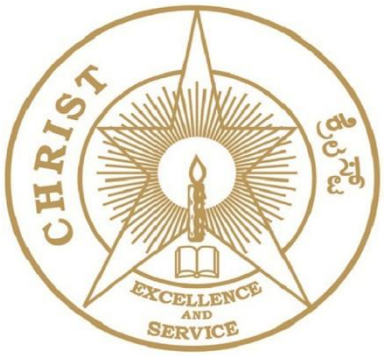

CHRIST

(DEEMED TO BE UNIVERSITY)

BANGALORE · INDIA

Principles of Data Science

MDS133

CIA 3

Case Study Report and Presentation

Analysis of Suicides in India from 2001 to 2012 and Prediction

Harsh Seksaria (2048011)

Harshita (2048035)

Jerlyn S (2048037)

TABLE OF CONTENTS

1. INTRODUCTION

Suicide is an essential issue in India and is not rare. More than a lakh people end their lives every year. Now and then, we get to hear about new people committing suicide. Be it school students, teenagers, or an adult individual, not only one group commits suicide. We always get to hear that a student has committed suicide because he/she got fewer marks or failed in a subject, a farmer committing suicide because of a loan and other problems, etc. The Government of India classifies a death as suicide if it is an unnatural death, the intent to die originated within the person, and there's a reason for the person to end his or her life, which may be specified in a suicide note or unspecified. According to an article in The Hindu, published on January 29, 2020, it claims that students' suicides were rising and that 28 lives were lost every day. Almost 800000 people die due to suicide worldwide every year, out of which 17% are Indians.

In this project, we analyze various reasons people have committed suicide under different age groups with the dataset of suicide records from 2001 to 2012 using multiple computational methods like machine learning, visualization, etc., and predict the number of cases that can arise a given year. We analyze the recorded cases of suicide to get the inferences rather than explore the various factors under which a person commits suicide.

2. PROBLEM STATEMENT

Suicide is a topic that needs severe attention, especially in the Indian context. Although suicide is a deeply personal and individual act, many societal factors may be affected. We analyze the dataset to get various answers like the gender that tends more to suicide, means adopted to commit suicide, the number of suicide due to a particular reason, etc.

3. METHODOLOGY

This section represents the research methodology that has been developed for this particular study. Machine learning is a programming technique which provides a computer the ability to learn automatically and improve from that knowledge and experience without being explicitly programmed to do so. Various machine learning methods help us to create a workable data model. The ML method that we have incorporated in our analysis is the Supervised Learning methodology. Supervised Learning is based on the outcomes of a similar process in the past; i.e., it predicts an outcome based on historical patterns.

To understand the behaviour and trend in suicide data, two logical methods effectively deal with our research problem. Following are the two approaches for our research work:

1. **Descriptive and Statistic Approach**: This method is used to find out the pattern of suicides concerning age group, gender, marital status, social status, education, along with professional occupation.
2. **Predictive Approach**: In this method, data will be used to generate a model to predict the total number of cases for every Year for a particular state by utilizing the information present in the existing data.

Machine Learning steps

- **Data Acquisition**: We gathered the data from Kaggle.
  The dataset contains the yearly suicide detail of India's states and union territories by various parameters from 2001 to 2012. It is crucial to identify the attributes and characteristics present in a dataset. The dataset contains the information of total suicide in a particular state and other meaningful information, which is as follows: State: This column includes the name of the state in India like West Bengal, Andhra Pradesh, etc.
  Year: This column contains information from 2001 - 2012.
  i) State: This column contains the name of the state in India like West Bengal, Andhra Pradesh, etc. The
  ii) Year: This column contains information from 2001 - 2012
  iii) Gender: This column specifies the gender of the person.
  iv) Age Group: This column contains different age groups from 0-14 to 60+.

v) Total: This column contains the sum of the total number of suicides in a particular state according to its gender, age, and state.
vi) Type/Cause: It tells us about the reason for attempting suicide like illness, Family Problems, unemployment, etc.
vii) Type/Marital Status: It represents whether the person who committed suicide was married, unmarried, divorced, or a widow.
viii) Type/Professional Occupation: It represents whether the victim was a student, employer, housewife, etc.
ix) Type/Educational Level: It contains information regarding the educational background of a victim

- **Data Pre-Processing**

In the data pre-processing step, we alter the data to convert it into a more usable and suitable format to suit our needs. Here, we convert the raw dataset into a cleaned dataset. There are multiple tools to prepare our data for use, such as Python, Tableau Prep, etc. In our analysis, we are using Python for the pre-processing data purpose. We perform the following operations on data to prepare it for analysis:

i) No missing values were found
ii) Removed the records where age is 0-100+ because it's unclear
iii) Then we remove the rows where the reason for suicide is Illegitimate Pregnancy for males because it is obviously illogical
iv) Also, the values of age as 60+ is removed and replaced with a random integer between 60 and 100
v) Removed rows where no reason is specified for suicide
vi) Removed rows where no profession of the person is given
vii) Removed the rows where gender for housewives was given as males, which again is illogical

In the raw dataset, we had 237519 rows, but after all the pre-processing and cleaning, we were left with 215163 rows. The benefit here is that whatever analysis or prediction we do will be close to accurate and will not lead to inconsistencies due to illogical data rows.

- **Exploratory Data Analysis (EDA)**

The Exploratory Data Analysis uses more in-depth insight into the data with graphical techniques and interactions between the variables. This step is basically about exploring the data, which means getting to know the data better. We summarize the characteristics of

data usually by plotting on some graph. We used Python to perform our Exploratory Data Analysis.

i. The dataset has more than 2 lakh rows and seven columns, namely, State, Year, Type Code, Type, Gender, Age Group, and Total.
ii. The State column contains the names of the states of India. The year column contains the years 2001-2012. We also find the unique states in the dataset.
iii. Type Code and Type are related to reason and type of suicide.
iv. The total column contains the number of people who committed suicide in that Year belonging to that state due to stated reasons in that row.
v. The dataset info gives us the total number of columns, their data types, memory consumption of the data, etc.
vi. To get a visual of the data, we plotted a line graph for the number of suicide cases in 12 years. We infer that the number of cases was on a constant rise till 2011 and was at peak between 2010 and 2011. The number of people committing suicide started decreasing in 2012. The Year 2011 recorded the highest number of suicides with 360331 cases. This is a 23.84% increase compared to 2001.
vii. We also found that twice the males than females committed suicide due to various reasons.
viii. Maharashtra had the highest number of recorded suicide cases followed by Andhra Pradesh and Tamil Nadu.
ix. Unemployment is a huge problem in our country. Andaman and Nicobar Islands had the most suicide cases due to unemployment.
x. We plotted the distribution of males and females in different age groups who committed suicide and found out that males in the age group 30-44 were almost twice in number as females and more than double in the age group 45-59. In the age group 0-14, the number of males and females were almost the same.
xi. Hanging was the most chosen means to commit suicide, which has taken over 4,00,000 lives. Family problems were the primary cause of committing suicide, followed by prolonged illness, mental illness, and love affairs. Upon finding the causes where females committed more suicide than men, dowry dispute, divorce, and illegitimate pregnancy were the primary reasons.

- **Model Building and Prediction**

In the dataset, we tend to predict the values for a year already in dataset to verify with the value we have and predict the number of cases in future years based on historical data provided using the Linear Regression model of Supervised learning methodology.

We have used linear regression for predicting the total number of cases for every Year for a particular state by utilizing the information present in the existing data. Linear regression is a linear approach used for modeling the relationship between a scalar dependent variable y and at least one independent variable denoted X. Linear regression has many practical uses. If the goal is prediction or forecasting, linear regression can be used to fit a predictive model to an observed dataset 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 to predict the value of y.

- **Result Interpretation**

We got multiple observations and inferences from all the EDA and analysis parts we did, and they provided some intriguing information.

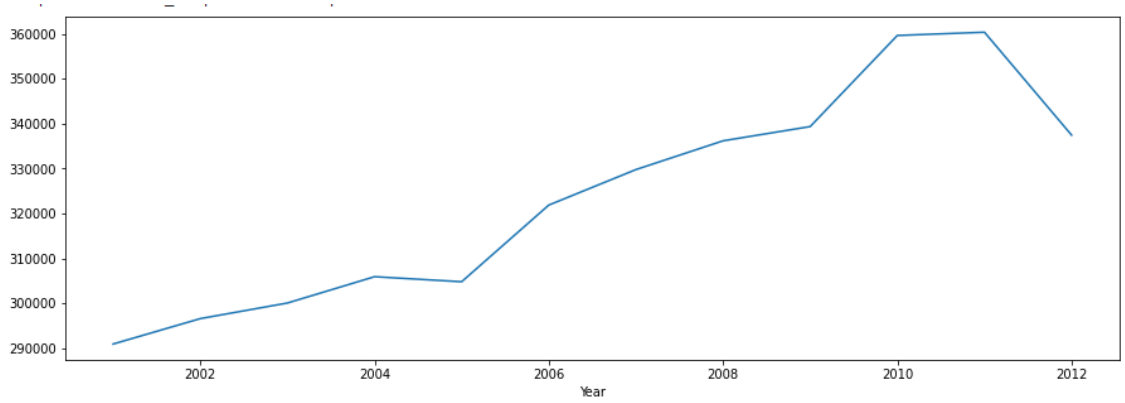

Figure 1: Line Graph Suicides per Year

The above graph shows the number of cases each year. The number of cases was on a constant rise till 2011 and was at peak between 2010 and 2011. The number of people committing suicide started decreasing in 2012.

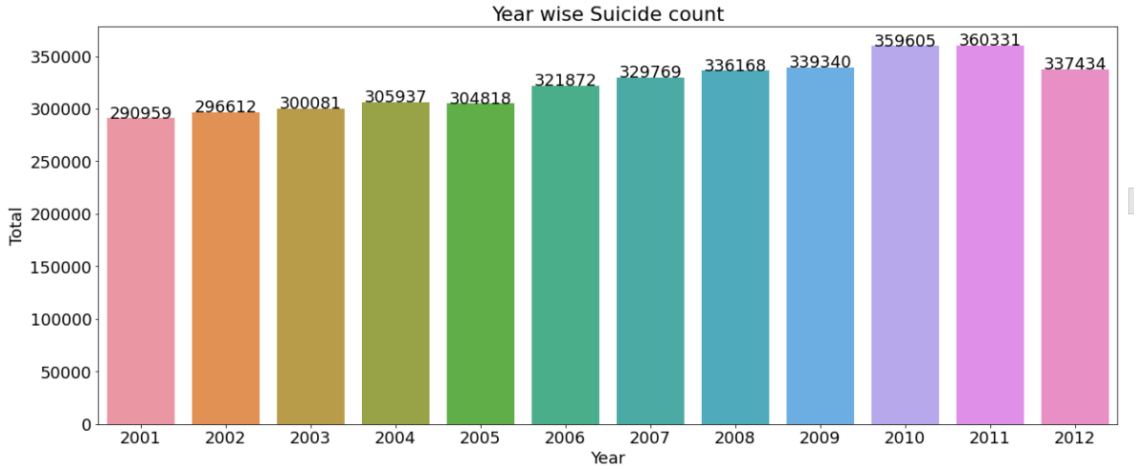

Figure 2: Exact number of cases per Year

The above graph shows a histogram for the number of suicides every year.

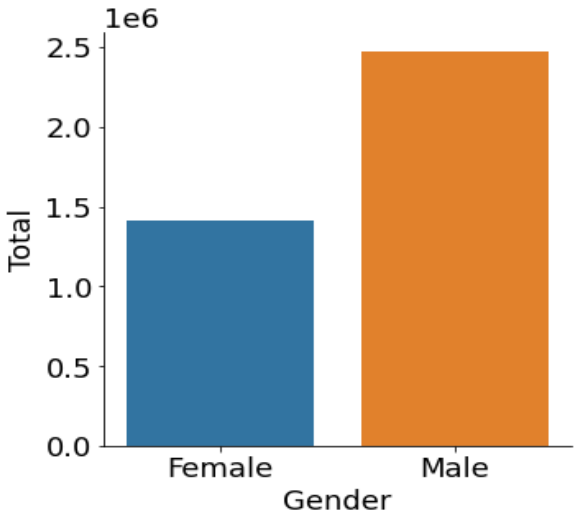

Figure 3: Number of female vs. male who committed suicide

The above graph plots the number of males and females who committed suicide, and we find that twice as males as females commit suicide.

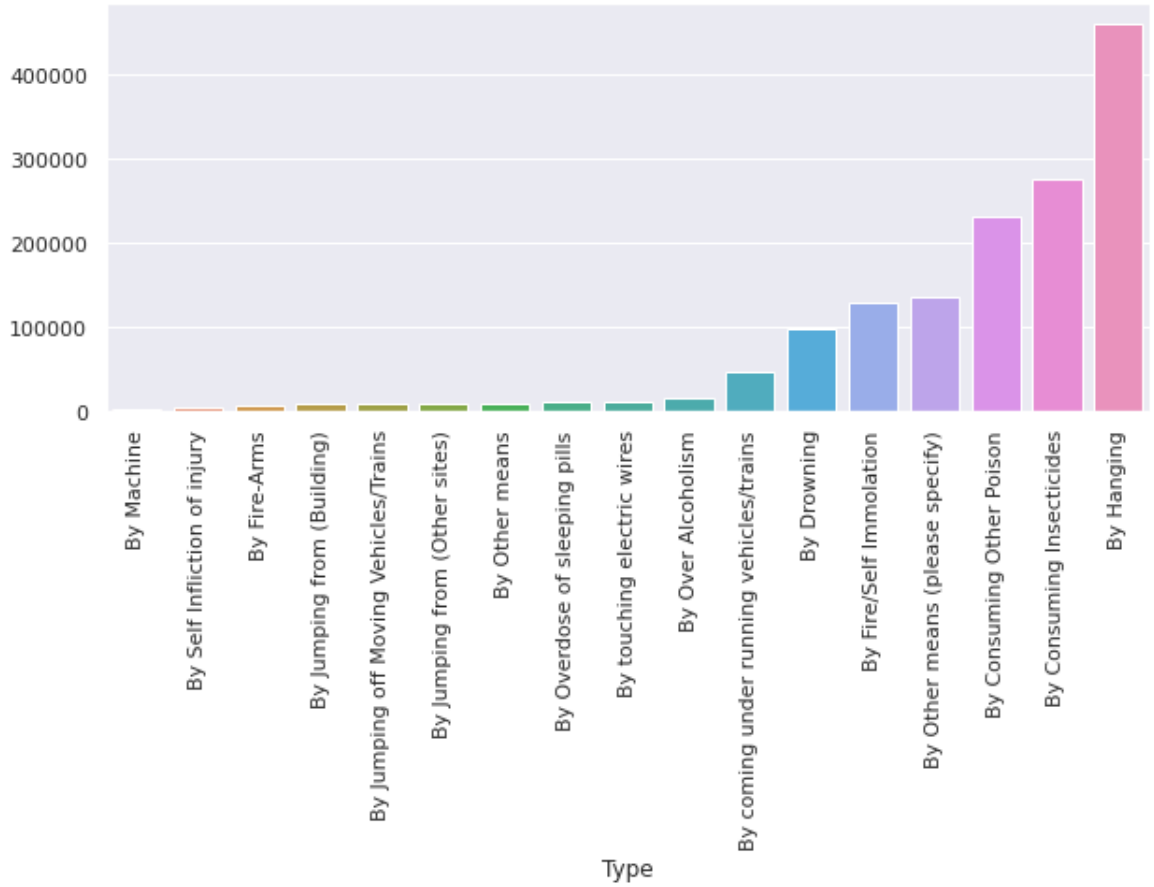

Figure 4: Number of suicides by a particular means

The above graph indicates the number of people who committed suicide by the stated means. We can observe that hanging was the very general used means of suicide.

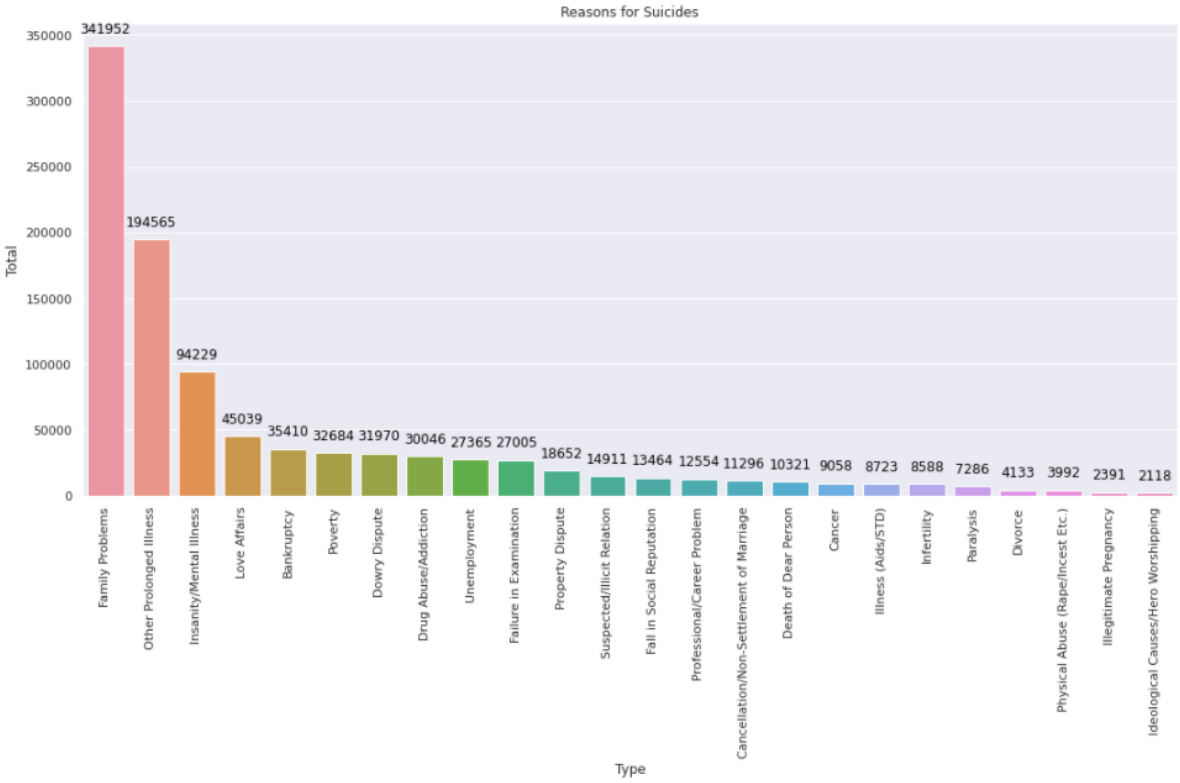

Figure 5: Number of suicides due to particular reasons

The above figure shows number of suicides due to given reason.

4. CONCLUSION

This study was carried out to analyse the suicide trend in India and the major influencers for the increased suicide rate in recent times. As this study was limited to suicide cases from year 2001 to 2012, we were able to predict the future trends with the application of various data science methodology. Data science has helped us to concentrate on broader aspects that can act as a great aid for our analysis. ML application was used in data preparation like data gathering, filtering, and cleaning data. Exploratory Data Analysis was done to perform initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics, graphical representations and other valuable insights. It also helped us summarize the characteristics of data and to plot trends. We incorporated model building to construct a workable model that was used to predict conclusions. Supervised learning was used to train the labelled data to the model. It also helped us to predict outcomes for unforeseen data as our dataset consisted of data pertaining to suicide cases from year 2001 to 2012. Linear regression methodology was used to predict the total count of suicide each year, and by implementing our model we obtained almost similar actual and the predicted values which resulted in maximum accuracy. Thereby were able to draw suicide counts for the future years. We also drew various inferences which would help a great way in placing a full stop to the growing trend of suicide rate in India.

5. REFERENCES

1. https://www.kaggle.com/rajanand/suicides-in-india
2. https://www.thehindu.com/news/national/student-suicides-rising-28-lives-lost-every-day/article30685085.ece
3. https://en.wikipedia.org/wiki/Suicide_in_India#:~:text=In%202016%20the%20number%20of,with%2017.5%25%20of%20world%20population.
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2917089/