Suicide Forecast System over Linear Regression, Decision Tree, Naïve Bayesian Networks and Precision Recall

Pragya Prashar
Amity School of Engineering
and Technology
Amity University
UP, India
pragyaprashar490@gmail.com

Tanupriya Choudhury
Amity School of Engineering
And Technology
Amity University
UP, India
tchoudhury@amity.edu

ABSTRACT: **Suicide is actually the act to hurt oneself in a manner that it may result into death. India is the nation where the rate of suicide is increasing immensely. India comprises of 17.5% of the world population. Also worldwide, there are 800,000 people who entrust suicide every year. Of which 135,000 i.e. 17% are the one that belong to nation India. There are several causes that may result into Suicide. Also the cause of suicide varies from person to person. Terms like Suicide signify as an unsuccessful attempt to kill oneself. The main objective of the analysis is to find out and classify the major cause of Suicide in different states of India so as to provide with preventive measures on the basis of the set of different parameters such as Gender, Age, and Cause. In order to prevent it in the Future.**

Keywords:- Linear Regression, Decision Tree, Bayesian Networks, Precision Recall.

I. INTRODUCTION

Suicides are the act wherein an individual harms oneself. Moreover, these days the rate of suicide is increasing, therefore there is an urgent need for preventive measures to be taken so as to reduce the amount of suicides that are taking place. In accordance to the study, there are several causes of Suicides. Furthermore according to the study there can be many causes of Suicides some are grouped under: Economic Cause (Unemployment, Bankruptcy, etc), Social Cause (Dowry Dispute, Love Affairs, Divorce, etc), Un curable Diseases (Cancer, Paralysis, AIDs, etc), Education (Failure in Examination, etc), etc[1]. It is also seen that Suicide is more or less related to Psychology. In other words what you think, how you tackle with the situation is also counted as the major cause of suicide other than internal and external problems that one faces in family or faces at an organization. Data Set group several causes of Suicide that majorly takes place under Economic, Social, Incurable Diseases, Education Failure and psychology so one can conclude that Suicide is not an act that is influenced only by socio-economic factors and psychological status of the person but also by several other parameters such as age, state or gender wherein they belong to plays a major role in a particular cause of Suicide. Furthermore demographic factors also play a vital role. Moreover it is also captured what all can be the means of suicide. Few means are listed below which an individual make use of to commit Suicide such as Hanging, Jumping, Drugs, Alcohol, Self-immolation. To prevent the rate of suicide in future. Data Analysis is carried out, in order to classify data so as to provide a set of preventive measures to control them in future. Data Set on which the analysis will be carried out will consist 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. The Data Analysis can provide information about the cause of suicide taken in a particular state followed by in a particular year. The Data Set can also provide information about whether the Suicide Rate for a particular cause has increased or not. Followed by a prediction what can be the Suicide Rate in Future? It can also provide information on what all areas need to be looked upon to reduce the Suicide Rate in different state as a result of different causes. 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. Followed by answer to question whether it may increase or decrease in the next coming years [2].

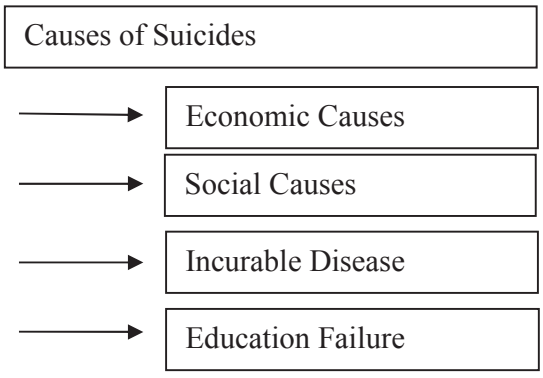

Fig.1. Causes of Suicides

II. THEORITICAL BACKGROUND

There is various application of Data Mining Field in the study of Suicides. Several techniques of Data Mining that can be used to analyze and classify the data under several parameters in order to find the trends or patterns in large dataset are as follows Regression Algorithm, Clustering Algorithm, Classification Algorithm, etc. Use of such Data

Mining Technique would help an individual to carry out a successful analysis from the dataset

Table I. Data Mining Algorithms in the study of Suicide

YEAR	NAME OF THE AUTHOR(S)	APPLICATION	DATA MINING METHODS
2008	Dr. Pooja Rastogi	Suicide in Youth: Shifting Paradigm	Data Mining
2011	Amitendu Palit	Suicides in India: The Economics at Work	Clustering
2016	Marouane Birjali	Prediction of Suicidal Ideation in Twitter Data	Machine Learning Algorithm
2017	Colin G.Walsh	Risk of Suicide Attempts	Classification

III. METHODOLOGY

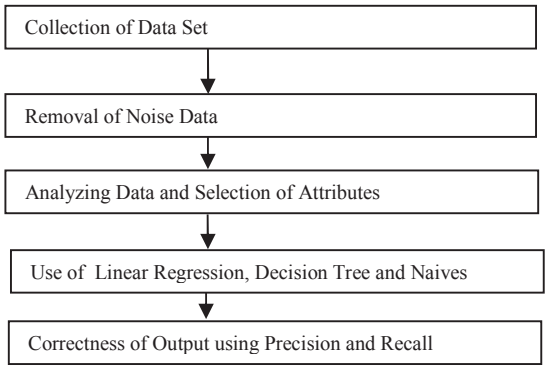

Fig .2. Flow Chart depicting steps undertaken

Collection of Data Set is done through various websites such as data.gov.in, Kaggle. Followed by cleaning of data ie removal of all missing values. Then clean data is to be analysed using Techniques such as Linear Regression, and further classified using Decision Tree, Naives Bayesian Network.

A. Data Set

The Data Set is collected from several websites. Different Attributes in Data sets are: State Name, Year, Age, Gender, etc. Figure 2 consists a screenshot of Suicide Data Set. Data Analysis will be done on this Data Set so as to prevent the Suicide Rate in Future by taking preventive measures against Suicide Act by classifying the major cause of suicide that has occur in different set of states of India during different years among different age group. The analysis carried out with dataset listed below will be of great help to young generation to reduce the amount of suicides that occur in day to day life by taking appropriate measures against them. Comparison will be carried in order to find out whether the Suicide Rate has increased or decreased in several years. Followed by answer to question whether it may increase or decrease in the next coming years.

	1	2	3	4	5	6	7	8	9	10	11
26	ANDHRA F	2001	Causes Nc	35	234	254	107	47	677	17	155
27	ANDHRA F	2001	Other Cau	54	280	293	285	48	960	61	196
28	ANDHRA F	2001	Total	173	2058	2213	1566	369	6379	185	1871
29	ANDHRA F	2001	Total Illne	49	600	791	689	181	2310	54	534
30	ANDHRA F	2002	Bankruptc	0	26	56	40	6	128	0	4
31	ANDHRA F	2002	Suspectec	0	18	10	3	0	31	0	12
32	ANDHRA F	2002	Cancellati	0	19	4	1	0	24	0	24
33	ANDHRA F	2002	Not havin	0	2	2	0	0	4	0	12
34	ANDHRA F	2002	Illness (Ai	3	47	99	44	18	211	2	43
35	ANDHRA F	2002	Cancer	1	2	10	18	14	45	0	6

Fig 3. Data Set of Suicide in India State Wise

B. Techniques for Implementation

1. Linear Regression

Linear Regression Technique is basically a Linear Approach which is used in order to model the relationship that exists between scalar dependent variables and also between variables more than pone termed as independent variable.

Equation of Linear Regression Form expressed as-

$$yi = \beta o1 + \beta 1 xi1 + \cdots + \beta pxip + \varepsilon i, i = 1, \ldots, n \qquad (1)$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$\boldsymbol{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Fig.4. Vector form of Linear Regression

2. Decision Tree

It is basically a flow chart. It consists of internal nodes that represent sets of tests that are performed on an attribute; set of branches basically represents the outcome from the tests performed on the attribute, leaf node present depict decisions that are to be performed after the computations on the attributes. There are classification rules that are represented with the help of path that exists between root node and the leaf node [9]. It has three nodes such as:- Decision Node, Chances Node and End Node.

1. Decision nodes –they have square representation.

2. Chance nodes –these nodes are represented with the help of circles.

3. End nodes – these nodes are represented with the help of triangle.

Equation representing Classification under Decision Tree is given below-

$$I(s1, s2, \dots, sm) = -\sum_{i=1}^{m} pi\ log2(pi) \quad (1)$$
$$pi = si/s$$

Equation representing Decision Tree partitioning by Attributes

$$E(Ai) = \sum_{j=1}^{q} \frac{sij + sij + \cdots + smj}{s} I(sij, \dots smj)$$
$$I(s1j, s2j, \dots, smj) = -\sum_{i=1}^{m} pij\ log2(pij) \quad (2)$$

3. Naïve Bayesian Network

This technique is very simple and is basically used as a classifier. It focuses on the classification of a certain conditional problem represented by vector having n features. Equation representing Bayesian classifier

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (1)$$

Equation representing Bayesian classifier with Conditional Probability

$$P(x1, x2, \dots., xn) = \prod_{i=1} n\ P(xi|parents(yi)) \quad (2)$$

C. Implementation

Firstly sum of people committing suicide under a certain age group is calculated using Linear Regression Technique; then again Linear Regression is used to calculate suicide committed by individual of certain age under a particular cause followed by Regression to find out at which state and at which year. Moreover a manual matrix will be prepared to classify the preventive measures on the basis of several parameters of Suicides. Then the analysis obtained after several Regression will be classified with the help of Decision Tree and Naïve Bayesian Network in order to know what preventive measures need to be taken in order to prevent Suicide from happening in Future. Moreover Results obtained will further be checked for proof of correctness with the of Precision and Recall Technique

IV. CONSTRAINTS

There were several constraints that were encountered some of them are listed below such as main challenge was the collection of Data Set followed by cleaning of Data Set that require removal of null data or redundant data from the data set. Another challenge was to purpose a new idea to carry out something effective and yet different from other analysis. Moreover the classification of data set also requires a detailed study of algorithms such as Decision Tree, Naïve Bayesian Network which is also a challenging task.

V. CONCLUSION

It has been discussed about the several causes of Suicides that take place in different states of India under different age group in different year. Also the Analysis that will be carried out will help to reduce the rate of Suicide that takes place in different states due different cause by providing those ways to prevent it in future. As the results obtained through the analysis will help the government, individual and an organization by providing them with an area that requires an improvement to be made in order to prevent suicide in Future. Moreover in order to have a proof of correctness. Precision and Recall Technique will be used which is indirectly related to reverse Engineering.

VI. FUTURE SCOPE

The analysis carried out will help the youth from taking preventive measures against suicide so as to reduce the rate of suicide and suicide attempts in Future. The analysis carried out will also provide knowledge about areas of improvement to the government. So that effective steps can be taken by the government and organization in order to prevent suicides in India.

VII. REFERENCES

[1] Dr Amitendu Palit, "Suicides in India: The Economics at Work", ISAS Insights No. 131, 25 August 2011.

[2] Gupta S.C. and Singh H. Psychiatric illness in suicide attempters; Indian Journal of Psychiatry1981; 23 (1):69-74.

[3] Marouane Birjali, Abderrahim Beni-Hssane1,and Mohammed Erritali, "Prediction of Suicidal Ideation in Twitter Data using Machine Learning algorithms", ACIT,2016.

[4] Dr. Pooja Rastogi, "Suicide in Youth: Shifting Paradigm", ISSN 0971-0973.

[5] Sharareh R. Niakan Kalhori, Xiao-Jun Zeng "Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course", Journal of Intelligent Learning Systems and Applications, 2013.

[6] Marouane Birjali, Abderrahim Beni-Hssane, and MohammedErritali, "Prediction of Suicidal Ideation in Twitter Data using Machine Learning algorithms", ACIT,2016..

[7] Jian Hua Yeh, FeiJieJoung, Jia Chi Lin, "CDV Index: A Validity Index for Better Clustering Quality Measurement", Journal of Computer and Communications, 2014, 2, 163-171

[8] Cha, C. B., Najmi, S., Park, J. M., Finn, C. T., & Nock, M. K.. "Attentional bias toward suicide-related stimuli Predicts suicidal behaviour". Journal of Abnormal Psychology, 119, 616–622, 2010.

[9] Delgado-Gomez, D., Blasco-Fontecilla, H., Sukno, F., Socorro Ramos-Plasencia, M., & Baca-Garcia, E. "Suicide attempters classification: Toward predictive models of suicidal behaviour", Neurocomputing, 92, 3–8, 2012.

[10] Fox, K. R., Franklin, J. C., Ribeiro, J. D., Kleiman, E. M., Bentley, K. H., & Nock, M. K., "Meta-analysis of risk factors for nonsuicidal self-injury",Clinical Psychology Review, 42, 156–167,2015.

[11] Phillips M.R., Yang G., Zhang Y et al. Risk Factors for suicide in China: A National Case Control Psychological Autopsy Study, Lancet 2002; 360: 1728-1736.
[12] Vijay Kumar L. Psycho Social Risk Factors for Suicide in India and Suicide prevention- Meeting the challenges together, Orient Longman-2003; 49-162.
[13] Jancloes M. The Poorest First: W.H.O. activities to help the people in greatest needs 1998; 19(2: 182-187).