Insight about Suicide of Worldwide and Korea: Classification according to suicide rate & Predict a suicide rate

Kim Dong Uk
Korea University
tigerrhs@naver.com

Nam Jeong Jae
Korea University
njj97@naver.com

Park Seong Ho
Korea University
rkwlgh@naver.com

Abstract

In this project, classify countries according to suicide rate and predict the number of suicides. According to the World Health Organization (WHO), about 1 million people commit suicide every year. In Korea, the seriousness of suicide is gradually increasing as celebrities commit suicide. We selected Suicide Rates Overview data set to recognize the seriousness of suicide and to suggest suicide prevention. The data is organized by year (1985 -2016), by country (100 or more countries). To find a better insight, would like to find the factors that affect the risk of suicide through ML analysis. We would like to proceed with the analysis in two ways in total. First is classification. A total of 4 labels are given based on the normalized suicide count. Through this, the classification model is trained. Second is regression. Estimating the number of suicides by country or the world. Additionally, we will conduct data analysis on Korea, which has the highest suicide rate among OECD countries.

1. Introduction

Looking at recent news and media reports, one can see that articles about suicide are constantly coming out. Whenever we encounter these articles, we wonder whether we take suicide seriously and are actually interested in it, and whether we are on the right path to prevent it.

South Korea, which has the highest suicide rate among OECD countries. In order to understand the reality of suicide in Korea and the seriousness of it, the above topics were selected, and related data were dealt with and felt directly. Also, suicide, not only in Korea, is a serious problem all over the world. We wanted to learn the seriousness of suicide worldwide through visualization. They also wanted to find out what was related to suicide. What are some things that are closely related to the suicide rate? Through the process so far, we have been able to visualize the number of suicides worldwide to understand the reality and seriousness of suicide.

1.1. World Suicide Status

First, we graph the changes in the number of suicides around the world each year to understand the reality of suicide worldwide. The annual number of suicides worldwide is as follows:

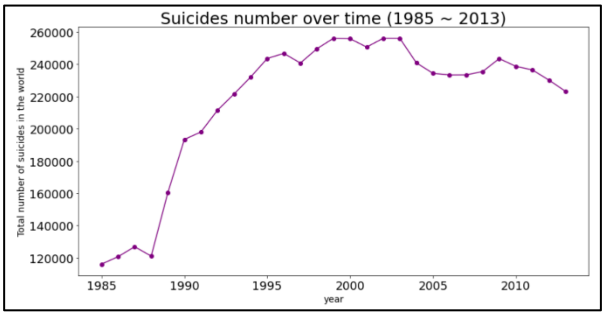


Fig 1 : Number of suicides worldwide by year 1985-2013

As you can see from the graph above, it increased significantly from the late 1980s to the 2000s. And the problem is that there has been a slight decline since 2000, but still a fairly high number of suicides for a long time.

Next is the pie chart of suicide rates by gender and age, and the graph of the number of suicides by gender and age. It also graphs the number of suicides according to the ratio of men and women in each age group.

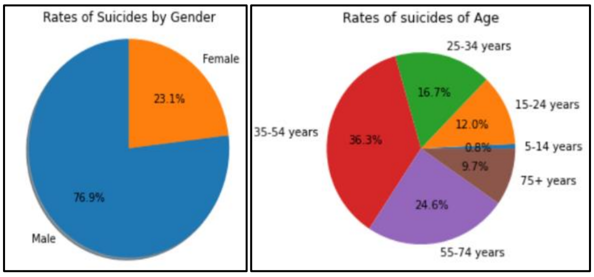


Fig 2 : The pie chart of suicide rates by gender and age

The chart above shows that the number of suicides among men is about three times higher than that of women. It can also be confirmed that those aged 35-54 and those aged 55-74 make up more than 60 percent of the suicide rate.

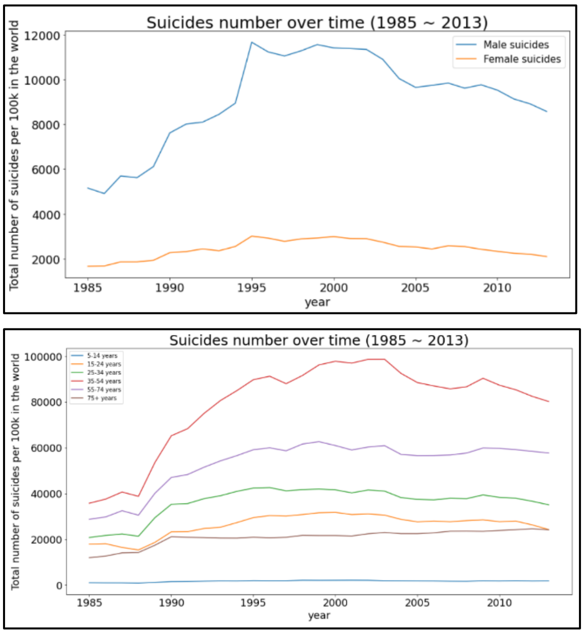


Fig 3 : The graph of the number of suicides by age and gender

The first graph is a graph of time series of suicides divided by gender. The graph shows that men and women have a much higher number of suicides than women, although both men and women seem to have similar tendencies.

The second graph is a time series of suicides divided by age group. The age group with the highest number of suicides is 35-54, and the lowest is 5-14 years old. Data from ages 35 to 54 show a slight decrease since 2000, which has affected the overall number of suicides. Except for 35-54, there is no significant decline in other age groups, and the number of age groups over 75 is increasing.

In summary, worldwide data visualizations above can be summarized into a total of 3 facts below.

1. Compared to the same age group, men are more suicidal than women.

2. The highest number of suicides is in the order of '35-54', '55-74', ‘25-34', '15-24', '75+' and '5-14'.

3. In particular, men between '35-74' show an overwhelmingly high number of suicides compared to other age groups.

1.2. Korea Suicide Status

As mentioned earlier, the status of suicide in Korea is very serious among countries around the world, and the articles can be easily found. So, using the suicide data we have, we decided to get information about how serious the situation of suicide in Korea is.

Below, the average number of suicides in the top 15 countries were displayed graphically.

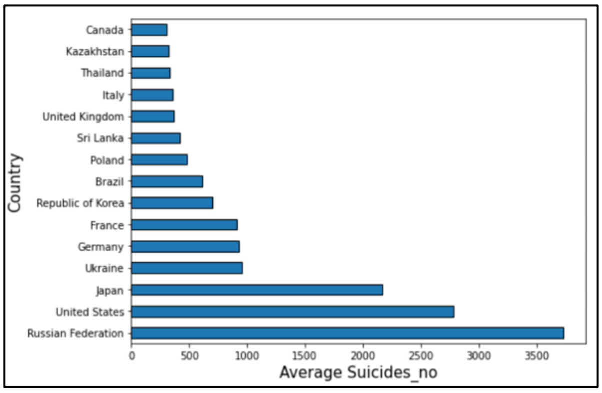


Fig 4 : Average number of suicides in the top 15 countries

The above graphs show the top 10 countries and their corresponding statistics based on 'Average Suicides no', 'Total Suicides no' and 'Total Suicides/100k', respectively. The results using ‘Suicides no’ determined that the impact of the country's population was significant. So, we also made the graph that represents the top 30 countries in the order in which the population by country is large.

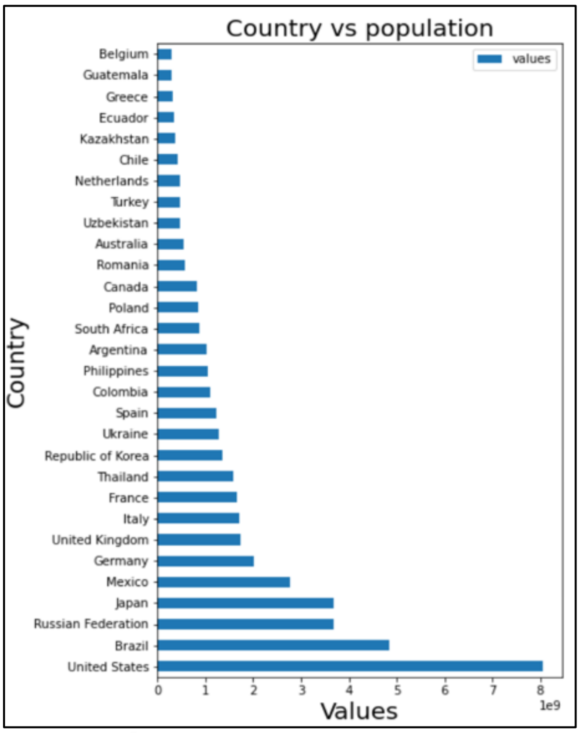


Fig 5 : The top 30 countries in population by country

According to this graph, the United States had the largest population, followed by Brazil, Russia, Japan, and Mexico. South Korea ranked 11th overall. Here's the question. Korea ranked 11th in population ranking, and the average suicide rate was 7th, 7th, and 11th on average.

This shows that Korea has a higher number of suicides and suicide rates compared to its population. we could once again feel the urgency of taking measures against suicide in Korea.

You can see that the number of suicides is constantly increasing. From visualization above, we felt the severity of suicide rates in Korea one more time. So, we decided to analyze more about the Korean suicide data in section 5.

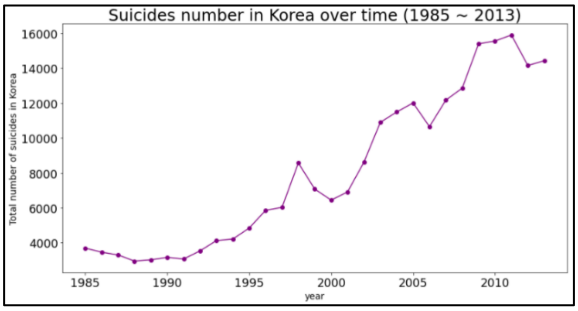


Fig 6 : Number of suicides in Korea by year 1985-2013

2. Related Works

We divide our discussion of related literature into two sections: Extraction of Insight through in-depth analysis of Korean data as well as global data, and Optimal Classification and Regression Model Selection.

2.1. In-depth Analysis of Korea

Preceding study of suicide, there is a study through the statistical approach of Durkheim and positivists. In their study, the analysis was applied to Europe and the United States to compare and analyze rural-urban residence, marital status, gender, race, occupation, etc. variables and suicide rates. Durkheim tried to establish sociology as empirical knowledge of revealing the laws of society through suicide research. Durkheim recognized suicide as a social fact arising from a socially determined structure and identified the causes of suicide as a framework for social regulation and integration (Durkheim, 1897). However, the above study was conducted based on the Western environment, such as Europe and the United States, without considering the cultural differences between the East and the West. On the other hand, we wanted to conduct an in-depth analysis of Korea to compare the reality of the world with that of Korea at the same time.

2.2. Optimal Model Selection

Preceding study of suicide-related national, there is a cross-country comparative analysis study of the social causes of suicide. In their study, based on the theory of anomie and institutional anomie, the hypothesis of causal relationship between suicide rate and social integration, relative deprivation, and anomie was established and comparative studies were conducted among countries (Dong-joon Shin, 2012). In carrying out the above study, the Gini coefficient and statistical significance were examined and the countries were compared and classified. On the other hand, our research conducted classification and prediction using ML method, measured performance by using the corresponding indicators for each, and selected the optimal model. We would like to mention in this regard that predictions are made through regression methods as well as classification.

3. Dataset and Features

3.1. Dataset Description

This compiled dataset pulled from four other datasets linked by time and place, and was built to find signals correlated to increased suicide rates among different cohorts globally, across the socio-economic spectrum. Features are shown in the following table.

Columns	Description
country (string)	Country name
year (numeric)	1985-2016
sex (string)	Male or Female
Age (string)	5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, 75+ years
suicides_no (numeric)	Number of suicides
population (numeric)	Number of population
suicides/100k pop (numeric)	Suicide Count per 100,000 people
country-year (string)	Country Name - 1985-2016
HDI for year (numeric)	Human Development Index (calculated by combining life expectancy, educational indicators, and living standards)
gdp_for_year ($) (numeric)	Annual gross domestic product
gdp_per_capita ($) (numeric)	GDP per capita (GDP per country / total population)
Generation (string)	Boomers, G.I. Generation, Generation X, Generation Z, Millenials, Silent

Table 1 : Description of Dataset

There are currently no labels in this dataset. Therefore, we cannot proceed with classification. However, we are going to do the analysis in two ways. The first is classification. Based on the Suicides_100k_pop column, we will create a total of 4 classes through categorization and treat them as labels for classification. The second is regression. we want to understand the trend of the number of suicides by year.

3.2. Feature Visualization

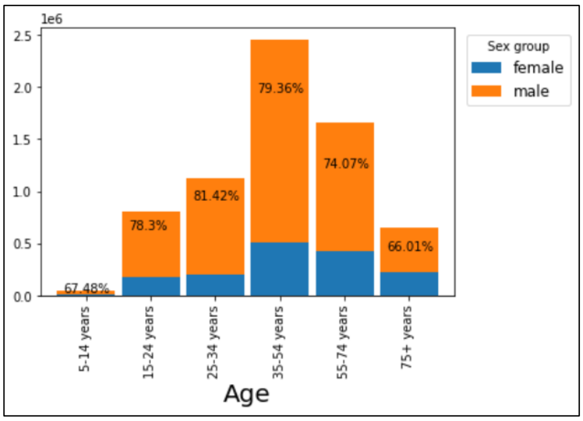


Figure 7 : The proportion of males and females in each age group

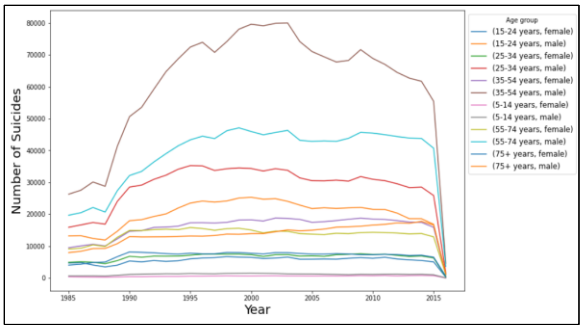
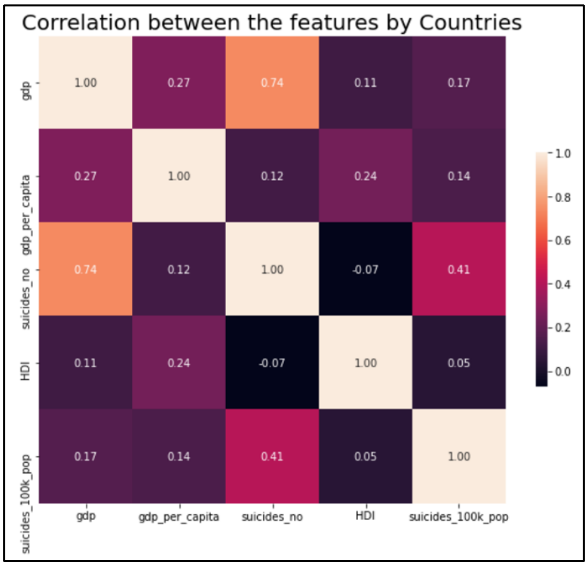


Figure 8 : Number of suicides by year, age and gender

This graph is by gender, age, and year. When it comes to gender, men have a much higher suicide rate than women. In the age group, both men and women in their 30s to 40s have the highest suicide rate. And the higher the age group, the higher the suicide rate is. Compared to other features, it can be seen that especially age and gender show a big difference according to the number of suicides.



	gdp	gdp_per_capita	suicides_no	HDI	suicides_100k_pop
gdp	1.00	0.27	0.74	0.11	0.17
gdp_per_capita	0.27	1.00	0.12	0.24	0.14
suicides_no	0.74	0.12	1.00	-0.07	0.41
HDI	0.11	0.24	-0.07	1.00	0.05
suicides_100k_pop	0.17	0.14	0.41	0.05	1.00

Figure 9 : Correlation between features

Looking at the correlation, it can be seen that gdp and suicide count have a high correlation with 0.74.

4. Methods

4.1. Classification

There are a total of 5 models to be used for classification training, and f1 score was used as a performance measure. The model is as follows.

4.1.1. Preprocessing

First is the processing of missing values. The NA values of the HDI_for_year column were treated as 0. Secondly, one hot encoding was performed for the sex, age, and year columns. Third is feature selection. These columns were excluded for the reasons of the visualization and for the following reasons.

Drop Features	Reason
Country, Country_year	ID value
Generation	Can be estimated by age and year
Suicide_100k_pop	Suicides / population, Used as a standard when creating labels
Gdp_for_year	Only certain countries have high gdp, and the gap is large

Table 2 : Drop columns

Finally, categorization for classification. These graphs represent the distribution of suicide rates, the frequency of the categorized suicide rate labels, and the frequency of the suicide rate labels resampled to match the frequency differences, respectively.

Labels	Suicides_100k_pop
Z0	0 ~ 19
Z1	20 ~ 39
Z2	40 ~ 79
Z3	80 ~

Table 3 : Categorization for classification

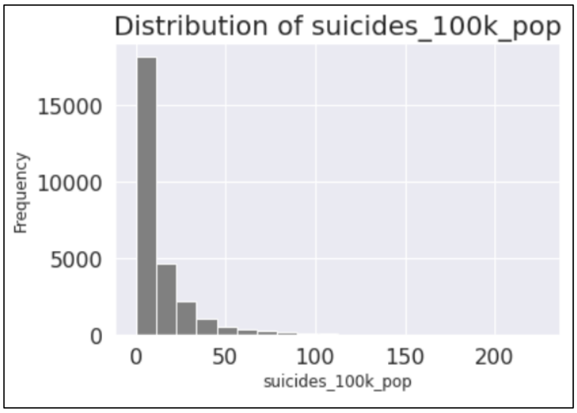


Figure 10 : Distribution of ‘suicides_100k_pop’

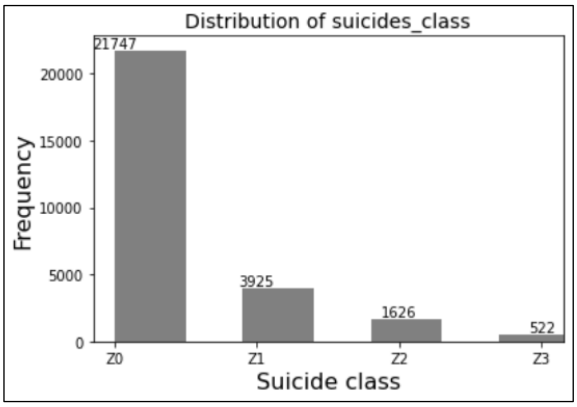


Figure 11 : Numbers of each suicides_class

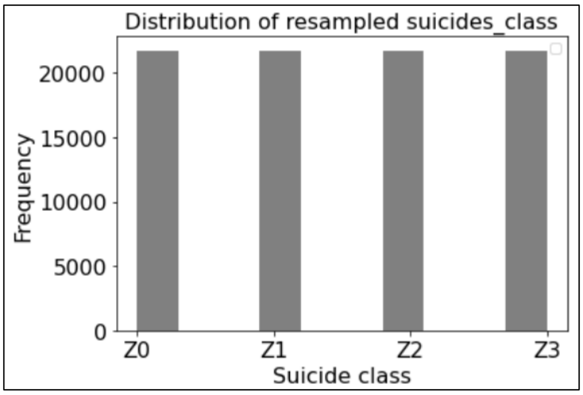


Figure 12 : resampled suicides_class

4.1.2. Random Forest

The parameters of the best model are as follows.

Parameters	Value
Criterion	Entropy
Max_depth	None
N_estimators	32
Max_features	6

Table 4 : Best parameters

Dataset	Score
Training F1-score	0.978
Test F1-score	0.980

Table 5 : F1-score

The first model is random forest. Looking at the confusion matrix, it can be seen that the normalized values for each label are well classified. Training and test data also showed high f1 score.

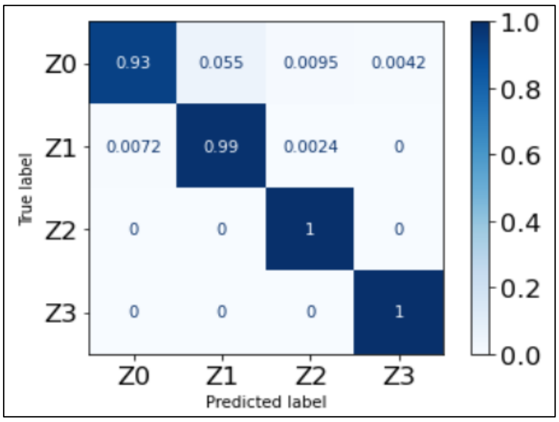
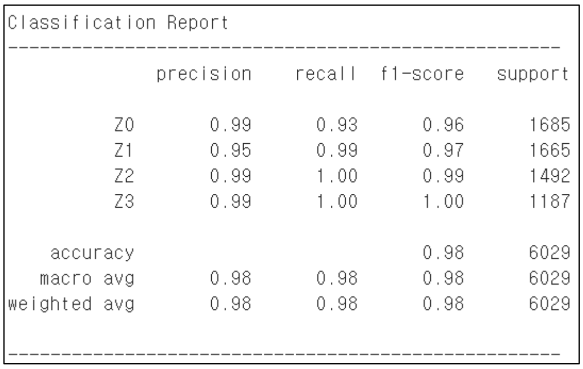


Figure 13 : Confusion matrix



```
Classification Report
----------------------------------------------------
               precision    recall  f1-score   support

          Z0       0.99      0.93      0.96      1685
          Z1       0.95      0.99      0.97      1665
          Z2       0.99      1.00      0.99      1492
          Z3       0.99      1.00      1.00      1187

    accuracy                           0.98      6029
   macro avg       0.98      0.98      0.98      6029
weighted avg       0.98      0.98      0.98      6029

----------------------------------------------------
```

Figure 14 : Random Forest report

4.1.3. Decision Tree

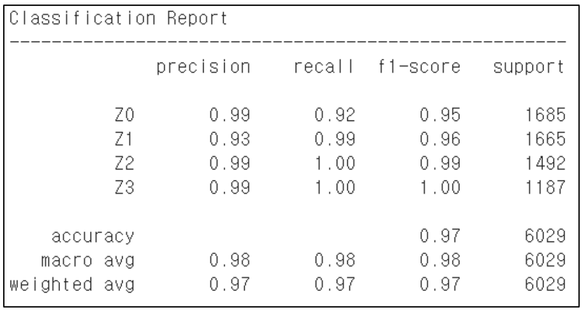
The second model is the decision tree. The best f1 score also shows a high value of 0.97, but it is slightly lower than random forest.

Parameters	Value
Criterion	Entropy
Max_depth	32
Splitter	Best
Max_features	None

Table 6 : Best parameters

Dataset	Score
Training F1-score	0.972
Test F1-score	0.976

Table 7 : F1-score



```
Classification Report
----------------------------------------------------
               precision    recall  f1-score   support

          Z0       0.99      0.92      0.95      1685
          Z1       0.93      0.99      0.96      1665
          Z2       0.99      1.00      0.99      1492
          Z3       0.99      1.00      1.00      1187

    accuracy                           0.97      6029
   macro avg       0.98      0.98      0.98      6029
weighted avg       0.97      0.97      0.97      6029
```

Figure 15 : Decision Tree report

4.1.4. KNN

The third model is KNN. As N_neighbor increases, the f1 score decreases, and 1 is the best model. The F1 score is about 0.93, which is slightly inferior to the previous models.

Parameters	Value
N_neighbors	1
Weights	Uniform
Metric	Manhattan

Table 8 : Best parameters

Dataset	Score
Training F1-score	0.924
Test F1-score	0.933

Table 9 : F1-score

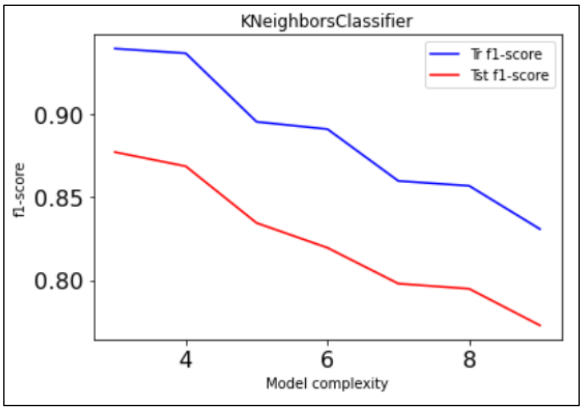


Figure 16 : F1-score according to neighbors

4.1.5. SVC

The fourth model is svc. It shows the best f1 score among the models performed.

Parameters	Value
Kernel	Rbf
C	1
Gamma	0.1

Table 10 : Best parameters

Dataset	Score
Training F1-score	0.991
Test F1-score	0.995

Table 11 : F1-score

4.1.6. Logistic Regression

The last model is logistic regression. The training data did not show good linearity, and because it is a multiple classification, the f1 score did not come out well.

Parameters	Value
Penalty	L1
C	0.616
Solver	Liblinear

Table 12 : Best parameters

Dataset	Score
Training F1-score	0.577
Test F1-score	0.568

Table 13 : F1-score

4.2. Regression

There are a total of 5 models to be used for regression training, and accuracy and RMSE score was used as a performance measure. The model is as follows.

4.2.1. Preprocessing

First is the processing of missing values. HDI _for_year has 19456 null values out of 27820 samples which is approximately 70% of the column. This may tamper the model performance so, dropping the HDI for year column from the dataset. The column country-year is just a combination of country and year columns. So, we drop that column. And drop all the null rows from the dataset. The non-numerical labeled columns, country, year, gender, age_group and generation are to be converted to numerical labels that can be done by using SkLearn's LabelEncoder. Standardization of a dataset is a common requirement for many machine learning estimators: they might behave badly if the individual features do not more or less look like standard normally distributed data. So, the numerical columns, population, gdp_for_year & gdp_per_capita are being standardized using SkLearn's RobustScalar. Finally, the data is split into train & test sets, 80-20 split.

4.2.2. KNN

First of all, it is the model of KNN. A simple implementation of KNN regression is to calculate the average of the numerical target of the k nearest neighbors. The gridsearchCV was used to find the optimal parameter value. Evaluating training and testing set performance with different numbers of neighbors from 1 to 30. Accuracy for training data was 1.000, and accuracy for test data was 0.803. The plot shows the training and test set accuracy on the y-axis against the setting of n_neighbors on the x-axis. This discrepancy between performance on the training set and the testing set for n_neighbors < 5 is a clear sign of overfitting. After that, the performance is not so great so, moving on to the other models.

Measure	Train	Test
Accuracy	1.000	0.803
RMSE	0.0	0.528

Table 14 : KNN Performance Measure

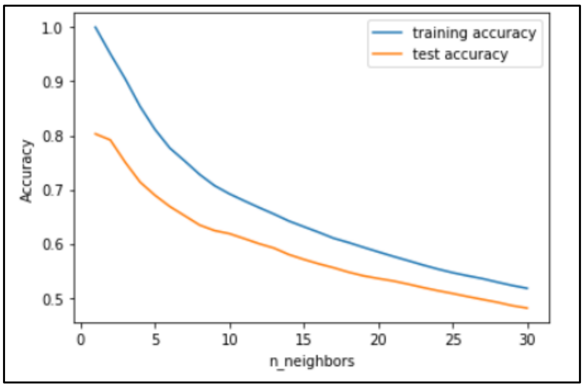


Figure 17 : KNN Performance Plot

4.2.3. Linear Regression

Linear regression finds the parameters w and b that minimize the mean squared error between predictions and the true regression targets, y, on the training set. The model performance is not very good, but we can see that the scores on the training and test sets are very close together. This means we are likely underfitting, not overfitting.

Measure	Train	Test
Accuracy	0.289	0.291
RMSE	1.021	1.002

Table 15 : Linear Regression Performance Measure

4.2.4. Decision Tree

The following is the decision tree regression model. Evaluating training and testing set performance with different numbers of max_depth from 1 to 30. The following are the performance results of the decision tree. Accuracy for training data was 0.969, and accuracy for test data was 0.952. The plot shows the training and test set accuracy on the y-axis against the setting of max_depth on the x-axis. The model performance is gradually increased on increasing the max_depth parameter. But after max_depth = 9, the model overfits. So, the model is considered with max_depth = 9 which has an accuracy of 95.2%.

Measure	Train	Test
Accuracy	0.969	0.952
RMSE	0.213	0.261

Table 16 : Decision Performance Measure

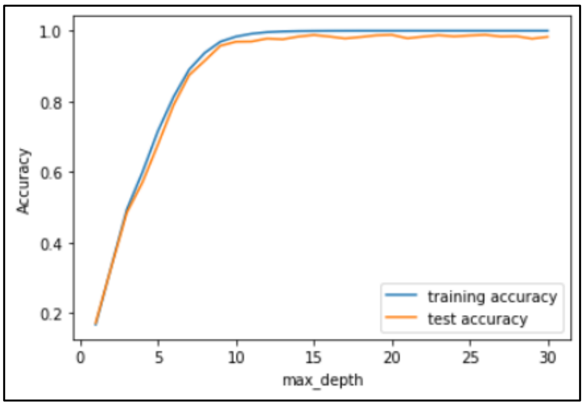


Figure 18 : Decision Performance Plot

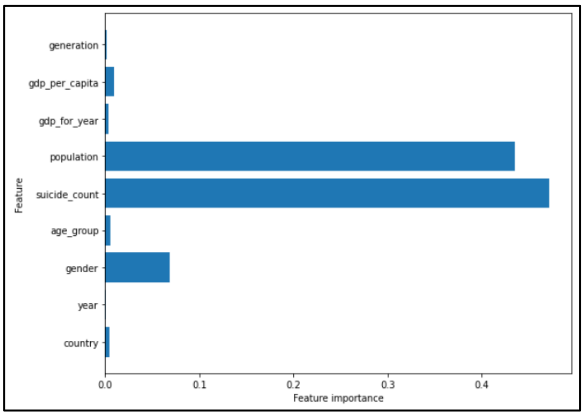


Figure 19 : Decision Feature Importance

4.2.5. Random Forest

Random forest regression model insert, train, and predict data in the same way. max_depth is set to 9. Accuracy for training data was 0.987, and accuracy for test data was 0.980.

The plot shows the training and test set accuracy on the y-axis against the setting of max_depth on the x-axis. The random forest gives us an accuracy of 98%, better than the linear models or a single decision tree, without tuning any parameters. But this might also be a case of overfitting. So, the parameter are tuned and the finalized model has an accuracy of 98% which is better than the linear & decision tree models.

Measure	Train	Test
Accuracy	0.987	0.980
RMSE	0.139	0.170

Table 17 : Random Forest Performance Measure

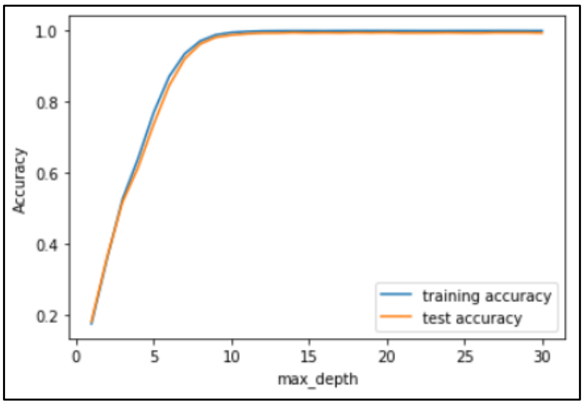


Figure 20 : Random Forest Performance Plot

4.2.6. MLPs

The last one is multi-layer perceptron. MLPs can be viewed as generalizations of linear models that perform multiple stages of processing to come to a decision. The hidden layer size is set to [100,100]. Accuracy for training data was 0.934, and accuracy for test data was 0.925. Hyperparameter tuning is performed for the model. The tuned parameters are number of hidden layers and the hidden_units of each layer with default values of alpha. The optimized Gradient Boosted model gives us an accuracy of 93.4%, with parameter tuning.

Measure	Train	Test
Accuracy	0.934	0.925
RMSE	0.310	0.326

Table 18 : MLPs Performance Measure

5. Korea Analysis

As mentioned in section 1, we analyze the number of suicides by age and gender to get more meaningful insights about the Korean suicide data.

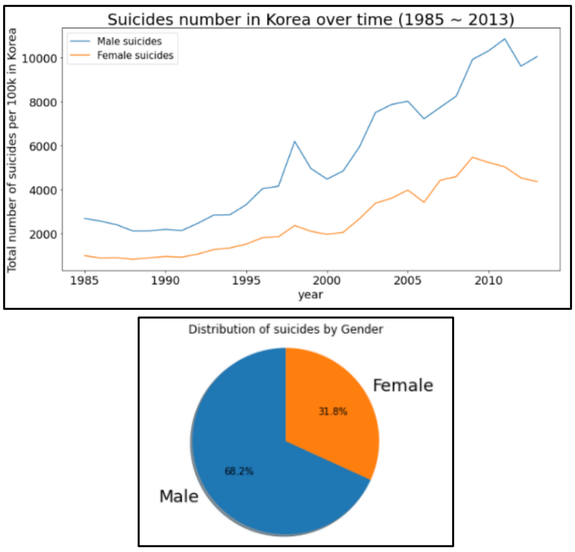


Fig 21 : The graph and pie chart of the number of suicides by gender

First, this is an analysis of suicide rates by gender. As you can see from the graph and pie chart, the number of suicides of male is more than twice as many as those of female. However, compared to the male to female ratio of about 5:1 in the world data, the female ratio can be said to be very high.

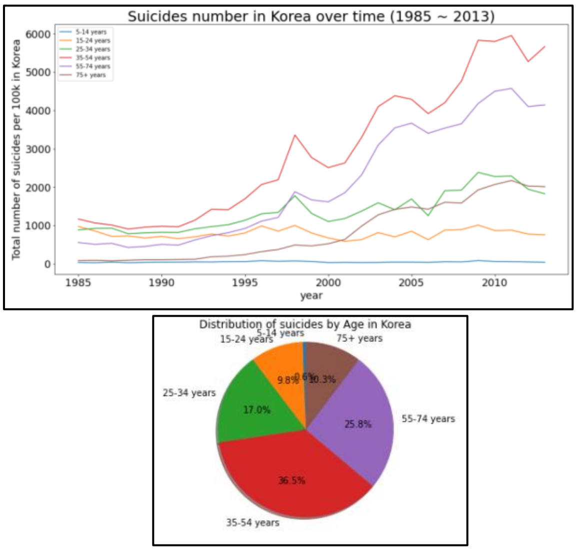


Fig 22 : The graph and pie chart of the number of suicides by age

The above graph and chart is an analysis of suicide rates by age. The age group with the highest number of suicides is the 35-54 group, and the second is the 55-74 group. An additional point to note is the suicide trend in the 75+ group. If you look at the graph, you can see that the number of suicides was very small in the past, but in recent years it has risen rapidly.

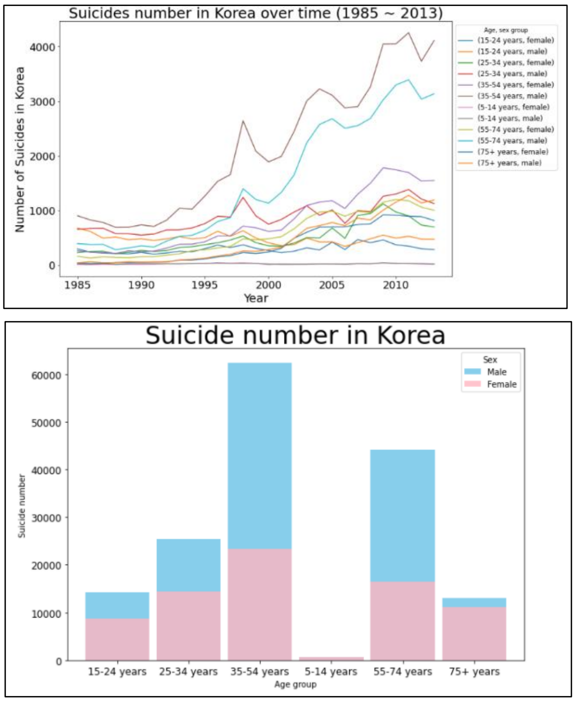


Fig 23 : The line and bar graph of the number of suicides by gender and age

Finally, this is the analysis of suicide rates by gender & age in Korea. If you look at the graph, you can see that the number of suicides continues to increase over time for all groups equally. Also, looking at the bar graph, as mentioned above, not only the proportion of female overall is quite high, but in the group excluding the 35-54 group and the 55-74 group, the proportion of female is even greater than that of male.

6. Experiments and Results

6.1. Classification Results

The classification model results are as follows. The model with the highest f1 score is svc. In terms of feature importance, population, gdp, and HDI are very high, followed by sex and age. In the case of age, it shows a particularly high value for those over 75 years old.

Model	Train F1-score	Test F1-score
Random Forest	0.978	0.980
Decision Tree	0.972	0.976
KNN	0.924	0.933
SVC	0.991	0.995
LinearRegression	0.577	0.568

Table 19 : Model F1-score comparison

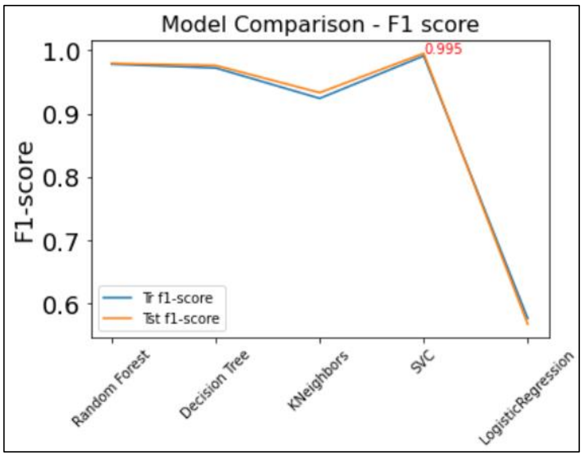


Figure 24 : Model comparison

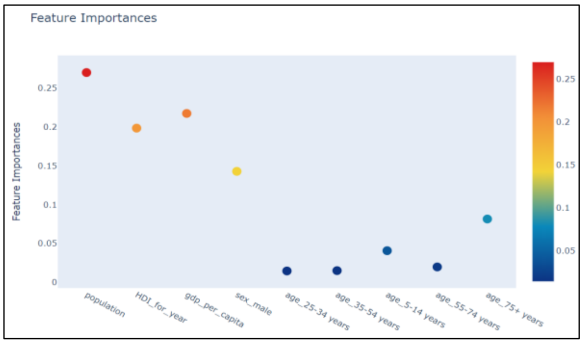


Figure 25 : Random Forest Classification Feature Importance

6.2. Regression Results

Among all the trained models, Random Forest performance is better. This is the result of using the decision tree as an ensemble technique and this model is very good in execution Speed & model performance. In terms of feature importance, population is very high, followed by gender, gdp_per_capita, age, gdp_for_year.

Model	Train Accuracy	Test Accuracy	Train RMSE	Test RMSE
Random Forest	**0.987**	**0.980**	**0.139**	**0.170**
Decision Tree	0.969	0.952	0.213	0.262
MLPs	0.934	0.925	0.310	0.327
KNN	1.000	0.803	0.000	0.529
Linear Regression	0.289	0.291	1.022	1.002

Table 20 : Model Accuracy & RMSE score Comparison

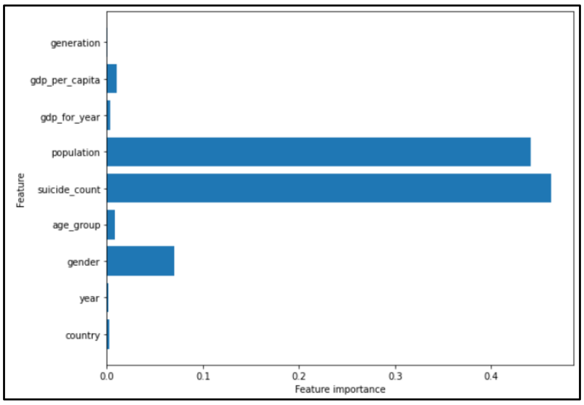


Figure 26 : Random Forest Regression Feature Importance

7. Conclusion

7.1. Insight

In classification and regression, the best models are SVC and random forest, respectively, and the important features obtained through classification are population, gdp and HDI.

Insight for Korea data can be organized into 4 categories as shown below.

- Suicide number is increasing constantly.
- Female's ratio is higher than world.
- Recently, 75+ group's ratio is increasing rapidly.
- 15-24 group's suicide is increasing less than other groups.

7.2. Suggestion

The following suggestions are made to prevent suicide in Korea, which is the world's No. 1 suicide rate.

- **Child care service -** As the suicide rate for women in their 30s and 50s is increasing, the 'child care service' relieves the burden of care focused.
- **Social network support -** As the suicide rate for women over 75 is considered to be higher than the global trend, support for social network support programs to alleviate social isolation.
- **Employment support system** – It is possible to infer that the unemployment problem in Korea causes suicide from the fact that the suicide rate of men aged 35-74 is quite high.

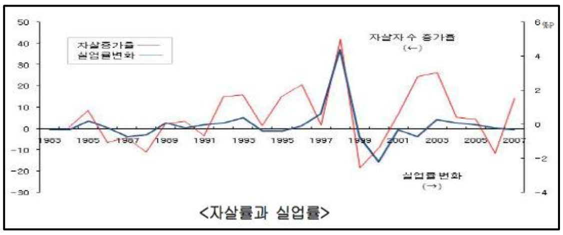


Figure 27 : Suicide rate and unemployment rate (2013.05. National Disaster Safety Research Institute)

- **Suicide prevention education -** Suicide prevention education is conducted mainly by companies and schools through the annual "Suicide Prevention Paper" issued by the Ministry of Health and Welfare.

8. Future Work

We could learn the seriousness of suicide not only in Korea but also around the world. While analyzing data, we realized again that suicide prevention is very necessary every time we extract insight. In particular, it was a good experience to analyze global data and Korean data through this project, and to think about both macro and micro aspects.

Also, through this opportunity, we realized how important the process of data preprocessing and visualization is before machine learning. Again, we felt the importance of filling the evidence based on data. When we removed the column, we thought was important before the analysis, we got better scores. So next time, we will put more effort and time into these processes.

Next time if we get the opportunity to analyze the same subject, we would like to use more features and use various models to conduct more in-depth analysis.

9. References

[1] Durkheim, E. 1897. La Sucide (by Hee-seop Lee. 1994. [Suicide Theory/Social Division Theory] Samsung Publishing Co., Ltd.)

[2] Dong-joon Shin. 2012. A Cross-National Analysis on Social Causes of Suicide -The Effects of Social Integration, Economic Inequality, and Economic Success Culture.

[3] VoA.'Suicide Is Not Just a US Problem, It's a Global Issue'. 2018. https://www.voanews.com/science-health/suicide-not-just-us-problem-its-global-issue

[4] Business Group on Health. "Suicide: An Increasing Concern for Global Employers". 2020. https://www.businessgrouphealth.org/en/resources/suicide-an-increasing-concern-for-global-employers

[5] Current Affairs Journal. "Embarrassed suicide rate number one in OECD"... An average of 38 lives a day. 2020. https://www.sisajournal.com/news/articleView.html?idxno=205522

[6] Hankyoreh. "Turning suicide rates last year...Back to number one in the OECD". 2019. http://www.hani.co.kr/arti/economy/economy_general/910665.html

[7] Young Doctor. "An increasing number of elderly people are taking their own lives because of loneliness.". 2019. http://www.docdocdoc.co.kr/news/articleView.html?idxno=1069704

[8] Hankyoreh. "The first measure of 'quiet slaughter' of women in their 20s has been announced". 2020. http://www.hani.co.kr/arti/society/women/972163.html

[9] Hankyoreh. "The pain of women in their 20s should start with social". 2020. http://www.hani.co.kr/arti/society/society_general/972605.html

[10] Joong-Ang. "The budget is one-tenth as much as those killed by suicide and those killed in traffic accidents". 2020. https://news.joins.com/article/23930870

[11] Ministry of Health and Welfare. "2020 Manual on Suicide Prevention". 2020. https://www.mdon.co.kr/news/article.html?no=27665