Solent University

Faculty of Business, Law and Digital Technologies

MSc APPLIED AI AND DATA SCIENCE

2021/22

Andrew Akinosho

"Using machine learning and a predictive model; Can signs of depression be predicted in young adults within education."

Supervisor : Dr Peyman Heydarian
Date of submission : September 2022

Contents

Table of Figures

Tables

Acknowledgements

I want to start by thanking my supervisor Dr Peyman Heydarian, who supported me from the start and taught me the importance of research and organisation of a thesis project. He has guided me with feedback which has helped me better understand and research writing.

I would also like to thank my lecturers and peers in this MSc programme for the support and teaching of AI and Data science.

Finally, my siblings, family and friends have supported me and guided me toward succeeding and doing my best. Without your encouragement and motivation, things would be more problematic during this challenging time.

Abstract

Mental health is becoming more known among individuals still there is a gap in knowledge and a stigma around it. Depression is a common mental health diagnosis, with around 5% of adults who suffer from this. (WHO, 2021). Two weeks of consistent sadness and other factors can lead to depression. This research uses machine learning to predict signs of depression in young adults in education. Machine Learning is a method which can predict the outcome from data. This research will show how data was collected, inspired by the Depression Anxiety Stress Survey (DASS-21), to find critical variables. From this survey, a questionnaire was created and distributed to the target audience via student platforms and social media to complete. This research specifically targets young adults between 18-35 in education. Data were pre-processed using Excel and Jupyter Notebook (Python) to work with data to build a suitable prediction model. A key finding from this research was that life was meaningless and scared without reason with high social media usage. It was noticed that a participant who said they feel like this most of the time had medium or high signs of depression. A classification model, Random Forest Classifier, proved best out of eight classifiers tested with an accuracy of 78%. Many studies and research have been conducted with uniqueness to their aims and methodology but close in accuracy scores. It will be good to get a larger dataset from the same data collection source, try to improve the accuracy, and launch this application to education establishments.

1 Introduction

Compared to other research topics, mental health is fascinating since it attracted personal interest and sparked a sense of wonder. Tayo (mother), who has worked as a mental health nurse in the public health sector for more than ten years, is the source of the interest. The goal is to carry out research that relates the two, having learnt a lot, become a champion for eradicating the stigma surrounding mental health, and developed a passion for AI and Data Science. We all understand how important technology is to our daily lives. It is incredible to comprehend how machine learning functions and the various forecasting models that can be applied to just about anything where data is available. Mental health and machine learning are two related concepts.

A minimum of two weeks of consistent sadness can indicate depression, a psychological condition. It makes it difficult for people to carry out daily tasks, and depressed people stop finding pleasure in the activities they once enjoyed (World Health Organization, 2021). Anxiety and depression have a wide range of multifactorial reasons, including those that are biological, economic, social, environmental, and cultural. Psychiatrists, psychologists, and other mental health medical experts typically make diagnoses. Our increased awareness motivates us to learn more about these two topics. Mental health has various facets, including anxiety, depression, bipolar disorder, personality disorder, and many others. Mental health is an umbrella term that covers a wide range of topics. A key idea in AI prediction that is currently gaining popularity and interest is machine learning.

The research topic and the query I am attempting to solve or respond to is:

Using machine learning and a predictive model; Can signs of depression be predicted in young adults within education.

This study focuses on young adults in the UK's educational system between the ages of 18 and 35. The target audience for this research project will be one of the primary objectives. This is a research gap related to my topic. Whom this is aimed at and precisely examining depression and anxiety makes this unique. While earlier research has been completed, it has been noticed that the term "mental health" for elderly individuals or a particular group of people is broad.

Other research has been a pinnacle and inspiration to create a tool and conduct analysis on other studies and projects. As a focus, one of the key features is the unique target group. In the initial stage, the research approach will utilise questionnaires to gather data and machine learning algorithms to forecast a possible score with the possibility of having one of these illnesses. It should be noted that this is not a diagnosis.

1.1 Aims and Objectives

By doing this research and thesis, personally would like to achieve greater understanding, build knowledge, and further my development in mental health and AI/Data Science. The aims and objectives are:

1. To create a questionnaire to generate of a dataset
2. Build an artefact where a predictive model is used to make a prediction and give a rating or percentage score.
3. Evaluate several machine learning algorithms and feature selection techniques to effectively detect the signs of depression with an accuracy score of between 75-90%
4. To build a user-friendly application which will hopefully be able to be live and used to get a score from my predictive model.

The remainder of the paper is organised in the manner described in the following structure: the literature review is described in Chapter 2. The entire methodology is explained in Chapter 3. Chapter 4 exhibits the study results with a segment of the discussion. The conclusion and summary with Limitations of the study will be outlined in chapter 5.

2 Literature Review

Most of the research has focused on older people who already have a mental health diagnosis. Additionally, studies have demonstrated that projects that focus on patients and persons who have decided to participate in the project are watched in clinical settings by medical professionals who employ practices that demonstrate signs in addition to assessment. With this, the authors have demonstrated that the project's machine learning component produces a result that is either true or false or yes or no. The examined papers revealed a pattern of events occurring in a hospital setting. Because of the controlled setting, this could be a advantage that they can push for this outcome.

The authors *Mohd and Mutalib (2020)* have taught us about the research done in Malaysia. A variety of factors influence higher education for young adults. On the one hand, family issues, hazy plans and aspirations for future career prospects, money issues, and living away from home raise the risk of mental health issues for students and manage the pressures of both university life and other aspects of daily living. These variables' results displayed a familiar pattern. They used a supervised machine learning algorithm, the Support Vector Machine model (SVM), with an accuracy of 70–90%.

Jain et al. (2019) conducted research to see the ability to predict depressed moods is still an open subject despite the extensive research on understanding individual moods, including depression, anxiety, and stress-backed activity data gathered by ubiquitous computing devices like smartphones. In order to anticipate suicidal behaviours based on the amount of depression, we have suggested a depression analysis and suicidal ideation detection system in this study. Parents and pupils completed questionnaires modelled after the PHQ-9 (Parent Health Questionnaire), which included questions like What is your age? to provide real-time data, Do you attend classes regularly? Then transformed the responses into valuable data that included relevant characteristics like age, sex, frequent attendance at school, and more. Then, classification machine algorithms are utilised to train and categorise the data into five severity-based phases of depression: minimal or none, mild, moderate, fairly severe, and severe. Using the XGBoost classifier, the maximum

accuracy of 83.87% was attained in this dataset. Additionally, information was gathered in the form of tweets and using classification algorithms, and it was determined whether the tweeter was depressed or not. The Logistic Regression classifier provided the highest accuracy, or 86.45%, for the same.

Additionally, research demonstrates the role social media plays in depression. *Immanuel et al. (2022)* concluded the research. The writers noted young adults' ability to observe the symptoms of depression via Twitter. Examining Twitter and associated tweets that promote early recognition and understanding of depression aligns with this thesis. It was unusual to see a time series model Long Short-Term Memory (LSTM), applied to identify depression by discovering and training the tweets in a dataset that showed depression-related symptoms. Understanding the authors' strategy will increase awareness since readers can spot long-term patterns and trends.

Students attending universities in Northern Island were the subject of a 2019 study by *McLafferty et al.* on mental health. They employed a method of data collecting that will be comparable. They used the World Mental Health International College questionnaire surveys to create their dataset. This has motivated me to examine official mental health evaluations and identify the variables I can exploit to accomplish my objective. It was exciting to observe that the authors' data analysis, rather than machine learning, allowed them to achieve the goals they set out to achieve. This paper does provide some beneficial aspects, but the machine learning aspect section is missing.

In this study, *Uddin et al.* seek to find critical trends in the performance and application of various supervised machine learning algorithms for health risk prediction. Machine learning is developing quickly, and these methods provide solid predictive capabilities. As the authors delve deeper, it will become clear that machine learning may also help identify those at risk for mental health issues. Hidden mental health problems may result from numerous factors, including genes, surroundings including poverty, hunger, and childhood adversity, as well as the interplay between the two. We discovered that the Nave Bayes technique and Support Vector Machine (SVM) algorithm are used most frequently (in 29 studies each) (in 23 studies). Comparatively speaking, the Random Forest (RF) algorithm demonstrated greater accuracy. In nine of the seventeen studies where it was used, or 53% of them, RF demonstrated the highest accuracy. SVM, which outperformed 41% of the studies, came in second.

Data from the Medical College and Hospital of Kolkata, West Bengal, were gathered by *Sau et al. (2017)* on 630 elderly patients, 520 of whom were receiving exceptional care. Random forest gave the best accuracy rates of 91% and 89%, respectively, among the two data sets of 110 and 520 individuals after they applied several classification methods and algorithms, including logistic, Naive Bayes, random forest, and random tree, and J48. WEKA was the most effective tool for feature categorisation and selection.

Doubtlessly, a person's emotions, intellect, and ability to communicate with others are all affected by mental illness, which is a health issue. In order to educate clinical treatment, the team wants to summarise the most recent research on machine learning techniques for foretelling mental health issues. This review study will identify the sorts of machine learning algorithms that have been extensively employed in this field. This review study contains a total of 30 research papers. Most research articles demonstrate that machine learning models have achieved more than 70% accuracy levels. In machine learning for mental health, plenty of issues still need to be identified and tested in several scenarios. Some of the authors' findings may be consistent with earlier research in this field: Random Forest and support vector machines have been the most often used machine learning models in the studies. According to *Teo et al.(2022),* this is due to their capacity to deliver outstanding performance in terms of accuracy. They argue that this systematic literature review research will cover recent developments. The study offers a significant summary of the knowledge gaps regarding machine learning applications in mental health. It also identifies future directions for further investigation.

According to *Na et al. (2020),* the Republic of Korea would soon experience a depression. The Random Forest classifier was used to build the predictive model. SMOTE was utilised to address the problems of class disparities. SMOTE is a statistical technique called Synthetic Minority Oversampling Technique which can be used to increase the number of cases in a dataset evenly. The accuracy of this study was 86.20%. According to this study, the two main factors influencing the start of depression are satisfaction with one's health and one's satisfaction with one's socio-familial relationships.

Garriga et al.(2022) wrote in "Machine learning model to predict mental health crises using electronic health records" that early detection of individuals at risk for a mental health crisis can result in better outcomes and the reduction of burdens and expenses. They used machine learning algorithms on longitudinally gathered EHR data to show that it is possible to forecast mental health crises. 17,122 patients' EHR data was collected over seven years (2012-2018) into a machine learning algorithm, and the researchers created a mental crisis risk model. Four distinct groups of clinicians received the crisis predictions on a bimonthly basis, and they assessed if and how they improved their ability to manage workload priorities. The research involved sixty clinicians. "Main limitation is the known and potentially unknown specificity of the single-centre cohort," the article states when discussing potential drawbacks. The authors restricted the use of our approach to patients with a history of relapse. They acknowledge that the idea of using an algorithm to identify early crises needs exploring further. They suggest that the physicians stated that the prediction model assisted 19% of cases in averting a crisis. Because it would have meant that the clinicians did not respond to the forecasts, which would have been unethical and illegal, this scenario was not seen

Hatton et al. (2019) used 284 older individuals' psychometric and demographic data to forecast the prevalence of depression. To predict the persistence of depression, they used the Extreme Gradient Boosting method and evaluated its effectiveness against the Logistic Regression model. They claimed that Extreme Gradient Boosting surpassed Logistic Regression in terms of performance.

Using machine learning algorithms, this study assessed five different anxiety levels, depression, and stress intensity. A basic questionnaire measuring the common signs of stress, depression, and anxiety was used to gather the data. Following that, Decision Tree, Random Forest Tree, Nave Bayes, Support Vector Machine, and K-Nearest Neighbour were used as five different algorithms. Despite Random Forest being the best model in this study, naive Bayes was shown to have the highest accuracy of 85.50%. . The best-model selection was based on the f1 score, employed in imbalanced partitioning cases because this problem resulted in unbalanced classes. For the scales of anxiety, depression, and stress, respectively, the key variables were "scared without any good reason," "Life was meaningless," and "Difficult to relax." the following variables were thought to be the most crucial in identifying psychological conditions *Priya et al. (2020)* revealed that predictions of anxiety, depression, and stress were noticed using machine learning algorithms in their article "Predicting Anxiety, Depression, and Stress in Modern Life using Machine Learning Algorithms." The Depression, Anxiety, and Stress Scale questionnaire gathered information from employed and unemployed people from various cultures and communities. Stress, depression, and other psychological health conditions have grown widespread among the public in today's fast-paced environment. Five distinct machine learning algorithms each predicted depression and stress to develop at five distinct levels of severity.

Zarandi et al. (2019) deployed type-2 fuzzy logic to determine the severity of depression. They used the Mutual Information Feature Selection (MIFS) method to increase the study's accuracy and forecast the patients' level of depression with fewer questions. Their suggested method had an accuracy of 84.00% and only required fifteen questions to predict the severity of depression.

Using the Long Short-Term Memory (LSTM) and Natural Language Processor model, *Zeberga et al. (2022)* investigated a unique text mining approach for mental health prediction. The analysis comprised 100,000 persons. A clever, context-aware deep learning system based on bidirectional encoder representations from transformers is proposed in this paper. After numerous hyperparameter tweaks, the model outperforms the compared approaches and offers an accuracy of 98%. The system detecting mental health issues uses the most recent deep learning-based text embedding method. In addition to being a significant cause of disability, mental health issues also significantly increase the global illness burden. It has been determined that depressive disorders are one of the leading causes of non-fatal health loss. With a suicide rate of 10.5 per 100,000 persons, suicide has emerged as a leading cause of mortality in young people. Zeberga and colleagues contend that in subsequent research, multiple uses of models for depression detection

systems can be built to use a broader range of data, such as text, image, and behavioural aspects.

Choudhury et al. (2019) researched to identify depression among Bangladeshi undergraduate students. They initially gathered data on 935 students. They used the data of 577 students for their study after purifying the data and using various data pre-processing techniques. They attained the best accuracy by using the Random Forest classifier, which was 75%.

The life steps, from childhood to adulthood, place a high value on mental health. The accuracy of five machine learning techniques in this study's identification of mental health concerns was evaluated using a variety of accuracy metrics. It is important to be psychologically, emotionally, and socially well. One's thoughts, feelings, and actions are influenced by their mental health. This study used five machine learning methods: random forest, decision tree stacking, logistic regression, and k closest neighbour classifier. The group evaluated how well they were able to spot mental health problems. A person's level of mental health acts as a gauge for how to treat their illnesses effectively. The researchers reviewed the fifty-five transdisciplinary studies.

In "Natural language processing applied to mental disease identification," *Zhang et al. (2022)* noted that mental health issues are widespread worldwide and have long been a significant public health concern. They can gather text data about mental health from various sources, including social media posts, screening surveys, narrative writing, interviews, and EHRs. Most of the methods covered in this review used supervised learning models. They offer a narrative review of NLP's use in the previous ten years to detect mental illness. The majority of the sources are social media posts, followed by interviews, electronic health records, screening surveys, and narrative writing. Both machine learning and deep learning techniques have produced promising outcomes. The authors examined 399 studies. They say their findings support earlier research on the topic: "Rutowski et al. used transfer learning to pre-train a model using an open dataset. The outcomes demonstrated the value of pre-training, according to *Zhang et al.* They claim that they are still viewed as mysterious black boxes and that the forecasts are left unanswered. The precision of deep learning models will be a crucial area of research in the future.

Shatte et al.(2019) at Federation University used a scoping review process in 2019 to quickly map the field of machine learning in mental health. There were found to be three hundred studies devoted to utilising ML to improve mental health. The literature identified four key application domains: prognosis, treatment, detection and diagnosis, and support. This study aims to synthesise the research on big data and machine learning applications for mental health. Support vector machines, decision trees, neural networks, latent Dirichlet allocation, and clustering were among the ML methods used. Depression, schizophrenia, and Alzheimer's were discussed among the most prevalent mental health issues.

Tate et al.(2020) conducted research at the Department of Medical Epidemiology and Biostatics that published a study on using machine learning to predict adolescent mental health issues. Psychopathology with a childhood onset can have severe consequences that last through adolescence and adulthood. There is currently no model available to screen the general population for the likelihood of acquiring mental health issues. The most accurate characteristics for intervention targets may be more clearly identified with research on the most informative mental health indicators. For early intervention and to avoid severe negative consequences in the future, it is essential to predict which kids will go on to develop mental health symptoms as teenagers. They used 474 predictors in 7,638 pairs of twins from the Child and Adolescent Twin Study in Sweden, which employed data from parental reports and registers. The top-performing model establishes the crucial groundwork for future models attempting to predict outcomes for general mental health, even though it is not appropriate for clinical application. The study involved 15156 sets of twins. Their findings seem to confirm what was already understood in this field: "Registry data on neighbourhood quality, parity, and gestational age of delivery were also considered essential. Tate states, "These findings are consistent with the literature and may be used by clinicians, parents, or educators to pinpoint children at risk." The researchers state, "Several limitations must be considered when interpreting the data. According to prior research, there is not much of a difference between twins and singletons in terms of mental health. Zygosity did not have a high importance ranking, showing that the twin resemblance was not the basis for the logistic regression model. They argue that, depending on desire, future research may employ cut-offs that have been verified for their nation or the original study. This proves that the more severe instances do not constitute a separate severe class. Although there was a lot to learn in this text, it was outside the scope of the research that needed to be accomplished.

Khondoker et al. (2013), from King's College London, under the direction of, examined a comparison of machine learning approaches for categorisation utilising simulation and several actual data examples from mental health studies. Linear Discriminant Analysis (LDA), Random Forests (RF), Support Vector Machines (SVM), and k-Nearest Neighbour were contrasted. LDA was discovered to be the best approach in terms of average generalisation errors and stability of error estimates. SVM performs significantly better than LDA, Random Forest, and kNN. Noticing this performs better in a few cases does not mean it will be the case on average or for the entire population. When there are fewer linked characteristics—generally, fewer features than half the sample size—Linear Discriminant Analysis was shown to be the method of choice. Even though there was a lot to learn within this text, this was out of scope for the research looking to be completed.

From its beginning to February 8, 2018, *Lee et al. (2018)* searched Ovid MEDLINE/PubMed for relevant studies. They evaluated the therapeutic results with a pharmaceutical or manual-based psychotherapy intervention for depression. Both a regression analysis and a random forest analysis of proportions were carried out. An integrated approach may more accurately describe the symptoms of mental

conditions as functional modules of pathology nested within the complex social dynamics. Classification algorithms might predict therapeutic outcomes with an overall accuracy of 0.82.

The prediction of depression in breast cancer patients was carried out by *Cvetkovi et al. (2017)*. Through a two-phase interview process, the information of eighty-four patients between the ages of 30-78 was gathered for this study. The patients' socio-demographic data was gathered in the initial phase. The second part involved administering the patients the standardised Beck Depression Inventory (BDI) test. The BDI test was used to determine the patients' actual depression range. The depression range obtained from the BDI test was the goal variable and attributes obtained from the socio-demographic data served as the predictor factors. This study assessed the performance of three different algorithms: a fuzzy genetic algorithm, an artificial neural network (ANN), and an ANN with a backpropagation learning method algorithm. The ANN using the extreme learning algorithm performed the best in this situation.

Chekroud et al. (2016) stated that antidepressant treatment efficacy is low but might be increased by matching patients to treatments in their paper, "Cross-trial prediction of treatment outcome in depression." They sought to create an algorithm to determine whether patients would experience symptomatic remission after a 12-week citalopram therapy. Out of 164 patient-reported characteristics, the group found twenty-five to be the most reliable indicators of therapy success. According to internal cross-validation, the Sequenced Treatment Alternatives to Relieve Depression cohort's outcomes were predicted by the model with an accuracy significantly above chance (59.4% accuracy). The study included 4041 depressed patients.

According to the number of analysis mentioned above, most of the research that has been done thus far has attempted to predict depression in particular populations, such as patients with a particular diagnosis or people in a specific age range. By considering several ways of living and life conditions, this study strives to take an innovative approach and fill a gap in research aimed at young adults between the ages of 18 and 35 within the UK in education. Data on people of various ages, health issues, and work situations, to name a few, will be collected to aid in this study.

3 Methodology

3.1 Dataset

Below you see the dataset consisting of 292 rows and twenty-two columns

Figure 1. Dataset imported as csv from questionnaire platform (Zoho)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	How old a	Where are	Are you cu	How many	Are you cu	Do you ha	Have you	How much	How many	How often	How often	How often	Do you ex	Do you ha	Do you ex	Do you fin
2	29	United Kin	Yes	10	Employed,	No	No	1-2 hours	3-4 hours	1	2	1	0	2	1	1
3	28	United Kin	Yes	10	Employed,	No	No	3-4 hours	0	1	1	1	0	0	0	2
4	28	United Kin	Yes	20	Employed,	No	No	3-4 hours	3-4 hours	3	2	2	0	1	0	0
5	26	United Kin	Yes	6	Not emplc	No	Prefer not	3-4 hours	0	1	1	1	0	1	0	1
6	33	United Kin	Yes	2	Employed,	No	No	5+ hours	1-2 hours	0	0	2	0	2	0	2
7	24	United Kin	Yes	15	Employed,	No	No	5+ hours	1-2 hours	2	1	1	0	0	0	0
8	25	United Kin	Yes	20	Not emplc	No	No	1-2 hours	3-4 hours	2	0	1	0	2	0	3
9	18	United Kin	Yes	24	Employed,	No	No	5+ hours	1-2 hours	3	0	1	1	1	2	2
10	28	United Kin	Yes	8	Employed,	No	No	5+ hours	3-4 hours	1	1	1	0	1	0	2
11	25	United Kin	Yes	9	Employed,	No	Yes	3-4 hours	3-4 hours	2	1	2	0	1	0	3
12	25	United Kin	Yes	13	Employed,	No	Yes	3-4 hours	0	2	2	3	0	2	2	2
13	31	United Kin	Yes	20	Employed,	Yes	No	3-4 hours	0	1	1	0	0	1	0	2
14	28	United Kin	Yes	5	Employed,	No	No	3-4 hours	3-4 hours	1	1	1	0	1	0	2
15	29	United Kin	Yes	15	Employed,	No	Yes	5+ hours	3-4 hours	1	0	2	1	2	1	3
16	30	United Kin	Yes	10	Employed,	No	No	5+ hours	5+ hours	3	2	2	2	1	1	1
17	21	United Kin	Yes	25	Employed,	Yes	No	1-2 hours	5+ hours	1	1	0	0	1	0	0

Table 1. Dataset breakdown

Column Name	Variable Name (python)	Data Type	Value
How old are you?	age	Int	18 – 35
'Where are you currently based/living?	address	Int	United Kingdom
Are you currently in education?	education	Object	Yes No
How many hours a week do you spend studying?	stud_hr	Object	1-60 hours
Are you currently employed?	employed	Object	Employed, full time Employed, part time Not employed
Do you have any physical health issues or a disability?	h-disab	Object	'Yes' 'No' 'Prefer not to say'
Have you been previously diagnosed with a mental health condition?	ment_cond	Object	'Yes' 'No' 'Prefer not to say'
How much time a day do you spend on social media?	social_hr	Int	1-2 hours 3-4 hours 5+ hours
How many hours a week do you take in health/wellbeing	fit_hr	Int	1-2 hours 3-4 hours 5+ hours

activities? (e.g. gym, fitness)			
How often do you find it hard to wind down?'	wind	Int	*0 1 2 3
How often do you experience dryness of my mouth?	dry_mouth	Int	*0 1 2 3
How often do you struggle to experience feeling positive?	positive	Int	*0 1 2 3
'Do you experience breathing difficulties (eg, excessively rapid breathing, breathlessness in the absence of physical exertion)?	breath_diff	Int	*0 1 2 3
Do you have difficulties working up the initiative to do things?	initiate	Int	*0 1 2 3
Do you experience trembling (eg, in the hands, legs)?	tremb	Int	*0 1 2 3
Do you find yourself worrying about situations in which you might panic and make a fool of yourself?	worry	Int	*0 1 2 3
Do you find yourself feeling that you have nothing to look forward to?	look_fwd	Int	*0 1 2 3
Do you find yourself feeling down-hearted?'	down	Int	*0 1 2 3
'Do you find yourself unable to become enthusiastic about anything?'	enthuse	Int	*0 1 2 3

'Do you find yourself feeling that life can be meaningless?'	life_mean	int	*0 1 2 3
Do you find yourself feeling scared without any good reason?	scared	Int	*0 1 2 3
Outcome	outcome	Object	High signs of depression Low signs of depression Medium/High No signs of depression very Low/low Signs

*0 -Never, 1- Sometimes, 2- Often & 3 – highly frequently/almost always.

3.2 Data Collection

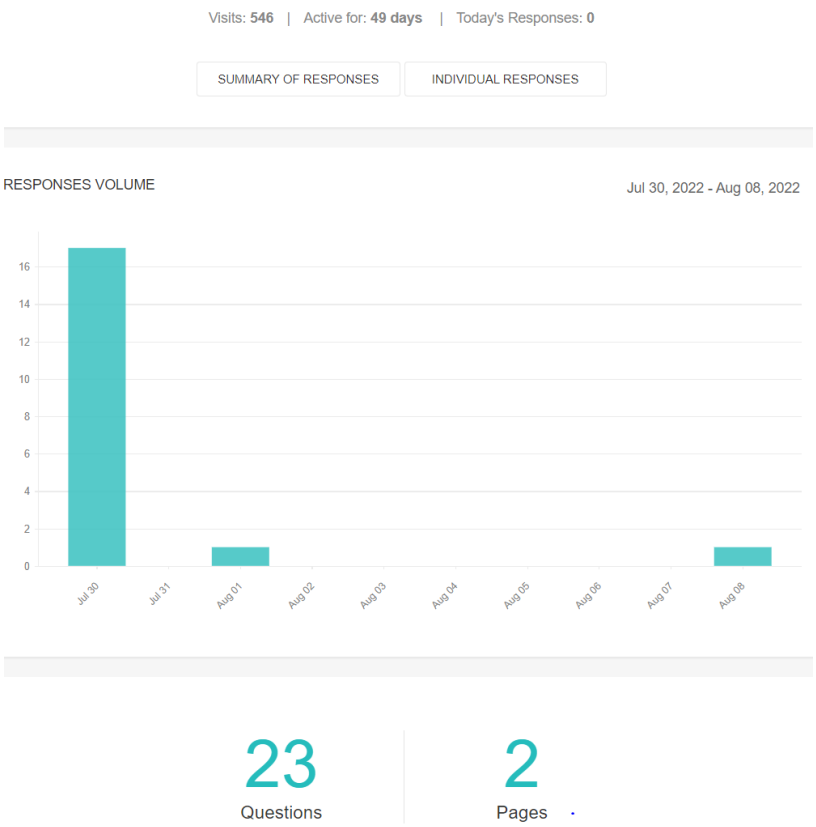
The practice of obtaining information from numerous sources is known as data collection. Just as important as the accuracy of a predictive model is the collection of appropriate data for AI initiatives.

The research focuses on mental health, focusing on depression and anxiety. Data was collected using a primary method which was in the form of a questionnaire. The reason for conducting primary data collection was influenced by previous research. Most researchers' data had been collected in live/real-time clinics and controlled environments. Making it difficult reviewing these datasets which makes this became hard to understand the whole picture and use within this research. Also, other researchers had not focused on a target group within education.

A questionnaire was the best form of data collection for building the dataset, organising and time constraints. The questionnaire was designed with the inspiration of the Depression Anxiety Stress Survey. Questions were implemented with a scoring system of 0 -Never, 1- Sometimes, 2- Often & 3 – highly frequently/almost always. These questions were situational based, and the experiences of an individual felt. Questions regarding everyday life, such as age, location, education, and employment status, were included. This was for eligibility of the research. No other personal or sensitive details was collected. [Appendix 3 – Research Zoho questionnaire]

The questionnaire was built using Zoho, a suitable platform for collecting data and created with twenty-three questions with simple, drop-down, and scale-based answers. This was with the aim of the questionnaire to be user-friendly, quick, and concise. Collecting the data in this format seemed best to fit my research purpose. The questionnaire was tailored to the data needed to be collected. This allowed for part of the aim and objectives to be met.

Figure 2. Question summary 546 visits and completion within 49 days



The questionnaire was initially distributed to student platforms and forums as this directly hit the target audience for this research. Once most students completed the questionnaire on these platforms, the input data was assessed, and the number of participants was low. The questionnaire was then shared on social media platforms, where there was a greater outreach. It was understandable that the questionnaire would be answered by people who did not meet the research requirements. However, the data pre-processing could mould the dataset with the necessary data. From this, and before any data pre-processing took place, 536 participants completed the questionnaire. The questionnaire was available for 20 days and reached a remarkably diverse crowd—submissions from South and West Africa to the USA, North America, and the UK. As the research is focused on the UK, I could not take these submissions further.

The completed questionnaire was exported as a CSV format and imported into Excel. A preliminary review of the data in Excel was done to determine its quality and how much of it could be used. This is assessed to see if the data serves the

research's objective and is divided into four parts timeliness, completeness, consistency, and accuracy *(A. Haug, F. Zachariassen, and D. Van Liempd, 2011)*.

Once the data was gathered and shared with professionals and experts in the mental health field. A psychologist and a mental health nurse, both of whom do not want their names used. They examined the data to see if there was enough of it to identify mental health indicators. They both concurred that the questionnaire could help distinguish between potential indicators of depression. After consulting with both professionals, they decided to check the data that complied with the study's requirements, evaluate each row, and report the results; by doing so, the expertise would define the target variable.

One target variable and twenty-two predictor variables comprise the dataset acquired from the survey. The justification above was used to derive the target variable. Below are each variable's potential values, variable types, and variable descriptions.

3.3 Exploratory Data Analysis

EDA is a detailed analysis created to unearth a dataset's underlying structure. It is significant for an organisation or research because it reveals trends, patterns, and relationships that are not immediately obvious. A vast amount of data cannot be reliably analysed by just skimming over it; instead, it must be thoroughly examined methodically via an analytical lens. Developing a "feel" for this vital information will make it easier for you to spot errors, disprove presumptions, and comprehend the connections between essential elements. A suitable prediction model may eventually be chosen due to such discoveries.

The following are reasons why EDA is crucial for data analysis:

- Aids in finding dataset errors.
- Provides more insight into the dataset.
- Helps find abnormal events or outliers.
- Aids in understanding the variables in a data set and their relationships.

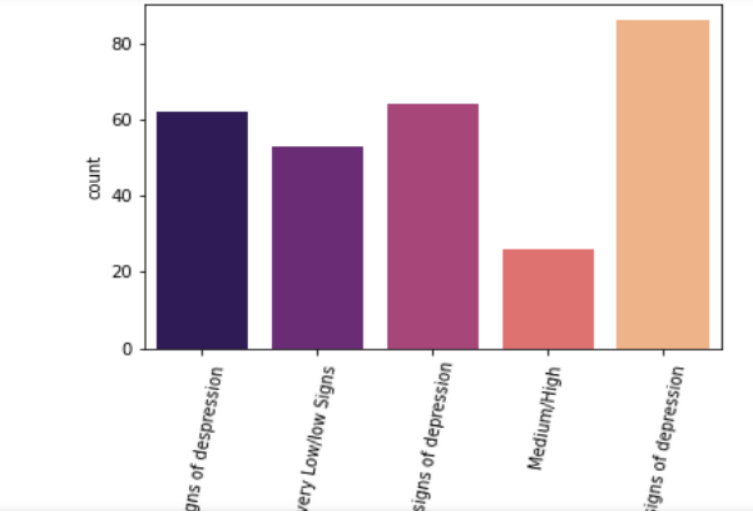
3.3.1 Univariate Analysis

The most fundamental method of statistical data analysis is known as univariate analysis. To summarise or describe a variable, a "univariate analysis" is defined as an analysis that uses only one("uni") variable ("variate") (Babbie, 2007; Trochim, 2006).

When doing univariate analysis, this was done by looking at categorical and then numerical variables so that we could see the data clearly in a visual form. As we know, all these participants are currently in education and studying. Looking at the categorical variables to start with, which were in the form of bar charts, it was interesting to see a pattern which includes the majority of the participants working full-time, and the majority had high signs of depression. This pattern was instantly noticed and could be vital to this research. Further exploration will be conducted to see if there is a correlation. The data also showed that most people stated that they

have no health issues or mental health diagnoses, even though signs of depression were at a high level overall. From studies on depression, we learn that people live or suffer from depression and do not know this due to stigmas and lack of education around mental health. (Kovac, 2007). Appendix 4. Displays distribution Histogram plots from the dataset which we get an overview of how the collected data is diverse.

Figure 3. univariate analysis showing visualisation of outcome variable

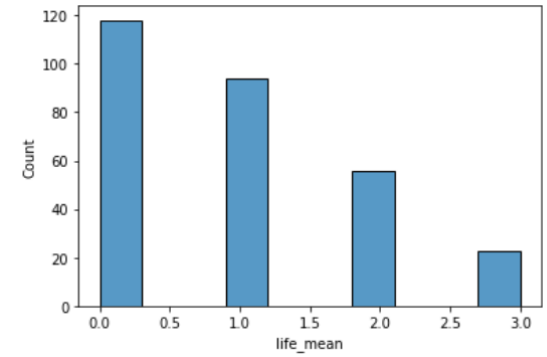


Still focusing on categorical variables, a key trend appeared in high social media usage and little to no time spent on fitness, health, and well-being. Studies show a regular correlation linked to much time spent on social media can lead to depression as we are consuming much information that may not be real and a persona that one carries. It is said that health and fitness can mean better mental well-being, even a 30-minute walk daily. (Sharma et al. 2006).

Analysis was also conducted on the numerical variables to see the average results of the questions asked. A primary factor of depression in this study is seeing life with no meaning and being scared for no reason. Most participants said, ‘this does not apply to them’(120) or ‘some of the time’ (80). It would be interesting if the result of the research took a turn due to the importance of these factors.

Figure 4. univariate analysis showing visualisation of a numerical variable (life_mean)

```
# Distribution of Numerical Varriable

df1 = df.select_dtypes(include=['int64'])
for i, col in enumerate(df1.columns):
    plt.figure(i)
    sns.histplot(x=col, data=df1,bins=10)
    plt.show()
```



3.3.2 Bivariate analysis

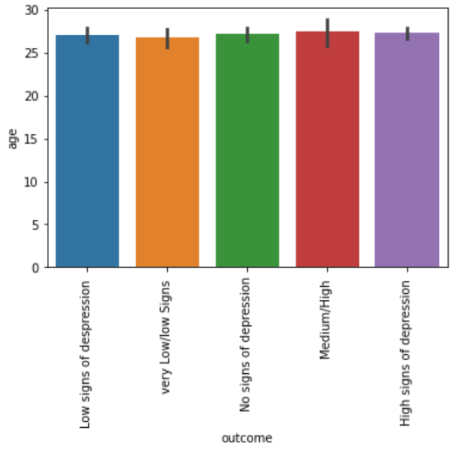
Two variables are compared to one another in a statistical method known as bivariate analysis. The dependent and independent variables will be the same. The letters X and Y stand for the variables. The differences between the two variables are evaluated to determine the magnitude of the change. (Sarangam, 2021).

The target variable ‘outcome was use as the X variable, whilst only changing the Y. As the research target group is age 18-35, trying to examine the correlation between age and the outcome was not easy as it showed to be balanced throughout. As mentioned previously in univariate section, the analysis went further to investigate the outcome variable against life without meaning and feeling scared. It distinctively shows a high correlation that many had have high signs of depression where they pick always for these two variables. this is key for feature importance and machine learning. From basic understanding of depression, it is known for believing life is meaningless, scared, and helpless are symptoms of depression. (NHS, 2019).

Figure 5. Bivariate Analysis; age and outcome

Bivariate Analysis

```
In [26]: sns.barplot(data=df, x="outcome", y="age");
         plt.yticks(size=10)
         plt.xticks(rotation=90, size=10)

Out[26]: (array([0, 1, 2, 3, 4]),
          [Text(0, 0, 'Low signs of despression '),
           Text(1, 0, 'very Low/low Signs '),
           Text(2, 0, 'No signs of depression'),
           Text(3, 0, 'Medium/High '),
           Text(4, 0, 'High signs of depression ')])
```



3.3.3 Multivariate Analysis

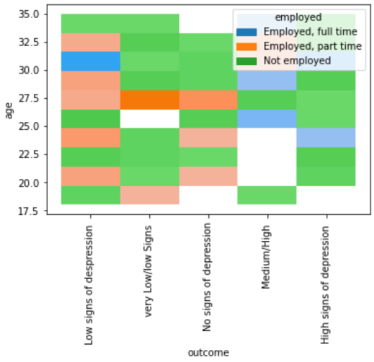
Numerous variables are considered in multivariate analysis. This makes it both a challenging and necessary tool. The fact that such a model considers as many variables as feasible is its greatest strength. This dramatically reduces bias and produces the most accurate result. (Vighnesh, 2021).

Key factors were taken to complete a multivariate analysis from the previous analysis. Here we focused on age, employment, and outcome. The analysis was trying to find a correlation between the variables. It was hard to tell if new patterns were shown that had not been mentioned in the previous analysis description. As we saw a strong correlation between these variables, Principal Component Analysis (PCA) did not need to be conducted.

Figure 6. Multivariate Analysis: age, outcome and employed

Multivariate Analysis

```
In [27]: sns.histplot(
             df,x="outcome", y="age", hue="employed", legend=True,bins=10
         );
         plt.yticks(size=10)
         plt.xticks(rotation=90, size=10)

Out[27]: ([0, 1, 2, 3, 4],
          [Text(0, 0, ''),
           Text(0, 0, ''),
           Text(0, 0, ''),
           Text(0, 0, ''),
           Text(0, 0, '')])
```



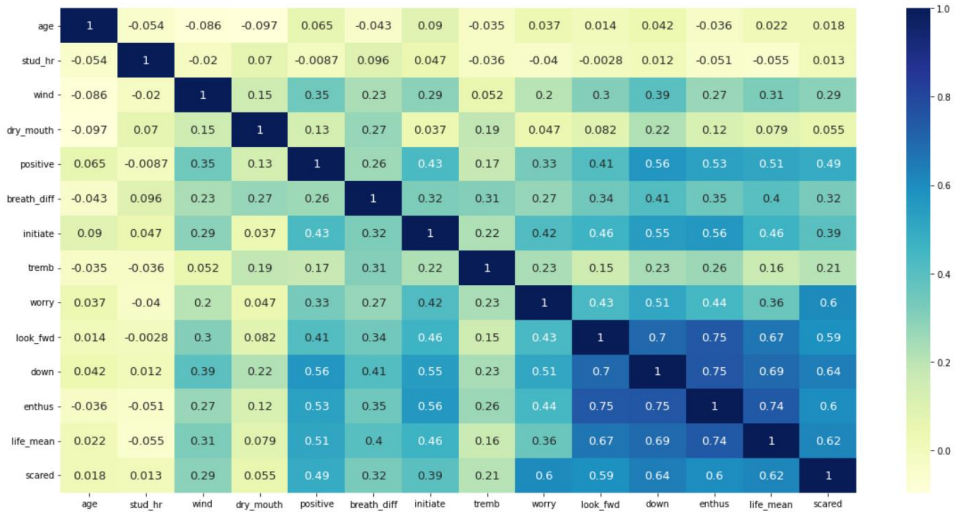
3.3.4 Correlation Analysis

Correlation analysis is a statistical practice used to assess the link between two quantitative variables. A high correlation indicates a significant association between two or more variables, whereas a low correlation indicates a minimal relationship between the variables. The coefficient of correlation is a metric that quantifies the relationship between two variables and the degree to which changes in one variable affect changes in the other. A visualisation of a correlation matrix that shows the relationship between various variables is called a correlation heatmap. Correlation can have any value between -1 and 1. It is not always the case that a correlation between two random variables or bivariate data indicates a causal connection. (Kumar, 2022). A correlation matrix and heatmap was completed on the dataset, which displayed the relationship between these six variables.

1. 'Do you find yourself worrying about situations in which you might panic and make a fool of yourself?'
2. 'Do you find yourself feeling that you have nothing to look forward to?'
3. 'Do you find yourself feeling down-hearted?'
4. 'Do you find yourself unable to become enthusiastic about anything?'
5. 'Do you find yourself feeling that life can be meaningless?'
6. 'Do you find yourself feeling scared without any good reason?'

As we have learnt this new information, it was understood that these features will have an impact on feature selection when we build the prediction model.

Figure 7. Correlation Heatmap

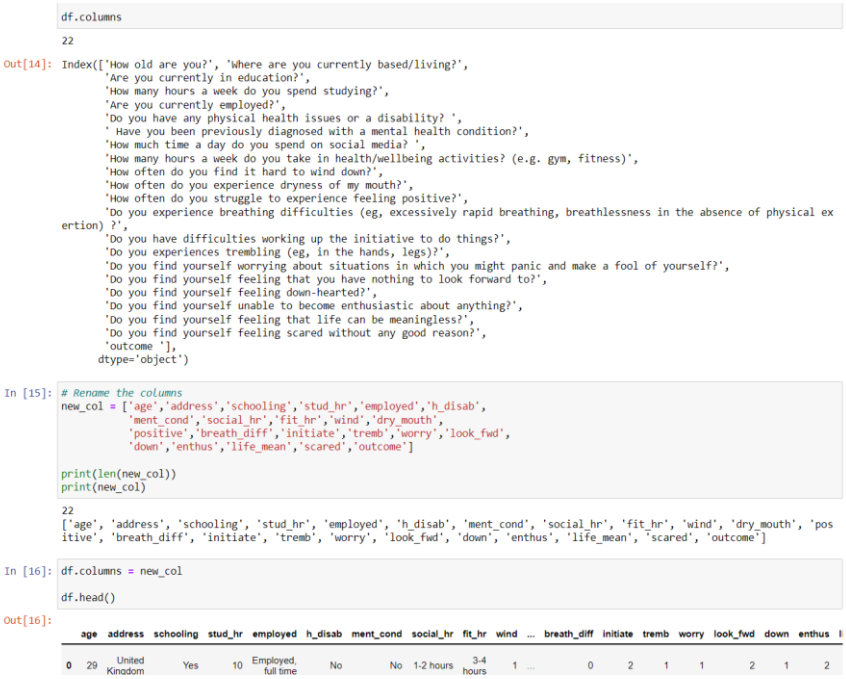


3.4 Data Pre-processing

Real-world data frequently has incomplete, inconsistent, inaccurate, and missing properties or values. As a result, data pre-processing is used to clean, prepare, and organise raw data so machine learning models can utilise it. (Sashikanta et al, 2022). Using Excel, participants that stated they were not in education, in the UK or over 35, were instantly removed from the dataset. Following this step, the dataset was imported into Jupyter Notebook for further processing. Then was first checked for null values, which null values returned. It was good to know that participants completed the questionnaire fully, as all questions were marked as mandatory, so we can get a complete picture of what is trying to be achieved. With this noted, the data cleaning completed in Excel meant that the data we could not use contained missing values as these were due to uncompleted questionnaires. With the type of data collected, there would be no capacity to fill null and empty fields with average values.

The column names were changed to make working with the data more manageable. The lengthy questionnaire questions were the names of the original columns. Column names are shortened to the feature's keyword, entered into a Python dictionary, and then used and replaced in the data frame. With the target variable in the dataset known as "outcome," this procedure was repeated. The two experts who contributed to the dataset had different spellings, and the field's incorrect values have all been renamed and replaced.

Figure 8. Column variable change in Jupyter Notebook



```
df.columns
```

```
22
```

```
Out[14]: Index(['How old are you?', 'Where are you currently based/living?',
       'Are you currently in education?',
       'How many hours a week do you spend studying?',
       'Are you currently employed?',
       'Do you have any physical health issues or a disability? ',
       ' Have you been previously diagnosed with a mental health condition?',
       'How much time a day do you spend on social media? ',
       'How many hours a week do you take in health/wellbeing activities? (e.g. gym, fitness)',
       'How often do you find it hard to wind down?',
       'How often do you experience dryness of my mouth?',
       'How often do you struggle to experience feeling positive?',
       'Do you experience breathing difficulties (eg, excessively rapid breathing, breathlessness in the absence of physical ex
ertion) ?',
       'Do you have difficulties working up the initiative to do things?',
       'Do you experiences trembling (eg, in the hands, legs)?',
       'Do you find yourself worrying about situations in which you might panic and make a fool of yourself?',
       'Do you find yourself feeling that you have nothing to look forward to?',
       'Do you find yourself feeling down-hearted?',
       'Do you find yourself unable to become enthusiastic about anything?',
       'Do you find yourself feeling that life can be meaningless?',
       'Do you find yourself feeling scared without any good reason?',
       'outcome '],
      dtype='object')
```

```
In [15]: # Rename the columns
new_col = ['age','address','schooling','stud_hr','employed','h_disab',
           'ment_cond','social_hr','fit_hr','wind','dry_mouth',
           'positive','breath_diff','initiate','tremb','worry','look_fwd',
           'down','enthus','life_mean','scared','outcome']

print(len(new_col))
print(new_col)
```

```
22
['age', 'address', 'schooling', 'stud_hr', 'employed', 'h_disab', 'ment_cond', 'social_hr', 'fit_hr', 'wind', 'dry_mouth', 'pos
itive', 'breath_diff', 'initiate', 'tremb', 'worry', 'look_fwd', 'down', 'enthus', 'life_mean', 'scared', 'outcome']
```

```
In [16]: df.columns = new_col

df.head()
```

Out[16]:

	age	address	schooling	stud_hr	employed	h_disab	ment_cond	social_hr	fit_hr	wind	...	breath_diff	initiate	tremb	worry	look_fwd	down	enthus	l
0	29	United Kingdom	Yes	10	Employed, full time	No	No	1-2 hours	3-4 hours	1	...	0	2	1	1	2	1	2	

It was crucial to examine the value count of all columns and responses as part of the pre-processing to determine which columns were to be dropped. Displaying the *'th'* count was the best practice for this. This showed that for education and location, just one form of response, 'United Kingdom' and 'yes' to indicate they are in education, was recorded, which is one of the critical criteria used in the research. Knowing this allows for dropping these two columns because, going forward, they could not be helpful or relevant, and know this has met part of the research requirements.

Figure 9. Analysing the 'th' value count

```
# To display th Value counts

for i in df.columns:
    display(df.groupby(i).count().T.head(1).reset_index().drop('index',axis=1))
```

age	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
0	4	2	4	12	14	12	14	23	15	26	65	45	15	11	14	9	4	2

address	United Kingdom
0	291

schooling	Yes
0	291

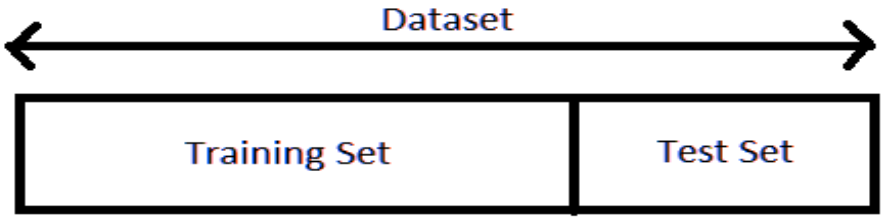
stud_hr	1	2	3	4	5	6	7	8	9	10	...	34	35	37	40	42	43	44	45	50	60
0	4	9	5	11	14	8	5	6	4	31	...	1	6	1	12	1	1	1	7	4	3

3.5 Feature scaling

Feature scaling was performed to normalise the data using label encoding. This was to transform the non-numerical labels as there were more than two answers

available for questions one-hot encoding will not be correct for this process and will cause issues when applying the data to a model. Label encoding allows different integers to represent data. From previous research which has been reviewed, it was noticed that the train and test splits remained between 0.2-0.3. Due to the dataset size, a split of 0.25 was decided best.

Figure 10. Dataset split



Stratified random sampling was employed to guarantee that the results were equally distributed among the sets. The cross-validation object 'StratifiedKFold is a KFold variant that produces stratified folds. The folds are created by keeping track of the sample percentages for each class.' (Sashikanta et al., 2022). KFold: Split the dataset into k consecutive folds. In the perspective of this research, we have the splits at five kFolds. When it is necessary to balance the percentage of each class in training and testing, StratifiedKFold is utilised and built with a pipeline that uses the final estimator to construct the fit and transform method. The pipelines vary but typically include the following steps: feature reduction, model training, classification, and performance evaluation.

Figure 11. Import of Starified KFold

```
# Starified k cross validation
Kfold = StratifiedKFold(n_splits=5)
```

Cross-validation is usually used to generalise the training process. There are several types of cross-validation, including k-fold, leave-one-out, and holdout. For k-fold Cross-validation, the training data are divided into k equal-sized groups. Then, each one of the k groups is treated as testing data and reiterated for k-iterations. The latter two techniques can be considered as variants of k-fold Cross-validation.

Figure 12. Cross validation usage

```
# Cross Validation.

def cv_fit_models():
    train_acc_results = []
    cv_scores = {classifier_name: [] for classifier_name in classifiers_names}
    for classifier_name, pipeline in zip(classifiers_names, pipelines):
        cv_score = cross_validate(pipeline,
                                  X_train,
                                  y_train,
                                  scoring=scoring,
                                  cv=Kfold,
                                  return_train_score=True,
                                  return_estimator=True)
```

3.6 Feature selection

The selection of features for a machine learning model should only include essential ones. The performance of the model may be harmed by choosing irrelevant features. Feature selection aids in eliminating redundant and pointless features that do not improve the model's performance. After pre-processing and data analysis was done. Since all entries from participants on the dataset had to be from the UK and in education, location and education was removed from the dataset. Of importance we have now learnt and seen that the physical elements employment, social media usage, health and wellbeing hours and the mental feeling elements of scared with no reason, life is meaningless and worrying and important features which makes a difference moving forward with building the model.

3.7 Models

Machine learning techniques for depression detection

In this research, eight different machine learning classifiers were utilised to predict the presence of depression, namely: Logistic Regression, Decision Tree Classifier, Support Vector Machine (SVM), Random Forest Classifier, K-Nearest Neighbour (KNN), Adaptive Boosting (AdaBoost), Gradient Boosting Classifier (GB), and Gaussian Naïve Bayes (NB). These eight have been chosen due to a mix of published research showing what has worked well and what has been missed within research. The details of these classifiers are described below.

Figure 13. import and usage of classifier models

```
classifiers_names = ['Logistic Regression',
                     'Decision Tree Classifier',
                     'Support Vector Machine',
                     'Random Forest Classifier',
                     'AdaBoost Classifier',
                     'Gradient Boosting Classifier',
                     'K Neighbors Classifier',
                     'Gaussian Naive Bayes']
```

The details of these classifiers are described below:

3.7.1 KNN

A popular distance-based approach for handling classification and regression issues is KNN, which is non-parametric. A new instance is classified by contrasting it with the instances in the training set that are the most similar, as KNN is also an instance-based learning algorithm. Distance measurements can estimate how similar the examples are to one another. The K value, or the number of nearest neighbours, is the main deciding factor in this classifier. If K is 1, the new instance is assigned to the class of its closest neighbour (Ali et al., 2019).

3.7.2 Gradient Boosting

The Gradient Boosting (GB) classifier successively builds new models out of a sequence of weak models. Every new model makes an effort to reduce the loss function. GB calculates the loss function using the gradient descent technique. Boosting should be halted promptly using stopping criteria to prevent overfitting issues. The stopping criteria can be a maximum number of models developed or a limit on the predicted accuracy (Rahman et al., 2020).

3.7.3 Decision Tree

Because they are simple to understand and have a stable structure, decision trees used in machine learning are ideal for prediction issues. Decision trees make choices at several levels using tree data structures. It covers regression (a volatile objective for an infinite set of values) as well as classification (tree models with a volatile target for a specific set of values). (Li and Zhang, 2010)

3.7.4 Random Forest

From a randomly chosen portion of the training dataset, the Random forest classifier generates several decision trees. The final class of test objects is then determined by averaging the votes from various decision trees presented a random forest classification with fewer trees. (Rodriguez-Galiano et al., 2012; Paul et al, 2018)

3.7.5 Support Vector Machine

A machine learning method known as a Support Vector Machine is primarily used in classification but can also be utilised for regression. Due to its excellent classification abilities and presentation quality, this classifier (Hamad et al. 2014) has recently been used in many applications. It divides the data linearly into two distinct classes (also known as hyperplanes), with the maximum distance between the two classes.

3.7.6 Naïve Bayes

This classifier relies on the Bayes theorem to supervised machine learning algorithms and operates under the assumption that features are analytically independent. (Martinez-Arroyo and Sucar, 2016) The naive assumption that the input factors are independent underlies the theorem. (Cheng and Griner, 1999)

3.7.7 Logistic Regression

Predictive analytics and categorization are frequently employed. Based on a given dataset of independent variables, logistic regression calculates the likelihood that an event will occur, such as voting or not voting. Because the result is a probability, the dependent variable's range is limited to 0 and 1.

3.7.8 AdaBoost

An AdaBoost classifier is a meta-estimator that first fits a classifier to the original dataset and then fits additional copies of the classifier to the same dataset with the weights of instances that were incorrectly classified being changed so that subsequent classifiers concentrate more on challenging cases.

3.8 Method

A range of tools was used within this research and achieved the aims and objectives set at the start of this study. From Excel to Python Notebook, to pre-process data to the prediction model. Python was also used to create a basic app to make this prediction model a user-friendly application, using frameworks such as Flask for backend development and JavaScript for the front end https://depress-status.herokuapp.com/ (for the user-friendly application.

4 Results

The test and train accuracy was run on all the eight models —AdaBoost, Decision Tree, Random Forest Tree, Gaussian Nave Bayes, Support Vector Machine, Gradient Boosting Classifier, Logistic Regression, and K-Nearest Neighbour to get the results to see what would be best to proceed with.

Figure 14. Accuracy Results (Train, Test, Precision and Recall)

	Model	train_accuracy	test_accuracy	test_precision	test_recall
0	Logistic Regression	87.157964	75.342466	71.995749	75.342466
1	Decision Tree Classifier	67.086371	64.383562	48.184309	64.383562
2	Support Vector Machine	93.349097	76.712329	73.306648	76.712329
3	Random Forest Classifier	90.021675	79.452055	79.570658	79.452055
4	AdaBoost Classifier	69.142857	72.602740	73.707347	72.602740
5	Gradient Boosting Classifier	79.359606	67.123288	62.945791	67.123288
6	K Neighbors Classifier	74.424959	68.493151	70.225969	68.493151
7	Gaussian Naive Bayes	82.568801	78.082192	78.745285	78.082192

Random Forest Tree showed the best with 79% accuracy.

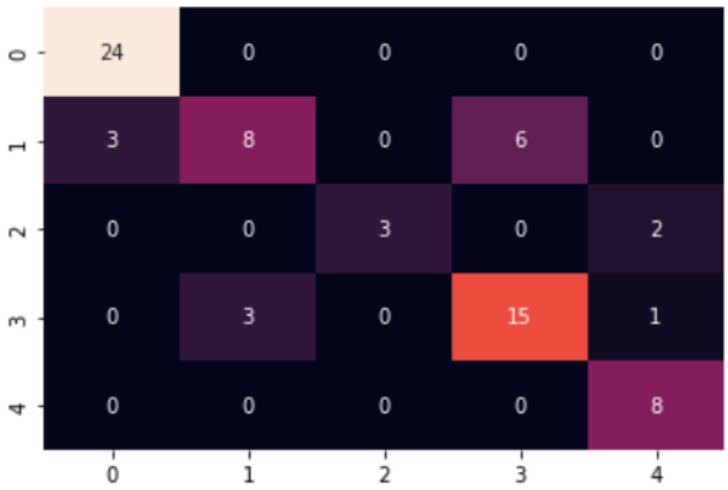
Figure 15. Classification report

```
Accuracy_score is: 0.7945205479452054
Classification report
                          precision    recall  f1-score   support

High signs of depression       0.89      1.00      0.94        24
Low signs of despression       0.73      0.47      0.57        17
             Medium/High       1.00      0.60      0.75         5
   No signs of depression      0.71      0.79      0.75        19
      very Low/low Signs       0.73      1.00      0.84         8

                accuracy                           0.79        73
               macro avg       0.81      0.77      0.77        73
            weighted avg       0.80      0.79      0.78        73
```

The confusion matrix result from applying the Random Forest classifier to the Depression classification. The rows and columns in which the digits 1, 2, 3, and 4

appear to stand for: Depression Symptoms: High Symptoms, Low Symptoms, Medium/High, No Symptoms, and Very Low/Low Signs

Figure 16. Confusion Matrix



4.1 Discussion

Proceeding with only the implementation of Random Forest Classifier. The results indicate that Test accuracy is 90%, Train accuracy is 79%, and Precision and Recall are 79%. The precision scores are within a reasonable range, which gives confidence in the chosen model.
We can answer the research topic with confidence and say yes! Machine learning can predict signs of depression.

From the results displayed above, it is clear that Random Forest proved best from the test to train to the precision, the accuracy scores showed to perform. Support Vector Machine was robust with the test accuracy, but train accuracy dropped significantly, which led to believe overfitting was the case.

According to *Pereira et al. (2003),* overfitting can be brought on by small sample sizes, high-dimensional features, complex models with too many parameters, and poor generalisation to independent datasets. Overfitting can produce a performance on training data but an inferior performance on testing data. This strategy delivers additional data to the learning method's training stage, which is linked to high variance, which could impair generalisation performance and result in overfitting for cross-validation (Elisseeff and Pontil, 2003). The confusion matrix was a surprise as the numbers were low except for true negative, which showed as 15. It would be good to see what determined this.

Referring back to *Priya et al.(2019)* study, they implemented all models and had more classifications whilst using a dataset that used all aspects of the DASS-21 survey mentioned. They demonstrated results from five classifiers with 62-80%

accuracy. It was insightful to know that both studies have similar variables, which proved significant. As this research focuses more on depression than *Priya et al.*, we have similar accuracy scores when looking at their Depression classification.

Figure 17. Priya et al. classification results

Classifier	Mental illness	Accuracy
Decision	Anxiety	0.733
Tree	Depression	0.778
	Stress	0.628
Random	Anxiety	0.714
Forest	Depression	0.798
	Stress	0.723
Naive	Anxiety	0.733
Bayes	Depression	0.855
	Stress	0.742
Support	Anxiety	0.678
Vector	Depression	0.803
Machine	Stress	0.667
K Nearest	Anxiety	0.698
Neighbour	Depression	0.721
	Stress	0.714

5 Conclusion

Since numerous machine learning approaches are accessible, it is crucial to compare them all and choose the one that best fits the target domain. Today, numerous specialised programmes in the medical profession can diagnose mental illness and diseases with extreme accuracy in advance, allowing for effective and fast treatment. In this proposed study, we have examined eight machine learning approaches to categorise the dataset on various aspects of mental health. Data were obtained using a standard questionnaire evaluating the common symptoms of depression inspired by the DASS-21 survey.

The classifier that was selected has accuracy levels above 79%. The research only employed a small amount of data; in the future, a more extensive data set might be used, and the research might then be applied to that more extensive data set for more accuracy. Although Support Vector Machine was judged to be a good model with the highest accuracy, there were overfitting indications. The f1 score, utilised in imbalanced partitioning, was used to pick the optimal model because this problem resulted in unbalanced classes. The variables "scared without reason" and "Life

meaningless" were identified as being crucial, and as a result, these variables were thought to be the most crucial in identifying depression.

Looking into future work, a larger dataset will be something that will be focused on. Having more time to collect the data will resolve this. It would be great if the research could grow in terms of location scope and see if different results are produced. Collaborating with a machine learning professional will be great to develop the models and code more tightly and see if this could result in a higher accuracy score. Developing the application artefact so it can be used with educational establishments.

5.1 Summary

Comparing the findings and contribution of this study to other published works are required to assess the quality of the study. Most earlier studies have been conducted to forecast depression among people with a specific age group, employment, or medical condition. A select few have identified the psychological and sociodemographic elements contributing to depression. However, this study aimed to identify depression among UK-educated individuals of various ages, and all rounded variables with published authors have produced. The most crucial elements from an inspired Depression Anxiety Stress Survey that contribute to or precede depressive symptoms have also been discovered by this study with a developed predicting model with an accuracy of 79%.

5.2 Limitations

Limitations have been shown whilst conducting this research. The dataset size is limited to under three hundred rows. This has shown in the data analysis that there were few numerical data to analyse. We noticed overfitting when testing and training the model when we applied SVM. When researching and looking into published work, comments were noticed on how it has been a struggle to collect data and focus groups were essential. This may be a challenge. Time was a constraint in this research. As the data collection was done when most UK education establishments were on summer break, most would not be active online on student platforms where the questionnaire was shared.

6 References and Bibliography

[1] Sharma Amita, Verbeke Willem J. M. I. Improving Diagnosis of Depression With XGBOOST Machine Learning Model and a Large Biomarkers Dutch Dataset (n = 11,081), Frontiers in Big Data 3, 2020, https://www.frontiersin.org/article/10.3389/fdata.2020.00015, DOI=10.3389/fdata.2020.00015 ISSN=2624-909X

[2] Mohd Shafiee, N. and Mutalib, S., 2020. Prediction of Mental Health Problems among Higher Education Student Using Machine Learning. International Journal of Education and Management Engineering, 10(6), pp.1-9.

[3] Vaishnavi, K., Nikhitha Kamath, U., Ashwath Rao, B. and Subba Reddy, N., 2022. Predicting Mental Health Illness using Machine Learning Algorithms. Journal of Physics: Conference Series, 2161(1), p.012021.

[4] Cho, G., Yim, J., Choi, Y., Ko, J. and Lee, S., 2019. Review of Machine Learning Algorithms for Diagnosing Mental Illness. Psychiatry Investigation, 16(4), pp.262-269.

[5] J. D. Immanuel, H. M. Ragavan, P. G. Rani, K. Niveditaa and G. Manikandan, "AI to Detect Social Media users Depression Polarity Score," *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 2022, pp. 415-418, doi: 10.1109/ICSCDS53736.2022.9761007.

[6] M. Usman, S. Haris and A. C. M. Fong, "Prediction of Depression using Machine Learning Techniques: A Review of Existing Literature," 2020 IEEE 2nd International Workshop on System Biology and Biomedical Systems (SBBS), 2020, pp. 1-3, doi: 10.1109/SBBS50483.2020.9314940.

[7] McLafferty, M., Lapsley, C. R., Ennis, E., Armour, C., Murphy, S., Bunting, B. P., … O'Neill, S. M. (2017). Mental health, behavioural problems and treatment seeking among students commencing university in Northern Ireland. PLoS ONE, 12(12), 1– 14.

[8] Vidourek, R. A., & Burbage, M. (2019). Positive mental health and mental health stigma: A qualitative study assessing student attitudes. Mental Health and Prevention, 13(October 2018), 1–6.

[9] A. Sau and I. Bhakta, "Predicting anxiety and depression in elderly patients using machine learning technology", Healthcare Technology Letters, vol. 4, no. 6, pp. 238-243, 2017.

[10] J. Choi, J. Choi and H.T. Jung, "Applying machine-learning techniques to build self-reported depression prediction models", Computers Informatics Nursing Kluwer Health, vol. 36, no. 7, pp. 317-321, 2018.

[11] Uddin, S., Khan, A., Hossain, M. et al. Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19, 281 (2019). https://doi.org/10.1186/s12911-019-1004-8

[12] Sau, A., Bhakta, I. (2017)"Predicting anxiety and depression in elderly patients using machine learning technology. “Healthcare Technology Letters 4 (6): 238-43

[13] Priya, A., Garg, S. and Tigga, N., 2020. Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms. Procedia Computer Science, 167, pp.1258-1267.

[14] A. Haug, F. Zachariassen, and D. Van Liempd, “The costs of poor data quality,” J. Ind. Eng. Manag., vol. 4, no. 2, pp. 168–193, 2011.

[15] Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. Nature Medicine, 28(6), 1240–1248.

[16] Konda Vaishnavi, U Nikhitha Kamath, B Ashwath Rao and N V Subba Red, https://iopscience.iop.org/article/10.1088/1742-6596/2161/1/012021/pdf, accessed on August 2022

[17] Jetli Chung; Jason Teo (2022). Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges. Applied Computational Intelligence and Soft Computing 2022. https://www.hindawi.com/journals/acisc/2022/9970363/

[18] Priya, A., Garg, S., & Tigga, N. P. (2020). Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms. Procedia Computer Science, 167, 1258–1267. https://doi.org/10.1016/j.procs.2020.03.442

[19] Zeberga, K., Attique, M., Shah, B., Ali, F., Jembre, Y. Z., & Chung, T.-S. (2022). A Novel Text Mining Approach for Mental Health Prediction Using Bi-LSTM and BERT Model. Computational Intelligence and Neuroscience, 2022, 1–18. https://doi.org/10.1155/2022/7893775

[20] Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. Npj Digital Medicine, 5(1). https://doi.org/10.1038/s41746-022-00589-7

[21] Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. Psychological Medicine, 49(09), 1426–1448. https://doi.org/10.1017/s0033291719000151

[22] Tate, A. E., McCabe, R. C., Larsson, H., Lundström, S., Lichtenstein, P., & Kuja-Halkola, R. (2020). Predicting mental health problems in adolescence using machine learning techniques. PLOS ONE, 15(4), e0230389. https://doi.org/10.1371/journal.pone.0230389

[23] Khondoker, M., Dobson, R., Skirrow, C., Simmons, A., & Stahl, D. (2016). A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. Statistical Methods in Medical Research, 25(5), 1804–1823. https://doi.org/10.1177/0962280213502437

[24] Md. Sabab Zulfiker, Nasrin Kabir, Al Amin Biswas, Tahmina Nazneen, Mohammad Shorif Uddin, An in-depth analysis of machine learning approaches to

predict depression, Current Research in Behavioral Sciences, Volume 2, 2021, 100044, ISSN 2666-5182, https://doi.org/10.1016/j.crbeha.2021.100044. (https://www.sciencedirect.com/science/article/pii/S2666518221000310)

[25] A.A. Choudhury, M.R.H. Khan, N.Z. Nahim, S.R. Tulon, S. Islam, A Chakrabarty. Predicting depression in Bangladeshi undergraduates using machine learning. Proceedings of the 2019 IEEE Region 10 Symposium (TENSYMP), IEEE (2019), pp. 789-794

[26] M.F. Zarandi, S. Soltanzadeh, A. Mohammadi, O. Castillo. Designing a general type-2 fuzzy expert system for diagnosis of depression. Appl. Soft Comput., 80 (2019), pp. 329-341

[27] C.M. Hatton, L.W. Paton, D. McMillan, J. Cussens, S. Gilbody, P.A. Tiffin. Predicting persistent depressive symptoms in older adults: a machine learning approach to personalised mental healthcare. J. Affect. Disord., 246 (2019), pp. 857-860

[28] K.S. Na, S.E. Cho, Z.W. Geem, Y.K. Kim. Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm. Neurosci. Lett., 721 (2020), Article 134804

[29] http://www.emro.who.int/mnh/what-you-can-do/index.html#accordionpan4 , accessed July 2022

[30] Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: a machine learning approach. The Lancet Psychiatry, 3(3), 243–250. https://doi.org/10.1016/s2215-0366(15)00471-x

[31] Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C.-H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. Journal of Affective Disorders, 241, 519–532. https://doi.org/10.1016/j.jad.2018.08.073

[32] J. Cvetković. Breast cancer patients' depression prediction by machine learning approach Cancer Investigation., 35 (8) (2017), pp. 569-572

[33] Jain, S., Narayan, S. P., Dewang, R. K., Bhartiya, U., Meena, N., & Kumar, V. (2019). A Machine Learning based Depression Analysis and Suicidal Ideation Detection System using Questionnaires and Twitter. 2019 IEEE Students Conference on Engineering and Systems (SCES). https://doi.org/10.1109/sces46477.2019.8977211

[34] Sashikanta. P, Srikanta. P, Kumar SD. SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer, Frontiers in Nanotechnology 4,

2022, https://www.frontiersin.org/articles/10.3389/fnano.2022.972421 DOI=10.3389/fnano.2022.972421

[35] Bhaumik R, Jenkins LM, Gowins JR, et al. Multivariate pattern analysis strategies in detection of remitted major depressive disorder using resting state functional connectivity. Neuroimage Clinical. 2017;16(C):390.

[36] Hilbert K, Lueken U, Muehlhan M, Beesdobaum K. Separating generalized anxiety disorder from major depression using clinical, hormonal, and structural MRI data: a multimodal machine learning study. Brain Behav. 2017;7(3):e00633.

[37] https://www.datadecisionsgroup.com/blog/bid/176827/a-closer-look-at-exploratory-data-analysis-what-and-why 21 August 2022

[38] https://businessanalyst.techcanvass.com/objective-of-exploratory-data-analysis/#:~:text=Importance%20of%20using%20EDA%20for,and%20the%20relationship%20among%20them. Accessed on 22 August 2022

[39] jenny Kovacs https://www.webmd.com/depression/features/could-you-be-depressed-not-know Accessed on 22 August 2022

[40] Sharma A, Madaan V, Petty FD. Exercise for mental health. Prim Care Companion J Clin Psychiatry. 2006;8(2):106. doi: 10.4088/pcc.v08n0208a. PMID: 16862239; PMCID: PMC1470658.

[41] https://www.jigsawacademy.com/blogs/business-analytics/bivariate-analysis/#:~:text=Bivariate%20analysis%20is%20a%20kind,extent%20the%20change%20has%20occurred. Accessed on 25 August 2022

[42] https://www.nhs.uk/mental-health/conditions/clinical-depression/symptoms/ Accessed on 25 August 2022

[43] https://s4be.cochrane.org/blog/2021/09/09/multivariate-analysis-an-overview/#:~:text=Multivariate%20analysis%20is%20defined%20as,and%20their%20structure%20are%20important Accessed on 25 August 2022

[44] https://vitalflux.com/correlation-heatmap-with-seaborn-pandas/ Correlation Concepts, Matrix & Heatmap using Seaborn April 16, 2022 by Ajitesh Kumar

[45] N. Ali, D. Neagu, P. Trundle. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Appl. Sci., 1 (12) (2019), pp. 1-15

[46] S. Rahman, M. Irfan, M. Raza, K. Moyeezullah Ghori, S. Yaqoob, M. Awais. Performance analysis of boosting classifiers in recognizing activities of daily living. Int. J. Environ. Res. Public Health, 17 (3) (2020), p. 1082

[47] Li, L., Zhang, X. (2010) "Study of data mining algorithm based on decision tree." In2010 International Conference On Computer Design and Applications IEEE 1: V1-155

[48] Paul, A., Mukherjee, D. P., Das, P., Gangopadhyay, A., Chintha, A. R., Kundu, S. (2018)"Improved random forest for classification."IEEE. Transactions on Image Processing 27 (8): 4012-4024.

[49] Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J. P. (2012) “An assessment of the effectiveness of a random forest classifier for land-cover classification.”ISPRS Journal of Photogrammetry and Remote Sensing 67: 93-104.

[50 Martinez-Arroyo, M., Sucar, L. E. (2006) "Learning an optimal naive bayes classifier."In18th International Conference on Pattern

[51] Hamed, T., Dara, R., Kremer, S. C. (2014) "An accurate, fast embedded feature selection for SVMs." In2014 13th International Conference on Machine Learning and ApplicationsIEEE :135-140.

[52] Cheng, J., Greiner, R. (1999) "Comparing Bayesian network classifiers."InProceedings of the Fifteenth conference on Uncertainty in artificial intelligence Morgan Kaufmann Publishers Inc : 101-108.

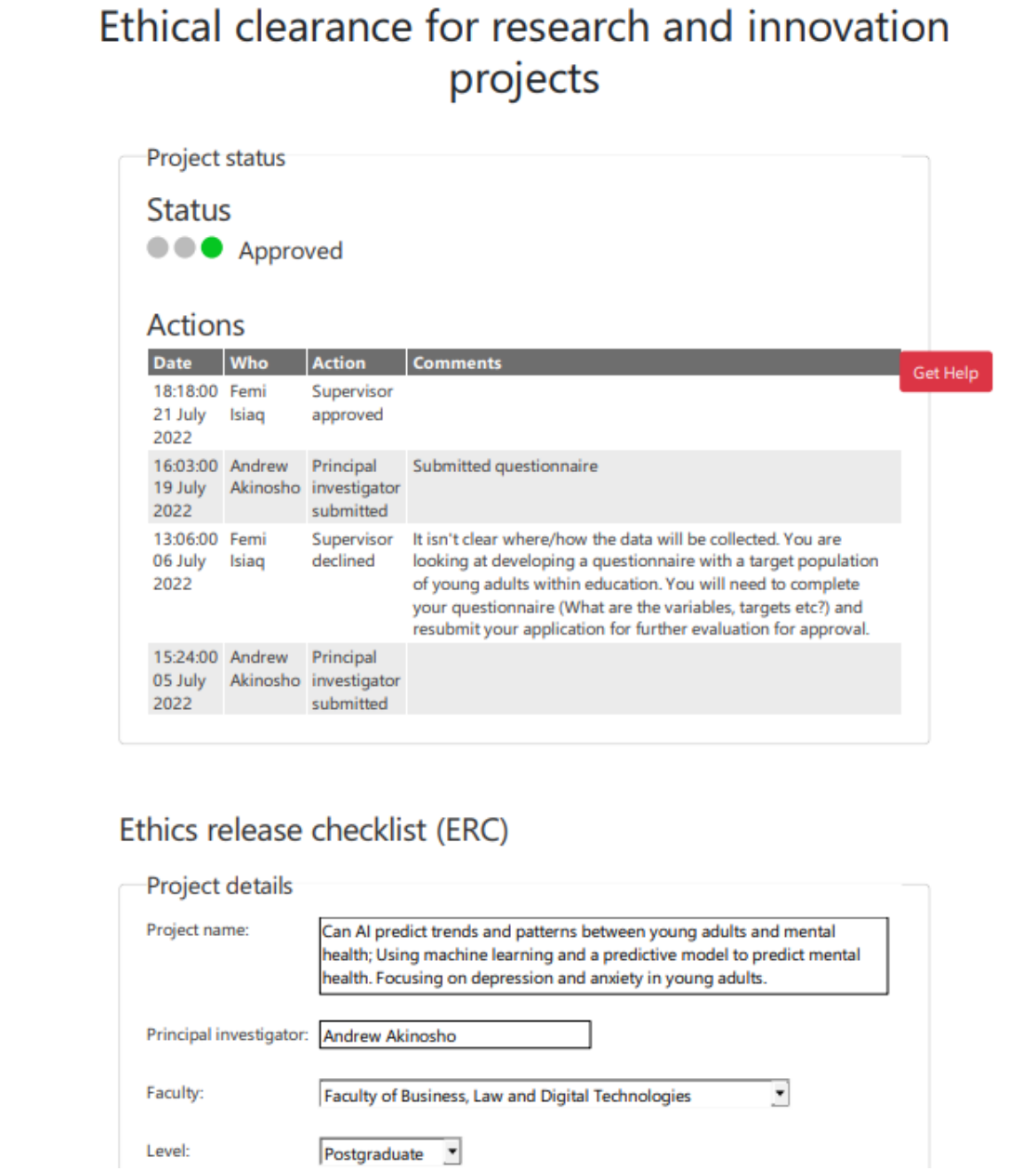
[53] https://www.ibm.com/uk-en/topics/logistic-regression#:~:text=Logistic%20regression%20estimates%20the%20probability,bounded%20between%200%20and%201. Accessed on 03 September 2022

[54] Elisseeff A, Pontil M. Leave-one-out error and stability of learning algorithms with applications. Adv Learn Theory Method Model Appl. 2003;190:111-130.

[55] Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage. 2009;45(1 Suppl):S199.

7 Appendices

7.1 Appendix 1. Ethics approval



Ethical clearance for research and innovation projects

Project status

Status

Approved

Actions

Date	Who	Action	Comments
18:18:00 21 July 2022	Femi Isiaq	Supervisor approved	
16:03:00 19 July 2022	Andrew Akinosho	Principal investigator submitted	Submitted questionnaire
13:06:00 06 July 2022	Femi Isiaq	Supervisor declined	It isn't clear where/how the data will be collected. You are looking at developing a questionnaire with a target population of young adults within education. You will need to complete your questionnaire (What are the variables, targets etc?) and resubmit your application for further evaluation for approval.
15:24:00 05 July 2022	Andrew Akinosho	Principal investigator submitted	

Get Help

Ethics release checklist (ERC)

Project details

Project name: Can AI predict trends and patterns between young adults and mental health; Using machine learning and a predictive model to predict mental health. Focusing on depression and anxiety in young adults.

Principal investigator: Andrew Akinosho

Faculty: Faculty of Business, Law and Digital Technologies

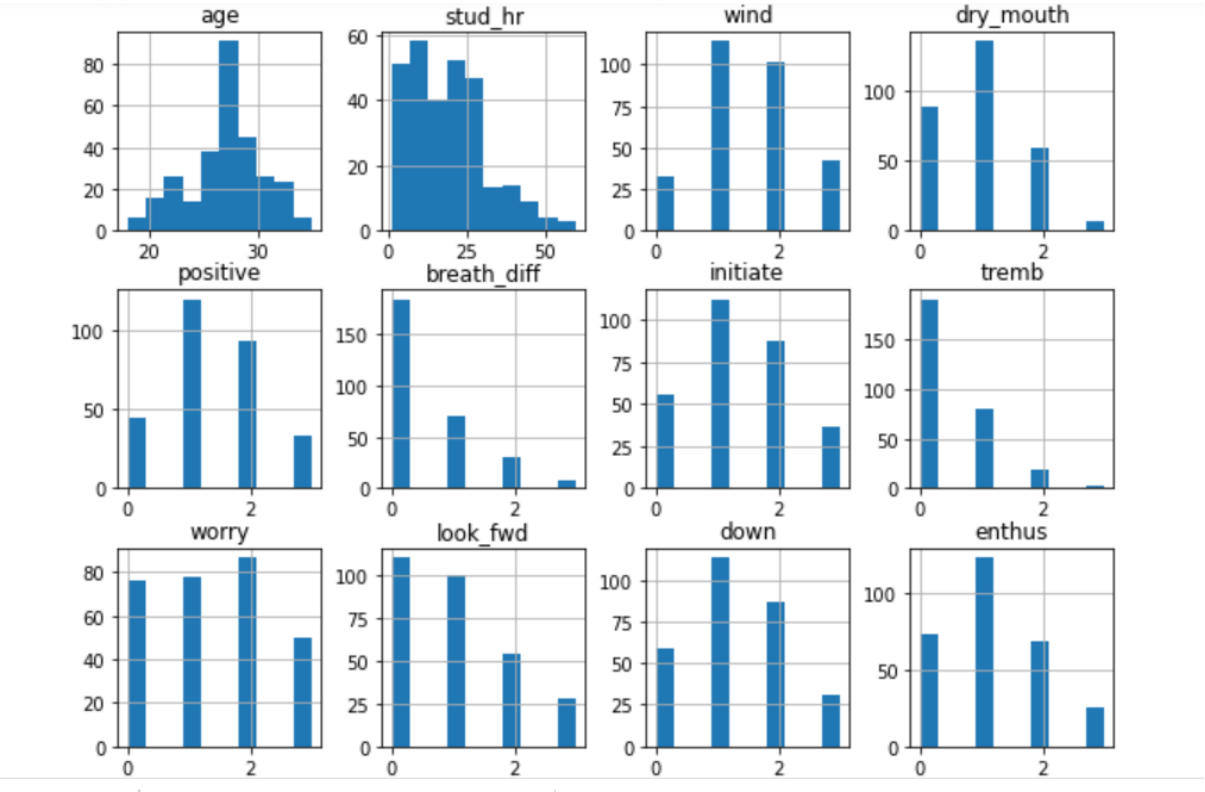
Level: Postgraduate

7.2 Appendix 2. Source code

file:///C:/Users/andre/Downloads/Depression%20predictor.html

7.3 Appendix 3. Research Zoho Questionnaire

7.4 Appendix 4. Distribution Histogram plots fo data columns.



7.5 Appendix 5. Link to application – Depression predictor.

https://depress-status.herokuapp.com/