Project Report

Predicting Heart Disease Using Personal Health and Lifestyle Indicators

1. Introduction

Cardiovascular disease remains one of the leading causes of mortality globally. Early identification of individuals at risk of developing heart disease can significantly improve health outcomes through timely interventions and preventive care.
The primary objective of this study is to explore whether an individual's personal health and lifestyle characteristics can be used to predict their likelihood of developing heart disease.

The research question guiding this project is:

Can we accurately predict whether an individual is at risk of developing heart disease based on their personal health and lifestyle indicators?

The goal is to develop a predictive model that identifies individuals at high risk of heart disease, allowing healthcare systems to focus on prevention and awareness efforts.

2. Data Source and Quality Assessment

The dataset used for this study is "Personal Key Indicators of Heart Disease" (heart_2020_cleaned.csv), obtained from Kaggle, originally sourced from the CDC's Behavioral Risk Factor Surveillance System (BRFSS).
It consists of 319,795 survey responses from adults across the United States. Each record captures a combination of health, lifestyle, and demographic indicators such as:

- Health factors: BMI, PhysicalHealth, MentalHealth, SleepTime
- Lifestyle factors: Smoking, AlcoholDrinking, PhysicalActivity
- Demographics: Sex, AgeCategory, Race
- Chronic conditions: Diabetes, KidneyDisease, Asthma, etc.

The target variable is HeartDisease (Yes/No).

Reliability and Validity:

The BRFSS dataset is highly reliable and valid due to its large, representative sample size and standardized data collection procedures.

However, since it relies on self-reported data, potential limitations include:

- Recall bias or underreporting of health behaviors
- Misinterpretation of survey questions
- Class imbalance (only 8.6% reported heart disease), which may affect model performance

Despite these caveats, the dataset remains a credible and comprehensive resource for population-level health analysis.

3. Type of Machine Learning Problem

This is a supervised machine learning problem, where the goal is to predict a binary outcome — whether an individual has heart disease (“Yes”) or not (“No”).
Hence, it is a binary classification task.

Two models were considered:

1. **Logistic Regression –** chosen for its simplicity, interpretability, and probabilistic output.
2. **Decision Tree Classifier –** selected for its ability to model non-linear relationships and provide clear, rule-based decision paths.

These models enable both predictive insights and interpretability into which features most strongly influence heart disease risk.

4. Data Understanding and Exploration

The dataset contains 18 columns with 4 numerical and 14 categorical variables.
There were no missing values, simplifying data preparation.

Numerical Variables:

- BMI: Body Mass Index
- PhysicalHealth: Number of poor physical health days in the past 30 days
- MentalHealth: Number of poor mental health days in the past 30 days
- SleepTime: Average hours of sleep per day

Categorical Variables:

Examples include Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, AgeCategory, Race, GenHealth, Asthma, KidneyDisease, and SkinCancer.

Summary Statistics

The summary statistics provide an overview of key health-related variables in the dataset and reveal several important patterns.

The average Body Mass Index (BMI) is approximately 28.3, indicating that overweight and obesity are common within this population. This aligns with broader public health trends observed globally. The standard deviation of 6.36 and a wide range (12 to 94) demonstrate considerable variability in BMI values, suggesting that the dataset includes individuals from underweight to severely obese categories. This diversity in BMI levels is valuable for understanding how different body compositions relate to heart disease risk.

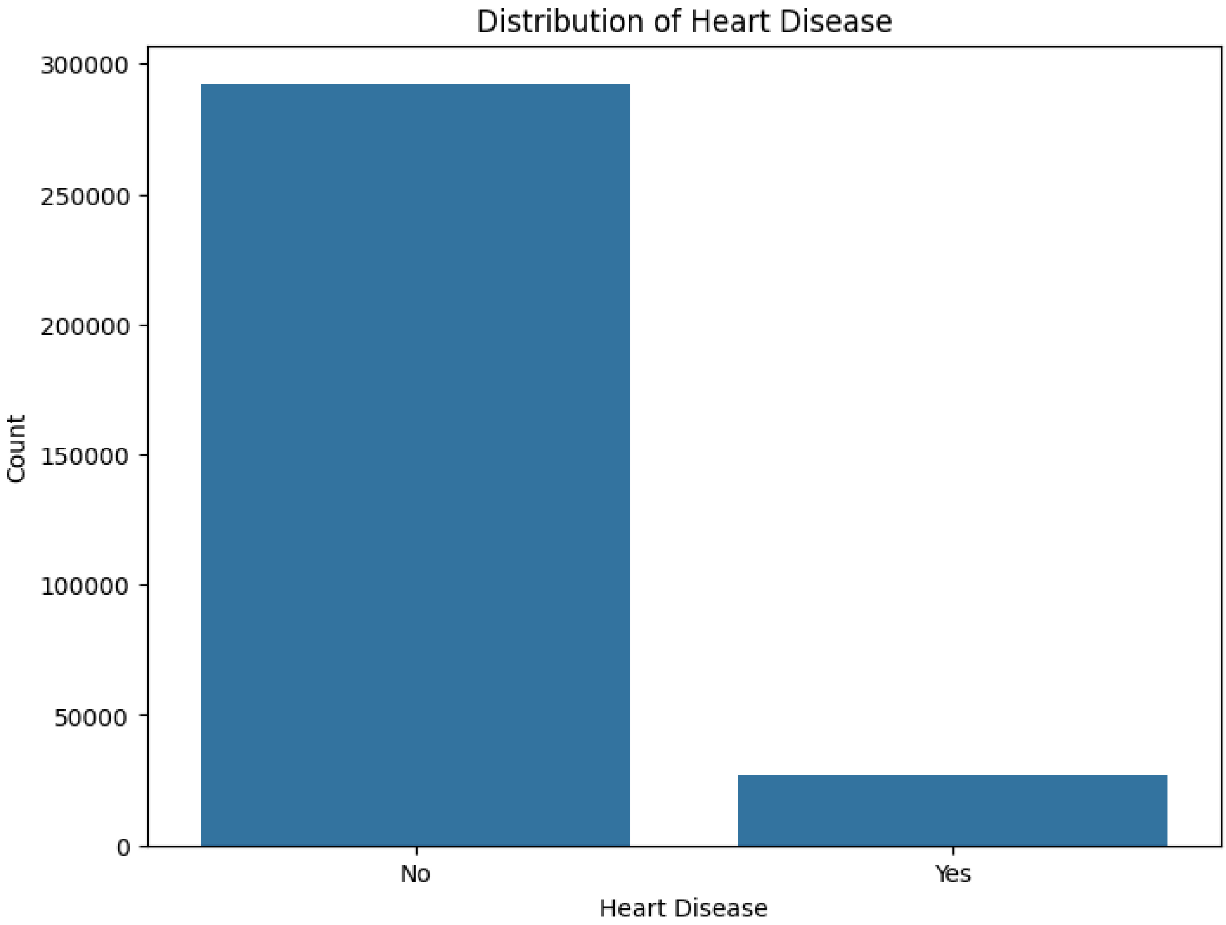
For both PhysicalHealth and MentalHealth, the median value of 0 indicates that most respondents reported no poor health days in the past month. However, the relatively high standard deviation (approximately 7.95) and the maximum of 30 days highlight that a subset of the population experiences consistent physical or mental health challenges. These variables exhibit right-skewed distributions, with the majority appearing healthy but a small portion reporting significant health concerns. Such extremes may represent individuals with chronic illnesses or ongoing mental health struggles. However, these results should be interpreted cautiously as self-reported data can sometimes reflect subjective perceptions or reporting biases.

The SleepTime variable shows greater consistency, with an average of around 7 hours and a standard deviation of about 1.4. This pattern aligns closely with recommended sleep guidelines for adults. Nevertheless, extreme values such as 1 hour or 24 hours may indicate potential data entry errors, misreporting (such as recording time spent in bed instead of actual sleep duration), or rare cases of sleep disorders.

Overall, the summary statistics depict realistic health patterns across the population while also revealing skewness and outliers in several variables. These irregularities may highlight meaningful subgroups, such as individuals with chronic conditions, but they also underscore the need for data cleaning and careful interpretation to ensure the accuracy and reliability of further analyses.

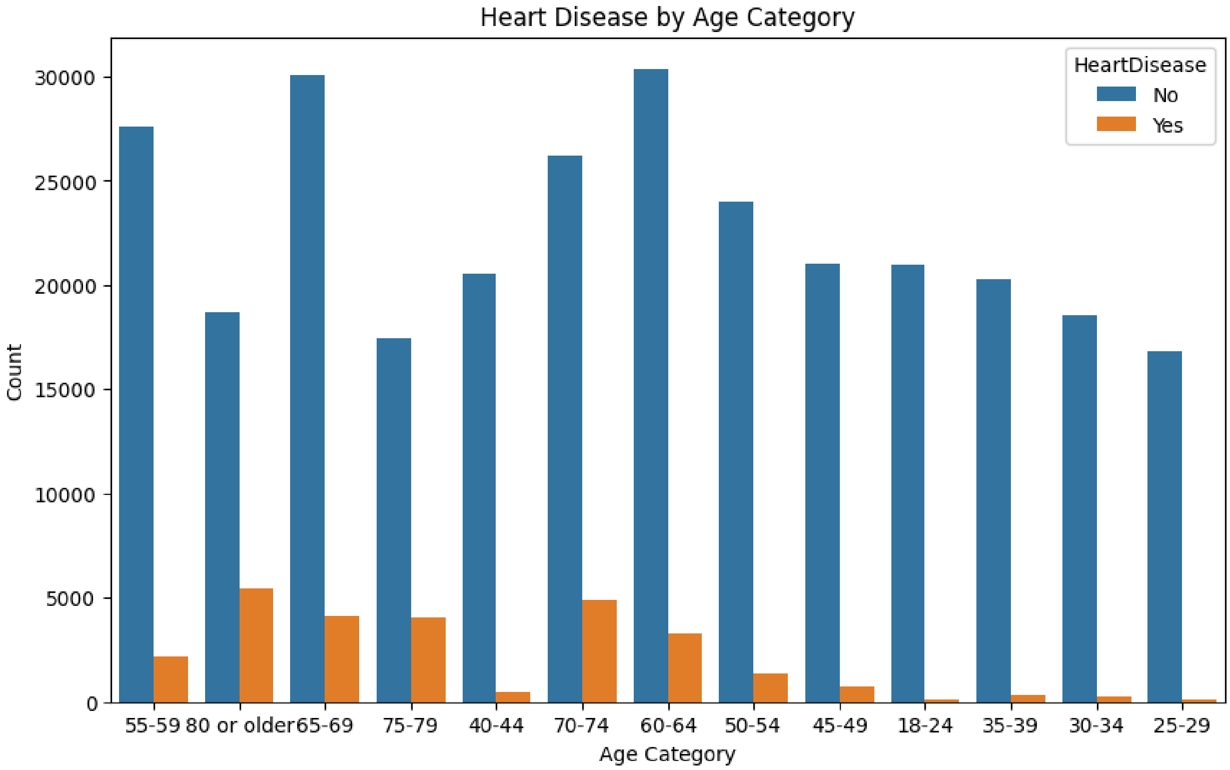
5. Key Findings from Exploratory Data Analysis

5.1 Distribution of Heart Disease



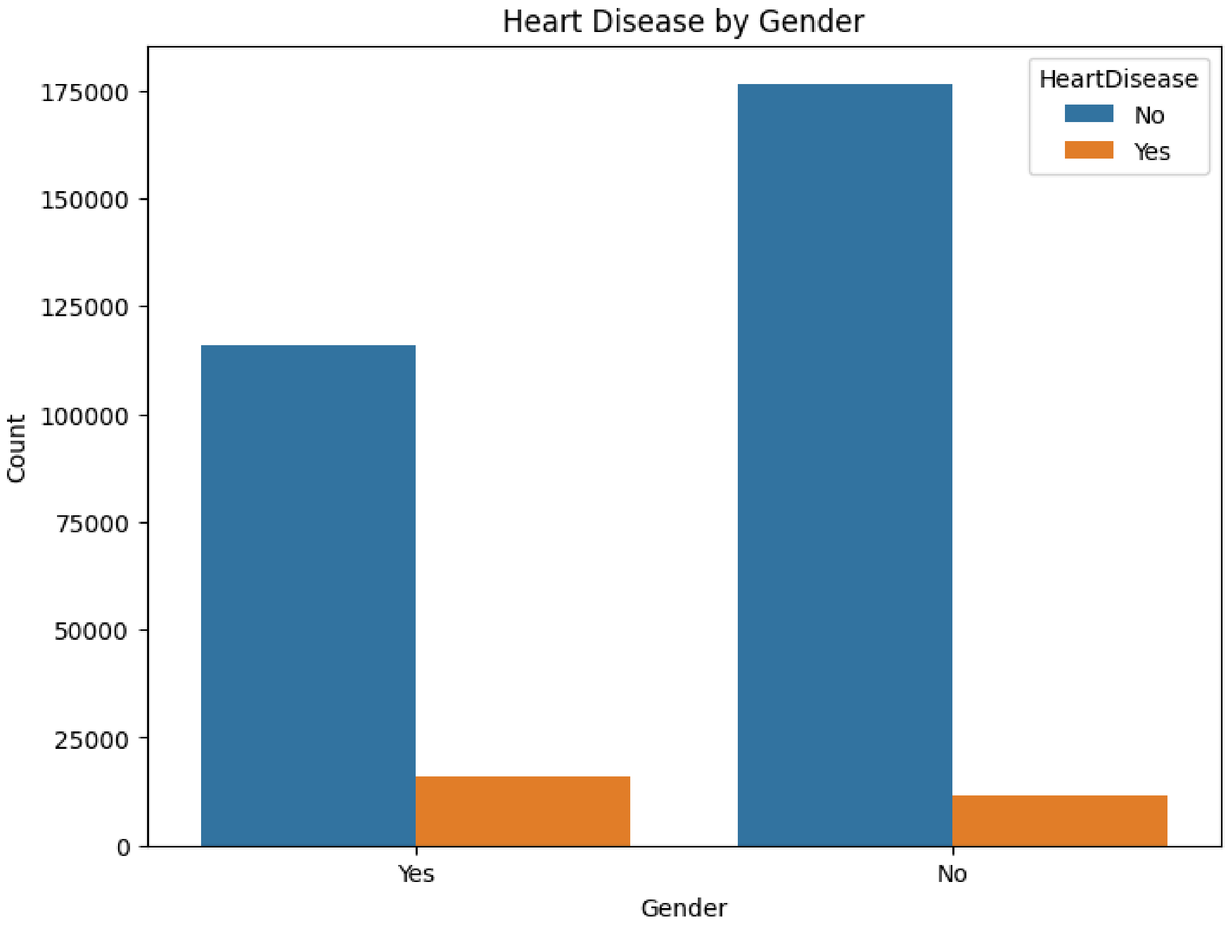
- The count plot shows the proportion of individuals with and without heart disease.
- A large imbalance is observed — the number of people without heart disease is much higher than those with it.
- This indicates class imbalance, which is an important factor to consider before model training.
- It also helps understand the prevalence of heart disease in the dataset.

5.2 Heart Disease by Age Category



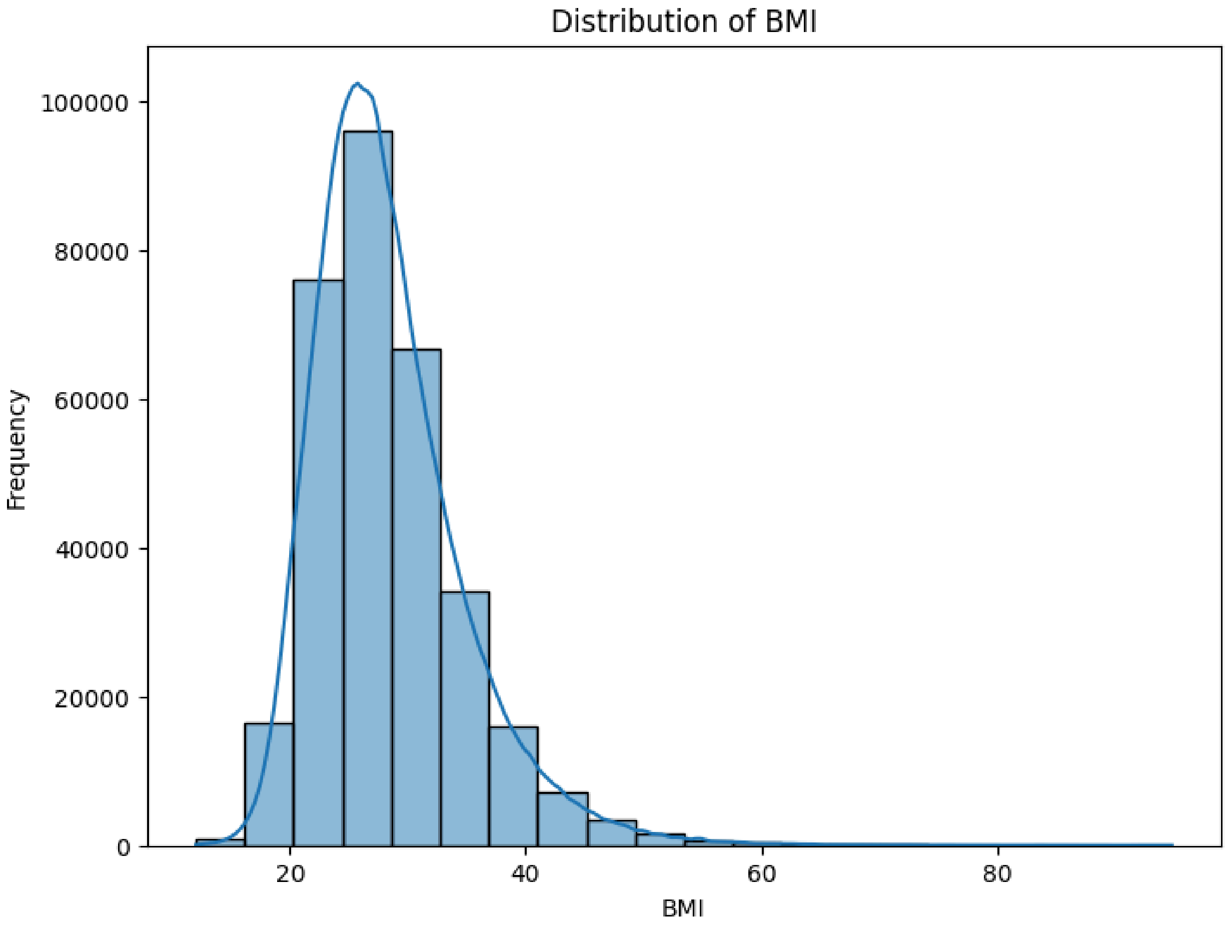
- The count plot compares the occurrence of heart disease across different age groups.
- Older age categories show noticeably higher counts of heart disease cases.
- Younger groups (like 18–24, 25–29) show very few cases, whereas older groups (65–69, 70–74, etc.) have more.
- This reveals that age is a significant risk factor for heart disease.

5.3 Heart Disease by Gender



- The visualization shows the distribution of heart disease across genders.
- Males tend to show a slightly higher proportion of heart disease compared to females.
- This aligns with medical studies suggesting men are at higher risk for heart disease at earlier ages.
- However, the difference may vary based on lifestyle or other contributing factors.

5.4 Distribution of BMI



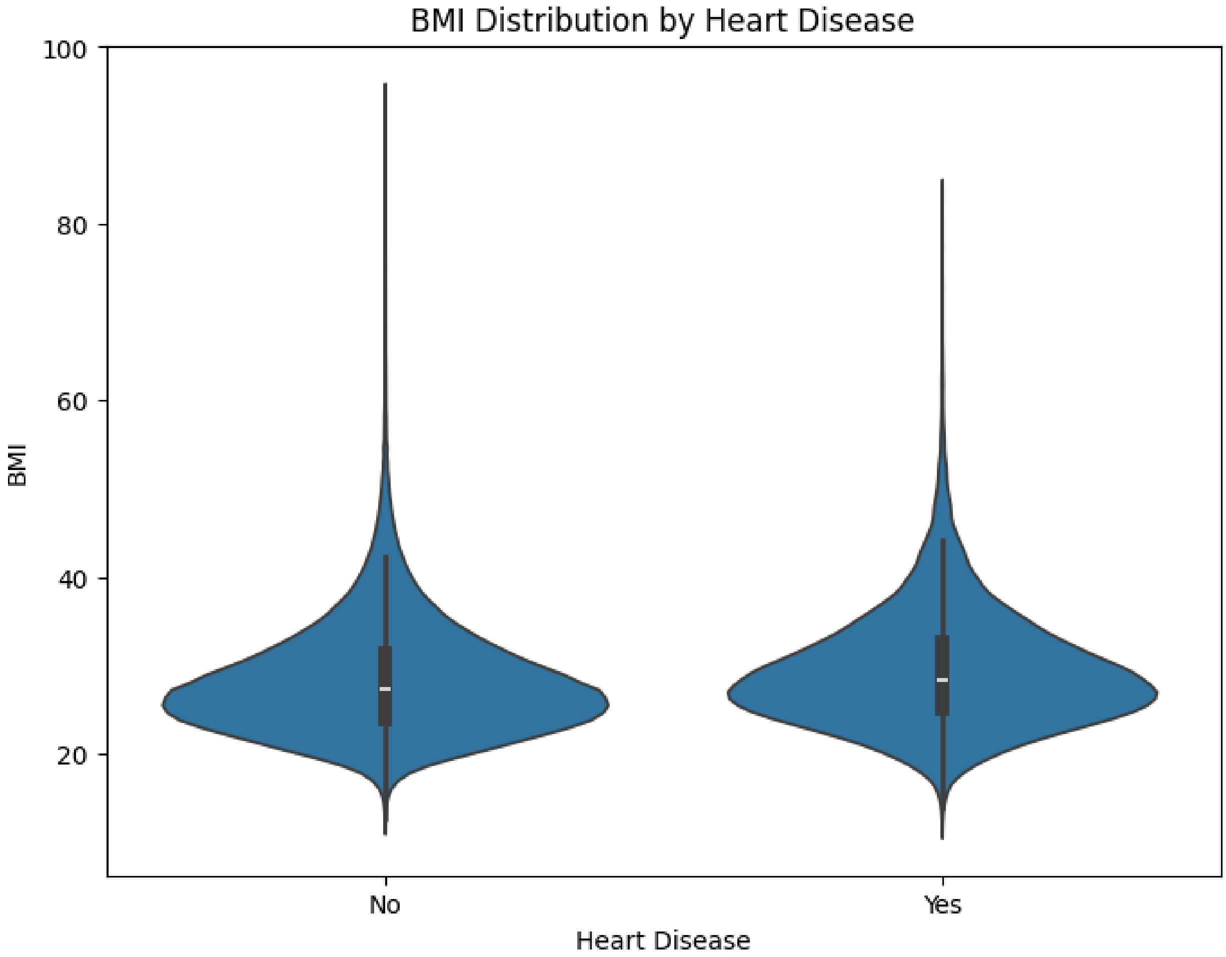
- The histogram illustrates how BMI values are distributed among individuals.

- Most values cluster between 20–35, indicating that a majority of the participants fall into the normal to overweight range.

- There are few extreme BMI values, suggesting limited outliers.

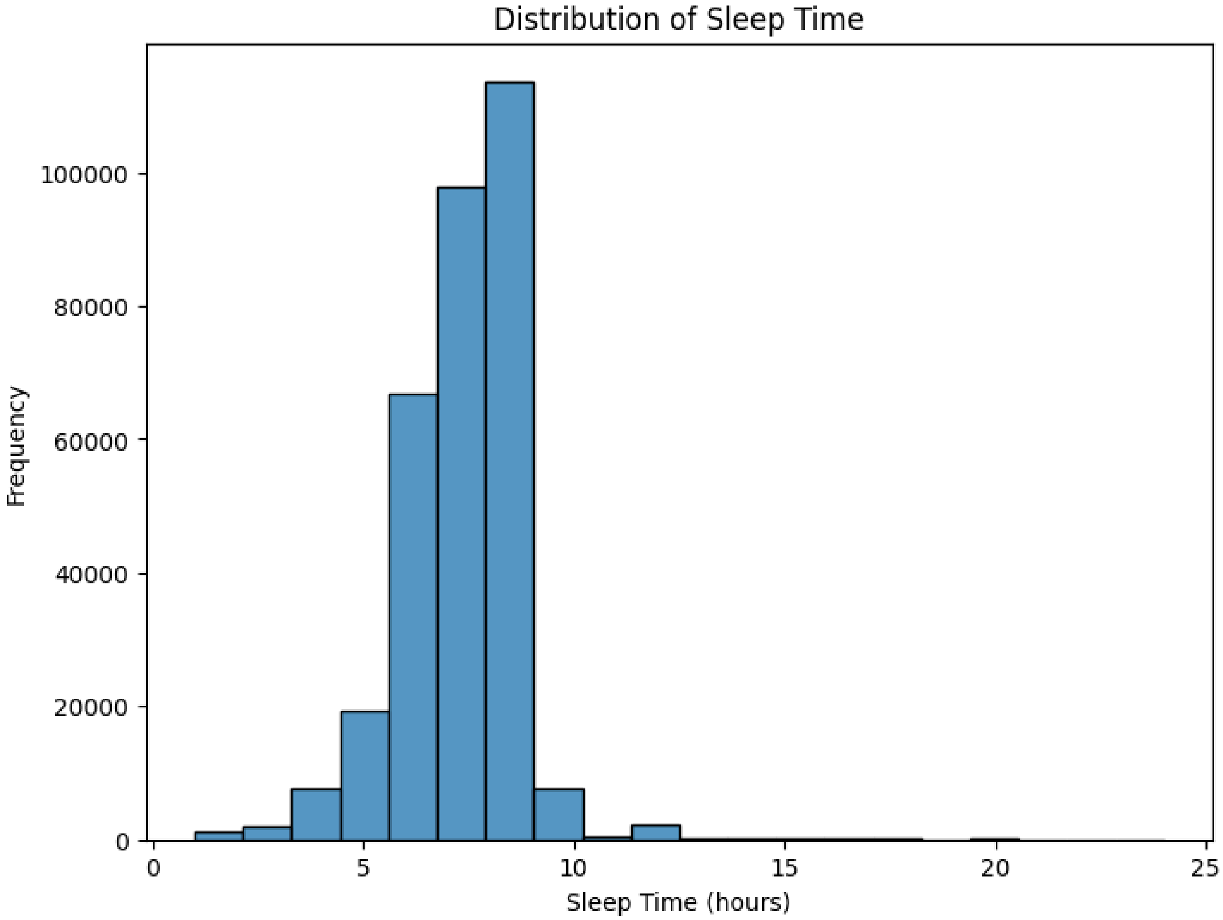
- The distribution helps check for skewness and identify potential health patterns in the data.

5.5 BMI Distribution by Heart Disease



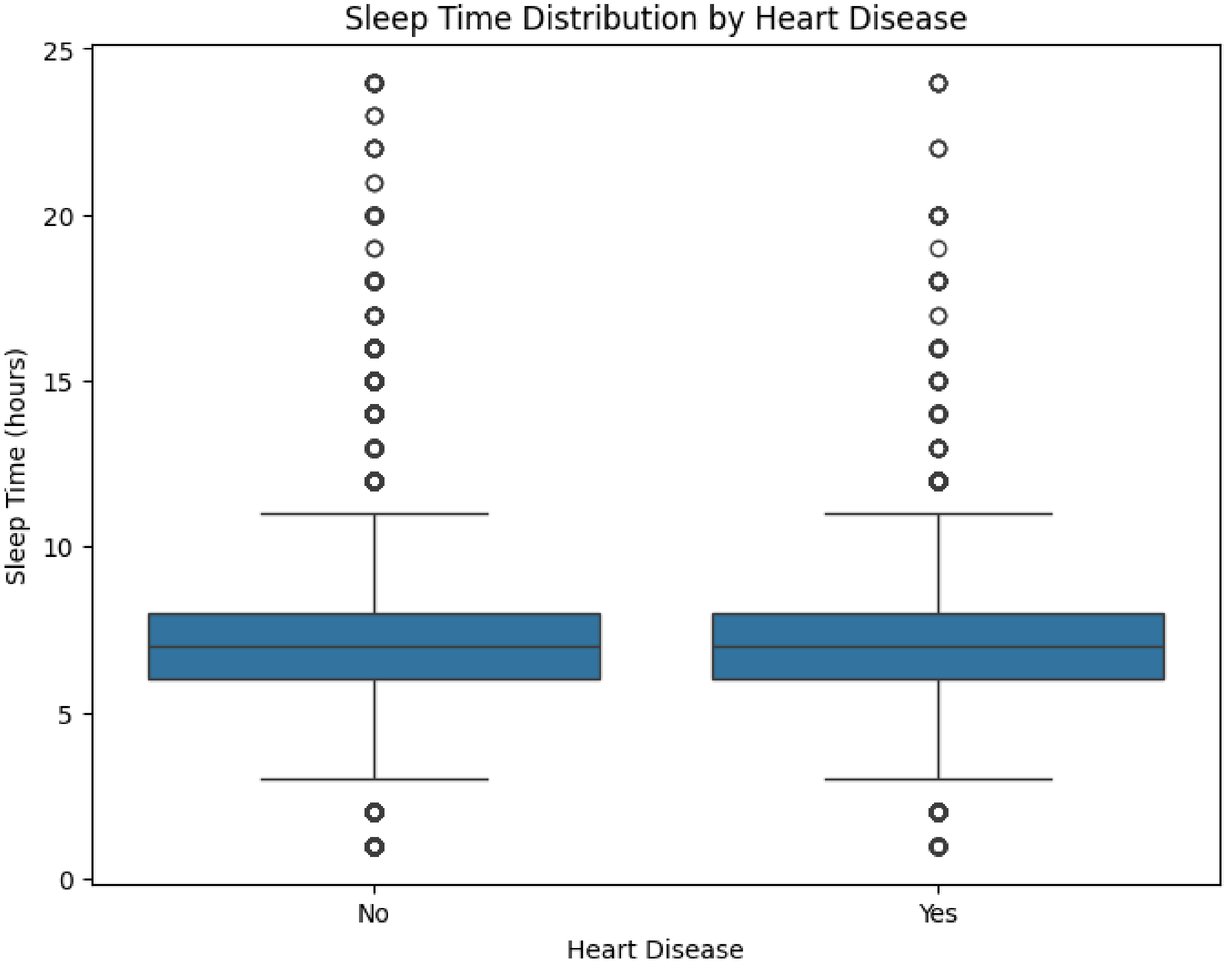
- The violin plot compares BMI values for people with and without heart disease.
- Individuals with heart disease tend to have slightly higher median BMI.
- The spread (interquartile range) appears similar for both groups, suggesting that although BMI plays a role, it’s not the sole determinant.
- This indicates that overweight and obesity could contribute to heart disease risk.

5.6 Distribution of Sleep Time



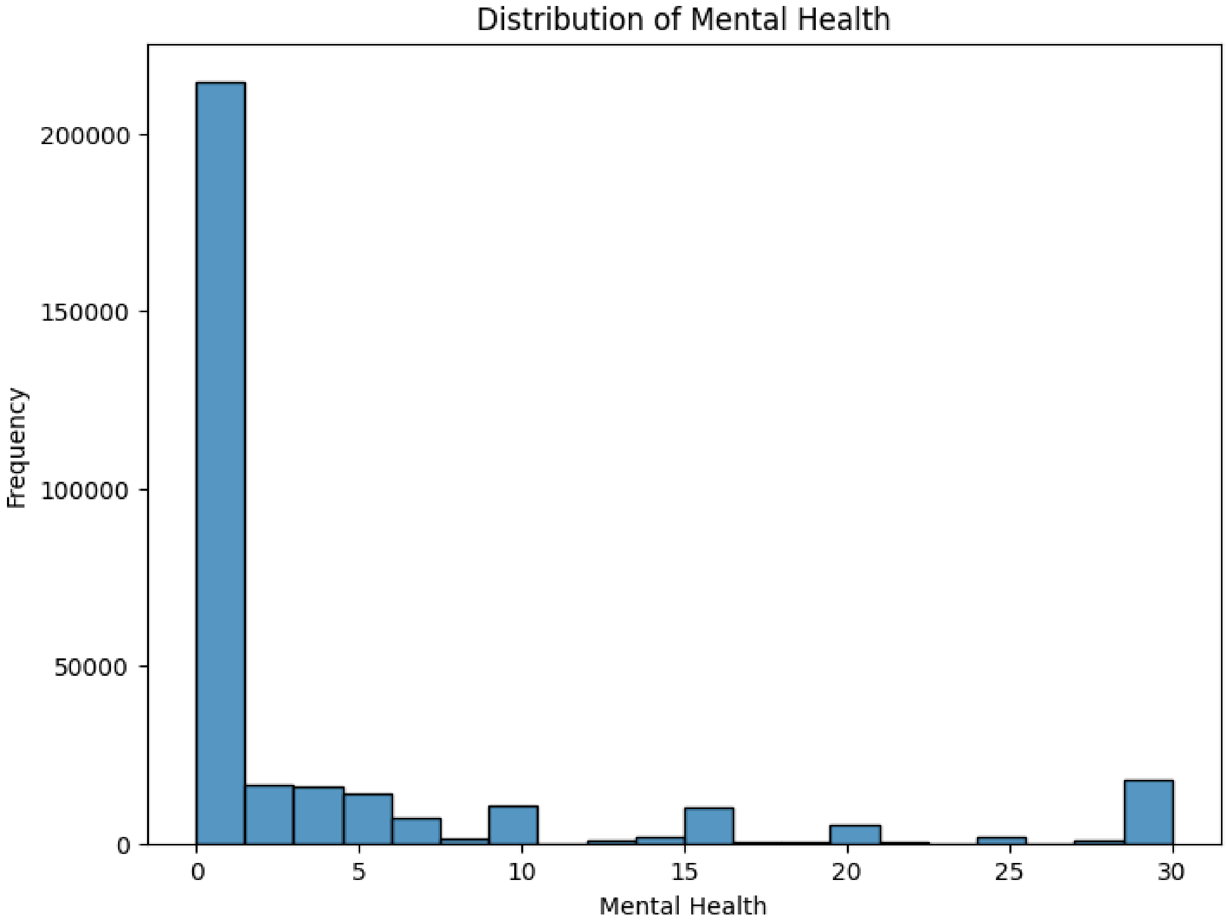
- The histogram shows how many hours individuals sleep per night.
- Most people report sleeping between 6–8 hours, which is within the recommended range.
- Few individuals report sleeping very little (<4 hours) or excessively (>10 hours).
- This helps understand lifestyle factors potentially linked to heart disease.

5.7 Sleep Time Distribution by Heart Disease



- The box plot compares sleep duration for individuals with and without heart disease.
- Those with heart disease appear to have a slightly lower median sleep time.
- The variability in sleep time is wider for people without heart disease.
- Poor or inadequate sleep could be associated with heart health risks.

5.8 Distribution of Mental Health



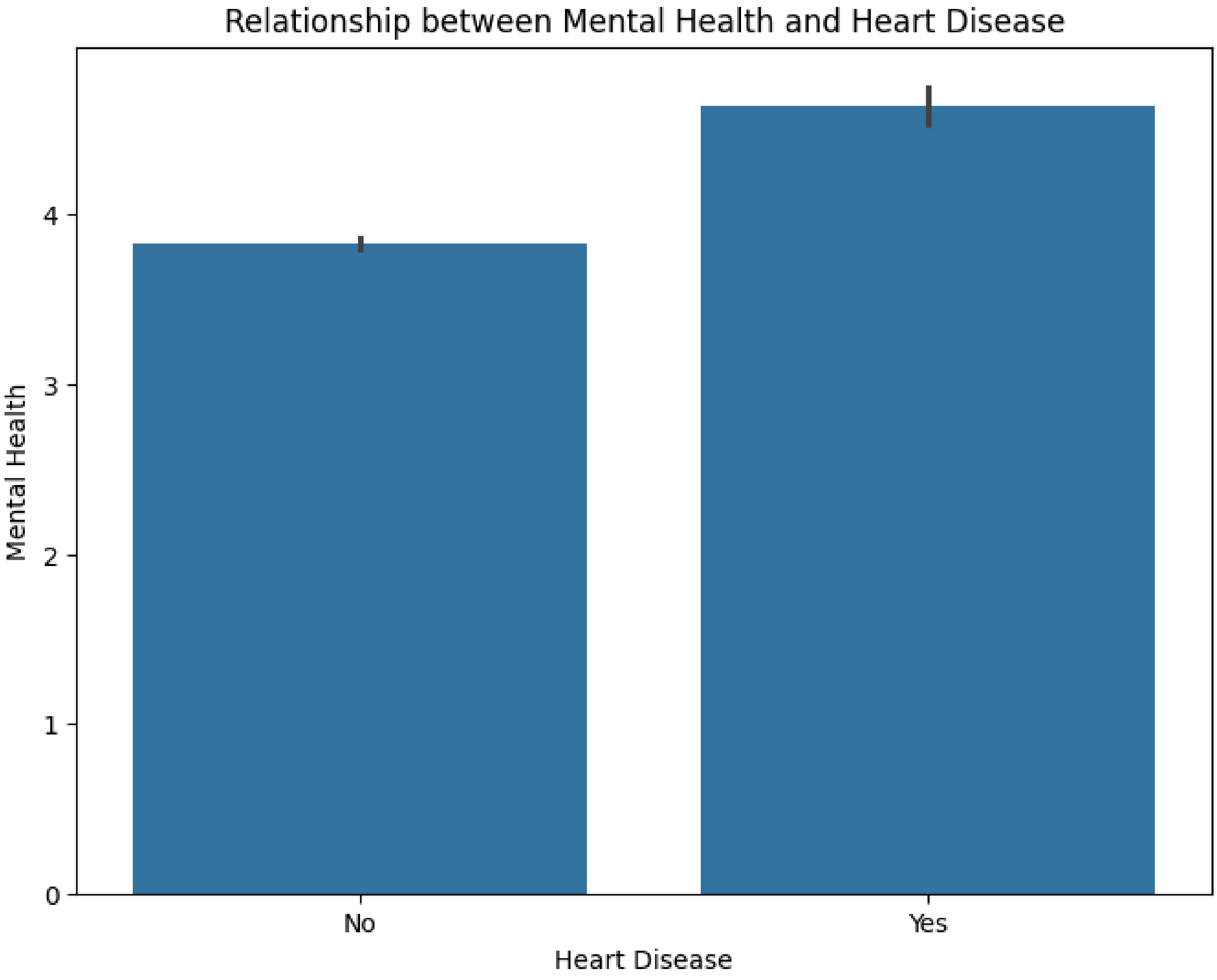
- The histogram shows how many days individuals reported poor mental health in the past month.

- Most people reported 0–5 days, meaning generally good mental health in the population.

- However, there are noticeable counts at higher values, indicating a subset facing significant mental health challenges.

- The plot helps identify population well-being patterns.

5.9 Relationship between Mental Health and Heart Disease



- The bar plot depicts the average number of poor mental health days for individuals with and without heart disease.
- Those with heart disease report higher average poor mental health days, suggesting a strong connection between emotional well-being and cardiovascular conditions.
- This emphasizes the mind–heart link, indicating that mental stress may contribute to heart problems.

Key Interpretations

The Distribution of Heart Disease graph highlights a clear imbalance, where a small fraction of individuals have heart disease compared to the healthy population. This finding emphasizes the importance of handling class imbalance while building predictive models.

The Heart Disease by Age Category chart demonstrates a clear upward trend, heart disease prevalence increases sharply with age. This supports existing medical knowledge that aging is a key risk factor for cardiovascular issues, making age an important predictive variable in the dataset.

In the Heart Disease by Gender visualization, a slightly higher proportion of males with heart disease can be observed. This aligns with clinical evidence that men are more prone to heart disease at earlier ages than women, potentially due to biological and lifestyle factors.

The Distribution of BMI plot shows that most individuals have BMI values in the normal to moderately overweight range, with few outliers. When combined with the BMI Distribution by Heart Disease violin plot, it becomes evident that individuals with higher BMI tend to have more heart disease cases, reaffirming the link between body weight and cardiovascular risk.

The Distribution of Sleep Time and Sleep Time by Heart Disease visualizations reveal that shorter sleep duration might correlate with higher heart disease risk. Adequate rest seems to be a protective factor, while both insufficient and excessive sleep can indicate underlying health issues.

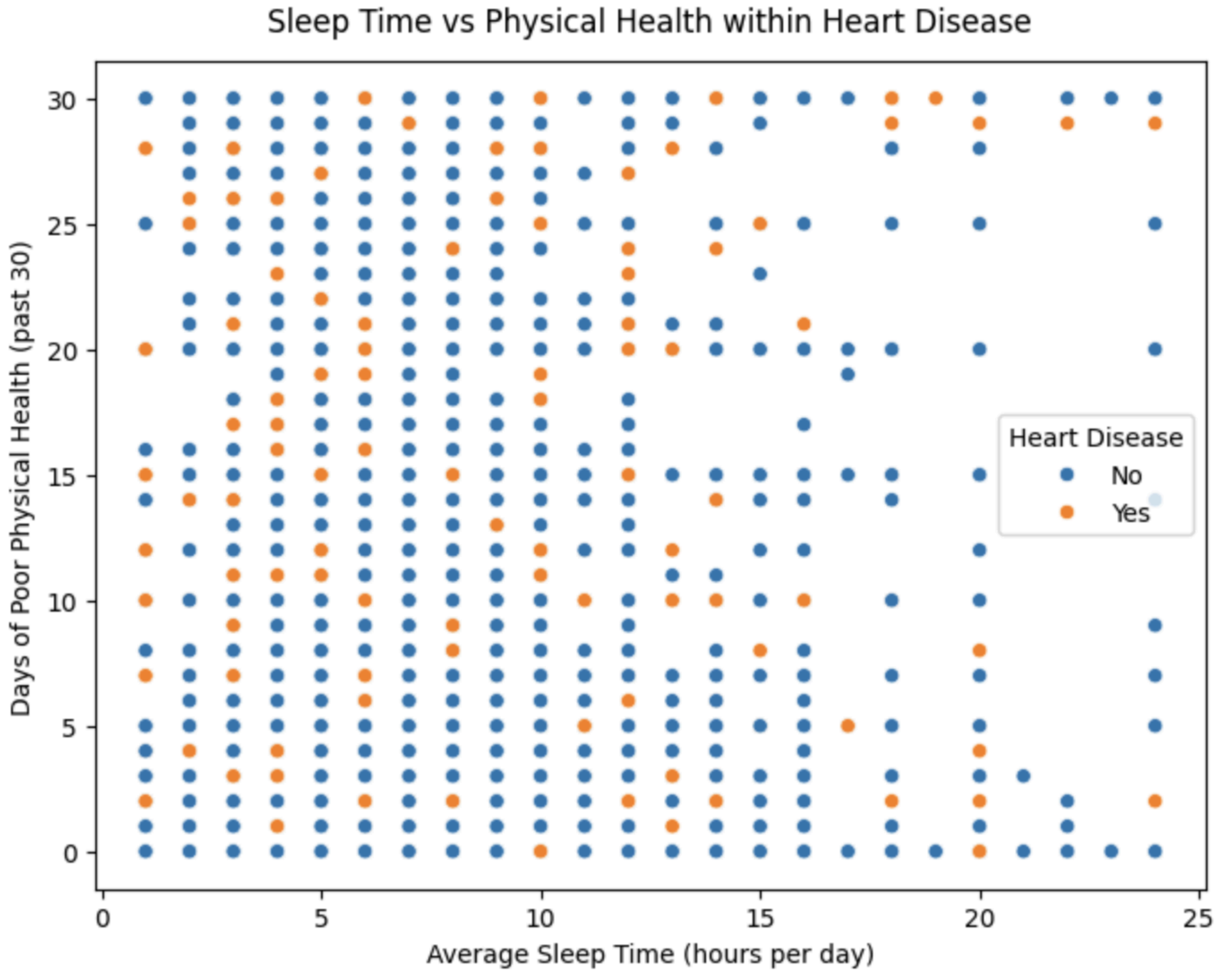
Finally, the Mental Health plots uncover an interesting relationship — individuals with heart disease experience more poor mental health days on average. This suggests that psychological stress, anxiety, or depression could be associated with heart conditions, highlighting the holistic connection between physical and mental health.

VISUALIZING RELATIONSHIP BETWEEN KEY VARIABLES:

5.10 Sleep time vsPhysical health

In the previous section, we examined visualizations of key individual variables such as BMI, Sleep Time, Physical Health, and Mental Health to understand their overall distributions.
Now, we shift our focus to exploring relationships between these variables, especially in the context of Heart Disease.
The following scatter plots allow us to see how lifestyle and health indicators interact specifically, how sleep duration relates to physical health, and how BMI relates to mental health, while distinguishing individuals with and without heart disease.

Here, we're looking at the relationship between average hours of sleep per day and days of poor physical health in the past month, with each point representing an individual.



The color coding separates people with heart disease (orange) from those without (blue).

The data shows a very weak or no visible linear trend between sleep duration and poor physical health. Most individuals, regardless of heart disease status, report varied physical health across all sleep durations, though a slight clustering appears around 6 to 9 hours of sleep, which aligns with typical sleep recommendations.

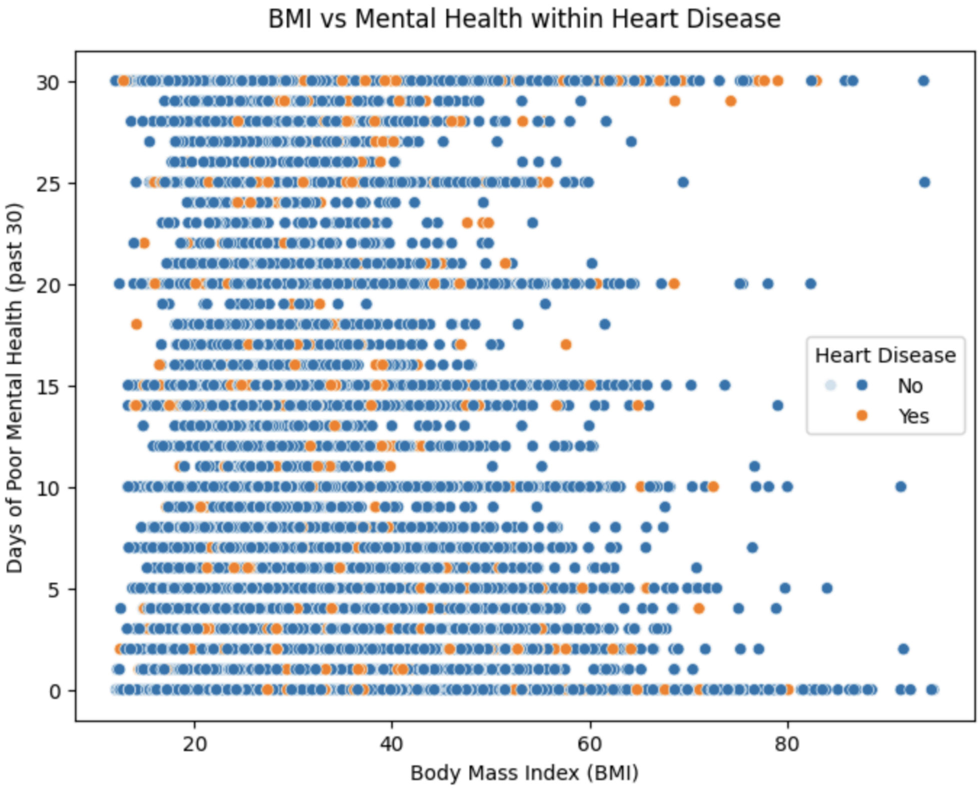
People with heart disease are distributed similarly to those without heart disease, though they appear slightly denser among individuals who report more poor-health days, suggesting that poorer physical health may coincide with a higher likelihood of heart disease.

Overall, the plot indicates that while sleep time alone is not a strong predictor of physical health, those with heart disease may experience more frequent days of poor health, independent of sleep duration.

Component	Description
Data	Derived from the Heart Disease 2020 dataset, including numeric variables SleepTime and PhysicalHealth, and categorical variable HeartDisease.
Graph	A scatter plot visualizing the bivariate relationship between sleep time (x-axis) and poor physical health days (y-axis).
Label	Axes are clearly labeled with descriptive units (“Average Sleep Time (hours per day)” and “Days of Poor Physical Health (past 30)”), and the title precisely states the comparison context.
Aesthetic	Distinct hues for HeartDisease status enhance clarity (blue = No, orange = Yes). Point transparency reduces overplotting, ensuring readability.
Ethical	The visualization maintains objectivity—no misleading color associations or exaggeration of relationships. The data are anonymized and used solely for academic analysis of health trends.

5.11 BMI vs Mental health

In this plot, we examine how Body Mass Index (BMI) relates to days of poor mental health, again distinguishing individuals by heart disease status.



What stands out is the wide scatter of points, there’s no clear relationship between BMI and mental health. People with both low and high BMI values report a wide range of mental health experiences.

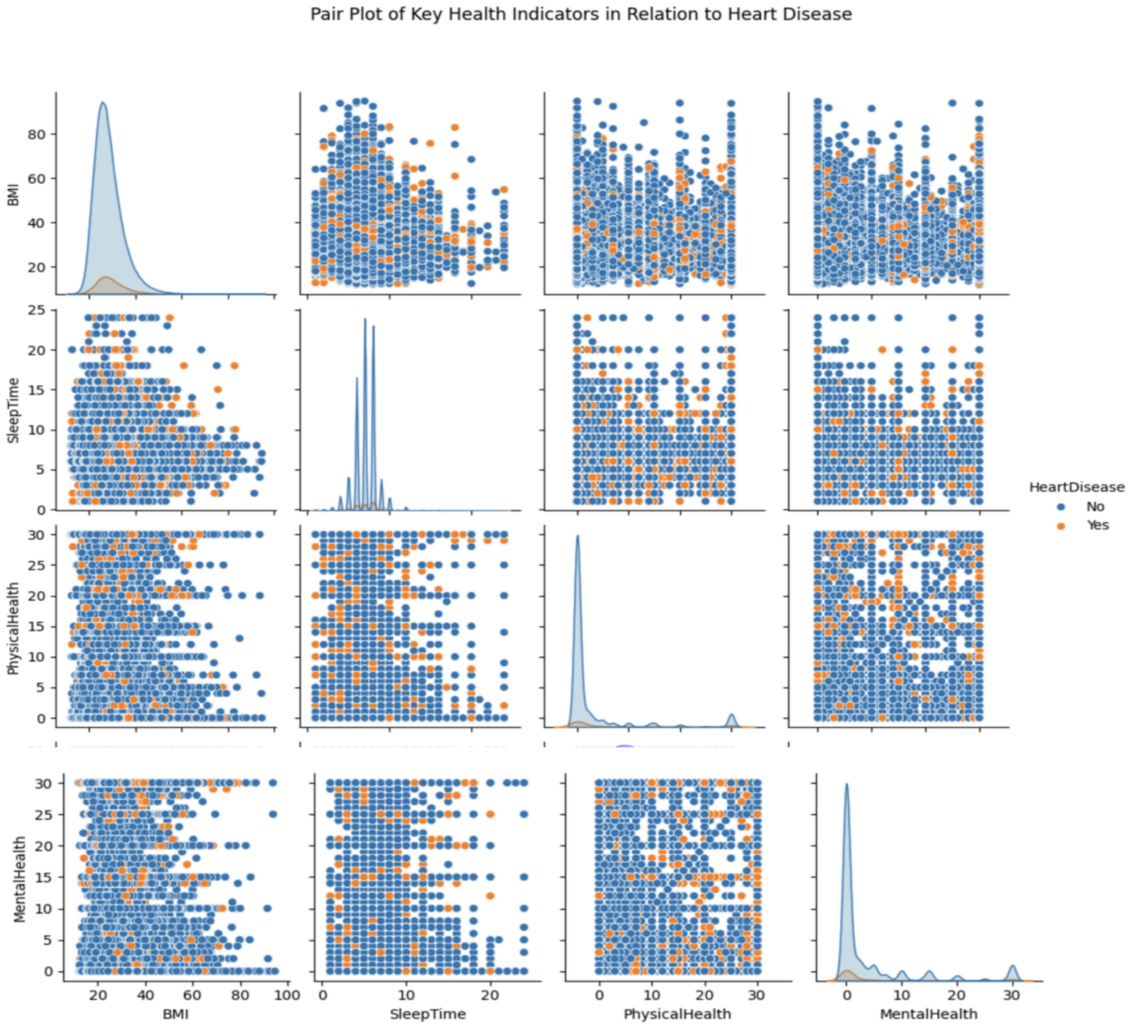
Although individuals with heart disease appear slightly more common in higher BMI ranges, the relationship between BMI and mental health remains weak and non-linear. This suggests that mental well-being is influenced by more complex factors such as stress, lifestyle, and socioeconomic conditions beyond body weight alone.

In summary, while higher BMI may coincide with a slightly increased presence of heart disease, BMI is not a reliable predictor of mental health status in this dataset.

Component	**Description**
Data	Subset from the Heart Disease 2020 dataset, including BMI, MentalHealth and HeartDisease
Graph	Scatter plot visualizing BMI (x-axis) against days of poor mental health (y-axis), categorized by heart disease presence.
Label	Titles, axes, and legends are clear and descriptive ensuring that the message is interpretable even without technical expertise.
Aesthetic	The use of Seaborn’s consistent palette ensures accessibility, while semi-transparent points manage dense data effectively.
Ethical	The visualization avoids stigmatizing weight or mental health conditions. It presents observed correlations responsibly, without suggesting direct medical causation.

5.12 PAIR PLOT

Now that we've explored the relationships between individual variables, let's take a step back and look at the broader picture using a *pair plot*.
This visualization helps us understand how four key health indicators being BMI, Sleep Time, Physical Health, and Mental Health interact with each other, while also highlighting differences between individuals with and without heart disease.



Across the grid, each scatter plot compares two variables, and the diagonal shows the distribution (KDE curve) for each metric. The blue points represent people without heart disease, while orange points represent those diagnosed with it.

What stands out here is that there are no strong linear relationships among the variables, meaning that one factor alone does not clearly predict another. However, a few subtle trends are visible:

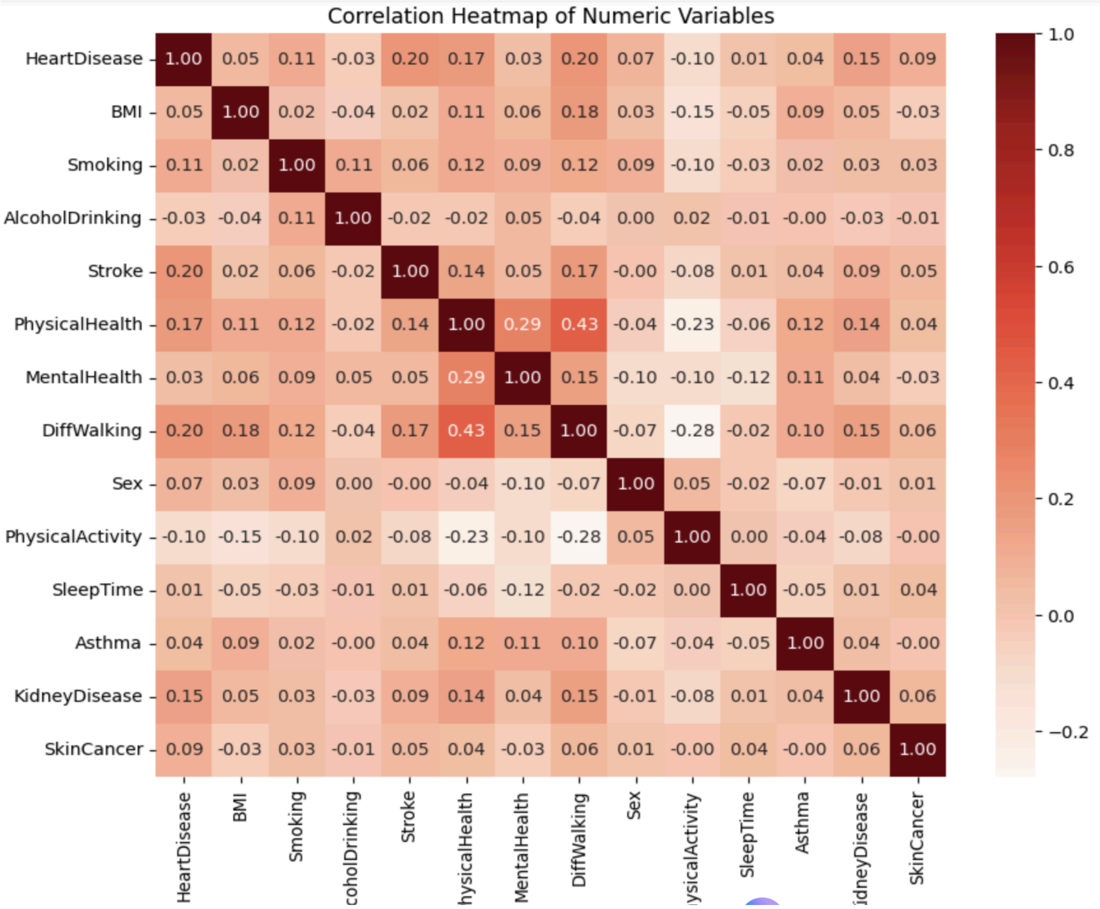
- Individuals with higher BMI or more days of poor physical and mental health appear slightly more likely to fall into the heart disease category.

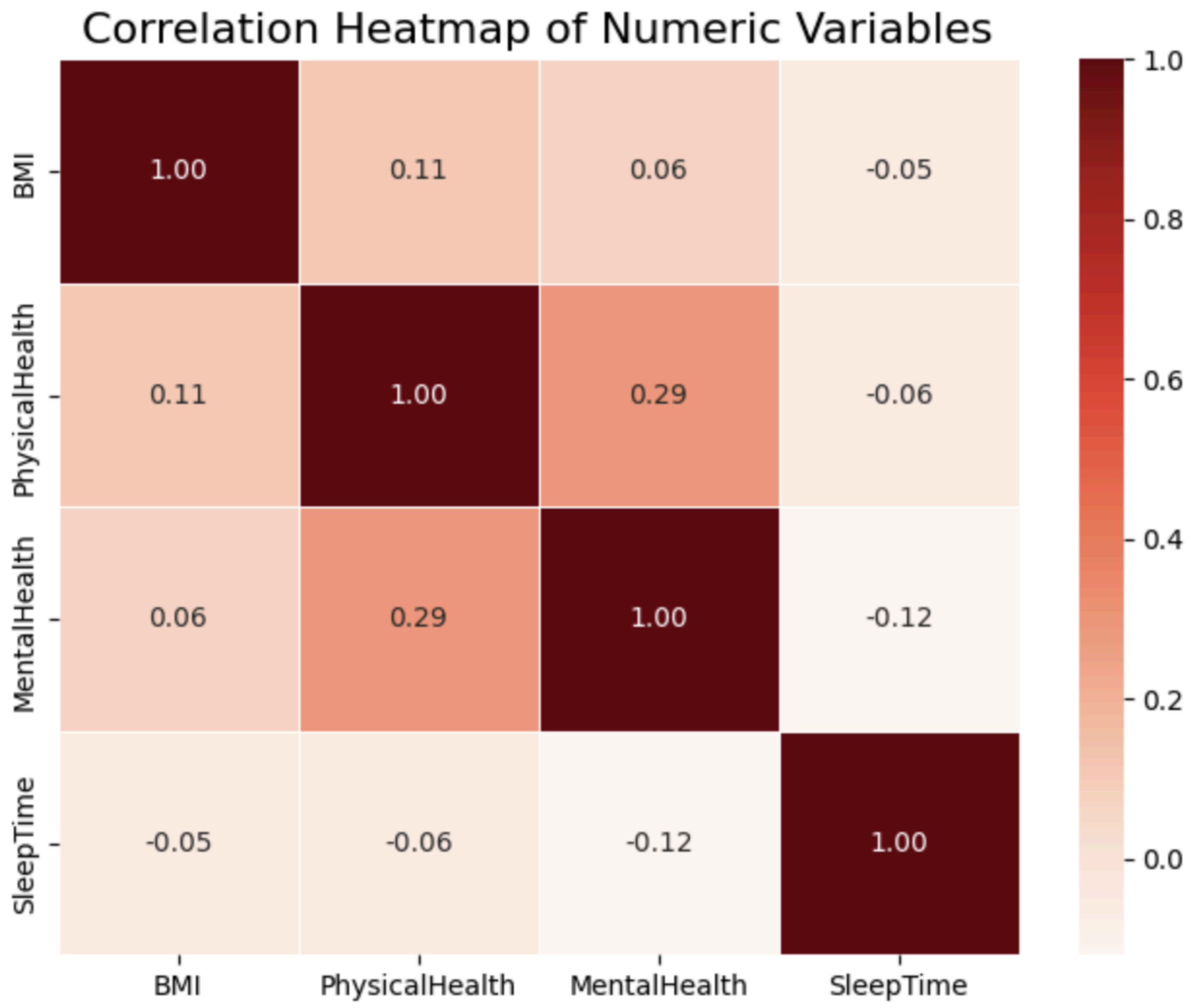
- Sleep Time is fairly consistent across both groups, with most individuals averaging between 6 and 9 hours of sleep, suggesting that sleep duration alone doesn't distinguish heart disease presence.

- The density curves show that most participants report relatively few poor-health days, though a small cluster with higher counts overlaps more with heart disease cases.

In summary, this pair plot gives us a holistic view: while these health indicators show mild interconnections, heart disease seems to emerge from the combined influence of multiple small factors rather than a single dominant one. This reinforces the importance of looking at health holistically considering physical, mental, and lifestyle aspects together rather than in isolation.

5.13 Interpretation of the Correlation Heatmap




Correlation Heatmap of Numeric Variables

In the previous section, we looked at the pair plot to understand how our key health indicators interact visually. Now, we move one step further with this correlation heatmap, which quantifies the strength and direction of relationships among all numeric variables in our dataset including health conditions, lifestyle factors, and demographic indicators.

Each cell in this matrix shows a correlation coefficient ranging from –1 to +1:

- Values closer to +1 indicate a strong positive relationship (both variables increase together),
- Values closer to –1 indicate a negative relationship (one increases while the other decreases),
- Values near 0 suggest little to no linear relationship.

1. **Heart Disease Relationships**
   The HeartDisease variable shows moderate positive correlations with a few health-related features:

   a. Stroke (r = 0.20), Difficulty Walking (r = 0.20), and PhysicalHealth (r = 0.17). This indicates that individuals who have had a stroke, report more physical health issues, or have mobility challenges are more likely to have heart disease. KidneyDisease (r = 0.15) also shows a mild correlation, which aligns with known medical patterns of comorbidity between cardiovascular and renal issues.

2. **Physical and Mental Health**
   a. The strongest relationship observed is between PhysicalHealth and Difficulty Walking (r = 0.43) which makes intuitive sense, as reduced mobility often reflects poorer overall physical well-being.
   Additionally, PhysicalHealth and MentalHealth (r = 0.29) exhibit a moderate correlation, suggesting that poor physical health tends to be accompanied by reduced mental well-being.

3. **Lifestyle Factors**

   a. BMI, Smoking, and AlcoholDrinking show weak or negligible relationships with heart disease (all below r = 0.11), indicating that while these factors may contribute to heart risk, they do not exhibit a strong *linear* relationship in this self-reported dataset.

   b. PhysicalActivity is weakly negatively correlated with both PhysicalHealth (r = –0.23) and Difficulty Walking (r = –0.28), which suggests that individuals engaging in more physical activity tend to report fewer health problems and better mobility.

4. **Other Observations**

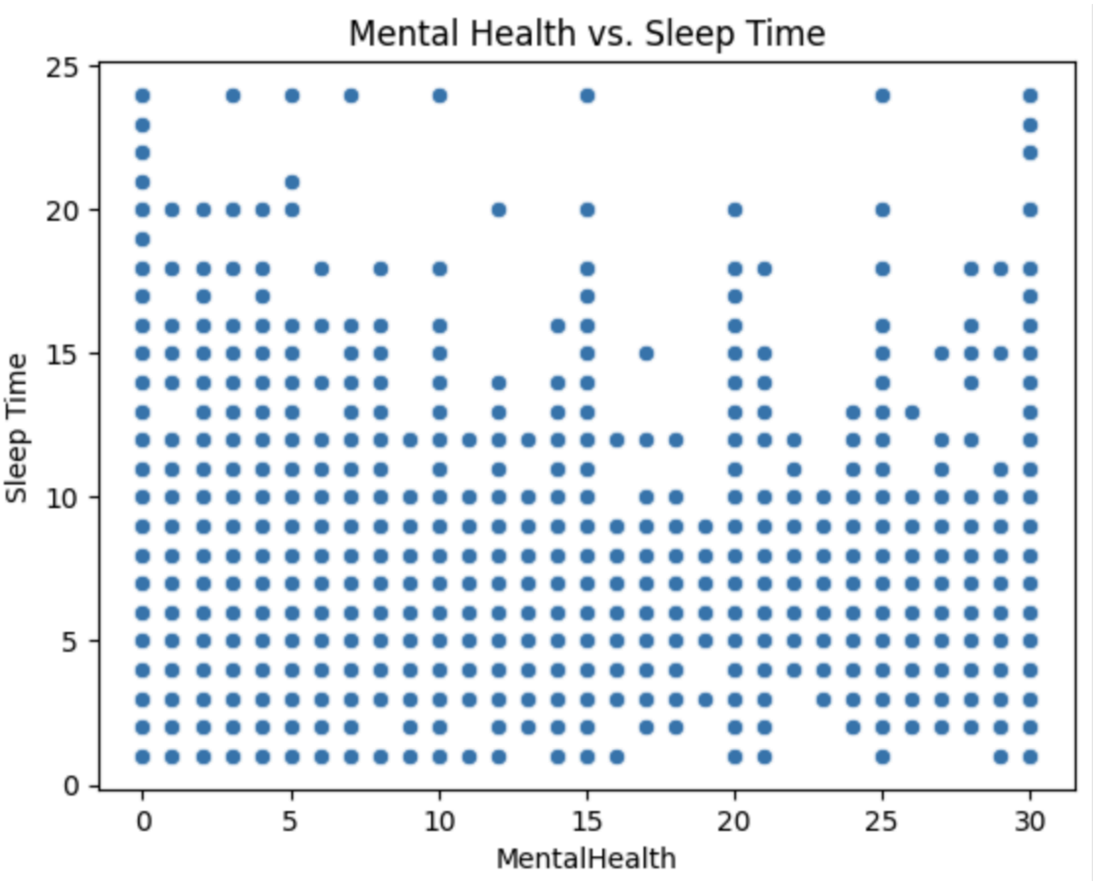
   a. SleepTime displays almost no correlation with any health variable, implying that sleep duration alone may not significantly distinguish health outcomes in this population.

   b. Most other variables show low inter-correlations, emphasizing that heart disease and general health outcomes are multifactorial rather than dependent on a single dominant feature.

5.14 Interpretation of Scatter Plots

In the previous section, the correlation heatmap provided a quantitative view of how various health and lifestyle variables relate to one another.
To complement that, we now look at two scatter plots that give a more visual understanding of specific variable relationships Mental Health vs. Sleep Time, and Physical Health vs. BMI.

A. Mental Health vs. Sleep Time

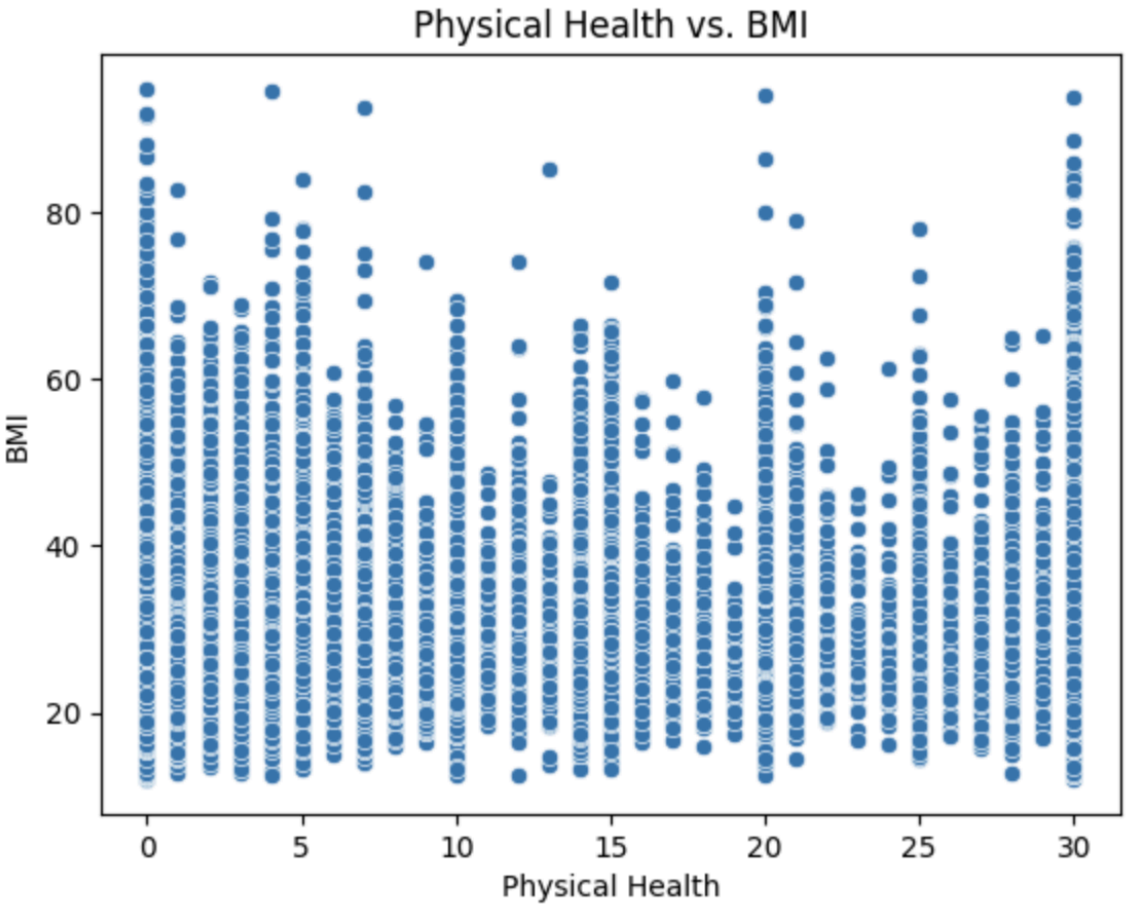


This plot examines how the number of days of poor mental health relates to the average hours of sleep per day.
From the visualization, there is no clear linear pattern between the two variables. Individuals across all ranges of sleep duration from very little sleep to over 10 hours, report varying levels of poor mental health. However, a slight visual trend can be observed where most points are clustered around 6–9 hours of sleep and fewer poor mental health days, suggesting that moderate sleep duration may correspond with relatively better mental well-being.

The absence of a strong trend also indicates that sleep time alone does not fully explain differences in mental health status. Mental well-being is likely influenced by a complex mix of factors such as stress, work-life balance, and underlying medical or lifestyle conditions, rather than by sleep duration in isolation.

B. Physical Health vs. BMI

The second plot explores the relationship between Body Mass Index (BMI) and days of poor physical health.



Here again, we see a widely scattered distribution, suggesting only a weak relationship. Most individuals, regardless of BMI, report low to moderate numbers of poor physical health days. However, there is a slight tendency for individuals with higher BMI values (above 30) to experience more days of poor physical health, consistent with the general understanding that excess body weight can contribute to chronic health issues.

The clustering near the lower end of the PhysicalHealth axis shows that a large portion of respondents report good overall physical well-being, which may dilute stronger trends within specific BMI groups. This pattern also aligns with the heatmap analysis, where BMI had only a minor positive correlation with physical health problems (r = 0.11).

Both scatter plots reinforce the earlier findings from the correlation analysis:

- The relationships among these health indicators are weak to moderately associated, with no single factor acting as a strong determinant of well-being.

- Mental health and sleep patterns appear relatively independent, while physical health and BMI show a mild, intuitive relationship where higher BMI aligns with slightly poorer health.

6. Data Preprocessing

Before training our predictive models, the dataset underwent several essential preprocessing steps to ensure it was clean, consistent, and ready for machine learning. This process involved identifying feature types, encoding categorical variables, scaling numeric features, and performing a balanced train–test split.

6.1. Identification of Feature Types

- The dataset was first examined to differentiate numeric and categorical variables.
- The numeric columns included: BMI, PhysicalHealth, MentalHealth, and SleepTime.
- All remaining variables were categorical in nature and were further divided based on the number of unique categories into:
- Binary categorical columns (with two unique values): HeartDisease, Smoking, AlcoholDrinking, Stroke, DiffWalking, Sex, PhysicalActivity, Asthma, KidneyDisease, and SkinCancer.
- Multi-class categorical columns (with more than two categories): AgeCategory, Race, Diabetic, and GenHealth.

This distinction was crucial as different encoding strategies were required for binary and multi-class features.

6.2. Encoding Categorical Variables

Since machine learning algorithms can only process numerical inputs, categorical variables were systematically converted into numeric form.

Binary Categorical Columns:

For binary features, Label Encoding was applied to convert text values such as “Yes/No” or “Male/Female” into 0 and 1. This simplified representation retains interpretability while ensuring the model can utilize these features effectively.

Multi-Class Categorical Columns:

For variables with more than two categories, One-Hot Encoding was applied using pd.get_dummies() with the parameter drop_first=True. This approach created new binary

columns for each category while dropping one category per variable to prevent the dummy variable trap (multicollinearity).
For instance, the column GenHealth was expanded into indicators such as GenHealth_Fair, GenHealth_Good, and GenHealth_Very good, with one category serving as the reference level.

This encoding ensured that categorical features were numerically represented without introducing artificial order or redundancy.

6.3. Scaling Numeric Features

The numeric features were then standardized using StandardScaler, which transforms values to have a mean of 0 and a standard deviation of 1.
This step ensures that all numeric features are on a comparable scale, preventing features with larger magnitudes from dominating the model's learning process. Standardization also facilitates faster convergence and improved performance, especially for algorithms such as logistic regression and K-nearest neighbors.

6.4. Train–Test Split

Following preprocessing, the dataset was divided into training and testing subsets to enable fair model evaluation.

Target variable (y): HeartDisease
Feature matrix (X): all remaining columns (after encoding)

The split ratio was 70% training and 30% testing, implemented using train_test_split() with stratification on the target variable.

The resulting sizes were:
Training set: 223,856 rows × 37 columns
Testing set: 95,939 rows × 37 columns

Stratification was applied to maintain the same proportion of positive and negative heart disease cases in both subsets, which is especially important because the dataset is highly imbalanced (only a small fraction of individuals reported heart disease). The use of a fixed random_state ensured reproducibility of results.

6.5. Outcome of Preprocessing

After preprocessing, all features were numeric, well-scaled, and free from redundancy. The final dataset was clean, balanced in terms of representation, and fully suitable for input into machine learning models. Each transformation was carefully chosen to preserve data integrity, enhance interpretability, and ensure fair, unbiased model evaluation.

7. Model Building and Interpretation

To predict the likelihood of heart disease based on self-reported health indicators, two supervised machine learning models were developed a Decision Tree Classifier and a Logistic Regression model. These models were trained and evaluated on the preprocessed dataset, which had been standardized, encoded, and split into 70% training and 30% testing subsets using stratified sampling to maintain the original proportion of heart disease cases.

7.1. Decision Tree Classifier

The Decision Tree Classifier was trained on the training data to learn patterns and relationships between the predictors and the target variable, HeartDisease. The model works by recursively splitting the dataset into subgroups based on feature thresholds that best separate individuals with and without heart disease. Once trained, the model was used to predict outcomes on the testing set.

The Decision Tree model was chosen because of its interpretability and its ability to capture complex, non-linear relationships between features. It provides clear decision rules that make it easier to understand which factors most strongly influence classification outcomes.

From the feature importance analysis, the Decision Tree primarily relied on variables representing physical and lifestyle conditions. Among these, Body Mass Index (BMI) emerged as the most influential predictor, followed by average sleep duration (SleepTime), days of poor physical health, and days of poor mental health. This indicates that individuals with higher BMI values, shorter sleep duration, and poorer physical or mental well-being were more likely to be classified as having heart disease.

Additional factors such as difficulty walking, lower levels of physical activity, smoking history, and previous stroke incidence also contributed to the model's predictions, though to a lesser extent. Overall, the Decision Tree emphasized lifestyle and health-quality measures, suggesting that these variables play a significant role in predicting heart disease risk. Its rule-based structure makes it particularly useful for identifying cut-off thresholds and interactions between variables that may not be visible through linear modeling approaches.

7.2. Logistic Regression

The Logistic Regression model was also trained on the same dataset to estimate the probability of heart disease occurrence. Because the target variable is binary (Yes/No), Logistic Regression is well-suited for this task, as it predicts the likelihood of an event occurring based on a set of input features.

This model was selected for its simplicity, efficiency, and interpretability. It provides a clear understanding of how each feature influences the odds of heart disease whether positively or negatively. After training, the model was evaluated on the test set to assess its performance and determine which predictors had the strongest influence.

The feature importance results from Logistic Regression revealed that age and general health perception were the most dominant predictors of heart disease. Older individuals, particularly those aged 65 and above, had significantly higher predicted probabilities of heart disease compared to younger respondents. Similarly, individuals who reported their general health as poor or fair were far more likely to be classified as at risk.

Other variables, such as history of stroke, sex, and presence of kidney disease, also contributed to the model's predictions but to a lesser degree. These findings align with established medical knowledge, reinforcing the strong correlation between age, perceived health status, and heart disease prevalence.

While both models addressed the same prediction task, their approaches and insights complemented each other. The Decision Tree captured non-linear and interaction effects between lifestyle and physical health variables, making it effective at identifying sharp thresholds where risk increases for example, when BMI or sleep duration crosses a certain level. In contrast, the Logistic Regression model captured broader, linear relationships, showing that the likelihood of heart disease increases steadily with age and decreases with better self-reported health.

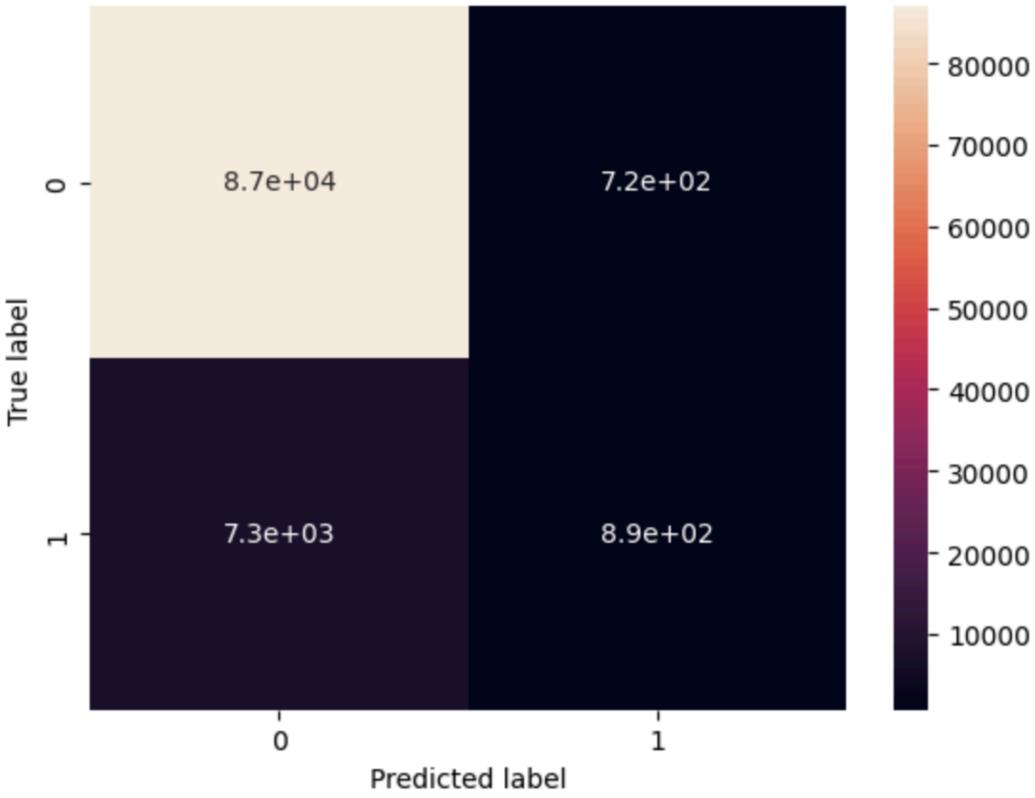
Together, these models provide a comprehensive understanding of the key factors influencing heart disease. The Decision Tree emphasizes behavioral and lifestyle factors, while the Logistic Regression model highlights demographic and self-perceived health factors. Both perspectives reinforce the conclusion that heart disease risk is shaped by a combination of lifestyle choices, physical well-being, and age-related health decline.

8. Evaluation of performance of models

In this analysis, two machine learning models Logistic Regression and Decision Tree were trained to predict the likelihood of heart disease using a dataset that is highly imbalanced, with only about 8.6% of cases representing patients with heart disease. This imbalance poses a common challenge in healthcare prediction models, as it can cause models to favor the majority "no disease" class and overlook true positive cases.

8.1 LOGISTIC REGRESSION:

Starting with the Logistic Regression model, the results reveal a distinct conservative behavior. Out of 95,939 test samples, the model correctly predicted 87,008 true negatives and 887 true positives, with only 721 false positives and 7,325 false negatives. These numbers translate to an overall accuracy of approximately 92%, precision of 55%, and recall of 11%. The high accuracy largely reflects the dominance of the "no disease" class, while the precision value indicates that more than half of the cases flagged as “disease” were indeed correct demonstrating that when the model raises an alert, it is often reliable. However, the low recall highlights that the model identifies only about one in ten actual heart disease cases, meaning many true cases remain undetected.

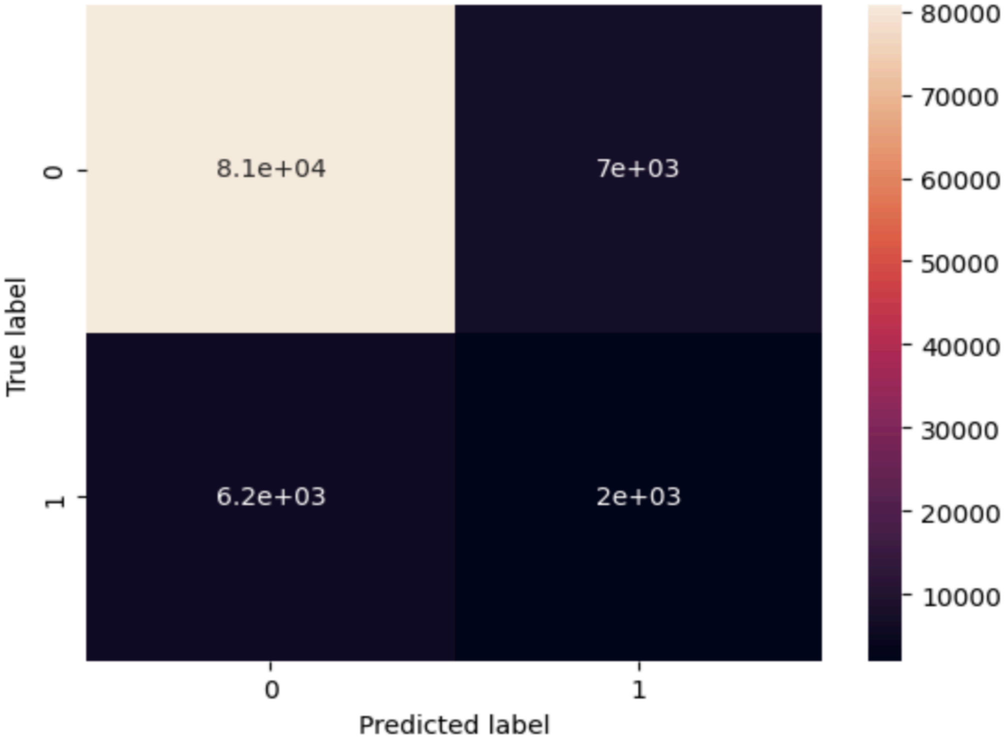


From an error perspective, the Logistic Regression model maintains a Mean Squared Error (MSE) of around 0.084 for both the training and testing data, corresponding to a Root Mean Squared Error (RMSE) of roughly 0.29. The close alignment between the train and test MSE values signals excellent generalization, showing that the model is neither underfitting nor overfitting.

This stability is reinforced by the confusion matrix heatmap, where the bright upper-left block (representing true negatives) dominates, and the smaller lower-right block (true positives) reflects the model’s high specificity but limited sensitivity. In simpler terms, the Logistic Regression model performs well at confirming the absence of disease but struggles to catch all positive cases. Such behavior makes it ideal for confirmatory diagnostics, where minimizing false alarms is crucial but less suitable for early screening, where missing a potential case could have serious consequences.

8.2 DECISION TREE

The Decision Tree model, in contrast, tells a more complex story. Its confusion matrix reports 80,740 true negatives, 2,058 true positives, 6,938 false positives, and 6,154 false negatives, yielding an accuracy of about 86%, precision of 23%, and recall of 25%. Compared to Logistic Regression, the Decision Tree is more aggressive in identifying heart disease cases, correctly catching a greater number of true positives, which leads to improved recall. However, this comes at a cost. It also raises more false alarms, reducing precision. In healthcare applications, such a trade-off might be acceptable in screening contexts where the goal is to detect as many potential cases as possible, even if some healthy individuals are flagged for further testing.



A deeper look at the model's errors, however, reveals a significant concern thai is overfitting. The Decision Tree achieves an extremely low training MSE of 0.003 (RMSE = 0.05), indicating that it fits the training data almost perfectly. Yet, its testing MSE rises sharply to 0.137 (RMSE = 0.37), showing that it fails to generalize to unseen data. This discrepancy means the model memorized patterns specific to the training dataset rather than learning general, predictive relationships. The heatmap visualization supports this observation. It shows larger positive blocks than Logistic Regression, reflecting more true detections, but also a bigger cluster of false positives, illustrating the instability typical of an overfitted model. The Decision Tree provides broader detection and can be used for screening but requires fine-tuning to manage false alarms and overfitting.

9. CONCLUSION

Our project set out to answer a central question, can heart disease be predicted using self-reported health indicators? Through the development and evaluation of two machine-learning models Logistic Regression and Decision Tree we were able to gain a nuanced understanding of how self-reported measures such as BMI, Sleep Time, Physical Health, and Mental Health relate to the likelihood of heart disease.

The results show that while self-reported health indicators hold predictive value, their standalone strength is modest.

The Logistic Regression model demonstrated excellent generalization with nearly identical training and testing errors (MSE = 0.084), achieving high accuracy (92%) and good precision (55%) but very low recall (11%). This indicates that the model performs reliably when predicting “no disease” and is cautious in labeling positive cases which is ideal for confirmatory diagnostics but limited for proactive screening.

The Decision Tree model, on the other hand, was more flexible and captured non-linear relationships, leading to better recall (25%) but at the cost of lower precision (23%) and clear signs of overfitting (train MSE = 0.003 vs test MSE = 0.137).

Together, these results suggest that while we can indeed model heart-disease risk from self-reported indicators, the current features and model complexity are insufficient for high-sensitivity medical prediction.

In essence, we did achieve the project’s goal which is to demonstrate that machine learning can identify patterns linking self-reported lifestyle and health data to heart disease, but with important limitations. The models confirm that these indicators carry useful signals, yet they alone cannot predict heart disease with clinical reliability. The low recall across both models reveals that many true cases remain undetected, likely because self-reported data, while accessible, lack the precision of clinical biomarkers such as cholesterol levels, blood pressure, or ECG results.

Overall, our analysis demonstrates that while self-reported health indicators alone cannot yet serve as a reliable diagnostic tool, they provide a valuable foundation for early-risk screening and public-health monitoring. By integrating clinical data such as laboratory results, blood pressure readings, and medical imaging alongside these self-reported features, future iterations of the model could significantly enhance predictive accuracy and interpretability. This integration would bridge personal health perception with objective physiological measures, transforming the model into a more holistic, data-driven tool for early detection and preventive care. In that sense, our project represents an important first step demonstrating the feasibility of predictive modeling from self-reported indicators while paving the way toward a more comprehensive and clinically integrated approach to heart-disease risk prediction.