Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

A Machine Learning Approach to Analyze and Predict Suicide Attempts

Mrs. B Ida Seraphim
Assistant Professor,
Dept. of Computer Science & Engineering
SRM Institute of Science & Technology, Kattankulathur
Chennai, India-603203

Subroto Das
4th year,
B.Tech Computer Science & Engineering Student
SRM Institute of Science & Technology, Kattankulathur
Chennai, India-603203

Apoorv Ranjan
4th year,
B.Tech Computer Science & Engineering student
SRM Institute of Science & Technology, Kattankulathur
Chennai, India-603203

***Abstract*— The issue of suicides has become increasingly worrisome and has received much attention in the today's society. Depression is thought of as the most common factor for suicides. However, there can be several other causes such as economic cause (unemployment), social cause (dowry dispute) , un-curable diseases (AIDs) etc. AI based chat-bots have been developed to prevent people from committing suicides but the accuracy is close to only 75%.To prevent the rate of suicide in future we use machine learning algorithms for successful prediction of suicide attempts with significant precision. Preliminary data analysis would give us insights about suicide statistics and the correlation between the various factors and the extent to which they contribute. Graphical representation is provided to understand the trends in suicide attempts. This paper presents different existing techniques that are used for developing suicide prediction models. We take a closer look at the pros and cons of these techniques. A comparative study is made on the effectiveness of these algorithms.**

***Keywords*— Suicide, machine learning, data analysis**

I. INTRODUCTION

Suicide is increasingly becoming a serious concern for the society. Close to 800 000 people die due to suicide every year, which is one person every 40 seconds. Suicide is a global phenomenon and occurs throughout the lifespan. There are indications that for each adult who died by suicide there may have been more than 20 others attempting suicide. Suicide is a global phenomenon; in fact, 79% of suicides occurred in low- and middle-income countries in 2016. Suicide accounted for 1.4% of all deaths worldwide, making it the 18th leading cause of death in 2016

Today, people experience severe physical disorders and psychological stress due to a variety of internal and external factors. Although depression is mainly found in people in their 30s and 40s, it is often detected in juveniles due to academic stress and interpersonal relationship and in elderly persons. Since there is a socially negative view on persons who suffer a mental illness, such patients often hide their illness. Economic conditions, drug and alcohol abuse also contribute to self - harm and suicide attempts. In India there are various social causes as well that promote suicide attempts, most common being dowry disputes.

Data Analysis is carried out, in order to classify data so as to provide a set of preventive measures to control them in future. This can provide information about the cause of suicide taken in a particular state followed by in a particular year. The dataset can also provide information about whether the suicide rate for a particular cause has increased or not. Analysis and Classification will not only provide preventive measures but will also led to comparison that will provide information whether suicide rate has increased or decreased in several years. Our aim is to find a machine learning model for the prediction of suicide attempts. Logistic regression, Decision trees , Gradient Boosted Decision Trees, Support Vector Machines and Artificial Neural Networks are some of the models used. A comparative study is made on the effectiveness of these algorithms.

II. STATE OF THE ART (LITERATURE SURVEY)

Gen Men Lin et al. [1] worked on suicide ideation prediction for military personal as they are more vulnerable to psychological stress[2] because of physical training, multiple deployments and responsibilities. They have utilized a large sample of the military members for several machine learning techniques by taking the psychological stress dimensions (BSRS - 5)into consideration to predict the presence of suicide ideation. A binary probabilistic classifier of machine learning algorithm can determine whether the military persons, through their questionnaires, have suicide ideations. The six input factors of psychological stress for their machine learning model include BSRS-5 score, anxiety, depression, hostility, interpersonal sensitivity and insomnia. They have used six machine learning techniques including logistic regression (LR), decision tree (DT), random forest (RF), gradient boosting decision tree (GBDT), support vector machine (SVM) and multilayer perceptron (MLP) for the prediction of the presence of suicide ideation of the military members.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

To solve the phenomenon of different dynamic ranges for the six input variables, they have applied the normalization of Min-Max scaling to normalize input data into the interval 0~1. They have utilised 10 fold cross validation in their model. The minimum accuracy is 98.4% for decision tree and the highest is 100% for Support Vector Machine and Multilayer Perceptron. Coming to some of the shortcomings, this paper is focussed mainly on military personnel and all of them are aged between 20 and 40 with an average age of 40 years. This caters to only a small demographic. Moreover the complexity is high when we factor in the 10 fold cross validation with that of multilayer perceptron and Adam optimizer.

Ji-Won Baek and Kyungyong Chung[3] have used a Context Deep Neural Network Model for predicting depression risk using multiple regression. They have used a Feed-Forward Network for the model and used backpropagation for model optimization. For the extraction of data features, data is entered in DNN, and then data features are extracted and displayed in the hidden layer. Data output is classified. They have used the raw data offered by the Korea National Health and Nutrition Examination Survey in their research. After pre-processing the data ,they have performed multiple regression analysis and selected context variables that are most related to depression. They have selected depression as the target variable. Lastly they perform prediction of depression. Their model uses Adam as an optimization function, relu as an activation function, and Mean Absolute Error as a loss function in hidden layers. The DNN model had an accuracy of 85.46% and the context- DNN model yielded an accuracy of 94.57%.

S.A.S.A. Kulasinghe et al. [4] have proposed a model called "AISA: Artificial Intelligence based Suicide Avert" whose main objective is to provide the facility for users to express their feelings and emotions freely in private and confidential environment, to an artificial companion(chatbot) who listens and monitors their behaviour through social media as well. They have discussed about the differences between extroverts and introverts' mental health issues and that there is a positive relationship between introversion and depression. They have implemented a working generative Chat-bot to cater the user's needs; to become a friendly companion to talk anything with. Their Chat-bot is using a Seq2Seq (Sequence-to-Sequence) model. They perform sentiment analysis on the user's social media feeds like Facebook posts and web browser history. They also perform voice analysis to understand the emotional state of the user. For the data collection, the data from Facebook will be tokenized and they have selected those tokens which are more frequent and use those tokens for the analyzing process. This process follows LDA (Latent Dirichlet Allocation) method in topic modeling to determine the tokens which will be used. After each session with the user, they send the conversations to the NLP(Natural Language Process) module for opinion mining. Based on this, the bot changes its conversational style to match that of the user's emotional state. The model achieved an accuracy of 75%. Some of the cons include a very small dataset (53 rows), talking with a "machine" may further alienate some individuals and increase their risk of attempting a suicide. Also, depression is not the only factor that forces an individual to commit suicide.

Jun Shen et al. [5] have used a linear regression model for the prediction of suicide. Their model uses past data from the year 1985 to 2016 provided by WHO[6]. The Human Development Index (HDI) [7] is missing from their research but the variation of HDI is extremely subtle so they have used an algorithm for the missing values and used the same value from the year 1985 to 2016. They have performed data analysis which reveals many underlying trends such as the male suicide rate is higher than that of women, areas with extremely low and high GDP have higher suicide rates, housing prices have an inverse relationship with suicides. They have also used Sharpio test[8], Wald Wolfowitz runs test[9] and Breusch- Pagan test[10]. All three of them indicate that the linear model is invalid. Then they used a generalized addictive model (GAM)[11] which can fit an unknown parametric form. Their model yields an R squared value of 0.7054 and GCV of 3.71 .

Tarun Agarwal et al. [12] made a comparison among 3 machine learning algorithms: - logistic regression, random forest, and Naïve-Bayes for suicide prediction. They made use of the "Forever Alone" dataset available on kaggle. In their analysis they found that of the total people who attempt suicide, 77.65% are the ones having 0-5 friends. Hence this attribute plays an important factor in the prediction and hence was taken into consideration. This trend has not been found before. Also of total people who attempt suicide, 87.06% are from the age group of 15-30. Coming to their models, Logistic Regression yielded 94.68% accurate predictions. Naïve Bayes yielded 64.89% accurate predictions. Random Forest resulted in 92.55 accurate predictions. The accuracy of Naives bayes was very less and that was attributed to the class distribution and small size of dataset.

Pragya Prashar et al.[13] carried out analysis to find major cause of suicide. They found that some of the causes of suicide are economic causes like unemploymet or bankruptcy, social causes like dowry disputes and education failures among others. Their Data Set on which the analysis is carried out consists of parameters listed below:- State, Year, Type(Cause), Gender, Age Group which would help to provide preventive measure for suicide to an individual belonging to certain age group, of a particular state or of a particular gender. Collection of Data Set is done through various websites such as data.gov.in, Kaggle. They have then cleaned the dataset and used three machine learning algorithms- Linear regression , Naive bayes and Decision trees. The accuracy of their Naive Bayes model is only 64.28%.

Shaoxiong Ji et al.[14] have used deep learning techniques such as CNN and RNN as well as content analysis and feature engineering in their work. They have applied keyword filtering to the user's social media posts to gain a better understanding of his/her mental health. They have also used a "reach-out" model in which a mental

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

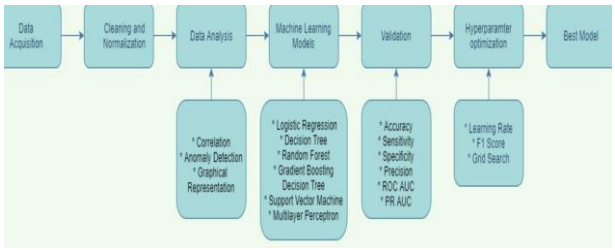
health organization is automatically informed if the person has high probability of committing suicides. Some of the cons include data deficiency and lack of intention understanding.

Data from social networking sites especially twitter has been extensively considered for research to automate the process of suicide prediction by using various machine learning and text mining techniques. Apart from the social media analysis, Ranjitha Korrapati et al.[15] also studied socio-economic and cultural factors. They found high correlation between suicide and maritial status and suicide and educational status. They found that among all the algorithms used, SVM classifier got the best performance with F1 68.3%, Precision 78.9%, and Recall 60.3%.

Tanupriya Choudhury et al.[16] made a comparative study on three machine learning algorithms- linear regression, decision tree, and naive bayes. For validation they have used precision and recall. Some of the cons of include low accuracy of Naive bayes(64.28%) and a small dataset.

Rifat Zahan et al.[17] used machine learning to identify a model to classify suicide and non-suicide related deaths from DNA methylation data. The two popular methods used for dimensionality reduction are: Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t- SNE). The output from that can be used as input for Support Vector Machine (SVM). They found that while reducing high-dimensional data, it is better to use t-SNE as an input for SVM rather than PCA. Some of the shortcoming of this paper include data deficiency (only 40 rows in the dataset), overfitting and spectrum effect.

III. PRPOSED WORK



1. DATA ACQUISITION

We are compiling our dataset from different sources including data.world, kaggle , data.gov.in etc.

2. CLEANING AND NORMALIZATION

After acquiring the dataset, we will remove null values or redundant rows and perform min-max normalization or standardization

3. DATA ANALYSIS

In this step, we will use data mining techniques to uncover hidden trends in the dataset and find the correlation between the variables, plot different graphs state-wise to uncover trends in suicide rates, and find the different factors behind suicides.

4. MACHINE LEARNING MODELS

Now, we will train our dataset with various machine learning models and use validation techniques to see the overall fit. Hyperparameter optimization would be done to ensure optimal accuracy. Finally, we will report the best model for the prediction of suicide.

IV. IMPLEMENTATION

There are 26 columns in our dataset. Some of the columns are numerical types which include GDP per capita, HDI for year, number of suicides, while others like country, age, sex, generation etc. are categorical. Using various Machine Learning algorithms, we wanted to investigate the potential triggers that could increase the risk of suicides in the societies. The Kaggle dataset we collected includes data from over 100 countries from 1985 to 2016. To make the study more insightful, we agreed to reduce the number of nations. We selected 40 countries from around the world that we believe are good sample of the various regions.

The main component of our dataset now that we have it is selecting the features that have the most differentiating effect. This is essentially feature engineering for training accurate prediction. Collinearity refers to the fact that certain features are extremely similar to an output class. We will use exploratory data analysis to understand the provided features before selecting the main features.

A. Exploratory analysis for feature selection

We started by making a bar plot that shows the number of suicides per 100,000 people. It also displays the distribution of the different age groups as well as the individual's gender. One trend that stands out is that the ratio of males to females who commit suicide is considerably higher. The age range 35-54 years has the highest ratio. The age range 35-54 years is the most likely to commit suicide. Another fascinating finding is that women over the age of 75 are more likely to commit suicide, which is also a trend reversal from what we've seen so far.

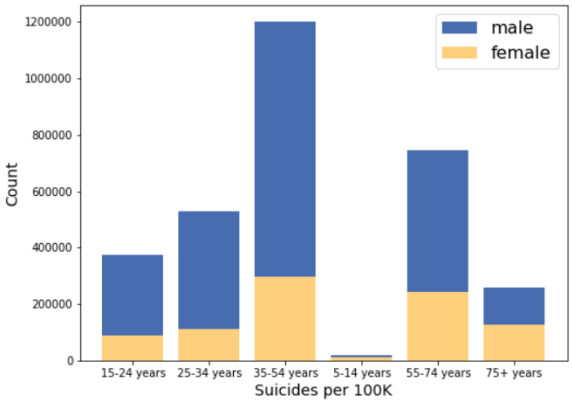


Fig 1. Bar plot showing no of suicide committed per 100k population

Following that, we look at the top ten countries with the largest number of suicides. The top three countries are Russia, the United States, and Japan, with the United Kingdom having the fewest suicides.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

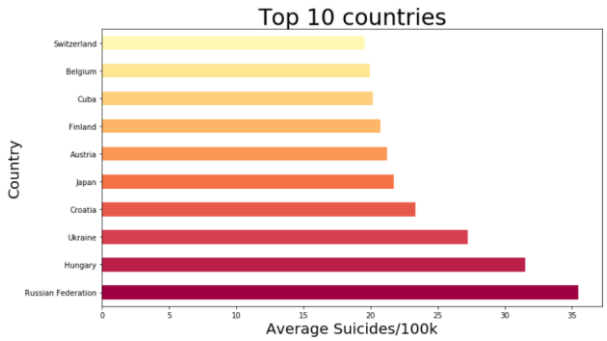


Fig 2. Country wise plot of suicide cases per 100k population.

After that, we look at the annual trend in the number of suicides. Suicide rates have steadily risen over time, as can be seen in the graph below. There are some dips in 1997 and a steady decline from 2002 to 2008, after which the rate jumps and then drop from 2009 to 2015.

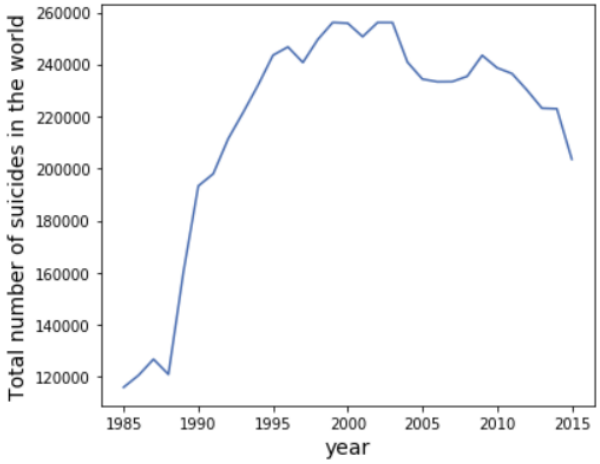


Fig 3. Suicides committed every year from 1985-2015

B. Binary Classification

We decided to conduct a binary classification on the suicide data, assigning high/low suicide risk groups based on the suicide incidences per 100,000 people. 'Risk' is added as an extra column to the "complete" data frame.

- Suicides<mean(Suicides) --> low risk --> class 0
- Suicides>mean(Suicides) --> high risk --> class 1

C. Model Training

Our datasets were divided into three categories: training, testing, and validation. Our Machine Learning algorithms include Decision Tree, AdaBoost, XGBoost, and Neural Network.

V. RESULTS DISCUSSION

Data analysis helped us understand several underlying trends in suicide attempts over the years 1985 and 2016. We found that Russia was the leading contributor to deaths occurring from suicides per 100K of the population. Talking about the country with the least deaths per 100K of population was Switzerland with less than 20 deaths per 100K. Coming to the performance of the four machine learning models - Ada-boost gave us an accuracy of 97.6% followed by xg-boost at 96.35%. Decision tree had an accuracy of 91.67% and finally the mlp classifier gave us an accuracy of 90%.

VI. CONCLUSION

We have taken a close look at the research done in the area of suicide prediction. Many successful high accuracy models have been made but there is still room for improvement. Preliminary data analysis revealed several hidden results such as the impact of GDP on suicides, the positive relationship between introversion and depression and that teen males have a higher tendency for suicides.The analysis carried out will also provide knowledge about areas of improvement to the government and other organizations working towards suicide prevention and counselling so that effective steps can be taken. Machine learning algorithms like gradient boosted decision tree and neural networks consistently outperformed other algorithms and had the highest accuracy and precision. Chatbots yielded a very low accuracy but its still early days of AI and there is much to be done in the field of human emotion understanding.

REFERENCES

[1] Gen-Min Lin, Masanori Nagamine, Szu-Nian Yang, Yueh-Ming Tai, Chin Lin, Hiroshi Sato, “Machine Learning Based Suicide Ideation Prediction for Military Personnel”, IEEE Journal of Biomedical and Health Informatics, vol. 24, issue: 7, July 2020.
[2] L. K. Richardson, B. C. Frueh, and R. Acierno, “Prevalence estimates of combat-related PTSD: critical review,” Australian and New Zealand Journal of Psychiatry, vol. 44, no. 1, pp.4-19, January 2010.
[3] Ji-Won Baek and Kyungyong Chung, “Context Deep Neural Network Model for Predicting Depression Risk Using Multiple Regression”,IEEE Access, vol. 8, January 2020.
[4] S.A.S.A. Kulasinghe; A. Jayasinghe; R.M.A. Rathnayaka; P.B.M.M.D. Karunarathne; P.D. Suranjini Silva; J.A.D.C. Anuradha Jayakodi, “AI Based Depression and Suicide Prevention System”,IEEE, 2019 International Conference on Advancements in Computing (ICAC), December 2019.
[5] Jun Shen; Shihui Zhao; Mingzi Ye,”Suicide Prediction Analysis with Generalized Addictive Model”,2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), October 2019.
[6] Max.R. (2014) Human development index(hdi). [Online]Available:https://ourworldindata.org/human-development- index
[7] Rusty. (2018) Suicide rates overview 1985 to 2016.[Online]. Available:https://wiki.math.uwaterloo.ca/statwiki/index.php?title=stat 841f11
[8] Y. W. NM Razali, “Power comparisons of shapiro-wilk,kolmogorov-smirnov, lilliefors and anderson-darling tests,”Journal of Statistical Modelling: Theory and Applications,2011.
[9] Bradley, Distribution-Free Statistical Tests. Prentice-Hall,1968.
[10] Econometrica, “A simple test for heteroskedasticity and random coefficient variation,” R package version, 2017,1979
[11] R. J. Hastie, T. J.; Tibshirani, “Generalized additive models,”Chapman & Hall/CRC, 1990.
[12] Tarun Agarwal; Anshul Dhawan; Apoorv Jain; Adeeshwar Jain; Shilpa Gupta, “Analysis and Prediction of Suicide Attempts”,2019 International Conference on Computing, Power and Communication Technologies (GUCON), September 2019.
[13] Pragya Prashar; Tanupriya Choudhury; Praveen Kumar; Karshin Khatri,”Analysis & Counter Measures paradigm associated with Suicide”,2018 International Conference on Advances in Computing and Communication Engineering (ICACCE),June 2018.
[14] Shaoxiong Ji; Shirui Pan; Xue Li; Erik Cambria; Guodong Long; Zi Huang,”Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications”,IEEE Transactions on Computational Social Systems, September 2020.
[15] Ranjitha Korrapati, Kranthi Nuthalapati, S.Thenmalar,”A Survey Paper on Suicide Analysis”,International Journal of Pure and Applied Mathematics”,Volume 118 No. 22 2018, 239-244.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 04, April-2021

[16] Pragya Prashar; Tanupriya Choudhury,"Suicide Forecast System Over Linear Regression, Decision Tree, Naïve Bayesian Networks and Precision Recall",2018 8th International Conference on Cloud Computing, Data Science & Engineering (Confluence)
[17] Rifat Zahan; Ian McQuillan; Nathaniel Osgood,"DNA Methylation Data to Predict Suicidal and Non-Suicidal Deaths: A Machine Learning Approach", 2018 IEEE International Conference on Healthcare Informatics (ICHI), June 2018.