Contents

1.INTRODUCTION

Suicide is defined as the act of ending people's own life. The causes of suicide are generally caused by psychological factors and there are many different situations that cause these factors. The financial situation of the people, the desperation and depressions brought by the cultural structure experienced, and being desperate towards the future can be the reasons that cause people to commit suicide. Apart from these, depression, psychosis, substance use disorders, bipolar disorder and post-traumatic stress disorder can be given among the frequently observed reasons. According to WHO statistics, close to 800.000 people die by suicide every year. Furthermore, for each suicide, there are more than 20 suicide attempts and suicide occurs in all regions of the world. In fact, 79% of global suicides happen in low- and middle-income countries [1]. The causes of suicide should be explored and tried to be eliminated for a healthy social environment. The aim of this study is first to evaluate suicide rates in terms of countries years and other attributes which in the dataset then to create a classification model and to divide the data into suitable classes.

2. TOOLS

The Python programming language of version 3.7.5, which contains many useful data tools, was used to analyze and manipulate the dataset. Jupyter Notebook 6.0.2 is used as programming editor. Pandas, numpy, sklearn modules to process data and make calculations, and IPython, pydotplus, seaborn and matplotlib modules were used to visualize the data.

3. DATASET AND PREPROCESSING

The dataset on the https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016 site was used to investigate suicide rates. There are 27820 records in the dataset and each record has 12 features. These features are “country”, “year”, “sex”, “age”, “suicides_no”, “population”, “suicides/100k pop”, “country-year”, “HDI for year”, “gdp_for_year ($)”, “gdp_per_capita ($)” and “generation”.

FEATURE	TYPE	CLASSES	NULL
country	object	101 different country	NO
year	integer	1985…2016	NO
sex	object	Female, Male	NO
age	object	5-14, 15-24, 25-34, 35-54, 55-74, 75+	NO
suicides_no	integer	continuous	NO
population	integer	continuous	NO

country-year	object	Unified feature	NO
suicides/100k pop	float	continuous	NO
HDI for year	float	continuous	NO
gdp_for_year ($)	object	Same value for same year and country couple	YES
gdp_per_capita ($)	integer	continious	NO
generation	object	X,Silent,G.I, Boomers,Millenials,Z	NO

Table 3.1. Dataset Features and Its Specifics

The “country” feature indicates the country where the suicide was made, and “year” indicates the year. “suicides_no” shows the number of suicides that occurred that year in the that country. The “age” attribute indicates the age range of suicides, “sex” for gender and “generation” indicates which generation they belong to. “population” shows the population of the country that year and “suicides/100k pop” shows the suicide rate per hundred thousand people. “HDI for year” refers to that year's Human Development Index. The Human Development Index (HDI) is a statistic composite index of life expectancy, education, and per capita income indicators, which are used to rank countries into four tiers of human development [4]. “gdp_for_year ($)” refers to that year's Gross domestic product and it is one of the most common indicators used to track the health of a country’s economy. The ratio of this value to the total population is expressed as “gdp_per_capita ($)”.

In some records, the gdp_for_year ($) attribute is empty, and the number of these records constitutes a large portion of the dataset as 19456. it is not the right approach to extract information from the remaining records, as many truth-reflecting data are deleted. Also, since the country-year property is not useful for extracting information, gdp_for_year ($) and country-year properties should be dropped from the dataset. Rather than doing a separate analysis for each year, evaluating suicide rates for every 10 years can help us have more general and accurate information. If there were information such as salary amount, marital status, number of children for each year, important inferences could be made for each year. In the last case, there are 3 different year categories in the dataset, these are 1985-1995, 1996-2006,2007-2016.

3.1. Information About Dataset

First, the names of the features have been rewritten as it provides ease of programming. The new features’ names are “Country”, “Year”, “Gender”, “Age”, “SuicidesNo”,

"Suicides100kPop", "CountryYear", "HDIForYear", "GdpForYearMoney", "GdpPerCapitalMoney" and "Generation". Inferences can be made considering the relationship of each feature in the dataset with the other. For example, by evaluating the distribution of age and gender in suicide rates in the dataset, it can be concluded which gender and age group are more prone to suicide. As seen in figure 3.1.1, suicides occurred between 1985 and 2016 were mostly done by males. According to the graph, the proportion of a man who commits suicide among every 100 thousand people is about 3.75 times more than women.

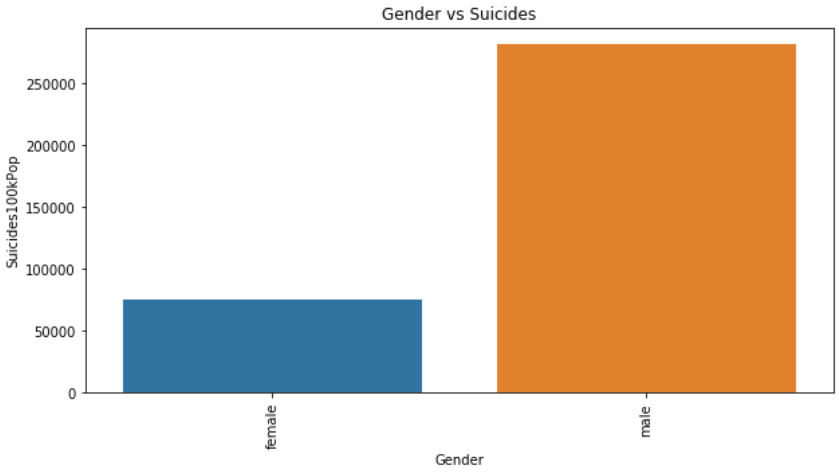

Figure 3.1.1. Distribution of Suicide Rates Per 100k By Gender

Considering the age of those who committed suicide between 1985 and 2016, as seen in figure 3.1.2, the suicide of people 35-54 is much higher than in other age groups. Also, considering the number of suicides per 100 thousand people, the group over 75 is the most suicidal group and it can also be said that suicide rates per 100 thousand increase with increasing age.

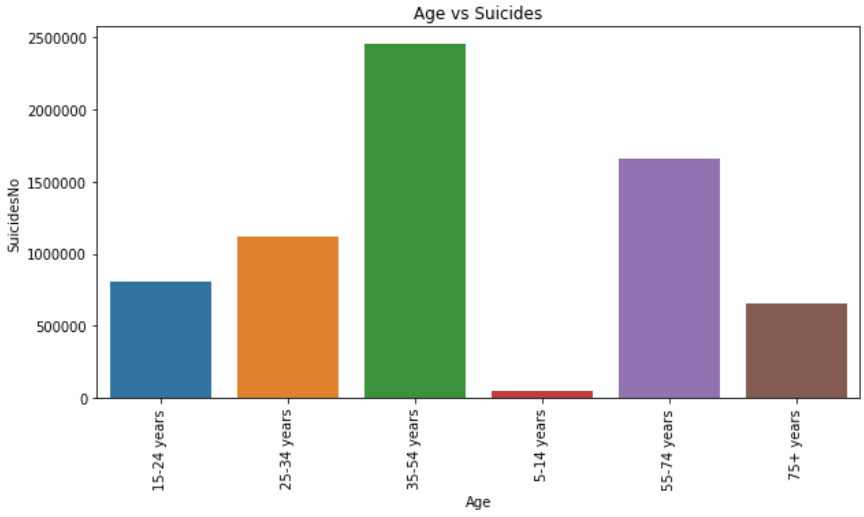

Figure 3.1.2. Distribution of Suicide Numbers by Age

Considering the distribution of suicide rates by years, it is seen that the most suicide occurred between 1996-206. In fact, suicide rates were lower between the years 1996-2006 than 1985-1995, and then increased from 2007-2016. It can be examined why suicide rates increased between 1996-2006 and later declined. This could be explained by many features not included in the dataset, but starting from dataset, this situation can be examined with the Human Development Index, Country and GDP per capita.

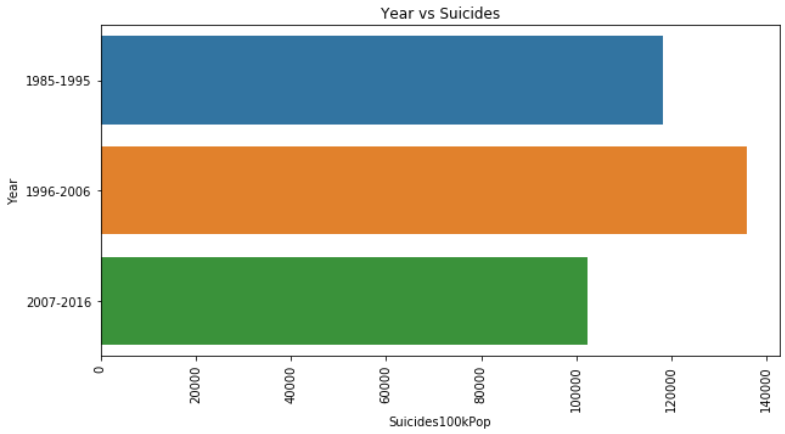

Figure 3.1.3. Distribution of Suicide Rates by Years

After observing that suicide rates had the highest value between 1996-2006, countries can be interpreted by looking at which countries have committed the most suicide. It can be evaluated whether the country had situations that could lead to suicide in those years.

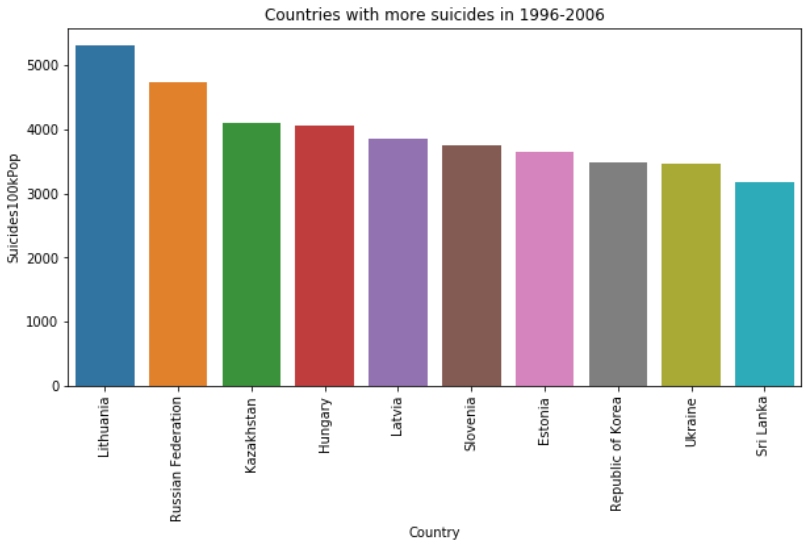

Figure 3.1.4. Countries with More Suicide In 1996-2006

Considering the rate of suicides per 100.000 people committed, the top three numbers belong to Lithuania, Sri Lanka and Russia as seen in figure 3.1.5. The reasons for this can be varied

according to the countries, but similar reasons can be found and prevented by doing various researches.

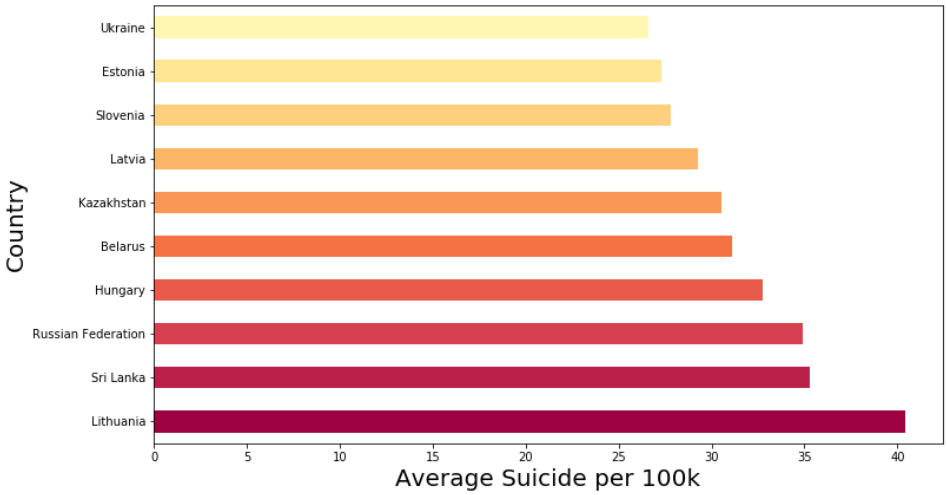

Figure 3.1.5. 10 Countries with The Highest Suicide Rates

3.2. Preprocessing

Some preprocesses are required before creating a model for the classification process. These preprocesses are made in terms of programming and for the implementation of algorithms. these preprocesses are:

1. “Country”, “Year”, “CountryYear” and “HDIForYear” are dropped.
2. Males and females are expressed as 0 and 1 respectively.
3. 'Boomers',' Generation X ',' Generation Z ',' G.I. Generation ',' Millennials', 'Silent' generations are expressed with 0,1,2,3,4,5 respectively.
4. Commas in the “GdpForYearMoney” feature have been removed.
5. '15 -24 years', '25 -34 years', '35 -54 years', '5-14 years', '55 -74 years', '75 + years' are expressed with 0,1,2,3,4,5 respectively.

The format of the data after preprocessing is presented in Figure 3.2.1 with samples.

	Gender	Age	SuicidesNo	Population	Suicides100kPop	GdpForYearMoney	GdpPerCapitalMoney	Generation
0	0	0	21	312900	6.71	2156624900	796	1
1	0	2	16	308000	5.19	2156624900	796	5
2	1	0	14	289700	4.83	2156624900	796	1
3	0	5	1	21800	4.59	2156624900	796	3
4	0	1	9	274300	3.28	2156624900	796	0

Figure 3.2.1. Preprocessed Dataset Samples

4. CLASSIFICATION

Many preprocesses have been made to model the dataset. The features to be used for the model at the end of these processes are “Gender”, “Age”, “SuicidesNo”, “Population”, “GdpForYearMoney”, “GdpPerCapitalMoney” and “Generation”. An extra feature is required to classify the data in the dataset. This feature is called a label and can be set as a risk level. The average of suicide rate per hundred thousand people was calculated as a 12.81 to assign a class to the risk level feature. Because this rate is high, it is assumed that the population of the country is suicidal. Values below the average are labeled "0" as low risk group and values above the average are labeled as high risk as "1". In Figure 4.1, an example of dataset ready for model training is given. The data set was divided into train and test, with 20% being the test and 80% being the training data. Since the “Suicides100kPop” attribute is already used for the labelling, this feature is not included in the training set.

	Gender	Age	SuicidesNo	Population	Suicides100kPop	GdpForYearMoney	GdpPerCapitalMoney	Generation	at_risk
1100	1	2	3	563626	0.53	2118467913	587	0	0
23643	1	2	168	4829500	3.48	529121577320	14223	0	0
23089	0	0	35	149200	23.46	22689994990	12041	1	1
22326	0	3	0	7017	0.00	847397850	10157	2	0
9362	0	5	1346	1985275	67.80	2918382891460	49901	5	1

Figure 4.1. Samples from The Dataset to be Modeled

4.1. Decision Tree Method

ID3 and its successor C4.5 algorithms use the information gain for decision tree branching. In the decision tree, according to the Information Gain or Gini index concepts, branching can be done with the attributes of the tree. In the study, two different decision trees are created by calculating the Information Gain and then using the Gini index. To create a tree with information gain, first total entropy and entropy of each feature must be calculated. Information gain favors splits with small counts but many unique values. The entropy of each feature is subtracted from the total entropy and the information gain of each feature is calculated and the feature with the highest information gain is selected as the node. In the first calculation, the biggest information gain belongs to the “Gender” feature with 0.9. So, the root of the tree will be the “Gender” feature. There are two classes in the Gender feature, male and female. Therefore, subsequent information gains are calculated for cases where this class is male and female, respectively.

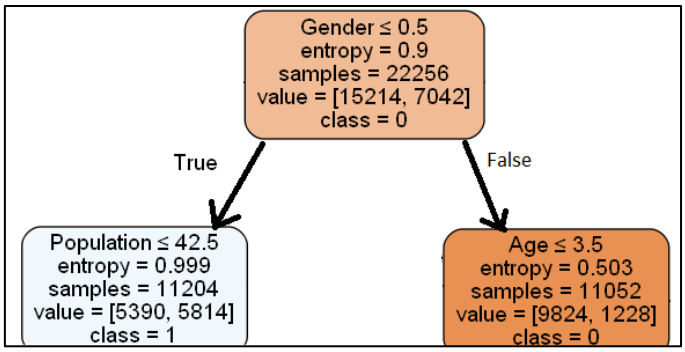

Figure 4.1.1. Branching of The Decision Tree According to Information Gain

The accuracy rate of the model created was measured as 0.9874. Figure 54 shows the correct and incorrect estimation rates of classes. According to this matrix, 3820 out of 3847 low-risk groups were estimated correctly. Likewise, 1674 out of 1717 high risk groups were estimated correctly. Using these numbers, f1-score, precision and recall metrics can be calculated that provide different inferences about the accuracy rate. Performance metrics calculated with the Python's "sklearn" module are shown in figure 4.2.3.

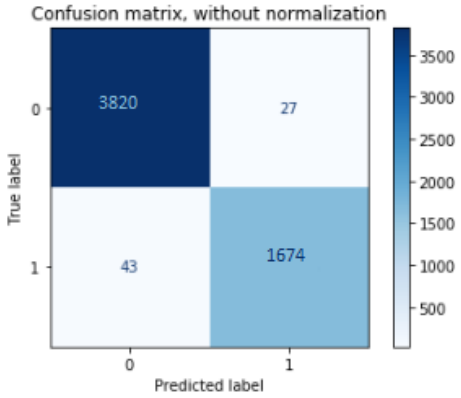

Figure 4.1.2. Confusion Matrix for Decision Tree Model

	precision	recall	f1-score
0	0.99	0.99	0.99
1	0.97	0.98	0.98
accuracy			0.99
macro avg	0.98	0.99	0.99
weighted avg	0.99	0.99	0.99

Figure 4.1.3. Performance Metrics for Information Gain Used Decision Tree Model

Another decision tree algorithm CART (Classification and Regression Tree) uses the Gini index method to branching the decision tree. Gini index method favors larger partitions and

uses squared proportion of classes. When creating a tree with the gini index, the gini index of all properties is calculated and the feature with the lowest result is selected as the root node. Gini Index is a metric to measure how often a randomly chosen element would be incorrectly identified. When using the gini index method to branch the decision tree, the root node has been "SuicidesNo" as figure 4.2.4 and the gini index is 0.433.

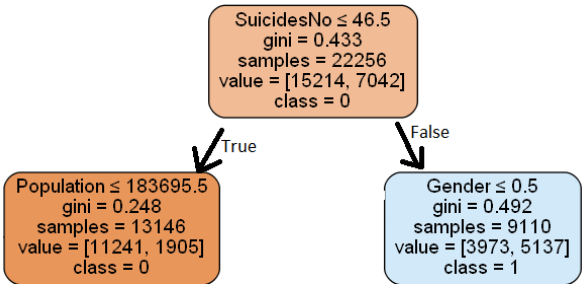

Figure 4.1.4. Branching of The Decision Tree According to Gini Index

The accuracy rate of the classification model created using gini index is 0.9906. The classification success of both models has been calculated very high and close to each other. According to confusion matrix in figure 4.2.4, 3823 out of 3847 low-risk groups were estimated correctly. Likewise, 1689 out of 1717 high risk groups were estimated correctly. Also, the performance metrics precision, recall and f1-score are shown in figure 4.2.5.

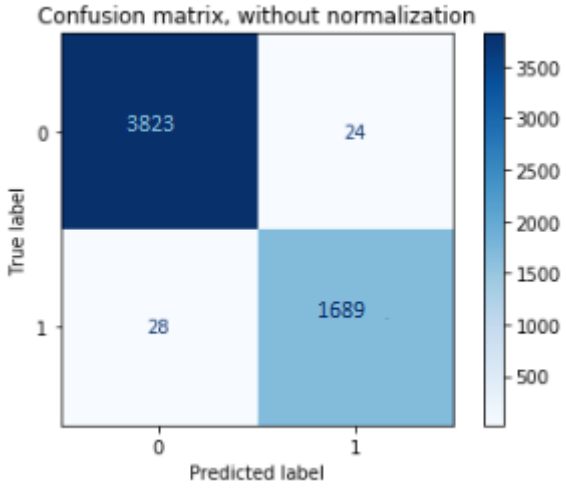

Figure 4.1.4. Confusion Matrix for Decision Tree Model

	precision	recall	f1-score
0	0.99	0.99	0.99
1	0.98	0.99	0.98
accuracy			0.99
macro avg	0.99	0.99	0.99
weighted avg	0.99	0.99	0.99

Figure 4.1.5. Performance Metrics for Gini Index Used Decision Tree Model

5. CONCULISION

Suicide rates dataset accessed through the internet were examined and classified using proper data mining tools. While creating the classification model, two different branching methods of decision trees which are gini index and information gain, were used and very close accuracy rates were obtained. The accuracy of the final model was calculated as % 99. As a result of this classification, countries should determine whether they are in the risk group and seek corrections such as making improvements in order to increase the national income per capita, increasing the welfare of the people. Countries that are at high risk in suicide tendency not to be psychologically healthy and strong, and since this affects every life from working life to social life, the underlying causes of suicides should be found and decreased for the development of nations.

SOURCES

[1]. Internet: https://www.who.int/health-topics/suicide#tab=tab_1 Accessed: 20.05.2020

[2]. Internet: https://www.datacamp.com/community/tutorials/decision-tree-classification-python Accessed: 20.05.2020

[3]. Internet: https://scikitlearn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

Accessed: 21.05.2020

[4]. Internet: https://en.wikipedia.org/wiki/Human_Development_Index Accessed: 20.05.2020