

[ B N C S 4 1 1 ] T e r m P r o j e c t

# Insight about Suicide of Worldwide and Korea

(Classification according to suicide rate & Predict a suicide rate)

Kim Dong Uk

Nam Jeong Jae

Park Seong Ho

Suicide Rescue Team! SRT!

# CONTENTS

01

Team

02

Introduction

03

Dataset

04

Classification

05

Regression

06

Experiments  
& Results

07

Insights  
for Korea

08

Conclusion

09

Feedback

10

Reference

01

---

Team

01

# “ Team ”

{ Suicide Rescue Team ! SRT ! }



Kim Dong Uk

[Korea Univ.]

Physics

tigerrhs@naver.com



Nam Jeong Jae

[Korea Univ.]

Applied Statistics

njj97@naver.com



Park Seong Ho

[Korea Univ.]

Industrial Management Engineering

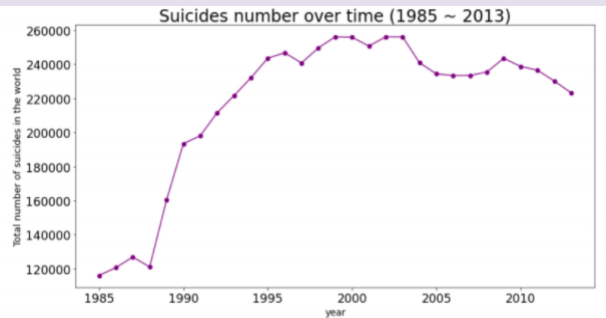
rkwlgh@naver.com

02

---

# Introduction

# “ Why this subject? ”



Science & Health

## Suicide Is Not Just a US Problem, It's a Global Issue

By Carol Pearson  
June 15, 2018 09:56 AM

### Suicide: An Increasing Concern for Global Employers

Approximately 800,000 people commit suicide each year, and many more make suicide attempts. In fact, suicide is the 15th leading cause of death globally for all age groups, the second for young people aged 15-29 and the fifth for those aged 30-49.

Suicide is not a national problem,  
but a global problem.



Feature extraction between  
High vs Low suicide rate countries  
->We can know how prevent suicide



So, subject :

"Insight about Suicide of  
Worldwide and Korea"

## “ Flow (Purpose of Analysis) ”

### Visualization

&

### Data preprocessing

Identification of data form and preparing for classification and regression

### Classification

1. RandomForest
2. Decision Tree
3. KNN
4. SVC
5. Logistic Regression

Classification for many countries into 4 classes based on suicide rate

### Regression

1. KNN
2. Linear Regression
3. Decision Tree
4. Random Forest
5. MLP

Prediction of suicide rate and selecting the best regression model

### Insights

for

### Korea

Visualization and draw insights about Korea suicide rate

03

---

Dataset



## 03

“

## Dataset

”

## # Content

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.

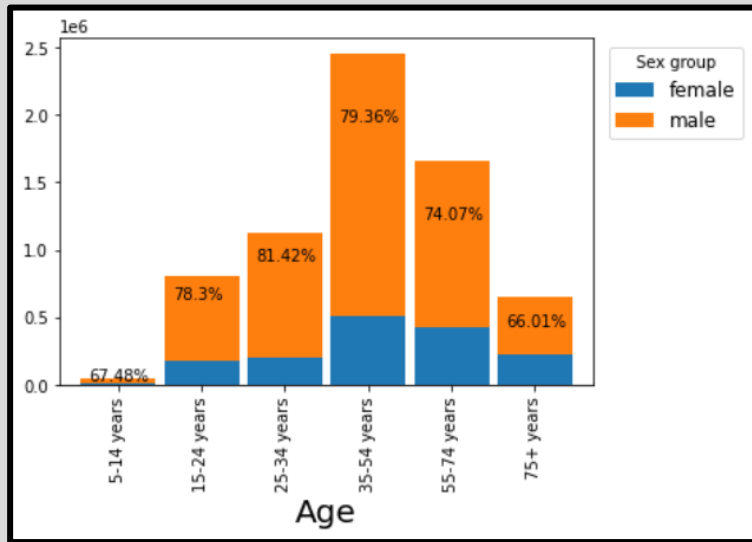
## # Shape

( 27820, 12 )

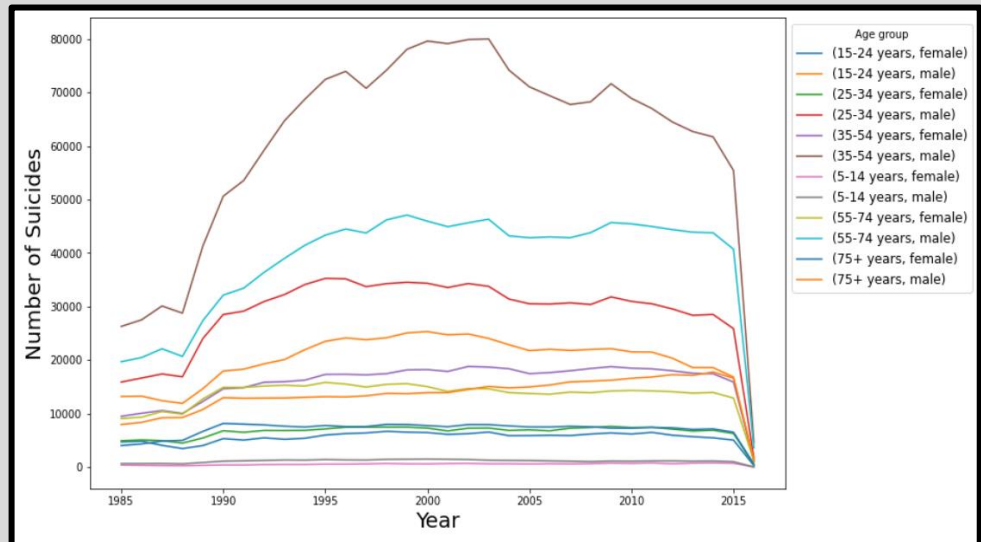
Column	Description
country (string)	Country name
year (numeric)	1985-2016
sex (string)	Male or Female
Age (string)	5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, 75+ years
suicides_no (numeric)	Number of suicides
population (numeric)	Number of population
suicides/100k pop (numeric)	Suicide 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

## 03

## “ Visualization ”



Bar graph of suicide rate by age and gender



Line graph of suicide rate by year, age and gender

- ✓ The male ratio is higher in all age groups, accounting for 70-80%.
- ✓ Both men and women have the highest suicide rate in their 30s and 40s.
- ✓ The older, the higher the female percentage.

04

---

# Classification

## 04

## “ Preprocessing ”

## # NA

'HDL\_for\_year' NA value  $\rightarrow$  0

## # One Hot Encoding

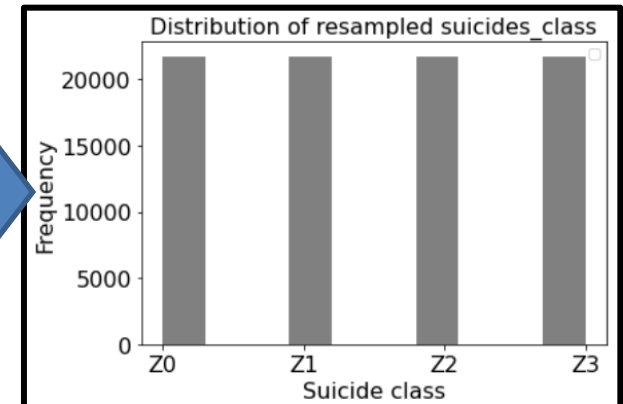
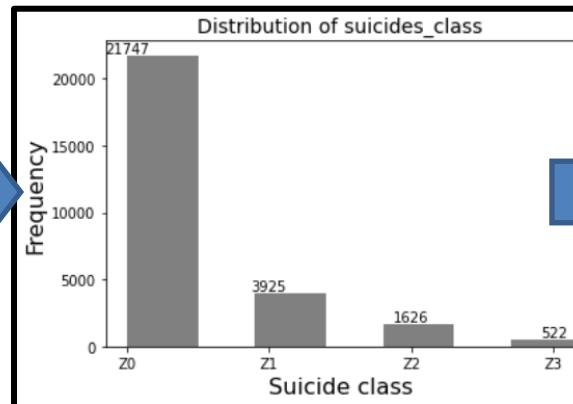
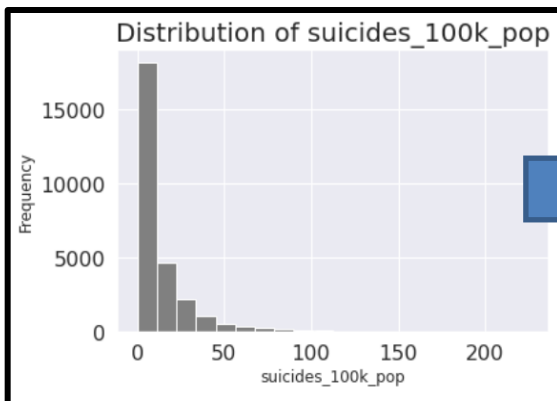
sex, age, year columns

## # Create 4 Labels according to Suicide Rate

Z0 = 0~19 / Z1 = 20~39 / Z2 = 40~79 / Z3 = 80~

## # Feature Selection : Drop Columns

- Country, country\_year : ID value
- Generation : can be estimated by age and year
- Suicide\_100k\_pop = suicides\_no/population
- Gdp\_for\_year : only certain countries have high gdp, and the gap is large



Resampled labels

04

# “ Model Building & Training ”

1

**Random Forest (Ensemble of Decision Tree)**

2

**Decision Tree**

3

**K - Nearest Neighbors Classification (KNN)**

4

**Support Vector Classification**

5

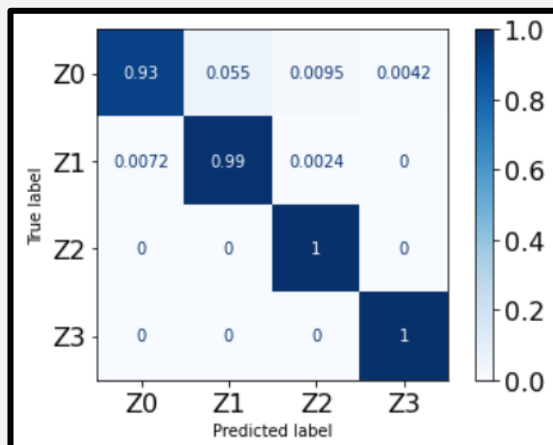
**Logistic Regression**

# “ Random Forest ”

## # Best model

Parameters	Best value
Criterion	Entropy
Max_depth	None
N_estimators	32
Max_features	6

## # Confusion Matrix



## # Classification Report

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

## # Best score

Dataset	F1-score
Train	0.978
Test	0.980

## 04

## “ Decision Tree ”

## # Best model

Parameters	Best value
Criterion	Entropy
Max_depth	32
Splitter	Best
Max_features	None

## # Best score

Dataset	F1-score
Train	0.972
Test	0.976

## # Classification Report

## 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

## 04

“

## KNN

”

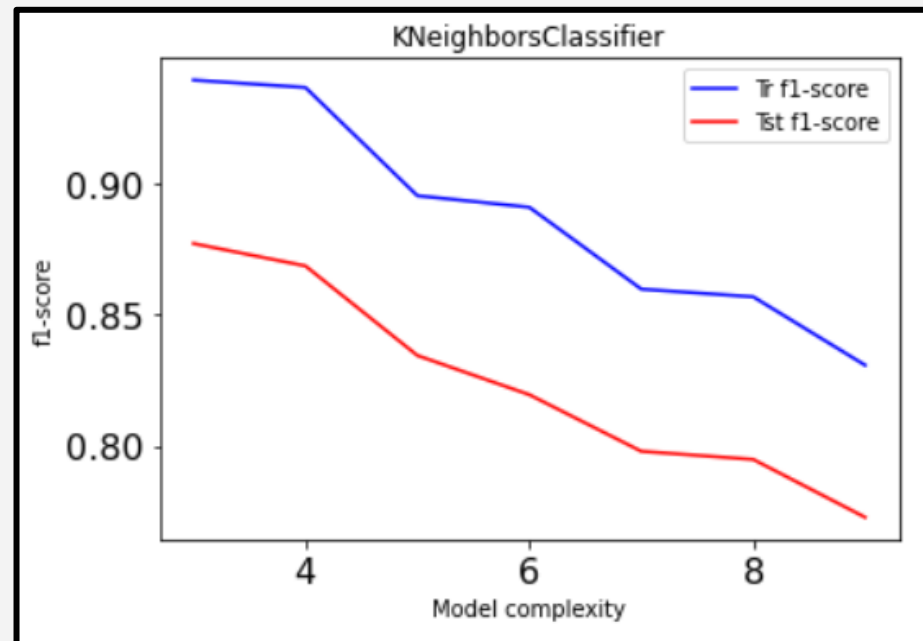
## # Best model

Parameters	Best value
N_neighbors	1
Weights	Uniform
Metric	manhattan

## # Best score

Dataset	F1-score
Train	0.924
Test	0.933

## # F1-score with increasing 'n\_neighbors'





## “ SVC & Logistic Regression ”

### # SVC Best model


Parameters	Best value
Kernel	rbf
C	1
Gamma	0.1

### # LR Best model

Parameters	Best value
Penalty	L1
C	0.616
Solver	liblinear

### # SVC Best score

Dataset	F1-score
Train	0.991
Test	0.995

 Highest score

### # LR Best score

Dataset	F1-score
Train	0.577
Test	0.568

05

---

# Regression

# “ Preprocessing ”

## # Drop

```
# dropping the HDI for year column

data = data.drop(['HDI for year'], axis = 1)
data.shape

(27820, 11)
```

```
#dropping the country-year for year column

data = data.drop(['country-year'], axis = 1)
data.shape

(27820, 10)
```

```
# droppinf off any null rows (is any)

data = data.dropna()
data.shape

(27820, 10)
```

## # Convert type

```
# Converting the column 'gdp_for_year' to float from object

data['gdp_for_year'] = data['gdp_for_year'].str.replace(',', '').astype(float)
```

## # LableEncoder

```
# encoding the categorical features with LabelEncoder

from sklearn.preprocessing import LabelEncoder
categorical = ['country', 'year', 'age_group', 'gender', 'generation']
le = sklearn.preprocessing.LabelEncoder()

for column in categorical:
    data[column] = le.fit_transform(data[column])
```

## # RobustScaler

```
# Scaling the numerical data columns with RobustScaler

numerical = ['suicide_count', 'population', 'suicide_rate',
             'gdp_for_year', 'gdp_per_capita']

from sklearn.preprocessing import RobustScaler

rc = RobustScaler()
data[numerical] = rc.fit_transform(data[numerical])
```

# “ Model Building & Training ”

1

**K - Nearest Neighbors Regression (KNN)**

2

**Linear Regression**

3

**Decision Tree**

4

**Random Forest (Ensemble of Decision Tree)**

5

**Multilayer Perceptron (MLP)**

# “ KNN & Linear Regression ”

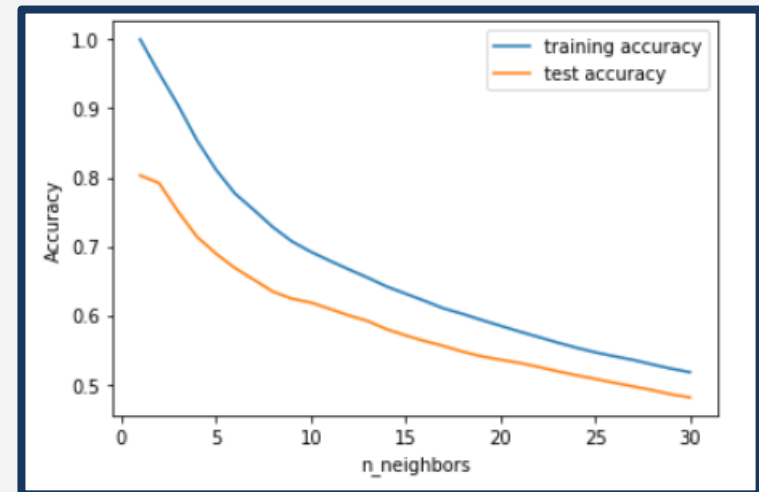
## # KNN

KNN: Accuracy on training Data: 1.000

KNN: Accuracy on test Data: 0.803

KNN: The RMSE of the training set is: 0.0

KNN: The RMSE of the testing set is: 0.528625104062679



## # Linear Regression

Linear Regression: Accuracy on training Data: 0.289

Linear Regression: Accuracy on test Data: 0.291

Linear Regression: The RMSE of the training set is: 1.021551293216169

Linear Regression: The RMSE of the testing set is: 1.0023713268564818

## # KNN Plot

Training & Testing accuracy  
for n\_neighbors from 1 to 30

# “ Decision Tree ”

## # Decision Tree

Decision Tree: Accuracy on training Data: 0.969

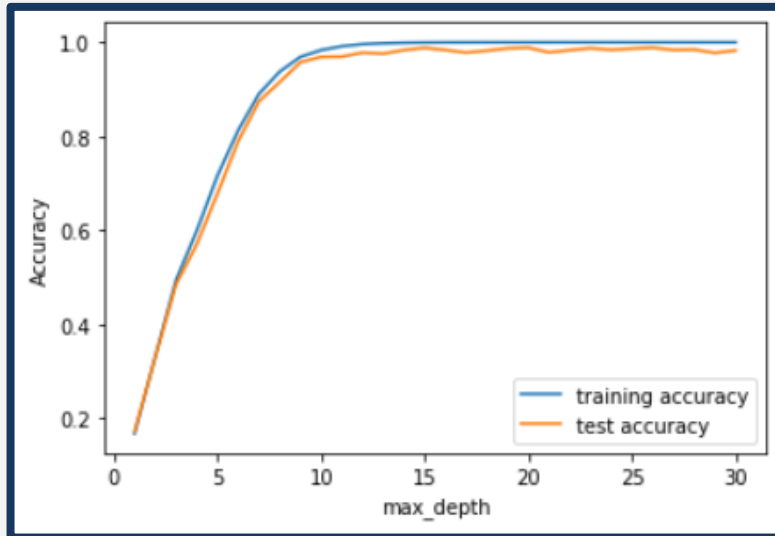
Decision Tree: Accuracy on test Data: 0.952

Decision Tree: The RMSE of the training set is: 0.21323575750994775

Decision Tree: The RMSE of the testing set is: 0.26179817997957533

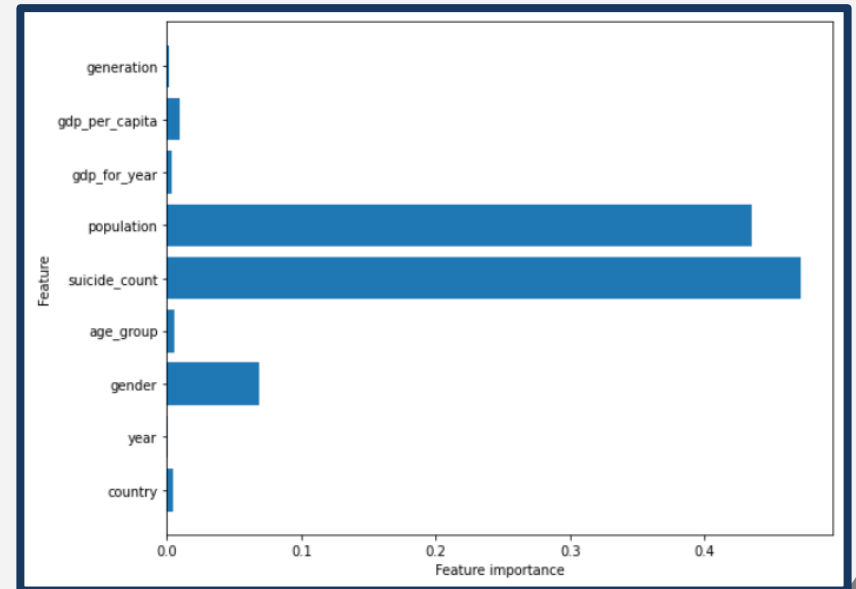
## # Decision Tree Plot

Training & Testing accuracy for max\_depth from 1 to 30



## # Decision Tree Plot

Graph of importance for each feature  
{ X-axis : Feature Importance, Y-axis : Feature }



# “ Random Forest & MLPs ”

## # Random Forest

Random Forest: Accuracy on training Data: 0.987

Random Forest: Accuracy on test Data: 0.980

Random Forest: The RMSE of the training set is: 0.13924040621142714

Random Forest: The RMSE of the testing set is: 0.17042857286372656

## # MLPs

Multilayer Perceptron Regression: Accuracy on training Data: 0.934

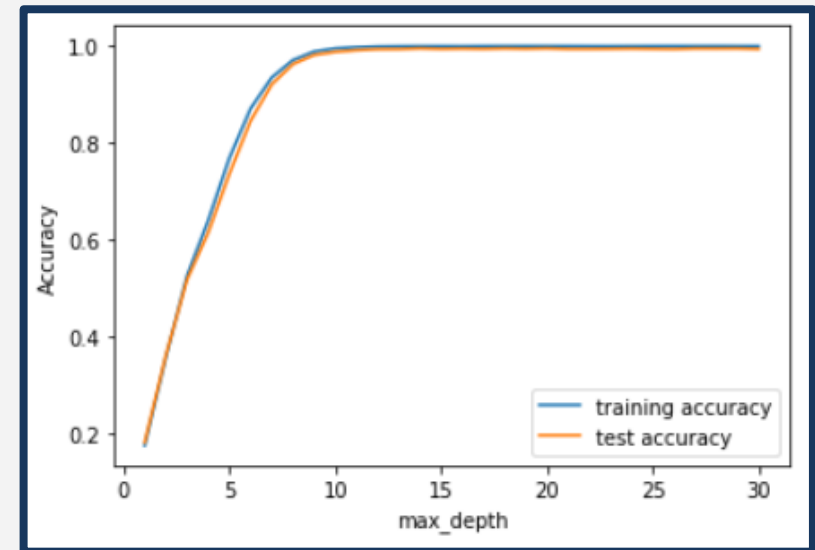
Multilayer Perceptron Regression: Accuracy on test Data: 0.925

Multilayer Perceptron Regression: The RMSE of the training set is: 0.3102405366573578

Multilayer Perceptron Regression: The RMSE of the testing set is: 0.32667890890643425

## # Random Forest Plot

Training & Testing accuracy for max\_depth from 1 to 30



06

---

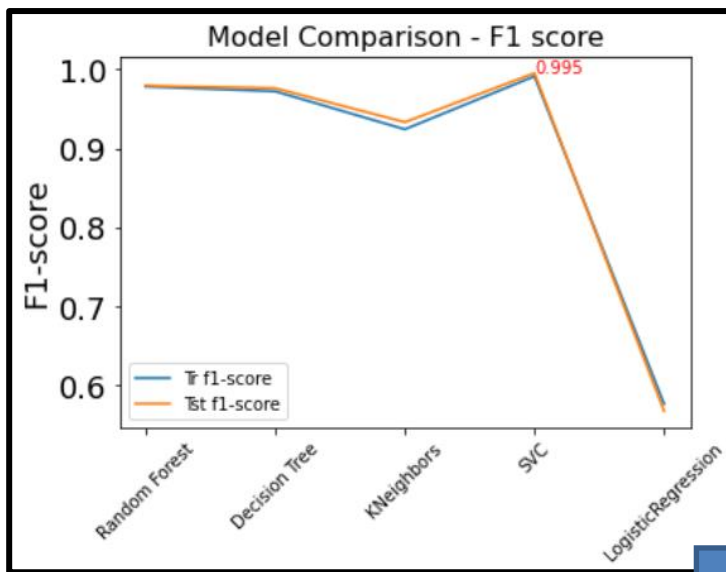
Results



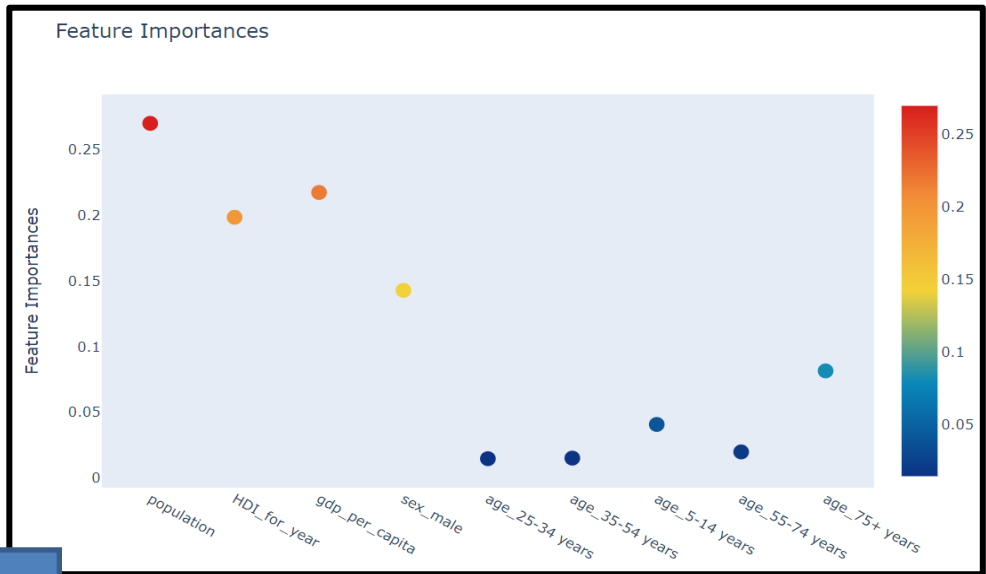
## 06

## “ Classification Results ”

## # Results Table



## # Feature Importances



- ✓ The best model is SVC and F1-score is 0.995
- ✓ Feature importances are in the order of population, gdp, HDI, sex, and age.
- ✓ It is much higher in the age group, especially when over 75

## 06

# “ Regression Results ”

## # Results Table

	ML Model	Train Accuracy	Test Accuracy	Train RMSE	Test RMSE
0	k-Nearest Neighbors Regression	1.000	0.803	0.000	0.529
1	Linear Regression	0.289	0.291	1.022	1.002
2	Decision Tree	0.969	0.952	0.213	0.262
3	Random Forest	0.987	0.980	0.139	0.170
4	Multilayer Perceptron Regression	0.934	0.925	0.310	0.327

## # Sorting the Data Frame on Accuracy



	ML Model	Train Accuracy	Test Accuracy	Train RMSE	Test RMSE
3	Random Forest	0.987	0.980	0.139	0.170
2	Decision Tree	0.969	0.952	0.213	0.262
4	Multilayer Perceptron Regression	0.934	0.925	0.310	0.327
0	k-Nearest Neighbors Regression	1.000	0.803	0.000	0.529
1	Linear Regression	0.289	0.291	1.022	1.002

07

---

**Insights  
for Korea**

# “The severity of suicide rates in Korea”

## From news

### 부끄러운 '자살률 OECD 1위'...하루 평균 38명 목숨 끊어

ⓒ 서지민 객원기자 (sisajournal.com) | ⓒ 승인 2020.09.22 14:32



작년 자살률 10만 명당 26.9명으로 OECD 중 1위  
연령 높을수록 자살률도 높아져...40대부터 30명 이상 70대 이상은 45명 이상

### 지난해 자살률 증가세 전환...다시 OECD 1위로

통계청 '2018년 사망원인통계' 발표

등록 : 2019-09-24 11:59 수정 : 2019-09-24 16:34

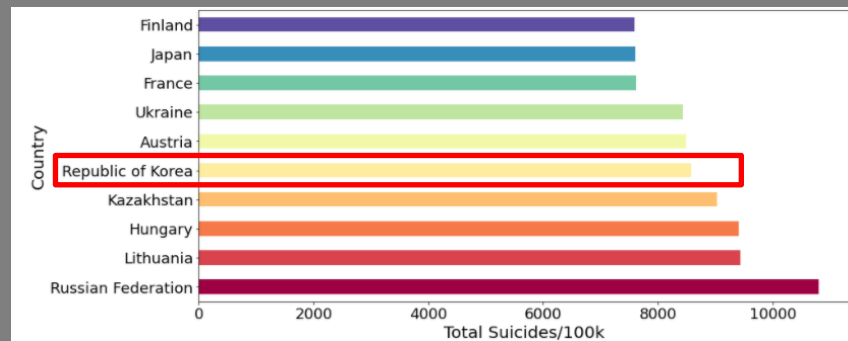
인구 10만명당 자살자 24.3명→26.6명 증가  
10대·30대·40대 자살률 증가 폭 특히 커  
10~30대 사망원인 1위 자살...40~50대는 2위  
OECD 국가 중 자살률 2위→1위 복귀  
“스스로 목숨 끊은 유명인 많아 베르테르 효과”

### [창간특집]외로움'에 스스로 목숨 끊는 노인들이 늘어난다

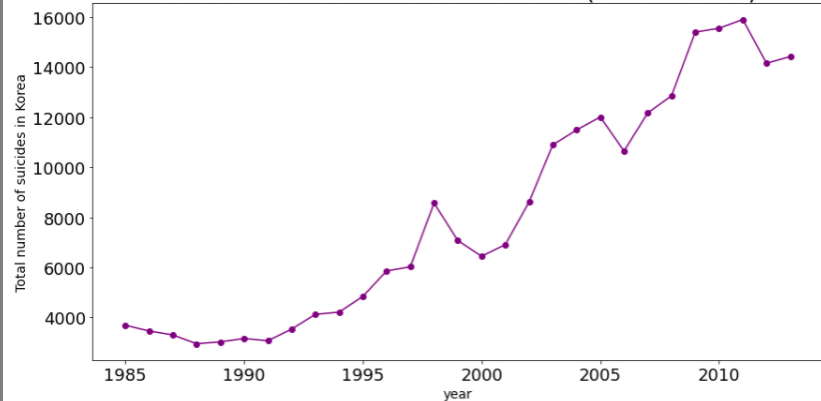
A 김은영 기자 | ⓒ 일력 2019.06.27 06:00 | ⓒ 수정 2019.07.01 07:03 | ● 댓글 0

2019년 자살예방백서, 65세 이상 노인 자살률 OECD 국가 중 1위...평균 보다 3배 높아  
백종우 중앙자살예방센터장 “외로움 달래는 사회공동체 복원 및 부처간 협의 필요”

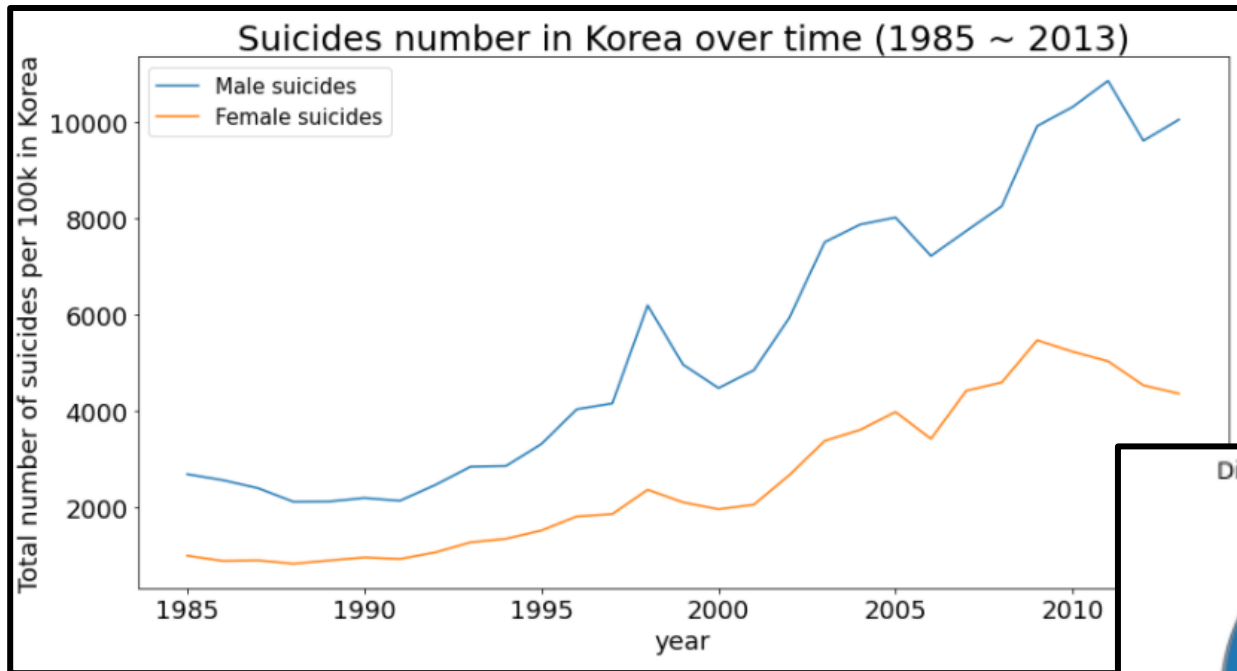
## From data



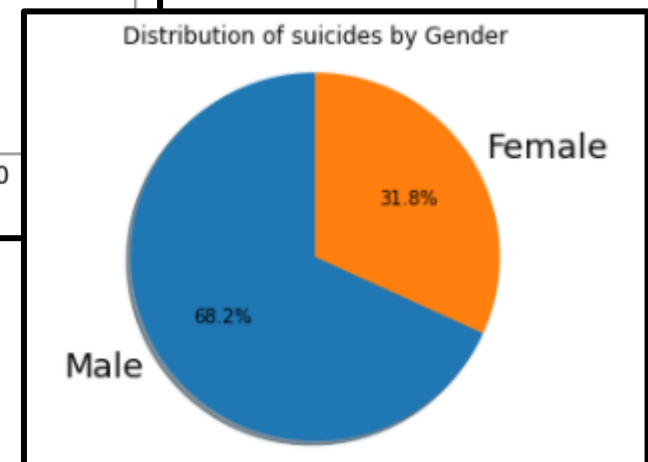
Suicides number in Korea over time (1985 ~ 2013)



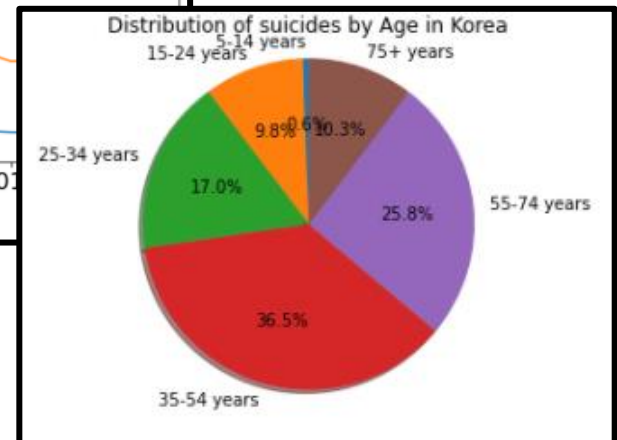
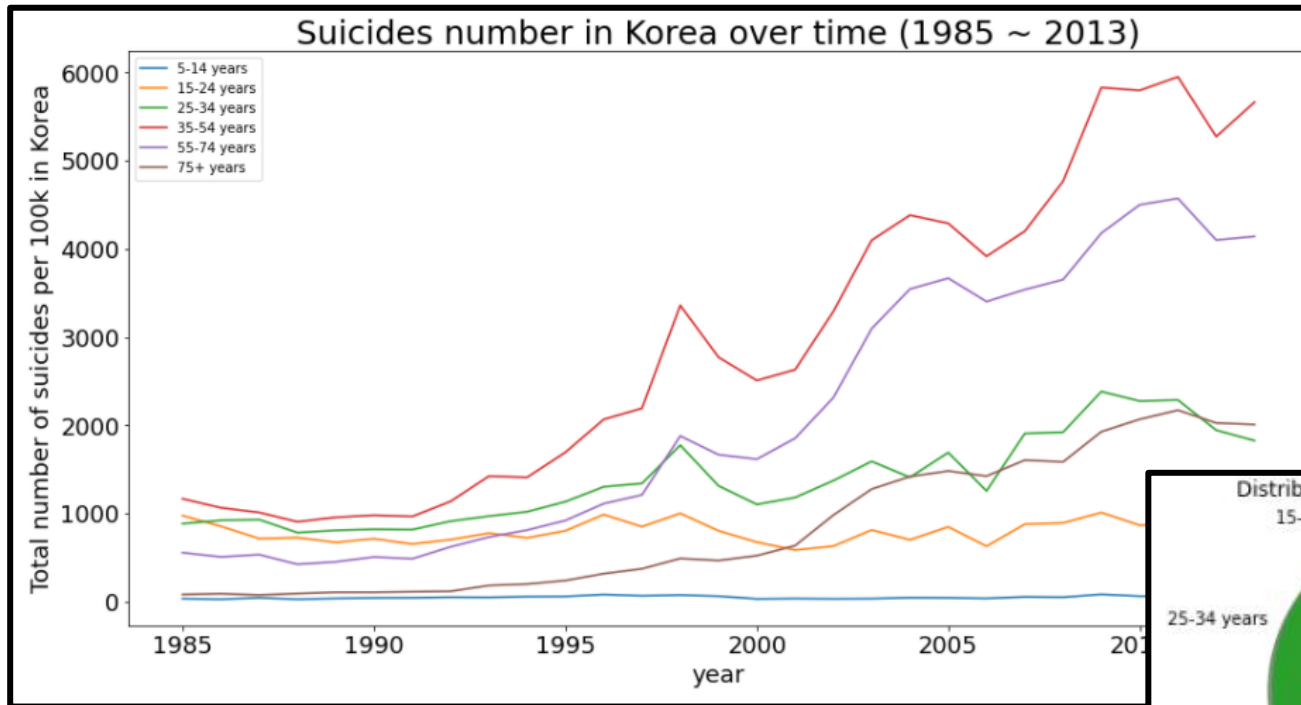
## “Suicide rates by gender in Korea”



1. **Male > Female**
2. **Female ↑** [comparing with worldwide data]

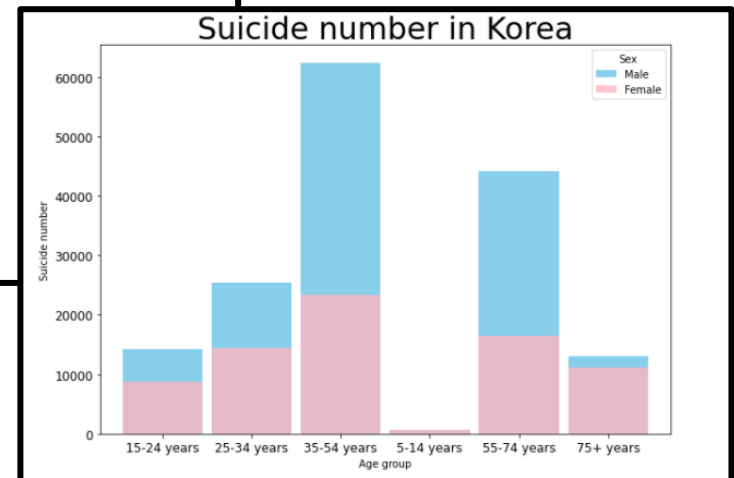
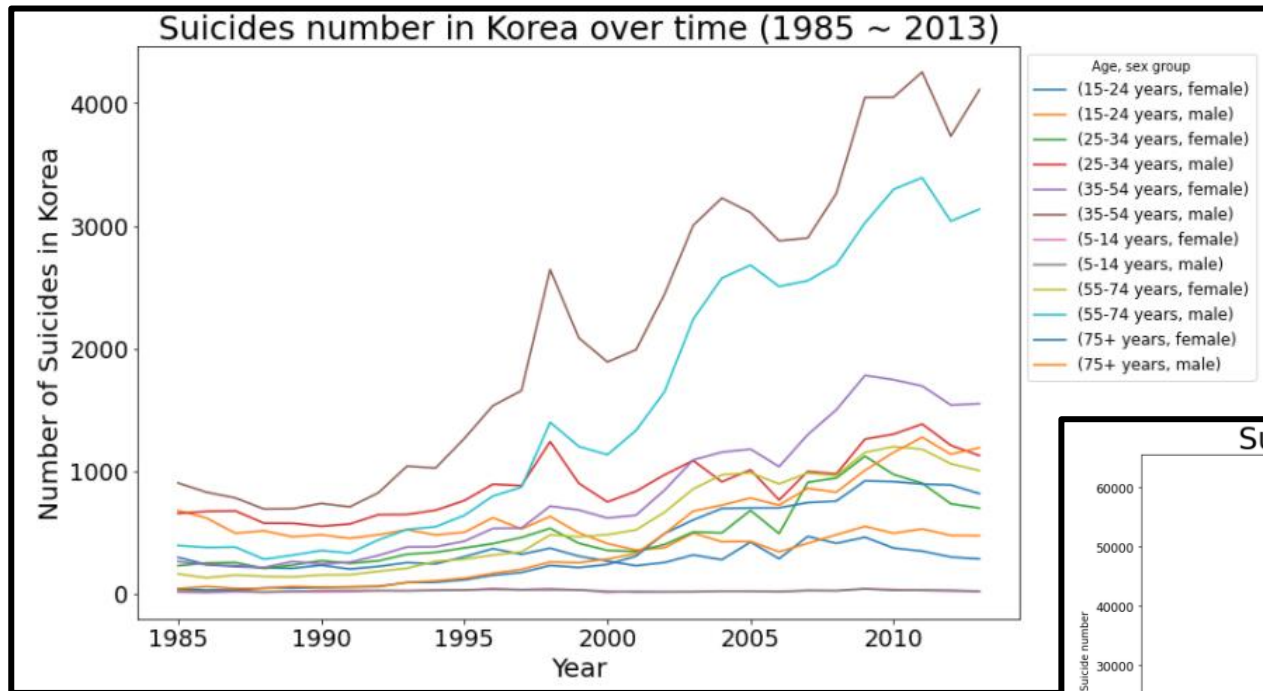


# “Suicide rates by age in Korea”



1. **1<sup>st</sup> : 35-54, 2<sup>nd</sup> : 55-74**
2. **75+ ↑ [comparing with worldwide data]**

## “Suicide rates by gender & age in Korea”



1. Suicide number constantly increase in all group
2. Female > Male in 4 age group (5-14, 15-24, 25-34, 75+)

08

---

# Conclusion



## “ Conclusion (insight) ”

### Classification & Regression

- ▷ Classification – Best model : SVC (f1-score = 99.5 %)  
Feature importance: Population, HDI, GDP
- ▷ Regression – Best model : Random Forest (RMSE = 0.170)

### Insight for Korea (vs worldwide data)

1. Suicide number is increasing constantly
2. Female's ratio is high
3. Recently, 75+ group's ratio is increasing rapidly
4. 15-24 group's suicide is increasing less than other groups.

## “Conclusion (suggestion)”

### Suggestion for Korea

1. **‘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.
2. **‘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.
3. **‘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.
4. **‘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.

09

---

**Feedback**

# “ Feedback ”

## Kim dong Uk

1. Through this opportunity, I realized how important the process of data preprocessing and visualization is before machine learning.
2. Suicide issue is a global issue, and especially when looking at the seriousness of suicide in Korea, I once again felt that suicide prevention was urgent.
3. I was able to develop the ability to express and coordinate opinions on team projects.
4. Next time if I get the opportunity to analyze the same subject, I would like to use more features and use various models to conduct more in-depth analysis.

## Nam Jeong Jae

1. We could learn the seriousness of suicide not only in Korea but also around the world.
2. While analyzing data, I realized again that suicide prevention is very necessary every time I extract insight.
3. I felt that my skills increased as I directly handled data and modelled it while doing the project.
4. In particular, I felt a great sense of accomplishment when I extracted Insight and learned unexpected facts and fresh information.

## Park Seong Ho

1. Again, I felt the importance of filling the evidence based on data. When I removed the column I thought was important before the analysis, I got better scores.
2. It was a great experience to analyze global and Korean data and to think about it in both macro and micro terms.

10

---

Reference

## “ Reference ”

- <https://www.voanews.com/science-health/suicide-not-just-us-problem-its-global-issue>
- <https://www.businessgrouphealth.org/en/resources/suicide-an-increasing-concern-for-global-employers>
- <https://www.sisajournal.com/news/articleView.html?idxno=205522>
- [http://www.hani.co.kr/arti/economy/economy\\_general/910665.html](http://www.hani.co.kr/arti/economy/economy_general/910665.html)
- <http://www.docdocdoc.co.kr/news/articleView.html?idxno=1069704>
- <http://www.hani.co.kr/arti/society/women/972163.html>
- [http://www.hani.co.kr/arti/society/society\\_general/972605.html](http://www.hani.co.kr/arti/society/society_general/972605.html)
- <https://news.join.com/article/23930870>
- <https://www.mdon.co.kr/news/article.html?no=27665>

Suicide Rescue Team! SRT!

**THANK  
YOU**