**PROJECT NAME: CLASSIFICATION OF TREATMENT OF MENTAL HEALTH IN THE TECH INDUSTRY**

**STUDENT: ROBIN KHAOYA WAFULA**

**SUPERVISOR: SAMUEL WAIYAKI**

**ABSTRACT**

Mental health is a very important aspect of the overall health of any individual. There are different mental illnesses eg depression, anxiety and bipolar disorder just to list but a few. These conditions, like any other illness, have the ability to affect all parts of an individual’s life, from social to economic and even physical health.

The aim of this project was to use classification models to determine wether or not people in the tech industry should seek treatment for mental illness based on different parameters.

From the study, the Ada Boost Classifier proved to be the most robust model as will be shown in later sections of this report.

**ACKNOWLEDGEMENT**

I would like to acknowledge every one whose contributions made this project a success.

I would also like to appreciate the online community of developers and data analysts whose tutorials and forum posts helped me tackle the various problems I encountered.

My heartfelt gratitude to my supervisor, Mr. Waiyaki, for his guidance and constructive feedback that enabled me to complete this project in a timely manner, meeting all the requirements of the poject scope.

**1.0 CHAPTER ONE**

**1.1 INTRODUCTION**

About 14% of the global burden of disease has been attributed to neuropsychiatric disorders, mostly due to the chronically disabling nature of depression and other common mental disorders (Prince, Martin, 2007). With the ever increasing complexity of work, and the pressure to keep up with new technologies, people in the tech industry often suffer from mental health issues. Cases of mental ilnessess have been steadily increasing over the years and it is only getting worse as time goes on.

Mental health is affected by various aspects of an individual’s life, be it social, professional or personal. Just as there are many causes of mental illness, there are also a variety of treatments and coping mechanisms to improve the quality of life of the individual. Mental disorders also manifest in different ways and in varying degrees, from truancy to aggressive behaviour.

This study focused on mental health in the tech industry and the aim was to classify whether individuals should seek treatment or not based on a number of factors for example, whether their mental health interferes with their physical health or their profession. A variety of machine learning classification models were used to achieve this.

The dataset used contained a variety of columns, like the type of companies the respondents were working in, and whether or not those companies had wellness programs for mental and physical health of their employees.

**1.2 Project Scope and Methodology**

1.2.1 Project Scope

The aim of this study was to clean the data, perform exploratory data analysis, create machine learning models to be able to classify whether or not individuals should seek treatment.

1.2.2 Methodology

The above diagram shows the steps and methodology used in this study. Each step will be discused in great detail in the chapters and sections to come.

**1.3 Tools, Programs and Requirements**

The requirements for this study were ;

* A computer / laptop
* Internet connection

There were various programs and libraries used in this project. They include ;

* An IDE (VS Code was used in this project but any other ide eg Jupyter can be used)
* Anaconda
* Python (version 3.9.2 was used in this study but any version above 2.7 will suffice)
* Python libraries for data science

An installation guide for each of the mentioned programs will be included in the appendix of this document.

**2.0 CHAPTER TWO: DATA COLLECTION**

**2.1 COLLECTION**

The dataset used in this study was not collected by me, but was provided freely by Kaggle ([www.kaggle.com](http://www.kaggle.com/)). A link to the exact dataset will be found in the appendix of this document.

The data was compiled after a survey was carried out in several countries all over the world. The survey was targeted at people working in tech. Some of the respondent worked in tech companies eg. Google, Facebook etc, some worked in non-tech companies and others were self employed.

The survey also included the work environment of the respondents, the number of collegues, welness programs and benefits provided by the companies they work for and how their mental health issues have affected their work and families.

**3.0 CHAPTER THREE: DATA CLEANING**

**3.1 Introduction**

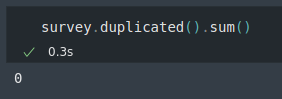
Data cleaning is used to refer to all kinds of tasks and activities to detect and repair errors in the data. (Ilyas, I. F., Chu, X. 2019). When data is collected, a lot of errors may be made in the process.

Missing values and structural errors are the most common types of errors. A respondent may choose not fill their age or gender for example, resulting in missing values in the colleected data.

Structural errors may occur for example, when one respondent fills their gender as ‘Female’ and another respondent fills ‘F’ or ‘woman’. These three values are different words that all mean the same thing. This is just one of the many ways structural errors may occur.

There is no agreed upon procedure to clean data as the nature of data varies greatly for every dataset. In the following sections, the steps followed in cleaning the data used in this study will be addressed.

**3.2 Handling of duplicates**

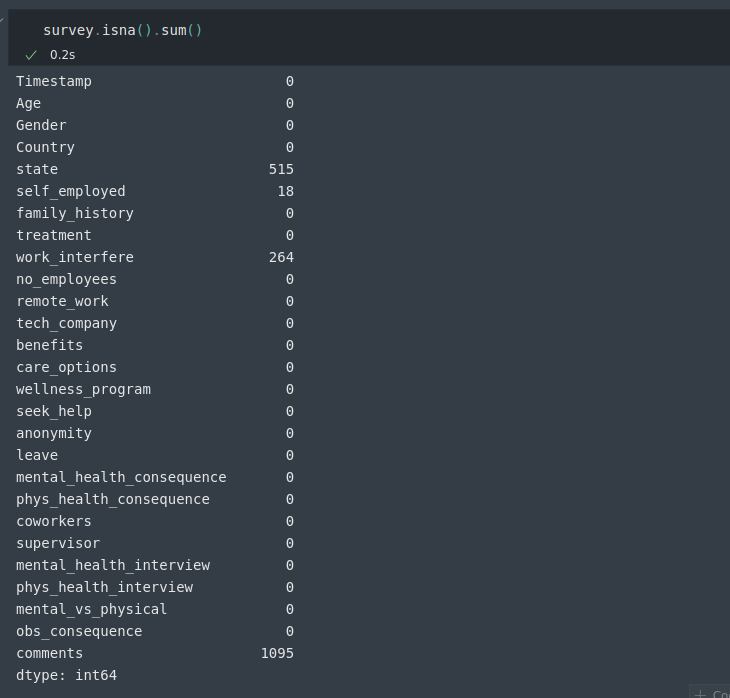
When cleaning the data, no duplicated records were found as shown in the diagram below

The duplicated() method is a built-in method included in the pandas library.

The line of code shown above returns the total number of duplicate records found in the dataset. The result is zero meaning no duplicates were found.

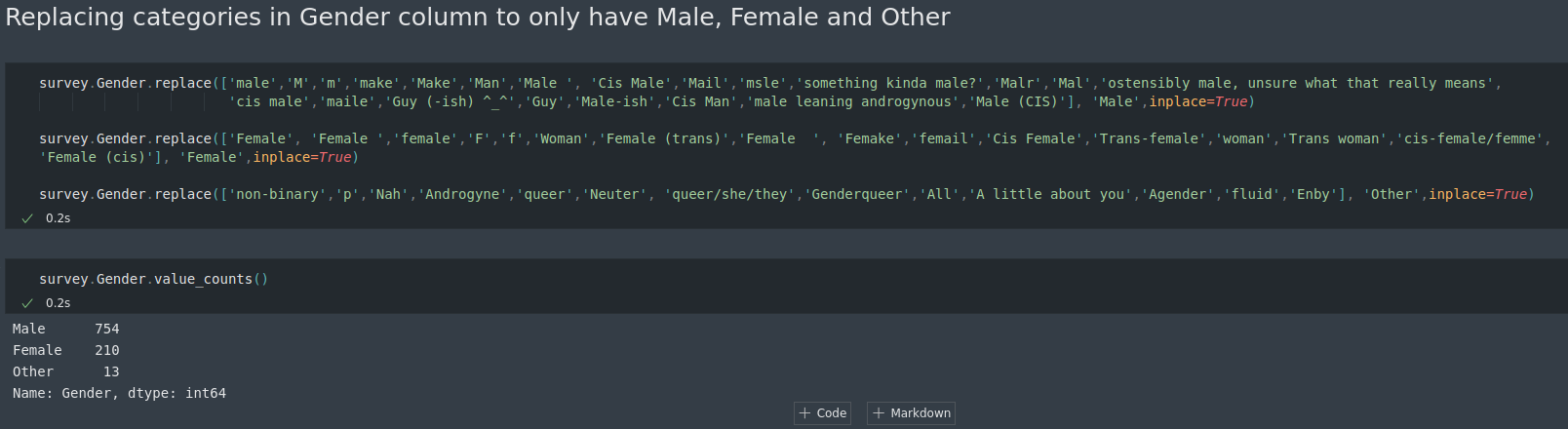
**3.3 Missing values**

There was a significant amount of missing values in the dataset as seen in the image below.

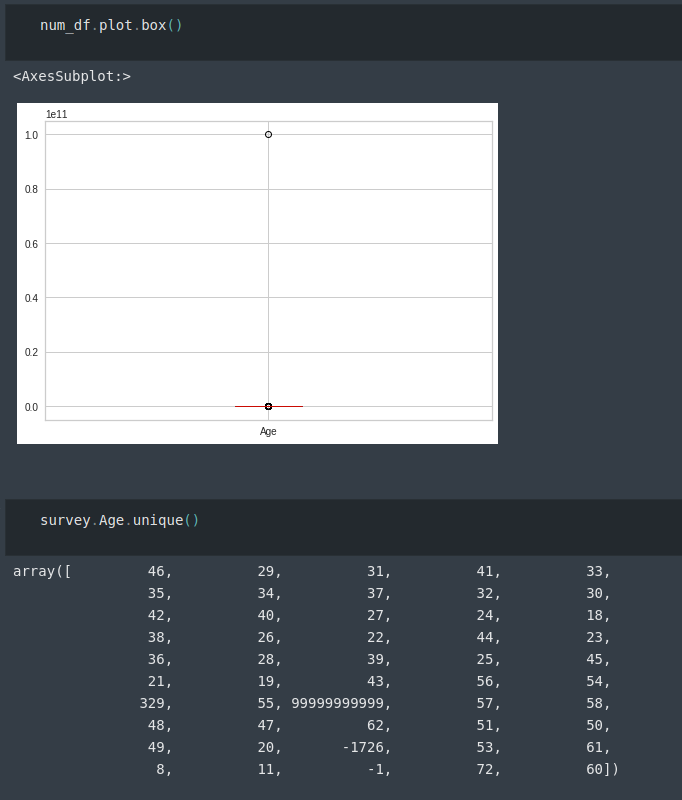
 Some columns, namely ‘state’ and ‘comments’ had too many missing values that using those entire columns had to be deleted.For the remaining columns, only the rows with missing values were deleted.

**3.4 Structural Errors**

The ‘Gender’ column was the only column with issues. Different spellings and cases were used for words like male and female. Those values were replaced with one of three categories, ‘Male’, ‘Female’ and ‘Other’, as shown below.

.

**3.5 Outliers**

Outleirs were found in the ‘Age’ column only as it was the only numeric column in the dataset. A boxplot of age shows the presence of outliers and using the unique() method, we can see some extreme values in the column.

These outliers were removed using the interquartile range method.

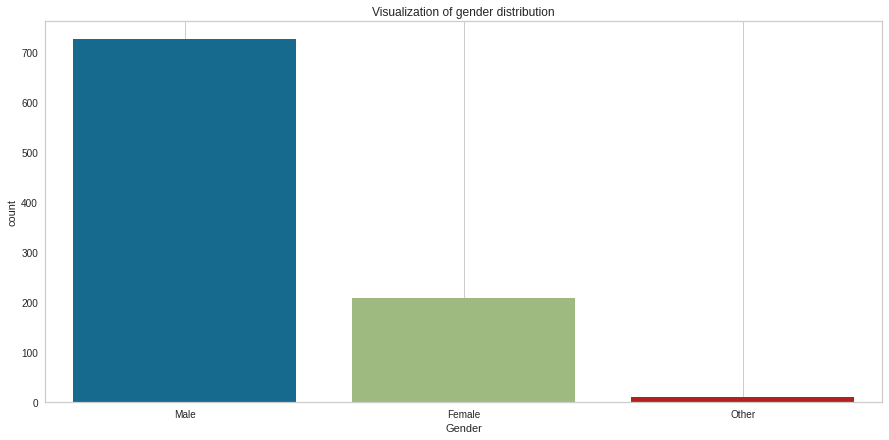
**4.0 CHAPTER FOUR: EXPLORATORY DATA ANALYSIS (EDA)**

**4.1 Introduction**

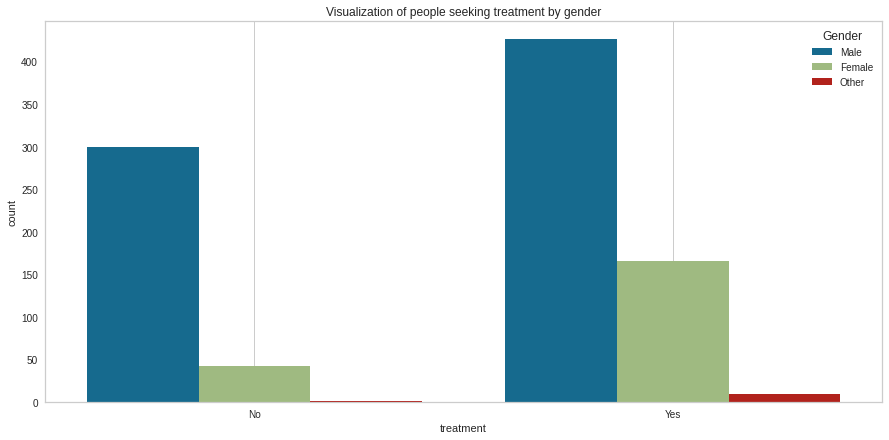
EDA is done not only to familiarize yourself with all the data you have collected, but also   
to reduce the workload during analysis (Cox, Victoria, 2017). In this section, graphs of different natures will be used to visualize different columns in the dataset and their relationship to each other. Python libraries such as *pandas, seaborn* and *matplotlib* will be used in the visualization process.

**4.2 Visualiztion**

4.2.1 Visualization of the distribution of gender

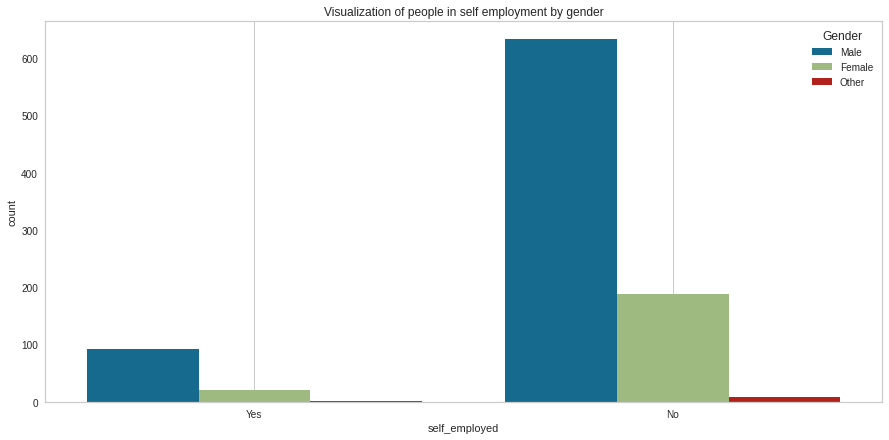
.

As seen in the image above, the data contained more males than females or other. This may be due to the fact that the tech industry is heavily male dominated.

4.2.2 Visualiztion of people seeking treatment classified by gender

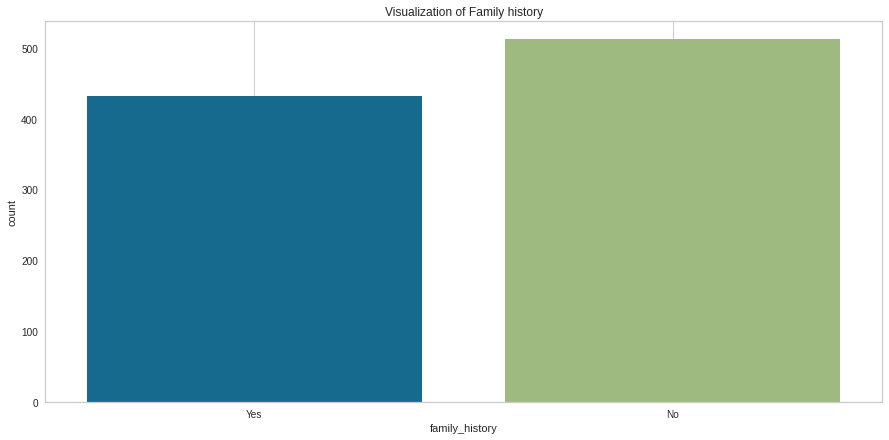
In all gender categories, there are more respondents seeking treatment as oppossed to those who do not. The number of males not seeking treatment is significantly higher than the other gender categories (Female and Other)

4.2.3 Visualization of people in self employment based on gender

.

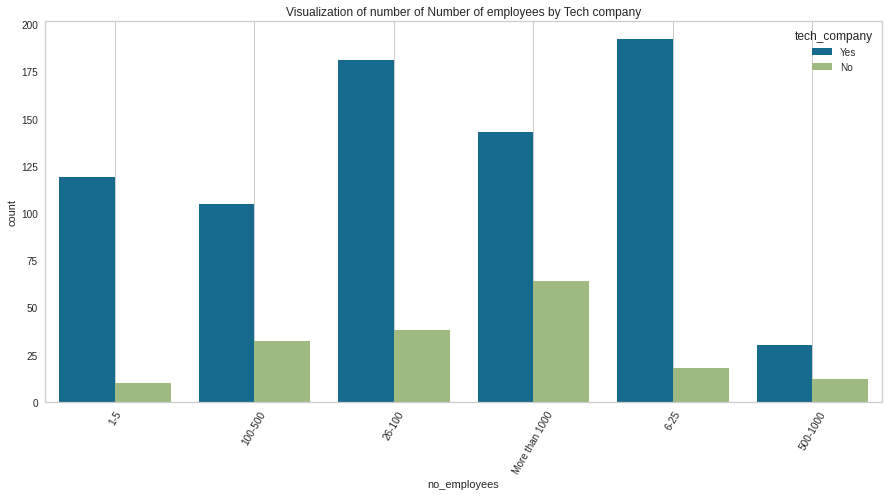
While most people are not self employed, a majority of those who are, are male.

4.2.4 Visualization of mental health issues in family of the respondents



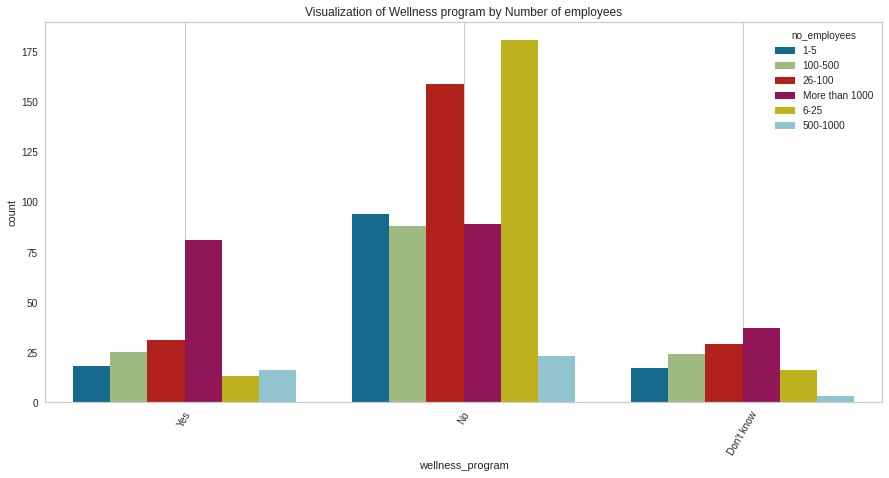
A slim majority of the respondents do not have a history of mental health issues in their families. Some mental health issues can be hereditary thus affect the respondent.

4.2.5 Visualization of number of employees based on type of company



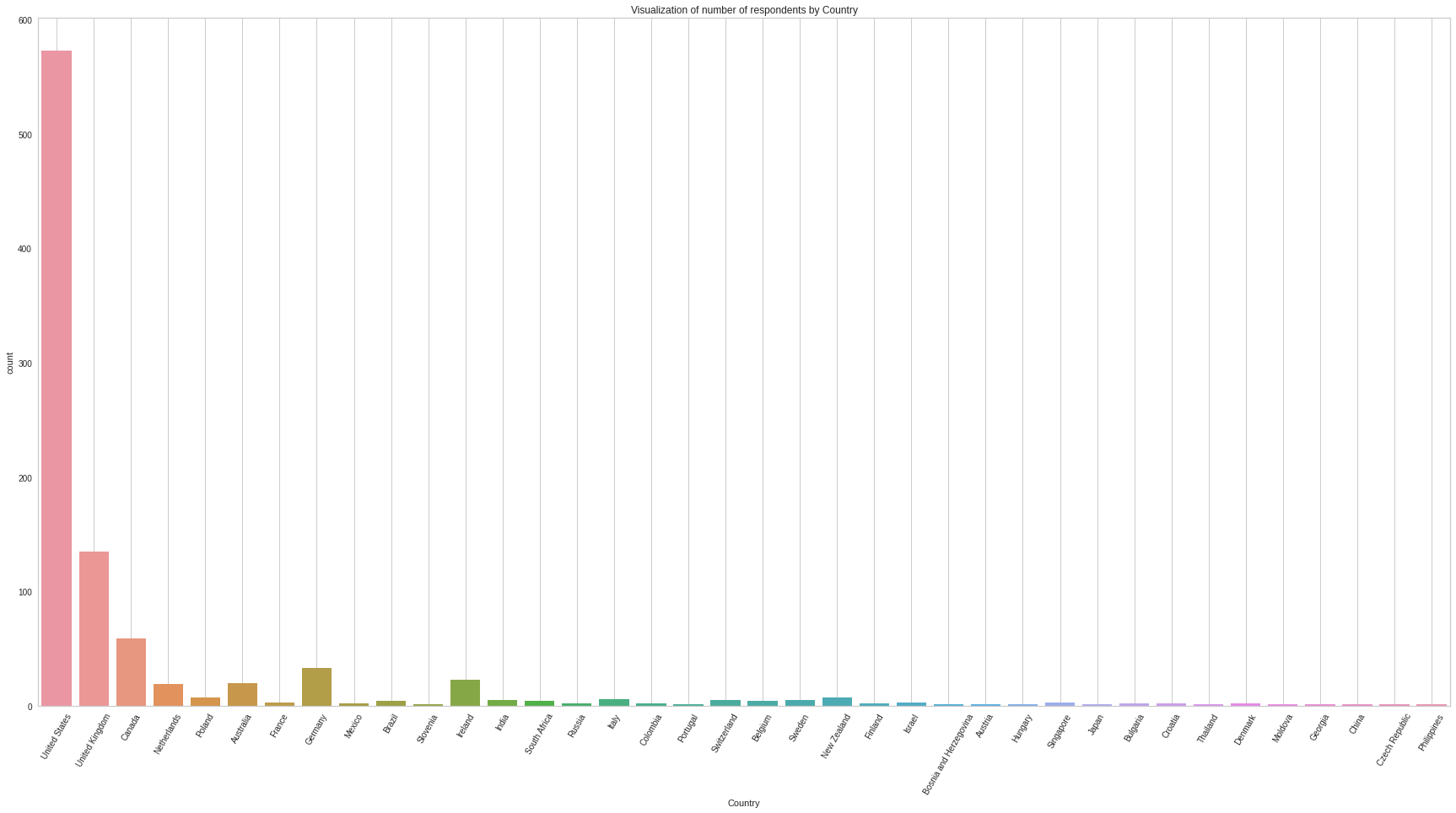
For each category of number of employees, tech companies are the biggest employers.

4.2.6 Visualization of wellness program based on number of employees



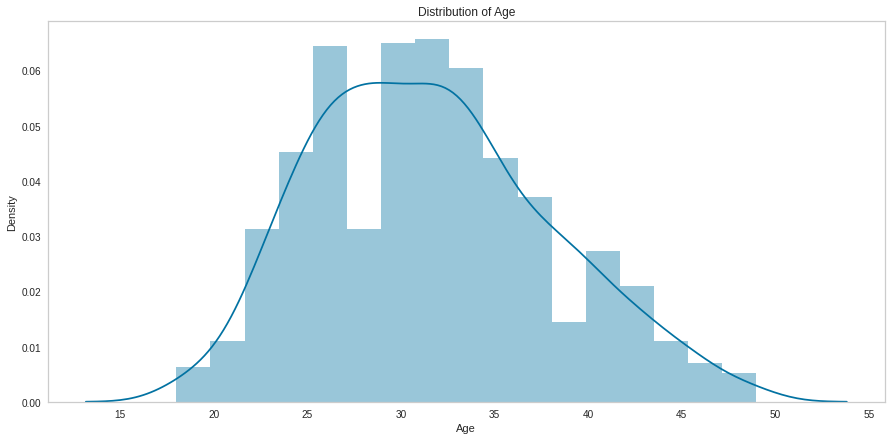
Larger companies (more than 1000 employees) are the majority in companies that offer wellness programs for their employees. This can be attributed to the fact that larger companies have more resources thus can invest more into the health of their employees.

4.2.7 Visualization of number of respondents by country

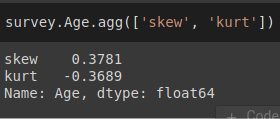


An overwhelming majority of the respondents are from the United States followed by the United Kingdom. This is because a majority of world leading tech companies are based in the US eg Amazon, Facebook, Google etc.

4.2.8 Visualization of age



Age seems to be slightly positively skewed as the right tail is slightly longer than the left tail.



It is also normally distributed because the skewness as it falls between -0.5 and 0.5.

**5.0 CHAPTER FIVE: MACHINE LEARNING (CLASSIFICATION)**

**5.1 Introduction**

The aim of machine learning for this study was to classify whether the respondent should seek treatment for their mental health issues based on a variety of factors included in the dataset. The procedures followed are as listed below;

1. Feature engineering
2. Separation of predictor and dependent variables
3. Encoding the predictor variable
4. Splitting the data into training and testing sets
5. Building the classification models and crossvalidation
6. Feature selection
7. Scaling the data
8. Splitting the scaled data into training and testing sets
9. Building the classification models and crossvalidation using the scaled data
10. Hyperparameter tuning

**5.2 Feature engineering**

The timestamp column had to be dropped as it does affect the machine learning classification that will be done and has no significance. The country column also had to be dropped as the work people in tech do is really similar all over the world. This reduced the number of coulumns from 25 to 23.

**5.3 Separation of predictor and dependent variables**

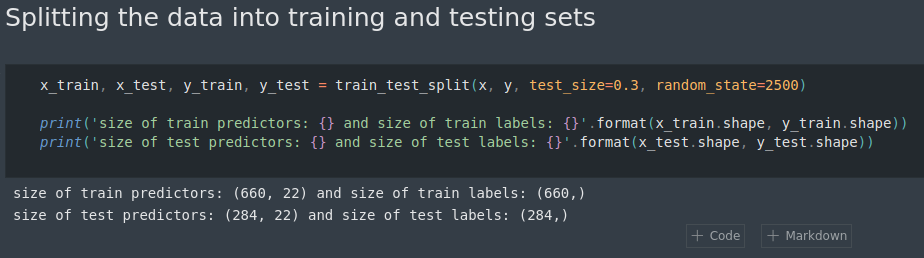
The treatment column is the dependent/response variable while the rest of the columns are the independent/predictor variable. The predictor variable is labeled x while the response variable is y. The x variable has 22 columns.

**5.4 Encoding the predictor variable**

Machine learning models use data in numeric form. Since the data had categorical columns, it was neccessary to convert the data into a format the models could utilise. Encoding was done using LabelEncoder instead of dummies as dummies would have increased the number of columns in the predictor variable. This would have made feature selection very difficult.

**5.5 Splitting the data into training and testing sets**

The predictor variable was split into two uneven portions, a train set and a test set. The test set was 30% of the entire predicor variable. The code is as shown below



**5.6 Building the classification models and crossvalidation**

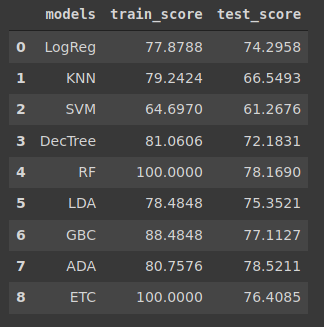
5.6.1 Building the models

The classification models that were used in this study were as listed below;

* Logistic regression classifier
* KNeighbours classifier
* Support vector machine classifier
* Decision tree classifier
* Random forest classifier
* Linear discriminant analysis
* Gradient boost classifier
* Ada boost classifier
* Extra trees classifier

5.6.2 Performance of the models

The accuracy of the models built are highlighted in the image below.



Logistic regression and Ada boost models were the least affected by overfitting. This can be seen as the difference in the accuracy of the train score and the test score was not as huge as in the other models.

5.6.3 Crossvalidation