**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

Data Collection

Data Cleaning

Exploratory Data Analysis

Building and Crossvalidating Machine Learning Models

Tuning Optimal Machine Learning Model

Deploying the Model

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.

**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 DATA CLEANING**

**3.1 Introduction**

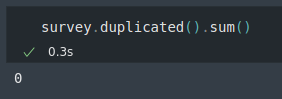
Data cleaning is used to refer to all kinds of tasks and activities to detect and repair errors in the data. (Ilyas, I. F., Chu, X. 2019). When data is collected, a lot of errors may be made in the process.

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**

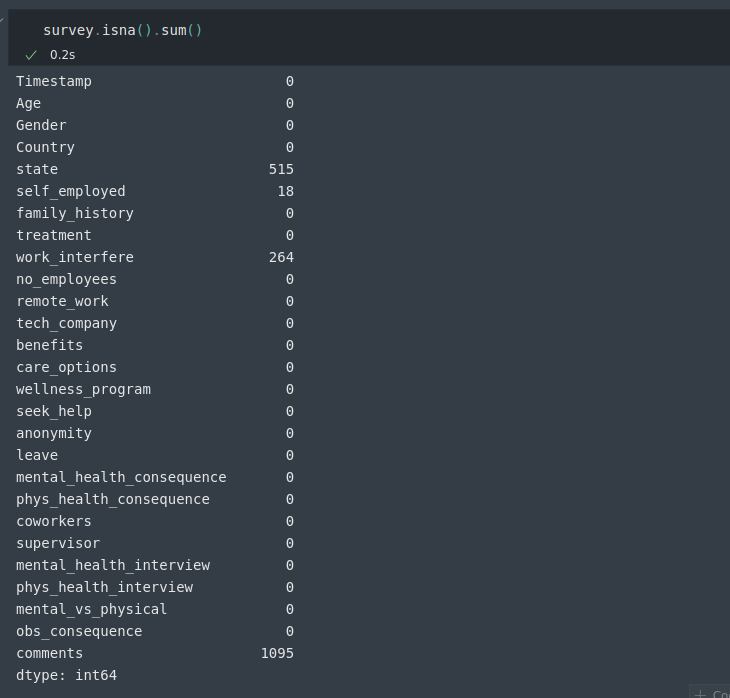
When cleaning the data, no duplicated records were found as shown in the diagram below

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.

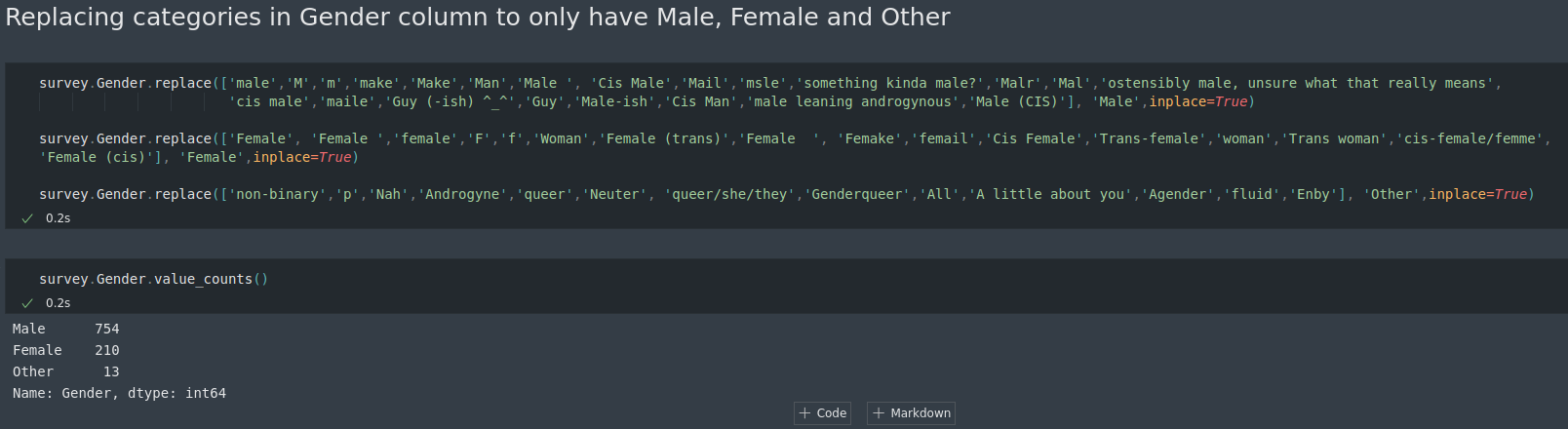
**3.3 Missing values**

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**