Statistical Analysis of Factors Related to Suicide Records in the World Between 1985 and 2020

João Pedro Barbosa de Almeida[1*], Matheus Carvalho Nascimento de Souza[1], Yuri Papaterra Fonseca[1], Felipe Emmanouil Martires Stamoglou[1], Márcio Renê Brandão Soussa[1]

[1]Senai Cimatec University Center, Computing Department; Salvador, Bahia, Brazil

Suicide is considered by the World Health Organization (WHO) as a public health problem that afflicts the whole society and it counts more deaths than many diseases. Therefore, the current objective is to analyze the suicide deaths between 1985 and 2020 and classify them according to the act. Through KNN, this study presents suicide cases grouped by sex (male and female) and associate them to factors by country, pointing out information that allows us to understand what influences the act the most directly or indirectly. The results showed that, unlike other researches, the rate of suicide does not have differences based on sex. However, further studies are needed.
Keywords: Suicide. Victim. Data. Influences.

Introduction

Suicide is a voluntary act that aims to end one's life and is considered an act of violence [1], affecting families, communities, countries, leaving deep marks on those who stay [2]. Therefore, the World Health Organization (WHO) considers suiciding a public health problem [3].

The WHO [3] reports that suicide remains one of the leading causes of death in the world (one in a hundred) and that annually counts more deaths than diseases such as HIV, malaria, breast cancer, as well as homicides and wars. In 2019 alone, more than 700,000 people committed suicide worldwide (1 case every 40 seconds) [4], and many others tried but failed. According to the Secretaria de Saúde da Bahia [4], in Brazil, about 12,000 people take their own lives each year, corresponding to approximately 6% of the population. There is one case every 46 minutes, being more recurrent in the male black population aged between 10 and 29 years. It is estimated that more than 90% of suicide cases are related to mental illnesses, being depression first, followed by bipolar disorder and drug abuse.

Received on 10 September 2021; revised 27 October 2021.
Address for correspondence: João Pedro Barbosa de Almeida. Avenida Orlando Gomes, 1845, Piatã, Zip Code: 41650-010, Salvador, BA, Brazil. Phone: +55 71 3879-5677. E-mail: jpbarbosinha@gmail.com.

J Bioeng. Tech. Appl. Health 2021;4(4):152-156.

Given the pandemic scenario due to COVID-19, the need to stay at home increased the rate of suicide worldwide. According to OPAS [5], the anguish, anxiety, and depression of isolation, combined with cases of violence, alcohol consumption disorders, substance abuse, and feelings of loss, can become factors that may increase the chances of a person taking their own life. With the severity of the scenario, several measures have been used to prevent the consummation of the act, such as the creation and operation of social assistance centers, like Psychosocial Care Centers (CAPS) and the Centers for The Valorization of Life (CVV). The WHO [4] states that suicide is preventable in 90% of cases, and towards that goal, there is also an awareness campaign established for suicide prevention, the month of September, also viewed as the month of suicide prevention. Initiatives that make use of technologies have also played an important role in suicide prevention, in example "the algorithm of life", developed by SAP and Amazon Web Services, which monitors tweets seeking to identify profiles at high risk of depression, forwarding the messages to specialized centers [6].

Seunghyong Ryu and colleagues [7] propose a tool that makes use of the random-forest machine learning algorithm, a computational model for predicting suicidal ideation in individuals through Machine Learning. The study aimed to develop a tool capable of perceiving how likely individuals are to proceed with the act, based on samples of ideals. In its completion, it was concluded that,

although its accuracy is acceptable, working on some factors is necessary to increase it throughout the technique.

Taking into account the previous study, another was also carried out involving the n-gram linear regression method and later random-forest machine learning within the Facebook social network looking for texts or other types of publications possibly related to the act of suicide[8]. These types of studies show that, through social media and its resources, it could be possible to prevent the suicide act from simple messages and posts interchange.

Therefore, this research aims at predicting and correlating factors involved in suicides to understand which of them most affect and influence the execution of the action, when those factors are seen from a scope.

Materials and Methods

In this work, a knowledge discovery in databases (KDD) process was carried out, containing the following steps: Database Definition, Pre-processing, Data Mining, and Data Analysis. They are detailed below.

Database Definition

We used two public databases extracted from the World Health Organization (WHO) [9], one (deaths database) containing records of deaths worldwide during the period from 1985 to 2020, and another (population database) containing the total population of each country during the same period. The first consists of 5 files, totaling 4.209.751 records, and 5 attributes the number of deaths grouped by age group, country, year, sex, population, and cause of death. The second consists of a single file grouped by mid-year population and live births, with 9.719 records.

Pre-Processing

The first step was to change the identification of the country attribute from numbers to their actual names. The database is formatted with only an ID of each country, and that has to be referenced with the list of names given by the same organization. As the objective of this project was to analyze a correlation between possible factors that lead to suicide, it was necessary to filter the deaths database. To proceed, a filtering of the column "cause" that represents the causes of death, was done, selecting only those whose codes are indicative of suicides and self-inflicted injuries, thus ensuring to work only with records related to the focus of the study, suicide. Given value of total deaths and populations were absolute, it was necessary to normalize them, so we used the MinMaxScaler method of Python 3's SKLearn module. Following that, the population database was joined with the deaths database generating a single dataset that was used in the data mining and data analysis steps.

Data Mining

We did a comparison between the columns of the dataset to check if the numbers of suicides and total populations have any perceivable relation. Then, to apply the K Nearest Neighbours (KNN) method, a range of K-values were analyzed based on the accuracy with which they could better fit the classification method. This research followed a branch of studies, namely CVV and OPAS, which reported cases of suicides being more present in men than women, justifying the main factors used for the classification: sex, number of suicides, and population. Therefore, we analyzed the results by the K-Nearest Neighbors (KNN) method, in which sex was used as the target of the classification. Thus it could be demonstrated if the cases would present a uniformity based on the sex of the victim.

Data Analysis

We used the Python 3 language to perform the data analysis with the libraries: numpy, pandas, seaborn, and matplotlib. The results were plotted

in the form of a Heatmap to allow visualization of the relations between each data. Then, after the best value for the KNN classification was found in a graph showing the accuracy, the classification was executed by itself and its decision regions were plotted, so their homogeneity could be analyzed.

Results and Discussion

Heatmap

The heatmap produced presents the correlation between the values of total deaths and total populations in each country and the years they were noted (Figure 1). The lighter points represent high correlation, and the darker ones represent the low correlation. Each label is separated with two hidden attributes between them.

The correlation showed no relevant data to the study, having the values of the deaths correlate with themselves. It, however, demonstrates that closer age groups have a higher correlation with themselves, except for the age groups in the extremities, the younger and older ones. It also shows in the darker areas a lesser correlation of the younger age groups to suicide rates when compared to the older age groups, as children are usually less connected with deaths by suicide and its causes.

The K Nearest Neighbour Method (KNN)

Cross-Validation

For the realization of KNN, the best value of K Nearest Neighbors was verified through cross-validation, using a graph for a more accurate demonstration of the value range chosen. We use another module called SciKit Learn, also from Python 3, to do the process.

The accuracy was measured between the values 1 through 30. The value of 2 was the most accurate for the classification (Figure 2). No further testing was needed as accuracy only decreased as the value increased.

Classification

The data was then formatted using Python's PCA module to fit the format of the KNN method, implementing it with the values 0 for the missing data, 1 for men, and 2 for women. With all the requirements mentioned, the decision regions were plotted to analyze the homogeneity of the data as shown in Figure 3.

In the plotting regions, a low homogeneity was demonstrated for both sexes, with the female showing to be the most concrete. About the regions, none seems to be consistent enough to classify the data, and multiple outliers are present across the regions, which is in the jagged borders of the regions and the invading contrasting areas. It is also worth noting that the value used for K is relatively small, both decreasing the bias of the classification and increasing its variation, which could explain the outliers.

Conclusion

After observing the results of the KNN methodology in standardized data, we concluded that there are no significant clusters among the evaluated data, demonstrating a lack of similarity in the characteristics presented, hindering the discovery of factors related to the rate of incidents. However, we verified that female victim does not differ much from males, except for a few outliers, contradicting a few existing studies. Although it can be theorized that there is a correlation between the size of the population and the number of suicides, as seen at the beginning of the study, subsequent checks through the methods bring contrary information, especially at years when the rate of calculated suicides does not differ much from country to country. With the data restructured, the plot created did not have enough information to answer what factors lead to the person committing the act but opened the possibility for new studies that seek different factors from those used here, which can determine what leads someone to exceed the limit

Figure 1. Heatmap of the correlation between the data.

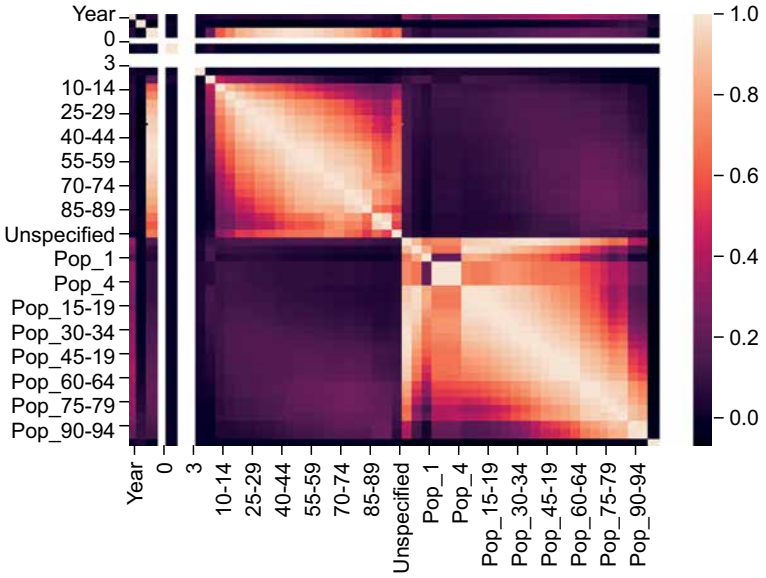


Figure 2. Graph of the values analyzed and their accuracy.

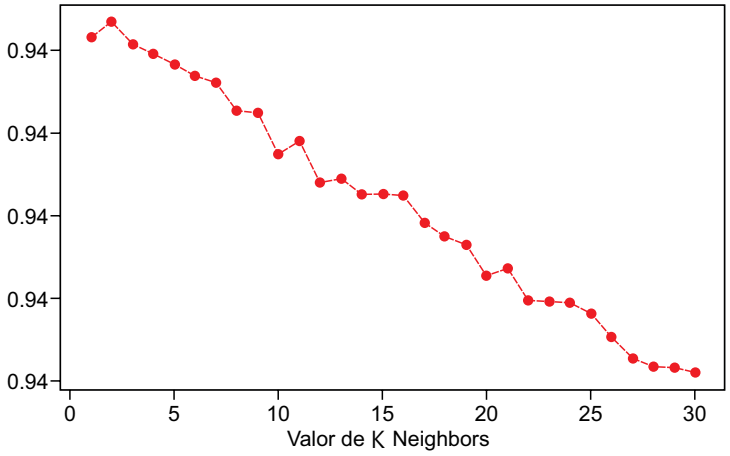
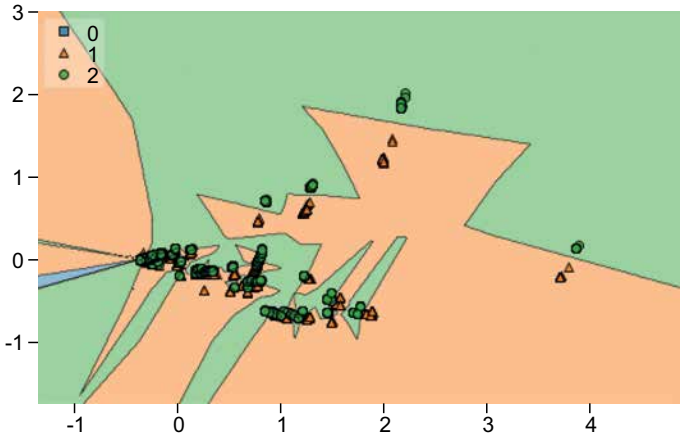


Figure 3. KNN decision-making regions carried out.



of "thinking" about suicide to "act", to point out new ways to reduce or prevent the in point of suicides in countries.

References

1 Ribeiro NM, et al. Análise da tendência temporal do suicídio e de sistemas de informações em saúde em relação às tentativas de suicídio. Texto & Contexto-Enfermagem 2018;27.

2. WHO - World Health Organization. Notícias. Available at: <https://www.who.int/news/item/17-06-2021-one-in-100-deaths-is-by-suicide>. Accessed Aug. 4, 2021.
3. WHO - World Health Organization. Departamento de Saúde Mental e de Abuso de Substâncias. 2006, 18p. Available at: <https://www.who.int/mental_health/media/counsellors_portuguese.pdf>. Accessed: Aug. 4, 2021.
4. Governo do Estado, Secretária de Saúde. Notícias. Available at: <http://www.saude.ba.gov.br/2020/09/10/oms-alerta-suicidio-e-a-3a-causa-de-mortede- jovens-brasileiros-entre-15-e-29-anos>. Accessed Aug 4, 2021.
5. OPAS - Organização Pan-Americana de Saúde. Notícias. Available at: <https://www.paho.org/pt/noticias/10-9-2020-pandemia-covid-19-aumenta-fatores-risco-para-suicidio>. Accessed Jul. 16, 2021.
6. SAP Brasil, Tecnologia como ferramenta de prevenção ao suicidio. 2019, Available at:<https://news.sap.com/brazil/2019/10/tecnologia-como-ferramenta-de-prevencao-aosuicidio-bl0g/>. Accessed Aug 4, 2021.
7. Ryu S, Lee H, Lee D, Park K. Use of a machine learning algorithm to predict individuals with suicide ideation in the general population. National Center for Biotechnology Information 2018. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6258996/>. Accessed Aug 4, 2021.
8. Andrade NNG, et al. Ética e inteligência artificial: prevenção do suicídio no Facebook. Filosofia e Tecnologia 2018;(31)4:669-684.<https://link.springer.com/content/pdf/10.1007/s13347-018-0336-0.pdf>. Accessed Aug 22, 2021.
9 WHO - World Health Organization. WHO Mortality Database. Available at: <https://www.who.int/data/data-collection-tools/who-mortality-database>. Accessed Jul. 9, 2021.