Machine learning prediction model for suicidal rate

Among the Continent

Submitted By

Amit Chowdhury

ID: 181-35-2376

Batch: 25th

Department of Software Engineering

Daffodil International University

Supervised By

Khalid Been Md. Badruzzaman Biplob

Senior Lecturer

Department of Software Engineering

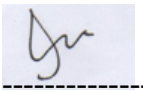
Daffodil International University

This thesis report was submitted in order to meet the requirements for a Bachelor of Science in Software Engineering degree.

APPROVAL

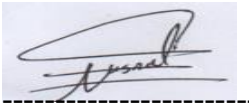
This thesis titled on "**Machine Learning prediction model for suicidal rate Among the Continent**", submitted by **Amit Chowdhury**, ID: **181-35-2475** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS



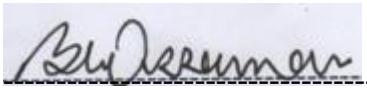
Chairman

Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University



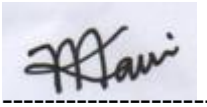
Internal Examine 1

Nusrat Jahan
Assistant Professor
Department of Software Engineering
Daffodil International University



Internal Examine 2

Khalid Been Badruzzaman Biplob
Senior Lecturer
Department of Software Engineering
Daffodil International University



External Examiner

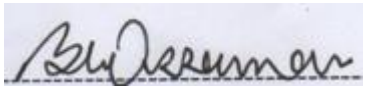
Professor Dr M Shamim Kaiser,
Professor
Institute of Information Technology
Jahangirnagar University

DECLARATION

I hereby state that I have taken this thesis under the supervision of Khalid Been Md. Badruzzaman Biplob, Senior Lecturer, Department of Software Engineering, Daffodil International University. I also acknowledge that neither this thesis nor any part of this has been submitted elsewhere for the award of any degree previously by others.

Amit Chowdhury
ID: 181-35-2475
Batch 25 th
Department of Software Engineering
Faculty of Science & Information Technology
Daffodil International University

Certified by



Khalid Been Md. Badruzzaman Biplob
Senior Lecturer
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

ACKNOWLEDGEMENT

This thesis I am representing was only possible to complete with guidance from some conscientious people. I want to thank each of them. Especially obligated to Daffodil International University for the direction and constant supervision by my honourable teacher Khalid Been Md. Badruzzaman Biplob. I would like to be thankful to my supervisor for his kind support, guidance, and encouragement and I want to express my gratitude towards my parents, teachers, batch mates, and my seniors of DIU for their kind assistance and advice to complete my study.

Table of Contents

Abstract

Suicide is frequently induced by a variety of factors. Above all, it's regarded as a mental illness. Because there is a link between mental illness and suicide. People commit suicide because they are depressed and anxious about their lives. Artificial intelligence and machine learning are two terms that are often used interchangeably. This is regarded as a watershed moment in the history of computing. We'll use machine learning to figure out what percentage of people die by suicide around the world. In this study, we use a variety of classification algorithms to identify each country's suicide predictor level. The Random Forest Classifier is the most accurate model. In our model, the Low class has 100 percent precision and recall, the Medium class has 96 percent precision and recall, and the High class has 90 percent precision and 88 percent recall.

Keywords: Suicide, Machine Learning, Classification Algorithms, Random Forest

CHAPTER 1

INTRODUCTION

Suicide is a deadly disease in today's society. As long as people are destroying their lives by their own hands. This loss of life is called suicide. According to the study, about one million people commit suicide each year. Adolescents under the age of 35 have the highest number of suicides in the world. This is because frustration and anxiety are more prevalent among them than others. In this case, the suicide rate of men is higher than that of women. This amount is about 3-4 times more than others. Then the question may arise why people destroy such a beautiful life with their own hands? Chose the path of suicide? Suicide is usually caused by different people. But above all, it is considered a mental illness. Because there is a connection between suicide and mental illness. People commit suicide out of extreme emotional frustration and anxiety about life. Those who commit suicide may also have complex depression, other mental illnesses such as bipolar disorder, which increase the risk of suicide by more than 20 times.

Machine learning and artificial intelligence. Which is one of the milestones in the computer world. We all know that machine means instrument, learning means education. Machine learning is a special kind of artificial intelligence. Which helps a machine or software to learn or operate something on its own. In other words, machine learning is to make the computer think like a human being.[23]

The main topic of our research was about suicide. We will determine the percentage of suicides in the world through machine learning. Machine learning gives results depending on how much data. When we want to find out the value of something, we create a model through

machine learning with some data before and after that subject. And by analysing the models later we get our desired result. However, different algorithms have to be used before making these models. In the same way in this study we will collect some data related to suicide and find out a specific value through machine learning. Most of the study use regression and classification as a predictor and we also use as predictor.

Regression could be a measurable strategy utilized in back, contributing, and other disciplines that endeavours to decide the quality and character of the relationship between one subordinate variable and a arrangement of other factors. Regression makes a difference venture and money related directors to esteem resources and get it the connections between factors, such as product costs and the stocks of businesses managing in those commodities. Regression analysis is utilized once you need to predict a persistent subordinate variable from a number of free factors. On the off chance that the subordinate variable is dichotomous, at that point calculated regression should be utilized. (In the event that the part between the two levels of the subordinate variable is near to 50-50, at that point both calculated and straight regression will conclusion up giving you comparable comes about.) The independent factors utilized in regression can be either nonstop or dichotomous. Free factors with more than two levels can moreover be utilized in regression analyses, but they to begin with must be changed over into factors that have only two levels. This can be called sham coding and will be talked about afterward. More often than not, relapse examination is used with naturally-occurring factors, as contradicted to tentatively controlled variables, although you'll be able utilize relapse with tentatively controlled factors. One point to be beyond any doubt with regression investigation is that causal connections among the factors cannot be decided. [24][25]

Classification may be a handle of categorizing a given set of information into classes, it can be performed on both organized or unstructured information. The method begins with

foreseeing the lesson of given information focuses. The classes are frequently alluded to as target, name or categories. The task of approximating the mapping job from input elements to discrete yield factors is classified predictive modelling. The most objective is to distinguish which class/category the unused information will drop into.

Classification is a directed learning concept in machine learning that simply categorizes a set of data into classes. The foremost common classification issues are – discourse acknowledgment, confront discovery, penmanship acknowledgment, record classification, etc. It can be either a twofold classification issue or a multi-class issue as well. There are a bunch of machine learning calculations for classification in machine learning. Let us take a see at those classification calculations in machine learning.[26]

1.1 Research Question

Our Research Question is:

- RQ 1: Which Machine learning Model fitted best for Predicting Suicide.
- RQ 2: What are the key features that affect suicidal rate among the continent.

1.2 Research Objectives

- To evaluate the best fitted prediction model for suicide rate among the continent.
- To visualize the difference in key feature that affect suicidal rate.
- To estimate the best fitted prediction model for suicidal rate.

1.3 Organization of the Thesis

- Chapter 1: Chapter one produces the introduction of the thesis. Define the study objectives as well as the research question in this section.
- Chapter 2: This chapter provides background information, a literature review, and examples of past work relating to this study.
- Chapter 3: This chapter displays the whole model and architecture that has been suggested.

- Chapter 4: The experiment, as well as the results and evaluation of the study, are presented in this chapter.
- Chapter 5: The final section of this chapter discusses future scope and restrictions.

CHAPTER 2

LITERATURE REVIEW

One of utmost pivotal way in the software development process is conducting a literature review. It's important to determine the time element, the frugality, and the company's strength before erecting the tool. Once these conditions have been met, the coming 10 way are to establish which operating system and programming language can be used to produce the tool. As soon as the programmers begin working on the tool, it'll be ready to use. Programmers bear a great deal of outside backing. Elderly programmers, books, and websites can all give this backing. The antedating considerations are taken into account when developing the suggested system before it's erected. The machine literacy model's explicatory delicacy is original to that of self-murder threat assessment tools used in internal health settings.[1][2]

Belsher et al. (2019) variety of population assessment characteristics Numerous adolescents who will no way essay self-murder have threat factors. In fact, a recent methodical review of self-murder attempt and death prediction models plant that the delicacy of prognosticating unborn self-murder events was near zero." Likewise, 99 of every 100 individualities anticipated to die by self-murder will not, “the authors continued. They will be subordinated to a potentially stigmatizing threat bracket as well as interventions of dubious efficacity and felicitousness.[3]

Mason Marks (2019) reminds While self-murder prediction in medical systems occurs within the healthcare system and is overseen by legislation similar as the HIPAA in the United States, as well as regulations, that safeguard the protection of mortal study subjects, as well as general guidelines AI- grounded self-murder prediction on social media, medical ethics Platforms are most generally plant outside of the healthcare system. nearly entirely limited, and pots constantly maintain their independence. Styles of prediction are considered exclusive trade secrets.[4]

Psychological risk factors include depression with psychosis, schizophrenia, suicidal ideation, and prior suicide attempts. D'Hotman and Loh (2020) found a significant potential to improve suicide prediction and prevention by combining novel analytical techniques and tools (e.g., leveraging machine learning algorithms and data science) with the opportunity presented in contact and assessment by health services in a qualitative narrative review. They cited a longitudinal study that found that the majority of people who die by suicide (83 percent) have

contacted health services in the year leading up to their death, and 45 percent have contacted these services in the month leading up to their death, highlighting the need for better screening and tracking.[5]

swin et al. 2020 use time series data over the last 50 years for forecasting suicide prediction and use ARIMA According to a study, financial incentives for public health, welfare, education, and employment may have led to a decrease in suicide mortality in India.[6]

Linthicum et al. (2019) wrote that Supervised learning entails training the model on labelled data before applying it to new data to make predictions. It entails dividing data into two sets: a training set and a testing set. The model is first trained on the training set, and then its performance is evaluated on the testing set. Performance metrics can be used to assess the model's performance. The problem of supervised learning can be classified as a classification or regression problem. The labelled value in supervised classification is a discrete value. The algorithms in this section are used to classify the problem into which class or category it belongs. On the other hand, supervised regression learning uses models to predict outcomes based on continuous (numeric) data.[7]

Walsh et al. (2017) Machine learning could be very useful in combining the many minor to moderate impacts seen in the field from risk variables and correlations. Long-term and short-term risk can both be identified using predictive analytical models, but each requires a distinct methodology. [8]

Several attempts have been undertaken to identify risk factors for suicidal thoughts and acts, as well as to forecast the likelihood of future suicide attempts. Although a few high-performing models were reported in studies where cohorts were enriched for cases, prediction accuracy was limited when applied to a general population with an exceptionally low rate of suicide attempts.[8][9][10]

Iliou et al. (2016) The major focus of this research is to build and propose a machine learning technique for predicting if a depressed patient is likely to commit suicide. One of the study's primary conclusions is that depression symptoms in adolescents as young as 14-15 years old can predict suicide behaviour. They used hybrid approaches (PCA/Evolutionary search feature selection - ANN, RBF, Random forest, Decision tree, IB) and discovered that IB and ANN, which used a new feature selection strategy as a pre-processing step, had the best accuracy of 93.75 and 92.18 percent, respectively.[11]

Jiang et al. (2021) researches suicide prediction in men and women, with the goal of predicting suicide among those who are depressed. employed data from Danish medical and social registries, and used the Classification trees and Random forest models to discover that

men and women have different perceptions of suicide when using depression as a dependent variable to determine risk. And it was discovered in men (n=96, risk=81%) and women (n=338, risk=58%). [12]

Zhang et al. (2021) The major goal of this study is to develop and verify a prediction model for the individual risk of suicide after a lung cancer diagnosis. The data came from the SEER (Surveillance, Epidemiology, and End Results) database, which is based on patients diagnosed with lung cancer between 2007 and 2016. They used univariate and multivariate Cox regression models to predict the outcome, and the results showed that males with advanced tumor growth, diagnosis of small carcinoma, and refusal of radiation have a higher suicide risk.[13]

A multi-level classification and regression tree (CART) was used to predict proximally developing suicidal ideation, operationalized as next-day suicidal ideation, during a high-risk post-discharge phase in this daily diary research of psychiatrically hospitalized teenagers. Because of its capacity to deliver easily interpretable results, the multi-level CART was chosen above other machine learning algorithms (e.g. random forest, elastic net, neural network). Boudreaux et al. (2021) [14]

Suicide is one of the most serious problems in the world. With each passing year, the total number of people who have committed suicide rises. It is expected to happen. that roughly 800 people died as a result of various factors Thousands of people die each year when attempting suicide, and an estimation model was constructed using their data. By supplying the nine qualities, the researchers were able to predict a linear relationship. Estimation accuracy of 99 percent. Podlogar et al. (2018) [15]

2.1 SUMMARY OF MOSTLY RELATED WORK

A few research papers are closely related to this work stated in table 2.2.1.

2.1.1 Table of mostly related work

Authors	Title	Objective	Year

Linthicum et al.	Machine learning in suicide science: Applications and ethics	comparing Supervised ml against traditional method	2018
Swain et al.	Forecasting suicide rates in India: An empirical exposition	To predict a reasonably consistent pattern of suicide in India	2021
Belsher et al.	Prediction Models for Suicide Attempts and Deaths A Systematic Review and Simulation	reviewed a total 7306 abstract and 64 unique prediction model and found classification result is good	2019

This paper concentrates on finding the best classification model for predicting suicide and the proper attribute to predict suicide using simple supervised machine learning technique (classification and regression) in Machine Learning algorithms.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Methodology model

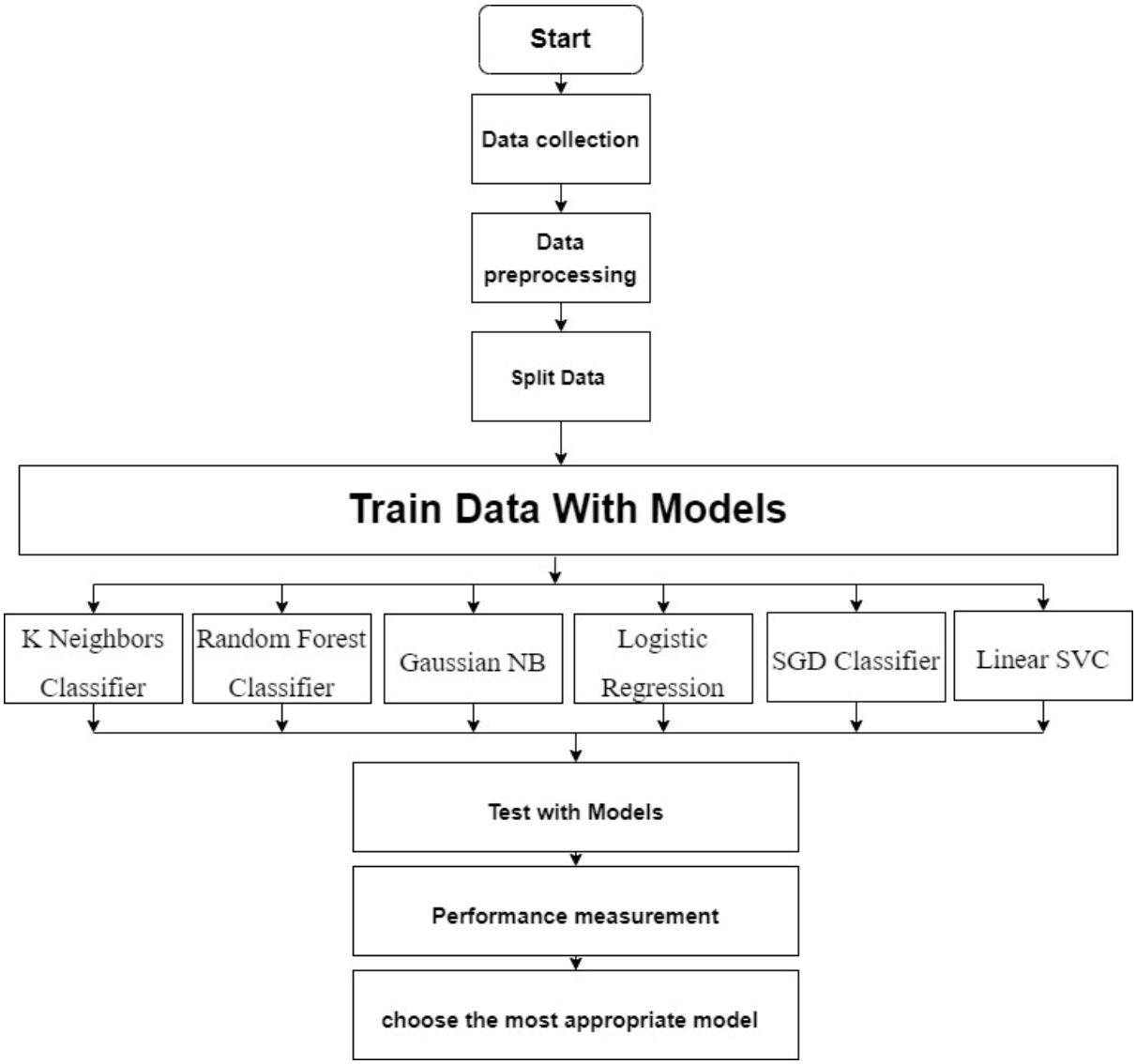


Figure 1: Methodology Model

3.2 DATASET

The data was taken from Kaggle. From 1985 to 2016, this dataset was built from four different datasets that were linked by time and place. WHO, World Bank, UNDP, and Kaggle dataset are the sources of the datasets This dataset contains 27820 Number of Instances and 13 features which are continent, country, year, gender, age_group, suicide_count, population, suicide_rate, country-year, HDI for year, gdp_for_year, gdp_per_capita, generation. The dataset contains data from numerous country like Argentina, Australia, Belgium, Brazil, Canada, Colombia, Denmark, Finland, France, Germany, Italy, japan, Netherland, New Zealand, Norway, Poland, Russia, South Africa, Spain, Sri Lanka, Sweden, Switzerland, Thailand, Turkey, United Arab Emirates, United Kingdom, United States and many more also the dataset contains

3.3 PRE-PROCESSING

Machine learning requires pre-processing, which includes data cleansing and standardization, noisy data filtering, and missing information management. The data was obtained from the Kaggle website. Combine all of the traits that have been connected to the prediction of suicide. when we start checking of null value, we get HDI for year has 19456 null values out of 27820 samples which is approximately 70% of the column data. This may tamper the model performance so, dropping the HDI for year column from dataset also, we observed in the dataset is ‘country-year’ column. This is just a combination of country and year columns which doesn’t have a significance to the model. So, dropping the ‘country-year’ also, Cleaning data in gdp_for_year column & converting it to float and Standardizing data with Robust Scalar and Encoding data with categorical Label Encoder. The data collection is separated when independent and dependent features are defined. 80 percent of the data should be used for training and 20% for testing.

3.4 Level Generation:

For better prediction result and accuracy, we do an level on data set in this leveling process we use Suicide rate. In level processing time we define it in three part. we assume those level as low, Medium and High and the level partition rate is <2 as Low, 3-5 as medium and >5 as High.

3.5 DATA VISUALIZATION AND STATISTICS

The visual depiction of data and information is known as data visualization. Data visualization tools make it easy to examine and comprehend trends, oddities, and relationships in the data by employing visual cues like charts, graphs, and maps.[28]

The dataset is shown using the well-known matplotlib and seaborn tools to create a few graphs/plots. The plots are displayed below. Individual bar graphs are created to visualize the distribution of all attributes in a dataset.

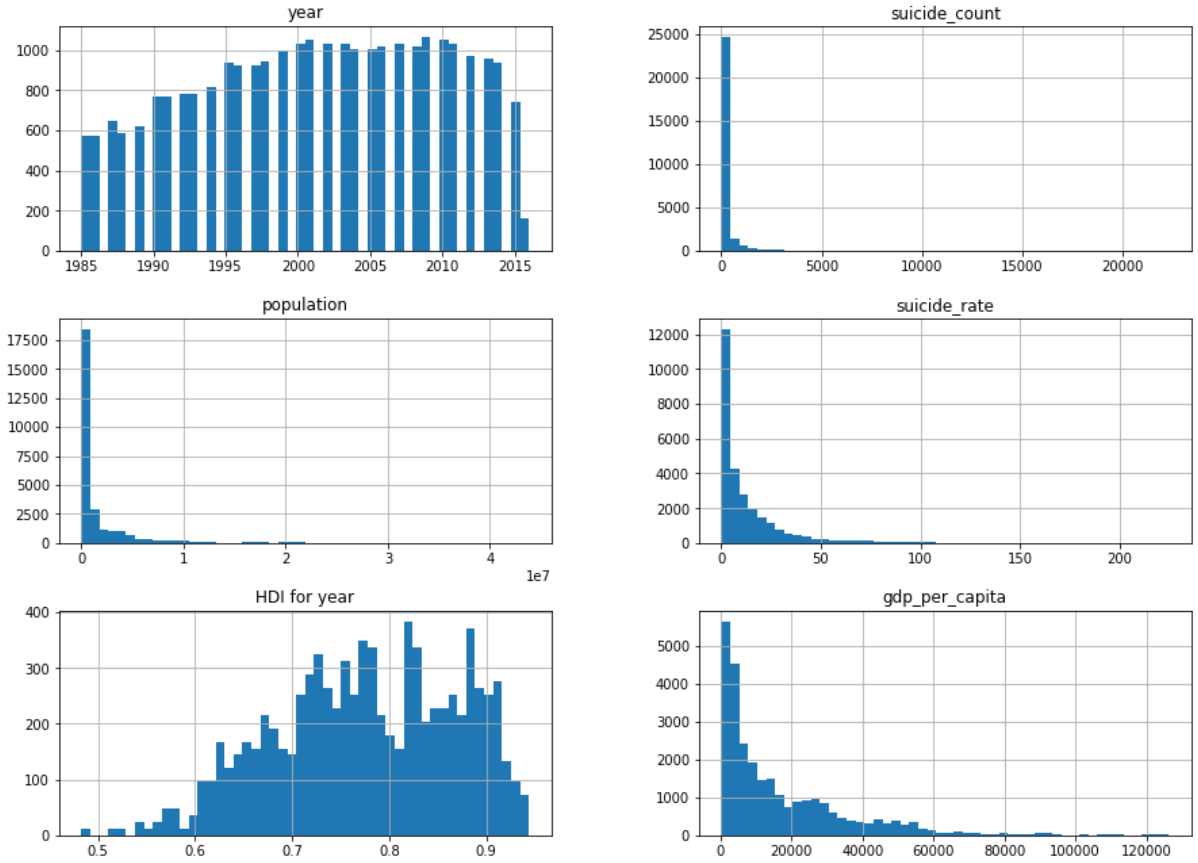


Figure 2: Distribution graphs of features in the dataset

A correlation heatmap seems to be a heatmap which depicts a two-dimensional correlation matrix among two separate dimensions, with coloured cells representing data on a monochromatic level. The very first dimension's numbers display as rows in the table, while the other dimension's numbers show as columns. For this data set A correlation heatmap is created to visualize the relationship between each dataset attribute.[27]

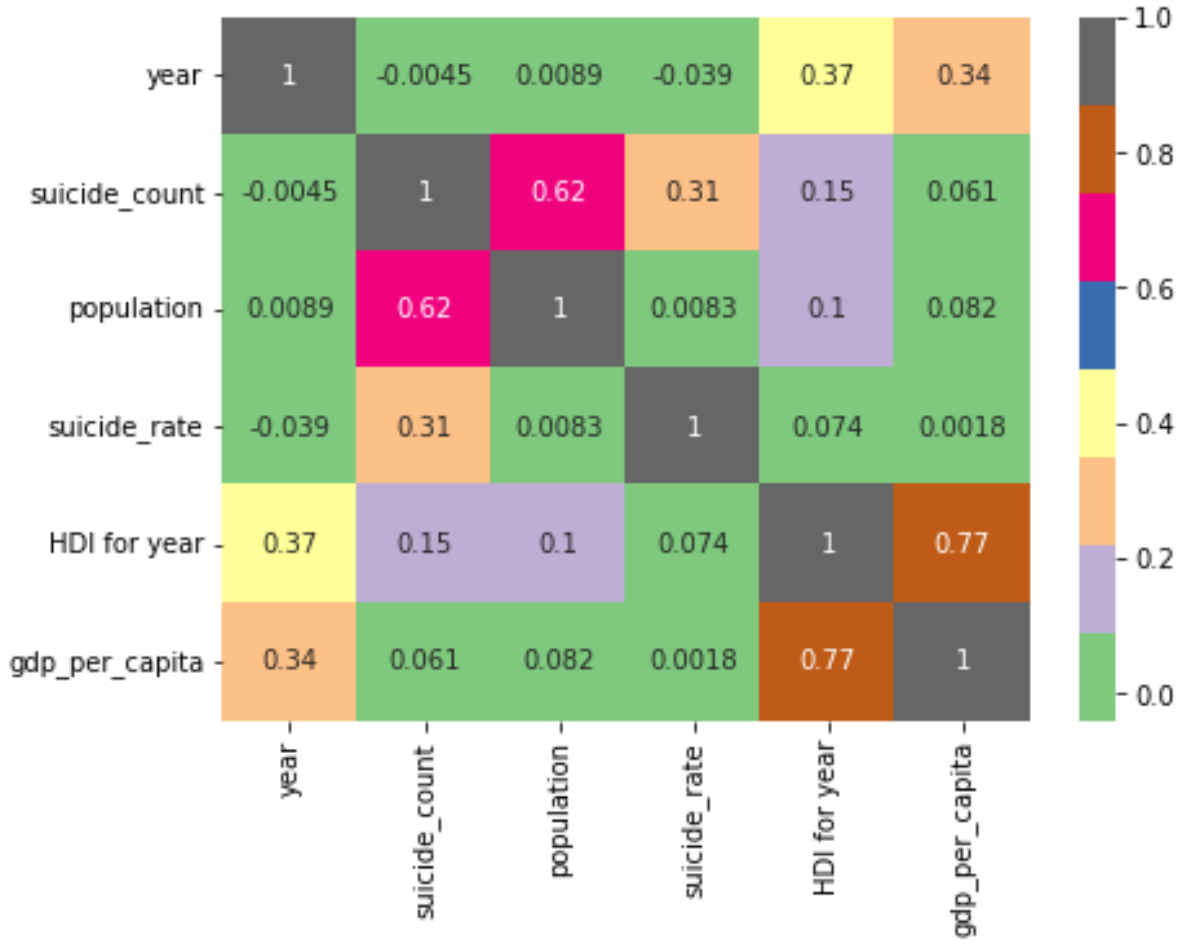


Figure 3: Heatmap of Dataset

The bar plot depicts the number of victims among male and female populations, and that we can deduce that men are far more likely to commit suicide than females.

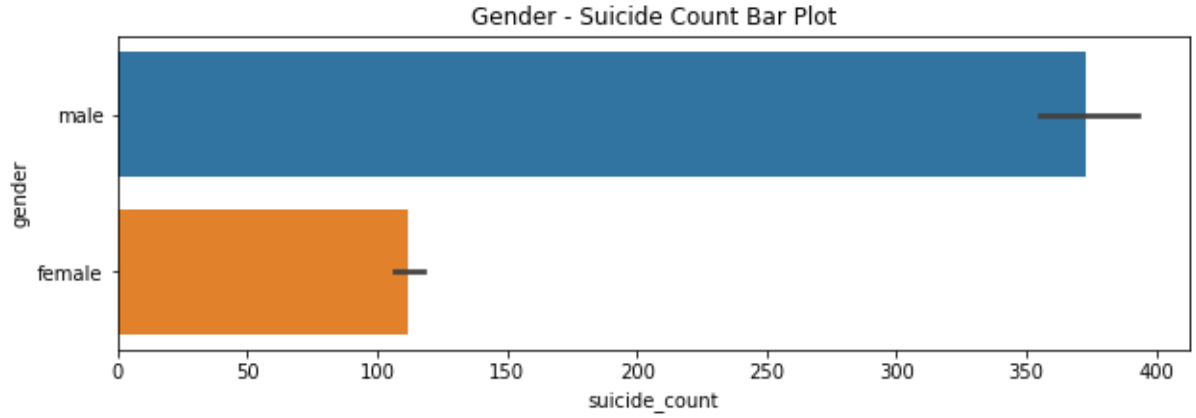


Figure 4: Bar plot Gender – Suicide count

The boxplot illustrates that the age group 35-54 years has the highest number of suicide cases, followed by 55-74 years. Surprisingly, even while the number of suicide cases in the 5-14-year age group is quite low, it is still in the tens. Let's look at how the number of suicides has changed through time.

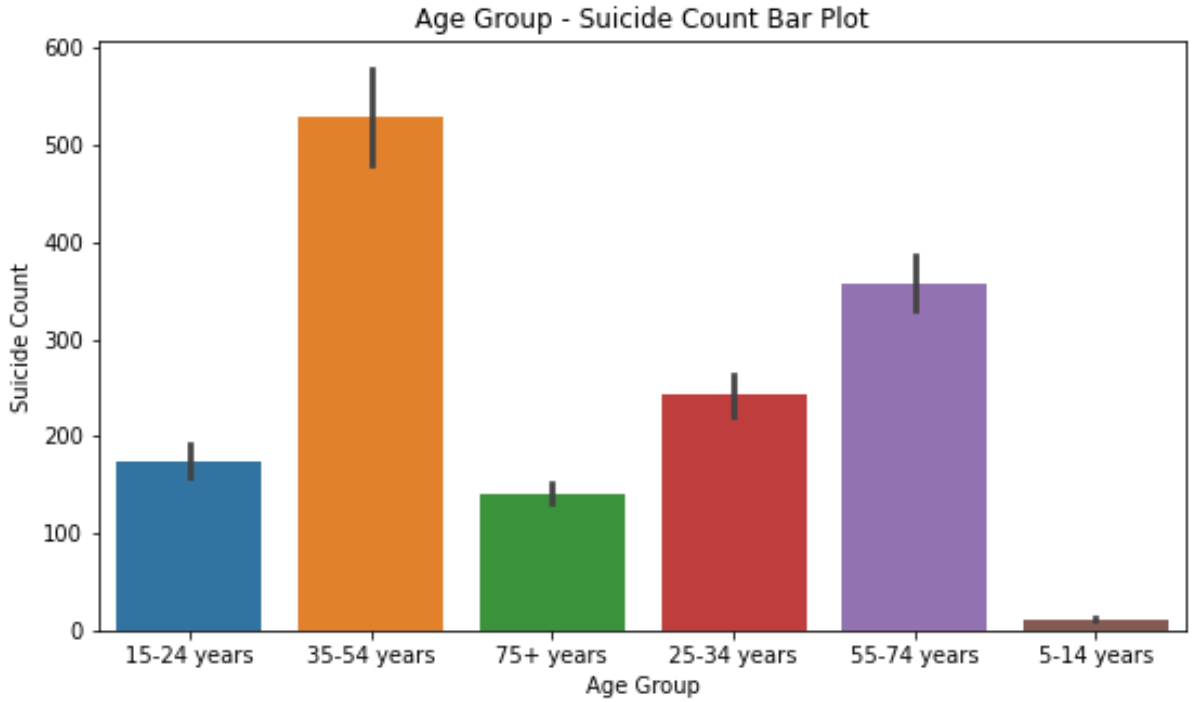


Figure 5: Bar plot of Generation & suicide count grouped by gender

The following are some observations made from the plot. The boomer, silent, and X generations are more affected. Based on the information supplied, these generations are made up of people born between 1946 and 1976. Upon further examination, these generations are the ones in which the majority of their members are in the age period in which the majority of suicides occur.

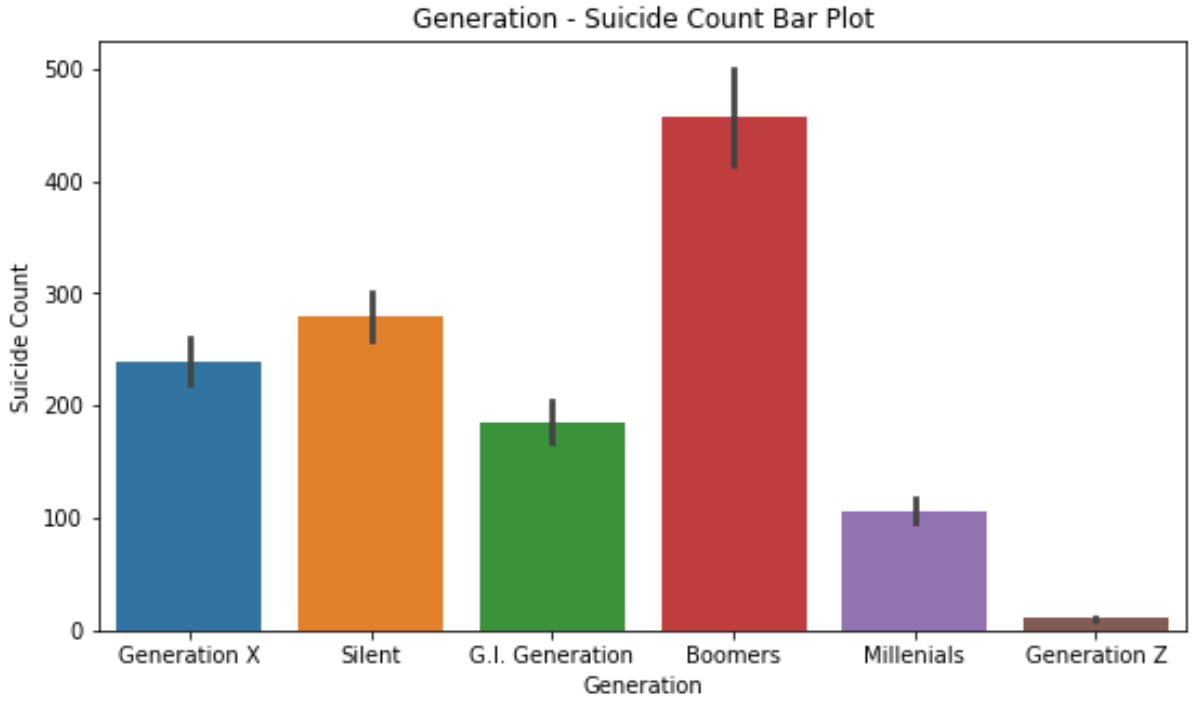


Figure 6: Bar plot of Generation & suicide count grouped by gender

Let's examine if the above-mentioned pattern holds true across all age groups, generations, and genders. According to the bar plot, the 35-54-year age group is more prone to suicides, regardless of gender, followed by the 55-74-year age group, regardless of gender.

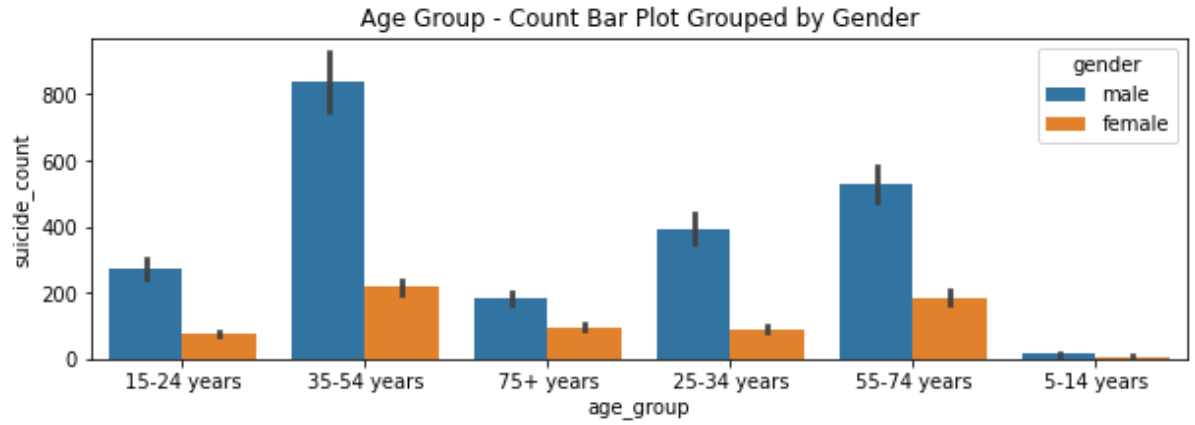


Figure 7: Bar plot of Age group & suicide count grouped by

According to the bar graph below, the Boomers generation has the most suicide cases, followed by the Silent generation, regardless of gender. It is evident from the above four bar plots that men commit suicide far more frequently than women, regardless of their age group or generation.

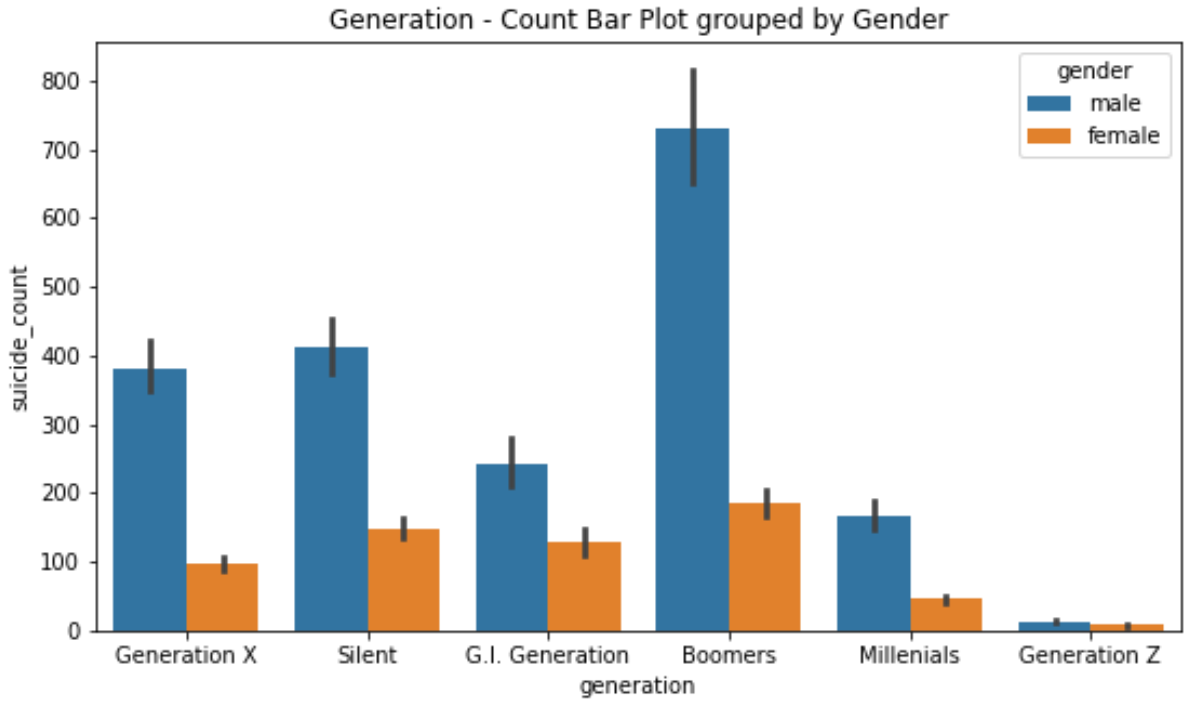


Figure 8: Bar plot of Generation & suicide count grouped by gender

The bar plot shows suicide rate in continent and it shows that Europe continent have more suicide rate than others continent.

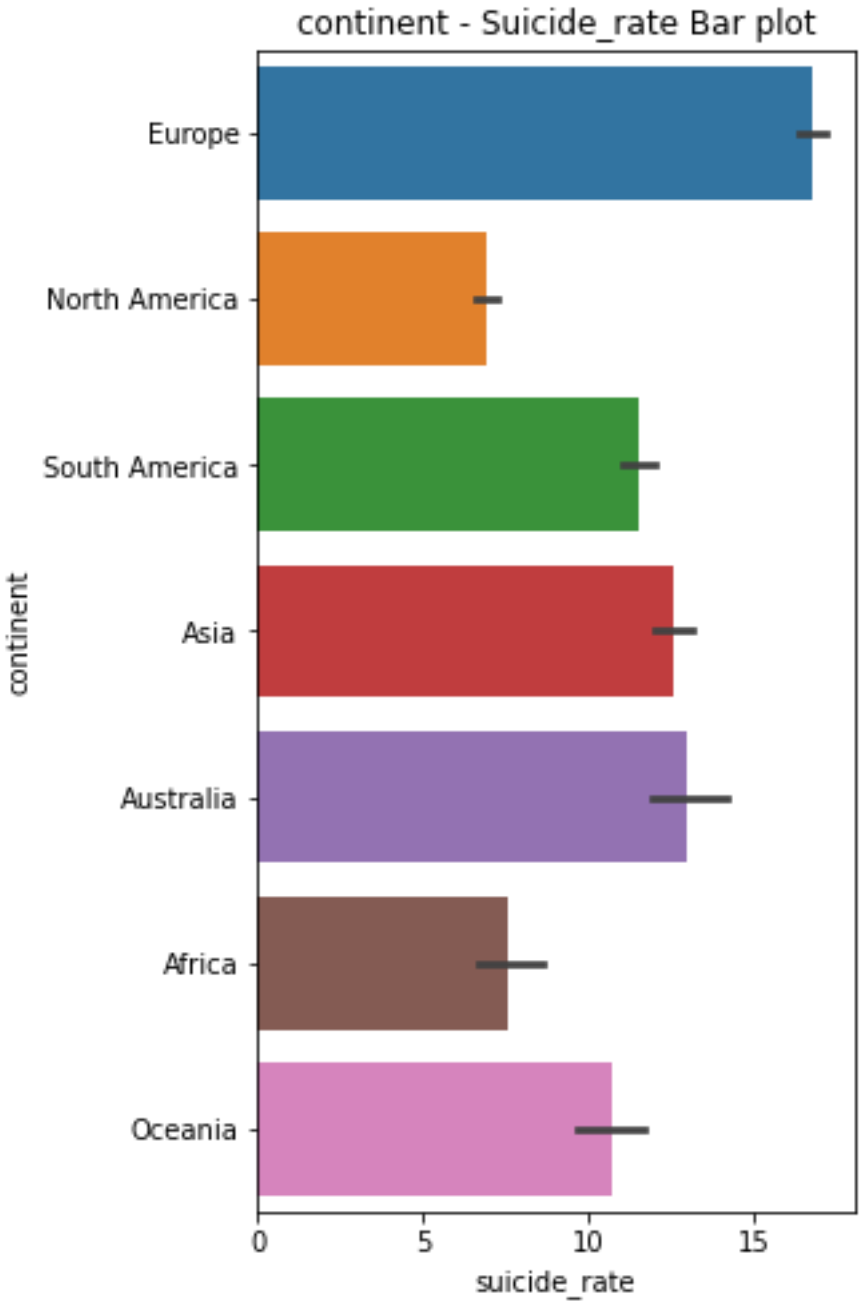


Figure 9: Bar plot of Continent & suicide rate

The bar plot show that Asia continent have most suicide count which is generate by gender it shows that Asia continent male suicide count than others.

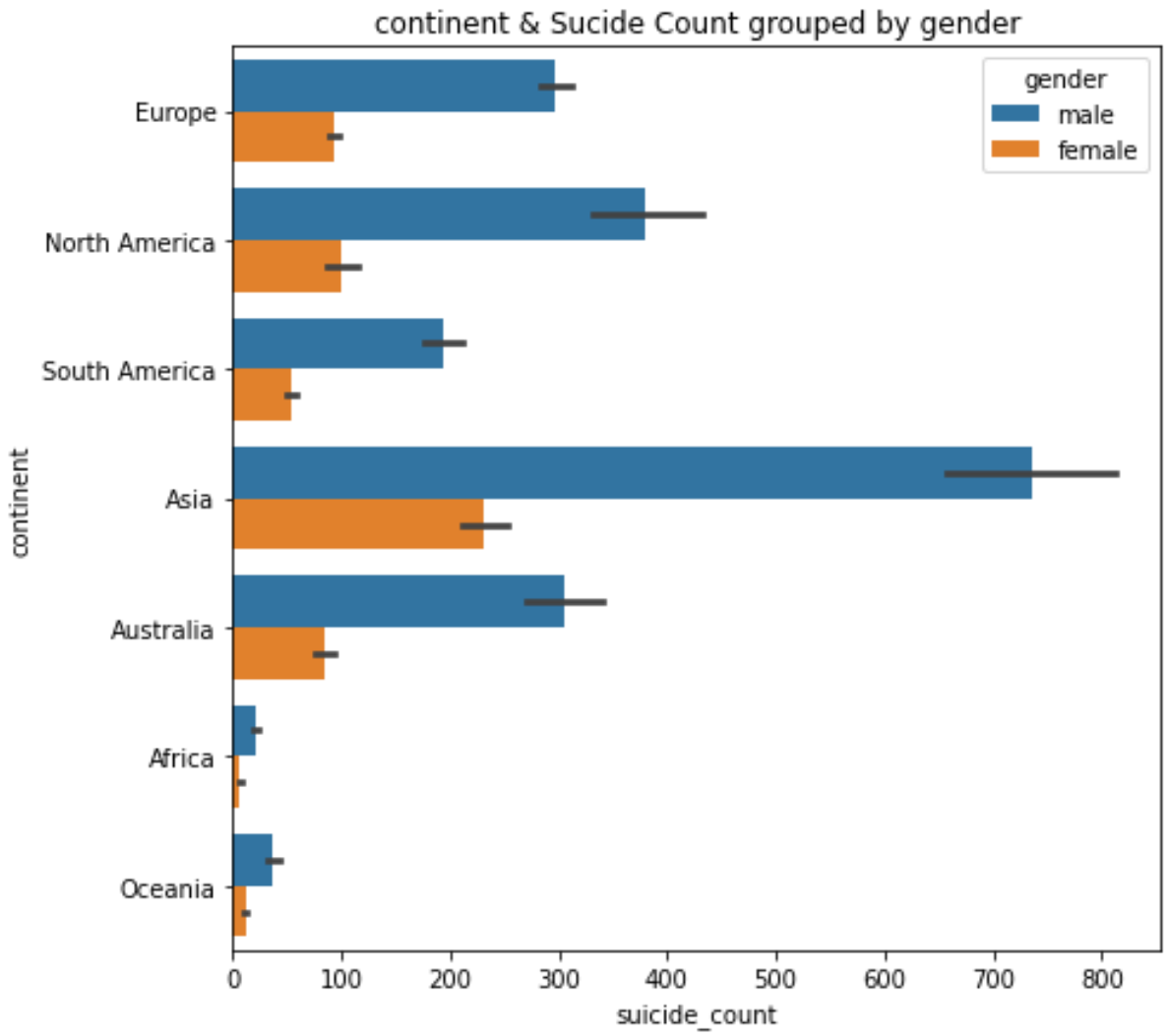


Figure 10: Bar diagram of Continent and Suicide grouped by Gender

The bar plot shows the continent and suicide count grouped by age group and it shows that Asia continent have the 35-54-year age group is more prone to suicides than others age group

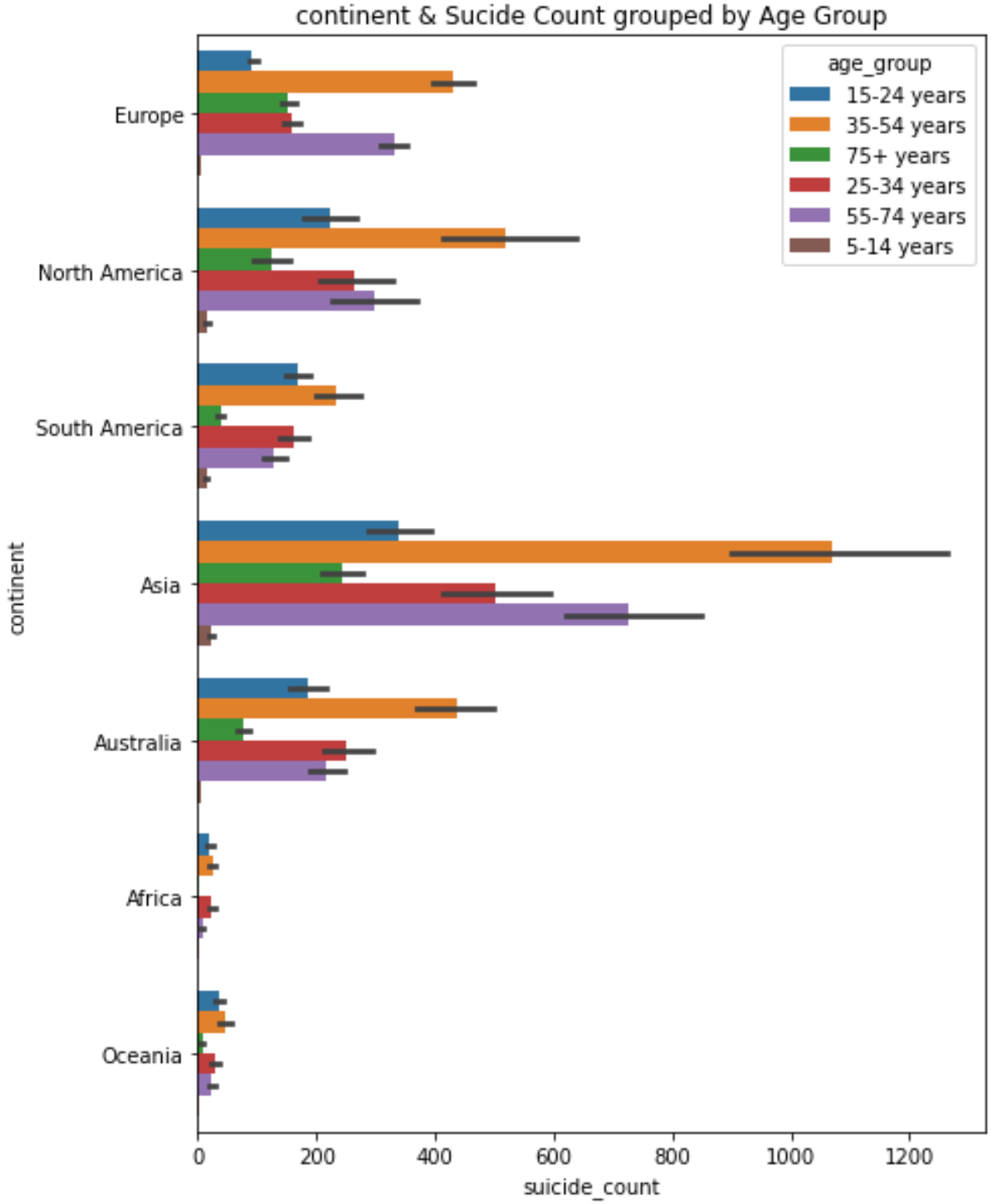


Figure 11: Continent and Suicide count grouped by Age group

The plot is about the countries and their suicide rate. Lithuania has the highest suicide rate, followed by Sri Lanka, as shown in the bar graph below.

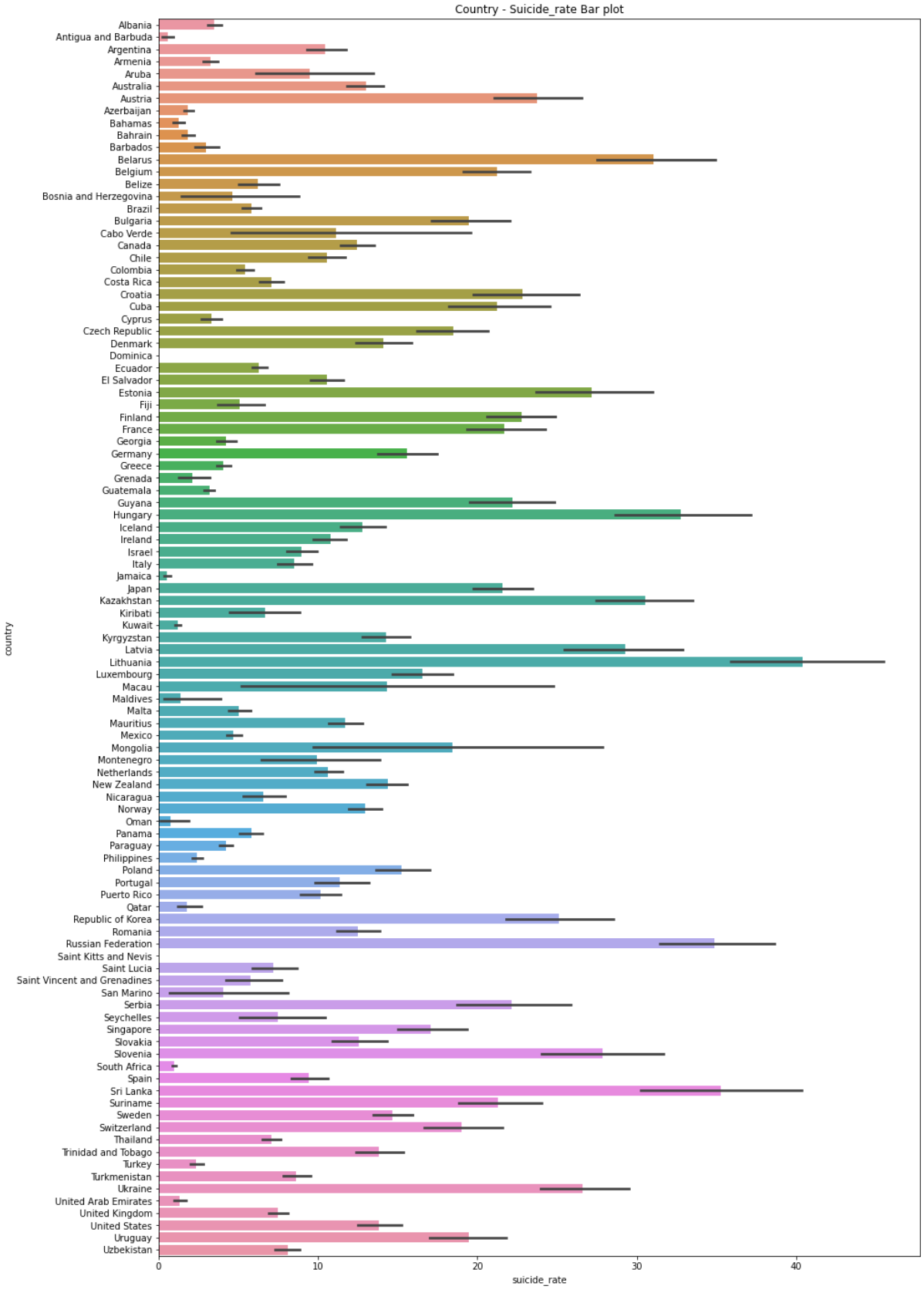


Figure 12: Bar plot of Countries & Suicide rate

This below bar plot shows the Continent wise Suicide rate which is level by High, Medium and low and 0-6 are continent.

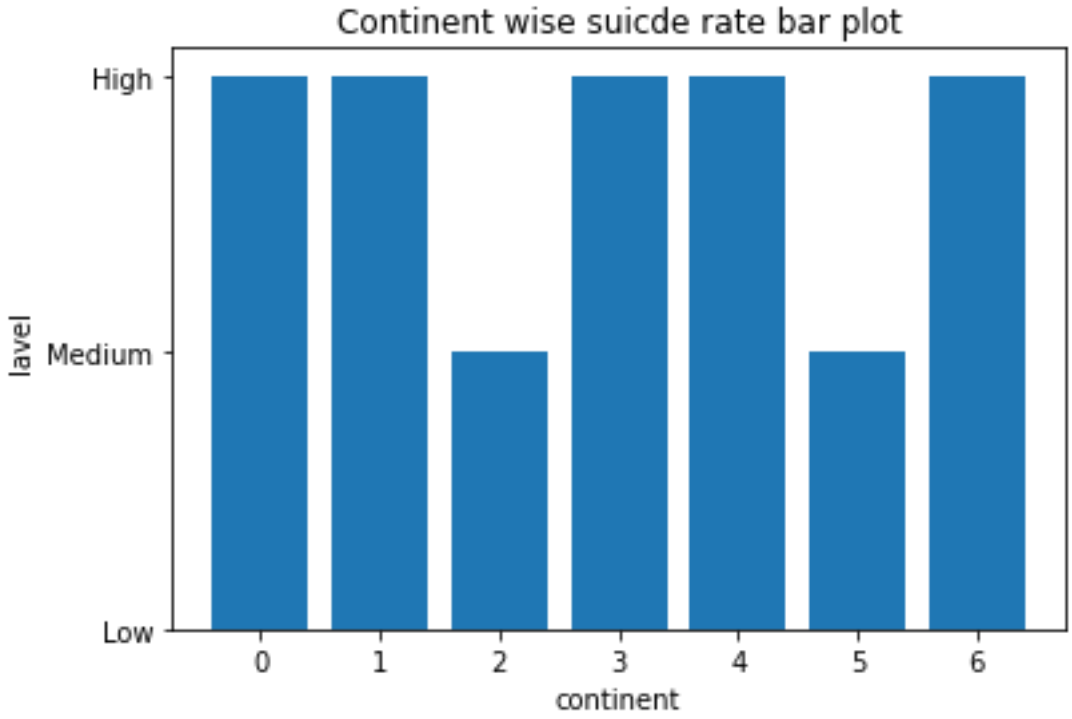


Figure 13: Continent wise Suicide rate

According to the following graph, the suicide rate has been steadily increasing since 1990, and it has been steadily decreasing since 2016. The information was gathered in early 2016. As a result, the dataset does not contain all of the 2016 suicide cases.

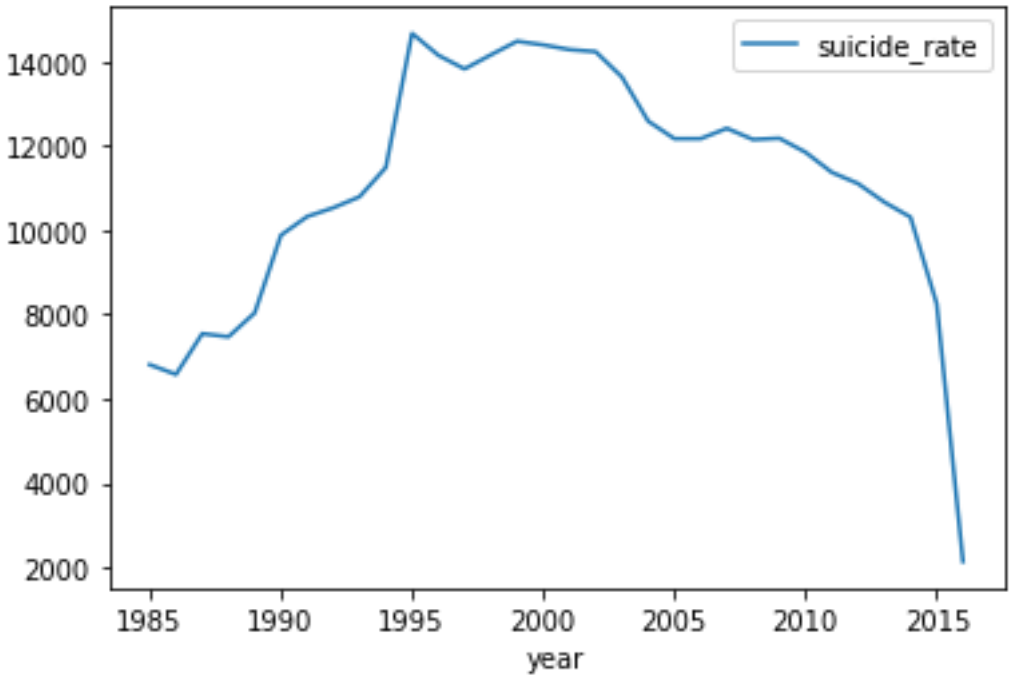


Figure 14: Line plot of Years & Suicide rate

It is clear from the scatter matrix that the data contains outliers, which are dealt with during data pre-processing by scaling and encoding the features.

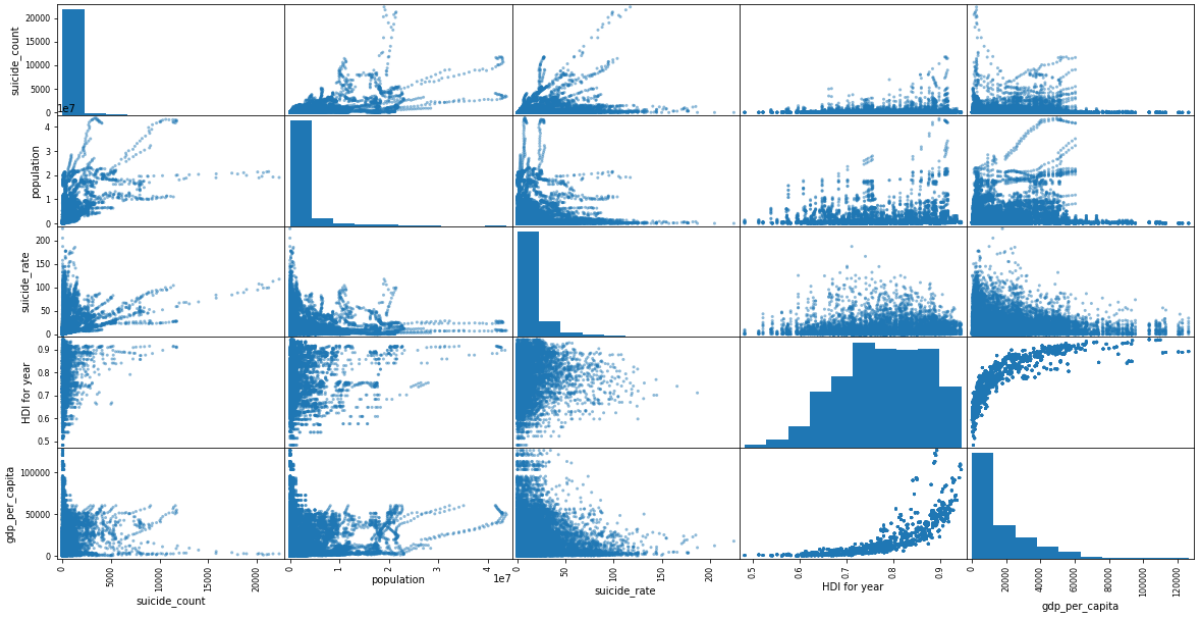


Figure 15: Scatter Matrix of Dataset

3.6 Estimator Selection:

We selected six machine-learning algorithms in our research. This is shown in the table below.

K Neighbors Classifier
Random Forest Classifier
Gaussian NB
Logistic Regression
SGD Classifier
Linear SVC

Table 1: Estimator Selection

3.7 K Neighbour:

The K-Nearest Neighbour algorithm is based on the Supervised Learning technique and is one of the most basic Machine Learning algorithms. K-NN method assumes that the new case/data and existing cases are similar and places the new case in the category that is most similar to the existing categories. K-NN method stores all available data and classifies a new data point based on its similarity to the existing data. This means that new data can be quickly sorted into a well-defined category using the K-NN method. K-NN approach can be used for both regression and classification, but it is more commonly utilized for classification tasks. for K Neighbors Classifier Assume that x is the independent variable and that y is the dependent variable. As a result, given x and y, the linear regression is

3.8 Random Forest Classifier:

The Xbox Kinect data was utilized to create a random forest, which is a collection of randomized decision trees used to distinguish human position [27]. A random selection of training instances and a random subset of characteristics are used to learn each decision tree in the forest. The results from each decision tree are averaged to determine the total outcome when classifying a test example. Each tree is walked through until it reaches a leaf node. The ratio of training instances of each activity type that belong to the leaf node is used to generate a likelihood score. To get an overall probability score for the example, these probability scores are averaged over each tree in the forest. Finally, for that case, the activity type with the highest likelihood is projected.

3.9 Gaussian NB:

The Gaussian Naive Bayes law is one of the most popular and fascinating theorems to study when working with statistics and probability. When working with statistics, the Bayesian theorem allows you to determine the likelihood that an event will occur if you have prior knowledge and information about the occurrence. The Gaussian Naive Bayes theorem appears as follows when converted from text to a mathematical representation [28]

3.10 Logistic Regression:

Logistic regression is a statistical analysis approach for predicting a data value based on previous data set observations. In the field of machine learning, logistic regression has become an important technique. The method enables a machine learning application to

classify incoming data using an algorithm based on historical data. The program should get better at guessing classes within data sets as more relevant data comes in. Logistic regression can also help with data preparation by allowing data sets to be placed into specified buckets during the extract, transform, and load (ETL) process, allowing the information to be staged for analysis.

3.11 SGD Classifier

SGD (Stochastic Gradient Descent) is a straightforward but effective optimization approach for determining the values of parameters/coefficients of functions that minimize a cost function. In other words, it's utilized to learn discriminative linear classifiers with convex loss functions like SVM and Logistic regression. Because the update to the coefficients is conducted for each training instance rather than at the end of examples, it has been successfully used to large-scale datasets. The Stochastic Gradient Descent (SGD) classifier is just a conventional SGD learning method that supports various loss functions and classification penalties. To implement SGD classification, Scikit-learn provides the SGD Classifier module.[29]

3.12 Linear SVC

A Linear SVC (Support Vector Classifier) is designed to fit to the data you provide and provide a "best fit" hyperplane that divides or categorizes your data. Following that, you may input some features to your classifier to check what the "predicted" class is after you've obtained the hyperplane. As a result, this algorithm is a good fit for our needs.

3.13 Performance Calculation

For each class, we examine Accuracy score, Precision value, F1 Score, Recall, and Support in order to select the best model for our proposed system “(1-4)”.

1 Precision = True Positives / (True Positives + False Positives)
2 Recall = True Positives / (True Positives + False Negatives)
3 F1 value = (2 * Precision * Recall) / (Precision + Recall)
4 Accuracy= (prediction of correct / prediction of all)

CHAPTER 4

4.1 Analysis Technique

In a web application called Google Colab Notebook. Python is the programming language I'm utilizing for the mentioned language, library, and visualization tools. All of this is built into the Colab notebook, which is a free Jupyter notebook environment. The script or markdown code language is incorporated in the final content of each cell of a document. Features, tables, charts, and graphs are common outputs. This technology makes it easier to interchange and reproduce scientific works because tests and results are presented in a self-contained format. The Google Colab is an initiative focused to machine learning research and education. Google Colab works in a similar way to Microsoft Office items in that it can be shared and numerous users can work on the same notebook at once. Pandas, numpy, seaborn, matplotlib, and sklearn are a few of the machine learning and AI libraries that no need to install. The virtual machine becomes dormant after a period of time under runtime (VM), and all user data and configuration are lost. The notebook, on the other hand, is secure, and files can be moved from the VM hard disk to the user's Google Drive account at Daffodil International University. Finally, this google colab is a GPU-accelerated runtime that fully integrates the previously stated technologies.

4.2 Training process

About 80% of the data from my dataset is used as training data, while the remaining 20% is used to test my models and demonstrate model accuracy for this dataset. These validation results show how well my model performed in this dataset and how accurate its predictions were.

4.3 Model Result:

In this section, I will discuss the results of my models before deciding which model is the best for malware detection.

Models name	Train Accuracy	Test Accuracy
K-Neighbors Classifier	1.0	0.977
Random Forest Classifier	0.998	0.993
Gaussian Naive Bayes	0.968	0.966
Logistic Regression	0.972	0.970
Stochastic Gradient Descent Classifier	0.981	0.978
Linear Support Vector Classifier	0.982	0.980

Table 2: Accuracy Table

[Table:2, Figure 16] shows that the Random Forest Classifier has the highest accuracy, and there is no doubt that this model is one of the finest among the classification methods. As a result, I declare that this model is the most accurate in predicting accuracy. Furthermore, this model outperforms others in terms of prediction accuracy based on all available features.

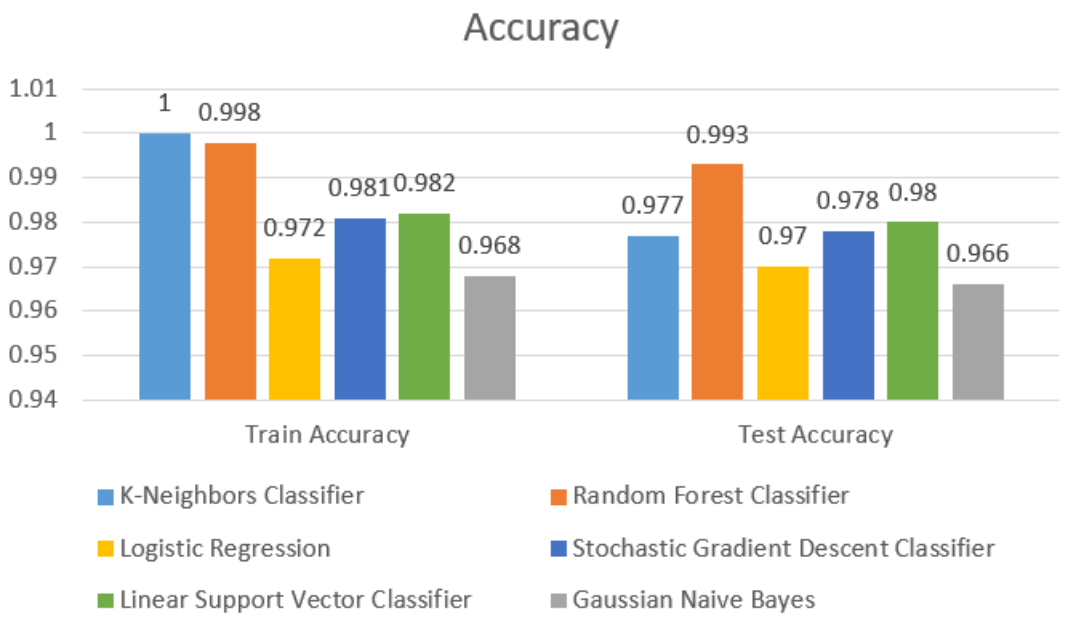


Figure 16: Train and Test Accuracy

4.4 Model evaluation:

In this section, I'll discuss how accurate suicide prediction is. All evaluation results are listed below, and they were discovered after training all of the specified models:

Models name	Precision	Recall	f1-score	Class
K-Neighbors Classifier	0.99	0.99	0.99	Low
	0.86	0.82	0.84	Medium
	0.82	0.81	0.81	High
Random Forest Classifier	1.00	1.00	1.00	Low
	0.96	0.96	0.96	Medium
	0.90	0.88	0.89	High
Gaussian Naive Bayes	1.00	0.98	0.99	Low
	0.72	0.88	0.79	Medium
	0.70	0.87	0.77	High
Logistic Regression	0.99	0.99	0.99	Low
	0.79	0.80	0.80	Medium
	0.72	0.70	0.71	High
Stochastic Gradient Descent Classifier	0.99	1.00	0.99	Low
	0.85	0.86	0.85	Medium
	0.93	0.56	0.70	High
Linear Support Vector Classifier	0.98	1.00	0.99	Low
	0.98	0.74	0.84	Medium
	0.89	0.94	0.92	High

Table 3: Model Evaluation Table

From the test dataset, we can see all precision, recall, and fi-scores [Table 4.3.2] for each class for suicide prediction. Here it is also stated that the Random Forest Classifier is the most accurate of the models. Here in this model we have 100% precision and 100% recall for the Low class also we have 96% precision and 96% recall for the Medium class and 90% precision with 88% recall for the High class. That why this is the best model with 99% accuracy [Table 4.3.1] for suicide prediction.

CHAPTER 5

CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

In today's world, suicide is a lethal sickness. As long as people continue killing their lives with their own hands, nothing will change. Approximately one million people commit suicide each year, according to the report. Suicides among adolescents under the age of 35 are the highest in the world. The Random Forest Classifier is the most accurate of the classification algorithms, and there is little doubt that this model is one of the best. The most accurate model is the Random Forest Classifier. We have 100 percent precision and 100 percent recall for the Low class, 96 percent precision and 96 percent recall for the Medium class, and 90 percent precision and 88 percent recall for the High class in our model.

5.2 Feature work

I plan to continue optimizing those classification models in the future, as well as build the deep learning classification model algorithm. In addition, I want to use time complexity to determine the model's performance.

Reference

[1] Bolton, J. M., Gunnell, D., & Turecki, G. (2015). Suicide risk assessment and intervention in people with mental illness. Bmj, 351.

[2] Ghasemi, P., Shaghaghi, A., & Allahverdipour, H. (2015). Measurement scales of suicidal ideation and attitudes: a systematic review article. Health promotion perspectives, 5(3), 156.

[3] Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., ... & Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: a systematic review and simulation. JAMA psychiatry, 76(6), 642-651.

[4] Marks, M. (2019). Artificial intelligence-based suicide prediction.

[5] D'Hotman, D., & Loh, E. (2020). AI enabled suicide prediction tools: a qualitative narrative review. BMJ Health & Care Informatics, 27(3).

[6] Swain, P. K., Tripathy, M. R., Priyadarshini, S., & Acharya, S. K. (2021). Forecasting suicide rates in India: An empirical exposition. PLoS one, 16(7), e0255342.

[7] Linthicum, K. P., Schafer, K. M., & Ribeiro, J. D. (2019). Machine learning in suicide science: Applications and ethics. Behavioral sciences & the law, 37(3), 214-222.

[8] Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. Clinical Psychological Science, 5(3), 457-469

[9] Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., ... & Reis, B. Y. (2017). Predicting suicidal behavior from longitudinal electronic health records. American journal of psychiatry, 174(2), 154-162.

[10] Ryu, S., Lee, H., Lee, D. K., & Park, K. (2018). Use of a machine learning algorithm to predict individuals with suicide ideation in the general population. Psychiatry investigation, 15(11), 1030.

[11] Iliou, T., Konstantopoulou, G., Ntekouli, M., Lymberopoulos, D., Assimakopoulos, K., Galiatsatos, D., & Anastassopoulos, G. (2016, September). Machine learning preprocessing method for suicide prediction. In IFIP International Conference on Artificial Intelligence Applications and Innovations (pp. 53-60). Springer, Cham.

[12] Jiang, T., Nagy, D., Rosellini, A. J., Horváth-Puhó, E., Keyes, K. M., Lash, T. L., ... & Gradus, J. L. (2021). Suicide prediction among men and women with depression: A population-based study. Journal of psychiatric research, 142, 275-282.

[13] Zhang, X., Sun, S., Peng, P., Ma, F., & Tang, F. (2021). Prediction of risk of suicide death among lung cancer patients after the cancer diagnosis. Journal of affective disorders.

[14] Boudreaux, E. D., Rundensteiner, E., Liu, F., Wang, B., Larkin, C., Agu, E., ... & Davis-Martin, R. E. (2021). Applying machine learning approaches to suicide prediction using healthcare data: Overview and future directions. Frontiers in psychiatry, 1301.

[15] Podlogar, M. C., Gai, A. R., Schneider, M., Hagan, C. R., & Joiner, T. E. (2018). Advancing the prediction and prevention of murder-suicide. Journal of Aggression, Conflict and Peace Research.

[16] Müller, A. C. (2016, October 1). Introduction to Machine Learning with Python. O'Reilly Online Learning. https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/

[17] Brownlee, J. (2020, August 27). How to Grid Search Hyperparameters for Deep Learning Models in Python With Keras. Machine Learning Mastery. https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/

[18] Brownlee, J. (2020a, April 24). 7 Step Mini-Course to Get Started with XGBoost in Python. Machine Learning Mastery. https://machinelearningmastery.com/xgboost-python-mini-course/

[19] D. (2018, April 15). REGRESSION PROBLEMS IN PYTHON | Data Vedas. Www.Datavedas.Com. https://www.datavedas.com/regression-problems-in-python/

[20] Brownlee, J. (2021, August 24). How to Develop Super Learner Ensembles in Python. Machine Learning Mastery. https://machinelearningmastery.com/super-learner-ensemble-in-python/

[21] Bevans, R. (2021, September 16). Statistical tests: which one should you use? Scribbr. https://www.scribbr.com/statistics/statistical-tests/

[22] Tran, N. (2021, December 8). Fundamental of The Chi Square in Statistics - Nhan Tran. Medium. https://medium.com/@nhan.tran/the-chi-square-statistic-p-1-37a8eb2f27bb

[23] Dipu, M. (2021, November 15). What is machine learning? The way the brain of the machine will change the world. DroidXplore. https://droidxplore.com/%E0%A6%AE%E0%A7%87%E0%A6%B6%E0%A6%BF%E0%A6%A8-%E0%A6%B2%E0%A6%BE%E0%A6%B0%E0%A7%8D%E0%A6%A8%E0%A6%BF%E0%A6%82/

[24] What Regression Measures. (2021, October 30). Investopedia. https://www.investopedia.com/terms/r/regression.asp

[25] DSS - Introduction to Regression. (n.d.). Dss.Princeton.Edu. https://dss.princeton.edu/online_help/analysis/regression_intro.htm

[26] Waseem, M. (2021, December 16). How To Implement Classification In Machine Learning? Edureka. https://www.edureka.co/blog/classification-in-machine-learning/?fbclid=IwAR2EMqS-UoIvpcXJvUsNJqwEX9FkTBI6uWdGDahuqkagtRFcAYxBjDQk42Y#algo

PLAGARISM REPORT