11-2022

Societal Impact Assessment of Depression

Mohammed Salah Almazrouei
msa5837@rit.edu

Societal Impact Assessment of Depression

by

Mohammed Salah Almazrouei

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology - DUBAI

November 2022

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Mohammed Salah Almazrouei**

Graduate Capstone Title**: Societal Impact Assessment of Depression**

Graduate Capstone Committee:

Name: **Dr.Sanjay Modak** **Date:**

Chair of committee

Name: **Dr.Ioannis Karamitsos** **Date:**

Member of committee

Acknowledgments

I would like to express my appreciation to everyone who was involved in the completion of my capstone project. The completion of this project would never be possible without the people who supported me throughout this journey, my parents, friends, family members, and doctors. They supported me when I needed them the most and they kept telling me that I can get my master's degree. I would like to thank Dr. Ioannis Karamitsos who was my mentor, he gave all his knowledge and answered all my questions without tiring, and helped me through my capstone journey. Special thanks to Dr. Snajay who gave me this opportunity to present my capstone & gave me clear guidance about how to submit a proper capstone.

ABSTRACT

Depression is a mental problem characterised by unusual changes in the mood of a person, accompanied by a temporary different emotional response to everyday changes in the phenomena (WHO, 2020). Affected people suffer greatly, hence functioning poorly in the workplace, school, or at home in a family setup. Even though the public is well aware of its existence, there is little going on to combat this condition. Amongst the various stumbling blocks towards achieving an effective approach to address the problem include limited resources and associated well-trained personnel to help the victims.

A mental disorder associated with social stigma and misdiagnosis is the greatest challenge that hinders the steps to address depression. This is a common challenge in developing countries especially low-income families where depressed persons are wrongly diagnosed or misdiagnosed. As a consequence, they are given the wrong prescriptions for antidepressants. Due to the steady rise in the case of depression reported in the recent past, many efforts have been made to allow country-level attention to the disorder in a coordinated and comprehensive approach.

Keywords: mental wellbeing, analytics, machine learning, depression

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

1.1 Background Information

Depression, a mental health disorder, has been an existing challenge for a long time. Due to its increasing rate of reported cases of depression and suicide, it is important to find a means to address the issue. According to WHO, depression is among the top priorities to be covered through their mental health Gap Action Programme. The target groups include workers and middle-aged youths who are the most hit, by numbers. This is because a larger portion of this population is lost every day to suicide causing deaths. As a result, the strong energetic workforce of the industry is lost whilst increasing the dependency ratio in the society.

There exists a close relationship between mental health with social interactions, work performance, and physical health. Whenever one is depressed, there are high chances of poor work performance, communication, coordination with colleagues, and poor judgement. Similarly, one's physical health suffers the most. Diseases such as cardiovascular complications are closely associated with depression, and vice versa. Some of the contributing factors of depression include adverse life situations such as psychological disruption, loss of employment, loss of a loved one, or financial crisis.

Stress and social stigma are among the killer accomplices of depression than even depression itself. As a common problem among the young generation, the responsible parties should move with speed to handle the crisis. This can be achieved through sufficient deployment of community-based education programs to enlighten the public on the symptoms and possible preventive measures. For the elderly, regular exercises help prevent depression while for children, guardian intervention is critical in observing and monitoring their social responses to curb the problem at its early stage.

1.2 Statement of the Problem

Depression is one of the killer diseases affecting millions of people in the world. According to a new report by the World Health Organization, more than 264 million people have suffered the effects of depression in the recent past (WHO, 2020). It grows to be a serious health condition with items, adversely affecting the social, health, and economic well-being of a person. It is a common threat that spreads across all ages, with its dominance being on middle-aged people. In a worst-case scenario, it may lead to suicide, leading to death, a common trend affecting youths and middle-aged group people between 15 and 29 years of age. Consequently, close to 1 million people die of suicide daily due to negligence about their mental health. There is a need to emphasise on preventive approach rather than treatment.

1.3 Project Aim and Objectives

The proposed project aims to assess and analyse the impact of depression on society. The objectives for this project are as follows:

i. Using machine learning models to detect depression in individuals at an earlier stage.
ii. Develop a model to associate the respective symptoms of depression and their control measures.

1.4 Limitation of the study

Depression score on the first data was measured using the DASS scale, which assigns depression categories as follows; Normal(0-9), Mild depression (10-13), Moderate depression(13-20), severe depression(21-27), extremely severe(28+), unfortunately, nobody had depression score less than 10. This means that all the sample respondents had mild to very severe depression scores made it hard to capture patterns of normal cases from this dataset.

CHAPTER 2: LITERATURE REVIEW

Social profiling can influence one's social well-being, sense of belonging, and good interaction between people (Steger & Kardashian, 2009). Positive social interactions foster good mental health and consequently a sense of belonging and performance. On the other hand, depressed people tend to behave with biassed judgments and perceptions toward life issues. Their cycle of thoughts and attention are majorly over-occupied with negativity. Previous research has shown that depressed people are much more preoccupied with finding a sense of belonging in any social interaction.

Social media has a significant contribution to the mental wellness of individuals too. Cases like cyberbullying and attacks have put young adults and youths in danger of developing mental disorders. Violence associated with social dating and scammers is one of the many factors derailing the mental wellness of individuals on social media platforms (Paat & Markham, 2021).

Recent studies have shown that the COVID-19 pandemic and its related lockdowns have contributed much to the rise in the cases of depression among family individuals. Through cross-sectional research using REDCap data, a secure web-based platform for data capture, data was processed and manipulated using automated procedures into statistical packages for analysis (Jacques-Aviñó, 2020). Using questionnaires, the living conditions, socioeconomic status, and COVID-19 related behaviour changes were studied and analysed. The study utilised a multivariable ordinal regression model to develop the relationship between the variable and depression. The study showed that more than a third of women experienced depression and anxiety during the period. On the other hand, about 17% of men were also affected by symptoms of anxiety and depression. An increase in the cases of anxiety, despair, insomnia, drug abuse, and isolation was also noticed during this period (Ustun, 2020).

In their paper, Shuang G. and team studies about one of the greatest causes of disability and morbidity through Major Depression Disorder or MDD. Due to the lack of a proper method to determine MDD in its early stages the team has explored the use of translational biomarkers of mood disorders based on Machine Learning which has one of the greatest potential to understand these disorders. They review popular machine learning models which can be used for brain image classification and predictions especially in the case of MDD. (Shuang, 2018).

Another research performed by Adam Mourad C. and team used patient-reported data from different patients who have depressions which start from level 1 to relieve Depression in order to identify variables which are most predictive of treatment outcome, later these variables are used to train a machine learning model to predict clinical remission. During the process, 25 variables were determined to have the best

predictive power in terms of treatment outcome from 164 patient-reportable variables. The model was also found to perform significantly compared to escitalopram-bupropion treatment groups. (Adam M.C., 2016)

The geriatric depression scale used to evaluate the relationship between mental health, depression, and retirement showed a direct relationship between the two aspects in question (Shiba, 2017b). The study used developed regression models of the first difference stratified based on gender to analyse the phenomena under study. Time-dependent variables such as marital status dependence, income lifestyle, and diseases were a great consideration too. The study revealed a great relationship between the people from the low economic class and the incidence of experiencing depression after retirement. COVID-19, though not alone, has played a role in the rising reported cases of mental disorders in the current times (Salari, 2020). With the help of the Comprehensive Meta-Analysis platform, a random-effects model used showed that the impact of the pandemic varied differently in various communities depending on their cultures and settings.

In the paper by Purude N. (2020), a data collection approach has been used to accumulate rich data to understand and predict depression in individuals. Some of the data collection mechanisms include questionnaires from people, social media posts and text messages through verbal communication or facial expressions. The team used different machine learning models like Decision Trees, Naive Bayes Classifier, Logistic Regression as well as KNN Classifier to perform a comparative analysis between all these models. They also used a Twitter scraping tool which helps in scraping tweets and identify if the message has a depression intent or not.

According to new research, a person's social environment highly affects mental status and wellness in the long run. Positive environments create a positive perception and reaction towards phenomena and vice versa. Studies based on laboratory experiments show a consistent continuation of depression symptoms over the test groups. This showed that social value varies directly with social exclusion, as symptoms of depression channel the resources of attention to a negative direction. McLaughlin (2011) has developed a targeted prevention scheme selectively for those at high risk of developmental disorders. A linear multilevel regression model used to determine multiple discrimination impacts on depression in Europe also showed similar results. (Yena, 2018) The effect increased the chances of one developing a mental disorder, especially in those with low socioeconomic backgrounds (Alvarez-Galvez, 2019).

Leveraging Machine Learning algorithms to predict anxiety, depression and stress has become one of the most interesting and sought after studies with the rising trend in depressions due to COVID-19, work life imbalance and many other factors. Anu (2020) and her team collected data from both employed and unemployed people across different cultures and communities after which the above state of mind was

predicted as occurring on five different levels of severity by the machine learning algorithms. Due to the high accuracy, they were considered for predicting psychological problems in humans.

Most of the efforts to curb depression in the US have been treatment-oriented rather than prevention (Baskin, 2021). Approximately 15 % of the adult population in the US and about 11% of teenagers have been affected by depression once or more in their lifetime. The effective developed preventive measures have been targeted at potentially high-risk groups of people in society. Although much effort has been put towards control of the disease, its extent of reach and accessibility to the victims is still limited. The result from the efforts in place is appealing but not yet satisfactory, thus, there is a need to redirect the efforts towards prevention rather than control.

Findings from the Literature Review

Depression is a mental health disorder that has existed for a long time with diverse effects, including suicide, thus, creating a need to develop a practical approach to combat the issue. Early detection of depression among patients is critical in preventing adverse effects, including suicide, linked to depression. (Xiaowei, 2019) Machine learning techniques are critical in the early detection and prevention of depression among patients.

Depression is a mental health disorder that has existed for a long time with diverse effects, including suicide, thus, creating a need to develop a practical approach to combat the issue.

Early detection of depression among patients is critical in preventing adverse effects, including suicide, linked to depression. This showed that social value varies directly with social exclusion, as symptoms of depression channel the resources of attention to a negative direction. Violence associated with social dating and scammers is one of the many factors derailing the mental wellness of individuals on social media platforms. (Meenal, 2015)

Social profiling can influence one's social wellbeing, sense of belonging, and good interaction between people. The study utilised a multivariable ordinal regression model to develop the relationship between the variable and depression. A linear multilevel regression model used to determine multiple discrimination impacts on depression in Europe also showed similar results. Positive social interactions foster good mental health and consequently a sense of belonging and performance. According to new research, a person's social environment highly affects mental status and wellness in the long run. Researchers have been successfully able to prove by analysing social media posts to understand if the user was sad or depressed in any manner through text sentiment analysis using machine learning. (Raymond, 2021)

Machine learning techniques are critical in the early detection and prevention of depression among patients. Social media has a significant contribution to the mental wellness of individuals too. Recent studies have shown that the COVID-19 pandemic and its related lockdowns have contributed much to the rise in the cases of depression among family individuals. The study showed that more than a third of women experienced depression and anxiety during the period.

CHAPTER 3: RESEARCH METHODOLOGY

In this project, the CRISP-DM methodology has been selected.

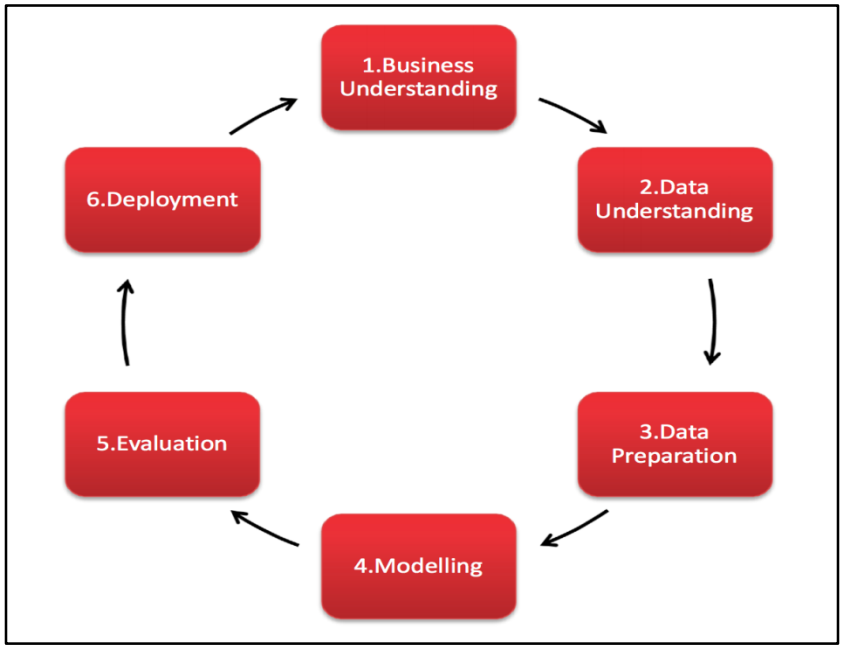


Figure 1:Crisp DM Process

2.1 Business understanding

This stage aims to understand the objectives of the study from a business perspective. The study will try to uncover the candidate factors contributing to depression in the source population as well as preventive factors of the same.

2.2 Data understanding

This study proposes the integration of new techniques of data analytics to develop insights for future planning and prevention execution of depression. It uses the data from the high-risk target groups such as middle-aged youths, teenagers, and the elderly. Adults too are a victim of the same, but their prevalence is

much less than the rest. The insights from the data give the comparative difference in the effectiveness of each treatment and also assess the impact of emphasising the preventive strategies rather than treatment.

2.3 Data Preparation

Data collection and preparation are key encounters in the study. This is because the required data should be reliable and credible enough to reflect the real situation on the ground. Large data sets required for analysis and evaluation can be obtained from organisations such as WHO, to help in the research. The data will also play an important role in the determination of the effectiveness of treatment of a preventive measure rolled out to a target group. Data is cleaned to remove invalid sections and parameters before using them. Data is then validated and split into two sets, 80% for training the models and the remainder for testing the accuracy and reliability of the models. (Smart Vision, 2019)

2.4 Data modelling

Data modelling and analytics are essential steps in data presentation. tools such as Tableau, R programming, and Excel which are used during data processing. Predictive data analytics methods are the candidate techniques that give useful information from the collected data. They clean, process, and evaluate data with its associated models making the work simpler. These machine learning models help in the recognition of patterns and trends in the data, monitoring behavioural changes and response of the test groups to different phenomena.

The study utilises the data analysis techniques such as logistic regression, random forests, decision trees, and k-nearest neighbour's methods. Neural networks will also be used to recognize the relationships between the datasets and pattern recognition. The choice of parameters in each model will rely on parameter tuning by cross-validation of the parameters. The results are important in predicting the viability of preventive and control measures applied, and methods to implement, based on the past responses (Data Analyst Tools - Best Software for Data Analysts, 2020).

2.5 Model Evaluation

The project will evaluate the development of an affordable detection and preventive model from the data. The interventions to be developed are geared towards reducing the occurrences of depression and the consequential impact on the population. After the prediction using the testing data, the model accuracy will be calculated and the results compared to determine the most efficient model to use in the analysis of the data. (Arkaprabha, 2017) As opposed to the existing multilevel techniques, this approach gives a universal model that can handle a variety of test data and give accurate reliable results. This project proposes a

promising approach to the development of preventive actions against depression as it employs the power of technology in data processing and analysis.

2.6 Deployment

At this stage the aim will be to explore the evaluation results and identify deployment strategy, this will be done through a conclusion derived from a comparison between the achieved results and the theoretical expectations obtained from a detailed review of related literature.

CHAPTER 4: DATA ANALYSIS

4.1 Data description

The study will use two datasets available on Kaggle machine learning repository. The first data consists of 39775 observations collected during an online survey, hosted on OpenPsychometrics.org. Respondent's demographic characteristics and Depression, Anxiety, and Stress score as measured using Depression Anxiety Stress Scales (DASS) were recorded.

The data from OpenPsychometrics.org had a few instances of missing cases, variable orientation had the highest level (7.82%) followed by education which had 1.25% missing cases. The rest of the variables had less than 1%. The second data had relatively a higher number of missing cases. About 15 of the 91 variables had over 30% missing cases. With the highest being with med_u5_deaths 94.84%. *see appendix 1 for variable description.*

The second data comes from a survey conducted by Busara Centre in rural Siaya County, Kenya in the year 2015. A total of 2286 respondents participated. Unlike the first datasets, this data mostly consists of financial status and income level variables. It was chosen to bring in the effect of the financial situation on depression.

4.2 Data cleaning

For the second dataset, all variables with more than 30% missingness were dropped to avoid losing much data for analysis methods that use listwise deletion in the handling of missing cases. Value labels were also added to both datasets. The dependent variable on the first dataset was recorded into a categorical variable based on the 4 DASS cut-offs described in the data description part. The two datasets were also split into 80% training set and 20% testing set.

4.3 Exploratory data analysis

Exploratory data analysis was conducted using graphical representation and summary statistics, to give an overview of trends and patterns of relationships in this data.

4.3.1 Sample distribution across categories of independent variables

The results of the survey hosted at OpenPsychometrics.org show that about 68.08% of the respondents had very severe depression, while 16.64% and 15.28% had severe and mild depression simultaneously. The respondents indicated their education levels as; less than high school (38.80%), high school degree (37.58%), less than high school degree (10.50%), and a graduate degree (13.12%). The sample was less

balanced across gender, with 75.63% being male and 22.86% being male. About 11% of the respondents were married people while 2.79% were previously married. The larger proportion of the respondents come from the urban locality (44.76%), followed by those from suburban areas (34.50%), rural residents contribute the least number (20.73%). Figure 1 and 2 shows the percentage distribution of respondents across categories of independent variables.

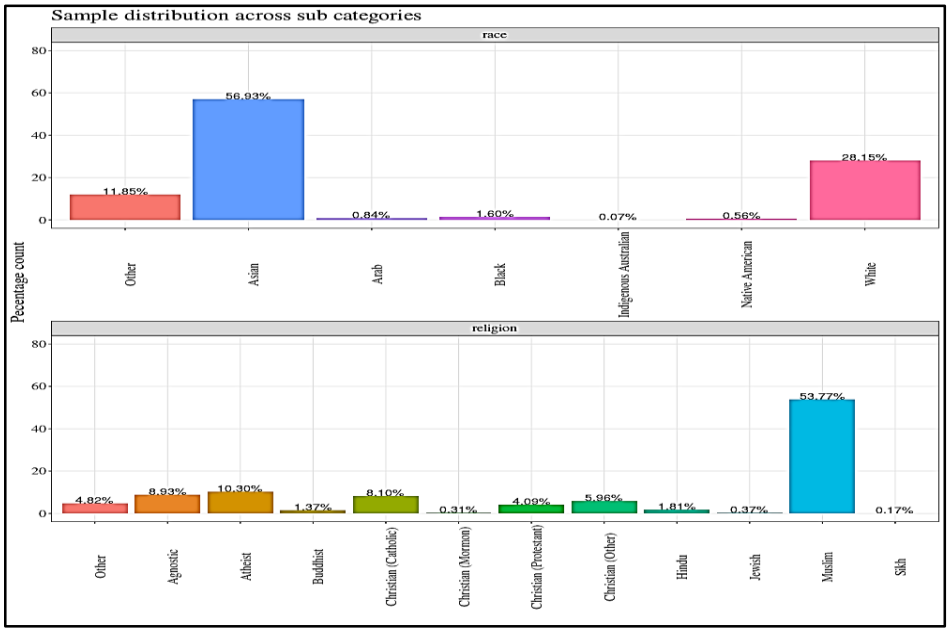


Figure 2: Sample distribution across categories of the independent variables

Figure 3 Sample distribution across categories of the independent variables

4.3.2 Distribution of total score across categories of independent variables

For the data collected via OpenPsychometrics.org, the total depression scores a normal distribution across the categories of independent variables, native Americans show a relatively higher depression score (μ=37.79, σ=12.53), followed by Arabs (μ =37.56, σ=12.69). Asians had the least depression score of all races surveyed races (μ =34.41, σ=12.2). Conversely, Mormons showed the highest depression score (μ=37.32, σ=13.08) while protestants had the least average score (μ=33.91, σ=12.43). See *table 1* below and appendix 2.

Table 1: Summary statistics for total depression score across categories of independent variables.

Variable	Category	n	mean(sd)	Median (IQR)	skewness	min	max
education	Less than high school	2965	38.82(12.21)	40(21)	-0.3	14	56
	High School	10614	36.42(12.22)	37(21)	-0.09	14	56
	University degree	10957	33.85(12.14)	33(20)	0.15	14	56
	Graduate degree	3705	32.05(12.43)	30(21)	0.32	14	56
engnat	Yes	10429	35.74(12.6)	36(22)	-0.02	14	56
	No	17812	34.73(12.23)	34(21)	0.06	14	56
gender	Male	6456	34.28(12.65)	34(22)	0.09	14	56
	Female	21359	35.23(12.27)	35(21)	0.03	14	56
	Other	426	41.39(11.4)	43(20)	-0.44	14	56
Hand	Left	24774	35.03(12.37)	35(21)	0.04	14	56
	Right	2958	35.12(12.3)	35(21)	0.03	14	56
	Both	509	38.42(12.82)	41(22)	-0.3	14	56
married	Never married	24319	35.77(12.26)	36(21)	-0.02	14	56
	Currently married	3133	30.16(12.07)	28(19)	0.5	14	56
	Previously married	789	34.16(12.69)	33(22)	0.14	14	56
orientation	Other	2680	36.2(12.25)	36(21)	-0.05	14	56

			34.04(12.31				
	Heterosexual	18695	)	33(20)	0.13	14	56
	Bisexual	3862	38.4(12.05)	39(20)	-0.27	14	56
			36.79(12.17				
	Homosexual	1501	)	37(20)	-0.12	14	56
	Asexual	1503	36.2(12.51)	37(22)	-0.07	14	56
			35.15(12.42				
Race	Other	3347	)	35(22)	0.02	14	56
	Asian	16074	34.41(12.2)	34(21)	0.1	14	56
			37.56(12.69				
	Arab	238	)	40(20)	-0.27	14	56
	Black	453	36(13.24)	36(23)	-0.11	14	56
	Indigenous						
	Australian	19	35.16(12.4)	38(23)	-0.37	14	52
			37.79(12.53				
	Native American	159	)	39(20)	-0.29	14	56
	White	7951	36.3(12.54)	37(21)	-0.07	14	56
religion	Other	1362	37(12.56)	38(21)	-0.16	14	56
			37.16(12.37				
	Agnostic	2523	)	38(21)	-0.16	14	56
			37.99(12.51				
	Atheist	2908	)	39(21)	-0.24	14	56
			32.17(12.16				
	Buddhist	387	)	31(20)	0.3	14	56
			35.18(12.55				
	Christian (Catholic)	2288	)	35(22)	0.01	14	56
			37.32(13.08				
	Christian (Mormon)	88	)	39.5(23.25)	-0.13	14	56
	Christian		33.91(12.43				
	(Protestant)	1154	)	33(21.75)	0.14	14	56
			34.97(12.57				
	Christian (Other)	1684	)	35(22)	0.04	14	56
	Hindu	511	34.4(13.14)	34(22)	0.1	14	56

	Jewish	104	33.83(13.48)	33(25.25)	0.14	14	56
	Muslim	15185	34.23(12.1)	33(20)	0.12	14	56
	Sikh	47	34.13(12.42)	35(18.5)	-0.05	14	56
Urban	Rural (countryside)	5855	34.87(12.44)	34(22)	0.07	14	56
	Suburban	9744	34.82(12.32)	34(21)	0.06	14	56
	Urban (town, city)	12642	35.43(12.39)	35(21)	0	14	56
Voted	Yes	8141	33.18(12.39)	32(21)	0.2	14	56
	No	20100	35.88(12.29)	36(20)	-0.03	14	56

Summary statistics for total depression score across categories of independent variables. Standard deviations and interquartile range are in brackets.

4.3.3 Relationship between Total score family size and age.

A weak negative correlation was found to exist between Total score, family size, and age. This means that the variables have very little impact on depression see figure 4 below.

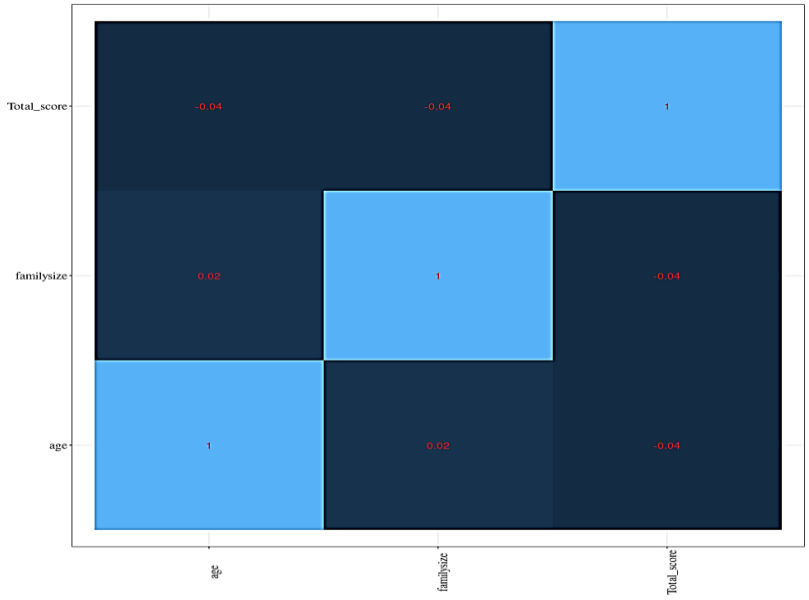


Figure 4: correlation plot for variables total depression score, Family size, and age

For the Busara Centre dataset, 83.11% of the respondents came from the depressed. The sample is composed of 91.69% women and about 72.55% married people, around 77.87% have no income from casual or wage labour primary source of income. *Figure 4* and *Appendix 3* below report respondents' distribution across independent variables categories.

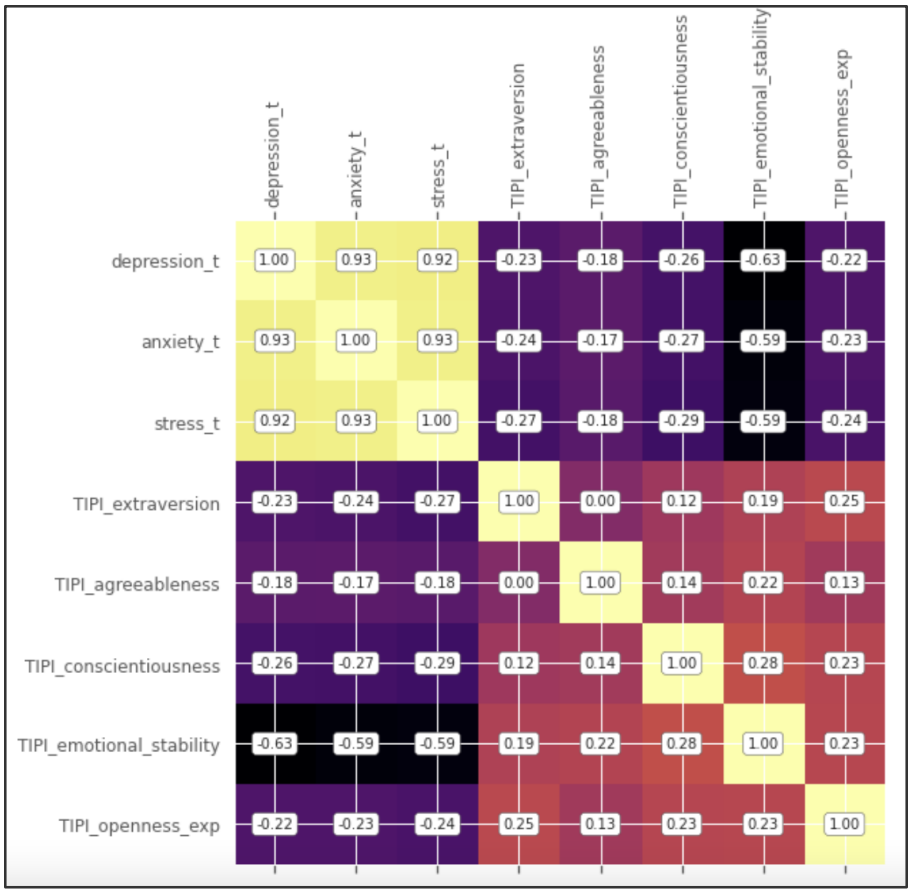


Figure 5: Correlation plot of all variables from the open psychometrics dataset

T*able2* below reports summary statistics for all continuous variables in this survey, standard deviations and the interquartile ranges are reported in brackets. Most of the variables have a skewness coefficient above 1 which means that data are positively skewed

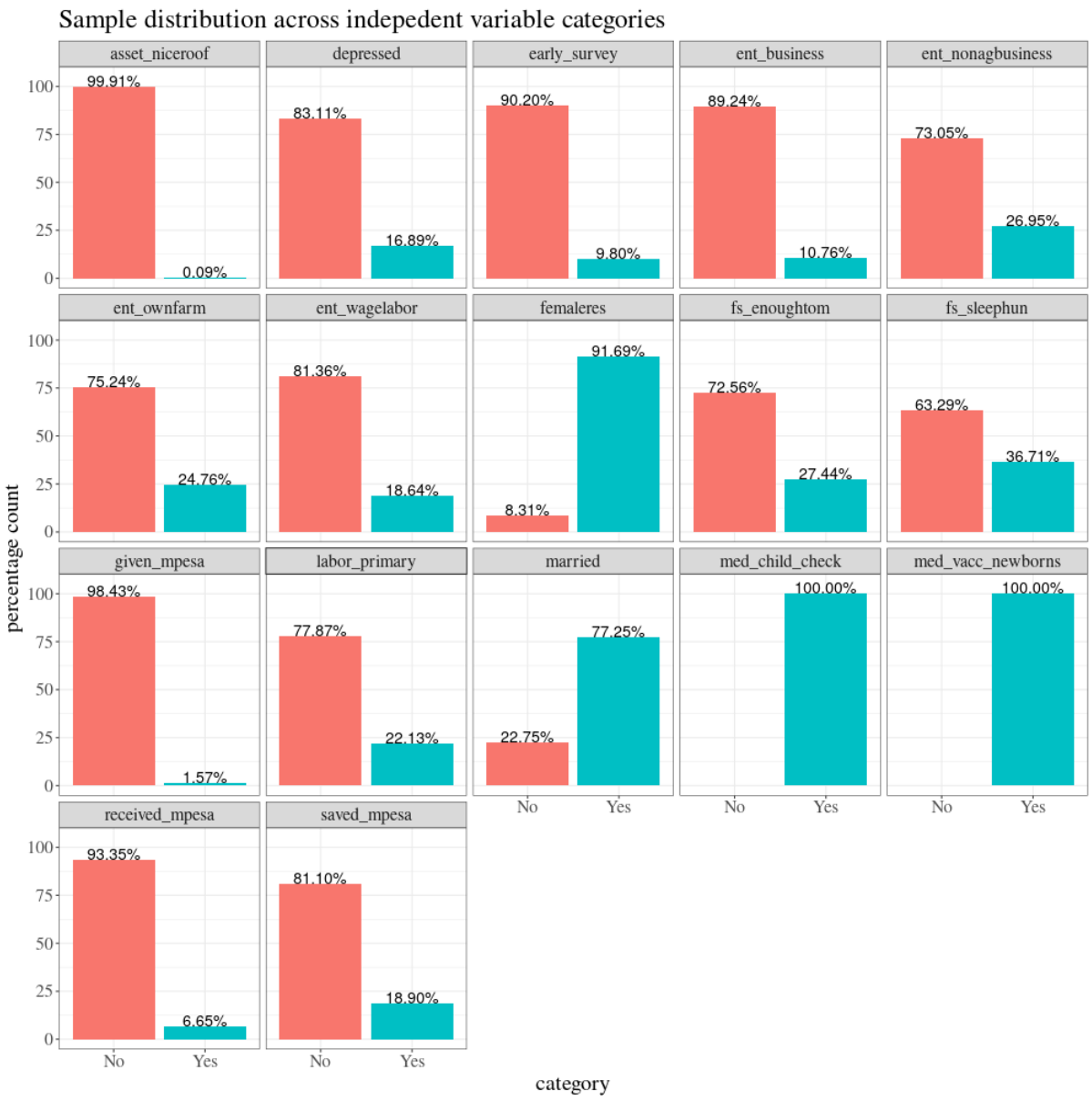


Figure 6: Sample distribution across categories of independent variables.

Table 2: Summary statistics for all numeric independent variables.

Variable	n	mean(sd)	median(IQR)	skewnes s	min	max
age	1143	34.54(13.74)	30(17)	1.28	17	91
amount_given_mpesa	1143	0.55(6.49)	0(0)	18.78	0	160.15
amount_received_mpesa	1143	3.56(24.95)	0(0)	10.07	0	352.34

amount_saved_mpesa	1143	2.27(19.68)	0(0)	17.05	0	488.47
asset_durable	1143	148.57(198.42)	121.72(221.89)	7.06	0	3720.37
asset_land_owned_total	1143	0.93(1.66)	0(1.5)	5.29	0	27
asset_livestock	1143	113.7(239.7)	16.52(102.57)	4.42	0	2754.53
asset_phone	1143	19.66(28.05)	0(32.03)	1.88	0	192.18
asset_savings	1143	10.25(79.81)	0(0)	22.04	0	2242.15
children	1143	2.86(1.85)	3(2)	0.52	0	10
cons_alcohol	1100	1.18(6.74)	0(0)	9.04	0	104.67
cons_allfood	1143	95.81(114.49)	77.11(136.84)	3.39	0	1386.97
cons_ed	1143	2.79(8.33)	0.47(2.4)	8.1	0	133.46
cons_med_total	1143	2.69(12.14)	0(0)	9.36	0	206.6
cons_nondurable	1143	128.41(140.16)	107.39(188.53)	2.45	0	1431.61
cons_other	1143	21.71(28.4)	13.77(30.91)	2.96	0	289.09
cons_ownfood	1143	8.32(15.15)	3.2(9.82)	4.41	0	181.57
cons_social	1143	3.65(7.57)	1.28(3.82)	7.19	0	140.13
cons_tobacco	1123	0.63(2.75)	0(0)	7.46	0	41.87
durable_investment	1143	288.5(385.35)	188.81(399.12)	3.25	0	3782.33
edu	1143	8.74(2.87)	9(2)	-0.77	1	19
ent_animalstockrev	1143	3.93(12.2)	0(2.74)	6.87	0	176.17
ent_employees	1143	0.03(0.39)	0(0)	21	0	11
ent_farmexpenses	1143	1.87(3.51)	0.49(2.16)	4.23	0	38.06
ent_farmrevenue	1143	4.52(8.63)	2.14(5.75)	8.04	0	161.35
ent_nonag_flowcost	1143	17.32(100.53)	0(0)	13.96	0	2067.58
ent_nonag_revenue	1143	34.85(257.96)	0(0)	23.96	0	7687.38
ent_total_cost	1143	21.41(101.39)	1.24(10.22)	13.6	0	2067.58
fs_adskipm_often	1143	4.04(6.25)	1(5.25)	1.79	0	20
fs_adwholed_often	1143	0.91(2.59)	0(0)	5.08	0	20
fs_meat	809	3.07(2.21)	3(2)	2.22	0	22
hh_children	1143	2.02(2.02)	2(3)	0.76	0	10
hh_totalmembers	809	4.91(2.1)	5(3)	0.49	1	12

hhsize	1143	4.87(2.12)	5(3)	0.44	1	12
med_portion_sickinjured	809	0.52(0.32)	0.5(0.5)	0.15	0	2
med_sickdays_hhave	809	1.96(3.36)	1(2.4)	5.02	0	31
net_mpesa	1143	3.01(25.84)	0(0)	8.8	-160.15	352.34
nondurable_investment	1143	34.46(134.06)	4.58(21.92)	11.31	0	2275.47
wage_expenditures	1143	24.36(798.99)	0(0)	33.72	0	27000

CHAPTER 5: DATA MODELLING

5.1 Modelling

Five classification models were trained on both datasets to predict depression based on a demographic characteristic of the respondents (dataset 1) and also based on respondent's income data (dataset 2). For all the 5 models, hyperparameters were tuned using cross-validation to ensure optimal results.

5.5.1 Logistic regression

Multinomial logistic regression was used to model the relation relationship between demographic variables and depression, the basic idea behind the model is the assumption of a linear relationship between the log of relative risk ratio and the independent variables. For the OpenPsychometrics.org survey, category Mild was used as the base category for the model. The prediction equations are as follows;

$$ln\,\square\square\left(\frac{\square(\square\square\square\square\square\square\square\square\square\square\square=\square\square\square\square\square\square)}{\square(\square\square\square\square\square\square\square\square\square\square\square=\square\square\square\square)}\right)$$
$$=\square_0+\square_1\square\square\square\square\square\square\square\square\square\square\square\square h\,\square\square h\square\square\square+\cdots+\square_{\square}\square\square\square\square\square\square\square\square\square\square\square$$

$$ln\,\square\square\left(\frac{\square(\square\square\square\square\square\square\square\square\square\square\square=\square\square\square\square\;\square\square\square\square\square\square)}{\square(\square\square\square\square\square\square\square\square\square\square\square=\square\square\square\square)}\right)$$
$$=\square_0+\square_1\square\square\square\square\square\square\square\square\square\square\square\square h\,\square\square h\square\square\square+\cdots+\square_{\square}\square\square\square\square\square\square\square\square\square\square\square$$

Where; $\beta_{\square}{}'\square$ are the estimated effects of the independent variables on a log of risk ratio.

Education was found to have a significant impact on the relative risk ratio at a 5% level, the risk of having severe compared to mild depression was estimated to be 1% times higher in people who have a high school as the highest level of education compared to those with less than high school education. Conversely, With a University degree, the risk of having severe compared to mild depression was estimated to be 11% less compared to when one has less than a high school education. This trend means that in this population, the risk of depression decreases significantly with the level of education. For gender, the results from this model suggest that the risk of severe depression compared to mild is 27.2% times higher in women than in men. English natives were found to be 1% times less at risk compared to other people. Religion, race, marital status, and whether one voted in the previous election had also a significant impact on depression ***Table 3*** below reports the estimates of log relative risk ratio (Estimate) and the relative risk ratios (Exp(B)).

Table 3: Estimates of multinomial logistic regression

Variable	*term*	*Severe*		*Very Severe*	
		Estimate	*Exp(B)*	*Estimate*	*Exp(B)*
	(Intercept)	0.088	1.092	1.572*	4.816
		(0.156)		(0.126)	
Education	High School	-0.01	0.99	-0.164*	0.849
		(0.091)		(0.073)	
	University degree	-0.118	0.889	-0.442*	0.643
		(0.094)		(0.075)	
	Graduate degree	-0.297*	0.743	-0.709*	0.492
		(0.105)		(0.084)	
Urban	Sub	-0.045	0.956	0.015	1.015
		(0.059)		(0.048)	
	urban(town, city)	0.004	1.004	0.068	1.07
		(0.056)		(0.046)	
Gender	Female	0.241*	1.272	0.336*	1.399
		(0.051)		(0.041)	
	Other	0.995*	2.706	1.307*	3.696
		(0.294)		(0.258)	
Engnat	No	0.07	1.072	0.111*	1.117
		(0.055)		(0.044)	
Age		-0.001	0.999	-0.001	0.999
		(0.001)		(0.001)	
Hand	handRight	-0.001		0.008	1.008
		(0.069)		(0.056)	
	handBoth	-0.294	0.745	0.067	1.07
		(0.185)		(0.138)	
Religion	Atheist	-0.101	0.904	-0.011	0.989
		(0.109)		(0.087)	
	Buddhist	-0.3	0.741	-0.611*	0.543
		(0.187)		(0.15)	
	Christian (Catholic)	-0.117	0.89	-0.302*	0.739

		(0.11)		(0.09)	
	Christian (Mormon)	0.189	1.208	0.085	1.089
		(0.407)		(0.341)	
	Christian (Protestant)	-0.132	0.877	-0.244*	0.783
		(0.129)		(0.103)	
	Christian (Other)	-0.166	0.847	-0.289*	0.749
		(0.119)		(0.095)	
	Hindu	-0.093	0.911	-0.299*	0.742
		(0.176)		(0.143)	
	Jewish	-0.461	0.631	-0.901*	0.406
		(0.324)		(0.259)	
	Muslim	-0.092	0.912	-0.28*	0.756
		(0.105)		(0.084)	
	Sikh	-0.647	0.524	-0.234	0.791
		(0.565)		(0.396)	
	Other	-0.089	0.915	-0.097	0.907
		(0.132)		(0.106)	
Orientation	Bisexual	0.098	1.102	0.506*	1.659
		(0.075)		(0.06)	
	Homosexual	0.179	1.196	0.414*	1.513
		(0.105)		(0.085)	
	Asexual	0.002	1.002	0.166*	1.181
		(0.1)		(0.08)	
	Other	0.078	1.081	0.276*	1.318
		(0.079)		(0.064)	
Race	Arab	-0.45	0.638	0.149	1.161
		(0.268)		(0.191)	
	Black	-0.372*	0.689	-0.201	0.818
		(0.188)		(0.141)	
	Indigenous Australian	0.04	1.041	-0.003	0.997
		(0.829)		(0.67)	
	Native American	-0.237	0.789	0.224	1.251
		(0.347)		(0.258)	

	White	0.088	1.092	0.149*	1.16
		(0.082)		(0.066)	
	Other	0.027	1.027	-0.003	0.997
		(0.07)		(0.057)	
Voted	No	0.038	1.038	0.168*	1.182
		(0.049)		(0.04)	
Married	Currently	-0.323*	0.724	-0.734*	0.48
		(0.063)		(0.051)	
	Previously	-0.096	0.908	-0.251*	0.778
		(0.126)		(0.101)	
Family size		0	1	-0.006	0.994
		(0.01)		(0.008)	

Notes: 1) standard deviations are in brackets

2) Exp(B) is a transformation of the estimates to relative risk ratios. These estimates can be interpreted by

computing the percentage change in relative risk ratio for each unit change independent variable.e

100(Exp(B)-1) when the estimate is positive and 100*(1-Exp(B) when negative*

3) significance codes 0.05 '', 0.01, '**',0.001*

4.4.2 Binary logistic regression

For the second dataset, the outcome variable had only two categories, this makes a binary logistic regression more appropriate. Similar to multinomial logistic, the model assumes a linear relationship between the independent variables and the log of odds ratio as follows:

$$ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Where; $ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$ is the log of odds ratio and β_p' are the estimated effects of the independent variables on log of odds ratio. The sample results show that the Odds of depression in this population decreased by 3.15% of the time for every additional person in the household (hh) while holding other factors constant and increasing by 10.7% times for every extra year of education (edu). For every dollar increase in Non-ag business flow expenses (ent_nonag_flowcost), as the other factors remain

constant, the odds of being depressed increase by 0.06% of the time. On the other hand, for every dollar increase in farm expenses (ent_farmexpenses), depression was found to decrease by 3.3% times. ***Table 4*** below reports estimates of logistic regression, standard errors, and the odds ratio for this sample data.

Table 4: Logistic regression estimates

term	estimate	std.error	statistic	p.value	Exp(B)
				0.0016	
(Intercept)	2.791	2.791	3.148	*	16.297
femaleres	0.087	0.087	0.182	0.8558	1.091
age	0.008	0.008	0.711	0.4772	1.008
married	0.426	0.426	1.302	0.1929	1.531
children	0.252	0.252	1.421	0.1554	1.287
				0.0202	
hhsize	-0.379	-0.379	-2.322	*	0.685
				0.0211	
edu	0.102	0.102	2.307	*	1.107
hh_children	NA	NA	NA	NA	NA
hh_totalmembers	NA	NA	NA	NA	NA
		10294.02			
cons_nondurable	10294.021	1	1.003	0.3157	inf
asset_livestock	-0.003	-0.003	-1.053	0.2926	0.997
asset_durable	-0.001	-0.001	-0.499	0.6177	0.999
asset_phone	-0.002	-0.002	-0.537	0.591	0.998
		58174.72			
asset_savings	58174.721	1	1.593	0.1112	inf
asset_land_owned_total	-0.011	-0.011	-0.187	0.8515	0.989
					859116339
asset_niceroof	22.874	22.874	0.000	0.9999	7
cons_allfood	-10294.021	-10294.02	-1.003	0.3157	0
cons_ownfood	0.015	0.015	1.105	0.269	1.015
cons_alcohol	-10294.052	-10294.05	-1.003	0.3157	0
cons_tobacco	-10293.968	-10293.97	-1.003	0.3157	0
cons_med_total	-10294.03	-10294.03	-1.003	0.3157	0

		47880.70			
cons_ed	47880.707	7	1.243	0.2138	inf
cons_social	-10294.033	-10294.03	-1.003	0.3157	0
cons_other	-10294.02	-10294.02	-1.003	0.3157	0
ent_wagelabor	-0.136	-0.136	-0.246	0.806	0.873
ent_ownfarm	-0.314	-0.314	-0.884	0.3766	0.731
ent_business	-0.631	-0.631	-1.424	0.1543	0.532
ent_nonagbusiness	-0.102	-0.102	-0.357	0.7209	0.903
ent_employees	0.06	0.06	0.129	0.8974	1.062
ent_nonag_revenue	0	0	-0.526	0.5988	1
				0.0265	
ent_nonag_flowcost	-0.063	-0.063	-2.219	*	0.939
ent_farmrevenue	-0.001	-0.001	-0.048	0.9614	0.999
				0.0031	
ent_farmexpenses	-0.143	-0.143	-2.960	*	0.867
ent_animalstockrev	-0.017	-0.017	-1.540	0.1235	0.983
		58174.79			
ent_total_cost	58174.792	2	1.593	0.1112	inf
fs_adskipm_often	-0.005	-0.005	-0.263	0.7925	0.995
fs_adwholed_often	-0.139	-0.139	-4.046	1e-04*	0.87
fs_meat	-0.081	-0.081	-1.769	0.0769	0.922
fs_enoughtom	-0.113	-0.113	-0.431	0.6666	0.893
fs_sleephun	0.179	0.179	0.680	0.4965	1.196
med_portion_sickinjured	-0.709	-0.709	-1.721	0.0853	0.492
med_sickdays_hhave	-0.019	-0.019	-0.559	0.5764	0.981
med_vacc_newborns	NA	NA	NA	NANA	NA
med_child_check	NA	NA	NA	NANA	NA
labor_primary	-0.37	-0.37	-0.641	0.5214	0.691
wage_expenditures	0.011	0.011	0.000	0.9998	1.011
durable_investment	0.002	0.002	0.619	0.5359	1.002
nondurable_investment	-58174.723	-58174.72	-1.593	0.1112	0
given_mpesa	-0.936	-0.936	-0.879	0.3796	0.392

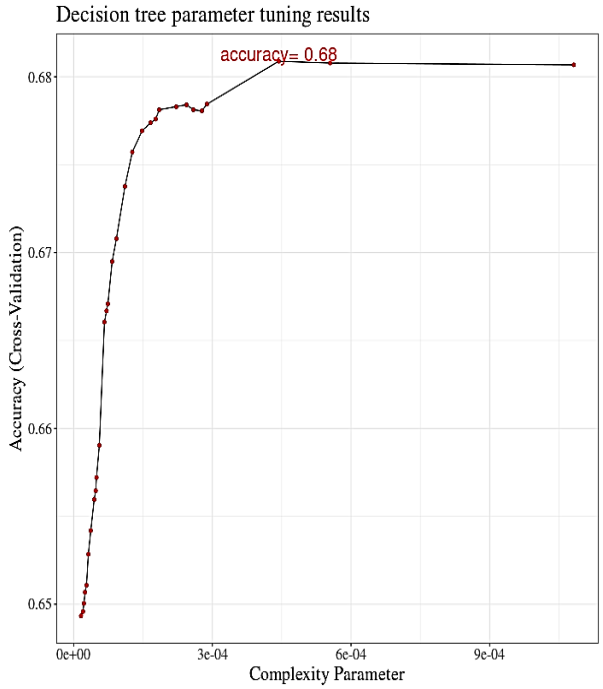
amount_given_mpesa	-15710015.95	-15710016	-0.275	0.7837	0
received_mpesa	0.101	0.101	0.172	0.8631	1.106
amount_received_mpesa	15710015.95	15710016	0.275	0.7837	inf
net_mpesa	-15710015.95	-15710016	-0.275	0.7837	0
saved_mpesa	-0.153	-0.153	-0.516	0.606	0.858
amount_saved_mpesa	0.037	0.037	1.408	0.159	1.038
early_survey	-0.194	-0.194	-0.606	0.5448	0.824

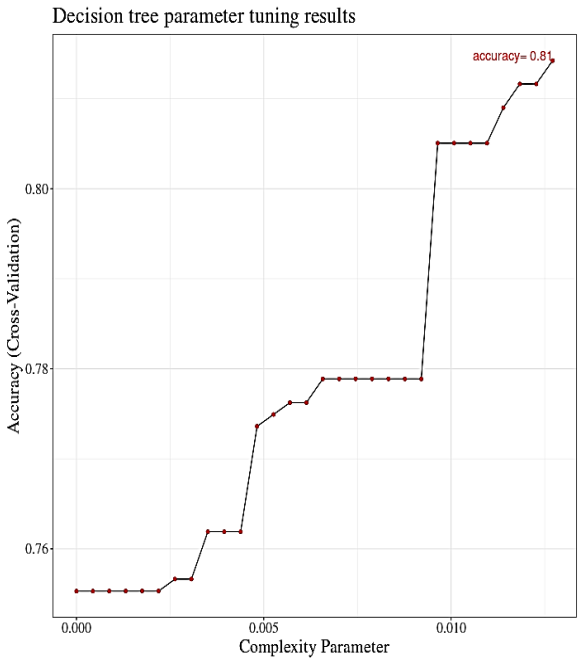
Notes : 2) significance codes 0.05 '', 0.01, '**',0.001*

4.4.3 Decision trees

A decision tree algorithm works by binary partitioning to split the dependent variable into subgroups based on the independent variables. At the first step, the algorithm establishes which of the independent variables is most related to the dependent variable, splits the data based on it; and tries to see how depression cases are distributed within the subgroups of the independent variable. This is repeated with the second most related variable to the independent variable and so on. This continues until all nodes become pure or the desired minimum number of cases on the node is reached or the resulting group has results to minimum subgroups with less than the desired count.

During the training, the models and hyperparameters were tuned using cross-validation to ensure the best combination was found. For the first dataset, the optimal complexity parameter was found to be 0.0004438034 which corresponds to a cross-validation accuracy of 68%. For the rest of the hyperparameters the r default setting of; 20 for the number of cases that must exist on a node for splitting to be tried (minsplit), 7 for the least count of the number of observations in any leaf node (min bucket), 4 for the maximum number of competitors retained(maxcompete), 5 for the maximum number of surrogates (maxsurrogate),2 for not surrogating the process(use surrogate), 10 for cross-validation splits (xval), and 10 for maximum depth of the nodes.



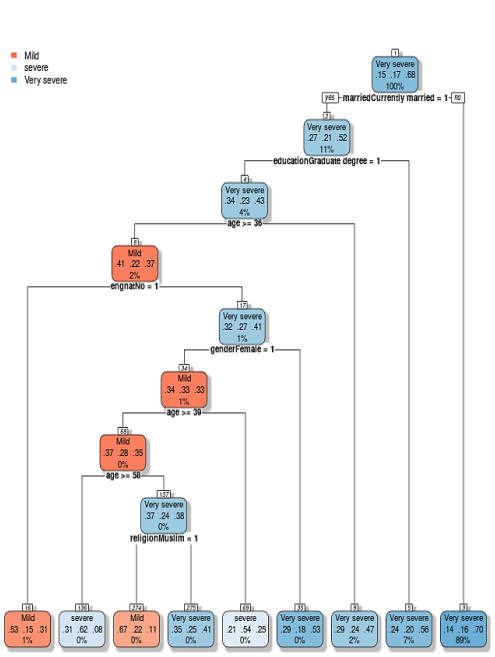


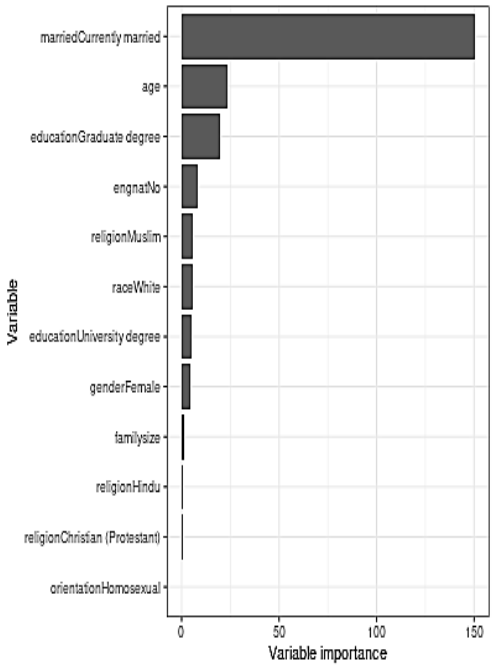
For the Busara centre dataset, the optimal complexity parameter is 0.01271186 which corresponds to a cross-validation accuracy of 81%. The above default was preserved for the rest of the hyperparameter too. *Figure 7* below shows the cross-validation results for the decision trees.

Figure 8 below visualises the splitting of the DT model on the OpenPsychometrics.org dataset. At the start the outcome variable was split based on marital status, according to the model, there is an 11% probability of very severe depression in married people and 89% in the other marital groups. For married people with a graduate degree, there is a 4% probability of very severe depression, while for married people with other education levels there is 7%. Going deeper down we see that the estimated probability of severe depression for married people, with a graduate degree and aged 30 years and below, is 2%, and the probability for mild depression is 2% in the same group. The right panel of figure 6 below shows Gini variable importance for

OpenPsychometrics.org dataset. Marital status is seen to be the most important variable in predicting depression according to the graph followed by age and education. Education is the least important.

For the Busara centre dataset, the tree couldn't be split past the root node; this is to mean that the root node doesn't meet the splitting rules.



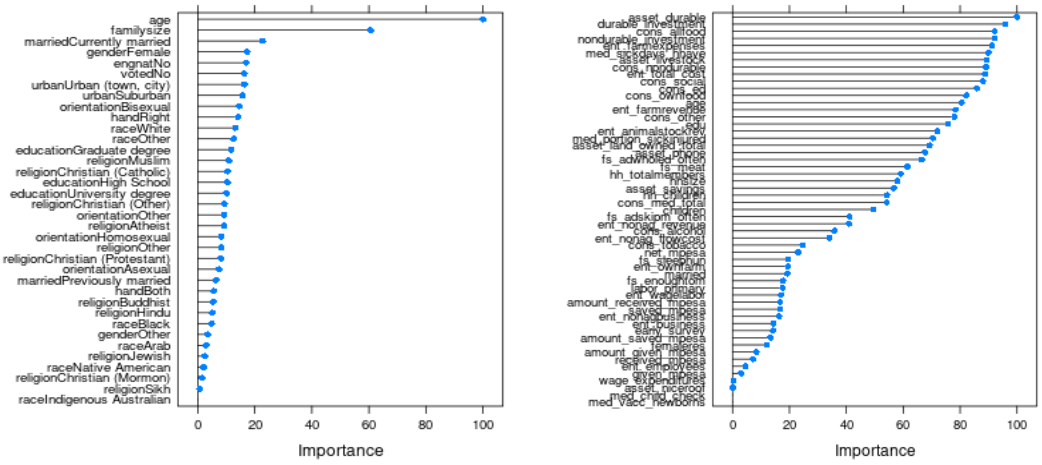


4.4.4 Random forest

A random forest is an extension of decision trees where, instead of building a single tree, the model builds multiple decision trees. During prediction all the trees emit predictions, and the most voted category is said to be the predicted value. The trees are built by taking bootstrap samples from the training data. By default, R takes 500 samples from the training data. Different values for the minimum number of variables were randomly included in the subsamples (mtry). Figure 7 the trend of cross-validation accuracy as mtry parameter is changed. The left panel features the model built on OpenPsychometrics.org dataset, while the right panel features variable importance for the model built on the Busara centre dataset.

For the OpenPsychometrics.org dataset, the model shows that age has the highest value of GINI importance followed by family size and marriage. On the other hand, whether one is of Sikh religion or indigenous Australian religion had the least importance in predicting depression. For Busara centre dataset, Ownership

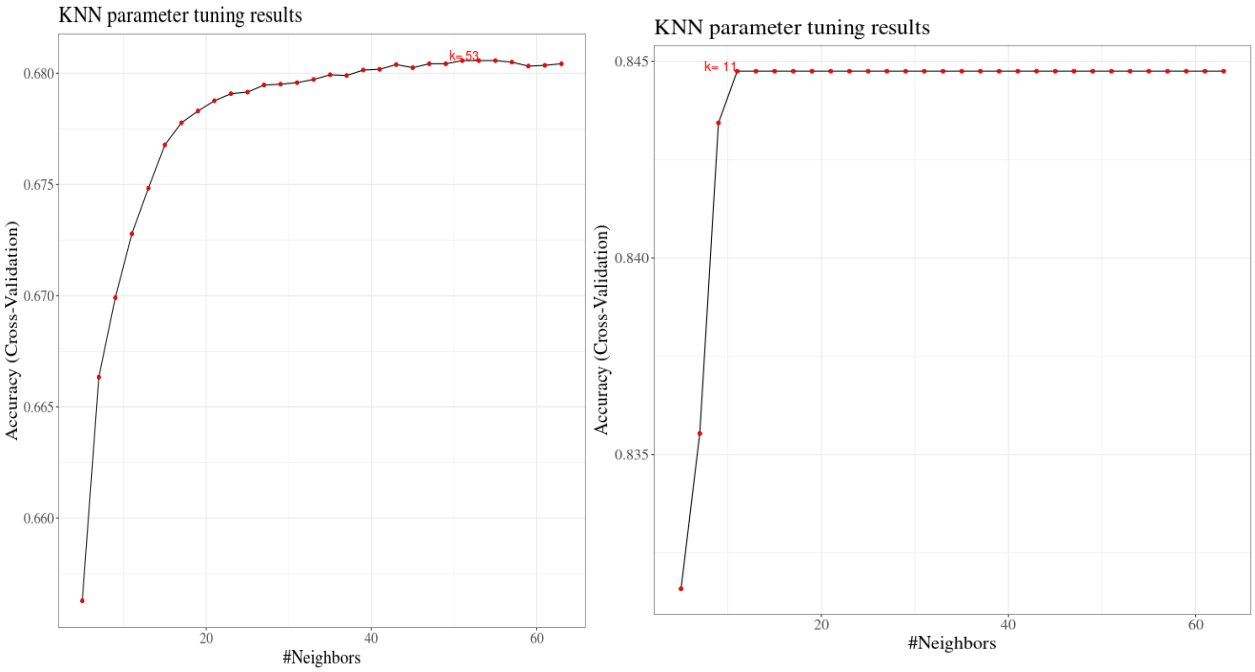
of durable assets and non-durable investments are the most important features in



4.4.5 K nearest neighbour

A K nearest neighbour model is a distance-based classifier. During training, a distance metric between every two points on the training data is computed. The distance metric acts as a measure of similarity between points. Each training case is then assigned a predicted class based on the k nearest points. During prediction, the distance between the prediction case and the training is calculated; the new case is assigned a predicted value based on the testing data.

The only hyperparameter on the model is k, which measures the number of neighbours on the data to consider, the distance metric used in this analysis is Euclidean distance (similarity metric). For the OpenPsychometrics.org dataset, people with more similar demographic variables are expected to have very little difference in Euclidean distance. For each person on the training data, the algorithm draws k most similar people, if most of them are depressed, then the person is predicted to be depressed. Different values of k were tried to find the optimal value of k. A similar approach is used on the second data. ***Figure 9*** below visualises the different values of k tried against the cross-validation accuracy. OpenPsychometrics.org dataset (left panel), k =53, was found to be optimal with 68% accuracy while for Busara centre dataset(right panel) k=11, was optimal with 84% accuracy.



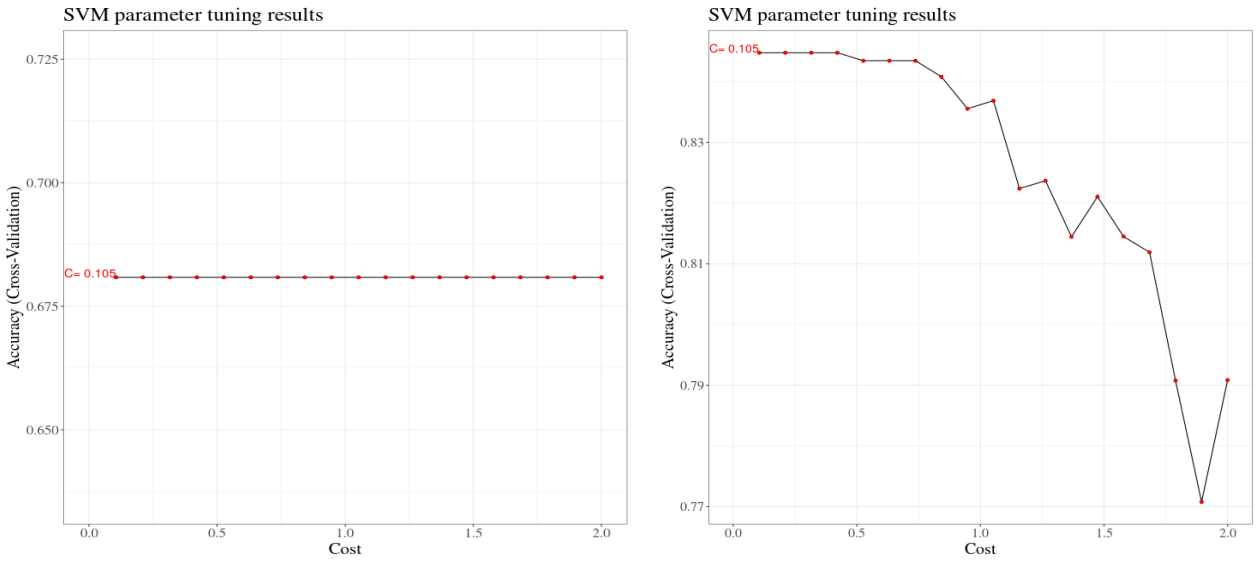
4.4.6 Support Vector Machine

A support vector machine works by using a kernel trick to first map the data into a high-dimensional feature space so that data points can be categorised, even when the data are not otherwise linearly separable. A linear kernel will be used for this analysis. The cost parameter which controls how fast the kernel is allowed to bend was tuned for both datasets. The optimal value for both datasets is 0.105 which leads to 68% cross-validation accuracy, on the OpenPsychometrics.org dataset and 84% on the Busara centre survey data. The

left panel of ***figure 10*** below shows results for the OpenPsychometrics.org dataset while the right shows the results on Busara centre dataset.

4.4.7 Comparison

By comparing the accuracy of the four models, the random forest model had the highest accuracy on the OpenPsychometrics.org dataset; it classified 68.45% of the evaluation cases correctly. Multinomial logistic regression came in second with 68.39% accuracy; the KNN model scored the poorest with 68.28% accuracy. For the second dataset, the random forest was the best with 99.08% accuracy. Binary logistic regression came in second with 84.61% accuracy. Decision tree, KNN, and SVM all scored 84.47%. These



metrics tell us the ability of each fitted model to predict one's depression status given their demographic characteristics and income data. High accuracy values mean that our system can be used to detect depressed people with precision and advise them to seek health care services on time.

Table *5* below reports this information.

Table 5: Comparing Model accuracies

Dataset	Logistic	DT	Rf	knn	SVM
OpenPsychometrics.org	0.6839	0.6836	0.6845	0.6828	0.6836
Busara Centre	0.8461	0.8447	0.9908	0.8447	0.8447

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

Depression is a mental disorder that has existed for decades and greatly affects most people across the world. The disease is associated with diverse health conditions that influence the social and economic well-being of the victims from different age groups, most significantly among middle-aged individuals. Early detection of depression cases is critical in preventing the adverse effects of the disease, including suicide among the victims. To understand and prevent such a problem in the future, we chose the problem statement so that Machine Learning is able to identify such trends and diagnose at an early stage of depression, anxiety or related mental disorder in patients.

Machine Learning techniques such as Random Forest, Logistic Regression, Decision Tree, and KNN are at the heart of curbing the effects of depression as far as early detection and prevention of depression are concerned. The techniques help develop a model that will help in the early detection of depression, thus, preventing the harmful effects linked to depression. The study utilised the four techniques in analysing two datasets which revealed that, in both datasets, Random Forest is the most effective Machine Learning model for detecting the causes of depression at an earlier stage to initiate control measures thus preventing adverse effects of depression.

The study was associated with various limitations. For instance, in the first dataset, all the people were found to be depressed, making it a challenge to capture the patterns related to depression.-means an imbalanced dataset, so how to solve this limitation. Besides, a comparison between the first and the second datasets could not be conducted. Such limitations greatly affect the study outcomes as far as the selection of the most effective model for detecting depression among patients is concerned. Besides, the datasets had instances of missing cases which greatly affected the study outcomes leading to inaccurate results.

5.2 Recommendations

Based on the literature review, it is evident that depression is a major killer of disease that affects the economic and social well-being of the victims. Thus, there is a need to leverage the benefits of Machine Learning techniques in developing the most effective model for detecting depression at an early stage among patients to help in preventing the adverse effects of the disease, including suicidal cases among the victims.

Regarding the analysis of the study outcomes, there is a need to capture the depression patterns on various datasets as well as compare the datasets to obtain more accurate results. Such approaches are critical in ensuring effective detection and prevention of depression among the associated victims. It is recommended

to use the primary dataset in the research study to obtain first-hand information thus improving the accuracy of the study outcomes.

5.3 Future Work

Since the current research study could not capture the patterns associated with depression as well as compare the outcomes between the first and the second datasets, there is a need to conduct another study in the future that addresses such shortcomings. By doing so it will ensure the most effective model is selected for detecting depression cases among patients at an early stage.

The current study relied on secondary data for conducting the analysis, thus, conducting another research study using primary data could give more reliable results.

The study employed Depression Anxiety Stress Scales (DASS) in measuring and recording the respondent's demographic characteristics and Depression, Anxiety, and Stress score, thus, conducting another study that utilises a different method could give more accurate results.

BIBLIOGRAPHY

[1] WHO. (2020, January 30). *Depression*. World Health Organisation. https://www.who.int/news-room/fact-sheets/detail/depression

[2] Steger, M. F., & Kashdan, T. B. (2009). Depression and Everyday Social Activity, Belonging, and Well-Being. *Journal of counselling psychology*, *56*(2), 289–300. https://doi.org/10.1037/a0015416

[3] Baskin, C., Zijlstra, G., McGrath, M., Lee, C., Duncan, F. H., Oliver, E. J., Osborn, D., Dykxhoorn, J., Kaner, E., LaFortune, L., Walters, K. R., Kirkbride, J., & Gnani, S. (2021). Community-centred interventions for improving public mental health among adults from ethnic minority populations in the UK: a scoping review. *BMJ Open*, *11*(4), e041102. https://doi.org/10.1136/bmjopen-2020-041102

[4] Top 12 Data Analyst Tools - Best Software For Data Analysts. (2020). Datapine. https://www.datapine.com/articles/data-analyst-tools-software

[5] James C. Coyne & Margaret M. Calarco (1995) Effects of the Experience of Depression: Application of Focus Group and Survey Methodologies, Psychiatry, 58:2, 149-163, DOI: 10.1080/00332747.1995.11024722

[6] Jacques-Aviñó, C. (2020, November 1). *Gender-based approach on the social impact and mental health in Spain during COVID-19 lockdown: a cross-sectional study*. BMJ Open. https://bmjopen.bmj.com/content/10/11/e044617

[7] Shiba, K. (2017, May 30). *Retirement and mental health: does social participation mitigate the association? A fixed-effects longitudinal analysis*. BMC Public Health. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-017-4427-0

[8] McLaughlin K. A. (2011). The public health impact of major depression: a call for interdisciplinary prevention efforts. *Prevention Science: the official journal of the Society for Prevention Research*, *12*(4), 361–371. https://doi.org/10.1007/s11121-011-0231-8

[9] Ustun, G. (2020). Determining depression and related factors in a society affected by the COVID-19 pandemic. *International Journal of Social Psychiatry*, *67*(1), 54–63. https://doi.org/10.1177/0020764020938807

[10] Yok-Fong Paat & Christine Markham (2021) Digital crime, trauma, and abuse: Internet safety and cyber risks for adolescents and emerging adults in the 21st century, Social Work in Mental Health, 19:1, 18-40, DOI: 10.1080/15332985.2020.1845281

[11] Salari, N. (2020, July 6). *Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis*. Globalization and Health. https://globalizationandhealth.biomedcentral.com/articles/10.1186/s12992-020-00589-w

[12] Alvarez-Galvez, J. (2019, April 25). *Measuring the impact of multiple discrimination on depression in Europe*. BMC Public Health. https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-6714-4

[13] Shuang Gao (2018, August 23). *Machine Learning in major depression: From classification to treatment outcome prediction.* CNS Neuroscience & Therapeutics. https://onlinelibrary.wiley.com/doi/full/10.1111/cns.13048

[14] Adam Mourad C. (2016, March). *Cross-trial prediction of treatment outcome in depression: a machine learning approach.* https://www.sciencedirect.com/science/article/abs/pii/S221503661500471X

[15] Anu Priya (2020, April). *Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms.* https://www.sciencedirect.com/science/article/pii/S1877050920309091

[16] Smart Vision (2019, March). *Crisp DM Methodology*. https://www.sv-europe.com/crisp-dm-methodology/

[17] Xiaowei Li (2019, August). *Depression recognition using machine learning methods with different feature generation strategies.* https://doi.org/10.1016/j.artmed.2019.07.004

[18] Arkaprabha Sau (2017, November). *Predicting anxiety and depression in elderly patients using machine learning technology.* https://doi.org/10.1049/htl.2016.0096

[19] Meenal J. Patel (2015, February). *Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction.* https://doi.org/10.1002/gps.4262

[20] Raymond Chiong (2021, August). *A textual-based featuring approach for depression detection using machine learning classifiers and social media texts.* https://doi.org/10.1016/j.compbiomed.2021.104499

[21] Rouzbeh Razavi (2020, January). *Depression screening using mobile phone usage metadata: a machine learning approach.* https://doi.org/10.1093/jamia/ocz221

[22] Yena Lee (2018, December). *Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review.* https://doi.org/10.1016/j.jad.2018.08.073

APPENDICES

Appendix 1: Data description tables

OpenPsychometrics.org data dictionary, the columns are variable, variable description, missing values and percentage of missing values.

Variable	Description	Type	missing	missing%
education	How much education have you completed?: 1=Less than high school, 2=High school, 3=University degree, 4=Graduate degree	Categorical	515	1.295%
urban	What type of area did you live when you were a child?:1=Rural (country side), 2=Suburban, 3=Urban (town, city)	Categorical	382	0.960%
gender	What is your gender?: 1=Male, 2=Female, 3=Other	Categorical	67	0.168%
engnat	Is English your native language?:1=Yes, 2=No	Categorical	52	0.131%
age	How many years old are you?	Numeric	0	0.000%
hand	What hand do you use to write with?:1=Right, 2=Left, 3=Both	Categorical	173	0.435%
religion	What is your religion?:1=Agnostic, 2=Atheist, 3=Buddhist, 4=Christian (Catholic), 5=Christian (Mormon), 6=Christian (Protestant), 7=Christian	Categorical	356	0.895%

	(Other), 8=Hindu, 9=Jewish, 10=Muslim, 11=Sikh, 12=Other			
orientation	What is your sexual orientation?:1=Heterosexual, 2=Bisexual, 3=Homosexual, 4=Asexual, 5=Other	Categorical	3109	7.816%
race	What is your race?:10=Asian, 20=Arab, 30=Black, 40=Indigenous Australian, 50=Native American, 60=White, 70=Other	Categorical	0	0.000%
voted	Have you voted in a national election in the past year?:1=Yes, 2=No	Categorical	327	0.822%
married	What is your marital status?:1=Never married, 2=Currently married, 3=Previously married	Categorical	195	0.490%
familysize	Including you, how many children did your mother have?	Numeric	0	0.000%
Total_score	Depression total score from DASS	Numeric	0	0.000%

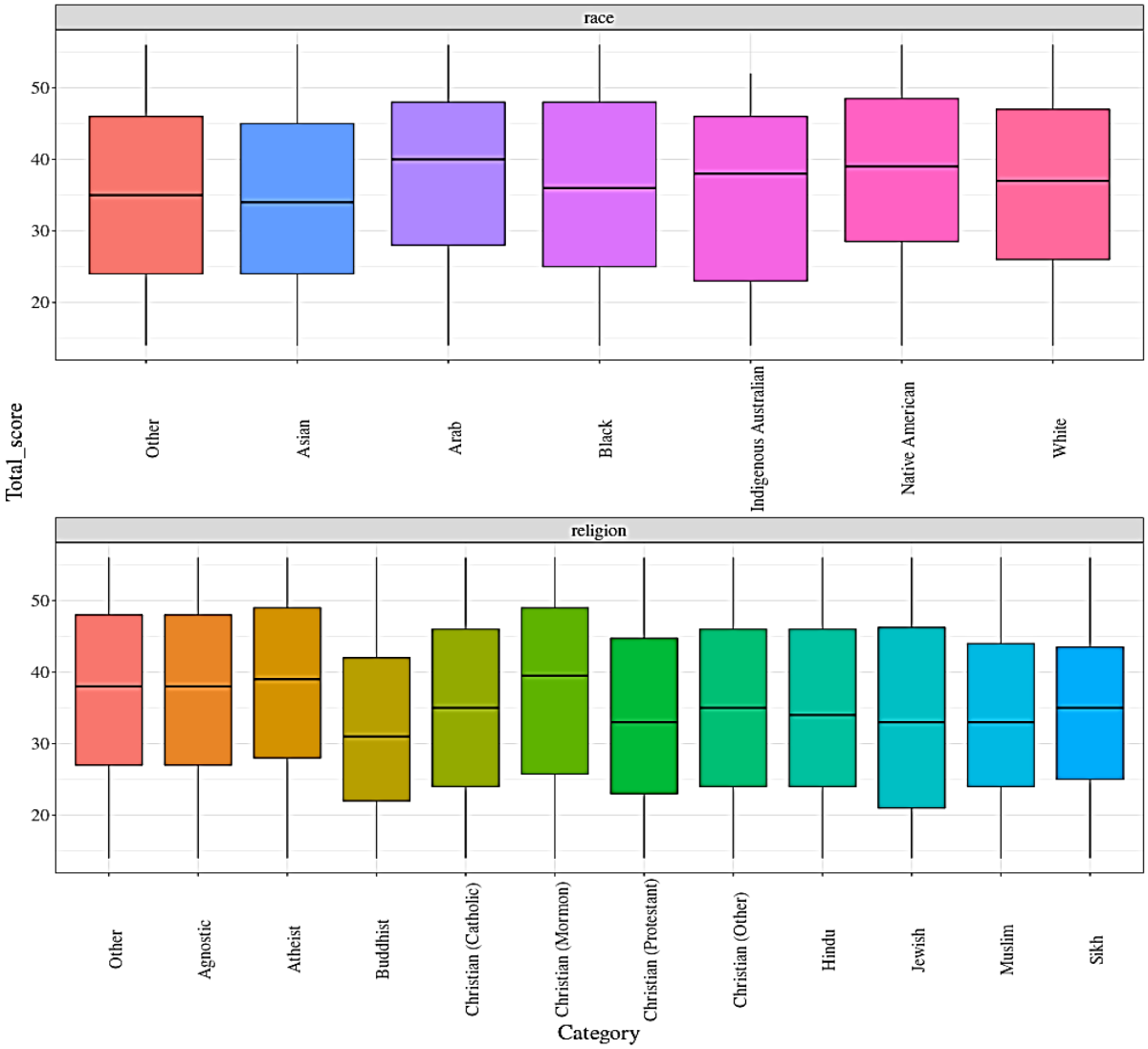
Data dictionary, the columns are variable, variable description, missing values and percentage of missing values.

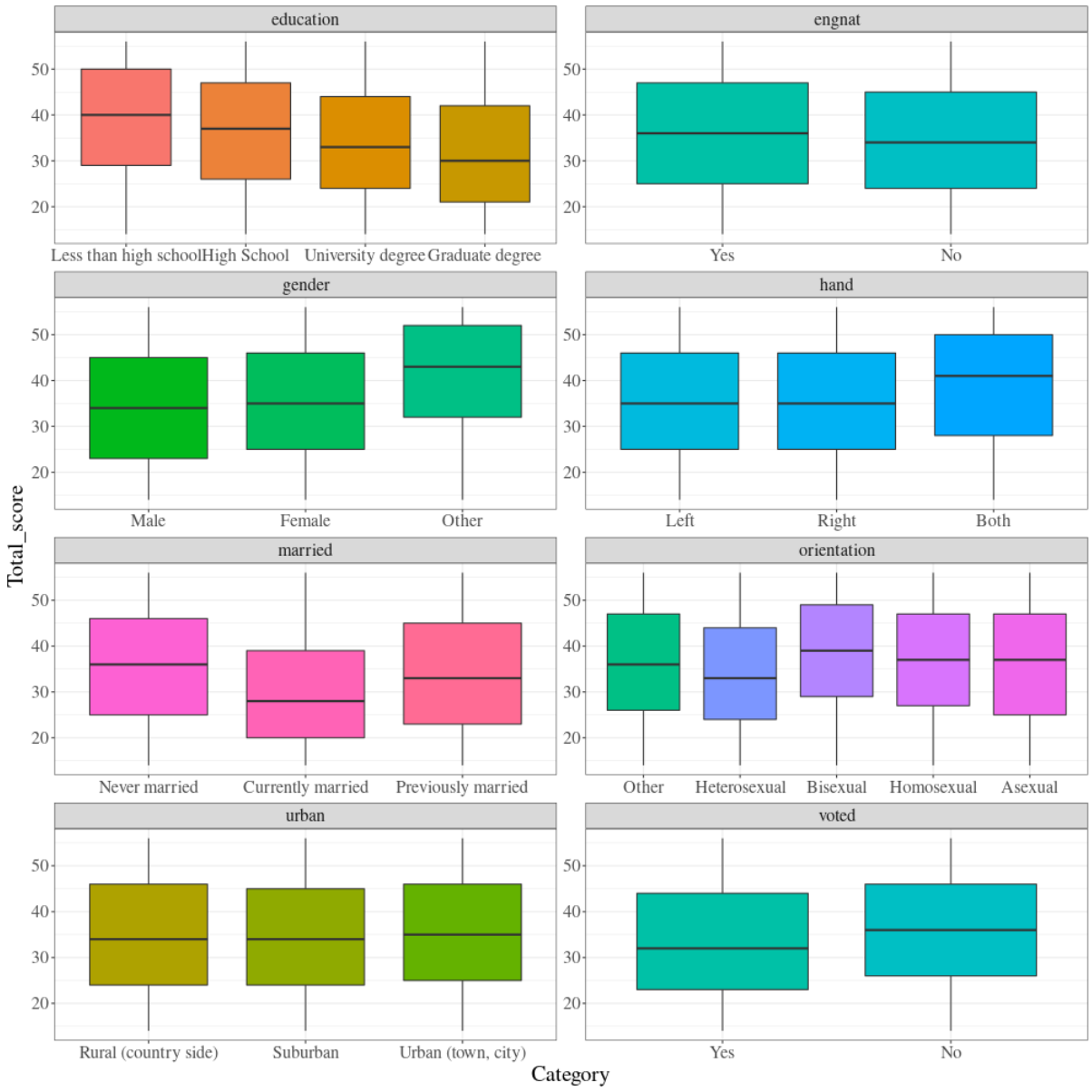
Variable	**description**	**missing %**	**Type**
surveyid	Individual Identifier	0.00%	Numeric
village	Village Identifier	0.00%	Numeric
survey_date	Date of Interview (days since Jan1 of first year)	0.00%	Date
femaleres	Female respondent	0.00%	Categorical

age	Age (respondent)	0.00%	Numeric
married	Marital status (respondent)	0.00%	Categorical
children	Number of children	0.00%	Numeric
hhsize	Household size	0.00%	Numeric
edu	Years of education completed (respondent)	0.00%	Numeric
hh_children	Number of children <=18 or younger in Household	0.00%	Numeric
hh_totalmembers	Household size	29.22%	Numeric
cons_nondurable	Non-durable expenditure (USD)	0.00%	Numeric
asset_livestock	Value of livestock (USD)	0.00%	Numeric
asset_durable	Value of durable goods (USD)	0.00%	Numeric
asset_phone	Value of cell phone (USD)	0.00%	Numeric
asset_savings	Value of savings (USD)	0.00%	Numeric
asset_land_owned_t otal	Land owned (acres)	0.00%	Numeric
asset_niceroof	Has non-thatched roof (dummy)	0.00%	Categorical
cons_allfood	Food total (USD)	0.00%	Numeric
cons_ownfood	Food own production (USD)	0.00%	Numeric
cons_alcohol	Alcohol (USD)	3.76%	Numeric
cons_tobacco	Tobacco (USD)	1.75%	Numeric
cons_med_total	Medical expenditure past month (USD)	0.00%	Numeric
cons_ed	Education expenditure (USD)	0.00%	Numeric
cons_social	Social expenditure (USD)	0.00%	Numeric
cons_other	Other expenditure (USD)	0.00%	Numeric
ent_wagelabor	Wage labor primary income (dummy)	0.00%	Categorical
ent_ownfarm	Own farm primary income (dummy)	0.00%	Categorical
ent_business	Non-ag business primary income (dummy)	0.00%	Categorical
ent_nonagbusiness	Non-agricultural business owner (dummy)	0.00%	Categorical
ent_employees	Number of employees working in non-ag business	0.00%	Numeric
ent_nonag_revenue	Non-ag business revenue, monthly (USD)	0.00%	Numeric
ent_nonag_flowcost	Non-ag business flow expenses, monthly (USD)	0.00%	Numeric
ent_farmrevenue	Farm revenue, monthly (USD)	0.00%	Numeric
ent_farmexpenses	Farm flow expenses, monthly (USD)	0.00%	Numeric
ent_animalstockrev	Livestock sales and meat revenue, monthly (USD)	0.00%	Numeric

ent_total_cost	Total expenses, monthly (USD)	0.00%	Numeric
fs_adskipm_often	Meals skipped (adults, \# last month)	0.00%	Numeric
fs_adwholed_often	Whole days without food (adults, \# last month)	0.00%	Numeric
fs_meat	Number of times ate meat or fish (last week)	29.22%	Numeric
fs_enoughtom	Enough food in the house for tomorrow? (dummy)	29.22%	Categorical
fs_sleephun	Respondent slept hungry (last week, dummy)	29.22%	Categorical
med_portion_sickinjured	Proportion of household sick/injured (1 month)	29.22%	Numeric
med_sickdays_hhave	Average number of sick days per HH member	29.22%	Numeric
med_vacc_newborns	Number of newborns vaccinated	0.00%	Categorical
med_child_check	Proportion of children <14 getting checkup (6 months)	0.00%	Categorical
labor_primary	Casual or Wage Labor Primary Source of Income	0.00%	Categorical
wage_expenditures	Expenditure on wages for HH enterprise	0.00%	Numeric
durable_investment	Durable Investments	0.00%	Numeric
nondurable_investment	Non-durable Investments	0.00%	Numeric
given_mpesa	Sent money using M-Pesa	0.00%	Categorical
amount_given_mpesa	Amount sent using M-Pesa	0.00%	Numeric
received_mpesa	Received money using M-Pesa	0.00%	Categorical
amount_received_mpesa	Amount received using M-Pesa	0.00%	Numeric
net_mpesa	Net Remittances using M-Pesa	0.00%	Numeric
saved_mpesa	Saved money using M-Pesa	0.00%	Categorical
amount_saved_mpesa	Amount saved using M-Pesa	0.00%	Numeric
early_survey	Psychology survey in 1st wave (dummy)	0.00%	Categorical
depressed	Meets epidemiological threshold for moderate depression (dummy)	0.00%	Categorical

Appendix 2:Box plots of total DASS Score


race
Total_score
50
40
30
20
Other
Asian
Arab
Black
Indigenous Australian
Native American
White
religion
50
40
30
20
Other
Agnostic
Atheist
Buddhist
Christian (Catholic)
Christian (Mormon)
Christian (Protestant)
Christian (Other)
Hindu
Jewish
Muslim
Sikh
Category


education
engnat
gender
hand
married
orientation
urban
voted
Total_score
50
40
30
20
Less than high school
High School
University degree
Graduate degree
Yes
No
Male
Female
Other
Left
Right
Both
Never married
Currently married
Previously married
Other
Heterosexual
Bisexual
Homosexual
Asexual
Rural (country side)
Suburban
Urban (town, city)
Yes
No
Category

Appendix 3: Frequency distribution for the Busara centre dataset

mean gives the percentage count for each variable. Standard deviations are in bracket.

Variable	Category	n	mean(sd)	median(IQR)	skewness	min	max
asset_niceroof	0	1142	99.91%(0.01%)	_	_	_	_
	1	1	0.09%(0.01%)	_	_	_	_
depressed	0	950	83.11%(0.11%)	_	_	_	_
	1	193	16.89%(0.11%)	_	_	_	_
early_survey	0	1031	90.20%(0.09%)	_	_	_	_
	1	112	9.80%(0.09%)	_	_	_	_
ent_business	0	1020	89.24%(0.09%)	_	_	_	_
	1	123	10.76%(0.09%)	_	_	_	_
ent_nonagbusiness	0	835	73.05%(0.13%)	_	_	_	_
	1	308	26.95%(0.13%)	_	_	_	_
ent_ownfarm	0	860	75.24%(0.13%)	_	_	_	_
	1	283	24.76%(0.13%)	_	_	_	_
ent_wagelabor	0	930	81.36%(0.12%)	_	_	_	_
	1	213	18.64%(0.12%)	_	_	_	_
femaleres	0	95	8.31%(0.08%)	_	_	_	_
	1	1048	91.69%(0.08%)	_	_	_	_

fs_enoughtom	0	587	72.56%(0.16%)	_	_	_	_
	1	222	27.44%(0.16%)	_	_	_	_
fs_sleephun	0	512	63.29%(0.17%)	_	_	_	_
	1	297	36.71%(0.17%)	_	_	_	_
given_mpesa	0	1125	98.43%(0.04%)	_	_	_	_
	1	18	1.57%(0.04%)	_	_	_	_
labor_primary	0	890	77.87%(0.12%)	_	_	_	_
	1	253	22.13%(0.12%)	_	_	_	_
married	0	260	22.75%(0.12%)	_	_	_	_
	1	883	77.25%(0.12%)	_	_	_	_
med_child_check	1	1143	100.00%(0%)	_	_	_	_
med_vacc_newborns	1	1143	100.00%(0%)	_	_	_	_
received_mpesa	0	1067	93.35%(0.07%)	_	_	_	_
	1	76	6.65%(0.07%)	_	_	_	_
saved_mpesa	0	927	81.10%(0.12%)	_	_	_	_
	1	216	18.90%(0.12%)	_	_	_	_