Prevention Worth a Pound of a Cure

What can we do differently to reach out to those in need of help?

An Analysis of Suicide Rates From 1985 to 2016

Prepared as the final project for Introduction to Data to Science (University of Waterloo)

Table of Contents

Introduction 3

2.0 Data Preparation *3*

2.1 Importing the Data **3**

2.2 Preparing the Data **5**
2.2.1 Handling Null Data 5
2.2.2 Analyzing Outliers and Incorrect Data 6

2.3 Feature Engineering **7**

3.0 Analysis of Suicides Attributes *8*

3.1 Exploratory Analysis **8**
3.1.1 Sex 8
3.1.2 Country 9
3.1.3 Age/Generation 10
3.1.4 GDP 10

4.0 Driving Factors of Suicide Based on Exploratory Analysis *10*

4.1 Data Analysis of Suicide Rates by Gender and Age **10**
4.1.1 Analysis of Suicide Rates by Gender 10
4.1.2 Analysis of Suicide Rate by Age 12

4.2 Outlier Analysis of Suicide Rates **12**

4.3 Data Analysis of Suicide Rates by GDP per Capita **13**

5.0 A Machine Learning Approach to Analyze and Predict Suicide Rate *14*

5.1 Feature importance **15**
5.1.1 Decision tree regression 15

5.1.2 Random Forest **15**

5.2 Performance of Random Forest Regression Model **16**

5.3 Performance of Linear Regression Model **18**

5.4 Performance of Model - Decision Tree Regression **19**

5.5 Comparison of Results by Different Algorithms **20**

Conclusions *20*

References *21*

Introduction

Suicide is a severe health and social problem whose incidence varies between genders, age groups, geography, and social structure, among many other variables. The development of a global strategy to target this issue must identify the descriptive characteristics of at-risk groups and understand how they correlate with increasing or decreasing suicide rates. The objective of our group project is to analyze worldwide suicide rates from 1985 to 2016 to identify such factors. Throughout this investigation, we will aim to identify the population groups that are the most affected and whether suicide rates could be affected by historical events in different regions. In addition to social factors, gross domestic product (GDP) and human development index (HDI) data will be used to investigate economic impacts on the global suicide rate.

2. Data Preparation

2.1 Importing the Data

The source of data for this report is "Suicide Rates Overview from 1985 to 2016," available from Kaggle.com (Rusty, 2021). It was developed to find signals correlated to increased suicide rates among different global cohorts across the socio-economic spectrum. This dataset was pulled from four other datasets linked by time and country. They are:

1. United Nations Development Program "Human Development Index (HDI)" (United Nations, 2021)
2. World Bank "World Development Indicators: GDP by Country: 1985 to 2016" (The World Bank, 12)
3. "Suicide in the Twenty-First Century" (Szamil, 2021)
4. World Health Organization, "Suicide Prevention" (World Health Organization, 2021)

The dataset comprises 12 attributes: country, year, sex, age group, count of suicides, population, suicide rate, country-year key, HDI, GDP, GDP per capita, and generation. By analyzing all features, we will explore what groups, regions, and segments of populations are the most vulnerable. Through these insights, we will achieve the final goal of his paper: increasing suicide prevention.

39]:

	0	1	2	3	4	5	6	7	8	9	...	22	23	24	25	26	27	28	29	30	31
year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	...	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
country	48	48	54	49	52	64	64	65	65	68	...	86	85	89	88	86	81	80	78	62	16

2 rows × 32 columns

Figure 1 - all years for which these are suicide data available

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27820 entries, 0 to 27819
Data columns (total 12 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   country             27820 non-null  object
 1   year                27820 non-null  int64
 2   sex                 27820 non-null  object
 3   age                 27820 non-null  object
 4   suicides_no         27820 non-null  int64
 5   population          27820 non-null  int64
 6   suicides/100k pop   27820 non-null  float64
 7   country-year        27820 non-null  object
 8   HDI for year        8364 non-null   float64
 9    gdp_for_year ($)   27820 non-null  object
 10  gdp_per_capita ($)  27820 non-null  int64
 11  generation          27820 non-null  object
dtypes: float64(2), int64(4), object(6)
memory usage: 2.5+ MB
```

Figure 2 – Suicide Dataset Info Table

Below is a brief description of the data columns:

- Country (Categorical): occupying a particular territory by a group of people
- Year (Numeric): 365 days starting from the first of January, used for reckoning time.
- Sex (Categorical): a trait that determines an individual's reproductive function, male or female.
- Age (Numeric): the length of time that a person has lived or a thing has existed.
- Suicide Number (Numeric): number of people or individuals that commit suicide (absolute value)
- Population (Numeric): number of people that live in a specific territory in a given time.
- Suicide/100K Population (Numeric): number of suicides committed for every 100K individuals
- HDI for Year (Numeric): a composite index measures average achievement in three basic dimensions of human development; long and healthy life, education, and standard of living.
- GDP For Year (Numeric): the total value of goods produced and services provided in a country during one year.
- GDP Per Capital (Numeric): the total value of goods produced and services provided in a territory divided by the number of individuals living in that territory
- Generation (Categorical): Generation classifications groups based on the year's people were born in and historical events in that territory.

	country	year	sex	age	suicides_no	population	suicides_100k	country_year	HDI_for_year	gdp_for_year	gdp_per_capita	generation
0	Albania	1987	male	15-24 years	21	312900	6.71	Albania1987	NaN	2,156,624,900	796	Generation X
1	Albania	1987	male	35-54 years	16	308000	5.19	Albania1987	NaN	2,156,624,900	796	Silent
2	Albania	1987	female	15-24 years	14	289700	4.83	Albania1987	NaN	2,156,624,900	796	Generation X
3	Albania	1987	male	75+ years	1	21800	4.59	Albania1987	NaN	2,156,624,900	796	G.I. Generation
4	Albania	1987	male	25-34 years	9	274300	3.28	Albania1987	NaN	2,156,624,900	796	Boomers

Figure 3 - Suicide rate data sample

2.2 Preparing the Data

Once the importing took place, meaningful data preparation was needed to understand the scope better and reach of the data used for this analysis. We focused our research on those countries with the most available data in our dataset due to a lack of consistency on the years and country data availability.

The following steps have been made to create a dataset that is valuable for the analysis:

- Renaming columns according to the same pattern (low-case, no spaces or special characters)
- Removing one of two columns that values have the same meaning ("generation" vs "age" and "country-year" as a concatenation of "country" and "year"; we can keep "generation" values as a possible Category value in a dictionary)
- Transforming column type according to its meaning (object to num64)

2.2.1 Handling Null Data

It's worth noting that there are some potential limitations with analyzing HDI by year because it is unavailable for most of the data points captured in the dataset, as can be seen in.

HDI will only be included for the countries that own this feature for the analysis. This attribute will be evaluated in isolation to identify the impact of this number in predicting suicides in those countries.

As was mentioned above and can be seen in the figure, the factor "HDI for Year" has only 30%, not NULL values, and because we want to use this factor in the analysis, we should get the data from another source. The best open official resource for HDI values is the United Nations HDI report (United Nations, 2021). This report has the data for 2020. We can use 2020 data in our analysis in 2 ways:

- use HDI data from 2020 only
- use HDI data from our initial dataset for existing values and set the NULL values equal to the 2020 data

Out[65]:

	Nulls
country	0
year	0
sex	0
age	0
suicides_no	0
population	0
suicides/100k pop	0
country-year	0
HDI for year	19456
gdp_for_year ($)	0
generation	0

*Figure SEQ Figure * ARABIC 4 - HDI null values*

We can check if HDI changes drastically in a country during the years. If it changes vastly, the data for one year cannot be applied; however, if the changes were insignificant, the 1-year information is statistically justified.

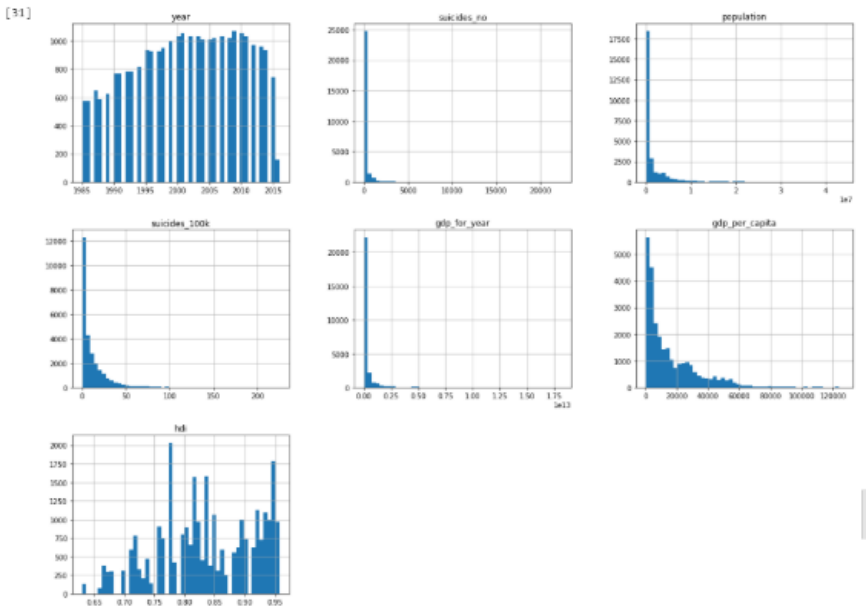
```
1. in provided years, the maximum difference of HDI by country is 0.202
2. in provided years, the average of HDI change is 0.097
```

Figure 5 - HDI max difference vs. average yearly change

As we can see, the HDI did not change in any country with known data by more than 20%. Moreover, the average changes were less than 10%. Therefore, having such low fullness of data in our master dataset, it would be better to use completer and more consistent 2020 data from (United Nations, 2021) in the HDI column instead of existing ones.

2.2.2 Analyzing Outliers and Incorrect Data

To check if there are misprint values in numeric columns, we can review maximum and minimum values using histograms or describe () function as seen in Figure 6, column "suicides no" has a spike value of 22338.000000.



To investigate is a possible value or a misprint, we need to identify other deals from this row - this provides an overall idea of how the numerical variables are distributed and potential errors or outliers' values that can affect the statistical analysis see.

	year	suicides_no	population	suicides_100k	gdp_for_year	gdp_per_capita	hdi
count	27820.000000	27820.000000	2.782000e+04	27820.000000	2.782000e+04	27820.000000	27820.000000
mean	2001.258375	242.574407	1.844794e+06	12.816097	4.455810e+11	16866.464414	0.837229
std	8.469055	902.047917	3.911779e+06	18.961511	1.453610e+12	18887.576472	0.079496
min	1985.000000	0.000000	2.780000e+02	0.000000	4.691962e+07	251.000000	0.630000
25%	1995.000000	3.000000	9.749850e+04	0.920000	8.985353e+09	3447.000000	0.779000
50%	2002.000000	25.000000	4.301500e+05	5.990000	4.811469e+10	9372.000000	0.837246
75%	2008.000000	131.000000	1.486143e+06	16.620000	2.602024e+11	24874.000000	0.916000
max	2016.000000	22338.000000	4.380521e+07	224.970000	1.812071e+13	126352.000000	0.957000

After comparing with other sources, e.g. WHO, we can see that the number of suicides in a specified year and country matches; therefore, the data looks plausible and will be kept for further analysis. This observation is highlighted in Figure 8 – Highest suicide data.

country	Russian Federation
year	1994
sex	male
age	35-54 years
suicides_no	22338
population	19044200
suicides_100k	117.3
gdp_for_year	395077301248
gdp_per_capita	2853
hdi	0.824

2.3 Feature Engineering

This step aimed to add attributes that could bring meaningful insights into the existing data. A "region" variable was introduced to categorize suicide rates by "continent" as well as by country, as it is currently designated. The region data was retrieved from GitHub and contains each country's region, code, sub-regions, country code, the intermediate region, and others (Duncalfe, 2021). Some exceptions were addressed manually due to the lack of consistent names between the suicide dataset and region datasets.

After merging both datasets, the data points available for the research are broken down by years and regions.

Out[60]:

year	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	...	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
region																					
Africa	24.0	24.0	24.0	12.0	12.0	12.0	12.0	12.0	12.0	12.0	...	36.0	36.0	36.0	36.0	48.0	36.0	36.0	36.0	36.0	10.0
Americas	264.0	264.0	276.0	252.0	276.0	300.0	276.0	252.0	264.0	264.0	...	312.0	324.0	336.0	336.0	324.0	312.0	312.0	264.0	204.0	10.0
Asia	96.0	84.0	96.0	72.0	60.0	144.0	144.0	156.0	144.0	168.0	...	216.0	192.0	228.0	240.0	216.0	192.0	192.0	204.0	156.0	50.0
Europe	168.0	180.0	228.0	228.0	252.0	288.0	300.0	324.0	324.0	336.0	...	432.0	432.0	432.0	420.0	408.0	396.0	396.0	420.0	336.0	90.0
Oceania	24.0	24.0	24.0	24.0	24.0	24.0	36.0	36.0	36.0	36.0	...	36.0	36.0	36.0	24.0	36.0	36.0	24.0	12.0	12.0	NaN

5 rows × 32 columns

Figure 9 - Suicide Date Per Continent

Regarding other attributes included in the merge, only those considered meaningful features within the datasets were kept. Therefore, only the region column was merged; other features were unnecessary for this research.

3. Analysis of Suicides Attributes

To gain a firm grasp of the attributes present in the dataset, it's ideal to look at the basic description, and the impact each has in predicting the number of suicides. The first step is to classify all attributes according to how they will be used in the model. Inputs: Country, Region, Year, Sex, Population, HDI for Year, GDP Per Capital, GDP For Year, Generation. Outputs: Suicide Number, Suicide/100K Population

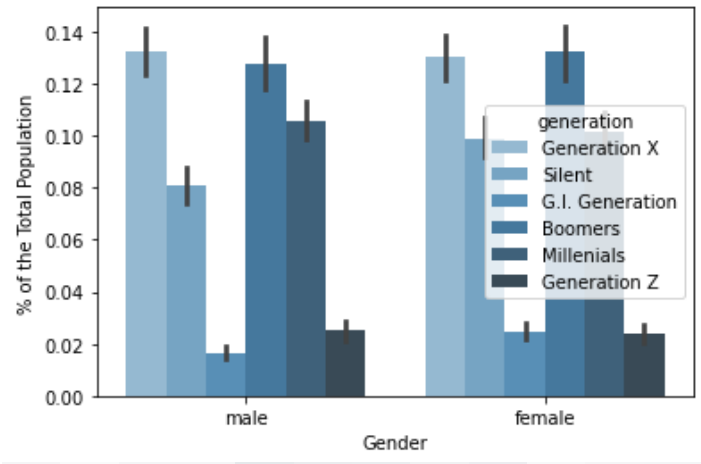
3.1 Exploratory Analysis

To understand the data better, the exploratory analysis will provide some basic details of the dataset, the size of the population, numbers of years, countries, rates and other information. All these factors are vital in having a diversified dataset, which will help build a more robust analysis and more accurate predictive model.

3.1.1 Sex

From a quick peek at the gender data across all the countries, there is a transparent, balanced distribution of the people captured in the dataset. As you can see in Figure 10 - Gender Makeup of Population across different age groups, the data is almost evenly spread gender by gender in all the age groups, from Generation X to Generation Z. The total distribution of the data points is 51.19% female and 48.80 %, male

therefore, we can conclude that Sex attributes in the (Suicide Rates Overview from 1985 to 2016) are a helpful feature that can be used for the analysis.



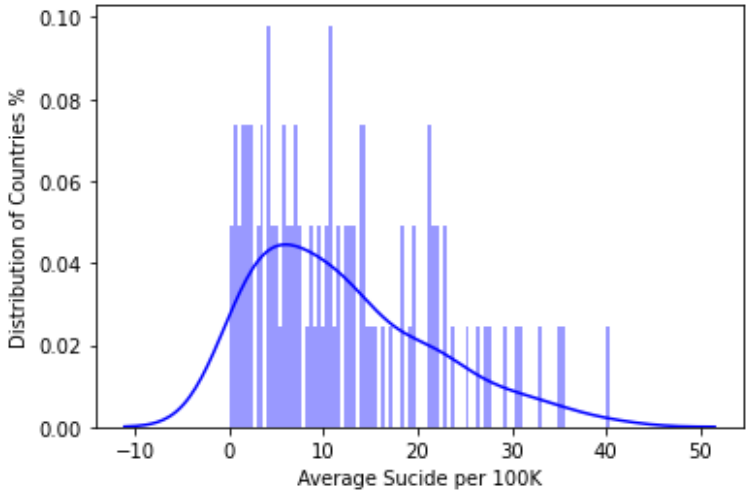
3.1.2 Country

To provide a micro and macro analysis of the individual countries and regions, we need to be fully aware of the number of countries present in the dataset, historical highest and lowest rates, and average per

territory. We will use suicide rates per 100K pop for this fundamental analysis. This will allow us to be more consistent regardless of the other factors that might affect the suicide rate.

The distribution chart in Figure 11 shows that the suicide rate per 100K historical has been around 12 people with a distribution skewed to the right. Most countries have been under the average except for a few outliers. Those outliers drew the national average to 12. The countries with the lowest historical average suicide rates per 100K pop are Dominica, Jamaica, Oman, Antigua, Kuwait, and United Arab Emirates. These countries have national average suicide rates across all ages and generations under two people per 100K. More details will be provided when all these attributes are analyzed together.

3.1.3 Age/Generation



To create efficient prevention per country, we need to understand how specific age groups are affected by suicides. The dataset has a total of six different ages groups from 0 to 75 +.
It's worth mentioning that other factors can affect how each group is affected by suicide, such as family culture, government support, programs available and many more. Still, for this exploration exercise, we will look at each group globally to understand the overall average rates of each group. There is a well-established trend here by a single inspection on the average suicides per 100k population by groups and age. It seems like as people grow older, there are more likely to be susceptible to commit suicide. This is a fundamental analysis of the period and suicide rate; more details will be provided.

3.1.4 GDP

A correlation analysis was completed, and GDP was determined not to significantly impact the number of suicides occurring in a country. In fact, and perhaps counter-intuitive to most, GDP was one of the least relevant parameters in determining the suicide rate. This observation is illustrated in

	year	suicides_no	population	suicides/100k pop	HDI for year	gdp_per_capita
year	1.000000	-0.004546	0.008850	-0.039037	0.366786	0.339134
suicides_no	-0.004546	1.000000	0.616162	0.306604	0.151399	0.061330
population	0.008850	0.616162	1.000000	0.008285	0.102943	0.081510
suicides/100k pop	-0.039037	0.306604	0.008285	1.000000	0.074279	0.001785
HDI for year	0.366786	0.151399	0.102943	0.074279	1.000000	0.771228
gdp_per_capita	0.339134	0.061330	0.081510	0.001785	0.771228	1.000000

Figure 12 - Variable correlation coefficients

4. Driving Factors of Suicide Based on Exploratory Analysis

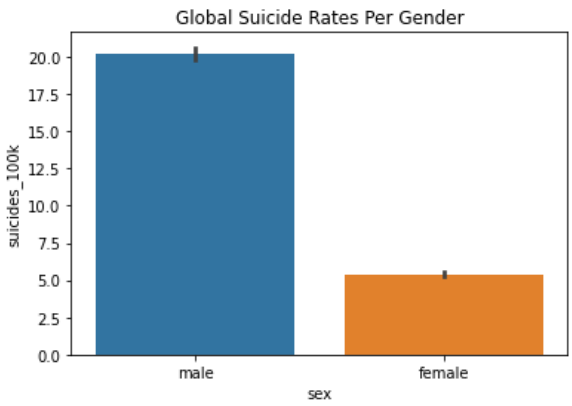
4.1 Data Analysis of Suicide Rates by Gender and Age

To develop a strategy with the end goal of reducing global suicide rates, one must understand how specific subgroups are affected. This section will focus on how different age groups and genders are represented in global suicide rates. Additionally, an outlier analysis will be performed to determine if major international or national events can further affect these groups

4.1.1 Analysis of Suicide Rates by Gender

Are suicide rates higher among males than females?

To understand how different genders are affected, we will first analyze global suicide rates over the timeframe for which data is available irrespective of geography. Our analysis shows that men are far more prone to be victims of suicide, representing ~ 80% of global suicides from 1985-2016.

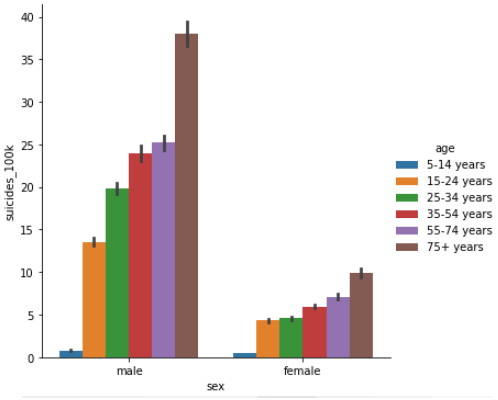


Additionally, we investigated how men and women in different age groups were affected. The age group categories are: 5-14, 15-24, 25-34, 35-54, 55-74, and 75+. The trend is relatively flat for females, hovering around an average of 5 suicides per 100k population. For males, there is a definitive increase in suicide rates with age; male and middle-aged adult males are at very high risk of suicide.

The top five countries with the highest suicide rates (#/100k people) for males are (in decreasing order): Lithuania (68), Russia (58), Sri Lanka (55), Belarus (53), and Hungary (51). The top countries for women are Sri Lanka (15), South Korea (15), Hungary (51), Japan (14), and Lithuania (13).

*Figure SEQ Figure * ARABIC 13 - Global Suicide Rates v. Gender*

In reviewing possible explanations of the data, it was discovered that alcohol consumption rates are among the highest in the world in Eastern Europe (Editorial Alcohol and Suicide in Eastern Europe, n.d.). Also, alcohol consumption rates among males are much higher than females. Any initiative to decrease global suicide rates among men should address combatting alcoholism and the underlying factors that may cause it (family problems, cultural stigmas, etc.) The significant overrepresentation of Eastern European countries among male suicide rates indicates that alcohol consumption is essential.



Interestingly, when female suicide rates are investigated, Sri Lanka jumps to the top of the list. In reviewing possible explanations, it was noted that Sri Lanka continues to suffer from significant gender inequality issues (Women's Rights in Sri Lanka, n.d.). Female engagement in the workforce is shallow, and many women are economically dependent on their spouses for survival. This could lead to many women being trapped in abusive relationships from which they do not have the means to escape or many women simply being unable to pursue their passions in life. A global aim to reduce suicide rates must include balancing global gender inequality and remarkably increasing women's participation in the workforce.

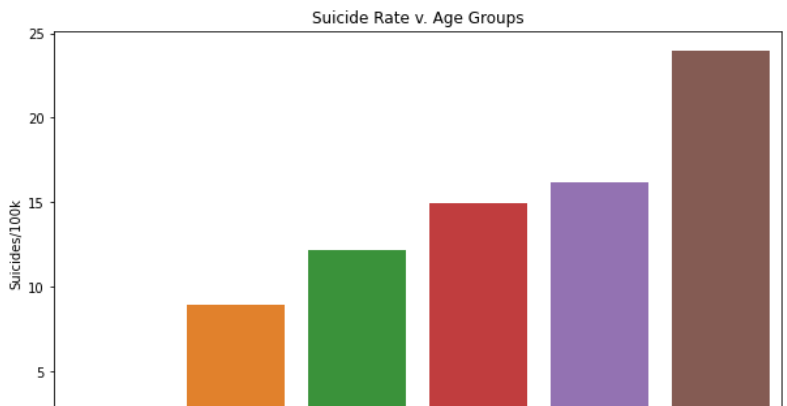
		mean
sex	**country**	
male	**Lithuania**	67.956947
	Russian Federation	58.183704
	Sri Lanka	55.091667
	Belarus	52.757619
	Hungary	51.419355
female	**Sri Lanka**	15.498636
	Republic of Korea	14.812527
	Hungary	14.103677
	Japan	13.692688
	Lithuania	12.874198

*Figure SEQ Figure * ARABIC 15 - Global suicide rates per gender by country*

4.1.2 Analysis of Suicide Rate by Age

Do older people have a higher likelihood of committing suicide?

In the earlier section investigating the impact of suicide rates v. gender, it was noted that there was a general increase in suicide rates with age for both genders. The following data is re-imaged in, irrespective of gender.



When exploring the gender impacts on suicide rates, we can investigate the distribution of suicide rates by age for different

countries. This information is shown in Figure 16 - Suicide rates v. age

As expected, Eastern European countries are represented in all age groups. The age-group category where they aren't the majority is 5-14 years. Kiribati and Cabo Verde are in the top 3 for this age group. As they are highly impoverished countries, this could indicate that poverty has a more significant impact on suicide rates among children.

4.2 Outlier Analysis of Suicide Rates

Figure SEQ Figure * ARABIC 16 - Suicide rates v. age

An outlier analysis of the suicide rate was completed to determine which years had notably higher rates and if there could be any historical events responsible for these rates. The mean and standard deviation of the suicide rate for each country was calculated. A particular year was flagged as "high" if the suicide rate for that year was more significant than the mean + 1 standard deviation (on a country-specific basis). This data is shown below in year. It is interesting to note that suicide rates were at their highest before and following the collapse of the Soviet Union (1989). This fact, combined with the earlier observation that the former Soviet States have been overly represented in suicide rates across different genders and age ranges, could indicate that the hardships leading up to and following the collapse of the USSR could have been a significant contributor to global rates of suicide.

1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	...	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
16	17	23	22	16	21	11	15	16	20	...	4	7	11	7	9	2	6	7	5	2

Figure 17 - Number of countries with higher-than-average suicides in a given year

4.3 Data Analysis of Suicide Rates by GDP per Capita

Are Suicidal rates higher in developed countries than in developing countries?

To understand how increasing GDP per capita over the years is playing a role in decreasing the suicide rates, we did a correlation of the suicide rate vs GDP per capita for each country. We found that there are countries where even the increase in GDP per capita is not bringing the suicide rates down. This analysis was done by isolating four critical variables in the dataset: country, Suicide rate, Year and GDP per capita. After isolating these variables, we calculated the correlation of suicide rate and GDP per capita over the years for each country.

From our analysis, below are some countries where even though the GDP per capita increased over the years, their suicide rates kept on growing. Most of the countries below are not part of the top 10 GDP per Capita, except for the United States. And, below are the top 10 countries where an increase in GDP per Capita did create a positive impact on the suicide rates. An observation that can be made here is that most of the countries in the graph below are either top or middle of the pack if the countries are ranked from highest to lowest GDP per Capita. Hence, the correlation indicates that GDP per capita can only mean a drop in suicides if the GDP per capita is not the lowest group among all the countries.

*Figure SEQ Figure * ARABIC 18 -Top 10 countries with a negative correlation between suicide rate and GDP*

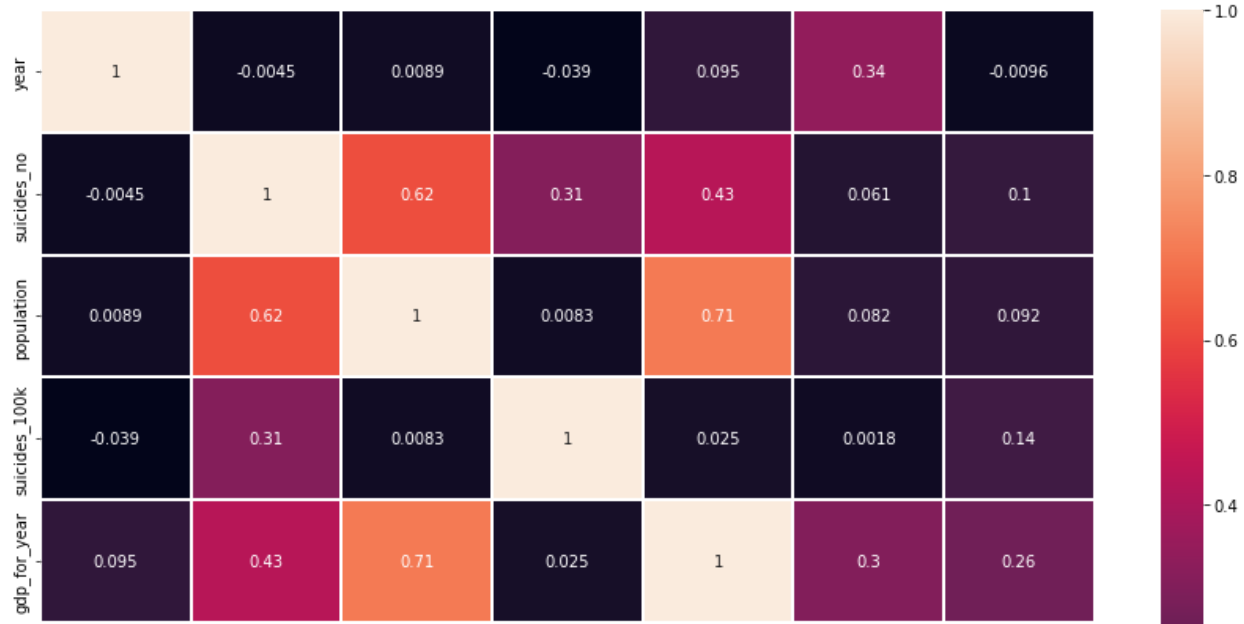
5. A

Our a
Decis
on the effectiveness of these algorithms.

Our model follows Supervised Learning, which consists in learning the link between two datasets: the observed data X and an external variable y that we are trying to predict, usually called “target.” Our goal is to make accurate predictions for each algorithm.

All ML models were deployed with the same target variable, ‘Suicides/100k’.

In the first part, we visualized the data to help a better understanding of how suicide related to the HDI, age, sex, country and population. We correlated each column and had a better experience of how the attribute groups are connected.



From the Correlation Heatmap, we see that the number of suicides is related to the population. And the GDP per capita is highly related to the HDI for the year. We can see also that the suicides no have less correlation with the GDP per capita. Before doing this correlation, we thought the suicide number would have a high correlation with GDP in that higher GDP would have fewer suicide numbers; it turns out that's not true.

5.1 Feature importance

5.1.1 Decision tree regression

The most important feature for the Decision Tree model to predict the suicide number is HDI and age, followed by sex. Other essential features include country and population.

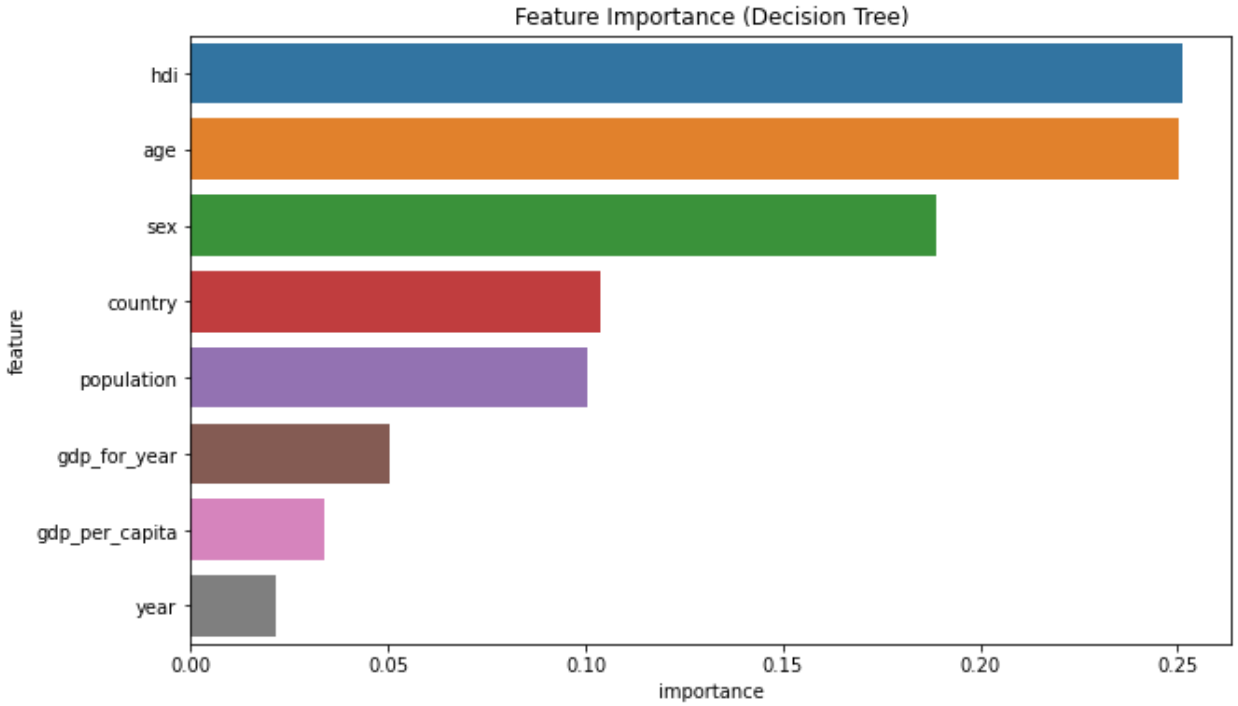


Figure 20 - Decision tree feature importance

5.1.2 Random Forest

The most important feature for the Random Forest model to predict the suicide number is HDI and age, followed by sex. Other essential features include country and population.

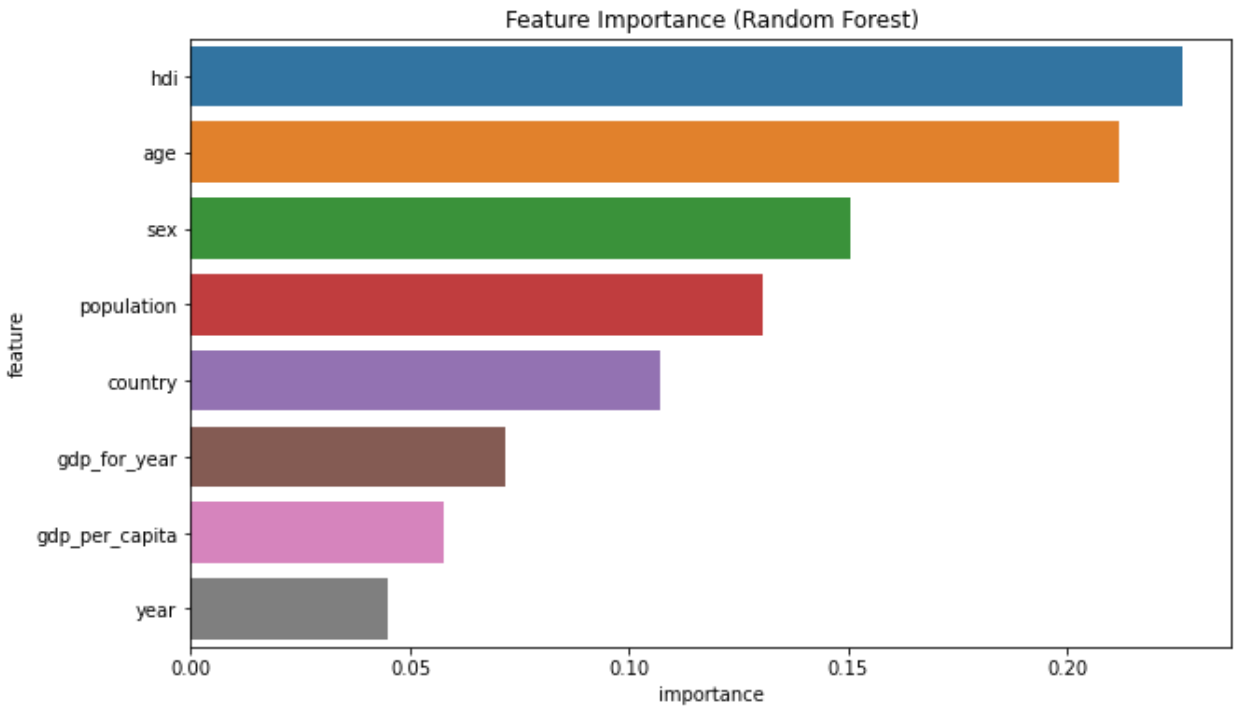


Figure 21 - Random Forest feature importance

5.2 Performance of Random Forest Regression Model

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks.

Various metrics can be used to evaluate the performance of a Linear Regression model. We will use the **RMSE (Root Mean Squared Error)** value, a frequently used measure of the differences between values (sample and population values) predicted by a model and the observed values.

```
Random Forest: Accuracy on training Data: 0.983
Random Forest: Accuracy on test Data: 0.890
Random Forest: The RMSE of the training set is: 2.444818497248533
Random Forest: The RMSE of the testing set is: 6.316115248096933
```

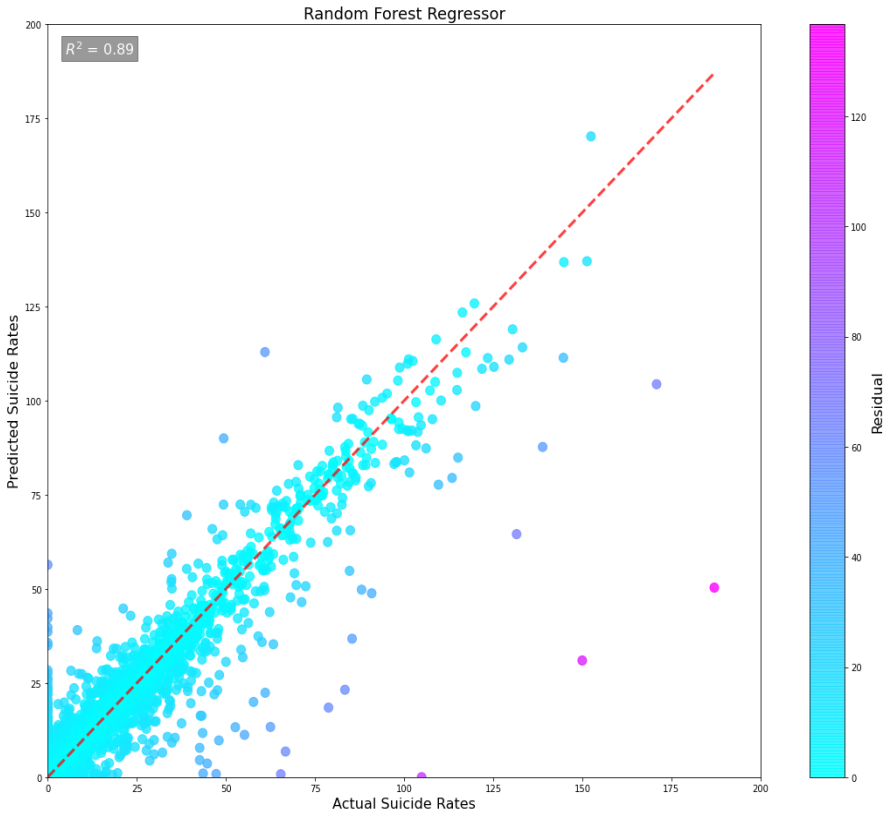
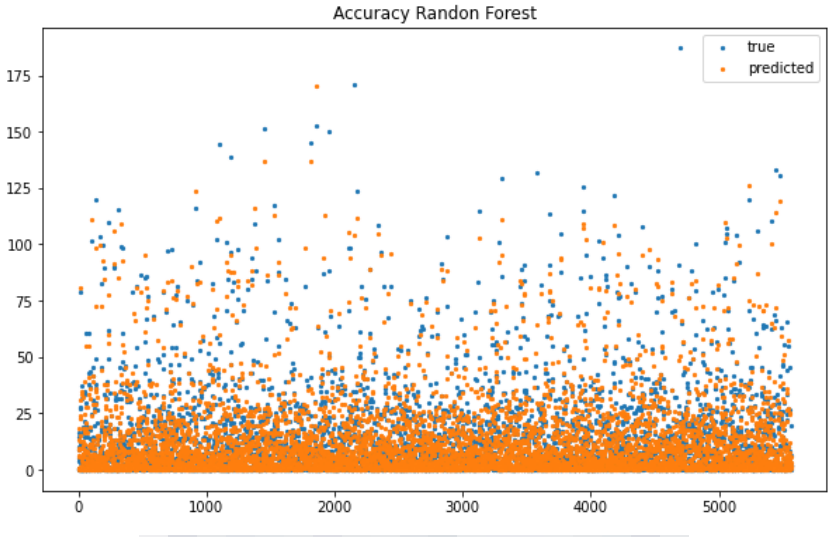


Figure 22 - RFR performance



5.3 Performance of Linear Regression Model

Linear regression, or ordinary least squares (OLS), is the simplest and most classic linear method for regression. Linear regression finds the parameters w and b that minimize the mean squared error between predictions and the actual regression targets, y, on the training set. The accuracy of training Data and test Data is just 0.2. Therefore, the performance of this model is not good. In addition, we observed that the score on the training set and test sets are very close. That means we are underfitting, not overfitting.

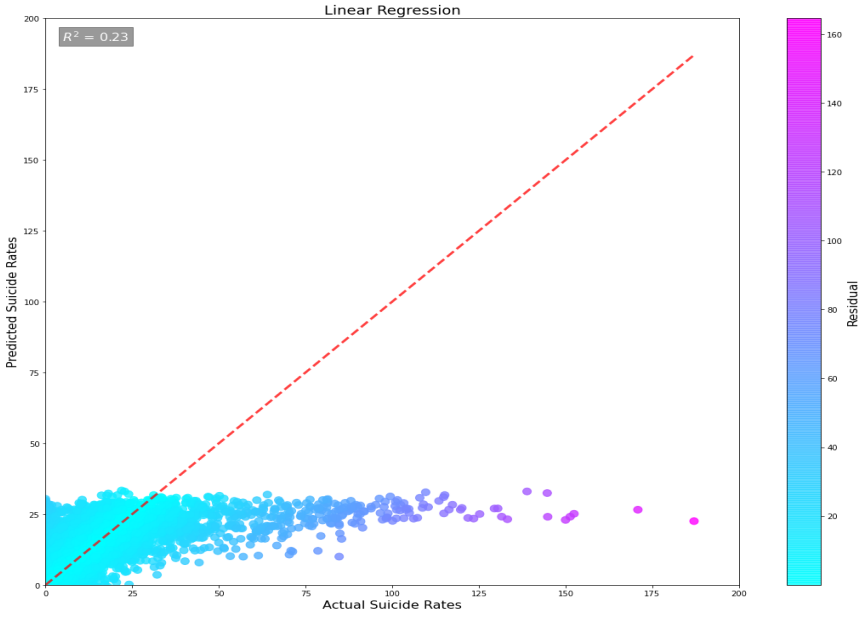


Figure 24 - Linear Regression Performance

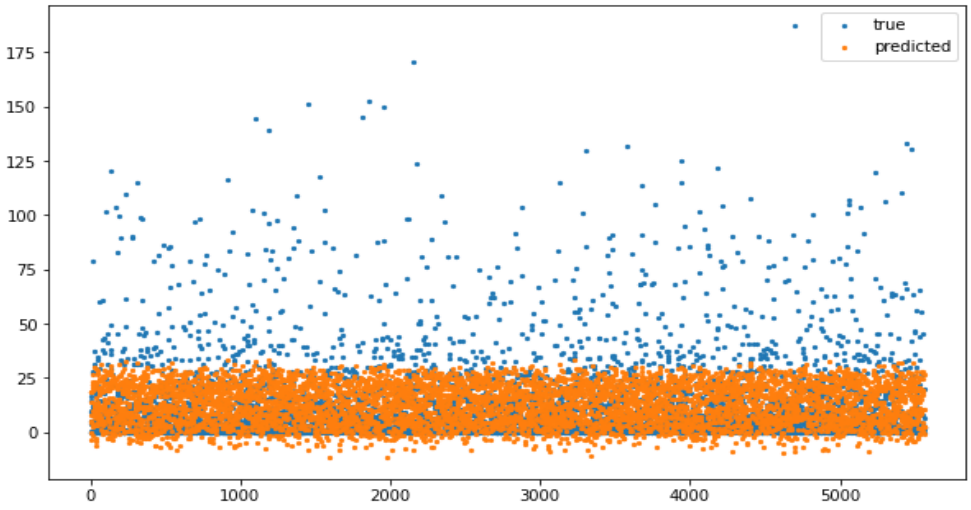


Figure 25 - Linear Regression Accuracy

5.4 Performance of Model - Decision Tree Regression

Decision trees are widely used models for classification and regression tasks. Essentially, they learn a hierarchy of if/else questions, leading to a decision.

As can be seen below, this model is entirely accurate. For the Decision Tree, the accuracy for training data is 0.802 and 0.733 for test data. This is more accurate than the Linear Regression.

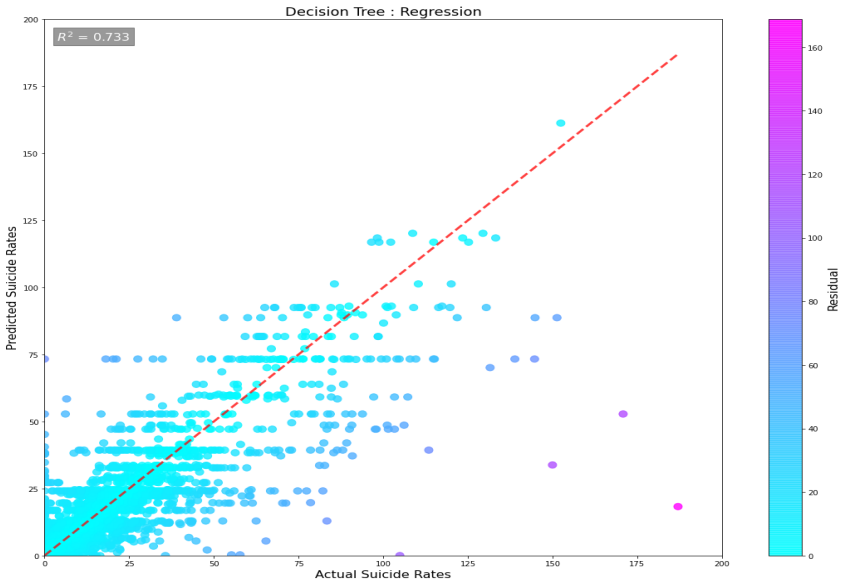


Figure 26 - Decision tree regression performance

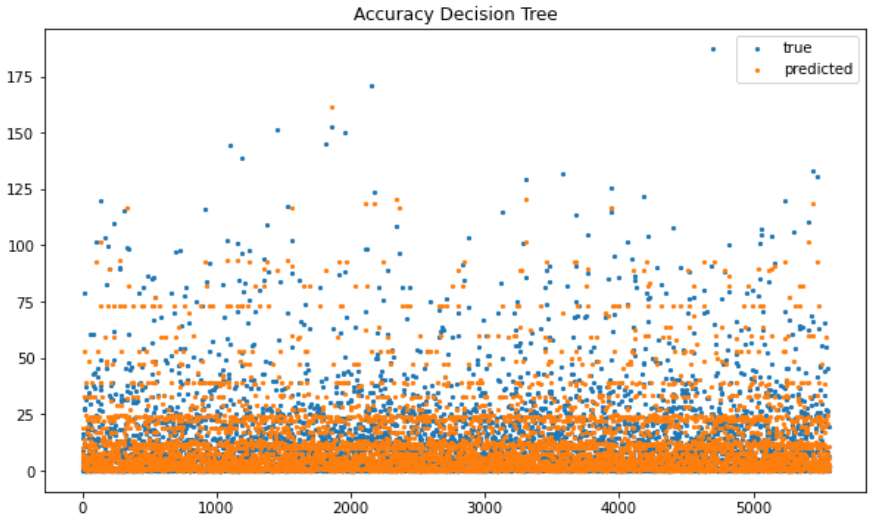


Figure 27 - Decision tree regression accuracy

5.5 Comparison of Results by Different Algorithms

The proposed system is implemented with a few regression algorithms to finalize which regression algorithm will give better accuracy, and the comparison results are listed in the table above. This result shows that the RMSE value is the least for the Random Forest Regression for the testing set and quite an optimal value for the training set. Thus, it performs the best for our model. Also, the accuracy of the Random Forest model is the best one, followed by the decision tree model.

	Model	Accuracy train	Accuracy test	RMSE train	RMSE test
0	Linear regression	0.22	0.23	16.68	16.72
1	Decision tree regression	0.80	0.73	8.41	9.85
2	Random Forest	0.98	0.89	2.44	6.31

Conclusions

As shown in the data exploration and machine learning analysis, the suicide rate between countries has a large variability over the years. We can see that the male population is more prone to suicide than females. Also, elderly populations tend to be more at risk than young ones.

An unexpecting and essential point is that even though the GDP per capita increased over the years, the suicide rates kept growing among the top 10 GDP per Capita (except for the United States) and substantial positive impact on the suicide rates below top 10 GDP per Capita.

The analysis shows the highest risk of suicide in the countries with :

- dramatical social changes
- gender inequality
- kids poverty

Interestingly, that social changes and gender inequality have different influence on sex (males depends on social changes more than women).

Therefore, we can see that the impact of social factors, such as gender inequity or social changes, on the suicide rates is much more important than economic factors have.

HDI is the most critical factor to predict suicide rate, and the best result to indicate Random Forest model gives the rate.

References

Duncalfe, L. (2021, 12 1). *ISO 3166 Countries with Regional Codes*. Retrieved from GitHub: https://raw.githubusercontent.com/lukes/ISO-3166-Countries-with-Regional-Codes/master/all/all.csv

Editorial Alcohol and Suicide in Eastern Europe. (n.d.). Retrieved from peertechzpublications: https://www.peertechzpublications.com/Addiction-Medicine-Therapeutic-Science/JAMTS-1-101.php#:~:text=There%20are%20a%20number%20of,in%20the%20world%20%5B4%5D.

Rusty. (2021, 12 5). *Suicide Rates Overview 1985 to 2016*. Retrieved from Kaggle.com: https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016

Szamil. (2021, 11 25). *Suicide in the Twenty-First Century*. Retrieved from Kaggle.com: https://www.kaggle.com/szamil/suicide-in-the-twenty-first-century/notebook

The World Bank. (12, 5 2021). *World Bank Development Index*. Retrieved from Development Indicators, GDP by Country: https://databank.worldbank.org/source/world-development-indicators#

United Nations. (2021, 12 1). *United Nation Human Development Reports*. Retrieved from United Nations Development Program: http://hdr.undp.org/en/indicators/137506

Women's Rights in Sri Lanka. (n.d.). Retrieved from Borgern Project: https://www.google.ca/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjgq8Ot39n0AhUSVc0KHdy8C90QFnoECBYQAw&url=https%3A%2F%2Fborgenproject.org%2Fwomens-rights-in-sri-lanka%2F%23%3A~%3Atext%3DGovernment%2520Representation%253A%2520Women%2

World Health Organization. (2021, 12 4). *Suicide Prevention*. Retrieved from WHO Health Topics: https://www.who.int/health-topics/suicide#tab=tab_1