Predictive Modelling of Suicide Rate Using Probabilistic Methods

Ranu Parate
x17161452
MSc in Data Analytics

8th April 2019

Abstract

Suicide is basically an act of hurting and causing one's own death. Suicide is the most painful area and global phenomenon around the world, it has been considered as a genuine health problem. Over the past decades, the suicide rate around the world has increased and it is still increasing. According to the World Health Organization, various approaches for preventing suicide rates has been launched. Though, there is not much change in the suicide rates. India is the country where the rate of suicide is increasing enormously. There are different causes due to which the individual commits suicide and this keeps on varying from individual to individual. There is much more need for predicting the suicide rate and for analyzing the various causes and types of suicide in order to increase the survival rate. However, many researchers have done research on the suicide rate but still, the rate of suicide has not decreased. In this study, the main objective is to forecast the suicide rate of India for the year 2001 to the year 2014 according to the gender, age, types, and causes of suicides using the time series analysis (ARIMA and ARIMAX) and the probabilistic techniques (SVM, SVR, and RF).

Keyword: suicide, prediction, India, ARIMA, ARIMAX, SVM, SVR, RF, data mining,regression, classification algorithm

Contents

1 Introduction

Across the world, suicide is one of the major causes of death among all the age groups. For quite a while suicide has been viewed as a general health problem and medical issue OLIVEN (1954). To prevent the suicide many strategies were propelled worldwide WHO (2017), yet regardless of expanding efforts to diminish suicide through the enhanced treatment, alertness and awareness operations and many other services, the suicide rate has not improved over the decades. Apart from the strategies, many clinical instruments were efficiently used for suicide prediction Carter et al. (2017) by clinical instruments have not been observed to be helpful in terms of clinically while categorizing the individual with high-risk. This study will endeavor to predict and analyze the causes, factors and the age group of different years contributing to the suicide rate in India by applying time series techniques.

1.1 Background

Around the globe, death because of suicide is the crucial concern. Suicide is an individual misfortune that is expanding significantly influencing the lives of near and dear. According to the author, in the world, India is among the most contributors to increase the number of suicides (Varnik; 2012).

According to National Crime Records Bureau (NCRB), in the period of the year 2004 to the year 2014 there was an increase of 16 % in the number of suicides and for the same decade, the total population was increased by 15 % (NCRB; 2014).

Different states of India are contributing to the rise in suicide growth. As per the National Crime Records Bureau (NCRB), the top three states having the largest number of suicides are Maharashtra, Tamil Nadu, and West Bengal. Among these three states, Maharashtra has the highest incidents of suicide with 16,300, followed by Tamil Nadu with 16,100 and West Bengal with 14,300 (NCRB; 2014).

Suicides are committed by many causes. Some of the causes of suicides are Marriage Related Issues', Love Affairs', Drug Abuse/Addiction', Bankruptcy or Indebtedness', Failure in Examination', Unemployment', Poverty', Property Dispute' and Death of Dear Person'. Out of these major causes of suicide are illnesses' and Other Family Problems' (NCRB; 2014).

In this study, the analysis and prediction will be done by time series as there is a change in the rate of suicides over time. The suicide rate goes on increasing and decreasing as there is a change in the year.

1.2 Motivation

According to the researcher K.Sai Teja and Basha (2018), almost 5.5 percent of suicide victims are farmers and students. Due to the disappointing grades and failure in exam students are committing suicides and on the other hand farmers. Agriculture is considered as the backbone of India. Around 50 % population is dependent on agriculture directly or indirectly. The rise in the rate of suicides in India is increasing enormously. India includes 18 % of the total world population. Additionally across the world, 800,000 numbers of people entrust suicides every year out of which 17 percent belongs to India (Prashar and Choudhury; 2018).

In 2014, over-all 5,700 number of farmers have committed suicides, which contributes 4 percent of total suicides in the country (NCRB; 2014). The splitting of the farmer's suicide among the male and female is 90 percent and 10 percent respectively. The state contributing highest in the numbers of suicides in India are Maharashtra, Telangana and Madhya Pradesh (NCRB; 2014).

Suicide being the problematic part that needs to be resolved and needs to be prohibited because due to suicide there is a huge loss of resources, humans and the economy of the country. Thus, the key motto of this research will be to analyze and predict the factors and causes such as love affairs', family disputes', failure of crops' because of which suicide rate is increasing in various age groups considering students and farmers in a major way.

1.3 Area of Research and Value

Suicide is a general medical problem across the world. Many awareness camps and strategies to prevent suicide are created. Though, after creating all this the suicide rate has not decreased over the decades. Kunawut Boonkwang and Youdkang (2018)suicide rate got increased to 6.47 per 100,000 population in the year 2015 compared to the suicide rate in the year 2014 that was 6.07 per 100,000 population. Glen Coppersmith and Fine (2018) the suicide rates have increased by 24% in the United States from the last 20 years. Due to this increase in the suicide rates, suicide has contributed to the among the top 10 death causes. There are many different reasons due to which an individual commits suicide. Those reasons are unemployment, problems in the family, and the crisis that occurs among the individual.

> *Research Question : In comparison with other regression model will time series model give the more accurate result to predict the suicide rate? To what extent supervised machine learning techniques can provide the best accuracy for predicting the suicides rates?*

1.4 General Findings on this research

Most of the studies carried out in the past years show the analysis of the suicide rate. This analysis is done mostly on the age factor, gender, causes. Also, related to suicides the research on the classification of text is done. This text includes the post and comments provided on social media. Amayas Abboute and PONCELET (2014) classified the tweets that are related to the suicides. The clinical data plays an important role in the suicide analysis. Some of the researchers have provided the analysis and classification of the clinical data. The research is done on the probability of suicide death by using baseline characteristics (Soo Beom Choi and Kim; 2018). Ryu S (2018) developed a model to predict the suicidal ideation of the people considering the total population of Korea. Therefore, this study will predict the suicide rates of Indian States based on age, gender, types and caused.

2 Literature Review

Different research has been carried out utilizing the various methods of probabilistic. Literature reviews are divided into 5 different parts: 1) How Data Mining helps in prediction of suicide among different age group, 2) Review of the suicidal ideation, 3) Identifying sentiments of suicidal notes and clinical data, 4) Classification of post, messages and tweets related to suicides on social media and 5) Review on the main reason and the way of committing suicides.

2.1 Data Mining Techniques used in Prediction of Suicides

Different data mining techniques and procedures are created by the data scientist for predicting the suicide rate and help in increasing the survival rate around the world.

K.Sai Teja and Basha (2018) has mainly focused to identify the factors and causes in committing suicides between the farmers and students in India. The research has also focused on the age group of 0-14 and 15-29. The dataset used in this research is the suicide dataset which has around 2, 00,000 observation and seven variables. The data gives the information about the number of suicides in different states of India, in different age groups and it also shows the causes and type of suicide, that is, in the way the suicide is committed. In this research K.Sai Teja and Basha (2018) the case studies are done on the BCom first year student committing suicides and on the students of IIT Delhi committing suicides. Various machines learning model has applied in the data and the performance of the model is evaluated stating the model whose performance is best. Following are the techniques used in this research: Random Forest, Logistic Regression, Decision Tree, Nave Bayes, and ADA Boost. After applying the techniques, it is concluded that the Random Forest algorithm gives the highest accuracy with 93.18 % and this model can be used for the prediction. Here two cases are considered for predicting the suicide rate, the first case is by considering the type of doing the suicide and the other case is prediction by the age group. Here the researchers have suggested that suicides can be controlled by endlessly improving the suicides thoughts and also by improving the etiological thought of an individual problem. The conclusion made in this research is that most of the suicides are committed by the people having age group 15-29.

Furthermore, Prashar and Choudhury (2018) main objective of the study is to categorize the various types and causes of suicides in various states of India. The data is collected from Kaggle and data.gov.in sites. The categorizing of the data is also done on various parameters like Age and Gender. Further, the research also gives information about the means like Drugs, Hanging, Jumping, Alcohol used by an individual for committing suicide. Data used for this research consists of the field like State, Year, Type, Causes, Gender, Age. The comparison is made in this research to check whether the suicide rates have increased or decreased among the years. The linear regression is used for the prediction of the suicide rate considering the age, year and the causes for committing suicide. Moreover, Decision Tree and Nave Bayesian Network were used for further analysis to know the steps required to take to prevent suicide. Suicide is the most serious problem and the major cause of death. In Brazil, 5.5% is the suicide rate for the year 2015.

Rodrigues CD and TCRO (2018) researchers have built a model to analyze the time trends of suicide rates in Brazil and to estimate the rate for the year 2020. The main objective of the research is to analyze the suicide rate for the year 1997 to the year 2015 respective of sex and the age group. To gather the data the study was carried out to analyze more than 200 units. The data regarding the number of suicides was taken from the Mortality Information System database that is handled and maintained by the Brazilian Ministry of Health. The population for the respective year is collected from the Brazilian Institute of Geography and Statistics. The model used for evaluating the research is the polynomial regression model. The dependent variable used here is the suicide rate and the independent variables used are the years. As per the authorized data 164,276 number of suicides occurred of an individual having age group 15 years and above. After considering the 224 records, it is concluded that there is a decrease of 9.4 % in suicide rate, 42.4 % increase and 48.2 % stable. When compared to the World Health Organization target it is shown that 67 % will not be able to meet the target of the year 2020.

Across the world, suicide is the major problem S Selva Priyanka and Srinivasa (2016) Almost, 800,000 individual die because of self-harm injuries and suicides. Researcher S Selva Priyanka and Srinivasa (2016) have aimed to discover the features and highlights of the states in India corresponding to the rate of suicides. For this research, the dataset regarding marital status is collected for all the states. This data consists of the number of individuals in different marital categories. Regarding professional data, the data is gathered from the census that consist of the number of working and non-working male and female. The dataset consists of 21 different features. For this research twelve main states are considered. The techniques used for this study are Pearson Correlation and Regression modeling. The accuracy obtained from this research was 99.8 %age. This provides more confirmation that different classes of individuals have added the contribution to the number of suicides to a large extent.

2.2 Review of the Suicidal Ideation

Andrea C.Fernandes and Chandran (2018) researchers have aimed to identify the idea of suicides and suicides attempts. The data for identifying the suicidal ideation is taken from the Psychiatric Clinical Research. In this database, the suicide attempts and the idea of suicide are documented in the form of free-text. The technique used here is Natural Language Processing (NLP). The paper has used two novel NLP approaches. The first approach is the rule-based method and hybrid machine learning that is used to analyze the existence of the idea of suicide and the second approach is the rule-based method for identification of suicide attempts on the database of the psychiatric clinic. From this research, it has been concluded that the precision for identification of suicide ideation is 91.7 % and for suicide attempts, it is 82.8 %.

Sinisa Colic and Hasey (2018) researcher attempts to study the prediction of the suicidal idea. The dataset used here was of 738 patients having 224 different variables that include the data about the general health of the patients, demographic information and Post-Traumatic Stress Disorder (PTSD) scale. This dataset was taken from the Parkwood Operational Stress Injury (OSI) Clinic present in London. The machine learning algorithm is used to study the suicidal ideation. The Random Forest is used here for the prediction. Researchers evaluated the sensitivity and specificity by making use of receiver operating characteristics (ROC) curve for every cross-validation. This result of this research suggests that the prediction of suicidal ideation can be done without asking any kind of question.

Kunawut Boonkwang and Youdkang (2018) has aimed to compare the different classification algorithms. Here the comparison is made between ID3, C4.5 and nave Bayes. Also, the researchers have done the prediction of the characteristics of suicidal ideation by using data mining techniques. The researchers also implemented the synthetic minority over-sampling techniques (SMOTE) for bringing the stability in the suicidal ideation. The suicide rate calculated in the year 2015 is 6.47 per 100,000 population compared to the suicide rate of the year 2014 it was 6.07 per 100,000 population. The dataset for this research is taken from the report (RP.506S) of self-harm surveillance of the hospital namely Khon Kaen Rajanagarindra Psychiatric for predicting the suicide attempters characteristics. The data is of the year 2008 to the year 2016 having 200,000 records. The experimental result of this research is that it has achieved the accuracy of 90.60 % by C4.5 algorithm.

2.3 Identification of Suicidal Notes and Clinical Data

Schoene and Dethlefs (2018) has proposed a specific method for identifying the sentiments in the data having textual nature. Here the researchers have analyzed particularly the suicide notes, notes written in depression and the love notes. The sentiment analysis is done here. The technique used by the researchers is Natural Language Processing

(NLP). For this research, the data is collected in the form of notes. The genuine 33 number of suicide notes and 33 fake suicide notes were collected. They have concluded from both the different sets, that there exists the language difference in fake and genuine notes. The genuine data was collected and segregated between love post (LP), depression note (DN) and genuine suicides note (GSN). GSN was collected from various sources one of the sources mentioned in the paper is the newspaper. LP and DN were collected from the posts that are publicly available on the experienced project website. Also, for this research, the researchers have not only focused on sentiment analysis of the note but also, they have focused on the content feature and the linguistic nature of the note. Learning model used here is the Long Short-Term Memory (LSTM). The conclusion made by the researchers is that the data can be classified accurately using methodologies of unsupervised deep learning compared to traditional machine learning. The 72 % accuracy is achieved when the classification is dependent on text and sentiment features and 69 % accuracy is achieved when the analysis is done on the words present only in the notes.

John Pestian (2018) according to the researchers, different machine learning techniques are utilized for identification of suicide risk by making use of verbal or written thoughts, or audibility or facial feature. The study done here is focused on clinical data. For this research, the data was gathered and collected from 253 patients who are the victim of suicidal. The data was collected for the period of October 2013 and March 2015 The technique used here are Natural Language Processing (NLP) and Machine Learning (ML) for the prevention of suicides. The three equal-sized groups were created for the analysis: First, the group of suicides that are with the complaint of attempting suicide, idea of suicide or psychiatric evaluation of suicide; Second, the group of mental health having mental illness but don't have the idea of suicide or the attempt; Finally, the group of non-suicidal matter also not having any mental illness. Every group was measured by using the "Hamilton Depression Rating Scale", "Young Mania Scale" and "the Columbia Suicide Severity Rating (C-SSRS)". In this research, only suicide and cohorts were given more important and the same was considered. For the hypothesis test, ML models were built for analyzing the linguistic behavior of the patient's voice, speech and to deliver a result ad suicidal or non-suicidal. The conclusion of this research is that there is no significant linguistic difference between the first and the second visit of the patients. Furthermore, other than clinical data and notes the research is done on the use of technology for prediction of the suicide rate.

Larsen ME and H. (2015) researchers have presented an outline on the existing development of technology that is encouraging exploration for the prevention of suicides, including various methods such as the collected connectivity data of cell-phone is proceed under network analysis, recognizing the content regarding suicide from social media automatically, and detection and discovery of any crisis from audio changeability in speech patterns. Also, researchers have given the summary of existing technological development and scientific progress that comes for the prevention of suicides. The research provides the explanation of detection and discovery of suicide prevention with the help of sensor data collection, analyzing the content of social media and regarding speech the paralinguistic processing is described. It is concluded that the development of the tool for the prevention of suicide is a challenging task.

2.4 Classification of Text Messages, Post and Tweets related to suicides

The tweeter is one of the famous social media sites where people are free to share their views and opinions. Amayas Abboute and PONCELET (2014) the researchers have defined a whole process that can collect and gather the tweets according to the language the person used to talk about suicidal. These automatic capturing of tweets shows the dangerous behavior of suicides by using classification methods. The researchers have recognized the nine topics the suicidal people mostly speak about "Sadness/psychological injury, mental state, depression, fear, loneliness, description of the suicidal attempt, insults, cyberbullying, and anorexia". With the help of API, the tweets were collected automatically. Later these tweets were classified as "risky" and "non-risky" by Waikato Environment for Knowledge Analysis (WEKA). Various machine learning algorithms were used for comparing the results; JRIP, IBK, IB1, J48, Naive Bayes, SMO.

Bridianne O'Dea and Christensen (2015) the level of concern studied by the researchers here is just basis on the content of the post related to the suicide on Twitter. The twitter post was arbitrated by the human coders and the same was then replicated by machine learning. The tweets consist of suicide-related phrases and the words or terms are checked using the Application Program Interface (API), the tweets which are identical in nature are stored in the tool established by the Commonwealth Scientific and Industrial Research Organization (CSIRO). In the data the total 14,701 tweets were collected, out of the selected tweets 14 % of the tweets were then classified into three different levels; "strongly concerning", "possibly concerning" and "safe to ignore". The Support Vector Machine (SVM) was used as a classifier. It has recognized that 80 % of the "strongly concerning" tweets have shown an increase in the accuracy. The study concludes that it is likely to differ in the level of concern between the suicide-related tweets by making use of machine classifier and the human coders.

Astoveza et al. (2018) according to the researcher, around the world suicide is the major leading cause of death among people with age group 15-29 years. In the Philippines the suicide rate among male is 5.8, among female is 1.9 and both sexes contribute 3.8. The research study aims to build a model that will classify the suicidal tweets by making use of an Artificial Neural Network (ANN). For this study the data was collected from Twitter, using Twitter Advanced Search (TAS). One of the objectives of this study is to define the complete procedure of categorizing the collected tweets by making use of the assumed keyword in the tweets. Overall 5,174 tweets were collected, out of these

3,055 tweets were in English and 2,119 in Filipino or Taglish. The collected tweets were categorized as ”non-risky” and ”risky” by the psychologist. Once the data pre-processing is completed the feature selection is done by SelectKbest and chi2 method to select the top 1,500 features out of 5,000. The use of two different validation approach is done in this research; Cross-Validation and 70-30 Split. Also, three different learning rates were used; Constant, Inv-scaling and Adaptive learning rate. The outcome for this research is that the inv-scaling learning rate was the best among the learning rates.

Nobles et al. (2018) the analysis, design, and collection of text messages are carried out in this research. This text messages are collected from individuals for identification of suicidal idea, suicidality. The study uses potential research for understanding whether communication through text will able to forecast the stages of suicidality in comparison with depression. The main purpose of this study is to check whether the text mining will able to identify the risky states such as depression to suicidality upon the day to day communications. The data for this study is collected from the SMS, text messages, emails and the call history, some data is collected from the social media, new-papers and mental health history. Before the collection of data, the online survey was conducted for evaluating the way of communication among the undergraduates of the Department of Psychology by different electronic services. Around 800 students participated in this survey. These surveys were conducted in the form of an online screen, phone screen, and laboratory study. The technique used here is supervised machine learning for building the classifier. SVM and DNN models were used and it is concluded that the performance of DNN model is good compared to others.

2.5 Reasons and the way of Committing Suicide

Prashar and Choudhury (2018) For committing suicides various means are used by an individual some of the means are; Drugs, Hanging, Jumping, Alcohol. According to the World Health Organization, suicide by consuming the pesticides has contributed more to the suicide rates significantly.

The researchers Shil Cha et al. (2019) has aimed to examine long-lasting trends in pesticides suicides rates and factors related to these trends in South Korea. The data for this analysis is provided by the South Korea government. Data contains the information on the age, sex, date of death and the place of residence. The graphical approach and the joinpoint regression were used for examining the trends in total pesticides suicide rates and the factors related to other trends. Also, the time series model to study the relations among pesticides suicide rate, factors related to agriculture and socioeconomic. The result of this research is that the suicide rate is varied between the year 1929 and the year 2000, from the year 2000 to the year 2003 there was increasing in the rates, going high to 9 % in the year 2004, afterward, there was a slow decrease in the suicide rate. Considering other tends, it is concluded that among male the pesticide suicide rates are more compared to female. Similarly, in rural areas, the suicide rates using pesticides as a mean is higher in comparison to non-rural areas.

As discussed by K.Sai Teja and Basha (2018) suicide victims are mostly farmers and students. According to the Ramadas and Kuttichira (2017) around 90 % of an individual has mental disorder those who are committing suicide or attempt to do suicide. Farmers are considered with high suicide risk. In this research, the comparison is done on the farmer's suicide that was conducted by the professionals of mental health disorders and other professionals. The research is done on the country India and Australia. Among farmers, common mental disorders are the use of alcohol, alcoholism and some other issues considered are problems in the family, marital life, economy, and earnings. For this research, two databases were used; firstly, for the proper medical field data with the keyword "farmers" and "suicide" articles from PubMed were used and secondly, for both medical and non-medical fields the articles were collected from Google Scholar. The data was collected for the period of January 2002 to January 2012. From the PubMed database,13 articles were downloaded out of that seven were on Australia, four on India and one from Korea and Scotland each. From Google Scholar around 25,000 articles were downloaded on the farmer's suicides across the world and then by using "farmers" and "suicide" as a keyword the articles were segregated to India and Australia. From this research, it is concluded that the collective and collaborative way in mental health, economics, political science, and agriculture will give the understanding in a more efficient way for farmer's suicides.

Fountoulakis et al. (2016) the researchers have studied on the relationship between climate and economic variables with the suicide rates in the Europe Country for the year 2000 to the year 2012. The data here was collected from 29 countries of Europe for the year 2000 to 2012. The data include the suicide rates according to the gender, as per the World Bank the data have the economic attributes and the climate attributes. The techniques used for this study is clustering; that is done independently for both economic attributes and climate attributes and principal component analysis is used to identify the attributes that are in correlation with economic and climate attributes. From this study it is concluded that the number of male suicides is highly correlated with the large unemployment rate, this unemployment rate consists of the large growth rate and increase and decrease of GDP per capita while the number of female suicides rate is not that much correlated with the inflation of GDP. The study also tells that for both the number of female and male have a correlation with the climate considering both high and low temperature.

Kposowa (2001) the main objective of this research is to study the relation of unemployment and the suicide rates in the United States of America. To evaluate the impact of joblessness on suicide the National Longitudinal Mortality Study (NLMS) was used in this research. Another data used for this research is taken from the Current Population Survey, the data consists of the survey led in the month of March 1979, April, August, and December in the year 1980 and March 1981. A statistical model was used for this research, Cox's proportional model was used on NLMS data to make a comparison with suicide data with respective of employment status. The research is done and concluded in two ways; firstly, to analyze the impact of work status that is employment and different other factors of suicide risk and secondly, to describe if there are exists any contrast or difference by gender. The conclusion is made that unemployment is related to the suicides and it is concluded that the suicide risk is more in the female because of unemployment.

The study by Tsai and Cho (2012) have inspected the connections among the suicidal and the climate and weather change as well as the economic attributes. The data required for this inspection were collected from the Department of Health of Taiwan of the year January 1991 to December 2010. The climate-related data is taken from six different weather stations, for representing the monthly meteorological attributes the average of data was used. The time series analysis is carried out in this research. Autoregressive Integrated Moving Average (ARIMA) was used. The result of this research is that the increase in temperature is the most significant factor that is emphatically related with suicide, the contribution of rainfall and sunlight are not significantly associated with the suicide rate considering total population. Considering the labor force and the unemployment the researchers concluded that both the factors are not significantly related to the suicide rates.

H. K. Seah and Jin Shim (2018) around the world, more than 900 social media sites like Facebook, Twitter, and Reddit have drastically increased for sharing the thoughts and views. By these sites, the state and feelings of individual are shared. Here the posts and comments based on the suicides on Reddit in Singapore are analyzed. Natural Language Processing technique is used for analyzing the text. Using Reddit API data crawling is performed. From Reddit around 385 posts and 21,000 comments were extracted for the year 2010 to the year 2018 that contains suicide and depression related phrases. The analysis of the research describes the characteristics and features of the text related to depression and suicide rates.

2.6 Conclusion

In all the related work mentioned above, the time series analysis has not been done for the prediction of suicide rates in India. Most of the research is done using classification and regression techniques. Therefore, based upon the gap noticed it is required to predict the suicide rates in India using time series analysis.

3 Methodology

3.1 Methodology Approach

Considering the related work discussed above by different researchers and various approaches used by them. In this section the overall design of the approach is discussed. In this research, the Knowledge Discovery Databases (KDD) methodology will be used. Fernndez-Arteaga V and Y (2016) KDD is the methodology approach that is used to extract information from the database. KDD is used among the clinical database such as prediction of lung cancer, breast cancer and strokes. The use of KDD approach for studying the suicide prediction has done by only a few researchers (Fernndez-Arteaga V and Y; 2016).

Figure 1 shows the steps carried out in KDD approach.

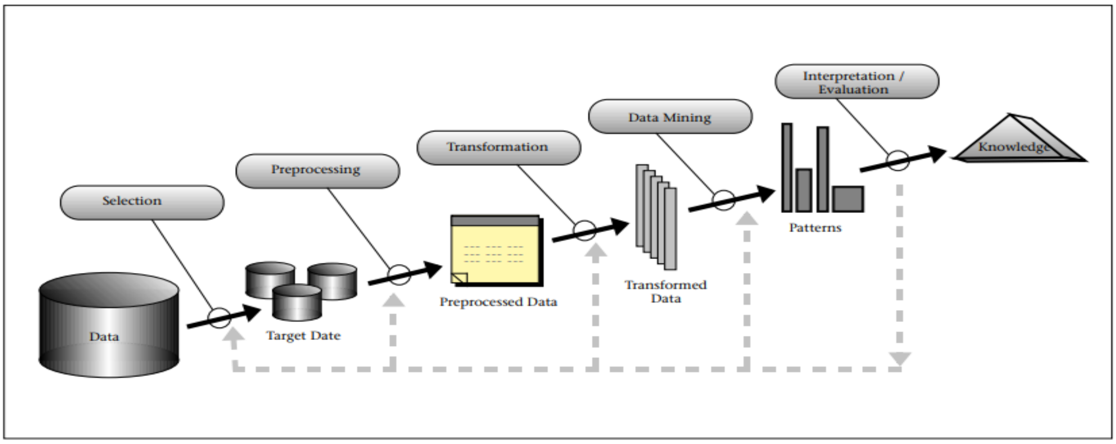

Figure 1: An Outline of the Steps That Compose the KDD Process (Usama Fayyad and Smyth; 1996)

KDD approach follows certain steps (Usama Fayyad and Smyth; 1996):

Understanding and development of the field of application: This is the first phase of the KDD process. This phase investigates the implication of previous knowledge and focuses on the aim and objectives of the end-users.

Creating a target dataset: This is the second phase. In this step, the data is selected and verified to achieve the insights of it and to sort out it properly. This step also includes the selection of proper attributes in the data. The data for this research is taken from different sources like data.world, data.gov.in.

Data Cleaning and Pre-processing: In this step, the data cleaning and preprocessing are carried out. This includes an operation such as removing of noise if exists, the missing data and null values are removed. The main motive of this step is to improve the consistency of the data.

Data Transformation: The next important step is data transformation. This step includes the dimension reduction, feature selection, and transformation methods.

Data Mining: This step includes the selection of appropriate and suitable data mining task. In this research, the algorithm used will be Autoregressive integrated moving average (ARIMA), Autoregressive integrated moving average with explanatory variable (ARIMAX) and Support Vector Regression (SVR).

Data Evaluation and Integration: In this step, the integrating and evaluation of the data are carried out using the algorithms. The evaluation of the data will be done using k-fold cross validation.

Knowledge Discovery: This is the final step in the KDD approach. The different purpose can use the acquired knowledge, or the simple documentation can be made that can use by other parties.

3.2 Key to methodology

The approach for implementation of the research is explained with the below flow chart:

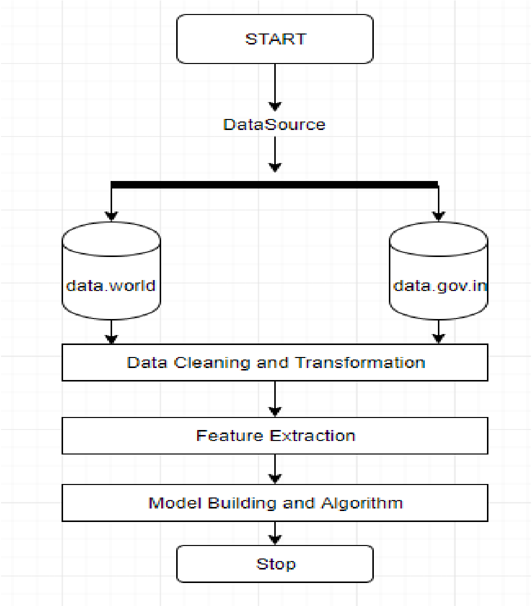
START
DataSource
data.world
data.gov.in
Data Cleaning and Transformation
Feature Extraction
Model Building and Algorithm
Stop

Figure 2: Process Flow Diagram

Selecting the Data: The data selected for this project is extracted from two different sources: data.world and data.gov.in

Date Cleaning and Transformation: Once the data is extracted the cleaning and transformation of data carried out.

Feature extraction: Essential Features will be separated from the dataset.

Building model and Algorithm: Building the suitable model and implementing the algorithms on the dataset.

3.2.1 Data Source:

The First step of the process flow is the extracting and entering the data. In this research, the data will be downloaded from two sources.

Data Source 1 - data.world: data.world is the open repository. The data extracted from this website have 237520 records and 7 attributes about the number of suicides. Some of the attributes are the name of states in India, the year, types of suicides, causes of suicides, gender, age group and the total number of suicides. The data present in the file is from the year 2001 to 2012. [1]

Data Source 2 - data.gov.in: Data.gov.in is the website of the Indian government. This is the open website. The data extracted from this website contains information about the number of suicides in India for the year 2013 and 2014. The data includes the number of suicide according to the gender, age, states of India. Url: [2]

3.2.2 Data Preprocessing:

Second step of the process flow is the data preprocessing. This includes the cleaning and transformation of the data. Once the data is extracted from all the different sources, it is necessary to do the preprocessing of data. The data may consist of some URLs, emoticons, symbols which are not required in the data. If the unwanted data is not removed, then it may affect the accuracy of the model.

Data Cleaning: Data cleaning for this study will be carried out in R.
The process of data cleaning includes:

- Renaming of column names.
- Removing null values or replacing the null values .
- Deleting unwanted rows and columns if required.
- Removing unwanted information from the data.

Data Transformation: Data transformation is the transforming on data whenever required. Data transformation can be:

- Splitting the column.
- Changing of data-types.
- Binary encoding.

[1] https://data.world/rajanand/suicides-in-india

[2] https://data.gov.in/

In this study, the csv from two different sources are merged to make single csv. This is another type of data transformation.

3.2.3 Feature Extraction:

The intense effect that suicide has on encompassing network combined with the absence of feature extraction. The Feature extraction is the important process in the data preprocessing. In feature extractions the factors or attributes that are essential is extracted using various techniques. As more than one source will be used in this project there are chances of having the different words. This prioritization of attributes will be done using different ranker techniques.

3.2.4 Explotary data analysis:

Once the cleaning and transformation of data is completed. Then the data can procced to the Exploratory Data Analysis (EDA). EDA is a method that analyze the dataset for summarizing the features, characteristics and mostly with the visualization methods. EDA highlights the general outcomes. The exploratory data analysis will be carried out in SPSS toll and RStudio.

3.2.5 Building Model:

For building this model time series techniques such as ARIMA, ARIMAX and regression model like Random Forest, Support Vector Regression, Support Vector Machine will be used for predicting the suicide rate and the cause of it. As to acquire an ideal outcome and check the accuracy, the method used will be k-fold cross-validation.

4 Proposed Model

4.1 Motivation of Model:

India is among one of the countries that are contributing high in suicide rates. There are various causes that lead to suicide. And this varies from individual to individual. Prashar and Choudhury (2018) India is among one of the countries that are contributing high in suicide rates. There are various causes that lead to suicide. And this varies from individual to individual. Prashar and Choudhury (2018) in this research, the data of India regarding suicides for the year 2001 to 2012 is used. The technique used here is Decision tree, Linear Regression, and Bayesian Networks. The same data can be analyzed using the time series models such as ARIMA, and ARIMAX and the result of the time series analysis and the regression model can be compared. As the data consists of the information according to the year. The time series model will be suitable for the prediction of suicide rates and the causes and types. The data is multivariate data that will be best suited for the time series analysis.in this research, the data of India regarding suicides for the year 2001 to 2012 is used. The technique used here is Decision tree, Linear Regression, and Bayesian Networks. The same data can be analyzed using the time series models such as ARIMA, and ARIMAX and the result of the time series analysis and the regression model can be compared. As the data consists of the information according to the year. The time series model will be suitable for the prediction of suicide rates and the causes and types. The data is multivariate data that will be best suited for the time series analysis.

4.2 Explanation of Model

The model that will be used in this study is explained below: Time Series: Time series analysis includes the techniques for analyzing the data that is in the interval or seasonal form and it includes the characterizing of the data. Time series forecasting is used for predicting future values dependent on the previous values.
Finance (2018) Time series consists of four different components:
Seasonal Variations: In a seasonal variation, the data repeats over a specific period such as year, quarter, month, week, days.
Trend Variations: In trend variation, the data shows the tendency of increasing and decreasing over a period.
Cyclic Variation: In a cyclic variation, it operates over a span of more than one year.
Random Movements: It does not fall in any of the above three.

Time series model that will be used in this study are:
1. ARIMA: Auto-Regressive Integrated Moving Average (ARIMA) for time series data it captures the different standard time-based structure. ARIMA is also known as Box-Jenkins. Taylor (2016) A univariate time series model defines as a mixture of autoregressive (AR) and moving average (MA). In ARIMA for forecasting the seasonal and non-seasonal model can be used.
ARIMA(p,d,q) forecasting equation:
Kongcharoen and Kruangpradit (2013) In terms of y the general forecasting equation is:
where,

$$\varphi(L)(1-L)^{d}Y_t = \theta(L)\varepsilon_t$$

Yt = stationary series,
ϵt = white noise,

d is the dth difference operator

2. ARIMAX: Autoregressive integrated moving average this is extended ARIMA with an explanatory variable (X).
Kongcharoen and Kruangpradit (2013) The equation of ARIMAX (p,d,q) is given as: where,

$$\varphi(L)(1-L)^{d}Y_t = \Theta(L)X_t + \theta(L)\varepsilon_t$$

Xt = Input Series

The dataset that will be used in this study contains the causes and types of suicides, this means that the suicide rate is dependent on the types and causes of the suicides. Hence, for predicting the suicide rates ARIMAX is considered as suitable techniques. Also, the comparison of the accuracy of ARIMA and ARIMAX can be done.

3. Support Vector Machine (SVM) and Support Vector Regression (SVR): It is a supervised machine learning algorithm that can be used for regression and classification. The use of SVM is generally done for classification (Ray; 2017).
The SVM model can be used to classify the suicide rate according to the types and causes. Another model that can be used for classification is Support Vector Regression (SVR).
SVR make use of kernel, spare solution and VC control and the number of support vector (Awad and Khanna; 2015). Using SVR the suicide rate can be predicted by types and causes of suicides in India. SVR performs linear regression in higher dimensional space.
Key terms of SVR:
Kernel: the lower dimensional attributes or data is mapped into higher dimensional data.
Hyper Plane: In SVM, between the data classes the separation line is the hyper plane. While in SVR the hyper plane is the line that will predict the target value or the continuous value.
Boundary line: Other than the hyper plane, the two lines that create the margin among the two classes is the boundary line.
Support vectors: Data points that are closest to the boundary is the support vectors. The distance here is very less.
4. Random Forest: Random Forest is the technique that can be used for classification by making use of multiple decision tree and regression. It is also called as Bootstrap Aggregation. It is a supervised machine learning algorithm. It creates numerous decision

trees and combines all of them together to achieve a more precise and steadier forecast. Overfitting is prevented by building the features of random subsets and small decision trees.

5 Evaluation

To determine the performance of the model, evaluation of the model plays a significant role. The model will be focusing more in analyzing the execution of the algorithm that is measured in terms of specificity, sensitivity and accuracy. Evaluating the model helps to recognize the classifier that have the more accuracy and good performance. Firstly, for analyzing the outcome of the classification, the confusion matrix will be studied. For further evaluation in this study, the k-fold cross- validation will be performed. By doing so the data will be divided into identical part and after each iteration the different portions of data will be selected for the further process. Moreover, by evaluation the Absolute Percentage Error (APE), Root Mean Square Error (RMSE) and Residual Error (RE) will be analyzed. This value will be compared and studied with the original value.

6 Future Plan

This section gives the detailed plan of the project.

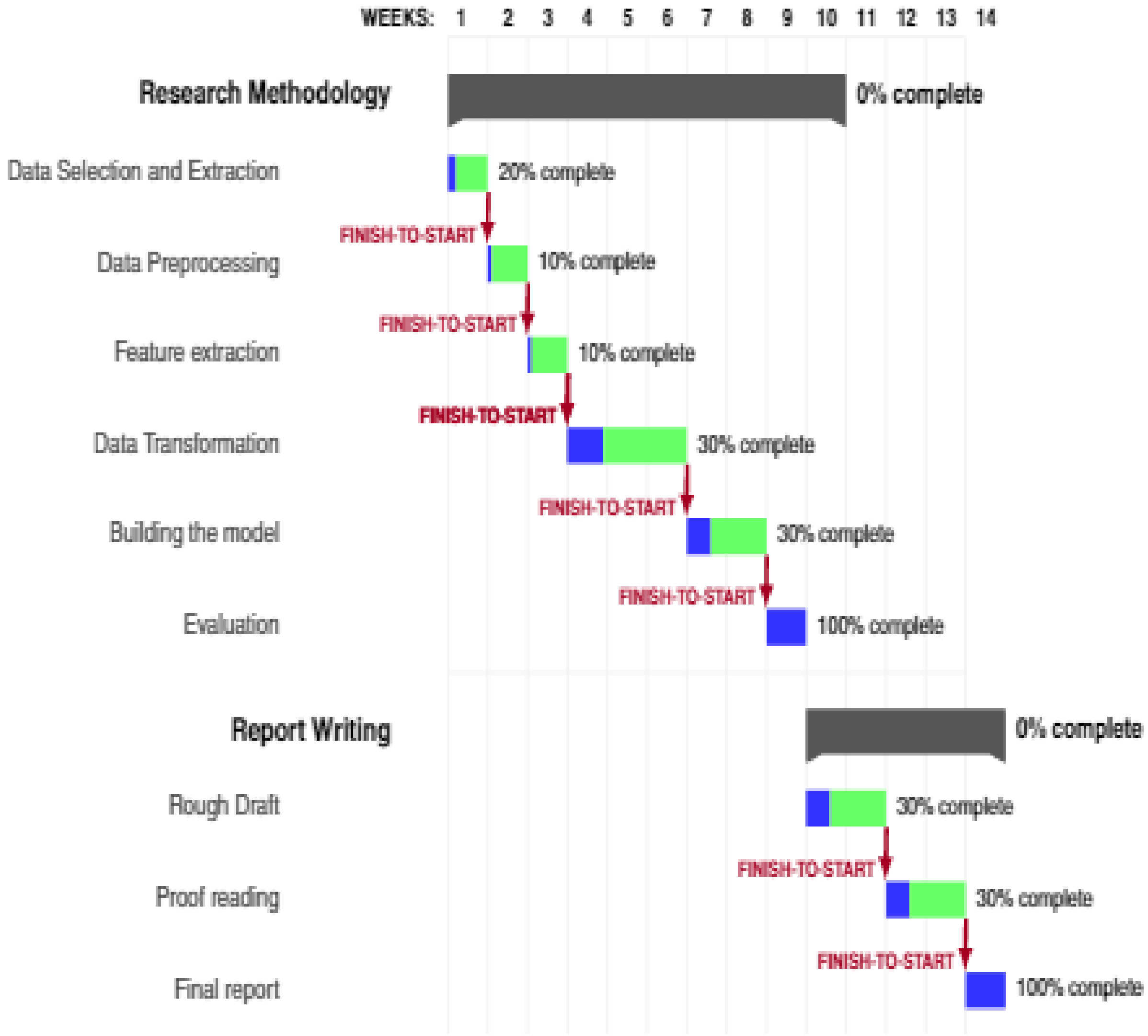

Figure 3: Gantt Chart

References

Amayas Abboute, Yasser Boudjeriou, G. E.-J. A. S. B. and PONCELET, P. (2014). Mining twitter for suicide prevention, *19th International Conference on Application of Natural Language to Information Systems* pp. 250–253.

Andrea C.Fernandes, Rina Dutta, S. V.-J. S. R. S. and Chandran, D. (2018). Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing, **8**(1).

Astoveza, G., Jay P. Obias, R., Jed L. Palcon, R., Rodriguez, R., Fabito, B. and V. Octaviano, M. (2018). Suicidal behavior detection on twitter using neural network, pp. 0657–0662.

Awad, M. and Khanna, R. (2015). Support vector regression, pp. 67–80.

Bridianne O'Dea, Stephen Wan, P. A. L. C. C. P. and Christensen, H. (2015). Detecting suicidality on twitter, *19th International Conference on Application of Natural Language to Information Systems* **2**(2): 183–188.

Carter, G., Milner, A., McGill, K., Pirkis, J., Kapur, N. and J. Spittal, M. (2017). Predicting suicidal behaviours using clinical instruments: Systematic review and meta-analysis of positive predictive values for risk scales, *The British Journal of Psychiatry* **210**: 385–395.

Fernndez-Arteaga V, Tovilla-Zrate CA, F. A. G.-C. T. J.-R. I. L.-N. L. and Y, H.-D. (2016). Association between completed suicide and environmental temperature in a mexican population, using the knowledge discovery in database approach, *Computer Methods and Programs in Biomedicine* **135**: 219–224.

Finance, W. (2018). Time series.

Fountoulakis, K., Chatzikosta, I., Pastiadis, K., Zanis, P., Kawohl, W., Kerkhof, A., Navickas, A., Höschl, C., Lecic-Tosevski, D., Sorel, E., Rancans, E., Palova, E., Juckel, G., Isacsson, G., Jagodic, H., Botezat-Antonescu, I., Rybakowski, J., Azorin, J., Cookson, J., Waddington, J., Pregelj, P., Demyttenaere, K., Hranov, L., Stevovic, L., Pezawas, L., Adida, M., Figuera, M., Jakovljevi, M., Vichi, M., Perugi, G., Andreassen, O., Vukovic, O., Mavrogiorgou, P., Varnik, P., Dome, P., Winkler, P., Salokangas, R., From, T., Danileviciute, V., Gonda, X., Rihmer, Z., Forsman, J., Grady, A., Hyphantis, T., Dieset, I., Soendergaard, S., Pompili, M. and Bech, P. (2016). Relationship of suicide rates with climate and economic variables in europe during 20002012, *Annals of General Psychiatry* **15**: 1–6.

Glen Coppersmith, Ryan Leary, P. C. and Fine, A. (2018). Natural language processing of social media as screening for suicide risk, *Biomedical Informatics Insights* **10**: 1–11.

H. K. Seah, J. and Jin Shim, K. (2018). Data mining approach to the detection of suicide in social media: A case study of singapore, pp. 5442–5444.

John Pestian, Daniel Santel, M. S. U. B. B. C. T. G.-M. D. S. T. K. C. (2018). A machine learning approach to identifying future suicide risk.

Kongcharoen, C. and Kruangpradit, T. (2013). Autoregressive integrated moving average with explanatory variable (arimax) model for thailand export, pp. 1–8.

Kposowa, A. (2001). Unemployment and suicide: A cohort analysis of social factors predicting suicide in the us national longitudinal mortality study, *Psychological medicine* **31**: 127–38.

K.Sai Teja, S.Pravalika, G. and Basha, S. M. (2018). Classification of suicidal deaths caused in india through various supervised machine learning techniques, *International Journal of Engineering Research in Computer Science and Engineering*, Vol. 5, pp. 237–242.

Kunawut Boonkwang, Sumonta Kasemvilas, S. K. and Youdkang, O. (2018). A comparison of data mining techniques for suicide attempt characteristics mapping and prediction, *International Seminar on Application for Technology of Information and Communication (iSemantic)* pp. 488–493.

Larsen ME, Cummins N, B. T. O. B. T. J. N. J.-S. F. E. J. and H., C. (2015). The use of technology in suicide prevention, *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* .

NCRB (2014). Suicides in india farmer suicides in india.

Nobles, A., Glenn, J., Kowsari, K., Teachman, B. and Barnes, L. (2018). Identification of imminent suicide risk among young adults using text messages.

OLIVEN, J. F. (1954). Suicide prevention as a public health problem, Vol. 44, pp. 1419–1425.

Prashar, P. and Choudhury, T. (2018). Suicide forecast system over linear regression, decision tree, nave bayesian networks and precision recall, *2018 8th International Conference on Cloud Computing, Data Science Engineering (Confluence)* pp. 310–313.

Ramadas, S. and Kuttichira, P. (2017). Farmers suicide and mental disorders perspectives in research approaches-comparison between- india and australia, *International Journal Of Community Medicine And Public Health* pp. 300–306.

Ray, S. (2017). Understanding support vector machine algorithm from examples (along with code).

Rodrigues CD, de Souza DS, R. H. and TCRO, K. (2018). Trends in suicide rates in brazil from 1997 to 2015, *Brazilian Journal of Psychiatry* **41**.

Ryu S, Lee H, L. D. P. K. (2018). Use of a machine learning algorithm to predict individuals with suicide ideation in the general population, *Psychiatry investigation* **15**: 1030–1036.

S Selva Priyanka, Sudeep Galgali, B. R. S. and Srinivasa, K. G. (2016). Analysis of suicide victim data for the prediction of number of suicides in india, *2016 International Conference on Circuits, Controls, Communications and Computing (I4C)* .

Schoene, A. M. and Dethlefs, N. (2018). Unsupervised suicide note classification.

Shil Cha, E., Chang, S.-S., Choi, Y. and Jin Lee, W. (2019). Trends in pesticide suicide in south korea, 19832014, *Epidemiology and Psychiatric Sciences* pp. 1–9.

Sinisa Colic, J D. Richardson, J. P. R. and Hasey, G. M. (2018). Using machine learning algorithms to enhance the management of suicide ideation, *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* pp. 4936–4939.

Soo Beom Choi, Wanhyung Lee, J.-H. Y. J.-U. W. and Kim, D. W. (2018). Ten-year prediction of suicide death using cox regression and machine learning in a nationwide retrospective cohort study in south korea, *Journal of Affective Disorders* **231**: 8–14.

Taylor, R. (2016). Arima models.

Tsai, J.-F. and Cho, W. (2012). Temperature change dominates the suicidal seasonality in taiwan: A time-series analysis, *Journal of Affective Disorders* **136**: 412 – 418.

Usama Fayyad, G. P.-S. and Smyth, P. (1996). From data mining to knowledge discovery in databases, *American Association for Artificial Intelligence* **17**: 37–54.

Varnik, P. (2012). Suicide in the world, *International journal of environmental research and public health* **9**: 760–71.

WHO (2017). Regional strategy on preventing suicide.