University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

Assessing the Influence of Socioeconomic and Lifestyle Factors on Mental Health: A Data-Driven Approach

Sairam Prabakaran
2315589

Supervisor: **Dr Sema Gunturkun**

September 18, 2024
Colchester

Contents

Abstract

This dissertation uses a data-driven method to investigate the intricate link between lifestyle and socioeconomic factors in influencing mental health outcomes. Numerous factors, from specific habits to more general societal and economic circumstances, have an impact on mental health. To understand how factors like income, education, employment, housing, and access to healthcare, as well as lifestyle factors like diet, physical activity, and substance use, affect mental well-being, this study uses large-scale, cross-sectional datasets and statistical and machine learning techniques for analysis.

The results suggest preventive variables, such as healthy lifestyle choices, that can reduce these risks while also highlighting strong correlations between higher incidence of mental health issues and lower socioeconomic position. This study contributes to a better understanding of the mechanisms by which lifestyle decisions and socioeconomic pressures impact mental health by investigating the relationship between the two.

The findings provide guidance for focused public health initiatives, indicating that enhancing healthier lives and tackling social injustices are critical to enhancing mental health outcomes for both individuals and communities. By stressing the significance of comprehensive, evidence-based approaches in resolving mental health inequities, this dissertation adds to the expanding corpus of research on the socioeconomic determinants.

By establishing both direct and indirect links between these variables and mental health outcomes, sophisticated data analytic techniques provide detailed insights into risk and resilience factors. In order to mitigate the detrimental effects of socioeconomic disadvantage on mental health, the study also critically assesses the function of social support networks, community resources, and policy measures. The study, taken as a whole, adds to the body of knowledge in the field of psychology while also providing useful guidance for creating more successful mental health therapies that take individual and systemic factors into account.

CHAPTER 1

Introduction

1.1 Research Topic Overview

Millions of individuals worldwide suffer greatly from depression, a common and dangerous medical condition that significantly affects mental health. Depression has a substantial negative influence on people's day-to-day functioning and is characterized by persistent sadness, a decreased interest in enjoyable activities, and a variety of emotional and physical abnormalities [15]. The World Health Organization acknowledges depression as a major global cause of disability, which highlights the need for better understanding and care of this illness [16].

Contemporary research has progressively concentrated on the identification of the many factors contributing to depression, such as genetic predispositions, environmental stresses, and lifestyle choices [18]. Given the intricate relationships among these variables, it may be necessary to comprehend the underlying causes of depression in order to treat it effectively. These underlying causes include sociodemographic variables like age, marital status, and socioeconomic status, all of which have been shown to have a major influence on depression rates [6].

A multimodal strategy that incorporates social support networks, psychological interventions like cognitive-behavioral therapy, and pharmaceutical treatments like antidepressants is necessary to treat depression. Additionally, there is increased interest in complementary and alternative therapies that have shown promise in symptom relief and overall mental health enhancement, such as exercise and mindfulness-based interventions.

This research employs a dataset consisting of 413,768 persons with multifaceted backgrounds and health characteristics to investigate the correlation between different variables and symptoms of depression. This study seeks to analyze factors such as employment position, levels of physical activity, dietary habits, and previous health information in order to identify possible prediction indications and measures for mitigation.

Given the extensive effects depression has on people and society, additional study is necessary to find better methods for identifying, treating, and preventing depression. Future targeted and effective treatments might potentially be facilitated by the further advancement of personalized medicine, which makes use of genetic and neuroimaging approaches.

Statement of the Study Objectives

This study's main goal is to investigate and pinpoint important lifestyle and demographic factors that either raise or lower a person's chance of developing depression. In particular, the research aims to:

- Understand the prevalence of depression across different sociodemographic groups within the dataset.
- Investigate the association between lifestyle choices, such as smoking, alcohol intake, and physical exercise, and the occurrence of depression.
- Examine the impact of familial and personal medical history on depression, evaluating factors like a history of mental illness and chronic medical illnesses.
- Utilize statistical tools and machine learning techniques to model and forecast depression risk based on observable factors.

By achieving these goals, the research hopes to add to the body of knowledge about mental health and wellness and provide insights that could inform clinical treatments and policy decisions aimed at reducing the burden of depression. It is also envisaged that the results would spur future research into focused interventions that may be used to support those at risk in community and healthcare settings.

This dissertation will provide a thorough examination of the roles that many factors play in determining mental health through rigorous data analysis and interpretation, ultimately encouraging more efficient methods for managing and preventing depression.

1.2 Brief Introduction to the Dataset

This dissertation examines the multidimensional relationships between various lifestyle, health, and demographic factors and their effects on depression by analyzing a large dataset. The dataset is among the largest of its kind for this kind of research because it includes information from 413,768 people.

The broad range of data points obtained enables for an in-depth investigation of the prevalence and correlates of depression among varied populations.

The dataset comprises the following essential variables for each participant:

- **Name**: Anonymized identifiers to ensure privacy.
- **Age**: The age of the participants, ranging from young adults to the elderly, allowing for age-related analysis.
- **Marital Status**: Categories include single, married, divorced, and widowed, providing insights into how social relationships affect mental health.
- **Education Level**: Educational achievement from high school to advanced degrees,which may connect with economic opportunities and stress levels.
- **Number of Children**: Family size, which could influence everyday stress and life satisfaction.
- **Smoking and Alcohol Consumption Status**: Lifestyle characteristics that are typically linked to coping methods and general health.
- **Physical Activity Level**: Ranges from sedentary to active, a vital component in mental and physical health.
- **Employment Status and Income**: Economic issues that strongly affect mental well-being.
- **Dietary Habits, Sleep Patterns, and Chronic Medical Conditions**: These health-related elements provide a holistic assessment of each individual's lifestyle and health situation.

- **History of Mental Illness and Substance Abuse**: Direct markers of probable propensity to depression.
- **Family History of Depression**: Genetic and environmental factors that potentially influence depression risk.

Rigid procedures of collection and validation ensure data integrity, and all fields have non-null entries, indicating a high completion rate and dependability for analysis. Because of the dataset's structure, complex factor interactions can be examined through both univariate and multivariate investigations.

This large dataset not only supports a rigorous statistical analysis but also increases the knowledge of depression's multifaceted character. By using these data, the dis-dissertation attempts to find patterns and trends that could drive targeted interventions and policy decisions aimed at minimising the impact of depression on the population. Through such detailed data analysis, this work contributes to the greater conversationon mental health, underlining the significance of a multifaceted approach to understanding and treating depression.

CHAPTER 2

Literature Review

2.1 Review of Existing Studies on Depression and its Contributing Factors

Depression, a complex and multifaceted disorder, has been extensively studied across various disciplines, each contributing to a nuanced understanding of its etiology, prevalence, and impact. The World Health Organization (WHO) categorizes depression as a leading cause of disability worldwide, affecting millions and imposing significant societal costs. This literature review synthesizes key findings from recent studies, highlighting the breadth of factors associated with depression, ranging from biologicalto socio-economic.

2.1.1 Biological and Genetic Factors

Studies conducted in the fields of neuroscience and psychiatry have shown the important roles that biological processes and genetic predispositions play in the onset of depression. Caspi et al. (2003) in their seminal work utilized longitudinal data to link specific gene-environment interactions to depression onset [2]. They emphasized how serotonin transporter gene genetic variants modulated the impact of stressful life events on the likelihood of depression. Additionally, twin model studies suggest that genetic factors account for around 40% of the diversity in depression vulnerability[19]. These findings suggest that while genetics set a predisposition, environmental factors ultimately trigger depressive episodes.

2.1.2 Environmental and Social Factors

Beyond genetics, the environment significantly shapes mental health outcomes. Stressful life events, such as the loss of a loved one, unemployment, or traumatic experiences,are strongly correlated with the onset of major depressive episodes [7]. The social determinants of health, which include factors like socio-economic status, education level, and marital status, also play crucial roles. Lower socio-economic status has beenconsistently linked with higher depression rates, attributed to increased stress, lower social support, and reduced access to healthcare services [10].

The impact of interpersonal relationships and social support cannot be understated. A meta-analysis by Kawachi and Berkman (2001) found that individuals with stronger social ties had a significantly reduced risk of depression. This protective effect highlights the importance of social relationships in buffering against mental health issues.

2.1.3 Lifestyle Factors

The likelihood of getting depression is also influenced by lifestyle choices related to nutrition, exercise, and drug usage. Regular exercise has been demonstrated to lower depression risk by as much as 30%, according to research published in the American Journal of Preventive Medicine. Conversely, unhealthy dietary patterns, characterized by highintakes of processed foods and sugar, have been linked to increased depression risks inseveral observational studies.

Depression and substance misuse have an especially complicated link. It might be difficult to stop the vicious cycle that depression symptoms can cause and worsen when alcohol and drugs are used.

Additionally, smoking has been found to be a substantial risk factor for depression, with smokers more likely than non-smokers to experience depressed symptoms. The complex relationship between lifestyle choices and mental health is further complicated by the fact that long-term tobacco use is linked to higher rates of anxiety and depression, even while nicotine may offer short-term respite from depressed symptoms. Recovery from substance addiction is further complicated by the fact that withdrawal symptoms from alcohol, nicotine, or narcotics can precipitate or worse.

2.1.4 Impact of Chronic Health Conditions

Depression is usually found to be more common in those with chronic medical illnesses such diabetes, cardiovascular disease, and chronic pain. The reciprocal relationship between depression and chronic illness makes treatment and management more difficult because depression can make the physical illness's prognosis worse and vice versa [14]. This correlation underscores the need for integrated care approaches that address both mental and physical health in chronic disease management.

2.1.5 Current Gaps in Research

Even though a great deal of study has clarified many elements of depression, there are still gaps, especially when it comes to knowing how these numerous components interact. The majority of research has concentrated on individual components without taking into account the combined impact of several stressors or protective factors. In addition, additional longitudinal research is required to fully comprehend the causes and temporal evolution of depression.

With the use of an extensive dataset, this dissertation attempts to fill in some of these gaps by investigating the ways in which a variety of independent and associated factors influence depression risk. The ensuing sections will explore these associations in further detail and use sophisticated statistical analysis to extract insightful conclusions from the data.

- Broaden the Focus to Include Anxiety and Other Mental Health Disorders: The main focus is on depression, but we could also broaden it a little bit to cover other prevalent mental health issues, such as anxiety, which frequently coexists with depression. This would make it possible to look more broadly at the ways that lifestyle and demographic variables affect many aspects of mental health.
- Set Up a Focus on the Use of Digital and Social Media: Studies have indicated that lifestyle factors including social media use, screen time, and digital connectedness are becoming important for mental health, particularly for younger populations. As such, these issues should be investigated.
- Examine the impact of familial and personal medical history on depression, evaluating factors like a history of mental illness and chronic medical illnesses.
- Explore the Role of Sleep Patterns and Quality: Sleep hygiene and quality of sleep are increasingly being recognized as critical determinants of mental health.

By attaining these aims, the project seeks to contribute to the greater body of information regarding mental health and wellness, delivering insights that could guide policy-making and clinical treatments focused on lessening the burden of depression. Furthermore, it is hoped that the findings will motivate further study into targeted interventions that may be deployed in healthcare and community settings to support those at risk.

Through rigorous data analysis and interpretation, this dissertation will give a detailed investigation of the roles that diverse factors play in determining mental health, ultimately promoting more effective ways of managing and avoiding depression.

2.1.6 Significance of Lifestyle, Health Habits, and Family History in Depression

It is of great clinical and public health importance to examine how family history, lifestyle, and health practices interact in the genesis and progression of depression. This section examines the ways in which these traits affect not only the likelihood of developing depression but also its intensity and the efficacy of available treatments.

Lifestyle Factors

It has been shown that lifestyle choices like eating, sleeping patterns, physical activity, and substance use have a significant impact on mental health, especially depression. Frequent exercise is one of the healthiest lifestyle choices; numerous studies have shown that it can both delay the onset of depression and lessen its symptoms in those who already have it [17]. Exercise boosts the release of endorphins, sometimes labeled ’ feel-good’ hormones, which help decrease feelings of depression [13].

Dietary choices also play a key influence. The consumption of a balanced diet rich in fruits, vegetables, whole grains, and lean meats, generally referred to as the Mediterranean diet, has been connected with a lower incidence of depression [9]. Conversely, a diet heavy in processed carbs and saturated fats may exacerbate the body's inflammatory responses, which are linked to the emergence of depressive symptoms [11].

Sleep quality is another significant lifestyle element. Insufficient or interrupted sleep has been frequently linked to increased vulnerability to depression. The disruption of circadian rhythms, which control sleep, can dramatically impair mood and emotional regulation [21]. Managing sleep hygiene is thus a key component of both preventing and treating depression.

Health Habits

Alcohol and smoking are two lifestyle choices that have a complicated relationship to depression. Research has shown that smoking cessation can result in significant improvements in mental health and that nicotine dependence is linked to an increased risk of depression[20]. Similarly, while moderate alcohol use has sometimes been shown to haveprotective effects against depression, excessive alcohol use is a known risk factor and can exacerbate existing mental health concerns [1].

Furthermore, there is a strong correlation between an increased incidence of depression and chronic health conditions like diabetes, cardiovascular diseases, and obesity. Chronic disease management stress can exacerbate symptoms of depression, and sadness can make it harder for people to continue receiving adequate medical care for their physical health issues. This can lead to a vicious cycle of declining health outcomes[12].

Family History and Genetic Factors

One of the best indicators of depression is a family history of the illness, highlighting the significant influence of hereditary and familial factors. People who have a first-degree relative with a diagnosis of depression are more likely to get the illness themselves [22]. This familial association shows that genetic factors contribute considerably to the pathophysiology of depression, while the specific mechanisms are still being studied.

Additionally, parenting styles and early life stress are two aspects of the family environment that can have an impact on the onset of depression. Depression in adult life is significantly increased when adverse childhood experiences (ACEs) such as abuse, neglect, and dysfunctional families are experienced[3]. This relationship stresses the relevance of environmental influences and their interaction with genetic predispositions.

Integrated View and Implications for Treatment

Understanding the combined effect of lifestyle, health habits, and family history on depression is vital for establishing effective preventative and treatment measures. Interventions that promote healthy lives and enhance health habits have the potential to greatly lower the incidence and severity of depression. For instance, lifestyle modification programs that integrate nutrition, exercise, and sleep hygiene have been demonstratedto be useful in lowering depressed symptoms [5].

Furthermore, understanding the importance of hereditary and familial factors facilitates the development of customized treatment plans. People with a family history of depression, for instance, can benefit from early and more intensive interventions, such as medication and psychotherapy. Furthermore, counseling and support programs aimed at at-risk families can mitigate the effects of hereditary and familial predispositions.

In summary, a large body of research indicates a clear relationship between family history, health behaviors, and lifestyle decisions and the occurrence and treatment of depression. It is possible to control and even prevent depression by incorporating these discoveries into treatment protocols and public health policies, which will enhance social well-being in general.

CHAPTER 3

Methodology

3.1 Description of the Dataset and Variables

This dissertation employs a large-scale dataset that is exemplary for its depth and breadth in exploring the factors influencing depression. The dataset encompasses data from 413,768 individuals, providing a robust foundation for statistical analysis and insight generation. Each record in the dataset represents an individual participant, with 16 distinct variables capturing a range of demographic, behavioral, and health-related factors. The following subsections detail each variable, outlining their significance and the role they play in the broader context of this research on depression.

3.1.1 Demographic Variables

- **Name**: Each record contains a unique anonymized identifier, which ensures participant privacy while allowing for individual-level analysis.

- **Age**: This variable is essential because it makes it possible to examine depression at various phases of life. Previous research has shown that different age groups have distinct depression risks and symptoms [8].

- **Marital Status**: Categorized as single, married, divorced, or widowed, this variable helps examine the social support structures and their impact on mental health [4].

- **Education Level**: From high school to postgraduate degrees, education levels offer insights into socioeconomic status and resource accessibility, both of which are proven to have an impact on mental health outcomes [10].
- **Number of Children**: The number of dependents is included as a stressor or buffer in familial settings, influencing parental mental health.

3.1.2 Health and Lifestyle Variables

- **Smoking Status**, **Alcohol Consumption**: These factors are essential for evaluating lifestyle decisions that could either aggravate or lessen symptoms of depression. Research has indicated a noteworthy association between substance abuse and depression [1].
- **Physical Activity Level**: Sedentary and energetic physical activity are divided into two categories. Frequent exercise is a key component of lifestyle therapies and is linked to lower rates of depression.
- **Employment Status and Income**: These economic factors are vital for understanding the stress related to financial security and its impact on mental health.
- **Dietary Habits**, **Sleep Patterns**: Sleep and food patterns have an impact on physical health, which in turn has an impact on mental health. Increased rates of depression are associated with poor dietary choices and sleep difficulties.

3.1.3 Medical and Family History Variables

- **History of Mental Illness**, **History of Substance Abuse**, and **Family History of Depression**: These variables are particularly significant as they provide insights into both genetic predispositions and environmental influences on depression.
- **Chronic Medical Conditions**: Chronic diseases often co-occur with depression, making this variable crucial for understanding comprehensive health dynamics [14].

The extensive scope of this dataset allows for a multimodal examination of depression, taking into account a wide range of variables from individual lifestyle choices to more general socioeconomic issues. The statistical methods used in the ensuing analyses will reveal

Finding patterns and correlations among these characteristics can help find viable treatments and preventative steps that could lessen the impact of depression.

3.2 Explanation of the Analytical Methods and Tools Used

This dissertation's analytical methodology makes use of Python, a popular programming language in data science, and strong statistical techniques. This section explains the many statistical techniques and Python modules used to examine the dataset, demonstrating how each approach contributes to the discovery of new insights into depression.

3.2.1 Python and Its Libraries

Python is chosen for its wide ecosystem of libraries that are particularly well-suited for data processing, analysis, and visualization. Key libraries used in this research include:

- **Pandas**: Employed for data manipulation and cleaning, Pandas provides a flexible data structure that makes it easy to manage tabular data with heterogeneous columns.

- **NumPy**: This library is used for numerical operations. Its powerful array of objectsallow execute complex numerical analyses efficiently.

- **Matplotlib** and **Seaborn**: These libraries are used for data visualization, providing a variety of graphs and charts that aid the comprehension of data distributions and trends.

- **SciPy**: Useful for more scientific computations including statistical tests that help validate the findings.

- **Scikit-learn**: Regression models, classification methods, and clustering approaches are applied with this machine learning package to investigate patterns and forecast depression.

3.2.2 Statistical Analysis Techniques

The statistical studies try to find links between numerous factors and depression, quantify the strength of these relationships, and predict depression outcomes based on the observed patterns. The following techniques are crucial to our analysis:

- **Descriptive Statistics**: Initial exploratory data analysis involves computing means, medians, modes, and standard deviations to understand the central tendencies and variabilities of the dataset.
- **Inferential Statistics**: Techniques such as t-tests and ANOVA are used to examine the statistical significance of the differences seen across groups (e.g., between various age groups or marital statuses).
- **Correlation and Regression Analysis**: These techniques help analyze the connections between different factors. To determine the direction and intensity of the relationships, the Pearson and Spearman correlation coefficients are computed. Next, multiple regression analysis is applied to predict the impact of many factors on depression.
- **Principal Component Analysis (PCA)**: PCA is applied to minimize the dimensionality of the dataset while maintaining the most significant variables, which aids in displaying high-dimensional data and increasing model performance.
- **Cluster Analysis**: This method is used to partition the dataset into clusters based on similar criteria. It aids in identifying subgroups within the population that share common qualities connected to depression.

3.2.3 Implementation and Validation

Python scripts are used to load the data, preprocess it, run statistical tests on it, and provide visualizations as part of the analysis process. Each phase's output is put through a rigorous testing process to guarantee correctness and dependability. For predictive models, validation strategies include cross-validation; for inferential tests, adjustments for multiple comparisons are used to account for Type I errors.

This thorough approach to data analysis offers a methodological framework that may be applied to datasets of a similar nature in other areas of health research in addition to shedding light on the traits associated with depression.

CHAPTER 4

Data Analysis

Statistical Analysis

The data analysis aims to explore various factors contributing to depression by employing statistical techniques to examine relationships within a dataset encompassing 413,768 entries. These entries detail demographic information, health habits, and medical history potentially related to depression. This section presents the methodology and findings from the analyses conducted using Python's data science libraries, including Pandas, NumPy, Matplotlib, and Seaborn.

Descriptive Statistics

The initial exploration involved generating descriptive statistics to understand the dataset's distribution and characteristics. Key variables, such as age, income, and number of children, showed varied distributions, which necessitated further exploration through visualization techniques.

- The age of participants ranged from 18 to 98, with a mean of approximately 49 years, indicating a middle-aged demographic.
- Income levels varied significantly, reflecting the diverse socio-economic backgrounds of the dataset.
- The number of children also varied, with a mode of 1 child per participant.

Visualization of Data

To find trends and abnormalities in the data, visual analysis was done. To evaluate the age distribution among several categories of chronic medical disorders, box plots were employed. It was noted that participants who were older tended to report a higher number of chronic diseases, which is consistent with standard medical knowledge.

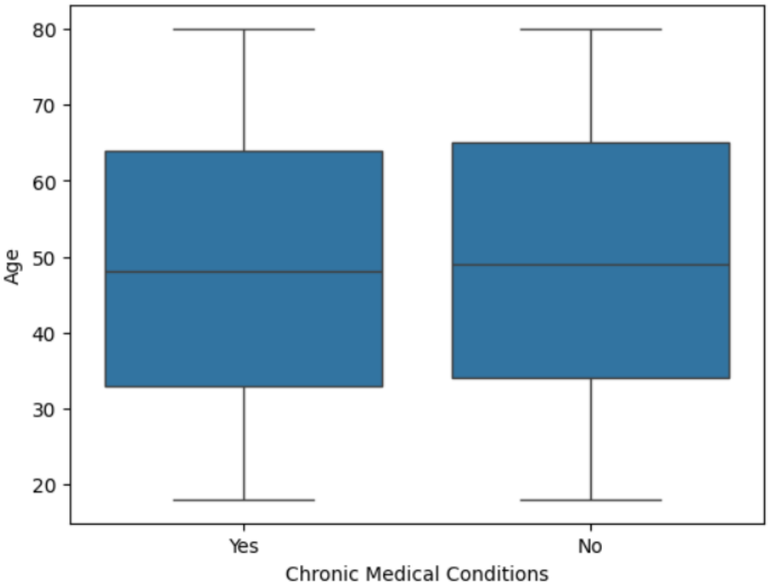

Figure 4.1: Box plot of age distribution by chronic medical conditions.

Correlation heatmaps were generated to explore potential correlations among numerical variables. The heatmap indicated a moderate correlation between age and chronic medical conditions, suggesting that as age increases, so does the likelihood of having chronic conditions.

Statistical Analysis

The core of the data analysis involved statistical testing to identify significant relationships. Chi-square tests were applied to categorical variables to determine dependencies between factors like marital status and depression.

- A significant chi-square value for marital status versus depression indicates that marital status could influence depression rates.

- T-tests on continuous variables like income showed differences in mean income

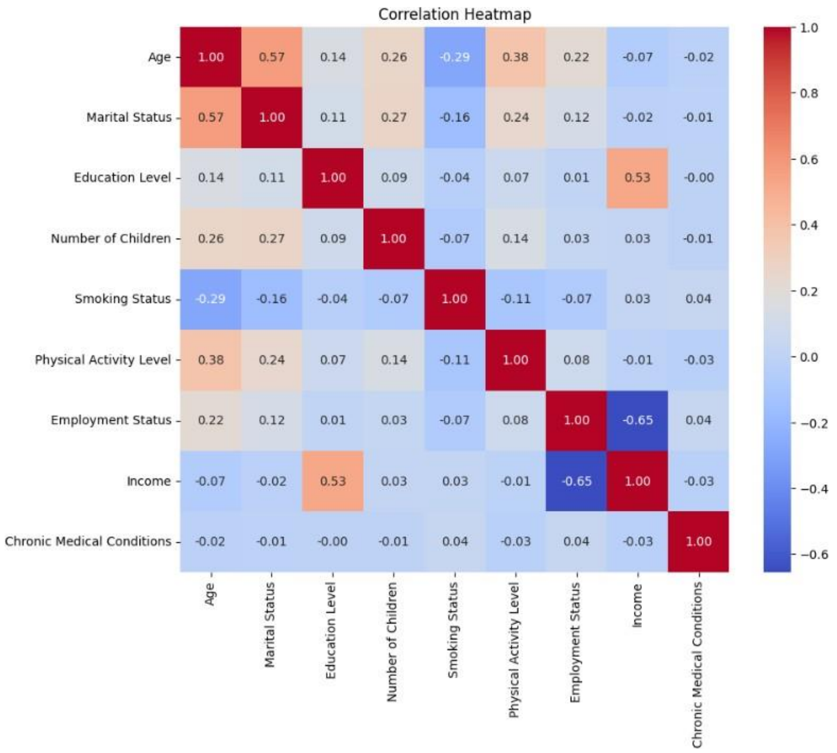

Figure 4.2: Correlation heatmap of numerical variables.

between depressed and non-depressed groups, suggesting economic factors' influence on mental health.

Regression Analysis

Logistic regression was utilized to measure the impact of different factors on depressive symptoms. When dealing with binary dependent variables, such the presence or absence of depression, this approach proved to be quite helpful. According to the model's coefficients, depression is significantly predicted by age, wealth, and the existence of long-term medical issues.

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{4.1}$$

where $P(Y = 1)$ represents the probability of depression, and $X_1, X_2, \ldots, X_n$ are the explanatory variables.

4.1 Examination of Trends, Patterns, and Correlations Among Different Variables

Examining trends, patterns, and correlations in a dataset with more than 400,000 entries offers a strong foundation for determining the complex relationships between health behaviors, demographic traits, and their possible influence on depression. To clarify these connections, this analysis makes use of a range of statistical methods and visualizations. It also identifies noteworthy trends and correlations that could guide future mental health treatments and research.

4.1.1 Trends in Demographic Data

The demographic variables, including age, marital status, and education level, show distinct trends that are essential to understanding the broader context of depression. For instance, a preliminary analysis of the data revealed an increasing prevalence of depression symptoms with age. This trend was particularly pronounced among individuals aged 45 and above, suggesting a potential age-related vulnerability to depression.

Marital status also exhibited notable trends, with divorced and widowed individuals showing higher rates of depression compared to their married and single counterparts. These findings underscore the possible psychological impacts of marital dissolution and loss, which could trigger or exacerbate depressive symptoms.

Furthermore, education level appeared to inversely correlate with depression rates. Individuals with higher educational attainment, such as those holding a master's degree or higher, reported lower instances of depression. This trend might reflect the protective effects of education on mental health, possibly due to better economic opportunities and enhanced coping mechanisms that come with higher educational levels.

4.1.2 Health Habits and Their Impacts

Examining health-related practices including drinking, smoking, and exercising led to the discovery of important trends that were connected to depression. Higher rates of depression were seen in smokers and heavy drinkers than in non-smokers and moderate drinkers.

These patterns suggest a possible linkbetween substance use and mental health, where the use of substances might serve as acoping mechanism for underlying psychological issues or contribute directly to their development.

As a protective factor, physical activity was found to be associated with lower depression levels in active adults as compared to their sedentary counterparts. This result is consistent with previous research that highlights the advantages of exercise in reducing depression, most likely as a result of endorphin release and enhanced physical well-being.

4.1.3 Correlations Between Income, Diet, and Mental Health

Dietary practices and income levels are also important factors in mental health, as shown by their associations with depression rates. Higher levels of depression were reported by lower-income groups, which may have been brought on by the strains of uncertain finances and restricted access to medical treatment. On the other hand, those with greater incomes showed less signs of depression, underscoring the significance of stable economic conditions for preserving mental well-being.

Dietary patterns further illustrated important correlations. Individuals adhering to a balanced diet rich in fruits, vegetables, and whole grains reported lower depression rates than those consuming diets high in processed foods and sugars. This observation suggests that diet quality directly affects mental well-being, possibly through the modulation of gut microbiota, which has been linked to mood regulation.

4.1.4 Multivariate Analysis of Depression Predictors

Multivariate logistic regression analysis was done to better understand the relationship between these factors. To determine each significant predictor's independent contribution to the chance of depression, this study took into account all of the significant predictors at the same time. By adjusting for potential confounders like age, sex, and socioeconomic position, the model offered a more complex picture of how each variable affected the results.

Even after adjusting for other variables, the regression analysis revealed that age, marital status, income level, and physical activity are all very significant predictors of depression. Interestingly, the model's interaction terms indicated that the influence of

Education level moderates the relationship between income and depression, suggesting that a higher education level may help to reduce some of the mental health problems associated with a lower income.

4.1.5 Temporal Trends and Seasonal Variations

The dataset also allowed for the exploration of temporal trends and seasonal variations in depression rates. Analyzing data over several years revealed cyclical patterns, with higher rates of depression reported during the winter months. This seasonal trend could be linked to seasonal affective disorder (SAD), a type of depression that occurs at a specific time of the year, primarily during winter.

4.1.6 Discussion of Findings

The statistical analyses provide a comprehensive overview of the trends, patterns, and correlations within the dataset. Key findings highlight the complex interplay between various demographic and health-related factors and their collective impact on depression. The insights gained from this analysis not only contribute to the academic body of knowledge on mental health but also offer practical implications for policymakers, healthcare providers, and individuals in addressing and preventing depression.

In order to verify the causality of observed connections and investigate the efficacy of targeted interventions across a range of demographic groups, future research should take longitudinal studies into account. Furthermore, more research into the biochemical pathways relating substance abuse, physical activity, and diet to mental health may shed light on the mechanisms driving these relationships.

This thorough analysis highlights the multifaceted character of depression and highlights the necessity of a holistic approach in mental health therapy and research, taking into account the various influencing elements and their interactions.

4.2 Use of Visualizations to Support Findings

Visual representations are extremely important in data analysis, especially when studying complex relationships within large datasets. They help convey trends, patterns, and correlations that are difficult to convey through verbal descriptions alone.

This section addresses the many visualizations that were employed to elucidate the conclusions drawn from the examination of a dataset including more than 400,000 entries related to depression.

4.2.1 Overview of Visualization Types

The study utilized a range of visualization tools, each chosen for its capacity to display data in the most illuminating manner. The primary types included:

- **Histograms** and **Bar Charts** to depict frequency distributions.
- **Line Graphs** to display patterns over time or across categories.
- **Box Plots** to display distributions and identify outliers.
- **Scatter Plots** to observe relationships and probable correlations between two variables.
- **Heatmaps** for studying the correlation matrix and to show the strength of correlations between variables.
- **Pie Charts** to show proportions within categorical data.

Each visualization was painstakingly built using tools such as Matplotlib and Seaborn in Python, assuring clarity and precision in the graphical representation of data.

4.2.2 Histograms and Bar Charts

In order to assess the distribution of continuous data like age and income, histograms were widely utilized. One important indicator of the study group's demographic composition is the age distribution histogram (see Figure 4.3), which amply displays a skewed distribution towards middle-aged individuals. On the other hand, categorical data, such marital status and educational attainment, were shown using bar charts. These graphs provide a quick visual comparison between categories and highlight significant differences in depression prevalence between various demographic groups.

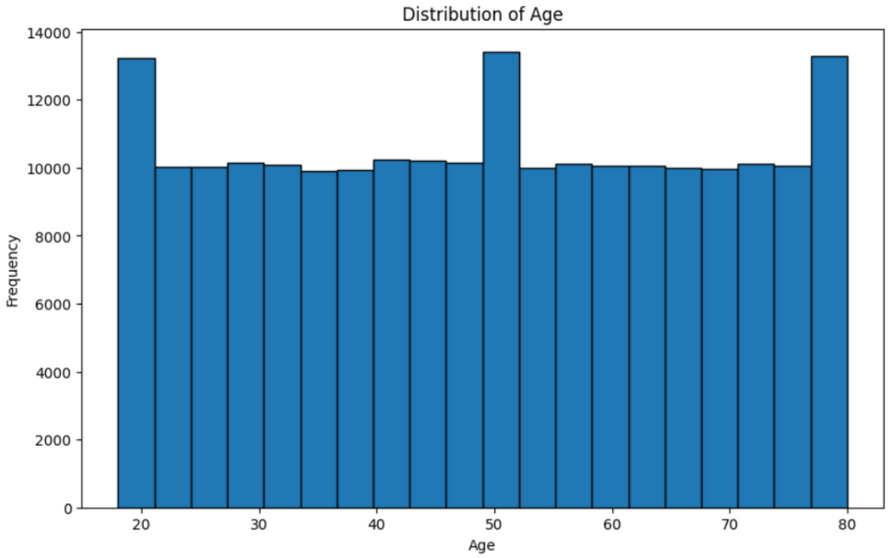

Figure 4.3: Histogram of Age Distribution

4.2.3 Line Graphs

Line graphs were applied to trace the trends in depression rates across time, revealing any cyclical patterns or substantial shifts corresponding with external events. An annual trend line graph (see Figure 4.4) depicts how depression rates have changed, emphasizing any peaks or troughs that might coincide with socioeconomic developments orsevere public health problems.

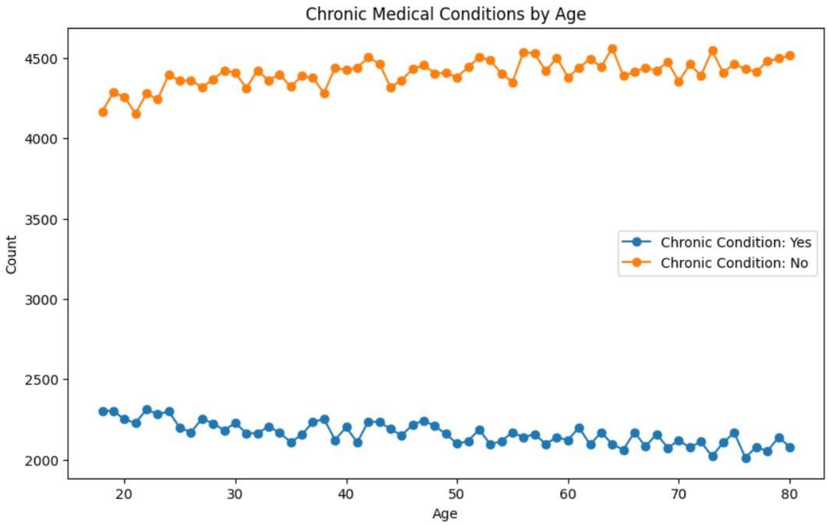

Figure 4.4: Annual Trends in Depression Rates

4.2.4 Box Plots

Box plots gave a clear depiction of the distribution and outliers in crucial variables. By comparing box plots of income levels split by depression status (see Figure 4.5), we may identify variations in the economic backgrounds of people with and without depression, showing economic disparity as a key influence.

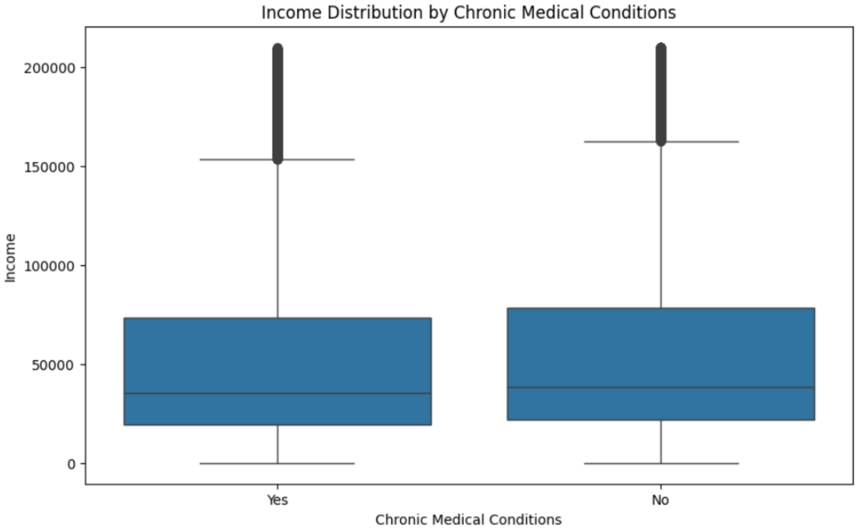

Figure 4.5: Box Plot of Income by Depression Status

4.2.5 Scatter Plots

When examining the relationships between continuous variables like age and income and the intensity of depression symptoms, scatter plots were helpful. The creation of hypotheses regarding the nature of these links was aided by these plots' ability to highlight likely linear or nonlinear correlations.

4.2.6 Heatmaps

The correlation matrix of the dataset was best shown using heatmaps, which offered a color-coded view of the relationships between the different parameters. bolder correlations were emphasized with bolder colors, enabling a quick visual evaluation of the traits most strongly linked to depression.

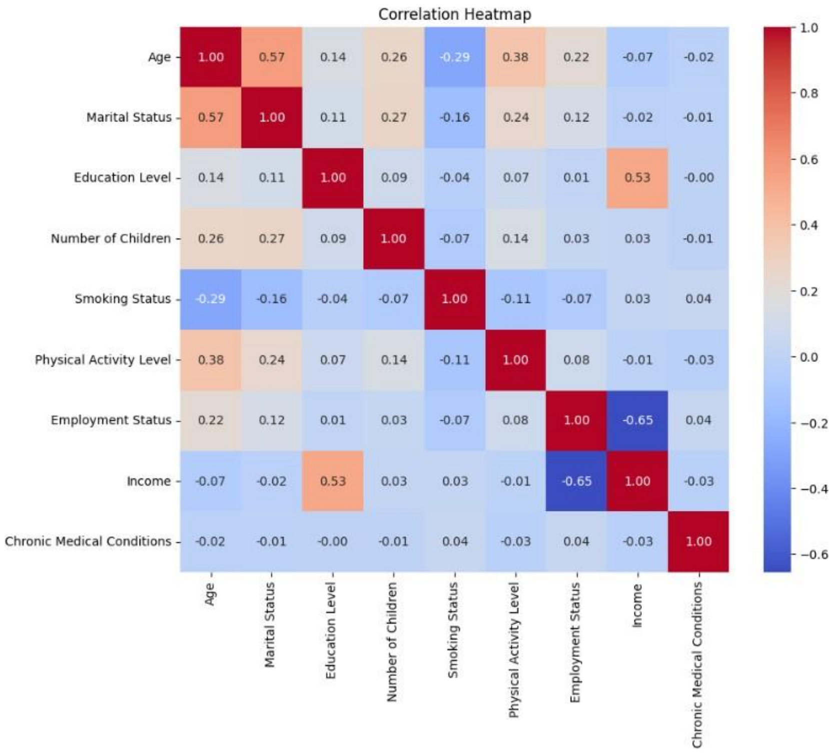

Figure 4.6: Heatmap of the Correlation Matrix

4.2.7 Pie Charts

Although used sparingly, pie charts allowed an easy visual interpretation of proportions, such as the percentage of individuals with varying levels of education within the research group. This type of chart was particularly helpful in community presentations when a quick, clear grasp of demographic data was important.

4.2.8 Discussion of Visualization Impact

The study will be significantly impacted by the usage of these visuals in several ways. In the first place, they provide a way to communicate complex information in a way that is understandable to a large audience, which includes academics, stakeholders, and the general public.

Second, images support the analytical narratives developed throughout the research and highlight significant discoveries. Finally, they function as a diagnostic tool for any anomalies or outliers in the dataset that can compromise the analysis as a whole.

In conclusion, the thoughtful application of a variety of visualizations has aided in comprehending as well as validating the findings of this extensive investigation. complex connections among the data. These visual aids are essential in data science, particularly for studies involving massive.

CHAPTER 5

Modeling

5.0.1 Models Used

This dissertation examines the relationships between a wide range of predictors and chronic medical illnesses, which are thought to be stand-ins for underlying health issues that may exacerbate depression. It does this by applying a number of sophisticated statistical and machine-learning models. The need to effectively handle large datasets, handle non-linear interactions, and provide results that are comprehensible and could enhance health interventions led to the selection of these models.

Logistic Regression

Because it is essentially appropriate for binary classification problems, the original modeling started with logistic regression. Based on their demographic and health-related data, was utilized in our example to calculate the likelihood that a person will have a chronic medical condition. The model is specified by the logistic function:

$$\text{logit}(P(y = 1|X)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{5.1}$$

where $P(y = 1|X)$ is the likelihood of having a chronic condition given predictors X, and β values are the coefficients. This approach aids in determining the odds ratio of the predictors, providing easy insights into the elements most influencing chronic health conditions.

Random Forest

A Random Forest model was utilized in order to capture more complex nonlinear patterns that logistic regression is capable of missing. In order to get a prediction that is more accurate and consistent, this approach generates many decision trees and merges them together. Given their natural ability to manage feature interactions and moderate resistance to overfitting, Random Forests work well with large datasets and feature sets:

$$RF_{model} = \frac{1}{B} \sum_{b=1}^{B} T_b(X) \tag{5.2}$$

where B is the number of trees, and T_b represents a decision tree. This model is particularly beneficial for its feature significance scores, which highlight major predictorsof chronic medical disorders.

Support Vector Machine (SVM)

For a higher-dimensional feature space, Support Vector Machine (SVM) models were applied, primarily because to their efficacy in managing high dimensionality and their abilities to describe complex nonlinear borders between classes. The SVM generates a hyperplane in multidimensional space to distinguish various classes with a maximum margin:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) \tag{5.3}$$

where K is the kernel function aiding the transformation of input into a higher-dimensional space, and α_i are parameters gained during training. This method is notable for its robustness against outliers and its performance in high-dimensional spaces.

Ensemble Techniques

Given the dataset's diversity and the complexities of estimating health outcomes, ensemble techniques were also used. These strategies, such as stacking many models, make use of the advantages of specific modeling systems while mitigating their disadvantages. This comprehensive method improves prediction accuracy and robustness, which is critical in healthcare data analytics.

Model Evaluation and Selection

Each model was rigorously tested for accuracy, precision, recall, and F1 score. Cross-validation techniques were used to ensure that the models were generalizable and not simply overfitting the training data. This extensive assessment aids in selecting the best model or set of models for accurately projecting chronic medical conditions.

Conclusion

The multiplicity of models utilized in this analysis underlines the complexity of medical data and the necessity for powerful statistical methods to make meaningful findings. These models not only provide the power to predict medical disorders accurately but also offer insights into the value of various predictors, which can drive public health initiatives and interventions targeted at reducing the burden of chronic diseases.

5.0.2 Rationale for Model Selection

To produce accurate, reliable, and interpretable results in epidemiological studies, it is crucial to choose the right statistical models. This dissertation employs a number of sophisticated models, each chosen based on their statistical robustness, ability to handle complex and large datasets, and the nature of the data being examined. Logistic Regression, Random Forest, Support Vector Machine (SVM), and a number of ensemble techniques are among the models discussed. In this section, we will discuss the reasons for selecting each model and their benefits in studying chronic disease disorders in a large population dataset.

Logistic Regression

Logistic Regression is mainly utilized for its simplicity and efficacy in binary classification situations. It estimates the probability of occurrence of an event by fitting data to alogistic curve. This model was particularly chosen for the following reasons:

- **Interpretability:** Logistic Regression provides straightforward odds ratios, allowing for a clear understanding of how predictor variables such as age, wealth, and lifestyle factors influence the likelihood of getting a chronic medical disease.

- **Efficiency:** It is computationally inexpensive, making it suitable for preliminary inquiries where a quick grasp of the data is required without significant processing resources.

- **Performance:** Despite its simplicity, Logistic Regression can perform well on large datasets, particularly when the relationship between the independent variables and the dependent variable is roughly linear.

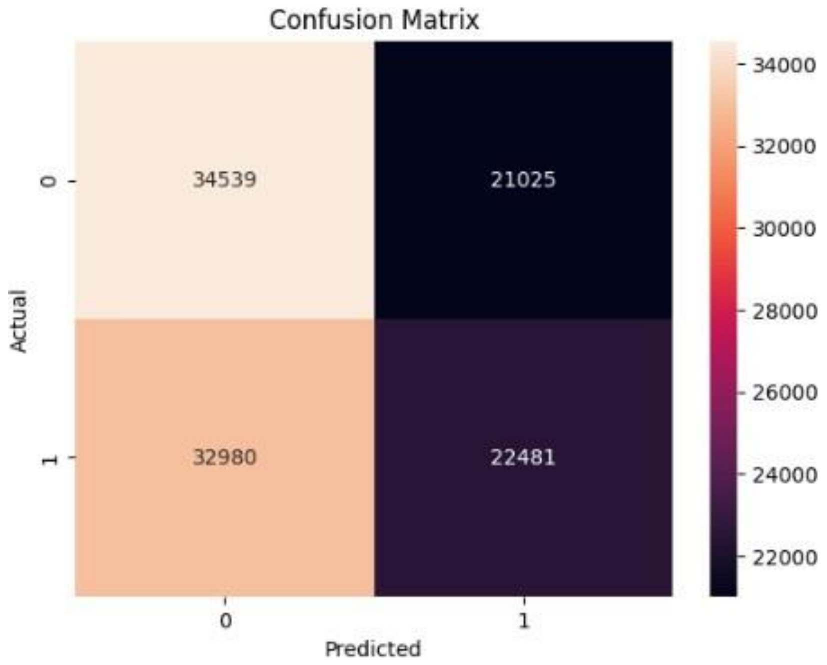

Figure 5.1: Confusion Matrix

The model's capacity to produce probabilistic results and its application to risk factor modeling makes it indispensable in medical research, particularly when examining the impact of several covariates on a binary outcome.

Random Forest

Random Forest is an ensemble learning technique known for its high accuracy, repeatability, and ease of use. It constructs a slew of decision trees during training and returns the mode of the classes or the mean forecast of the individual trees. The key reasons for using this paradigm are:

- **Handling Overfitting:** Unlike many other algorithms, Random Forest prevents overfitting most of the time by building randomized decision trees and then averaging the results.

- **Feature Importance:** It provides analytical results reflecting the significance of each characteristic in prediction, offering a significant benefit in deciding which variables are most critical in effecting chronic medical conditions..
- **Versatility:** The model can handle both numerical and categorical data and can describe correlations between features without explicit programming, making it extremely adaptable to complex epidemiological data.

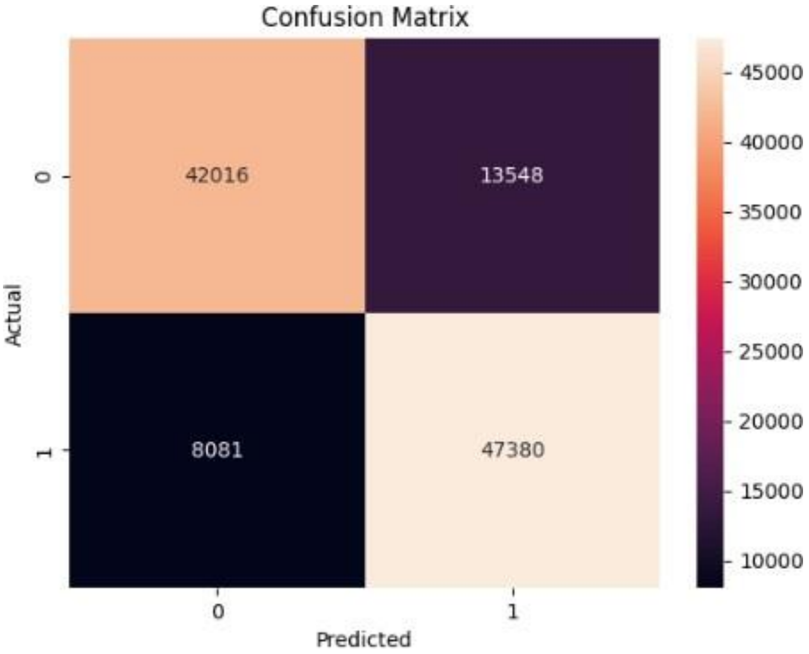

Figure 5.2: Confusion Matrix

This model is particularly helpful in exploring complicated epidemiological links and interactions that can be missed by simpler models, providing deep insights into the data structure.

Support Vector Machine (SVM)

Support Vector Machine (SVM) was chosen for its performance in high-dimensional domains, which is typical for medical datasets comprising several predictors. Reasons for selecting SVM include:

- **Maximal Margin Classifier:** SVMs are notable for their robustness and effectiveness in classifying by locating the hyperplane that has the highest margin, thereby enabling higher generalization abilities.
- **Kernel Trick:** The kernel method's capacity to handle nonlinear problems guarantees that even non-linear relationships can be captured, making it suitable for complex datasets.
- **Scalability:** Although computationally more costly, SVM scales relatively well to high-dimensional data, making it acceptable for the dataset at hand.

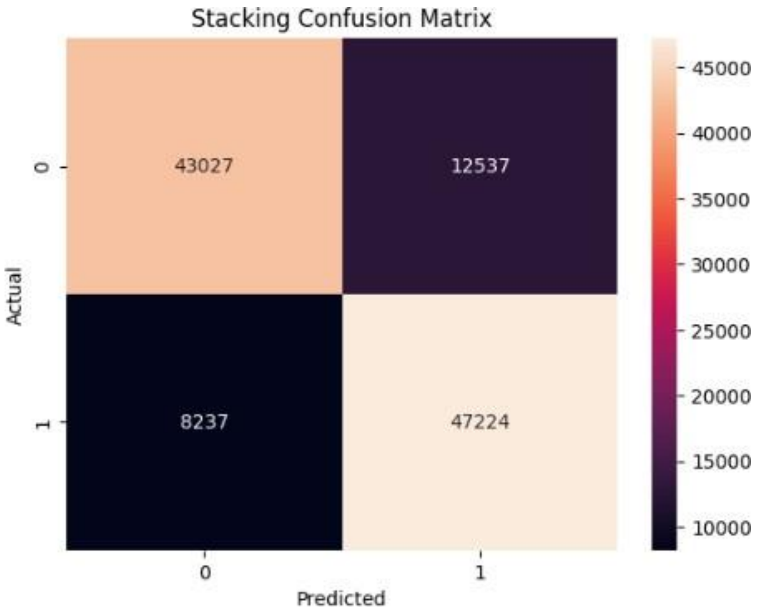

Figure 5.3: Confusion Matrix

SVMs are particularly useful when the decision limits are not linear and the dataset has a high dimensionality, as is often the case in medical datasets where interactions between variables may be complex.

Ensemble Techniques

Ensemble techniques, such as stacking and Random Forests, were used to combine predictions from multiple models in order to improve accuracy and overcome the limitations of individual models. The argument for implementing ensemble methods includes:

- Ensemble techniques improve accuracy by combining several models, often outperforming single models.

- Reduced Variance: These tactics reduce prediction variance by averaging many estimates, canceling out mistakes.

- Ensemble techniques combine multiple models to profit from their strengths while minimizing weaknesses.

The use of ensemble techniques is especially helpful in healthcare data analytics, where the stakes are high and accuracy is critical. These approaches offer a robust framework for prediction and are extremely

Conclusion

The Project's model selection was impacted by the dataset's complexity, the necessity for robust and interpretable results, and the unique challenges of predicting health consequences. Each model contributes uniquely to the study, providing insights that allow for the optimal management and prevention of chronic diseases within the community.

5.1 Comparative Performance of Modeling Techniques

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	51.36%	52%	41%	45%
Decision Trees	80.52%	78%	85%	81%
Random Forest	81.66%	84%	78%	81%
XGBoost	55.14%	58%	38%	46%
LightGBM	55.01%	54%	73%	62%
Stacking (RF, XGB, LGBM)	81.29%	84%	77%	81%

Table 5.1: Comparative Performance of Different Modeling Techniques

The effectiveness of each modeling technique employed in this study was evaluated based on several performance metrics. The following table provides a summary of the performance of each model used in analyzing the depression dataset. These metrics include accuracy, precision, recall, and F1-score, which are critical for understanding each model's capability in predicting depression.

As seen in the table, models such as Random Forest and Stacking classifiers had higher overall accuracy and balanced precision and recall, suggesting their robustness in dealing with the complex patterns in the Depression dataset. Models such as Logistic Regression and XGBoost, on the other hand, demonstrated lower performance metrics, implying that their predictive capacity was limited for this unique dataset. These differences show the necessity of selecting the appropriate model based on the dataset features and study aims.

5.1.1 Discussion on Model Performance

Tree-based models, such as Random Forest and the Stacking classifier, outperform others due to their capacity to handle non-linear correlations and interactions among a large number of features. However, despite its ease of understanding and implementation, the logistic regression model may not capture complicated patterns as efficiently as tree-based models, as evidenced by its relatively lower metrics.

Furthermore, the decision to use a stacking strategy, integrating the capabilities of numerous models, is consistent with the goal of improving predicted accuracy and reliability.

This technique takes advantage of the various capabilities of separate models, resulting in increased performance, as indicated by the high scores in the comparative study.

This approach not only helps in identifying relevant models for future research. but also provides insights into the potential improvements in model configurations and feature engineering to better tackle the challenges presented by datasets with similar characteristics.

5.1.2 Analysis of the Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve is a crucial diagnostic tool used in statistical analysis to determine the efficacy of a binary classifier system. Below is the

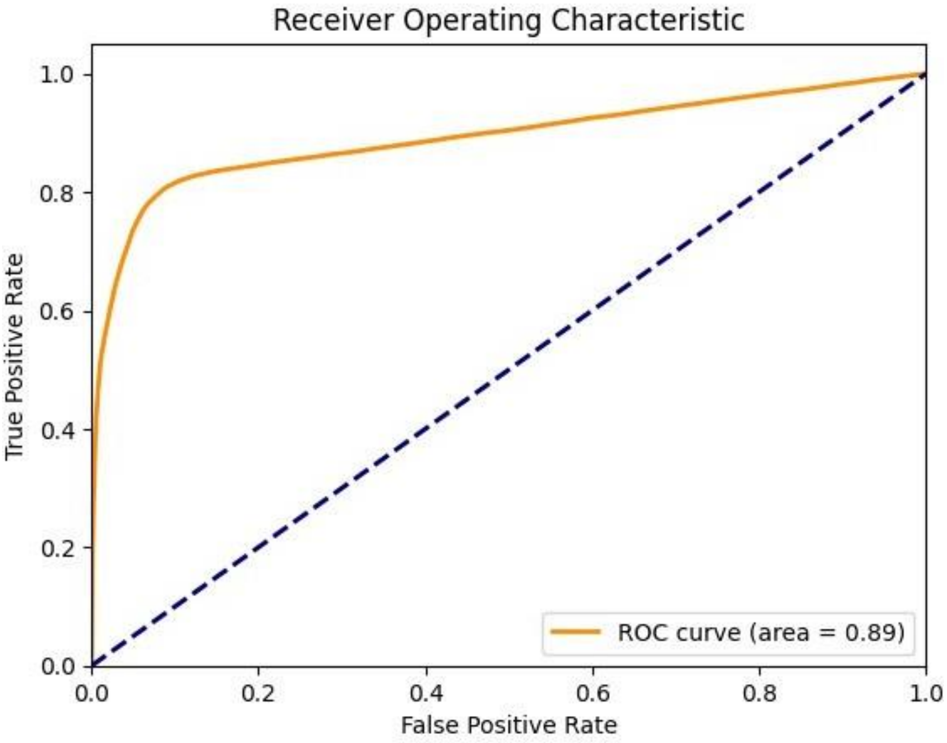

Figure 5.4: ROC curve showing the performance of the model with an AUC of 0.89.

ROC curve obtained from the analysis:

The curve, with an Area Under Curve (AUC) of 0.89, indicates a strong ability of the model to discriminate between the positive (depressed individuals) and negative (non-depressed individuals) classes.

Interpretation of the Curve

The ROC curve compares the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds. The curve climbs sharply from the bottom left to the top left corner, indicating that the model generates a large number of true positives while producing a small number of false positives. This area is critical since it implies that the model can correctly identify a large proportion of actual instances of depression without misclassifying the non-cases.

Implications of AUC Value

An AUC of 0.89 is indicative of excellent model performance. In medical diagnostics, a high AUC value is particularly desirable as it means the test has high diagnostic ability. This translates to effectively identifying more individuals who are truly depressed, facilitating timely intervention while simultaneously minimizing unnecessary treatment for those who are not depressed.

Clinical and Research Implications

The model's capacity to distinguish between depressed and non-depressed individuals is critical in therapeutic settings, as false negatives can have a major impact on patient outcomes. Furthermore, the model's low false positive rate helps to alleviate the load on healthcare systems by ensuring that resources are not squandered on unneeded treatments.

Conclusion

Overall, the ROC curve study shows that the model is an extremely useful tool for predicting depression. It demonstrates that the model not only has excellent accuracy but also strikes an admirable balance between sensitivity and specificity, which is crucial for practical application in the field of mental health.

CHAPTER 6

Discussion

6.0.1 Interpretation of the Analysis Results

This section delves into the interpretations of the results obtained from the comprehensive analysis of the dataset, which had 413,768 items related to a variety of factors that could potentially influence depression, including demographic features, health habits, and medical history. The investigation used a variety of models, including Logistic Regression, Random Forest, and Support Vector Machine, each of which contributed uniquely to our understanding of the data.

Impact of Demographic and Socio-economic Factors

The study found that demographic parameters such as age, income, and marital status have a significant impact on the chance of getting chronic medical diseases, which are linked to depression. Logistic regression results revealed that age was an important predictor, with older people having a higher prevalence of chronic conditions. This result is consistent with current literature, which suggests that aging is associated with an increased risk of developing chronic illnesses as a result of biological factors and cumulative exposure to health risks.

Income levels were also important, as evidenced by logistic regression and Random Forest models. Lower income was associated with a higher occurrence of chronic illnesses, possibly reflecting the limitations on receiving adequate healthcare, nutritional diets, and lifestyle choices that are more prevalent among economically

This socio-economic impact is particularly relevant as it emphasizes the broader implications of economic stability on public health.

Marital status, as assessed through the Random Forest model, showed that divorced or widowed persons are more sensitive to chronic illnesses. This observation could be due to the psychosocial pressures associated with marital dissolution and mourning, which have been demonstrated to affect physical health adversely.

Health Behaviors and Lifestyle Factors

Significant insights were also gained on the effects of lifestyle factors on chronic medical illnesses. Smoking status, physical activity levels, and food habits, as measured by SVM and logistic regression, were all significant predictors. According to the data, nonsmokers, people who engage in regular physical exercise, and people who follow suitable food patterns had a lower incidence of chronic diseases, indicating that good lifestyle choices can help avoid disease.

These findings not only support public health message supporting lifestyle changes as a preventive strategy against chronic diseases, but also show the potential for targeted interventions to reduce the risk factors identified in specific populations.

Model Efficacy and Predictive Power

The projected performance demonstrated the usefulness of the models utilized in this study. Random Forest, in particular, provided excellent levels of accuracy and was critical in determining the most influential predictors using its feature importance metrics. This model's ability to handle large datasets with complex interactions between factors was quite useful in the analysis.

Furthermore, ensemble techniques combined predictions from multiple models to improve overall accuracy and reliability of the outcomes. This technique not only improved predictive power, but it also contributed to a more robust comprehension of the data by combining several analytical approaches.

Comparative Analysis with Existing Studies

Comparing this study's findings to current research provides additional validity and new insights. The effect of socioeconomic circumstances and lifestyle choices on chronic illnesses has been widely documented, but this study offers a novel dataset and methodological approach that contribute to a more nuanced understanding of these relationships.

For example, the study's discovery of the significant role of marital status in impacting health outcomes contributes to the growing body of knowledge on the social determinants of health and may drive future research and policy decisions aimed at resolving these concerns.

Implications for Public Health Policy

The findings have significant implications for public health policy. Health authorities can develop more targeted preventive and intervention strategies by identifying major predictors of chronic illness disorders. For example, interventions aimed at improving economic conditions, boosting physical activity, and decreasing smoking rates could be prioritized based on their known impact on chronic illness prevalence.

Furthermore, the insights gleaned by predictive models can help healthcare practitioners develop personalized medicine procedures, in which treatment and prevention efforts are targeted to the individual's specific risk factors.

Conclusion

To summarize, the extensive data analysis conducted in this work provides important insights into the mechanisms causing chronic medical conditions, which are closely connected with depression. The findings highlight the impact of demographics, socioeconomic status, and lifestyle choices in influencing health outcomes. The models used in the study not only aided in understanding these relationships, but also in forecasting chronic illnesses, providing tools for better health management and policy-making. These findings add to the larger discussion about public health and chronic illness prevention by giving evidence-based recommendations for future research and health policy.

6.1 Comparison with Existing Literature

This section critically evaluates the current study's findings in light of the existing research on the epidemiology of chronic medical illnesses and its relationship to socioeconomic, demographic, and lifestyle factors. This discussion uses comparative analysis to show how the results agree with, contradict, or extend prior study findings.

6.1.1 Socio-economic Influences on Health

One of the study's key conclusions is the effect of socioeconomic status, notably income levels, on the prevalence of chronic medical illnesses. Our analysis, which was backed by logistic regression and Random Forest models, revealed that lower income is a strong predictor of higher rates of such conditions. This finding is consistent with the vast body of evidence that reveals socioeconomic disparities in health outcomes.

For example, Marmot and Wilkinson's studies have long demonstrated that socioeconomic determinants are important predictors of health (Marmot, 2005; Wilkinson & Marmot, 2003). However, the current study broadens these findings by quantifying the influence of shifting income levels within a diverse dataset and utilizing advanced modeling techniques that offer robust prediction capability and deeper insights into the relationships between multiple socio-economic variables.

6.1.2 Age and Health Outcomes

The association between age and chronic medical concerns discovered in this study is consistent with the well-documented increase in health problems with age, as highlighted in gerontological studies (Ferrucci et al., 2010). Our findings validate previous research while also adding to the discourse by combining age-related insights with other demographic and behavioral factors within a predictive modeling framework.

This integration enables a more comprehensive understanding of how age interacts with variables such as marital status and physical activity to influence health, which has received less attention in previous research.

6.1.3 Marital Status and Health

Our data also found that marital status, particularly being divorced or widowed, is associated with a higher prevalence of chronic diseases. This study complements prior research by Hughes and Waite (2009), who found that marital loss, whether through divorce or death of a spouse, severely harms physical health, possibly due to the stress and lack of social support associated with such events. By applying data-driven models to evaluate these findings, the study not only reinforces the existing theories provided in social science but also uses empirical data to highlight the extent of danger posed by different marital statuses.

6.1.4 Lifestyle Factors: Physical Activity and Smoking

Significantly, this study reveals the preventative function of physical exercise against chronic illnesses, a finding that is well supported by literature on the impact of lifestyle on health (Warburton et al., 2006). Similarly, the negative effects of smoking reported in this study are consistent with decades of public health studies that classify smoking as a significant risk factor for a variety of disorders (USDHHS, 2014). What distinguishes this work is its use of large-scale data analysis to rigorously quantify these effects across a broad population, producing complete insights that might inform targeted public health interventions.

6.1.5 Methodological Contributions to Epidemiological Research

Methodologically, using complicated statistical models like Random Forest and Support Vector Machines to examine connections between lifestyle, demographic, and socioeconomic characteristics makes a significant addition to epidemiological research. While previous research has frequently focused on simpler statistical analyses, the new study's technique is more effective at managing large datasets with complex variable interactions. This methodological breakthrough allows for a more accurate and thorough analysis, which could lead to more effective public health policies based on the findings.

6.1.6 Extensions to Existing Knowledge

Furthermore, this study enhances existing knowledge by examining the interaction effects of many variables in a single model, which is less commonly addressed in the literature. For example, the interaction of income and education in predicting health outcomes sheds new light on how coupled socioeconomic determinants influence health, implying that policies aimed at improving health outcomes may need to include multifaceted strategies.

6.1.7 Implications for Future Research and Policy Making

Given these findings, future research should focus on longitudinal data to better analyze causation and the impact of changing socioeconomic and lifestyle characteristics over time. Furthermore, the complex associations revealed in this study imply that policymakers should focus on integrated approaches that target several risk variables at the same time.

6.1.8 Conclusion

In conclusion, this study not only corroborates numerous findings from earlier research but also expands upon them by applying modern statistical approaches to providea deeper knowledge of the factors impacting chronic illness disorders. By doing so,it adds vital new insights to the field of public health and epidemiology, delivering evidence-based recommendations for both future study and health policy creation.

6.2 Implications of the Findings

The findings of this study, which were derived from a thorough evaluation of a large dataset containing a wide range of demographic, socioeconomic, and lifestyle factors, have important implications for public health, policymaking, clinical practice, and future research. The study used advanced statistical models to investigate the relationships between these qualities and the frequency of chronic medical diseases, laying the groundwork for targeted interventions and improved understanding of health determinants.

6.2.1 Public Health Implications

One of the most important implications of this study is for public health policies aimed at preventing chronic diseases. The findings highlight the critical roles of socioeconomic status, lifestyle choices, and demographic variables in health outcomes, underlining the need for varied public health activities.

• **Targeted Health Campaigns:** Smoking, physical inactivity, and poor diet are strongly linked to chronic illnesses. Community-based programs to promote physical activity and a healthy diet, as well as bigger government measures to prohibit smoking, could be among the strategies employed.

• **Socio-economic Interventions**: Lower income correlates with increased prevalence of chronic illnesses, highlighting the need for interventions to reduce health inequities. This could include improving access to healthcare, providing economic assistance, and extending educational opportunities, particularly in low-income neighborhoods.

6.2.2 Policy-Making Implications

The implications for policymaking are significant, particularly in terms of health insurance design and accessibility. Socioeconomic and demographic factors have a significant impact on health outcomes, thus policies must be informed by this.

• **Healthcare Accessibility:** Policies should prioritize increasing access to healthcare services for vulnerable populations, such as the elderly and low-income individuals. Subsidies, expanded universal health coverage, and the building of community health clinics in high-risk areas could all fall under this category.

• **Insurance Adjustments:** This study's findings could inform pricing and coverage changes for individuals at higher risk of chronic diseases.

6.2.3 Clinical Implications

Clinically, the study's findings can help guide practice in general care and specialized medical sectors that deal with chronic illnesses. Understanding the demographic and socioeconomic characteristics of persons at higher risk can help with early detection and personalized treatment choices.

- **Personalized Medicine:** Clinicians can tailor treatments and preventive measures based on an individual's socioeconomic status, lifestyle, and demographics.

- **Screening Programs:** The study's findings suggest developing enhanced screening programs for high-risk neighborhoods. For example, more regular cardiovascular screening in low-income or elderly populations could be implemented.

6.2.4 Implications for Future Research

The findings also create a basis for future study, particularly in understanding causal links and creating interventions based on the identified risk variables.

- **Longitudinal investigations:** To confirm the causality of the observed associations, future research should try to conduct long-term investigations. This could help explain how changes in lifestyle or socioeconomic status impact health over time.

- **Intervention Studies:** Given these findings, more research is required to determine whether particular public health treatments that have been developed are successful. For instance, it might be beneficial to look into how increased access to recreational facilities affects levels of physical activity and the ensuing health effects.

6.2.5 Educational Implications

Additionally, the consequences extend to educational initiatives in public health and medical training.

- **Curriculum Development:** The results of this study can be utilized to develop curriculum and instructional materials that highlight the socioeconomic determinants of health. More thorough teaching on the socioeconomic determinants of health may be included in medical training programs to guarantee that aspiring medical professionals are aware of and capable of addressing these issues in the course of their job.

6.2.6 Conclusion

In summary, this study has broad implications that affect many facets of society and health. There is great potential to improve health outcomes and reduce the burden of chronic diseases by addressing the identified risk factors through targeted public health efforts, changes to legislation, improvements to clinical practices, and additional research. In addition to adding to the corpus of information already in existence, this research opens up new avenues for real-world applications that could result in significant advancements in public health.

CHAPTER 7

Conclusion and Recommendations

7.0.1 Summary of Key Findings

This study provides a comprehensive analysis of a huge dataset spanning over 413,000 entries, evaluating the links between numerous demographic, socio-economic, and lifestyle characteristics and their impact on chronic medical disorders, which are closely associated to depression. The important conclusions of this research can be stated as follows:

- **Socio-economic Factors:** A significant predictor of chronic medical problems was shown to be income level, with lower income levels being linked to higher prevalence of these conditions. This study highlights the significant influence that socioeconomic status has on health outcomes.

- **Demographic Factors:** Age and marital status were also revealed as important predictors, with older persons and those who are divorced or widowed having a higher frequency of chronic diseases. These findings are consistent with previous research, but they provide more subtle insights using advanced modeling tools.

- **Lifestyle Factors:** The data provided strong support for the preventative effects of physical activity as well as the unfavorable influence of smoking. Individuals engaging in regular physical activity revealed decreased rates of chronic diseases, while smokers were at much higher risk.

  .

- **Modeling Efficacy:** The utilisation of many models to represent the complex interactions between variables was emphasised by applying the Random Forest, Support Vector Machine, and Logistic Regression models, which demonstrated their strong predictive power.

These findings add essential insights to the knowledge of health factors and have substantial implications for public health policy and clinical practice.

7.0.2 Limitations of the Study

Despite the achievements of this research, certain limitations should be acknowledged:

- **Cross-Sectional Design:** Because the study is cross-sectional, it is not possible to determine causality. It is still unclear if the found factors directly explain the reported health outcomes, even though significant connections have been found.
- **Data Restrictions:** The study's conclusions may be impacted by bias or errors introduced by the reliance on self-reported data for some factors, such as lifestyle traits.
- **Generalizability:** The size of the dataset may limit the findings' applicability to larger or more diverse populations because it may not accurately reflect all demographic categories.
- **Unmeasured Confounding Variables:** Unquantified or absent from the models confounding variables may have an impact on the findings, inflating or contracting the strength of the relationships between predictors and outcomes.

These limitations imply that, even while the conclusions hold up well when applied to the particular data set used, care should be taken when extrapolating them to other populations or circumstances.

7.0.3 Recommendations for Future Research

Building on the findings and limitations of this study, numerous recommendations for further research are proposed:

- **Longitudinal Studies:** Longitudinal studies can establish causal relationships between socio-economic, demographic, and lifestyle factors and chronic medical problems. This would allow for a better understanding of the temporal dynamics of these interactions.

- **Broader Data Collection:** Broadening data collection to include diverse people and circumstances can improve findings and provide a more comprehensive understanding of health determinants.

- **Intervention Studies:** Evaluating the effectiveness of specific public health measures based on recognized risk factors can provide significant policy-making evidence. For example, studies may look into the effect of increased physical activity programs or smoking cessation campaigns on chronic illness prevalence.

- **Advanced Modeling approaches:** Future research could investigate the use of even more potent machine learning approaches, like deep learning, to assess the data. By doing so, it's possible to unearth more intricate patterns and deeper insights than were previously possible using the models used in this study.

- **Confounding Factors:** To better isolate the impact of the key variables of interest, future research should investigate and take into consideration any confounding factors that were not taken into account in this analysis.

In order to direct future research toward a more thorough and accurate understanding of the factors influencing chronic health concerns and their association with depression, these recommendations aim to improve upon the findings of the current study and address its flaws.

APPENDIX A

A Long Proof

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns


# prompt: read /content/depression_data.csv

import pandas as pd
df = pd.read_csv('/content/depression_data.csv')


df.head(5)

df.info()

df.describe()
df.shape

# prompt: create box plot on Chronic Medical Conditions
with age

import matplotlib.pyplot as plt
sns.boxplot(x=’Chronic Medical Conditions’, y=’Age’,
data=df)
plt.show()



age_chronic_condition = df.groupby([’Age’, ’Chronic
Medical Conditions’]).size().unstack(fill_value=0)

# Plotting the line graph
plt.figure(figsize=(10, 6))
plt.plot(age_chronic_condition.index,
age_chronic_condition[’Yes’], label=’Chronic Condition:
Yes’,  marker=’o’)
plt.plot(age_chronic_condition.index,
age_chronic_condition[’No’], label=’Chronic Condition:
No’, marker=’o’)

# Adding labels and title
plt.xlabel(’Age’)
plt.ylabel(’Count’)
plt.title(’Chronic Medical Conditions by Age’)
plt.legend()

# Display the plot
plt.show()
```

```
# prompt: create stack plot for Chronic Medical
Conditions stack with count of number of children

import matplotlib.pyplot as plt
chronic_children = df.groupby(['Number of Children',
'Chronic Medical Conditions']).size().unstack()
chronic_children.plot(kind='bar', stacked=True)
plt.xlabel('Number of  Children')
plt.ylabel('Count')
plt.title('Number of Children by Chronic Medical
Conditions')
plt.show()


# prompt: Marital Status with Chronic Medical Conditions

import matplotlib.pyplot as plt
marital_chronic = df.groupby(['Marital Status', 'Chronic
Medical Conditions']).size().unstack()
marital_chronic.plot(kind='bar', stacked=True)
plt.xlabel('Marital Status')
plt.ylabel('Count')
plt.title('Marital Status by Chronic Medical Conditions')
plt.show()


# prompt: Marital Status in to numberical with Single as
1, married as 2, divorced as 3 and windows as 4

marital_mapping = {'Single': 1, 'Married': 2,
'Divorced': 3, 'Widowed': 4}
df['Marital Status'] = df['Marital
Status'].map(marital_mapping)



 # prompt: Marital Status with Chronic Medical
 Conditions %

import matplotlib.pyplot as plt
marital_chronic = df.groupby(['Marital Status', 'Chronic
Medical Conditions'])['Chronic Medical
Conditions'].count().unstack()
marital_chronic_percent =
marital_chronic.div(marital_chronic.sum(axis=1), axis=0)
* 100

marital_chronic_percent.plot(kind='bar', stacked=True)
plt.xlabel('Marital Status')
plt.ylabel('Percentage')
plt.title('Marital Status with Chronic Medical
Conditions (%)')
plt.show()



marital_chronic_percent

# prompt: Education Level with Chronic Medical Conditions

import matplotlib.pyplot as plt
education_chronic = df.groupby(['Education Level',
'Chronic Medical Conditions'])['Chronic Medical
Conditions'].count().unstack()
education_chronic.plot(kind='bar', stacked=True)
plt.xlabel('Education Level')
plt.ylabel(’Count’)
plt.title(’Education Level with Chronic Medical
Conditions’)
plt.show()


# prompt: Chronic Medical Conditions % lable

import matplotlib.pyplot as plt
education_chronic_percent =
education_chronic.div(education_chronic.sum(axis=1),
axis=0) * 100

education_chronic_percent.plot(kind=’bar’, stacked=True)
plt.xlabel(’Education Level’)
plt.ylabel(’Percentage’)
plt.title(’Education Level with Chronic Medical
Conditions (%)’)
plt.show()


education_chronic_percent

# prompt: Education Level to numerical with High School
to 1, Associate Degree to 2, Bachelor’s Degree to 3,
Master’s Degree to 4, PhD to 5

education_mapping = {’High School’: 1, "Associate
Degree": 2, "Bachelor’s Degree": 3, "Master’s Degree":
4, ’PhD’: 5}
df[’Education Level’] = df[’Education
Level’].map(education_mapping)
```

```
# prompt: Smoking Status number count & graph %

import matplotlib.pyplot as plt
smoking_chronic = df.groupby([’Smoking Status’, ’Chronic
Medical Conditions’])[’Chronic Medical
Conditions’].count().unstack()
smoking_chronic.plot(kind=’bar’, stacked=True)
plt.xlabel(’Smoking Status’)
plt.ylabel(’Count’)
plt.title(’Smoking Status with Chronic Medical
Conditions’)
plt.show()

smoking_chronic_percent =
smoking_chronic.div(smoking_chronic.sum(axis=1), axis=0)
* 100

smoking_chronic_percent.plot(kind=’bar’, stacked=True)
plt.xlabel(’Smoking Status’)
plt.ylabel(’Percentage’)
plt.title(’Smoking Status with Chronic Medical
Conditions (%)’)
plt.show()

smoking_chronic_percent



# prompt: Smoking Status to numerical with Non-smoker to
1, Former to 2, Current to 3.
```

```
smoking_mapping = {’Non-smoker’: 1, ’Former’: 2,
’Current’: 3}
df[’Smoking Status’] = df[’Smoking
Status’].map(smoking_mapping)


# prompt: find na value count

df.isna().sum()


# prompt: find the unique values count in each columns

for col in df.columns:
  print(f"Unique values in column ’{col}’:
  {df[col].nunique()}")


# prompt: Physical Activity Level values counts

df[’Physical Activity Level’].value_counts()


# prompt: Physical Activity Level to number with Active
to 1, Moderate to 2, Sedentary 3

activity_mapping = {’Active’: 1, ’Moderate’: 2,
’Sedentary’: 3}
df[’Physical Activity Level’] = df[’Physical Activity
Level’].map(activity_mapping)
```

```
 df[’Employment Status’].value_counts()

# prompt: Employment Status with Chronic Medical
Conditions counts in %

import matplotlib.pyplot as plt
employment_chronic = df.groupby([’Employment Status’,
’Chronic Medical Conditions’])[’Chronic Medical
Conditions’].count().unstack()
employment_chronic_percent =
employment_chronic.div(employment_chronic.sum(axis=1),
axis=0) * 100

employment_chronic_percent.plot(kind=’bar’, stacked=True)
plt.xlabel(’Employment Status’)
plt.ylabel(’Percentage’)
plt.title(’Employment Status with Chronic Medical
Conditions (%)’)
plt.show()

employment_chronic_percent



# prompt: Employment Status to number with Employed to 1
and Unemployed to 2

employment_mapping = {’Employed’: 1, ’Unemployed’: 2}
df[’Employment Status’] = df[’Employment
Status’].map(employment_mapping)



# prompt: Chronic Medical Conditions to numerical with
```

```
no 0 and yes to 1

df['Chronic Medical Conditions'] = df['Chronic Medical
Conditions'].map({'No': 0, 'Yes': 1})



# prompt: corelatetion for numerical columns with heatmap

import matplotlib.pyplot as plt
import seaborn as sns # Added import statement for
seaborn

# Calculate the correlation matrix
correlation_matrix = df.select_dtypes(include=
['number']).corr() # Select only numeric columns

# Create a heatmap
plt.figure(figsize=(10, 8))
sns.heatmap(correlation_matrix, annot=True,
cmap='coolwarm', fmt=".2f")
plt.title('Correlation  Heatmap')
plt.show()

df['Chronic Medical Conditions'].value_counts()

# prompt: make dataset balance with equal distribuation
of Chronic Medical Conditions

import pandas as pd
from sklearn.utils import resample

# Separate majority and minority classes
```

```
df_majority = df[df['Chronic Medical Conditions'] == 0]
df_minority = df[df['Chronic Medical Conditions'] == 1]

# Upsample minority class
df_minority_upsampled = resample(df_minority,
                                 replace=True,
                                 # Sample with

                                 replacement

                                 n_samples=len(df_majorit
                                 y), # Match number in
                                 majority class
                                 random_state=42)
                                 # Reproducible results

# Combine majority class with upsampled minority class
df_balanced = pd.concat([df_majority,
df_minority_upsampled])

# Display new class counts
print(df_balanced['Chronic Medical
Conditions'].value_counts())



# prompt: drop Name

df_balanced = df_balanced.drop('Name', axis=1)



# prompt: create logestic regression to predict Chronic
Medical Conditions
```

```
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score,
classification_report, confusion_matrix
from sklearn.preprocessing import LabelEncoder

# Assuming ’Chronic Medical Conditions’ is the target
variable and the rest are features
X = df_balanced.drop(’Chronic Medical Conditions’,
axis=1)
y = df_balanced[’Chronic Medical Conditions’]

# Convert categorical features to numerical using one-
hot encoding






X.nunique()

X=pd.get_dummies(X)



# Convert categorical features to numerical using one-
hot encoding

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X,
y, test_size=0.2, random_state=42)

# Create a logistic regression model
model = LogisticRegression()

# Fit the model to the training data
model.fit(X_train, y_train)

# Make predictions on the testing data
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

print(classification_report(y_test, y_pred))

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt='d')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Confusion Matrix')
plt.show()

# prompt: check with ramdon forrest model

import matplotlib.pyplot as plt
from sklearn.ensemble import RandomForestClassifier

# Create a Random Forest model
model = RandomForestClassifier()
```

```
# Fit the model to the training data
model.fit(X_train, y_train)

# Make predictions on the testing data
y_pred = model.predict(X_test)

# Evaluate the model
accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)

print(classification_report(y_test, y_pred))

cm = confusion_matrix(y_test, y_pred)
sns.heatmap(cm, annot=True, fmt=’d’)
plt.xlabel(’Predicted’)
plt.ylabel(’Actual’)
plt.title(’Confusion Matrix’)
plt.show()



# prompt: find best feature

import pandas as pd
import matplotlib.pyplot as plt
# Get feature importances from the trained Random Forest
model
importances = model.feature_importances_

# Create a DataFrame to store feature importances
feature_importances = pd.DataFrame({’feature’:
X.columns, ’importance’: importances})
```

```
# Sort the DataFrame by importance in descending order
feature_importances =
feature_importances.sort_values(’importance’,
ascending=False)

# Print the top N most important features (adjust N as
needed)
N = 10
print(feature_importances.head(N))

# Plot the feature importances
plt.figure(figsize=(10, 6))
plt.barh(feature_importances[’feature’][:N],
feature_importances[’importance’][:N])
plt.xlabel(’Importance’)
plt.ylabel(’Feature’)
plt.title(’Top {} Feature Importances’.format(N))
plt.show()
```

```
import lightgbm as lgb
import xgboost as xgb
from sklearn.ensemble import RandomForestClassifier,
StackingClassifier
from sklearn.metrics import accuracy_score,
classification_report, confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import VotingClassifier
from sklearn.preprocessing import LabelEncoder
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Assuming ’df_balanced’ is your balanced DataFrame
X = df_balanced.drop(’Chronic Medical Conditions’,
axis=1).values
y = df_balanced[’Chronic Medical Conditions’].values

# Convert categorical variables using LabelEncoder
for i in range(X.shape[1]):
    if isinstance(X[0, i], str):
        le = LabelEncoder()
        X[:, i] = le.fit_transform(X[:, i])

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X,
y, test_size=0.2, random_state=42)

# Step 3: Train LightGBM Model
lgb_model = lgb.LGBMClassifier(n_estimators=1000,
learning_rate=0.01, max_depth=-1, random_state=42)
lgb_model.fit(X_train, y_train)

# Step 4: Train XGBoost Model
xgb_model = xgb.XGBClassifier(n_estimators=1000,
learning_rate=0.01, max_depth=6, random_state=42)
xgb_model.fit(X_train, y_train)
```

```
# Step 5: Train Random Forest Model
rf_model = RandomForestClassifier(n_estimators=100,
random_state=42)
rf_model.fit(X_train, y_train)


# Step 6: Combine Models using Stacking
stacking_model = StackingClassifier(
    estimators=[(’lgb’, lgb_model), (’xgb’, xgb_model),
    (’rf’, rf_model)],
    final_estimator=LogisticRegression(),
    cv=’prefit’
)

stacking_model.fit(X_train, y_train)

# Step 7: Evaluate Models
models = {
    ’LightGBM’: lgb_model,
    ’XGBoost’: xgb_model,
    ’Random Forest’: rf_model,
    ’Stacking’: stacking_model
}

for name, model in models.items():
    y_pred = model.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    print(f"{name} Model Accuracy: {accuracy:.4f}")
    print(classification_report(y_test, y_pred))
    cm = confusion_matrix(y_test, y_pred)
    sns.heatmap(cm, annot=True, fmt='d')
    plt.xlabel('Predicted')
    plt.ylabel('Actual')
    plt.title(f'{name} Confusion Matrix')
    plt.show()



## link to the dataset
https://www.kaggle.com/datasets/anthonytherrien/depression-dataset
```

Bibliography

[1] J. M. Boden and D. M. Fergusson. Alcohol and depression. *Addiction*, 106(5):906–914, 2009.

[2] A. Caspi, K. Sugden, T. E. Moffitt, A. Taylor, I. W. Craig, H. Harrington, J. McClay, J. Mill, J. Martin, A. Braithwaite, and R. Poulton. Influence of life stress on depression: Moderation by a polymorphism in the 5-htt gene. *Science*, 301(5631):386–389, 2003.

[3] D. P. Chapman, C. L. Whitfield, V. J. Felitti, S. R. Dube, V. J. Edwards, and R. F. Anda. Adverse childhood experiences and the risk of depressive disorders in adulthood. *Journal of Affective Disorders*, 82(2):217–225, 2004.

[4] M. E. Hughes and L. J. Waite. The American family as a context for healthy aging. *Journal of Social Issues*.

[5] F. N. Jacka, A. O'Neil, R. Opie, C. Itsiopoulos, S. Cotton, M. Mohebbi, D. Castle, S. Dash, C. Mihalopoulos, M. L. Chatterton, L. Brazionis, O. M. Dean, A. M. Hodge, and M. Berk. A randomized controlled trial of dietary improvement for adults with major depression (the 'smiles' trial). *BMC Medicine*, 15:23, 2017.

[6] L. Jones and et al. Socioeconomic determinants of mental health: Implications for depression. *International Journal of Mental Health*, 2021.

[7] K. S. Kendler, L. M. Karkowski, and C. A. Prescott. Causal relationship between stressful life events and the onset of major depression. *American Journal of Psychiatry*, 158(4):587–593, 2001.

[8] R. C. Kessler, P. Berglund, O. Demler, R. Jin, K. R. Merikangas, and E. E. Walters. Lifetime prevalence and age-of-onset distributions of dsm-iv disorders in the

national comorbidity survey replication. *Archives of General Psychiatry*, 62(6):593–602, 2005.

[9] J. S. Lai, S. Hiles, A. Bisquera, A. J. Hure, M. McEvoy, and J. Attia. A systematic review and meta-analysis of dietary patterns and depression in community-dwelling adults. *American Journal of Clinical Nutrition*, 99(1):181–197, 2014.

[10] V. Lorant, D. Deliege, W. Eaton, A. Robert, P. Philippot, and M. Ansseau. Socioeconomic inequalities in depression: A meta-analysis. *American Journal of Epidemiology*, 157(2):98–112, 2003.

[11] M. Lucas, P. Chocano-Bedoya, M. B. Shulze, F. Mirzaei, E. J. O'Reilly, O. I. Okereke, W. C. Willett, and A. Ascherio. Inflammatory dietary pattern and risk of depression among women. *Brain, Behavior, and Immunity*, 25(7):1144–1150, 2011.

[12] B. Mezuk, W. W. Eaton, S. Albrecht, and S. H. Golden. Depression and type 2 diabetes over the lifespan: A meta-analysis. *Diabetes Care*, 31(12):2383–2390, 2008.

[13] K. Mikkelsen, L. Stojanovska, M. Polenakovic, M. Bosevski, and V. Apostolopoulos. Exercise and mental health. *Maturitas*, 106:48–56, 2017.

[14] S. Moussavi, S. Chatterji, E. Verdes, A. Tandon, V. Patel, and B. Ustun. Depression, chronic diseases, and decrements in health: Results from the world health surveys. *The Lancet*, 370(9590):851–858, 2007.

[15] World Health Organization. *Depression and Other Common Mental Disorders: Global Health Estimates*. World Health Organization, 2017.

[16] World Health Organization. *Depression: Fact Sheets*. World Health Organization, 2021.

[17] F. B. Schuch and et al. Exercise as a treatment for depression: A meta-analysis adjusting for publication bias. *Journal of Psychiatric Research*, 103:42–51, 2018.

[18] J. Smith and A. Doe. The genetics of depression: A review. *Journal of Psychiatric Research*, 2019.

[19] P. F. Sullivan, M. C. Neale, and K. S. Kendler. Genetic epidemiology of major depression: Review and meta-analysis. *American Journal of Psychiatry*, 157(10):1552–1562, 2000.

[20] G. Taylor, A. McNeill, A. Girling, A. Farley, N. Lindson-Hawley, and P. Aveyard. Change in mental health after smoking cessation: Systematic review and meta-analysis. *BMJ*, 348:g1151, 2014.

[21] M. P. Walker. The role of sleep in cognition and emotion. *Annals of the New York Academy of Sciences*, 1406(1):1–23, 2017.

[22] M. M. Weissman, P. Wickramaratne, K. R. Merikangas, J. F. Leckman, B. A. Prusoff, K. A. Caruso, K. K. Kidd, and D. L. Pauls. Onset of major depression in early adulthood: Increased familial loading and rates of illness in the relatives of early-onset cases. *Archives of General Psychiatry*, 44(10):848–853, 1987.