ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Classification of Suicidal Deaths Caused in India through Various Supervised Machine Learning Techniques

[1] K.Sai Teja , [2] S.Pravalika, [3] G.Varshitha, [4] Syed Muzamil Basha
[1, 2, 3] Sreenidhi Institute of Science and Technology,Yamnampet, Ghatkesar, Hyderabad, Telangana 501301
[4] Sri venkateswara college of engineering and technology, Chittoor, India

Abstract: The prevalence of suicides registered among students and farmers show a discrepancy across countries and socio demographic populations. Whereas in India, In 2012, over 2200 students committed suicide, due to failure in exams. There are several risk factors which causes the suicides i.e. Professional reasons, Personal problems, Depression and other reasons. Understanding the causes and classifying the data considering the several factors such as age, cause, state and the year of the incident and performing classification of the causes and specifying the counter measures to overcome the cause. Considering dataset having 236583 observations and seven variables form year 2001-2012 describing out the suicide cases registered in India. The main focus, in our research is to address the risk factor in committing suicides among students and farmers in India. The findings in our research shows that, the Impact of committing suicides is more with the Age group of 0-14 and 15-29.

Keywords: — prevalence, suicides, India, Classification.

I. INTRODUCTION

According to the survey made by NDTV on 2013, It was declared that for every hour there are 15 suicides getting registered across India. In fact, more than 371 suicides every day, the reason behind all this suicides are instability in thinking, lack of motivation. Among all the suicide cases, 5.5% of all suicide victims are students and farmers. Medical care claims that "person mental illness will attempts to commit suicide". In India, addressing this issue, a Mental Health care Bill got introduced in 2013 towards decriminalizing suicides. As per the records collected form 2001-2012, mostly students and farmers are vulnerable to suicides. Students committed suicide, due to failure in exams. The other side, the farmers. It is commonly know that the agriculture is the backbone of India. But, many are unaware of pledge of the Indian farmer. Maharashtra a state in India constitutes for half of the farmer suicides in India [14]. The government has been battling farmer suicides over 15 years.. Vidarbha is a village in Maharashtra, where dams are constructed for irrigation. But, the outlet is not started. This is the fundamental problem for the former as they are dependent on rain for their crops to grow. The BJP government as lunched a 15,000 crore scheme called "Jalyukta Shivaar Yojana" aims to make Maharashtra a drought-free state by 2019. While government is attempting to help, they are not even close to meeting the needs of the farmers. The life of the farmer is hard, working 12 hours a day, with limited resources and no guarantee of a yield due to the in consistent water supply. Adding to this the cost of production for cotton farming has increased with the evolution of Bt cotton. However, this is the only crop whose rate is determined by the world market. This suicides as there is no proper training in this field. The central and state government of India should take necessary steps in stigmatizing illness. In the way to provide justification for the above statement made, we are conducting experiment on suicide dataset. The description of the dataset consider in our experiments are as follows. It contains 236583 observations and seven variables. The complete summary is provided in the Table 1.

II. LITERATURE REVIEW

In [1] the author used weighted fuzzy logic to assign weights in training the data to extract sentiments from the labeled tweets and achieved good F-score. where as in [2] prediction is made on time series data. In [3] the author performed deep analysis on PIMA diabetes. where as in [4] the author used gradient ascent algorithm in finding out the exact weights of the terms used in determining the sentiment of tweet and used Boosting approach to improve the accuracy of linear classifier. In [5], the author provide a novel way of performing prediction on

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

Breast cancer dataset, compared the performance of three different feature selection algorithm and proved that genetic algorithm is giving best result in selecting the best feature among all the available feature. SVM algorithms gives the best result in predicting the level of certainty in breast cancer. In [6], the author made an attempt to develop an recommender system, helping in searching the item, that might out found by themselves, In which precision and recall measures are used in measuring the performance of proposed model. In [7], the author made an research in solving the problem in Diabetic Retinopathy. In which, the author proposed a Model, which can capable of calculating the weights, that gives severity level of the patient's eye by using weighted Fuzzy C-means algorithm. In [8], the author proposed a model for airlines, that can extract sentiments from customer feedback and achieved Vital accuracy.

III. BACKGROUND KNOWLEDGE

Random Forest is considered to be as supervised Machine learning algorithm. It is used in addressing problems of regression and classification.

State	Year	Type	Age_Group	Total
Karnataka : 6792	Min. :2001	Others (Please Specify) : 7263	0-100+:10920	Min. : 0.00
Madhya Pradesh: 6792	1st Qu.:2004	By Coming Under Running Vehicles/Trains : 4200	0-14 :45027	1st Qu.: 0.00
Maharashtra : 6792	Median :2007	By Consuming Insecticides : 4200	15-29 :45223	Median : 0.00
Andhra Pradesh: 6791	Mean :2007	By Consuming Other Poison : 4200	30-44 :45193	Mean : 30.64
Odisha : 6791	3rd Qu.:2010	By Drowning : 4200	45-59 :45146	3rd Qu.: 6.00
Rajasthan : 6791	Max. :2012	By Fire-Arms : 4200	60+ :45074	Max. :8756.00
(Other) :195834	Na	(Other) :208320	Na	Na

Table 1. Description of the suicide Dataset in India from 2001-2012.

It is a ensemble classifier i.e it is a ensemble of Decision Tree. In prediction it is used in finding out the relative importance of features. Random Forest classifier considers different features and construct a decision tree based on that features as in Figure 1. It constructs a Decision Tress and then based on the majority voting and accuracy decision tree is selected. Major advantage of random forest is it considers each and every attribute gives the best result. As Random forest classifier can also be used for cause of death prediction[9].

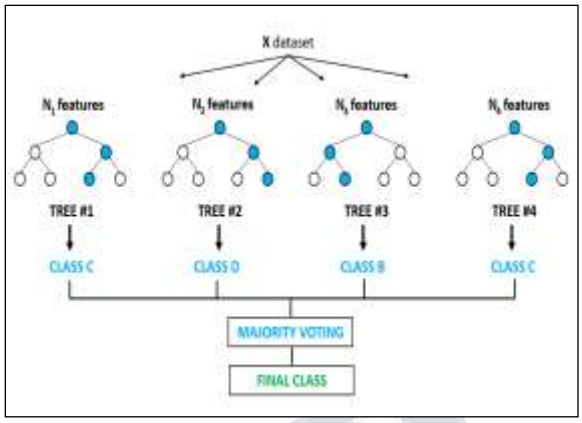

Figure 1. Random Forest Classifier

No of Decision Trees to be taken as estimators in Random forest and the Criterion are Entropy: The entropy (very common in Information Theory) characterizes the impurity of an arbitrary collection of attributes and Gini index: Gini ratio or a normalized Gini index.

Logistic regression measures the dependency of nominal independent variables, as in Figure 2. According to the research related to the Logistic [10]. In the logistic regression the constant (b0) moves the curve left and right and the slope (b1) defines the steepness of the curve as in Equation 1.

$$\frac{p}{1-p} = \exp(b_0 + b_1 x) \tag{1}$$

Finally, the coefficient (b1) is the amount the logit (log-odds) changes with a one unit change in x as in Equation 2 and 3.

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x \tag{2}$$

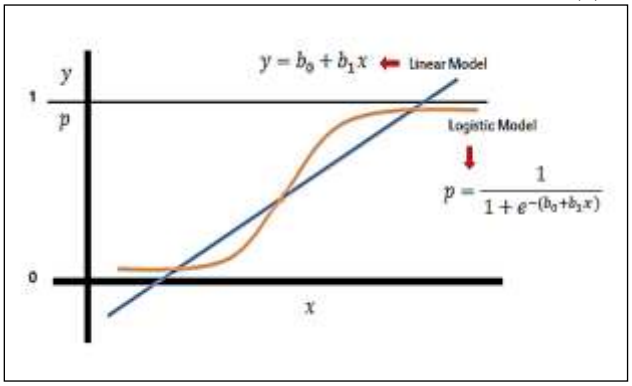

Figure 2. Logistic and linear model curves

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

The idea of a decision tree is to divide the problem into smaller sub problems until you reach a unique solution as in Figure 3. The decision of splitting is based on entropy reduction [11].

$$p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}} \quad (3)$$

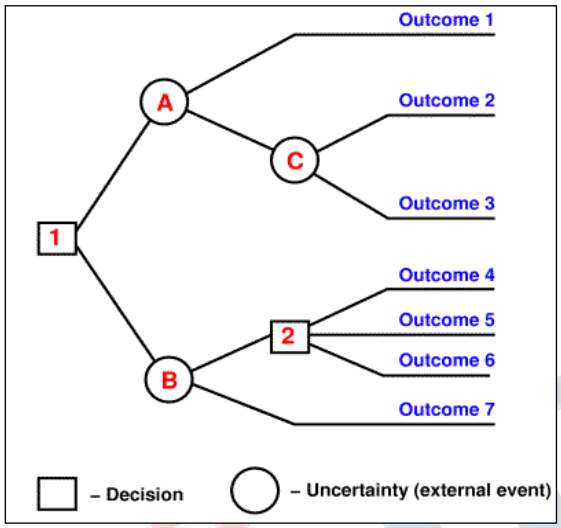

Figure3. Decision Tree

ID3 algorithm uses entropy to calculate the homogeneity of a sample. The aim of constructing Decision tree is to gain highest information gain. Bayesian Classification calculates explicit probabilities as in Equation 5.

$$P(h/D) = P(D/h)\, P(h)\, P(D) \quad (5)$$

Weak classifier algorithm can be combined using Ada Boost to form strong classifier. we can have good accuracy score for overall classifier [12].

4. Major Causes and Precautions

Gender: Female, Male
Type of attempt:By Drowning, By Consuming Poison, Health issues
Reasons:
1) Children not themselves attempt to suicide but parent make them to take sleeping pills or poison due there financial issues or personal issue, Total family attempt to suicide
2)Second reason is due the health issues like cancer since there is no cure, they attempt for suicide or parents make them to do it
Case studies:
1) Man Allegedly Commits Suicide; Wife, Two Daughters Found Dead
2) six Members Of Family Found Dead In Telangana, Suicide Suspected

Classified Type Code: Causes, Means adoption
Countermeasures:
1) Formulation of "implementation models" with local governments to prevent suicide
2) Personal support to prevent depression and suicide, Creation of an internet-based crisis intervention model

15-29 Age Group
Gender: Male, Female
Types of attempts/reasons: Failure in Examination, Fall in Social Reputation, Family Problems, Cancellation/Non-Settlement of Marriage, Love Affairs

Reasons:
1) Academic performance effects the youth a lot, If they could not perform well in academics which lead to mental stress and leads to addiction of drugs and alcohol
2)Physical abusing and harassment, another major cause in love failure

Case studies:
1)Bhopal: B.Com First Year Student Commits Suicide After Being Scolded by Father

2) IIT-Delhi student commits suicide in hostel room

Classified Type Code: Causes, Means adoption,Education_Status
Countermeasures:
1)Make them to attend counselling class inorder to grow mentally strong
2)Government should create employment opportunities to the youth
3)Those who addicted to drugs and alcohol should be taken to Rehabitation centers

30-44+ Age Group

Gender:Male,Female
Types of attempts/reasons:Farming/Agriculture Activity[3],Self-employed (Business activity),Professional Activity.

Reasons:
1)Major cause of suicides in this age is professional problems,unemployment,depts. Which indeed leads to poverty which causes depression

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

2)Most of registered suicides cases in this age group are Farmers.

Classified Type Code:
Education_Status, Professional_Profile, Social_Status

Case Studies:
1)After Selling Crop At A Loss, Telangana Farmers Return Home With Pesticide

2)Burdened By Debt, Indian-Origin Student Killed Himself In UK

Countermeasures:

1. Govt. should make institutional finance free to every farmer, provided guidelines on economical methods of cultivation. Encouragement in developing alternative sources of income and our government can take up responsibility to teach them new skills.
2. Don't Let Fear Of Failure Stop You From Success – Take action on those dreams!(Business failures)

Don't Let A Temporary Failure Kill Long Term Potential

V. METHODOLOGY

Using various machine learning Algorithms and evaluating the models and choosing the algorithm which gives the best perform metrics.Lets first understand the Algorithmic flow of the procedure refer to Figure 4. Extact the dataset from the database and perform some dimensionality reduction by feature selection and feature extraction.One of the interesting analysis is Univarient and Multivarient Analysis in order to find the relation among the attributes present in the dataset.

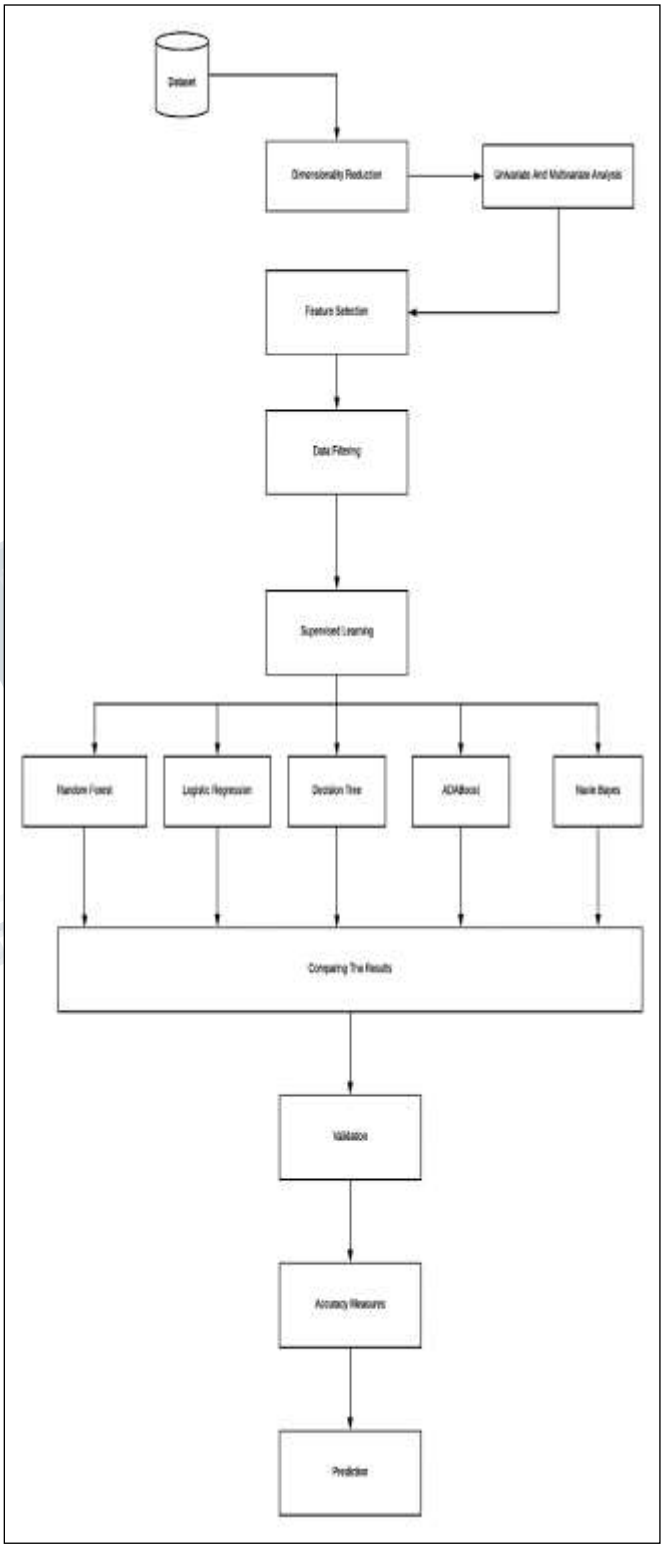

Figure 4. Algorithmic Dataflow

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 5, Issue 4, April 2018

After performing the uni-varient and multi-varient analysis we would like to extract the features which are useful for our classification and then perform the data filtering process inorder to remove the outliers and noisy data from the dataset and then comes the major task of applying the supervised machine learning algorithms to the filtered dataset

VI. RESULTS

After the comparison of various machine learning algorithms on the dataset ,Random Forest Classifier gives the highest Accuracy and can be used for the predictions. In Random Forest classifier 'Type' plays a major role to determine the Major reason to suicide and the Second major role is played by the "Age_group" Attribute. Through this classification we found that Age group between 15-29suicides are caused due Personal causes i.e family problems, academics as in Table 2.

Classifier	Accuracy
Decision Tree	**49.07**
Random Forest	**93.18**
Naïve Bayes	**81.67**
Logistic Regression	**78.54**
ADA Boost	**85.67**

Table 2. Comparison of MLA algorithm performance

VII. CONCLUSION

When a person thinks about killing themselves, the person is described as suicidal.In this study we classified the age groups and found the results that age 15-29 people are committing suicides,we found the countermeasures and discussed. In our research, We found that there is a need to do much research on ways of controlling suicide by continuously improving the conceptualization of suicides thoughts and behaviors and by improvising etiological understanding of a individual problem. In this study we classified the age groups and found the results that age 15-29 people are committing suicides, we found the countermeasures and discussed.

VIII. ACKNOWLEDGEMENT

we sincerely thank to our Executive Director Dr.P.Narasimha Reddy, HoD Of IT Dr.V.V.S.S.S.Balaram and Supervisor Dr.N.Ch.S.N.Iyengar for their support and encouragement.

REFERENCES

[1] Basha, Syed Muzamil, Yang Zhenning, Dharmendra Singh Rajput, N. Iyengar, and D. R. Caytiles. "Weighted Fuzzy Rule Based Sentiment Prediction Analysis on Tweets." International Journal of Grid and Distributed Computing 10, no. 6 (2017), pp. 41-54. . DOI: 10.14257/ijgdc.2017.10.6.04

[2] Basha, Syed Muzamil, Yang Zhenning, Dharmendra Singh Rajput, Ronnie D. Caytiles, and N. Ch SN Iyengar. "Comparative Study on Performance Analysis of Time Series Predictive Models." International Journal of Grid and Distributed Computing, Vol. 10, No. 8 (2017), pp.37-48. DOI: 10.14257/ijgdc.2017.10.8.04

[3] Basha, Syed Muzamil, H. Balaji, N. Ch SN Iyengar, and Ronnie D. Caytiles. "A Soft Computing Approach to Provide Recommendation on PIMA Diabetes." International Journal of Advanced Science and Technology 106 (2017): 19-32. DOI: 10.14257/ijast.2017.106.03

[4] Basha, Syed Muzamil, Dharmendra Singh Rajput, and Vishnu Vandhan. "Impact of Gradient Ascent and Boosting Algorithm in Classification."International Journal of Intelligent Engineering and Systems (IJIES) 11, no.1,(2018): 41-49. DOI: 10.22266/ijies2018.0228.05

[5] Basha, Syed Muzamil, Dharmendra Singh Rajput, N. Iyengar, and D. R. Caytiles. "A Novel Approach to Perform Analysis and Prediction on Breast Cancer Dataset using R." International Journal of Grid and Distributed Computing, Vol. 11, No. 2 (2018), pp.41-54, http://dx.doi.org/10.14257/ijgdc.2018.11.2.05

[
6] Vivek P. Khadse, Basha, Syed Muzamil, N. Iyengar, and D. R. Caytiles," Recommendation Engine for Predicting Best Rated Movies " International Journal of Advanced Science and Technology, Vol.110 (2018), pp.65-76, http://dx.doi.org/10.14257/ijast.2018.110.07

[7] Suvajit Dutta, Basha, Syed Muzamil, N. Iyengar, and D. R. Caytiles. " Classification of Diabetic Retinopathy Images by Using DeepLearning Models." International Journal of Grid and Distributed Computing,Vol. 11, No. 1 (2018), pp.89-106, http://dx.doi.org/10.14257/ijgdc.2018.11.1.09

[8] Divisha khaturia, ,Basha, Syed Muzamil, N. Iyengar, and D. R. Caytiles. " A Comparative study on Airline Recommendation System Using Sentimental Analysis on Customer Tweets." International Journal of Advanced Science and Technology, Vol.111 (2018), pp.107-114, http://dx.doi.org/10.14257/ijast.2018.111.10

[9] Sawant, Rutweek. "ANALYTICS DRIVEN INFLUENCING METHODS FOR ELECTION STRATEGIES." International Education and Research Journal, Vol. 3, No. 12 (2017).

[10] Peng, Chao-Ying Joanne, Kuk Lida Lee, and Gary M. Ingersoll. "An introduction to logistic regression analysis and reporting." The journal of educational research 96, no. 1 (2002), pp. 3-14.

[11] Sharma, Himani, and Sunil Kumar. "A survey on decision tree algorithms of classification in data mining." Vol.5 No. 4, (2016).

[12] Freund, Yoav, Robert Schapire, and Naoki Abe. "A short introduction to boosting." Journal-Japanese Society For Artificial Intelligence vol. 14, No. 1, (1999), pp. 771-780.

[13] Swami D, Dave P, Parthasarathy D. Agricultural susceptibility to monsoon variability: A district level analysis of Maharashtra, India. Science of The Total Environment, Vol. 1, No. 619 (2018), pp. 559-77.