International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Analysis of Suicide Attempts and Its Prediction

Rasika Mahadik[1], Shubham Salunkhe[2], Sneha[3], Prof. Vijaya Sagvekar[4]

[1]*Rasika Mahadik Mumbai University*
[2]*Shubham Salunkhe Mumbai University*
[3]*Sneha Mumbai University*
[4]*Professor Vijaya Sagvekar, Dept. of Information Technology, PVPPCOE, Maharashtra, India*

Abstract - *Suicide is becoming serious concern as it is contributing to major deaths happening worldwide. Each year thousands of people are suffering from depression and a few receives adequate treatment. The paper considers various attributes or features responsible for suicide attempts. These factors are analyzed from available dataset. Using various machine learning algorithms like Random Forest, XG Boost and Naïve Bayes suicide prediction has been made here. The aim of this research is to understand the importance of these algorithms for decreasing future suicides. In this research paper we mainly focused on developing a prediction model which will predict the suicide attempt. The model is first trained using dataset and then tested. Various features are considered from dataset which have significant linear relationship with number of suicides.*

Key Words: **Machine Learning, Naive Bayes, XG Boost, Random Forest**.

1. INTRODUCTION

Suicide is a major problem that affects millions of people worldwide. Suicide behavior could be conceptualized as a phenotype continuum ranging from suicide ideation to suicide attempt and eventually leading to suicide. This machine Learning approach provides way for predicting suicide attempts and taking measures accordingly. We have used various machine learning algorithms which provides way for predicting suicide classified as Yes or No.

Depression is one of the reasons for suicide attempts happening. Depression is nothing but a low mood, aversion to activity that affects a person's attitude, behavior & feeling of well-being. Depression gives rise to so many other problems that includes loss of interest, helpless, hopeless feelings & mental disorder. Mental illness is a leading cause of disability worldwide. The study in this paper includes analyzing the dataset and understanding various attributes contributing towards suicide attempts through various visualizations.

2. EXISTING SYSTEM

Many contributions have been made in the field of suicide prediction and prevention in recent years. These works focused on one or few attributes or features. Mostly the Demo-graphical features like education, marital status or gender were considered for creating the model. In researches specific people or category of people are taken into considerations. The accuracies in the predictions are also affected by change in the size of data in few of the existing systems. The Disadvantages of the Existing systems are:

- One or few features or attributes are considered as factors that lead to suicide attempts for analyzing and creating the models.
- Selected group of individuals or people from a particular region are considered for making the model or system.

3. PROPOSED SYSTEM

In today's rapidly changing world where competition between individuals is more, people tend to be stressed and depressed.

The proposed system tries to overcome the drawbacks of the existing system. In the proposed system, a model is developed for the significant features to predict the suicide. The system will first do the analysis of the features that contribute to a person for attempting suicide and then will use algorithms like XG Boost, Random Forest and Naïve Bayes for suicide prediction and accuracy. It will then consider the results and accuracy of different algorithms to predict future such attempts with precision. Our system will have modules like:

- GUI (Graphical user interface) which will contain modules like individual **user** profile and **Admin.**
- Questionnaire.

In the first phase data preprocessing, data is cleaned by removing null values, unwanted data, etc. before using it in different algorithms for prediction. After the analysis, it was found that gender, sexuality, age, income, race, bodyweight, virgin, friends, social _fear, and depressed as the most

important contributing factors (independent variable) for predicting the target(dependent) variable: attempt_suicide. Other features were dropped. Since the machine learning models can only work on numerical data, data was checked null and NA values.

Questionnaire are asked to the user, who needs to register and login first, for more accurate prediction and to understand the individual's state of mind to prevent suicide attempt by giving appropriate guidance after the questionnaire are answered by the user.

The Admin will get details of all the users with their questionnaire data and can help to create awareness among people.

In the last phase of the model the classification result is analyzed and it is displayed to the admin to keep a track of the data of individuals.

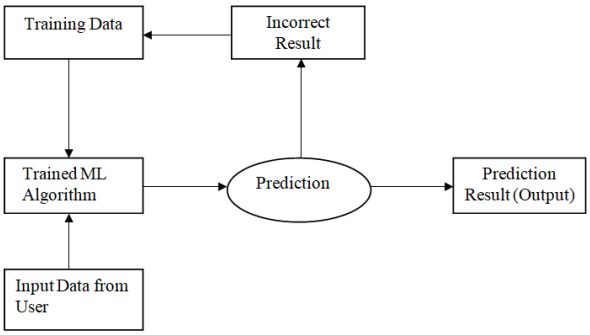

Figure 1:- System Flow Architecture

3.1 DATASET USED

A dataset is a collection of data arranged in some order. In a relational database management system data is stored in tables which consists of rows and columns. The dataset used in this project contains various features such as demographics data such as gender, sexuality, age, race, employability of person, weight of person etc. This data is real and is available on Kaggle (open-source dataset repository) under the name 'Forever Alone'.

As dataset plays vary important role in any prediction model. A part of dataset is used for training and other part is for testing. So, dataset used has to be accurate and data is preprocessed for cleansing data and removing missing values. Based on the dependent variables binary classification is performed on target variable as Yes or No indicating whether the person will attempt suicide or not.

If it is classified as Yes them it will be consider as Possibility of person may commit suicide otherwise no.

3.2 TECHNOLOGY USED

Anaconda: Anaconda v3.5 was used for the implementation.

Language: Python 3.8

Flask Web application framework.

3.3 ALGORITHMS

RANDOM FOREST

Random forest is a supervised learning algorithm and ensemble. Ensemble in the context of machine learning means that aggregating the set of weak learners and make them work together to develop a strong learner. The general idea of the bagging method is that a combination of learning models increases the overall result. Random forest simply builds multiple decision trees and merges them together to get more accurate and stable prediction. Decision Trees (weak learners) are ensemble to build a Random Forest (strong learner) to perform various tasks like classification a regression. Random forest is very efficient on huge datasets. One big advantage of Random forest is that it can be used for both classification and regression problems.

The difference between random forests and decision trees is that while random forest is a collection of decision trees, there are some differences. If one inputs a training dataset with features and labels into a decision tree, it will formulate some set of rules, which will be used to make the predictions. Most of the time, random forest prevents this by creating random subsets of the features and building smaller trees using those subsets. Afterwards, it combines the subtrees.

- The **hyperparameters** in random forest are used to increase the predictive power of the model or to make it faster.
- The **n_estimators** hyperparameter, which is the number of trees the algorithm builds before taking the maximum voting or taking the averages of predictions.
- The **max_features** hyperparameter is the maximum number of features random forest considers to split a node.

Advantage and disadvantage:
Random forest is good algorithm because the default hyperparameters it uses often produce a good prediction result.

The main limitation of random forest is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions. This algorithm is fast to train, but quite slow to create predictions once trained. For more accurate prediction requires more trees, which results in a slower model.

NAÏVE BAYES

Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
The formula for Bayes theorem is given as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

It is a good and mostly used classification Algorithm. However it provides less accuracy for this model.

XGBOOST

XGBoost or extreme gradient boosting is one of the well-known gradient boosting techniques(ensemble) having enhanced performance and speed in tree-based (sequential decision trees) machine learning algorithms. It is the most common algorithm used for applied machine learning in competitions and has gained popularity through winning solutions in structured and tabular data. XGBoost falls under the category of Boosting techniques in Ensemble Learning. Ensemble learning consists of a collection of predictors which are multiple models to provide better prediction accuracy. In Boosting technique, the errors made by previous models are tried to be corrected by succeeding models by adding some weights to the models. In proposed system, the accuracy and efficiency of this algorithm is good.

4. CONCLUSIONS

Suicide is never a solution to any problem. The Mental health issues are considered as insignificant and we don't pay much attention to it, but on a greater extent these issues are needed to be openly discussed and proper diagnosis should be given to the individuals. In recent times with this increasing rate of suicide hopefully this proposed system can contribute in reducing causalities. The main objective of this paper is to predict accurately. Preventive measures focused on these classes of people will help in bringing down the number of suicides in the country. Government policies and preventive measures must be focused at these classes of people.

In this paper, by the help of questioners we took input from users. Analyzing the data collected from users and providing help accordingly. By further comparison of results of these three algorithms, it is found that accuracy of Random Forest is higher, accuracy of XG Boost is medium and accuracy of Naive Bayes is lower. The accuracies for prediction could be further improved if Class Distribution of the target variable gets balanced.

5. REFERENCES

[1] https://ncrb.gov.in/en/accidental-deaths-suicides-india-2019

[2] Vijayakumar L. Indian research on suicide. Indian J Psychiatry. 2010;52: S291–S296. doi:10.4103/0019-5545.69255.

[3] Priyanka, S. S., Galgali, S., Priya, S. S., Shashank, B. R., & Srinivasa, K. G. (2016). Analysis of suicide victim data for the prediction of number of suicides in India. 2016 International Conference on Circuits, Controls, Communications and Computing (I4C).

[4] Colic, S., J Richardson, D., James Reilly, P., & Gary Hasey, M. (2018). Using Machine Learning Algorithms to Enhance the Management of Suicide Ideation. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).