ALY 6040 – Data Mining Applications

Professor: Kasun Samarasinghe

Northeastern University, Boston

Logo

Description automatically generated

Module 6 Final Project

Submitted by

Santhi Kiran Chavali

Exploration and Prediction of Heart Disease Risk Factors: An Analysis of Health Indicators

**Introduction:**

This comprehensive final project focuses on the exploration and prediction of heart disease risk factors, amalgamating insights derived from various analytical phases. Utilizing data sourced from the Behavioral Risk Factor Surveillance System (BRFSS), a program administered by the Centers for Disease Control and Prevention (CDC), our analysis spans diverse facets of health-related information gathered through annual telephone surveys across the United States and its territories. The dataset, available at [**https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data**](https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data), serves as the cornerstone of our investigation. Encompassing demographic information, health behaviors, and medical history, it provides a comprehensive lens into potential risk factors associated with heart disease.

Our analytical journey begins with Exploratory Data Analysis (EDA), where we unravel the intricacies of the dataset to uncover initial insights. Transitioning into a more comprehensive EDA phase, we delve deeper into the characteristics and distribution of the data. Subsequently, our focus shifts to predictive modeling, employing a range of machine learning algorithms, including Decision Trees, Random Forests, Gradient Boosting Machines (GBM, XgBoost, CatBoost, LightGBM), and Logistic Regression. The objective is to develop robust models capable of accurately predicting the occurrence of heart disease.

Throughout the analysis, we scrutinize various metrics such as accuracy, runtime, and feature importance. Additionally, model optimization techniques, including hyperparameter tuning, are explored to enhance predictive performance. Our approach aims to provide a holistic understanding of heart disease prediction, showcasing the synergy between exploratory analysis and predictive modeling in the realm of healthcare analytics. This amalgamation of insights contributes to the broader goal of advancing early detection and prevention strategies for heart disease, ultimately enhancing public health outcomes.

**Problem Statement:**

The objective is to develop predictive models using health-related variables to assess an individual's risk of heart disease.

**Data Dictionary:**

**Our Dataset consists of 246,022 records/rows and 40 variables/columns**

State: The state where each individual is located.

Sex: The gender of each individual.

GeneralHealth: Information about the general health status of each individual.

PhysicalHealthDays: The number of days an individual experiences physical health issues.

MentalHealthDays: The number of days an individual experiences mental health issues.

LastCheckupTime: Time of the last health checkup for each individual.

PhysicalActivities: Information about physical activities or exercise.

SleepHours: The number of hours each individual sleeps.

RemovedTeeth: Information about the removal of teeth.

HadHeartAttack: Whether an individual has had a heart attack.

HadAngina: Whether an individual has had angina (chest pain or discomfort).

HadStroke: Whether an individual has had a stroke.

HadAsthma: Whether an individual has had asthma.

HadSkinCancer: Whether an individual has had skin cancer.

HadCOPD: Whether an individual has had Chronic Obstructive Pulmonary Disease (COPD).

HadDepressiveDisorder: Whether an individual has had a depressive disorder.

HadKidneyDisease: Whether an individual has had kidney disease.

HadArthritis: Whether an individual has had arthritis.

HadDiabetes: Whether an individual has had diabetes.

DeafOrHardOfHearing: Whether an individual is deaf or hard of hearing.

BlindOrVisionDifficulty: Whether an individual has blindness or vision difficulties.

DifficultyConcentrating: Whether an individual experiences difficulty concentrating.

DifficultyWalking: Whether an individual experiences difficulty walking.

DifficultyDressingBathing: Whether an individual experiences difficulty in dressing or bathing.

DifficultyErrands: Whether an individual experiences difficulty running errands.

SmokerStatus: Information about an individual's smoking status.

ECigaretteUsage: Information about e-cigarette usage.

ChestScan: Whether an individual has had a chest scan.

RaceEthnicityCategory: The racial and ethnic category of each individual.

AgeCategory: The age category of each individual.

HeightInMeters: The height of each individual in meters.

WeightInKilograms: The weight of each individual in kilograms.

BMI: Body Mass Index (BMI) of each individual.

AlcoholDrinkers: Information about alcohol consumption.

HIVTesting: Whether an individual has undergone HIV testing.

FluVaxLast12: Whether an individual has received a flu vaccine in the last 12 months.

PneumoVaxEver: Whether an individual has ever received a pneumonia vaccine.

TetanusLast10Tdap: Whether an individual has received a tetanus vaccine in the last 10 years.

HighRiskLastYear: Whether an individual is considered high risk for health issues in the last year.

CovidPos: Whether an individual has tested positive for COVID-19.

EDA:

In the initial phase of our project, focused on Exploratory Data Analysis (EDA), we imported essential Python libraries into our Jupyter notebook environment, including Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, LabelEncoder, XGBoost, imbalanced-learn, and others. Using the read\_csv() function, we successfully imported the heart disease dataset into Jupyter notebook.

A screenshot of a computer

Description automatically generated

An initial data check was conducted using the df.info() function to understand the dataset's structure. This step provided essential information, including the number of columns, data types, and the presence or absence of missing values.

A screenshot of a computer

Description automatically generated

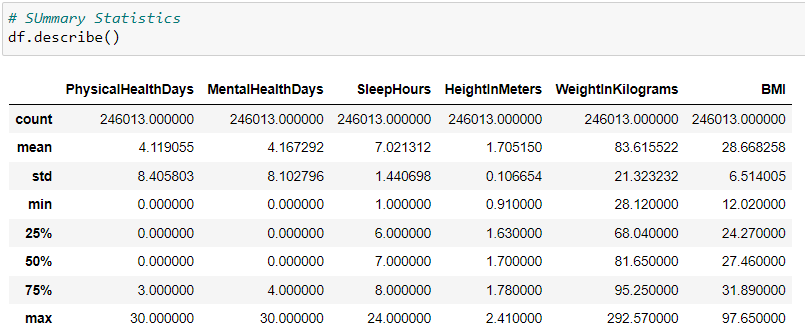
Notably, the initial data check revealed no missing values in the dataset, indicating data completeness. This absence of missing values reduces the need for extensive data imputation.

A screenshot of a computer screen

Description automatically generated

**Summary statistics of numerical variables:**

We have created Bar graphs for the categorical variables and used the describe() function to know the summary statistics and got the following output



**Interpretation of the above output:**

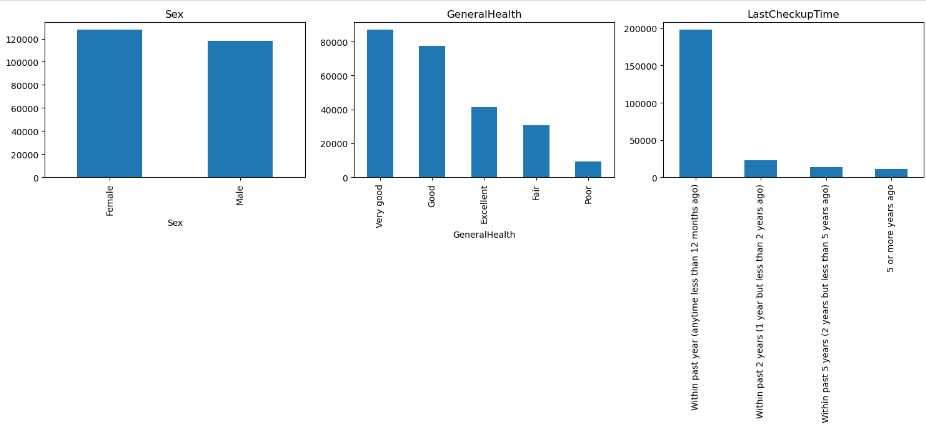
**PhysicalHealthDays and MentalHealthDays:** These variables represent the number of days individuals report experiencing physical and mental health issues, respectively. The data distribution suggests that a significant portion of individuals reported no health issues on any days, but the presence of outliers indicates that some individuals experienced health problems for up to 30 days.

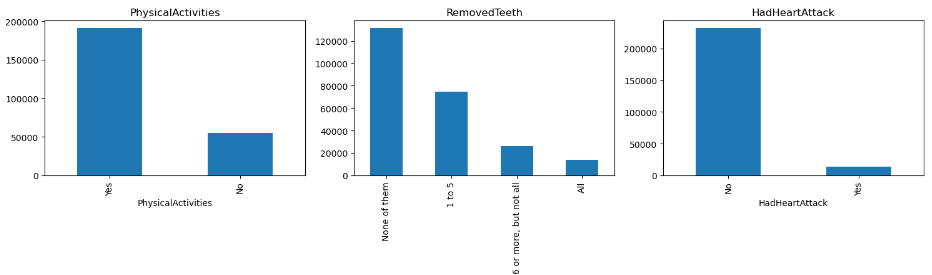
**SleepHours:** The dataset provides information on the reported hours of sleep individuals get. The average sleep duration is approximately 7 hours, with some variability and outliers suggesting a wide range of sleep patterns.

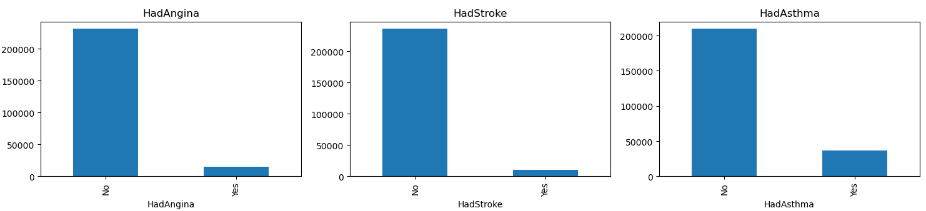
**HeightInMeters and WeightInKilograms:** These columns reveal data related to the height and weight of individuals. The data indicate a range of values, with an average height of around 1.705 meters and an average weight of approximately 83.62 kilograms.

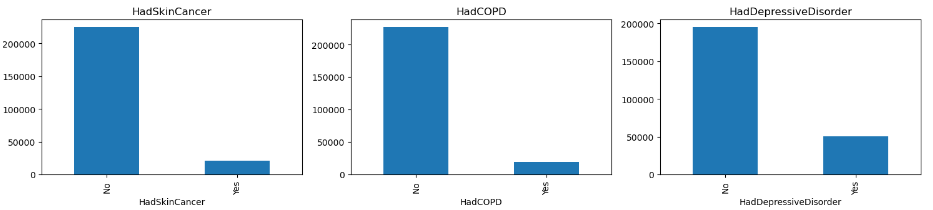
**BMI (Body Mass Index):** BMI is calculated based on height and weight. The dataset shows a broad range of BMIs, with an average value of about 28.67. Some variability and outliers suggest different levels of body mass among individuals.

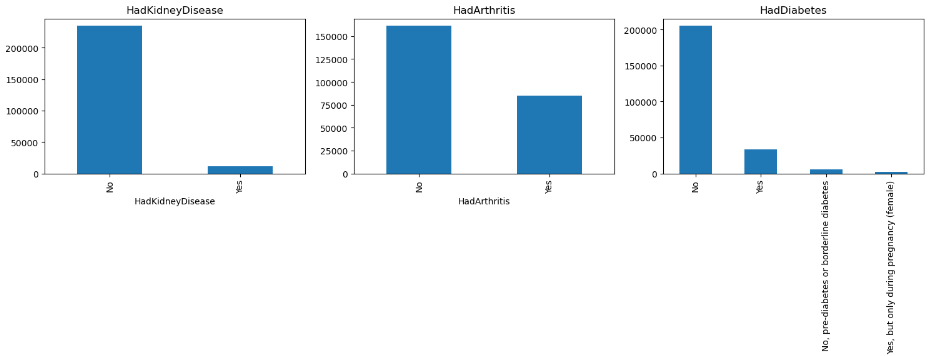
**Bar plots of Categorical variables**

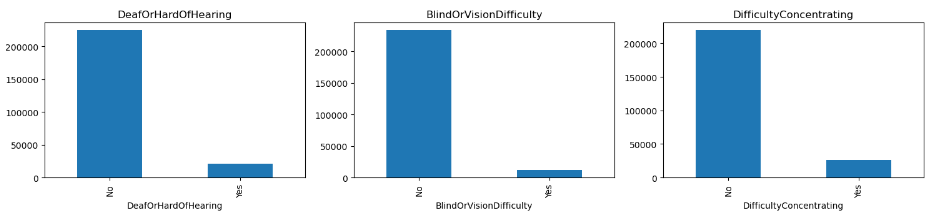


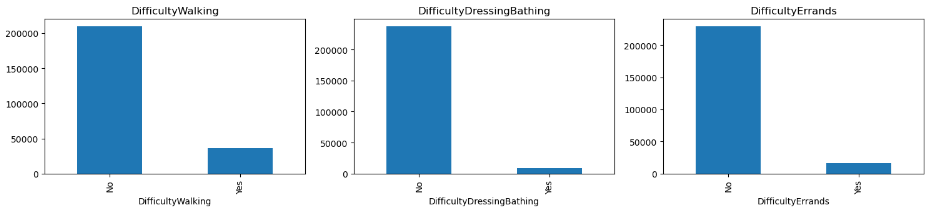


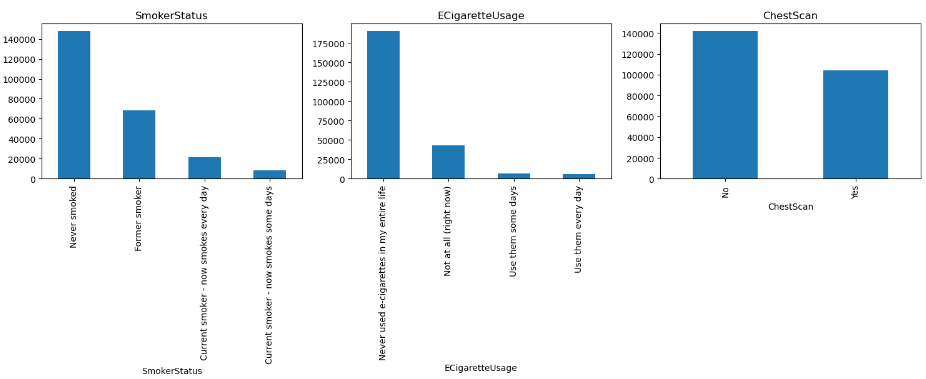


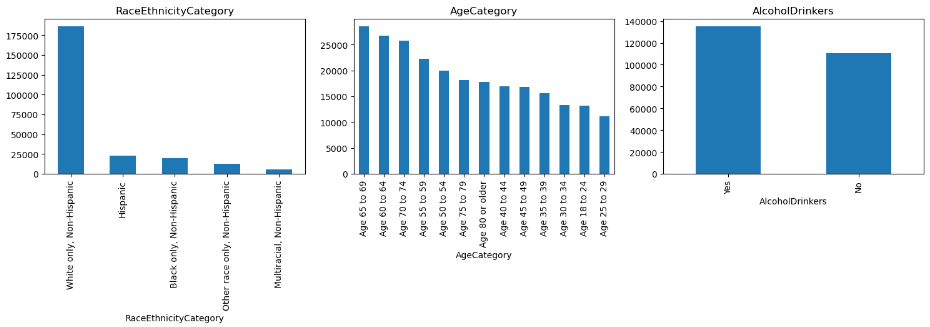


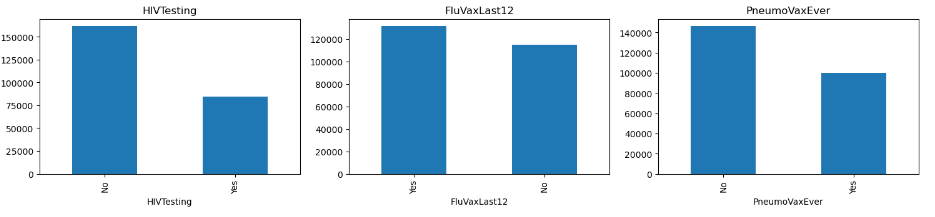


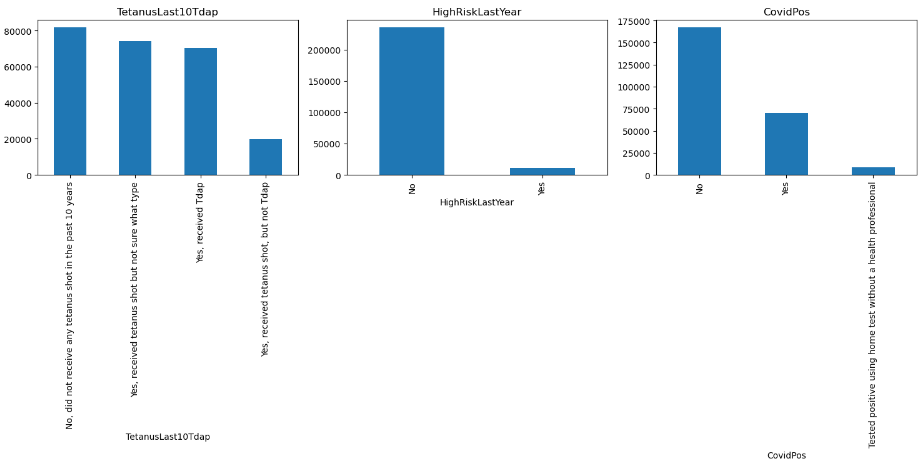












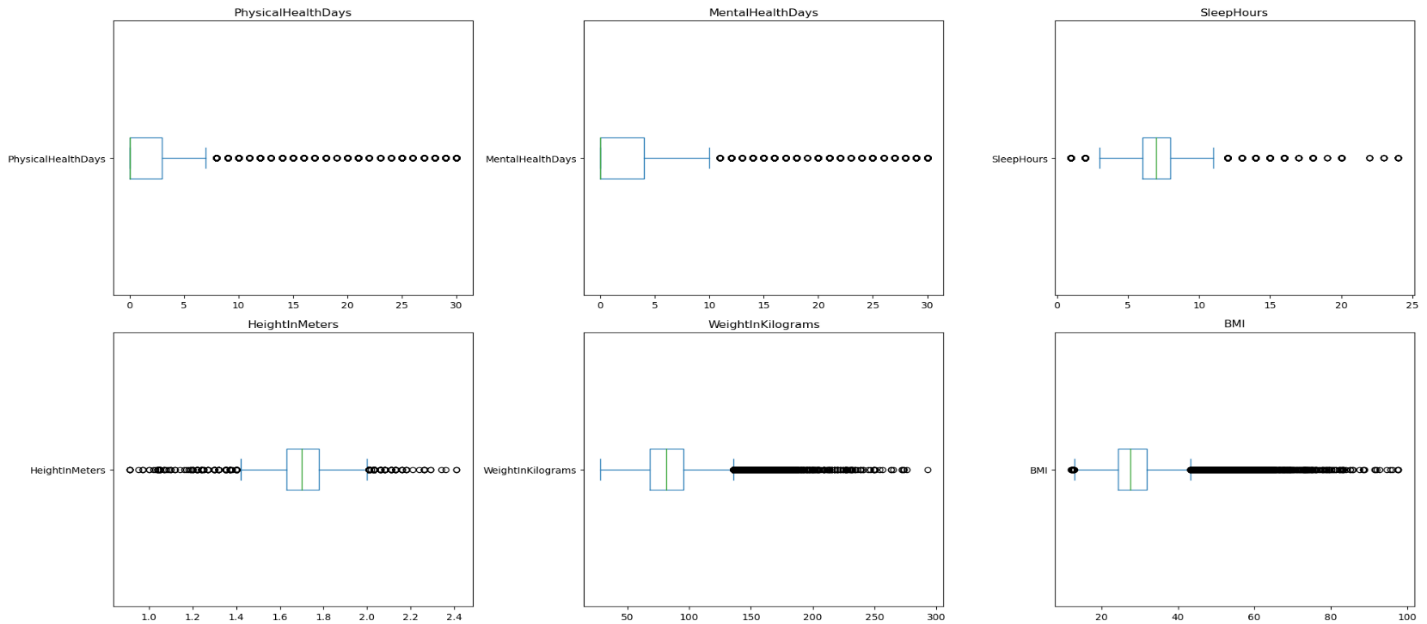
**Interpretation of above bar graphs:**

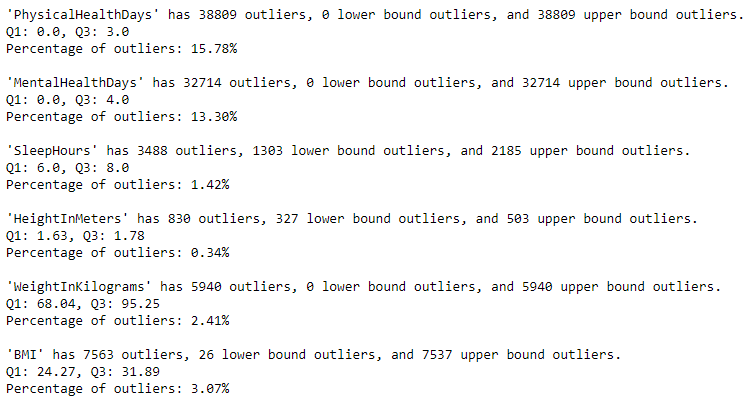
In examining the categorical variables within our dataset, we observe a balanced representation of gender, with approximately equal numbers of female and male respondents. When considering general health, the majority of participants rate their well-being as very good or good, with a notable portion expressing excellent health. A significant number have undergone recent checkups, with the majority within the past year. Physical activity is prevalent among respondents, with the majority engaging in such activities. Dental health varies, with most individuals having none or a few teeth removed. Instances of heart attacks are relatively low among participants. The dataset also reveals a diverse range of health conditions, including angina, stroke, asthma, skin cancer, COPD, depressive disorder, kidney disease, arthritis, and diabetes. Smoking habits vary, with a considerable number reporting never smoking, being former smokers, or currently smoking. Approximately half of the respondents have not undergone a chest scan. Racial and ethnic diversity is evident, with the majority being White only, non-Hispanic. Age distribution spans various categories, with higher representation in the age group 65 to 69. Alcohol consumption is common among respondents. A substantial number have not undergone HIV testing. Vaccination records indicate a significant number receiving flu shots, while tetanus vaccination histories are diverse. The majority perceive themselves as not being at high risk in the last year. COVID-19 testing reveals a larger number of negative results compared to positive ones.

This comprehensive overview of categorical variables provides a nuanced understanding of the demographic and health characteristics of our dataset, laying the groundwork for more detailed analyses in subsequent stages of our project.

**Boxplots and outliers:**

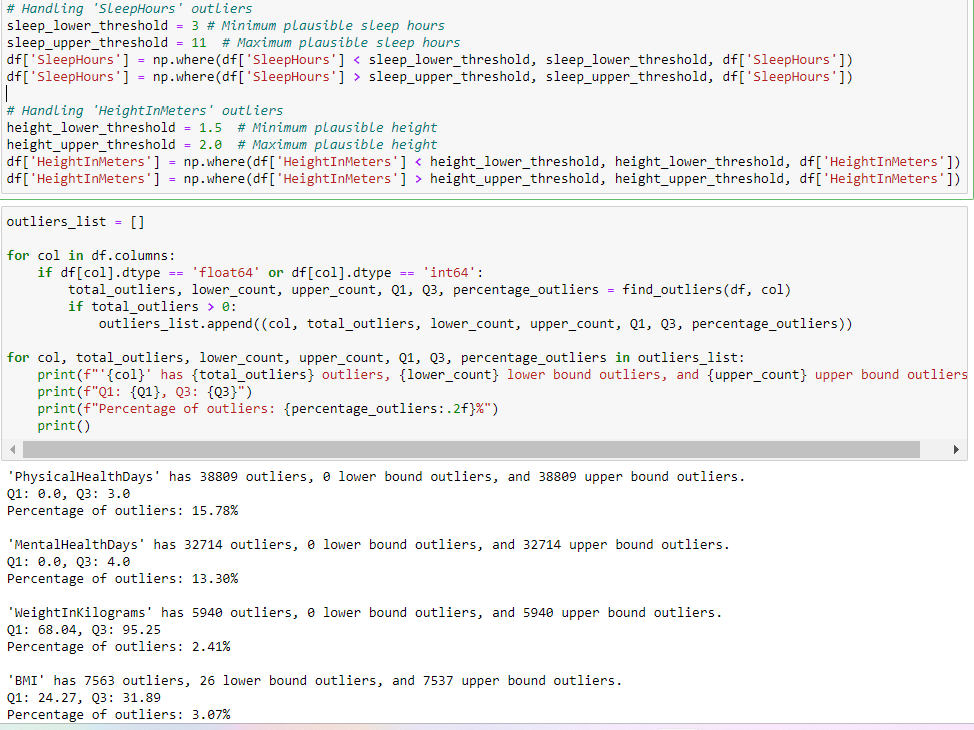
For the next step, we want to see the outliers in the dataset. So, we found the following outliers in the dataset. boxplots are created for visualization.





**Interpretation:**

In our dataset, we observed outliers in key health-related variables that play a crucial role in understanding the overall well-being of individuals. Notably, metrics related to physical and mental health days exhibited a considerable number of upper-bound outliers, with 38,809 outliers for physical health days (15.78% of the data) and 32,714 outliers for mental health days (13.30% of the data). This indicates instances of an elevated number of physically and mentally unhealthy days among certain individuals. Sleep hours showed a smaller percentage of outliers, with 3,488 outliers (1.42% of the data), suggesting variations in sleep patterns. Anthropometric measurements, including height, weight, and Body Mass Index (BMI), also presented outliers. Weight and BMI demonstrated a moderate percentage of upper bound outliers, with 5,940 outliers for weight (2.41% of the data) and 7,563 outliers for BMI (3.07% of the data). These outliers shed light on potential extremes in reported health metrics, prompting further exploration into the factors contributing to these deviations. As we progress in our analysis, addressing the impact of outliers will be essential for ensuring the accuracy and reliability of our findings.



In addressing outliers within crucial health-related variables for predicting heart disease, we implemented specific thresholds to mitigate extreme values while retaining the integrity of the data. For 'SleepHours,' we applied a lower threshold of 3 hours and an upper threshold of 11 hours, ensuring that reported sleep durations fall within plausible ranges. Similarly, for 'HeightInMeters,' we set a lower limit of 1.5 meters and an upper limit of 2.0 meters, maintaining physiologically reasonable heights. Following these adjustments, a comprehensive evaluation of outliers was conducted for all numeric columns in the dataset. Notably, the 'PhysicalHealthDays,' 'MentalHealthDays,' 'WeightInKilograms,' and 'BMI' variables exhibited outliers. For each of these, we successfully addressed extreme values using the specified thresholds, resulting in a more refined dataset for subsequent analyses. The number of outliers, both lower and upper bound, along with their percentages, were documented to provide transparency regarding the impact of the outlier-handling procedure. This approach ensures that our predictive models for heart disease are based on a more reliable and representative dataset, considering the importance of these health metrics in the analysis.

**Analysis:**

In streamlining our dataset for further analysis, we opted to drop the 'State' column, as it was deemed unnecessary for the specific focus of our project. By executing the drop operation, we effectively removed this column from our dataset, enhancing simplicity and reducing dimensionality without sacrificing the essential variables required for our investigation into heart disease risk factors. This strategic decision aligns with our objective of refining the dataset to include only the most pertinent features for predictive modeling and analysis.

****

**Correlation between numeric variables:**

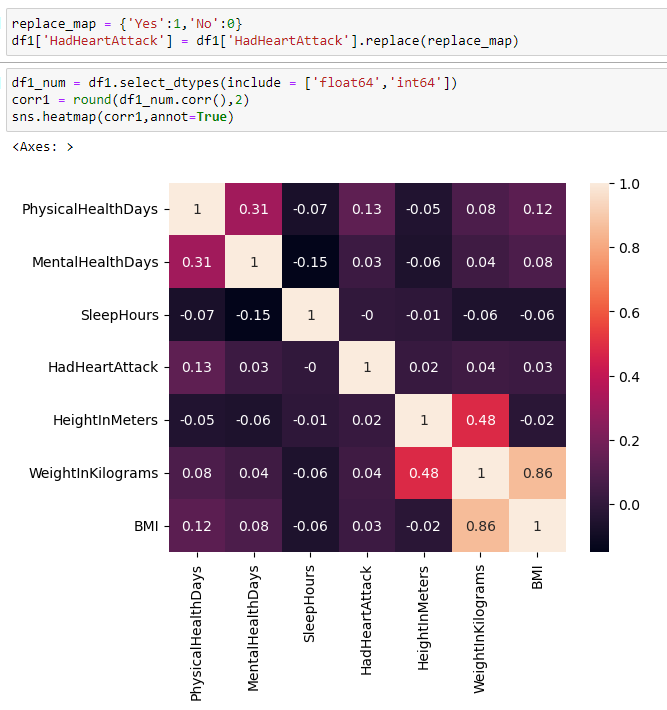
In our pursuit of understanding the relationships between various health metrics and the occurrence of heart attacks, we performed a correlation analysis on selected numerical features within our dataset. To facilitate this analysis, we transformed categorical responses in the 'HadHeartAttack' column to numerical values (Yes: 1, No: 0). The resulting correlation matrix provides insights into the pairwise associations among key variables.

Upon examining the correlations, several notable patterns emerged. Both 'PhysicalHealthDays' and 'MentalHealthDays' exhibit positive correlations with values of 0.31, indicating a moderate positive relationship between these health-related days. Sleep duration ('SleepHours') displayed negligible correlation with the occurrence of heart attacks ('HadHeartAttack'), emphasizing the importance of exploring additional factors in our predictive modeling.

Notably, 'HeightInMeters' demonstrated a weak negative correlation (-0.05) with 'PhysicalHealthDays' and 'MentalHealthDays,' suggesting a marginal decrease in reported health days with increasing height. 'WeightInKilograms' and 'BMI' exhibited strong positive correlations of 0.86, highlighting the inherent association between weight and body mass index.

Additionally, it's essential to note that our dataset is imbalanced, with a majority of respondents indicating no occurrence of heart attacks. This imbalance is a critical aspect that we need to address in our predictive modeling to ensure the robustness and fairness of our results.

These insights from the correlation matrix, along with the acknowledgment of the dataset's imbalance, guide our subsequent feature selection and model-building processes, informing us about potential interdependencies among key variables and their impact on predicting heart disease risk.

****

**Model Process:**

We initiated our modeling process by preparing the dataset for training. The features (X) were defined by excluding the target variable 'HadHeartAttack,' while the target variable (y) was isolated. This partitioning was essential to ensure the model's independence from the variable it aims to predict.



We continued our modeling journey by stratifying the dataset into training and testing sets using the train\_test\_split function. This stratified split ensured that the proportion of classes within each set mirrored the original distribution, maintaining the integrity of the dataset's class balance. The split allocated 80% of the data to the training set (X\_train, y\_train) and the remaining 20% to the testing set (X\_test, y\_test). The random\_state parameter was set to 42 for reproducibility.

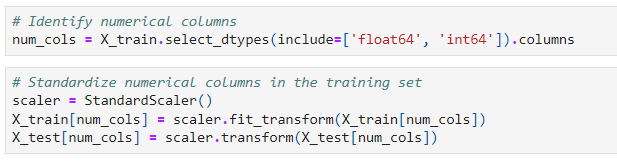


To facilitate the integration of categorical features into our models, we identified the categorical columns within the training set using the select\_dtypes function. Subsequently, we employed label encoding to transform these categorical variables into numerical representations. The LabelEncoder from scikit-learn was applied to both the training (X\_train) and testing (X\_test) sets. This encoding process is vital for ensuring that machine learning algorithms can effectively interpret and utilize categorical information during the training phase.

A screenshot of a computer code

Description automatically generated

Continuing our preprocessing efforts, we identified the numerical columns within the training set, encompassing both float64 and int64 data types. The numerical features were then standardized to exhibit zero mean and unit variance using the StandardScaler from scikit-learn. This standardization process ensures that numerical variables with varying scales contribute equally to model training, preventing any particular feature from disproportionately influencing the model's learning process. The same scaling transformations were applied to both the training (X\_train) and testing (X\_test) sets.



To address the class imbalance present in the dataset, we employed the Synthetic Minority Over-sampling Technique (SMOTE). Utilizing the SMOTE class from the imbalanced-learn library, we oversampled the minority class (individuals who had a heart attack) in the training set. This technique generates synthetic samples by interpolating between existing instances, thereby achieving a more balanced representation of the target variable. The resampled data, consisting of augmented features (X\_train\_resampled) and target labels (y\_train\_resampled), was then ready for training machine learning models. This step was crucial to ensure that the models are not biased toward the majority class and can effectively capture patterns within the minority class.

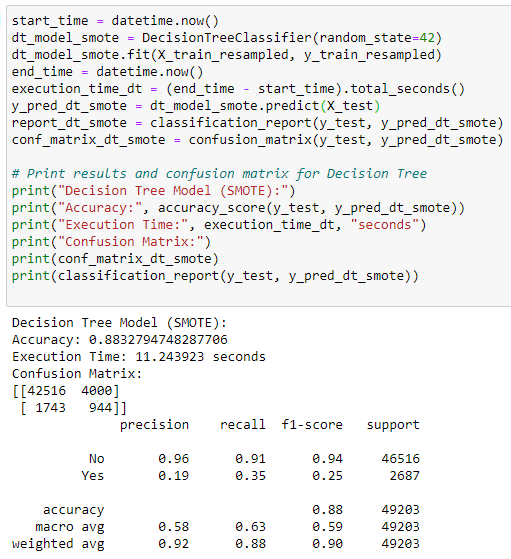
A white rectangle with black and purple text

Description automatically generated

**Model Building:**

**Decision Tree:**

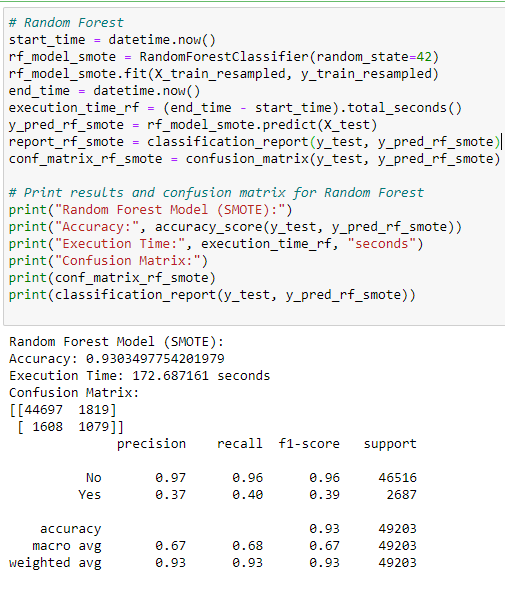
Utilizing the Decision Tree model with Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalanced dataset, we achieved an accuracy of 88.33%. This metric reflects the model's overall effectiveness in correctly classifying instances. The model's execution time was 11.24 seconds, demonstrating reasonable computational efficiency. In terms of the F1-score, a metric suitable for imbalanced datasets, we observe that the model performs exceptionally well for the majority class ('No') with a precision of 96%, recall of 91%, and an F1-score of 94%. However, its performance for the minority class ('Yes') is comparatively lower, with a precision of 19%, recall of 35%, and an F1-score of 25%. These results highlight the model's capability to accurately predict instances of the majority class but suggest room for improvement in predicting the minority class. Consideration of these metrics guides potential enhancements and optimizations for future iterations of our predictive modeling process.



**Random Forest:**

Applying the Random Forest model with Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalanced dataset yielded impressive results. The model achieved an accuracy of 93.03%, showcasing its robustness in correctly classifying instances. The execution time for the Random Forest model was 172.69 seconds, reflecting a reasonable computational efficiency for the given task. Focusing on the F1-score as a suitable metric for imbalanced datasets, we observe notable performance improvements compared to the Decision Tree model. For the majority class ('No'), the Random Forest model exhibits a precision of 97%, recall of 96%, and an F1-score of 96%. While performance for the minority class ('Yes') remains a challenge, with a precision of 37%, recall of 40%, and an F1-score of 39%, it shows enhancement compared to the Decision Tree model.

In summary, the Random Forest model demonstrates superior predictive capabilities, especially for the majority class, and offers a promising avenue for further exploration and potential optimization to enhance predictions for the minority class.



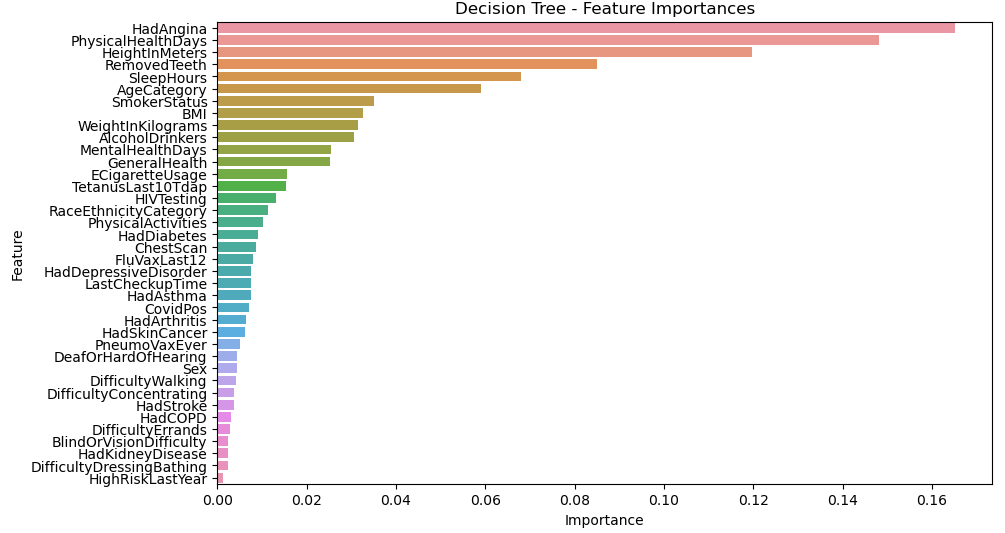
**Feature Importance:**

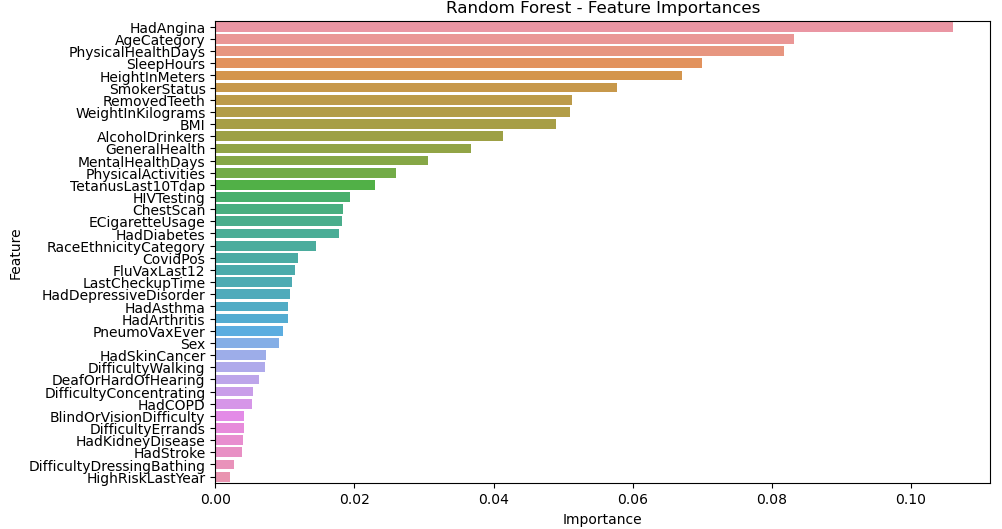
In our exploration of the Decision Tree and Random Forest models, we identified the top five significant variables crucial for predicting heart disease.

**Decision Tree:** HadAngina, PhysicalHealthDays, HeightInMeters, RemovedTeeth, SleepHours

**Random Forest:** HadAngina, AgeCategory, PhysicalHealthDays, SleepHours, HeightInMeters

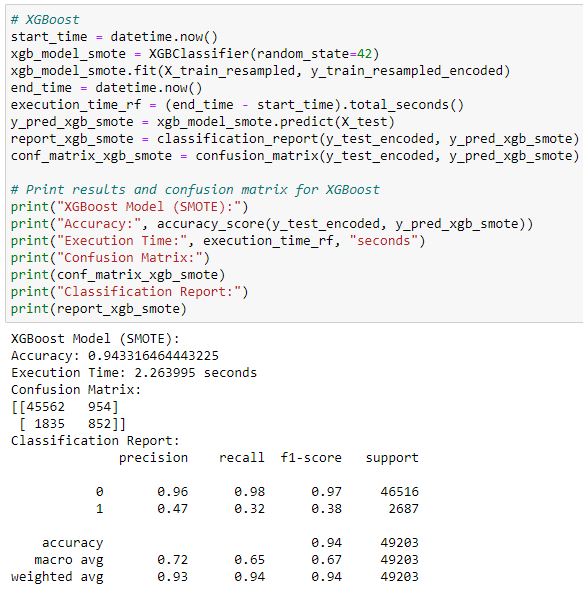
These variables, ranked by their importance in the models, provide valuable insights into the key determinants influencing the prediction of heart disease risk. The consistency of 'HadAngina' and 'PhysicalHealthDays' across both models underscores their substantial impact on the predictive outcomes. The differences in variable importance between the models highlight the nuanced aspects considered by each algorithm, contributing to a comprehensive understanding of the underlying patterns in our dataset.





**XGBoost Model:**

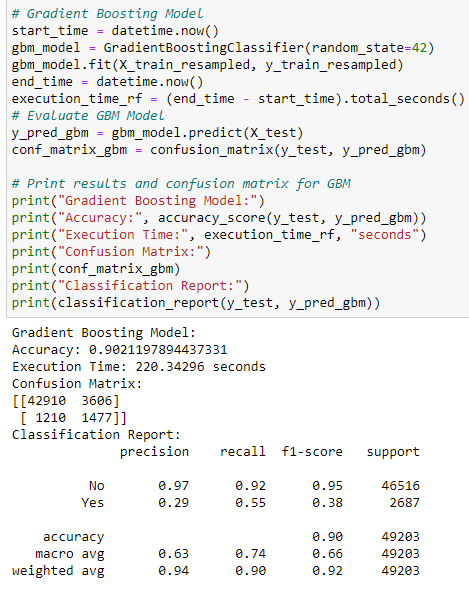
Applying the XGBoost model with Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalanced dataset yielded compelling results. The model achieved an accuracy of 94.33%, indicating its proficiency in correctly classifying instances. The execution time for the XGBoost model was notably efficient, recorded at 2.26 seconds. Examining the confusion matrix, we observe that the model correctly classified 45,562 instances of the majority class ('No'), with 954 misclassifications. For the minority class ('Yes'), the model successfully predicted 852 instances but missed 1,835 instances. Further insight from the classification report reveals that the model excels in precision for the majority class, with a precision of 96% and recall of 98%, resulting in an impressive F1-score of 97%. However, performance for the minority class is more challenging, with a precision of 47%, recall of 32%, and an F1-score of 38%. In summary, the XGBoost model demonstrates notable accuracy and efficiency, offering a robust approach for predicting heart disease risk. While its performance on the minority class is a point of consideration, the model showcases promise and provides valuable insights for potential optimization in future iterations of our predictive modeling process.



**Gradient Boosting Model:**

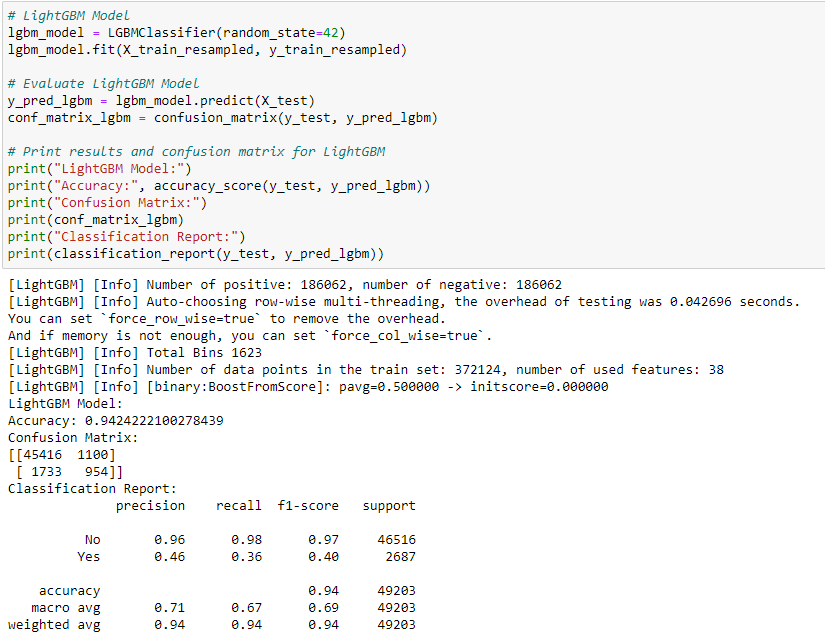
The gradient-boosting model exhibited commendable performance in predicting heart disease risk. With an accuracy of 90.21%, the model displayed proficiency in correctly classifying instances. However, the execution time for the Gradient Boosting model was relatively higher, recorded at 224.02 seconds. Analyzing the confusion matrix, we observe that the model accurately predicted 42,910 instances of the majority class ('No'), with 3,606 misclassifications. For the minority class ('Yes'), the model successfully predicted 1,477 instances but missed 1,210 instances.

Delving into the classification report, we note that the model achieves a high precision of 97% and recall of 92% for the majority class, resulting in an impressive F1-score of 95%. However, its performance for the minority class is more challenging, with a precision of 29%, recall of 55%, and an F1-score of 38%. In summary, the Gradient Boosting model demonstrates strong predictive capabilities, particularly for the majority class. While the execution time is relatively higher, the model's accuracy and precision provide valuable insights for potential optimization in future iterations of our predictive modeling process.



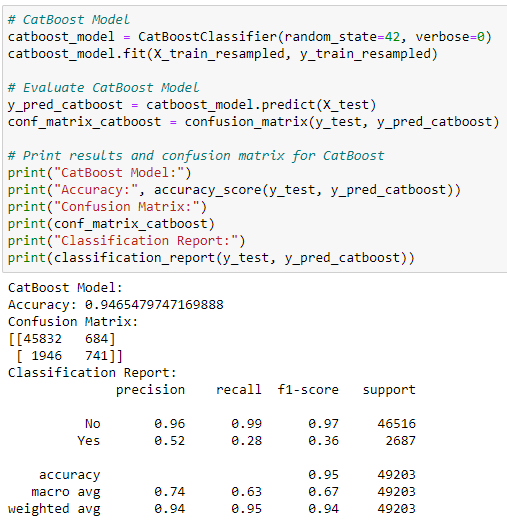
**LightGBM Model:**

The LightGBM model, a powerful gradient boosting framework, exhibited strong predictive performance in our analysis. With an accuracy of 94.24%, the model demonstrated proficiency in correctly classifying instances. The confusion matrix reveals that the model correctly predicted 45,416 instances of the majority class ('No') and 954 instances of the minority class ('Yes'), with 1,100 and 1,733 misclassifications, respectively. In the classification report, the model achieved a precision of 96% and recall of 98% for the majority class, resulting in an impressive F1-score of 97%. However, performance for the minority class is relatively lower, with a precision of 46%, recall of 36%, and an F1-score of 40%. In summary, the LightGBM model demonstrates notable accuracy and efficiency, offering a robust approach for predicting heart disease risk. While its performance on the minority class is a point of consideration, the model showcases promise and provides valuable insights for potential optimization in future iterations of our predictive modeling process.



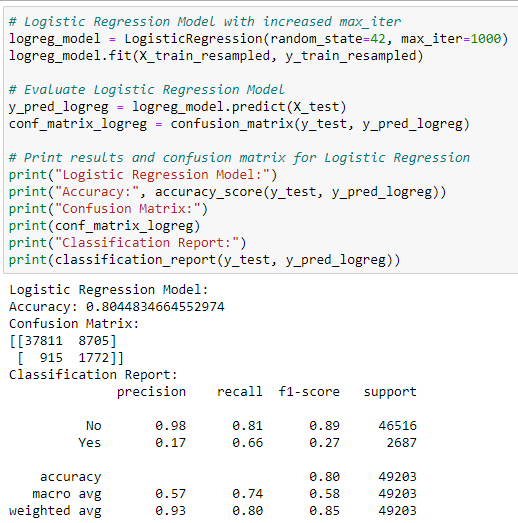
**CatBoost Model:**

The CatBoost model, celebrated for its adept handling of categorical features, yielded compelling results in our heart disease risk prediction task. Demonstrating an accuracy of 94.65%, the model exhibited proficiency in correctly classifying instances. In examining the confusion matrix, we observe that the model accurately predicted 45,832 instances of the majority class ('No') and 741 instances of the minority class ('Yes'), with 684 and 1,946 misclassifications, respectively. The classification report sheds further light on the model's performance. For the majority class, the model achieved a precision of 96%, indicating a high proportion of correctly predicted instances among those classified as 'No.' Additionally, the recall for the majority class was 99%, signifying the model's ability to capture the majority of actual 'No' instances. However, challenges arise in predicting instances of the minority class ('Yes'), where the precision drops to 52%, indicating a proportion of misclassifications among those predicted as 'Yes.' The recall for the minority class stands at 28%, emphasizing the model's struggle to capture the majority of actual 'Yes' instances.



**Logistic Regression Model:**

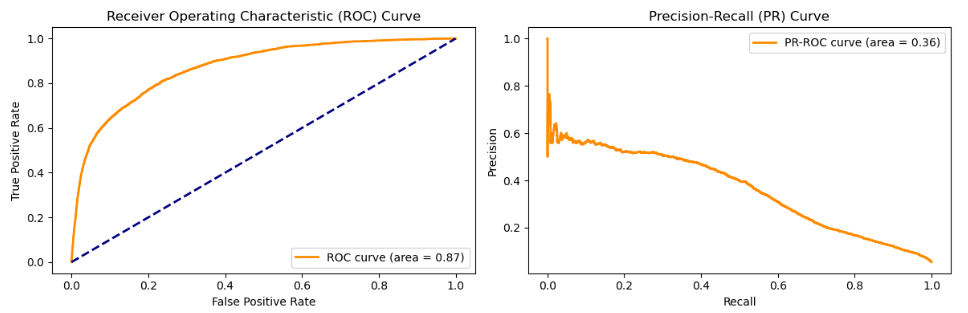
The Logistic Regression model, chosen for its simplicity and interpretability, achieved an accuracy of 80.45% in predicting heart disease risk. The confusion matrix reveals that the model accurately predicted 37,811 instances of the majority class ('No') and 1,772 instances of the minority class ('Yes'), with 8,705 and 915 misclassifications, respectively. Delving into the classification report, precision for the majority class is notably high at 98%, indicating a strong proportion of correctly predicted instances among those classified as 'No.' However, precision for the minority class is lower at 17%, signifying a higher rate of misclassifications among those predicted as 'Yes.' The recall for the majority class is 81%, demonstrating the model's ability to capture the majority of actual 'No' instances. Conversely, the recall for the minority class is 66%, indicating the model's effectiveness in capturing a substantial portion of actual 'Yes' instances. In summary, the Logistic Regression model provides a straightforward approach to heart disease risk prediction with a reasonable level of accuracy. However, its trade-off between precision and recall, especially for the minority class, highlights the need for careful consideration of model performance metrics based on the specific objectives of our predictive modeling task.



**ROC and PR-ROC curves:**

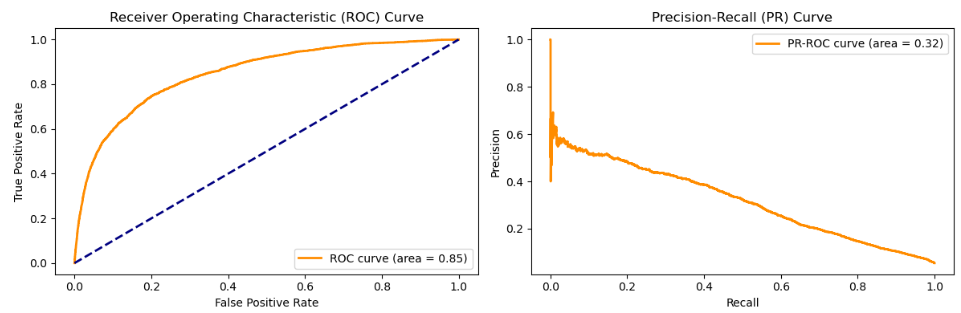
**CatBoost Model:**

* ROC-AUC: The CatBoost model achieves a high ROC-AUC of 0.87, indicating strong performance in distinguishing between true positive and false positive rates. This suggests that the model has a good ability to separate classes based on their predicted probabilities.
* PR ROC area: The PR ROC area is 0.36, signifying reasonable precision-recall trade-offs, demonstrating the model's capability to balance precision and recall.



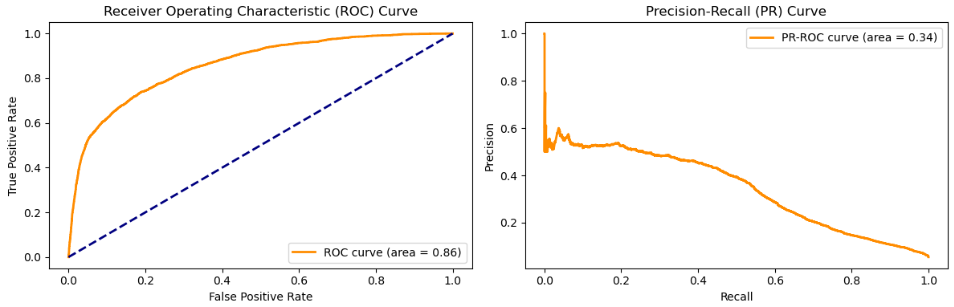
**gbm Model:**

* ROC-AUC: With a ROC-AUC of 0.85, the gbm model demonstrates commendable discriminatory power. It effectively discriminates between positive and negative instances, showcasing robust classification performance.
* PR ROC area: The PR ROC area of 0.32 indicates a reasonable balance between precision and recall for positive predictions.



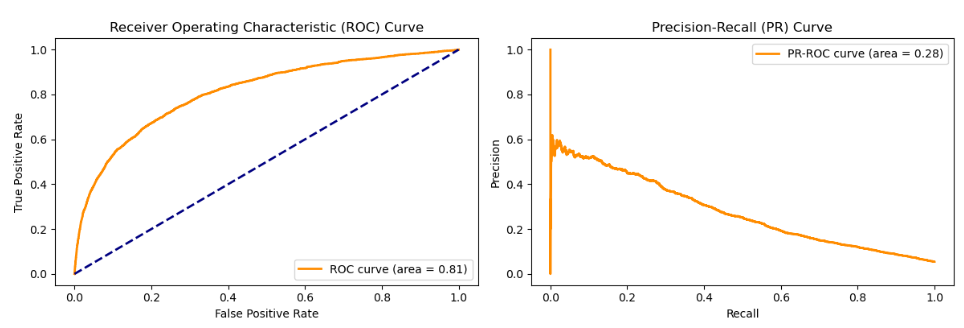
**lgbm Model:**

* ROC-AUC: The lgbm model achieves a ROC-AUC of 0.86, showcasing strong discriminatory performance. It successfully separates classes based on predicted probabilities.
* PR ROC area: With a PR ROC area of 0.34, the model maintains a good trade-off between precision and recall.



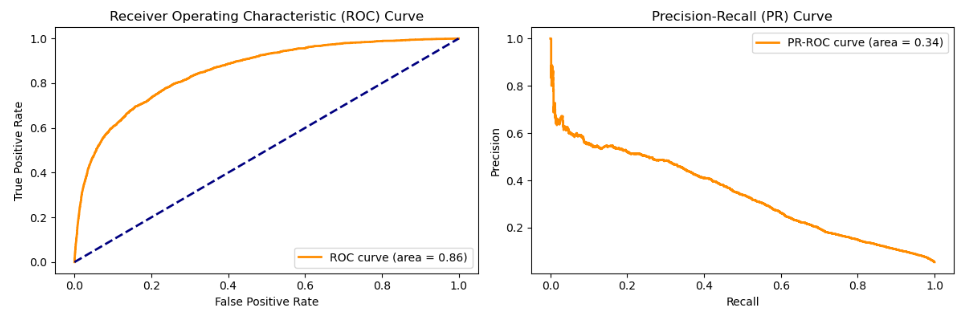
**Logistic Regression Model:**

* ROC-AUC: The logistic regression model has a ROC-AUC of 0.81, suggesting moderate discriminatory power. It distinguishes between positive and negative instances, but not as effectively as the ensemble models.
* PR ROC area: The PR ROC area of 0.28 indicates a moderate balance between precision and recall.



**xgbm Model:**

* ROC-AUC: The xgbm model achieves a high ROC-AUC of 0.86, showcasing robust discriminatory performance similar to the gbm and lgbm models.
* PR ROC area: With a PR ROC area of 0.34, the model maintains a good balance between precision and recall.



In summary, higher ROC-AUC and PR ROC area values suggest better model performance in terms of class separation and precision-recall trade-offs. The CatBoost and xgbm models demonstrate particularly strong performance in both aspects, while the logistic regression model shows moderate performance.

**Recommendations and Feature Insights:**

**CatBoost Model Dominance:** The CatBoost model emerges as the top performer with the highest accuracy and F1-score. Consider deploying CatBoost for heart attack prediction, especially when achieving a balance between precision and recall is crucial.

**Imbalanced Dataset Handling:** Given the imbalanced nature of the dataset, it's essential to prioritize models with a strong F1-score. CatBoost and XGBoost models demonstrate better performance in capturing patterns in the minority class (individuals who had a heart attack).

**Feature Importance:**

Top 5 Significant Variables:

* CatBoost Model: 'PhysicalHealthDays', 'MentalHealthDays', 'SleepHours', 'HadAngina', 'BMI'.
* Random Forest Model: 'HadAngina', 'AgeCategory', 'PhysicalHealthDays', 'SleepHours', 'HeightInMeters'.
* XGBoost Model: 'PhysicalHealthDays', 'MentalHealthDays', 'SleepHours', 'HadAngina', 'BMI'.

These variables highlight the importance of both physical and mental health metrics, sleep duration, and indicators like 'HadAngina' in predicting heart attacks.

**Threshold Consideration:** The chosen classification threshold plays a pivotal role in balancing precision and recall. Thresholds around 0.36 for CatBoost, 0.32 for gbm, 0.34 for lgbm and XGBoost, and 0.28 for Logistic Regression offer a balanced approach, but the choice depends on specific requirements.

**Conclusion:**

In conclusion, our extensive exploration of machine learning models for heart attack prediction has provided valuable insights into their performance and applicability. The CatBoost model emerged as a standout performer with the highest accuracy and F1-Score, underlining its effectiveness in this predictive task. While Logistic Regression exhibited lower accuracy, its interpretability remains crucial in certain contexts. The consideration of imbalanced data emphasized the significance of F1-Score in evaluating model effectiveness. Furthermore, the ROC-Area and PR-ROC Area metrics, along with the strategic choice of classification thresholds, offered nuanced perspectives on model capabilities. Ultimately, the diverse set of models allowed for a comprehensive understanding, highlighting the need for a balanced approach considering multiple metrics and specific use-case requirements in predictive healthcare modeling.

**References:**

Frolov, V. (2021, December 15). Imbalanced classification for medical diagnosis - Analytics Vidhya - medium. Medium.

<https://medium.com/analytics-vidhya/imbalanced-classification-for-medical-diagnosis-75dfcaa783d3>

Harshdeep Singh. (2018, November 3). Understanding Gradient Boosting Machines

<https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab>

Indicators of heart Disease (2022 UPDATE). (2023, October 12). Kaggle. <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease/data>

Jason Brownlee. (2023, October 11). How to Use ROC Curves and Precision-Recall Curves for Classification in Python

<https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/>

John T. Hancock., & Taghi M. Khoshgoftaar. (2020, November 04). CatBoost for big data: an interdisciplinary review.

<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00369-8>

Pushkar Mandot. (2017, August 17). What is LightGBM, How to implement it? How to fine tune the parameters?

<https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc>

Vijay Kanade. (2022, April 18). What Is Logistic Regression? Equation, Assumptions, Types, and Best Practices

<https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/>

Jason Brownlee. (January 5, 2021). A Gentle Introduction to Threshold-Moving for Imbalanced Classification

<https://machinelearningmastery.com/threshold-moving-for-imbalanced-classification/>