Analyzing data of suicide rates to discover factors associated with death

Erlend Kristensen
(Dated: January 26, 2023)

In this report i went ahead and looked at data over suicide rates for over 100 different countries over different years. I first tested *multilayer perceptron* (MLP) and random forest classifiers to see which performed better, and found random forest performed best, where it produced an F1-score of 0.99 at highest, while MLP produced a highest score of 0.61. I also looked into the different features, and found the countries with the highest suicide rates. Here i saw that South Korea had a suicide rate almost twice as high as the country with the second largest suicide rate, meaning suicides is a massive problem there, indicating very poor mental health among its population. I also found out that men have about a 4 times higher suicide rate than women, and the data indicated that the suicide rates increase with age. I also studied GDP, both yearly and per capita, and found strong indications that GDP per capita does affect the suicide rate, where higher GDP is good. However the data also suggested that too much GDP made for a worse suicide rate again, indicating that there is a sweet spot in the amount of money the population should have for a good mental health.

I. INTRODUCTION

Suicides is an ever increasing problem in our modern society, with on average 800000 yearly suicides [Our World in Data VII]. As our society seems to become more and more technological and modernized, there also seems to be an indication of a decrease in mental health and well being among the population. Seeing as mental health is such a complex field, it is difficult to pinpoint exactly what to blame for the decrease in mental health. however, i will in this report use a large data set of suicides across the world to get an indication of where most suicides tend to happen, and why this might be.

I will primarily utilize two different machine learning algorithms: The *Multilayer Perceptron* (MLP) and *Random Forest*. The algorithms are trained with a data set of suicides in different age groups, from different countries, to try to classify which group are likely to be above the median suicide rate and which groups are not. Once the algorithms are trained, we will analyze the models to extract the relative importance of each feature. We will then analyze each individual feature to further see what plays a role into the suicide rate.

II. THEORY AND METHODS

A. Data set

I will use a publicly found data set (A) to study factors contributing to suicides. It is a data set compiled of four other data sets from the United Nations Development Program, World bank and World Health Organization. The data set consists of features such as total amount of suicides for a certain age group, amount of suicides per 100000 citizens for this age group, which country they are from, what year and how the GDP was at that time.

Since the data set consists of number of suicides across different age groups, it is inherently a regression problem. To make it into a classification problem, which is easier for our machine learning algorithms, we calculate a median number of suicides, and give the value 0 to the instances lower than this number, and 1 to the instances higher. This means we will classify whether or not a group from a certain country is above the median or not.

To investigate which relevance the different features have in contributing to the suicide rates, we will use the Random Forest Classifier, which has a built in feature importance method which can analyze this for us.

B. Classifiers

1. *Multilayer Perceptron*

MLP is a type of neural network with both input and output layers, whose neurons can use any arbitrary activation function, such as *ReLU* or *sigmoid* [Carolina B. (VII)]:

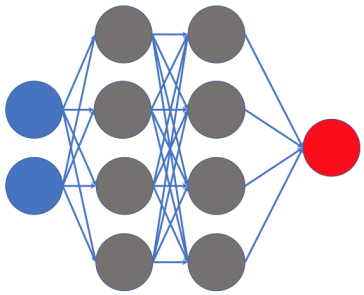

FIG. 1. A representation of what an MLP classifier could look like.

MLP is often used for classification problems and for situations where we don't know the complexity of our

given problem. For more info on feed forward neural networks, have a look at an earlier report I did, found here.

2. Random Forest

To describe the random forest classifier, we start off by looking at what it is built upon, which are *decision trees*. Decision trees work by asking simple questions, which for this data set would be questions such as "which age group is it, which country are they from" etc, and use these facts that they gather to determine an overlying question, which in our case would be "is this group of people above the median in number of suicides?".
With decision trees we can choose how many nodes we want, which means how many questions we want the classifier to ask.

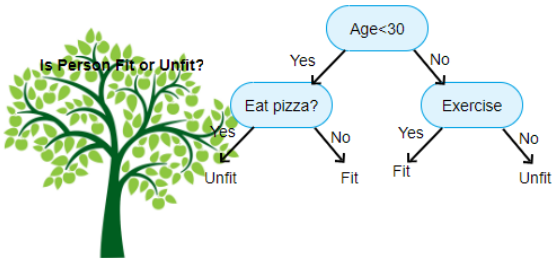

FIG. 2. A representation of what a decision tree would look like, asking the question "is this person fit?".

A random forest classifier consists of multiple independent decision trees. Each tree makes a decision and our classifier chooses the class which the majority of the trees voted for. It is the fact that each tree is different which ensures us that our classifier performs well. Random forest creates this randomness by allowing each individual tree to sample randomly from the data [Tony Yiu (VII)].

C. Evaluation

Since the data set consists of a lot of different features and variables within these, making it extremely complex, we will need to look at more than one form of performance measurement other than accuracy. We will therefore also look at *F1-score*, which can be given by this formula:

$$F_1 = 2 * \frac{precision * recall}{precision + recall} = \frac{2tp}{2tp + fp + fn} \qquad (1)$$

where tp is true positives, fp is false positives, and fn is false negatives.

III. IMPLEMENTATION

For this report i will use built in MLP and random forest classifiers from *scikit-learn*. These classifiers are very flexible, and allow us to change number of hidden layers, which activation function we want to use, and so on. So I will start off by analyzing the results from different choices of parameters and hyper parameters (such as learning rate) to find which classifiers are most suited for this problem, and what they can tell us.

For MLP we will test out three different activation functions, namely ReLU, logistic (sigmoid) and *tanh*. More info on these can be found in the report cited above in section (II B 1).

Since the data set is not too large for us to use, we will use the entirety of it when testing. However, we will split it into a training part, consisting of a random sample of 75%, and a test part which is the remaining 25%. This is to ensure that our classifiers do not end up with overfitting.

IV. RESULTS AND DISCUSSION

As stated previously, I start of by analyzing each activation function of different hyper parameters for the MLP classifier using the training and testing parts.

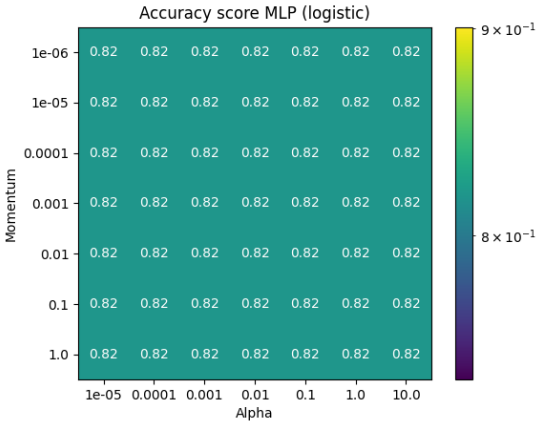

FIG. 3. Plot of accuracy made using logistic as an activation function, and changing the alpha and momentum parameters.

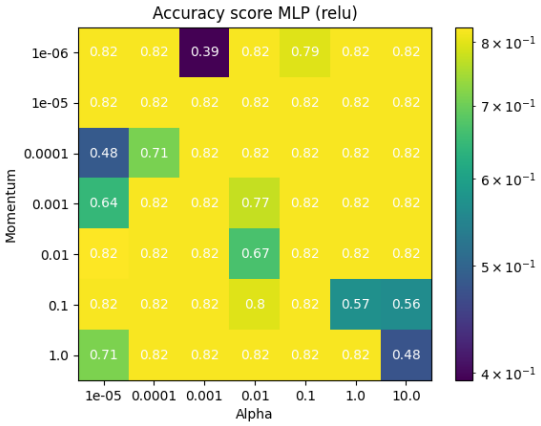

FIG. 4. Plot of accuracy made using ReLU as an activation function, and changing the alpha and momentum parameters.

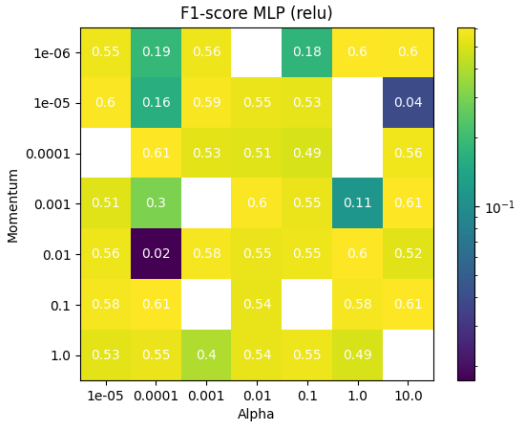

FIG. 6. Plot of F1-score made using ReLU as an activation function, and changing the alpha and momentum parameters.

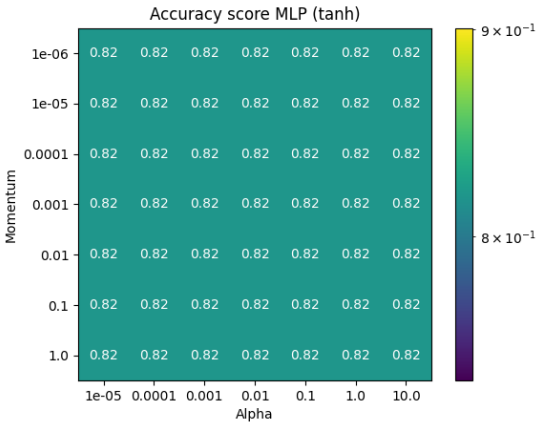

FIG. 5. Plot of accuracy made using tanh as an activation function, and changing the alpha and momentum parameters.

As we see, the ReLU activation function seems to be the only one that changes with α and momentum changes. This could however be due to random initialization where some seeds make the classifiers stuck in one solution. However, looking at these results it seems to indicate that the classifiers work fine, but as stated earlier we need to also look into the F1-scores. We do this for the ReLU activation function then, since it was the one varying the most and could give us the most information.

Here we see that the results aren't as great anymore. The best F1-score is about 0.61, which is not bad, but not really good either, we also have a lot of low F1-scores overall, with the white spots being 0.
I also run a test of F1-scores for different sample sizes for all the activation functions:

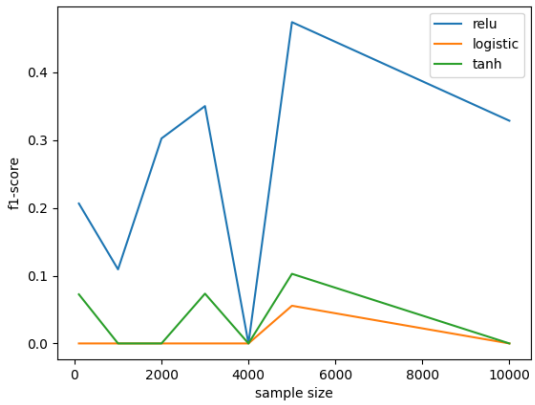

FIG. 7. Plot of F1-score made using the three different activation functions for the MLP classifier, with different sample sizes.

Here we see ReLU outperforming the other two activation functions greatly, with almost four times the performance. However the results aren't enough to be satisfied with, so I will therefore see if using the random forest classifier can provide better results. We therefore run the same sampling test using our random forest classifier.

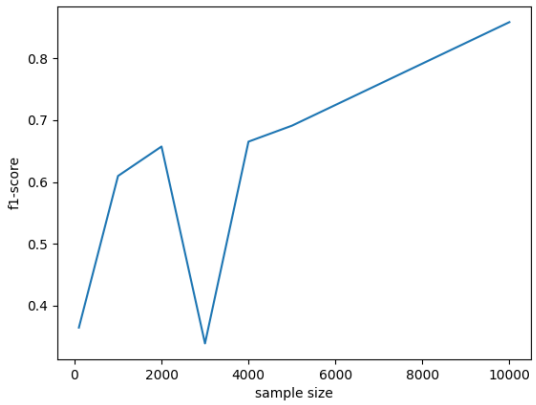

FIG. 8. Plot of F1-score made using different sample sizes with the random forest classifier.

Here we see much greater results, with about twice the score as what MLP provided us with ReLU activation function. I also tested for the entire data set for random forest, and got a score of above 0.95 on average doing multiple runs with highest score being almost 0.99, where as MLP gave us 0.61 as the highest score. This indicates to us that random forest classifier is better suited for this problem compared to MLP.

To now investigate the causes for the suicide rates, i will use the random forest classifiers built in feature importance method to see which features matter more than others.

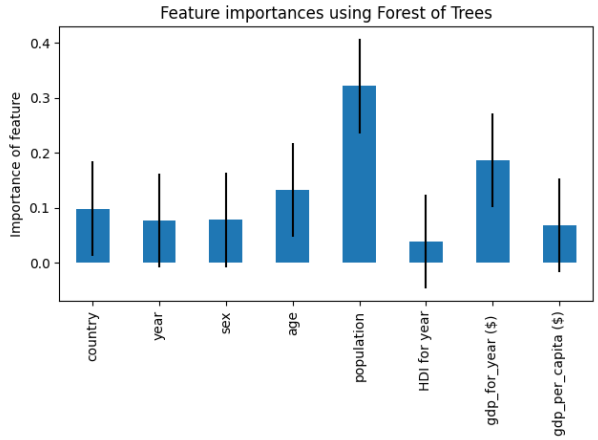

FIG. 9. Plot of feature importance for the data set, made using the random forest classifiers built in feature importance method.

Here we see that population is the biggest contributor to suicide deaths. This is not unexpected, seeing as the higher the population, the higher the total number of suicides will be. However, we also see country, age and GDP per year playing a significant role. We will therefore study these more, in addition to the sex of the person.

We start of by looking at which countries have the highest suicide numbers. We base this off by using suicides per 100000 population to ensure that the number of population does not matter.

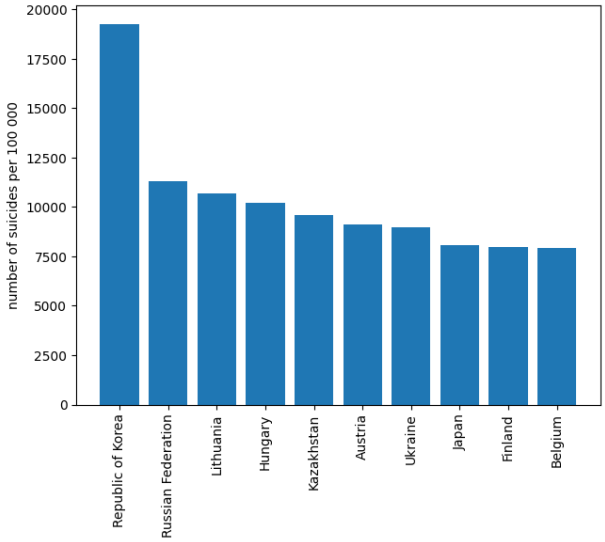

FIG. 10. Plot of top 10 countries with the highest suicide rate.

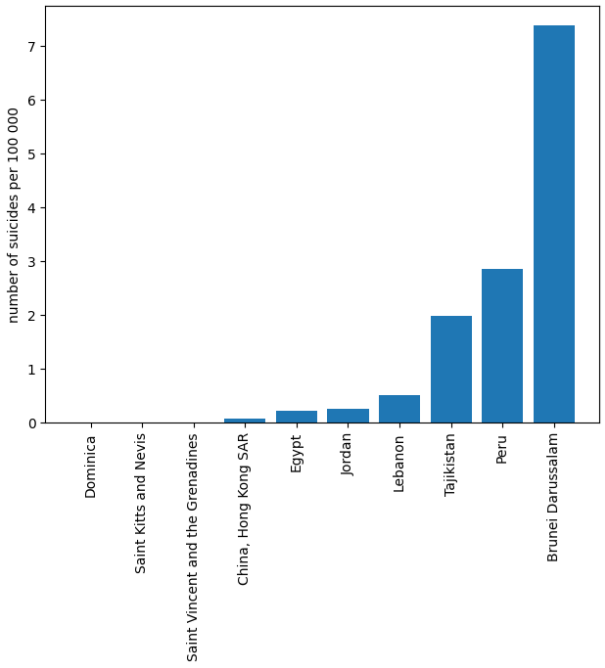

FIG. 11. Plot of bottom 10 countries for suicide rate.

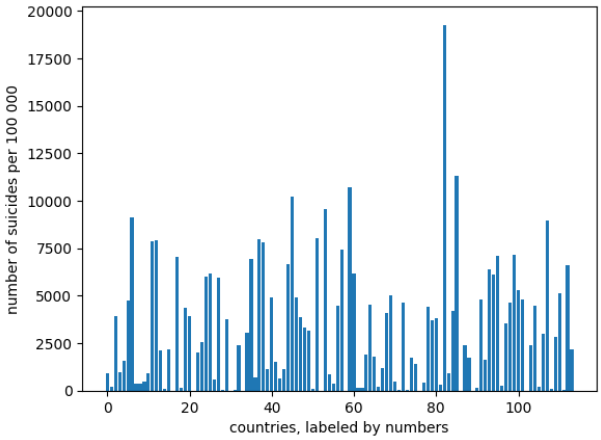

FIG. 12. Plot of suicide rate for all countries. Since there are too many countries to label, each are represented by a number given by the order that they appear in the data set.

As we can see, there is a huge difference in the 10 countries with the highest number of suicides, and the 10 countries with the lowest. Looking at the top 10 we see South Korea having a lot more suicides than the other countries, indicating to us that mental health in Korea is way worse than the rest of the world. The bottom 10 however seem to be a bit small. This could be due to a lack of data, or if the total population of some of those countries are so small that it doesn't register, seeing as we look at suicides per 100000 citizens.
When looking at the graph for all the countries, we can clearly again see how much South Korea stands out on the suicide numbers, which further proves to us how insanely high the suicide rate there is.

As stated, I also wanted to look into which age groups and which genders are more affected by suicides.

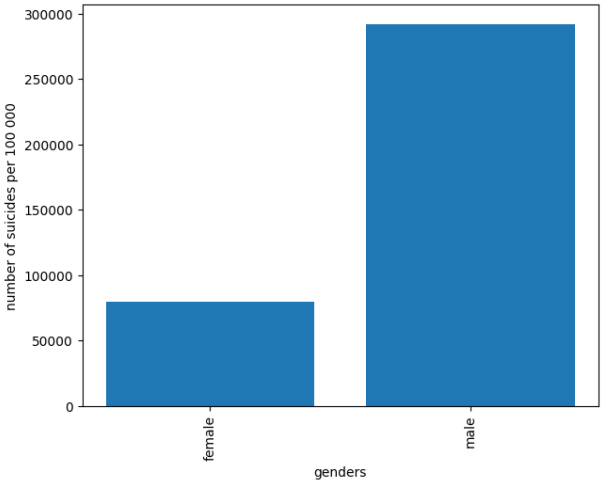

FIG. 13. Plot of suicide rates for both genders.

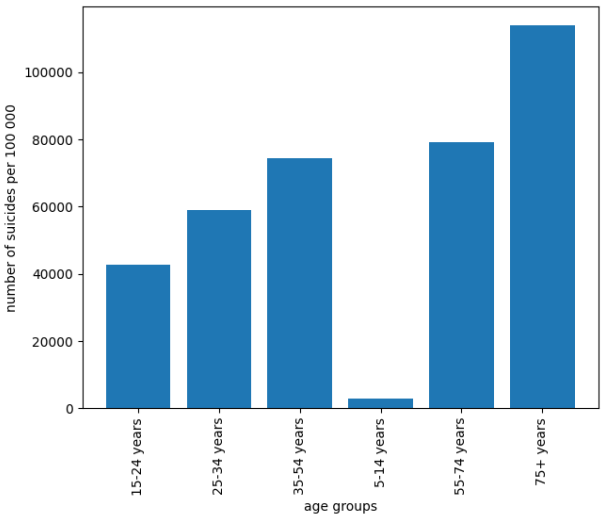

FIG. 14. Plot of suicide rates for different age groups.

Here we see a clear indication that men's mental health is worse of than women's, where the number of suicides seem to be about 4 times larger. We also see that the number of suicides seem to increase the older the age group we look at, indicating to us that mental health among elders is worse of than that of the younger population.

Finally i wanted to look into GDP, both yearly and per capita to see the effects this has on the suicide rates, and what we can learn from it.

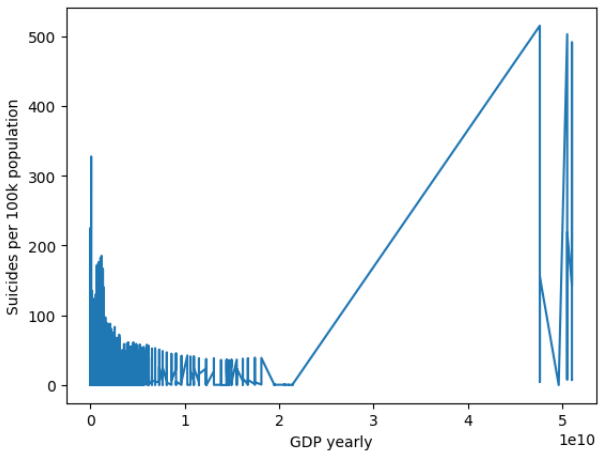

FIG. 15. Plot of suicide rates for different countries based on yearly GDP.

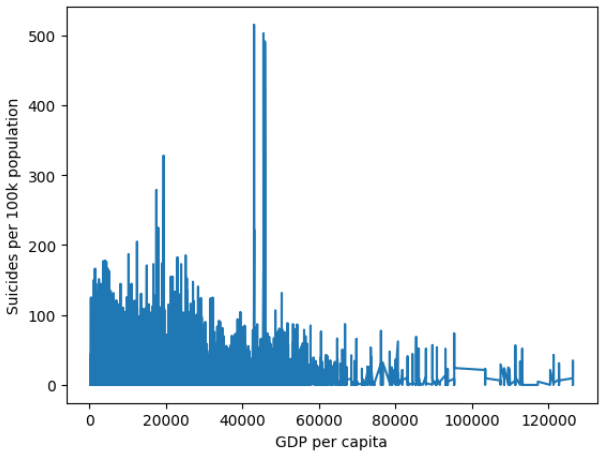

FIG. 16. Plot of suicide rates for different countries based on GDP per capita.

As we see on both figures, there seems to be a trend that higher GDP results in a lower suicide rate. The yearly GDP plot is a bit more difficult to analyze than the GDP per capita plot. To further investigate whether or not GDP matters, I used a linear regression analysis of the data for GDP per capita.

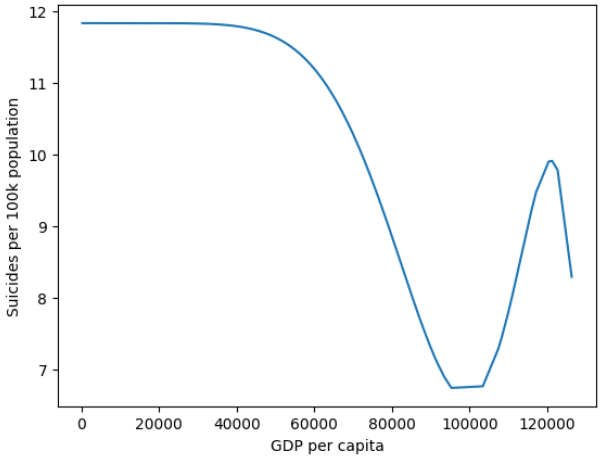

FIG. 17. Plot of suicide rates for different countries based on GDP per capita made using linear regression analysis.

As we can see from this plot, there seems to be a strong indication that suicide rate does fall when increasing GDP. However, there also seems to be an increase when countries get even richer, which might indicate that strictly more money does not always result in a better mental health among the population. The data set seems to indicate that at about 100000 GDP per capita is the sweet spot for mental health. Any less will make life too difficult, but any more could perhaps have effects on the culture, such as in South Korea, and result in worse mental health.

V. CONCLUSION

I have gone through and tested both MLP and random forest classifiers and found that random forest performs a lot better when predicting which countries have a high suicide rate. I also looked into the features, and found that South Korea as an extremely large suicide rate compared to the rest of the countries. I also found out that suicide rate among men is almost 4 times higher than that among women, and that the suicide rate seems to increase with age. Finally i looked at GDP and found indications that higher GDP results in lower suicide rates, meaning money does in a way bring happiness.

VI. CODE

The code used in this project can be found in this Github repository.

VII. REFERENCES

Hannah Ritchie, Max Roser and Esteban Ortiz-Ospina Suicide (Our world in data, 2015)

Carolina Bento Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis (Towards Data Science, 2015)

Tony Yiu Understanding Random Forest (Towards Data Science, Jun 12, 2019)

Appendix A: Dataset

The dataset was downloaded from: Suicide Rates Overview