Analysis of suicide rates and suicide number prediction

Manchikanti Rajesh
computer science engineering
PES UNIVERSITY
rajeshmanchikanti10@gmail.com

Tejas v
computer science engineering
PES UNIVERSITY
tejas.v.reddy07@gmail.com

Charan raj k
computer science engineering
PES UNIVERSITY
charanraj.k888@gmail.com

Abstract:
Study on suicide rates is an important topic to be analysed, because knowing the factors for suicides Could help government or higher authorities in reducing the suicide rates .The dataset which we have used is suicide rates overview 1985-2016, which has been taken from Kaggle, it contains around 28000 instances. The overall purpose of the project is to analyse the trends in suicides over the year, and check the trends of suicides with all other attributes in the dataset ,and predicting the suicides count using Machine learning models such as Multi linear regression, K-Nearest Nearest Neighbor's , decision tree regressor, and random forest regressor.

Keywords: Machine learning, decision tree regressor ,K-Nearest Neighbor's (KNN) regressor, Random Forest Regressor

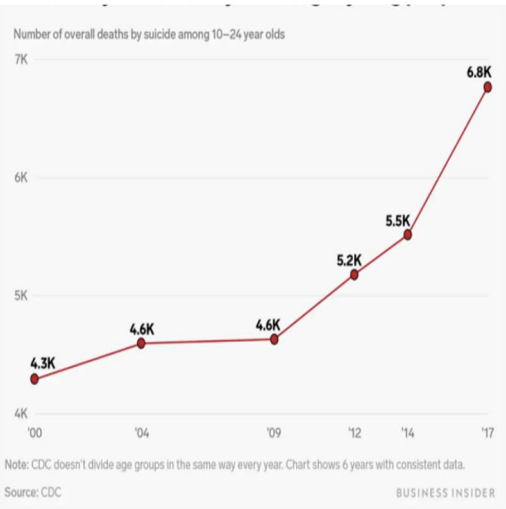
I .Introduction and background:

Suicides is one of the top causes for death among middle aged people and youth. WHO Claims that one death occurs for every 40secs, and it also claims that around 1 million people die by suicide and more than 20 million people attempt suicide or one attempt for every 3 seconds. Suicides Contribute 1.8% of total deaths world wide ,It as raised to 2.4% in 2020

According to recent data of world health organization ,suicide rates ranges from 0.7/100000 in Maldives to 67.3/100000 in Belarus.. India had suicide rate of 10.9/100000 in 2009.

In recent times suicides among youth has increased exponentially. Youth are more prone to commit suicides in developing countries. Suicide rate in this internet era has increased rapidly.

Understanding region-specific factors would help in preventing or reducing suicides globally. This qualitative review explores the historical and epidemiological aspects of suicide like HDI .countries with high HDI also have high suicide rates.



Fig[1] shows the suicide trends among youth for 6 years

From above graph we can see that suicides among youth have gone up exponentially

The main aim is to predict suicide number using various attributes which has some good significant relation with suicides_no

- Attributes like generation are less significant (low correlation)

Above all limitations are addressed in our work

II. literature Review:

Worldwide, Suicide rate is one of the most important problems. The total number of individuals who committed suicide is increasing with each passing year. It is projected that because of the various causes, around eight hundred thousand individuals expires while attempting suicide [5]. Suicide is considered as a disease and according to the report of WHO (World Health Organization), 17 percent residents of the global suicide sufferers belongs to India [5].

According to the CDC-2015, in the last few years, researchers have focused on recognizing, understanding,
curing and impediment of suicidal patterns and behaviour. Regardless of all the efforts and studies, the rate of suicide is not decreasing [6].

For this reason it is important to make better prediction which would be helpful in reducing the suicide rates across the globe

A. Title: Studying Suicides rates 1985 to 2016:

1. Claims:

- Some of the main claims made by author are

Countries with high HDI's also have higher
suicide rates

- Suicides rates among middle aged people is high
- Suicide rate of male is more compared to female

The author has use KNN regressor to predict suicide numbers using input attributes population, GDP per year, HDI, generation and many more

2. Limitations:

- Their exists multicollinearity among input attributes
- The KNN regressor accuracy is low

III. Methodology/Implementation:

The main aim is predicting the suicide number with given set of the input attributes. The approach is Same as any analytics project. First step is pre-processing which includes data cleaning and standardization of data and then removing duplicate columns or redundant columns, second step is checking and dealing with multi-collinearity in the data which would be useful for multi linear regression model, Third step is doing Exploratory data analysis, and last step is to applying different Models and evaluating those models using standard statistical methods

Steps:

- Pre-processing
- EDA
- Applying different models
- Evaluating models
- Conclusions

Pre-processing further includes some steps they are as follows:

- Data cleaning
- Removing redundant or duplicate columns
- Dealing multi-collinearity
- standardization

Next follows Exploratory data Analysis(EDA),Some useful insights which came from EDA are as follows;

- Male suicide rates is higher compared to female suicides
- Suicide rates of people belonging to age 35-54 has higher suicide rates ,then follows
  People belonging to 15-24
- Higher economic countries also had higher suicide rates

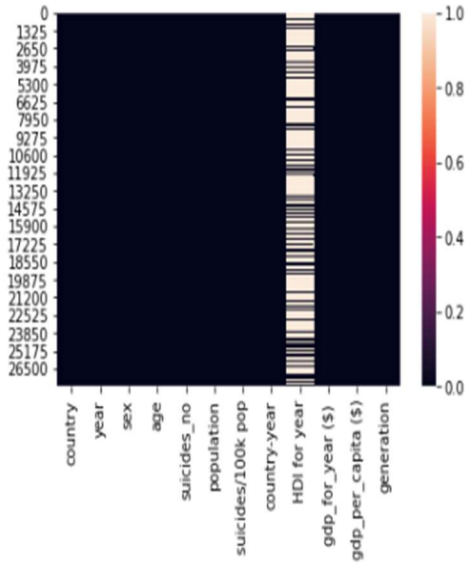
After performing EDA we go for model building and evaluating models using statical methods From our work Decision tree

regressor and random forest regressor had pretty good accuracy

A. Pre-processing:

1. *Data cleaning:*

Attributes or columns of our dataset are country, year, sex, age, suicide number, population, suicides/100k pop, country-year, HDI for year, GDP for year, GDP per captia , Generation. Other than HDI column no other attributes has NAN values, As HDI is calculated per country ,so NAN value of particular country is replaced with mean value of that country.



Fig[2] shows that only HDI column has null values

2. *Removing redundant columns*:

Country-year is redundant column of country, year and suicides/100k is also an redundant information , so those columns can be removed for our further analysis.

3. *Dealing with multi-collinearity:*

Multi-collinear can change the results of our analysis so we need to remove columns which are causing multi-collinearity.
GDP-for year and GDP per captia are multicollinear with HDI for year . These columns are Removed to prevent wrong results in our prediction

```
def correlation(df,threshold):
    col_corr=set()
    corr_matrix=df.corr();
    for i in range(len(corr_matrix.columns)):
        for j in range(i):
            if(abs(corr_matrix.iloc[i,j])>threshold):
                col_corr.add(corr_matrix.columns[i])
    return col_corr
```

```
extra_columns=correlation(new_df,0.7)
```

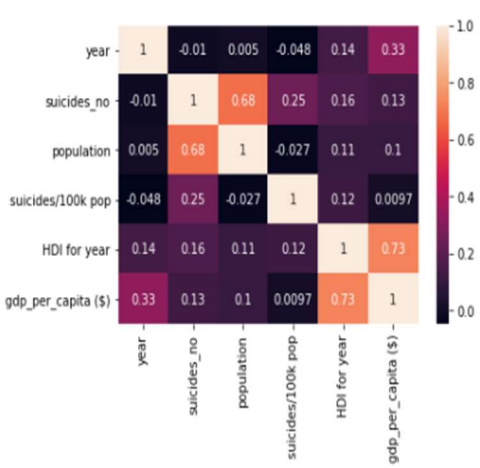
```
extra_columns
```

```
{'gdp_for_year', 'gdp_per_capita'}
```

Fig[3] shows that are GDP for year, GDP per captia are causing multi-collinearity

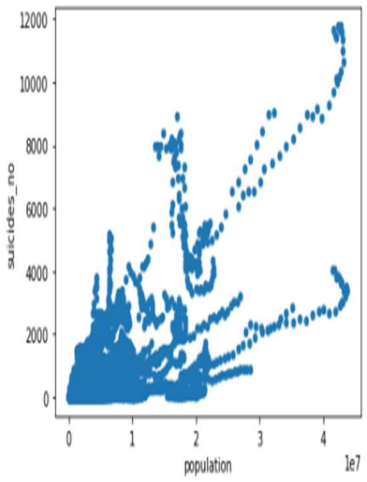
B. Exploratory Data Analysis(EDA):

The below plot shows correlation among all attributes in the dataset

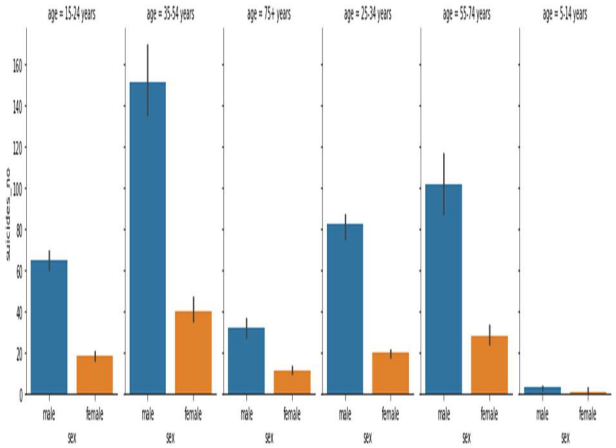


Fig[4] :correlation matrix

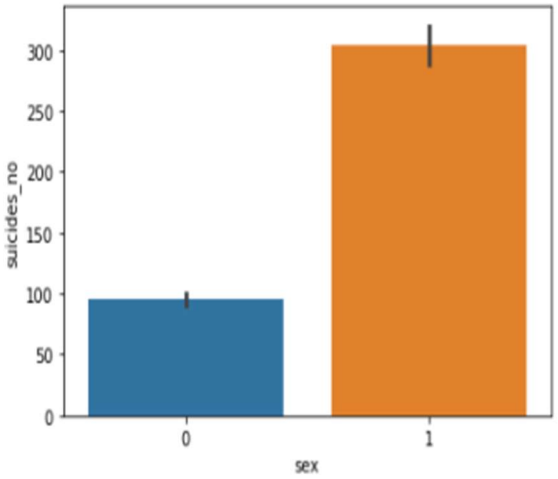
From above we can see that suicides number doesn't have any significant relationship with most of the attributes except the population and a bit significant with HDI



Fig[5] :plot b/w suicides_no and population

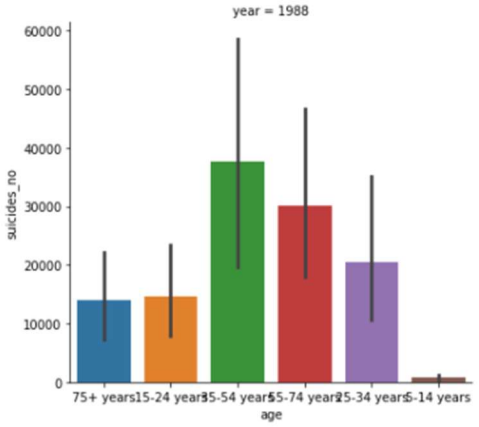


Fig[6]:sex vs suicide_no grouped by age



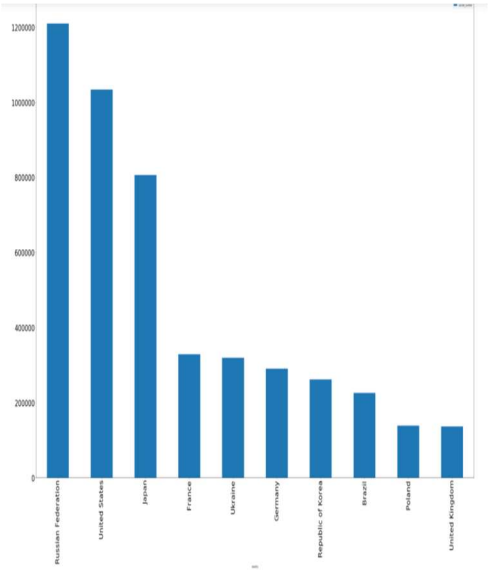
Fig[7]:sex vs suicide_no

From the above two plots we can see that males are prone to more commit suicides than female in all age groups



Fig[8]. Age vs suicide number

From above plot we can see that people belonging to the age group 35-54 are prone to commit more suicides



The above plot shows total suicides number of top 10 countries till 2016

Fig[9] total suicide number vs country

C. Model building and Evaluation :

Problem statement: suicide number prediction

Models applied:

- Multi linear regression
- KNN regressor
- Decision tree Regressor
- Random forest Regressor

From exploratory data analysis we have seen that suicide number doesn't have any variable with Significant correlation except population and a bit with HDI, but we can't use population alone to predict suicide number ,because it is spurious correlation

So we will be using backward elimination to build multi linear regression model

1. Multi linear regression:

Using backward elimination we find final input variables to our models, the input variables are as follows:

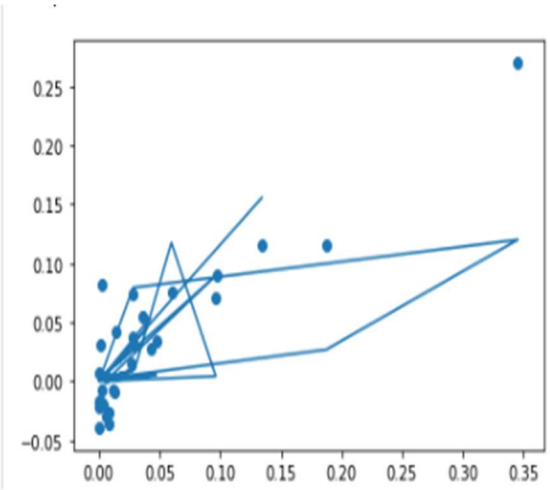
Input variable:

HDI for year, age, sex, population ,year

Output variable:

Suicide number

The accuracy of the model is 44.6% and RMSE is 0.0732



Fig[10]

As the accuracy is not good we go for other models

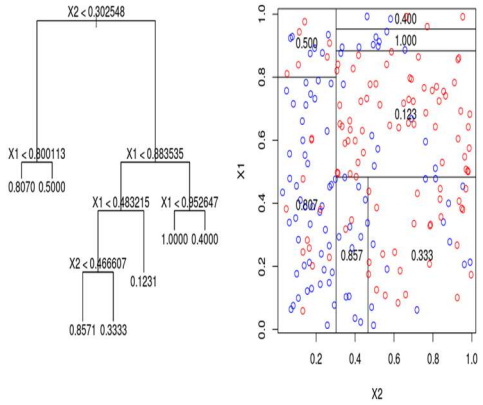
2. KNN regressor:

KNN regressor is combination of classification and regression whose output would be average of k nearest neighbors

Accuracy was 64.31% and RMSE is 0.058 ,Accuracy of model has increased but we need more accuracy, so we go for decision tree regressor

3. *Decision Tree Regressor:*

Decision tree regressor is also classification, regression model ,where split is based on continuous variable



Fig[11]

Accuracy:88.26%

RMSE:0.033

We can see that accuracy as increased a lot and RMSE also reduced from 0.05 to 0.03.We can use Decision tree regressor for prediction of suicide number, but we can use one more model called random forest regressor which is combination of decision trees.

4. Random forest Regressor:

It is bunch of decision trees who accuracy to our problem is 92.1% and RMSE is 0.026

IV. Conclusion:

- Population ,HDI per year, sex, age are the variables used in prediction of suicide number
- Decision tree Regressor and Random Forest Regressor are models which gave pretty good accuracy ,but random forest regressor requires more memory , so according to requirements and restriction, we use choose one the model for prediction

V. Contributions:

Every one of us worked together for every step in our approach ,so everyone have contributed equally