|  |
| --- |
| Suicide Rate Prediction  Project Report |

|  |
| --- |
| Electrical & Computer Engineering & Computer Science (ECECS) |

|  |  |
| --- | --- |
| SPRING 22 |  |



Contents

[Motivation 2](#_Toc101747807)

[Methodology 2](#_Toc101747808)

[Implementation 3](#_Toc101747809)

[Discussion 7](#_Toc101747810)

[Conclusion 8](#_Toc101747811)

[Contributions/References 8](#_Toc101747812)

|  |
| --- |
|  |

## Motivation

The issue of suicide rate has become increasingly rigorous and has received extensive attention in the contemporary society. Suicide has been a serious global public health issue throughout the lifespan.

Suicide rates data and economic development indicators during 1987 to 2016 collected by World Health Organization were dissected. WHO estimates that every year close to 800000 people take their own life, which is one person every 40 seconds. The suicide data in this research were classified by 101 countries, gender, age group, population, HDI, GDP, Region, and GDP per capita.

The main objective of this project is to predict the suicide rates using different Machine learning algorithms. We mainly focus on analyzing what features are responsible for increase in the suicide rates worldwide.

## Methodology

In this project, we predict the suicide rates using the machine learning techniques and analyze the different patterns which cause the increase in suicide rates globally.

* Initially we download the data from [Kaggle](https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016) where the data is collected globally from different sources which has features like country, year, sex, age, suicides\_no, population, suicides/100k pop, country-year, HDI for year, gdp\_for\_year, gdp\_per\_capita, generation.
* Next step is to familiarize with the loaded data.
* In data preprocessing we clean the data apply the techniques to transform the data.
* The next step is visualizing the data to find the relationship between the features and target variables and check how the data is distributed.
* The non-numerical features will be converted to numeric by using Label Encoder which is one of the encoding techniques. We scale the numerical data columns with the RobustScaler.
* After the data preprocessing and visualization, we split the data and fit data to the various classifications models and evaluate the accuracy and Root mean square error.
* The Supervised Machine Learning Algorithms used in our project are:
  + - K- Nearest neighbors
    - Linear Regression – Given in paper
    - Decision Tree
    - Random Forest
    - Gradient Boosting
    - Bagging Regressor

## Implementation

**Data Collection**

Import all the required libraries like Pandas, NumPy, sklearn, seaborn. The Dataset is extracted from [Kaggle](https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016) which contains around 27820 instances with 12 attributes. Loading the data into the Data frame:

Graphical user interface, text, application

Description automatically generated

**Familiarizing with Data:**

Here we can see the information about the shape of the data, columns of the data and the data types of the given data. Graphical user interface, text, application

Description automatically generated

We can observe that HDI for year has null values. There are 6 different age groups in age column. There are total 101 countries in the data set.

**Data Preprocessing**

Here we clean the data and transform the data. In the HDI for year column there are 19456 null values. We replace the null values with the mean of the HDI for year.

Graphical user interface, text, application, email

Description automatically generated

We can also observe that Country-year column is a combination of both country and year columns, so we are dropping the Country-year column from the data. The shape of data changes to (27820,11) after dropping the country-year column.

Graphical user interface, text, application, email

Description automatically generated

We use one of the encoding techniques (**LabelEncoder)** to convert data into numerical form for the columns Country, year, sex, generation.A picture containing graphical user interface

Description automatically generated

**Data Visualization**

After visualizing the data, we have some observations like:

* The number of suicides in male population is higher than that of the female population.
* The age group of people between 35-54 years are committing more suicides.
* In generations, boomers generation has a greater number of suicides.
* From the bar chart we can see Lithuania country has a greater number of suicides

**Splitting Data**

Here we split the data into training and testing sets. We consider Suicides/100K pop column as target feature and assign it to y. For x drop the target feature and assign rest of the data to X. After splitting the data, we can observe Train data has 22256 instances and Test data has 5564 instances.

Graphical user interface, text, application

Description automatically generated

**Model Building**

After splitting the data, we train the models using different algorithms. Here we can look at the different models and the accuracy and Root mean square error of both the Train and Test data sets.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Machine Learning Model | Accuracy of Training data | Accuracy of Testing data | Root Mean Square Error of Train Data | Root Mean Square Error of Test Data |
| K- Nearest Neighbor | 0.94 | 0.80 | 0.29 | 0.56 |
| Decision Tree | 0.98 | 0.97 | 0.15 | 0.22 |
| Random Forest | 0.99 | 0.99 | 0.10 | 0.14 |
| Gradient Boosting | 0.99 | 0.98 | 0.14 | 0.19 |
| Bagging Regression | 0.99 | 0.98 | 0.11 | 0.18 |
| Linear Regression | 0.29 | 0.30 | 1.01 | 1.04 |

## Discussion

Techniques using Machine Learning model in predicting the suicide rate predictions shows that Male population has more suicides that female. Also, the suicides in developed countries are less than that of the developing countries. Among the Machine learning techniques used Random Forest is better. It has the better accuracy than the other models.

In this approach the missing of HDI values might affect the performance as there were 19456 null values out of 27820 values. We used the mean value of the HDI column to replace in place of the missing values as per the understanding of the domain knowledge.

The accuracy can be further improved by using additional features and good amount of data. We can also use algorithms like Multilayer Perceptrons (MLP), XGBoost etc.

## Conclusion

The Model is build to predict the suicide rates. We used K- nearest neighbor, Decision Tree, Random Forest, Linear Regression, Gradient Boosting, Bagging regression.

Random Forest Classifier gives better performance with:

* Train Accuracy: 0.99
* Test Accuracy: 0.99
* Train RMSE: 0.10
* Test RMSE: 0.14

## Contributions/References

1. “Suicide Prediction Analysis with Generalized Addictive Model – Authors - Jun Shen;Shihui Zhao;Mingzi Ye” <https://ieeexplore-ieee-org.unh-proxy01.newhaven.edu/document/8993091>
2. <https://www.kaggle.com/datasets/russellyates88/suicide-rates-overview-1985-to-2016>