Design of Machine Learning based Suicide Rate Prediction System

Jhansi lakshmi Durga Nunna[#1], Akila Rani M [*2], B V Ram Kumar[#3]

[#*]*Department of CSE,BVC Institute of Technology and Science, JNTUK, Andhra Pradesh*

***Abstract*—There are two most important days in our life the day we born and the day we find out why, but now a day's most of the people are leaving the world without finding the answer for that question.People who can't get through life are just easily give up on it may make an easy but irrational decision just to end their own life. Everyone has set of their own problems. It is really depend on how the people are emotionally strong. Support from their family and friends also play an important role to prevent them to make such a bad decision. Youth are back bone to the nation. They can change the future of the society with their well being and courageous behaviour. But most of the youngsters are being affected by suicide. The proposed system concentrated on predicting the suicide Rate at a particular Country. It can be helpful to take the preventive action to reduce suicide risk rate based on some criteria and also it can predict the suicide risk rate more accurate than the existing system by applying Random Forest Regression Algorithm in machine learning. The suicide rate Prediction is very important factor for the government because if we already predict potential suicidal conditions through surveys we can try to stop them.**

***Keywords*— Regression Algorithm, Prediction, machine learning, suicide rate**

I. INTRODUCTION

Machine Learning (ML) is expected to bring heavy changes to the world of technology. Machine learning is a subfield of Artificial Intelligence and computer science that allows software applications to be more accurate in predicting results. The prime objective of machine learning technology is to build algorithms that can get input data and leverage statistical analysis to predict an acceptable output value.

In this section, we will present an overview of popular supervised Machine Learning techniques, for its subsets of classification and regression.

Supervised learning is the task of learning a function that maps input data to output data based on example input-to-output pairs. Classification happens when the out-put is a category, whereas regression happens when the output is a continuous number.

A. Generalized linear models

Generalized Linear Models are a set of regression methods for which the output value is assumed to be a linear combination of all the input values .Generalized Linear Models are a popular technique as they are easy to implement and, in many classification or regression problems, assuming linearity between predictor variables and the outcome variable is sufficient to generate robust predictions.

B. Decision trees

Decision Trees are a popular Machine Learning technique to link input variables, represented in the tree's branches and nodes, with an output value represented in the tree's leaves. Trees can both be used in classification problems, by outputting a category label, or in regression problems, by outputting a real number. Decision Trees can be fitted using different algorithms, including the CART or ID3 decision tree algorithms which are the most popular. However, decision trees can often become inaccurate, especially when exposed to a large amount of training data as the tree will fall victim to over fitting.

C. Random forest

Random Forests operate by building a large amount of decision trees during training, taking a different part of the dataset as the training set for each tree. For classification problems, the final output is the mode of the outputs of each decision tree, whereas for regression problems, the mean is taken. This result in a model with much better performance compared to a simple decision tree, thanks to less over fitting, but the model is less interpretable as the decisions at the nodes of the trees are different for each tree.

D. Support vector machines

Support Vector Machines (SVMs) are Machine Learning models for both classification and regression. An SVM model represents the training data as points in space so that examples falling in different categories are divided by a hyperplane that is as far as possible from the nearest data point.

II. LITERATURE SURVEY

Literature survey is the most important step in software development process. Before developing the tool it is necessary to determine the time factor, economy and company strength. Once these things are satisfied, ten next steps are to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool the programmers need lot of external support. This support can be obtained from senior programmers, from book or from websites. Before building the system the above consideration are taken into account for developing the proposed system.

The predictive capability of the machine learning model is comparable to that of suicide risk assessment tools used in the clinical background [1][2].Suicide rate has been found to be related with clinically significant symptoms of mental illnesses such as depression and bipolar disorder [3][4].

Observational studies have also time after time corroborated this 47–74% of suicide rates finding [6][7]. Other suicidal behaviours, such as past try, or psychiatric hospitalization indicate particularly high risk and take place in an estimated 25–65% of suicide attempts internationally [8][9].

Worldwide in 2015, about 788 000 people died by suicide at a worldwide rate of 10.7 per 100 000 person-years (or about 1 in 9350 people per annum). That year, national suicide rates were lowest in the small Caribbean nations of Barbados and Grenada, Antigua and Barbuda, each of which had suicide rates below 1 per 100 000 person years among populations of fewer than a quarter of a million. However, in 2015, suicide rates also speckled remarkably between populous nations, ranging from 1.4 per 100 000 person years in Jamaica to 34.6 per 100 000 person-years in Sri Lanka. While some of this international variation might be a result of differences in the definition suicide or methods of data collection, there is little doubt that there are large and real differences in national suicide rates. Decades of work standardizing the reporting of suicide has not resulted in converging rates, and national suicide rates are notably stable on a year-to-year basis. Hence, suicide rates between nations can vary by more than an order of magnitude.

This suggests that some preventative measures might be justified in nations that have a higher suicide rate, but not in lower suicide rate nations. For example, although two nations might have similar problems with agricultural pests, the overall benefit of restricting access to toxic pesticides might clear in high suicide rate countries like Sri Lanka (where 1 in 2900 die by suicide each year) but might be less obvious in low suicide rate countries like Jamaica (where as few as 1 in 74 500 die by suicide each year).

The reasons for the marked heterogeneity in international suicide rates are not fully understood. One important observation is that national suicide rates by particular lethal methods (such as hanging, poisoning, gassing, shooting, jumping, and drowning) vary greatly between nations but tend to be stable within nations on a year-to-year basis.3 This predictability of method specific suicide rates underpins most universal measures to prevent suicide. Well-known examples include the substitution of natural for coal gas in the United Kingdom in the 1960s, the regulation of firearms in Australia in the 1990s, and the trend towards bans on highly hazardous pesticides in many countries. Each of these universal measures resulted in reductions in both cause specific suicide mortality and a drop in suicide rates. Other universal preventative measures are the reduction in analgesic pack size, the substitution of barbiturates with benzodiazepines, the placement of barriers at jumping hotspots, measures to decrease alcohol consumption, and changes to media reporting of suicides. In each of these cases (with the slightly contentious exception of the regulations in firearms) suicides rates have been reduced at little or no cost or inconvenience to the whole population.

III.PROPOSED SYSTEM DESIGN

The overall system design consists of following modules

A. Data Collection

The dataset (suicide-rates-overview-1985-2016) used in this proposed system was an open source dataset from KaggleInc. It consists of 27820 records with 12 parameters that have the possibility of affecting the suicide rate. However out of these 12 parameters only 8 were chosen which are bound to affect to predict the suicide rate. Parameters such as Country, Year, Gender, Age, PopulationGdpPerCapitalmoney, Generation. Suicide number is a dependent variable on several other independent variables.

B. Preprocessing

It is a process of transforming the raw, complex data into systematic understandable knowledge. It involves the process of finding out missing and redundant data in the dataset. Entire dataset is checked for Not a Null(NaN) and whichever observation

consists of NaN will be deleted. Thus, this brings uniformity in the dataset. However in our dataset, there was no missing values found meaning that every record was constituted its corresponding feature values.

C. Data Classification

Classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

D. Data regression

Regression is basically a statistical approach to find the relationship between variables. In machine learning, this is used to predict the outcome of an event based on the relationship between variables obtained from the data-set. Linear regression is one type regression used in Machine Learning.

E. Prediction of Output

Output can be predicted by using Machine Learning algorithms.
Data is collected and stored in NoSQL / SQL format. That data is divided into two parts

i) Training data
ii) Testing data.

Training data is used for training the model and then that model is tested using testing data. After this, the trained model is used for predicting house price given feature set.

Flowchart for suicide rate prediction

Fig 1 outlines the selection of input and the procedure to get the output that in this study is the suicide rate prediction.

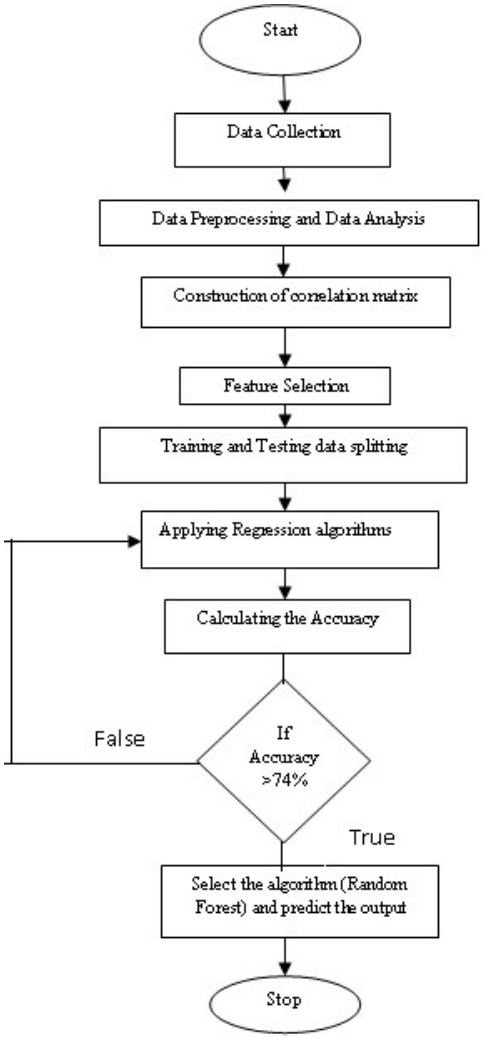

Fig 1 Flowchart for suicide rate prediction

The flowchart illustrates the first step as a selection of inputs that are parameters, processing of these inputs and completing training, testing and validation for accurate (<74%) and precise output that is rain forecast. Psychological autopsy studies have now established that psychiatric disorders account for 47–74% of suicides at a population level [5]. So the accuracy is validated with the value <74%.

During the last decades, there has been an incredible growth in our capabilities of generating and storing data. In general, there is a competitive edge in being able to properly use the abundance of data that is being collected in industry and society today. Efficient analysis of collected data can provide significant increases in productivity through better business and production process understanding and highly useful applications for e. g. decision support, surveillance and diagnosis

IV. RESULTS AND DISCUSSION

The proposed system is implemented with few regression[10] algorithms in order to finalize which regression algorithm will give better accuracy and the comparison results were listed in Table 1. From this result, the proposed system was implemented with random forest algorithm for regression with better suicide rate prediction accuracy than the existing system.

TABLE I
PREDICTION ACCURACY COMPARISON

S.NO	NAME OF THE REGRESSION ALGORITHM	PREDICTION ACCURACY
1	Linear regression	3.9%
2	Ridge regression	4%
3	Lasso regression	4%
4	Elasticnet regression	4%
5	Decision tree regression	90%
6	Random forest regression	98%

V. CONCLUSIONS

This proposed system design provides an opportunity with the help of Machine learning to predict the suicide rate in the world by country wise. The suicide rate Prediction is very important factor for the government because if we already predict potential suicidal conditions through surveys we can try to stop them. By applying random forest regression in machine learning the proposed system attained 98% accuracy which is comparatively better than the existing system.

REFERENCES

[1] Bolton JM, Gunnell D, Turecki G., “Suicide risk assessment and intervention in people with mental illness”, BMJ 2015;351:h4978..
[2] Ghasemi P, Shaghaghi A, Allahverdipour H, “Measurement scales of suicidal ideation and attitudes: a systematic review article”, Health Promot Perspect 2015;5:156-168.
[3] Chellappa SL, Araujo JF. Sleep disorders and suicidal ideation in patients with depressive disorder. Psychiatry Res 2007;153:131-136.
[4] Valtonen H, Suominen K, Mantere O, Leppamaki S, Arvilommi P, Isometsa ET. Suicidal ideation and attempts in bipolar I and II disorders. J Clin Psychiatry 2005;66:1456-1462.
[5] Cavanagh J, Carson A, Sharpe M, Lawrie S, “Psychological autopsy studies of suicide: a systematic review”. Psychol. Med. 33(3), 395–405 (2003).
[6] Nordentoft M, Mortensen PB, Pedersen CB. , “Absolute risk of suicide after first hospital contact in mental disorder”. Arch. Gen. Psychiatry 68(10), 1058–1064 (2011).
[7] Tidemalm D, Långström N, Lichtenstein P, Runeson B. , “Risk of suicide after suicide attempt according to coexisting psychiatric disorder: Swedish cohort study with long term follow-up”. BMJ 337(8), 2205 (2008).
[8] Qin P, Nordentoft M., “Suicide risk in relation to psychiatric hospitalization: evidence based on longitudinal registers”. Arch. Gen. Psychiatry 62(4), 427–432 (2005).
[9] Beck AT, Steer RA., “Clinical predictors of eventual suicide: a 5- to 10-year prospective study of suicide attempters”. J. Affect. Disord. 17(3), 203–209 (1989).
[10] Tri Doan, jugal kalita, “Selecting Machine Learning Algorithms using Regression Models”, IEEE International Conference on DataMining Workshop,41-17 Nov, 2015.