Advanced Machine Learning Approaches for Predicting Mental Health Disorders Following Long COVID Diagnosis

Manoj Purohit
Marquette University

Follow this and additional works at: https://epublications.marquette.edu/theses_open

Part of the Computer Sciences Commons

Recommended Citation
Purohit, Manoj, "Advanced Machine Learning Approaches for Predicting Mental Health Disorders Following Long COVID Diagnosis" (2024). Master's Theses (2009 -). 796.
https://epublications.marquette.edu/theses_open/796
ADVANCED MACHINE LEARNING APPROACHES FOR PREDICTING MENTAL HEALTH DISORDERS FOLLOWING LONG COVID DIAGNOSIS

By

Manoj Purohit

A Thesis submitted to the Faculty of the Graduate School, Marquette University, in Partial Fulfillment of the Requirements for the Degree of Master of Science

Milwaukee, Wisconsin

May 2024
ABSTRACT

After the global spread of COVID-19, the enduring effects of Long COVID and its health implications have emerged as a significant global issue, affecting people worldwide. The lingering symptoms post a COVID-19 infection can significantly affect individuals who had previously contracted the virus, exerting considerable influence over their mental well-being. Prolonged recuperation associated with Long COVID has been connected with the emergence of symptoms such as depression and anxiety, all of which can have adverse effects on emotional health. This project delves into an in-depth analysis of healthcare data pertaining to Long COVID from the Froedtert Health (FH) Medical System in Wisconsin, United States. Through the application of advanced Machine Learning (ML) techniques, we present predictive models aimed at assessing the risk of developing Mental Health Disorders (MHD) in patients diagnosed with Long COVID. Our study also encompasses the identification of pivotal features impacting MHD. To thoroughly investigate the factors that have a substantial impact on MHD, we employed the Recursive Feature Elimination (RFE) technique to carefully pick out essential attributes from our dataset. Given the dataset’s inherent imbalance, we have employed the Synthetic Minority Over-sampling Technique and Edited Nearest Neighbors (SMOTEEN) technique to effectively address this issue. Multiple ML models have been meticulously constructed and validated using cross-validation methodologies. The results indicate that Random Forest (RF) Classifier shows better performance in comparison to other models with an area under the ROC curve (AUC) of 0.97, precision of 0.90, and recall of 0.89. Remarkably, the XGBoost Classifier also demonstrates strong predictive abilities for MHD, achieving an AUC of 0.90, precision of 0.79, and recall of 0.82. Ultimately, the crucial features identified through our predictive models hold the potential to identify individuals at risk of MHD, facilitating the delivery of targeted preventive care and essential resources.
ACKNOWLEDGMENTS

I extend my heartfelt gratitude to my advisor, Dr. Praveen Madiraju, for his unwavering mentorship and profound guidance throughout my academic journey. His wisdom has been a beacon of inspiration, profoundly shaping my growth as a scholar.

My sincere appreciation also goes to Dr. Keke Chen and Dr. Walter Bialkowski, esteemed members of my dissertation committee. Their invaluable contributions, continuous encouragement, and insightful feedback have been instrumental in the success of my thesis. Their teachings have significantly enriched my academic prowess.

I am thankful to Dr. Priyanka Annapureddy for her expert assistance during the critical data preprocessing phase of my research. Her expertise was pivotal in laying the groundwork for my study.

I extend my deepest thanks to my wife, Priya, who believed in the value of education and supported my pursuit of higher learning. Her constant encouragement and understanding made it possible for me to achieve my academic goals and complete this thesis.

I would like to express my love and thanks to my son, Aariv, whose boundless positivity and joy have been a constant source of motivation and light.

I am also thankful for the Computational Sciences Summer Research Fellowships (CSSRF) provided by Marquette University, which made this project possible.
# TABLE OF CONTENTS

Acknowledgments ................................................................. iii

List of Tables ................................................................. vii

List of Figures ................................................................. viii

Chapter 1: Introduction ....................................................... 1

  1.1 Background ............................................................. 1
  
  1.2 Aims and Objectives ................................................... 2
  
  1.3 Problem Statements ................................................... 2
      1.3.1 Problem Statement 1: .......................................... 2
      1.3.2 Problem Statement 2: .......................................... 2
  
  1.4 Special Emphasis: ..................................................... 3

Chapter 2: Related work ....................................................... 4

  2.0.1 Introduction: .......................................................... 4
  
  2.0.2 Recent Study on MHD Due to Chronic Condition of Long COVID . 4
  
  2.0.3 Machine Learning, Deep Learning, and Artificial Intelligence in MHD Prediction ................................................. 5
  
  2.0.4 Innovative Approach and Contribution ............................. 6
Chapter 3: Research Methodology ................................. 7
  3.0.1 Data Source ............................................. 8
  3.0.2 Data Preparation and Transformation ..................... 9
  3.0.3 Patient Categorization: Mental Health Diagnoses After Long COVID 9
  3.0.4 Data Characteristics ................................... 10
  3.0.5 Correlation Analysis .................................... 12
  3.0.6 Feature selection ....................................... 14
  3.0.7 Normalization and Data Balancing ......................... 15

Chapter 4: Exploring Machine Learning Approach .............. 17
  4.1 Machine Learning Algorithms ............................... 17
    4.1.1 Random Forest Algorithm (RF) .......................... 17
    4.1.2 Decision Tree (DT) .................................... 18
    4.1.3 Naive Bayes Algorithm (NB) ............................ 20
    4.1.4 Gradient Boosting Classifier (GB) ...................... 20
    4.1.5 Support Vector Machine (SVM) ........................ 21
    4.1.6 K Nearest Neighbour (KNN) ............................ 21
    4.1.7 XGBoost ................................................ 22
    4.1.8 Multi-Layer Perceptron (MLP) .......................... 22

Chapter 5: Model Evaluation ................................. 25
  5.1 Domain-Specific Evaluation Metrics for Mental Health Disorder Prediction 25
    5.1.1 Accuracy (ACC) ....................................... 25
    5.1.2 Precision .............................................. 25
LIST OF TABLES

3.1 The list of ICD Codes related to MHD and Long COVID ............... 8
3.2 The type of encounters in EHR ............................................. 9
3.3 Statistics for final MHD study population ............................. 13
3.4 Features correlation with MHD ............................................. 14
### LIST OF FIGURES

3.1 Methodological framework for predicting MHDs after the diagnosis of Long COVID .................................................. 7

3.2 Flowchart illustrating the selection of the sample used in study ........... 11

3.3 REF Score Vs Number of selected features .................................. 15

4.1 Machine-learning algorithms for prediction ................................. 18

4.2 Random Forest algorithm ......................................................... 19

4.3 Decision Tree algorithm for classification ................................. 19

4.4 Naïve Bayes algorithm ............................................................ 20

4.5 Gradient Boosting classifier ...................................................... 21

4.6 Support Vector Machine classifier ............................................ 22

4.7 K Nearest Neighbour .............................................................. 23

4.8 A general architecture of XGBoost ......................................... 23

4.9 Multi-Layer Perceptron ........................................................... 24

6.1 Heatmap for the performance measure of all classification methods ...... 29

6.2 Average Receiver Operating Characteristic (ROC) curves ............... 30

6.3 Decision tree predicting the Mental Health Disorder with Decision Tree classifier (max depth=3) .......................... 31
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4</td>
<td>Decision tree predicting the risk of Mental Health Disorder with Random Forest classifier (max depth=3)</td>
<td>31</td>
</tr>
<tr>
<td>6.5</td>
<td>Importance of RF classifiers (Feature importance attribute)</td>
<td>32</td>
</tr>
<tr>
<td>6.6</td>
<td>Feature Importance of XGBoost classifiers (Gain Function)</td>
<td>33</td>
</tr>
<tr>
<td>6.7</td>
<td>Confusion Matrix RF classifiers</td>
<td>33</td>
</tr>
<tr>
<td>6.8</td>
<td>Confusion Matrix XGBoost classifier</td>
<td>34</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

1.1 Background

Long COVID, affecting a subset of COVID-19 survivors with symptoms extending beyond four weeks, is linked to diverse complications and constitutes a significant global health concern [1], persisting even after the initial clearance of the COVID-19 infection. A concerning estimate suggests that around 65 million people globally are dealing with Long COVID and its lasting effects[2], highlighting the pressing need to fully comprehend its consequences [3]. While the focus has largely been on the mental health aspects, it is crucial to recognize that Long COVID has engendered an array of chronic conditions, spanning from pulmonary fibrosis and chronic fatigue syndrome to hypertension and chronic heart diseases [4], imposing a dual physical and psychological burden that necessitates a comprehensive approach [5]. This increasing burden of chronic ailments not only takes a toll on individuals but reverberates across economies, leading to substantial healthcare costs, reduced productivity, and an overall diminished quality of life.

The significant role of ML techniques in forecasting the outcomes of COVID-19, as well as in identifying the susceptibility to Long COVID and its associated MHDs, is indeed noteworthy [6] [7]. In line with this trajectory, this study embarks on leveraging ML to meticulously analyze Electronic Health Records (EHR) data, discern risk factors associated with Long COVID-related MHDs, and devise targeted interventions to mitigate their impact. This research marks a paradigm shift in the management of Long COVID-related MHDs, harnessing the prowess of ML algorithms [8]. The fundamental foundation of this work is the assimilation of diverse data sources from EHR, a holistic approach aimed at constructing predictive models of utmost accuracy. Through the synergy of ML methodologies and advanced feature engineering, the objective is to unravel intricate patterns and identify pivotal variables that contribute to the development of MHD in the aftermath of Long COVID. The main goal is to outline patient subgroups that exhibit increased vul-
nerability, thereby facilitating tailored treatment plans that enable timely interventions and ultimately enhance patient outcomes, thereby alleviating the escalating healthcare costs.

1.2 Aims and Objectives

The objective of this research is to construct a predictive model using ML techniques capable of accurately forecasting the probability of individuals with Long COVID developing MHDs. This will involve analyzing large-scale electronic health records (EHR) to identify patterns and risk factors associated with these disorders using supervised and ensemble ML algorithms, with the aim of improving patient outcomes. The research question will be investigated by evaluating the performance of various ML algorithms, with the ultimate goal of identifying the most accurate and reliable predictive model. This model will inform healthcare decision-making and enhance the quality of care for patients with mental health disorders related to Long COVID.

1.3 Problem Statements

1.3.1 Problem Statement 1:

Determining which ML model(s) can be applied to forecast the likelihood of developing mental health disorders related to Long COVID by analyzing large-scale EHR data.

Research Question:

Can ML models be developed using large-scale EHR data to accurately predict which patients are at high risk of developing Long COVID-related mental disorders, considering demographics such as age, gender, race, ethnicity, and underlying medical conditions ?.

1.3.2 Problem Statement 2:

Assessing the predictive performance of ML models in accurately predicting MHDs related to Long COVID.
Research Question:

What are the key performance metrics, including accuracy, sensitivity, specificity, F-1 score, and area under the curve (AUC), that can be used to evaluate the predictive performance of ML models in accurately predicting the likelihood of MHDs related to Long COVID?

1.4 Special Emphasis:

This project focuses on predicting MHDs and related disabilities due to Long COVID, aligning with one of the research priorities from the Department of Health and Human Services and the National Research Action Plan on Long COVID
CHAPTER 2
RELATED WORK

2.0.1 Introduction:
In recent years, the application of ML, Deep Learning (DL), and Artificial Intelligence (AI) techniques to predict MHDs has gained significant attention. Researchers have explored various methodologies to improve the accuracy and early detection of MHD, leveraging the capabilities of advanced computational algorithms. This section provides a comprehensive review of the existing literature, highlighting key studies, methodologies, and contributions in the domain of MHD prediction.

2.0.2 Recent Study on MHD Due to Chronic Condition of Long COVID
Emerging research has delved into the intricate relationship between persistent COVID-19 symptoms and subsequent mental health challenges. Notably, a comprehensive study by Liu et al.[9] involving 3973 Australian university students revealed significant predictors of their psychological wellbeing, highlighting negative factors such as ethnicity, stress, dietary changes, and social isolation, while emphasizing the positive impact of physical health, emotional support, and resilience. To effectively enhance student mental health, universities should focus on the targeted internet and tele-based interventions, particularly addressing perceived social isolation, as suggested by the findings.

Moreover, another recent study by O’Connor et al.[10] conducted a longitudinal analysis of mental health and well-being of adults in the UK during the first 6 weeks of lockdown during COVID-19 pandemic. By employing a combination of electronic health records and self-reported surveys, the study revealed distinct trajectories of mental health symptoms over time, emphasizing the need for personalized intervention strategies.

Furthermore, an investigation by Fancourt et al.[11] delved into the trajectories of depressive and anxiety symptoms before and after SARS-CoV-2 infection, comparing Long COVID and Short COVID groups. Analyzing the COVID-19 social study longitudinal
data from the University College London (UCL), the study finds that both groups experience immediate increases in symptoms, with the Long COVID group showing greater and sustained depressive symptoms over 22 months. While initial anxiety increases are similar, only the Short COVID group improves over time, emphasizing the need for integrated mental health support alongside physical treatment for Long COVID. The findings underscored the importance of considering sociodemographic factors when predicting and addressing MHDs in the context of Long COVID.

2.0.3 Machine Learning, Deep Learning, and Artificial Intelligence in MHD Prediction

The application of ML, DL, and AI techniques has substantially accelerated the progress in predicting MHD. Cacheda et al.[12], address the critical need for early detection of Major Depressive Disorder (MDD) using social media data. Leveraging ML techniques, specifically a dual model combining two RF classifiers, the study demonstrates improved performance in early detection compared to singleton models. By utilizing textual, semantic, and writing similarity features, this research contributes to advancing the field of early depression detection on social networks.

ML-based methodologies have harnessed various data sources for predicting MHDs. Alam, et al. [13] utilized imaging with genetic data to predict schizophrenia onset, while Birjali et al. [14] employed social media data and sentiment analysis to predict suicide risk. DL algorithms have proven effective in capturing intricate patterns for MHD prediction. Zhang et al. [15] leveraged DL models on functional magnetic resonance imaging (fMRI) data to predict severe mental illness, showcasing the power of Multiple Instance Learning (MIL) in extracting meaningful features.

Temporal and longitudinal analysis has gained prominence in MHD prediction, recognizing the dynamic nature of these disorders. Barak, et al.[16] proposed a longitudinal approach utilizing EHRs to predict patients’ future risk of suicidal behavior, capitalizing on temporal trends. Additionally, the fusion of multi-modal data sources has been explored to improve prediction reliability. Rahaman et al. [17] introduced a method of neuroimaging and genomic data for predicting risk of mental illness, providing a holistic view of MHDs.
2.0.4 Innovative Approach and Contribution

In contrast to prior methodologies, the present study adopts an innovative approach by combining a diverse array of datasets including social history, lifestyle, demographics, problem lists, immunization records, medication orders, vitals, and diagnostic results for MHD prediction. This comprehensive dataset integration offers a holistic understanding of an individual’s health profile. Particularly distinctive is the utilization of a multitude of ML algorithms for both feature selection and prediction, showcasing a holistic and integrated approach that enhances predictive accuracy.
CHAPTER 3
RESEARCH METHODOLOGY

This section elucidates the primary research methodology, as illustrated in Figure 3.1. The initial step involves a selection of patients diagnosed with Long COVID, alongside their corresponding medical records. This selection adheres rigorously to stringent criteria, with the patient’s medical histories and pertinent lab measurements merged for comprehensive investigation. The identification of Long COVID cases adheres to well-defined criteria aligned with prevailing medical standards. Furthermore, patients diagnosed with MHD after the diagnosis of Long COVID cohort are identified using the International Classification of Diseases (ICD-10) code standards.

Figure 3.1: The methodological framework for predicting Mental Health Disorders after the diagnosis of Long COVID

Table 3.1 presents a list of ICD codes that are associated with different MHD and Long COVID conditions from the EHR. All major categories and subcategories of ICD codes are considered in this study. The subsequent refinement of the data involves meticulous procedures for cleaning. Imputation has been employed to address missing values, while outliers are systematically removed to enhance the integrity of the dataset. Subsequently, an extensive correlation analysis has been conducted to unveil potential interrelationships
among variables. Following this, the sophisticated RFE technique is invoked, systemati-
cally discarding less impactful features to enhance the predictive capacity of the model.

To reinforce analytical robustness, data transformation techniques have been deployed. Notably, the fundamental process of normalization is employed, aligning variable ranges to enable an unbiased comparison and evaluation of their individual impacts. Given the inherent imbalance in the dataset, pivotal measures have been implemented to rectify this asymmetry. Leveraging the SMOTEEN technique, equilibrium is introduced to the data, ensuring a proportional representation of minority classes. Finally, we fine-tune the classification models and assess the predictive power of the chosen variables.

<table>
<thead>
<tr>
<th>ICD code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>F32</td>
<td>Major depressive disorder, single episode</td>
</tr>
<tr>
<td>F33</td>
<td>Major depressive disorder, recurrent</td>
</tr>
<tr>
<td>F41</td>
<td>Other anxiety disorders</td>
</tr>
<tr>
<td>F43</td>
<td>Reaction to severe stress and adjustment disorder</td>
</tr>
<tr>
<td>G47</td>
<td>Sleep disorders</td>
</tr>
<tr>
<td>Z63</td>
<td>Other problems family circumstances</td>
</tr>
<tr>
<td>F34</td>
<td>Persistent mood disorders</td>
</tr>
<tr>
<td>U09</td>
<td>Long COVID</td>
</tr>
</tbody>
</table>

### 3.0.1 Data Source

The investigation uses de-identified hospital-level data sourced from Froedtert Hospital (FH) in Milwaukee, WI, USA, spanning three years from June 2021 to January 2023. This study was conducted with the appropriate Institutional Review Board (IRB) approvals. The implementation of de-identification protocols guarantees the anonymization of the patient information, thus facilitating a comprehensive analytical exploration of the dataset. This methodology preserves the integrity of individual privacy, making it possible to conduct a rigorous investigation into various health-related parameters and outcomes within the stipulated time frame.
3.0.2 Data Preparation and Transformation

The dataset obtained from Froedtert Hospital encompasses a comprehensive range of data categories, including diagnostic results, medication, diagnoses, encounters, demographics, problem lists, procedures, and social history/lifestyle information. The foundational hospitalization details are captured in the diagnoses and encounters tables from the datasets, providing information such as admission and discharge dates, departments, diagnosis codes, etc. Additional details, such as diagnostic outcomes, vital measurements, prescribed medications, and other relevant diagnostic entries, are recorded in diagnostic results, medication orders, and problem lists. The social history/lifestyle dataset contains important information such as substance use, dietary habits, occupation and work environment, living environment, marital status and social support, education and socioeconomic status.

3.0.3 Patient Categorization: Mental Health Diagnoses After Long COVID

Within this dataset, ICD-10 codes encode a diverse array of diagnoses, with a subset of interest comprising instances related to Long COVID, which are marked by codes starting with "U09". At the beginning of our data preparation, we sorted through the diagnosis dataset to find unique patient IDs while removing any duplicate encounter records. The diagnosis dataset originally contained 17 million patient encounters, which boiled down to 67,881, unique patient records. The list of encounters from EHR are shown in Table 3.2. After that, we employed a filtering process to identify patients affected by Long COVID

<table>
<thead>
<tr>
<th>Encounter Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AV</td>
<td>Ambulatory visit</td>
</tr>
<tr>
<td>ED</td>
<td>Emergency visit</td>
</tr>
<tr>
<td>EI</td>
<td>ED to inpatient</td>
</tr>
<tr>
<td>IP</td>
<td>Inpatient stay</td>
</tr>
<tr>
<td>IS</td>
<td>Non-acute Institutional stay</td>
</tr>
<tr>
<td>NI</td>
<td>No information</td>
</tr>
<tr>
<td>OA</td>
<td>Other Ambulatory</td>
</tr>
<tr>
<td>OS</td>
<td>Observation stay</td>
</tr>
<tr>
<td>TH</td>
<td>Telehealth</td>
</tr>
</tbody>
</table>
conditions using the ICD code "U09". We then divided these patients into two groups: those without any MHD and those who were diagnosed with MHD after Long COVID was identified. The identification of patients with MHD was based on the ICD code list provided in Table 3.1. This careful approach was aimed at accurately identifying patients who were facing mental health problems due to Long COVID. After combining data from patients with and without MHD, we integrated additional datasets from EHR. We achieved this integration by matching the patient and encounter identifiers shared between different datasets. The final dataset size was 998 rows and 63 columns. A visual representation of the sample selection process, depicted in Figure 3.2, provides a clear overview of the methodology adopted in this study.

In the final dataset, the presence of mental health disorders was encoded as binary values, resulting in two distinct groups. It is important to note that our primary objective revolves around predicting the occurrence of MHD, constituting a classification task. The discretization of the target attribute offers several benefits, including facilitating the streamlined model training and potentially enhancing model performance.

### 3.0.4 Data Characteristics

Table 3.3 illustrates a significant class imbalance within the study population. The majority of patients are non-Hispanic (94.29%) and White or Caucasian (76.15%), respectively. This demographic alignment with Wisconsin’s racial and ethnic composition is noteworthy limitation of study. The dominant employment statuses are retired, accounting for 32.87%, and full-time positions, making up 35.57%. Approximately 44% of the total population holds Private insurance as their primary payer. Once again, the population is predominantly Non-Hispanic, with White individuals constituting a substantial portion. While White patients make up less than two-thirds of the undiagnosed population, they represent 75% of the MHD diagnosed population. Notably, a larger percentage of diagnosed patients belong to the retired and employed categories in both the groups. The age distribution reveals that the entire population has a median age of 57 years, with a standard deviation of 17.44. Interestingly, the median age remains consistent across both non-MHD (58 years) and MHD (57
years) patients, suggesting relative stability in age demographics between two categories. When examining gender distribution, females comprise a substantial majority, accounting for 63.53% of the total population. In the non-MHD group, 63.75% are female, while in the MHD group, 63.22% are female. Males constitute 36.47% of the population, with 36.43% from the non-MHD group and 36.54% from MHD individuals. The immunization status underscores the significance of immunization awareness. Approximately 40.08% of the population has been immunized. This includes 36.25% of individuals from the non-MHD group and 45.43% of individuals from the MHD group. However, the immunization status is unknown for 59.92% of the population. Notably, a higher proportion of 66.49% from the non-MHD group and 50.72% from MHD individuals lack this information.
The distribution of marital statuses provides valuable insights into the relationship status of the population. For instance, 50.80% of the population is classified as "Married," with 53.26% from the non-MHD group and 47.36% from MHD individuals. Significantly, 28.46% are categorized as "Single," consisting of 25.09% from the non-MHD group and 33.17% from MHD individuals. Additionally, the proportions of "Divorced," "Widowed," and "Significant Other" categories exhibit distinct disparities between the non-MHD and MHD groups.

Table 3.3 provides a comprehensive breakdown of various healthcare encounters. Notably, "Admission Visits" (AV) account for 48.40% of the total encounters, with 49.14% from the non-MHD group and 47.36% from MHD individuals. Outpatient Appointments (OA) represent 22.65% of the encounters; however, there is a notable discrepancy in the distribution, with 32.99% from the non-MHD group and 8.17% from MHD individuals.

3.0.5 Correlation Analysis

Correlation analysis is a crucial technique for uncovering valuable insights in datasets, revealing intricate relationships that shape the data’s structure. These insights are essential for understanding the relevance of attributes in predicting the target class, a critical aspect in various analytical pursuits. By combining correlation analysis with the machine learning, we can effectively identify related attributes in the dataset. These attributes are integral in accurately classifying an individual’s mental health status, thereby enhancing our ability to provide precise and timely assessments. In Table 3.4, we present the top 20 correlated pairs within the dataset, along with their correlation coefficients in relation to the target variable. This table visually captures the complex interplay between variables, highlighting how specific features may influence the MHD outcome. It’s imperative to acknowledge that while several features exhibit robust correlations with MHD, there are also notable instances of features bearing negligible correlations. As an illustration, 'Employment Status Student' manifests a correlation coefficient of 0.0381, signaling a tenuous link with MHD. Similarly, 'Insurance Payer Medicare' demonstrates a correlation coefficient of 0.0406, implying a marginal impact on mental health outcomes. These features, characterized by their modest correlation coefficients, might offer limited contributions to the predictive model-
Table 3.3: Statistics for final MHD study population

<table>
<thead>
<tr>
<th>Feature</th>
<th>Total Population (n %)</th>
<th>Non-MHD, n (%)</th>
<th>MHD, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>998 (100.0)</td>
<td>582 (58.32)</td>
<td>416 (41.68)</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (SD)</td>
<td>57 (17.44)</td>
<td>58 (17.22)</td>
<td>57 (17.44)</td>
</tr>
<tr>
<td>Encounters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AV</td>
<td>483 (48.40)</td>
<td>286 (49.14)</td>
<td>197 (47.36)</td>
</tr>
<tr>
<td>ED</td>
<td>59 (5.91)</td>
<td>42 (7.22)</td>
<td>17 (4.09)</td>
</tr>
<tr>
<td>EI</td>
<td>86 (8.62)</td>
<td>35 (6.01)</td>
<td>51 (12.26)</td>
</tr>
<tr>
<td>IP</td>
<td>43 (4.31)</td>
<td>25 (4.30)</td>
<td>18 (4.33)</td>
</tr>
<tr>
<td>OA</td>
<td>226 (22.65)</td>
<td>192 (32.99)</td>
<td>34 (8.17)</td>
</tr>
<tr>
<td>OS</td>
<td>13 (1.30)</td>
<td>4 (0.69)</td>
<td>9 (2.16)</td>
</tr>
<tr>
<td>TH</td>
<td>95 (9.52)</td>
<td>66 (11.34)</td>
<td>29 (6.97)</td>
</tr>
<tr>
<td>Insurance Type</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>17 (1.70)</td>
<td>5 (0.86)</td>
<td>12 (2.88)</td>
</tr>
<tr>
<td>Government</td>
<td>10 (1.00)</td>
<td>2 (0.34)</td>
<td>8 (1.92)</td>
</tr>
<tr>
<td>Medicate</td>
<td>120 (12.02)</td>
<td>55 (9.45)</td>
<td>65 (15.63)</td>
</tr>
<tr>
<td>Medicare</td>
<td>269 (26.95)</td>
<td>148 (25.43)</td>
<td>121 (29.09)</td>
</tr>
<tr>
<td>Private</td>
<td>439 (43.99)</td>
<td>256 (43.99)</td>
<td>183 (43.99)</td>
</tr>
<tr>
<td>Unknown</td>
<td>172 (17.23)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Employment Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disabled</td>
<td>81 (8.12)</td>
<td>31 (5.33)</td>
<td>50 (12.02)</td>
</tr>
<tr>
<td>Full Time</td>
<td>355 (35.57)</td>
<td>218 (37.46)</td>
<td>137 (32.93)</td>
</tr>
<tr>
<td>Not Employed</td>
<td>128 (12.83)</td>
<td>66 (11.34)</td>
<td>62 (14.90)</td>
</tr>
<tr>
<td>Part Time</td>
<td>75 (7.52)</td>
<td>41 (7.04)</td>
<td>34 (8.17)</td>
</tr>
<tr>
<td>Retired</td>
<td>328 (32.87)</td>
<td>183 (31.44)</td>
<td>145 (34.86)</td>
</tr>
<tr>
<td>Self Employed</td>
<td>50 (5.01)</td>
<td>31 (5.33)</td>
<td>19 (4.57)</td>
</tr>
<tr>
<td>Student</td>
<td>18 (1.80)</td>
<td>8 (1.37)</td>
<td>10 (2.40)</td>
</tr>
<tr>
<td>Marital Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Divorced</td>
<td>119 (11.92)</td>
<td>58 (9.97)</td>
<td>61 (14.66)</td>
</tr>
<tr>
<td>Married</td>
<td>507 (50.80)</td>
<td>310 (53.26)</td>
<td>197 (47.36)</td>
</tr>
<tr>
<td>Significant Other</td>
<td>25 (2.51)</td>
<td>12 (2.06)</td>
<td>13 (3.13)</td>
</tr>
<tr>
<td>Single</td>
<td>284 (28.46)</td>
<td>146 (25.09)</td>
<td>138 (33.17)</td>
</tr>
<tr>
<td>Widowed</td>
<td>90 (9.02)</td>
<td>55 (9.45)</td>
<td>35 (8.41)</td>
</tr>
<tr>
<td>Unknown</td>
<td>1 (0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>634 (63.53)</td>
<td>371 (63.75)</td>
<td>263 (63.22)</td>
</tr>
<tr>
<td>Male</td>
<td>364 (36.47)</td>
<td>212 (36.43)</td>
<td>152 (36.54)</td>
</tr>
<tr>
<td>Race</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White or Caucasian</td>
<td>760 (76.15)</td>
<td>447 (76.80)</td>
<td>313 (75.24)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>182 (18.24)</td>
<td>95 (16.32)</td>
<td>87 (20.91)</td>
</tr>
<tr>
<td>Multiracial</td>
<td>46 (4.61)</td>
<td>23 (3.95)</td>
<td>23 (5.53)</td>
</tr>
<tr>
<td>Asian</td>
<td>22 (2.20)</td>
<td>11 (1.89)</td>
<td>11 (2.64)</td>
</tr>
<tr>
<td>American Indian or Alaska Native</td>
<td>9 (0.90)</td>
<td>2 (0.34)</td>
<td>7 (1.68)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non Hispanic</td>
<td>941 (94.29)</td>
<td>533 (91.58)</td>
<td>408 (98.08)</td>
</tr>
<tr>
<td>Hispanic</td>
<td>57 (5.71)</td>
<td>30 (5.15)</td>
<td>27 (6.49)</td>
</tr>
<tr>
<td>NI</td>
<td>1 (0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Immunisation Status</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Given</td>
<td>400 (40.08)</td>
<td>211 (36.25)</td>
<td>189 (45.43)</td>
</tr>
<tr>
<td>Unknown</td>
<td>598 (59.92)</td>
<td>387 (66.49)</td>
<td>211 (50.72)</td>
</tr>
</tbody>
</table>
ing of an individual’s mental health status, and could wield diminished influence within our analytical models.

Table 3.4: Features correlation with MHD

<table>
<thead>
<tr>
<th>Feature</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>MHD</td>
<td>1.0000</td>
</tr>
<tr>
<td>Encounter Type TH</td>
<td>0.1709</td>
</tr>
<tr>
<td>Encounter Type EI</td>
<td>0.1557</td>
</tr>
<tr>
<td>employment status Disabled</td>
<td>0.1208</td>
</tr>
<tr>
<td>Encounter Type IP</td>
<td>0.1094</td>
</tr>
<tr>
<td>Encounter Type OA</td>
<td>0.1073</td>
</tr>
<tr>
<td>Encounter Type OS</td>
<td>0.1000</td>
</tr>
<tr>
<td>Insurance Payer Medicaid</td>
<td>0.0936</td>
</tr>
<tr>
<td>Marital Status Single</td>
<td>0.0883</td>
</tr>
<tr>
<td>Encounter Type ED</td>
<td>0.0859</td>
</tr>
<tr>
<td>Insurance Payer Govt</td>
<td>0.0781</td>
</tr>
<tr>
<td>Insurance Payer Commercial</td>
<td>0.0771</td>
</tr>
<tr>
<td>Marital Status Divorced</td>
<td>0.0714</td>
</tr>
<tr>
<td>Race American Indian</td>
<td>0.0698</td>
</tr>
<tr>
<td>Marital Status Married</td>
<td>0.0582</td>
</tr>
<tr>
<td>Race Black</td>
<td>0.0562</td>
</tr>
<tr>
<td>Employment status Unemployed</td>
<td>0.0525</td>
</tr>
<tr>
<td>Employment Status Full time</td>
<td>0.0465</td>
</tr>
<tr>
<td>Insurance Payer Medicare</td>
<td>0.0406</td>
</tr>
<tr>
<td>Employment Status Student</td>
<td>0.0381</td>
</tr>
</tbody>
</table>

3.0.6 Feature selection

In predictive model construction, feature selection is pivotal due to complexities stemming from a high number of features. Such complexities increase computation demands and affect overall model performance. Our approach began with calculating correlation scores for feature pairs, discarding those with correlations over 0.7 to address multicollinearity. We integrated RF classifier in RFE technique, which systematically identified significant features based on their assigned weights [18]. The outcome, illustrated in Figure 3.3, highlighted 15 essential features selected by RFE. The curve’s maximum score aligns with the optimal features according to RFE. This peak score of 0.96 with 15 features suggests that
additional features might not significantly enhance model performance, implying RFE effectively captured vital information.

![Figure 3.3: REF score Vs Number of selected features](image)

3.0.7 Normalization and Data Balancing

As illustrated in Figure 3.2, the original dataset consisted of $N = 998$ instances. Among these instances, 582 represented patients without pre-existing mental health disorders, while 416 instances corresponded to patients who developed mental health disorders after their Long Covid diagnosis. Due to the uneven distribution of data, a deliberate strategy was employed. The SMOTEEN technique was applied to address this imbalance. By leveraging this technique, the dataset underwent a transformation that resulted in a more balanced distribution. Specifically, the technique yielded an equal representation of 353 instances in each subset. This meticulous rebalancing process was conducted prior to the involvement of ML algorithms. The incorporation of the SMOTEEN technique seamlessly integrated equitable representation into the model training process. This integration not only enhanced
prediction accuracy, but also ensured fairness in the research outcomes, as previously cited [19].
4.1 Machine Learning Algorithms

Following the steps of normalization and data balancing, we proceeded to utilize a diverse array of ML models, including logistic regression (LR), random forest (RF), support vector machine (SVM), AdaBoost, XGBoost, k-nearest neighbors (KNN), Gaussian Naive Bayes (NB), and Multi-Layer Perceptron. Each of these models offers distinct strengths suited for various predictive tasks. For instance, RF employs a bagging technique to consolidate multiple decision trees into a robust and accurate model [20]. On the other hand, SVM excels in defining decision boundaries in multi-dimensional spaces to classify data points, especially within datasets containing varying features [21]. Boosting algorithms like AdaBoost and XGBoost sequentially combine weak learners to create a powerful model with significantly improved predictive capabilities [22]. XGBoost, in particular, stands out for its exceptional speed and performance among boosting algorithms [23]. To comprehensively evaluate model performance, we meticulously divided the dataset into training and testing subsets, allocating 70% of instances for training and the remaining 30% for testing. This approach empowers models to learn from a substantial portion of the dataset while also providing a separate testing subset for assessing their generalization prowess and accuracy on previously unseen data. The flow of the machine learning algorithms for prediction as shown in Figure 4.1.

4.1.1 Random Forest Algorithm (RF)

Random Forests, also known as random decision forests, are an ensemble learning method for classification, regression, and other tasks. This method operates by constructing a mul-
Figure 4.1: Machine-learning algorithms for prediction

titude of decision trees during training time and outputting the class that is the mode of
classes for classification tasks or mean prediction for regression tasks. The fundamental
principle behind this algorithm is the wisdom of the crowd, where the collective decision
from a group of individuals is preferred over a single individual’s decision. Figure 4.2
shows the RF algorithm.

4.1.2 Decision Tree (DT)

A decision tree is a flowchart-like structure in which each internal node represents a feature
(or attribute), each branch represents a decision rule, and each leaf node represents an
outcome. The topmost node in a decision tree is known as the root node. It learns to
partition based on the attribute value. It partitions recursively in such a manner called
recursive partitioning. Figure 4.3 shows the DT algorithm.
Figure 4.2: Random Forest algorithm (Image credit: https://medium.com)

Figure 4.3: Decision Tree algorithm for classification (Image credit: https://www.shiksha.com)
4.1.3 Naive Bayes Algorithm (NB)

NB classifiers are a family of probabilistic classifiers based on applying Bayes’ theorem with strong (naive) independence assumptions between the features. Despite their naive design and oversimplified assumptions, naive Bayes classifiers often work much better in many complex real-world situations than one might expect. Figure 4.4 shows the NB algorithm.

![Naive Bayes algorithm](https://medium.com)

Figure 4.4: Naïve Bayes algorithm (Image credit: https://medium.com)

4.1.4 Gradient Boosting Classifier (GB)

GB classifiers are a group of ML algorithms that combine many weak learning models to create a strong predictive model. Decision trees are usually used when doing gradient boosting. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Figure 4.5 shows the GB algorithm.
4.1.5 Support Vector Machine (SVM)

SVMs are supervised learning models with associated learning algorithms that analyze data for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. SVMs can also be used for regression tasks. They are known for their robustness in high-dimensional spaces and their effectiveness when the number of dimensions is greater than the number of samples. Figure 4.6 shows the SVM algorithm.

4.1.6 K Nearest Neighbour (KNN)

KNN is a method of learning that falls under the category of instance-based learning, or what is sometimes referred to as lazy learning. In this approach, computations are not performed until the function is evaluated. This method is applicable to both classification and regression tasks. A common strategy in KNN is to assign different weights to the neighbors’ contributions. This means that neighbors that are closer to the query point have a greater influence on the outcome than those further away. This technique enhances the precision of the local approximation of the function. Figure 4.7 shows the KNN algorithm.
4.1.7 XGBoost

XGBoost is a highly efficient distributed gradient boosting library that has been optimized for maximum performance, flexibility, and portability. It employs ML algorithms within the context of the Gradient Boosting framework. XGBoost offers a solution to numerous data science challenges by providing a parallel tree boosting technique, also known as GBDT or GBM, which delivers fast and precise results. Figure 4.8 shows the XGBoost algorithm.

4.1.8 Multi-Layer Perceptron (MLP)

A MLP is a form of artificial neural network that is specifically designed for supervised learning tasks. It is composed of several layers of nodes, also known as neurons, and utilizes a method called backpropagation for its training process. MLPs have the ability to learn intricate patterns, making them a popular choice for a wide range of machine learning tasks.
Figure 4.7: K Nearest Neighbour (Image credit: www.geeksforgeeks.org)

Figure 4.8: A general architecture of XGBoost (Image credit: www.researchgate.net)
applications. Figure 4.9 shows the MLP algorithm.

Figure 4.9: Multi-Layer Perceptron (Image credit: towardsdatascience.com)
5.1 Domain-Specific Evaluation Metrics for Mental Health Disorder Prediction

The evaluation metrics play a pivotal role in assessing the performance of ML models designed for predicting the risk of MHD within the healthcare domain. Each metric provides unique insights into the model’s ability to make accurate predictions and is tailored to address specific challenges associated with MHD prediction.

5.1.1 Accuracy (ACC)

In the context of mental health prediction, accuracy holds significance as it quantifies the overall correctness of the model’s predictions. Achieving high accuracy is crucial for healthcare professionals to trust the model’s ability to correctly identify individuals at risk of MHD. A high accuracy rate ensures that both positive and negative predictions are reliable, fostering confidence in the model’s utility in real-world healthcare applications.

\[
\text{Accuracy (ACC)} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Instances}}
\]

5.1.2 Precision

Precision gains prominence in healthcare applications, especially when predicting the risk of MHDs. In this context, precision represents the model’s accuracy in identifying individuals at risk among those predicted as positive cases. Minimizing false positives is paramount in mental health prediction, as misidentifying a healthy individual as at risk could lead to unnecessary interventions. Precision ensures that positive predictions are reliable and relevant, supporting healthcare practitioners in targeted interventions.

\[
\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}
\]
5.1.3 Recall (Sensitivity)

Recall assumes a critical role in mental health prediction, particularly in identifying individuals at risk among the actual positive cases. Focusing on recall is essential to prevent false negatives, where the model fails to identify individuals with actual mental health disorders. In healthcare, prioritizing recall ensures that individuals who truly need attention and intervention are not overlooked, facilitating early diagnosis and timely support.

\[
Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}
\]

5.1.4 F1 Score

The F1 Score, a measure that balances precision and recall, is particularly valuable in healthcare applications where the equilibrium between false positives and false negatives is vital. This metric provides a single value that encapsulates the trade-off between identifying as many true cases as possible (recall) and minimizing the risk of false alarms (precision), which is a critical consideration in medical diagnostics and treatment planning. Achieving a high F1 Score is imperative in mental health prediction to ensure a holistic and accurate assessment, allowing healthcare practitioners to make informed decisions based on a well-balanced model.

\[
\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

5.1.5 Area Under the ROC Curve (AUC)

In mental health prediction, the AUC metric is of particular importance as it evaluates the model’s ability to distinguish between individuals at risk and those not at risk. A higher AUC suggests superior discrimination capabilities, crucial in identifying subtle patterns indicative of mental health disorders. A well-discriminating model aids healthcare professionals in confidently interpreting the predictions, facilitating early intervention and personalized care for individuals at risk.

In summary, each evaluation metric serves a unique purpose in the healthcare domain,
contributing to the overall effectiveness of ML models in predicting the risk of MHDs. The combination of these metrics ensures a comprehensive assessment, supporting healthcare practitioners in making informed and reliable decisions for patient well-being.
CHAPTER 6
RESULTS DISCUSSION

6.1 Analysis Results

The performance of various classification methods was evaluated to discern their effectiveness in diagnosing MHDs. The Heatmap in Figure 6.1 presents a comprehensive overview of the evaluation metrics for each classifier. The RF classifier displayed significant performance, achieving an accuracy of 0.9150. Furthermore, it exhibited notable recall (0.8888) and precision (0.9090) values, indicating its proficiency in correctly identifying true positive cases while minimizing false positives. The associated F1-score (0.8988), a balanced measure of precision and recall, underscores the classifier’s overall capability in outstanding a harmonious equilibrium between these vital metrics. The area under the ROC curve (AUC) for RF (0.9686) further affirms its robustness in distinguishing between positive and negative samples. Figure 6.2 illustrates the AUC curves for all the classifiers. The mean ROC AUC is 0.88. XGBoost classifier also demonstrated a commendable performance with an accuracy of 0.8301. Its recall (0.8222) and precision (0.7872) values, coupled with a competitive F1-score (0.8043), suggest its effectiveness in classifying positive samples while maintaining a balance between precision and recall. As shown in Figure 6.2 the AUC value for XGBoost (0.8987) underscores its aptitude in differentiating between classes. Other classifiers exhibited varying levels of performance in terms of accuracy, recall, precision, F1-score, and AUC. Notably, Gaussian NB demonstrated higher precision (0.8571), but relatively lower recall (0.5333), resulting in a trade-off between false positives and false negatives. The evaluation outcomes showcased that RF and XGBoost classifiers outperformed others in terms of multiple metrics, demonstrating their potential for accurate MHD classification. These findings underscore the significance of utilizing ensemble methods in medical diagnostic tasks, as they combine the strengths of multiple models to yield enhanced predictive performance.

The evaluation of tree based classifier using decision tree plots are crucial for interpret-
Figure 6.1: Heatmap for the performance measure of all classification methods

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Tree</td>
<td>0.821</td>
<td>0.800</td>
<td>0.783</td>
<td>0.791</td>
<td>0.818</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.915</td>
<td>0.889</td>
<td>0.909</td>
<td>0.899</td>
<td>0.909</td>
</tr>
<tr>
<td>GradientBoosting</td>
<td>0.811</td>
<td>0.822</td>
<td>0.755</td>
<td>0.787</td>
<td>0.902</td>
</tr>
<tr>
<td>Support Vector</td>
<td>0.792</td>
<td>0.644</td>
<td>0.829</td>
<td>0.725</td>
<td>0.854</td>
</tr>
<tr>
<td>K Nearest Neighbors</td>
<td>0.783</td>
<td>0.800</td>
<td>0.720</td>
<td>0.758</td>
<td>0.866</td>
</tr>
<tr>
<td>Gaussian NB</td>
<td>0.764</td>
<td>0.533</td>
<td>0.857</td>
<td>0.657</td>
<td>0.770</td>
</tr>
<tr>
<td>Multilayer Perceptron</td>
<td>0.820</td>
<td>0.733</td>
<td>0.846</td>
<td>0.785</td>
<td>0.923</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.830</td>
<td>0.822</td>
<td>0.787</td>
<td>0.804</td>
<td>0.899</td>
</tr>
</tbody>
</table>

The decision tree plots offer a straightforward and intuitive depiction of how input features contribute to predictions, making them highly accessible for interpretation by both experts and non-experts alike. The decision tree rules for the DT classifier outline the predictive factors associated with MHD. As shown in Figure 6.3 the initial split is based on the variable encoded type TH, suggesting that certain encoded types are indicative of a higher likelihood of mental health disorders. Subsequent splits occur based on features such as payer type Medicare, bmi, employment status Disabled, marital status Single, race Black, age at visit in years, and sex (gender). These splits delineate distinct subgroups within the data, each characterized by different risk profiles for MHDs. For instance, individuals with specific combinations of these features, such as being of a certain race or having a particular employment status, may exhibit a higher or lower likelihood of MHDs. Overall, this decision tree provides valuable insights into the multifaceted nature of factors influencing mental health outcomes. Figure 6.4 shows the output represents decision tree models generated by
random forest classifiers for predicting MHDs based on demographic and socioeconomic factors. Each decision tree in the random forest consists of a series of conditions that partition the data into subsets with similar characteristics, leading to predictions for each subset. The conditions in the trees, such as age, race, employment status, and insurance type, serve as indicators of potential risk factors for mental health disorders. By combining predictions from multiple trees, random forest classifiers provide robust estimates while mitigating the risk of overfitting to the training data. In the context of MHD prediction, these models offer insights into which demographic and socioeconomic factors are most influential in determining an individual’s likelihood of experiencing such disorders.

In our analysis, we also employed the feature importance attribute method to identify the top 10 features within the scope of tree-based algorithms, such as the RF classifier (Figure 6.5). In contrast, for the XGBoost algorithm, a distinct approach is utilized to determine feature importance—utilizing the “gain” concept explicitly (Figure 6.6). This concept quantifies the contribution of each feature to enhancing the model’s predictive
Figure 6.3: Decision tree predicting the Mental Health Disorder with Decision Tree classifier (max depth=3)

Figure 6.4: Decision tree predicting the risk of Mental Health Disorder with Random Forest classifier (max depth=3)
performance. These values are normalized to ensure their sum equals 1, allowing for a relative comparison of feature importance. This attribute is computed during the model’s training process and relies on the frequency of each feature’s usage in node splitting within the constructed trees.

The resulting selection of the top 10 features for predicting MHDs encompasses various attributes, including encounter type (specifically TH, EI, or ED), employment status indicating “disabled,” “retired,” and “full-time,” body mass index (BMI), marital status “single,” insurance payer “Medicare,” racial category “Black,” and gender “sex.” This compilation of significant attributes underscores the identification of key variables contributing to the predictive capabilities of the models [24].

![Figure 6.5: Importance of RF classifiers (Feature importance attribute)](image)

Figure 6.5 exhibits the confusion matrix associated with the RF classifier, while Figure 6.7 portrays the confusion matrix attributed to the XGBoost classifier. Notably, the RF classifier achieved a higher number of accurate predictions for patients with MHD compared to the XGBoost classifier. This indicates that the RF classifier demonstrated a greater proficiency in identifying MHD patients with precision when contrasted with the XGBoost classifier.
Figure 6.6: Feature Importance of XGBoost classifiers (Gain Function)

enc_type = encounter type

Figure 6.7: Confusion Matrix RF classifiers (MHD = 1, No-MHD = 0)
Figure 6.8: Confusion Matrix XGBoost classifier (MHD = 1, No-MHD = 0)
CHAPTER 7
CONCLUSIONS AND FUTURE DIRECTIONS

7.1 Conclusions

In this study, advanced ML techniques were employed to address the intricate link between Long COVID and mental health. Diverse ML algorithms were rigorously assessed for predicting MHD following Long COVID exposure. The identified crucial predictors not only shed light on the post-Long COVID mental health trajectory but also equip healthcare practitioners with a potent predictive tool. This model can proactively identify individuals at MHD risk, enabling targeted interventions. These findings go beyond data analysis, offering hope to those dealing with Long COVID’s aftermath and showcasing technology’s transformative potential in healthcare paradigms.

7.2 Limitations

The dataset primarily encompasses patients diagnosed with Long COVID within a specific medical system, potentially leading to selection bias and limiting the generalizability of findings. Additionally, the study focuses on patients who have received medical care within this system, potentially excluding individuals with less severe symptoms or those seeking care elsewhere. While ML models have exhibited promising performance, they rely on retrospective data and might not capture the entire spectrum of factors influencing mental health outcomes. Moreover, Long COVID patients can vary in their demographics, clinical presentation, and comorbidities. This variability may pose a challenge in developing a predictive model that can accurately identify patients at risk of developing chronic diseases. These limitations underscore the need for cautious interpretation of findings and encourage further research involving diverse datasets and collaboration across healthcare institutions.
7.3 Future Work

To enhance the study’s robustness, prospective data collection could be considered like the National Institute of Health (NIH) All of Us or the National COVID Cohort Collaborative (N3C) dataset. This approach would allow for a more comprehensive understanding of the evolution of mental health symptoms over time in individuals affected by Long COVID. Moreover, incorporating additional data sources such as wearable devices and patient-reported outcomes could provide real-time insights into patients’ mental health status, enabling timely interventions. Partnering with mental health professionals can provide valuable insights and contribute to the refinement of models by integrating specialized knowledge from the field. This collaboration can enhance the effectiveness and accuracy of the models in addressing mental health issues.

In conclusion, this research embarks on a significant journey toward understanding the complex interplay between Long COVID and mental health using advanced ML methodologies. By acknowledging limitations and setting sights on future advancements, this study not only contributes to the evolving landscape of healthcare research, but also echoes a call for comprehensive, interdisciplinary approaches to address the multifaceted challenges posed by Long COVID on mental well-being.
REFERENCES


