Analysis of Suicide Database to Reduce Number of Suicides in India

Yuvraj Singh Chouhan [1], Suraj Kaushik [2], Nagendra Sharma [3], Shreyansh Singh [4], Shamela Rizwana [5]
[1, 2, 3, 4] Department of Computer Science and Engineering,SRM Institute of Science and Technology, Ramapuram, Chennai, India.
[5] Assistant Professor, Department of Computer Science and Engineering,SRM Institute of Science and Technology, Ramapuram, Chennai, India.

Abstract – A study is presented at analyzing the major factors that affect the number of suicides in different parts of India from year 2000 to 2012 and subsequently using them to predict the number of suicides in the future in different parts of India. By analyzing the data and predicting the major causes of suicides it can help government to know which part of population is most affected by this problem so that government can take the required steps to reduce them. The Indian government keeps the database of each suicides that happens in India. Along with the age-group, cause of death, state of victim. This data was made public by crime branch bureau of the data analytics purpose. Relationship will be made between the different features of suicide so that a linear relationship can be formed and then linear regression will be used to develop a model for the prediction of number of suicides in rear future. Through this study the Indian government will come to know that which part of population is most affected by suicides so that government can work on preventive measures for different parts of the country.

Index Terms – Machine Learning, Prediction, Data Analysis.

1. INTRODUCTION

Suicides is one of the major problem that Indian government is facing. About 800000 people commit suicide worldwide every year, of these 17% are residents of India. The male to female suicide ratio has been about 2:1 in India. On an average a total number of suicides in India per day is 300. According to the Suicides reports in India and National Crime Records Bureau, the total number of suicides in India as per 2014 statistics is 1,09,456. Suicide is the act of deliberately killing oneself or, more specifically, an act deliberately initiated and performed by the person concerned in the full knowledge, or expectation, of its fatal outcome. The Indian Government classifies a death as suicide if it meets the following three criteria: it is an unnatural death, the intent to die originated within the person, there is a reason for the person to end his or her life. In most cases the person writes the reason in a suicide note or it remains unspecified.

Suicide prevention is a term for the collective efforts of local citizen organizations, health professionals and related professionals to reduce the incidence of suicide. Other than direct interventions to stop an impending suicide, methods also involve treating the psychological symptoms of depression, providing counselling to the person, improving the coping strategies of persons who would otherwise seriously consider suicide, reducing the prevalence of conditions believed to constitute risk factors for suicide, and giving people hope for a better life by resolving current problems.

The first step in public health approach to suicide prevention is to identify those who are at the risk of suicide attempts. To identify this that which population is at greater risk this study would be useful. This can be done by making a correlation between different features and the number of people committing suicides. The relationship can be used to develop a linear relationship so that number of suicides due to a particular cause in particular age group within a particular state can be predicted.

2. DATABASE CREATION

A dataset is a collection of data. Mostly a data set refers to the contents of a single database table, or a single statistical data matrix, where every table column represents a particular variable, and each table row corresponds to a given member of the data set. The objective of project is to find the relations between the dataset to predict the future dataset.

The data set used in this project contains yearly suicide detail of all the states/u.t of India by various parameters from 2001 to 2012. This data is real and National Crime Records Bureau (NCRB), Govt of India has shared this dataset under Govt. Open Data License - India. NCRB has also shared the historical data on their website. The data contains various fields like State, year, reason, gender, age group and total.

There are various reasons due to which a person may commit suicide. It may be due to education burden, family problem, financial problem, health status, etc. There may be certain states where particular reason for suicide is higher than other reason. We need to find relation between reason of suicide and the state in which it is committed. So that government will focus on particular reason resulting in minimizing that factor of suicide. For example, if in particular state the maximum suicides are committed by farmers due to low agriculture production or higher input and lower output in agriculture than

the government can focus more on particular field thus making more schemes that would help farmers of particular state resulting in reduced number of suicides in that state.

Another feature is age group. The dataset is divided in different age groups. So that we can find relationship between age and number of suicides which helps us to know that particular age group population is mostly affected by which suicide reason. For example, if in particular age group the major reason of suicide is failure in examination than more counselling institutes should be open in that state to do counselling of all people belonging to that age group resulting in reduction of suicides

The existing considered these states: -

- Maharashtra
- West Bengal
- Tamil Nadu
- Andhra Pradesh
- Karnataka
- Kerala
- Madhya Pradesh
- Gujarat
- Rajasthan
- Uttar Pradesh
- Punjab
- Bihar

While the proposed system considers some extra places too: -

- Chhattisgarh
- Odisha
- Assam
- Haryana
- Delhi (Ut)
- Jharkhand
- Tripura
- Puducherry
- Himachal Pradesh
- Uttarakhand
- Goa
- Jammu & Kashmir
- Sikkim
- A & N Islands
- Arunachal Pradesh
- Meghalaya
- Chandigarh
- Mizoram
- D & N Haveli
- Manipur
- Nagaland
- Daman & Diu
- Lakshadweep

3. EXISTING SYSTEM AND DRAWBACKS

Existing System: -

- Analysis of Suicide Victim Data for the Prediction of Number of Suicides in India (Publisher IEEE 2017)
- Suicide Victim Data was analyzed for the Prediction of Number of Suicides in India.
- Linear regression algorithm was used for prediction.
- Data of different places was used for prediction.
- Data set used was of made public by crime branch bureau.

Drawbacks of Existing system: -

- The existing system has considered only 12 states and given results on basis of them.
- Union Territories were not taken into consideration during study.
- Age group between 15-29 was only considered for study.
- No proper simulation and results were explained.
- Only 2011 data has been used for prediction which is less efficient.

4. TECHNOLOGY USED

Machine Learning

Machine learning is a study of computer science that provides computers the ability to learn without being explicitly programmed. Machine learning is used to study algorithms that learn from and make predictions on data. Machine learning is related to computational statistics, which also focuses on prediction making.

Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction; in commercial use, this is known as predictive analytics. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The learning process begins with observations or data, examples, direct experience, or instruction, in order to find patterns in data and make better decisions in the future based on the examples provided. The objective is to allow the computers learn automatically without human assistance and adjust actions accordingly.

5. MODULES

5.1 Data analytics

Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making. Data mining is a data analysis technique which focuses on modeling and knowledge discovery typically for predicting the future. Data analytics is performed on the suicide database so that the data can be cleaned and data that is not required can be deleted. It is used to model the complex suicide data to a simpler form so that it can be used as input for prediction process.

It refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain. Data is categorized and extracted to identify and analyze similar behavioral data and patterns, and techniques vary according to organizational requirements. Data analysis is linked to data visualization. It is used to make relationship between different columns of database so that it can be used for prediction and visualization.

Data visualization is the way of presenting data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, to understand difficult concepts or to identify new patterns. Visualization helps to make charts and graphs for more detail, thereby changing what data you see and how it's processed. This helps us to understand which features of data have strong relations between them.

5.2 Linear Regression

Linear regression is a linear approach used for modelling the relationship between a scalar dependent variable y and one or more independent variables denoted X.

Linear regression has many practical uses. If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of y and X values. After developing this model, when an additional value of X is given without its accompanying value of *y*, the model can be used for prediction of the value of *y*.

Given number of variables X_1 and a variable y , ..., X_p that may be related to *y*, linear regression analysis can be applied to quantify the strength of the relationship between *y* and the X_j, to assess which X_j may have no relationship with *y* at all, and to identify which subsets of the X_j contain redundant information about *y*.

With simple linear regression as shown in Fig1 we can model our data as follows:

- y = C + M* X
- Here y is the output variable we want to predict, X is the input variable we know and C and M are coefficients we need to estimate.
- C is called the intercept because it determines where the line intercepts the y axis. The M term is called the slope because it defines the slope of the line or how X translates into a y value before we add our bias.

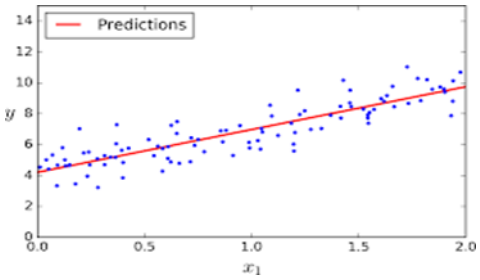

Fig 1: Linear Regression

6. RESULTS AND DISCUSSION

Number of suicides appear to be concentrated towards Lower education level. Most of the people who have committed suicides have education level below Matriculate/Secondary. This is represented in fig 2.

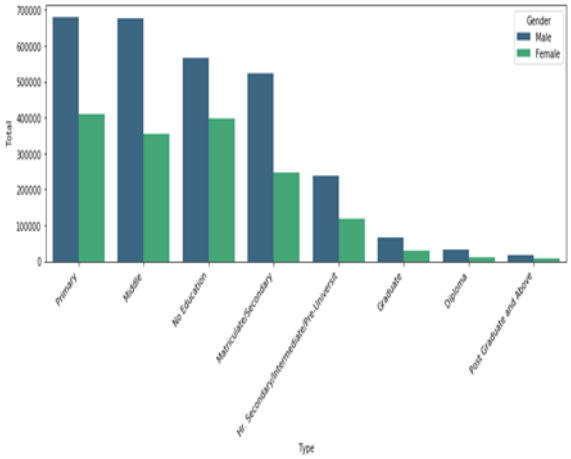

Fig 2: Bar Graph representing Number of suicides of different education level people

- Social status also tends to have great affect over suicides. Married people commit more suicides as compared to never married, widower, separated and divorced.
- While most of the causes of the suicides are not known, the three major causes among the known cases are Family problems, prolonged illness and mental illness. This is represented in fig 3

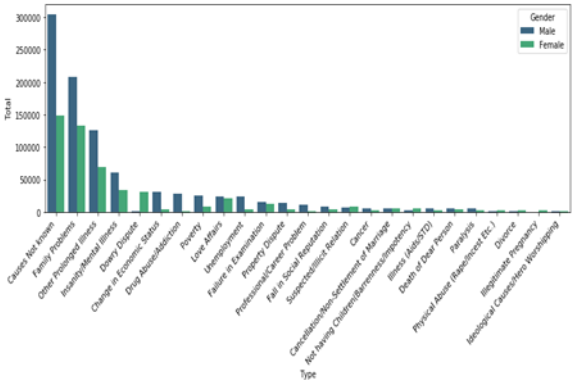

Fig 3: Bar Graph representing Number of suicides because of different causes

- According to the data men seem to be badly effected by unemployment, property dispute, poverty, drug abuse or addiction and change in economic status than women.
- Another major concern is number of women who have committed suicides due to dowry disputes is much higher when compared to men. This is represented in fig 4.

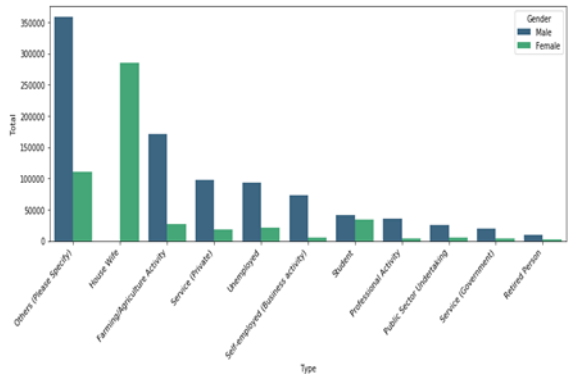

Fig 4: Bar Graph representing Number of suicides because of dowry disputes

- According to the data, most of the females who have committed suicides are house wives. The percentage of female suicides in all other categories is much lower than the percentage of male suicides. While this is an indication of lower representation by women in professional careers, it also reiterates the importance of girl education and women empowerment. The graph shows that financially independent women are much mentally stronger.
- It is disheartening to see that farmers who feed the rest of the country are the ones who are more committing suicides than any other profession. Followed by farmers, it's the unemployed and private sector employees who are most effected. It's surprising to know that the number of suicides among the unemployed and the private sector employees is almost the same. This also might be due to higher pressure in private sector when compared to government sector jobs. This is represented in fig 5.

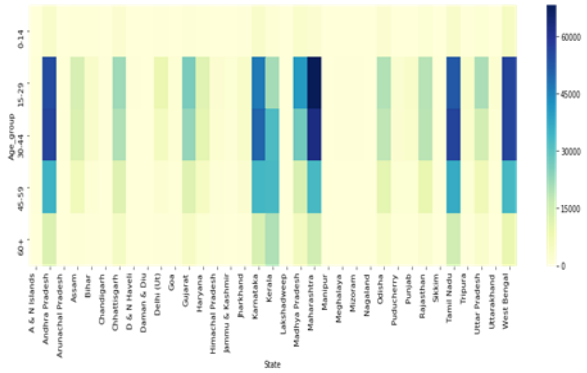

Fig 5: Heat Map representing Number of suicides of different age group in different states

- Based on state and age group we can see 15-29 is the most vulnerable age in all states except Kerala. Maharashtra is the state with most number of suicides.
- Union territories whose area is much smaller compared other states have higher number of suicides per square kilometer. Again, Kerala is an exception here as it larger compared all other union territories. Kerala is followed by West Bengal and Tamil Nadu among the Indian states to register higher number of suicides. This is represented in fig 6.

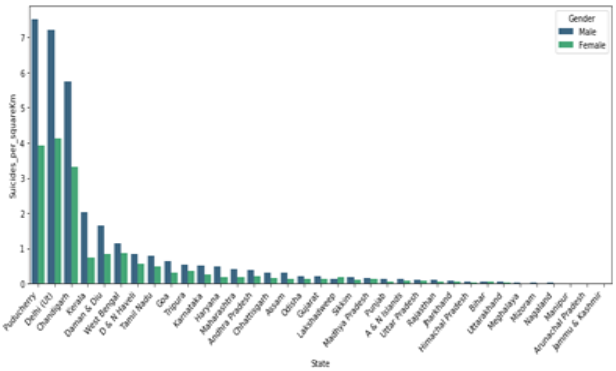

Fig 6: Bar Graph representing Number of suicides in different states per square kilometer

- The total number of suicides in the country are increasing with time. From 2001 to 2012 the percentage increase in suicides is 24.8% which is very scary. This is represented in fig 7.

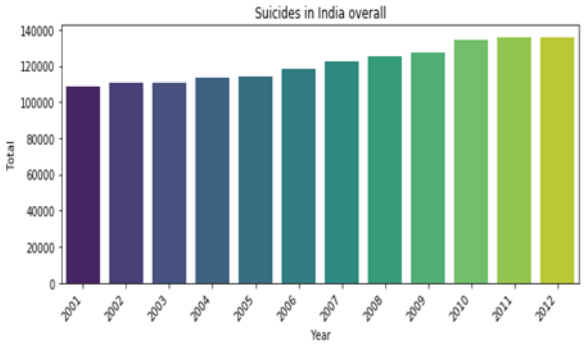

Fig 7: Bar Graph representing Total Number of suicides increasing with time

7. CONCLUSION

The results obtained gives us the clear vision about what type of population is highly affected by this problem. Government should take the preventive measures in bringing down number of suicides in our country by giving more attention on the population that is mostly affected in their respective states. It's not only the government but even it is job of us also to work hand in hand with the government and help them in reducing suicides from our nation by providing counselling to the population of the respective states who are greatly affected by it.

In future by this study for any group of population in given states the number of suicides can be predicted. This can be used as a reference for evaluating the effectiveness of the preventive measures and policies that government took for reduction of suicides.

REFERENCES

[1] CDC, "Suicide Facts at a Glance", online 2012
[2] Lakshmi Vijaykumar, "Suicide and its prevention: The urgent need in India" , Indian Journal of Psychiatry, 2007 Apr Jun;49(2):81 84
[3] Omprakash Mandge, "A Data Mining Tool for Prediction of Suicides among Suicides", National Conference on New Horizonsin IT (NCNHIT) 2013
[4] Census India 2011
[5] National Crime Reports Bureau, ADSI Report Annual 2014 Government of India, p. 242, table 2.11
[6] David A. Freedman, Statistical Models: Theory and Practice, Cambridge University Press. P. 26, 2009.Richard Taylor, Interpretation of the correlation coefficient:A Basic Review , JDMS 1:35-39, January/February 1990
[7] Lomax, Richard G. (2007). Statistical Concepts: A Second Course. p. 10. ISBN 0-8058-5850-4
[8] Gareth James, Daniela Witten, Trevor Hastie Robert Tibshirani (8th printing 2017). "An Introduction to Statistical Learning". ISBN 978-1-4614-7138-7 (eBook)