SUICIDE RATES ANALYSIS AND PREDICTIONS

Hasmitha Bhutham (G01205552)

December 9th, 2020.

Abstract

Suicide has become one of the most serious public health problems in the world and is one of the top ten causes of death in the United States and no progress has been seen with respect to this problem. Every suicide is a loss which affects families, communities and may cause several mental health issues to people close to the person. To tackle this problem, we need data on how, why and when these suicides happen. The data for this project has been used from two informative datasets, one with the data of suicide rates all over the world and the other one with just the suicide rates of India. This data is used for exploratory data analysis, to perform predictions, analyze correlations, understand trends etc. Several python libraries are used to perform the said objectives and various models are implemented to test which model has the least error. Also, sentiment analysis has been done on tweets to understand what kind of vocabulary people who're suffering from suicidal tendencies tend to use if they're about to commit suicide. This assists us by taking advantage of the social networking sites and their extensive usage. The main motivation of this project is to understand how various factors affect suicide rates and analyze tweet data to address patterns and trends related to suicide because this kind of experiments would be of assistance for governments or communities to perhaps develop strategies that might help prevent some of them.

1. Introduction

Suicide, the act of killing oneself, is one of most horrific moves but still a common occurrence. There could be a number of reasons one might choose this way of ending things. No matter what our experience with anything related to suicide, it is important to understand the science behind suicidal behavior. And for that, we need data. So why care about data? The data collected would be very useful to find trends, patterns and can give insights into what is causing people to end their lives by themselves. Recent reports from CDC claims that suicide rates are on the rise in young Americans around the ages 15-24. Hence, these few experiments can help us make our first step into the problem and raise awareness. This project is developed for anyone who wants to have a better understanding on the information regarding suicides. The results we get from this analysis could be used to work on multiple ways to address the problem. All these demonstrations are done with the help of numerous python libraries and classification algorithms. A comprehensive research has been done with respect to the global suicide information as well as the technical aspects of implementation. All the coming chapters are written meticulously to give the reader an ordered interpretation.

2. Problem Statement

To have a better understanding of these rates which are recorded to be progressing from 1985 to 2016, we need new technological approaches. Hence, we use different Data Mining techniques to address the problems. The project aims to predict the suicide rates for 2021 using Support Vector Regressor, Linear Regression, Random Forest Regression and to perform EDA and find correlations to have better demonstration of the statistics which provide useful information. The model with the least error is used to perform predictions.

Sentiment Analysis on the tweet data is done to have a clear picture of what tweets have the most tendency to be related to suicide or depression. The sentences or phrases are classified to be positive or negative, positive if it is related to suicide or depression.

2.1 Notations

- EDA: Exploratory Data Analysis.
- SVR: Support Vector Regressor.
- RMSE: Root Mean Squared Error.
- SGDClassifier: Stochastic Gradient Descent Classifier.
- NLP: Natural Language Processing.
- NLTK: Natural Language Toolkit.

3. Literature Review

The research done regarding this problem resulted in finding out a number of papers with studies done on suicides and its factors. One of the papers mentions about the correlations of suicides and homicides in Japan and the United States from 1953-1982. The paper discussed about how the problems like unemployment, divorce rates etc., are correlated with violence and suicides. The results for both the countries significantly varied. Another paper discussed about how the technological advancements are related to suicide. The paper showed a survey of middle school children which demonstrated that children who encountered bullying or cyberbullying are more likely to attempt suicide. A paper published by the Cambridge University Press discussed about the relationship between increasing age and the suicide rates using multinational data from the World Health Organization. The results gave out many correlations between age, suicide rates and how the rates differ for different age-groups and how varied the results are for different countries. A Brazilian paper talked about how socioeconomic factors affect the suicide rates in adolescents. The results showed that social inequality is positively associated with the suicide rates in adolescents. This has been assessed using the Gini Index values. The last paper I found is about the socioeconomic correlations in Taiwan which showed that economic statistics seem to have a bigger influence on suicide rates than social factors. This study is done with the statistics from 1983-1993. It was also found out that the suicide rates decreased as the per capita GDP increased.

Other works done with the same dataset as this project only focused on analyzing the WEKA tool has been used. Some of them tried to understand correlations between the GDP of a country and the suicide rates. Other projects were implemented in R language to analyze the data.

4. Methods and Techniques

- **Exploratory Data Analysis:** EDA is an approach to analyze the data and to make sense of it. EDA has been performed on the data to get proper insights into the rates. Several python libraries such as NumPy, seaborn, matplotlib etc. have been used. Various plots have been shown to better visualize the data. Correlations are shown with the help of heatmaps and scatterplots.
  All the graphs and plots visualized are shown in section 5.
- **Feature Engineering:** The columns of dataset are categorized into numeric and categorical values and categorical values are encoded to fit into the model. To do this, sklearn's OneHotEncoder, ColumnTransformer have been used. In **OneHotEncoder**, each categorical value is converted to a column and is given a binary notation (0 or 1). **ColumnTransformer** allows different columns to be transformed independently and the results of each transformer are concatenated finally into a single feature. Sklearn's **pipeline** has been used to integrate different steps to cross-validate then together.
  After this feature engineering, the features are fit into the models.
- **Model Building:** SVR, Random Forest Regressor, Linear Regression are implemented and are evaluated using RMSE scores. The model with the least RMSE error is used to make predictions.

SVR: Support vector Regressor model uses regression in SVMs which gives us the flexibility of choosing which error rate is acceptable. It minimizes the coefficients with the acceptable error rate or we can also tune it to satisfy our needs.

Random Forest Regression: It is a supervised learning algorithm which uses ensemble learning method for classification and regression. It is a bagging technique in which the 'random forests' are run in parallel independently i.e., there is no interaction between the trees while building them. A random forest regressor is considered to be a meta estimator i.e., it aggregates many trees in which each tree picks random samples from the dataset which prevents overfitting. Also, the number of features which are to be split at each node are limited which prevents too much reliability of independent features.

Linear Regression: Linear Regression is one of the oldest and most used supervised learning algorithms for predictive analysis. It is simply a method to predict the dependent variable by fitting the best possible linear relationship with the independent variables. This fit is done with the help of Least Squares Method which makes sure that the distance between the observations and shape is as little as possible.

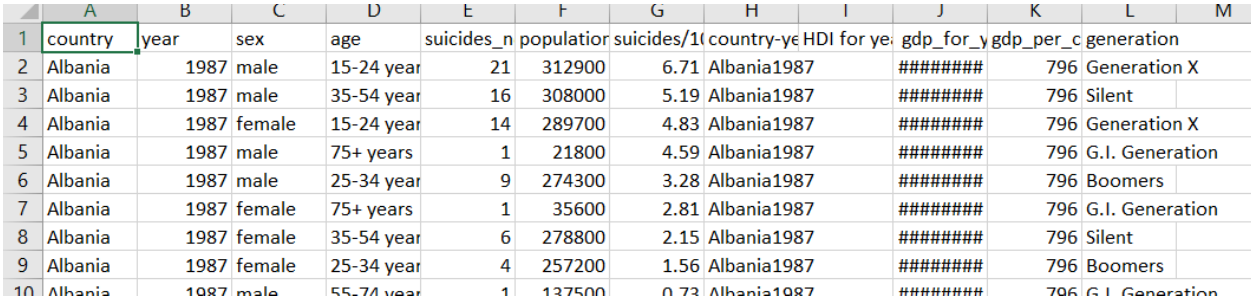
- **2021 Predictions:** Here, predictions are performed on the data of United States, but we can input any country's data and get the prediction for that country. Some of the data required for predicting is taken from 2019 samples and some data is taken from the projected data available on the internet.
- **Sentiment Analysis:** Sentiment Analysis has been done on the tweet data which has been downloaded in the form of a csv file from a website from which about 2000 recent tweets can be extracted by using hashtags or keywords. Preprocessing is done by removing special characters and turning the text into lowecase. PorterStemmer is used to stem the words for better analyzations. Tokenizing and Vectorizing is done to fit the data into the model. SGDClassifier has been used to predict if a certain text is positive or negative. From genism, Word2Vec is used to vectorize the words.

  SGDClassifier: It is an approach to fit classifiers to sparse problems which are often encountered in text classification and NLP.

5. Discussion and Results

5.1 Datasets

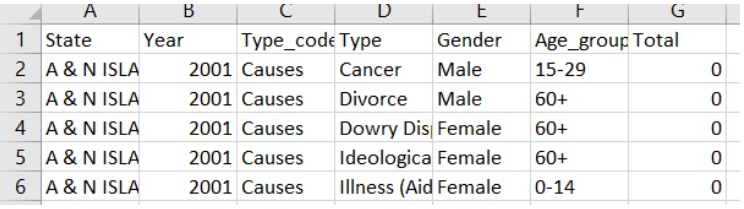
The first dataset has been taken from Kaggle. (https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016). It has 27,821 entries with 12 columns. The snapshot of the dataset is shown below:



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	country	year	sex	age	suicides_n	populatior	suicides/1(	country-ye	HDI for ye:	gdp_for_y	gdp_per_c	generation	
2	Albania	1987	male	15-24 year	21	312900	6.71	Albania1987		########	796	Generation X	
3	Albania	1987	male	35-54 year	16	308000	5.19	Albania1987		########	796	Silent	
4	Albania	1987	female	15-24 year	14	289700	4.83	Albania1987		########	796	Generation X	
5	Albania	1987	male	75+ years	1	21800	4.59	Albania1987		########	796	G.I. Generation	
6	Albania	1987	male	25-34 year	9	274300	3.28	Albania1987		########	796	Boomers	
7	Albania	1987	female	75+ years	1	35600	2.81	Albania1987		########	796	G.I. Generation	
8	Albania	1987	female	35-54 year	6	278800	2.15	Albania1987		########	796	Silent	
9	Albania	1987	female	25-34 year	4	257200	1.56	Albania1987		########	796	Boomers	
10	Albania	1987	male	55-74 year	1	137500	0.73	Albania1987		########	796	G.I. Generation	

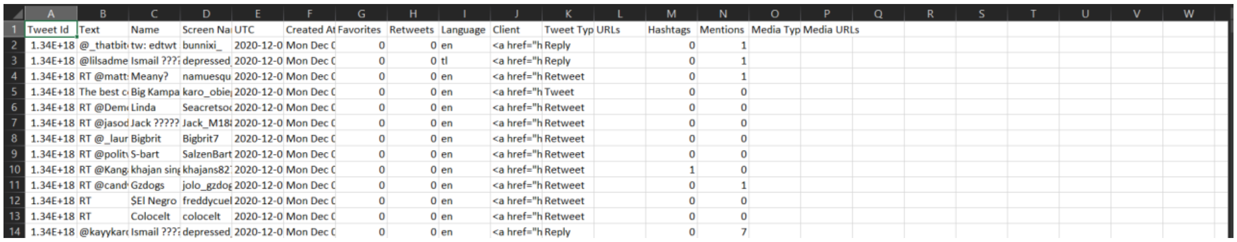
The independent variables are country, year, sex, age, suicides_no, population, country-year, HDI for year, gdp_for_year, gdp_per_capita and generation and the dependent/predictable variable is suicides/100k population.

The second dataset has been taken from data.world. (https://data.world/rajanand/suicides-in-india/workspace/file?filename=Suicides_in_India.csv). It has 237530 entries. The snapshot of the data is shown below:



	A	B	C	D	E	F	G
1	State	Year	Type_code	Type	Gender	Age_group	Total
2	A & N ISLA	2001	Causes	Cancer	Male	15-29	0
3	A & N ISLA	2001	Causes	Divorce	Male	60+	0
4	A & N ISLA	2001	Causes	Dowry Dis	Female	60+	0
5	A & N ISLA	2001	Causes	Ideologica	Female	60+	0
6	A & N ISLA	2001	Causes	Illness (Aid	Female	0-14	0

The data for sentiment analysis has been taken from (https://www.vicinitas.io/free-tools/download-search-tweets) and downloaded into a csv file. This is done by inserting hashtags and keywords related to suicide in the website.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	Tweet Id	Text	Name	Screen Na	UTC	Created At	Favorites	Retweets	Language	Client	Tweet Typ	URLs	Hashtags	Mentions	Media Typ	Media URLs							
2	1.34E+18	@_thatbit	tw: edtwt	bunnixi_	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Reply		0	1									
3	1.34E+18	@lilsadme	Ismail ????	depressed	2020-12-0	Mon Dec 0	0	0	tl	<a href="h	Reply		0	1									
4	1.34E+18	RT @matt	Meany?	namuesqu	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Retweet		0	1									
5	1.34E+18	The best c	Big Kampa	karo_obie	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Tweet		0	0									
6	1.34E+18	RT @Dem	Linda	Seacretso	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Retweet		0	0									
7	1.34E+18	RT @jasod	Jack ?????	Jack_M18	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Retweet		0	0									
8	1.34E+18	RT @_laur	Bigbrit	Bigbrit7	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Retweet		0	0									
9	1.34E+18	RT @polit	S-bart	SalzenBart	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Retweet		0	0									
10	1.34E+18	RT @Kang	khajan sing	khajans82	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Retweet		1	0									
11	1.34E+18	RT @cand	Gzdogs	jolo_gzdog	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Retweet		0	1									
12	1.34E+18	RT	$El Negro	freddycuel	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Retweet		0	0									
13	1.34E+18	RT	Colocelt	colocelt	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Retweet		0	0									
14	1.34E+18	@kayykar	Ismail ????	depressed	2020-12-0	Mon Dec 0	0	0	en	<a href="h	Reply		0	7									

5.2 Evaluation Metrics

The Evaluation Metrics used are the RMSE scores. It is one of the most common metrics used to measure the accuracy for continuous variables. The quadratic equation can be given as:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

RMSE has been used instead of MAE because RMSE gives a relatively high weight to large errors i.e., RMSE should be more useful when large errors are undesirable. The errors attained for all the models mentioned in section 4 are as follows:

```
SVM RMS Value: 15.48396759687091
Linear Regression RMS Value: 12.358634913567636
Random Forest RMS Value: 5.766325720614945
```

Hence, it is evident that Random Forest has the least error when compared to SVM Regression and Linear Regression.

5.3 Experimental Results

EDA: The EDA performed provided us with the following visualizations.

Figure 1: This plot shows the suicides per 100k population from 1985 to 2016. It can be observed that 1994 has the highest rates.

```
Out[145]: <matplotlib.axes._subplots.AxesSubplot at 0x220bcb874c8>
```

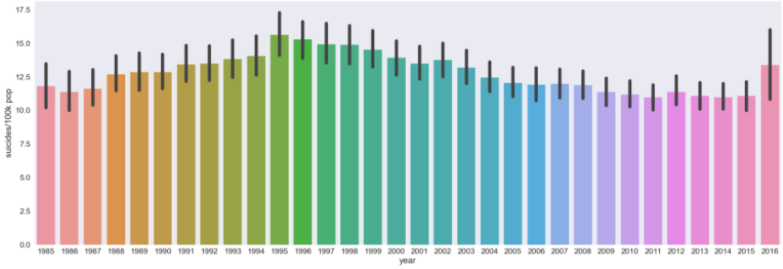
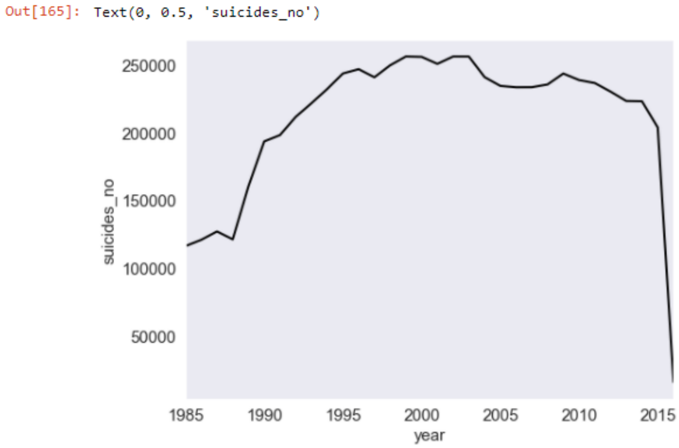
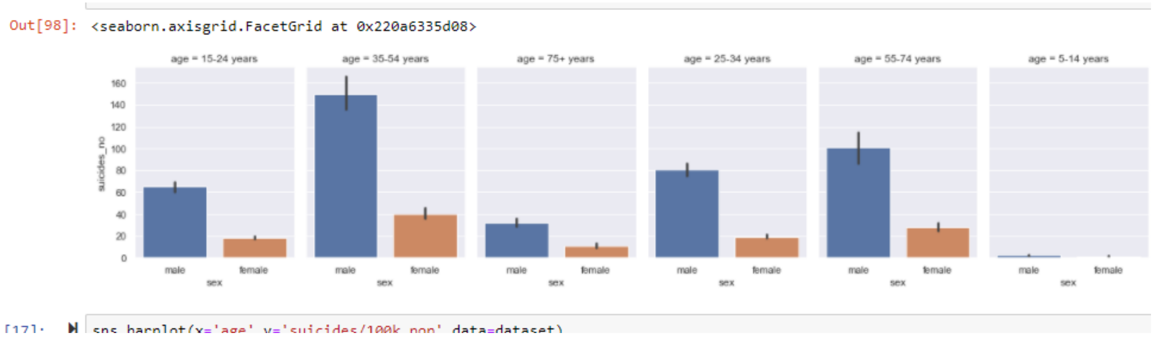


Figure 2: It shows that the suicides have been decreasing over the years.



```
Out[165]: Text(0, 0.5, 'suicides_no')
```

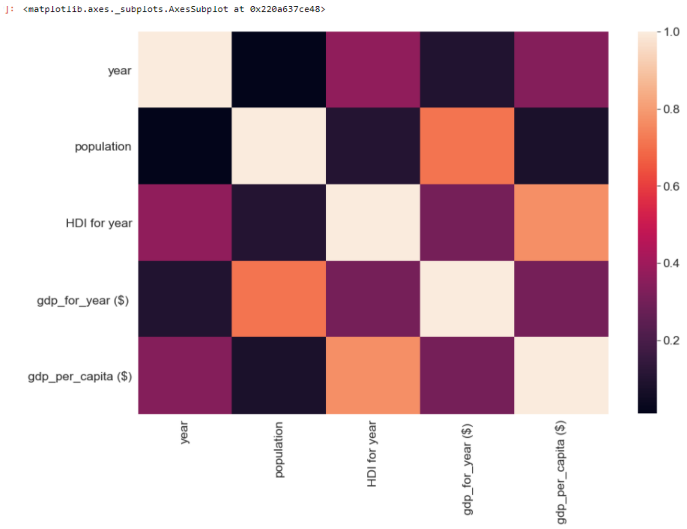
Figure 3: It shows males vs females for different age groups. It can be observed that males tend to commit suicide more than females even around the age of 15.



```
Out[98]: <seaborn.axisgrid.FacetGrid at 0x220a6335d08>
```

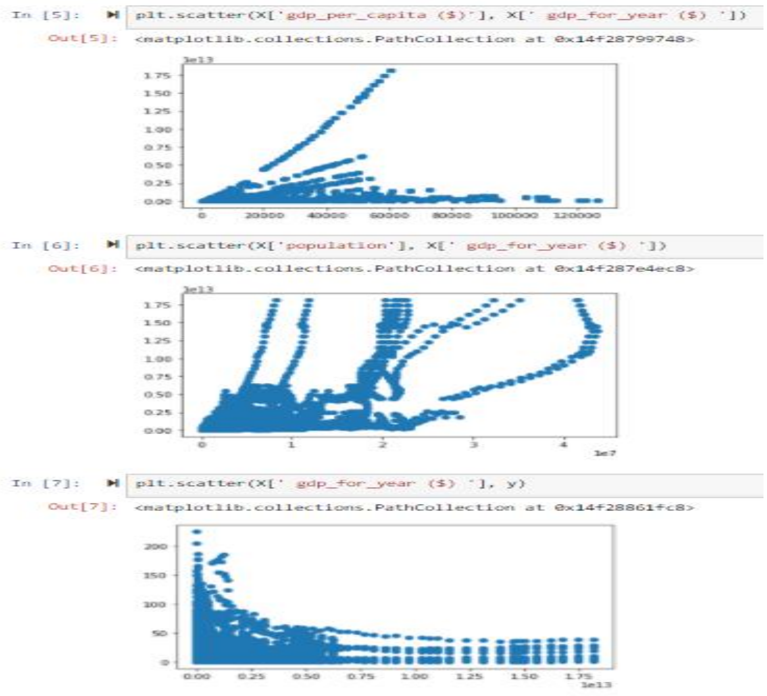
```
[17]:  sns.barplot(x='age', y='suicides/100k pop', data=dataset)
```

Figure 4: The heatmap shows the correlations between different attributes with each other. It can be observed that population & gdp_for_year have higher effect on each other.



```
]: <matplotlib.axes._subplots.AxesSubplot at 0x220a637ce48>
```

Figure 5: We can also observe scatter plots to address the correlations.



```
In [5]:   plt.scatter(X['gdp_per_capita ($)'], X[' gdp_for_year ($) '])
 Out[5]:  <matplotlib.collections.PathCollection at 0x14f28799748>
```

```
In [6]:   plt.scatter(X['population'], X[' gdp_for_year ($) '])
 Out[6]:  <matplotlib.collections.PathCollection at 0x14f287e4ec8>
```

```
In [7]:   plt.scatter(X[' gdp_for_year ($) '], y)
 Out[7]:  <matplotlib.collections.PathCollection at 0x14f28861fc8>
```

Figure 6: It can be observed from this bar plot that boomers (75+ years) have more suicide rates.

```
<matplotlib.axes._subplots.AxesSubplot at 0x220a16b49c8>
```

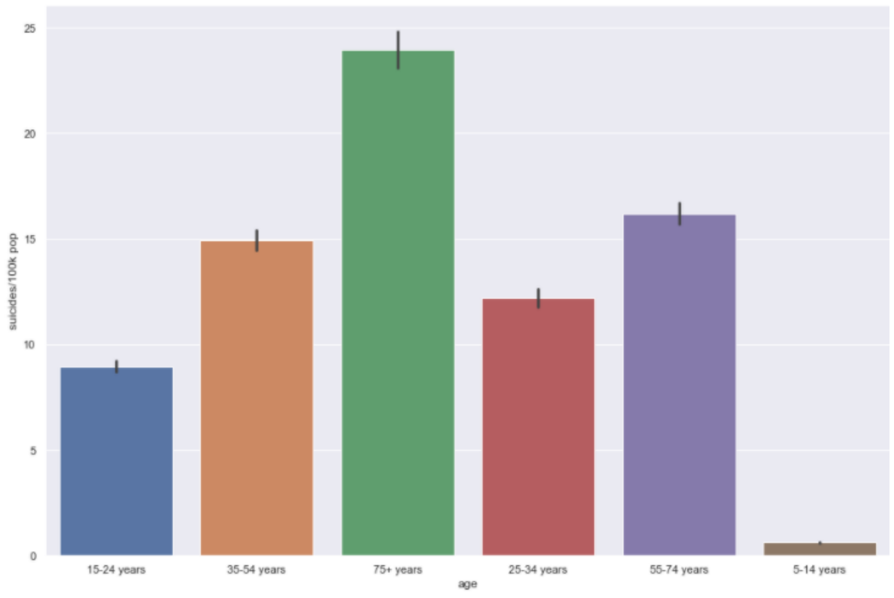
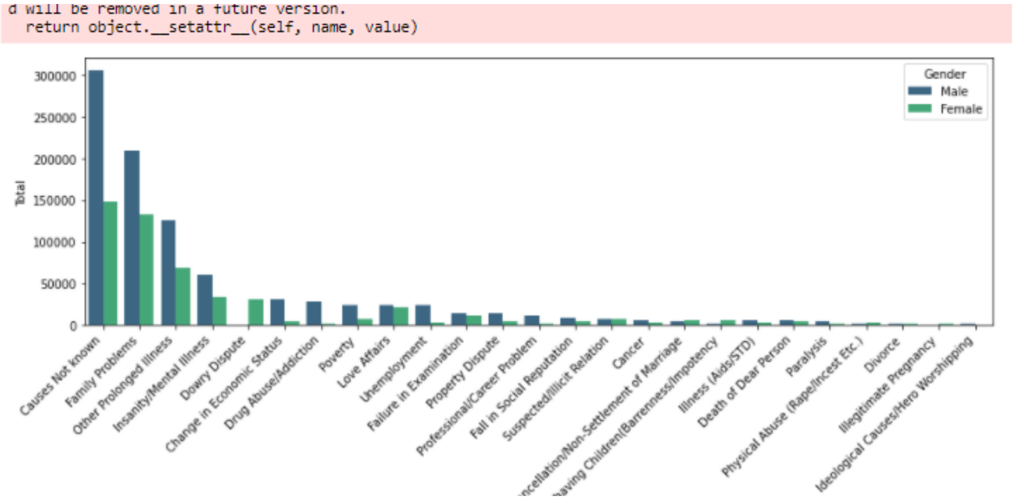


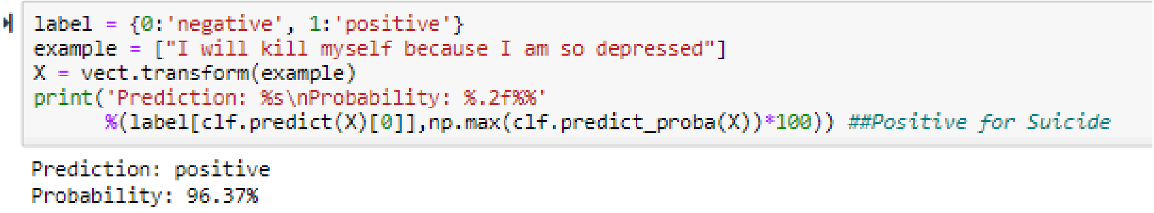
Figure 7: This bar graph is made from the second data set and it shows that the highest type “causes not known”, if we can ignore that and look at other types, the highest value is for “family problems” followed by prolonged and mental illnesses.



```
d will be removed in a future version.
  return object.__setattr__(self, name, value)
```

Predictions: The suicides per 100k population of the year 2021 for United States for males from age 15-24 are predicted as `2021 Prediciton for Males under 15-24 years: 10.642999999999997` Similarly, the suicides per 100k population of the year 2021 for United States for females from age 15-24 are predicted as `2021 Prediciton for Females under 15-24 years: 3.8269999999999995`

Sentiment Analysis: The accuracy for the model is 0.84. Prediction is done by giving an example sentence as input and the output gives if the given sentence is negative or positive (positive is for suicide). It also shows the probability of the prediction.



```
label = {0:'negative', 1:'positive'}
example = ["I will kill myself because I am so depressed"]
X = vect.transform(example)
print('Prediction: %s\nProbability: %.2f%%'
      %(label[clf.predict(X)[0]],np.max(clf.predict_proba(X))*100)) ##Positive for Suicide
```

```
Prediction: positive
Probability: 96.37%
```

6. Conclusion

To conclude, the steps I’ve went through include, Exploratory Data Analysis, finding correlations, Feature Engineering, predictions for 2021 after performing RMSE metrics on SVR, Random Forest and Linear Regression. Since the error for Random Forest turned out to be the least, the model has been used for predictions. These experiments answered a lot of questions related to the dataset and the predictions showed that males have a higher risk of suicide compared to females. The analysis also showed that boomers i.e., people above the age of 75 years have a higher rate of suicides. My own perception about this information after reading a number of articles is that people above that age may either face some illness which might lead to them exhausting their savings. This leads to money problems where sometimes people tend to not even have money to buy proper food and have shelter.

As for sentiment analysis, SGDClassifier has been used. Number of NLTK libraries have been used for preprocessing and to vectorize the texts to fit into the model. It has been observed that the words “depression” and “suicide” are being used very commonly on social media, sometimes casually, sometimes trying to seek help.

The difficulties I faced was with sentiment analysis where I wanted to SGDClassifier but had no proper background knowledge about it. It took some time to understand it but it was worth it. I have learnt a new method of dealing with text mining and would like to use it for future works.

6.1 Directions for future work

- Additionally, this project can be upgraded to predict the suicide rates with respect to the other attributes and predict the rates of India with the help of the second dataset. Sentiment analysis can be upgraded by using it on different social networking sites, to figure out in which social media, people feel more free to talk about their mental health.
- I would also like to work on different datasets of the same topic, which I encountered while researching for this project. These datasets also seem interesting and can answer specific problems.
- Also, I found out some papers about homicides, particularly serial killers in which I read about behavioral sequence analysis which really triggered my interest. I would like to learn more about it and use it to implement on related data.

References

[1] https://towardsdatascience.com/suicide-in-the-21st-century-part-1-904abe8e1f5c

[2] https://www.rpubs.com/lmurra38/509871

[3] https://www.populationpyramid.net/united-states-of-america/2021/

[4] https://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-44462019000500389

[5] https://www.cambridge.org/core/journals/international-psychogeriatrics/article/abs/relationship-between-suicide-rates-and-age-an-analysis-of-multinational-data-from-the-world-health-organization/8A15082759EDB1FDBD002F4218AB861A

[6] https://datasciences.org/kdd2018-tutorial-behavior-analytics-methods-and-applications/

[7] https://journals.sa gepub.com/doi/abs/10.1177/0886260518759655

[8] https://github.com/ParmenidesSartre/Suicide-Rates-Overview-1985-to-2016/blob/master/Suicide%20Rate.ipynb

[9] https://medium.com/yernagula-akanksha/suicide-rates-overview-1985-to-2016-data-analysis-using-orange-gui-f2b83f0272c7

[10] https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

[11] Lester, David, et al. “The Impact of the Economy on Suicide and Homicide Rates in Japan and the United States.” International Journal of Social Psychiatry, vol. 38, no. 4, Dec. 1992.

[12] FocusEconomics. “The Poorest Countries in the World (2019-2023).” FocusEconomics | Economic Forecasts from the World's Leading Economists,