obtained. On the other hand, for every completed suicide there are at least 10 suicide attempts [2].

Among the most serious consequences is the social impact, especially related to the economy, since a large number of suicides are performing some work activity, and are at a highly productive age [1].

As can be seen in figure 1, the ages of suicides are especially concentrated in the range of 20 and 24 years, with a decrease in cases as age increases.

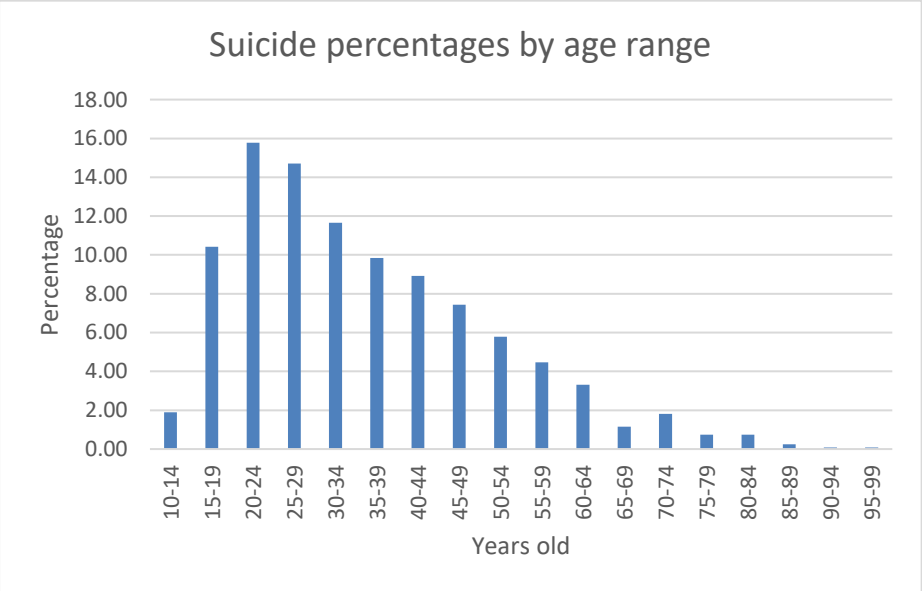

Fig. 1 Suicide percentages by age range [1].

On the other hand, as can be seen in figure 2 only 29.31% of the people who commit suicide are not working, thus causing a great problem by destabilizing the economy of the state, and the health of the companies since 70.69 % of suicidal are engaged in some economic activity

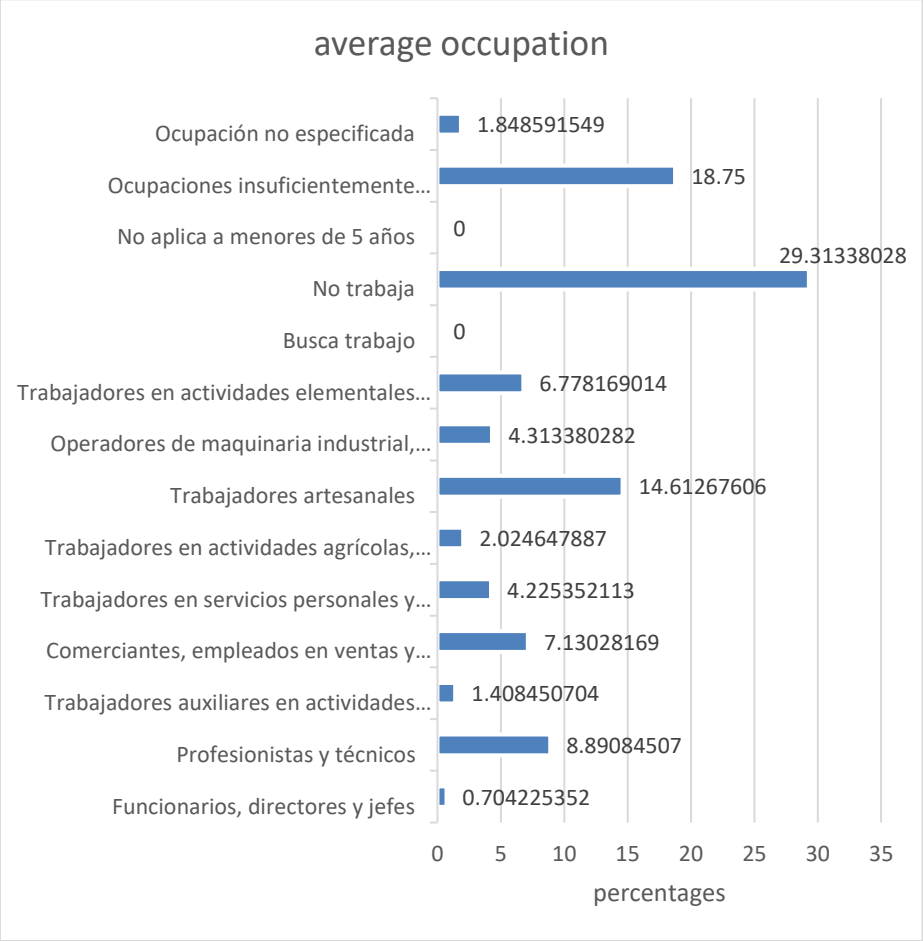

Fig. 2. Percentages in activities of suicidal people [1].

Using a computational tool based on CUDA [3], the set of all typical testors associated with a database with information on suicidal people and a control group have been identified. This database was compiled by the psychology department of the Autonomous University of Aguascalientes and contains information about mood and stress level.

1.1 Suicide

Suicide, as its definition indicates, is the act by which a person voluntarily ends his own existence [4]. In other words, this person presents the loss of neuroencephalic or cardiorespiratory constants in a definitive and irreversible way [2].

In this study, the features associated with the state of mind and the sociocultural environment experienced by the people surveyed are grouped together, features that according to Barajas [5] fit within the factors by which people commit suicide.

1.2 Combinatorial logic approach

Testor theory was formulated in the mid-1950s in the former Soviet Union of Socialist Republics (USSR) as one of the independent scientific directions of mathematical cybernetics [6]. Testors were first used to find faults in electrical circuits.

A testor is a set of features capable of distinguishing among classes, because no object of a certain class can be confused with one of another class [7]. And a typical testor is one that contains the minimum number of features, so losing one of them, implies that this group stops being a testor.

The informational weight tells us the number of times a feature appears in the total set of typical testors. So, this percentage gives us the level of importance of each feature.

1.3 Naive Bayes

Bayes' theorem expresses the conditional probability of a random event A given B in terms of the conditional probability distribution of event B given A and the marginal probability distribution of just A.

Where $\{A_1, A_2, \ldots, A_n\}$ is a set of mutually exclusive and exhaustive events, such that the probability of each of them is different from zero. Let B be any event for which the conditional probabilities $P(B|A_i)$ are known. Then the probability of $P(A_i|B)$ is given by the expression 1 [8]:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} \quad (1)$$

Where:
$P(A_i)$ are the prior probabilities
$P(B|A_i)$ is the probability of B in hypothesis A
$P(A_i|B)$ are the posterior probabilities

2 Methodology

To conduct this article, a methodology that combines obtaining typical testors and a Bayesian classifier was followed. This methodology is shows in Figure 3 and described below.

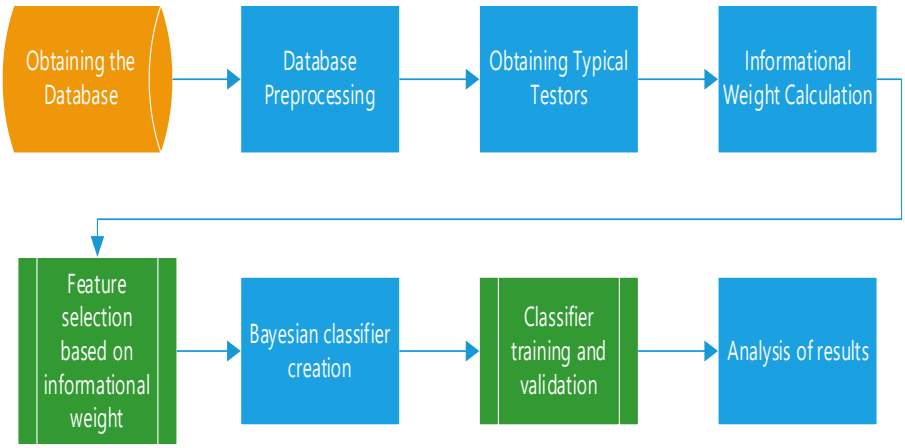

Fig. 3. Methodology

As a first step, a database was obtained, this consists of forty-one features that show the state of mind of a person, as well as their overcrowding, age, and occupation.

After this, the database was cleaned, eliminating incomplete and/or duplicate records, leaving a total of seventy suicidal people and 252 belonging to the control group.

Using a tool based on CUDA [3], the total set of typical testors associated with the database were obtained, and after this the variables whose informational weight exceeded 40% were selected.

A Bayesian classifier was created using the full set of features, training 80% of the data, and then evaluating it with the remaining 20%. This same step was conducted using only the features with an informational weight greater than 40%.

3 Results

As a result of the typical testors calculation, a total of 33,125 testors were obtained. Then the informational weight was obtained, resulting in nine features with an informational weight greater than 40%, which can be seen in Table 1.

Table 1. Informational weight

Feature		Informational weight
CD5	slept without rest	47.443
CD7	could not move on	41.221
CD8	nothing made me happy	41.387
CD9	felt that I was a bad person	73.966
CD11	slept more than usual	44.587
CD14	felt like I was dead	44.738
CD15	wanted to hurt myself	40.702
CD17	was disgusted with myself	42.942
CD18	lost weight without trying	40.744

A Bayesian classifier was created and feed with the set of all the features and another using only the features with an informational weight greater than 40%.

For the first case with all the features of the matrix, an effectiveness of 0.71 was obtained. The confusion matrix obtained is shown in table 2.

Table 2. Confusion matrix all features

confusion matrix			
predicted class	0	27	29
	1	22	99
		0	1
		true class	

The second experiment was conducted using only the features with informational weight greater than 40%. This times an effectiveness of 0.76 was reached. Results are shown in the confusion matrix in Table 3.

Table 3. Confusion matrix features with 40% informational weight

confusion matrix			
predicted class	0	25	19
	1	24	109
		0	1
		true class	

4 Conclusion

In the work it is possible to appreciate that the main features that influence the decision to commit suicide are those related to the sleep cycle (sleeping more than usual, and sleeping without resting) and weight loss without trying, aspects that are closely related to suicidal thoughts

The features of being a bad person has an extremely high informational weight. Which indicates that people with suicidal tendencies experience problems in self-assessment

Similarly, the use of typical testors as a mechanism for reducing features is particularly useful, especially when creating classifiers. In terms of effectiveness, it can be seen that with only a part of the features, the same or even better results are obtained, so it is possible to eliminate redundant features and only keep those with a high informational weight.

As future work, a greater number of samples will be collected, in order to train the classifier more exhaustively, and the creation of a multilayer neural network will be chosen to improve precision.

References

[1] Instituto Nacional de Estadística y Geografía, "Estadísticas Vitales Defunciones generales," 2022.

[2] N. Campos, "Diplomado en el Protocolo de Actuación (PROL-SMDIFAGS-SUIC/2016)." 2016.

[3] D. A. Barajas Aranda, A. Torres Soto, and M. D. Torres Soto, "DBT an Algorithm Based on CUDA for Reducing the Time to Obtain Typical Testors," *Res. Comput. Sci.*, vol. 150 (9), 2021.

[4] "Definición de suicidio - Diccionario del español jurídico - RAE." [Online]. Available: https://dej.rae.es/lema/suicidio. [Accessed: 26-Nov-2019].

[5] D. Barajas, "Identificación de Factores de Riesgo determinantes en el suicidio en Aguascalientes mediante la técnica de Testores Típicos," Universidad Autonoma de Aguascalites, Aguascalientes Mexico, 2017.

[6] A. N. Dmitriev, Y. I. Zhuravlev, and F. P. Krendeleiev, "On mathematical principles of object and phenomena classification, Discrete Analysis," pp. 3–15, 1966.

[7] J. R. Shulcloper, A. Guzmán, and J. F. Martínez, *Enfoque Lógico Combinatorio al Reconocimiento de Patrones*. Mexico: Instituto Politécnico Nacional, 1999.

[8] C. M. Bishop, *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.