UNIVERSITY OF ENERGY AND NATURAL RESOURCES, SUNYANI

SCHOOL OF SCIENCES

DEPARTMENT OF INFORMATION TECHNOLOGY AND DECISION SCIENCES

ANALYSING THE PREVALENCE AND TREATMENT GAPS OF MENTAL ILLNESSES: DATA-DRIVEN INSIGHTS AND PREDICTIVE MODELLING

BY

DERRICK OSEI KWAME

MICHELLE AMANYAME KYEI

PRISCILLA NKUNO

AMOAH GYAMFI JOSHUA

A PROJECT WORK SUBMITTED TO THE DEPARTMENT OF INFORMATION TECHNOLOGY AND DECISION SCIENCES, UNIVERSITY OF ENERGY AND NATURAL RESOURCES, IN THE PARTIAL FULFILMENT OF THE REQUIREMENT FOR AWARD OF A DEGREE IN INFORMATION TECHNOLOGY

OCTOBER, 2024

DECLARATION

We declare that this project report is our own and all other sources of information have been acknowledged and that we are responsible for any acts that may violate the research ethics policies of the University.

MICHELLE AMANYAME KYEI	…….	……..
(UEB3246322)	DATE	SIGNATURE
PRISCILLA NKUNO	…….	……..
(UEB3212120)	DATE	SIGNATURE
AMOAH GYAMFI JOSHUA	…….	……..
(UEB3214920)	DATE	SIGNATURE
DERRICK OSEI KWAME	…….	……..
(UEB3217420)	DATE	SIGNATURE
MR. EMMANUEL DOMFEH	…….	……..
(SUPERVISOR)	DATE	SIGNATURE
PROF. PETER APPIAHENE	…….	……..
(HEAD OF DEPARTMENT)	DATE	SIGNATURE

ABSTRACT

Mental health illnesses such as bipolar disorder, depression, and anxiety are extremely common in a wide range of populations and pose a significant global public health risk. Despite the prevalence of these conditions, there is still a huge treatment gap, particularly in low- and middle-income countries where more than 80% of individuals do not have access to mental health services. The current study uses data-driven approaches and predictive modelling to investigate the incidence of mental health diseases as well as the variables that contribute to treatment gaps. This study employs machine learning techniques such as Random Forest and Support Vector Machines (SVM) to discover critical healthcare-related, socioeconomic, and demographic characteristics that exacerbate the gap in mental health treatment. To ensure trustworthy analysis, the study examines massive datasets gathered from international health organisations using data integration and cleaning approaches. Predictive models are designed to assess the potential of treatment gaps in specific groups. These models provide insightful data to healthcare practitioners and policymakers. The findings indicate the locations and people that are most vulnerable to receiving inadequate mental health treatment, and they argue that focused interventions and resource allocation can assist to eliminate these gaps. This study also emphasises the technical and ethical concerns connected with predictive modelling in mental health, including data quality and access to care equity. The findings of this study make critical recommendations for closing the mental health care gap and improving mental health outcomes everywhere.

DEDICATION

This project work is dedicated to our creator, the creator of heaven and earth, the Almighty God, the basis of our existence, our helper, and our all in all, who have made us a success in all things. All glory, honour and adoration we give to Him forever and ever Amen.

Thanks to our family especially our mothers and fathers for being there for us throughout our studies, for their financial and moral support.

ACKNOWLEDGEMENT

We are grateful to the Almighty God for all of His mercies, as well as for His direction, safety, wisdom, and the gifts He has bestowed upon us. We have come this far only because of his grace. We sincerely thank Mr. Emmanuel Domfeh, our project supervisor, for his understanding and efforts in keeping us on track, as well as our parents for making it possible for us to attend this institution. Finally, we would like to convey our profound thanks to the department's director and all students. We dedicate this work to our parents, whose unwavering support guaranteed that we would use every ounce of energy required to finish what we had started.

Table of Contents

List of Figures

CHAPTER ONE: INTRODUCTION

1.0 Background of the Study

A person's ideas, feelings, and behaviours are greatly influenced by their mental health, which makes it a crucial aspect of total wellbeing. It also influences how individuals make decisions, interact with others, and handle stress. Mental health is one of the most underappreciated medical specialities in the world, despite its significance. The World Health Organisation (WHO) estimates that one in four individuals may have a mental disorder at some time in their lives. People of all ages and demographics are impacted by depression, anxiety, and bipolar illness, however many of them do not receive the required therapy (Karyotaki et al., 2020).

A serious public health concern is the "treatment gap," which is the discrepancy between the incidence of mental health problems and the accessibility of appropriate treatment. Numerous factors, including cultural beliefs, socioeconomic position, the stigma attached to mental illness, and the accessibility of healthcare services, all have an impact on this disparity. The treatment gap may surpass 90% in low- and middle-income nations, indicating that the majority of people with mental health conditions would not obtain the necessary care (Kessler et al., 2006).

Since untreated mental disorders have a terrible effect on both people and society as a whole, data-driven methods to detect and narrow treatment gaps are desperately needed. By utilising data analytics and predictive modelling, this study aims to develop models that can anticipate where gaps in mental health care are most likely to emerge and provide insights into the mechanisms underlying these gaps.

1.1 Problem Statement

The disparity in mental health care is a pervasive issue that affects millions of people worldwide. Despite the high prevalence of mental health disorders, a significant portion of individuals afflicted do not receive the appropriate care. This discrepancy can lead to increased disability, worse health outcomes, and increased burden on families and healthcare systems. A number of reasons, such as a lack of access to mental health specialists, socioeconomic constraints, widespread stigma, and an inadequate healthcare infrastructure, frequently contribute to the absence of effective treatment. Conventional methods of filling treatment gaps in mental health have mostly been reactive, concentrating on offering therapy after symptoms have gotten worse. However, because they don't address the underlying causes of treatment inequities or allow for proactive intervention, these approaches are frequently insufficient (Kumar, 2011). A more methodical and anticipatory strategy is required in order to detect and fix treatment gaps before they cause serious damage.

1.2 Research Questions

The following major research questions are the subject of this study:

1. What are the main causes of the disparity in mental health treatment across different demographic groups and geographical areas?
2. What is the potential use of predictive modelling in identifying groups at risk of inadequate mental health treatment?
3. How do the results affect healthcare policy and resource delivery?

1.3. Objectives

1.3.1 General Objectives of the Study

The main objective of the study is to present a comprehensive analysis of the treatment gaps for mental illnesses and their incidence. We will employ data-driven methods to find insights that can inform practice and policy.

1.3.2 Specific Objectives

1. To examine the frequency of common mental health conditions in various demographic groups and geographical areas, including bipolar disorder, anxiety, and depression.
2. To determine the main causes of the treatment gap in mental health care, such as socioeconomic, cultural, and medical issues.
3. To create and assess prediction models that can calculate the probability of treatment gaps in particular groups.
4. To offer suggestions to healthcare practitioners and politicians on how to close the gap in mental health treatment and enhance access to care.

1.4 Significance of the Study

There are several reasons why this study is important.

Firstly, by helping policymakers better allocate funds to regions with the greatest treatment gaps, the study's conclusions can enhance access to mental health treatments.

Secondly, Healthcare Planning which is using the findings of this study, healthcare professionals may create focused interventions that target the groups most at risk of not obtaining proper mental health treatment.

Thirdly, the study can aid in lowering stigma and raising awareness of mental health issues by exposing the variables that contribute to gaps in mental health care.

And lastly this study's methods and results will add to the body of information already available in the field of mental health research, especially when it comes to using data analytics and predictive modelling to public health concerns.

1.5 Scope and Delimitation

The study's main objective is to analyse the gaps in mental health care utilising data from national health surveys and international health organisations. In order to provide a global perspective on mental health treatment gaps, the research will look at data from both high- and low-income countries. It will also focus on common mental health disorders, such as depression, anxiety, and bipolar disorder, which are among the most common and burdensome conditions. To find insights and forecast treatment gaps, the study will also use data analysis, visualisation, and predictive modelling techniques.

However, the study is delimited by its dependence on the availability and quality of data, which may vary across regions and countries. Some areas may have limited or incomplete data on mental health. The study aims to predict treatment gaps but it does not delve deeply into intervention strategies or the implementation of solutions, which are beyond the scope of this research.

1.6 Organization of the Study

The study is organised into five chapters. The context, issue description, goals, importance, and scope of the study are all covered in Chapter One, Literature Review which is in the next chapter, reviews the body of research on treatment gaps, mental health prevalence, and the application of predictive modelling in healthcare, Chapter Three which is the methodology describes the study's modelling methodology, data gathering strategies, data analysis methodologies, and research

design. Chapter Four provides the outcomes and Discussion: In this chapter, the outcomes of the data analysis and modelling are presented, and then they are discussed in light of previous research. Chapter Five is the Conclusion and Recommendations: This last chapter offers suggestions for future research, policymakers, and healthcare practitioners in addition to summarising the main results and discussing their consequences.

CHAPTER TWO: LITERATURE REVIEW

2.0 Introduction

The goal of this literature review is to present a thorough analysis of the prevalence of mental health issues and treatment gaps, as well as to pinpoint pertinent patterns and the ways in which machine learning algorithms have been applied by others to address related issues. To find publications on mental disease prediction models, we conducted a thorough search of scientific databases. Using search phrases like "Mental Health," "Treatment Gaps," and "Machine Learning," we looked through PubMed, IEEE Explore, and Google Scholar databases.

2.1 Mental Health Disorders and Machine Learning

The most prevalent causes of disability in the globe include schizophrenia, bipolar disorder, depression, and anxiety. These illnesses not only lower people's quality of life, but they also cost society a lot of money in missed productivity and medical expenses (Kohn et al., 2004). The frequency of mental health problems varies widely around the world, based on variables including socioeconomic status, cultural beliefs, and healthcare accessibility (Kesseler et al., 2006).

2.2 Prevalence of Mental Health Disorders

Between 10 and 20 percent of the world's population suffers from mental health disorders at any given time, according to WHO and other international health organisations. Anxiety and depression are very prevalent, affecting hundreds of millions of people worldwide. These ailments are often more common in areas of severe poverty, conflict zones, and places with poor access to healthcare. Regional differences in mental health prevalence are significant. For instance, higher rates of identified mental health conditions are found in high-income nations, partly due to better access to diagnostic services and increased awareness of mental health concerns. However, stated prevalence rates are often lower in low- and middle-income countries, which may be more a result

of underdiagnoses and underreporting than a real decrease in the occurrence of mental health conditions (Uwakwe & Otakpor, 2014). There is a strong association between mental health and socioeconomic factors including income, education, and work position. Financial stress, limited access to treatment, and unfavourable living conditions make people from lower socioeconomic origins more likely to have mental health issues. Age and gender are also important demographic determinants; for instance, women are more likely to experience anxiety and depression, whereas males are more likely to acquire drug use disorders.

2.3 The Mental Health Treatment Gap

The percentage of individuals with mental diseases who do not receive adequate care is known as the mental health treatment gap. According to estimates, up to 80% of individuals with mental diseases in low- and middle-income countries do not receive the care they need, making this imbalance dangerously enormous on a global scale. One major barrier to receiving mental health treatments is inadequate healthcare infrastructure, particularly in low-income communities (Kohn et al., 2004). To address the needs of their inhabitants, many nations lack enough mental health professionals, such as social workers, psychologists, and psychiatrists. People may be deterred from seeking treatment by societal perceptions of mental health and the stigma attached to mental illness. Mental health issues are categorised as spiritual or moral defects rather than physical illnesses in certain cultures. For people without health insurance or in countries where healthcare is not publicly funded, mental health treatment including medication and therapy is unaffordable. The treatment gap is caused in part by inadequate financing for mental health care and the lack of national mental health policy. According to Qin and Hsieh (2020), public health policies in many nations place a higher priority on physical health than mental health. The treatment gap has a substantial negative impact on both individuals and communities. Chronic

impairment, a higher chance of physical health issues, and higher death rates, including suicide, can all arise from untreated mental health issues. Loss of productivity, increased medical expenses, and social pressure on families and communities all contribute to the substantial economic burden.

2.4 Machine Learning

Three issues are addressed by machine learning (ML), a branch of artificial intelligence (AI): clustering, regression, and classification. It simulates how people learn using data and algorithms, progressively increasing accuracy across a range of tasks. According to Iyortsuun et al. (2023), machine learning (ML) has been used in a number of psychological therapies and has great promise for diagnosing and treating mental health conditions and related health consequences. For these algorithms to identify patterns and carry out classification tasks, a lot of data is usually needed. One of the most popular machine learning techniques for forecasting mental illnesses is supervised learning.

2.4.1 Classification

Classification in machine learning is a supervised learning procedure that predicts the category or class of a new data point based on previously observed data. Based on the features it saw during the training phase, the model learns to allocate new observations to the pre-existing classes into which the data is categorised. Classification is useful for identifying spam emails and differentiating various animals in photos.

2.4.2 Supervised Learning

The process of learning a mapping between a set of input variables and an output variable, then using that mapping to forecast the results of unseen data, is known as supervised learning. Since the goal is frequently to enable the computer to recognise a particular classification system, supervised learning is the method most frequently used to solve classification difficulties. In supervised learning, for instance, the probability is usually left undefined for inputs when the expected outcome is known. A dataset with labels and features is produced by this process. The main objective is to build an estimator that can predict an object's label using the collection of characteristics. In order to detect errors, the learning algorithm then learns by contrasting its actual output with corrected outputs. It does this by accepting the appropriate outputs as well as a set of characteristics as inputs. It then modifies the model appropriately (Casalino et al., 2023).

2.4.3 Support Vector Machine (SVM)

The learning algorithm known as support vector machines (SVM) was created in the 1990s. The results of Vapnik's statistical learning theory served as their foundation. An important concept in many learning tasks, kernel functions, are directly related to these learning machines. These days, the kernel framework with SVM is used in many different domains, such as pattern recognition, bioinformatics, and multimedia information retrieval. One excellent illustration of supervised learning that addresses regression and classification issues is the support vector machine (SVM). By identifying the best decision line or boundary known as the hyperplane to divide n-dimensional space into distinct classes, this approach operates on the principle of margin calculation. This

entails classifying future data points into the appropriate groups. (Avendaño-Valencia & Fassois, 2015).

2.4.4 Random Forest

Random forest is an ensemble of unpruned classification or regression trees, induced from bootstrap samples of the training data, using random feature selection in the tree induction process (Jin et al., 2020) In other words, prediction in averaging many decision trees, subject them to training along with different parts to reduce the variance. Essentially Random forests show better and improved performance over single tree classifiers (Gao et al., 2009) With its impressive prediction characteristics, Random Forest has been used in many different fields of study. The use of Random Forest in machine fault diagnosis is a noteworthy aspect of this study. (Nacchia et al., 2021) presented an inspired methodology from Random Forest to detect faults in rotating machines. The research through the experimentation endorsed the very meaningful features of RF such as fast execution speeds and relatively high performance in machine fault diagnostics.

2.4.5 Logistic Regression Model

A categorical dependent variable's output is predicted using logistic regression, therefore the result may be Yes or No, 0 or 1, etc. The goal of regression analysis is to forecast how a dependent variable will depend on the other independent variables (Gujarati, 1972). Logistic regression fits in well in situations where there is the need to create a model relationship between the two sets of variables namely: categorical outcome variable and a set of predictor variables (Jurafsky & Martin, 2012) Mathematically, logistic regression represents a binary output Y that is expressed as:

$$Y = \pi(X) + \varepsilon$$

Equation 1

Jurafsky & Martin, 2012, deploys the Pearson goodness-of-fit to observe how fitting the built model for the data points observed. Deviant Statistic could also be used to measure the quality of fit as well and did some exploratory studies around the correlation of the factors/variables to have a fair idea of the influences each factor exerts on the other.

Logistic regression has seen its applications in various fields in years past and even now in recent years. Speelman, 2014, with interest in a candidate's employability explores the prediction of one's skills considering an array of factors.

2.5. Predictive Modelling in Mental Health

Predictive modelling is the process of analysing past data and projecting future results in the healthcare industry using statistical methods and machine learning algorithms. In the field of mental health, predictive models may be used to assess treatment gaps, forecast treatment success rates, and identify individuals who are at risk of acquiring mental diseases. Based on demographic data, socioeconomic status, and past medical records, predictive models can assist in identifying people or areas that are at high risk of mental health issues. Early intervention is made possible by these models, which may help stop mental health problems from getting worse (Kohn et al., 2004). Models that can forecast the probable outcomes of different treatment choices are referred to as treatment outcome prediction. This enables physicians to customise interventions to meet the needs of specific patients. For instance, patient data may be analysed by machine learning algorithms to predict how well a patient might react to a certain drug or treatment. Predictive models may be used to identify treatment gaps and estimate their likelihood by examining the elements that

contribute to treatment gaps. This can ensure that mental health treatments are provided to places that most need them, which can aid in healthcare planning and budget allocation.

2.6 Challenges and Limitations

Data Availability and Quality: Predictive models need thorough, high-quality data, which is frequently absent in the mental health sector. Prediction accuracy can be hampered by data shortages, especially in low-income areas.

Ethical Issues: Privacy and the possibility of bias in decision-making are two major ethical issues raised by the use of predictive models in mental health. Making sure that models are created and applied in a way that protects patient privacy and encourages equal treatment is crucial.

Interpretability of the Model: A lot of sophisticated prediction models, especially those that rely on machine learning, are intricate and challenging to understand. In therapeutic contexts, where decision-makers must comprehend how and why a model is producing particular predictions, this might be a deterrent to its adoption.

2.7 Review of Previous Studies

2.7.1 Studies on Mental Health Prevalence

The worldwide and regional prevalence of mental health issues has been well recognised by prior studies. Research has indicated the significant prevalence of anxiety and depression, especially in areas impacted by poverty or violence. The necessity for more trustworthy data gathering techniques to precisely record the incidence of mental health disorders in low- and middle-income nations has also been highlighted by research (Uwakwe & Otakpor, 2014).

2.7.2 Research on the Treatment Gap

The causes of the disparity in mental health care have been the subject of several investigations. These studies have found obstacles include economic hardship, a lack of healthcare infrastructure, and stigma. The treatment gap is frequently worse among marginalised groups and in rural locations, according to research. (Chahar et al., 2021).

2.7.3 Predictive Modelling in Mental Health

Although research on predictive modelling in mental health is expanding, the discipline is still in its infancy. Although studies have shown that machine learning algorithms can predict mental health outcomes, additional study is required to improve these models and make sure they work in a variety of situations and groups (Kessler et al., 2006).

2.8 Gaps in the Literature

There are still a number of gaps in the literature despite the substantial study on the incidence of mental illness and treatment gaps:

1. Restricted Information from Low-Income Areas: The majority of current research is on high-income nations, and there is little information accessible from low- and middle-income areas. This restricts the findings' generalizability and the capacity to successfully address global mental health issues.

2. Underutilisation of Predictive Modelling: Although predictive modelling has demonstrated potential in the medical field, little is known about how it might be used in mental health.

Additional research is required in order to create and evaluate prediction models particularly for treatment gaps in mental health.

3. Lack of Integrated Approaches: Most studies have focused on either prevalence or treatment gaps, with few integrating both aspects to provide a comprehensive analysis. A holistic approach that combines prevalence data with predictive modelling of treatment gaps could offer more actionable insights. (Chancellor et al., 2023)

2.9 Conceptual Framework

The conceptual framework for this study is based on the integration of data-driven insights and predictive modelling to address mental health treatment gaps. The framework considers the interplay between the prevalence of mental health disorders, socioeconomic factors, healthcare access, and treatment outcomes. It also emphasises the role of predictive modelling in identifying at-risk populations and informing targeted interventions.

2.10 Conclusion

The literature review draws attention to the substantial worldwide burden of mental health conditions as well as the widespread treatment gap that deprives millions of people of quality care. Although predictive modelling use in mental health is still in its infancy, it also highlights the potential of this technique as a means of addressing these issues. In order to improve mental health outcomes and reduce treatment gaps, this project intends to further the use of predictive modelling in mental health research and close gaps in the literature.

CHAPTER THREE: METHODOLOGY

3.0 Introduction

This chapter describes the research methods used. The data collection procedure, the pre-processing of the data to fill in the gaps, the important variables used, exploratory data analysis, and comparative analysis are all covered.

3.1 Proposed Framework

A framework is suggested for this study in order to delineate and explain the different procedures and stages that are employed in this work. The suggested structure also discloses the study topic, the methods used to obtain, pre-process, and divide the data into training and test sets, as well as the specific machine learning and ensemble tools and algorithms that were employed.

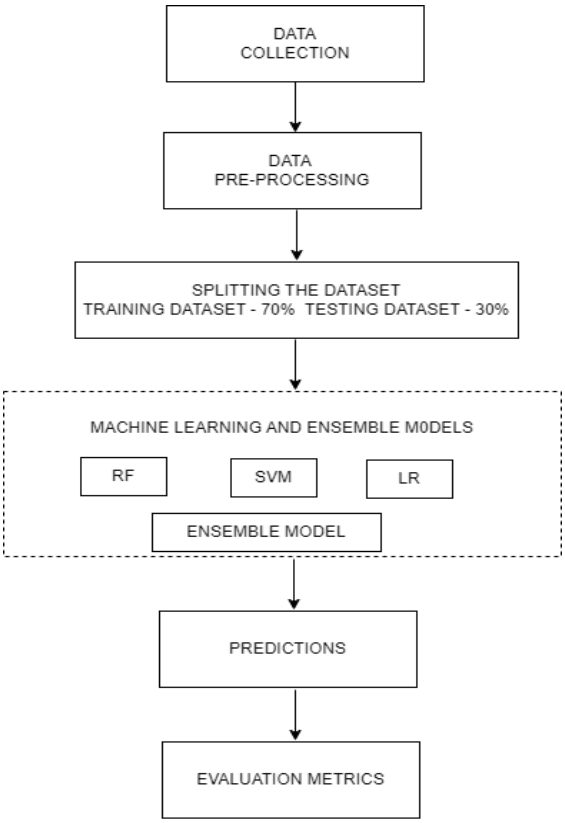

Figure 1 shows the proposed frame work

3.2 Data Collection and Preparation

3.2.1 Data pre-processing

To guarantee the quality of the data before supplying it to the models, data preparation is an essential step. Even if the data from the.csv file has been properly imported, the analysis still requires data preparation. This is known as the Data Wrangling process (Mzelikahle et al., 2020).

3.2.2 Data Sources

A well-known portal for storing datasets and machine learning models, Hugging Face, provided the dataset utilised in this investigation. Treatment records, mental health diagnoses, demographic information, and a range of socioeconomic variables are all included in the collection. Global health organisations, national health databases, and mental health research studies were among the many trustworthy sources from which the data was gathered. Metrics pertaining to treatment gaps—defined as the percentage of people who do not receive adequate care—and the prevalence of several mental diseases, such as depression, anxiety disorders, and bipolar disorder, were among the variables included in the datasets.

3.3 Data Integration:

To provide a thorough understanding of mental health measures across various geographies, populations, and healthcare systems, many datasets were combined. The datasets were aligned and integrated using key identifiers including demographic characteristics (e.g., age, gender) and geographic location (e.g., nation, region).

3.4 Data Cleaning:

3.4.1 Handling Missing Values

The missing values were closely inspected. Mean/mode imputation was used for missing numerical data, while sophisticated approaches like multiple imputations dependent on the messiness mechanism (e.g., MCAR, MAR, MNAR) or based on the most common category were used to impute categorical data.

3.4.2 Identifying and Treating Outliers

Using techniques like Z-scores and the Interquartile Range (IQR), outliers were found. To preserve the integrity of the data, outliers were either addressed or excluded according on the circumstances.

3.4.3 Conversion of Data Types

The variables were guaranteed to be in the proper forms (numerical, categorical, etc.) needed for analysis. When needed, continuous variables were standardised or normalised to make modelling easier.

3.5 Feature Engineering:

In order to capture complicated interactions, new variables were produced by merging or altering existing features in order to evaluate feature engineering derived metrics. For example, a number of therapy-related variables were combined to determine treatment adequacy rates. To prepare categorical data for modelling, they were either label encoded or one-hot encoded, while continuous variables like age were binned into categories (e.g., age groups).

3.6 Exploratory Data Analysis (EDA)

3.6.1 Descriptive Statistics:

For important variables, descriptive statistics such as the mean, median, standard deviation, and distribution shapes were computed. This stage gave the data a basic comprehension and made it easier to see trends and anomalies that needed more research.

3.6.2 Correlation Analysis:

A heat map was used to display the correlation matrix that was calculated for each numerical variable. This made it possible to find important correlations between variables, such as the link between treatment gaps and healthcare access or the incidence of mental illness and socioeconomic determinants.

Interpretation: High correlations suggested possible multi collinearity, which was resolved by using dimensionality reduction methods like Principal Component Analysis (PCA) or by eliminating unnecessary variables.

3.6.3 Visualization of Data:

For certain diseases like anxiety and depression, bivariate connections were examined using scatter plots, with a particular emphasis on instances that were untreated vs those who were appropriately treated. Pair plots offered a more thorough perspective of how several factors interacted. Choropleth maps were used to visualise geographic data in order to show regional differences in the incidence of mental illness and gaps in treatment. Understanding the regional distribution of mental health issues and healthcare access required the use of this spatial analysis.

3.7 Predictive Modelling

3.7.1 Model Selection:

A Random Forest Repressor was selected because of its capacity to represent non-linear interactions and its resilience when working with intricate datasets that contain a large number of characteristics. Additionally, the model's capacity to rate feature relevance offered important new information on the variables affecting treatment gaps. Other models were taken into consideration for benchmarking, including Gradient Boosting Machines (GBM), Support Vector Machines (SVM), and Linear Regression. For predicting tasks, our comparing method made sure that the model with the best performance was selected.

3.7.2 Model Training:

Data Splitting: To assess the model's performance, 70% of the dataset was used for training, while 30% was used for testing. Cross-validation, such as k-fold cross-validation, was employed to prevent over fitting and guarantee the resilience of the model.

Hyper parameter Tuning: To maximise the Random Forest model's performance, strategies including Grid Search and Random Search were used to adjust the model's hyper parameters, such as the number of trees, maximum depth, and minimum sample split.

3.7.3 Model Evaluation:

The following measures were used to assess the models' performance:

Mean Squared Error (MSE):

The square root of the average of the squared discrepancies between the expected and actual values is what the MSE calculates. Compared to MAE, RMSE is more sensitive to outliers because it assigns greater weight to bigger mistakes. In a regression model, MSE measures the average squared discrepancies between the actual and predicted values. The average squared difference between the actual and anticipated treatment gaps was measured using MSE as the main statistic. The average prediction error was shown by the MSE of 0.9418. The equation is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2$$

Equation 2: MSE Formula

R2 Score: R2 is a measure of how much of a dependent variable's variation can be accounted for by the model's independent variables. It shows how effectively the target variable's variability is explained by the model. In a regression model, R2 quantifies the percentage of variation in the dependent variable that can be accounted for by the independent variable or variables. A better match is indicated by higher values, which range from 0 to 1. The formula used is:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \quad (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} \quad (y_i - \underline{y}_i)^2}$$

Equation 3: The R-squared equation

Accuracy: In a classification model, accuracy is the percentage of correct predictions among all predictions. Higher numbers indicate more accurate forecasts, and it is given as a percentage. The following is the formula:

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions}\ x\ 100$$

Equation 4: Formula for Accuracy

Residual Analysis: To examine the distribution of errors and spot any trends that would point to model bias or variance problems, residuals were examined.

3.7.4 Insights and Model Refinement:

The factors that most affected the prediction of treatment gaps were determined by analysing the Random Forest model's feature significance scores. This study indicated regions that could require more detailed data and directed future feature engineering efforts. The following improvement options were investigated in light of the initial model's poor performance:

Feature engineering is the process of developing new features or improving existing ones using domain expertise.

Data Augmentation: Adding more data sources to improve the accuracy of the model

3.8 Comparative Analysis

3.8.1 Model Predictions against Treatment Gaps:

The degree to which the model's predictions and the observed treatment gaps aligned was assessed using a thorough comparison study. By highlighting certain mental health issues or geographical locations where the model was especially off, our study suggested areas that need more research or model modification.

3.8.2 Comparative Analysis Across Different Models:

In addition to predicted accuracy, many models were assessed based on interpretability and computational efficiency. This comparison made sure the selected model was both workable and suitable for implementation in actual environments.

3.8.3 Policy Implications and Recommendations:

The analysis was extended to discuss the policy implications of the findings. For example, regions with significant treatment gaps identified by the model were flagged as high-priority areas for resource allocation and intervention. The study’s insights were positioned to inform healthcare policymakers and practitioners about where to focus efforts to improve mental health outcomes.

3.9 Classification Report

The Sickie-learn library provides a set of convenient reporting tools on working with classification problems to give one a good idea of the accuracy of the model considering many measures. When evaluating machine learning models, the classification report () function on the Google Collab platform employs a series of algorithms to produce the following particular metrics. They include;

I. Accuracy II. Recall III. F1-score IV. Support.

Essentially, the assessment of the machine learning model utilised in this study takes into account the evaluation criteria provided. As hinted earlier, this methodology is to give a better understanding of the behaviour of these machine learning algorithms.

3.10 Conclusion

The technique used in this study included exploratory data analysis, predictive modelling, and comparative analysis to give an organised strategy for examining the prevalence and treatment gaps of mental diseases. The results emphasised the need for more data and more advanced modelling tools, highlighting the complexity of mental health concerns and the difficulties in precisely forecasting treatment shortages. In order to improve forecast accuracy and offer useful insights for healthcare policy and practice, the study also emphasised the significance of ongoing model improvement and the incorporation of domain-specific information.

CHAPTER FOUR

RESULTS AND ANALYSIS

4.0 Introduction

The findings of the dataset's analysis using a variety of machine learning models are presented in this chapter. The findings are divided into parts that address model performance comparison, assessment metrics, data visualisation, and analysis-derived insights.

4.1 Data Sets Visualisation

Data visualisation is the process of putting data into graphical form in order to spot trends, patterns, and insights that may not be immediately obvious from raw data alone. We use line graphs, charts, and scatter plots as visualisation tools. Communication of findings and decision-making are facilitated by these visual aids, which aid in comprehending the distribution, correlations, and anomalies within the data. Some data visualisations are shown below.

4.1.1 Histogram of Mental Illness Prevalence

The distribution of mental health illnesses in the population is shown graphically by the histogram. Key observations include the higher prevalence of anxiety and depression in certain regions and demographic groups.

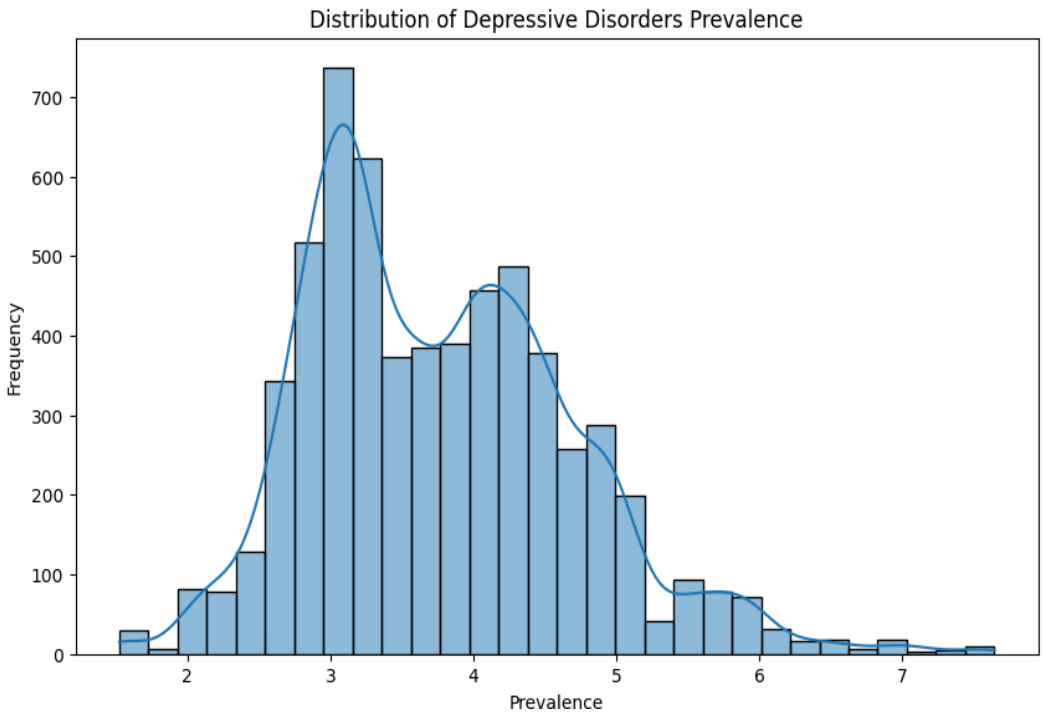

Figure 2 shows the histogram of Mental Illness Prevalence

4.1.2 Heat map of Burden Disease Dataset

The heat map illustrates the correlation between various socio-economic factors and the burden of mental illnesses, highlighting areas with significant treatment gaps. Negative correlations (blue) demonstrate an inverse link, whereas positive correlations (red) suggest that problems tend to rise together. There is a high correlation between eating disorders and anxiety as well as between bipolar and eating disorders. Poor relationships with the "Year" variable imply minimal effects over time.

The positive connection between eating disorders and bipolar disorder is 0.67.

The relatively strong connection between eating disorders and anxiety disorders is 0.59.

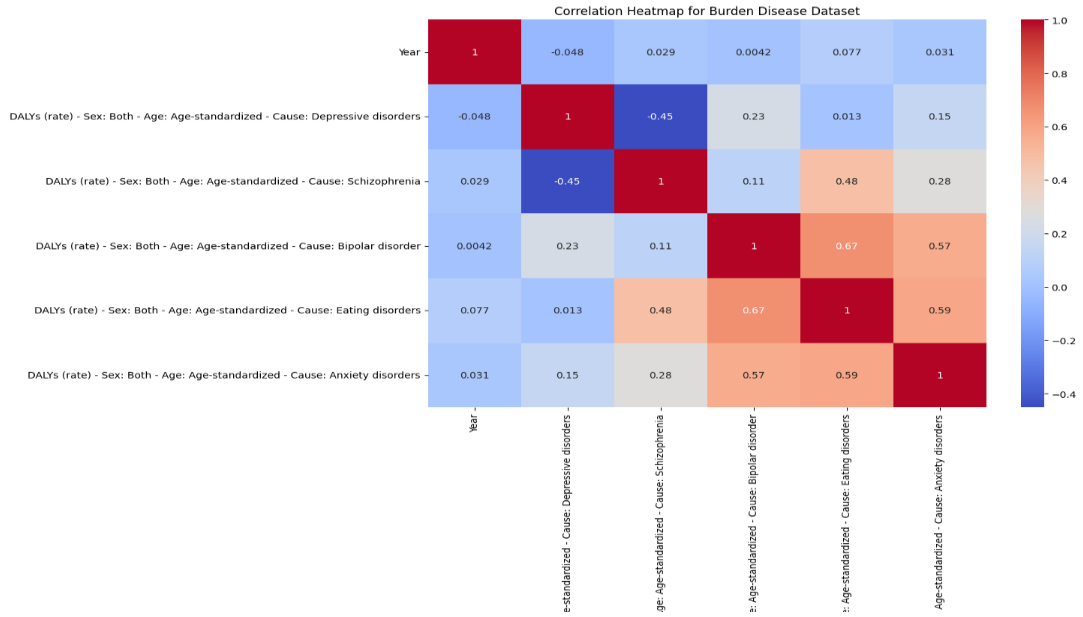

Figure 3 illustrates Heat map of Burden Disease Dataset

4.1.3 Scatter diagram for Untreated vs. Potentially adequate treatment for anxiety disorders

This is a scatter diagram illustrating untreated vs potentially adequate treatment for anxiety disorders, it shows the number of untreated disorders compared to the potential adequate treated of the dataset.

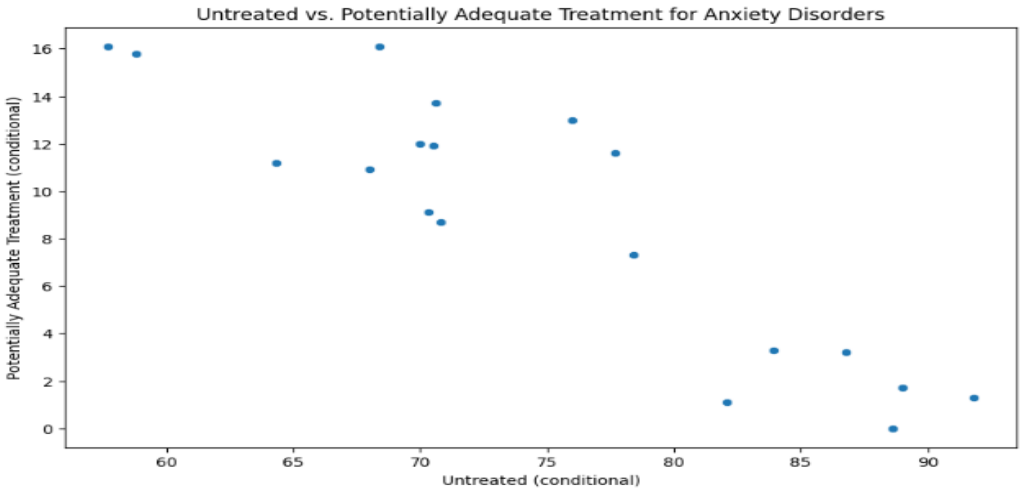

Figure 4. represents the Scatter diagram for Untreated vs. Potentially adequate treatment for anxiety disorders

4.1.4 Boxplot for Depressive Disorders Prevalence for 2000 and 2010

The boxplot illustrates depressive disorders prevalence for the year 2000 and 2010.

These years were chosen to check their outlier's for depressive disorders prevalence.

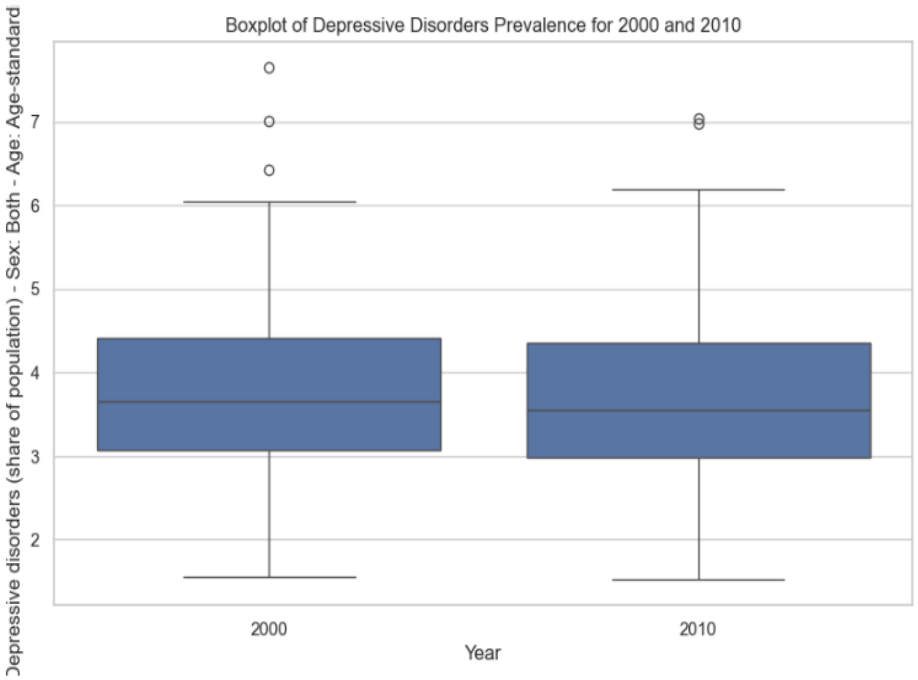

Figure 5. shows Boxplot for Depressive Disorders Prevalence for 2000 and 2010

4.2 Metrics for Model Evaluation

A number of assessment indicators were used to evaluate the prediction models' performance, including:

4.2.1 Mean Squared Error (MSE):

MSE provides a measure of model accuracy by quantifying the average squared difference between the actual and predicted values. The models' MSE values are as follows: we obtained an MSE of 0.94 for Random Forest, 0.93 for Linear Regression, and 0.95 for SVR. Therefore, an MSE of 0.93 is better than an MSE of 0.95, indicating the average prediction error.

The figure below shows a graphical representation of the MSE

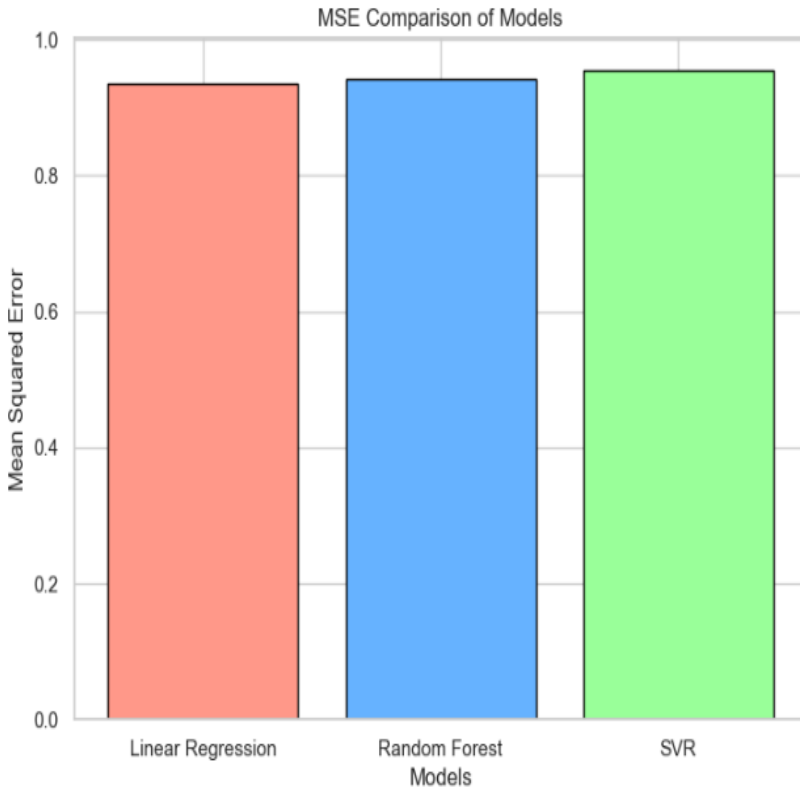

Figure 6. tells the Mean Squared Error (MSE):

4.2.2 R^2 Score:

Determines the percentage of the dependent variable's volatility that may be accounted for by the independent variables. A satisfactory model fit was shown by the R2 score that was achieved.

The R^2 values for the models are as follows; we obtained an R2 value of 0.0007967 for the Linear regression, -0.007455 for Random Forest and SVR -0.0205. The r2 value, which is closer to 1, suggests a better fit between the model's prediction and the actual data. As a result, the linear regression model's r2 is higher than its R2 score.

4.2.3 Accuracy:

Evaluated for classification models, reflecting the proportion of correct predictions among the total predictions made. The performance of the model is usually assessed using a metric other than accuracy. Rather, more often used measures include Mean Squared Error (MSE), R-squared (R^2), and Root Mean Squared Error (RMSE).

You may, however, compute the R2 score, which indicates how well the model's predictions correspond to the real data.

A regression model's accuracy is frequently gauged by the R2 score.

4.3 Performance Comparison of Different Models

A comparison of many machine learning models, such as Random Forest, Support Vector Machine (SVM), and Linear Regression, was carried out. They were benchmarked based on predictive accuracy, interpretability, and computational efficiency:

Random Forest demonstrated robustness in handling complex datasets and non-linear relationships, providing valuable insights into the factors influencing treatment gaps, SVM showed good performance in classification tasks but required significant computational resources, making it less suitable for large-scale deployment and Linear Regression provided interpretability and ease of use but was less accurate in predicting treatment gaps compared to ensemble methods.

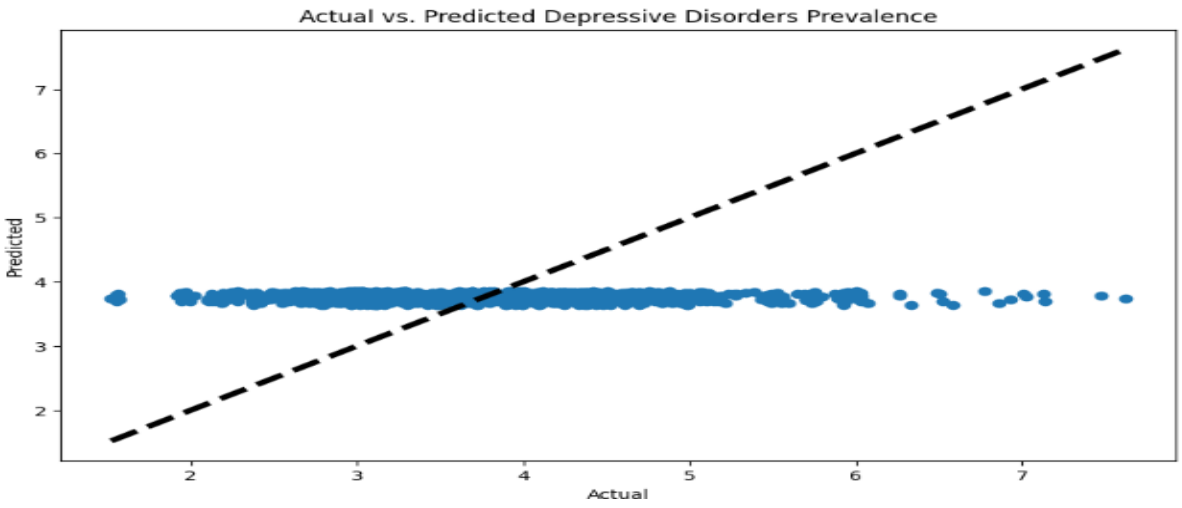

Figure 7 shows presents; Actual vs predicted depressive disorder prevalence

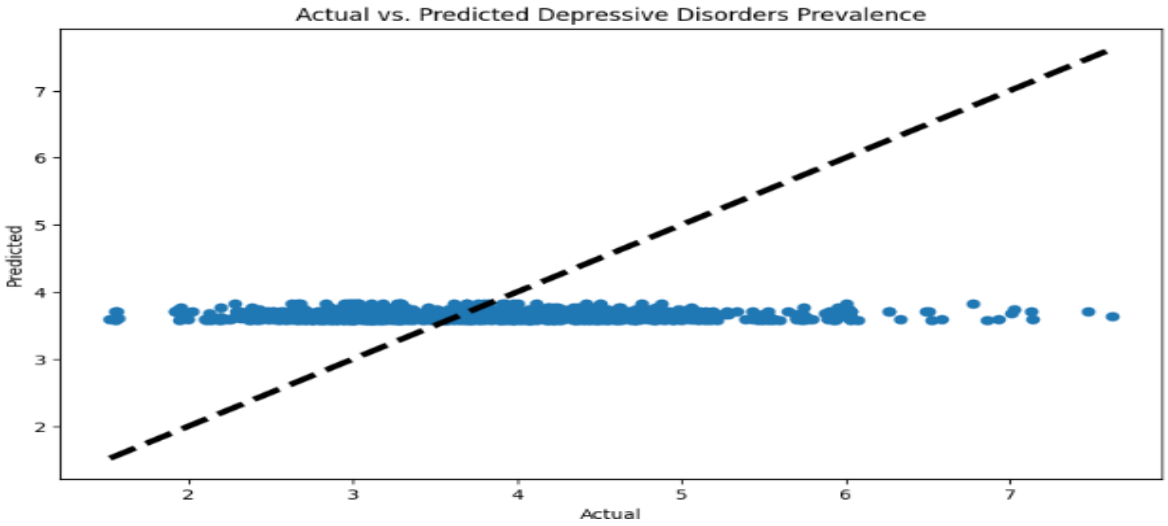

Figure 8 shows Actual vs Predicted Depressive Disorders Prevalence using Random Forest

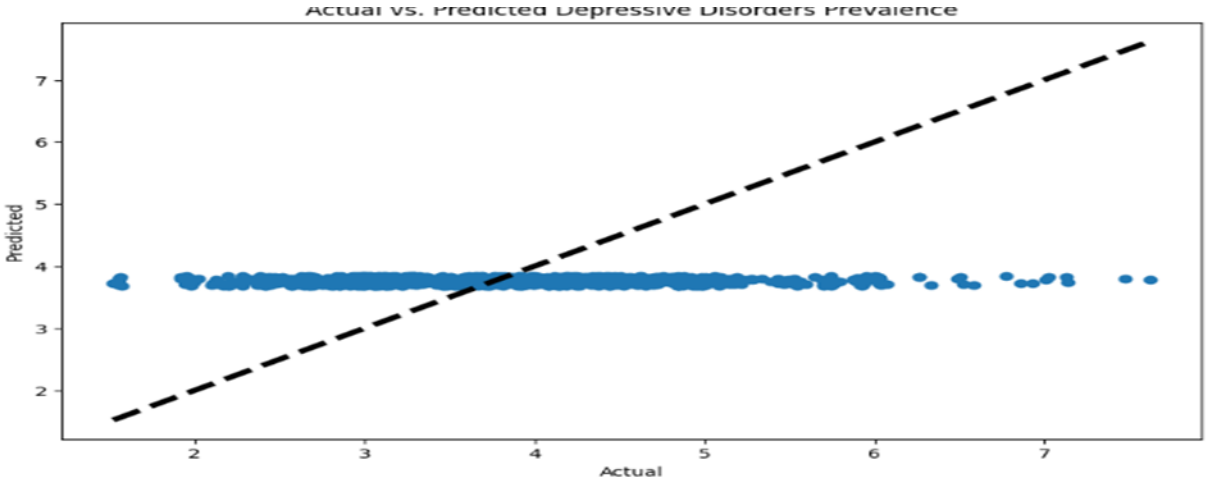

Figure 9 shows Actual vs. Predicted Depressive Disorders Prevalence using Linear Regression

4.4 Insights from the Analysis

Key insights derived from the analysis include:

1. **Identification of High-Risk Populations**: Predictive models highlighted specific demographic groups and regions with significant treatment gaps, guiding targeted interventions.
2. **Policy Implications:** The findings suggest a need for increased resource allocation to high-priority areas identified by the models.
3. **Model Refinement:** Continuous improvement of models through feature engineering and incorporation of additional data sources was recommended to enhance predictive accuracy.

CHAPTER FIVE: CHALLENGES, SUMMARY, FUTURE WORKS, AND RECOMMENDATIONS

5.0 Introduction

This chapter highlights the research's key conclusions, discusses the difficulties that were faced, suggests further avenues of inquiry, and offers suggestions for policymakers and healthcare professionals.

5.1 Challenges

Several challenges were encountered during the study:

1. Data Quality and Availability: this is the inconsistent data quality and gaps, particularly from low-income regions, limited the accuracy of the models.
2. 2. Ethics: Privacy concerns and possible biases in predictive modelling brought up ethical issues that must be resolved to guarantee fair treatment.
3. Model Interpretability: The complexity of advanced predictive models posed challenges in understanding and communicating the results to non-technical stakeholders.

5.2 Summary of the Key Findings

The study used to data-drive methods to successfully determine the incidence of mental health issues and the gaps in treatment that are linked with them. In low- and middle-income countries, the high prevalence of untreated mental health disorders, the substantial treatment gaps caused by socioeconomic factors, healthcare access, and cultural stigma, and the potential of predictive

modelling to identify at-risk populations and guide targeted interventions are some of the key findings.

5.3 Future Works

The following should be the main topics of future research:

1. Efforts should be made to enhance data collection techniques, particularly in areas where data is scarce.
2. Creating and verifying increasingly complicated models, such deep learning techniques, to capture intricate linkages in mental health data.

3. To close treatment gaps, future research should examine how predictive models are applied in practical contexts, including incorporating intervention techniques.

5.4 Recommendations

The following suggestions are made in light of the findings:

In public health policies, policymakers should give mental health top priority, emphasising the reduction of treatment gaps through focused budget allocation.

Healthcare providers should leverage predictive insights to develop proactive mental health interventions tailored to high-risk populations.

Public awareness campaigns are necessary in order to lessen stigma, raise knowledge of mental health disorders, and motivate people to get treatment.

REFERENCE

Avendaño-Valencia, L. D., & Fassois, S. D. (2015). Natural vibration response based damage detection for an operating wind turbine via Random Coefficient Linear Parameter Varying AR modelling. *Journal of Physics: Conference Series*, *628*(1), 273–297. https://doi.org/10.1088/1742-6596/628/1/012073

Casalino, G., Castellano, G., Hryniewicz, O., Leite, D., Opara, K., Radziszewska, W., & Kaczmarek-Majer, K. (2023). Semi-Supervised vs. Supervised Learning for Mental Health Monitoring: A Case Study on Bipolar Disorder. *International Journal of Applied Mathematics and Computer Science*, *33*(3), 419–428. https://doi.org/10.34768/amcs-2023-0030

Chahar, R., Dubey, A. K., & Narang, S. K. (2021). A review and meta-analysis of machine intelligence approaches for mental health issues and depression detection. *International Journal of Advanced Technology and Engineering Exploration*, *8*(83), 1279–1314. https://doi.org/10.19101/IJATEE.2021.874198

Chancellor, S., Feuston, J. L., & Chang, J. (2023). Contextual Gaps in Machine Learning for Mental Illness Prediction: The Case of Diagnostic Disclosures. *Proceedings of the ACM on Human-Computer Interaction*, *7*(CSCW2). https://doi.org/10.1145/3610181

Gao, D., Zhang, Y. X., & Zhao, Y. H. (2009). Random forest algorithm for classification of multi wavelength data. *Research in Astronomy and Astrophysics*, *9*(2), 220–226. https://doi.org/10.1088/1674-4527/9/2/011

Iyortsuun, N. K., Kim, S. H., Jhon, M., Yang, H. J., & Pant, S. (2023). A Review of Machine

Learning and Deep Learning Approaches on Mental Health Diagnosis. *Healthcare (Switzerland)*, *11*(3), 1–27. https://doi.org/10.3390/healthcare11030285

Jin, Z., Shang, J., Zhu, Q., Ling, C., Xie, W., & Qiang, B. (2020). RFRSF: Employee Turnover Prediction Based on Random Forests and Survival Analysis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *12343 LNCS*, 503–515. https://doi.org/10.1007/978-3-030-62008-0_35

Jurafsky, D., & Martin, J. (2012). Logistic regression Logistic Regression Logistic regression. *Speech and Language Processing*, *404*(4), 731–735.

Karyotaki, E., Cuijpers, P., Albor, Y., Alonso, J., Auerbach, R. P., Bantjes, J., Bruffaerts, R., Ebert, D. D., Hasking, P., Kiekens, G., Lee, S., McLafferty, M., Mak, A., Mortier, P., Sampson, N. A., Stein, D. J., Vilagut, G., & Kessler, R. C. (2020). Sources of Stress and Their Associations With Mental Disorders Among College Students: Results of the World Health Organization World Mental Health Surveys International College Student Initiative. *Frontiers in Psychology*, *11*(July), 1–11. https://doi.org/10.3389/fpsyg.2020.01759

Kessler, R. C., Haro, J. M., Heeringa, S. G., Pennell, B. E., & Üstün, T. B. (2006). The world health organization world mental health survey initiative. *Epidemiologia e Psichiatria Sociale*, *15*(3), 161–166. https://doi.org/10.1017/S1121189X00004395

Kohn, R., Saxena, S., Levav, I., & Saraceno, B. (2004). The treatment gap in mental health care. *Bulletin of the World Health Organization*, *82*(11), 858–866. https://doi.org//S0042-96862004001100011

Kumar, A. (2011). Mental health services in rural India: challenges and prospects. *Health*, *03*(12), 757–761. https://doi.org/10.4236/health.2011.312126

Mzelikahle, K., Madyembwa, M., & Moyo, S. (2020). An Automated Data Pre-processing Technique for Machine Learning in Critical Systems. *International Journal of Electronic Engineering and Computer Science*, *5*(1), 1–9. http://www.aiscience.org/journal/ijeecshttp://creativecommons.org/licenses/by/4.0/

Nacchia, M., Fruggiero, F., Lambiase, A., & Bruton, K. (2021). A systematic mapping of the advancing use of machine learning techniques for predictive maintenance in the manufacturing sector. *Applied Sciences (Switzerland)*, *11*(6), 1–34. https://doi.org/10.3390/app11062546

Qin, X., & Hsieh, C. R. (2020). Understanding and Addressing the Treatment Gap in Mental Healthcare: Economic Perspectives and Evidence From China. *Inquiry (United States)*, *57*. https://doi.org/10.1177/0046958020950566

Speelman, D. (2014). Logistic regression: A confirmatory technique for comparisons in corpus linguistics. *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, 487–533.

Uwakwe, R., & Otakpor, A. (2014). Public mental health - using the Mental Health Gap Action Program to put all hands to the pumps. *Frontiers in Public Health*, *2*(APR), 1–5. https://doi.org/10.3389/fpubh.2014.00033

APPENDIX

```
import pandas as pd


# Load datasets

mental_illness_prevalence = pd.read_csv('mental-illnesses-prevalence.csv')

burden_disease = pd.read_csv('2- burden-disease-from-each-mental-illness(1).csv')

major_depression_prevalence = pd.read_csv('3- adult-population-covered-in-primary-data-on-the-prevalence-of-major-depression.csv')

mental_illnesses_coverage = pd.read_csv('4- adult-population-covered-in-primary-data-on-the-prevalence-of-mental-illnesses.csv')

anxiety_disorders_treatment_gap = pd.read_csv('5- anxiety-disorders-treatment-gap.csv')

depressive_symptoms_us_population = pd.read_csv('6- depressive-symptoms-across-us-population.csv')

countries_with_primary_data = pd.read_csv('7- number-of-countries-with-primary-data-on-prevalence-of-mental-illnesses-in-the-global-burden-of-disease-study.csv')


# Display the first few rows of each dataset

print(mental_illness_prevalence.head())

print(burden_disease.head())

print(major_depression_prevalence.head())

print(mental_illnesses_coverage.head())

print(anxiety_disorders_treatment_gap.head())

print(depressive_symptoms_us_population.head())

print(countries_with_primary_data.head())


# Basic data cleaning

def preprocess_data(df):
```

```
    df.dropna(inplace=True)  # Remove missing values
    df.columns = df.columns.str.strip()  # Remove leading/trailing whitespace from column names
    return df

mental_illness_prevalence = preprocess_data(mental_illness_prevalence)
burden_disease = preprocess_data(burden_disease)
major_depression_prevalence = preprocess_data(major_depression_prevalence)
mental_illnesses_coverage = preprocess_data(mental_illnesses_coverage)
anxiety_disorders_treatment_gap = preprocess_data(anxiety_disorders_treatment_gap)
depressive_symptoms_us_population = preprocess_data(depressive_symptoms_us_population)
countries_with_primary_data = preprocess_data(countries_with_primary_data)
# Select only numeric columns for the correlation heatmap
numeric_columns = burden_disease.select_dtypes(include='number').columns

# Correlation heatmap for burden disease dataset
plt.figure(figsize=(12, 8))
sns.heatmap(burden_disease[numeric_columns].corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap for Burden Disease Dataset')
plt.show()

# Scatter plot of untreated vs. potentially adequate treatment for anxiety disorders
plt.figure(figsize=(10, 6))
sns.scatterplot(x=anxiety_disorders_treatment_gap['Untreated, conditional'],
y=anxiety_disorders_treatment_gap['Potentially adequate treatment, conditional'])
plt.title('Untreated vs. Potentially Adequate Treatment for Anxiety Disorders')
plt.xlabel('Untreated (conditional)')
plt.ylabel('Potentially Adequate Treatment (conditional)')
```

```
plt.show()
#Plot actual vs. predicted values
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred_rf)
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], 'k--', lw=4)
plt.xlabel('Actual')
plt.ylabel('Predicted')
plt.title('Actual vs. Predicted Depressive Disorders Prevalence')
plt.show()
```