UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

INDIVIDUAL THESIS

May 2025

Individual Diploma Thesis

PREDICTIVE AND CAUSAL MODELLING OF MENTAL HEALTH USING SUPERVISED LEARNING

Antonia Loizou

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

May 2025

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

PREDICTIVE AND CAUSAL MODELLING OF MENTAL HEALTH USING SUPERVISED LEARNING

Antonia Loizou

Supervisor Professor
George Pallis

The Individual Thesis was submitted for partial fulfilment of the requirements for obtaining a degree in Computer Science from the Department of Computer Science at the University of Cyprus.

May 2025

Acknowledgments

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor George Pallis, for his invaluable support and guidance throughout the entire duration of my thesis. From the very beginning, his willingness to assist and provide direction played a pivotal role in the completion of this work.

I would also like to extend my heartfelt thanks to the doctoral candidate, Mr. George Ioannou, for his continuous support and insightful guidance during the implementation phase. His availability and readiness to address any questions or concerns I encountered were truly appreciated.

I am deeply grateful to both of them for their contributions and unwavering support.

Finally, I would like to thank my family and friends for always being there for me during this journey.

Abstract

This thesis investigates the use of supervised machine learning and causal inference techniques to analyse mental health data and develop both predictive and causal models. The research addresses multiple aspects of mental health, focusing on the probability of individuals seeking treatment, identifying medication use for anxiety and depression, and estimating the risk of eating disorders at the population level based on coexisting mental health disorders. Furthermore, it explores the impact of the COVID-19 pandemic on mental health conditions through causal modelling.

Three real-world datasets were employed, and a comprehensive data preprocessing pipeline was implemented. This included imputation of missing values, encoding of categorical variables, feature scaling, dimensionality reduction, and techniques for handling class imbalance. Various machine learning algorithms were applied to construct and evaluate models for classification and regression tasks, with feature selection and hyperparameter tuning incorporated to enhance performance and generalization.

Causal inference was carried out using the DoWhy framework to estimate the effect of the COVID-19 pandemic on medication usage related to mental health. Visualisations and causal graphs were used to interpret and communicate the findings clearly.

The study demonstrates how machine learning can support mental health research by offering insights into treatment behaviour and condition prevalence. Additionally, it highlights the role of causal analysis in understanding broader social impacts. Ethical considerations and responsible models deployment are emphasized, especially due to the sensitivity of mental health data.

Contents

Chapter 1

Introduction

1.1 Motivation

Mental health has been a major issue for people's well-being for a long time, but even more in recent years. Analysing mental health data has become challenging but equally important area in the field of data analytics. This analysis is made possible through Machine Learning (ML) and Artificial Intelligence (AI) , which are increasingly being applied to areas related to mental health. The ability to analyse such data is crucial as it can help draw important conclusions that support the prevention of mental illnesses.

Personally, as an undergraduate computer science student focused on Artificial Intelligence, I realised that AI and ML can offer valuable insights into the field of mental health. Data analysis using these technologies can help identify important patterns, such as characteristics of people who suffer from mental illnesses which can raise awareness and assist in early prevention. Furthermore, combining mental health data with powerful machine learning models allows for new and effective ways to understand mental health issues and support those who are on medication.

This thesis explores the intersection of artificial intelligence, machine learning, and mental health. It aims to develop predictive models for individuals who are currently taking medication, those who may need medication in the future and to detect the development of eating disorders in a population level. In addition, this research applies causal inference techniques to examine whether the COVID-19 pandemic had an effect on anxiety and depression in relation to medication use.

1.2 Problem Statement

Mental health is a significant issue affecting a large proportion of the world's population. It is important to note that a person's mental health is just as important as their physical health. However, many people still questioned this, resulting in low levels of awareness and support for mental health issues.

Unlike physical illnesses, mental disorders are not always visible. By simply observing a person, it is difficult to know whether they are experiencing a mental health issue. However, by analysing some of their characteristics, it is possible to identify signs of mental disorders. The use of ML and AI, combined with mental health datasets, enables the analysis and prediction of mental health conditions.

In this thesis, several machine learning models were implemented using traditional techniques such as exploratory data analysis, preprocessing, and classification. In particular, these models were developed to predict whether a person would be asked to seek treatment for mental issues. For example, if people are currently taking medication for anxiety or depression, can be predicted at a population level whether someone is likely to develop an eating disorder given other mental health conditions. Additionally, causal modelling was implemented to identify the impact of COVID 19 pandemic to people’s mental health.

1.3 Study Objectives

The primary objective of this thesis is to analyse mental health data using machine learning and causal inference techniques in order to build predictive models and assess the broader impact of external factors on mental well-being. The study focuses on treatment seeking behaviour, medication usage, and the likelihood of developing eating disorders based on mental health disorders. Causal analysis is also applied to examine the effect of the COVID-19 pandemic on mental health outcomes.

The specific objectives of the study are:

1. **Develop Predictive Models:** Develop and evaluate machine learning models that predict whether an individual is likely to seek mental health treatment, based on demographic and workplace related characteristics.

2. **Identify Medication Use Patterns:** Build classification models to determine whether a person is currently taking medication for anxiety or depression behavioural data.
3. **Predict Eating Disorder Risk:** Analyse population-level patterns to predict the likelihood of developing an eating disorder given other existing mental health conditions.
4. **Apply Causal Inference:** Use causal modelling techniques to investigate the impact of the COVID-19 pandemic on the prevalence of anxiety and depression that is related to medication use.
5. **Perform Data Exploration and Feature Engineering:** To employ exploratory data analysis, preprocessing, and feature selection to improve model performance and interpretability.

1.4 Overview of the Thesis Structure

This thesis is organized into 5 chapters, each designed to build upon the information and analysis presented in the previous chapters:

Chapter 1: Introduction

This chapter displays the motivation behind my decision to select the thesis topic and the problem that I addressed. Moreover, it describes the study objectives and the overview of this thesis.

Chapter 2: Background

This chapter explains what is Artificial Intelligence (AI) and Machine Learning (ML). Furthermore, it defines the 3 types of ML, especially supervised learning. It also explains what data analysis tools are used such as Seaborn and NumPy. Also, it defines what is the causal inference technology. In addition, it summarizes a related work to this thesis. Especially, defines what the paper is about and differences between the paper [20] and this thesis.

Chapter 3: Methodology

The chapter 3 provides the research approach and design. Moreover, it explains the preprocessing methods applied to the data, as well as the data visualization techniques used. Additionally, the model's training and prediction processes define in this chapter. Lastly, it details the implementation of the causal inference model.

Chapter 4: Results

Chapter 4 presents the results of this thesis. Especially, the exploratory data analysis and the results are displayed for both classification and regression tasks. The results of the causal inference analysis are also discussed in this chapter.

Chapter 5: Conclusion

This final chapter provides a summary of the thesis, discusses its limitations and challenges, outlines ethical considerations and the responsible use of the developed models, and presents directions for future work.

Chapter 2

Background and Related Work

2.1 Background

Machine learning, which is one of the subsets of Artificial Intelligence, was used extensively in this work. In particular, there are three types of machine learning: supervised learning, unsupervised learning, and reinforcement learning. In this paper, supervised learning has been used, involves learning from labelled data that is divided into training and testing sets. The model is trained using the training data and evaluated using the testing data. Furthermore, one technique that was used is classification. With classification, an input value is taken and assigned to a class or category, depending on the training data provided. A supervised learning classification model predicts which category the data belongs to. Additionally, regression was used, which is applied to continuous values. The main difference between regression and classification models is that regression algorithms are used to predict continuous values, while classification algorithms predict discrete values. The package that is used for machine learning is scikit-learn.

Pandas is a very important Python package for data manipulation and analysis. By using this Python package we can do the significant step for the data analysis which is the exploratory data analysis. It can help the researcher to do several data transformations like sorting rows, taking subsets, calculating summary statistics such as mean, reshaping, DataFrames, and joining DataFrames together. Moreover, Pandas package can be used for importing datasets and clean them. Pandas package works well with other popular Python data science packages such as NumPy, Matplotlib, Seaborn and Plotly. NumPy package is also important for data analysis projects because offers an array data structure that gives more advantages over Python lists, including increased compactness, quicker access for reading and writing items, as well as greater convenience and efficiency.

Another powerful and popular library that was used in data analysis projects, is Matplotlib. Matplotlib library used for data visualization because it creates line plots, bar plots, scatter plots, heatmaps and other useful plots. These are some foundational plots that will allow the scientist to start understanding, visualizing and telling stories about data. Matplotlib offers a high degree of flexibility and customization for generating plots. It involves a significant amount of code to create simpler plots with minimal customizations. In environments where the primary objective is exploratory data analysis, there is a need for numerous rapidly created plots with less focus on visual appeal. Instead of using only Matplotlib, scientists using the library Seaborn which is a great option as it builds on top of Matplotlib to create visualizations more quickly.

Furthermore, another technology that used in this research is Causal Inference. Causal inference involves determining and measuring the causal impact of one variable on another. It entails employing statistical techniques, research designs, and conceptual frameworks to determine causality, considering confounding variables, possible biases, and the constraints of observational data. Moreover, Propensity Score Analysis was used which is a statistical matching technique used to estimate the effect of an intervention by attempting to isolate it from other variables.

2.2 Related Work

2.2.1 Previous Research on Mental Health

The paper [20] presents a critical evaluation review of mental health detection in Online Networks (OSNs) with regards to data sources, machine learning methods and methods of feature extraction. Further, the paper [20] relies on the understanding of the studies, the limitations and challenges. In particular, the paper [20] analyses 22 articles in order to compare both of them in relation to the methodology that was employed and the algorithms that were employed. In general, the paper referred to the need of early detection and treatment of mental health because millions of individuals all over the globe suffer from mental health problems.

2.2.2 Modelling and Prediction in Previous Work

The research [20] discussed above has the comparative nature of other research in the context of methodologies employed, machine learning models developed and model performance.
In detail, most widely employed machine learning models are Random Forests (RF), Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN). However, in the context of measures employed to examine the model performance are accuracy, F1-score, precision, recall and ROC-AUC.

2.2.3 Differences Between the Paper and This Thesis

The paper [20] compares several studies on mental health. In contrast, this thesis focuses mainly on treatment prediction and help-seeking, medication use and causal inference. However, this thesis uses both some identical machine learning algorithms such as Random Forest and other classifiers and evaluation metrics such as precision, recall, accuracy and F1-score.

Chapter 3

Methodology

3.1 Research Approach and Design

The methodology used in this research is based on supervised learning techniques for mental health data analysis. Especially, the research presents the following analytical goals. Firstly, to predict whether an individual is likely to experience mental health issues and seek help. Secondly, to predict whether an individual takes medication for anxiety and therefore, in conclusion, suffers from an anxiety disorder. Thirdly, to predict whether an individual takes medication for depression and therefore, in conclusion, suffers from a depression disorder. These analyses are classification problems. In addition, to estimate eating disorder prevalence based on other mental health conditions at the population level, which is a regression problem. Lastly, I used causal inference techniques to examine whether the COVID-19 pandemic had a causal impact on anxiety medication and whether the COVID-19 pandemic had a causal impact on depression medication. The aforementioned analyses were carried out using a combination of supervised machine learning models, statistical techniques, and a causal inference model to extract significant conclusions from various mental health datasets. Regarding the implementation, my individual thesis was implemented using the Python programming language while the development environment used was Jupyter Notebook.

3.2 Data Preprocessing

Data preprocessing is major for appropriate data analysis. Plenty of methods were used to preprocess the data and as a result different feature sets were created for each combination of methods. Below are the methods that I used to reduce the “noise” of the data:

1. **Simple Imputer:**
   The Simple Imputer is an univariate imputer that fills the missing values with simple strategies. Some of the strategies are descriptive statistics corresponding to mean, median and most frequent. The strategies are used along each column, or using a constant value. The strategy that I used for my data analysis is “most frequent”.

2. **Iterative Imputer:**
   The Iterative Imputer is a Multivariate imputer that estimates each feature from all the others. A strategy for imputing missing values by modelling each feature with missing values as a function of other features in a round-robin fashion.

3. **Label Encoder:**
   The Label Encoder encodes target labels with value between 0 and n_classes-1. I used a Label Encoder in my research because the target variable had non-consecutive numeric labels, and it needed to be transformed into a sequence of consecutive integers for compatibility with machine learning algorithms.

4. **Ordinal Encoder:**
   The Ordinal Encoder convert categorical features as an integer array. The features are converted to ordinal integers. This results in a single column of integers (0 to n_categories - 1) per feature.

5. **One Hot Encoder:**
   The One Hot Encoder creates a binary column for each category and returns a sparse matrix or dense array.

6. **Robust Scaler:**
   The Robust Scaler scales features using statistics that are robust to outliers. This Scaler removes the median and scales the data according to the quantile range. The IQR is the range between the 1st quartile and the 3rd quartile.

7. **Power Transformer:**

The Power Transformer applies a power transform feature wise to make data more Gaussian-like. PowerTransformer supports the Box-Cox transform and the Yeo-Johnson transform. The optimal parameter for stabilizing variance and minimizing skewness is estimated through maximum likelihood.

Box-Cox transform is given by:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} & , if\ \lambda \neq 0 \\ \log y \quad , if\ \lambda = 0 & \end{cases}$$

Yeo-Johnson transformation is given by:

$$\psi(\lambda, y) = \begin{cases} \dfrac{(y+1)^{\lambda} - 1}{\lambda}, & if\ \lambda \neq 0, y \geq 0 \\ \log(y+1), & if\ \lambda = 0, y \geq 0 \\ -\dfrac{[(-y+1)^{2-\lambda} - 1)]}{2-\lambda}, & if\ \lambda \neq 2, y < 0 \\ -\log(-y+1), & if\ \lambda = 2, y < 0 \end{cases}$$

8. **Standard Scaler:**

The Standard Scaler standardizes features by removing the mean and scaling to unit variance.

The standard score of a sample x is calculated as:

$$z = \frac{x - u}{s}$$

where u is the mean of the training samples or zero if with_mean=False, and s is the standard deviation of the training samples or one if with_std=False.

9. **Sequential Feature Selector:**

The Sequential Feature Selector adds or removes features to form a feature subset in a greedy fashion. At each stage, this estimator chooses the best feature to add or remove based on the cross-validation score of an estimator.

10. **Dimensionality Reduction – PCA:**

    Principal Component Analysis is a linear dimensionality reduction using Singular Value Decomposition of the data to project it to a lower dimensional space. The input data is centred but not scaled for each feature before applying the SVD.

11. **Random Over Sampler:**

    Object to over-sample the minority class(es) by picking samples at random with replacement.

12. **Random Under Sampler:**

    Under-sample the majority class(es) by randomly picking samples with or without replacement.

During data preprocessing different feature sets are created containing transformations of the data with the above mentioned techniques.

3.3 Data Visualization

Data visualization is one of the first steps in data analysis for understanding the dataset. In addition, it helps capture various semantic aspects such as the distribution of the data, the correlation between features, the presence of outliers, skewness, and imbalances in feature values. All of these, influenced later decisions regarding feature selection, transformations, and model choice. The following types of the visualizations were engaged:

3.3.1 Distribution Plots

Distribution Plots can be both histograms and count plots that displays if a feature has skewness and does not follow a normal distribution as usual. Moreover, show the distribution of numerical features and as well for categorical features. These helped identify irregularities such as unrealistic age values or unbalanced class distributions.

3.3.2 Correlation Heatmaps

Correlation heatmaps display the relationships between multiple variables. Correlation can be either positive or negative. A positive correlation between two variables means that when one variable increases, the other also increases. Conversely, a negative correlation means that when one variable increases, the other decreases.

3.3.4 Bar Plots

Bar Plots were used to compare class frequencies and highlight feature relevance, particularly in model evaluation. These were useful in determining which traits were effective or underrepresented in the sample.

3.3.5 Confusion Matrix

A confusion matrix is a simple table that shows how well a classification model is performing by comparing its predictions to the actual results. It breaks down the predictions into four categories: correct predictions for both classes (true positives and true negatives) and incorrect predictions (false positives and false negatives).

3.3.6 Learning Curve

A learning curve graphically depicts how a process improves through learning and increased proficiency.

3.3.7 Actual vs Predicted Plot

A Predicted vs Actual plot is a scatter plot used to visualize the performance of a regression model. The x-axis shows the actual values, while the y-axis shows the predicted values. If the model makes perfect predictions, all points will lie on a straight diagonal line with a slope of 1.

3.3.8 Residual Plot

A residual plot is a useful tool in regression that shows how far off the predictions are from the actual values. It plots the difference (called residuals) between the predicted and actual values. These differences are shown against the predicted values or input features. If the model is good, the points will be spread out randomly around the line y = 0. If you see a pattern, it might mean the model missed something in the data.

3.4 Model Training and Prediction

After creating the different feature sets different machine learning models were used for predictions for both classification problems and regression problems.

Classification Machine Learning Algorithms:

1. **Random Forest:**

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

2. **AdaBoost:**

An AdaBoost classifier is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

3. **XGBoost:**

XGBoost is a distributed gradient boosting toolkit that is optimized for efficiency, flexibility, and portability.It employs machine learning methods within the Gradient Boosting framework.XGBoost delivers parallel tree boosting to address numerous data science issues quickly and accurately.

4. **CatBoost:**

CatBoost is based on the gradient boosting technique, which builds decision trees consecu tively to reduce errors and improve predictions. The approach works by building a decision tree and determining how much error is present in forecasts.

5. **SVC**:

The Support Vector Classifier is a supervised machine learning model that utilizes the Su pport Vector Machine method. It is mostly used for classification jobs and works by determining the optimum hyperplane to segregate data points from distinct classes in a high-dimensional space.

6. **K-Nearest Neighbors:**

K-Nearest Neighbors is a simple way to classify things by looking at what’s nearby.

7. **LogisticRegression:**

Logistic regression is a supervised machine learning technique used in classification problems to predict whether an instance belongs to a specified class or not. Logistic regression is a statistical procedure that examines the relationship between two data variables.

8. **DecisionTree:**

A Decision Tree splits data into smaller groups based on feature values, forming a tree-like structure where each internal node represents a decision rule, and each leaf node represents a class label.

9. **GaussianNB:**

GaussianNB is a classification algorithm that uses Baye's Theorem and assumes that feat ures have a normal distribution.It is a member of the Naive Bayes family, which asserts that all features are independent of the class label.

Regression Problems Machine Learning Algorithms:

For regression problems, I employed models like Random Forest Regressor, AdaBoost, XGBoost Regressor, KNeighbors, DecisionTree, and Support Vector Regressor. These methods are essentially the regression counterparts to their corresponding classification models.While the underlying structure and training method are similar, the primary distinction is that regression models predict continuous numerical values rather than discrete class labels.

Afterwards, I applied all of the algorithms to each feature set to compare their performance before selecting the 2 top performing feature sets and algorithms. The training was performed using Scikit-learn pipelines, which allowed for a clean and efficient integration of preprocessing steps and model fitting. A pipeline enables the successive application of a list of transformers to preprocess data, with the option of concluding the sequence with a final predictor for predictive modelling. In my research, I used pipelines to produce the 2 best feature sets. Then, I set up a range of values for the hyperparameters of the 2 best-performing algorithms. To prevent data leakage and increase generalization, the training method employed k-fold cross-validation, usually with 10 folds. The models were evaluated using measures like as accuracy, precision, recall, and F1-score based on the type of prediction task. Later, I utilized GridSearchCV, which performs an exhaustive search over the specified parameter values for an estimator. GridSearchCV also includes methods for fitting and scoring. Once GridSearchCV was completed, I determined the best-performing combination and tested it on the test set by generating graphs such as the confusion matrix and actual vs. predicted values.

3.5 DoWhy: Causal Analysis Framework

"DoWhy" is a Python package designed to promote causal thinking and analysis. DoWhy offers a rigorous four-step interface for causal inference that emphasizes not only openly modelling causal assumptions but also validating them to the greatest extent possible. DoWhy enables estimating the average causal effect for backdoor, frontdoor, instrumental variable, and other identification approaches, as well as estimating the conditional effect (CATE) via an integration with the EconML library.

Specifically, in my research the purpose of the causal analysis was to examine the case that the COVID-19 pandemic affected both anxiety and depression medications. This approach allows the assessment of causal associations by incorporating causal graphical models and statistical estimations.

3.5.1 Data Preparation

Missing values of the features are impute by using Iterative Imputer. The anxiety and depression medications were compared before and after the COVID-19 pandemic.

3.5.2 Causal Model Description

Using DoWhy a CausalModel was defined, which is identified below:

1. **Treatment:**
   The feature that could affect the result is referred to as the treatment. The COVID-19 pandemic is the treatment in this case.

2. **Outcome:**
   The outcome refers to the feature we are examining whether it is affected by the treatment. The outcome in my research are anxiety medication and depression medication.

3. **Confounders:**
   Additional features that could affect outcome and treatment.

3.5.3 Considering and Estimating the Causal Effect

In order to choose the appropriate collection of variables for adjustment, I employed the causal graph and selected the backdoor criterion. Once observed, the Average Treatment Effect (ATE) of COVID-19 on anxiety and depression medication use, was calculated using linear regression, especially the backdoor.linear_regression method in DoWhy. Consequently, the results that came out from the causal model estimates measurements to the extent of the use of medications for depression and anxiety during and after the pandemic period.

3.5.4 Visualization and Analysis

The plots I created show the trends in medication use before and after the COVID-19 pandemic, the average medication that is used during these periods and finally they provide insights into how treatment for anxiety and depression were affected by the pandemic. A 95% confidence interval was included with the estimated causal effect, giving a range that the genuine effect is probably inside.

Chapter 4

Results

4.1 Datasets

4.1.1 Dataset 1 : Mental Health in Tech Dataset

The first dataset used in this research to predict whether an individual is likely to experience mental health issues and seek help is the 2014 Mental Health in Tech Survey, conducted by Open Sourcing Mental Illness (OSMI). The survey aims to measure attitudes toward mental health and the availability of mental health resources in the technology workplace. It is publicly available on Kaggle and includes 1,259 responses, primarily from individuals working in the tech sector. The dataset includes a wide range of variables related to employment context, mental health history and workplace culture. All responses are self-reported, and the dataset is cross-sectional. The following features were used for the prediction:

Age	Respondent's age.
Gender	Respondent's gender.
family_history	Indicates whether the respondent has a family history of mental illness. The responses are typically Yes and No.
Treatment	Indicates whether the respondent has sought treatment for a mental health condition. The responses are typically Yes and No.
Benefits	Indicates whether the respondent's employer provides mental health benefits. The responses are typically Yes, No

	and Don’t Know.
care_options	Indicates whether the respondent knows the options for mental health care that the employer provides. The responses are typically Yes, No and Not Sure.
wellness_program	Indicates whether the employer has ever discussed mental health as part of an employee wellness program. The responses are Yes, No and Don’t Know.
seek_help	Indicates whether the employer provides resources to learn more about mental health issues and how to seek help. The responses are Yes, No and Don’t Know.
Anonymity	Indicates whether the respondent's anonymity is protected if they choose to use mental health or substance abuse treatment resources. The responses are Yes, No and Don’t Know.
mental_vs_physical	Indicates whether the respondent feels that their employer takes mental health as seriously as physical health. The responses are Yes, No and Don’t Know.

Table1 : *Features of the Mental Health in Tech dataset summarized.*

The feature used as the target variable was treatment, as this provided the basis for the prediction. Preprocessing steps such as handling missing values and encoding categorical features were applied, as described in Chapter 3.

4.1.2 Dataset 2 : National Health Interview Survey (NHIS) Questionnaire

The second dataset I used for my research is a combination of five datasets from 2019 to 2023, representing five distinct years. This dataset is used to predict whether a person takes medication for anxiety, which serves as a proxy for having an anxiety disorder. It is also used to predict whether a person takes medication for depression, consequently has a depression disorder. Additionally, it is used in the causal inference analysis to explore the impact of the COVID-19 pandemic on mental health treatment. The NHIS is conducted yearly by the National Center for Health Statistics (NCHS), part of the Centers for Disease Control and Prevention (CDC). It is one of the principal sources of information on the health of the civilian, noninstitutionalized U.S. population. The survey uses a multistage probability sampling design and collects data via in-person interviews conducted by trained U.S. Census

Bureau interviewers. As a result, the dataset is nationally representative, enabling generalizable insights about U.S. adults. The variables selected for prediction were the most relevant to mental health. In 2019, the datasets contain 31,997 records, 31,568 for 2020, 29,482 for 2021, 27,651 for 2022 and 29,522 for 2023. The following features were used for the aforementioned analyses:

ANXEV_A	Indicates whether the respondent has ever been told by a doctor or other health professionals that had any type of anxiety disorder. The responses are Yes, No, Refused, Not ascertained, and Don't Know.
DEPEV_A	Indicates whether the respondent has ever been told by a doctor or other health professionals that had any type of depression disorder. The responses are Yes, No, Refused, Not ascertained, and Don't Know.
ANXFREQ_A	Indicates how often the respondent feels worried, nervous and anxious. The responses are Daily, Weekly, Monthly, A few times a year, and Never.
ANXMED_A	Indicates whether the respondent take prescription medication for anxiety feelings. The responses are Yes, No, Refused, Not ascertained, and Don't Know.
ANXLEVEL_A	Indicates the level of the feelings that the respondent felt the last time. The responses are a little, a lot, somewhere between a little and a lot, Refused, Not ascertained, and Don't Know.
DEPFREQ_A	Indicates how often the respondent feels depressed. The responses are Daily, Weekly, Monthly, A few times a year, Never, Refused, Not ascertained, and Don't Know.
DEPMED_A	Indicates whether the respondent takes prescription medication for depression. The

	responses are Yes, No, Refused, Not ascertained, and Don't Know.
DEPLEVEL_A	Indicates the level of the depression that the respondent felt the last time. The responses are a little, a lot, somewhere between a little and a lot, Refused, Not ascertained, and Don't Know.
MHRX_A	Indicates whether the respondent took prescription medication in the past 12 months to help him with any other emotions or with his concentration, behaviour or mental health. The responses are Yes, No, Refused, Not ascertained and Don't Know.
MHTHRPY_A	Indicates whether the respondent received counselling or therapy from a mental health professional such as a psychiatrist, psychologist, psychiatric nurse or clinical social worker in the past 12 months. The responses are Yes, No, Refused, Not ascertained, and Don't Know.
MHTHDLY_A	Indicates whether the respondent had delayed getting counselling or therapy from a mental health professional because of the cost in the past 12 months. The responses are Yes, No, Refused, Not ascertained, and Don't Know.
MHTHND_A	Indicated whether the respondent needed counselling or therapy from a mental health professional and did not get it because of the cost in the past 12 months. The responses are Yes, No, Refused, Not ascertained, and Don't Know.

Table 2: *Summary of the National Health Interview Survey features.*

Note: *The responses to each question in the dataset were 1, 2, 3, 4, 5, 7, 8, 9. These responses correspond to different values based on the question, as shown in Table 2.*

The features chosen as target variables were ANXMED_A and DEPMED_A, which served as the foundation for predictions. Preprocessing tasks, such as managing missing values and encoding categorical features, were carried out as outlined in Chapter 3. The treatment variable in the causal inference analysis was the COVID-19 pandemic indicator, which was utilized to analyse its impact on medicine use.

4.1.3 Dataset 3 : Global Mental Health Disorders

The third dataset used in this study aims to estimate the prevalence of eating disorders based on the occurrence of other mental health conditions at the population level. The dataset, which is available on Kaggle, compiles global statistics on various mental health disorders, including schizophrenia, bipolar disorder, eating disorders, anxiety disorders, drug use disorders, depression and alcohol use disorders. It contains data from several countries for the period 1990 to 2017. The data come from Our World in Data, which compiles information from the Global Burden of Disease (GBD) studies conducted by the Institute for Health Metrics and Evaluation (IHME). These studies provide standardized estimates of disease prevalence globally and are widely used in public health research. The original dataset consisted of 108,553 records, but a filtered subset of 6,486 records was used for this study. The selection focused on series containing prevalence rates of mental health conditions by country and year. Records that included only general population data over time were excluded, as they contained information on each country's population by year but not the corresponding mental disorder prevalence rates. As a result, they were not relevant to the forecasting task. This dataset supports a cross-national analysis of how the prevalence of different mental health conditions may correlate with the presence of eating disorders. The following features were used for the prediction:

Schizophrenia (%)	The percentage of people with schizophrenia in the country or region.
Bipolar disorder (%)	The percentage of people with bipolar disorder in the country or region.
Eating disorders (%)	The percentage of people with eating disorders in the country or region.
Anxiety disorders (%)	The percentage of people with anxiety disorders in the country or region.
Drug use disorders (%)	The percentage of people with drug use disorders in the country or region.
Depression (%)	The percentage of people with depression in

	the country or region.
Alcohol use disorders (%)	The percentage of people with alcohol use disorders in the country or region.

***Table 3**: Summary of Global Mental Health Disorders features.*

4.2 Classification Results

4.2.1 Prediction of Mental Health Issues and Help-Seeking Behaviour

The prediction is performed using Dataset 1, where the target variable is the “treatment” feature.

4.2.1.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is a valuable method for understanding the data and determining the techniques for preprocessing.

1. **Age Distribution Plot**

   The figure[1] displays the distribution of people across various age groups. The X-axis indicates the range of ages while the Y-axis shows the number of people in each age group. The graph indicates that the age distribution follows a right-skewed pattern. The ages that occur most frequently are 29 and 32. After age 40, the count of people at the subsequent ages declines, illustrating a long tail. The long tail in the graph indicates that older individuals are less frequently represented in the dataset.

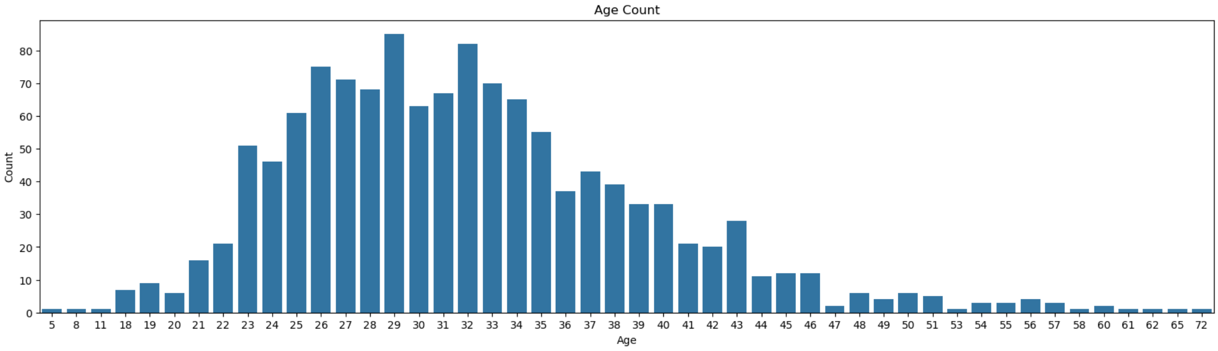


***Figure 1:** Age Distribution Plot*

2. Age Group Distribution by Gender

The figure [2] displays how individuals are distributed across different age groups, separated by gender. The X-axis is divided into 6 bins, each one of them representing a different age group, while the Y-axis shows the number of individuals in each group. The most frequent group is males in the 27–38 age range. Older age groups are less frequently represented in the dataset.

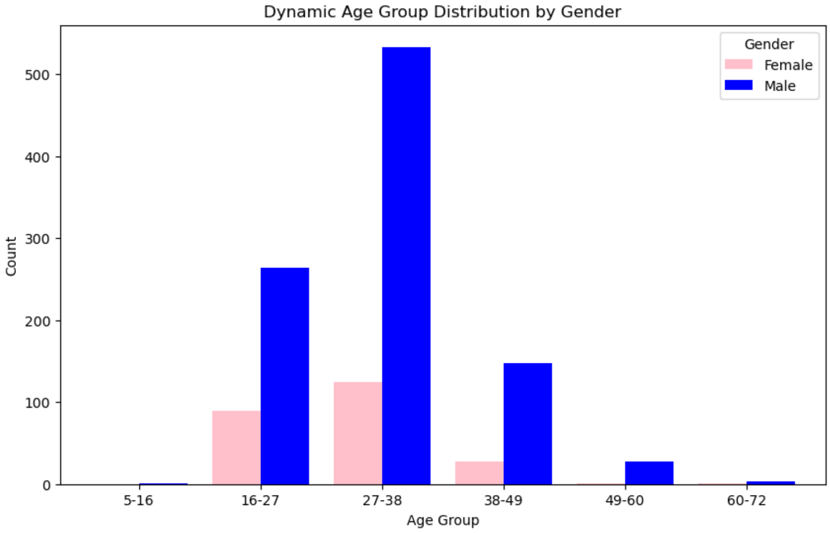


Figure 2: *Age Group Distribution by Gender Plot*

3. Boxplot of Age Distribution by Gender

The figure [3] below displays the age distribution for females and males, enabling for a comparison of means, distribution and outliers. The median age for women is below 30, while for men it exceeds 30. Additionally, the age distribution is comparable between genders. There are more outliers among males than females, indicating that older men outnumber older women.

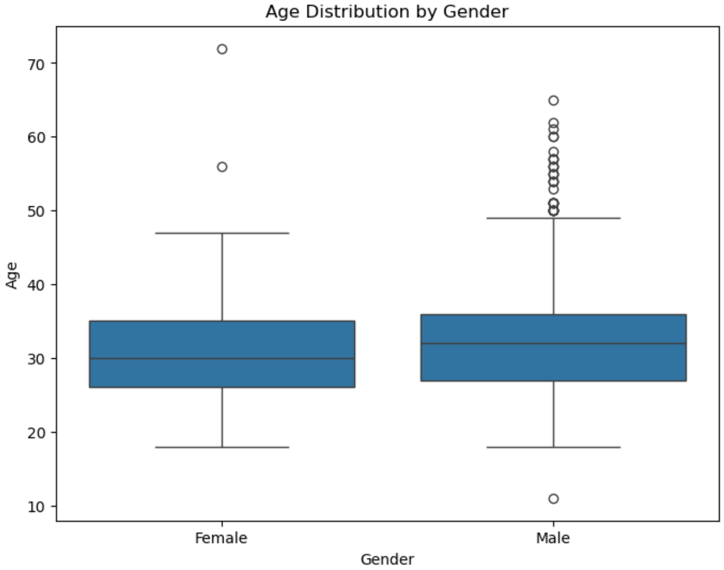


Figure 3: *Box Plot of Age Distribution by Gender Plot*

4. Age Group Distribution by Family History

The figure [4] displays how individuals are distributed across different age groups, based on their responses to whether they have a family history. The X-axis divided in 6 bins, each bin represents different age group. The Y-axis displays the count of individuals in each group. The age group 27-38 has the highest count for both genders. The least frequent bins are the 5-16 and 60-72 age ranges.

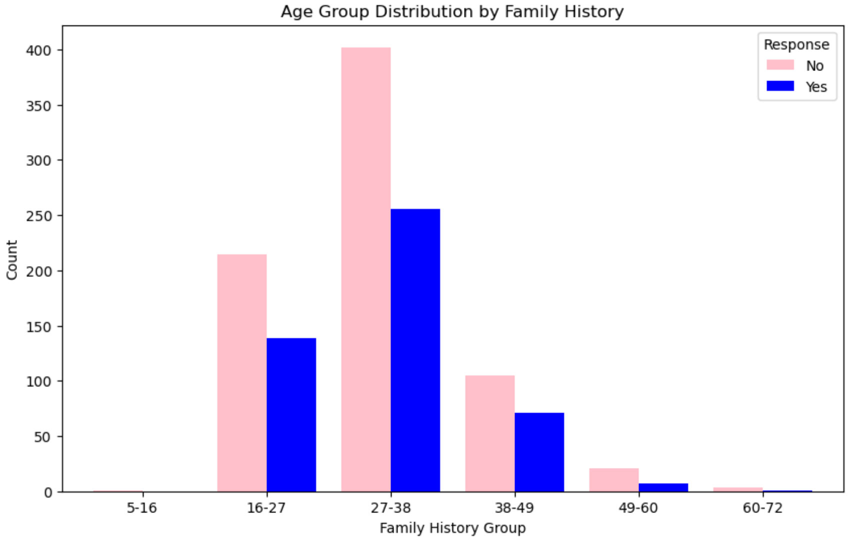


Figure 4: *Age Group Distribution by Family History Plot*

5. Age group Distribution by Treatment

The figure [5] displays how individuals are distributed across different age groups, based on their response to whether they have seek treatment for a mental health condition. The X-axis divided in 6 bins, each bin represents different age group. The Y-axis displays the count of individuals in each group. The age group 27-38 has the highest count for both genders. The least frequent bins are the 5-16 and 60-72 age ranges.

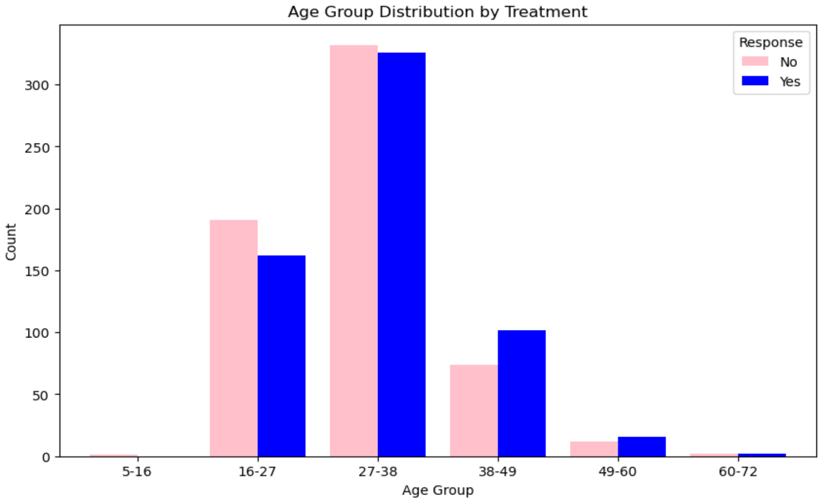


Figure 5: *Age Group Distribution by Treatment Plot*

4.2.1.2 Classifiers and Feature Sets Comparison

Before starting the process, I run all possible classifier and feature set combinations. The following graphs show the mean accuracy for classifiers and feature sets:

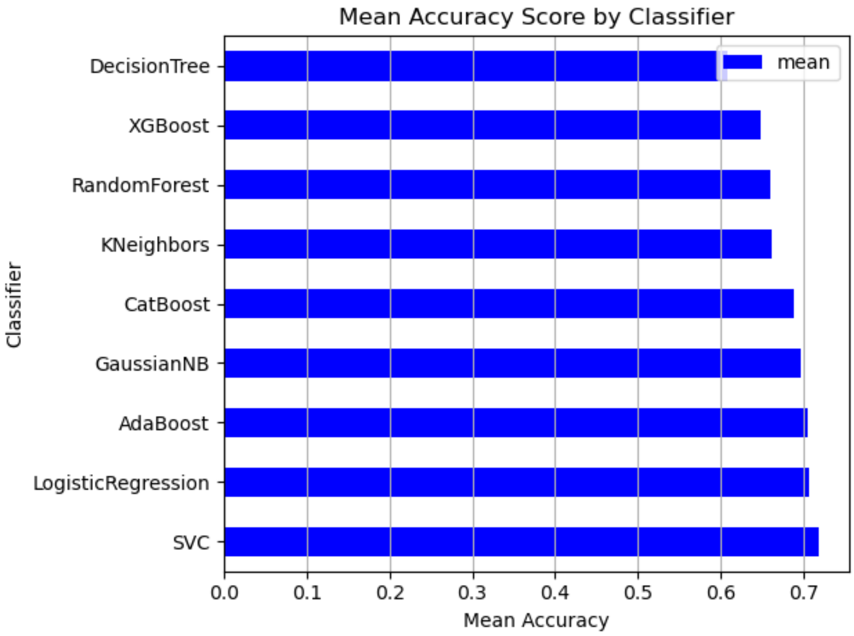


Figure 6: *Mean Accuracy By Classifiers Plot 1*

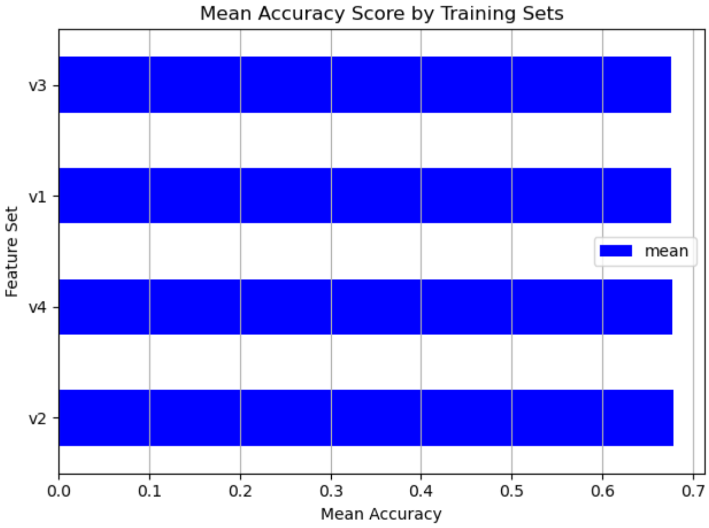


Figure 7: *Mean Accuracy By Feature Sets Plot 1*

The feature sets and classifiers that I chose were the top 2 of each. The two classifiers I used were the SVC and LogisticRegression. Furthermore, the two feature sets I chose were V2 and V4. The feature set V2 was pre-processed using PowerTransformer with the "yeo-johnson" approach for numerical features and Ordinal Encoding for categorical features. The feature set V4 was pre-processed using PowerTransformer with the "yeo-johnson" approach for numerical features, One Hot Encoding for categorical features and dimensionality reduction with PCA and 13 components because they explained 95% of variance. The PowerTransformer was used so as to normalize the skewed numerical features, which helps improve the performance of many models such as Logistic Regression and SVC that are sensitive to feature scale and distribution. Furthermore, Ordinal Encoding was used to convert the categorical features into numerical format and that might create an order where none exists. Additionally, One Hot Encoding was used for converting the categorical variables into a binary format which is more suitable for many classifiers. Lastly, the PCA dimensionality reduction was used for reducing the "noise" in the data but without losing information about the data and that helps the model generalize.

4.2.1.3 Model Performance of Selected Classifiers and Feature Sets

My next step was to use pipelines to construct feature sets V2 and V4, and then run GridSearchCV for both SVC and LogisticRegression classifiers on each of them. The following table contains the performance of the combined classifiers and feature sets:

GridSearchCV Results: Top Classifiers and Feature Set Combinations			
Classifier	**Feature Set**	**Cross-Validation Score**	**F1-score**
LogisticRegression	V2	0.71	0.69
SVC	V2	0.73	0.70
LogisticRegression	V4	0.72	0.69
SVC	V4	0.72	0.70

***Table 3:** GridSearchCV Results: Top Classifiers and Feature Set Combinations 1*

Among the above combinations, SVC with feature set V2 achieved the highest cross-validation score and F1-score. This model was accordingly chosen for the final classification pipeline and subsequent analysis. The final model is SVC with an RBF kernel and hyperparameters C=1, gamma = 'scale'. The model was trained on feature set V2 which was pre-processed with PowerTransformer on numerical features. That gave the ability to the model to capture patterns, especially for SVC because is sensitive to feature scale. Furthermore, it seems that the Ordinal Encoding worked properly in this case because appears to have preserved useful structure. The model's generalization ability was evaluated using both the training and test datasets, with F1-scores of **0.74** and **0.69**, respectively. These results indicate strong performance with no significant overfitting. I calculated the F1-score because it represents the harmonic mean of the precision and recall, where an F1-score reaches its best value at 1 and worst score at 0.

$$F1 = \frac{2 * TP}{2 * TP + FP + FN}$$

4.2.1.4 Comprehensive Evaluation of the Final Model (SVC with Feature Set V2)

i) Confusion Matrix

The confusion matrix [8] below displays the distribution of actual vs predicted labels for the test set. The model classified 83 true positives, 85 false positives, 83 true negatives, and 85 false negatives. It recognized successfully the majority of those seeking help, however some were classified incorrectly as not seeking treatment. These symmetrical misclassifications suggest that the model does not exhibit strong bias toward either class.

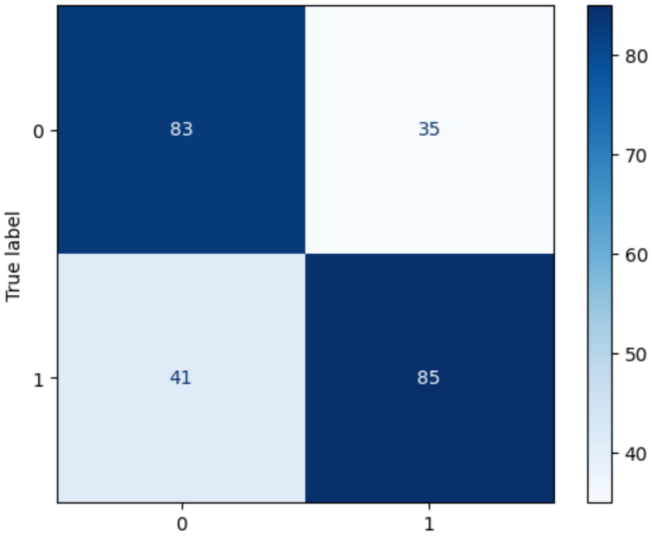
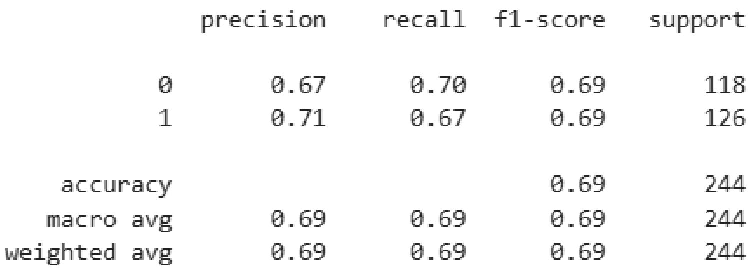


Figure 8: *Confusion Matrix*

ii) Classification Report

The classification report helped me to understand critical conclusions regarding the performance of the best model, particularly how well it classified individuals seeking assistance based on precision and recall.



```
              precision    recall  f1-score   support

           0       0.67      0.70      0.69       118
           1       0.71      0.67      0.69       126

    accuracy                           0.69       244
   macro avg       0.69      0.69      0.69       244
weighted avg       0.69      0.69      0.69       244
```

Figure 9: *Classification Report for the Prediction of Individuals Seeking Help*

The insights that had been understand from the classification report [9] was that the 2 classes are balanced based on the values of precision and recall are relatively similar. This balance suggests that the model is similarly effective at recognizing both categories without showing preference for either ,which is crucial in ensuring fair and unbiased predictions. Moreover, based on the recall 67% of individuals in class "Yes" were classified correctly. Respectively, for class "No", 71% of the individuals were classified correctly. Additionally, based on the precision for class "Yes", 67 out of 100 people predicted correctly. Similarly, based on the precision for class "No", 71 out of 100 people predicted correctly.

iii) Learning Curve

The learning curve graph [10] helped me understand how my model is performing. The model displays satisfactory generalization with a balanced performance on both training and validation sets. Besides, the training score is elevated when employing a small portion of the data, as the model can easily adapt to the restricted instances. On the other hand, when the size of the training set increases, the training score decreases due to increased difficulty in fitting more varied data. At first, the line of the validation score starts low because of the model has not seen enough data but steadily improves. As the 2 curves converge we can say that the model is generalizing well and does not overfit.

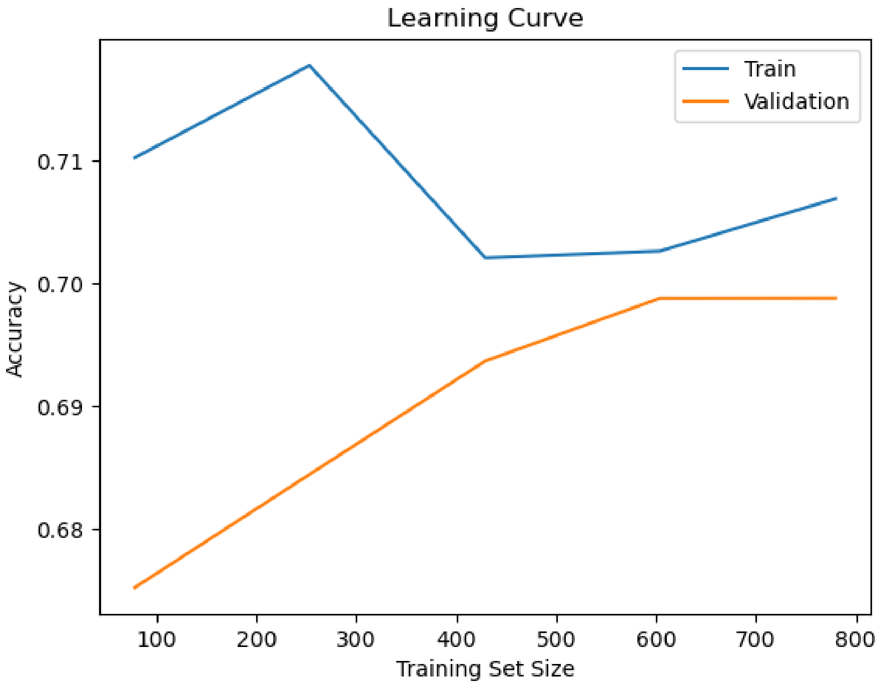


Figure 10: *Learning Curve Plot 1*

iv) Feature Importance

The feature importance plot [11] shows how each feature impacts the target variable in the prediction. In this case, the features that play an important role for the "treatment" prediction are family history, Gender and Age because of the high importance scores. This result is reasonable, as people with family history of mental health issues may be more likely to seek treatment because of the greater awareness and concerns. or worry. Similarly, gender and age may indicate underlying social or psychological elements that affect the tendency to seek help. The other features suggests that they do not influence significantly the target variable "treatment".

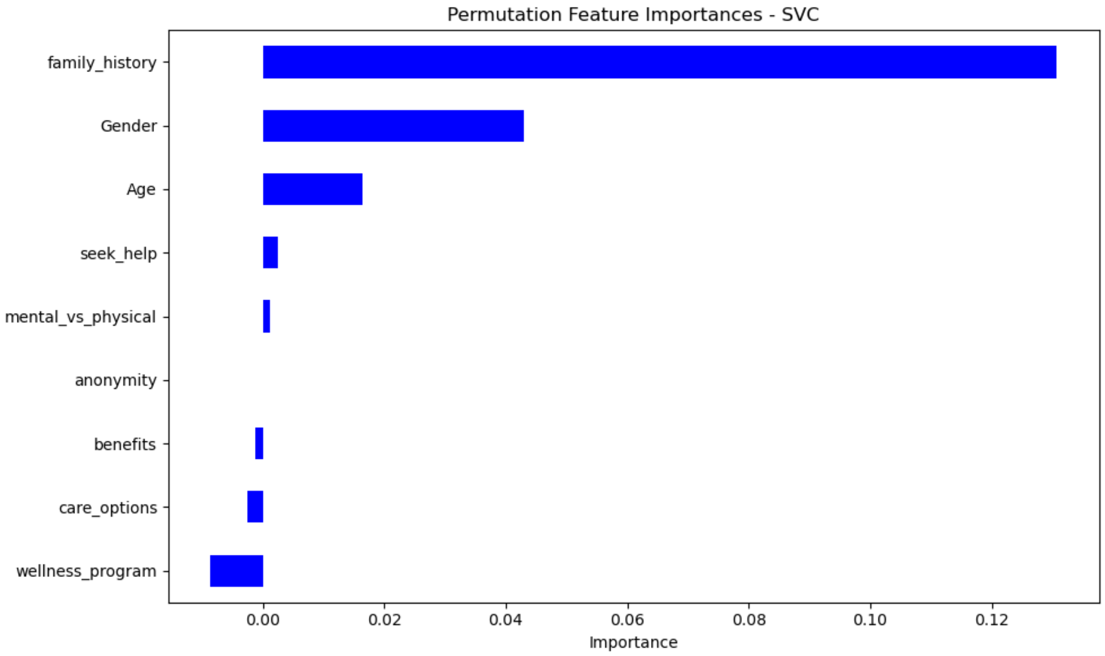


Figure 11: *Feature Importance Plot 1*

4.2.1.5 Significance of the Prediction

To sum up, being able to forecast if an individual will have mental issues and seek treatment is crucial, since issues pertaining to mental health are something that many people face across the globe. To this end, I forecast 69% that an individual will have issues and seek treatment. The importance of this prediction is that this person might need special interventions or support, particularly in the workplace where the mind tends to be ignored.

4.2.2 Prediction whether an individual takes medication for anxiety.

The prediction is performed using Dataset 2, where the target variable is "ANXMED_A" feature.

4.2.2.1 Exploratory Data Analysis (EDA)

1. **Anxiety Medication Usage Distribution**

   The figure [12] displays the anxiety medication usage distribution. The X-axis represents the types of the responses and the Y-axis shows the number of individuals corresponding to each response. The most common response is "No", suggesting that most people do not use anxiety medication. The significant imbalance between "Yes" and "No" answers might be affected by multiple factors, such as restricted access to care, underdiagnosis, or society stigma linked to mental health and the use of medication. These elements might lead individuals to not take medication or not to reveal it in the survey.

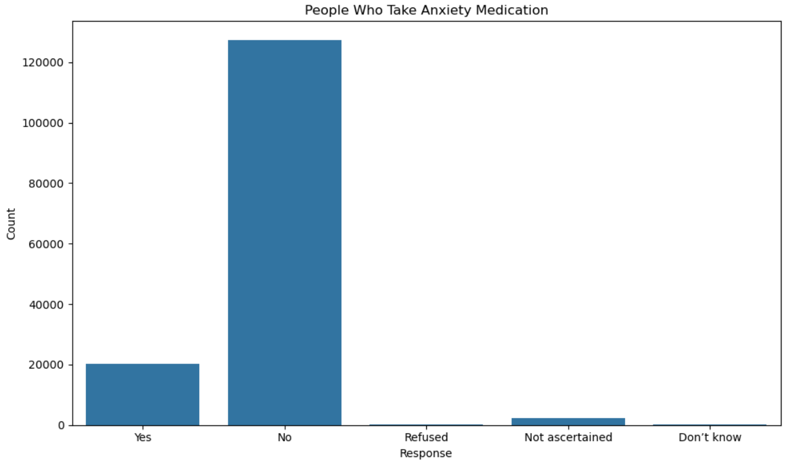


Figure 12: *Plot of Anxiety Medication Usage Distribution*

2. **Anxiety Frequency Distribution**

   The figure [13] below shows the anxiety frequency distribution. The X-axis corresponds to the possible answers and the Y-axis displays the number of individuals corresponding to each answer. The 2 most common answers to the question about how often individuals experience feelings of anxiety are "A few times a year" and "Never". Moreover, the responses "Daily" , "Weekly", "Monthly" have exhibit comparable numbers, indicating that a large segment of people do face anxiety feelings on a regular basis. This might indicate wider social issues like economic instability, stress and worries about the future.

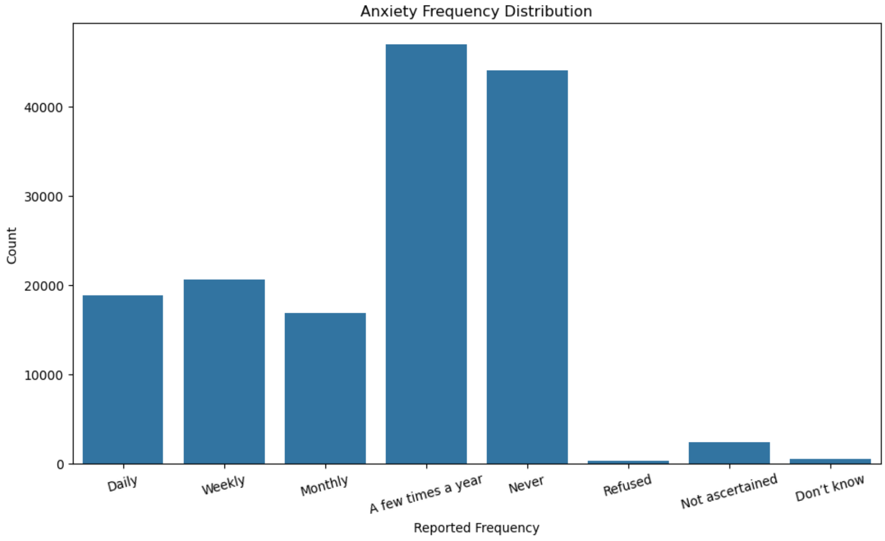


Figure 13: *Plot of Anxiety Frequency Distribution*

3. Anxiety Ever Distribution

The figure [14] displays the distribution of whether individuals have ever been diagnosed with an anxiety disorder. The X-axis corresponds to the response categories and the Y-axis corresponds to the number of individuals that who have ever been told by a mental health professional that they have any type of anxiety disorder. The most common response is "No" and the second most common response is "Yes". The other 3 responses indicates that most respondents gave a clear answer. The number of responses for response "No" may suggest an underdiagnosis of anxiety disorders.

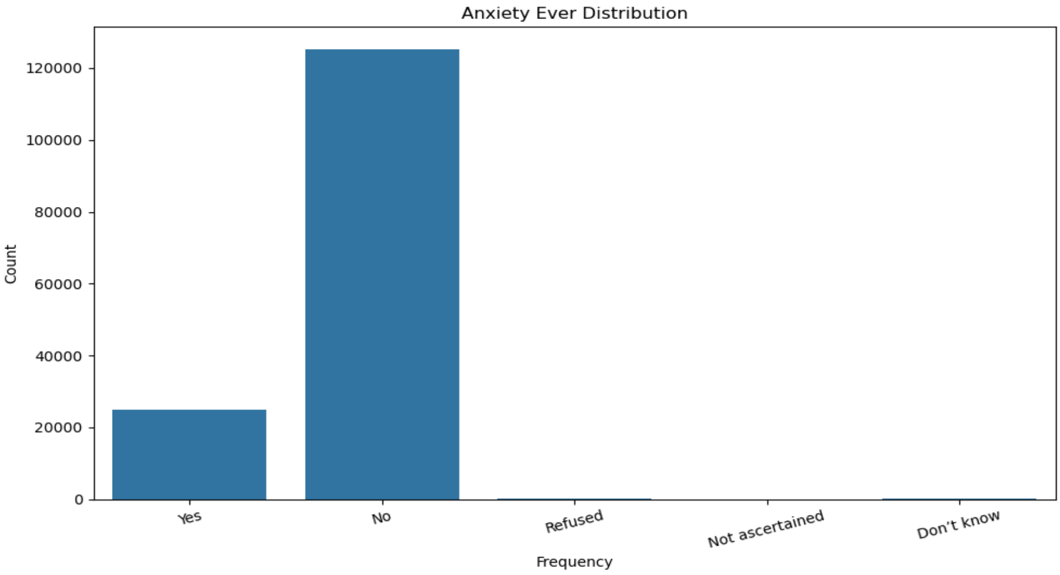


Figure 14: *Plot of Anxiety Ever Distribution*

4. **Anxiety Level Distribution**

The figure [15] shows the distribution of Anxiety Level. The X-axis presents the responses and the Y-axis presents the number of individuals of each response. The 2 more frequent responses are the "A little" and "Somewhere in between a little and a lot". This pattern might indicate the existence of persistent minor stressors in everyday life, including job demands, financial instability, or social difficulties, which results in moderate anxiety without necessarily causing clinical diagnoses.

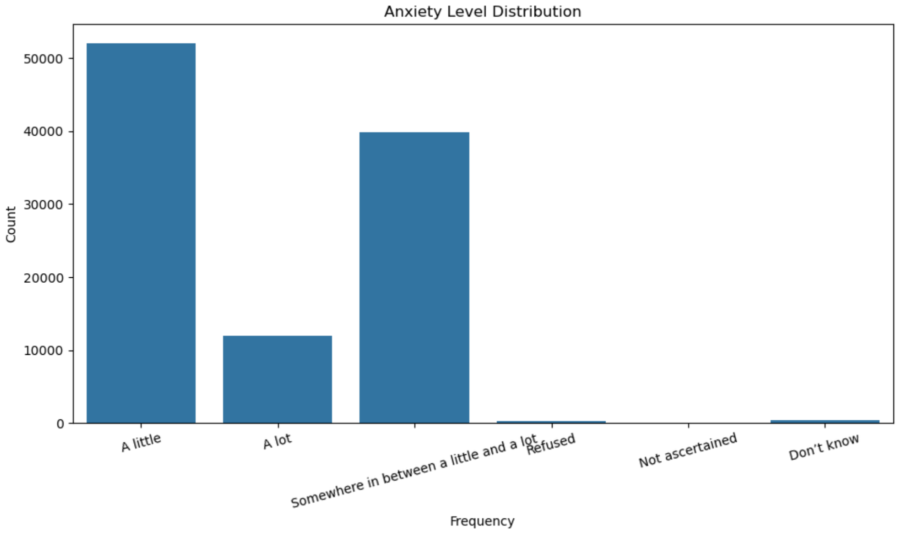


Figure 15: *Plot of Anxiety Level Distribution*

5. **Medication for emotions or Mental Health over the last year Distribution**

The figure [16] below shows the distribution plot of medication for emotions or Mental Health over the last year. The X-axis corresponds to the responses and the Y-axis corresponds to the number of individuals of each response. The majority class is "No" and that explains that there is low access to mental health care.

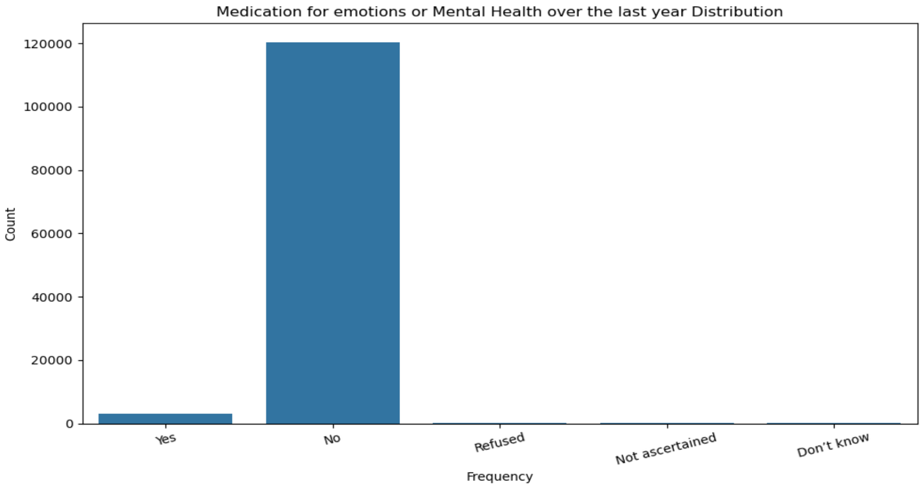


Figure 16: *Plot of medication for emotions or Mental Health over the last Distribution*

6. Therapy from a professional over the last year Distribution

The figure [17] represents the distribution of the individuals that take therapy from a professional over the last year. The common answer is “No”. The answer “Yes” had also more responses based on the other 3 answers. This imbalance might be a result of several reasons, specifically the cost of the therapy, the social stigma and the underdiagnosis.

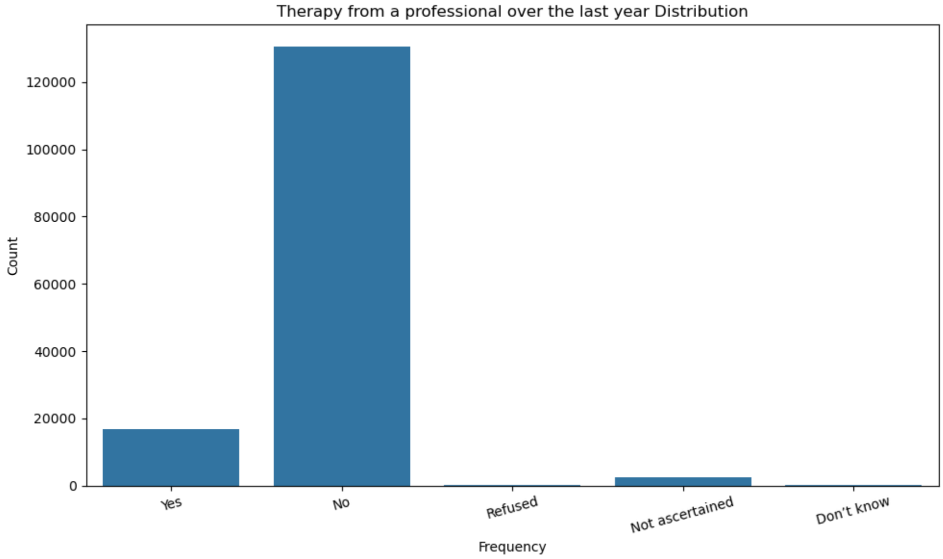


Figure 17: *Plot of Therapy from a professional over the last year Distribution*

7. Delayed Therapy from a professional over the last year Distribution

The below graph [18] shows the distribution of the responses regarding delayed therapy from a professional over the last year. The X-axis presents the types of the responses and Y-axis presents the number of individuals on each response. The most frequent response is “No” and the less frequent responses are “Yes”, “Refused” , “Not ascertained” and “Don’t Know”. This, indicate an insufficient mental health infrastructure world-wide.

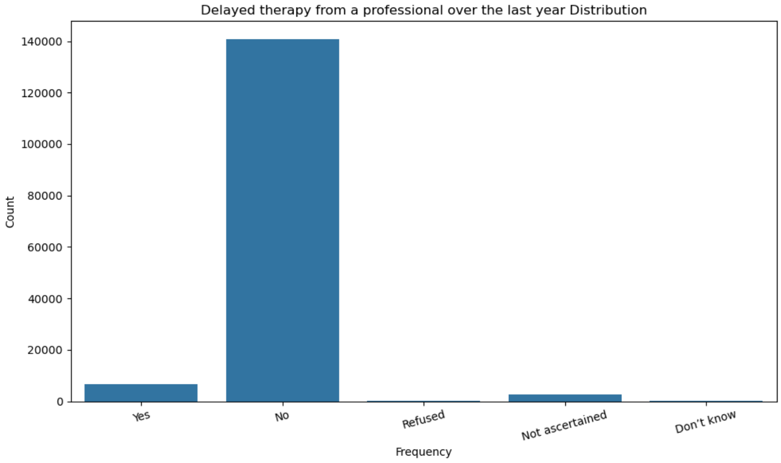


Figure 18: *Plot of Delayed Therapy from a professional over the past year Distribution*

8. Distribution of Responses Regarding No Therapy Due to Cost Over the Last Year

The figure [19] displays the distribution of the responses regarding no therapy due to cost over the past year. The X-axis corresponds to possible answers and the Y-axis corresponds to the number of individuals in each answer. The plot shows huge imbalance between the answers with the response "No" to be the most common answer. This implies that for the majority of people , expenses did not hinder their access to therapy. Nonetheless, the occurrence of "Yes" answers indicates that financial limitations may still hinder certain individuals from accessing mental health therapy.

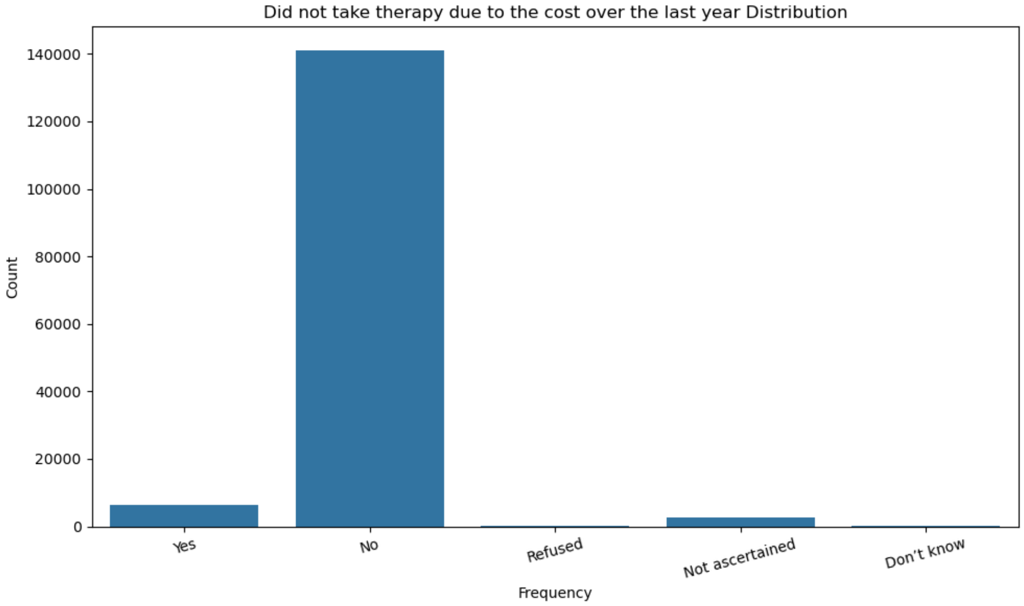


Figure 19: *Plot of Responses Regarding No Therapy Due to Cost Over the Last Year Distribution*

4.2.2.1 Classifiers and Feature Sets Comparison

I used numerous combinations of features and classifier to consider the best result. Also, the following graphs show the average accuracy of the features and classifiers:

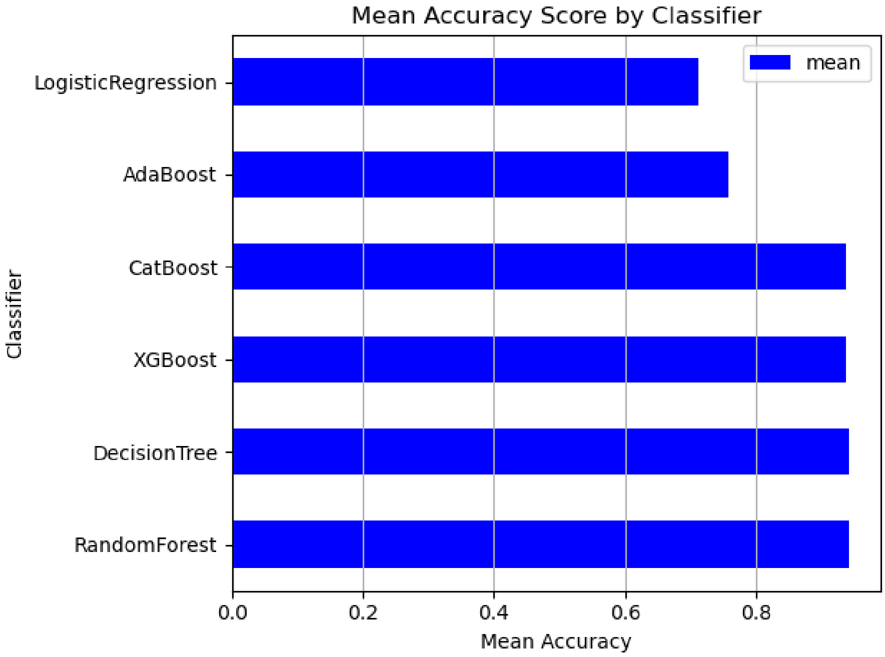


Figure 20: *Mean Accuracy By Classifiers 2*

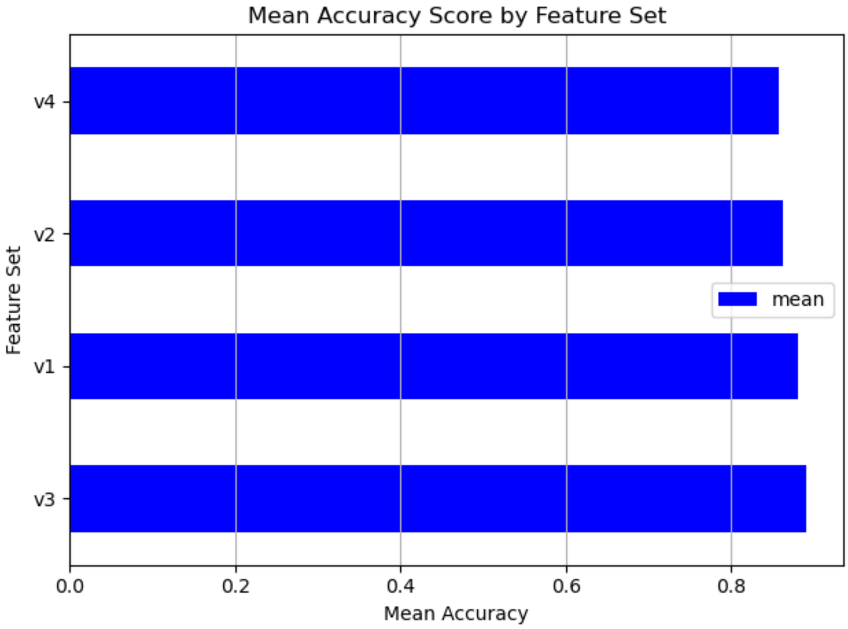


Figure 21: *Mean Accuracy By Feature Sets 2*

In this prediction task, I used the top 2 classifiers and feature sets. The classifiers selected were Random Forest and Decision Tree, while the top 2 feature sets were V1 and V3. These classifiers were selected because of their capacity to manage non-linear connections, their robustness to "noise" and their transparency. Random Forest classifier was performed better than Decision Tree due to its characteristics that lower variance and leads to more stable predictions. Feature set V1 was pre-processed using Simple Imputer with the "most-frequent" strategy to fill missing values in the columns 'ANXLEVEL_A', 'DEPLEVEL_A' and 'MHRX_A'. The decision behind the selection of the Simple Imputer with the “most-frequent” strategy lies to the fact that is a simple and effective way to fill the missing data assuming follow common patterns because it replaces the missing data with the most likely category. Additionally, feature set V3 was pre-processed using Iterative Imputer in the same columns to handle missing values. As another approach, I used Iterative Imputer because I wanted to fill the missing data with a way for underlying relationships in the data, especially when features are correlated. This is the reason of a better performance of the feature set V3, in contrast feature set V1 may have introduced some information loss. To deal with class imbalance, I used RandomOverSampler only in training set after the split of the dataset to ensure that the model learned from an even distribution of classes and did not bring any bias to the test set.

4.2.2.2 Model Performance of Selected Classifiers and Feature Sets

My next step was to use pipelines to construct feature sets V1 and V3, and then run GridSearchCV for both RandomForest and DecisionTree classifiers on each. The following table contains the performance of the combined classifiers and feature sets:

GridSearchCV Results: Top Classifiers and Feature Set Combinations			
Classifier	**Feature Set**	**Cross-Validation Score**	**F1-score**
RandomForest	V1	0.970	0.972
DecisionTree	V1	0.965	0.964
RandomForest	V3	0.966	0.968
DecisionTree	V3	0.965	0.965

***Table 4:** GridSearchCV Results: Top Classifiers and Feature Set Combinations 2*

From the above models, RandomForest using feature set V1 yielded the best cross-validation score and F1-score. This model was thus taken to the final classification pipeline and analysis. The last used model was RandomForest with hyperparameters 'max_depth': 10, 'min_samples_split': 2, 'n_estimators': 200. The model was then trained on feature set V1. The generalizability of the model was assessed on both the training and testing datasets and achieved F1-scores of **0.895** and **0.90**, respectively. This outcome can be linked to various factors. Initially, the Random Forest algorithm is ideally suited for data that includes categorical characteristics, and it performs well even when some certain features provide less information. Furthermore, employing a Simple Imputer using the most common strategy in V1 helped preserve dominant patterns in the data while avoiding noise from complex imputations.

4.2.2.3 Comprehensive Evaluation of the Final Model (RandomForest with Feature Set V1)

i) **Confusion Matrix**

The confusion matrix plot [22] below shows how the model performed for every class. Class 0, which corresponds to response "Yes," was correctly classified on 3,581 of the samples and misclassified on 578 of the samples. Additionally, class 1, which corresponds to response "No," was correctly classified on 22,783 and misclassified on 2,528 of the samples. These 2 are the most common classes in the dataset, and as expected the model here had a strong performance. Also, class 2, corresponds to response "Refused," was correctly classified on 25 and misclassified on 8 of the samples. Furthermore, class 3, corresponds to response "Not Ascertained," was correctly classified on 506 and misclassified on 3 of the samples. Last but not least, class 4, which corresponds to response "Don't Know," was correctly classified on 15 of the samples and misclassified on 17 of the samples. According to the aforementioned results, the model performed very well, both on frequent classes and on less frequent classes because of the class balancing technique that I used.

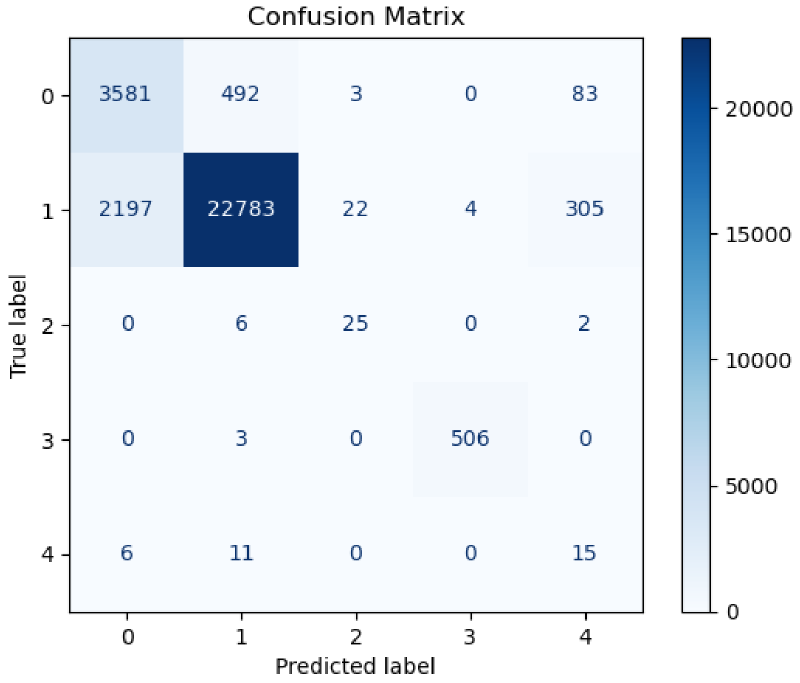
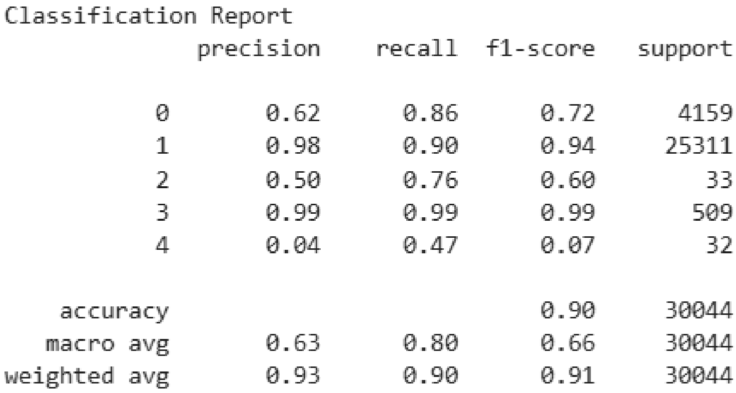


***Figure 22:** Confusion Matrix 2*

ii) Classification Report

The observations drawn from the classification report [23] demonstrate that with the exception of class 4, which has noticeably lower recall, the precision and recall levels for the five classes are relatively similar. Based on the recall, 86% of individuals in class "Yes" were classified correctly. Respectively, for class "No", 90% of the individuals were classified correctly. These high values are expected, as these are the most frequent classes in the dataset, and the model had sufficient examples during training to learn their patterns effectively. Additionally, for the class "Refused" 76% of individuals were classified correctly based on the recall, which is reasonably high due to its lower frequency. For the classes "Not ascertained" and "Don't Know" individuals were classified correctly based on the recall 99% and 47%, respectively. The high recall for class 3 indicates that its features are distinct and easily separable from other classes. Additionally, the low recall for class 4, based on the fact that the number of the samples of this class in training set is small. As a result, based on the precision for class "Yes", 62 out of 100 people predicted correctly. Similarly, based on the precision for class "No", 98 out of 100 people predicted correctly. For the class "Refused", 50 out of 100 people predicted correctly based on the precision. Lastly, for the classes "Not ascertained" and "Don't Know", 99 out of 100 people and 4 out of 100 people predicted correctly based on the precision, respectively.



```
Classification Report
              precision    recall  f1-score   support

           0       0.62      0.86      0.72      4159
           1       0.98      0.90      0.94     25311
           2       0.50      0.76      0.60        33
           3       0.99      0.99      0.99       509
           4       0.04      0.47      0.07        32

    accuracy                           0.90     30044
   macro avg       0.63      0.80      0.66     30044
weighted avg       0.93      0.90      0.91     30044
```

Figure 23: *Classification Report 2*

iii) Learning Curve

The learning curve [24] below illustrates how the model can generalize to previously unseen data. Also, the training accuracy slightly decreases with more data, while the validation accuracy steadily increases. This pattern indicates that the model is learning more generalizable patterns rather than overfitting to the training set. The convergence between training and validation accuracy suggests improved performance and stability, confirming that the model can effectively generalize beyond the training data.

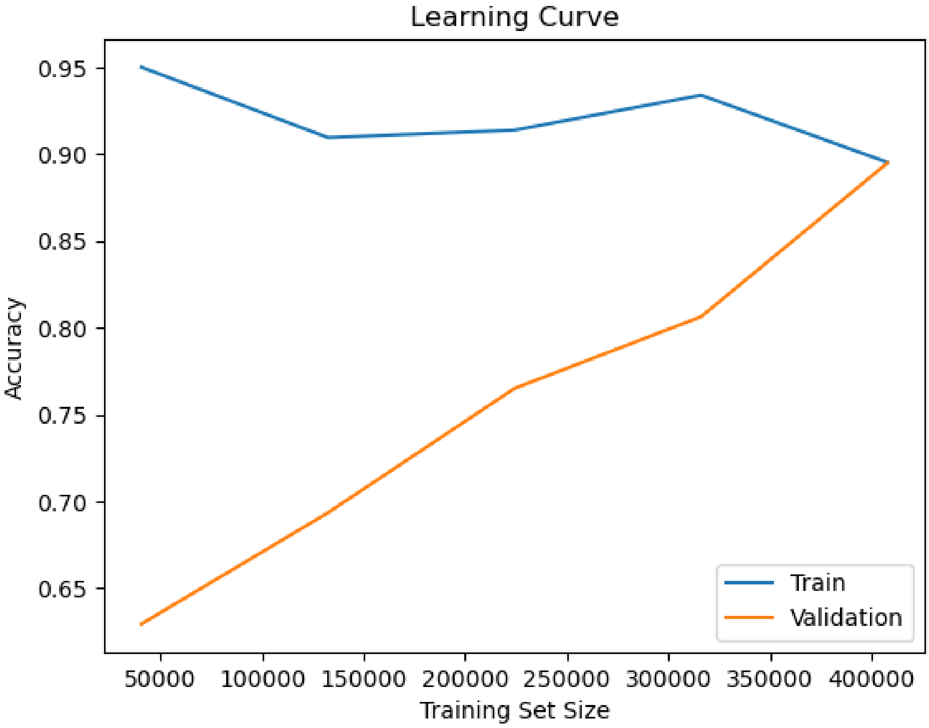


Figure 24: *Learning Curve 2*

iv) Feature Importance

The feature importance plot [25] below shows which features are significantly affect the ANXMED_A. For this prediction, the most important features are the DEPMED_A, ANXFREQ_A and MHTHDLY_A. The DEPMED_A is an important feature for the anxiety medication prediction because these 2 mental health disorders coexist most of the time. Furthermore, ANXFREQ_A it is an important feature because the frequency of anxiety reflects the severity of the condition. The feature MHTHDLY_A is also important because it indicates access issues, for medication uptake. The other features do not influence the ANXMED_A as significantly as the previously mentioned features.

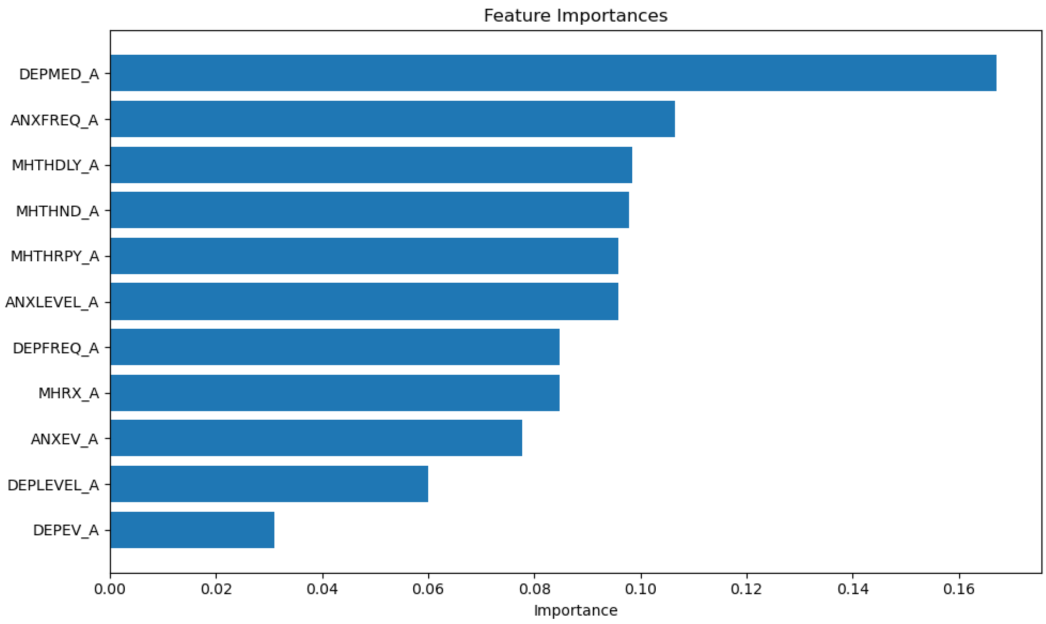


Figure 25: *Feature Importance 2*

4.2.2.4 Significance of the Prediction

In short, it is vital to predict if a person is prescribed medication for an anxiety disorder. The main reason for the necessity of prediction is that if one knows whether a person is prescribed medication, one can conclude without any doubt that the person is experiencing an anxiety disorder since a person is prescribed medication only if he or she is diagnosed by a medical specialist. The model predicts at a rate of 90% if a person receives prescription medication because of the person who has an anxiety disorder.

4.2.3 Prediction whether an individual takes medication for depression

The prediction task is performed using Dataset 2, with the feature "DEPMED_A" as the target variable.

4.2.3.1 Exploratory Data Analysis (EDA)

1. **Depression Medication Usage Distribution**

   The figure [26] displays the distribution of the depression medication usage. The X-axis represents the response categories and the Y-axis shows the number of individuals corresponding to each response. The most common response is "No", suggesting that most people do not use depression medication. The notable disparity between "Yes" and "No" responses may be influenced by various factors, including limited access to care, underdiagnosis, or social stigma associated with mental health and medication use. These factors could result in individuals not taking medication or not disclosing it in the survey.

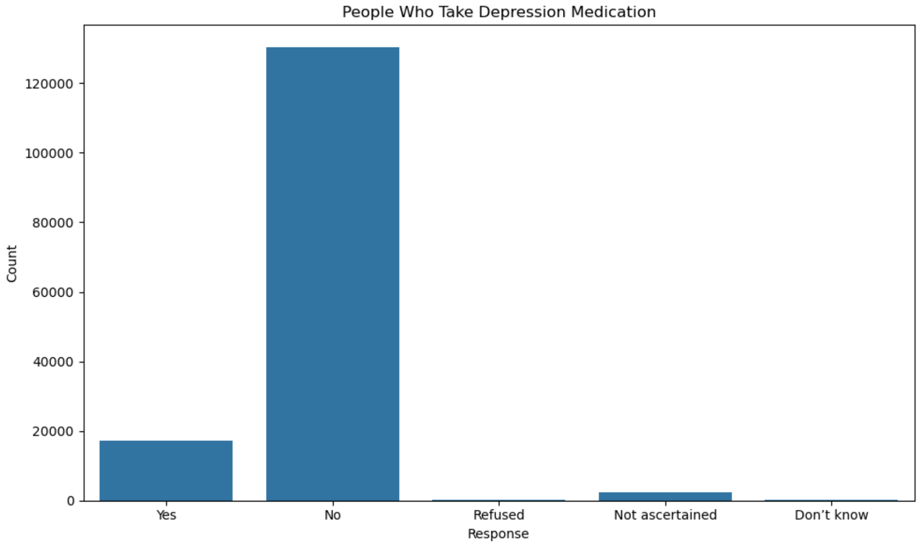


Figure 26: *Plot of Depression Medication Usage Distribution*

2. Depression Frequency Distribution

The figure below [27] shows the depression frequency distribution. The X-axis corresponds to the possible answers and the Y-axis displays the number of individuals corresponding to each answer. The 2 most common answers to the question about how often individuals experience feelings of depression are "Never" and "A few times a year". Moreover, the responses "Daily" , "Weekly", "Monthly" have exhibit comparable numbers, indicating that a large segment of people does face anxiety feelings on a regular basis. This could suggest broader social problems such as economic uncertainty, depression, and concerns about what lies ahead.

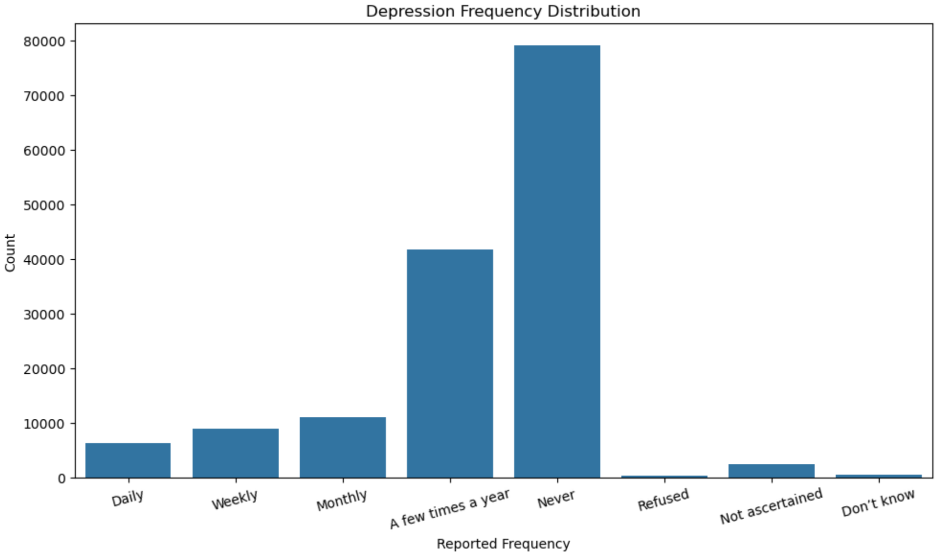


Figure 27: *Plot of Depression Frequency Distribution*

3. Depression Ever Distribution

The figure [28] displays the distribution of the depression ever. The X-axis corresponds to the types of the responses and the Y-axis corresponds to the number of individuals that had been ever told by a mental health professional that they have any type of depression disorder. The most common response is “No” and the second most common response is “Yes”. The other 3 responses are indicating that most respondents gave a clear answer. The number of responses for response “No” shows the underdiagnosis of anxiety disorders.

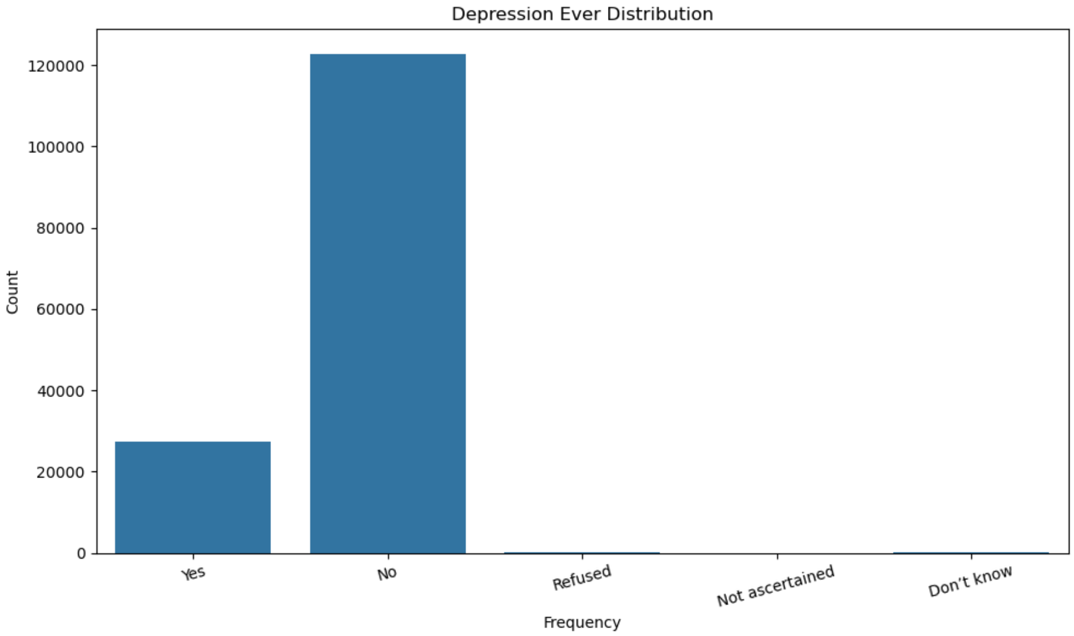


Figure 28: *Plot of Depression Ever Distribution*

4. Depression Level Distribution

The figure [29] shows the distribution of Depression Level. The X-axis presents the responses and the Y-axis presents the number of individuals of each response. The 2 more frequent responses are the “A little” and “Somewhere in between a little and a lot”. This pattern could indicate ongoing low-grade stressors in daily life like work pressures, financial uncertainty, or social challenges that lead to moderate depression levels without guaranteeing a clinical diagnosis.

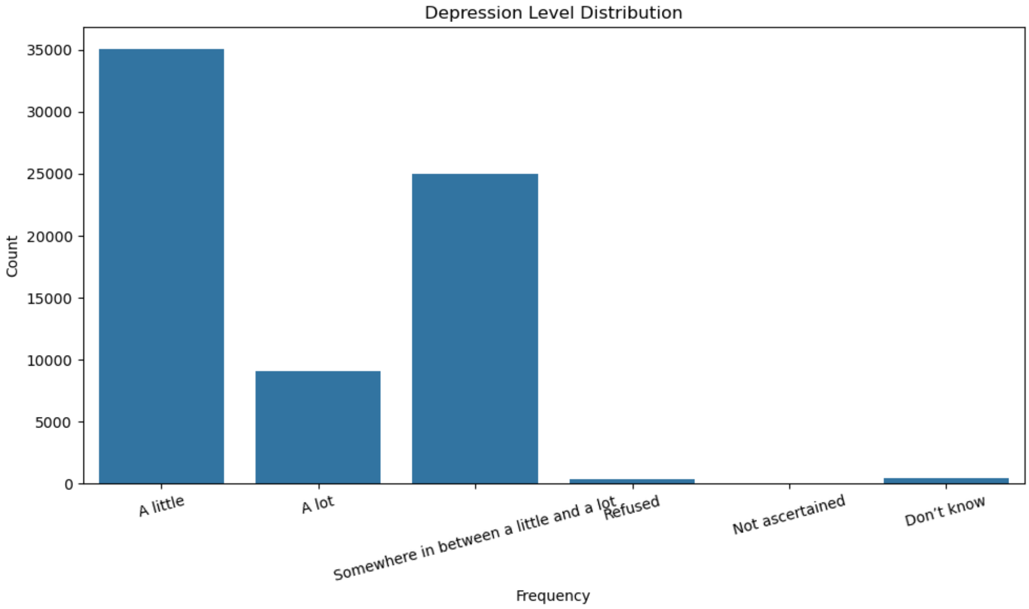


Figure 29: *Plot of Depression Level Distribution*

4.2.3.2 Classifiers and Feature Sets Comparison

To determine the optimal outcome, I experimented with a variety of feature and classifier combinations. The average accuracy of the features and classifiers is also displayed in the following graphs:

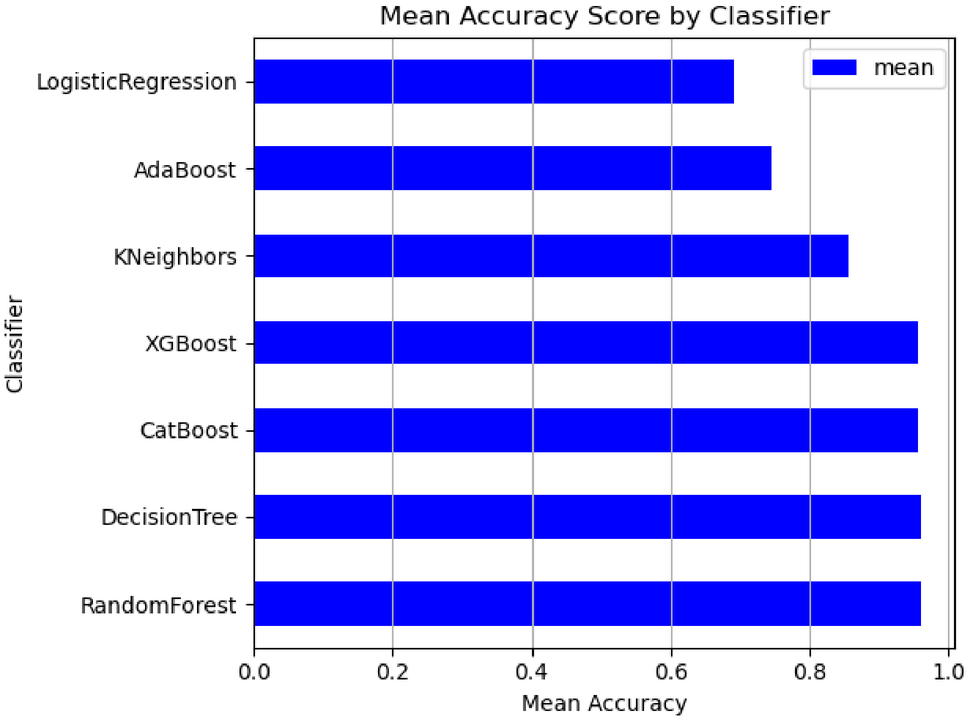


Figure 30: *Mean Accuracy By Classifiers 3*

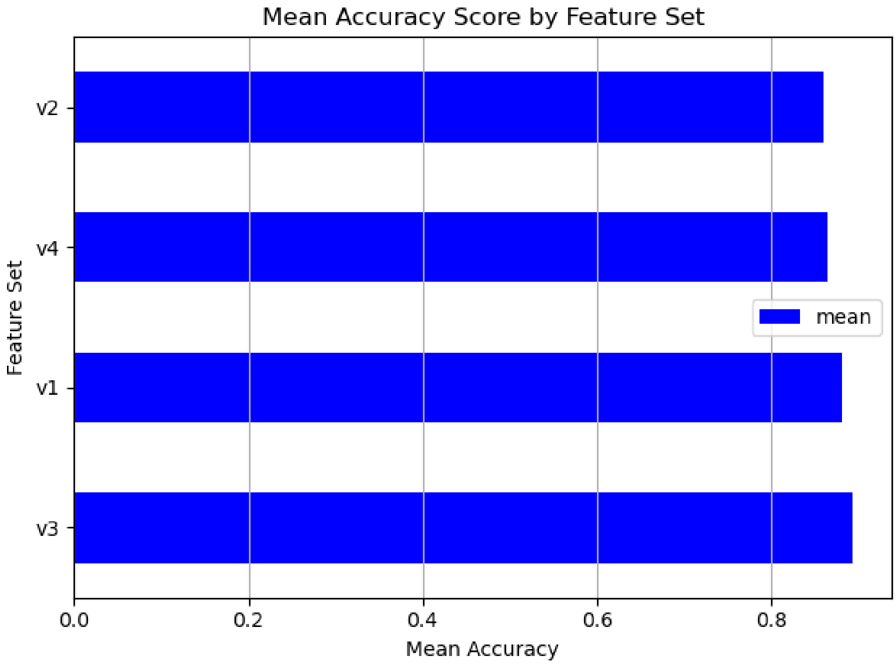


Figure 31: *Mean Accuracy By Feature sets 3*

For this prediction task, I used the top 2 results of both classifiers and feature sets. The 2 classifiers with the best performance are RandomForest and DecisionTree. These classifiers were selected because of their capacity to manage non-linear connections, their robustness to "noise" and their transparency. Random Forest classifier was performed better than Decision Tree due to its characteristics that lower variance and leads to more stable predictions. Moreover, the 2 feature sets with the best performance are the feature set V1 and feature set V3. The feature set V1 was pre-processed using Simple Imputer to fill in the missing values of the columns 'ANXLEVEL_A', 'DEPLEVEL_A' and 'MHRX_A' and the feature set V3 was pre-processed using Iterative Imputer to fill in the missing values of the same columns. The better performance of feature set V3 can be attributed to the use of the Iterative Imputer, which fills in missing data by capturing underlying relationships between features, especially when they are correlated. On the contrary, the Simple Imputer used in feature set V1 may have introduced some information loss by relying only on the most frequent value for imputation. To address class imbalance, I applied RandomOverSampler exclusively on the training set after splitting the dataset to guarantee that the model trained on a balanced distribution of classes and avoided introducing any bias to the test set.

4.2.3.3 Model Performance of Selected Classifiers and Feature Sets

After running all the combinations of classifiers and feature sets, I developed the pipelines to produce the feature sets and run a GridSearchCV for the best classifiers. The following table indicates the performance of each combination:

GridSearchCV Results: Top Classifiers and Feature Set Combinations			
Classifier	**Feature Set**	**Cross-Validation Score**	**F1-score**
RandomForest	V1	0.919	0.917
DecisionTree	V1	0.918	0.915
RandomForest	V3	0.970	0.968
DecisionTree	V3	0.966	0.964

Table 5: *GridSearchCV Results: Top Class Classifiers and Feature Set Combinations 3*

Among the models listed above, RandomForest with feature set V3 had the highest cross-validation result as well as F1-score. This model was therefore included in the final classification pipeline and subsequent analysis. The final model utilized was RandomForest with the hyperparameters 'max_depth': 10, 'min_samples_split': 10, 'n_estimators': 50. This model was trained using feature set V3. This model's generalizability was tested both for training data as well as testing data and had F1-scores of **0.928** as well as **0.955**, respectively. This outcome can be linked to various factors. Initially, the Random Forest algorithm is ideally suited for data that includes categorical characteristics and it performs well even when certain features provide less information. Furthermore, the application of an Iterative Imputer in feature set V3 aided the model's effectiveness by uncovering hidden relationships between features throughout the imputation procedure. This method is particularly advantageous when features are related, as it enables more precise and informed estimations of missing values, thereby enhancing the quality of the input data.

4.2.3.4 Comprehensive Evaluation of the Final Model (RandomForest with Feature Set V3)

i) Confusion Matrix

The following confusion matrix plot [32] illustrates how the model performed for each class. Class 0, corresponding to response “Yes,” was classified correctly for 3,248 of the samples and misclassified for 193 of them. Class 1, corresponding to response “No,” was classified correctly for 24,619 and misclassified for 1,398 of the samples. Also, class 2, corresponding to response "Refused," was classified correctly for 33 and misclassified for 4 of the samples. Further, class 3, corresponding to response "Not Ascertained," was classified correctly for all the 524 samples. Lastly, class 4, corresponding to response "Don't Know," was classified correctly for 12 of the samples while being misclassified for 13 of the samples. From the above results, we can see that the model performed exceptionally well on frequent classes as well as on less frequent classes because of the class balancing technique that I used.

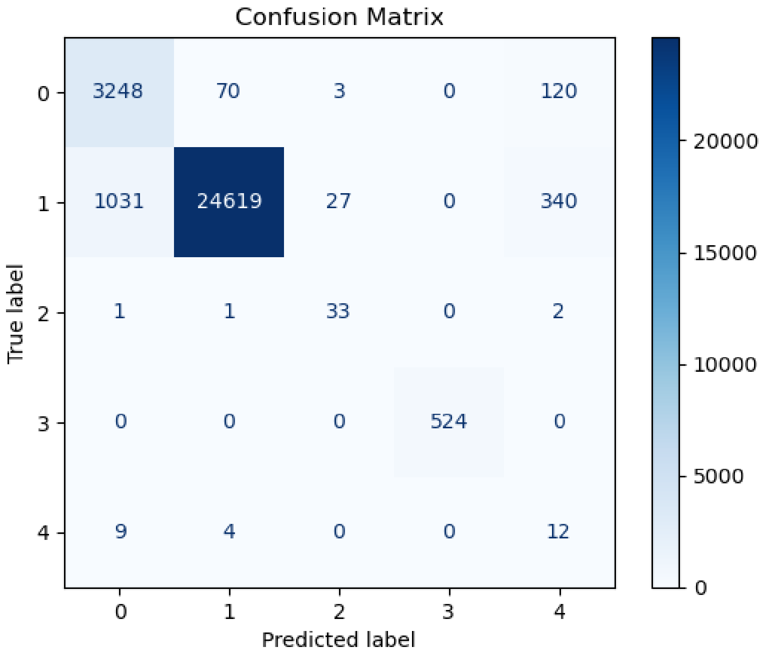
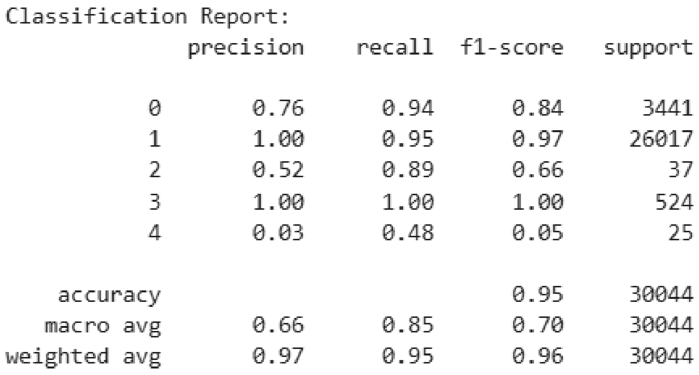


Figure 32: *Confusion Matrix Plot 3*

ii) Classification Report

Observations from the classification report [33] reveal that the classes 0,1,3 have high recall and precision, the class 2 has a relatively recall and precision and class 4 has significantly lower recall, precision. Additionally, according to the recall 94% of the people belonging to class "Yes" were correctly classified. Respectively to class "No", 95% of people were correctly classified. These high values are expected, as they are the most frequent classes in the dataset, and the model had sufficient examples during training to learn their patterns effectively. Also, for class "Refused", 89% of people were correctly classified according to the recall, which is reasonably high due to class lower frequency. For the categories "Not ascertained" and "Don't Know" people were correctly classified according to the recall 100% and 48%, respectively. The high recall for class 3 indicates that its features are distinct and easily separable from other classes. Additionally, class 4 still exhibits low recall, suggesting other factors such as feature similarity or noise might be affecting performance. Also, according to the precision for class “Yes”, 76 of 100 people predicted correctly. Similarly, according to the precision for class “No”, 100 of 100 people predicted correctly. For class “Refused”, 52 of 100 people predicted correctly according to the precision. Lastly, for class “Not ascertained” and class “Don’t Know”, 100 of 100 people and 3 of 100 people predicted correctly according to the precision, respectively.



```
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.94      0.84      3441
           1       1.00      0.95      0.97     26017
           2       0.52      0.89      0.66        37
           3       1.00      1.00      1.00       524
           4       0.03      0.48      0.05        25

    accuracy                           0.95     30044
   macro avg       0.66      0.85      0.70     30044
weighted avg       0.97      0.95      0.96     30044
```

Figure 33: *Classification Report 3*

iii) Learning Curve

The learning curve [34] demonstrates the model's generalization to new data with a growing training set. As predicted, training accuracy drops off slightly with increased data, showing decreased overfitting. At the same time, validation accuracy rises steadily, showing that model performance on new data improves with increased training examples. This trend is a sign of a more stable model with increased generalization ability as the data set grow.

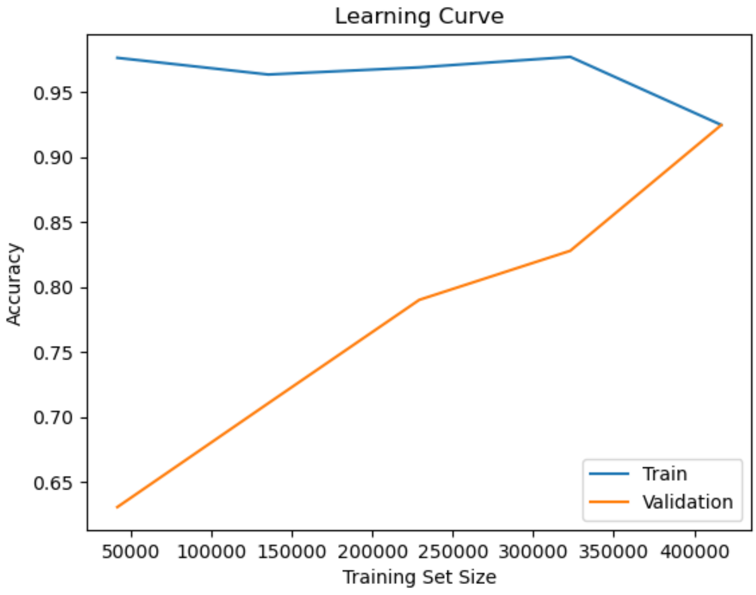


Figure 34: *Learning Curve 3*

iv) Feature Importance

The feature importance plot [35] that is given, indicates which of the features greatly impacts the DEPMED_A. For the given prediction, the DEPFREQ_A, DEPLEVEL_A, ANXMED_A and MHRX_A are the most important features. The DEPFEQ_A and DEPLEVEL_A features are important because they reflect the frequency and severity of depressive symptoms and feelings. Moreover, the feature ANXMED_A is important because most of the times depression and anxiety are coexist and the treatment for one condition often correlated with each other's treatment. Additionally, MHRX_A is a strong predictor because it is closely tied to medication use. All of the other features effected the prediction for DEPMED_A but not significantly.

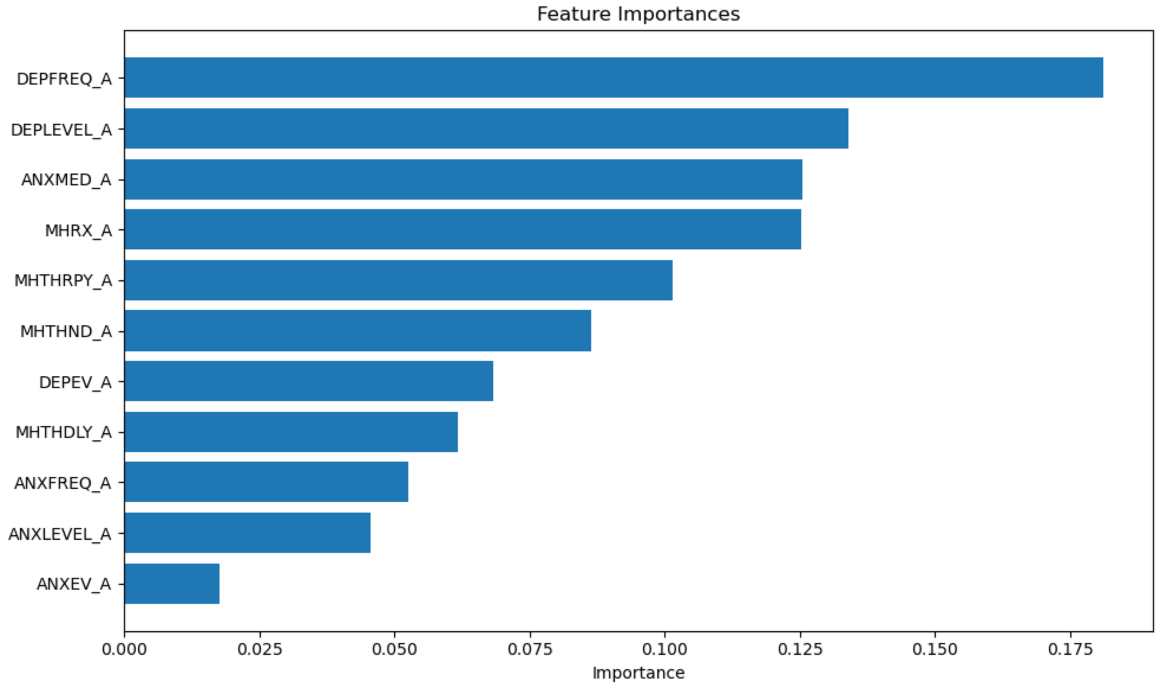


Figure 35: *Feature Importance Plot 3*

4.2.3.5 Significance of the Prediction

In brief, it is crucial to predict whether a person is prescribed medication for a depression disorder. This is because if one knows whether a person is being prescribed medication or not, one can definitely say without having a doubt that the person is suffering from a depression disorder because a person is only prescribed medication if he or she is diagnosed by a medical specialist. The model predicts with a success rate of 95.30% whether a person is given prescription medication due to which the person suffers from a depression disorder.

4.3 Regression Results

4.3.1 Prediction of Eating Disorder Prevalence Based on Other Mental Health Conditions at the Population Level

The prediction task is performed using Dataset 3 by using the feature "Eating (%) Disorders" as the target variable.

4.3.1.1 Exploratory Data Analysis

1. **Eating Disorders (%) Distribution**

The plot [36] illustrates the prevalence distribution of eating disorders. The X-axis shows the proportion of countries which are impacted by eating disorders. The Y-axis indicates the number of observations that fit into each percentage range. The plot's distribution is right-skewed, indicating that the majority of countries report low rates of eating disorders, typically ranging from 0.1 to 0.3. The typical prevalence range is approximately 0.13-0.15. The long tail shown in the graph below indicates that elevated prevalence rates correspond to a limited number of countries.

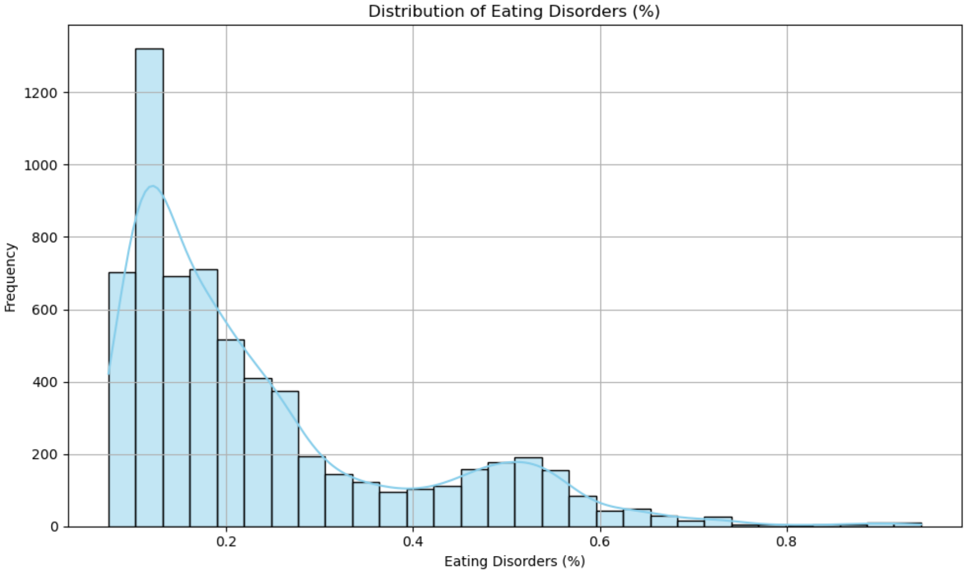


Figure 36: *Plot of Eating Disorders (%) Distribution*

2. **Schizophrenia (%) Distribution**

The plot [37] displays the prevalence distribution of Schizophrenia. The X-axis shows the proportion of countries which are impacted by Schizophrenia. The Y-axis indicates the number of observations that fit into each percentage range. The distribution of the plot is right-skewed with the greatest concentration of values ranging from 0.18% to 0.22%. The most frequent prevalence rate among the observations is around 0.20%. Furthermore, high prevalence of schizophrenia is rare because of the few cases risen above 0.30%. The long tail depicted in the graph suggests that high prevalence rates are associated with a small number of countries.

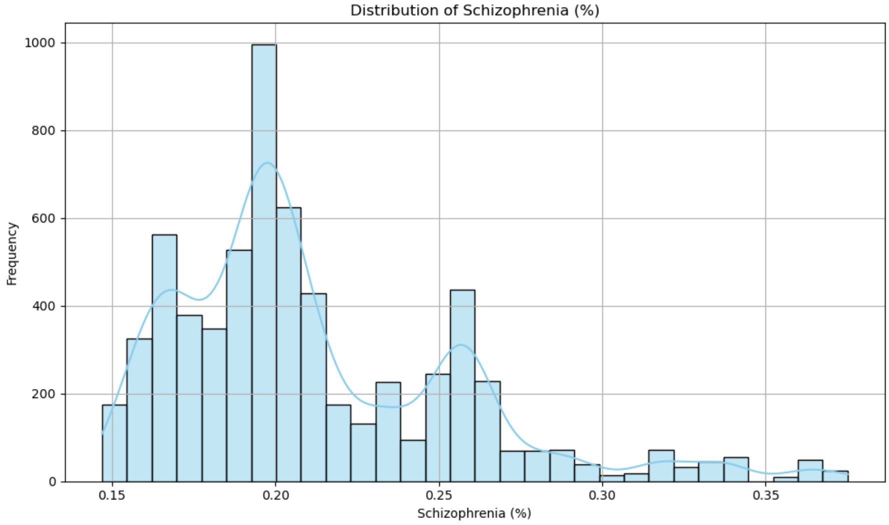


Figure 37: *Plot of Schizophrenia (%) Distribution*

3. Bipolar Disorder (%) Distribution

The plot [38] shows the prevalence distribution of Bipolar Disorder. The X-axis shows the proportion of countries which are impacted by Bipolar Disorder. The Y-axis indicates the number of observations that fit into each percentage range. The plot's distribution is right-skewed with the highest accumulation of prevalence values focused on 0.60%. The most common prevalence rates are between 0.55% and 0.65%. At the prevalence rate 1.0% a decrease in the frequency is observed, meaning that the number of cases is negligible.

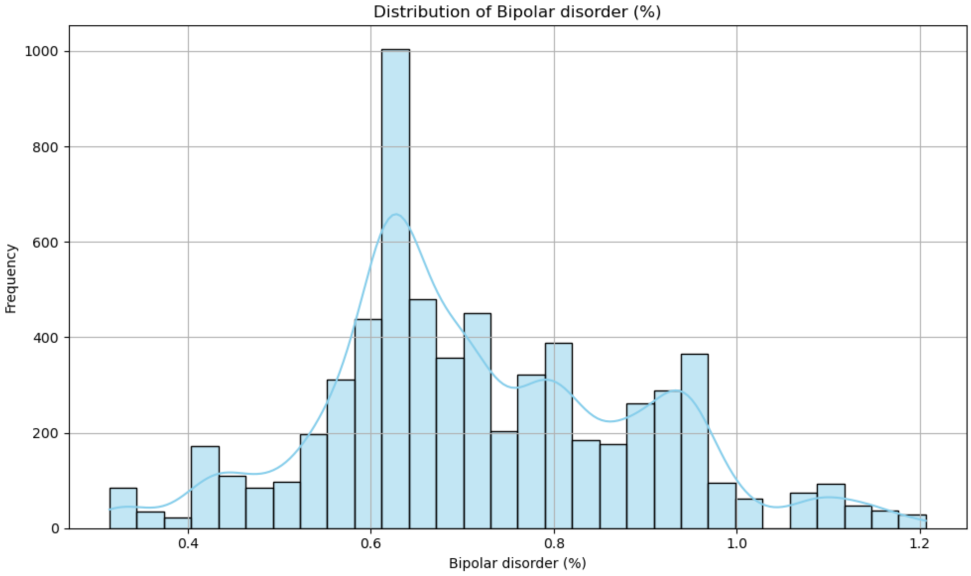


Figure 38: *Plot of Bipolar Disorder Distribution*

4. **Anxiety Disorders (%) Distribution**

The plot [39] represents the prevalence distribution of Anxiety Disorders. The X-axis shows the proportion of countries which are impacted by Anxiety Disorder and the Y-axis indicates the number of observations that fit into each percentage range. The plot's distribution is right-skewed with highest accumulation of prevalence values focused on 2.5% and between 3.4% and 4.6%. Additionally, at the prevalence rate 7.0% a decrease is observed in the frequency axis, meaning that the number of cases is negligible.

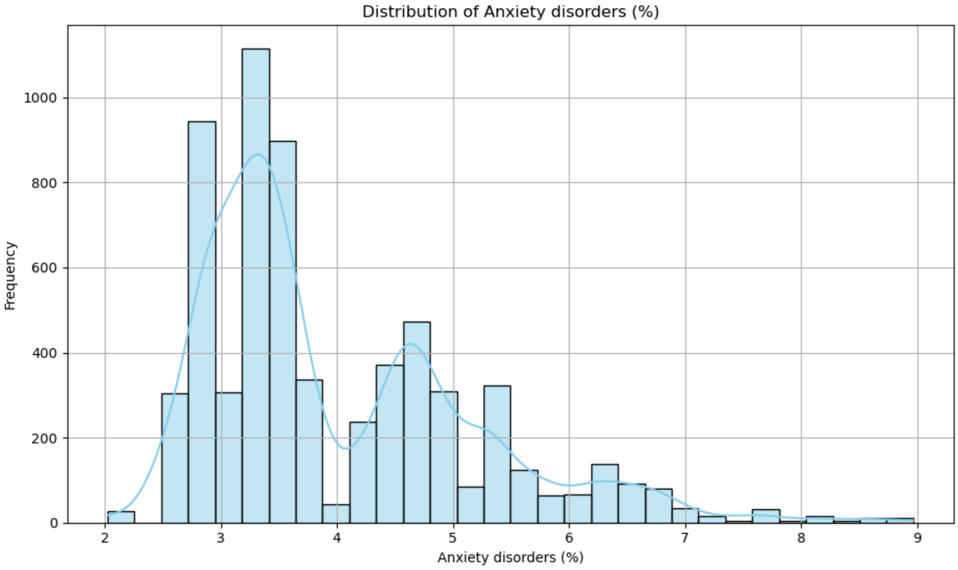


Figure 39: *Plot of Anxiety Disorders Distribution*

5. **Drug Use Disorders (%) Distribution**

The graph [40] displays the distribution of Drug Use Disorders. The X-axis shows the proportion of countries which are impacted by Drug Use Disorders and the Y-axis indicates the number of observations that fit into each percentage range. The distribution of the plot is right-skewed with a maximum prevalence values between 0.4% and 1%. Additionally, at the prevalence rate 2.0% a decrease is observed in the frequency axis, meaning that the number of cases is negligible. The long tail represented in the graph suggests that high prevalence rates are linked to a small number of countries.

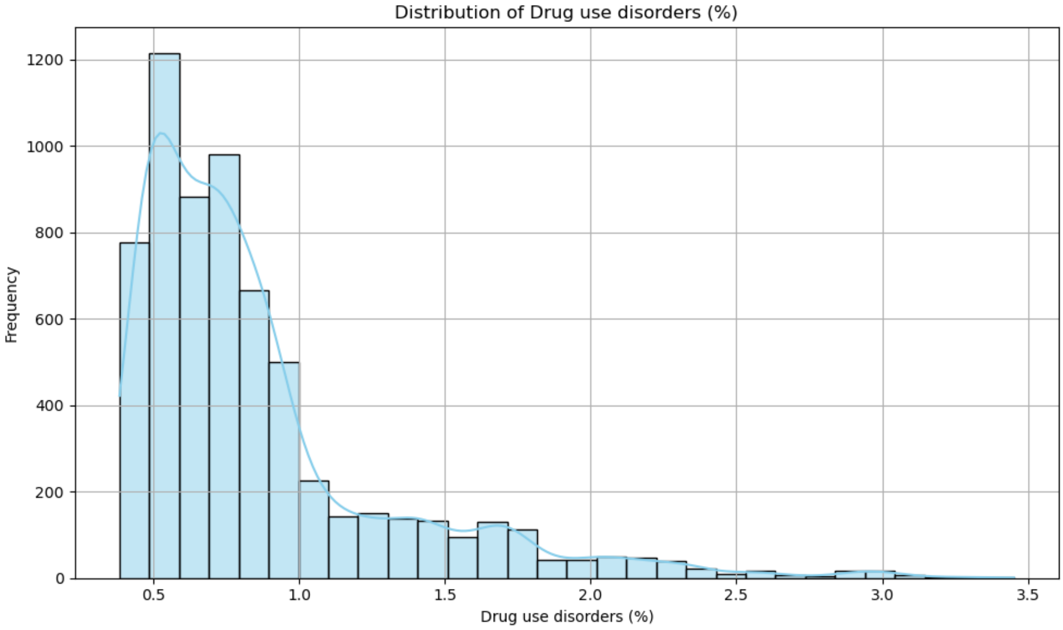


Figure 40: *Plot of Drug Use Disorders Distribution*

6. Depression (%) Distribution

The graph [41] displays the distribution of Depression. The X-axis shows the proportion of countries which are impacted by Depression and the Y-axis indicates the number of observations that fit into each percentage range. The distribution of the plot is slightly right-skewed with maximum prevalence values between 2.5% and 4.5%. Additionally, at the prevalence rate 5.0% a decrease is observed in the frequency axis, meaning that the number of cases is negligible. The long tail represented in the graph suggests that high prevalence rates are linked to a small number of countries.

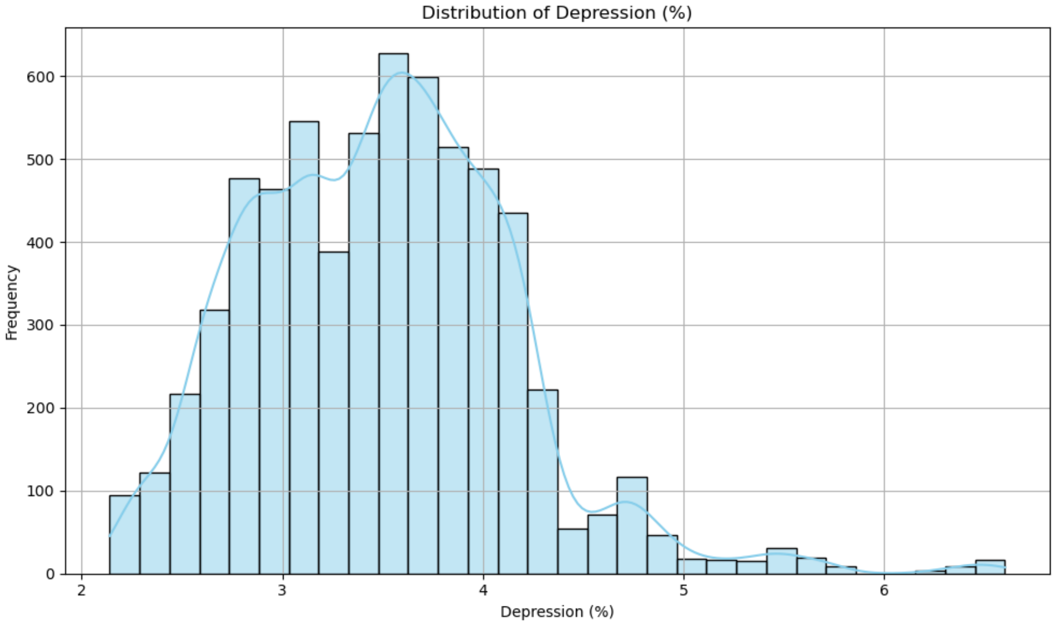


Figure 41: *Plot of Depression Distribution*

7. Alcohol Use Disorders (%) Distribution

The graph [42] displays the distribution of Alcohol Use Disorders. The X-axis shows the proportion of countries which are impacted by Alcohol Use Disorders and the Y-axis indicates the number of observations that fit into each percentage range. The distribution of the plot is right-skewed with maximum prevalence values between 0.5% and 1.6%. Once surpassing the 2.0% threshold, the occurrence steadily decreases indicating that greater prevalence rates are more rare. The long tail of the distribution shows that high rates of alcohol use disorders are focused in a limited number of countries.

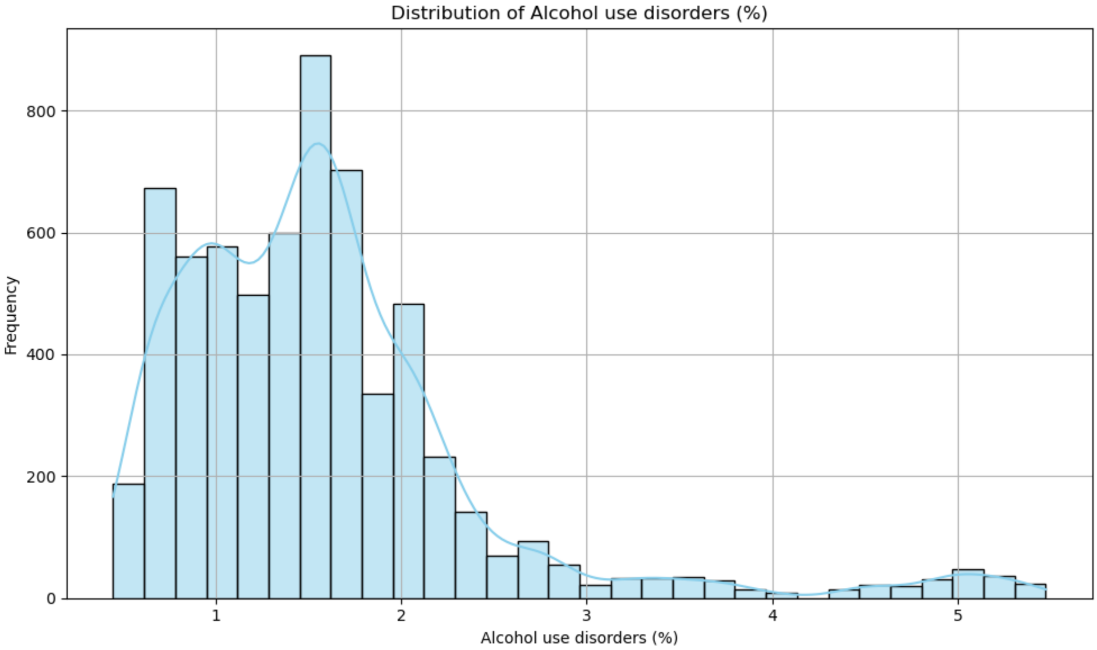


Figure 42: *Plot of Alcohol Use Disorders Distribution*

8. Skewness of Numeric Features

The graph [43] presents the skewness for every numerical features. The characteristics of Alcohol use disorders and drug use disorders display pronounced right-skewed distributions, whereas Bipolar Disorder exhibits behaviour that it is nearly symmetric.

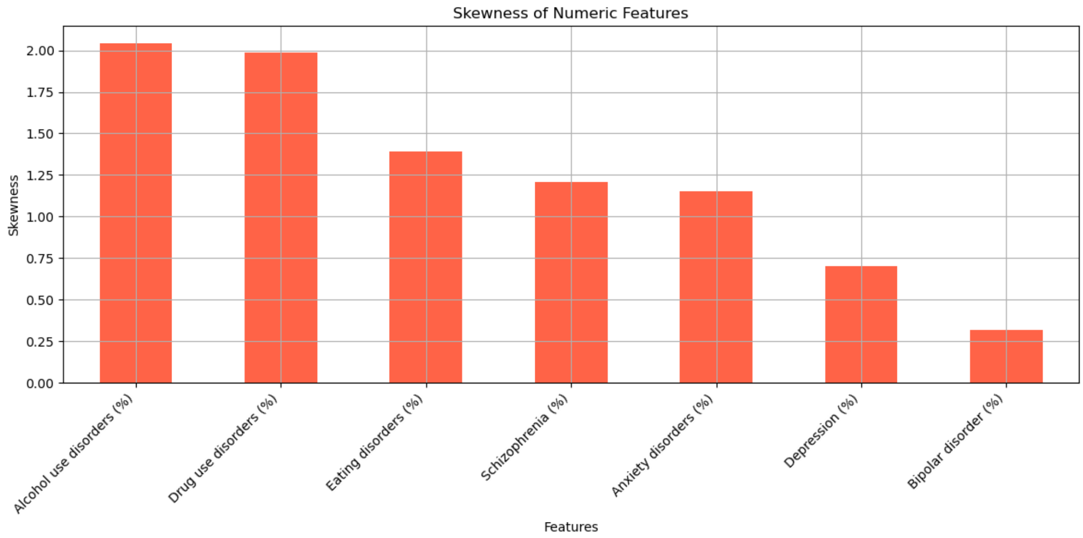


Figure 43: *Skewness of Numeric Features Plot*

9. Correlation Matrix of Numerical Features

The follow heatmap [44] displays the correlation between numerical features. The features that are highly correlated with Eating Disorders feature are the Schizophrenia (0.67), Bipolar Disorder (0.71) and Anxiety Disorder (0.70). This indicates that these variables could be significant predictors or key elements in models designed to forecast the prevalence of eating disorders.

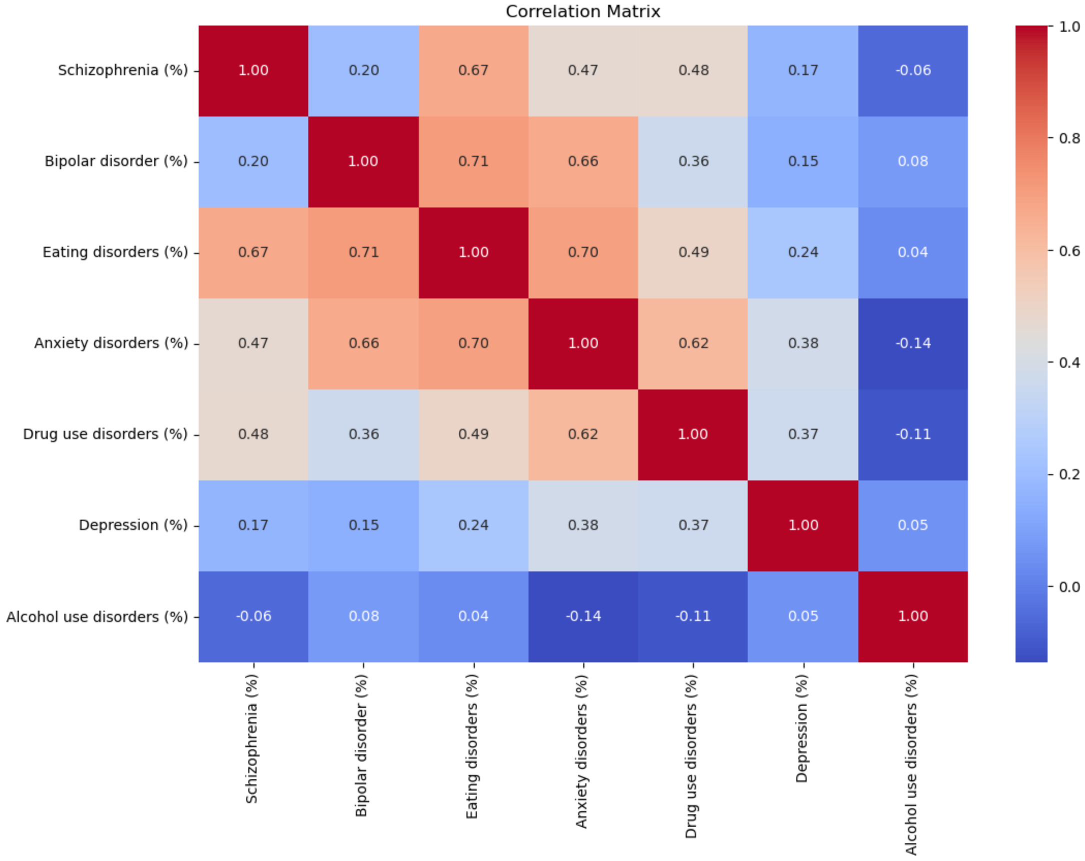


Figure 44: *Correlation Matrix Plot*

4.3.1.2 Regressors and Feature Sets Comparison

I tried a number of different feature sets and regressor pairs to find the best possible outcome. The average accuracy of the regressor and feature sets is also plotted in the next graphs:

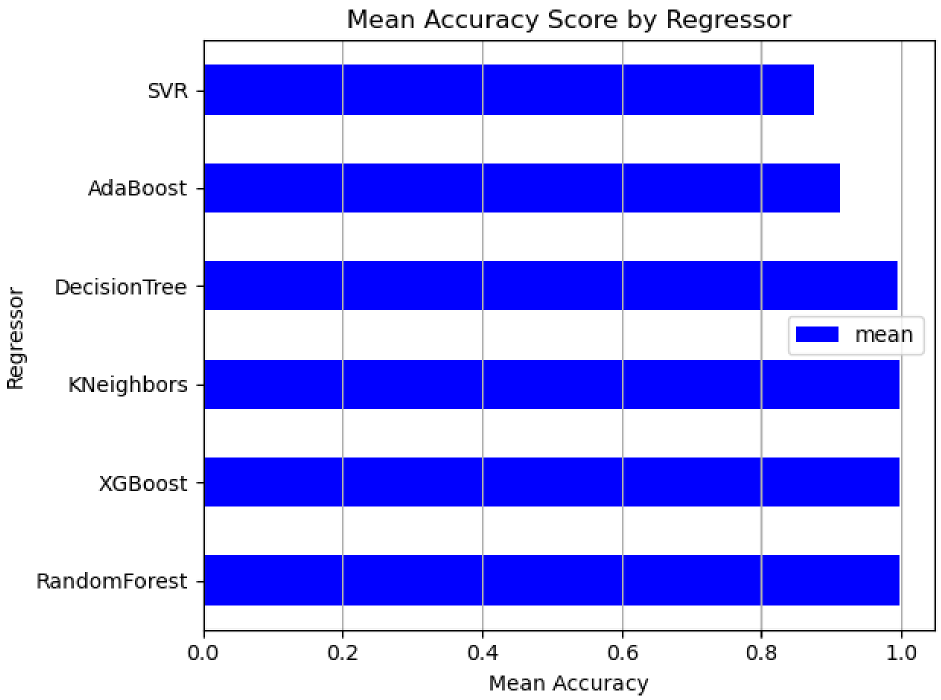


Figure 45: *Mean Accuracy By Regressors 1*

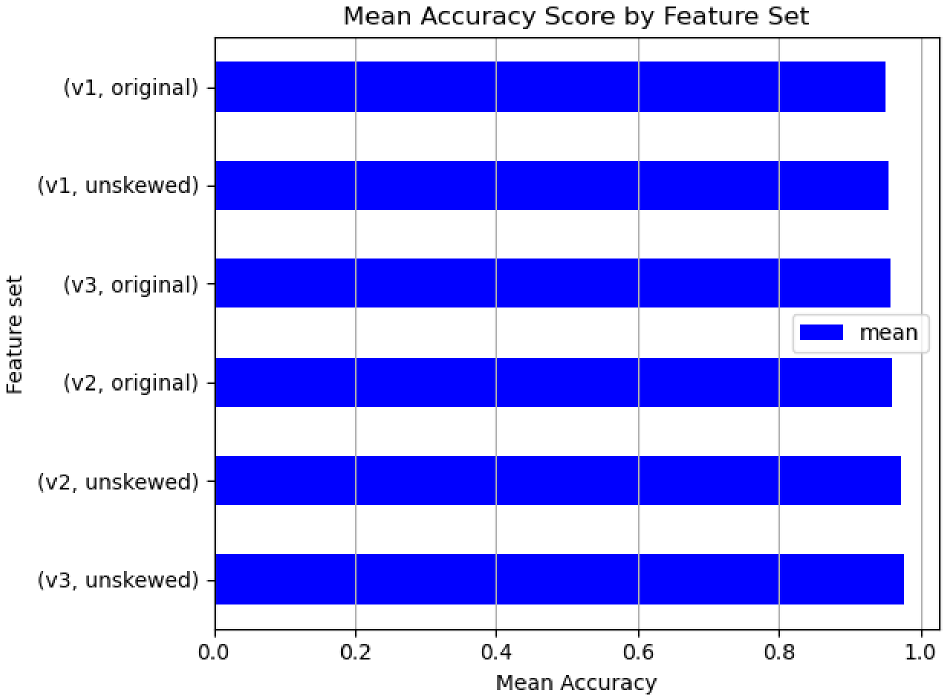


Figure 46: *Mean Accuracy By Feature Sets 4*

Based on this prediction and the above graphs I selected the 2 regressors as well as the 2 sets of features that have the best mean accuracy. The 2 best performing regressors are RandomForest and XGBRegressor. These regressors perform better due to their ability to capture non-linear relationships, their robustness to noise, their tendency to reduce overfitting, and their strong generalization capabilities through ensemble techniques. However, the 2 top performing sets of features, are V2 and V3. Feature set V2 is a build over feature set V1, where Sequential Forward Selection was applied. Extra preprocessing was applied by means of PowerTransformer on 'Alcohol use disorders (%), 'Drug use disorders (%), 'Schizophrenia (%), 'Anxiety disorders (%)' columns due to the values 2.043835,1.988651, 1.207086, 1.154393 respectively dealing with the issue of skewness. Feature set V3 being the build over feature set V1 was pre-processed by means of StandardScaler for all the numerical features. Finally, I have used two versions of the target variable 'Eating disorders (%)' here. The first one is the original and the second one a transformed variable because of the skewness value of 1.393398. I have used a PowerTransformer based on the 'yeo-johnson' method to resolve the issue of the skewness. As shown in the feature set comparison plot, the unskewed versions of V2 and V3 consistently outperformed their original counterparts, which supports the conclusion that addressing skewness in both input features and the target variable leads to improved model accuracy and generalization. Skewed distributions can negatively impact regression models because they violate the assumption of normally distributed input or output variables, which many of them rely on for accurate prediction. Skewness can also lead to biased error terms, distort feature relationships, and cause the model to focus too heavily on outliers or rare extreme values, ultimately reducing performance and generalization.

4.3.1.2 Model Performance of Selected Regressors and Feature Sets

GridSearchCV Results: Top Regressors and Feature Set Combinations			
Regressor	**Feature Set**	**Cross-Validation Score**	R^2 **score**
RandomForest	V2	0.9975	0.9956
XGBRegressor	V2	0.9976	0.9963
RandomForest	V3	0.9975	0.9958
XGBRegressor	V3	0.9976	0.9961

Table 6: *GridSearchCV Results: Top Regressors and Feature Set Combinations*

Out of the models summarized above, XGBRegressor for feature set V2 recorded the maximum cross-validation value as well as R^2 score. The model was thus integrated into the final regression pipeline as well as analysis. The model that was finally used was XGBRegressor with the hyperparameters colsample_bytree=1.0, gamma=0, n_estimators=800,learning_rate=0.1, max_depth=10, reg_alpha=0, reg_lambda=1, and subsample=0.7. The model was trained on feature set V2. The generalizability of this model was tested on the training as well as the testing data and produced R^2 scores of **0.9999** as well as **0.9955**, respectively. This outstanding performance can be attributed to several factors. Firstly, XGBRegressor is a gradient boosting ensemble method known for its ability to handle complex, non-linear relationships and to control overfitting through regularization parameters such as gamma, reg_alpha, and reg_lambda. Secondly, feature set V2 enhanced learning by reducing noise and irrelevant input through feature selection, while addressing skewness in critical variables improved the distribution of the data, making it easier for the model to learn effectively. The combination of a robust model architecture with carefully pre-processed, informative features likely explains the performance observed in both cross-validation and final evaluation.

4.3.1.3 Comprehensive Evaluation of the Final Model (XGBRegressor with Feature Set V2)

i) **Actual VS Predicted Values**

The graph [47] displays actual versus predicted values. The red dashed line indicates perfect correlation between predicted and actual values, while the blue dots represent predicted values. As can be observed, most predicted values are near the red dashed line, with some deviations above and below it, which is expected. The data range from roughly 0.1 to 0.9, indicating that the model performs well across the full range of eating disorder prevalence. Slightly larger deviations at higher values may suggest increased variability or data sparsity in those regions.

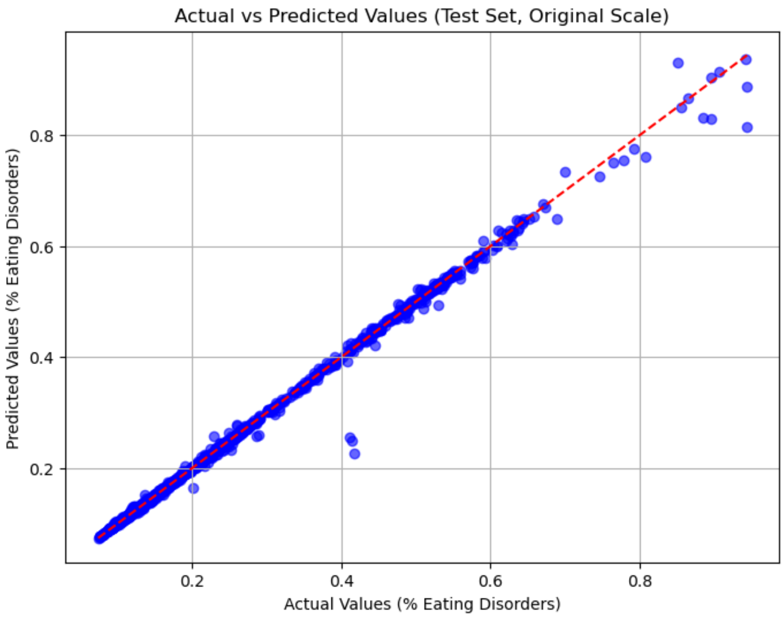


Figure 47: *Actual vs Predicted Values Plot*

ii) Residual Plot

The residual plot [48] shows the prediction errors of the test set compared to the predicted values from the model. The red dashed line indicates zero residual error which means perfect predictions. The majority of the residuals are close to zero, indicating that the model predictions are reasonable over the entire range of predicted values. There is no discrete pattern or structure involved in the residuals, suggesting that the model worked satisfactorily. Some residuals are scattered at higher predicted values above about 0.6, but this suggests slightly higher prediction error in higher parts of the range of eating disorder values. The overall residuals are small and randomly distributed, reassuring that the model performs well without significant biases.

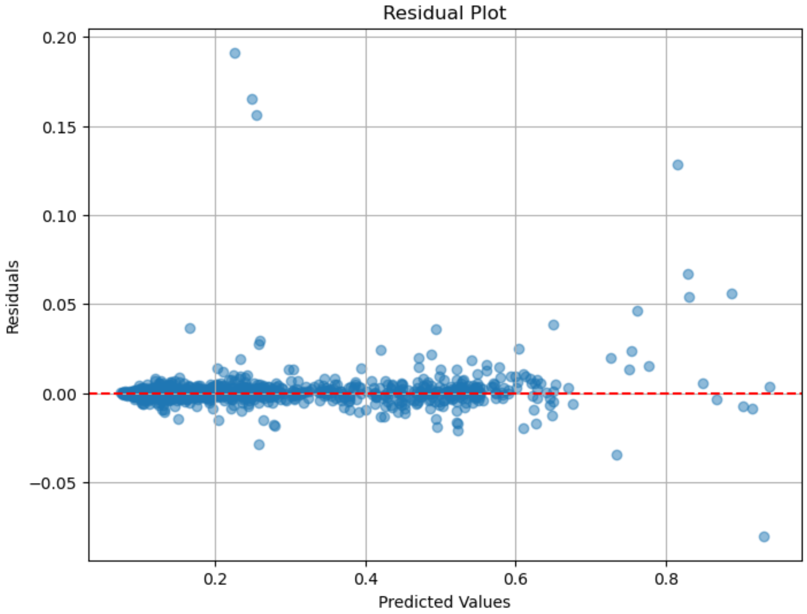


Figure 48: *Residual Plot*

iii) Feature Importance

The feature importance plot [49] is given that indicates which of the features greatly impacts the Eating disorder (%). For the given prediction, the Bipolar disorder (%), Schizophrenia (%) and Drug use disorders (%) are the most important features. These three features are often linked to disordered eating behaviours through shared underlying psychological or neurological mechanisms. All the other features have no effect significant to the Eating disorder (%).

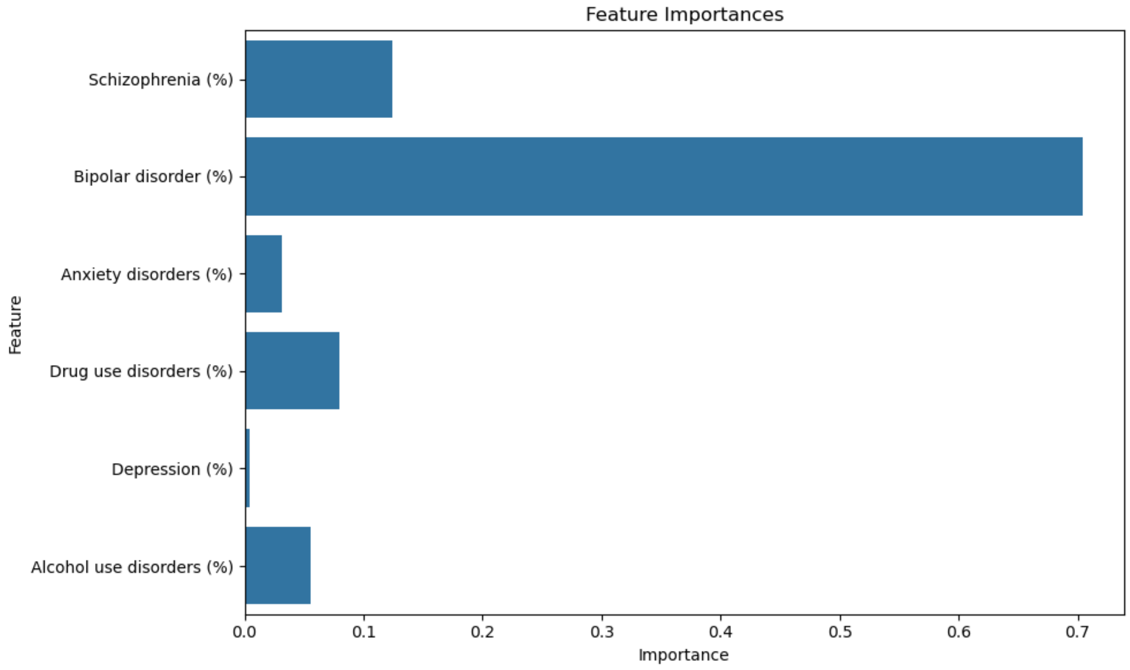


Figure 49: *Feature Importance 4*

4.3.1.4 Significance of the Prediction

In this study, a predictive model was developed to estimate the prevalence of eating disorders based on other mental health indicators at the population level. The model achieved an R² score of 0.9955, indicating a very strong ability to explain variations in eating disorder prevalence across different countries and years. This approach can meaningfully contribute to public health planning by enabling early identification of high-risk regions and supporting the development of effective prevention strategies.

4.4 Causal Inference Results

4.4.1 Causal Analysis of COVID-19's Impact on Anxiety Medication Use

In this part of my analysis, I used the DoWhy causal inference framework and Dataset 2 to investigate whether the COVID-19 pandemic impacted the use of anxiety medication.

4.4.1.1 Approach A : Imbalanced Data

In this approach of the causal inference, the data that I used were the data without any balancing techniques applied to the outcome variable 'ANXMED_A'.

4.4.1.1.1 Yearly Distribution Anxiety Medication Usage

The figure [50] shows the proportion of people taking anxiety medication between years 2019 and 2023. The vertical dashed red line displays the beginning of COVID-19 pandemic in 2020. In the year 2019, 25.4% of the population reported to be taking medication for anxiety. For the year 2021, 25% of the population reported to be taking medication for anxiety. Moreover, in years 2022 and 2023, above cited proportion is 26.6% and 27.2% in respect. Overall, trends of using anxiety medication seem to rise steadily during the years subsequent to COVID-19 pandemic.

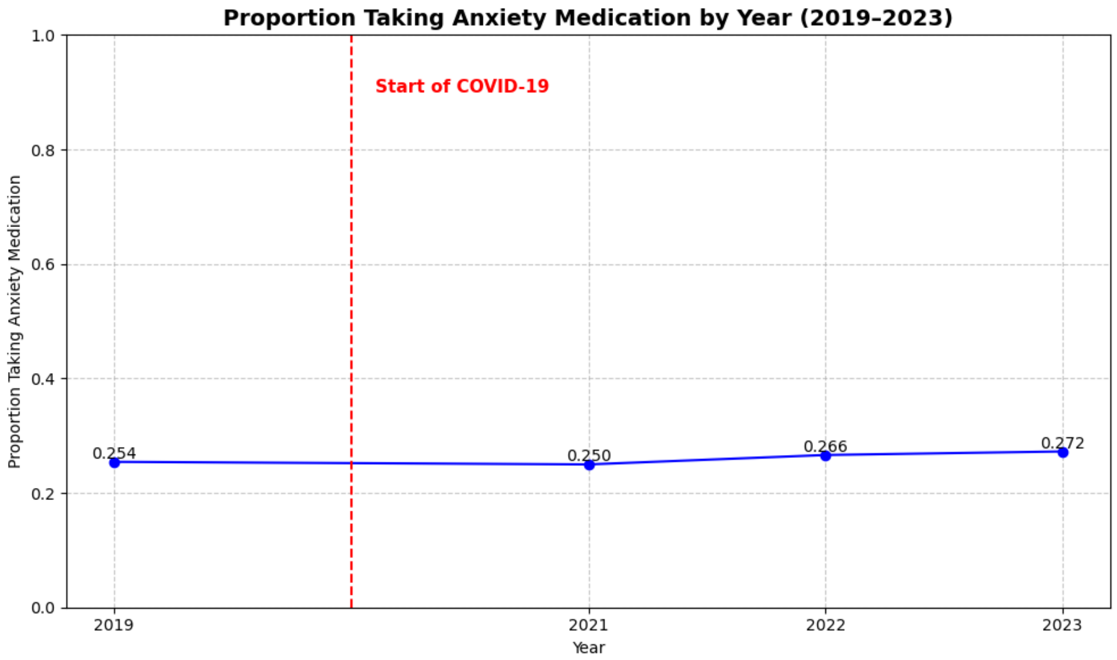


Figure 50: *Plot of Yearly Trends in Anxiety Medication Use (2019–2023) 1*

4.4.1.1.2 Anxiety Medication Use Before and After COVID-19

The plot [51] reveals the proportion of individuals taking anxiety medication before and after inception of the COVID-19 pandemic. The data is divided into 2 periods:

1) Pre-COVID : 2019
2) Post- COVID : 2021-2023

In Pre-COVID period, 25.4% of individuals reported to taking anxiety medication. In Post-COVID period, 26.3% of individuals reported to taking anxiety medication.

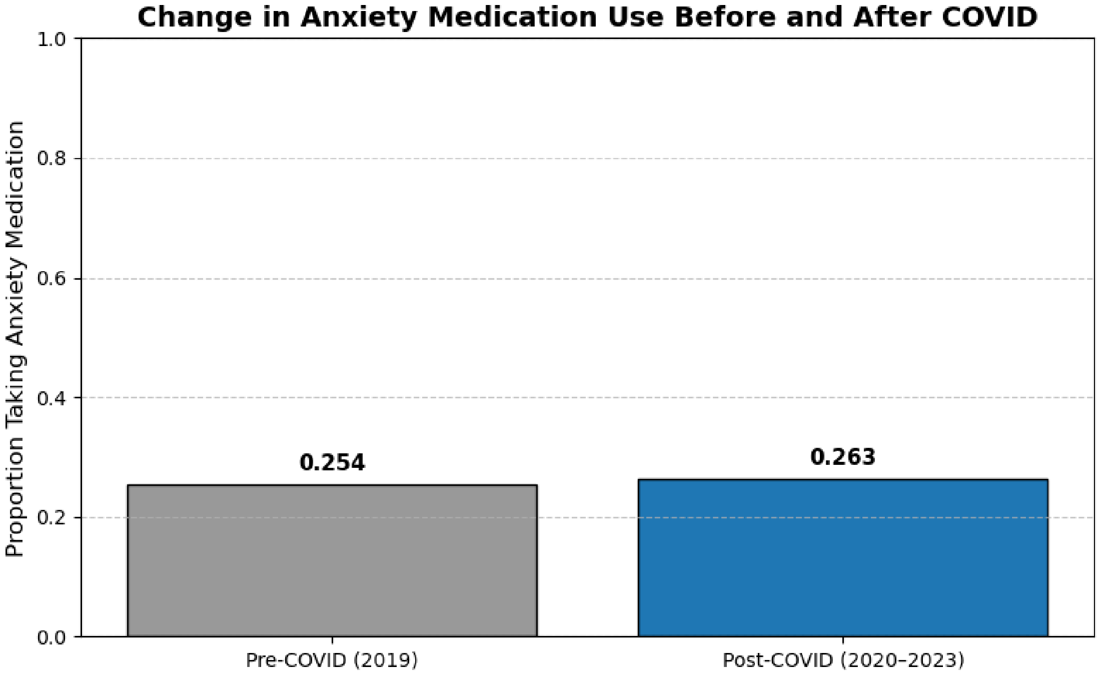


Figure 51: *Plot of Anxiety Medication Pre- and Post- COVID-19 1*

4.4.1.1.3 Explanation

A general insight, comes from the usage of the raw data. The insight is that anxiety medication shows a slight increase after the COVID-19 pandemic. This increase reflects an upward trend which possibly comes from several reasons such as uncertainty for the future and economy and stress. However, it is crucial to recognize that the dataset is not balanced, and that could influence the understanding of the findings. Although, the difference between the 2 periods appears.

4.4.1.2 Approach B : Balanced Data

In this approach of the causal inference, the data that I used were the data with under sampling balancing technique applied to the outcome variable ‘ANXMED_A’.

4.4.1.2.1 Yearly Distribution Anxiety Medication Usage

The figure [52] displays the proportion of individuals taking anxiety medication from 2019 to 2023. Vertical red line marks the start of COVID-19 pandemic in 2020. For the year 2019 the percentage is 25.4% of the people that reported taking anxiety medication. In 2021, the percentage of individuals that reported taking anxiety medication is 25%. Furthermore, in years 2022 and 2023 the aforementioned percentage is 26.2% for both of them. Generally, trends of anxiety medication use, appear to increase steadily the years following COVID-19.

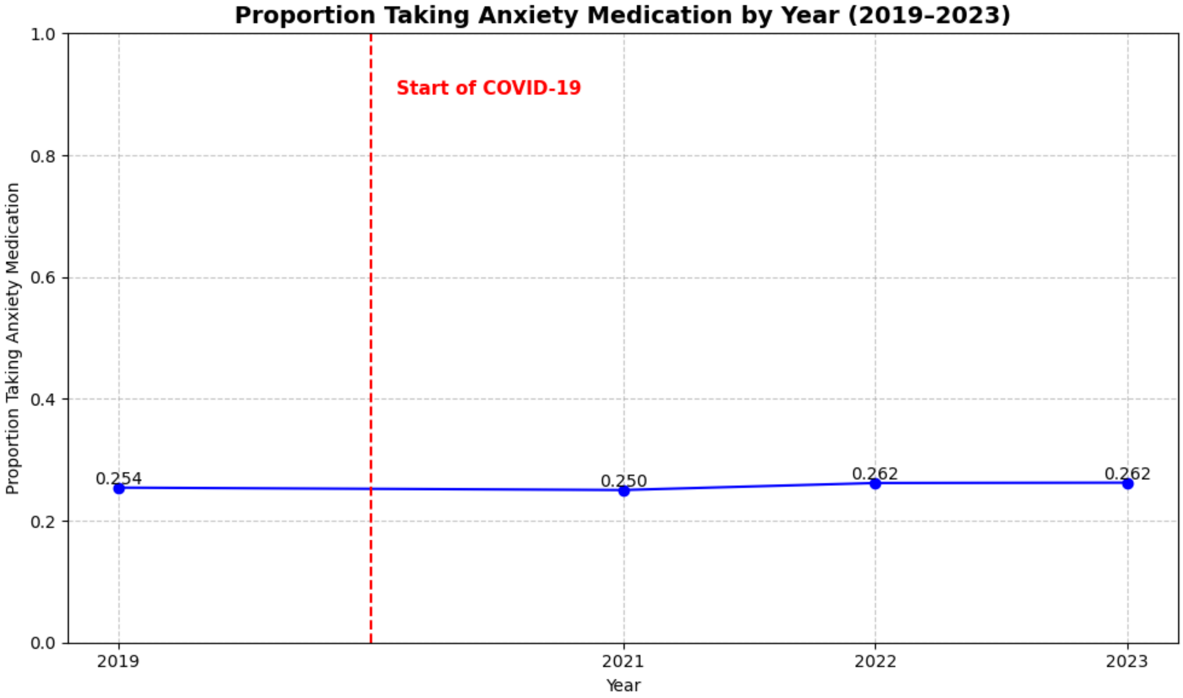


Figure 52: *Plot of Yearly Trends in Anxiety Medication Use (2019–2023) 2*

4.4.1.2.2 Anxiety Medication Use Before and After COVID-19

The graph [53] shows the proportion of people on anxiety medication prior to and after onset of COVID-19 pandemic. Data are split into 2 periods:

1) Pre-COVID : 2019
2) Post- COVID : 2021-2023

During Pre-COVID time, 25.4% of individuals reported taking anxiety medication. During Post- COVID time, 25.8% of individuals reported taking anxiety medication.

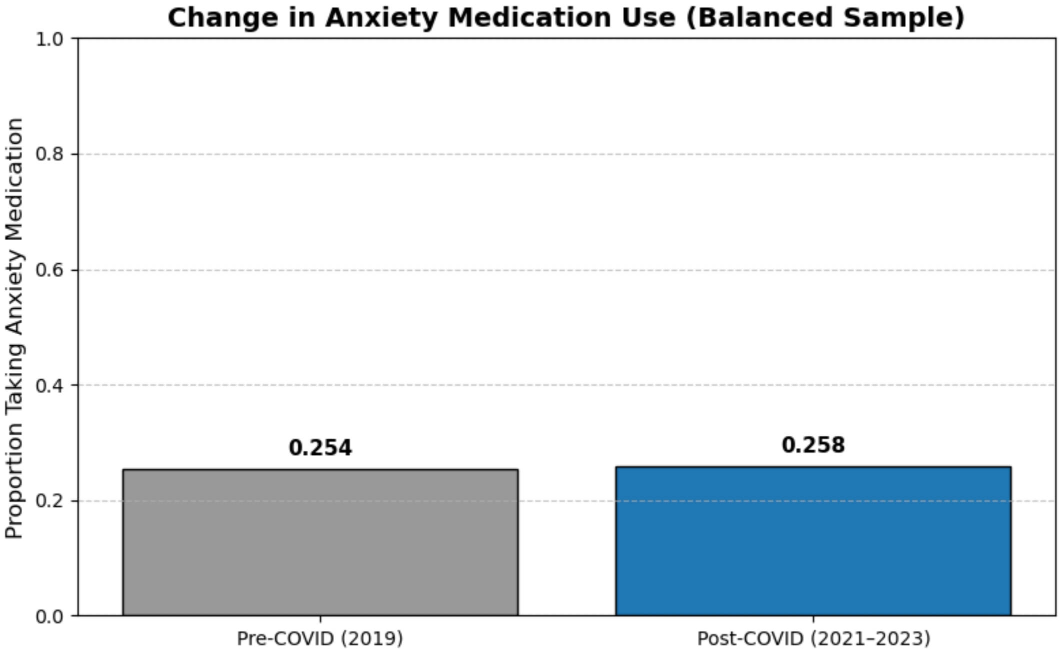


Figure 53: *Plot of Anxiety Medication Pre- and Post- COVID-19 2*

4.4.1.2.3 Causal Inference Explanation

To estimate the causal impact of the COVID-19 pandemic on the use of anxiety medication, I used the DoWhy framework. The linear regression and the propensity score matching were used to calculate the estimated causal impact. The figure [54] presented, linear regression provided the causal estimate of -0.0054. Conversely, propensity score matching yielded a slightly larger causal estimation of -0.0105. The above values enable me to conclude that when controlled variables were accounted, individuals were less likely to report use of anxiety medication following the outbreak of COVID-19 according to their small negative impact. The range of the 95% confidence interval of the estimate of linear regression was between -0.0123 and +0.0014, and it contains zero. This implies that at the level of 5%, the impact is not significant and we do not safely assume that the pandemic impacted positively or negatively the use of anxiety medication according to available information.

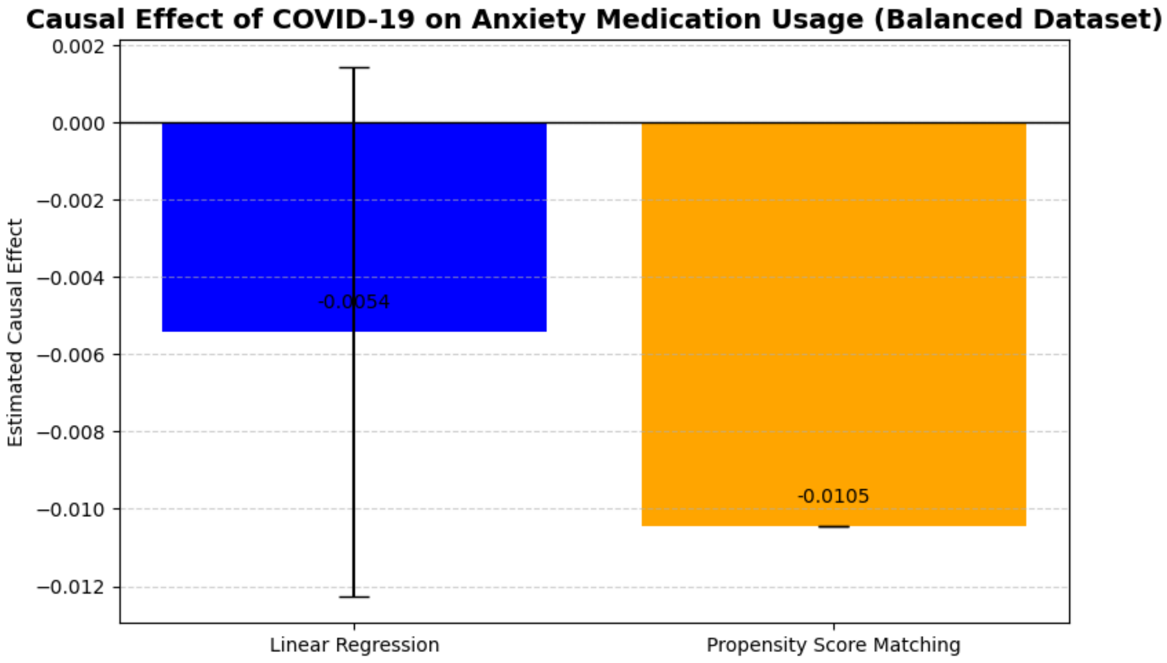


Figure 54: *Causal Effect of COVID-19 on Anxiety Medication Usage*

4.4.2 Causal Analysis of COVID-19's Impact on Depression Medication Use

In this section of my analysis, I used DoWhy Causal Inference framework to investigate if COVID-19 pandemic impacted the usage of depression medication.

4.4.2.1 Approach A : Imbalanced Data

In this approach of the causal inference, the data that I used was the data without any balancing techniques applied to the outcome variable 'DEPMED_A'.

4.4.2.1.1 Yearly Distribution Depression Medication Usage

The following figure [55] depicts proportion of population taking depression medication between years 2019 and 2023. The vertical dashed red line shows the start of COVID-19 pandemic in 2020. In 2019, reporting of 24.1% of population to take depression medication was reported. In case of the year 2021, reporting of 23% of population was reported to take depression medication. Moreover, in years 2022 and 2023, proportion is reported as 24.8% and 24.7%, respectively. Overall, trends of taking depression medication seem to increase steadily in years subsequent to pandemic of COVID-19.

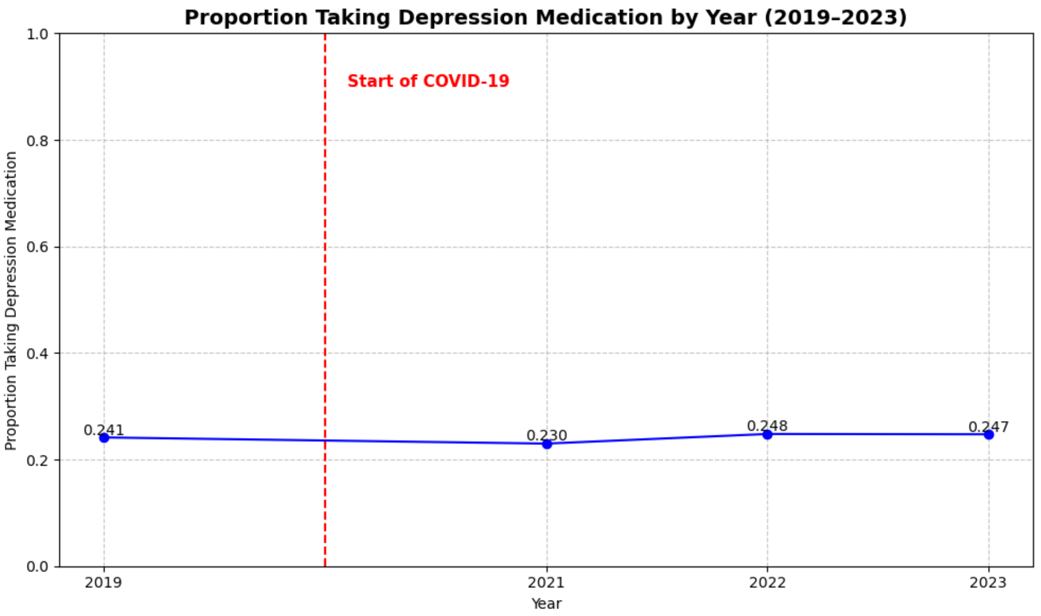


Figure 55*:* *Plot of Yearly Trends in Depression Medication Use (2019–2023) 1*

4.4.2.1.2 Depression Medication Use Before and After COVID-19

The plot [56] reveals the proportion of individuals taking depression medication before and after inception of the COVID-19 pandemic. The data is divided into 2 periods:

1) Pre-COVID : 2019
2) Post- COVID : 2021-2023

In Pre-COVID period, 24.1% of individuals reported to taking depression medication. In Post- COVID period, 24.2% of individuals reported to taking depression medication.

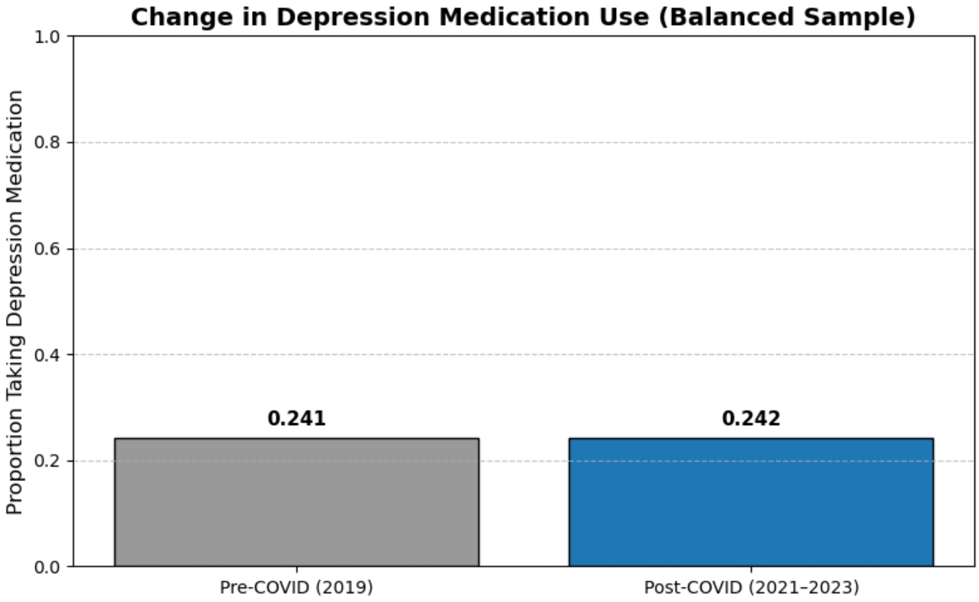


Figure 56: *Plot of Depression Medication Pre- and Post- COVID-19 1*

4.4.2.1.3 Explanation

A wider observation, is obtained with the use of the raw data. The observation is that antidepressant medication indicates a minimal, hardly perceivable increase, following the COVID-19 pandemic. Nonetheless, of paramount consideration is to note that the dataset is imbalanced and therefore may impact the interpretation of findings.

4.4.2.2 Approach B : Balanced Data

In this approach of the causal inference, the data that I used was the data with under sampling balancing technique applied to the outcome variable 'DEPMED_A'.

4.4.2.2.1 Yearly Distribution Anxiety Medication Usage

The figure [57] illustrates the percentage of people using depression medication from 2019 to 2023. The vertical red line indicates the beginning of the COVID-19 pandemic in 2020. In 2019, the percentage of individuals who reported using anxiety medication was 24.1%. In 2021, the proportion of people who indicated they were using depression medication was 23.3%. Additionally, in the years 2022 and 2023, the previously mentioned percentages are 25% and 24.7% correspondingly. Overall, the usage patterns of depression medication seem to rise consistently in the years after COVID-19.

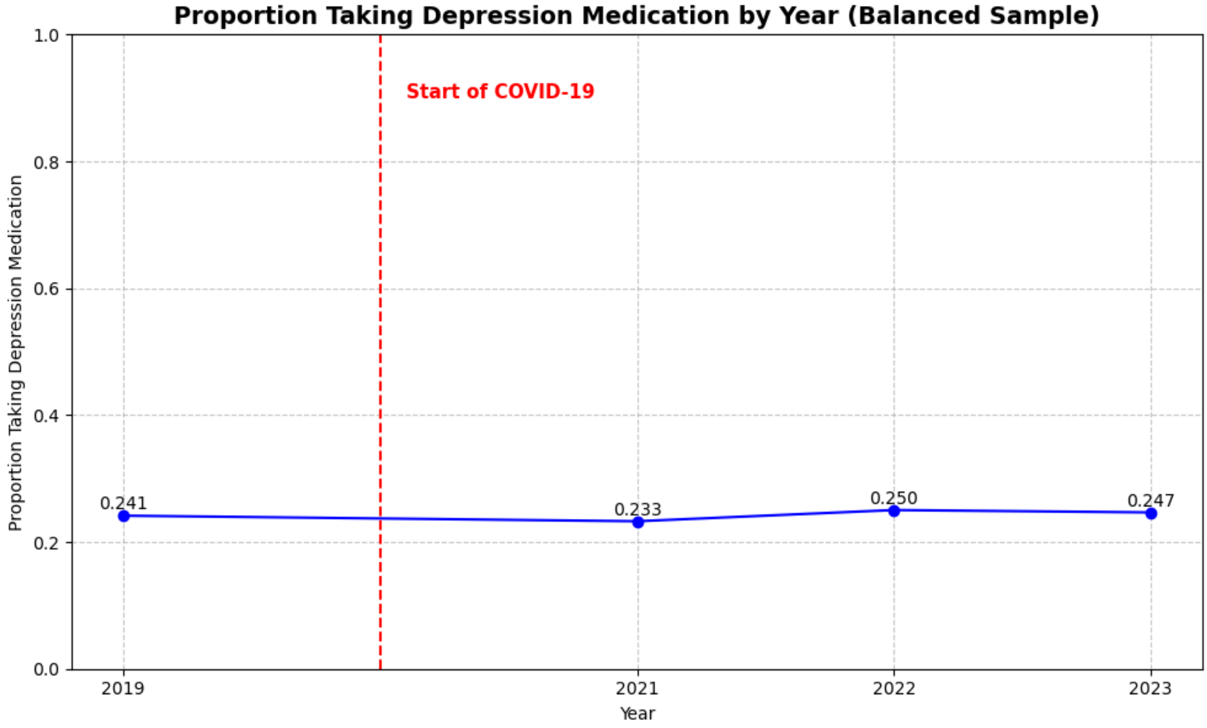


Figure 57: *Plot of Yearly Trends in Depression Medication Use (2019–2023) 3*

4.4.2.2.2 Depression Medication Use Before and After COVID-19.

The graph below shows the proportion of people on depression medication prior to and after onset of COVID-19 pandemic. Data are split into 2 periods:

1) Pre-COVID : 2019
2) Post- COVID : 2021-2023

During Pre-COVID time, 24.1% of individuals reported taking anxiety medication. During Post- COVID time, 24.3% of individuals reported taking anxiety medication.

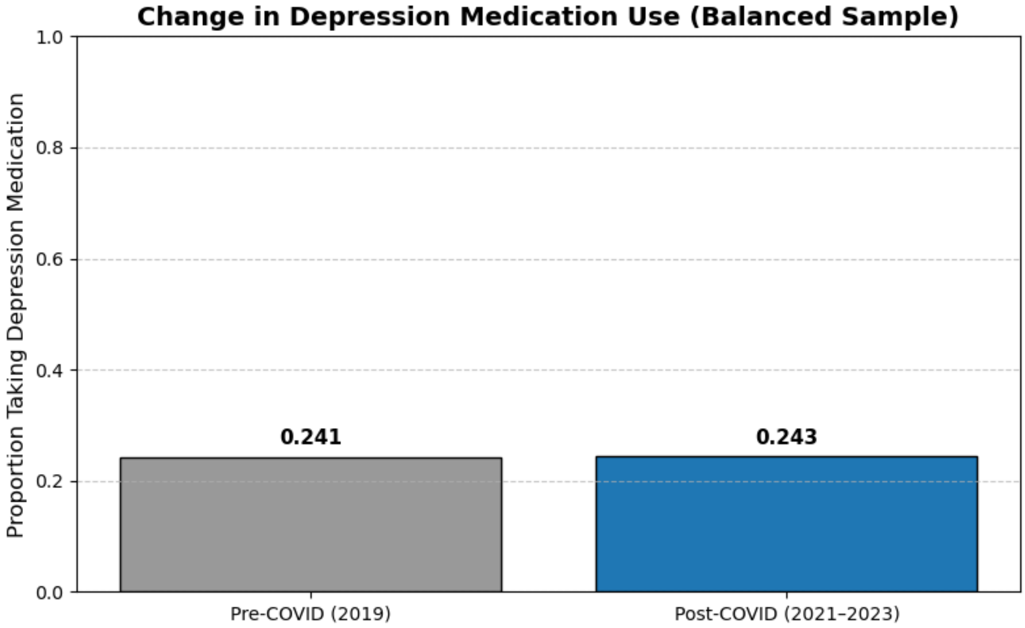


Figure 58: *Plot of Depression Medication Pre- and Post- COVID-19 2*

4.4.2.2.3 Causal Inference Explanation.

To evaluate the causal impact of the COVID-19 pandemic on the use of depression medication, I employed the DoWhy framework with a well-balanced dataset. Two methods were employed, linear regression and propensity score matching. As shown in figure [59], the linear regression method yielded a causal estimate of -0.0013, whereas the PSM technique led to a slightly more robust negative estimate of -0.0049. These marginally negative values indicate that once potential confounders are taken into account, individuals were a bit less inclined to disclose and they report taking depression medication following the COVID-19 pandemic's start. The 95% confidence interval for the linear regression estimate was roughly between -0.008 and +0.004, and as it encompasses zero, the effect is not statistically significant at the 5% level.

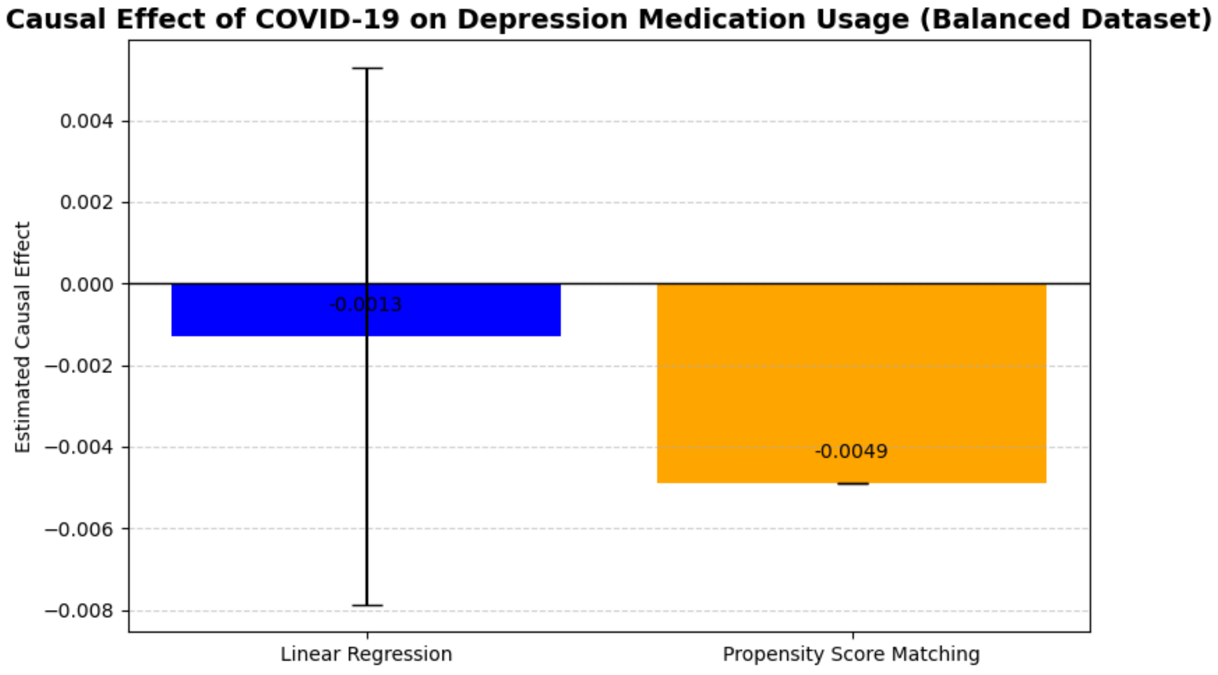


Figure 59: *Causal Effect of COVID-19 on Depression Medication Usage*

The confounders I used to infer causality can be verified by looking at the figure [60] below, as they are directly related to both the treatment, which is COVID_Indicator, and the outcome which was the ANXMED_A and DEPMED_A separately to each case of causal inference. The causal estimate thus illustrates the isolated effect of the COVID-19 pandemic on the use of anxiety and depression medication, making the analysis more valid.

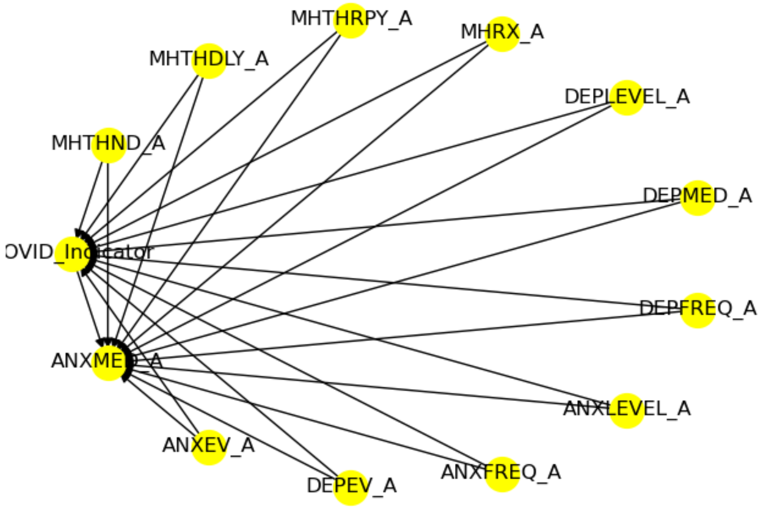


Figure 60: *Causal Network of COVID-19 Impact on Mental Health Variables*

Chapter 5

Conclusion

5.1 Final Summary

The purpose of this thesis was to investigate and apply data analysis and machine learning techniques to mental health data. As mental health is a critical global concern because a lot of people affected by several mental health disorders. These data analysis, gives several important insights about mental health that can support early detection and prevention strategies.

Throughout the study, several predictive models were developed to identify individuals likely to seek treatment for mental health issues. Those who are currently taking medication for anxiety or depression, or to predict the population-level prevalence of eating disorders based on other existing mental health conditions. In addition, causal inference techniques were used to examine the impact of the COVID-19 pandemic on mental health, specifically related to medication usage.

This thesis, may mobilise governments and organizations for spreading awareness for mental health. Moreover, people may consider and concern more about their mental health because it is as major as physical health.

5.2 Challenges and Limitations

The first limitation of this work was my knowledge and my inexperience based on machine learning and data analysis. However, through continuous research and practice during the thesis, I significantly improved my understanding and skills in these areas.

The second limitation was related to finding datasets. Initially, I found datasets but during the analysis process, I found out that some of them were not based on real world data, and many did not follow a normal distribution. There were few datasets which were clearly related to mental health because of the sensitivity of the subject. Eventually, I was able to identify and use reliable, real-world datasets that were suitable for the goals of this thesis.

5.3 Ethics and Responsible Use

The data utilized in this thesis was sourced from publicly accessible and anonymized datasets. No personal or identifiable data was present, guaranteeing that privacy and confidentiality were preserved during the study. Due to the delicate nature of mental health information, particular attention was given to handling the data ethically and with respect.

The machine learning models created in this thesis aim to enhance research and awareness within the mental health. These models are not intended for clinical diagnosis or treatment choices. It is crucial that the application of these models in real life settings is led by mental health experts and backed by suitable context and understanding.

Furthermore, it is acknowledged that mental health datasets may include biases related to gender, culture, or workplace conditions.

5.4 Future Work

Future work could focus on enhancing fairness, transparency, and inclusivity in the development of predictive models for mental health applications. Additionally, classes with limited representation could be merged based on semantic similarity to improve model performance. Further research could also explore expanded prediction of various mental health disorders and apply causal inference methods to better understand underlying factors.

References

[1] “sklearn.impute.SimpleImputer — scikit-learn 0.24.1 documentation,” *scikit-learn.org*.
Available:
https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html

[2] “sklearn.impute.IterativeImputer,” *scikit-learn*.
Available:
https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

[3] Scikit-learn, “sklearn.preprocessing.LabelEncoder — scikit-learn 0.22.1 documentation,” *Scikit-learn.org*, 2019. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html

[4] “sklearn.preprocessing.OrdinalEncoder,” *scikit-learn*.
Available:https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OrdinalEncoder.html

[5] Scikit-learn, “sklearn.preprocessing.OneHotEncoder” *Scikit-learn.org*, 2019. Available:https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html

[6] scikit-learn, “StandardScaler,” *scikit-learn.org*, 2019. Available: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[7] “Box Cox Transformation: Definition, Examples,” *statisticshowto.com*. Aug. 20, 2021. Available: https://www.statisticshowto.com/probability-and-statistics/normal-distributions/box-cox-transformation/

[8] scikit-learn, “sklearn.decomposition.PCA ,” *Scikit-learn.org*, 2009. Available: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[9] “Classifier comparison,” *scikit-learn*. Available: https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

[10] scikit-learn, “sklearn.metrics.f1_score” *Scikit-learn.org*, 2019. Available: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[11] “sklearn.pipeline.Pipeline — scikit-learn 0.24.1 documentation,” *scikit-learn.org*. Available: https://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

[12] Open Sourcing Mental Illness, LTD, “Mental Health in Tech Survey,” *Kaggle.com*, 2016. Available: https://www.kaggle.com/datasets/osmi/mental-health-in-tech-survey/data?select=survey.csv

[13] CDC, “2019 NHIS Questionnaires, Datasets, and Documentation,” *National Health Interview Survey*, Nov. 21, 2024.
Available: https://www.cdc.gov/nchs/nhis/documentation/2019-nhis.html

[14] CDC, “2020 NHIS Questionnaires, Datasets, and Documentation,” *National Health Interview Survey*, Nov. 21, 2024.
Available: https://www.cdc.gov/nchs/nhis/documentation/2020-nhis.html

[15] CDC, “2021 NHIS Questionnaires, Datasets, and Documentation,” *National Health Interview Survey*, Nov. 21, 2024.
Available: https://www.cdc.gov/nchs/nhis/documentation/2021-nhis.html

[16] CDC, “2022 NHIS Questionnaires, Datasets, and Documentation,” *National Health Interview Survey*, Nov. 21, 2024.
Available: https://www.cdc.gov/nchs/nhis/documentation/2022-nhis.html

[17] CDC, “2023 NHIS Questionnaires, Datasets, and Documentation,” *National Health Interview Survey*, Nov. 21, 2024.
Available: https://www.cdc.gov/nchs/nhis/documentation/2023-nhis.html

[18] CDC, “NHIS 2019 English Questionnaire,” *CDC FTP Server*. Available: https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Survey_Questionnaires/NHIS/2019/EnglishQuest.pdf

[19] The Devastator, “Global Mental Health Disorders,” *Kaggle.com*, 2017.
Available: https://www.kaggle.com/datasets/thedevastator/global-mental-health-disorders?select=Mental+health+Depression+disorder+Data.csv

[20] A. Jaiswal, P. C. Jha and R. Tripathi, “Mental Health Detection using Machine Learning,” in *Proc. 6th International Conf. Computing, Communication and Automation (ICCCA)*,2020.
Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9214815

[21] “What is supervised learning? Machine learning tasks,” *SuperAnnotate*, Oct. 28, 2022.
Available: https://www.superannotate.com/blog/supervised-learning-and-other-machine-learning-tasks#types-of-machine-learning-models

[22] V. Chugh, “Python pandas tutorial: The ultimate guide for beginners,” *www.datacamp.com*, May 30, 2023.
Available: https://www.datacamp.com/tutorial/pandas

[23] K. Babitz, “Matplotlib Tutorial: Python Plotting,” *www.datacamp.com*, May 30, 2023. Available: https://www.datacamp.com/tutorial/matplotlib-tutorial-python

[24] “Python NUMPY Array TUTORIAL,” *www.datacamp.com*.
Available: https://www.datacamp.com/tutorial/python-numpy-tutorial

[25] “Casual Inference - The Decision Lab,” *The Decision Lab*, 2025.. Available: https://thedecisionlab.com/reference-guide/statistics/casual-inference

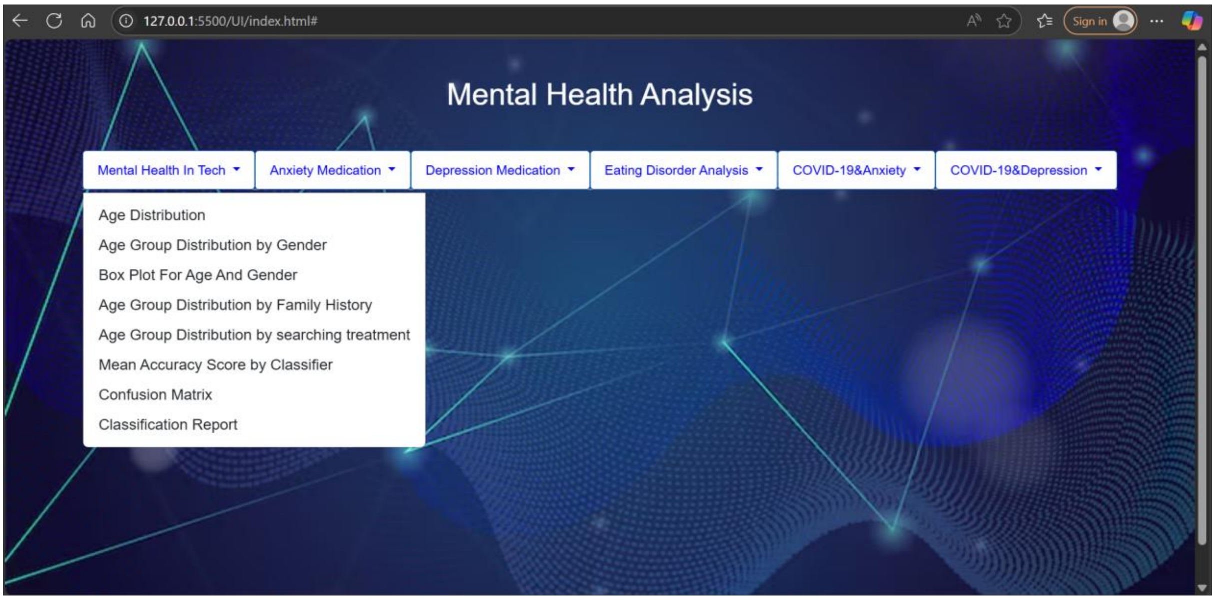
[26] Shriya Wakdevi Kuppa, K. Jadhav, and Shraddha Sonone, “Mental Health Analysis Using Machine Learning,” *International Journal of Scientific Research and Technology*, Dec. 2024, Available: https://doi.org/10.5281/zenodo.14365295

Appendix A

When all the data analysis experiments finished, a User Interface was created by displaying the exploratory data analysis and the results of the predictions. The User Interface was created by using HTML5, CSS and JavaScript.

A.1 Screens

The navigation bar displays all the aforementioned predictions and causal inference options. When one of these is selected, a list of possible buttons is displayed. Each button corresponds to a plot that is presented in Chapter 4.



127.0.0.1:5500/UI/index.html#
Sign in
Mental Health Analysis
Mental Health In Tech
Anxiety Medication
Depression Medication
Eating Disorder Analysis
COVID-19&Anxiety
COVID-19&Depression
Age Distribution
Age Group Distribution by Gender
Box Plot For Age And Gender
Age Group Distribution by Family History
Age Group Distribution by searching treatment
Mean Accuracy Score by Classifier
Confusion Matrix
Classification Report

Figure A.1.1: *Possible Buttons for Mental Health in Tech*

When a button is selected, the corresponding plot will be displayed.

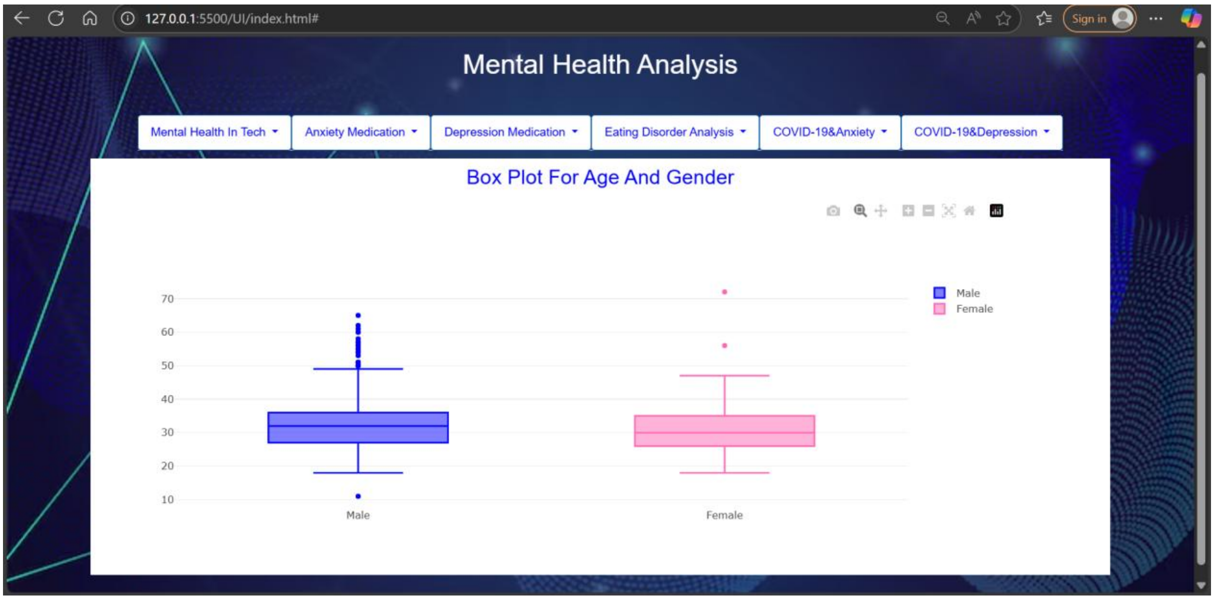


Figure A.1.2: *Example of selecting the "Box Plot For Age And Gender" button*

Below, all the possible buttons for every prediction and causal inference option are displayed:

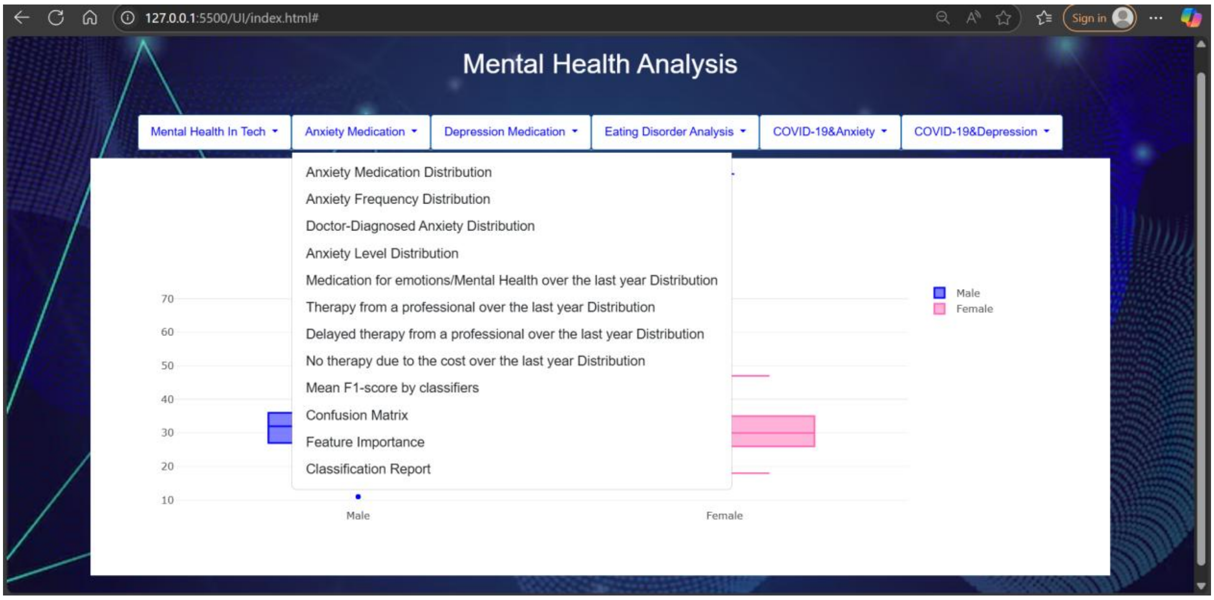

127.0.0.1:5500/UI/index.html#
Sign in
Mental Health Analysis
Mental Health In Tech
Anxiety Medication
Depression Medication
Eating Disorder Analysis
COVID-19&Anxiety
COVID-19&Depression
Anxiety Medication Distribution
Anxiety Frequency Distribution
Doctor-Diagnosed Anxiety Distribution
Anxiety Level Distribution
Medication for emotions/Mental Health over the last year Distribution
Therapy from a professional over the last year Distribution
Delayed therapy from a professional over the last year Distribution
No therapy due to the cost over the last year Distribution
Mean F1-score by classifiers
Confusion Matrix
Feature Importance
Classification Report
70
60
50
40
30
20
10
Male
Female
Male
Female

Figure A.1.3: *Possible Buttons for Anxiety Medication*

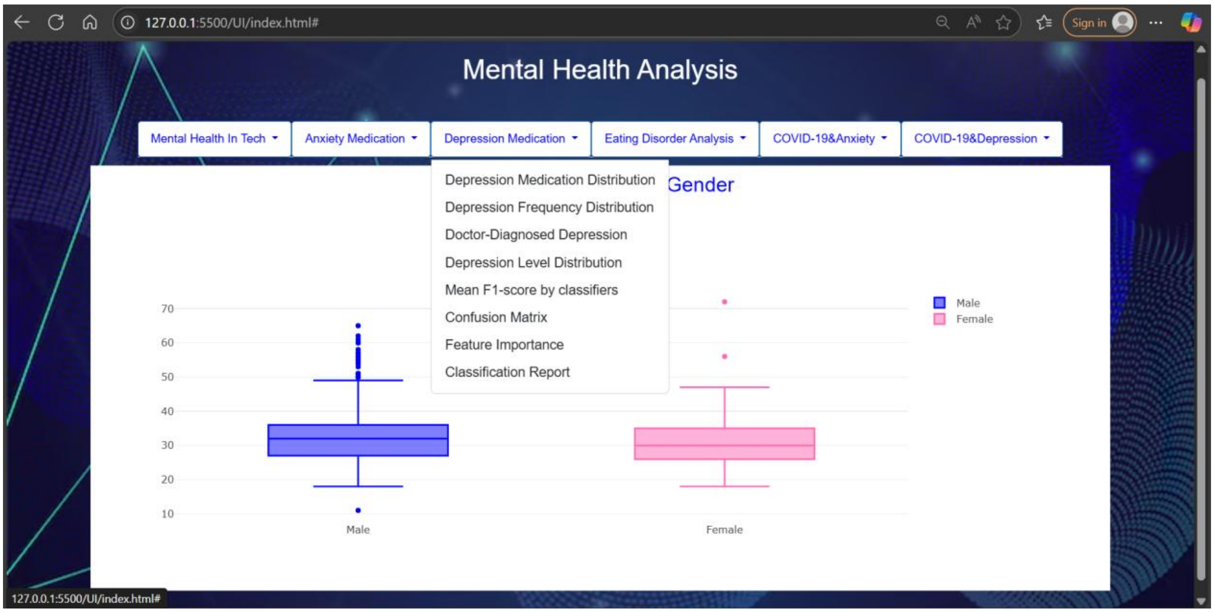


Figure A.1.4: *Possible Buttons for Depression Medication*

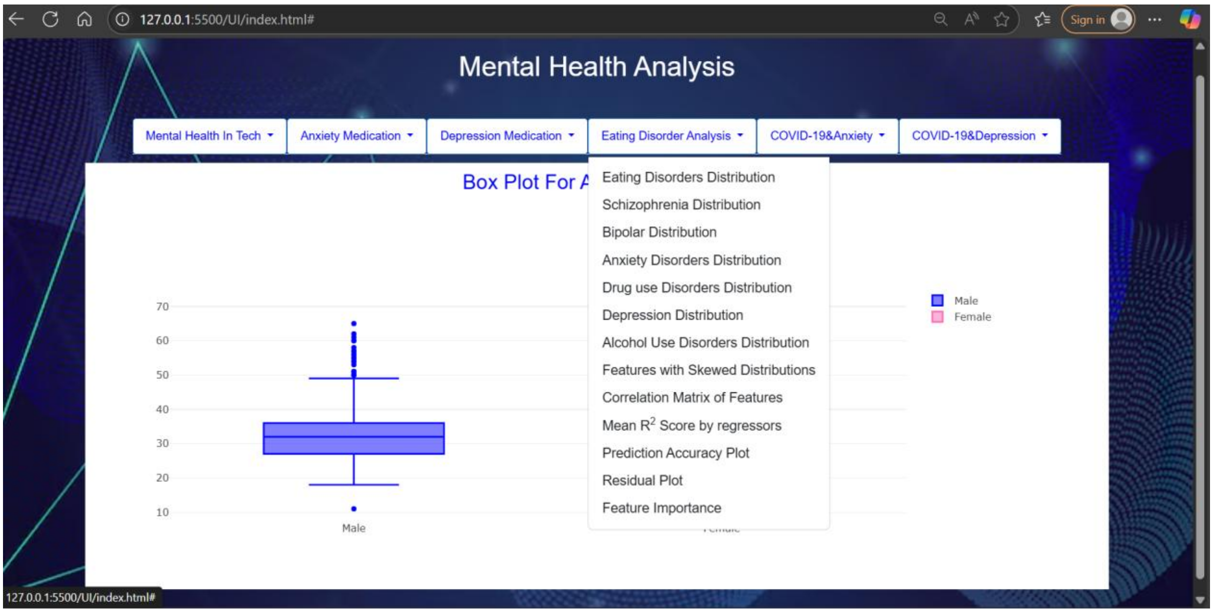

127.0.0.1:5500/UI/index.html#
Sign in
Mental Health Analysis
Mental Health In Tech
Anxiety Medication
Depression Medication
Eating Disorder Analysis
COVID-19&Anxiety
COVID-19&Depression
Box Plot For A
Eating Disorders Distribution
Schizophrenia Distribution
Bipolar Distribution
Anxiety Disorders Distribution
Drug use Disorders Distribution
Depression Distribution
Alcohol Use Disorders Distribution
Features with Skewed Distributions
Correlation Matrix of Features
Mean R² Score by regressors
Prediction Accuracy Plot
Residual Plot
Feature Importance
70
60
50
40
30
20
10
Male
Male
Female
127.0.0.1:5500/UI/index.html#

Figure A.1.5: *Possible Buttons for Eating Disorders Analysis*

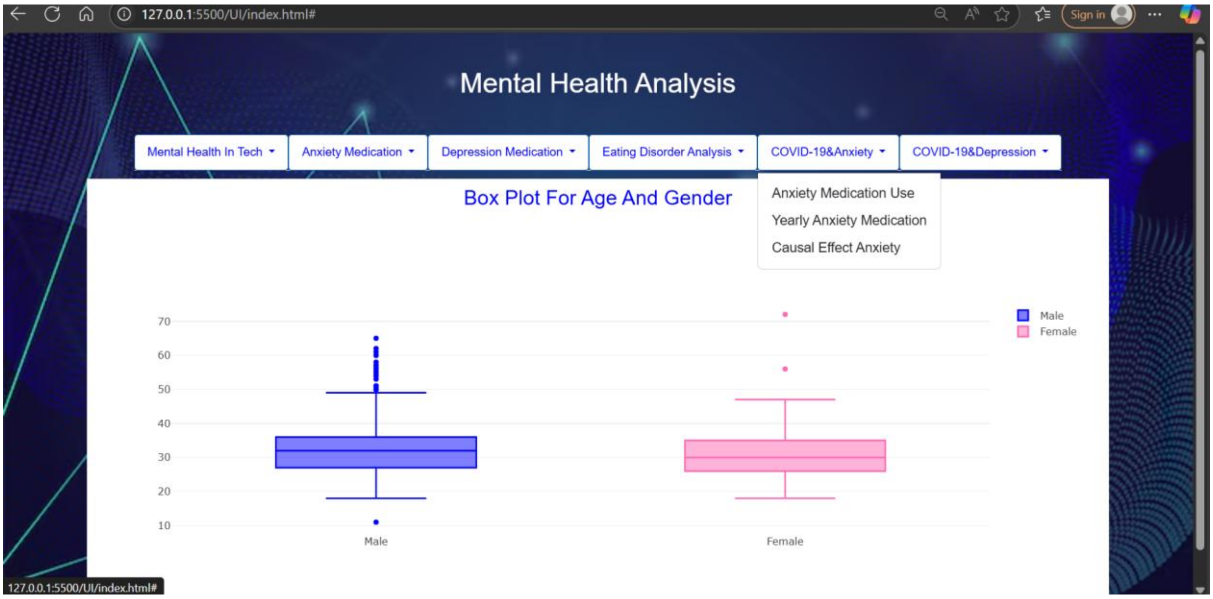


Figure A.1.6: *Possible Buttons for COVID-19 & Anxiety*

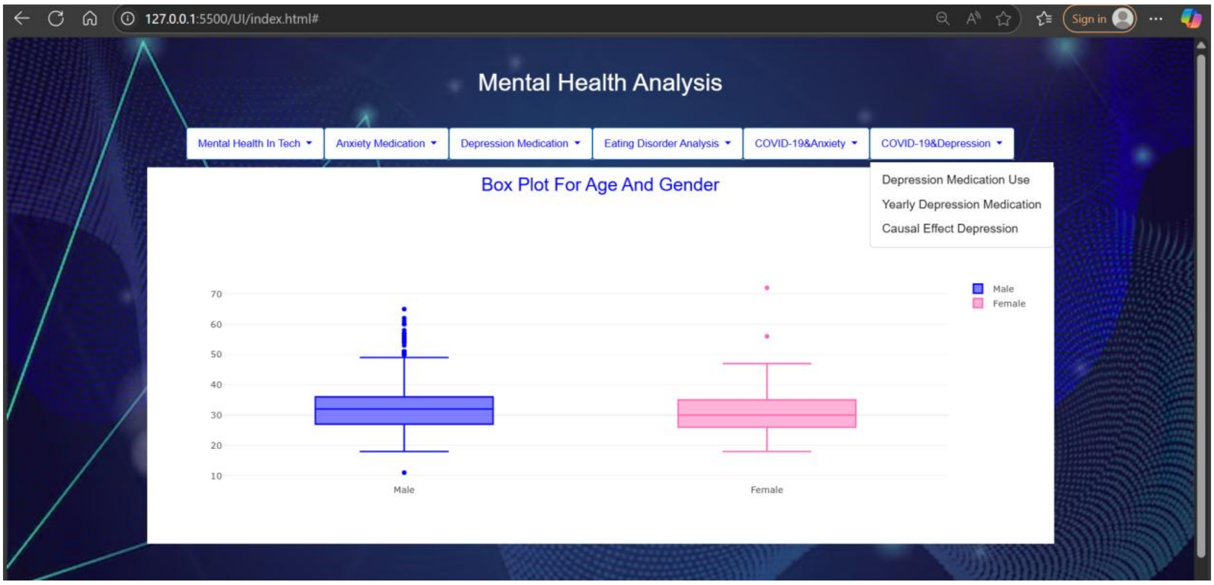


Figure A.1.7: *Possible Buttons for COVID-19 & Depression*