‘Knowledge Discovery and Data Mining’
Data Mining Project with R

Master 1 MLDM
Saint-Étienne, France

S G V Jayani Gunasekara
Student No: 18007140

Contents

Table of Figures

1. Problem Understanding

Suicide Rate (Self-inflicted death)

Suicide rate has become one of the famous topics in the world as reasonable number of people commit suicide all over the world every day. During this study the main objective is to analyze the number of suicides of each country, investigate the evaluation of generations over suicide rates and identify the countries which are at high risk. In each year the affecting individuals belong to the different cultures, regions, genders, educational background, etc.

The main reason behind suicide is unable to manage stress pressure. Since the victims come from different age groups this is leading to loss of talent which is an inevitable fact required to build a nation.

The ability of identifying a suicide behavior is bit difficult unless the personality traits and a close study of each person. Goal of this study is to discover group of people who really need counselling system and rescue them from suicidal tendencies and depression. According to the studies there are several stages of suiciding such as the idea of suicidal planning, preparation, ends with threatening, attempting and completing suicide. According to the researchers, suicide is preventable through the effective treatment of mental illnesses in terms of lot of community activities ranging from kids to elders.

2. Data understanding

The data set "suicide rate overview 1985 to 2016 in file size 2.40 MB from Kaggle website" https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016 is used for this project. There are 27820 records collected under 12 variables such as country, year, sex, age, suicide number, population, suicide / 100K population, country year, GDP_per_year ($), GDP_per_capita ($), generation and HDI for year. The data set contains lot of categorical data. HDI for year represents the human development index which represent the life expectancy, education and per capita income of the country. HDI is a factor developed by Pakistani economist "Mahbub Ul Haq" which is now used by united nations development program as a measure of country's development.

Gross domestic product (GDP) is a measure of purchasing power of a good or a service. Here the data is observed in terms of year. Further GDP does not reflect the inflation rates and cost of living of countries. GDP per capita evaluates a purchasing power of a person in a country. It is evaluated by dividing the GDP by the population of the country.

Overall evaluation of data understanding shows the data set consists with limited number of useful variables after excluding the variables, which can be derived by another two columns (Country year and GDP per capita). Though the data set is quite small, it carries out an important interesting relevant topic.
This data set can be explored with certain additional variables such as biological risk factors, psychological risk factors, warning signs, environmental risk factors, suicidal behavior etc.

3. Data Preparation

First task is to recognize the number of null values in each column. The column HDI for year was removed from the dataset as it contains 69.93% null values. Generally acceptable maximum missing value percentage is 25% to 30%. Another way of eliminating missing values is logically substitute values for those missing entities, but in this situation the percentage of missing values is too complicative. At the same time the duplicate entries were deleted.

Categorical data is converted to numeric values to simplify the calculations. In addition, the features relevant to the study are selected while excluding the unnecessary variables. Then the data set is subdivided into two categories as training and test data, where the training data recalls 75% of the whole population. Further the data is normalized.

4. Modeling the data

Initially the correlation between selected variables are checked.

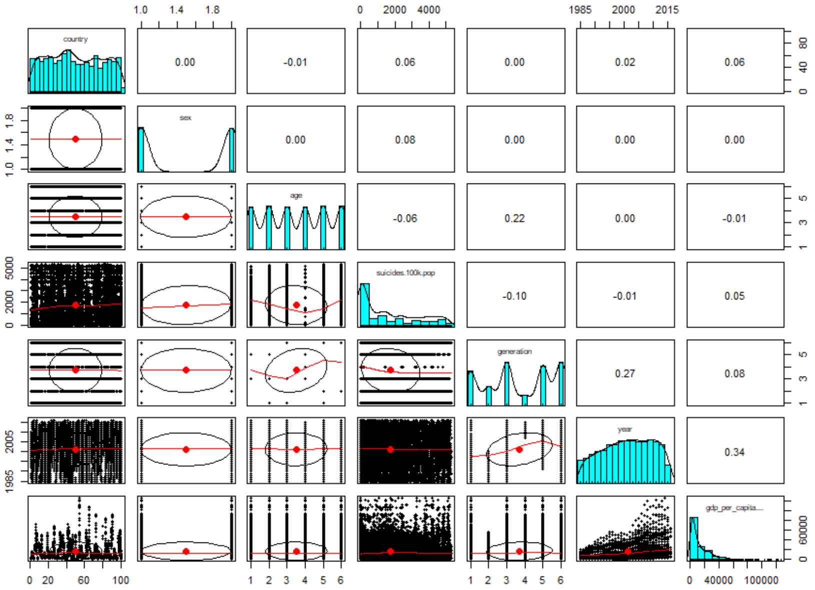
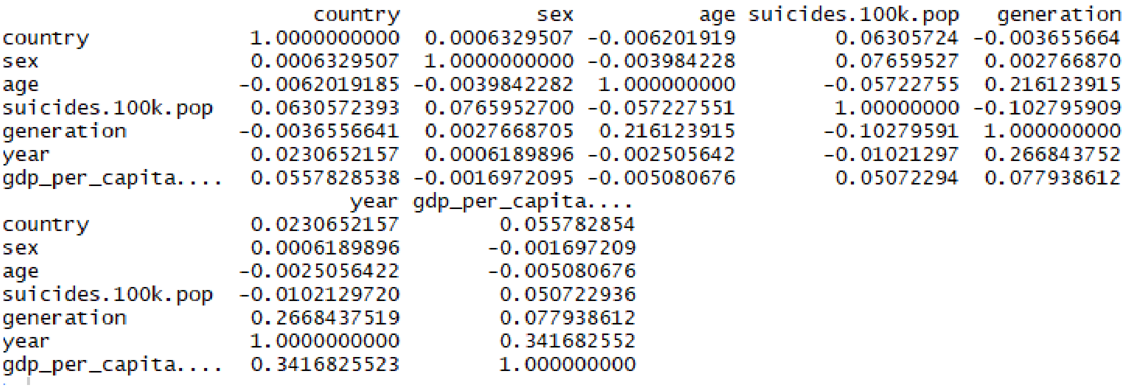


Figure 1 Correlation between variable



```
                      country           sex          age suicides.100k.pop   generation
country             1.0000000000  0.0006329507 -0.006201919        0.06305724 -0.003655664
sex                 0.0006329507  1.0000000000 -0.003984228        0.07659527  0.002766870
age                -0.0062019185 -0.0039842282  1.000000000       -0.05722755  0.216123915
suicides.100k.pop   0.0630572393  0.0765952700 -0.057227551        1.00000000 -0.102795909
generation         -0.0036556641  0.0027668705  0.216123915       -0.10279591  1.000000000
year                0.0230652157  0.0006189896 -0.002505642       -0.01021297  0.266843752
gdp_per_capita....  0.0557828538 -0.0016972095 -0.005080676        0.05072294  0.077938612
                           year gdp_per_capita....
country             0.0230652157        0.055782854
sex                 0.0006189896       -0.001697209
age                -0.0025056422       -0.005080676
suicides.100k.pop  -0.0102129720        0.050722936
generation          0.2668437519        0.077938612
year                1.0000000000        0.341682552
gdp_per_capita....  0.3416825523        1.000000000
```

Figure 2 Correlation between variables

There exists no strong correlation between variables where generation-year, gdp_per_capita-year, age-generation interpret a relatively small correlation. Moreover generation-suicides.100k.pop show a slightly negative correlation. Hence multiple linear regression algorithm cannot be used to predict the suicide rate.

Algorithm to classify the data according to the generation

The suicide data(suicides.100k.pop) is classified in terms of generation over the year 1985 to 2016. (Figure 3). Different colors of dots denote the six different generations named, Generation X, Silent G.I. Generation, Boomers, Millennials and Generation Z. Support vector Machine (SVM) was used for the analysis as it has strong theoretical properties.

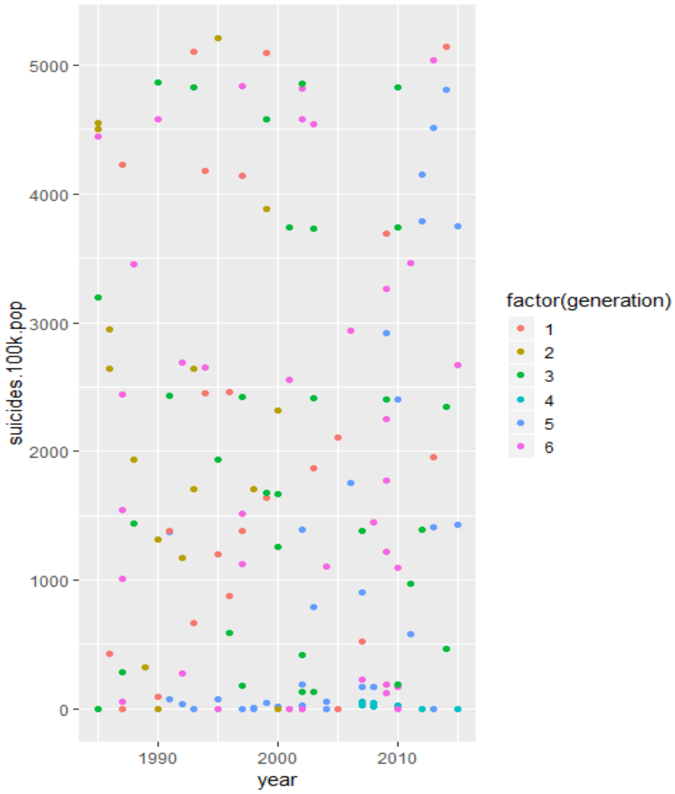
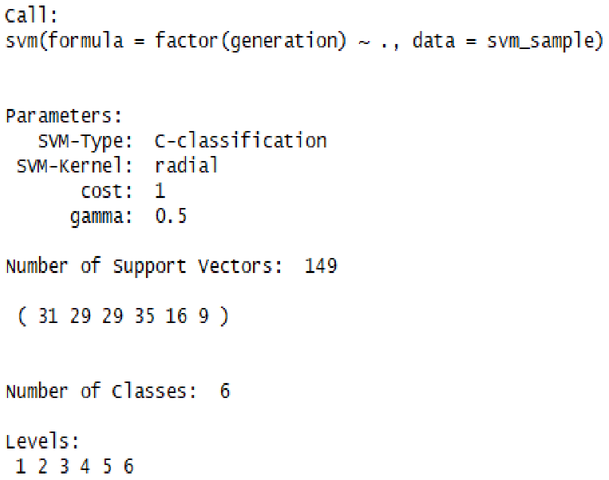


Figure 3 Plot of suicide rate over year



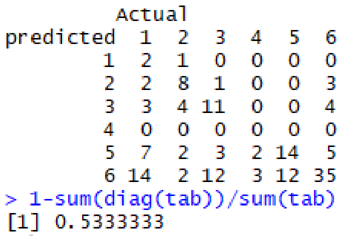
```
Call:
svm(formula = factor(generation) ~ ., data = svm_sample)


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.5

Number of Support Vectors:  149

 ( 31 29 29 35 16 9 )


Number of Classes:  6

Levels:
 1 2 3 4 5 6
```

Figure 4 Summary of SVM

Figure 4 illustrates SVM type as classification because the study is carried out for categorical variable. SVM kernel type ‘radial’ was selected as SVM-Kernel after checking the other modes of kernels ‘linear’ and ‘sigmoid’ where ‘radial’ performed well than the other occasions. At the beginning the cost was set to 1 and gamma =0.25. There were 149 support vectors group into 6 categories of generation.

Confusion Matrix and Misclassification

The misclassification error was recorded as 0.533333 which is quite high. The hyperparameter tuning was done in terms of different epsilon values and cost values to optimize the misclassification error.



```
           Actual
predicted  1  2  3  4  5  6
        1  2  1  0  0  0  0
        2  2  8  1  0  0  3
        3  3  4 11  0  0  4
        4  0  0  0  0  0  0
        5  7  2  3  2 14  5
        6 14  2 12  3 12 35
> 1-sum(diag(tab))/sum(tab)
[1] 0.5333333
```

Figure 5 Misclassification error

The regions of the plot of tuned model illustrates the cost variation. Out of the regions, dark regions specify the better results. If the cost is too high it turns out that high penalty for non-separable points. That means model stores too many support vectors and leads to overfitting. On the other hand, when cost value is too small, derives to underfitting which means the model is inadequate to explain the data set.

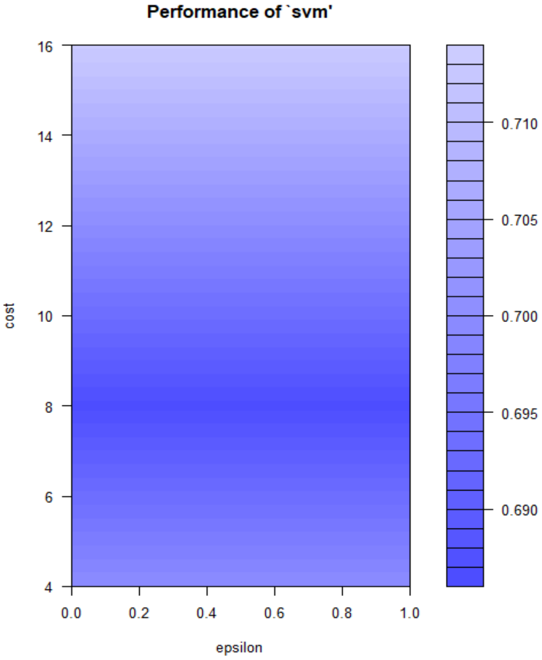
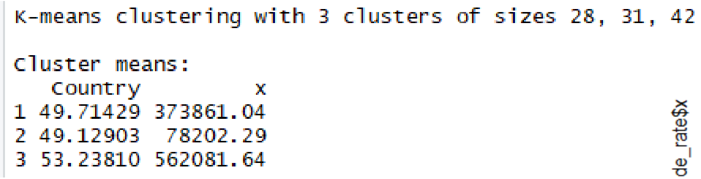


Figure 6 plot of tuned model

Risk of the suicidal rate around the world

Main objective of this study is to recognize the countries which are at high risk of number of suicides. Over the path of classification, I used K means clustering for the grouping algorithm. K means clustering is a well-known machine learning algorithm.



```
K-means clustering with 3 clusters of sizes 28, 31, 42

Cluster means:
   Country         x
1 49.71429 373861.04
2 49.12903  78202.29
3 53.23810 562081.64
```

Figure 7 Clustering

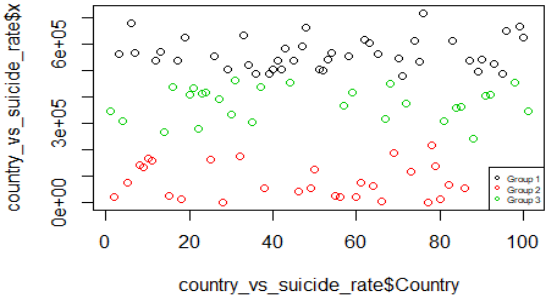
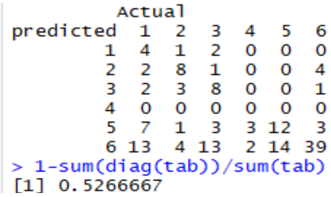


Figure 8 Plot of the clusters, country vs suicides_100k_population

There are 3 clusters varying according to the level of risk (Figure 8).

5. Evaluation and Deployment

After the tuning of the SVM algorithm, it performs 48% classification accuracy which is quite higher than the initial model.



```
            Actual
predicted  1  2  3  4  5  6
        1  4  1  2  0  0  0
        2  2  8  1  0  0  4
        3  2  3  8  0  0  1
        4  0  0  0  0  0  0
        5  7  1  3  3 12  3
        6 13  4 13  2 14 39
> 1-sum(diag(tab))/sum(tab)
[1] 0.5266667
```

Figure 10 Misclassification error after tuning

The graph exhibits the data which is classified over the 6 generations accordingly suicide rate 100k population vs year. The six different colors of crosses represent the generations. Since the misclassification rate is reasonable this graph doesn't show the exact edges of the kernels. Generation "Silent" records highest suicide rate over the year 1990 to 2005. Silent generation represent the people age 35-54 years.

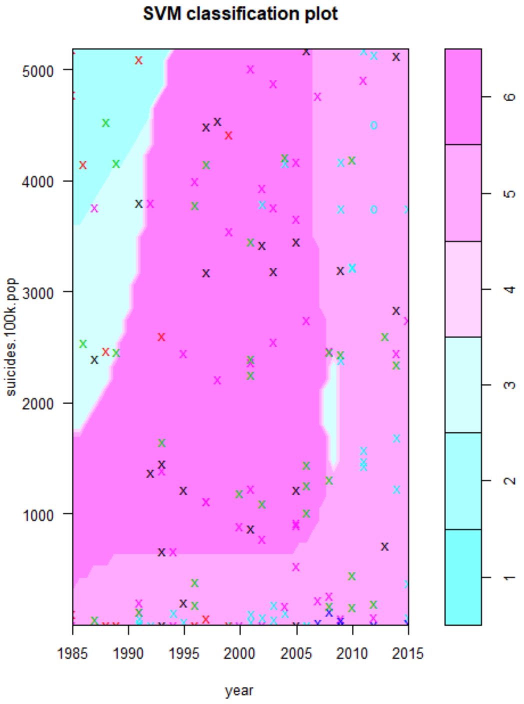
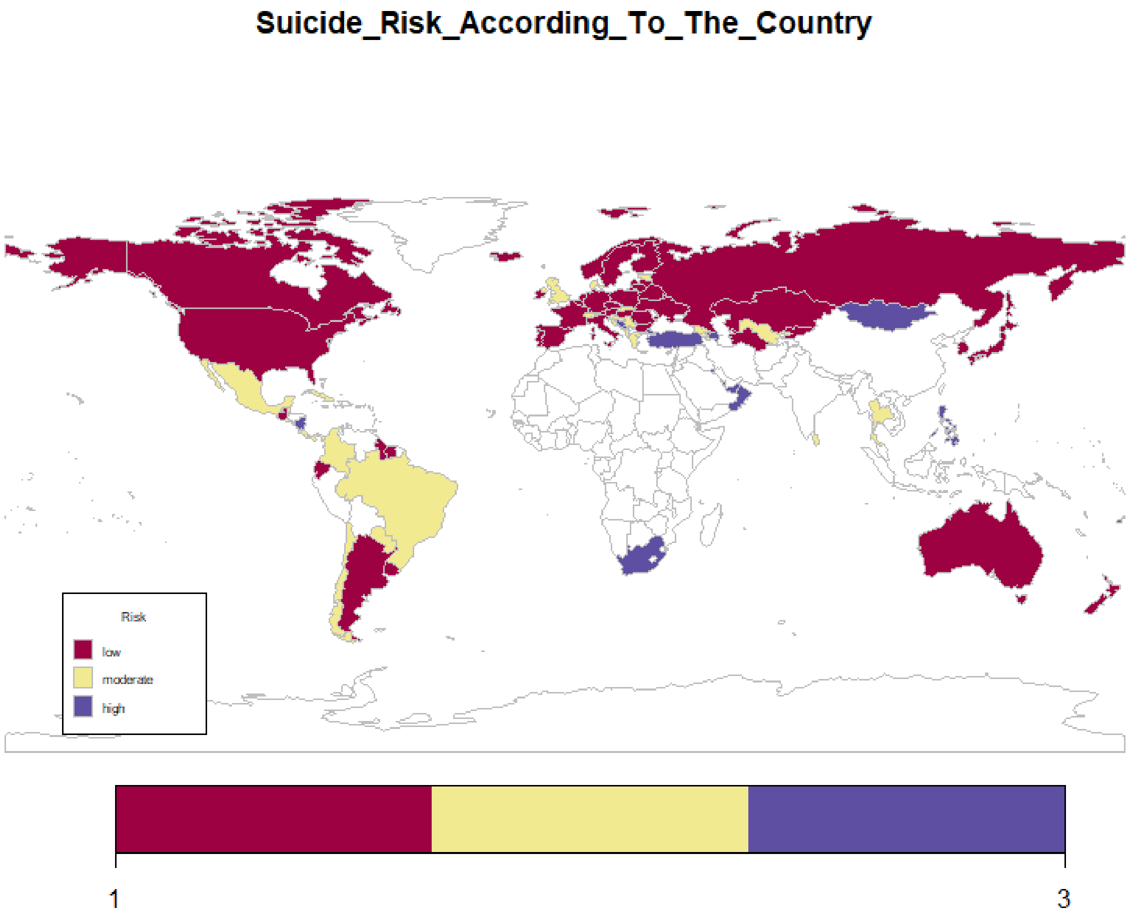


Figure 11 plot of SVM classified according to the generation

As an overall conclusion of two different evaluation problems the data classification done by SVM quite unclear as it was designed to all the entities without grouping them in terms of one variable. But in the K-mean clustering, I grouped and calculated the sum of suicide rate per 100k in each country to clearly investigate the risk associate with each country. The below world map shows the risk faced by each country. France records low risk of suicidal.


Suicide_Risk_According_To_The_Country

6. Reference

- Mandge,O.L,2013,A Data Mining Tool for Prediction of Suicides among Students http://www.conference.bonfring.org/papers/met_ncnhit2013/ncnhit39.pdf
- Bharatendra, R.,2015,https://www.youtube.com/watch?v=5eDqRysaico
- Patel,2017,https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72
- Berwick,R.,http://web.mit.edu/6.034/wwwbob/svm-notes-long-08.pdf