Tampere University of Applied Sciences

Suicide Prediction

The relationship between GDP and the number of suicides in each country.

Xiaomin Wu
2103776

Option: 2

Used algorithms:

1) Linear Regression
2) Decision Tree Regression
3) Random Forest

1 The goal of the work:

This assignment is test the relationship between GDP per capita and the number of suicides in each country, and predicted the number of suicides by new datas.
The datasets are continued data, there are ten regression algorithms for it, and three of those be studied this semester: Linear Regression, Decision Tree Regression and Random Forest. They are used in this project.

2 Original dataset:

There are two datasets in this project:

1) GDP figures in each country from 1985 to 2010,
2) Number of suicides and population in each country from 1985 to 2015, separate by age and gender.

3 Data preparation for the training:

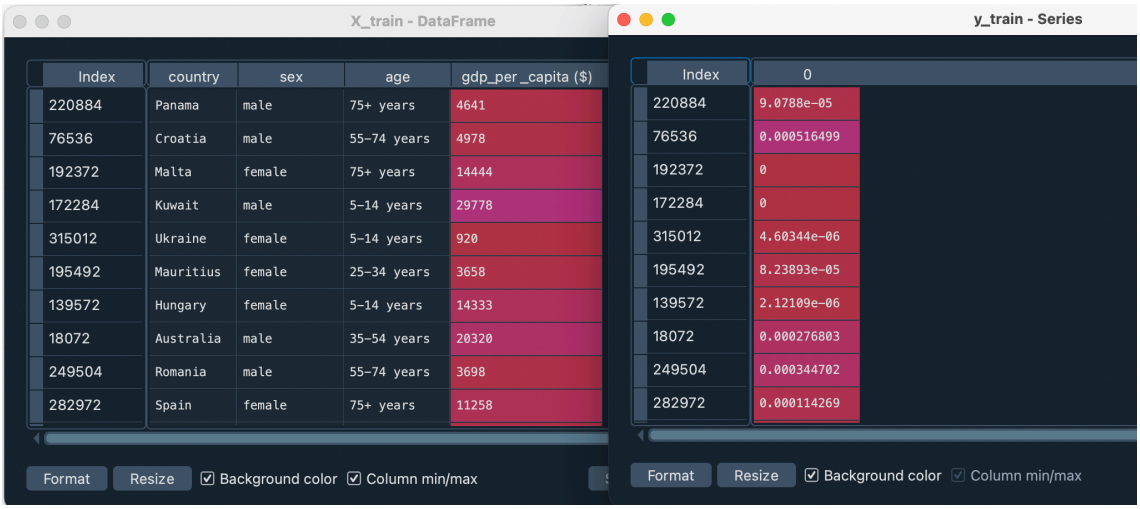
Combine those two datasets to make the data frame in use. Drop some rows that have empty values and duplicate with others.
The data frame includes:

1) Country: suicide data were missing for some countries, therefore discarded those countries, for example China.
2) Year: from 1985 to 2010.
3) Gender: male or female.
4) Age: there are 6 age groups: 5-14 years, 15-24 years, 25-34 years, 35-54 years, 55-74 years, and 75+ years.
5) Suicides number: in each gender and each age group.
6) Population: in each gender and each age group.
7) GDP for year
8) GDP per capita

Set independent variable (X) and dependent variable (y):

X) Country, gender, age, GDP per capita

y) Suicides as a percentage of total population, suicide number divide by population.



PICTURE 1: 10 rows of training data before dummy and scale values.

Dummy and scale values:
Dummy the countries, gender, and age groups, drop first line to avoid 'Dummy Variable Trap'.
Scale the GDP per capita values will be within range *[0, 1]*, divide by maximum value.

4 Relevant metrics for the case(s):

In this project, the values below are used to explain regression metrics:

1) Mean Absolute Error: In statistics, MAE is a measure of errors between paired observations expressing the same phenomenon. The lower of MAE means the model better.
2) Mean Squared Error: In statistics, MSE measures the average of the squares of the errors, it is always a positive value that decreases as the error approaches zero. The lower of MSE means the model better.
3) Root Mean Squared Error: RMSD is the square root of the MSE. So it is always a positive value that decreases as the error approaches zero. The lower of RMSE means the model better.
4) R2 score: R2 is used for evaluating the performance of linear regression models. It is between 0 and 1, the closer to 1 means the better the regression fit.

4.1 Linear Regression

As the first model in the project, country is discussed should it be an independent variable. According to picture 2 and 3 below, when there is country in the independent variables, the model fits better than discard it. In this project, different country means different culture, so it should be an important factor influencing if people suicide.

```
Test data metrics(with country as an independent variable ):
Mean Absolute Error is: 8.550976760282135e-05
Mean Squared Error is: 1.6320609441542055e-08
Root Mean Squared Error is:  0.00012775214065346245
R2 score of the model is: 0.5320630087696575
```

PICTURE 2: Relevant metric of Linear Regression with country as an independent variable.

```
Test data metrics(without country as an independent variable ):
Mean Absolute Error is: 0.00010188076677747862
Mean Squared Error is: 2.5576039010616777e-08
Root Mean Squared Error is:  0.00015992510437894602
R2 score of the model is: 0.2937672623853035
```

PICTURE 3: Relevant metric of Linear Regression without country in X.

4.2 Decision Tree Regression

In Decision Tree Regression, the depth is a factor influencing how better the regression fit, as the picture 4 and 5 below, the larger depth makes the model better.

```
Test data metrics(max depth is 20):
Mean Absolute Error is: 5.8544528017359925e-05
Mean Squared Error is: 1.1285476345521826e-08
Root Mean Squared Error is:  0.00010623312263847762
R2 score of the model is: 0.7135631029583787
```

PICTURE 4: Relevant metric of Decision Tree Regression with max depth is 20.

```
Test data metrics(max depth is 30):
Mean Absolute Error is: 4.623059620036823e-05
Mean Squared Error is: 7.311424101036696e-09
Root Mean Squared Error is:  8.550686581226501e-05
R2 score of the model is: 0.8208498124677265
```

PICTURE 5: Relevant metric of Decision Tree Regression with max depth is 30.

4.3 Random Forest

In Random Forest, the depth also positive correlation influencing the model fit. Compare with Decision Tree, it performs better with same depth.

```
Test data metrics(Random Forest with max depth is 20):
Mean Absolute Error is: 5.387210417258057e-05
Mean Squared Error is: 9.074129948931895e-09
Root Mean Squared Error is:  9.525822772302608e-05
R2 score of the model is: 0.7354778300029019
```

PICTURE 6: Relevant metric of Random Forest with max depth is 20.

```
Test data metrics(max depth is 30):
Mean Absolute Error is: 4.091373475113025e-05
Mean Squared Error is: 6.134511975346396e-09
Root Mean Squared Error is:  7.832312541865523e-05
R2 score of the model is: 0.8375790480120193
```

PICTURE 7: Relevant metric of Random Forest with max depth is 30.

5 Conclusions of the results:

Compare those three models, the accurate sequence is Random Forest > Decision Tree > Linear Regression. This is in line with conventional wisdom of machine learning, the project is good enough as a student's experiment.

But the R2 score of the best model (Random Forest) is still lower than 90%, not good enough for the planned purpose, the result with new data is not accurate enough to use.

This probably because there are more factors influencing the result except GDP and culture (country in independent variable). To improve the model, other indicators can be introduced, such as OECD Better Life Index and so on.