MENTAL HEALTH PREDICTION USING MACHINE LEARNING

BHUMIKA JAWA

Final Thesis Report

DECEMBER 2023

TABLE OF CONTENTS

DEDICATION

This thesis is dedicated to my parents who have constant conviction in my skills and has supported and motivated me throughout my academic career. The route to this success has been lighted by your love and motivation, and I will always be thankful for having you in my life.

ACKNOWLEDGEMENTS

I would like to extend my sincere gratitude to those who helped my master's thesis come to fruition,

I am extremely thankful of my family and friends for their consistent support and direction during this academic path. Their encouragement and support have been crucial throughout this difficult but worthwhile endeavor.

I would like to express my gratitude to the professors and specialists who kindly offered their knowledge and skills, enhancing the breadth and depth of my thesis.

Finally, I'd want to thank everyone who helped with this effort, whether they were actively involved in the study or not. We really value the donations you have made.

ABSTRACT

The effect of mental health problems on both the wellbeing of individuals and society makes them a worldwide public health concern. The use of machine learning to predict mental health has drawn a lot of interest recently and offers intriguing opportunities for early identification, risk assessment, and customized therapies.

Predictive modeling has a crucial significance in the field of mental health. It recognizes the value of early identification and intervention in enhancing outcomes for those who are dealing with mental health issues. Deep learning models and classification algorithms are only two examples of machine learning techniques that provide effective tools for forecasting outcomes related to mental health. These models have the potential to transform the way mental health services are provided by utilising a variety of data sources, such as clinical evaluations, demographic data, and behavioral data. To achieve responsible and successful deployment, ethical issues and privacy worries need to be properly considered.

In summary, machine learning-based mental health prediction has enormous potential for tackling the complex issues associated with mental health illnesses.

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AUC	Area under the Curve
KNN	K-Nearest Neighbors
ML	Machine Learning
AI	Artificial Intelligence
NLP	Natural Language Processing
NN	Neural Network
FPR	False Positive Rate
TPR	True Positive Rate
ROC	Receiver Operating Characteristics

CHAPTER 1

INTRODUCTION

1.1 Background

Mental health is a major public health concern across the world, with a rising number of people dealing with mental health issues. The importance of early identification and prompt intervention in enhancing outcomes and lowering the burden of mental health issues cannot be overstated. These illnesses have serious personal, social, and economic repercussions on people of all ages, ethnicities, and socioeconomic statuses. More than 450 million people worldwide suffer from mental health illnesses, making it one of the top causes of disability globally, according to the World Health Organisation (WHO).

Standardised tests, clinical observations, and self-assessments have been used as the backbone of traditional methods for diagnosing mental illness. However, these approaches frequently have drawbacks, such as inherent biases, prolonged assessment durations, and challenges in identifying early warning indications. Additionally, the stigma attached to mental health conditions may discourage some people from seeking treatment or from disclosing their genuine symptoms during professional assessments.

More innovative techniques for mental health prediction and intervention have recently become feasible thanks to the development of technology, the accessibility of large-scale datasets, and the rising interest in AI and ML. These models seek to make use of data from various sources, including wearable technology, social media activity, smart phone apps, genetic data, and electronic health records, in order to spot patterns, risk factors, and markers that can help with early detection and individualised treatment plans.

Models for predicting mental health have substantial potential advantages. Early mental health issue detection allows medical professionals to take appropriate action, preventing more serious illnesses and increasing treatment outcomes overall.

In order to predict mental health, machine learning methods such as deep learning, decision trees, and random forest are essential. Large and complicated datasets may be processed by these algorithms, which can also find hidden patterns and forecast the future using learnt patterns from prior data. The potential of machine learning models for mental health is growing along with the area of artificial intelligence.

To sum up, mental health prediction has the power to completely alter the way we approach mental health treatment. Healthcare practitioners may enhance early identification, intervention, and individualised treatment options for those with mental health difficulties by using the potential of artificial intelligence.

1.2 Problem Statement

With more people experiencing mental health concerns, mental health is a significant health concern worldwide. It is impossible to estimate the value of early detection and timely intervention in improving outcomes and minimising the burden of mental health concerns.

A rapidly developing subject, mental health prediction research uses a variety of machine learning and predictive modelling approaches to discover and forecast outcomes related to mental health. Here are some important areas of study and papers that go into further detail into mental health forecasting:

Machine learning was utilised in a study that was published in JAMA Psychiatry to forecast depression outcomes using information from electronic health records. The methodology successfully identified patients who were at high risk of developing depression, allowing for early intervention and individualised therapy. (Moshe et al., 2021)

Natural language processing (NLP) techniques have been used by researchers to analyse text data from social media and detect those who are at risk for anxiety disorders. Predictive algorithms can shed light on anxiety-related symptoms by looking for language patterns and emotional clues in social media posts. (Guntuku et al., 2017)

Research has used genetic information to forecast the likelihood of getting schizophrenia. Genetic markers that relate to schizophrenia susceptibility have been identified through polygenic risk score and genome-wide association studies (GWAS). (Cho et al., 2019)

Machine learning algorithms have been used in studies to find trends and risk variables related to suicide risk. To determine a person's chance of attempting suicide, predictive models have been created utilising demographic information, mental health history, and other clinical markers. (Lee & Pak, 2022)

After exposure to stressful events, researchers have focused on predicting the likelihood of developing PTSD. To identify those who are more likely to acquire PTSD, machine learning algorithms have been used to analyse physiological data, biomarkers, and behavioural patterns. (Cooper et al., 2021)

1.2 Aim and Objectives

In today's generation it is very common to observe cases of mental health. Mental illness have taken a shape of pandemic where every other person is dealing with some or the other mental health issue i.e. Anxiety, depression, eating disorders and addictive behaviours.
The aim of the study is to use data-driven approaches, such as machine learning and deep learning models, to identify patterns in data and to collect valuable insights from that so as to capture the trend behind these issues. Other major concern is the risk factors associated with the mental illness which has be life taking as well. So, this study will cover all the possible ways to develop a machine learning model so as to reduce the risk of mental illness by early warning signs related to mental health disorders.

Now to avoid such problems of mental illness one possible way could be to determine at an early stage so that problem could be cured before it even affects a person life and provide personalized solutions for the same to deal the problem. This will also help us in targeting the patients who are at high risk and need special care and attention. One of the important reason behind working on this is to bring awareness among people so that they can accept the problem and take action.

Objectives:

- The primary objective is to identify mental health illnesses early, even before visible signs appear
- The goal of mental health prediction models is to offer individualised treatment plans based on a person's unique traits, medical history, and anticipated reactions to therapies.
- A person's risk of having a mental health condition may be determined using predictive models based on a variety of variables, including genetic predisposition, environmental effects, and behavioural patterns. By being aware of these dangers, precautions can be done to lessen the possibility of mental health issues.

- By identifying high-risk patients who need more intense care and assistance, mental health prediction can help with the effective allocation of healthcare resources. This may result in enhanced healthcare system efficiency and optimal resource utilisation.
- By encouraging the idea of early intervention and preventative care, mental health prediction research can help to lessen stigma around mental health concerns. Predictive models can enhance overall mental health outcomes by proactively urging people to seek treatment.
- A better understanding of the complex connections between multiple risk factors and mental health outcomes might result from the creation and use of mental health prediction models. This may encourage more study and improvements in the field of mental health treatment.

1.4 Research Questions

- How accurate are these machine learning models in spotting early detection of mental health?
- How can ML models help in early detection of mental health disorders?
- Which factors play an important role in predicting the metal health condition?
- How can machine learning algorithms improve the accuracy of mental health prediction models

1.5 Significance of the Study

In the field of healthcare and mental health research, the study of mental health prediction is of utmost importance. Some of the reasons why the study of mental health prediction is important are:

- Early Intervention and Prevention: The early detection of those who are at risk of developing mental health issues is made possible by predictive models build for mental health prediction. Early intervention and preventative actions may prevent symptoms from getting worse and enhance the effectiveness of treatment.
- Individualised Care: Predictive models allow treatment regimens to be customised based on each patient's specific traits, improving the efficiency and patient focus of mental health care.

- Lowering the cost of healthcare: By avoiding more serious diseases and lowering hospitalisation rates, early identification and customised therapies may be able to lower the overall healthcare expenses related to mental health issues.
- Breaking Stigma: By advocating for early intervention and preventative care, mental health prediction research can aid in the de-stigmatization of mental health conditions. This can motivate people to actively seek assistance and lessen the social stigma associated with mental health difficulties.

1.6 Scope of the Study

- Data Collection: The research can investigate the utilisation of many data sources, including genetic data, wearable device data, social media activity, and electronic health records. It can concentrate on locating appropriate features and data preparation methods that aid in making precise mental health forecasts.
- Machine Learning Algorithms and Models: The goal of the study is to create prediction models for outcomes related to mental health by examining several machine learning techniques, including deep learning, SVM, decision trees, and ensemble approaches. Finding the best strategy for a given prediction problem may be determined by comparing the performance of multiple models.
- Model Outcome: The model will predict whether treatment is required by the individual or not.
- Model Evaluation: Using proper measures, like accuracy, precision & recall the performance of the models will be accessed. The results of all the models will be compared to find the best model for early detection of mental health.
- Future Research directions: By encouraging the idea of early intervention and preventative care, mental health prediction research can help to lessen stigma around mental health concerns. A better understanding of the complex connections between multiple risk factors and mental health outcomes might result from the creation and use of mental health prediction models. This may encourage more study and improvements in the field of mental health treatment.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Human well-being is fundamentally influenced by mental health, which has an effect on many facets of our lives, including our relationships, productivity at work, and general quality of life. However, millions of individuals worldwide are plagued by illnesses including depression, anxiety, schizophrenia, and bipolar disorder, and mental health issues are becoming more common. These disorders impose major emotional, social, and financial costs, underscoring the urgent need for new methods of providing mental healthcare.

In recent years, there has been a rising understanding of the need of early detection and prevention in tackling the complex issues of mental health. The subject of mental health prediction has emerged as a result of this realization, and it makes use of developments in technology, data science, and healthcare to anticipate the likelihood of mental health illnesses, identify at-risk patients, and improve treatment methods.

A fundamental change in how we see mental health may be seen in mental health prediction. It aims to actively identify those who may be at risk for mental health issues rather than waiting for symptoms to appear or crises to happen. The goal of mental health prediction is to usher in a new era of personalized and preventative mental healthcare by using the capabilities of predictive models, machine learning algorithms, clinical evaluations, and a variety of data sources.

This review delves into the methodology, applications, difficulties, and possible ramifications of this developing area as it investigates the varied landscape of mental health prediction. This review aims to emphasize the significance of mental health prediction in reshaping the way we approach mental well-being with the ultimate goal of lowering the burden of mental health disorders and enhancing the lives of individuals and communities by reviewing the body of existing research and highlighting emerging trends.

This chapter reviews the studies carried out on early prediction of mental health state and take necessary actions timely. In this field of early mental health prediction, we will be working on various machine learning models like Logistic Regression, Random forest, K-Neighbors Classifier and as well as neural networks. This chapter begins with liner models like logistic regression techniques followed by non linear models like random forest and then detailed review of the boosting models to get accurate results. Some extensively used models are Logistic Regression, K-Neighbors Classifier, Decision Tree, Random Forest, Boosting and Neural Network.

2.2 Logistic Regression

An important statistical modeling method in the field of mental health prediction is logistic regression. It is frequently used to evaluate the likelihood of binary outcomes, making it especially pertinent for determining whether or not a person is at risk of acquiring a mental health issue. We will address the use of logistic regression in the context of mental health prediction and its usefulness in enhancing early intervention and individualized treatment approaches in this session.

Logistic Regression is widely used to evaluate the risk variables for mental health disorders. Data on numerous predictors, including demographic details, family history, lifestyle characteristics, and psychological variables, are gathered by researchers and healthcare experts. These indicators are utilized in the logistic regression model as independent variables to forecast the probability that a person would experience a mental health problem like depression or anxiety. The likelihood score that is produced can be used to find high-risk people who might benefit from early intervention.

For prompt care, early mental health condition diagnosis is essential. On the basis of historical data from people with known outcomes in terms of their mental health, logistic regression models may be developed. The algorithm can learn to identify trends and risk factors linked to the beginning of particular mental health problems by analyzing the historical information. The model can offer early warnings or risk assessments when used with fresh data, enabling medical professionals to take action before symptoms get worse.

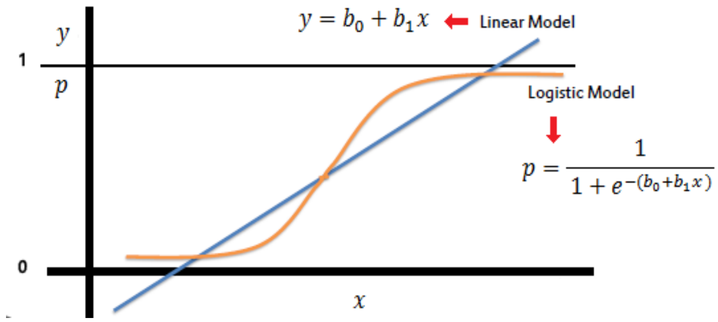


Figure 2.1 Working of Logistic Regression Model

For quick action, early diagnosis of mental health issues is essential. On historical data from people with known mental health outcomes, logistic regression models may be developed. These historical datasets may be used to train the model to identify patterns and risk factors related to the beginning of particular mental health problems. When used on fresh data, the model can offer risk evaluations or early warnings, enabling medical professionals to take action before symptoms get worse.

For people with mental health issues, the likelihood of particular treatment outcomes may also be forecast using logistic regression. Clinicians can determine the likelihood of a patient responding successfully to therapy or the risk of relapsing by looking at baseline features and treatment-related factors. This data aids in continuing monitoring and modifications while guiding treatment choices.

The use of logistic regression in mental health prediction aids in the establishment of evidence-based research and policy. It makes it possible for researchers to pinpoint important protective factors, risk factors, and treatment predictions for a range of mental health issues. In order to lessen the prevalence and effects of mental diseases, public health policies, preventative programmes, and focused treatments can be developed with this knowledge.

2.3 K Neighbours Classifier:

Due to its efficiency and simplicity, the K-Nearest Neighbour (K-NN) classifier, a machine learning technique, has drawn interest in the field of mental health prediction, especially when the data cannot be linearly separated. K-NN is a supervised learning method that is frequently used in mental health prediction modelling for a variety of reasons.

The similarity concept underlies the k-NN algorithm. The majority class among a data point's k-nearest neighbours in the feature space is used to classify it. According on the data and problem at hand, alternative distance metrics may be employed instead of the commonly used Euclidean distance to find the "nearest neighbours". Cross-validation is frequently used to establish the value of k, the number of neighbours to take into account, a critical parameter that affects the algorithm's performance.

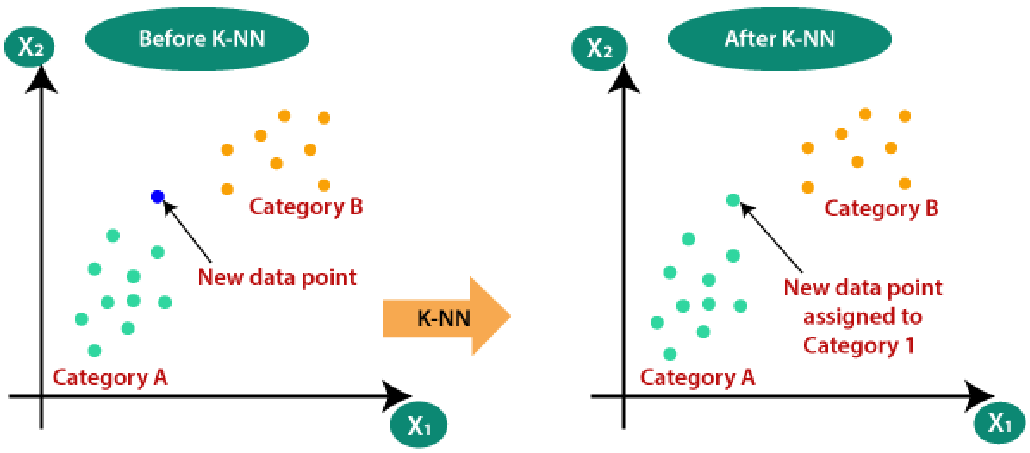


Figure 2.2 Working of KNN model

Based on a variety of input elements, such as demographic data, lifestyle factors, and psychosocial variables, K-NN may be used to forecast a person's chance of acquiring a mental health illness, such as depression or anxiety. For instance, a study used K-NN to predict a cohort of adolescent participants' likelihood of depression based on their online and social media behaviour. (Islam et al., 2018)

K-NN can be used to identify mental health problems early on. Proactive treatments and support can be given to persons who are at risk based on their resemblance to others with recognised diseases in order to stop the worsening of symptoms or the emergence of more serious problems.

For those with mental health disorders, personalised treatment regimens may be created using predictive models based on k-NN. Healthcare professionals can customise therapies to meet the needs of each patient by taking into account their commonalities.

Although the k-NN classifier has showed potential in predicting mental health, it is not without difficulties. When using k-NN for mental health prediction, researchers need to consider a number of difficulties, including selecting the best value for k, choosing the proper distance measure, and handling unbalanced datasets.

In conclusion, the k-NN classifier is a flexible technique used in mental health prediction to pinpoint those who are at risk of developing a variety of mental health issues. It is an important tool in the field of mental health research because of its capacity to anticipate outcomes based on similarities to nearby data points. To get accurate and therapeutically useful predictions, like with any machine learning strategy, thorough consideration of data quality and model adjustment are important.

2.4 Decision Trees

Decision trees, a flexible and understandable machine learning technique, can be useful in predicting mental health. By offering a clear and organised framework for comprehending the elements and characteristics that affect a person's mental health state, decision trees aid in the prediction of mental health. Decision trees can be helpful in this situation in the following ways:

Decision trees can highlight the significance of many characteristics or factors in predicting outcomes related to mental health. One may figure out which elements have the most bearing on a person's mental health condition by looking at the tree's structure and the locations of particular aspects within it. For instance, it could draw attention to the fact that certain demographic traits or recent life experiences are excellent predictors.

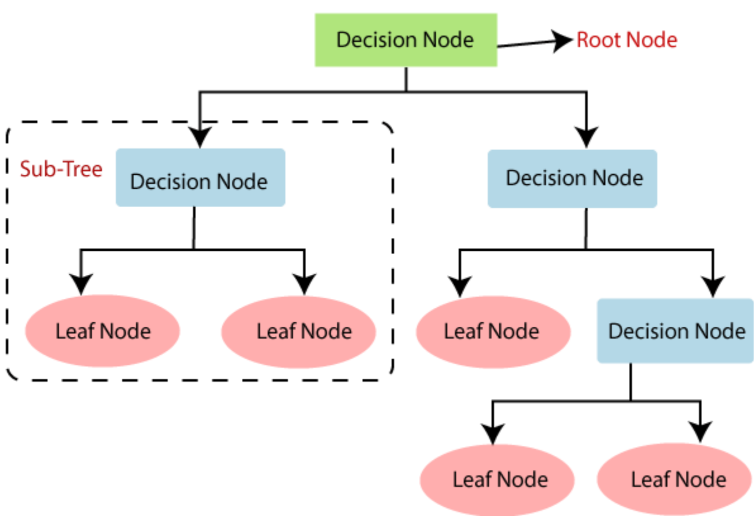


Figure 2.3 Working of Decision model

Decision trees are ideal for clinical and healthcare contexts because of their great interpretability. The decision-making process shown in the tree may be easily followed by healthcare practitioners, who can also quickly comprehend why a specific forecast was made. This openness can encourage communication and cooperation between the prediction model and the practitioners.

Decision trees can recognize risk factors and early indicators of mental health disorders. Healthcare professionals can identify people who have particular combinations of traits that raise their risk of developing mental health difficulties by analyzing the tree structure, enabling early intervention and prevention.

It is possible to utilize decision trees to direct the creation of individualized treatment regimens. The tree can recommend customized actions or therapies for people with recognized risk factors or symptoms based on their particular traits, improving treatment results.

Based on shared traits, decision trees may divide populations into several risk categories. By identifying populations who might need more focused therapies or closer monitoring, this might aid healthcare systems in allocating resources effectively.

A complex interaction of elements, including biological, psychological, social, and environmental influences, affects mental health. By dividing the data into subgroups depending on feature interactions, decision trees are able to capture these intricate linkages. Modeling of nonlinear and interactive effects is made possible by this.

When used in conjunction with ensemble techniques like Random Forests or Gradient Boosting, which increase the resilience of the model and minimize over fitting, decision trees can offer competitive prediction accuracy.

Challenges and Considerations:

- Decision trees can over fit, when the model takes into account data noise. To address this problem, proper hyper parameter trimming and tweaking are crucial.
- They might not always be as successful as other machine learning algorithms at detecting subtle or complicated patterns in the data.
- Decision trees are sensitive to even minor data changes, various tree architectures and predictions may result

In conclusion, decision trees are a useful tool for predicting mental health because of its openness, interpretability, and capacity to spot significant patterns and linkages in the data. Decision trees can assist healthcare practitioners in making knowledgeable choices about early intervention, treatment planning, and resource allocation for people who are at risk of mental health difficulties when used in conjunction with suitable data preparation and validation methodologies.

2.5 Random Forest

In order to predict mental health, Random Forest, an ensemble learning technique made up of several decision trees, might be a useful tool. Here is how Random Forest aids in predicting mental health:

Random Forest combines the predictions of multiple decision trees', which lessens over fitting and improves the resilience and accuracy of the model as a whole. This can produce more accurate forecasts in the field of mental health, where the data may be complicated and noisy.

Numerous characteristics, including demographic, clinical, behavioural, and psychosocial factors, are frequently used to predict mental health. High-dimensional data may be handled by Random Forest in an efficient manner, with essential characteristics being automatically selected and intricate connections between them being captured.

A measure of feature significance provided by Random Forest may be used to determine which elements have a big impact on mental health prediction. Clinicians and academics may find this material useful in understanding the major factors influencing the outcomes of mental health.

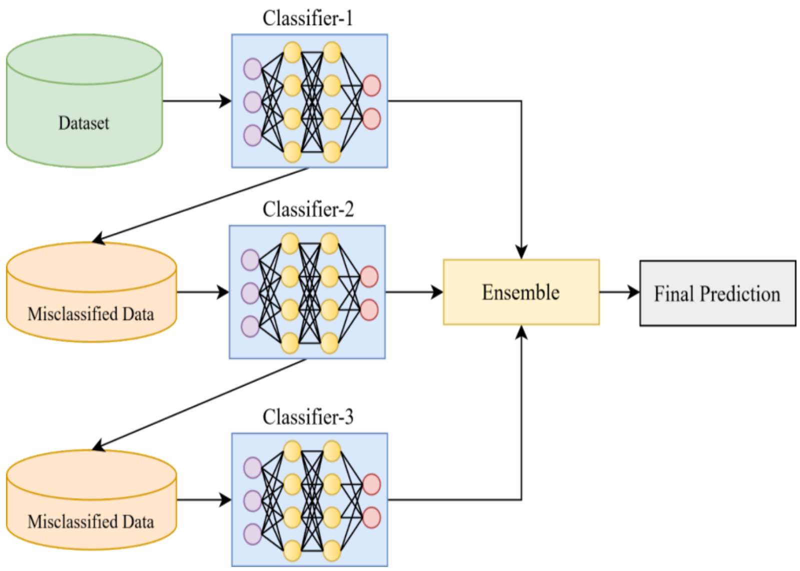


Figure 2.4 Working of Ensemble model

In mental health prediction, unbalanced datasets are typical, where there may be fewer positive instances (people with mental health disorders) than negative cases (people in good health). Unbalanced data may be handled by Random Forest by modifying class weights and improving predictions for the minority class.

Nonlinear connections exist between factors and outcomes when predicting mental health. Since Random Forest is capable of capturing these nonlinearities and intricate relationships between data, it can successfully mimic a variety of mental health issues.

Even if individual decision trees may be understood, the ensemble structure of Random Forest may make the model more difficult to understand. However, tools like partial dependency graphs and feature significance plots may be used to analyse the model's behaviour and forecasts.

Personalised mental health forecasts may be made using Random Forest. It can provide individualised insights and suggestions for early intervention and treatment planning by taking into account a person's particular collection of characteristics.

Out-of-bag (OOB) samples are one of the built-in procedures offered by Random Forest for model validation. This eliminates the requirement for a separate validation dataset and enables a trustworthy estimation of the model's generalisation performance.

Comparing the ensemble technique of Random Forest to individual decision trees, it helps to decrease bias and volatility. This can result in a model that is more accurate and stable.

Random Forest is less prone to overfitting by combining the predictions of many decision trees, which is an important factor when working with noisy mental health data.

The resilience, accuracy, feature significance analysis, and capacity to handle high-dimensional, unbalanced, and complicated data are only a few benefits that Random Forest offers for mental health prediction. Random Forest can lead to more precise and individualised mental health forecasts when used strategically and read correctly, enabling early intervention and better healthcare results.

2.6 Bagging

Bagging, also known as Bootstrap Aggregating, is a potent ensemble learning method that may be used to improve the precision and robustness of predictive models in the context of mental health. Creating models that can accurately detect and categorize people at risk of mental health illnesses based on a variety of characteristics and data points is frequently the aim in the field of mental health.

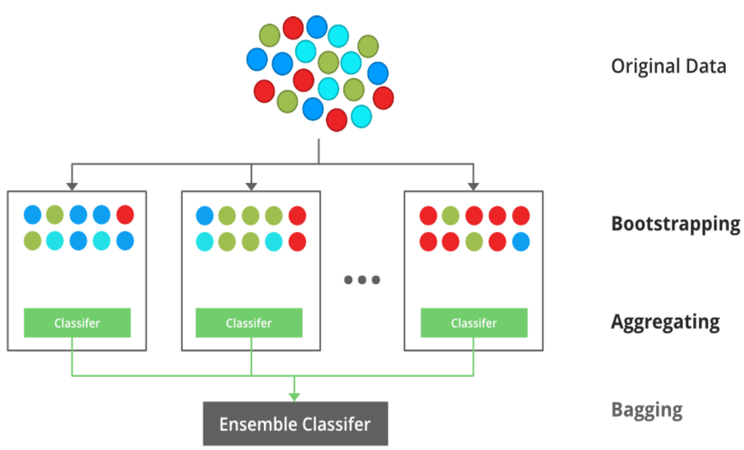


Figure 2.5 Working of Bagging model

This work can benefit from bagging in the following ways:

- Reducing Overfitting: Overfitting in mental health prediction models may be problematic, particularly when handling intricate associations between variables and mental health outcomes. By training many base models on various data subsets, bagging introduces variety into the learning process and helps reduce overfitting.

- Improving Broadness of View: Using bagging, which involves training on a variety of dataset subsets, encourages generalization. Predictive models for mental health must be able to accurately forecast on fresh, untested data. Combining many models into a bag makes the predictor more reliable and applicable to a wider range of situations.
- Handling Imbalanced Data: Class imbalance is a common problem in mental health statistics, where the proportion of people with mental health issues may be much lower than that of those without such illnesses. By resampling with replacement (bootstrap sampling), bagging may manage unbalanced data by making sure that every base model is exposed to several examples of each classes.
- Improving Model Stability: A factor in the model's stability is bagging. Models for predicting mental health issues shouldn't be unduly affected by little changes in the training set. Because bagging is ensemble-based, it lessens the effect of noise and outliers, increasing the stability and dependability of the model.
- Enabling Interpretability: Even though ensemble models are sometimes referred to as "black-box" models, combining several models may show trends or characteristics that are consistent contributors to several base models. This may improve the overall mental health prediction model's interpretability.
- Cross-Validation and Model Evaluation: Using methods like cross-validation, bagging enables robust model assessment. Researchers can get a more precise approximation of the model's performance by training and assessing it on various subsets of the data.

To summarise, the utilisation of bagging approaches in the context of mental health prediction has the potential to greatly enhance the dependability and precision of models, hence aiding in the creation of more potent instruments for detecting individuals who may be susceptible to mental health illnesses. To guarantee the bagging ensemble's success in their particular application, researchers and practitioners should, nonetheless, carefully choose base models, adjust hyperparameters, and properly assess model performance

2.7 Boosting

A potent ensemble learning method called "boosting" is used to improve machine learning models' capacity for prediction. By concentrating on cases that were incorrectly categorised in earlier iterations, boosting produces a powerful learner sequentially as opposed to bagging, which builds numerous models individually and aggregates their results. The main principle underlying boosting is to prioritise mistake repair, thereby enhancing the model's capacity to manage intricate patterns in the data.

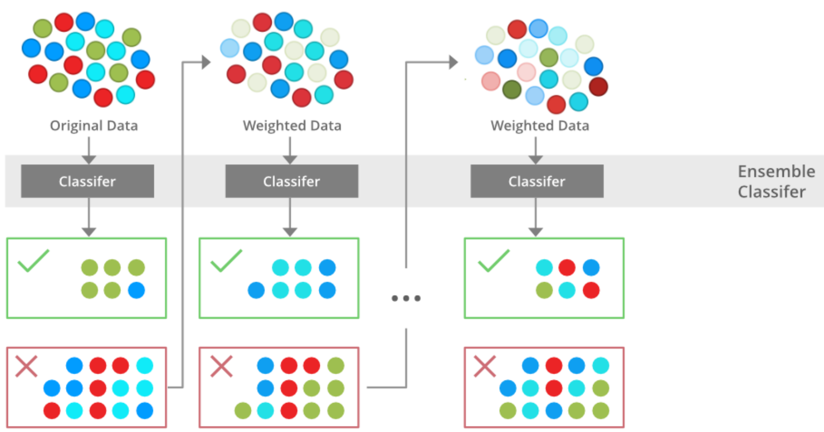


Figure 2.6 Working of Boosting model

The following are some essential traits and advantages of boosting:

1. Sequential Learning: Boosting creates the final model through a sequential process of training each weak learner to fix the errors of its predecessor. Over iterations, boosting algorithms may adjust to the properties of the data thanks to this sequential learning process.
2. Weighted Training: By giving occurrences in the dataset weights, the technique known as "boosting" gives those who were incorrectly categorised in earlier rounds more weight. This weight modification makes the model more efficient in scenarios when the data is noisy or shows complicated relationships by helping it concentrate on difficult cases.
3. Model Adaptability: Boosting algorithms are excellent for jobs where the links between attributes and outcomes are complicated or not immediately obvious since they can adapt

to the intricacies of the data. This flexibility aids in the model's capacity to identify complex patterns.

4. Combining Weak Learners: Boosting usually uses weak learners as building blocks, which are models that perform marginally better than random guessing. By adding these underperforming learners together in a weighted way, boosting builds a robust and precise prediction model.
5. Reducing Bias and Variance: Boosting is iterative in nature and frequently produces models with high prediction accuracy, By reducing bias and variance simultaneously. Because of this, boosting is appropriate for applications that worry about either overfitting or underfitting.
6. Feature Importance: Boosting algorithms shed light on the relative significance of various factors in prediction-making. The contributing components in the dataset may be better understood with the help of this feature significance analysis.
7. Ensemble of Diverse Models: Boosting produces a collection of unique models, each of which adds to the predictive power of the whole. By emphasising misclassified occurrences and using a sequential training method, variety is achieved.

Boosting has been effectively used in many other fields, such as regression, classification, and, as was previously noted, mental health prediction. Boosting is a useful technique for enhancing the precision and resilience of machine learning models because of its capacity to adjust to the complexities of the input and gradually improve the model.

2.8 Stacking

A meta-model, often referred to as a blender or meta-classifier, is also called as stacking is an ensemble learning approach, to combine the predictions of several individual models, or base models, that have been trained separately. Comparing this technique to individual models helps improve the forecasting performance overall. When many models identify various patterns in the data, stacking is especially helpful since combining their strengths can provide forecasts that are more accurate.

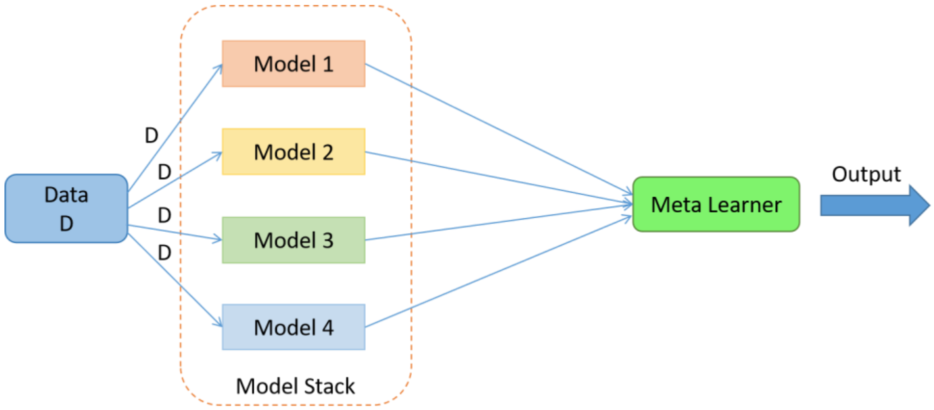


Figure 2.7 Working of stacking model

In the context of mental health prediction, stacking can be beneficial in several ways:

1. Model Diversity: The information pertaining to mental health can be complicated, and several models may be better than others at identifying particular features or trends in the data. Stacking makes use of the advantages of each base model by mixing a variety of them, which may increase the forecast accuracy as a whole.
2. Handling Heterogeneous Data: Numerous characteristics from various sources, including questionnaires, physiological assessments, and clinical records, are frequently included in mental health statistics. By combining multiple models that may be more adept at managing distinct feature sets, stacking strengthens the model's resistance to the heterogeneity of the data.
3. Improved Generalization: The goal of stacking is to provide a meta-model with good generalisation to novel, unobserved data. This is important in the field of mental health prediction, where the objective is frequently to create models that can precisely identify people who are at risk based on their particular traits.
4. Combining Strengths of Models: If certain models are better at capturing short-term patterns while others excel at long-term trends, stacking allows for the combination of these strengths. This can be valuable in mental health prediction, where both immediate and long-term factors may contribute to outcomes.
5. Model Interpretability: When applied to interpretable base models, stacking can reveal which patterns and characteristics are consistently important in several models. In

applications related to mental health, where it's critical to comprehend the variables influencing forecasts, interpretability is crucial.

6. Ensemble Learning Advantages: The benefits of ensemble learning, including less overfitting, enhanced stability, and superior performance over individual models, are carried over to stacking. This is especially helpful for mental health prediction, where model dependability and accuracy are crucial.

You may train many base models on the training data, utilise their predictions as input features for a meta-model, and then train the meta-model on the same dataset to use stacking for mental health prediction. Logistic regression, decision trees, and even more sophisticated models like neural networks are common candidates for meta-models.

It's crucial to remember that, even though stacking may be an effective strategy, the choice of base models, their variety, and the meta-model should all be carefully considered. To guarantee the stacking model's robustness and generalizability in mental health prediction situations, appropriate model assessment and validation approaches, including cross-validation, should be used.

2.9 Studies on Mental Health

Table 2.1 Studies done on Mental Health Prediction

Research Paper Title	Year	Methodology	Research Gap
Prediction of mental health treatment adherence using ML algo’s (Bajaj et al., 2023)	2023	In this research paper, they have worked on various ML algo’s to build a predictive model & the evaluated the model using accuracy & precision	There are many limitations on this study. The generalizability of the model may be impacted by intrinsic biases in the dataset. Predictions may be impacted by missing numbers and poor data quality.
Machine Learning for Mental Health Prediction in Social Media Data (Kim et al., 2022)	2022	This study used natural language processing (NLP) techniques to analyze social media posts and applied machine learning algorithms for mental health prediction.	Social media postings including biassed data, mental health labels without objective information, and user privacy ethics raised.
Machine Learning Techniques for Prediction of Mental Health (Jain et al., 2021)	2021	This paper provides a comprehensive review of various machine learning techniques and their applications in mental health prediction.	No original data collection or experimentation, potential bias in selected studies.
Deep Learning for Mental Health Diagnosis with Brain Imaging Data(Jain et al., 2021)	2021	This research paper has ethical problems about user privacy, data bias in social media posts, and the absence of ground truth labeling for mental health.	Large and varied brain imaging datasets are rare, and deep learning models have interpretability issues.
External validation	2023	A Smartphone app that	Ethical concerns about user

of prediction models for patient-reported outcome measurements collected using the selfback mobile app (Verma et al., 2023)		gathered user behavior data and using predictive modeling to identify signs of depression was created by researchers.	consent and data privacy, potential user bias.
A Survey on Wearable Sensors for Mental Health Monitoring (Verma et al., 2023)	2023	Wearable sensors were employed to track behavioural and physiological data, and machine learning techniques were utilised to forecast mental health.	Limited long-term data collection, potential discomfort or users wearing sensors.
Prediction of Public Mental Health by using Ml Algo's (Saibaba et al., 2022)	2022	This paper involves a very retrospective approach towards, having data of individuals with different mental health records and performed preprocessing and data analysis steps. Using basic pre build machine learning algorithms and tested for predictions. Performance was evaluated using different metrics like accuracy & F1 score.	The study is limited because it is completely dependent on electronic health records and the data is collected from very small population. There is a lack of long term follow up which is constraint for early prediction.
Predicting Adolescent Mental Health Outcomes Across Cultures: A ML Approach	2023	Adolescent mental health outcomes were predicted over time using statistical modeling and longitudinal data gathering.	Predictions are difficult due to the dynamic nature of mental health and the resource-intensive nature of long-term data collecting.

(Verma et al., 2023)			
A Study on Sentiment Analysis of Mental Illness Using ML Techniques (Tiwari et al., 2021)	2022	Social network analysis and sentiment analysis of social media postings were employed to forecast changes in mental health.	Data bias, social media posts without context, and difficulties determining causality.
Comparative Analysis of ML Techniques for Mental Health Prediction (Tiwari et al., 2021)	2021	In this research paper, they have worked on various ML algo's to build a predictive model & the evaluated the model using accuracy & precision	There are many limitations on this study. The generalizability of the model may be impacted by intrinsic biases in the dataset. Predictions may be impacted by missing numbers and poor data quality.
Enhancing Mental Illness Prediction using Tree based Machine Learning Approach (Tiwari et al., 2021)	2021	The study improved mental disease prediction by using a tree-based machine learning technique. It entailed gathering information from a variety of sources, such as biomarkers, self-reports, and electronic health records. To build the model, gradient boosting, decision trees, and random forests were used.	Problems with electronic health record data quality. Sample variety may restrict generalizability. Model hyperparameters and feature selection may have an impact on predictions. This research did not explore patterns in long-term mental health.
Mental Health Prediction Using ML: Taxonomy, Applications, and Challenges (Chung & Teo,	2022	In order to create a taxonomy of machine learning techniques for mental health prediction, this study surveyed the body of literature already in existence.	It lacks the qualities of empirical research and original facts. The quality and usability of the included research determine how accurate the taxonomy and classification

2022)		It categorised a range of uses, such as stress detection, mood disorders, and suicide risk. The standard procedures are part of the approach.	will be. The ever-evolving field of mental health and the swift advancement of machine learning methodologies might provide obstacles to maintaining the taxonomy current.
Prediction of Mental Health in Medical Workers During COVID-19 Based on ML (Wang et al., 2021)	2021	This study employed ML to predict mental health impacts of COVID-19 on medical workers . Demographics, psychological evaluations, and self-reported questionnaires were used in the data gathering process. The training and testing stages used predefined models. Model correctness was measured via performance measures.	Limitations include the possibility of self-reporting bias in survey data and difficulties in gathering data. The predictions made by the study might only apply to healthcare professionals during the COVID-19 pandemic, which would restrict their applicability to other demographics.
New Analytic Framework of Public Mental Health Prediction Using Data Science (Marrapu et al., 2022)	2022	The study presents data science-based public mental health prediction. Data is compiled from a variety of sources, including social media, electronic health records, and ambient circumstances. The process of constructing a model involves combining statistics and machine learning techniques.	Real-world data and empirical validation are absent from it. Prediction accuracy depends on the availability, quality, and efficacy of the selected data sources. Careful thought must be given to ethical issues related to bias in data selection, privacy, and data consent. Predictive modelling may provide implementation

		The framework was created to provide a thorough forecast of general population changes in mental health.	issues when combining numerous data streams.
Mental Health Prediction through Text Chat Conversations (Nouman, 2023)	2023	It uses text chat to forecast mental health issues. Preprocessed text data was gathered from chat sites for mental health. To extract features NLP, sentiment analysis, and topic modeling were used. Based on conversation, pre build ML algo's were used to forecast mental health conditions.	One of the limitations is that self-reported text data may contain misrepresentations since participants might not completely reveal their mental health state. The chat platform and user demographics may have an impact on how generalizable a model is.
Virtual Reality Data for Predicting Mental Health Conditions (Playne, 2022)	2022	This study explored the use of virtual reality data to predict mental health conditions. Data were collected through VR simulations and interactions. ML algos, including NN were used to build predictive models. Model performance metrics assessed accuracy.	One of its drawbacks is that VR simulations may seem more artificial. Those who are at comfortable with virtual reality technology may find it generalizable. Difficulties include the requirement for vast and diverse VR datasets for reliable forecasts.
Classification Algorithms based Mental Health Prediction using Data Mining (Laijawala et al.,	2020	The research paper employed classification algorithms to predict mental health conditions using data mining techniques. Psychological assessments and surveys were	It's possible that the study underrepresents the complexity of mental health issues. Particularly in the context of mental health research, ethical issues related

2020)		used in the data gathering process. We trained using pre build models. The accuracy of the models was evaluated using performance metrics .	to informed permission and data protection are crucial. The selection of classification methods and the demographics of the sample may have an impact on generalizability.

2.10 Discussion

Because of the many benefits that ensemble learning techniques specifically, stacking models offer over independent machine learning models, their application in the field of mental health prediction has gained traction. Data on mental health is complex and varied, and stacking is excellent at capturing this. The capacity of the ensemble to incorporate many base models, each of which may be particularly good at analysing different parts of the data, adds to a more thorough comprehension of the variables affecting mental health outcomes. Because stacking models are able to combine the best features of several base models into a single, more accurate prediction framework, they perform superior than other machine learning techniques. In contrast to standalone models, stacking makes use of the unique advantages of each component by combining the predictions of several models into a meta-model. This method works particularly well in domains where several factors influence outcomes, such as mental health prediction, since it captures complicated interactions within the data. Stacking is distinguished by its inherent interpretability through the integration of transparent base models, its flexibility to accommodate heterogeneous data sources, and its ability to enhance generalisation by learning from many model viewpoints. Moreover, the benefits of ensemble learning, such as less overfitting and increased stability, add to the higher performance of stacking models. These qualities are crucial for predicting mental health, as model accuracy and dependability are critical. To sum up, the use of stacking models in mental health prediction is a calculated decision that takes advantage of ensemble learning's advantages to help handle the complexity prevalent in mental health datasets. In order to assure the robustness and generalizability of the stacking strategy in mental health prediction applications, the discussion focused on the necessity of rigorous assessment procedures, intelligent feature engineering, and a broad selection of base models.

2.11 Summary

This chapter provides a thorough overview of modelling for mental health prediction. The linear and non-linear prediction frameworks, such as KNN, Logistic regression and different types of ensemble models are reviewed in this chapter. The benefits of a ensemble approach over stand-alone methods are discussed in this chapter. In terms of model creation and execution, this chapter covers in depth different types of models studied and why ensemble models are better than other linear models and why stacking is used over other models.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, we will study the use of machine learning to identify mental health condition and its early prediction. It goes over the theoretical foundations and research methodology. It aims to explain the scientific jargons used throughout the research while providing a comprehensive explanation of the techniques. The chapter begins with data selection, where a comprehensive dataset comprising instances of mental health is chosen. After that, the chapter discusses data pre-processing techniques such exploratory data analysis, missing value treatment, feature scaling and train-test separation. The importance of data visualisation is highlighted, and several scaling techniques are examined. The chapter concludes with a discussion of the benefits of the proposed machine learning models created specifically for the early identification of mental health problems. The performance metrics portion concludes in a comprehensive examination of error metrics, which evaluates the accuracy and efficacy of ML models in detecting mental health condition.

3.2 Research Approach

Research techniques in the field of mental health prediction comprise many systematic procedures intended to predict mental health outcomes. These methods seek to find trends and risk factors, evaluate and analyse pertinent data, and produce useful insights for prompt response. This research project is essential in expanding our knowledge of the dynamics of mental health and opening doors to more potent preventive and therapeutic approaches that can enhance the wellbeing of both people and communities.

The study technique for applying machine learning to detect mental health includes several important steps, such as data collection, pre-processing, building and deploying models, and evaluating them. These processes aim to do data analysis and develop the models specifically designed for the identification of mental health. The entire study process is shown in Figure 3.1.

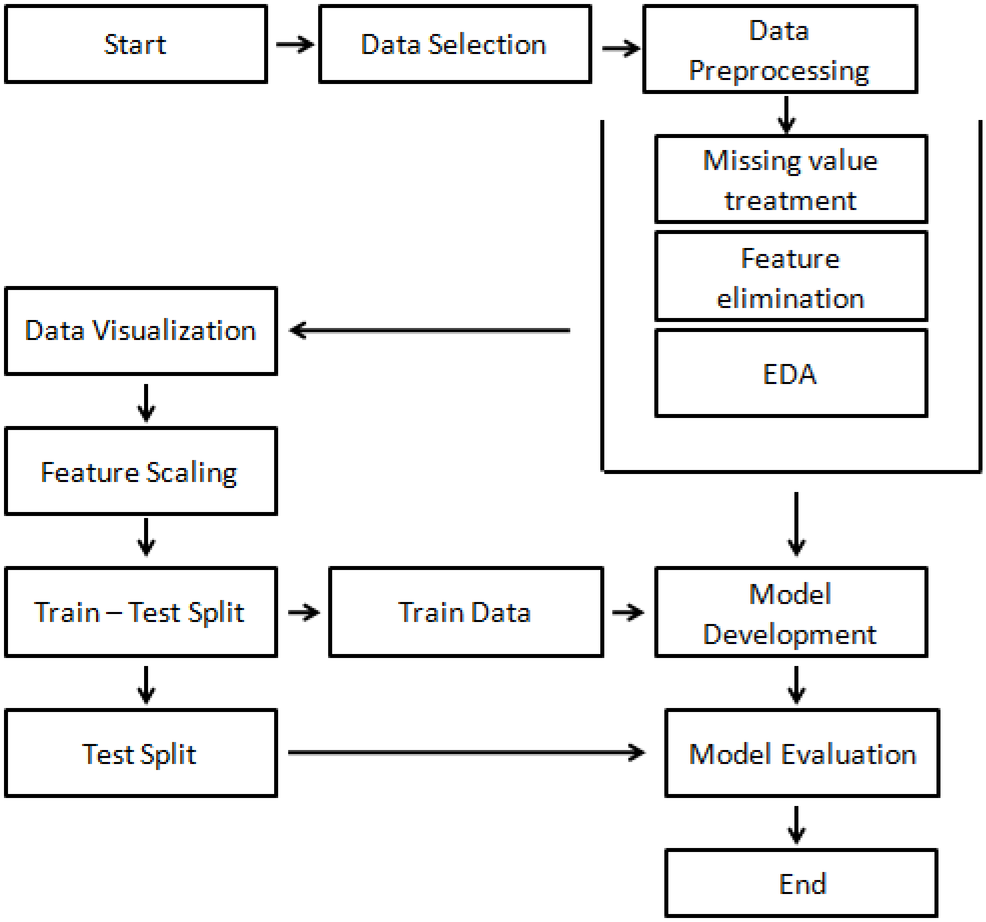


Figure 3.1 Research methodology plan

3.2.1 Data Selection

Choosing appropriate data for mental health prediction is essential for creating precise and functional predictive models. It entails compiling a range of representative and diverse datasets covering a range of symptoms, behavioral patterns, and demographic characteristics. These databases have to be sizable enough to encompass the diversity and intricacy of mental health issues. The data collected for building the model is collected from a survey where a working individual is asked to fill all the details regarding his professional and personal life and as well as covering some lifestyle questions. In the end, the foundation for developing strong prediction models that might help with early identification and intervention for mental health concerns is a well chosen dataset. We have chosen the data available on Kaggle under the same of survey which contains around 1200 + surveys of people collected over a period of time.

3.2.2 Data Preprocessing

Data preprocessing is a crucial step in the data analysis and machine learning pipeline. Numbers of tasks involved are cleaning, transforming and preparing the raw data for analysis and model training. These activities can involve feature selection to identify the most pertinent qualities, data normalization or scaling to bring variables to a consistent range, feature engineering to produce new, meaningful variables from the current data, and data cleaning to address mistakes, outliers, and missing values. In order to evaluate a model, data preparation may also include encoding categorical variables, handling unbalanced datasets, and dividing the data into training and testing sets. Because it may greatly affect the precision and efficacy of analytical or machine learning models, proper data preparation is crucial to ensure that these models operate with high-quality, trustworthy data in order to provide predictions or judgments that are well-informed.

3.2.2.1 Missing Value Treatment

During data pre-processing, any missing values, noise, and other irregularities in the raw data must be removed to prevent inconsistent findings. There may or may not be missing values in the raw or real-world data. Data corruption or unavailability might be the reason for missing values. Since many machine learning algorithms do not allow missing values or because the missing values may skew the model output, treating missing values is an extremely important task in data pre-processing. One of the two methods for treating missing data is either imputing the missing values or removing them. The mean, median, mode, and/or any arbitrary value that matches the domain can be used to impute the missing data. The sci-kit-learn library may also be used to infer the missing data. It is necessary to identify and handle the missing values in the chosen dataset appropriately.

3.2.2.2 Feature Elimination

A number of machine learning approaches are vulnerable to the dimensionality curse. The exponential computing effort needed to analyse and interpret high dimensional data is referred to as the "curse of dimensionality." Feature elimination provides a practical solution to problems with the curse of dimensionality, overfitting, computational time, and learning accuracy. Finding the key elements or variables needed to get the intended outcomes is essential when pre-processing data. Feature removal can be done automatically by an algorithm or by hand.

Recursive Feature Elimination (RFE) is one such method. Until the desired number of features is obtained, the weakest features are eliminated using the regression feature extraction (RFE) feature selection approach. Feature elimination will be used to guarantee the key features needed for the study's objectives.

3.2.2.3 Exploratory Data Analysis

A method for analysing data that makes use of visual aids is called exploratory data analysis (EDA). Utilising graphical representations and statistical summaries, it is used to find trends and patterns as well as outliers. Since the quality of the input data used to train the model determines its final result, EDA accounts for around 40–50% of the project's entire duration. Both univariate and bivariate analysis are used in EDA.

The most basic type of data analysis is called univariate analysis. Analysing each feature independently is necessary for univariate analysis. It may be descriptive or inferential. The descriptive statistics of each feature are provided by a descriptive Univariate analysis, which can aid in the identification of data outliers. An observation that deviates significantly from the rest of the data numerically is called an outlier. In specifically, boxplots are used to find outliers in a dataset. A boxplot provides a clear picture of the distribution of the data. Through its quartiles, a boxplot may visually represent the localization, dispersion, and skewness of the numerical data. At the median, a vertical line passes through the box.

A statistical technique known as bivariate analysis involves analysing two variables to ascertain their empirical connection. When it comes to mental health prediction, it evaluates the relationships between variables such as stress and depression. This helps with early identification and focused interventions for those who are at risk, improving general wellbeing and lessening the stigma attached to mental health problems.

3.2.2.4 Data Visualization

In the fields of machine learning and data analysis, data visualisation techniques are developing at an exponential rate. These methods improve comprehension of the data and offer an up-close look at the trend the data is following. The goal of data visualisation techniques is to improve understanding of data presented in graphical forms including maps, graphs, and charts. Trends, outliers, and patterns in data may be easily seen and identified with the use of data visualisation tools. Matplotlib and Seaborn are two of the most widely used libraries for data visualisation.

Seaborn is a Python data visualisation toolkit, based on matplotlib. It offers a sophisticated drawing tool for creating eye-catching and educational statistics visuals. It offers lovely colour schemes and default styles to enhance the visual appeal of statistics charts. Seaborn wants to make data exploration and comprehension revolve on visualisation.

3.2.2.5 Feature Scaling

Regardless of the unit of measurement, a machine learning algorithm generally weighs larger values as higher and smaller ones as lower. To manage the greatly fluctuating magnitudes or values of the characteristics, scaling techniques are applied. Techniques for scaling are frequently applied before the train-test split process. Standardisation and Min-Max Normalisation are two popular methods for feature scaling.

One of the most used techniques for normalising data is min-max normalisation. For each characteristic, the lowest value is converted to zero, the greatest value to one, and the remaining values are turned into a decimal number ranging from zero to one. The following is the formulation for Min-Max scaling.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

The process of scaling characteristics such that their mean becomes zero and their standard deviation becomes one is known as standardisation. Values undergo standardisation, which centers them around the mean and assigns a unit standard deviation. The following is the phrase for standardisation.

$$x_{scaled} = \frac{x - \mu}{\sigma}$$

3.2.2.6 Train-Test Split

A method for assessing a machine learning algorithm's performance is called train-test split. It is a quick and simple strategy that works with any supervised learning algorithm. Despite being easy to use and understand, there are instances in which the approach shouldn't be applied, such as when there is an imbalance in the classes or when the dataset is limited. When dealing with huge datasets, the train-test split technique makes sense. The dataset is divided into two subsets as part of the process. The first subset, known as the training dataset, is utilised to fit or train the model. The second subset, known as the test dataset, is supplied as input to the model.

3.2.3 Model Development

The methodical process of developing machine learning models to address certain issues is known as model development. When it comes to mental health prediction, this entails identifying the issue, gathering and preprocessing representative and diverse data, picking features, deciding on suitable machine learning algorithms, adjusting hyperparameters for best results, and validating the model with metrics that take into account the particularities of mental health applications. Important components that guarantee the model is not only accurate but also transparent and objective are interpretability and ethical concerns. The development process includes ongoing monitoring and enhancement, which is indicative of the dynamic character of predictive models and the changing terrain of data and research on mental health.

When developing a model, especially one that predicts mental health, sophisticated methods like the stacking model must be applied strategically in order to improve predicted accuracy and dependability. The procedure starts with the creation of the prediction task and a thorough grasp of the mental health area. After then, data is gathered from a variety of sources, including questionnaires, medical records, and physiological assessments. Steps in preprocessing take care of feature normalisation, class imbalances, and missing values. It takes feature engineering to identify and modify pertinent factors in order to represent the complex relationships between variables and mental health outcomes.

Selecting the right models is crucial, and the stacking model works well in this situation. By employing a meta-model, stacking leverages the benefits of several basic models—including decision trees, support vector machines, and neural networks—to create a strong prediction framework. Hyperparameter tuning maximises the effectiveness of each model by optimising its performance both individually and collectively.

Fairness and bias reduction are two ethical factors that are integrated into the model creation process. The stacking model's ongoing evaluation and refinement highlight how flexible it is to changing data dynamics and mental health research. In the end, stacking-based model building is a complex method that uses ensemble learning to overcome the difficulties in mental health prediction, offering a more nuanced understanding and precise identification of those who are at risk.

3.2.4 Model Evaluation

In order to guarantee the accuracy and dependability of predictive models, model evaluation for mental health prediction is a crucial step. Prioritising evaluation metrics that are in line with the particular objectives of mental health applications is essential, even above conventional metrics like accuracy. Aspects like precision, ROC, false positive rate are especially important when taking into account the possible effects of false positives and false negatives on people's health. Comprehensive cross-validation should also be a part of the assessment process in order to gauge the model's resilience to various data subsets. The predictability and interpretability of the model are crucial as it's critical to comprehend the variables affecting mental health outcomes.

Fairness and bias are two ethical issues that need to be carefully considered in order to prevent unforeseen outcomes or differences in the model's performance across various demographic groups. All things considered, a thorough model review approach guarantees that mental health prediction models are not only morally acceptable and reliable, but also have the potential to significantly improve people's quality of life.

3.3 Summary

This chapter gives a thorough description of the procedures involved, together with an outline of the research strategy that was used and the theoretical framework that supported it. This chapter uses a flow chart to demonstrate the research methodology used. The choice of survey dataset and the necessary data pre-processing procedures are covered in this chapter. How missing values have been treated and how features are selected after feature engineering are discussed in this chapter. It offers insights into the procedures that go into developing and assessing models. The two widely used data visualisation libraries, Matplotlib and Seaborn, are highlighted in this chapter. The performance measures which we have used to compare the models like accuracy, precision, ad ROC have been discussed in this chapter.

CHAPTER 4

ANALYSIS AND DESIGN

4.1 Introduction

This chapter covers the crucial steps in data pre-processing and model-designing for our ML-based mental health prediction model. This chapter has several sub-chapters, including data preparation, exploratory data analysis, data visualisation, model design, and model implementation. The first section of the chapter provides a detailed explanation of the dataset used in the research, which was specifically selected to document instances of mental health conditions. After then, pre-processing is done on the data to ensure its accuracy and suitability for further study. The dataset may be improved for modelling with the use of these techniques, which include feature engineering, train-test split, and feature elimination to understand the distribution of characteristics in the dataset. The chapter highlights how important it is to show the features using popular plotting programmes like Matplotlib and Seaborn in order to properly understand the dataset's properties. This thorough data pre-processing and exploratory analysis lays the groundwork for later model construction and execution, enabling precise identification of early signs of mental health prediction when using machine learning techniques.

4.2 Dataset Description

The dataset used in the study is available publicly and is obtained from Github. The dataset comprises of survey data i.e. data collected from people of different countries, age group and background. It has 1259 entries with 27 columns. A variety of behavioural, workplace-related and demographic characteristics are included in the dataset. The collection includes data on employee's mental health as well as a number of characteristics that may be useful in forecasting their mental health state. The goal is to create machine learning models for mental health prediction using this data.

4.3 Dataset Preparation

The data can be obtained from Github in a csv file. Pre-processing the raw data before to do train-test split is recommended practise even when the collected raw data is in structured format. Univariate and bivariate analysis, feature engineering, missing value treatment, and feature deletion are a few of the data pre-processing approaches.

4.3.1 Data Loading

Pandas is a Python library which simplifies the process of loading data into Dataframes. We can easily import the data into a DataFrame from CSV file using the read_csv function. We can even customise the parameters like delimiter, header, and data kinds in addition to providing the file path of the data source. Pandas offer a robust framework for data manipulation, analysis, and visualisation. It facilitates data-driven decision-making by streamlining processes including statistical analysis, data cleansing, and exploration.

4.3.2 Data Exploration

After exploring the data we saw that survey data contains a variety of behavioural, workplace-related and demographic characteristics of people of different countries. There are few columns in the data with missing values and categorical data as well. Further missing value treatment , feature elimination, one hot encoding and processes like binning have been performed on the data to prepare the data for modelling.

4.3.3 Missing Value Treatment

There may or may not be missing values in the raw or real-world data. Data corruption or unavailability might be the reason for missing values. Since many machine learning algorithms do not allow missing values or because the missing values may skew the model output, treating missing values is an extremely important task in data pre-processing.

We have 4 columns which have missing values in it. Comments and state have missing values greater than 30% and it won't be a right choice to treat this columns so we will drop these 2 columns and remaining columns i.e. work interface and self_employemt have a very small percent of missing values i.e. 20% so we consider treating the missing values. Both the columns have categorical data and both of them were treated using the majorty value because of a great difference in the data.

4.3.4 Data Visualisation

Data visualisation is a graphical way of representing the data and knowledge. Charts, maps, and graphs are examples of visual components that may be used in data visualisation. A simple approach to observe and identify patterns, trends, and outliers in data is through the use of data visualisation elements.

Furthermore, data visualisation may reveal hidden trends and abnormalities in data, assisting in the identification of problems or opportunities that could have gone missed otherwise. It helps with data-driven decision-making and the creation of engaging narratives for efficient information delivery.

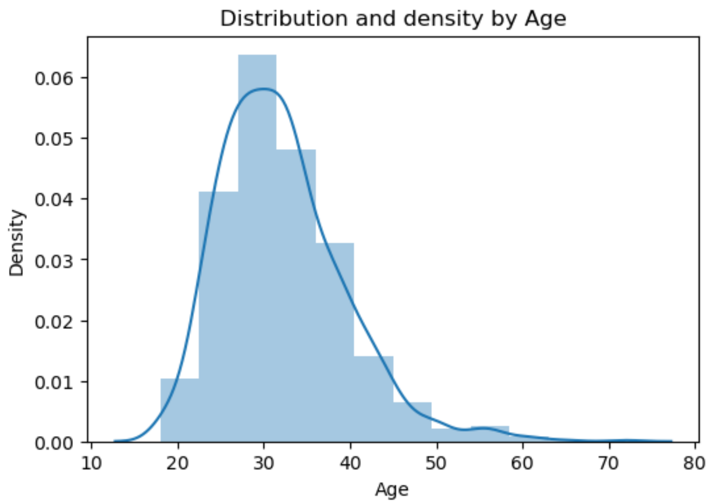


Figure 4.1 Data Visualisation : Distribution of Density by age

The above distplot show the graph distribution of density by age which shows that the maximum number of people fall under the 25 to 35 age category.

4.3.4.1 Univariate Analysis

A basic statistical method for analysing single variables can be done using univariate analysis. Using this approach, the properties and distribution of each variable are uncovered one at a time. It entails presenting the data using graphical tools like histograms, box plots, and bar charts in addition to summarising the data using metrics like mean, median, mode, and standard deviation.

Researchers and analysts can learn more about the central trends and variability of a single variable by using univariate analysis, which is an essential initial stage in the data exploration process. In order to establish the foundation and comprehend the fundamental characteristics of every variable in a dataset, univariate analysis is still necessary.

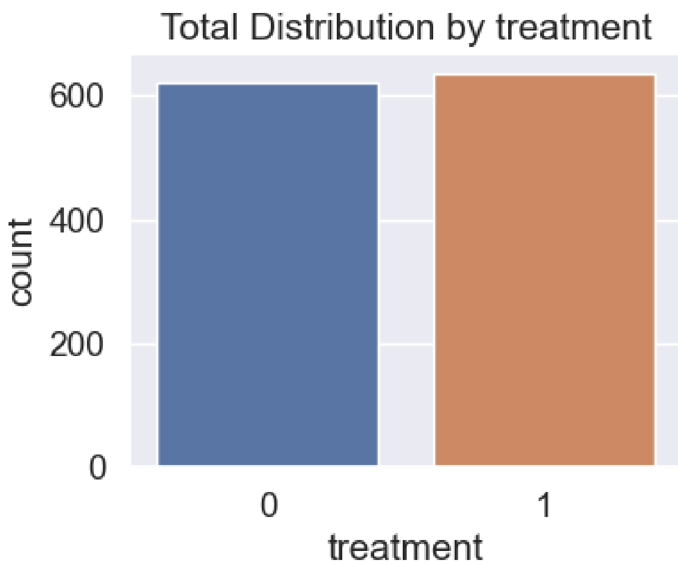


Figure 4.2 Univariate Analysis: Total Distribution by Treatment

This graph shows the distribution of treatment variable using a count plot where it shows that data is not imbalanced. We have almost equal count of people who require treatment and don’t require treatment.

4.3.4.2 Bivariate Analysis

A statistical technique called bivariate analysis is used to examine the connection between two variables in a dataset. Researchers and analysts can examine the relationship between changes in one variable and changes in another by using this method. Bivariate analysis uses a number of methods, including cross-tabulations, scatter plots, and correlation coefficients, to find patterns, trends, and dependencies between the two variables. It is a crucial tool that supports the identification of cause-and-effect links, greater comprehension of the data, and informed decision-making. The basis for more intricate multivariate studies, which examine the interactions between several variables, is laid by bivariate analysis

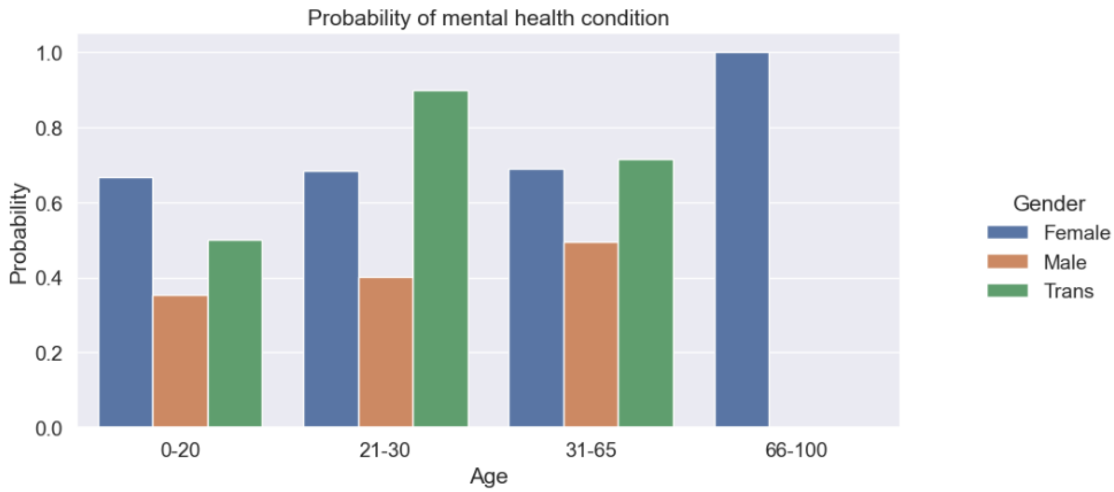


Figure 4.4 Bivariate Analysis: Probability of Mental health condition

This graph is basically a representation of age column with respect to gender plotting the probability that who is more prone to mental health problems. This graph is a clear representation of females being more prone to mental health conditions with respect to other genders. Whereas Trans of age in between 21-30 are more prone to mental health disorders.

4.3.4.3 Multivariate Analysis

A statistical method for concurrently examining the correlations between several variables is called multivariate analysis. It extends beyond bivariate analysis, which looks at the interactions between pairs of variables, and univariate analysis, which looks at each variable separately. Multivariate analysis provides a more thorough understanding of complicated data patterns by examining the interactions and influences between several variables. Through the use of multivariate analysis, researchers may generate more accurate predictions, discover important elements, and derive nuanced insights that improve decision-making in a variety of applications by revealing hidden patterns, dependencies, and correlations within datasets.

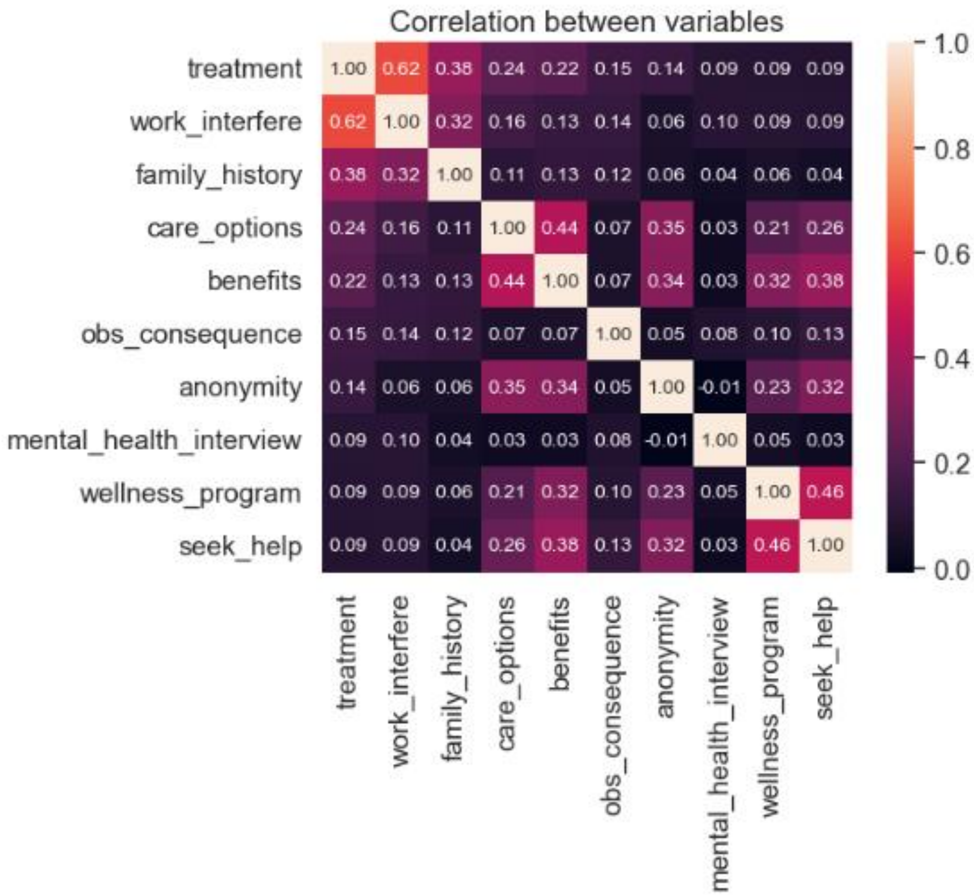


Figure 4.5 Multivariate Analysis: Correlation Matrix

This graph is called as heat map which basically gives the correlation between two variables. The higher the value between 2 variables the higher is the correlation between them.

4.3.5 Feature Elimination

A critical stage in machine learning and data analysis is feature removal, which is identifying the most relevant and instructive traits and removing the less useful ones. This procedure is crucial for increasing interpretability, decreasing overfitting, and boosting model performance. Techniques like univariate feature selection, recursive feature elimination, and feature significance ranking are examples of feature elimination approaches. We need to optimise the model and improve decision-making and prediction performance by selectively selecting which information to retain and reject. This process streamlines the models and increases their accuracy and efficiency.

After looking at the correlation between the columns and on understanding the importance of features we have dropped few columns like timestamp, employment type, wellness program and others which were not important for training the model or have high correlativity value.

4.3.6 Feature Engineering

A crucial component of data analysis and machine learning is feature engineering, which involves feature development and modification to improve model performance and data-driven insights. It seeks to enhance the model's capacity to identify patterns and correlations in the data, extract pertinent information, and minimise noise. Feature engineering encompasses a range of methods, including text pre-processing, scaling, one-hot encoding, and the development of new features through the combination or modification of pre-existing ones.

We did a lot of feature engineering as our data consists of lot of columns having categorical data and as we know that machine learning models accept numerical data only so we encoded the columns in such a way such that the columns are converted into integer data type. We even did the binning of columns example we binned the age column into 4 categories.

Example:

We have a column in our dataset called work_interface which have 5 different types of values:

Don't know, Never, Often, Rarely, Sometimes

Which are encoded into integer data type having values in range from 1 to 5. In similar way we have treated a lot of columns according to the requirement.

4.3.7 Feature Scaling

A key data preparation method in statistics and machine learning is feature scaling. In order to guarantee that every variable is on the same scale, it makes sure to standardise or normalise the range of numerical characteristics. This procedure is essential because the performance of many machine learning algorithms can be impacted by scale differences, as many algorithms are sensitive to the size of the features. Typical feature scaling techniques are standardisation, which centres the data around a mean of zero and a standard deviation of one, and Min-Max scaling, which transforms values to fall inside a given range. By scaling features, we may make models are more accurate and stable by accelerating their convergence, preventing the dominance of variables with bigger scales, and ensuring that each feature's contribution is proportionate.

In our dataset we have done standardisation on the data o the entire dataset so as to avoid any chances of inaccurate results in future.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

4.3.8 Train-Test Split

A key method in data analysis and machine learning for evaluating the effectiveness of prediction models is the train-test split. To train a model, a dataset is divided into two subsets: the training set and the testing set. The testing set is used to assess how well the model performs and how well it generalises to new, unknown data. Usually, the split uses a predetermined ratio—for example, 70% of the data for training and 30% for testing. By dividing the data, we can make sure that the model's performance is assessed on independent data, which aids in the detection of overfitting—a situation in which the model fits the training data too closely but performs poorly in generalization—and offers a trustworthy approximation of the model's expected performance in practical applications. A crucial stage in the creation and validation of a model is the train-test split.

4.4 Model Development and Implementation

A methodical approach is taken in the creation and application of mental health prediction models in order to provide a reliable and morally sound predicting instrument. Starting with a deep comprehension of mental health data, which is subsequently preprocessed to resolve problems such as missing values and unequal class distribution. Relevant patterns are extracted from the data using feature engineering. Through hyperparameter tweaking, appropriate machine learning models that are customised for mental health applications are chosen and made better. To ensure the proposed model's efficacy and dependability, it is carefully verified using measures unique to mental health. During the implementation phase, the model is smoothly incorporated into real-world applications, with an emphasis on ethical and interpretable issues. The model's performance is continuously monitored, enabling fast updates that guarantee its flexibility to changing mental health environments and offer a useful tool for early identification and intervention.

4.4.1 Logistic Regression

Logistic regression is a popular statistical technique for predicting mental health because of its efficacy, simplicity, and interpretability. In this case, logistic regression uses input feature-based modelling, we have passed 8 features with the size of training data to be 879 data points to predict the likelihood of a binary outcome, such the existence or absence of a mental health condition. It works especially well when there is a linear link between variables and outcomes or when interpretability is important.

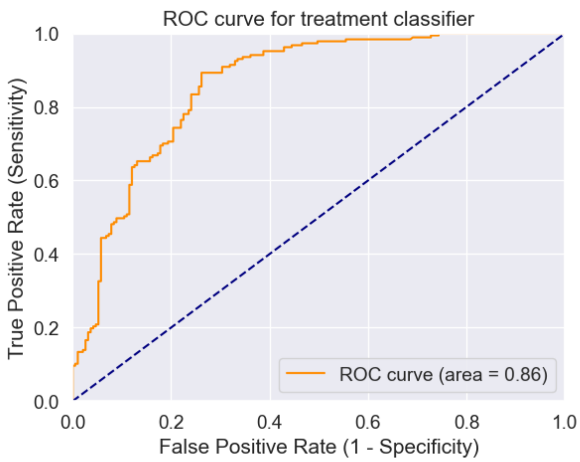


Figure 4.5 ROC curve for Logistic Regression model

Key contributors may be identified with the use of logistic regression, which offers insights into the impact of particular factors on mental health outcomes. Because of its simplicity, it may be used in mental health applications and offers a compromise between interpretability and prediction accuracy.

Table 4.1 Evaluation metrics results for Logistic Regression

Model	Accuracy	Precision	AUC	FPR	Sensitivity	Specificity
Logistic Regression	79.6%	76.4%	86%	25.6%	85%	74.3%

4.4.2 KNN

The goal of building a mental health predictive model with the k-Nearest Neighbours algorithm is to use data patterns to provide precise predictions. KNN is an instance-based learning technique that is non-parametric and uses the majority class of its k-nearest neighbours to classify data points. Here in our model we have used 27 nearest neighbours and the weights are uniformly distributed

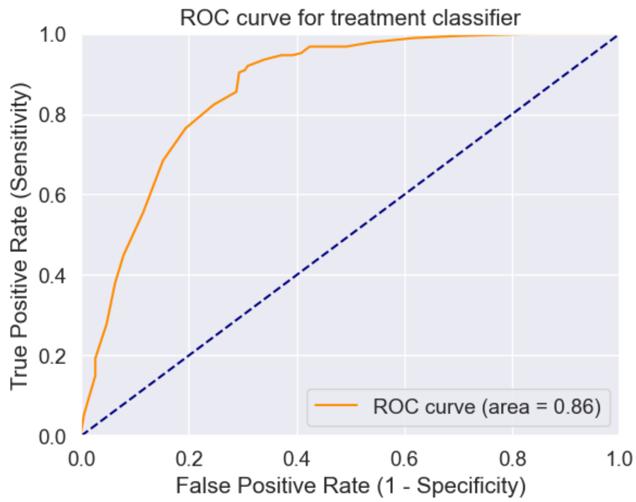


Figure 4.6 ROC curve for KNN model

.

This entails obtaining pertinent characteristics and traits that influence a person's mental health in the context of mental health prediction. Subsequently, the KNN model gains knowledge from the available data, recognising patterns among individuals and forecasting mental health consequences by analysing the traits of their closest neighbours. KNN's simplicity and flexibility in handling different datasets are its main advantages. To guarantee the model's dependability and efficacy in assisting with mental health evaluation and assistance, however, thorough consideration of feature selection, distance measurements, and model validation is essential.

Table 4.2 Evaluation metrics results for KNN

Model	Accuracy	Precision	AUC	FPR	Sensitivity	Specificity
KNN	80.4%	75.1%	86%	29.3%	90.3%	70.6%

4.4.3 Decision Trees

Using this machine learning approach, decision trees are used in model construction to analyse and comprehend complicated associations found in mental health data. Decision trees are very useful for this kind of work since they offer an understandable and transparent depiction of decision-making procedures. When building the model, pertinent characteristics are employed to divide the dataset into subgroups and iteratively create a tree structure. These elements include behavioural patterns, demographic data, and psychological evaluations. Every node in the tree denotes a choice made in response to a certain characteristic, which causes the leaf nodes to anticipate outcomes related to mental health. Decision trees' interpretability helps researchers and clinicians find important indicators and possible therapies by providing insight into the elements influencing mental health forecasts.

Decision trees are flexible for a variety of mental health datasets since they can handle both numerical and categorical data. Frequent validation and improvement of the model guarantee its precision and applicability, which aids in the creation of useful instruments for the early identification and treatment of mental health problems.

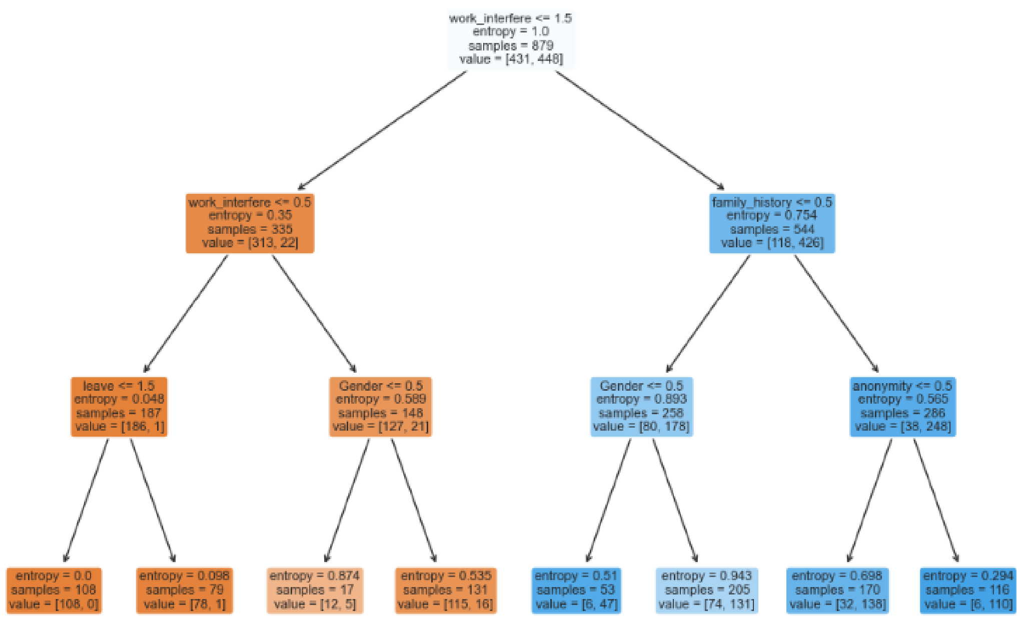


Figure 4.7 Splitting of Decision tree for mental health prediction

While building a decision tree model we obtained the best score of 83% with 93.5% sensitivity.

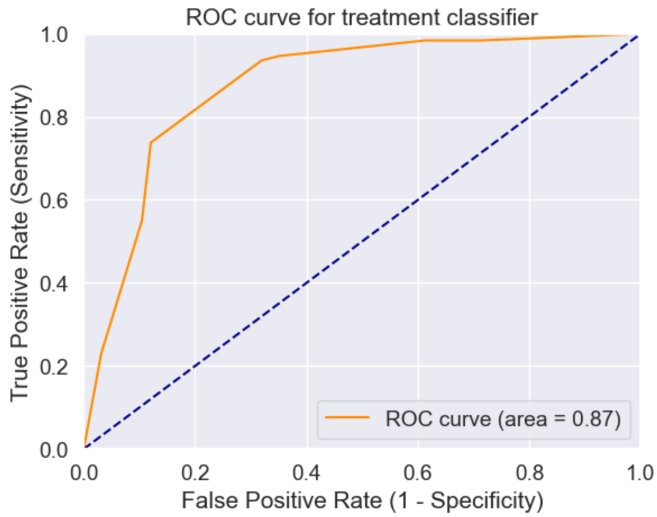


Figure 4.8 ROC curve for Decision Tree

To obtain these results the best parameters we obtained were the depth of the tree was 3 and the maximum number of features obtained from cross validation was 6, with minimum samples leaf of 7 and minimum split obtained was 8.

Table 4.3 Evaluation metrics results for Decision Tree

Model	Accuracy	Precision	AUC	FPR	Sensitivity	Specificity
Decision Trees	80.6%	74.1%	86%	31.9%	93.5%	68.06.3%

4.4.4 Random Forest

When building mental health prediction models, the Random Forest algorithm proves to be an effective and adaptable instrument. In order to function, Random Forest builds an ensemble of decision trees, each of which is trained using random feature picks on a portion of the dataset. Because of its intrinsic variety, the model is more resilient and less prone to overfitting. The algorithm can efficiently utilise a wide range of factors in the context of mental health prediction, including behavioural patterns, psychological evaluations, and demographic data. Because Random Forest can capture intricate associations in the data, it is very useful for detecting minute but important signs of mental health disorders. Furthermore, because of its natural interpretability, researchers and practitioners may learn more about the key variables influencing mental health outcomes.

We have obtained best score for sensitivity for this model i.e. 93.04%. now sensitivity indicates how sensitive is our model towards predicting the accurate results regarding the person who needs treatment or not.

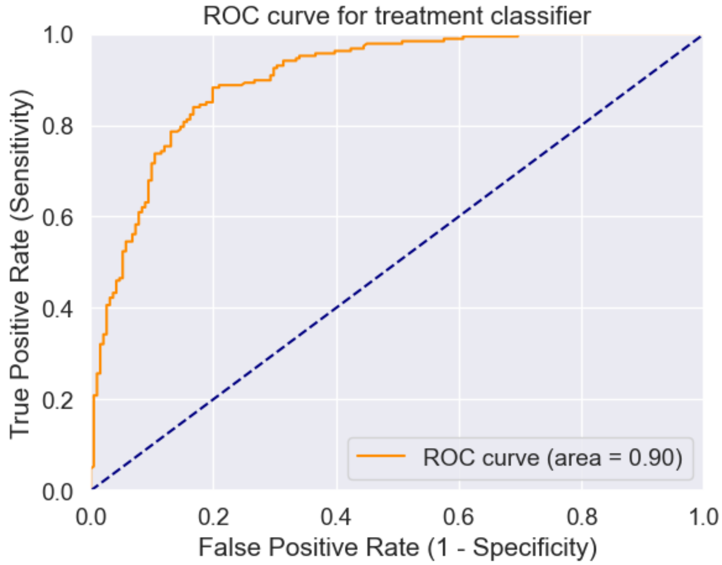


Figure 4.9 ROC curve for Random Forest

Random Forest models help provide accurate and dependable forecasts by utilising the combined power of multiple decision trees. This promotes improvements in early diagnosis, intervention, and individualised mental health treatment.

Table 4.4 Evaluation metrics results for Random Forest Tree

Model	Accuracy	Precision	AUC	FPR	Sensitivity	Specificity
Random Forest	81.2%	75%	90%	30.3%	93.04%	69.6%

4.4.5 Bagging

In the field of mental health prediction, bagging, or bootstrap aggregating, is a potent method for creating models. Bagging is the process of training several instances of the same predictive model using bootstrap sampled subsets of the dataset. This method helps in mental health prediction because it reduces overfitting and strengthens the model's resilience. As a result of combining predictions from many models that have been exposed to somewhat different aspects of the data, the ensemble model is better able to capture intricate patterns linked to indices of mental health.

As an example, bagging may be used to create a group of decision trees that together constitute a Random Forest. Taking into account the complex and varied nature of mental health, this method not only aids in the identification of pertinent characteristics but also offers a more accurate forecast of mental health issues. In the end, bagging helps the predictive model grow more skilled at managing the complexities of mental health data, making it a more sophisticated and useful tool for mental health evaluation and prediction.

.

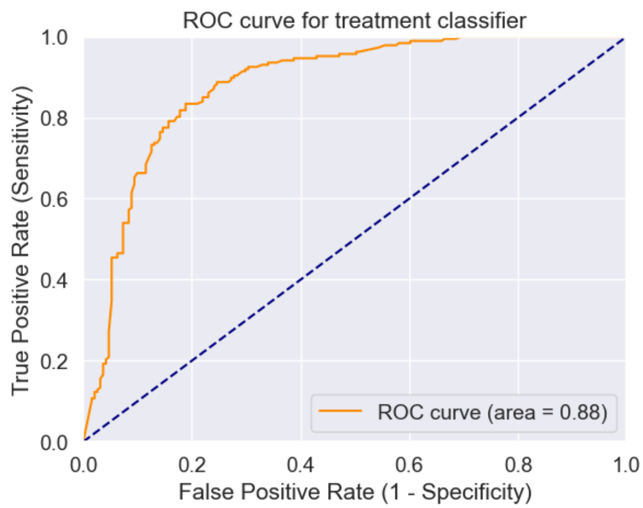


Figure 4.10 ROC curve for Bagging

The results obtained with bagging are quite not impressive when compared with other models. The accuracy is lowest in this model and the results are not that impressive

Table 4.3 Evaluation metrics results for Bagging Tree

Model	Accuracy	Precision	AUC	FPR	Sensitivity	Specificity
Bagging	77.7%	78.5%	86%	27.23%	82.89%	72.7%

4.4.6 Boosting

Boosting is a potent ensemble learning method used in the construction of mental health prediction models. Boosting algorithms, such AdaBoost or Gradient Boosting, combine the capabilities of several weak learners, usually decision trees, to iteratively create a powerful prediction model in the context of mental health. The final model is a weighted aggregate of these learners, with each weak learner educated to rectify the mistakes of its predecessor. This method is especially useful for predicting mental health because it reflects the intricate interactions between many aspects that affect mental health. Boosting is an effective way to reduce overfitting, improve generalisation of the model, and find subtle patterns in a variety of datasets.

The creation of proactive therapies and individualised techniques for mental well-being is supported by boosting algorithms, which use the collective wisdom of poor learners to provide accurate and robust predictions about mental health. A potent ensemble learning method called "boosting" is used to develop mental health prediction models. In the domain of mental health, boosting algorithms, such Gradient Boosting or AdaBoost, combine the capabilities of several weak learners, usually decision trees, to iteratively construct a powerful predictive model. The ultimate model is a weighted sum of these weak learners, each of which is taught to fix the mistakes of its predecessor. Because it captures the intricate interactions between numerous elements impacting mental well-being, this technique is very beneficial for mental health prediction.

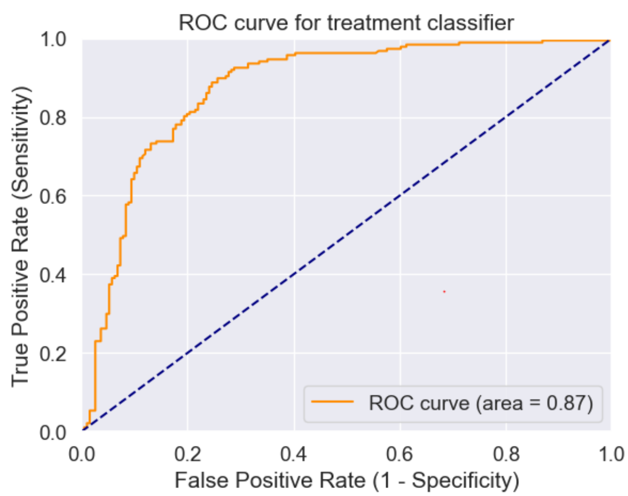


Figure 4.11 ROC curve for Boosting

The results obtained from Boosting are very impressive with accuracy being 81.7% and the highest sensitivity out of all the models i.e. 91.9%.

The best thing about boosting is how it works, how it trains himself with help of poor learners which makes it one of the best model for mental health prediction.

Table 4.3 Evaluation metrics results for Boosting

Model	Accuracy	Precision	AUC	FPR	Sensitivity	Specificity
Boosting	81.7%	76.10%	87%	28.2%	91.9%	71.7%

4.4.7 Stacking

Using a comprehensive approach to model construction is essential for accurate and dependable results in the field of mental health prediction. In this situation, stacking is a potent ensemble learning technique which proves very helpful to improve overall performance, the stacking process combines the predictions of several different models. Various factors, including physiological measures, behavioural patterns, and demographic data, can be used as input features for separate base models in the context of mental health prediction. These foundational models comprises of KNN, Random forest and Gaussian NB which represent different aspects of the intricate interactions related to mental health. Following that, the predictions from these several models are combined into a meta-model, also known as a "stacker," which gains the ability to balance and integrate the contributions of each base model.

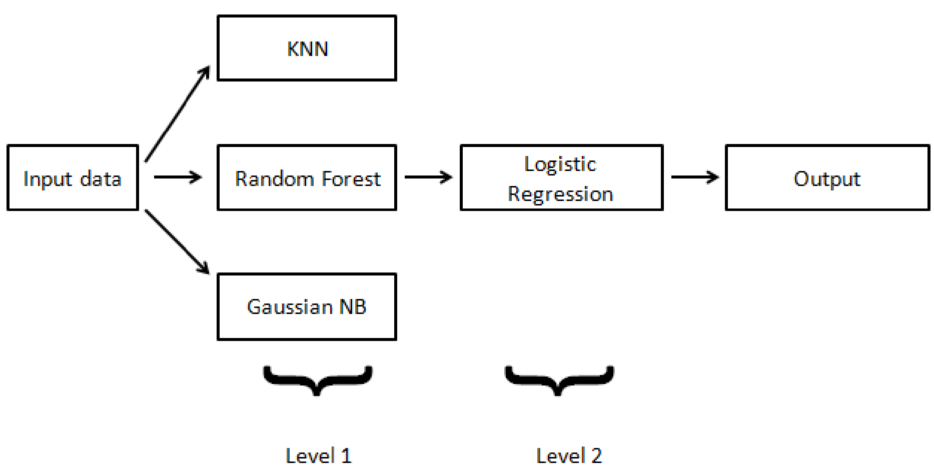


Figure 4.12 Architecture of stacking model

By combining the advantages of many models, this stacking strategy not only increases forecast accuracy but also strengthens the model's ability to handle the complex and diverse nature of mental health concerns. The end product is a thorough and sophisticated prediction model that has enormous potential to improve our comprehension of and ability to recognise mental health issues early on.

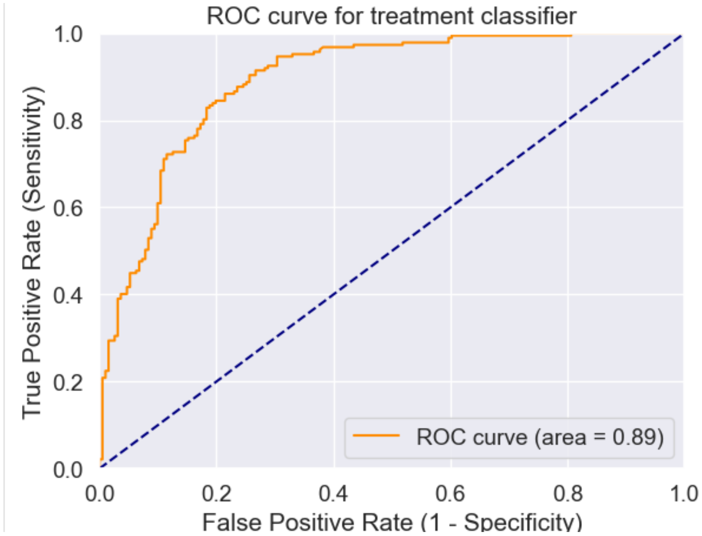


Figure 4.13 ROC curve for Stacking

The results obtained from the model were quite impressive, the best accuracy was obtained with this model i.e. 82% with best area under curve from the same model. The best point to note is that the false positive rate is quite low i.e. just 21.9% which is quite impressive wrt other models.

Table 4.4 Evaluation metrics results for Stacking

Model	Accuracy	Precision	AUC	FPR	Sensitivity	Specificity
Stacking	82.01%	79.3%	89%	21.9%	86.09%	78.01%

4.5 Summary

In this chapter, we provide a comprehensive review of model construction, application, and data pre-processing. Initially, a pre-processing and visualisation step is performed on the survey dataset. The dataset is then split up into 80% for training and 20% for testing.

Seven distinct models—Logistic regression , KNN, Decision trees , random forest , bagging , boosting and stacking —are used to evaluate performance. Each model's ideal parameters are used to make adjustments for improved results. By carefully reviewing and optimising these algorithms, we want to avoid the cases of mental health issues.

It takes careful planning to create a prediction model for mental health that works. This includes compiling a wide range of input elements, such as physiological data, behavioural patterns, and demographic information. The ensemble learning method known as "stackering" works well by aggregating predictions from these several models into a single meta-model. This method increases the model's resilience in handling the complex and multidimensional nature of mental health issues in addition to improving predicted accuracy. All things considered, developing a mental health prediction model entails carefully weighing the input features and integrating several modelling approaches in a purposeful way to provide a thorough and sophisticated instrument for the early detection and comprehension of mental health issues.

CHAPTER 5

RESULTS AND DISCUSSIONS

5.1 Introduction

This chapter on Results and Discussion is a turning point in our investigation of the complex field of mental health prediction. We explore the results of our painstakingly constructed prediction model with the aim of elucidating the subtleties behind the intricate interactions of many elements affecting mental health. This chapter seeks to demonstrate how well our model predicts future mental health issues by thoroughly examining physiological measures, behavioural patterns, and demographic data. Diverse machine learning models, such as decision trees, random forest, and stacking, have been integrated into our method to give it the adaptability required to understand the complex nature of mental health. We have combined the best features of separate models by using sophisticated ensemble learning strategies like stacking, which has allowed us to push the limits of resilience and forecast accuracy. A thorough analysis of our findings is provided, along with comparisons across models, performance measures, and practical applications of our findings. As we navigate this chapter, we set out to not only understand the predictive power of our model but also to make a significant contribution to the larger conversation about mental health prediction and intervention techniques.

5.2 Interpretations of Visualizations

The various visualizations exhibited in chapter 2 say a lot about the trends. We have plotted ROC curves and confusion matrix to present various parameters like accuracy, precision, sensitivity and specificity.

A confusion matrix provides a breakdown of true positive, true negative, false positive, and false negative predictions. Analyzing this matrix helps us understand the model's ability to correctly identify individuals with mental health issues (sensitivity) and those without (specificity). The trade-off between sensitivity and specificity at different categorization thresholds is represented visually by the ROC curve. The model's total discriminatory power is quantified by the area under the curve (AUC), with a steeper curve indicating greater model performance.

5.3 Model Evaluation

An essential first step in determining the efficacy and dependability of predictive models is model evaluation in the context of mental health prediction. The special difficulties and sensitivity related to mental health call for a careful analysis of model performance. Metrics like sensitivity, specificity, accuracy, and F1 score are essential for assessing how well the model can identify people with mental health issues while reducing false positives and negatives. Furthermore, the model's discriminative capacity and performance across various classification thresholds are elucidated by the Area under the Precision-Recall Curve (AUC-PR) and the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC-ROC).

In the context of mental health therapies, model evaluation also entails taking into account the possible fallout from inaccurate predictions. Maintaining a balance between specificity and sensitivity is frequently essential since incorrect classifications can have serious negative effects on people's health. The continuous enhancement of mental health prediction tools is facilitated by the validation and development of models on a regular basis, which is based on real-world results and expert feedback. This ensures the tools' ethical and practical application in many settings.

5.3.1 Logistic Regression

Logistic regression is a statistical method for binary classification It uses a logistic function to describe the likelihood that an instance belongs to a specific class. Because of its ease of interpretation, efficacy, and simplicity in situations where the dependent variable is categorical and has two possible outcomes, it is frequently utilised in machine learning.

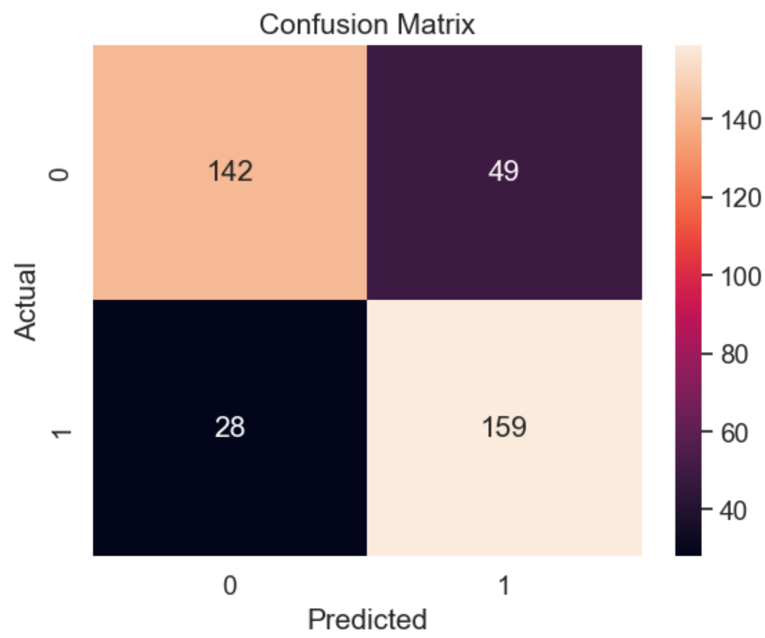


Figure 5.1 Confusion Matrix for Logistic Regression Model

The model gave an accuracy of 79.6% on training dataset and recall of 85%.

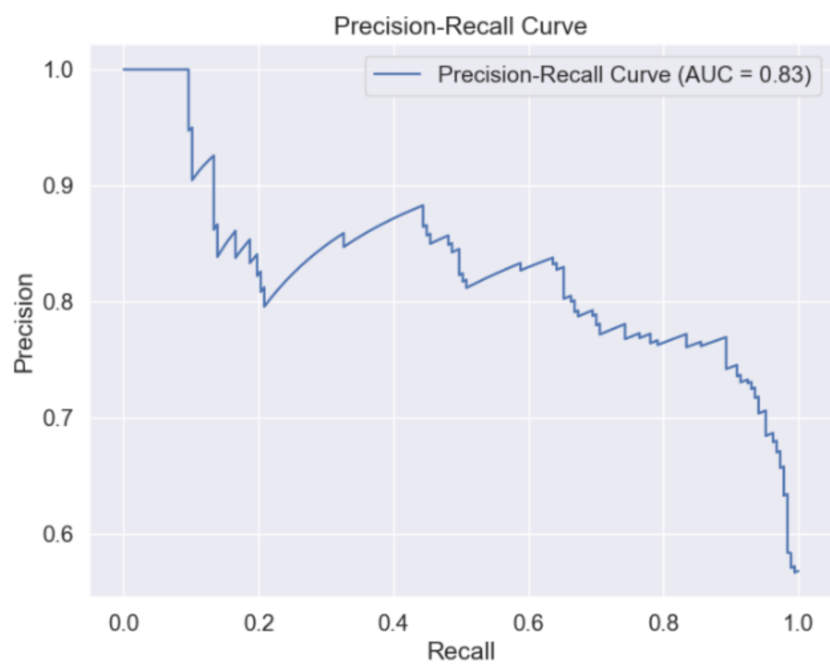


Figure 5.2 Precision Recall Curve for Logistic Regression Model

5.3.2 KNN

K-Nearest Neighbours (KNN) is a straightforward and efficient machine learning technique for problems involving regression and classification. A data point is given a label according to the majority class of its k-nearest neighbours. It is non-parametric and bases its predictions on how close together data points are.

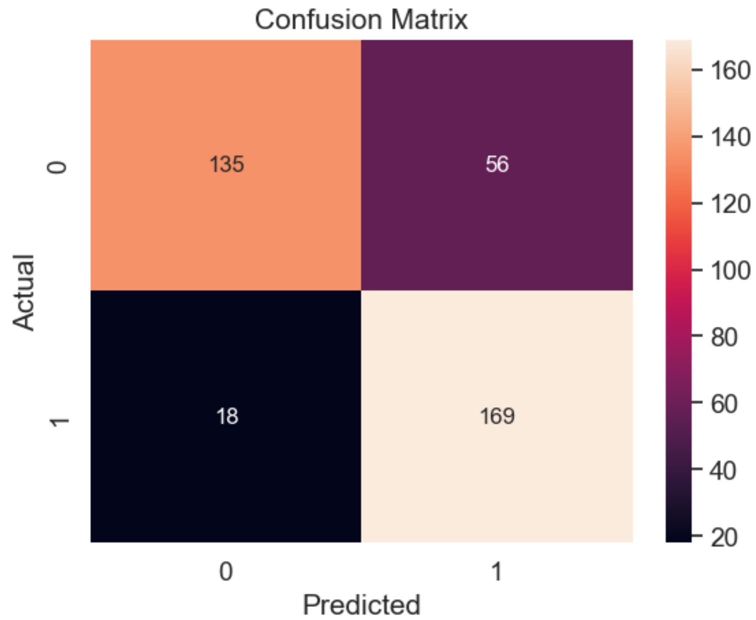


Figure 5.3 Confusion Matrix for KNN

The model gave an accuracy of 80.4% on training dataset and recall of 90.3%.

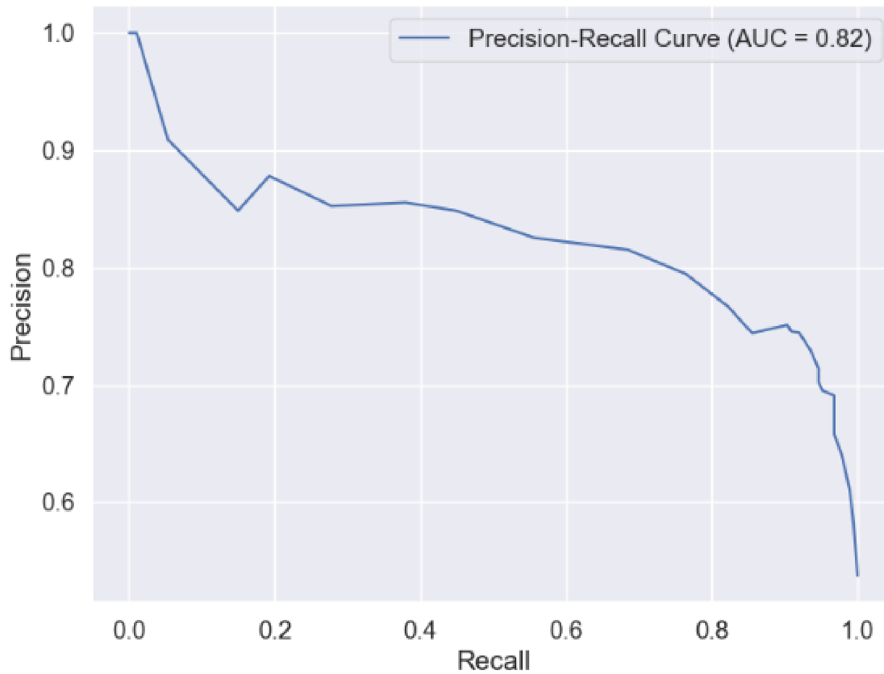


Figure 5.4 Precision Recall Curve for KNN

5.3.3 Decision Tree

A well-liked machine learning approach for tasks involving regression and classification is the decision tree. To make judgements, they partition the data recursively according to features. Every split maximises a certain criteria, such as entropy or Gini impurity. Decision trees have a variety of uses because of their interpretable tree-like structure.

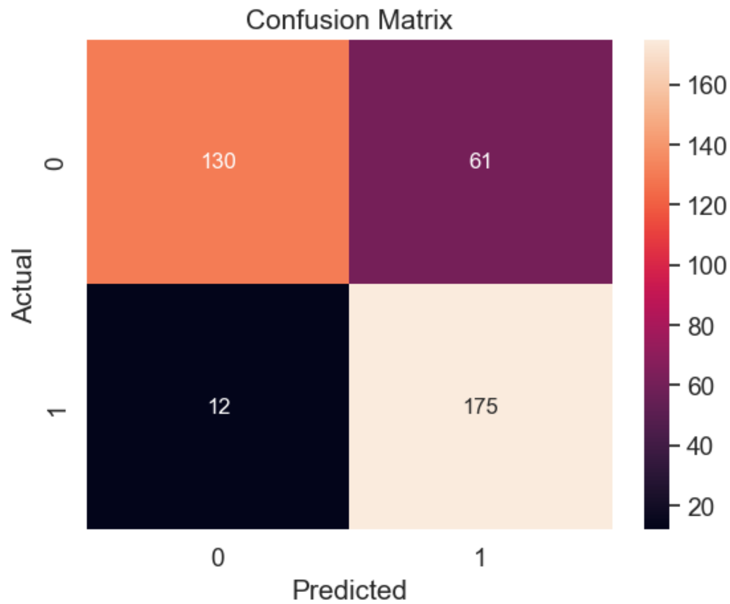


Figure 5.5 Confusion Matrix for Decision Tree

The model gave an accuracy of 79.6% on training dataset and recall of 85%.

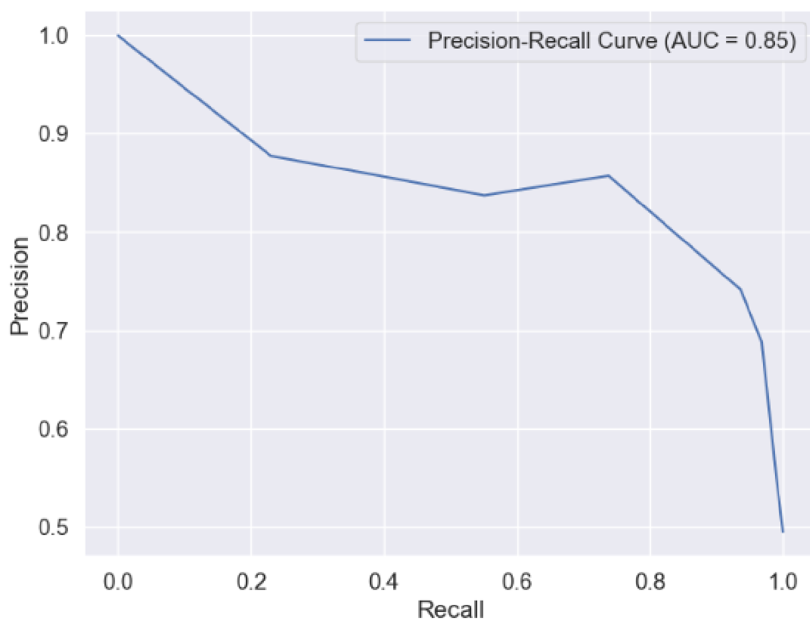


Figure 5.6 Precision Recall Curve for Decision Tree

5.3.4 Random Forest

An ensemble learning technique for anomaly detection, regression, and classification is called Random Forest. In order to increase precision and resilience, it constructs many decision trees and combines their predictions. By adding unpredictability to the tree-building process, it reduces overfitting and produces a strong and adaptable model for a range of machine learning applications.

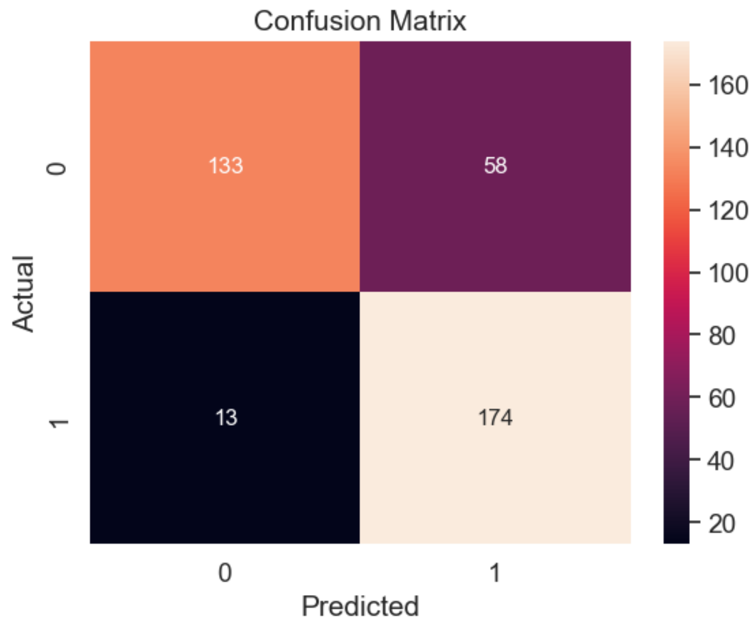


Figure 5.7 Confusion Matrix for Random Forest

The model gave an accuracy of 81.2% on training dataset and recall of 93.04%.

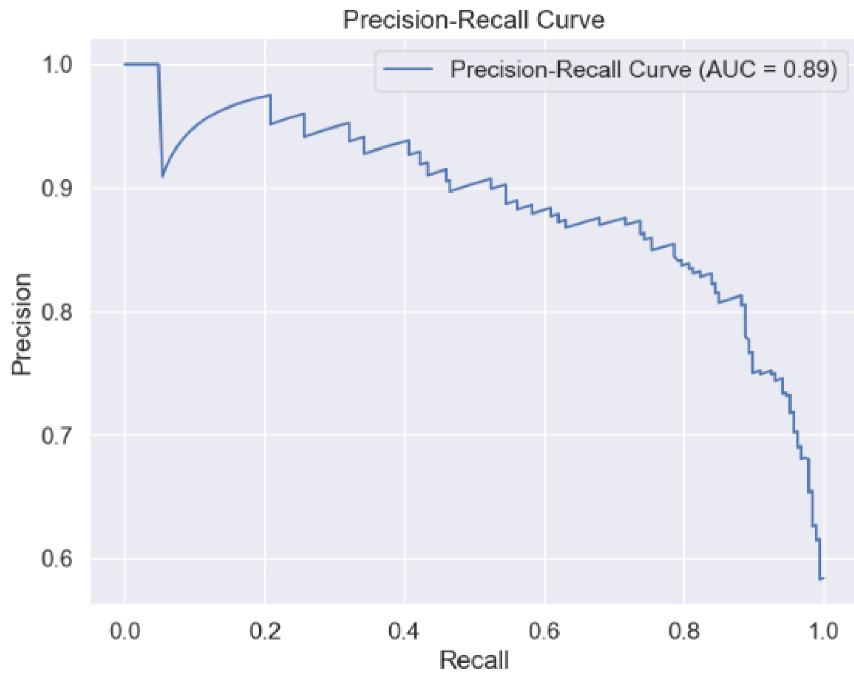


Figure 5.9 Precision Recall Curve for Random Forest

5.3.5 Bagging

Bootstrap Aggregating, or Bagging, is an ensemble learning strategy that trains several instances of a base model using bootstrap samples of the dataset, therefore improving the model's robustness and accuracy. A voting system or average is frequently used as the final forecast. Bagging improves prediction accuracy across a range of contexts and decreases overfitting.

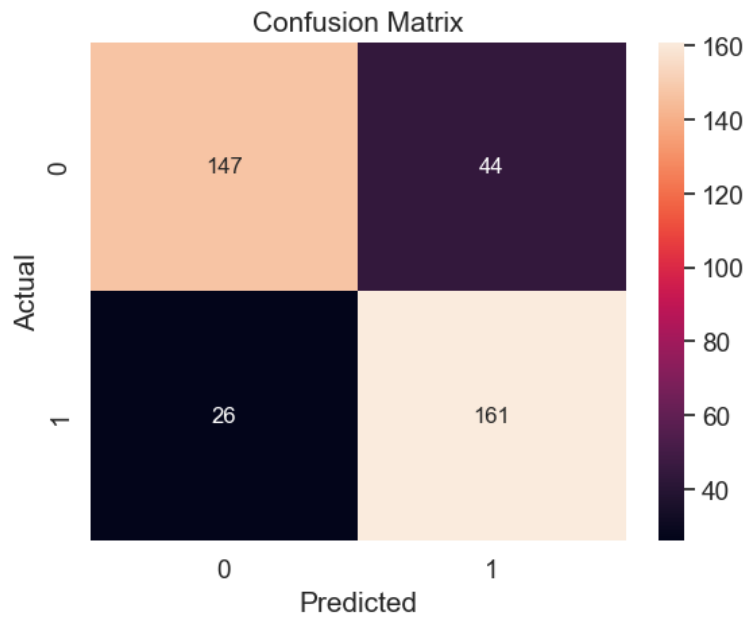


Figure 5.7 Confusion Matrix for Bagging

The model gave an accuracy of 77.70% on training dataset and recall of 82.89%.

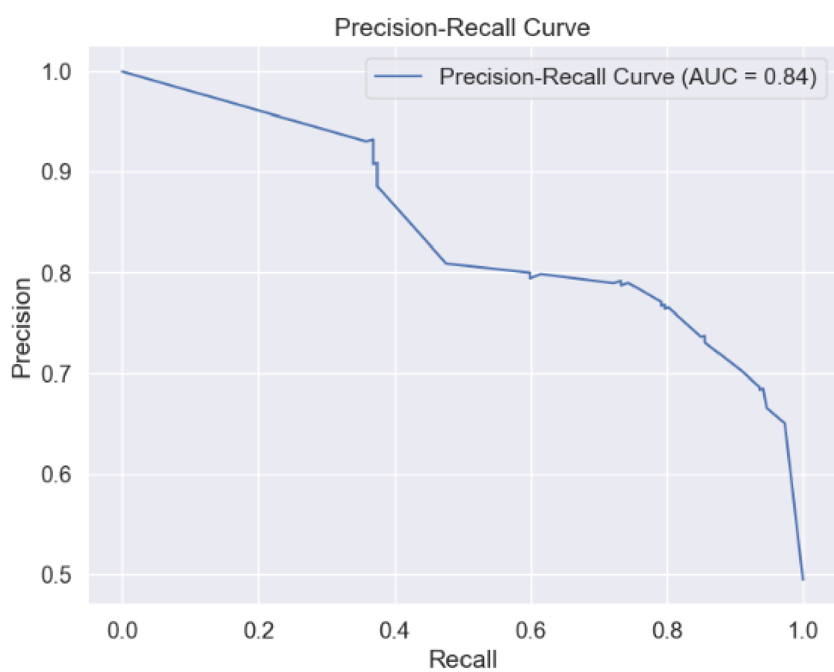


Figure 5.10 Precision Recall Curve for Bagging

5.3.6 Boosting

Boosting is an ensemble machine learning strategy that builds a reliable and accurate model by combining several weak learners. In order to improve the predictability of the misclassified occurrences in following rounds, it gives weights to them. Boosting is implemented by well-known algorithms like AdaBoost and Gradient Boosting, which improve prediction performance in a variety of applications.

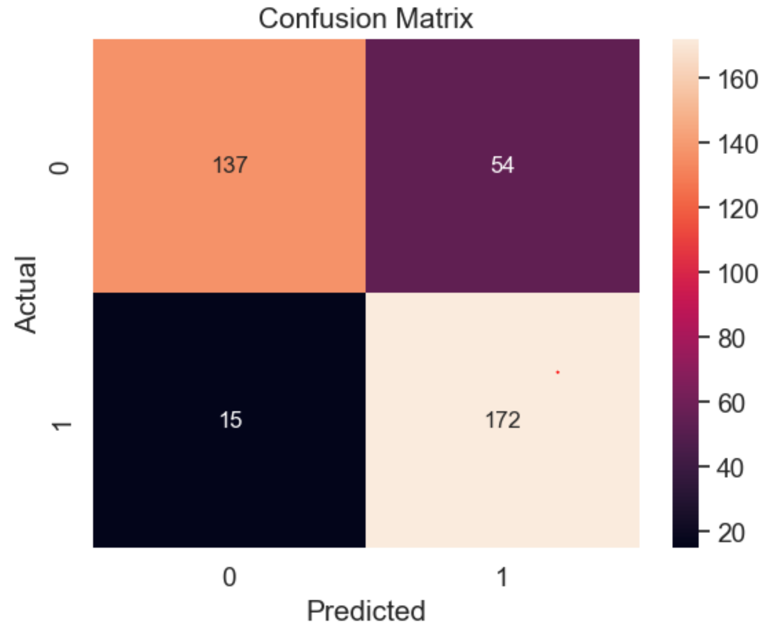


Figure 5.11 Confusion Matrix for boosting

The model gave an accuracy of 81.70% on training dataset and recall of 91.90%.

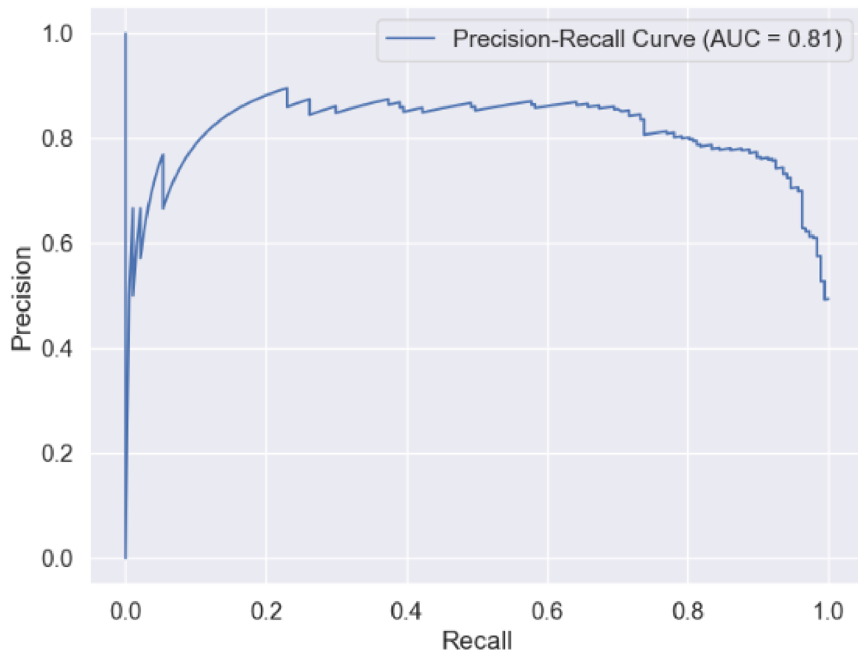


Figure 5.12 Precision Recall Curve for Boosting

5.3.7 Stacking

An ensemble learning method called stacking combines predictions from several models to improve performance as a whole. It entails training many base models, then combining their results with a meta-model. For complicated problems like mental health prediction, stacking increases predictive accuracy and resilience, making it an efficient method.

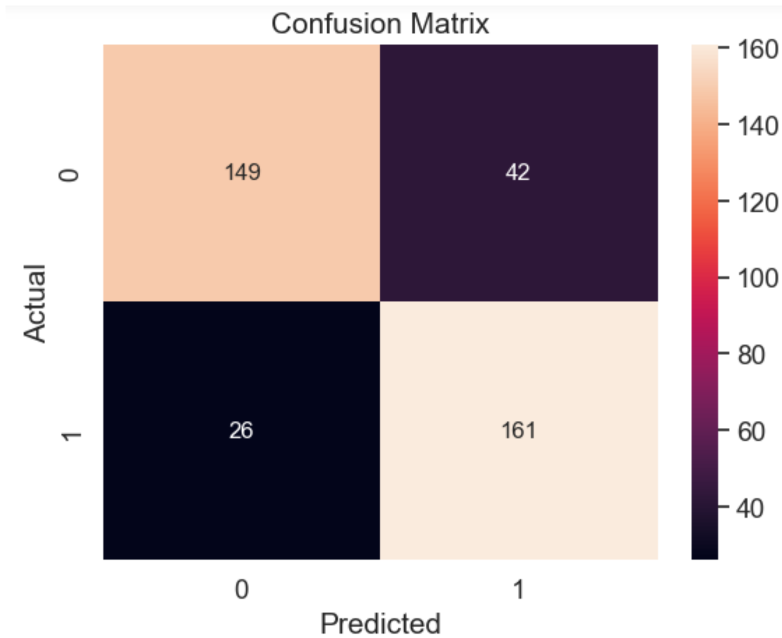


Figure 5.13 Confusion Matrix for Stacking

The model gave an accuracy of 82.01% on training dataset and recall of 86.09%.

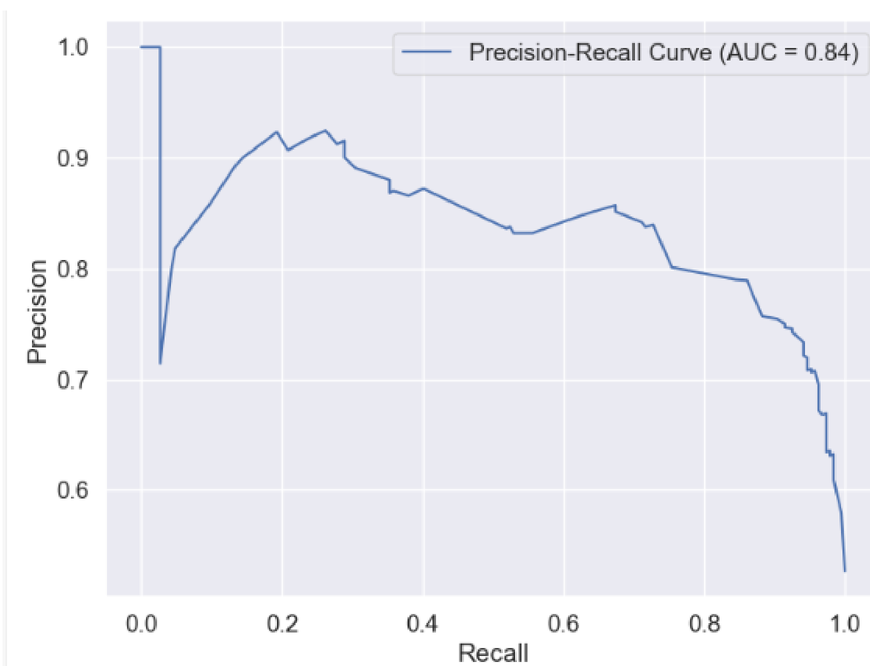


Figure 5.12 Precision Recall Curve for stacking

5.4 Visualizations for the Optimal Model

The optimal model for predicting mental health or to identify mental health illnesses early is by using stacking technique where we have used 3 different models i.e. KNN, Gaussian NB , Random forest. The visualizations for the optimal model are illustrated in figures below.

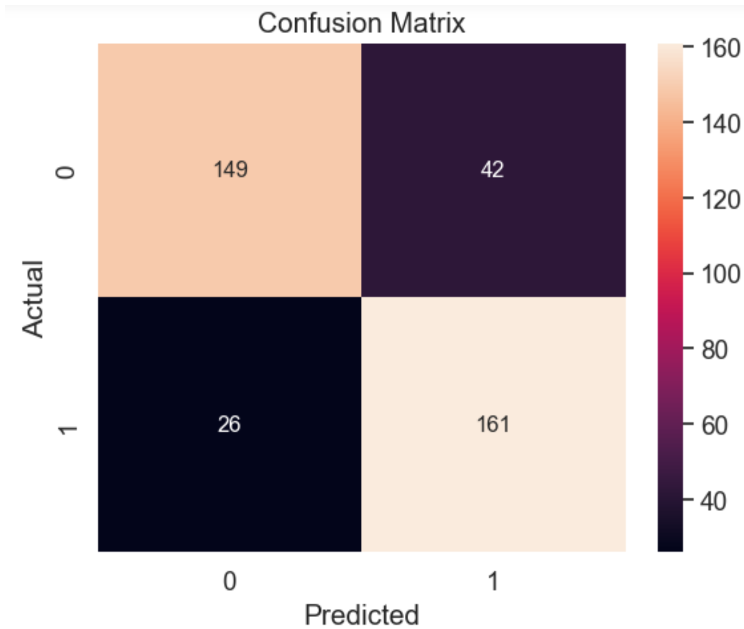


Figure 5.13 Final Model Confusion Matrix

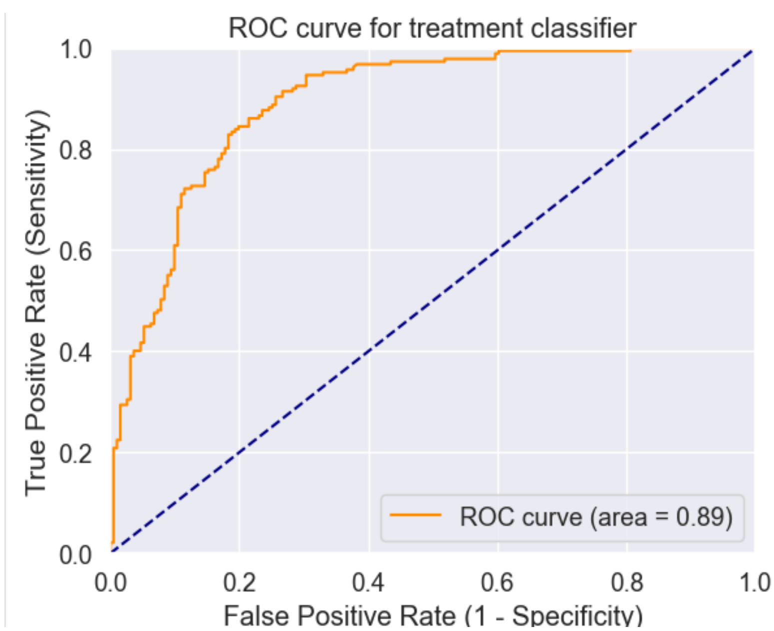


Figure 5.14 ROC of the Final model

5.5 Discussion

After proper scaling, training and testing of the machine learning models the results are evaluated. The results were compared on the basis of various factors like the values of precision and recall. As we can see the values obtained by the stacking models are quite impressive when compared to other pre built machine learning models. Diverse machine learning models, such as decision trees, random forest, and stacking, have been integrated into our method to give it the adaptability required to understand the complex nature of mental health. We have combined the best features of separate models by using sophisticated ensemble learning strategies like stacking, which has allowed us to push the limits of resilience and forecast accuracy. A thorough analysis of our findings is provided, along with comparisons across models, performance measures, and practical applications of our findings. We set out to not only understand the predictive power of our model but also to make a significant contribution to the larger conversation about mental health prediction and intervention techniques.

5.6 Summary

The analysis and findings of the study are interpreted in this chapter. The various machine learning models' visualizations are compared in this chapter. This chapter different classification models on the basis of their precision, recall and AUC values. Ensemble models have good precision values as compared to the Logistic Regression or KNN. This chapter examines and contrasts the findings with the most recent expectations.

CHAPTER 6

CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This chapter provides a brief overview of the study and makes conclusions based on the results of the experiment on mental health prediction. The major topic of discussion is the methods utilised to anticipate occurrences of mental health using the KNN, Logistic Regression, stacking, and other ensemble models. It delves further into the analysis conducted and the model's assessment. The chapter also examines the aims and objectives of the research and assesses the degree to which they were met. It also suggests future avenues for further research and suggests possible research subjects to help better understand and spot early signs for treatment of mental health condition.

6.2 Discussions and Conclusions

The most recent prediction frameworks for mental health have been evaluated in this paper. A publicly accessible dataset that was collected from Github is used in the study. The dataset includes details of people collected from different countries and includes there behavioural , psychological and lifestyle data. Prior to the train-test split, the dataset is pre-processed using feature scaling and deletion. We have considered only important features related to an individual. We have prepared a lot of models considering both linear and non linear models. In non linear models we have tried ensemble models and as well as hybrid models to obtain the best results. The models are compared on the basis of there precision and recall values to obtain the best model out of all the models.

The study comes to the conclusion that complicated non-linear patterns on healthcare data that are challenging to identify using linear algorithms may be revealed by combining multiple models or by using the techniques of ensemble. When it comes to predicting mental health of an individual stacking algorithms have outperformed linear regression models. Recent research, however, shows that hybrid approaches—which include several models—can perform better than solo machine learning models. According to the study's findings, a hybrid model more accurately than an individual model.

To sum up, the analysis of the Ensemble models in the field of healthcare give good results when compared to another models. The study highlights how important it is to use advanced prediction frameworks and hybrid models to improve mental health prediction and to determine the early signs to avoid the cases.

6.3 Contribution to Knowledge

In order to identify mental health, this study focuses on using many machine learninh models, such as logistic regression, KNN, stacking, bagging and boosting. The models get trained and evaluated using the Survey data. According to the study's analysis of these models' performance, both the stacking and boosting models are good at accurately recognising mental health cases. These models work effectively and accurately in classifying early signs of mental health behaviours. The primary focus of the project is on the hybrid models, which have shown encouraging results in accurately identifying the mental health. These models provide informative data that can direct further research and technological developments in the field of mental health.

6.4 Future Recommendations

In order to further improve the effectiveness and usefulness of predictive models, future developments in mental health prediction should concentrate on a number of important areas. First and foremost, more cooperation between data scientists, mental health specialists, and legislators is required to guarantee data usage that is morally righteous, privacy protection, and the appropriate application of predictive technology. Additionally, in order to promote confidence between healthcare practitioners and those in need of assistance, research activities should give priority to developing better interpretable and explainable models.

In addition, the integration of a varied array of data sources, such as lifestyle characteristics, social determinants of health, and real-time behavioural data from wearable devices, can enhance prediction models and provide a more all-encompassing comprehension of a person's mental health condition.

In terms of technology, textual data from diverse sources, such social media and electronic health records, may be analysed to provide deeper insights through the combination of artificial intelligence with natural language processing and sentiment analysis. The possibility of longitudinal studies and ongoing monitoring to record dynamic shifts in mental health conditions over time should also be investigated in ongoing research.

In summary, multidisciplinary cooperation, ethical issues, technical advancement, and a dedication to meeting the many requirements of people in various sociocultural situations are critical to the future of mental health prediction. By making progress in these areas, we can create an environment in which predictive models will not only increase diagnosis accuracy but also play a major role in early intervention, individualised treatment programmes, and the general improvement of mental health outcomes worldwide.

RERFERENCES

1. Angehrn, A., Teale Sapach, M.J.N., Ricciardelli, R., Macphee, R.S., Anderson, G.S. & Nicholas Carleton, R. (2020). Sleep quality and mental disorder symptoms among canadian public safety personnel. *International Journal of Environmental Research and Public Health*. [Online]. p.p. 14. Available from: https://pubmed.ncbi.nlm.nih.gov/32326489/.

2. Bajaj, M., Rawat, P., Diksha, Vats, S., Sharma, V. & Gopal, L. (2023). Prediction of Mental Health Treatment Adherence using Machine Learning Algorithms. *2023 International Conference on Computational Intelligence, Communication Technology and Networking, CICTN 2023*. [Online]. p.pp. 1–5. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10141520.

3. Chandana, C., Neha, P.N., Nisarga, S.M., Thanvi, P. & Balarengadurai, C. (2023). Mental Health Prediction Using Machine Learning. *Cognitive Science and Technology*. [Online]. p.p. 22. Available from: https://www.hindawi.com/journals/acisc/2022/9970363/.

4. Cho, G., Yim, J., Choi, Y., Ko, J. & Lee, S.H. (2019). Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investigation*. [Online]. p.pp. 1–8. Available from: https://www.psychiatryinvestigation.org/journal/view.php?doi=10.30773/pi.2018.12.21.2.

5. Chung, J. & Teo, J. (2022). Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges. *Applied Computational Intelligence and Soft Computing*. [Online]. p.p. 19. Available from: https://www.hindawi.com/journals/acisc/2022/9970363/.

6. Cooper, A., Horrocks, J., Goodday, S., Keown-Stoneman, C. & Duffy, A. (2021). Predicting the risk and timing of major mood disorder in offspring of bipolar parents: exploring the utility of a neural network approach. *International Journal of Bipolar Disorders*. [Online]. p.p. 9. Available from: https://doi.org/10.1186/s40345-021-00228-2.

7. Ebert, D.D., Buntrock, C., Mortier, P., Auerbach, R., Weisel, K.K., Kessler, R.C., Cuijpers, P., Green, J.G., Kiekens, G., Nock, M.K., Demyttenaere, K. & Bruffaerts, R. (2018). Prediction of major depressive disorder onset in college students. *Depression and Anxiety*. [Online]. p.p. 11. Available from: https://onlinelibrary.wiley.com/doi/10.1002/da.22867.

8. Gomes, N., Pato, M., Lourenço, A.R. & Datia, N. (2023). A Survey on Wearable Sensors for Mental Health Monitoring. *Sensors*. [Online]. p.pp. 1–16. Available from: https://www.mdpi.com/1424-8220/23/3/1330#:~:text=Tracking physiological parameters%3A Wearable sensors,stress or anxiety %5B5%5D.

9. Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H. & Eichstaedt, J.C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*. [Online]. p.p. 7. Available from: http://dx.doi.org/10.1016/j.cobeha.2017.07.005.

10. Islam, M.R., Kamal, A.R.M., Sultana, N., Islam, R., Moni, M.A. & Ulhaq, A. (2018). Detecting Depression Using K-Nearest Neighbors (KNN) Classification Technique. *International Conference on Computer, Communication, Chemical, Material and Electronic Engineering, IC4ME2 2018*. [Online]. (March 2019). p.pp. 1–4. Available from: https://www.researchgate.net/publication/327821019_Detecting_Depression_Using_K-Nearest_Neighbors_KNN_Classification_Technique.

11. Jain, T., Jain, A., Hada, P.S., Kumar, H., Verma, V.K. & Patni, A. (2021). Machine Learning Techniques for Prediction of Mental Health. *Proceedings of the 3rd International Conference on Inventive Research in Computing Applications, ICIRCA 2021*. [Online]. p.pp. 1606–1613. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9545061.

12. Kim, J., Lee, D. & Park, E. (2022). Machine learning for mental health in social media: Bibliometric study. *Journal of Medical Internet Research*. [Online]. p.pp. 1–17. Available from: https://www.researchgate.net/publication/349903899_Machine_Learning_for_Mental_Health_in_Social_Media_Bibliometric_Study.

13. Laijawala, V., Aachaliya, A., Jatta, H. & Pinjarkar, V. (2020). *Classification Algorithms based Mental Health Prediction using Data Mining*. [Online]. p.pp. 1–5. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9137856.

14. Lee, J. & Pak, T.Y. (2022). Machine learning prediction of suicidal ideation, planning, and attempt among Korean adults: A population-based study. *SSM - Population Health*. [Online]. p.pp. 1–9. Available from: https://doi.org/10.1016/j.ssmph.2022.101231.

15. Marrapu, H.K., Maram, B. & Reddi, P. (2022). New Analytic Framework of Public Mental Health Prediction Using Data Science. *1st IEEE International Conference on Smart Technologies and Systems for Next Generation Computing, ICSTSN 2022*. [Online]. p.pp. 1–6. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9761324.

16. Moshe, I., Terhorst, Y., Opoku Asare, K., Sander, L.B., Ferreira, D., Baumeister, H., Mohr, D.C. & Pulkki-Råback, L. (2021). Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. *Frontiers in Psychiatry*. [Online]. p.pp. 1–12. Available from: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.625247/full.

17. Naveen Paul, E. & Juliet, S. (2023). Comparative Analysis of Machine Learning Techniques for Mental Health Prediction. *Proceedings of the 8th International Conference on Communication and Electronics Systems, ICCES 2023*. [Online]. p.pp. 1–6. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10192763.

18. Nouman, M. (2023). Mental Health Prediction through Text Chat Conversations. *2023 International Joint Conference on Neural Networks (IJCNN)*. [Online]. p.pp. 1–6. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=10191849.

19. Papini, S., Pisner, D., Shumake, J., Powers, M.B., Beevers, C.G., Rainey, E.E., Smits, J.A.J. & Warren, A.M. (2018). Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization. *Journal of Anxiety Disorders*. [Online]. p.p. 8. Available from: https://doi.org/10.1016/j.janxdis.2018.10.004.

20. Playne, D. (2022). Virtual Reality Data for Predicting Mental Health Conditions. *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. [Online]. p.pp. 1–3. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9974469.

21. Rothenberg, W.A., Bizzego, A., Esposito, G., Lansford, J.E., Al-Hassan, S.M., Bacchini, D., Bornstein, M.H., Chang, L., Deater-Deckard, K., Di Giunta, L., Dodge, K.A., Gurdal, S., Liu, Q., Long, Q., Oburu, P., Pastorelli, C., Skinner, A.T., Sorbring, E., Tapanya, S., Steinberg, L., Tirado, L.M.U., Yotanyamaneewong, S. & Alampay, L.P. (2023). Predicting Adolescent Mental Health Outcomes Across Cultures: A Machine Learning Approach. *Journal of Youth and Adolescence*. [Online]. (8). p.pp. 1–25. Available from: https://link.springer.com/article/10.1007/s10964-023-01767-w.

22. Saibaba, C.M.H., Alekhya, K.V.K., Yeshwanth, K. & Tumuluru, P. (2022). Prediction of Public Mental Health by using Machine Learning Algorithms. *Proceedings of the 2nd International Conference on Artificial Intelligence and Smart Energy, ICAIS 2022*. [Online]. p.pp. 1–4. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10141520.

23. Smucny, J., Shi, G. & Davidson, I. (2022). Deep Learning in Neuroimaging: Overcoming Challenges With Emerging Approaches. *Frontiers in Psychiatry*. 13 (June). p.pp. 1–7.

24. Srinath, K.S., Kiran, K., Gagan, A.G., Jyothi, D.K., Shenoy, P.D. & Venugopal, K.R. (2022). Enhancing Mental Illness Prediction using Tree based Machine Learning Approach. *2022 IEEE International Conference on Electronics, Computing and Communication Technologies, CONECCT 2022*. [Online]. p.pp. 1–5. Available from: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9865689.

25. Tiwari, P.K., Sharma, M., Garg, P., Jain, T., Verma, V.K. & Hussain, A. (2021). A Study on Sentiment Analysis of Mental Illness Using Machine Learning Techniques. *IOP Conference Series: Materials Science and Engineering*. [Online]. p.pp. 1–10. Available from: https://iopscience.iop.org/article/10.1088/1757-899X/1099/1/012043/pdf.

26. Verma, D., Bach, K. & Mork, P.J. (2023). External validation of prediction models for patient-reported outcome measurements collected using the SELFBACK mobile app. *International Journal of Medical Informatics*. [Online]. p.pp. 1–5. Available from: https://doi.org/10.1016/j.ijmedinf.2022.104936.

27. Wang, X., Li, H., Sun, C., Zhang, X., Wang, T., Dong, C. & Guo, D. (2021). Prediction of Mental Health in Medical Workers During COVID-19 Based on Machine Learning. *Frontiers in Public Health*. [Online]. p.pp. 1–13. Available from: https://www.frontiersin.org/articles/10.3389/fpubh.2021.697850/full.

28. N. P. E and S. Juliet, "Comparative Analysis of Machine Learning Techniques for Mental Health Prediction," 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 1-6, doi: 10.1109/ICCES57224.2023.10192763.

29. R. Boina, S. Sangeethapriya, H. S. Pokhariya, A. Sharma, A. Rao and S. Singh, "Data Mining-based Classification Methods for Mental Health Prediction," 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2023, pp. 738-743, doi: 10.1109/ICPCSN58827.2023.00127.

30. Y. Li, "Application of Machine Learning to Predict Mental Health Disorders and Interpret Feature Importance," 2023 3rd International Symposium on Computer Technology and Information Science (ISCTIS), Chengdu, China, 2023, pp. 257-261, doi: 10.1109/ISCTIS58954.2023.10213032.

31. E. Mylona et al., "Explainable machine learning analysis of longitudinal mental health trajectories after breast cancer diagnosis," 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Ioannina, Greece, 2022, pp. 1-4, doi: 10.1109/BHI56158.2022.9926952.

32. Z. Yan, "Application of ID3 Algorithm in Mental Health Education of College Students," 2022 International Conference on Knowledge Engineering and Communication Systems (ICKES), Chickballapur, India, 2022, pp. 1-5, doi: 10.1109/ICKECS56523.2022.10060591.

33. S. O. Isiaq and L. Dawson, "Mental health predictive models for triaging young adults," 2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME), Maldives, Maldives, 2022, pp. 1-6, doi: 10.1109/ICECCME55909.2022.9988301.

34. A. A. Mir, N. G. G. Jerome, R. Akshara, S. K. Reddy, S. K. Mohapatra and J. Mohanty, "Comparative Analysis of Mental Health disorder in Higher Education Students using Predictive Algorithms," 2023 International Conference on Applied Intelligence and Sustainable Computing (ICAISC), Dharwad, India, 2023, pp. 1-5, doi: 10.1109/ICAISC58445.2023.10199947.

35. W. Santos, S. Yoon and I. Paraboni, "Mental Health Prediction from Social Media Text Using Mixture of Experts," in IEEE Latin America Transactions, vol. 21, no. 6, pp. 723-729, June 2023, doi: 10.1109/TLA.2023.10172137.

36. M. A. Abid, Z. Dehghan, T. Shinde and G. Narang, "Machine Learning based approaches for Identification and Prediction of diverse Mental Health Conditions," 2023 IEEE International Conference on Contemporary Computing and Communications (InC4), Bangalore, India, 2023, pp. 1-5, doi: 10.1109/InC457730.2023.10263141.

37. J. Park, J. Kim and S. -P. Kim, "A Study on the Development of a Day-to-Day Mental Stress Monitoring System using Personal Physiological Data," 2018 18th International Conference on Control, Automation and Systems (ICCAS), PyeongChang, Korea (South), 2018, pp. 900-903.

38. D. Naveen, P. Rachana, S. Swetha and S. Sarvashni, "Mental Health Monitor using Facial Recognition," 2023 2nd International Conference for Innovation in Technology (INOCON), Bangalore, India, 2023, pp. 1-3, doi: 10.1109/INOCON57975.2023.10101000.

39. P. Yadav, S. Shinde and R. Shedge, "Mental Health Disorder Detection Using Machine Learning and Deep Learning Tehniques," 2023 3rd Asian Conference on Innovation in Technology (ASIANCON), Ravet IN, India, 2023, pp. 1-6, doi: 10.1109/ASIANCON58793.2023.10270707.

40. N. Isa, N. Mohamad, A. I. A. Badri and H. B. Sa'aid, "Analyzing Factors of Mental Health Problems among Malaysian University Students using Clustering Analysis," 2022 3rd International Conference on Artificial Intelligence and Data Sciences (AiDAS), IPOH, Malaysia, 2022, pp. 164-169, doi: 10.1109/AiDAS56890.2022.9918810.

41. S. Salmani, P. Desale and F. Anthony, "Analysis of Anxiety and Depression in Gaming Individuals," 2021 International Conference on Communication information and Computing Technology (ICCICT), Mumbai, India, 2021, pp. 1-5, doi: 10.1109/ICCICT50803.2021.9510061.

42. M. P. Jain, S. Sribash Dasmohapatra and S. Correia, "Mental Health State Detection Using Open CV and Sentimental Analysis," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 465-470, doi: 10.1109/ICISS49785.2020.9315984.

43. D. Zhang, T. Guo, S. Han, S. Vahabli, M. Naseriparsa and F. Xia, "Predicting Mental Health Problems with Personality, Behavior, and Social Networks," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 4537-4546, doi: 10.1109/BigData52589.2021.9671987.

44. J. Nanavati and U. Patel, "Hybrid Model for Analysis of Social Media Posts for Identification of Depression and Measuring Its Severity," 2023 International Conference on Data Science and Network Security (ICDSNS), Tiptur, India, 2023, pp. 1-5, doi: 10.1109/ICDSNS58469.2023.10245441.

45. Y. Hata and H. Nakajima, "A Health Care Service to Disaster Survivors," 2012 Annual SRII Global Conference, San Jose, CA, USA, 2012, pp. 637-641, doi: 10.1109/SRII.2012.77.

46. S. Gangbo and G. Shidaganti, "Classification of Student Mental Health Prediction Using LSTM," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/GCAT55367.2022.9972061.

47. T. Rutowski, E. Shriberg, A. Harati, Y. Lu, P. Chlebek and R. Oliveira, "Depression and Anxiety Prediction Using Deep Language Models and Transfer Learning," 2020 7th International Conference on Behavioural and Social Computing (BESC), Bournemouth, United Kingdom, 2020, pp. 1-6, doi: 10.1109/BESC51023.2020.9348290.

48. V. M. Sanjay, A. I, D. Gc, A. HP, J. Malik and T. Bn, "Anxiety Prediction during Stressful Scenarios using Machine Learning," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1199-1202, doi: 10.1109/ICICCS53718.2022.9788151.

49. A. Beheshti, V. M. Hashemi and S. Wang, "Towards Predictive Analytics in Mental Health Care," 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 2021, pp. 1-7, doi: 10.1109/IJCNN52387.2021.9534233.

50. Z. Wang, "A Fine-Grained Data Mining Model Based on Psychometric Data," 2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE), Harbin, China, 2020, pp. 1139-1142, doi: 10.1109/ICMCCE51767.2020.00250.

Appendix B: Research Proposal

ABSTRACT

Mental health is a major public health concern across the world, with a rising number of people dealing with mental health issues. The importance of early identification and prompt intervention in enhancing outcomes and lowering the burden of mental health issues cannot be overstated. The rise in mental health disorders and the necessity for better medical health treatment have prompted an inquiry into machine learning applications in mental health. Diagnosing mental health is difficult because people aren't always willing to talk about their problems.

Machine learning & Deep learning are branches of artificial intelligence that are mostly used nowadays. We have collected data from different sources where we have information of an individual regarding their work, lifestyle, work load, family history, type of work etc and will predict whether mental health treatment is required or not. So, we will be using different machine learning models for the same.

Table of Contents

1. Background

Mental health is a major public health concern across the world, with a rising number of people dealing with mental health issues. The importance of early identification and prompt intervention in enhancing outcomes and lowering the burden of mental health issues cannot be overstated. These illnesses have serious personal, social, and economic repercussions on people of all ages, ethnicities, and socioeconomic statuses. More than 450 million people worldwide suffer from mental health illnesses, making it one of the top causes of disability globally, according to the World Health Organisation (WHO).

Standardised tests, clinical observations, and self-assessments have been used as the backbone of traditional methods for diagnosing mental illness. However, these approaches frequently have drawbacks, such as inherent biases, prolonged assessment durations, and challenges in identifying early warning indications. Additionally, the stigma attached to mental health conditions may discourage some people from seeking treatment or from disclosing their genuine symptoms during professional assessments.

More innovative techniques for mental health prediction and intervention have recently become feasible thanks to the development of technology, the accessibility of large-scale datasets, and the rising interest in AI and ML. These models seek to make use of data from various sources, including wearable technology, social media activity, smart phone apps, genetic data, and electronic health records, in order to spot patterns, risk factors, and markers that can help with early detection and individualised treatment plans.

Models for predicting mental health have substantial potential advantages. Early mental health issue detection allows medical professionals to take appropriate action, preventing more serious illnesses and increasing treatment outcomes overall.

In order to predict mental health, machine learning methods such as deep learning, decision trees, and random forest are essential. Large and complicated datasets may be processed by these algorithms, which can also find hidden patterns and forecast the future using learnt patterns from prior data. The potential of machine learning models for mental health is growing along with the area of artificial intelligence.

To sum up, mental health prediction has the power to completely alter the way we approach mental health treatment. Healthcare practitioners may enhance early identification, intervention, and individualised treatment options for those with mental health difficulties by using the potential of artificial intelligence.

2. Related Research

With more people experiencing mental health concerns, mental health is a significant health concern worldwide. It is impossible to estimate the value of early detection and timely intervention in improving outcomes and minimising the burden of mental health concerns.

A rapidly developing subject, mental health prediction research uses a variety of machine learning and predictive modelling approaches to discover and forecast outcomes related to mental health. Here are some important areas of study and papers that go into further detail into mental health forecasting:

Machine learning was utilised in a study that was published in JAMA Psychiatry to forecast depression outcomes using information from electronic health records. The methodology successfully identified patients who were at high risk of developing depression, allowing for early intervention and individualised therapy. (Moshe et al., 2021)

Natural language processing (NLP) techniques have been used by researchers to analyse text data from social media and detect those who are at risk for anxiety disorders. Predictive algorithms can shed light on anxiety-related symptoms by looking for language patterns and emotional clues in social media posts. (Guntuku et al., 2017)

Research has used genetic information to forecast the likelihood of getting schizophrenia. Genetic markers that relate to schizophrenia susceptibility have been identified through polygenic risk score and genome-wide association studies (GWAS). (Cho et al., 2019)

Machine learning algorithms have been used in studies to find trends and risk variables related to suicide risk. To determine a person's chance of attempting suicide, predictive models have been created utilising demographic information, mental health history, and other clinical markers. (Lee & Pak, 2022)

After exposure to stressful events, researchers have focused on predicting the likelihood of developing PTSD. To identify those who are more likely to acquire PTSD, machine learning algorithms have been used to analyse physiological data, biomarkers, and behavioural patterns. (Cooper et al., 2021)

3. Research Questions

- How accurate are these machine learning models in spotting early detection of mental health?
- How can ML models help in early detection of mental health disorders?
- Which factors play an important role in predicting the metal health condition?
- How can machine learning algorithms improve the accuracy of mental health prediction models?

4. Aim and Objectives

The aim and objective of mental health prediction is to use data-driven approaches, such as machine learning and deep learning, to identify patterns, risk factors, and early warning signs related to mental health disorders.

Objectives:

- The primary objective is to identify mental health illnesses early, even before visible signs appear
- The goal of mental health prediction models is to offer individualised treatment plans based on a person's unique traits, medical history, and anticipated reactions to therapies.

- A person's risk of having a mental health condition may be determined using predictive models based on a variety of variables, including genetic predisposition, environmental effects, and behavioural patterns. By being aware of these dangers, precautions can be done to lessen the possibility of mental health issues.
- By identifying high-risk patients who need more intense care and assistance, mental health prediction can help with the effective allocation of healthcare resources. This may result in enhanced healthcare system efficiency and optimal resource utilisation.
- By encouraging the idea of early intervention and preventative care, mental health prediction research can help to lessen stigma around mental health concerns. Predictive models can enhance overall mental health outcomes by proactively urging people to seek treatment.
- A better understanding of the complex connections between multiple risk factors and mental health outcomes might result from the creation and use of mental health prediction models. This may encourage more study and improvements in the field of mental health treatment.

5. Significance of the Study:

In the field of healthcare and mental health research, the study of mental health prediction is of utmost importance. Some of the reasons why the study of mental health prediction is important are:

- Early Intervention and Prevention: The early detection of those who are at risk of developing mental health issues is made possible by predictive models build for mental health prediction. Early intervention and preventative actions may prevent symptoms from getting worse and enhance the effectiveness of treatment.

- Individualised Care: Predictive models allow treatment regimens to be customised based on each patient's specific traits, improving the efficiency and patient focus of mental health care.

- Lowering the cost of healthcare: By avoiding more serious diseases and lowering hospitalisation rates, early identification and customised therapies may be able to lower the overall healthcare expenses related to mental health issues.

- Breaking Stigma: By advocating for early intervention and preventative care, mental health prediction research can aid in the de-stigmatization of mental health conditions. This can motivate people to actively seek assistance and lessen the social stigma associated with mental health difficulties.

6. Scope of the Study

- Data Collection: The research can investigate the utilisation of many data sources, including genetic data, wearable device data, social media activity, and electronic health records. It can concentrate on locating appropriate features and data preparation methods that aid in making precise mental health forecasts.

- Machine Learning Algorithms and Models: The goal of the study is to create prediction models for outcomes related to mental health by examining several machine learning techniques, including deep learning, SVM, decision trees, and ensemble approaches. Finding the best strategy for a given prediction problem may be determined by comparing the performance of multiple models.

- Model Outcome: The model will predict whether treatment is required by the individual or not.

- Model Evaluation: Using proper measures, like accuracy, precision & recall the performance of the models will be accessed. The results of all the models will be compared to find the best model for early detection of mental health.

- Future Research directions: By encouraging the idea of early intervention and preventative care, mental health prediction research can help to lessen stigma around mental health concerns. A better understanding of the complex connections between multiple risk factors and mental health outcomes might result from the creation and use of mental health prediction models. This may encourage more study and improvements in the field of mental health treatment.

7. Research Methodology

The development, validation, and evaluation of predictive models for outcomes related to mental health are all part of the research process for mental health prediction. The following are the main steps in the research methodology:

1. Problem Statement: The research topic and goals for the mental health prediction study should be clearly stated. Determine the target population for prediction as well as the particular mental health disorder of interest.

2. Data Collection: Collect relevant details from a variety of sources, including surveys, wearable technology, social media activity, wearable health records, and genetic data.

3. Data Pre-processing: To manage missing values, eliminate noise, and standardise data formats, clean and pre-process the data. To improve prediction performance, feature engineering involves choosing relevant characteristics, altering data, and developing new variables.

4. Model Selection: Considering the study topic and the features of the data, select the most suited machine learning methods and models. Think about different algorithms like deep learning, random forests, support vector machines, and logistic regression.

5. Training and Testing: To train the prediction model on one subset of the dataset and assess its performance on the other, divide the dataset into training and testing sets. Techniques for cross-validation may be used to evaluate resilient models.

6. Model Evaluation: Use appropriate evaluation criteria like accuracy, precision, recall, area under the ROC curve (AUC-ROC), etc., to rate the prediction models performance.

7. Interpretability: Make sure the prediction models can be understood, particularly in the context of mental health. Utilise approaches to explain models, such as feature significance analysis, to understand the variables influencing predictions.

8. Required Resources:

Python modules & Libraries required:

- Numpy
- Pandas
- Sklearn
- Matplotlib
- Seaborn
- Keras
- Tensorflow

Cloud:

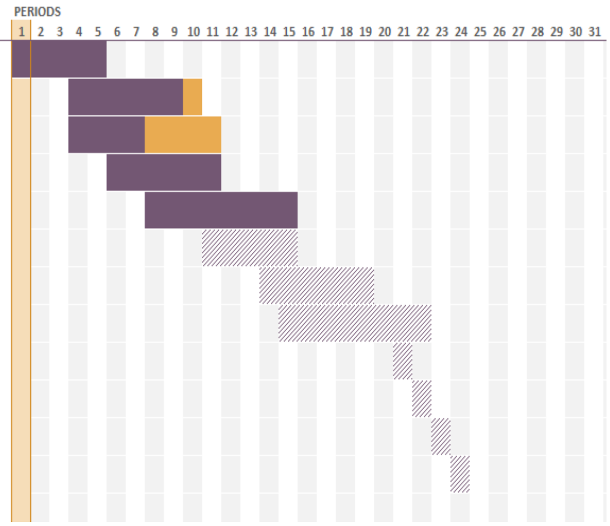
- Google colab
- Nimble Box

Local:

- Graphic card with atleast 4GB memory
- 16GB RAM
- Intel i5 10^{th} gen or higher

9. Research Plan:

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Research Interest	1	5	1	5	100%
Topic Selection	4	6	4	7	100%
Research	4	4	4	8	100%
Data Collection	6	6	6	6	100%
Research Proposal	8	8	8	8	100%
Data Pre processing	11	5	0	0	0%
Model Creation	14	6	0	0	0%
Parameter tuning	15	8	0	0	0%
Comparing Accuracy	21	1	0	0	0%
Selecting best model	22	1	0	0	0%
Final Submission	23	1	0	0	0%
Final Thesis	24	1	0	0	0%



References:

Angehrn, A., Teale Sapach, M.J.N., Ricciardelli, R., Macphee, R.S., Anderson, G.S. & Nicholas Carleton, R. (2020). Sleep quality and mental disorder symptoms among canadian public safety personnel. *International Journal of Environmental Research and Public Health*. [Online]. p.p. 14. Available from: https://pubmed.ncbi.nlm.nih.gov/32326489/.

Chandana, C., Neha, P.N., Nisarga, S.M., Thanvi, P. & Balarengadurai, C. (2023). Mental Health Prediction Using Machine Learning. *Cognitive Science and Technology*. [Online]. p.p. 22. Available from: https://www.hindawi.com/journals/acisc/2022/9970363/.

Cho, G., Yim, J., Choi, Y., Ko, J. & Lee, S.H. (2019). Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investigation*. [Online]. p.pp. 1–8. Available from: https://www.psychiatryinvestigation.org/journal/view.php?doi=10.30773/pi.2018.12.21.2.

Chung, J. & Teo, J. (2022). Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges. *Applied Computational Intelligence and Soft Computing*. [Online]. p.p. 19. Available from: https://www.hindawi.com/journals/acisc/2022/9970363/.

Cooper, A., Horrocks, J., Goodday, S., Keown-Stoneman, C. & Duffy, A. (2021). Predicting the risk and timing of major mood disorder in offspring of bipolar parents: exploring the utility of a neural network approach. *International Journal of Bipolar Disorders*. [Online]. p.p. 9. Available from: https://doi.org/10.1186/s40345-021-00228-2.

Ebert, D.D., Buntrock, C., Mortier, P., Auerbach, R., Weisel, K.K., Kessler, R.C., Cuijpers, P., Green, J.G., Kiekens, G., Nock, M.K., Demyttenaere, K. & Bruffaerts, R. (2018). Prediction of major depressive disorder onset in college students. *Depression and Anxiety*. [Online]. p.p. 11. Available from: https://onlinelibrary.wiley.com/doi/10.1002/da.22867.

Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H. & Eichstaedt, J.C. (2017). Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*. [Online]. p.p. 7. Available from: http://dx.doi.org/10.1016/j.cobeha.2017.07.005.

Lee, J. & Pak, T.Y. (2022). Machine learning prediction of suicidal ideation, planning, and attempt among Korean adults: A population-based study. *SSM - Population Health*. [Online]. p.pp. 1–9. Available from: https://doi.org/10.1016/j.ssmph.2022.101231.

Moshe, I., Terhorst, Y., Opoku Asare, K., Sander, L.B., Ferreira, D., Baumeister, H., Mohr, D.C. & Pulkki-Råback, L. (2021). Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. *Frontiers in Psychiatry*. [Online]. p.pp. 1–12. Available from: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.625247/full.

Papini, S., Pisner, D., Shumake, J., Powers, M.B., Beevers, C.G., Rainey, E.E., Smits, J.A.J. & Warren, A.M. (2018). Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization. *Journal of Anxiety Disorders*. [Online]. p.p. 8. Available from: https://doi.org/10.1016/j.janxdis.2018.10.004.

Wang, X., Li, H., Sun, C., Zhang, X., Wang, T., Dong, C. & Guo, D. (2021). Prediction of Mental Health in Medical Workers During COVID-19 Based on Machine Learning. *Frontiers in Public Health*. [Online]. p.pp. 1–13. Available from: https://www.frontiersin.org/articles/10.3389/fpubh.2021.697850/full.