Suicide Rates from 2000 - 2010

ChingWei Hsieh
University of Pittsburgh
PA, United States of America
chh182@pitt.edu

I.ABSTRACT

Suicide occurs all over the world and is a serious public health problem worldwide, touching individuals of all nations, cultures, genders and generations. The suicide rate like the number of suicides in population we consider to be an important indicator of general well-being of society.

The purpose of increasing awareness of the public health significance of suicide, as well as making suicide prevention a high priority on the global public health area. We believe that it is important to increase significance of Heath care organizations, non-profitable lifeline groups, life awareness groups, society.

The objective of analyzing this data is to contribute to an informed, and communities can have a better understanding about suicide rates. Studies regarding the characteristics that underlie differences in suicide rates across various countries have the potential to yield meaningful insights for interventions and related strategies.

INTRODUCTION

Data Analyzation on the given dataset

Our dataset records data from 1987 – 2016, and according to different countries, the amount of samples are different which implies we have missing values in our dataset and is MCAR(Missing Completely At Random).

The only consistent data and equal number of samples are throughout years from 2000 – 2010. Therefore, we could not preform analyzations on incomplete data, so we apply a filter that filters out all data that are not from years to 2000 – 2010.

Since we already have the output for suicides/100k pop, so we perform supervised MLRegression to estimate values that are already known and test which algorithm suits best and why.

Analyzing columns

This step is important because we want to check if the data is biased, for example: sampling females more than males or only considering people that are at a certain age.

- Country: The countries that participated in this data set are briefly shown in a figure below.

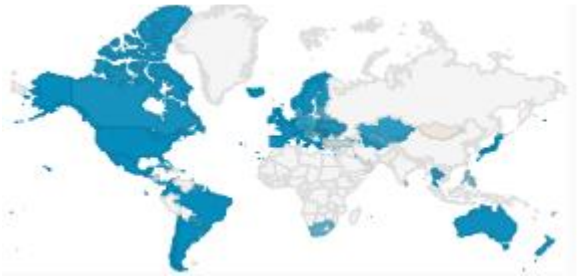
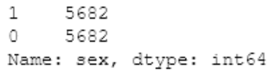


Figure 1: countries that are sampled

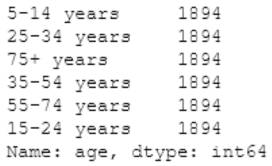
- Gender: Fifty percent of the portion sampled is female and the rest fifty percent is male.



```
1    5682
0    5682
Name: sex, dtype: int64
```

Figure 2: screenshot of program

- Age: Each age group is evenly sixteen percent of the sampled age group data.



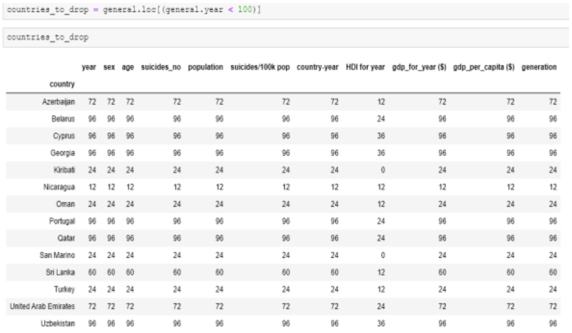
```
5-14 years     1894
25-34 years    1894
75+ years      1894
35-54 years    1894
55-74 years    1894
15-24 years    1894
Name: age, dtype: int64
```

Figure 3: screenshot of program

- Categorical Columns: Country, Year: 1985 to 2016, Sex: Male/female, Age, Generation
- Numerical Columns: Population, Number of Suicides, Suicides per 100k people, GDP per capita for year

From the overlook of data collected, we could see that most of the data sampled are evenly distributed, except for the distribution of countries. Unfortunately, since we do not have supplement data available, we would have to work with limited data.

- As a alternative method, we decided to leave out data that are way too less to be considered. So, we filtered out data that does not have at least 100 samples.



```
countries_to_drop = general.loc[(general.year < 100)]
```

```
countries_to_drop
```

country	year	sex	age	suicides_no	population	suicides/100k pop	country-year	HDI for year	gdp_for_year ($)	gdp_per_capita ($)	generation
Azerbaijan	72	72	72	72	72	72	72	12	72	72	72
Belarus	96	96	96	96	96	96	96	24	96	96	96
Cyprus	96	96	96	96	96	96	96	36	96	96	96
Georgia	96	96	96	96	96	96	96	36	96	96	96
Kiribati	24	24	24	24	24	24	24	0	24	24	24
Nicaragua	12	12	12	12	12	12	12	12	12	12	12
Oman	24	24	24	24	24	24	24	12	24	24	24
Portugal	96	96	96	96	96	96	96	24	96	96	96
Qatar	96	96	96	96	96	96	96	24	96	96	96
San Marino	24	24	24	24	24	24	24	0	24	24	24
Sri Lanka	60	60	60	60	60	60	60	12	60	60	60
Turkey	24	24	24	24	24	24	24	12	24	24	24
United Arab Emirates	72	72	72	72	72	72	72	24	72	72	72
Uzbekistan	96	96	96	96	96	96	96	36	96	96	96

Figure 4: Countries we should drop

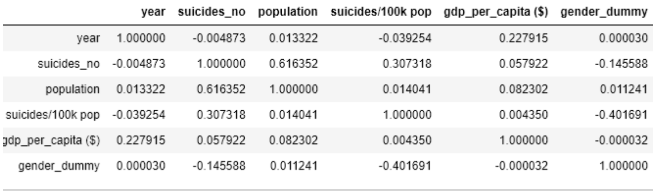
II.METHODOLOGY

We have different types of values recorded in our dataset, some columns are labels, and some are numerical. Therefore, it is important that we think about our desired output and result to decide on which algorithm to use.

Examine correlations

First of all, before performing any machine learning algorithm, we should examine our data carefully. Our goal is to predict values for suicides/100p population, so we should pick suitable features by using the “.corr()” method. We chose to convert the gender column from labels to dummy variables ('male':0, 'female':1) because we would also like to see the correlations between gender and the suicide rates.

- From the table we could also observe that Dimension Reduction is not necessary because we do not have highly correlated column that may affect our results



	year	suicides_no	population	suicides/100k pop	gdp_per_capita ($)	gender_dummy
year	1.000000	-0.004873	0.013322	-0.039254	0.227915	0.000030
suicides_no	-0.004873	1.000000	0.616352	0.307318	0.057922	-0.145588
population	0.013322	0.616352	1.000000	0.014041	0.082302	0.011241
suicides/100k pop	-0.039254	0.307318	0.014041	1.000000	0.004350	-0.401691
gdp_per_capita ($)	0.227915	0.057922	0.082302	0.004350	1.000000	-0.000032
gender_dummy	0.000030	-0.145588	0.011241	-0.401691	-0.000032	1.000000

Figure 5: Table of correlations

Deciding on Features

Else than depending on correlations and measurements on data visualizations, we should also think about which kind of data makes sense to be counted as features. The following bullet points show in detail description why we chose it or why not.

- Country: We believe that suicide rates differ in country due to many factors. There may be physical factors and also invisible ones which we could not easily visualize it with numbers. Physical factors such as cultural differences, annual salary, capital GDP and much more. Invisible factors such as peer pressure, cultural behaviors, human intimacy and much more.

- Gender: It could be shown that males have higher tendency in suicide rates, but after comparison in studies, we found that this dataset has an overrepresentation of males regarding suicide rates. According to a study [1], we could see that females and males have around 6% of the whole population difference. But in our data set we have sampled the same number of women and men. Therefore, the female gender has been sampled too much meanwhile the number of samples for men has been too less.

- Suicides_no (or population): The number of suicides will not be considered as a feature in our program due to several reasons.

- The insight of creating this program was meant to easily be able to estimate suicide rates. Features such as age, gender, country are easy to decide on, but it is not same for numbers of suicides occurred. The same also applies for population. It is almost impossible to ask for a user to come up with accurate/similar values of the numbers of suicide and population. Although the correlation graph shows high correlation between “suicide_no” and “suicides/100k pop” due to decreasing population, therefore decreasing suicide_no

- Also, the value of suicides_no and population could somehow related to suicide/100 k pop.

- GDP: Although GDP is an extremely important factor because it shows how wealthy the country is therefore how lifestyles of the citizens in the country is, but similar to above reasons, this knowing the accurate GDP of a country is very hard for a casual user, so we excluded this feature.

- Country_year: This represents the year of the data recorded in the corresponding country. Which basically is equivalent to the year of the suicide occurred.

In summarization, we will be using the following features to predict the target.

TABLE I.

Features	country	Country_year	sex	age
Target	Suicide/100 k pop			

Figure 6: Model Feature and Target

Classification VS Regression

With the given dataset, we now have two options to proceed to get our final model. Either we Convert all columns to numerical data or into categorical values. The following sections show arguments for both debates.

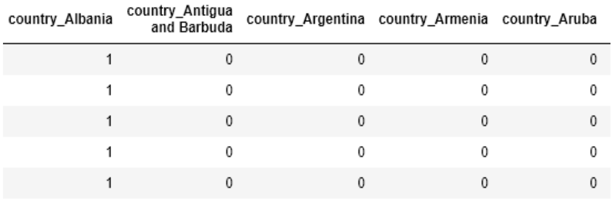
Suicide rate as numerical value(will have to ignore country differences)

We would like to predict values for suicides/100p population. so it would presumably fall into the regression section instead of classification. There are several algorithms that are suitable for our dataset and approach. Since we would like to predict numerical values, we should choose algorithms that are suitable for regression problems.

The idea of predicting suicide rates as a numerical doesn’t seem to be the best approach due to the following reasons:

- The features as extremely low coefficients as correlation with each other.

- The feature: country is a significant but categorical value, so we definitely have to convert country to numerical values in order to perform regression methods. The option of converting each country to an according column was considered but was withdrawn due to practical reasons. This will result in too many columns for countries (There are 96 countries in total), and would be very impractical to set all of them as individual features and train them, therefore we should change an approach.



country_Albania	country_Antigua and Barbuda	country_Argentina	country_Armenia	country_Aruba
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0
1	0	0	0	0

Figure 7: Laborious columns we should avoid

Therefore, we will focus on predicting suicide rates via classification methods instead.

Suicide rate as categorical outcome

Options to perform SVM (Support Vector Machines), KNN (k-Nearest Neighbors), Naive Bayes, Decision Tree, Random Forest

- Transform suicide rates to labels: we would transform the numerical data to label buckets in order to be able to predict categorical values.
- We chose to use the binning/bucketing discretization method.

Converting from “suicide/100k pop” to: labels= ["su_0_2", "su_2_5", "su_5_15", "su_15_200"]

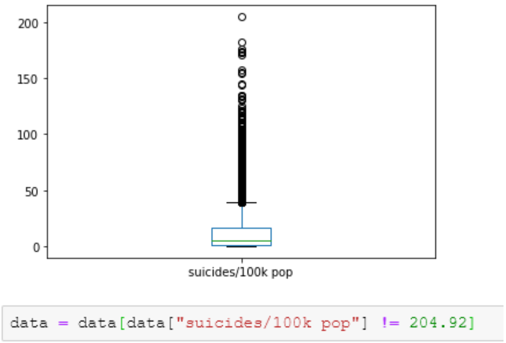
Note that it is important to make sure it is the best way to divide the numerical values into different ranges. We could use the “.describe()” function to see how the data is distributed and further on decide what range the label be able to cover.

We will test our model using predicted accuracy and Receiver Operating Characteristics (ROC) scores.

We encoded a function “multiclass” in order to be able to retrieve ROC scores. If this step was not brought out, the function “metrics.roc_auc_score” would throw an error. [2]

Dropping an outlier

It is very important to evaluate the outliers and decide upon whether to drop or not because outliers may be very important due to different usages of the program. In our case, we just wish to predict suicide rates as accurate as possible therefore we should drop one outlier that is significantly affecting our data.



```
data = data[data["suicides/100k pop"] != 204.92]
```

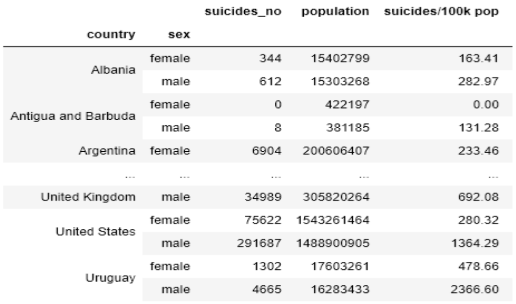
Figure 8: Dropping one outlier that is affecting our dataset

III. Simple Data Visualization

Data visualization was performed in order to have an idea about how the features may affect the target. It is important to understand that visualization should not be a decision maker which also means that the outcome does not represent anything. It is performed just to be able to understand how the data interact with each other.

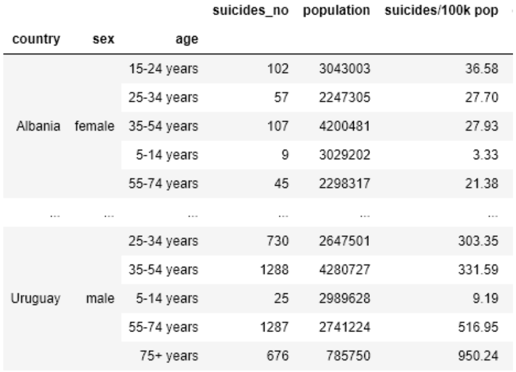
Using groupby() functions to see numerical summarizations

In our program we could expand and extract different groupby objects through code, but for a brief impression we would just look briefly at the first and last rows of data in our dataset.



		suicides_no	population	suicides/100k pop
country	sex			
Albania	female	344	15402799	163.41
	male	612	15303268	282.97
Antigua and Barbuda	female	0	422197	0.00
	male	8	381185	131.28
Argentina	female	6904	200606407	233.46
...	...	...	...	...
United Kingdom	male	34989	305820264	692.08
United States	female	75622	1543261464	280.32
	male	291687	1488900905	1364.29
Uruguay	female	1302	17603261	478.66
	male	4665	16283433	2366.60

Figure 9: Groupby on country, sex



			suicides_no	population	suicides/100k pop
country	sex	age			
Albania	female	15-24 years	102	3043003	36.58
		25-34 years	57	2247305	27.70
		35-54 years	107	4200481	27.93
		5-14 years	9	3029202	3.33
		55-74 years	45	2298317	21.38
...	...	...	...	...	...
Uruguay	male	25-34 years	730	2647501	303.35
		35-54 years	1288	4280727	331.59
		5-14 years	25	2989628	9.19
		55-74 years	1287	2741224	516.95
		75+ years	676	785750	950.24

Figure 10: Group by on country, sex, age

Using plot functions to have better interpretatin on data

This is a better and more advanced data visualization representation than the group by values. Plotting graphs enable us to be able to interpret and understand the relationships better.

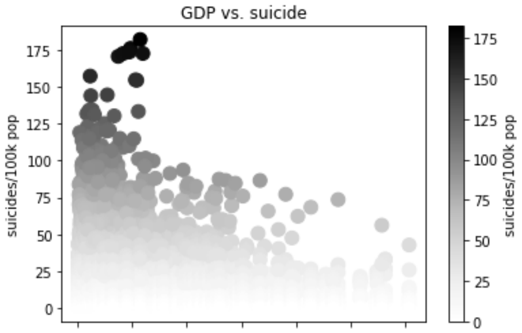


Figure 11: relation between GDP and suicide rates

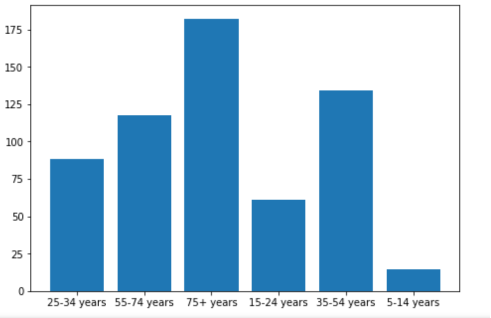


Figure 12: Relation between age and suicide rates globally

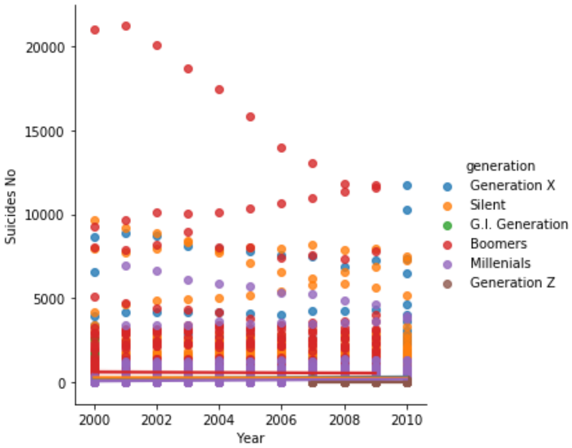


Figure 13: Different amount of generations that committed suicide through years

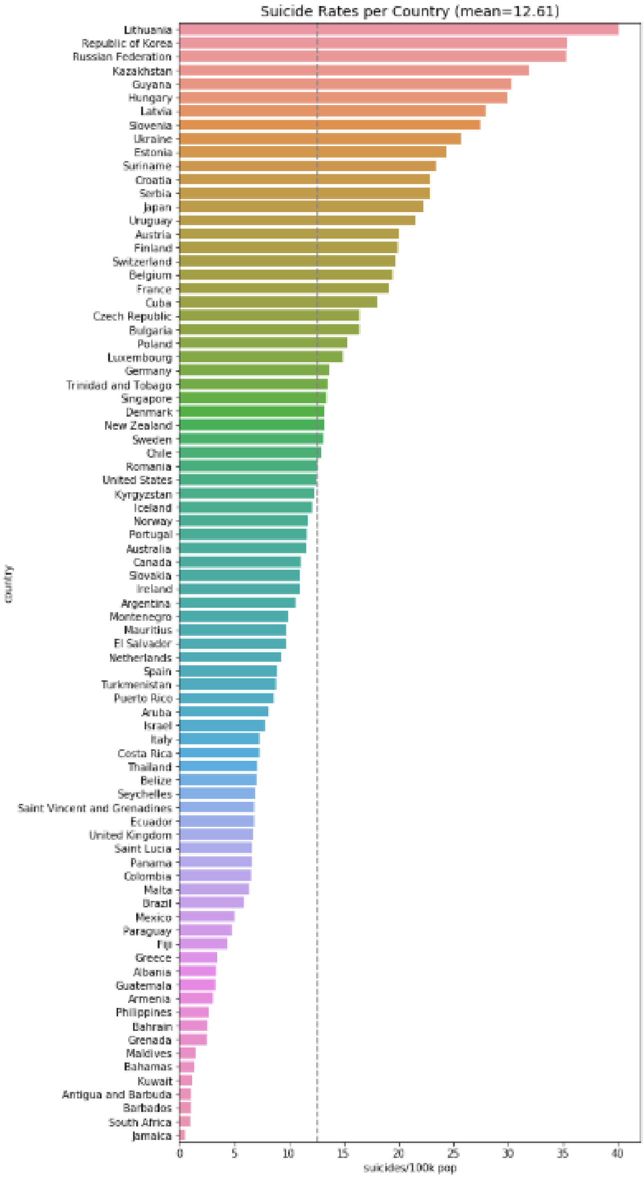


Figure 14: Countries with highest suicide rates from top to bottom

IV. Results

In our program, we implemented functions in order to perform the following classification methods and calculated each accuracy score and roc score relatively.

Classification Models

- Naïve Bayes: accuracy score: 0.46 ROC score: 0.61
- SVM: accuracy score: 0.64 ROC score: 0.74
- KNN:

  (k = 2) accuracy score: 0.59 ROC score: 0.69

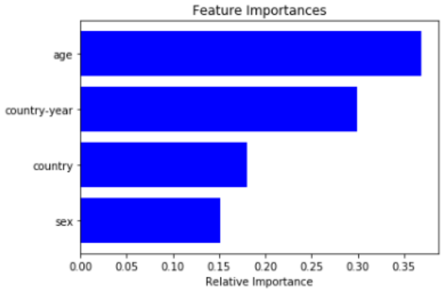
  (k = 3) accuracy score: 0.62 ROC score: 0.72

  (k = 4) accuracy score: 0.70 ROC score: 0.77

  (k = 5) accuracy score: 0.66 ROC score: 0.74

- Random Forest: accuracy score: 0.75 ROC score: 0.81 10-fold cross validation score: 0.52

We also obtained the Feature Importances to show how important the feature is to the response



-

Figure 15: Relative Importance

Model Analyzation

The highest score we could obtain from the above models would be using the Random Forest method. Although we also tried performing 10-fold cross validation on the Random Forest model, but it significantly reduced our accuracy for about 20% when the number of folds was 10.

Naïve Bayes is highly not recommended in our case because the features in our data set are uninevitablely related to each other because the globe is a whole interacting interface.

SVM in our case does not take a lot of computation time since we do not have enormous amount of data and also has a decent performance. But is not as effective due to our dataset only consisting of 4 dimensions while SVM works better on high dimension data.

The KNN model has a nice performance with the ROC score and may also be calculated fast. It reaches peak performance when K = 4.

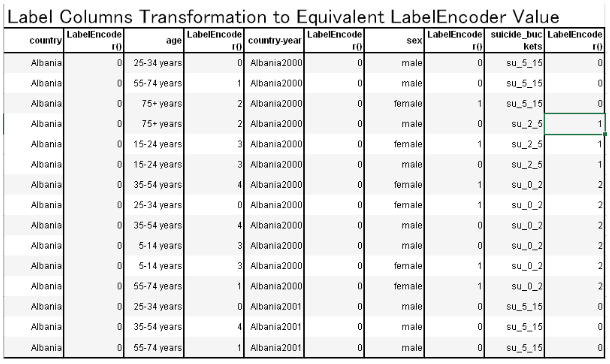
In conclusion, I would consider the Random Forest model to be the best model, it doesn't take super long calculation time and is also accurate and precise with many different categorical data. This algorithm decides which feature is important and weighs more on it. It is important to acknowledge that Random Forest models tend to underfit data, in our case which is extremely critical, since underfitting data may cause users to oversee or neglect to be able to detect/prevent a critical feature that might potentially cause suicide rates to increase.

Testing

After agreeing on a Random Forest would be our final best model, we would like to perform some tests specifically on our Random Forest model by manually inserting already known features to see how accurate the model is.

An inconvenient part about the program is that recall how we transformed categorical data to numerical data by using the LabelEncoder() function, therefore all of our inputted to the simulation was encoded to numerical values.

We collected the first 16 rows of the dataset and compared on how it would look after LabelEncoder().

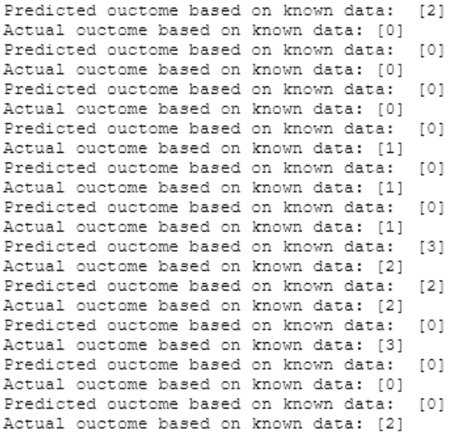


Label Columns Transformation to Equivalent LabelEncoder Value

country	LabelEncoder()	age	LabelEncoder()	country-year	LabelEncoder()	sex	LabelEncoder()	suicide_buckets	LabelEncoder()
Albania	0	25-34 years	0	Albania2000	0	male	0	su_5_15	0
Albania	0	55-74 years	1	Albania2000	0	male	0	su_5_15	0
Albania	0	75+ years	2	Albania2000	0	female	1	su_5_15	0
Albania	0	75+ years	2	Albania2000	0	male	0	su_2_5	1
Albania	0	15-24 years	3	Albania2000	0	female	1	su_2_5	1
Albania	0	15-24 years	3	Albania2000	0	male	0	su_2_5	1
Albania	0	35-54 years	4	Albania2000	0	female	1	su_0_2	2
Albania	0	25-34 years	0	Albania2000	0	female	1	su_0_2	2
Albania	0	35-54 years	4	Albania2000	0	male	0	su_0_2	2
Albania	0	5-14 years	3	Albania2000	0	male	0	su_0_2	2
Albania	0	5-14 years	3	Albania2000	0	female	1	su_0_2	2
Albania	0	55-74 years	1	Albania2000	0	female	1	su_0_2	2
Albania	0	25-34 years	0	Albania2001	0	male	0	su_5_15	0
Albania	0	35-54 years	4	Albania2001	0	male	0	su_5_15	0
Albania	0	55-74 years	1	Albania2001	0	male	0	su_5_15	0

Figure 16: Labels to Numbers

Finally, we could input features to see the outcome.



```
Predicted ouctome based on known data:  [2]
Actual ouctome based on known data: [0]
Predicted ouctome based on known data:  [0]
Actual ouctome based on known data: [0]
Predicted ouctome based on known data:  [0]
Actual ouctome based on known data: [0]
Predicted ouctome based on known data:  [0]
Actual ouctome based on known data: [1]
Predicted ouctome based on known data:  [0]
Actual ouctome based on known data: [1]
Predicted ouctome based on known data:  [0]
Actual ouctome based on known data: [1]
Predicted ouctome based on known data:  [3]
Actual ouctome based on known data: [2]
Predicted ouctome based on known data:  [2]
Actual ouctome based on known data: [2]
Predicted ouctome based on known data:  [0]
Actual ouctome based on known data: [3]
Predicted ouctome based on known data:  [0]
Actual ouctome based on known data: [0]
Predicted ouctome based on known data:  [0]
Actual ouctome based on known data: [2]
```

Figure 17: snippet of code output to proof that data was not made up

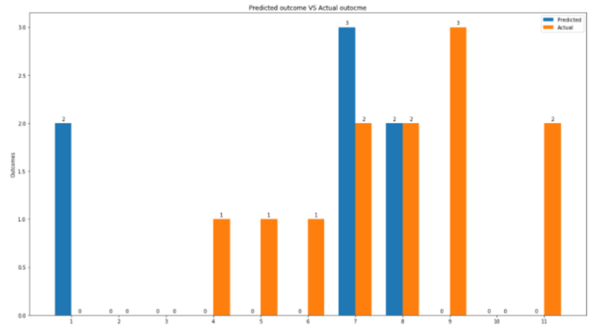


Figure 18: Bar graph for results of tests

The blue columns represent the predicted value meanwhile the orange one is the actual value.

As we could see, that the performance is not as accurate as we expected (accuracy score of 75%), we believe that it may be due to the limited amounts of test we ran. From the graph we could see that the predicted value somehow trends with the actual value, so it was considered as a successful model.

V. DISCUSSION

This model is dedicated to raise awareness in suicide rates across all kinds of factors. There may be many underlying factors that result in increasing suicide rates and we hope to use this model to predict and learn about what is are the factors causing it to increase and decrease. There are a few interesting problems and topics we came across while implementing and designing the model.

The problem of whether this model should be a user-friendly, where casual users could be able to easily get an experience on using this model, resulting in more inaccurate predictions?

Or should the aim for being used on analysis fields, using this model be as professional as possible in order to come up with the most accurate and wanted results

Depending on which goal we would like to achieve the program would be implemented completely differently

This model may be beneficial for various kinds of awareness groups and especially for all people around the world. It is important that we value this topic our goal should be constructing a friendly environment to prevent tragedies.

VI. REFERENCES

[1] https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS

[2] https://www.medcalc.org/manual/roc-curves.php

[3] https://towardsdatascience.com/choosing-the-right-encoding-method-label-vs-onehot-encoder-a4434493149b

[4] https://www.who.int/news-room/fact-sheets/detail/suicide

[5] https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

This report is completely written and programmed by ChingWei Hsieh

Thank you for reading