MEF UNIVERSITY

SUICIDE TENDENCY CLASSIFICATION AND SUICIDE NUMBER PREDICTION FOR POPULATION SUBGROUPS

Capstone Project

Mehmet Ak

İSTANBUL, 2019

MEF UNIVERSITY

SUICIDE TENDENCY CLASSIFICATION AND SUICIDE NUMBER PREDICTION FOR POPULATION SUBGROUPS

Capstone Project

Mehmet Ak

Advisor: Asst. Prof. Hande Küçükaydın

İSTANBUL, 2019

MEF UNIVERSITY

Name of the project: Suicide Tendency Classification and Suicide Number Prediction For Population Subgroups
Name/Last Name of the Student: Mehmet Ak
Date of Thesis Defense: 09/09/2019

I hereby state that the graduation project prepared by Your Name (Title Format) has been completed under my supervision. I accept this work as a "Graduation Project".

09/09/2019

Asst. Prof. Hande Küçükaydın

I hereby state that I have examined this graduation project by Your Name (Title Format) which is accepted by his supervisor. This work is accepTable as a graduation project and the student is eligible to take the graduation project examination.

09/09/2019

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Asst. Prof. Hande Küçükaydın	……..
2. ……..	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Mehmet Ak Date Signature

EXECUTIVE SUMMARY

SUICIDE TENDENCY CLASSIFICATION AND SUICIDE NUMBER PREDICTION FOR POPULATION SUBGROUPS

Mehmet Ak

Advisor: Asst. Prof. Hande Küçükaydın

AUGUST, 2019, 17 Pages

Suicide is becoming a bigger problem for the world day by day and detecting population subgroups who are more prone to suicide is seen as one of the most important steps for taking precautions to decrease the suicide rates. This study consists of five machine learning models for suicide tendency classification and three machine learning models for prediction of suicide numbers by population subgroups. The dataset provided by World Health Organization is used in the project. Obtained models classify population subgroups as suicide-prone or less suicide prone with 86% accuracy and explain 90 % of the variance in the suicide number per 100,000 population of specific countries.

Key Words: suicide tendency, population subgroups, classification, regression

ÖZET

POPÜLASYON ALT GRUPLARI BAZINDA İNTİHAR EĞİLİMİ SINIFLANDIRMASI VE İNTİHAR SAYISI TAHMİNLEMESİ

Öğrencinin Adı Mehmet Ak

Tez Danışmanı: Asst. Prof. Hande Küçükaydın

AĞUSTOS, 2019, 17 Sayfa

İntihar, gün geçtikçe dünya için daha büyük bir problem haline gelmekte ve intihara daha yatkın nüfus alt gruplarının tespit edilmesi intihar oranlarını azaltmak için önlem almanın en önemli adımlardan biri olarak görülmektedir. Bu çalışma, intihar eğilimi sınıflandırması için beş makine öğrenimi modeli ve popülasyon alt grupları tarafından intihar sayısının tahminlenmesi için üç makine öğrenimi modelinden oluşmaktadır. Bu projede Dünya Sağlık Örgütü tarafından sağlanan veri seti kullanılmıştır. Elde edilen modeller popülasyon alt gruplarını intihar eğilimli veya daha az intihar eğilimi olarak % 86 doğrulukla sınıflandırmakta ve bir ülkenin 100.000 populasyon başına düşen intihar sayısındaki varyansın % 90'ını açıklamaktadır.

Anahtar Kelimeler: intihar eğilimi, popülasyon alt grupları, sınıflandırma, regresyon

TABLE OF CONTENTS

.

1. INTRODUCTION

Suicide is getting a bigger problem for the world year by year and because of that it is very important to detect groups that are prone to suicide and understand reasons behind it. Using machine learning methods may provide us with more meaningful results to detect population subgroups which are prone to suicide and to decrease suicide rates with additional precautions.

In the project, we use a historical dataset which contains suicide rates between years 1985 and 2016. Features in the dataset are country, year, gender, age, suicides number, population, number of suicides per 100,000 population of specific country (suicides/100 K pop), human development index (HDI) for each year, gross domestic product (GDP) for each year (in $), GDP per capita (in $), and generation. Firstly, we determine the correlation between the features and the suicide rates and useful features are selected for classification model. After this phase, various classification and regression algorithms are used for classification of population subgroups and for prediction of number of suicides per 100,000 population of the specific country.

2. LITERATURE REVIEW

The main focus of the articles in the literature is generally to solve the relationship between financial changes and suicide ratio for different countries or age groups. Shah (2010) studies the impact of internet usage on elderly people's suicide ratios. The author found out that there is a significant positive correlation between internet usage and suicide rates regarding people over 65 ages for both genders. Peeter (2012) examines the general characteristics of the suicide rates related to different countries in the world. It turns out that the countries with highest suicide rates are India, Russia, USA, Japan, South Korea and China. Yin et al. (2016) address the relationship between suicide ratios, economic growth and stock market prices. They found out that there is a negative correlation between suicide ratios and economic growth. No significant correlation is detected between suicide rates and stock market prices. Jalal and Nahid (2016) study the relationship between smoking habit and suicide tendency. The study showed that there is a positive correlation between smoking and suicide tendency. The paper of Khazaei et al. (2017) focuses on the impact of different human development indices to suicide ratios in different countries. The authors point out that different human development indices have impact on the suicide ratio differences between countries. The study of Hyun et al. (2018) is related to the suicide rates of different 29 OECD countries and the authors indicate that being in the late adolescence period and being a man are the main factors resulting in relatively high suicide ratios in OECD countries. Johan (2018) examined the main risk factors of suicide of young people. The author determined that being in the late adolescence is the most important factor regarding young people suicide. Other sub-factors are determined as personality characteristics, family factors and mental disorders. The study of Deborah et al. (2018) addressed suicide rate trends in several states of USA. The study showed that suicide rates are significantly increased in 44 states of USA. Cora et al. (2018) examined the relationship between occupation groups and suicide rates. The authors showed that the construction sector is the most suicide-prone occupation group for men. Art, design and entertainment are found as the most suicide-prone occupation groups for women.

3. Exploratory Data Analysis

3.1 Data Cleaning

The source of the data is World Health Organization and it is obtained via Kaggle. Our analysis is built on the data set consisting of 27,820 suicide rates per country-gender-age group between years 1987 and 2016.

Column Name	Column Definition
Country	Represents the country of the observation
Year	Represents the year of observation
Gender	Represent the gender of the observation
Age	Represents the age group of the observation
Suicides No	Represents the total suicide number of specific gender-age group-country
Population	Represents the population of the country of measurement
Suicides/100 K Pop	Represents the suicide number of specific gender-age group-country per 100,000 population of the specific country
HDI for Year	Represent the human development index for the specific country in the measurement year
GDP for Year	Represent the gross domestic products for the specific country in the measurement year
GDP per Capita	Represents the ratio of gross domestic products per person in the country in the measurement year
Generation	Represents the specific terms for all of the people born and living at about the same time

Table 1: Definitions of columns in the dataset

Table 1 summarizes each column of the data set and its definition. Before applying classification techniques on data, country column is discarded since the aim of the analysis to classify suicide tendency independently from country. Moreover, human development index column is discarded, since 69% of the values are missing.

Statistical Metric Name	Statistical Metric Value
Mean	12.81
Standard Deviation	18.96
Minimum Value	0
25th Percentile Starting Point	0.92
50th Percentile Starting Point	5.99
75th Percentile Starting Point	16.62
Maximum Value	224.97

Table 2: Statistical metrics for number of suicides per 100,000 population of the specific country

Statistical metrics for the number of suicides per 100,000 population of the specific country is given in Table 2. Target column is the number of suicides/100 K population, which represents the number of suicides per 100,000 population of the specific country. Values are between zero and 224, but the mean is 12 and the 75th percentile is 16.62, which indicates that the number of suicides /100 K pop has a right skewed distribution.

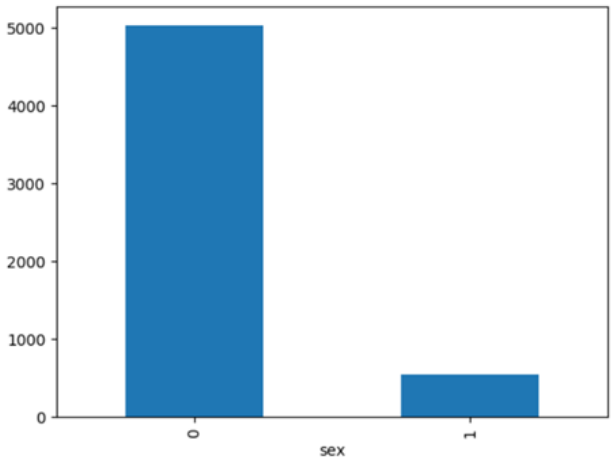

Figure 1: Histogram of suicide-prone groups by gender

Figure 1 shows the histogram of the target variable on gender distribution. It can be seen that men are more suicide-prone than women.

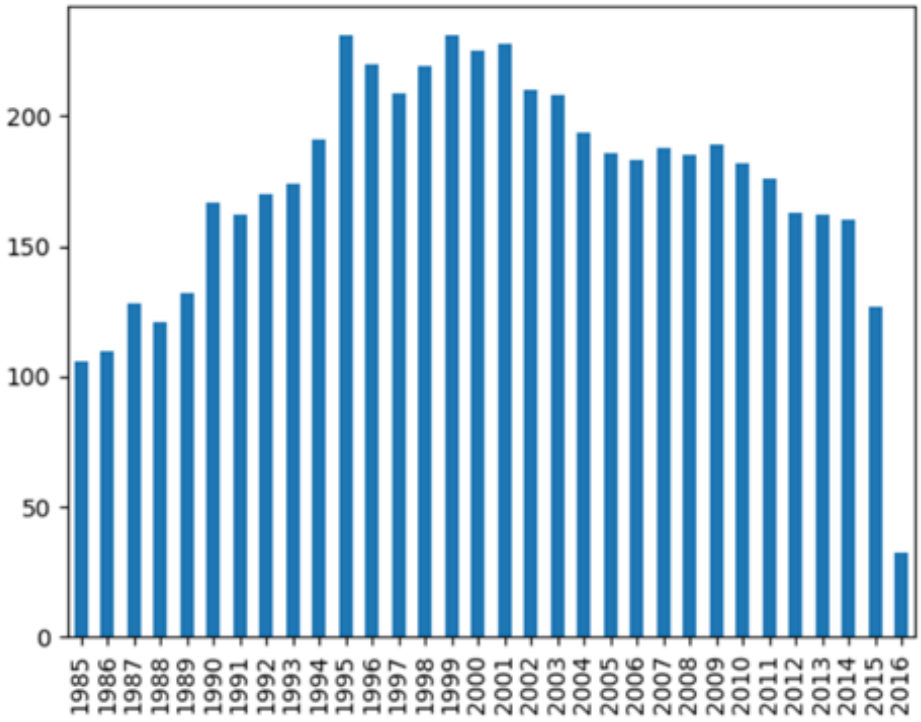

Figure 2: Histogram of suicide-prone groups by year

Figure 2 shows the histogram of the target variable on yearly distribution. It can be seen that there is an increasing trend from 1985 to 2001. However, after 2001 there is a decresing trend.

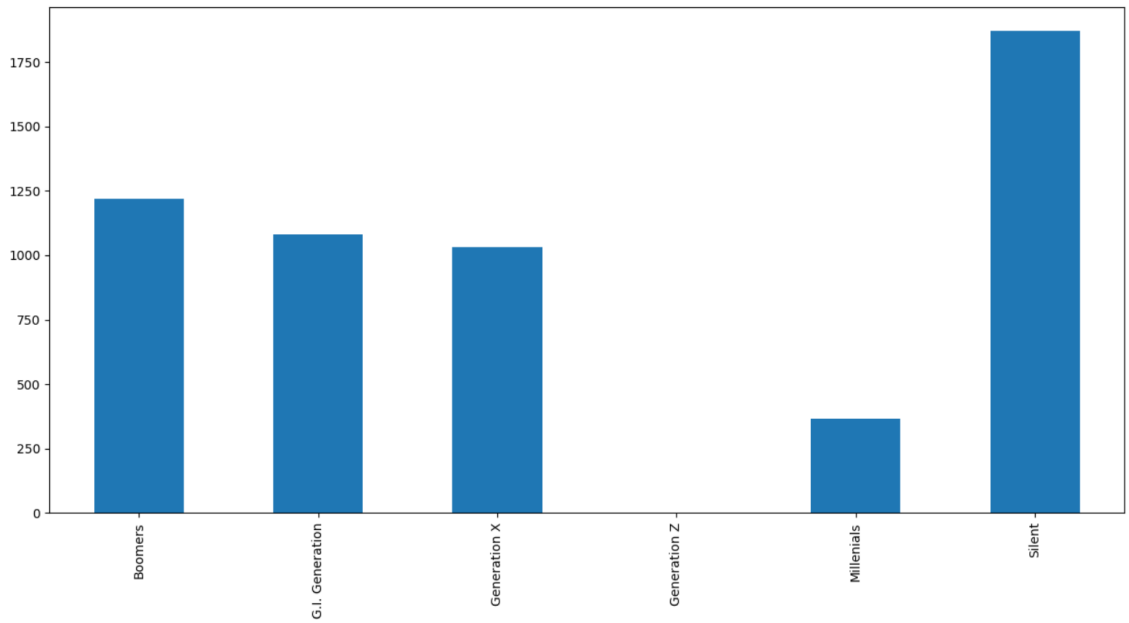

Figure 3: Histogram of suicide-prone groups by generation

In Figure 3, there is a histogram to understand suicide-prone groups by generation. We can see that there are different patterns for generations. It is clear that generation Z is less suicide prone group and generation Silent is the most suicide prone group, and as a result we can say that generation feature may help us for suicide tendency classification.

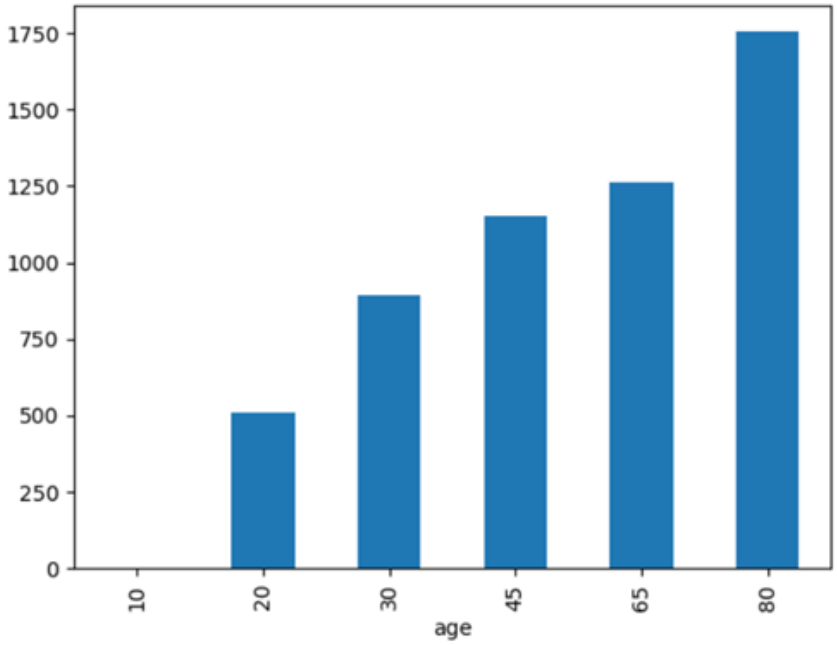

Figure 4: Histogram of suicide-prone groups by age

We can see the histogram related to the suicide-prone group distribution by age in Figure 4. The histogram shows that there is a positive correlation between suicide tendency and age, which indicates that the more people get older the more they are suicide-prone.

Feature name	Correlation coefficient between feature and target
Gender	-0.83
Age	0.36
Year	0.3
Generation	0.26
GDP per capita	-0.19

Table 3: Correlation coefficients of features and target value

A correlation matrix is given in Table 3 to understand the relation between target value and each feature. The strongest correlation is observed between target and gender feature with a 0.83 negative correlation coefficient. The correlation is negative, since a zero is assigned

to male gender, whereas a one is assigned to female gender. Men are more prone to suicide and as a result if the gender is male, in other words if the feature takes the value “zero”, the suicide number increases. In case it is woman, i.e. the feature takes the value “one”, the suicide number decreases and the situation creates a negative correlation between gender feature and suicide numbers. Gender feature helps us a lot for suicide tendency classification with correlation coefficient and the lowest correlation value is observed between the target and the GDP per capita.

3.2 Transformation of Data

After cleaning the data, five main features and target values are left in the dataset: year, gender, age group, GDP per capita, generation are features and the target column is selected as number of suicides per 100,000 population of the specific country. Gender, age group and generation features are categorical. One hot encoding is a process to create separate columns for all of the options in a feature. Newly created separate columns are assigned a one for correct option and zero for other options. One hot encoding is applied to gender, age group and generation features to help machine learning algorithms make sounder predictions.

Number of suicides per 100,000 population of the specific country values are turned into categorical values. The 80^{th} percentile is determined as 20.53 suicides per 100,000 population and suicide values over 20.53 are labelled as suicide tendency group one (more suicide-prone). Values under 20.53 are labelled as zero (less suicide-prone). The 70^{th} percentile is determined as 13.56 suicides per 100,000 population and suicide values over 13.56 are labelled as suicide tendency group one (more suicide-prone). Values under 13.56 are labelled as zero (less suicide-prone).

3.3 Data Scaling

After one hot encoding is applied, categorical features are transformed into columns which contain only zero or one. GDP per capitas vary between 251 dollars and 126,352 dollars and years vary between 1985 and 2016. Using features with different ranges may cause biased models, since model algorithms directly use these values related to features.

However, their importance is not the same. There are scaling techniques to prevent deficiencies related to imbalanced feature value ranges.

MinMax scaling transforms the values of the feautures and makes the values range between zero and one. MinMax scaling is performed to year and GDP per capita features to prevent problems due to scale differences between features.

4- METHODOLOGY

Our first aim in the project is to classify population subgroups as suicide prone or less suicide prone with an accuracy over 70%. In classification part of our study, we use five types of classification algorithms to see the result differences and select the best model for the case. These are logistic regression, boosted decision tree, decision forest, support vector machine classifier and neural networks.

Logistic regression is an algorithm that uses coefficient of logistics regression for binary classification. (Walker and Duncan, 1967). Boosted decision tree is an algorithm, in which new trees learn from the errors of the previous tree and prediction is made considering all of the tree results. Decision forest algorithm creates multiple decision trees and the result of the most accurate tree is selected as the model prediction (Quinlan, 1987). Support vector machine classifies the data via a hyperplane in two-dimensional space (De Costa, 2002). Neural Networks algorithm are inspired from human brain and uses nodes called artificial neurons (Riesenhuber & Poggio, 1999).

Our second aim in the project is to predict the number of suicides per 100,000 population of the specific country with a coefficient of determination over 70%. In prediction part of our study, we use three types of regression algorithms to predict the suicide number per 100,000 population of the specific country using country, year, gender, age, GDP per capita and generation features. These algorithms are boosted decision tree regression, linear regression, and neural networks regression.

Linear regression tries to predict the dependent variable via using multiple independent variables. It tries to create a formulation with proper coefficients related to independent

variables (Walker and Duncan, 1967). Boosted decision tree creates multiple regression trees, where each tree learns from the errors of the previous tree via a loss function defined by the user (Quinlan, 1987). Neural networks regression is a method that is inspired from human brain and uses nodes called artificial neurons to predict the dependent variables (Riesenhuber & Poggio, 1999).

Models which are used in the project contain a lot of hyperparameters and the predictions of the models are different for different hyperparameter values. Grid search is a method used in hyperparameter optimization. It is used to determine the optimal hyperparameters of a model, as a result we can obtain more accurate predictions (Bergstra and Bengio, 2012). We apply grid search to get more accurate results in all of the models in this project.

There are several metrics to evaluate for the classification model performance. These are accuracy, area under curve, F1 score, precision, and recall. Accuracy is the overall correctness ratio of the classifications made by model (Taylor, 1999). Area under curve is a metric that varies between 0.5 and 1. It represents the area over the diagonal in the correctness graph. Precision is the ratio of number of true positives to total of true positives and false positives. Recall is the ratio of true positives to total of true positives and false negatives. F1 score is a metric which considers recall and precision together to calculate the performance (Sasaki, 2007)

We select accuracy and area under curve as classification performance metrics, because we want to measure the overall classification correctness of the models. Furthermore, area under curve represents the total power of the model under different sensitivities related to false positive and false negative ratios.

There are also several different metrics to evaluate the regression model performance. These are mean absolute error, root mean squared error, relative absolute error and coefficient of determination.

Mean absolute error is the ratio between total of absolute errors and number of errors. Root mean squared error is the standard deviation of the prediction errors and coefficient of determination is a metric, which shows that how the model explains the variance in the label (Taylor, 1999).

We choose mean absolute error to see the average of errors and the coefficient of determination to understand, how well the model explains the variance in the target.

4.1 Classification Algorithms Results

The aim of the classification is to classify population subgroups as suicide-prone or less suicide-prone. We use boosted decision tree, decision forest, neural networks, logistic regression and support vector machine algorithms for the classification.

Rank	Algorithm for the 80th percentile	Accuracy	Area under curve	Precision	Recall	F1
1	Boosted decision tree	0.867	0.914	0.694	0.626	0.658
2	Decision forest	0.857	0.899	0.677	0.570	0.619
3	Neural networks	0.829	0.880	0.623	0.406	0.492
5	Logistic regression	0.825	0.872	0.610	0.397	0.481
4	Support vector machine	0.815	0.849	0.543	0.590	0.565

Table 4: Comparison of classification algorithms for 80th percentile label

Rank	Algorithm for the 70th percentile	Accuracy	Area under curve	Precision	Recall	F1
1	Boosted decision tree	0.844	0.915	0.745	0.736	0.741
2	Decision forest	0.823	0.894	0.713	0.696	0.704
3	Neural networks	0.802	0.877	0.692	0.621	0.655
4	Logistic regression	0.801	0.866	0.662	0.701	0.681
5	Support vector machine	0.800	0.823	0.651	0.731	0.688

Table 5: Comparison of classification algorithms for 70th percentile label

Performance comparison of the classification models are given in the Table 4 and the Table 5. Accuracy of the classification models vary between 0.815 and 0.867 for label over 80th percentile threshold and vary between 0.800 and 0.844 for label over 70th percentile threshold. This means that the accuracy of all classification models is more than 70%, which is the aim that we determine at the beginning.

The most successful model is boosted decision tree and its accuracy is 0.867, while the least successful model is support vector machine with an accuracy of 0.815. The performance of five classification models are very similar considering the accuracy difference between the most successful model and the least successful model.

Two class boosted decision tree may perform better than the support vector machine and logistic regression because of the complexity of the relationship between features. Error correction function of previous trees in boosted decision tree algorithm may help to provide us with more accurate results.

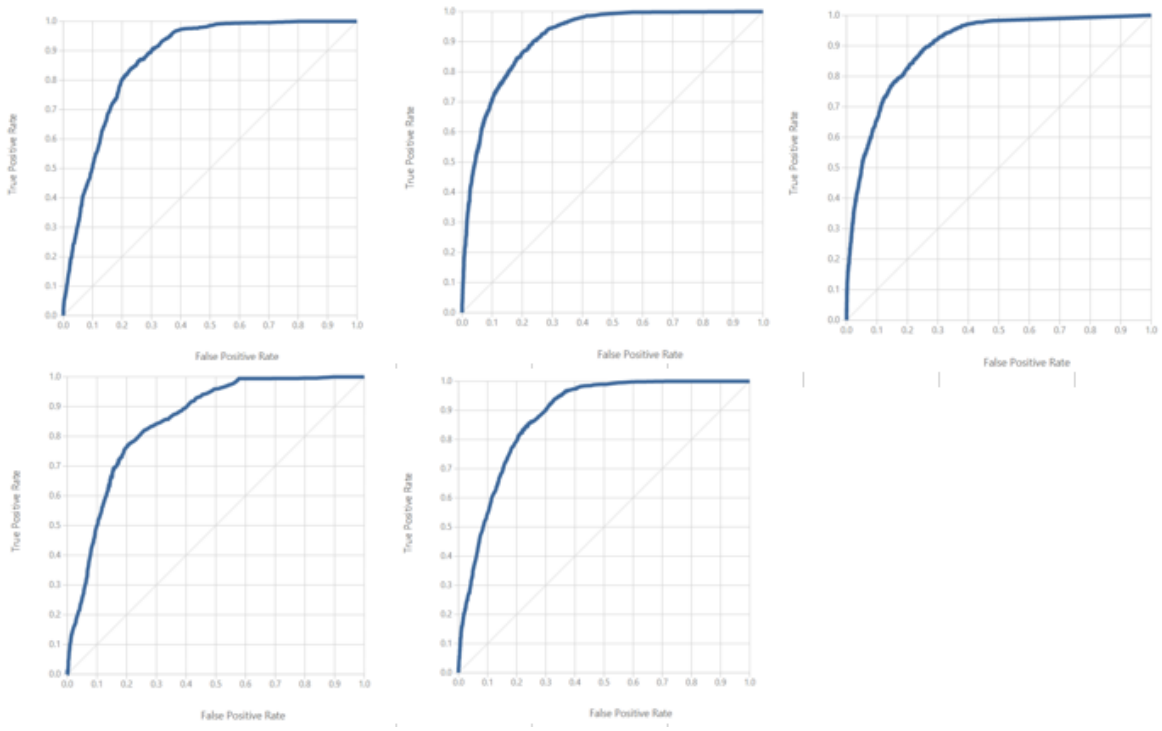

Figure 5: Area under curve graphs of classification models

Area under curve graphs related to classification models are given in Figure 5. It can be seen that graphs are very similar to each other, which suggests that the performances of the models regarding area under curve score is similar to each other. The more the graph is far away from the diagonal line, the more we can accept it as a better classifier. Furthermore, it can be seen that line of the classifier is far away from the diagonal line for all of the five classifiers.

Area under curve scores are also in line with the graph results. They vary between 0.849 and 0.914. Considering the best area under curve score as one, it can be said that models classify the population sub-groups as "suicide-prone" and "less suicide prone" with adequate success.

	Predicted Positive	Predicted Negative	
Real Positive	True Positive: 1067	False Negative:637	1704
Real Negative	False Positive:471	True Negative:6171	6642
	1538	6808	

	Predicted Positive	Predicted Negative	
Real Positive	True Positive: 971	False Negative:733	1704
Real Negative	False Positive:464	True Negative:6178	6642
	1435	6911	

Table 6: Confusion matrices of boosted decision tree and decision forest algorithms

Boosted decision tree and decision forest are the best two algorithms considering the classification accuracy results. Their confusion matrices are given in Table 6. It is seen that the performance of boosted decision tree is slightly better than the performance of the decision forest algorithm. Boosted decision tree correctly classifies 1067 suicide-prone observations as suicide-prone and 6171 less-suicide prone observations as less suicide-prone. On the other hand, it classifies 637 suicide-prone observation as less-suicide prone and 471 less suicide-prone observations as suicide-prone. Nevertheless, the misclassification ratio can be evaluated as low considering a total of 8346 observations in the test set.

4.2 Regression Algorithms Results

The aim of the regression is to predict the suicide number per 100,000 population of the specific country. We use boosted decision tree, neural networks and linear regression algorithms for regression.

Rank	Regression Algorithm	Mean Absolute Error	Coefficient of Determination
1	Boosted Decision Tree Regression	2.91	0.90
2	Neural Networks Regression	3.55	0.87
3	Linear Regression	8.64	0.51

Table 7: Performance comparison of regression models

The mean absolute errors and the coefficient of determinations related to regression models are given in Table 6.

Linear regression can only explain 51% of the variance in the suicide number per 100,000 population of the specific country. Our aim is to explain the variance with a coefficient of determination over 70%. Therefore, one can assume that linear regression model is not successful to predict and explain the variance in the label. Linear regression might be insufficient to explain the variance in case there is no adequate linear relationship between features and label.

Performance of boosted decision tree and neural networks are very close to each other considering the coefficient of determination values of the predictions. Boosted decision tree predicted the suicide number per 100,000 population of the specific country with 2.91 average error and explained 90% of the variance in the suicide number per 100,000 population of the specific country. Boosted decision tree regression results are better than the linear regression results and slightly better than the neural networks regression result. As a result, boosted decision tree is the most successful model in three models used for regression.

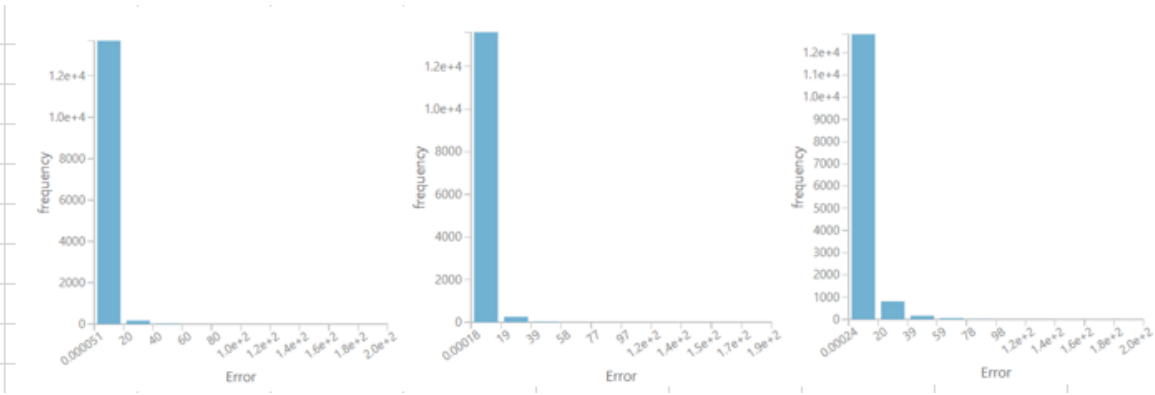

Figure 6: Histogram of regression models' prediction errors

Histogram of regression models' prediction errors are given in Figure 6. It is clear that most of the prediction errors are lower than 20 suicides per 100,000 population for all of the three models. Boosted decision tree and neural networks have almost no prediction errors over 40 suicides per 100,000 population. However, linear regression has errors over 40 suicides per 100,000 population and has more errors between 20 suicides per 100,000 population and 39 suicides per 100,000 population than the two other models. Prediction

error graphs also support the conclusion that boosted decision tree and neural networks are successful at predicting the suicide number, whereas linear regression model is not that much successful at predicting the suicide number.

5. Conclusion

In the study, we apply logistic regression, boosted decision tree, decision forest, support vector machine classifier and neural networks for suicide tendency classification and neural networks regression, boosted decision tree regression and linear regression to predict suicide number per 100,000 population of the specific country.

Performance metrics of the classification models are given in Table 7 for 80^{th} percentile label threshold and Table 8 for 70^{th} percentile label threshold. It is seen that boosted decision tree is the most successful model for both labels with 86% for 80^{th} percentile label threshold and 84% accuracy for 70^{th} percentile label threshold. Model using boosted decision tree correctly classify the population subgroups as suicide-prone or less suicide-prone with 86% accuracy.

Performance metrics of regression models are given in Table 6. It is seen that boosted decision tree regression is the most successful model for regression in our case. The model predicts the suicide number per 100,000 population of the specific country with 2.91 average error and successfully explains the 90% of the variance in suicide number per 100,000 population of the specific country with using country, year, gender, age, GDP per capita and generation features.

It is observed that gender and age are the most important two factors for having high suicide numbers. Men and old people are more suicide prone compared to women and young people.

6. References

Sasaki Y (2007). *The truth of the F-measure.* Teach Tutor mater 1 (5), 1-5

Bergstra J & Bengio Y (2012). *Random Search for Hyper-Parameter Optimization*. Journal of Machine Learning Research. 13: 281–305.

Walker S & Duncan D (1967). "Estimation of the probability of an event as a function of several independent variables". Biometrika. 54 (1/2): 167–178.

Quinlan R (1987). *Simplifying decision trees.* International Journal of Man-Machine Studies. 27 (3): 221–234.

Riesenhuber M.& Poggio T (1999). *Hierarchical models of object recognition in cortex*. Nature Neuroscience. 2 (11): 1019–1025.

DeCoste D (2002). *Training Invariant Support Vector Machines*. Machine Learning. 46: 161–190.

Taylor J (1999). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements.* University Science Books. pp. 128–129.

Shah A (2010). *The relationship between elderly suicide rates and the internet: a cross-national study.* International Journal of Social Psychiatry. 56(3):21-49.

Peeter V (2012). *Suicide in the World.* Int Journal Environ Public Health. 9(3): 760–771.

Yin H, Xu L, Shao Y, Li L & Wan C (2016). *Relationship between suicide rate and economic growth and stock market in the People's Republic of China: 2004-2013.* Neuropsychiatric Disease and Treatment 12: 3119-3128.

Jalal P. & Nahid D. (2016). *Smoking and Suicide: A Meta-Analysis.* PLoS One. 11(7):13-17.

Khazaei S, Armanmehr V, Nematollahi S & Rezaeian S (2017). *Suicide rate in relation to the Human Development Index and other health related factors: A global ecological study from 91 countries.* Journal of Epidemology and Global Health. 7(2):131-134.

Beop-R, Eun H & Hyun J (2018). *A Comparative Study of Suicide Rates among 10–19-Year-Olds in 29 OECD Countries*. Psychiatry Investigation. 15(4): 376–383.

Johan B. (2018). *Suicide and Youth: Risk Factors.* Front Psychiatry. 9: 540.

Deborah M, Thomas R, Katherine A, Scott R, Keming Y, Kristin M, Asha Z. & Alex E. (2018). *Vital Signs: Trends in State Suicide Rates — United States, 1999–2016 and Circumstances Contributing to Suicide.* Weekly. 67(22);617–624

Cora P, Deborah M, Suzanne M, Pamela K, Hope M, Wendy L, Colby N, Aimée R, Brad B, & Feijun L (2018). *Suicide Rates by Major Occupational Group — 17 States, 2012 and 2015.* Mortal Weekly Report. 67(45): 1253–1260.

7. TABLES AND FIGURES

Table List

Figure List