# Suicide rate Prediction

# Introduction

**Problem statement**

The rate of suicide is a major concern for public health. The World Health Organization (WHO) estimates that up to 800,000 individuals take their own lives each year, which equates to one person taking their own life every 40 seconds; in addition, there are many more people who attempt to take their own lives. It is possible to commit suicide at any point in one's life, and in 2016, suicide was the second largest cause of death among those aged 15 to 29 around the world. This project's goal is to analyze major patterns and features that contribute to an increase in suicide rates all over the world, as well as to use Machine Learning algorithms to make predictions about future suicide rates.

**Summary of the approach**

To predict the suicide rate in each region, we will use five machine learning algorithms that are support vector machine, decision tree regressor, random forest regressor, k neighbors and linear regression. First, we will preprocess to data and prepare it for modeling. A deep analysis about features is also planned to do to find out the relationship between different features. After that, data will be passed to machine learning models to predict the results. At the end the results will be evaluated based on Mean square error and R square error.

# Data

Dataset contains the facts about the suicide incidents such as suicide year, Gender and region of the person attempting suicide. Below image shows how the dataset looks like:

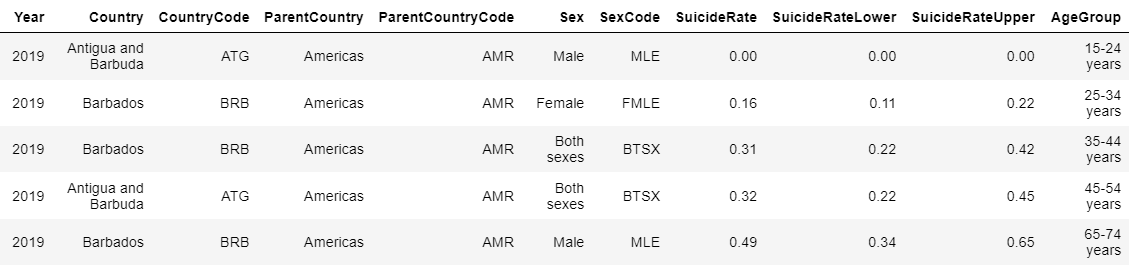


Figure : First view of data

By using the info function of the python, we got the overview of the features of the dataset as follows:

A screenshot of a computer

Description automatically generated with low confidence

Figure : Overview of features

There are total 11 features with 10980 records in this suicide dataset. This suicide dataset is collected for the years 2000 to 2019.Most of the features are object type whereas suicide rate has the floating-point values. The age values are not discrete but in ranges from 15 to 85+.

Data Analysis

Data Analysis used to find the useful information from the data using different statistical and visualization techniques. These insights provide the deep understanding of data that helps to choose the best model for dataset. This dataset has been analyzed through all features to find the most affected factors. Below is a detailed analysis of the features.

**Gender wise suicide rate**

The gender has three types of data: male, female and both sexes. For better understanding we have converted both sexes into other sex. The figure below shows the ratio of genders in this suicide dataset.

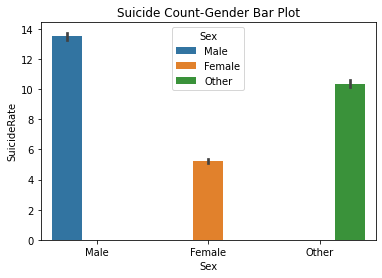


Figure : Gender wise suicide rate

We can see that male has the highest suicide rate and female gender has the lowest suicide rate. Whereas the other gender that is not specifically defined has more rate than female.

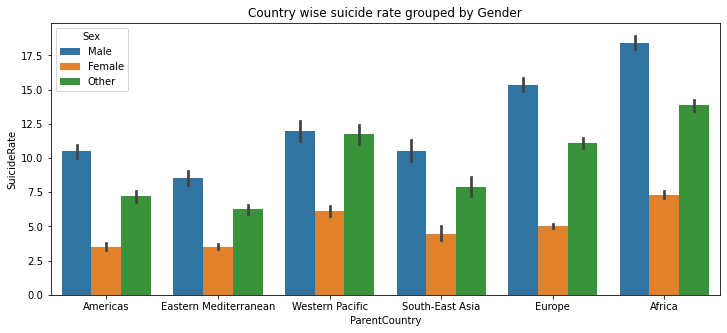


Figure : Country wise suicide rate for each gender

In above figure, region wise suicide ration has been shown for each gender. Male is again dominating here. Africa has the highest suicide rate in male followed by other and female. Europe is on second number and western pacific at third. Eastern Mediterranean has the lowest suicide rate. One thing is common that male has the highest suicide rates in all regions.

**Age wise suicide rate**

In this section, age wise suicide rates are analyzed for each region and gender. There is total eight age groups given in this dataset from 15 years to 85+ years.

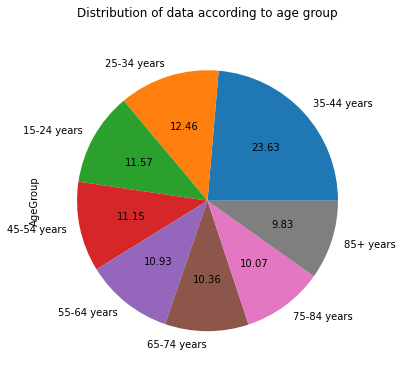


Figure : Distribution of age group

In above figure, age wise distribution has been displayed for all persons involved in this dataset. Distribution has been done in percentage and age group from 35 to 44 has the highest ratio. 85 + age group has the lowest percentage.

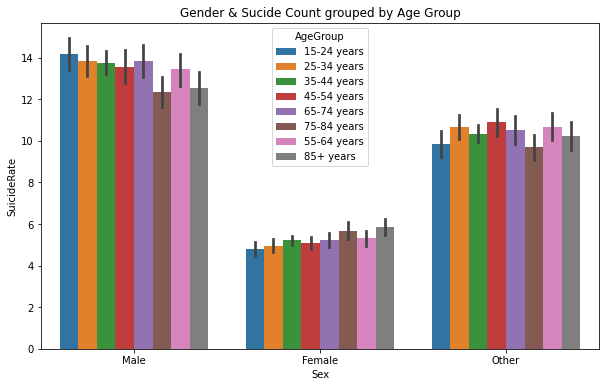


Figure : Gender wise suicide rate for all age groups

The male group is more affected group from suicide rated in each age group. The age group 15-24, 65-74, and 25-34 has the more male suicide affected ratio.

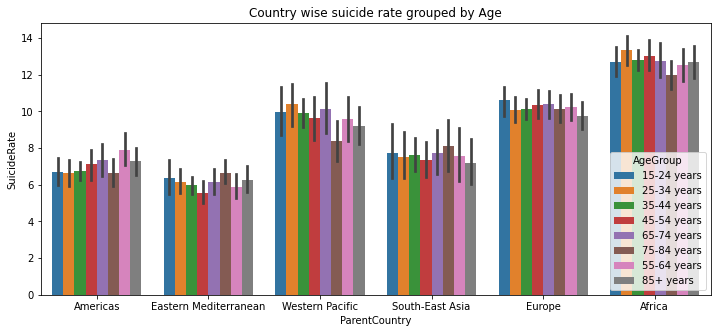


Figure : Suicide rate for each country

Different regions have varying suicide rates for each age groups. In Africa, 25 to 34 years age groups have highest suicide rate 75 to 84 has lowest rate. In Europe, 15 to 24 has the highest suicide ration whereas in America 55 to 64 years age group is leading with suicide rate.

**Year wise suicide rate analysis**

Each year has the varying suicide rates. We can see in below figure that suicide rates have been gradually declined each year.

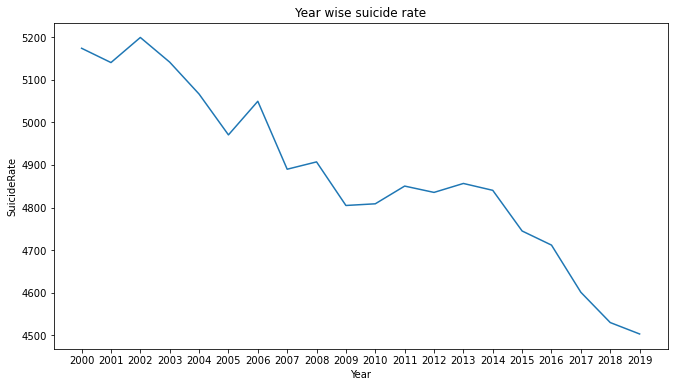


Figure : Year wise suicide rate

In early years, it was more than 5000 which went on decreasing each year and reached to around 4500 in 2019.

# Data cleaning and data Preparation

Data preparation is the process of transforming raw data into a format that data scientists and analysts can use to run it through machine learning algorithms to acquire insights or make predictions. "Data preparation" is another word for this procedure. Data that hasn't been adequately categorized or prepared. It's a crucial stage that comes before processing, and it usually include reformatting the data, correcting errors, and integrating multiple sets of data to improve the data.

In this dataset, first we checked for null values that are none. Then we applied function to detect and remove outliers from the dataset. Then the categorical features were converted into numerical to pass the data to a machine learning algorithm. Because machine learning algorithms cannot work with categorical data types. After preprocessing, the dataset was divided into training and testing sets with 80 and 20% ratio.

# Model Implementation

Model implementation is the process of training the machine learning models with training data. After training it is tested and evaluated with testing data. Below are the machine learning models to be used for prediction of students’ adaptability to online education.

i. Linear Regression

ii. Support Vector Classifier

iii. Decision Tree Classifier

iv. KNearest Neighbor classifier

v. Random Forest

**Description of models**

Linear regression analysis is used to investigate two aspects: (1) Can a group of independent variables that serve as predictors accurately forecast a group of variables that serve as dependents? (2) Which specific variables are significant predictors of the outcome variable, and how do these variables influence the outcome variable, as indicated by the magnitude and sign of the beta estimates? These regression estimations are utilized in the process of providing an explanation for the connection that exists between one dependent variable and one or more independent variables (Maulud, 2020).

Decision Tree is displayed as a recursive splitting of the instances. The nodes that make up the decision tree can be connected to one another to create what is referred to as a source tree. This demonstrates that the decision tree is a dispersed tree with a primary node that is referred to as the root node without incoming edges below (Charbuty, 2021).

The KNN algorithm is one of the more user-friendly examples of machine learning (ML) techniques. It's possible that the KNN's success is attributable to its A straightforward explanation and a rapid calculation time are also included. KNN begins by choosing k neighbors at random and then computing the distances between them. The distance function is what calculates the distance between k neighbors. This function is what is commonly utilized to refer a class among an object's k nearest neighbors (Jawthari, 2021).

The random forest algorithm can perform both regression and classification analysis. This is a group of decision trees that can be used together, as suggested by the name of the product, which can be found here. The benefit of RF is that one tree shields the others from the shortcomings that are unique to each individual tree (Abdulkareem, 2021).

Support Vector Machine: By using various classification strategies, the Support Vector Machine is able to provide solutions to situations that involve more than one set of categories. The SVM model can categorize newly faced text after the model has been loaded with sets of labelled data from all categories. Because of the unique way in which they select the decision boundary, support vector machines stand out among other classification techniques. SVMs are responsible for generating a decision boundary, also known as the maximum margin classifier or the maximum margin hyper plane. This boundary is the result of the SVMs' classification efforts (Campbell, 2011).

# Results

After loading models from sklearn library as shown in the figure below, we passed the training data to all models one by one.

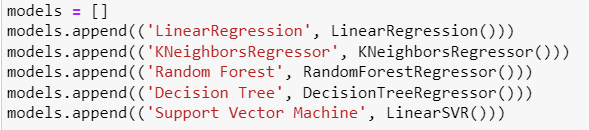


Figure : Loading models

After training, model was tested and evaluated with the testing set to check the performance of the models. As this is a regression problem and I used following evaluation metrices to evaluate my models.

* R2 Score:
* Mean Absolute Error
* Root Mean Squared Error

R2 or r2 and pronounced "R squared," is the proportion of the variation in the dependent variable that can be predicted based on the independent variable. This proportion is indicated by the letter R2 or r2 and written as "R squared" (s). Mean absolute error" (MAE) refers to a method for determining the degree of error that exists between paired data that describe the same phenomena. The root-mean-square deviation (RMSD), also known as the root-mean-square error (RMSE), is a measurement that is widely used to determine the discrepancies between values (sample or population values) that are predicted by a model or an estimator and the values that are observed. The root means square deviation, often known as the RMSD, is calculated by taking the square root of the second sample moment of the differences in values that were anticipated and those that were observed or calculating the quadratic mean of these differences.

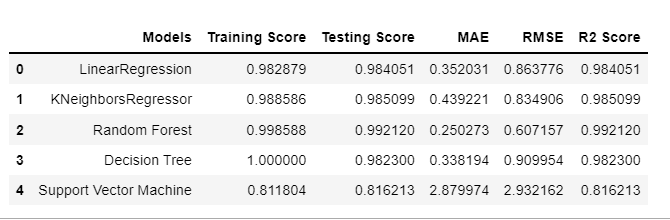


Figure : Performance results

Finally, we have achieved the results of all models with training and testing sets. The above figure is showing the results for all five machine learning models. We can see that with the random forest is showing good accuracy with highest r2 score of 0.992120. It’s training and testing score is also good. Although decision tree is giving 100% training accuracy at but in testing, random forest is leading.

# Conclusion

In this project we applied five machine learning models to predict the suicide rate in different regions around world. The random forest performed best with higher testing accuracy and r2 score. From the analysis done in this study, we concluded that male between the age of 34 to 45 are more prone to suicide. And Africa has the highest suicide rate in male group from 25 to 34. The reason may be the poverty and crisis in the African countries. This study can help the WHO to take the practical measures to stop the suicide ration in these regions.

# References

Abdulkareem, N. a. A. A., 2021. Machine learning classification based on Radom Forest Algorithm: A review. *International Journal of Science and Business,* 5(2), pp. 128-142.

Campbell, C. a. Y. Y., 2011. Learning with support vector machines.. *Synthesis lectures on artificial intelligence and machine learning,* 5(1), pp. 1-95.

Charbuty, B. a. A. A., 2021. Classification based on decision tree algorithm for machine learning.. *Journal of Applied Science and Technology Trend,* 2(1), pp. 20-28.

Jawthari, M. a. S. V., 2021. Predicting students’ academic performance using a modified kNN algorithm.. *Pollack Periodica,,* 16(3), pp. 20-26.

Maulud, D. a. A. A., 2020. A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends,* 1(4), pp. 140-147.