****

**Module 6 Assignment - Final Project**

Jessi Yang

Chengjun Jiang

Jinsheng Wei

Junchen Yi

ALY6040: Data Mining 80375

CPS, Northeastern University

Dr. Kasun Samarasinghe

May 14, 2024

**Introduction**

Heart disease is the leading cause of death globally and there is a need to enhance early detection and prevention methods. This project employs machine learning techniques to predict heart disease risk using the Behavioral Risk Factor Surveillance System (BRFSS) dataset. By developing multiple classification models, we aim to identify the major lifestyle and health predictors of heart disease. Our approach helps in early identification of at-risk populations, thus guiding preventive health measures more effectively. Through our analysis, we utilize advanced algorithms such as Decision Trees, Random Forests, XGBoost, LightGBM and CatBoost to improve our understanding and management of heart disease risk factors.

Our analysis is based on the BRFSS dataset, which provides a wealth of data on the health behaviors and conditions of U.S. residents. The dataset includes telephone survey data from more than 308,854 respondents across 19 variables and is an important resource for understanding cardiovascular risk factors. Key variables used include age, body mass index, health score, and unhealthy score, each selected for its relevance in predicting cardiovascular disease. The target variable "heart\_disease" was coded in binary, with 1 representing people diagnosed with heart disease and 0 representing people without heart disease, thus allowing our machine learning model to perform a clear, focused analysis. This setup provides a solid foundation for our predictive analytics, aiming to provide actionable insights into the prevention and management of heart disease based on lifestyle and health data.

**Methodology**

Our research utilizes a variety of machine learning classifiers to predict heart disease risk, each selected for its unique ability to handle complex data:

1. Decision Tree Classifier: Simple and interpretable, ideal for capturing nonlinear relationships.

2. Random Forest Classifier: An ensemble method that combines multiple decision trees to improve accuracy and stability, ideal for unbalanced datasets.

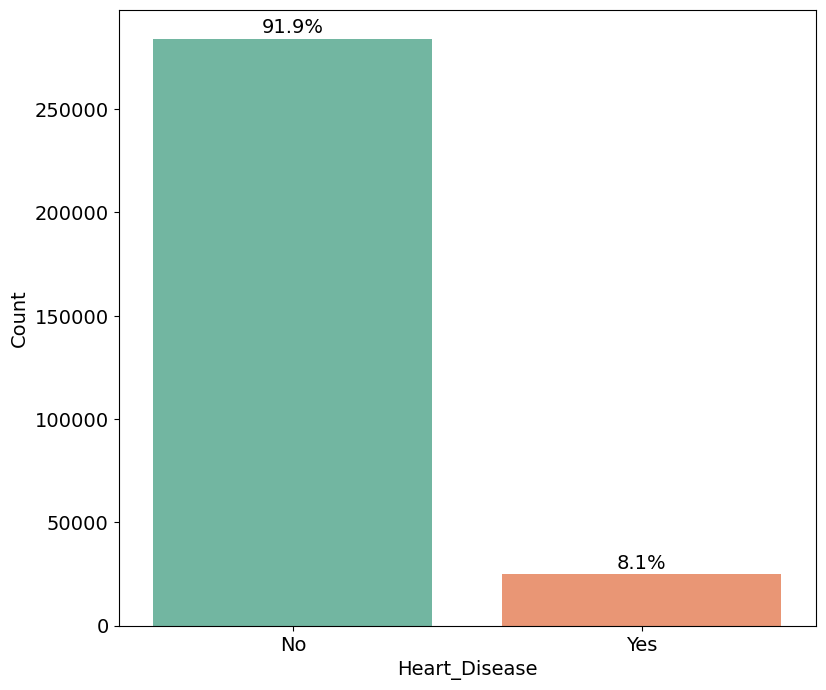
3. XGBoost Classifier: A gradient boosting framework known for its high performance and speed to optimize computational speed and model efficiency.

4. LightGBM Classifier: Optimized for speed and efficiency, this gradient boosting method can effectively handle large datasets with low memory usage.

5. CATBoost Classifier: Specializes in dealing with categorical variables with minimal preprocessing, increasing training and prediction speed while maintaining accuracy.

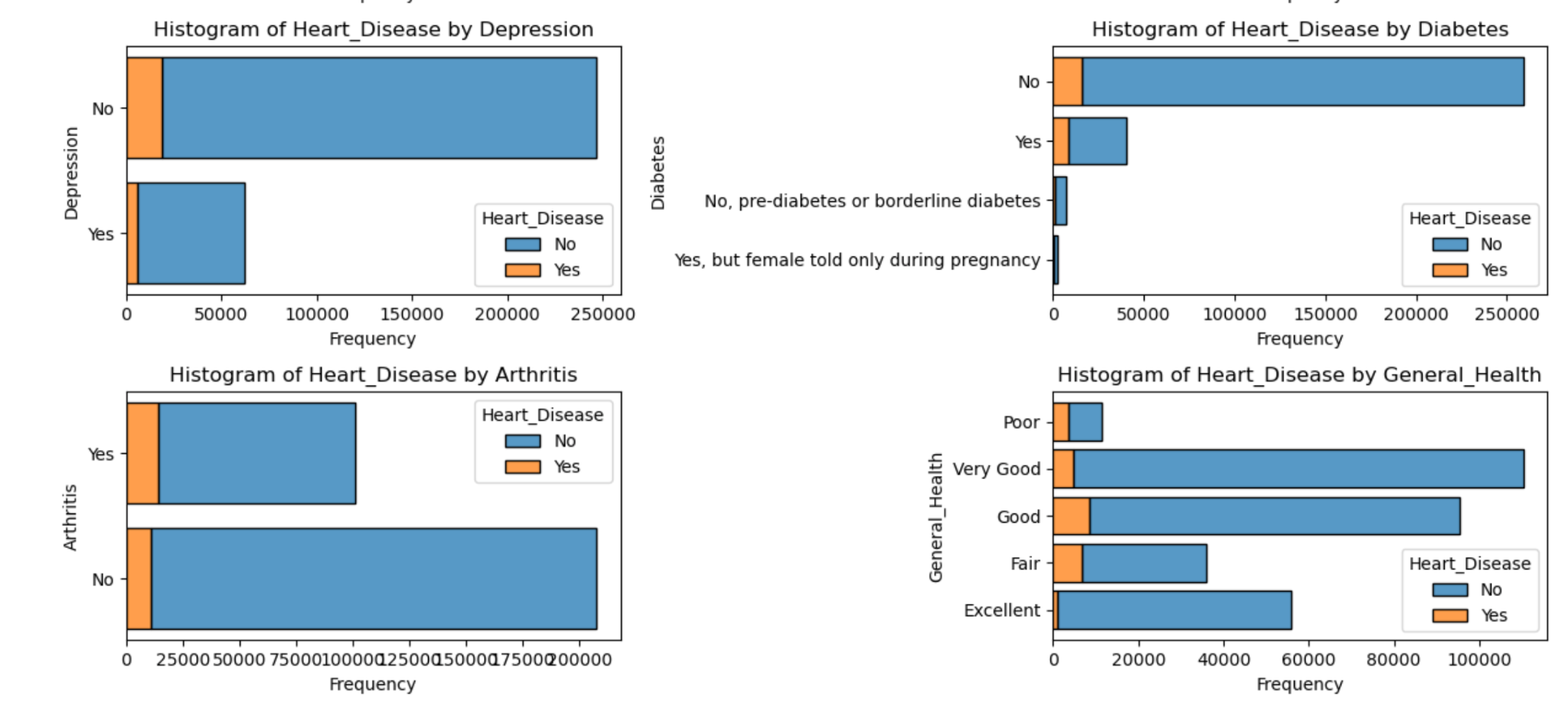
**Analysis**

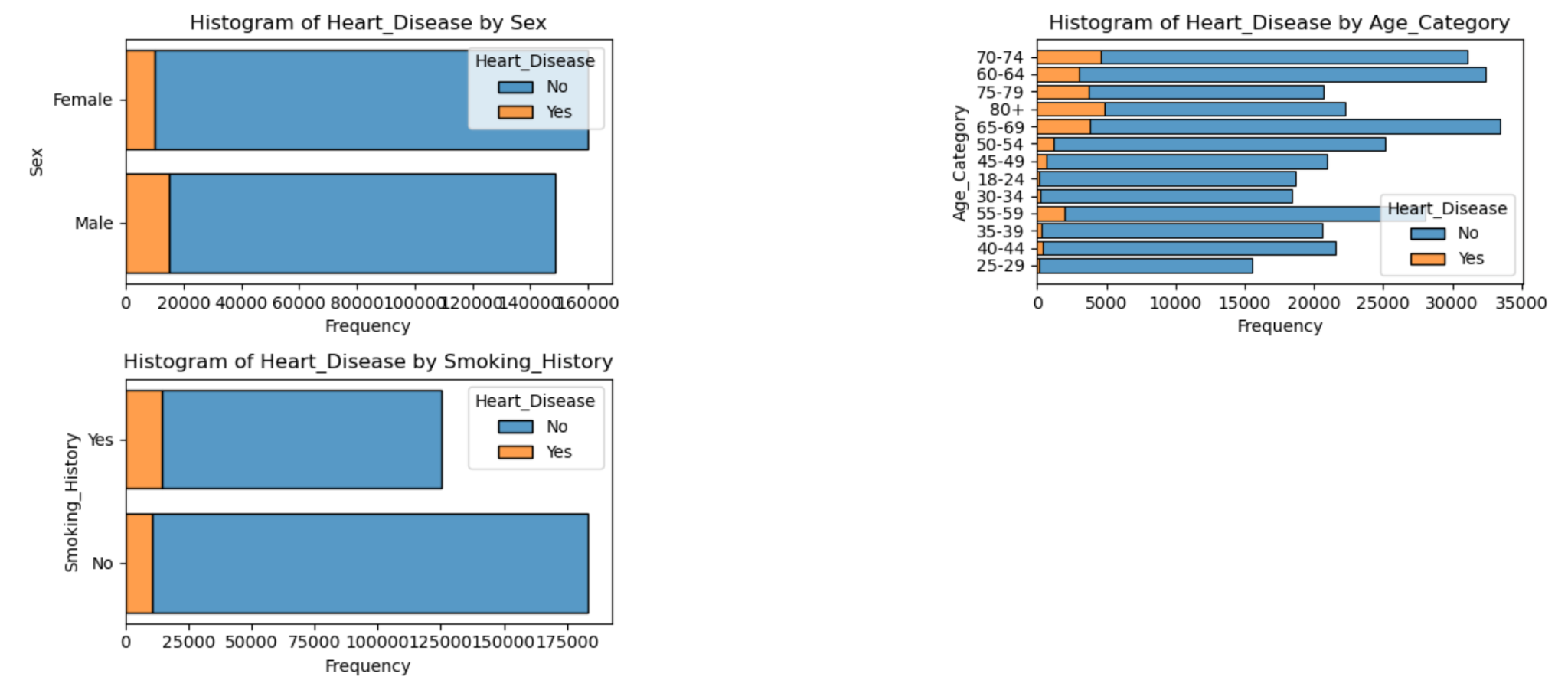
During the preliminary analysis phase, Exploratory Data Analysis (EDA) revealed a significant imbalance in the target variable "Heart Disease" in the BRFSS dataset. As shown in the bar chart, our findings indicate that 91.9% of respondents did not have heart disease, while only 8.1% of respondents were reported to have heart disease. This imbalance poses a challenge to predictive modeling because it can bias machine learning algorithms toward the majority, thereby potentially underestimating the risk of the minority.



Our in-depth exploratory data analysis reveals important associations between various lifestyle and health factors and heart disease:

* Age category: The data shows that the incidence of heart disease rises with age, making age an important risk factor.
* Smoking history: There is a significant correlation between smoking and elevated risk of heart disease, highlighting smoking as a major causative factor.
* General health status: Individuals who report poorer general health also have a higher incidence of heart disease, suggesting that self-assessed health status can be used as an indicator of cardiovascular risk.
* Diabetes: The strong link between diabetes and heart disease is clear, with diabetics having significantly higher rates of heart disease, highlighting the interconnectedness of the two conditions.





**Feature Engineering**

1. Reducing the size of the age profile:

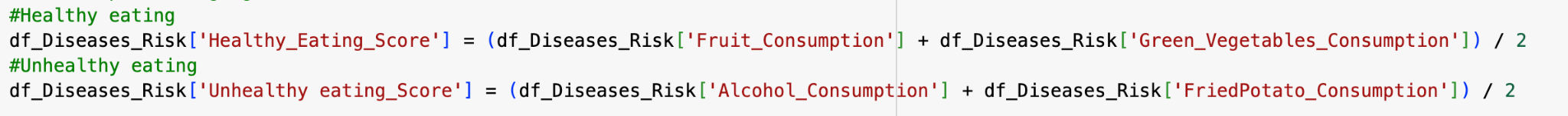
- In order to improve computational efficiency and reduce complexity, we reclassified age variables that originally spanned a wide range into broader categories. This step consisted of mapping the original detailed age ranges into generic groups such as "young", "adult", "middle-aged", "near-elderly", and "elderly". This categorization simplifies the dataset and allows for a clearer understanding of age-related trends in heart disease prevalence without the need for single coding for each year.



2. Creating new features - healthy and unhealthy diet scores:

- We innovatively developed two new variables - Healthy Diet Score and Unhealthy Diet Score - to quantitatively assess dietary habits. The Healthy Eating Score is calculated based on the average intake of fruits and green vegetables and represents a positive eating behavior. In contrast, the Unhealthy Eating Score is calculated based on average intake of alcohol and fried potatoes and reflects less healthy dietary choices. These scores are intended to provide a direct metric for assessing the impact of diet on heart disease risk, capturing the essence of dietary patterns in a form that is easy to use by machine learning models.

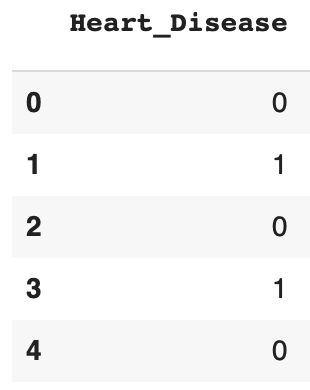




**Pre-Processing**

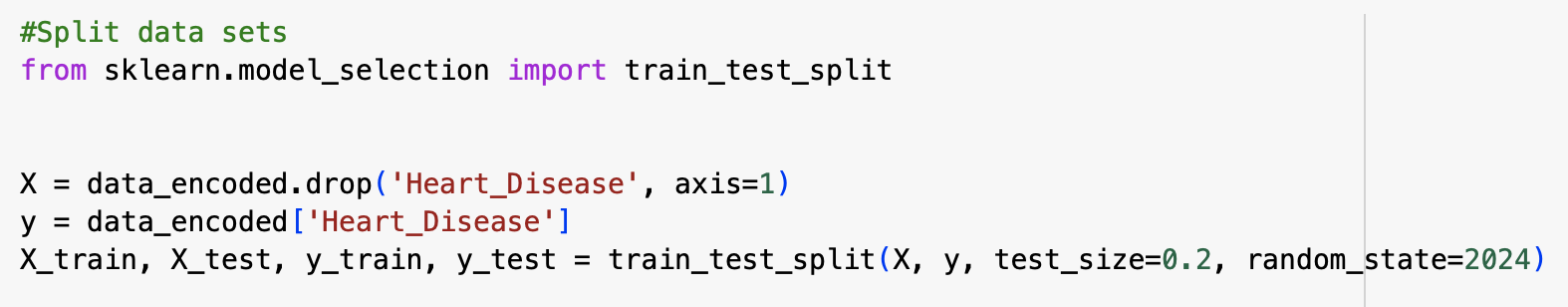
1. Feature encoding:

- The 'heart disease' variable was converted from categorical ('yes', 'no') to numeric format (1, 0) to facilitate processing by machine learning algorithms. This binary encoding simplifies the goal of the classification model, allowing it to learn and predict the presence of heart disease more efficiently.



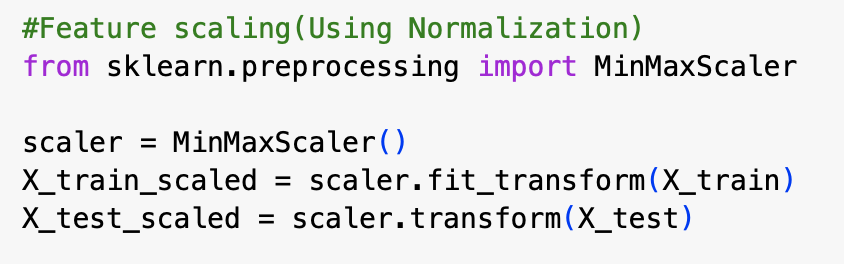
2. Splitting the dataset:

- The dataset is divided into a training set (80%) and a test set (20%). This division ensures that the model can be trained on a large portion of the data while keeping a separate subset for evaluating its performance. This division also helps to validate the model's ability to generalize to new and unseen data, reducing the risk of overfitting.



3. Feature scaling:

- We use MinMaxScaler to normalize feature values between 0 and 1. This step is crucial because it prevents any single feature from dominating the model due to its scale, thus ensuring that all features contribute equally to the prediction. Normalization is especially important for distance-based algorithms, as scale can significantly affect performance.



4. SMOTE for dealing with class imbalance:

Given the severe class imbalance found in the EDA process, where the majority of instances are negative (no heart disease), we employ the Synthetic Minority Oversampling Technique (SMOTE) to balance the classes.SMOTE works by creating synthetic samples instead of performing substitution oversampling, which can lead to overfitting. This technique improves the model's ability to recognize minority classes and improves overall classification performance on unbalanced datasets.



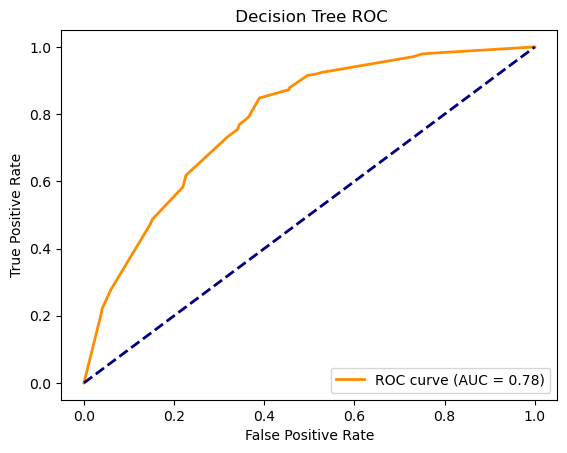
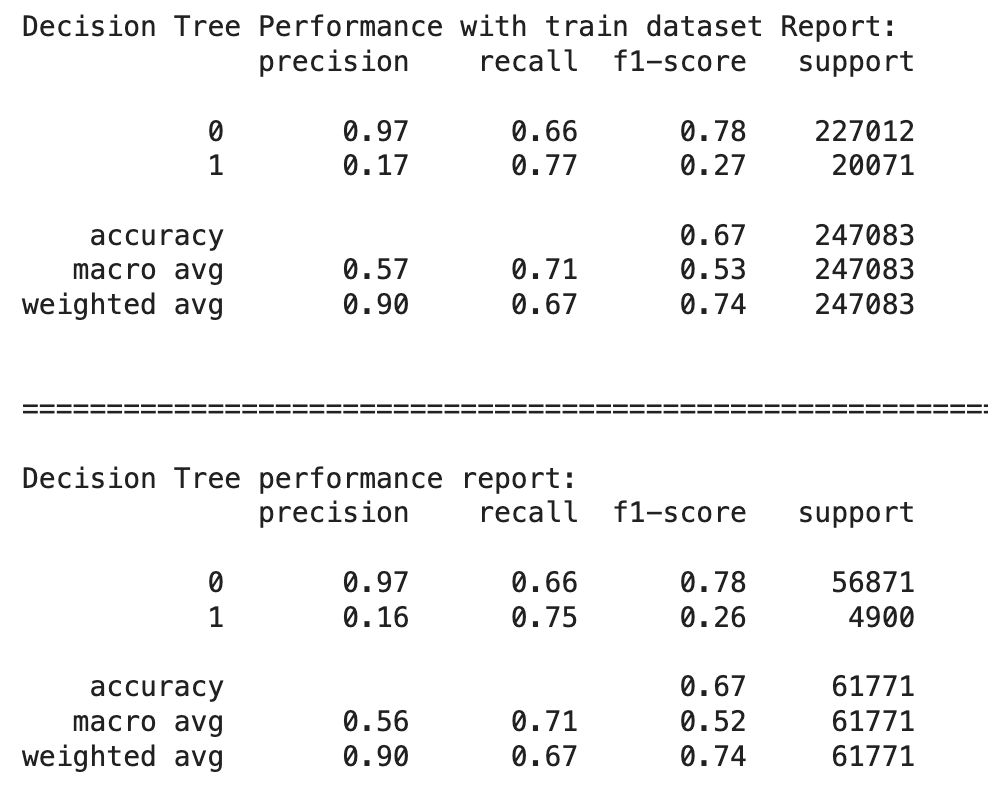
**Modelling**

To determine which model performs better on the imbalanced dataset where class 1 represents patients with heart disease and class 0 represents patients without heart disease, we need to focus on metrics that handle class imbalance effectively. Here are the key metrics to consider:

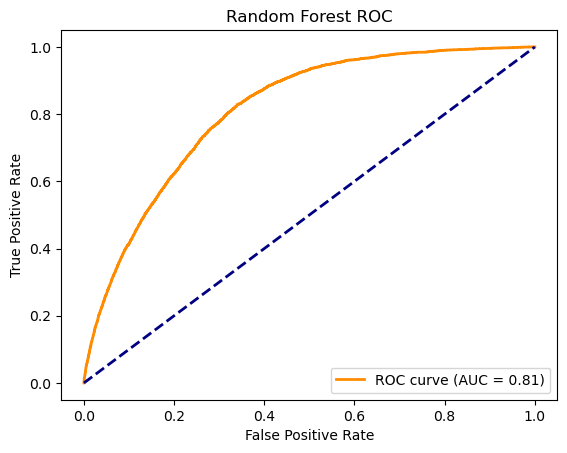
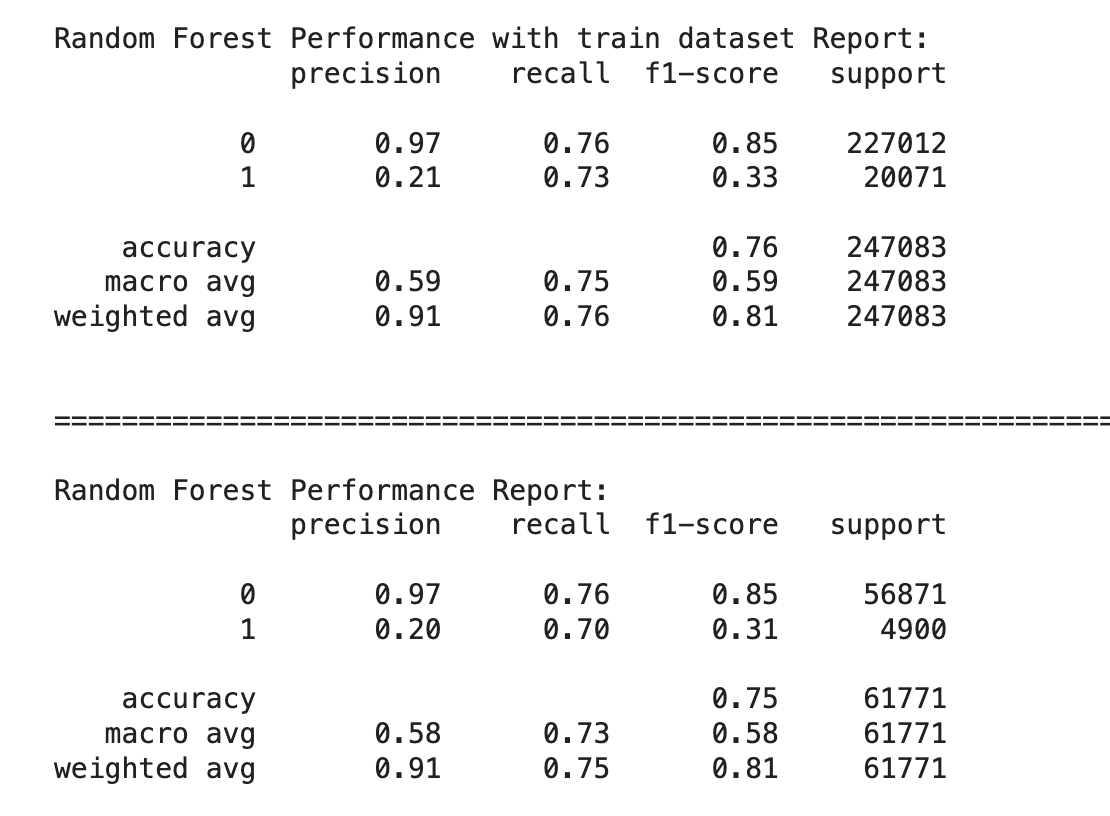
1. Precision for Class 1: This indicates the proportion of true positive predictions among all positive predictions made for class 1 (heart disease patients). It is crucial when the cost of false positives is high.
2. Recall (Sensitivity) for Class 1: This indicates the proportion of true positives among all actual positives for class 1. It is essential when the cost of false negatives is high.
3. F1-Score for Class 1: This is the harmonic mean of precision and recall for class 1 and provides a balanced measure of both precision and recall.
4. Area Under the ROC Curve (AUC-ROC): This measures the model's ability to discriminate between classes across all thresholds and is particularly useful for imbalanced datasets.

Here are our five models without hyperparameter tuning, which provide the basic performance of different models in addressing our problem.

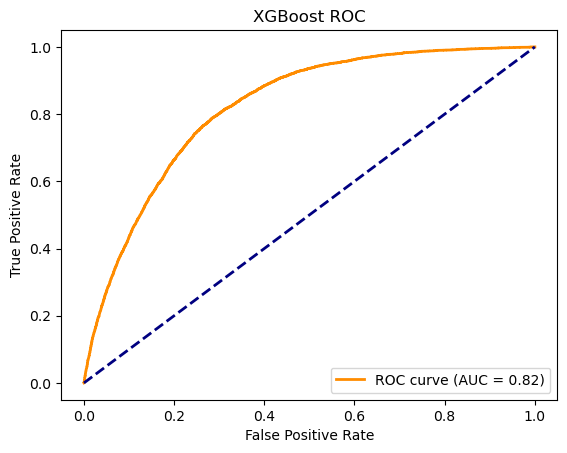
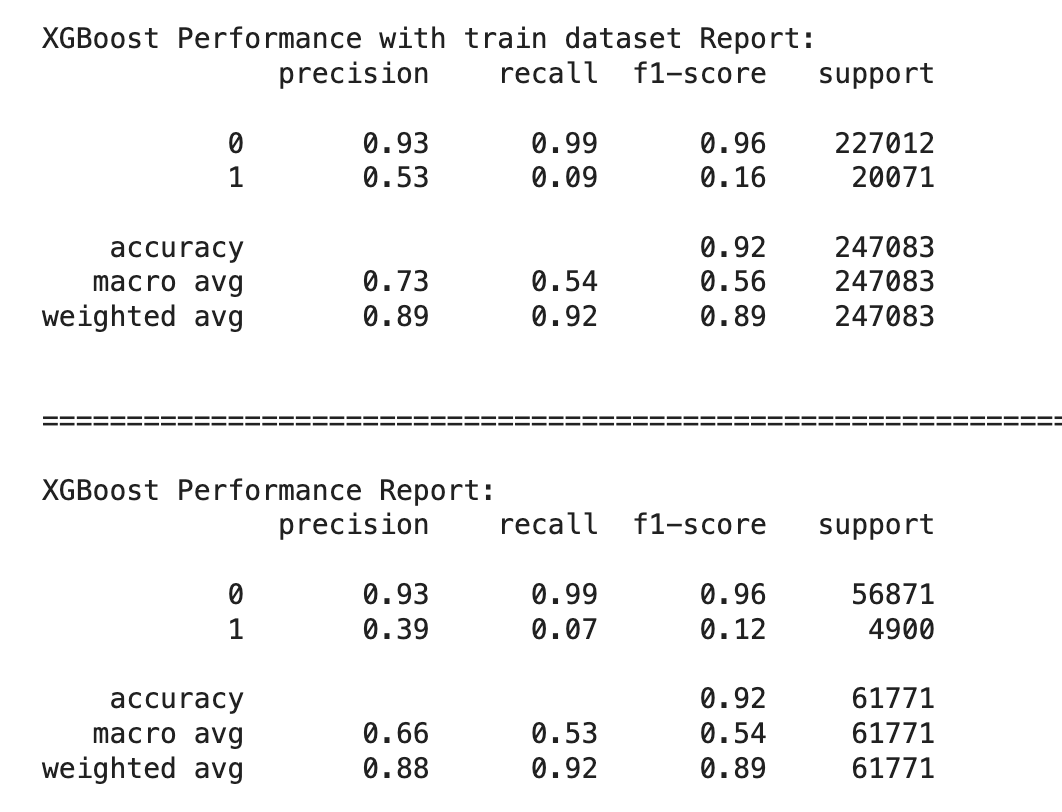
Decision tree:



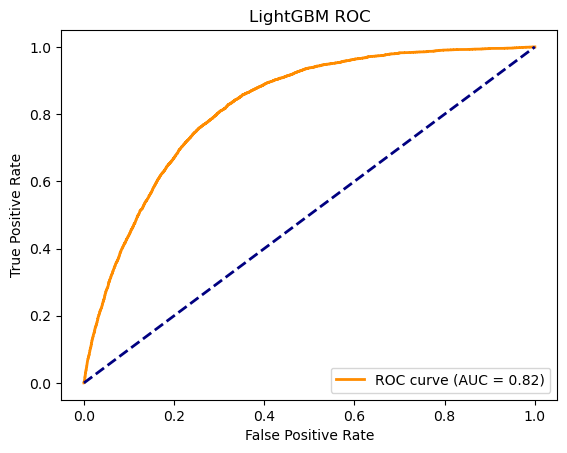
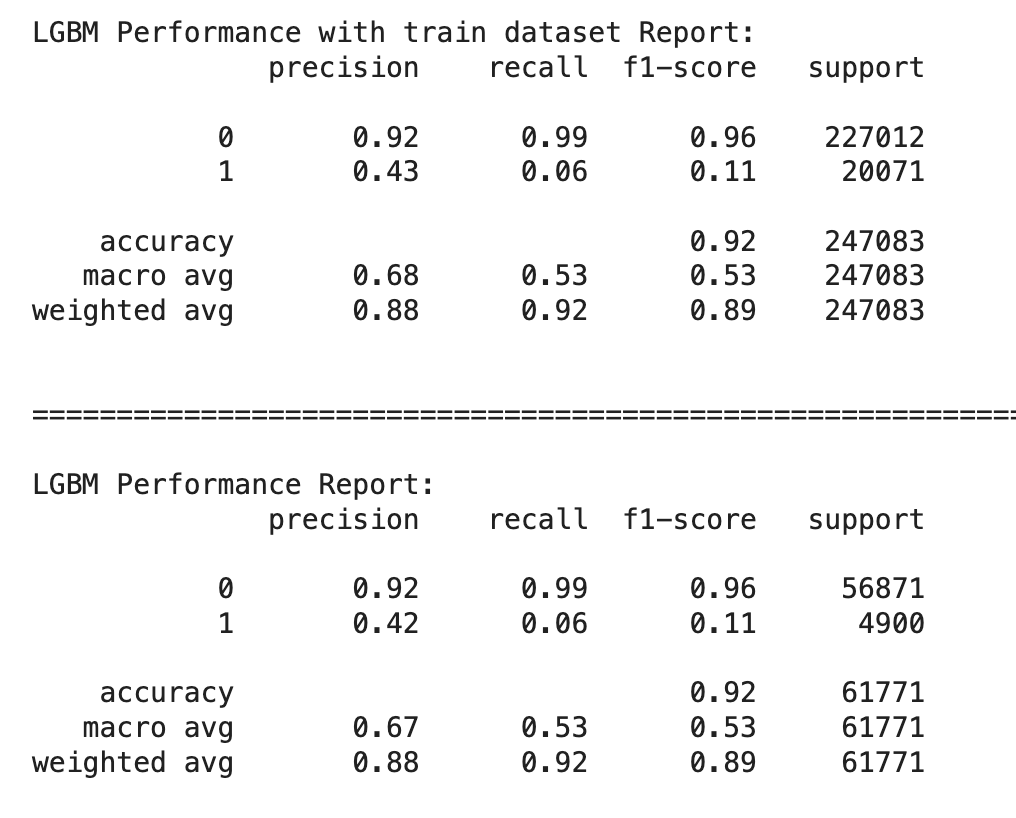
Random forest:



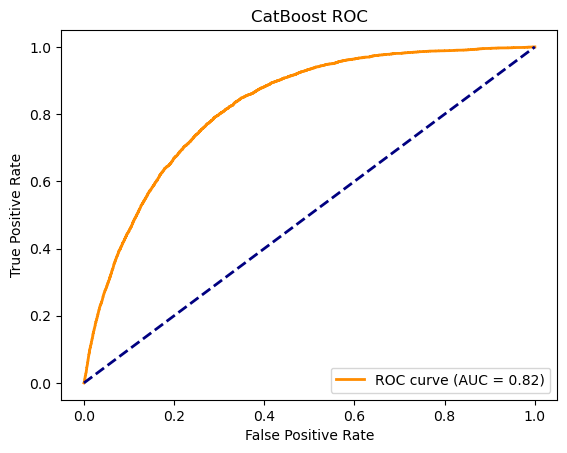
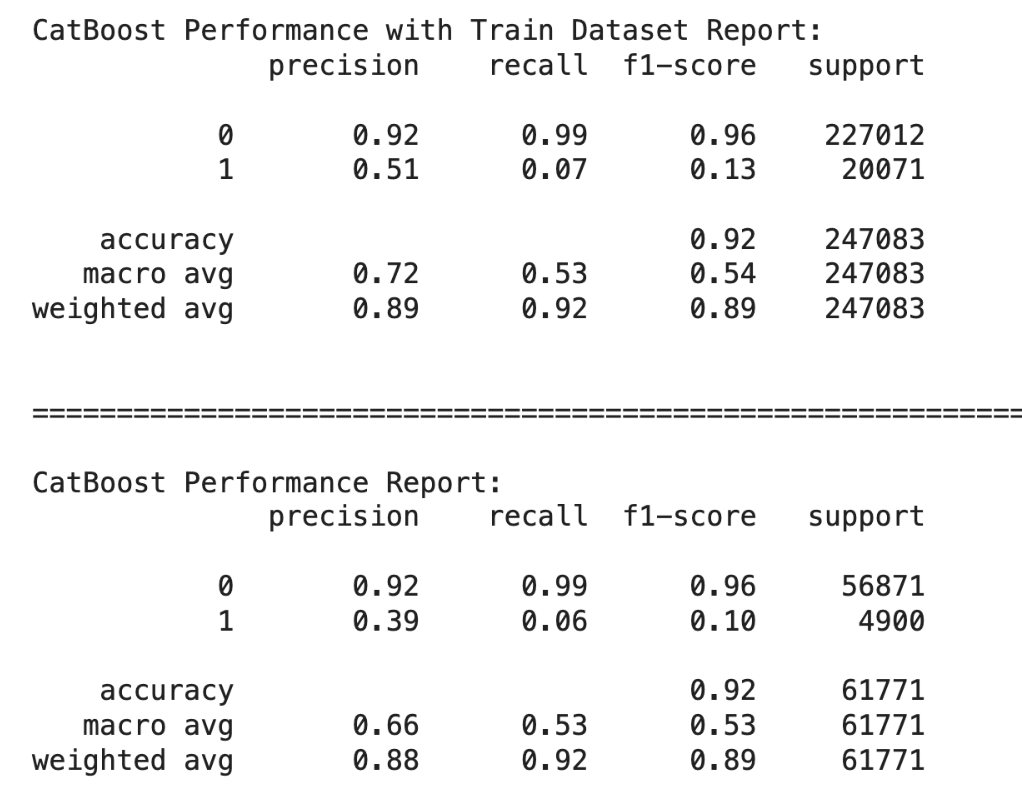
XGBoost:



LightGBM:

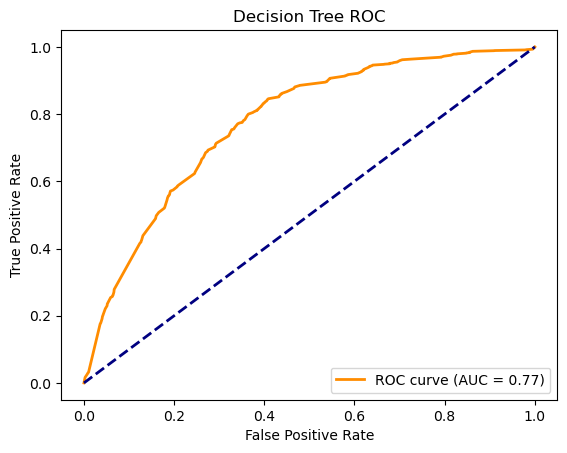
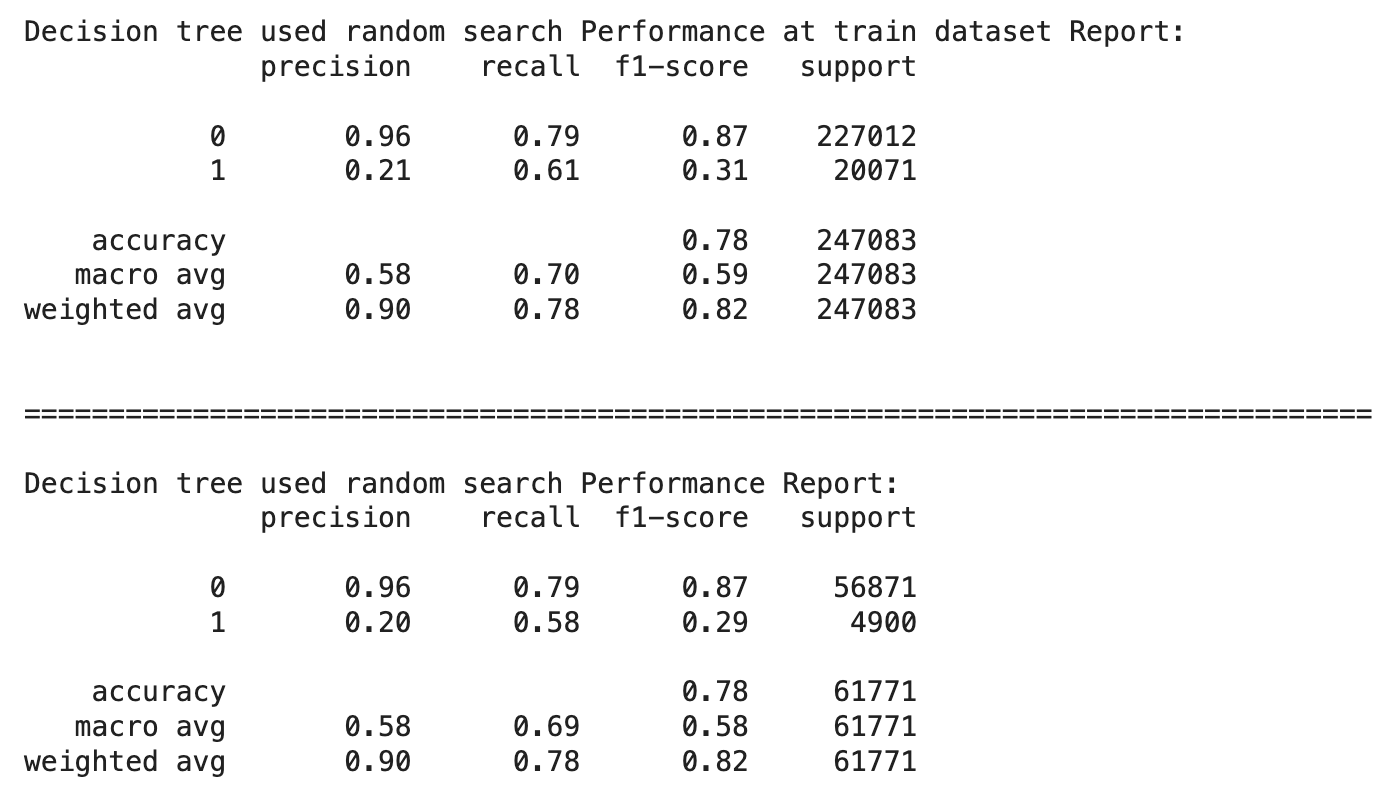


Catboost:



Next, I will list the five models after hyperparameter tuning to see how they perform on the metrics we are interested in.

1.Decision Tree Model



Precision for Class 1: 0.20

This indicates the proportion of true positive predictions among all positive predictions made for class 1 (heart disease patients). It is crucial when the cost of false positives is high.

Recall (Sensitivity) for Class 1: 0.58

This indicates the proportion of true positives among all actual positives for class 1. It is essential when the cost of false negatives is high.

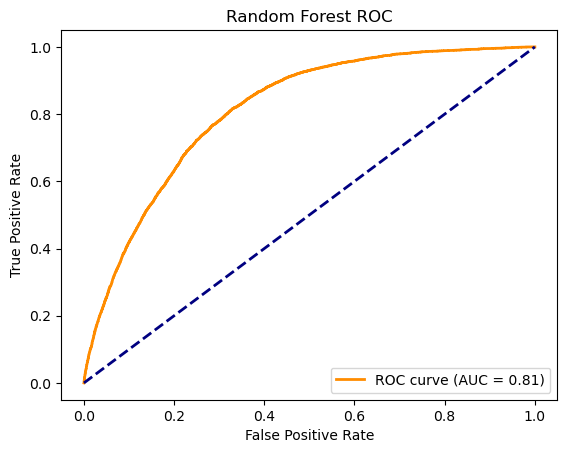
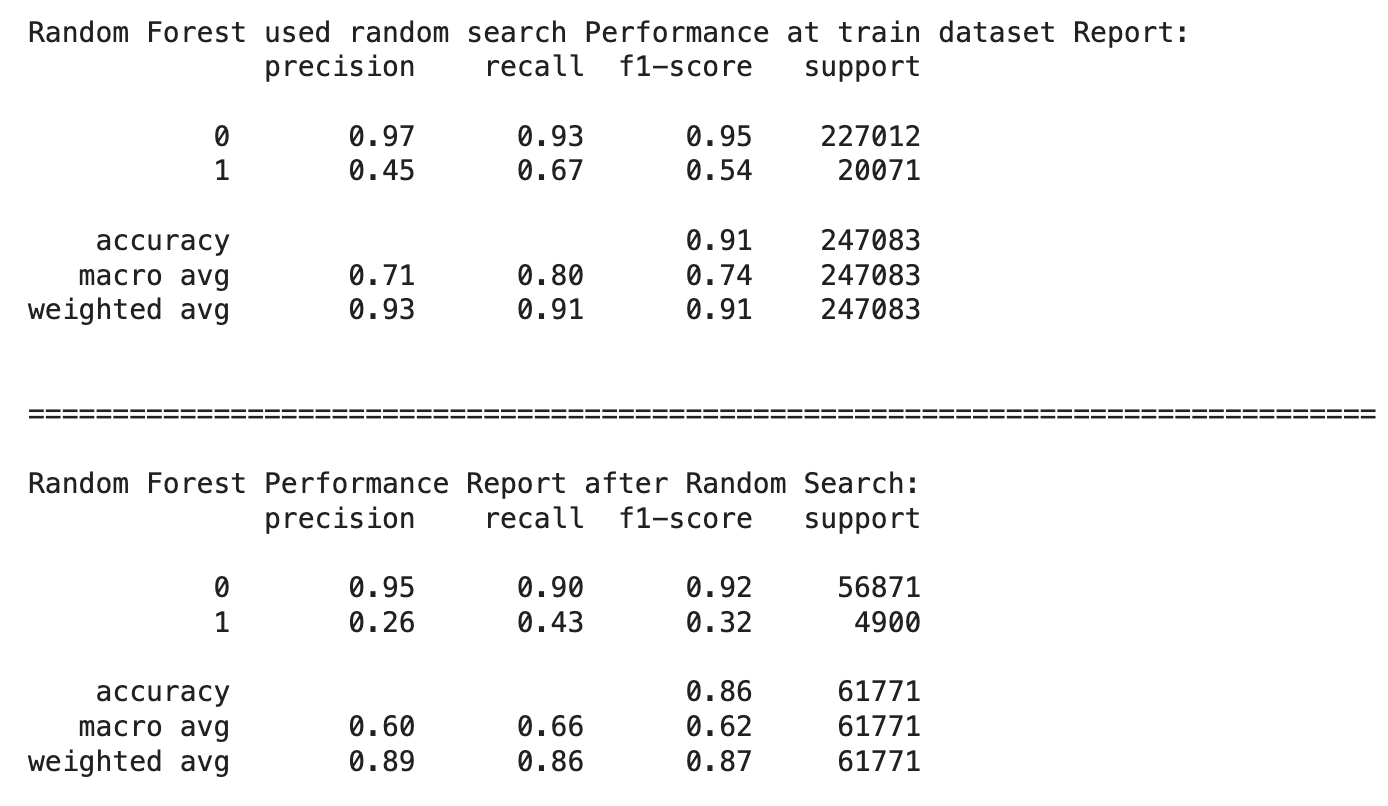
F1-Score for Class 1: 0.29

This is the harmonic mean of precision and recall for class 1 and provides a balanced measure of both precision and recall.

Area Under the ROC Curve (AUC-ROC): 0.77

This measures the model's ability to discriminate between classes across all thresholds and is particularly useful for imbalanced datasets.

2.Random Forest



Precision for Class 1: 0.26

This indicates the proportion of true positive predictions among all positive predictions made for class 1 (heart disease patients). It is crucial when the cost of false positives is high.

Recall (Sensitivity) for Class 1: 0.43

This indicates the proportion of true positives among all actual positives for class 1. It is essential when the cost of false negatives is high.

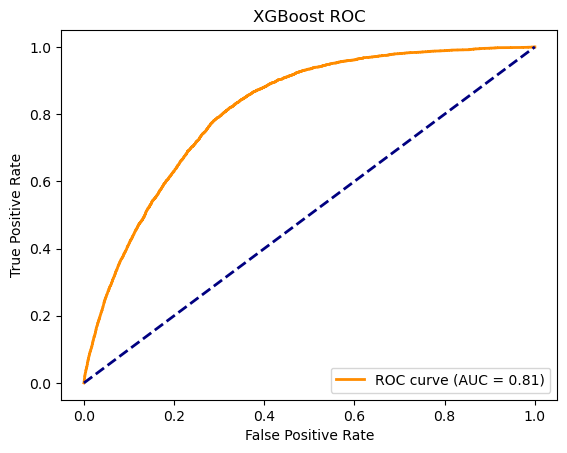
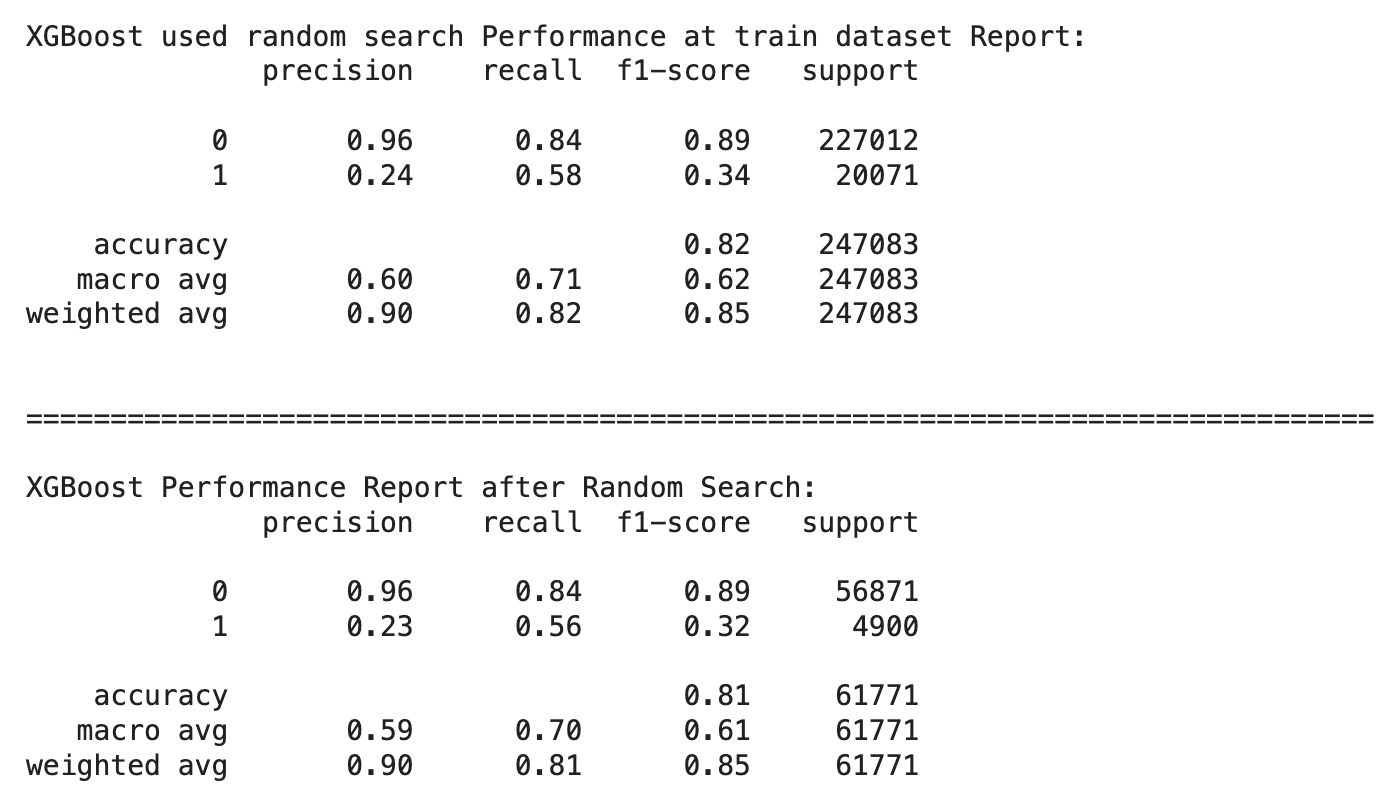
F1-Score for Class 1: 0.32

This is the harmonic mean of precision and recall for class 1 and provides a balanced measure of both precision and recall.

Area Under the ROC Curve (AUC-ROC): 0.81

This measures the model's ability to discriminate between classes across all thresholds and is particularly useful for imbalanced datasets.

3.XGBoost



Precision for Class 1: 0.23

This indicates the proportion of true positive predictions among all positive predictions made for class 1 (heart disease patients). It is crucial when the cost of false positives is high.

Recall (Sensitivity) for Class 1: 0.56

This indicates the proportion of true positives among all actual positives for class 1. It is essential when the cost of false negatives is high.

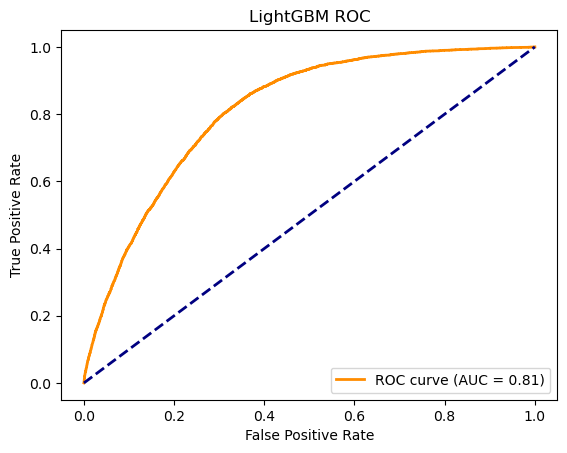
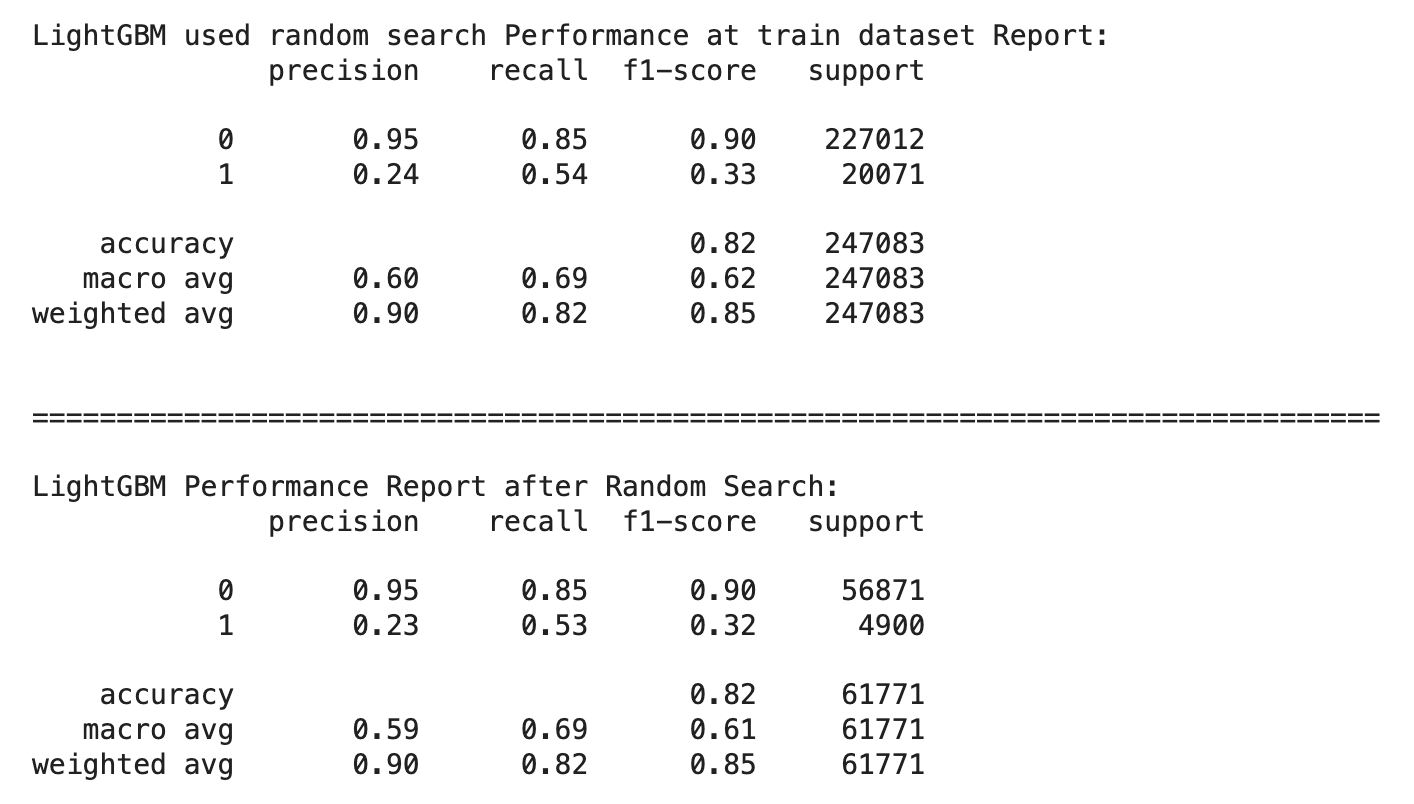
F1-Score for Class 1: 0.32

This is the harmonic mean of precision and recall for class 1 and provides a balanced measure of both precision and recall.

Area Under the ROC Curve (AUC-ROC): 0.81

This measures the model's ability to discriminate between classes across all thresholds and is particularly useful for imbalanced datasets.

4.LightGBM



Precision for Class 1: 0.23

This indicates the proportion of true positive predictions among all positive predictions made for class 1 (heart disease patients). It is crucial when the cost of false positives is high.

Recall (Sensitivity) for Class 1: 0.53

This indicates the proportion of true positives among all actual positives for class 1. It is essential when the cost of false negatives is high.

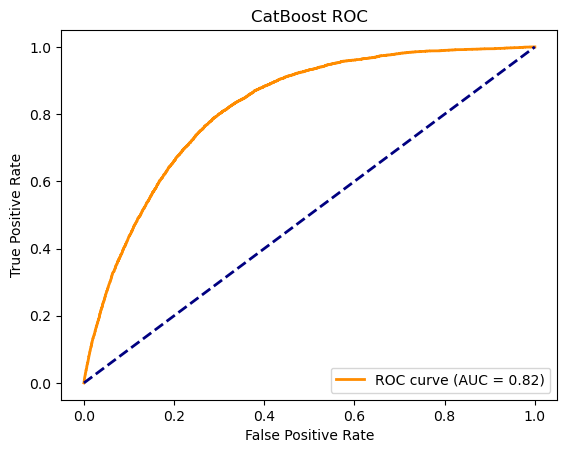
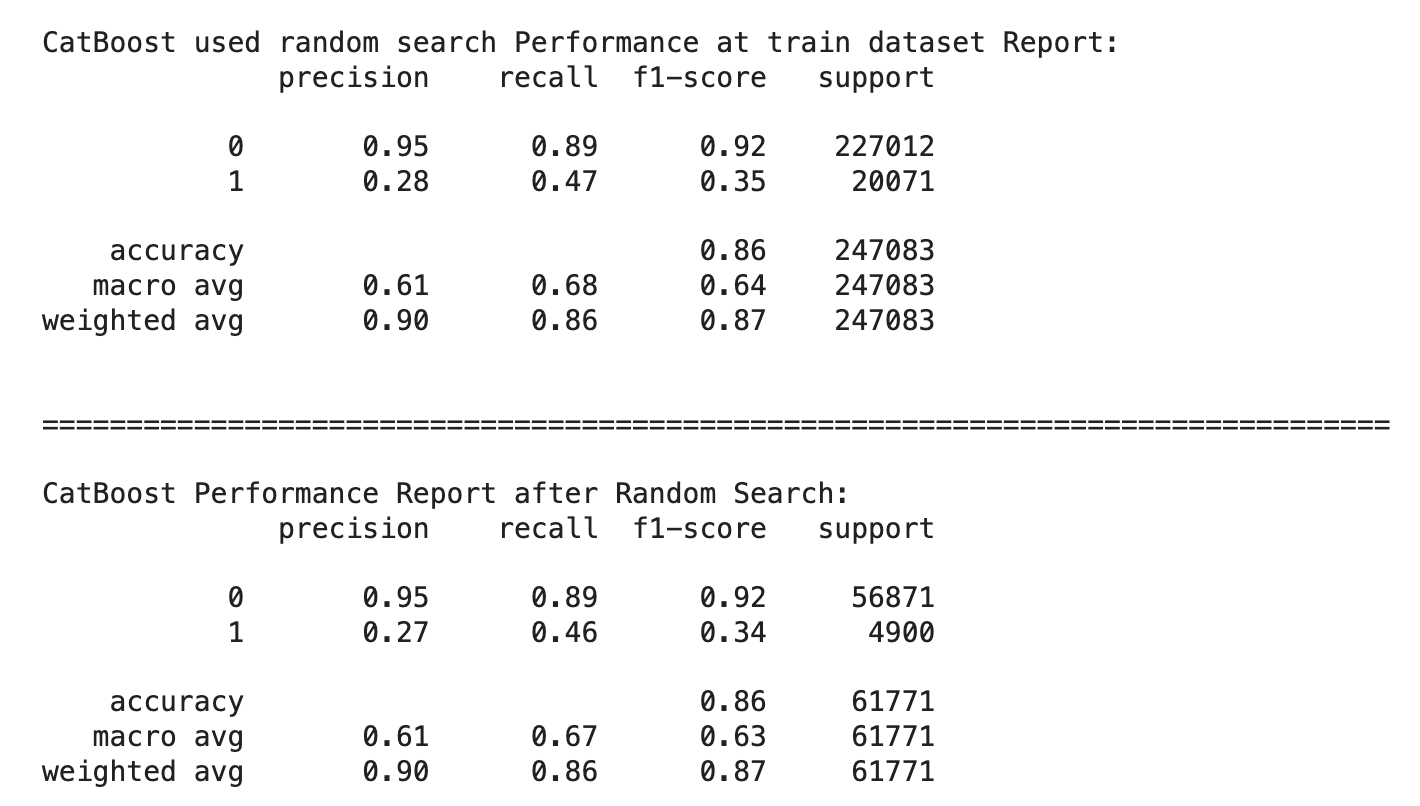
F1-Score for Class 1: 0.32

This is the harmonic mean of precision and recall for class 1 and provides a balanced measure of both precision and recall.

Area Under the ROC Curve (AUC-ROC): 0.81

This measures the model's ability to discriminate between classes across all thresholds and is particularly useful for imbalanced datasets.

5.Catboost



Precision for Class 1: 0.27

This indicates the proportion of true positive predictions among all positive predictions made for class 1 (heart disease patients). It is crucial when the cost of false positives is high.

Recall (Sensitivity) for Class 1: 0.46

This indicates the proportion of true positives among all actual positives for class 1. It is essential when the cost of false negatives is high.

F1-Score for Class 1: 0.34

This is the harmonic mean of precision and recall for class 1 and provides a balanced measure of both precision and recall.

Area Under the ROC Curve (AUC-ROC): 0.82

This measures the model's ability to discriminate between classes across all thresholds and is particularly useful for imbalanced datasets.

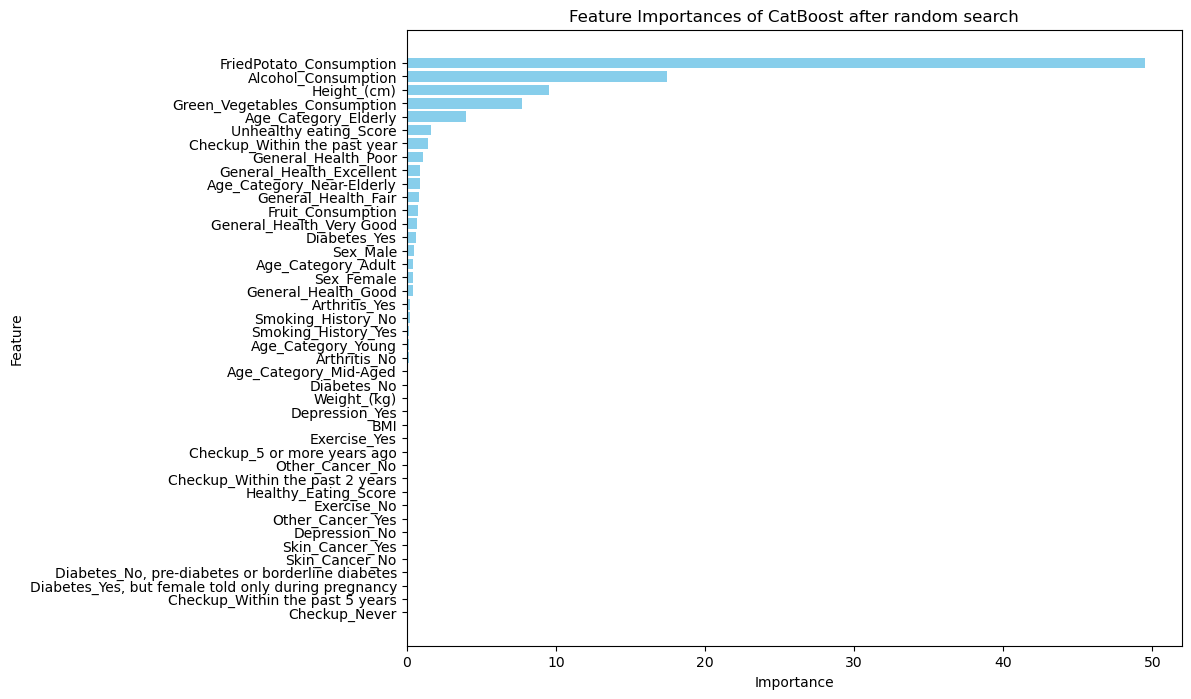
Based on the metrics provided:

CatBoost has the highest F1-Score (0.34) for class 1, indicating it has a good balance between precision and recall for predicting heart disease patients.

CatBoost also has the highest AUC-ROC (0.82), suggesting it performs the best in distinguishing between patients with and without heart disease across all thresholds.

Therefore, for this imbalanced dataset, CatBoost appears to be the best-performing model based on both the F1-Score and AUC-ROC.

We also want to understand which specific variables play a role in the CatBoost model, so we have plotted a feature importance table.

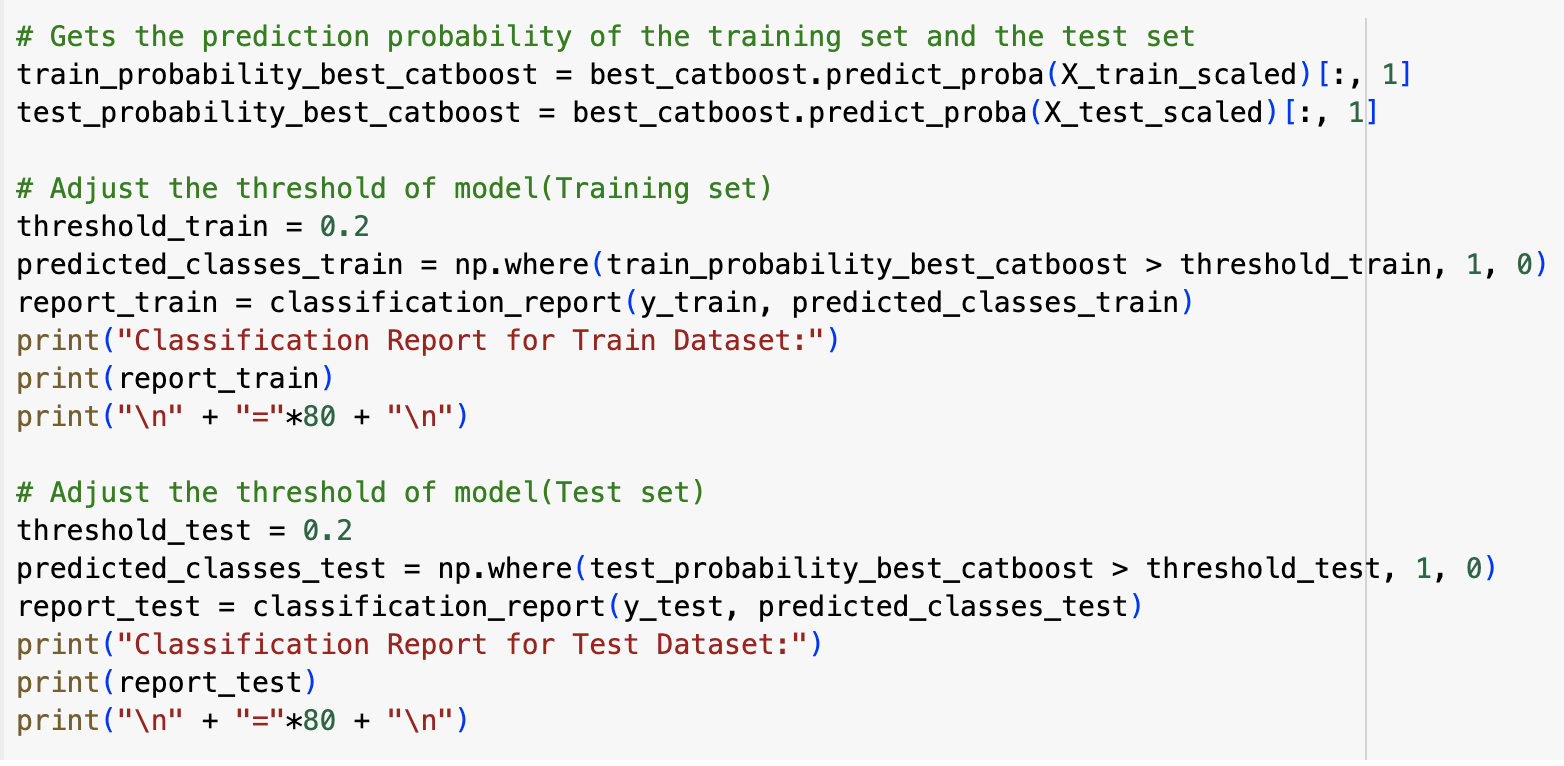


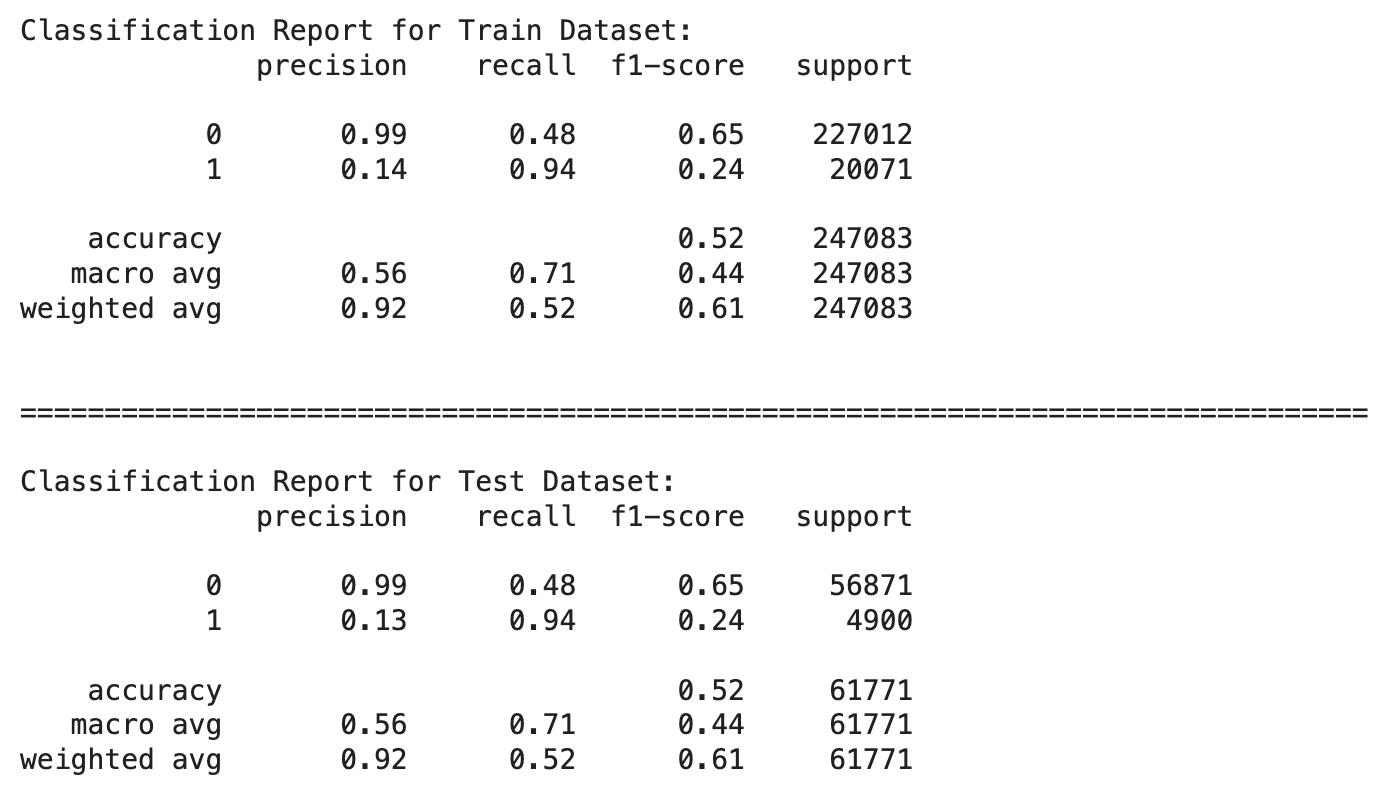
It shows that

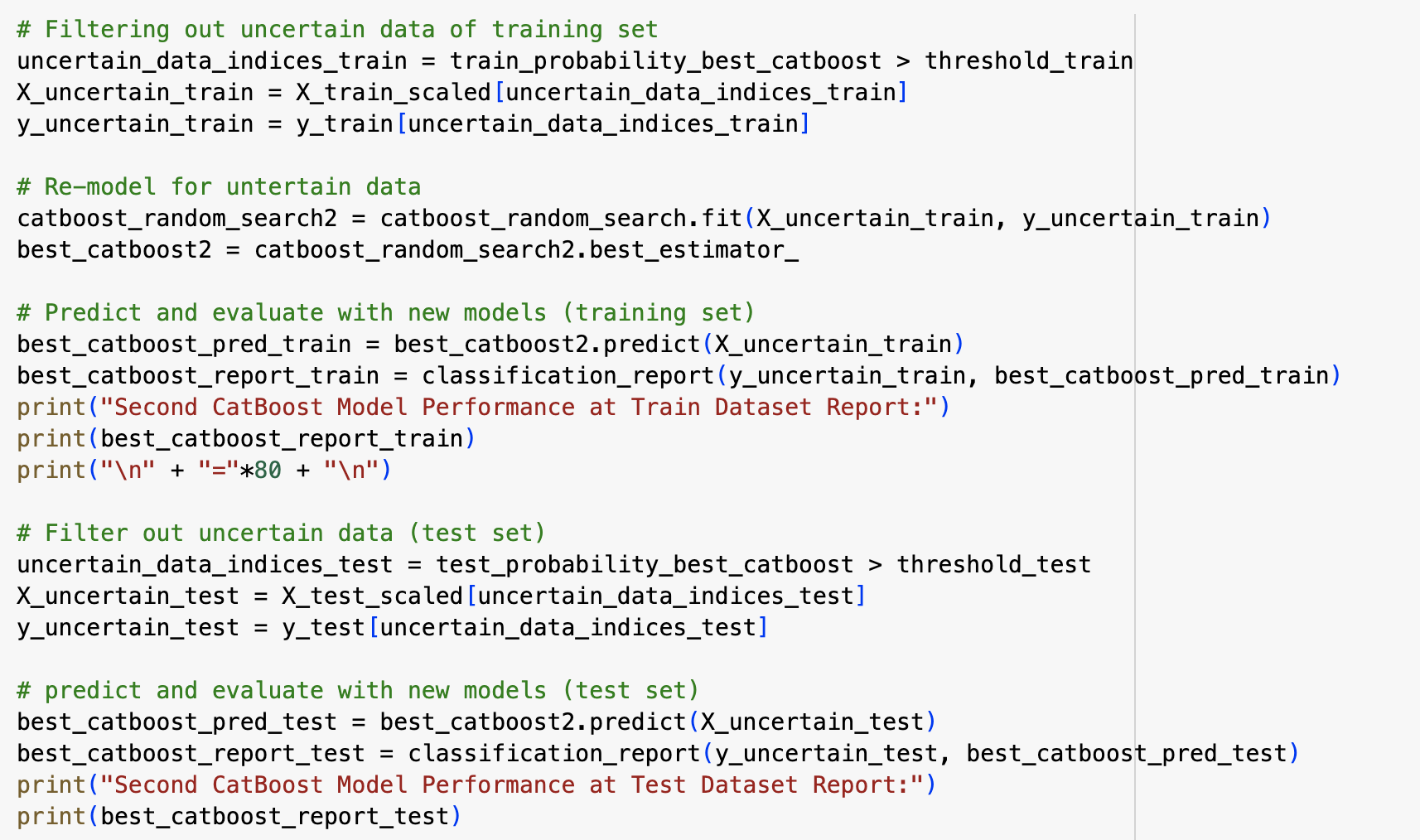
1. **FriedPotato\_Consumption:** High consumption of fried potatoes shows the impact of high-fat diets on heart health.
2. **Alcohol\_Consumption:** strong correlation between alcohol consumption and heart disease risk.
3. **Height (cm):** Height is linked to BMI, it impacts overall body health metrics.
4. **Green\_Vegetables\_Consumption:** highlights the importance of a nutrient-rich diet for heart health.
5. **Age\_Category\_Elderly:** Elderly age has a significantly higher risk of disease.

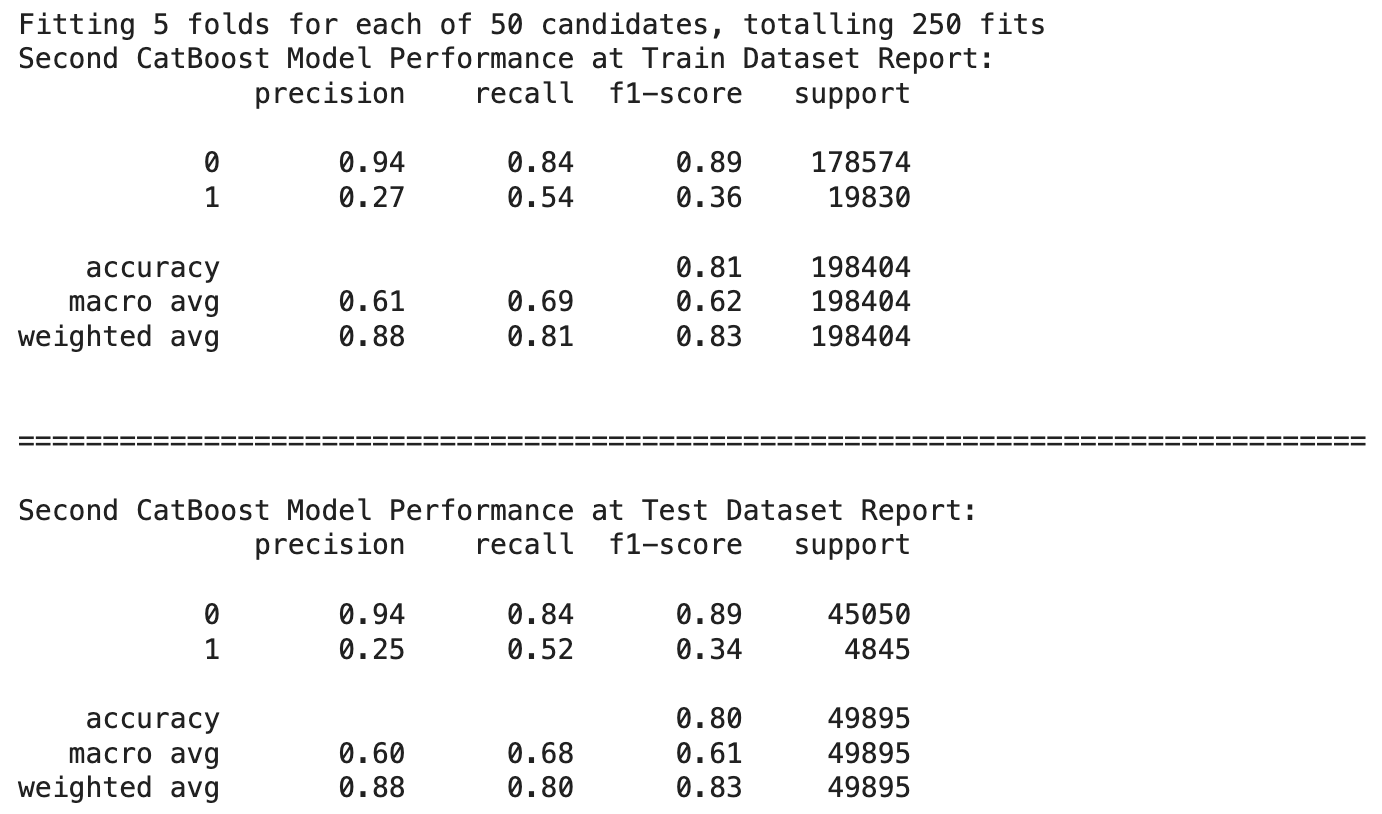
**Threshold adjustment and final model set up**

Here is the coding part:









In this process, we began by obtaining the prediction probabilities for both the training and test sets using our best CatBoost model. We then adjusted the threshold for classifying the predictions, setting it to 0.2 for both datasets. This involved converting the probabilities to binary class labels based on whether they exceeded the threshold. We generated and examined classification reports to assess the model's performance on the training and test sets. Subsequently, we filtered out uncertain data, specifically those with prediction probabilities exceeding the threshold, from both datasets. We re-modeled the uncertain data using a second round of hyperparameter tuning with CatBoost, obtaining a new best estimator. We evaluated the performance of this refined model on the filtered training and test sets, generating new classification reports. This iterative approach of threshold adjustment, uncertainty filtering, and re-modeling aimed to enhance the overall accuracy and reliability of our predictive model.

**Conclusion**

Throughout this project, our main goal was to enhance early detection and prevention of heart disease through sophisticated risk prediction modeling. We successfully utilized a variety of machine learning algorithms, including Decision Trees, Random Forest, XGBoost, LightGBM, and CatBoost, to analyze and predict heart disease based on a comprehensive dataset. After careful evaluation and tuning, the CatBoost model was the best performing model due to its superior accuracy. However, despite these advances, there is still room for improvement in the models, especially in terms of recall. This highlights the need for continuous improvement and optimization to ensure that the model not only accurately predicts, but also minimizes false negatives, which are critical in medical diagnosis. Future work will focus on further improving the recall of these models and exploring other data features that can improve predictive performance, ensuring that these models become more effective and useful tools in the fight against heart disease.

**Recommendations**

1. Enhance the dataset with more variables: To improve the predictive accuracy and robustness of the model, we recommend the inclusion of more health-related variables such as genetic markers, detailed lifestyle factors (e.g., sleep patterns and stress levels), and long-term health history. Expanding the dataset will provide a more nuanced understanding of the factors influencing heart disease.

2. Integration of predictive modeling into clinical workflows: Work with healthcare professionals to integrate these predictive tools into existing clinical workflows. This integration can facilitate real-time risk assessment, personalized patient management, and prospective care delivery, ultimately improving patient outcomes.

3. Expanded data collection on cardiac cases: In order to further validate and refine predictive models, it is critical to collect more comprehensive data on heart disease cases, including a broader demographic and geographic distribution. This expansion will help test the validity of the models in different populations and conditions, ensuring their applicability and reliability in a variety of clinical settings.

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## 

## **Reference**

## Duggal, N. (2023, February 20). *Advantages of decision trees : Everything you need to know: Simplilearn*. Simplilearn.com. https://www.simplilearn.com/advantages-of-decision-tree-article

GeeksforGeeks. (2024, February 15). *What are the advantages and disadvantages of Random Forest* https://www.geeksforgeeks.org/what-are-the-advantages-and-disadvantages-of-random-forest/

Hachcham, A. (2023, August 11). *XGBoost: Everything you need to know*. neptune.ai. https://neptune.ai/blog/xgboost-everything-you-need-to-know

Alphiree. (2023, July 3). Cardiovascular diseases risk prediction dataset. Kaggle. https://www.kaggle.com/datasets/alphiree/cardiovascular-diseases-risk-prediction-dataset/data?select=CVD\_cleaned.csv

*Hyperparameter tuning: GRIDSEARCHCV and RANDOMIZEDSEARCHCV, explained*. KDnuggets. (n.d.). https://www.kdnuggets.com/hyperparameter-tuning-gridsearchcv-and-randomizedsearchcv-explained

Kumar, Mr. B. (2023, August 4). *Difference between GRIDSEARCHCV and RANDOMIZEDSEARCHCV - 360digitmg*. 360digitmg.com. https://360digitmg.com/blog/comparison-gridsearchcv-and-randomsearchcv

*Sklearn.model\_selection.RANDOMIZEDSEARCHCV*. scikit. (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.RandomizedSearchCV.html

Brownlee, J. (2023, October 10). *How to use ROC curves and precision-recall curves for classification in Python*. MachineLearningMastery.com. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/