suicide Rates:

The dataset is about suicide rates[1] happened between 1982 to 2016, it is a compiled dataset which is built to find signals related to suicide rates. I have applied random forest algorithm to predict the suicide rates.

This dataset contains total of 27820 rows and 12 columns

Columns	Column Description	Data Types
Country	Country name	Object
Year	Year	Int 64
Sex	Male or female	Object
Age	Age of person	Object
Suicides_no	No of suicides	Int 64
Population	Population of country	Int 64
Suicides/100k pop	No of suicides for every 100k population	Float 64
Country-year	Country name & year of data collected	Object
HDI for year	Human development Index value	Float 64
Gdp_for_year	Gross domestic value per year	Object
Gdp_per_capita	GDP measure of country	Int 64
generation	Type of generation	object

Fig: 3

Data Preprocessing and Cleaning:

Few insights are driven from the suicide dataset to know the suicides is associated with different variables.

Below are the countries that shows the top & least 30 suicide countries

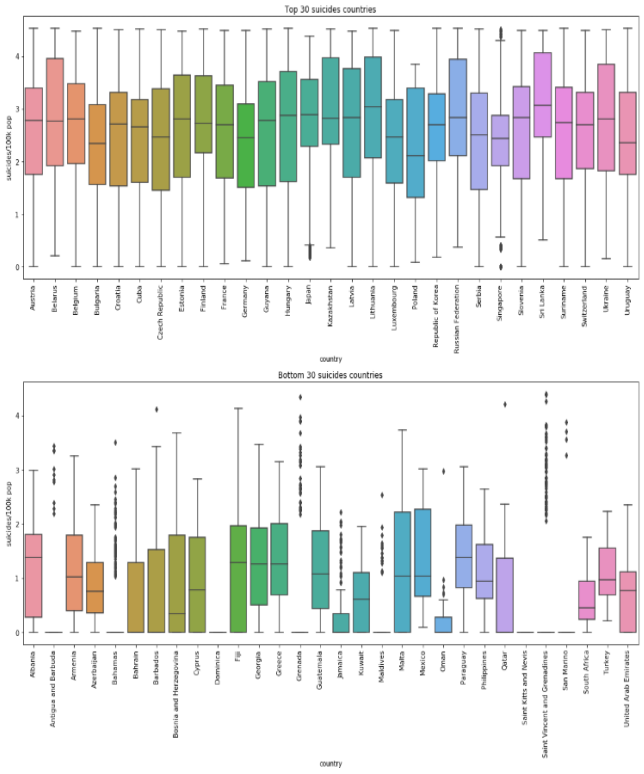

Fig: 23

From above figures we can observe that Sri Lanka, Lithuania, Kazakhsthan are the top 3 suicidal countries and Saint kitts and Nevis, Dominica, Maldives are the least three suicidal countries.

[1] [1] "Suicide Rates Overview 1985 to 2016." https://kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016 (accessed May 04, 2020).

Trend between suicide numbers & suicide population over all the years.

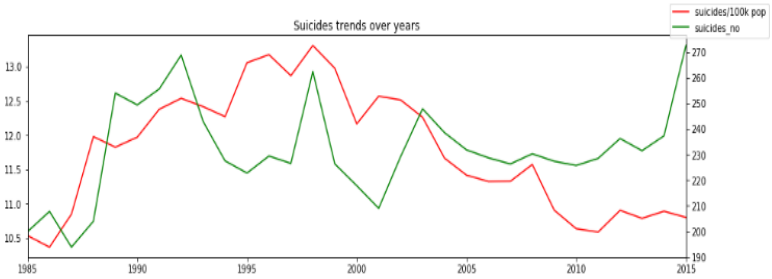

Fig: 24

There were a lot of suicides happened between the years 1995 to 2002. The suicides per 100K population has decreased eventually but the suicide count but the suicide numbers are increasing.

Plot to show average suicides per age:

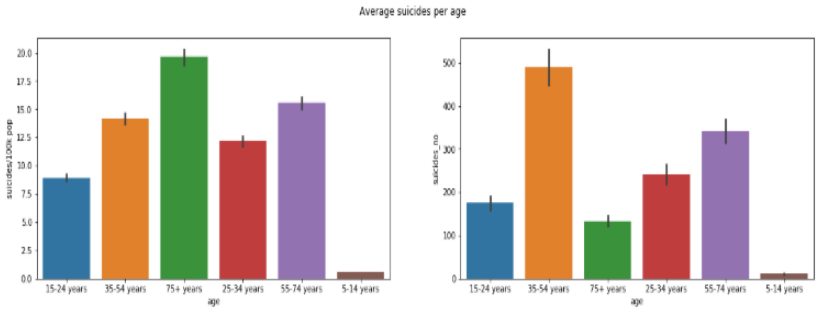

Fig: 25

The above figure talks about the distribution of suicide rates between all the ages. Left side of fig 25 tells about there are more number of suicides per 100k population for age 75+ & number of suicides at higher level for the age 35-54 years.

Plot of suicide rates based on 5 categories of generations:

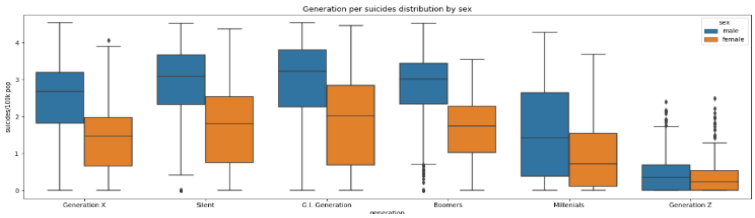

Fig: 26

The above plot tells how the suicide rates distributed between males & females among all five different generations. The generation G.I. have highest number of suicide rates while generation Z has the least count.

Correlation Plot:

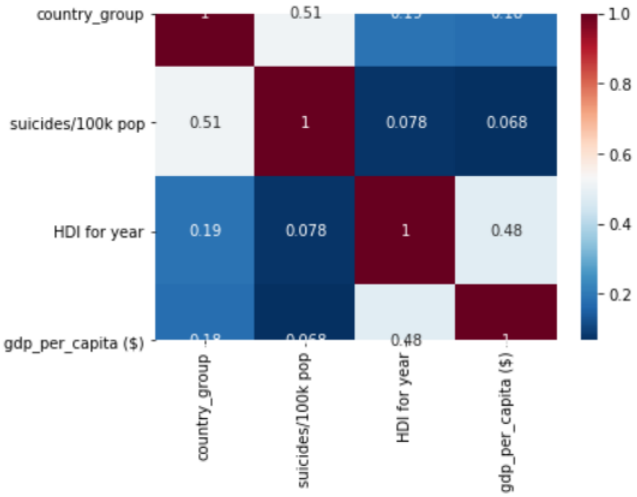

Fig: 27

From the above correlation plot all the variables are less correlated most of them are under 0.5 correlation. So I'm taking all the variables for the model prediction.

Applying the data mining algorithm:

Random Forest:

I have used random forest algorithm to predict the suicide rate and to know what are the most influential factors that for suicide rate. As my target variable is continuous we have the accuracy and RMSE values as follows:

```
RMSE: 7.566604004750796e-05 R2: 0.9999999964648535
```

The R square value is 0.9 which is a good accuracy in prediction of suicide rates and the RMSE(root mean square error) means the variance between the target and input variable is 7.5, To know the most influential variables from my dataset I have ordered them ascendingly.

	0
HDI for year	0.014766
year_range	0.031190
gdp_for_year ($)	0.060992
gdp_per_capita ($)	0.072469
sex	0.190849
age	0.303276
country_group	0.326457

Fig: 35

We can see the countyr_age, age and sex are the most important factors for the suicide rates.

Conclusion:

In conclusion, the random forest algorithm worked well in predicting the suicide rates that are going to happen and gave the top three influential variables that are country_group, age and sex. As the suicide rates prediction is much complex its better to use the black box machine learning algorithms like neural network machine learning models can help a lot in prediction with higher accuracy.