Interpretable Heart Failure Prediction Model: Leveraging Advanced Machine Learning Models and Interpret Clinical Insights

Mohammad Minhazul Amin

Abstract

Heart disease remains the leading cause of mortality worldwide, underscoring the critical need for predictive tools that integrate effectively into clinical settings. This study addresses the challenge posed by the "black box" nature of machine learning (ML) models in healthcare, focusing on enhancing their transparency and interpretability to facilitate clinical decision-making. Leveraging a comprehensive Cardiovascular Disease dataset, which includes detailed demographic, clinical, and lifestyle information from 70,000 patients, this study evaluated seven ML models: Logistic Regression (LR), K-Nearest Neighbors (KNN), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM). Among these, XGBoost and Logistic Regression emerged as the best-performing complex and simple models, respectively, with XGBoost achieving accuracy of 0.7386. Logistic Regression displayed robust performance with an accuracy of 0.7299, underscoring its utility in scenarios requiring straightforward interpretability. Advanced explainable artificial intelligence (XAI) techniques incorporated, including SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), Individual Conditional Expectation (ICE) plots, and Partial Dependence (PD) plots to bridge the gap between performance and clinical usability. These tools were instrumental in dissecting the decision-making processes of both simple and complex models that were utilized in this study to analyze randomly selected patient profiles from the dataset to compare the interpretability of ML models. This analysis highlighted where the straightforward interpretability of the Logistic Regression model was advantageous such as low risk situations and high risk patients where the detailed complexity was captured by the XGBoost model. Such findings provide essential insights for clinicians on selecting machine learning models that balance interpretability with predictive accuracy, crucial for patient-centered care in heart failure management. Additionally, this paper demonstrates the importance of interpretability in deploying ML models in healthcare, emphasizing that the choice of model—simple versus complex—should be tailored to the specific clinical context and patient profile because this will not only builds trust and reliance on ML predictions among healthcare providers but also enhances patient care by enabling more informed and personalized treatment decisions.

Table of contents

Chapter 1

Introduction

Heart failure represents a significant global health challenge, manifesting as a complex clinical syndrome where the heart fails to pump blood effectively to meet the body's needs. This condition often stems from various underlying causes such as coronary artery disease, hypertension, and diabetes. The multiplicity of these causes complicates diagnosis and treatment, making it imperative for healthcare providers to identify and stratify risks early to improve patient outcomes. The global burden of Heart Failure is immense, affecting over 26 million people worldwide, and this number is expected to rise due to aging populations and increasing prevalence of risk factors like CAD and diabetes [1].

Early detection and risk stratification are vital in managing Heart Failure, as they enable timely and targeted interventions that can enhance the quality of life and reduce mortality rates. Early intervention can significantly slow disease progression, reduce hospitalizations, and improve survival rates [2]. Recent advancements in machine learning and explainable artificial intelligence have shown promising potential in transforming heart failure prediction and management. Traditional statistical models, which often rely on linear assumptions and limited feature interactions, are increasingly being augmented or replaced by ML algorithms that can capture complex, non-linear relationships within vast datasets, thereby improving diagnostic accuracy and predictive power [3]. However, the "black box" nature of these models often poses a significant challenge in clinical settings, where understanding the rationale behind predictions is crucial for informed decision-making and patient safety [4].

Explainable AI (XAI) techniques, such as SHapley Additive exPlanations and Local Interpretable Model-agnostic Explanations , have emerged as pivotal tools in addressing the interpretability issue [5]. These methods provide insights into how individual features contribute to a model's predictions, making the decision-making process transparent and understandable for clinicians. SHAP values, for instance, offer a unified measure of feature importance by distributing the

prediction's contributions among its features based on cooperative game theory. LIME, on the other hand, approximates the model locally around the prediction to provide an interpretable representation [6]. Individual Conditional Expectation plots and Partial Dependence plots have been incorporated into the XAI toolkit to visualize the relationship between features and predictions for granular inspection on complex model performance and simple model performance [7]. Despite advancements, the deployment of machine learning models in the diagnosis of heart failure poses a significant challenge due to the varying levels of model complexity, which directly affects their interpretability and clinical usability [8]. Complex models provide nuanced interpretability that can offer deeper insights into disease prediction mechanisms but may be too intricate for routine clinical use [9]. On the other hand, simpler models provide straightforward interpretability, which ensures ease of understanding and application in clinical practices but might lack the depth required for certain diagnostic complexities [10].

This study aims to investigate the comparative effectiveness of complex versus simple machine learning models in predicting heart failure, focusing on their interpretability and practical utility in clinical settings. The research seeks to determine how the depth of model interpretability affects their suitability for varying patient profiles within clinical settings by integrating advanced ML algorithms with SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). The goal of this study is to develop an optimized predictive framework for heart failure that effectively balances model complexity with interpretability, tailored to specific patient needs in clinical settings. Practically, necessity of both model types arises from the diverse nature of heart failure cases—some may require the depth and accuracy of complex models, especially in patients with multiple comorbidities, while others might benefit from the simplicity and transparency of more straightforward models in cases where fewer variables are at play. This level of interpretability ensures that healthcare providers can understand and trust the model's predictions, facilitating targeted interventions that are appropriately customized for each patient's unique health profile. These interventions could include personalized lifestyle modifications, precise medication adjustments, or strategic monitoring plans, all designed based on a deep understanding of the factors driving a patient's Heart Failure risk as elucidated by the predictive model.

Chapter 2

Literature Review

Heart failure is a growing global health burden that represents a complex clinical syndrome when the heart is unable to pump blood sufficiently to meet the body's needs. Coronary artery disease, hypertension and diabetes these different causes make heart failure hard for doctors to diagnose the problem and provide relevant treatment. This literature review highlights the critical intersection of machine learning and explainable artificial intelligence within heart failure prediction, emphasizing the evolving capabilities and challenges of these technologies in enhancing diagnostic accuracy, ensuring clinical interpretability.

2.1 Machine Learning in Healthcare

2.1.1 Beyond Traditional Models

The application of machine learning in healthcare grew exponentially, driven by advancements in computational power and the availability of large-scale health data [11]. Traditional statistical models transition into Machine learning models, specifically deep learning algorithms have shown promising results through capturing non-linear complex relationships and nuance among variables in diagnosing disease, predict medical outcomes and outline a prescribed course of treatments [12].

2.1.2 Machine Learning Black Box

Machine learning models started showing promise in healthcare diagnosis and treatment prediction through identifying complex patterns in vast datasets with improved accuracy compared to traditional statistical methods [13]. However, the model became more complex and doctors sought to understand the rationale behind the model's predictions to make more informed decisions and patient safety [14].

2.1.3 Explainable AI

Interpretability is crucial for decision making in healthcare, AI integration into healthcare is placed to mitigate the challenge associated with explaining models and make it more transparent [15]. XAI techniques such as SHAP analysis to provide insights into relative impact of each feature on the model's output, Local Interpretable Model-agnostic Explanations (LIME) builds an interpretable model around a specific prediction to understand the complexity [16]. However, developing more sophisticated XAI techniques are tailored for complex healthcare models so researchers are exploring the potential of combining multiple XAI techniques to establish a more comprehensive understanding of heart failure prediction models [17].

2.1.4 Controversies and Diverse Viewpoints

A notable controversy emerges between achieving high predictive accuracy and ensuring model interpretability. Advocates for complex, deep learning models argue for their superior accuracy in diagnosing heart failure and predicting patient outcomes, potentially transforming patient care and reducing healthcare costs [18]. In contrast, proponents of simpler, more interpretable models emphasize the ethical need for transparency in clinical decision-making processes, fostering trust and informed decision-making among healthcare providers and patients [14]. Critical examination of existing literature reveals a nuanced landscape where complex models offer enhanced predictive capabilities but suffer from opacity that may acquire skepticism among healthcare professionals [19]. Meanwhile, models that incorporate interpretability, such as those using SHAP analysis, facilitate understanding of individual feature contributions to predictions, yet might not fully capture the data's nuances [20]. This debate reflects broader ethical considerations in AI's use in healthcare, underscoring the tension between technological advancement and the development of regulatory and ethical frameworks. Future research is directed toward methodologies that reconcile accuracy with transparency, aiming to fully harness AI's potential in healthcare while adhering to ethical standards and fostering user trust.

2.2 Prior exploration XAI in heart failure

Ke Wang's study [21] aimed to evaluate the performance of machine learning models and establish an explainable ML model to identify the cause of 3-year mortality in patients with heart failure caused by coronary heart disease. Author developed six ML models to predict 3-year cause of mortality including five frequently used models Logistic Regression (LR), K-Nearest Neighbor (KNN), Naive Bayesian (NB), and Multi-year Perceptron (MLP), also introduced XGBoost. SHapley Additive exPlanations incorporated to provide interpretations of the model's decisions and to calculate the individual mortality risk. Evaluation metrics implemented to assess using sensitivity, specificity, F-1 score and Area under the curve (AUC) metrics. This study utilized follow-up records of 1562 patients with heart failure due to cardiovascular disease from Shanxi Medical University in China. The dataset consists of demographics, medical history, physical status, echocardiography status, electrocardiography result and laboratory parameters. The research found Extreme Gradient Boosting (XGBoost) as a best performing model and SHAP values identified age, NT-proBNP levels, occupation, NYHA classification and nitrate drug use as important factors for predicting mortality in male and female. Authors pointed out the limitation as the study was geographically limited to two hospitals in Shanxi Province, China might indicate bias. A lack of validation with independent cohorts to further verify the mode's superiority is also explicitly mentioned. Incorporation of machine learning algorithms and SHAP for interpretability along with techniques for preprocessing data, model evaluation and interpretation can inform similar approaches in this present study for combining predictive performance with interpretability in heart failure models.

Pedro Moreno-Sanchez [22] embarked on a study to confront the prevalent challenge where models typically sacrifice either accuracy or interpretability. The objective was to create a model that maintains high accuracy while also providing a clear explanation for its predictions to offer healthcare professionals actionable insights and improve treatment and diagnosis. Author's approach involved utilizing the XGBoost classifier and undertaking a feature selection preprocessing to discern which features significantly influence the model's prediction. The study employed a dataset from the UCI Machine Learning Repository, consisting of medical records of 299 heart failure patients from two hospitals in Pakistan. This dataset consists of crucial

indicators such as age, anemia, blood pressure and others with both numerical and binary features. This study revealed that the XGBoost classifier was the most accurate with 83% accuracy, and features like 'anemia', 'time', 'ejection_fraction', and 'serum_creatinine' were highly significant in the model's results. The paper highlighted the potential bias into the model as the dataset is from only two hospitals and might not represent the larger population and also the absence of external validation with an independent cohort to further establish the model's applicability across diverse clinical settings. The combination of ensemble trees with feature importance techniques could serve as a strong foundation in the present research paper for developing predictive models that are both accurate and transparent.

Jili Li's research [23] centered on developing an interpretable machine learning model to predict mortality for patients with heart failure in intensive care units and utilized SHAP method to provide explanations for the predictions from machine learning model XGBoost. The research involved a retrospective cohort study and compared the performance of four machine learning models: XGBoost, Logistic Regression, Random Forest, and Support Vector Machine, with an emphasis on using SHAP for interpretability. The models were trained and validated with a split of 70% training and 30% validation from the dataset. This study extracted data of 2,789 patients diagnosed with primary heart failure from the eICU Collaborative Research Database version 2.0, a multicenter database that includes comprehensive records from ICU admissions across 208 hospitals in the United States from 2014-2015 and has demographics, vital signs, diagnosis and treatment information of the patient. According to the findings, XGBoost model outperformed other models with the highest AUC of 0.824, proving to be the most effective in predicting in-hospital mortality among the studied patients. SHAP analysis revealed the top predictors for mortality, with blood urea nitrogen being the most significant, followed by factors like patient age and various physiological measurements. However, the study acknowledges several limitations including potential bias due to missing data and the absence of external validation, which might affect the generalizability of the model. The use of data only from the first 24 hours of ICU admission could overlook changes in the patient's condition that affect mortality risk. The methodology of using SHAP for interpretability in conjunction with XGBoost can provide a blueprint for developing transparent and effective predictive tools in clinical settings.

Putri Sari Asiha's study [24] developed and evaluated performance of machine learning models that can predict heart disease accurately and explain predictions with SHAP and LIME to make them useful and understandable in clinical settings. The research implemented four machine learning models Support Vector Machine, Random Forest, XGBoost and K-Nearest Neighbor and evaluated F-1 score to determine their performance. Additionally, SHAP and LIME were utilized to provide interpretations of model's decisions. The research utilized Cleveland Heart Disease dataset consists of 303 patient observations with 76 features but only 14 were released for publication. Features such as age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, rest electrocardiographic results, maximum heart rate, exercise-induced angina, old peak, number of major vessels colored by fluoroscopy, thalassemia evaluated in this research paper. The SVM and XGBoost models demonstrated the highest performance with an F1-score of 88%, indicating their robustness in predicting heart disease. SHAP revealed that chest Pain Type, number of major vessels colored by fluoroscopy, thalassemia, maximum heart rate, and old peak had a more substantial impact on the model predictions. LIME provided local interpretable insights showing features like chest pain, number of major vessels colored by fluoroscopy, and old peak often prediction towards positive diagnosis. However, negative prediction features such as Fasting Blood Sugar and Cholesterol might not significantly influence the outcome. Additionally, The research highlights the challenge of making complex machine learning models interpretable while maintaining high accuracy. The use of SHAP and LIME to interpret machine learning predictions is a critical advancement that can be applied to this research paper to enhance patient reliability and user trust.

Linardatos [25] tackle the pressing questions surrounding the interpretability of complex machine learning models. The paper evaluated a broad spectrum of interpretability methods, including SHAP, LIME, gradient-based techniques, and surrogate models, highlighting their capabilities in making model decisions comprehensible. The authors found that while methods like SHAP and LIME excel in offering insights into individual predictions, they fall short in more complex model structures or massive datasets. Gradient-based methods offer a broader understanding of model behavior yet also face challenges with intricate interactions within models. The review underscored a significant gap in the current XAI landscape: the trade-off

between model accuracy and interpretability remains unresolved, with highly accurate models often being the least interpretable. This insight is crucial for advancing heart failure prediction models, suggesting a careful consideration of the interpretability techniques that can be realistically applied without overly compromising model performance.

Researcher Petch [26] critically explores the application of explainable machine learning techniques in cardiology, aiming to demystify complex ML models for clinical use. The paper addresses key research questions regarding the effectiveness of explainable ML methods in making these models more interpretable and identifies the limitations inherent in these approaches. The methods evaluated include variable importance methods, which quantify the influence of individual variables on model predictions; surrogate methods, such as Local Interpretable Model-Agnostic Explanations and Shapley Additive Explanations, which approximate complex models with simpler, understandable explanations; and visualization techniques like Partial Dependence Plots (PDPs) that visually depict the impact of variables. These methods were demonstrated using the Cleveland Heart Disease Data Set, a common benchmark dataset in cardiac health studies. Findings from the review highlight that while techniques like LIME and SHAP enhance transparency and could potentially foster greater trust in ML models among clinicians, they also caution that such explanations are simplifications and may not fully capture the intricate dynamics of the model's decision-making processes. The review underscores a significant gap: the potential for these explanations to mislead by oversimplifying complex interactions. For researchers developing interpretable models for heart failure, the study underlines the importance of selecting appropriate explainable techniques that align with the clinical need for accuracy and comprehensibility, while also being wary of the limitations that could impact clinical decision-making.

Author	Approach	Dataset	Key Findings
Ke Wang	Developed six ML models including LR, KNN, NB, MLP, and XGBoost	Follow-up records of 1562 patients with heart failure from Shanxi Medical University, China; includes extensive clinical parameters	XGBoost was the best performing model; key predictors were age, NT-proBNP, occupation, NYHA classification, nitrate drug use
Pedro Moreno-Sanchez	Utilized XGBoost and feature selection preprocessing	Medical records of 299 heart failure patients from two hospitals in Pakistan included age, anemia, blood pressure, etc.	XGBoost was most accurate; significant features included 'anaemia', 'time', 'ejection_fraction', 'serum_creatinine'
Jili Li	Compared four ML models: XGBoost, LR, RF, SVM using retrospective cohort study	Data of 2,789 patients from the eICU Collaborative Research Database, includes comprehensive ICU admission records from the US	XGBoost outperformed others with highest AUC; top predictors included blood urea nitrogen, patient age, physiological measurements
Putri Sari Asiha	Developed and evaluated SVM, RF, XGBoost, and KNN models	Cleveland Heart Disease dataset with 303 observations and 14 features including clinical and diagnostic information	SVM and XGBoost had the highest F1-score at 88%; important features included chest pain type, major vessels, thalassemia
Linardatos	Review of XAI techniques including SHAP, LIME, and gradient-based methods	General application across sectors; no specific datasets discussed for experiments	Effective in insights on individual predictions but struggle with complex model structures and large datasets
Petch	Permutation importance, surrogate decision trees, LIME, partial dependence plots; Random Forest and CNN models	Cleveland Heart Disease Data Set, EyePACS dataset (illustrative purposes)	While techniques increase transparency, they simplify complex behaviors, which could mislead users about model decisions

Table 2.1: Prior Studies Interpretable Heart Failure Prediction Models

2.3 Addressing the Gap of Complex and Simple Model Interpretability

Building on the findings of prior researchers, which highlighted the importance of model interpretability, this study further investigates how different levels of interpretability affect clinical decision-making in real-world settings. This research explores and evaluates the differential impacts of complex and simple machine learning models on heart failure prediction across different patient profiles by integrating advanced interpretability techniques such as SHAP and LIME. This integration aims to make the predictive outcomes not only transparent but also actionable for healthcare providers. Through assessing these models on varied patient datasets, the study determines which model complexity—simple or complex—is more appropriate for specific patient demographics and clinical scenarios. A key goal is to bridge the existing gap in clinical ML applications, where the opaque nature of complex models often hinders their practical use. By providing insights into how different models handle interpretability, this research informed the development of heart failure prediction tools that are both highly accurate and suitably transparent, ensuring they meet the needs of diverse clinical environments and support personalized patient care strategies.

Chapter 3

Methodology

This study aims to investigate and assess the varying impacts of complex and simple machine learning models on predicting heart failure, with a particular emphasis on how the complexity of these models influences their interpretability. It seeks to understand how the level of interpretability of each model type affects its appropriateness for different patient profiles in clinical environments. The research will analyze whether simpler or more complex models provide the clarity and depth needed for effective clinical application, thereby helping to identify which model characteristics best align with the needs of diverse patient groups.

3.1 Dataset Description

The study utilizes the Cardiovascular Disease dataset, which comprises 70,000 records and 13 features that include demographic and health-related information [27]. Many studies focus predominantly on clinical and physiological parameters and integrated lifestyle factors [28]. The dataset includes three categories of input features:

- **Objective Features:** Age, height, weight, and gender
- **Examination Features:** Systolic and diastolic blood pressure, cholesterol levels, and glucose levels
- **Subjective Features:** Smoking status, alcohol intake, and physical activity

The presence or absence of cardiovascular disease is the binary target variable. Data were collected at the time of the patients' medical examination, providing a snapshot of their current health status. This dataset was collected from health surveys and medical records, ensuring a comprehensive set of features. The attributes include:

- **Demographics:** Age (in days, later converted to years), gender
- **Physical Measurements:** Height (cm), weight (kg)

- **Blood Pressure:** Systolic (ap_hi) and diastolic (ap_lo)
- **Biochemical Parameters:** Cholesterol level, glucose level
- **Lifestyle Factors:** Smoking habits, alcohol intake, physical activity level
- **Target Variable:** Presence of cardiovascular disease (cardio)

3.2 Data Preprocessing

3.2.1 Data Cleaning and Normalization

Initial data cleaning involved removing or correcting erroneous entries, handling missing values, and ensuring consistency across the dataset.

- **Conversion of Age:** Age was initially recorded in days and was converted to years to enhance interpretability. This was achieved by dividing the age in days by 365.25 (accounting for leap years).
- **Handling Missing Values:** Missing values were checked and found no values to ensure no loss of information.

3.2.2 Outlier Detection and Treatment

Outliers in systolic and diastolic blood pressure readings were identified through visualizations such as box plots and statistical techniques like the interquartile range (IQR). Blood pressure readings were capped within physiological limits, systolic blood pressure was capped between 90 and 180 mmHg, and diastolic blood pressure was capped between 60 and 120 mmHg [29]. Values outside these ranges were adjusted to the nearest limit to avoid skewing the model training process.

- **Systolic Blood Pressure (90-180 mmHg):** This range includes typical physiological blood pressure values for adults. Values below 90 mmHg indicate hypotension, which might be due to measurement errors or rare medical conditions, while values above 180 mmHg indicate hypertension, which is clinically significant but extremely high values might indicate measurement errors [30].

- **Diastolic Blood Pressure (60-120 mmHg):** Similarly, this range covers the normal physiological limits. Diastolic values below 60 mmHg indicate hypotension, and values above 120 mmHg are indicative of severe hypertension, with extremely high values likely being erroneous [31].

3.2.3 Feature Scaling

Numerical features were standardized to have a mean of zero and a standard deviation of one. This was done using standardization techniques, which ensure that all features contribute equally to the model training process and improve the performance of algorithms sensitive to data scaling, such as logistic regression and support vector machines [32]. This process involves subtracting the mean of each feature and dividing by its standard deviation.

3.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to gain a comprehensive understanding of the dataset and its underlying patterns, which is crucial for identifying anomalies, understanding data distribution, and determining relationships between variables.

3.3.1 Descriptive Statistics

Mean, median, standard deviation, minimum, and maximum values were calculated for all features to understand the central tendency, dispersion, and shape of the data distribution.

- **Age:** The mean age (in years) was approximately 53, with a standard deviation of 6.7, indicating a middle-aged population.
- **Height and Weight:** Average height was 169 cm, and the average weight was 74 kg, with respective standard deviations indicating moderate variability.
- **Blood Pressure:** Mean systolic blood pressure was 128 mmHg, and the mean diastolic blood pressure was 82 mmHg. Both exhibited a normal range but with some outliers indicating potential hypertension.
- **Cholesterol and Glucose Levels:** Average cholesterol and glucose levels were within expected ranges, but the standard deviations suggested some variability, indicating the presence of individuals with high cholesterol and glucose levels.

3.3.2 Correlation Metrics

Systolic and diastolic blood pressure had a strong positive correlation (r = 0.64), as expected. Other significant correlations included weight and height (r = 0.47) and cholesterol and glucose levels (r = 0.21). These correlations were used to guide feature selection and engineering, ensuring that multicollinearity issues were minimized.

3.3.3 Class Distribution Analysis

Target variable (cardio) was analyzed by calculating the frequency of each class. The analysis revealed a nearly balanced distribution between classes, with 49.7% of the records indicating the presence of cardiovascular disease and 50.3% indicating its absence. This balanced class distribution ensured that the model would not be biased towards predicting the majority class, thereby improving its generalizability and robustness.

3.4 Model Development and Selection

3.4.1 Model Selection

Developed seven machine learning models using follow-up data to predict Heart Disease are Logistic Regression (LR), Decision Tree (DT), Naive Bayes (NB), K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) [33]. XGBoost is an optimized implementation of gradient boosting, based on the ensemble of weak learners, characterized by high bias and low variance. It uses a second-order Taylor series to approximate the value of the loss function and further reduces the possibility of overfitting through regularization. This algorithm is chosen for its robust performance in classification tasks across medical datasets, noted for handling large-scale data and complex feature interactions effectively. Numerous studies in the field of medical research have validated the effectiveness of XGBoost [34].

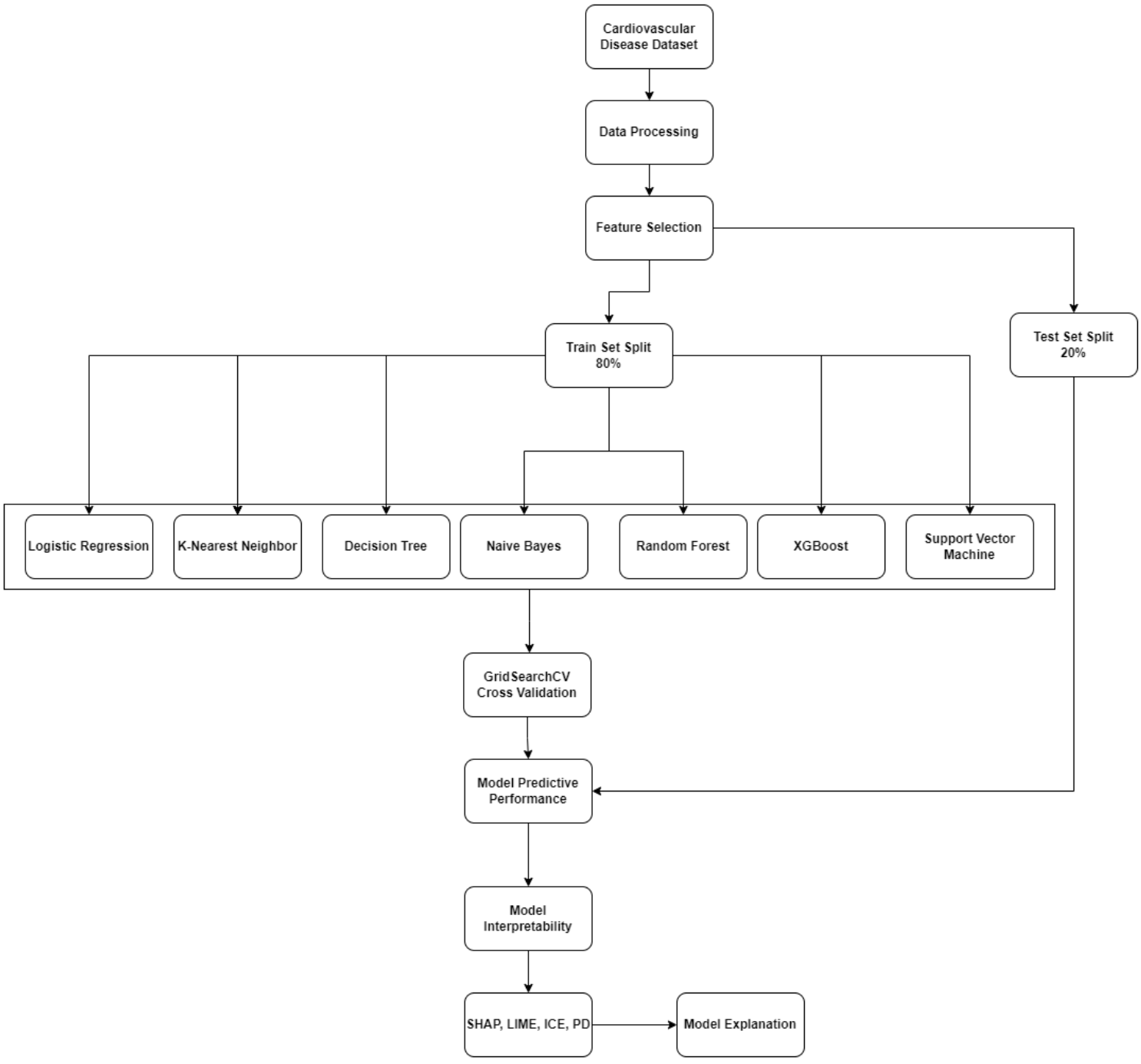


Figure 3.1: Evaluation Model Development Flow

3.4.2 Model Training and Testing

Hyperparameters are optimized through grid search with five-fold cross validation for finding the best parameters. The model was trained on an 80-20 split of the dataset and each model performance evaluated on a test set over Cross-validation score [35].

3.4.3 Evaluation Metrics

Accuracy, Precision, Recall, Classification Report, ROC-AUC score, ROC Curve and Precision-Recall Curve to choose the best balanced model five models further explained through interpretable techniques:

- **Accuracy:** Measures the proportion of true results (both true positives and true negatives) among the total number of cases examined. This metric is straightforward and gives a quick indication of overall model effectiveness. However, accuracy alone can be misleading, especially in imbalanced datasets where the number of negative cases significantly outnumbers the positives..

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

- **Precision:** Indicates the proportion of positive identifications that were actually correct. This is particularly important in medical diagnoses where false positives can lead to unnecessary stress and procedures for patients.

$$Precision = \frac{TP}{TP + FP}$$

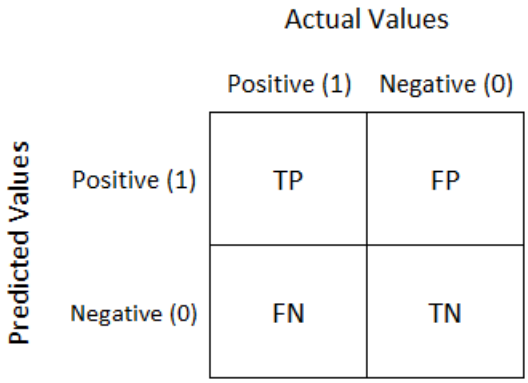
- **Recall:** Measures the ability of the model to correctly identify all relevant instances (true positives). High recall is critical in medical settings to ensure that cases of heart disease are not missed.

$$Recall = \frac{TP}{TP + FN}$$

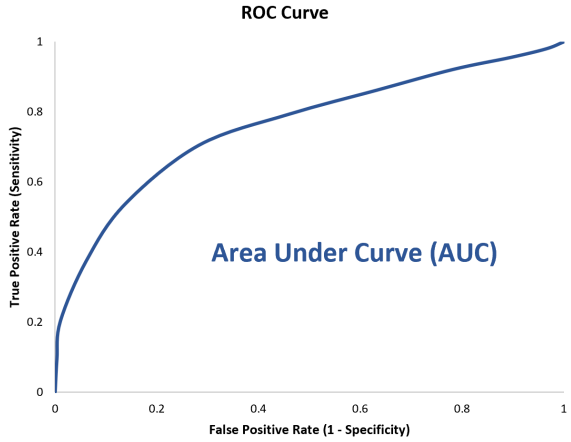
- **ROC-AUC Score:** The Receiver Operating Characteristic (ROC) curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) and provides a single value to summarize the performance of the model across all classification thresholds.

$$ROC - AUC = \int_0^1 TPR(FPR)\,\mathrm{d}FPR$$

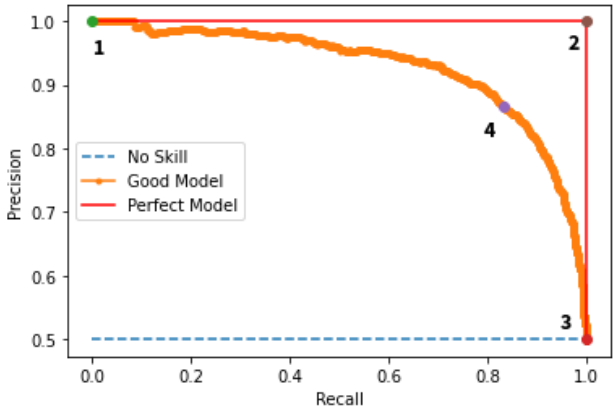
- **Classification Report:** Provides a detailed breakdown of precision, recall, F1 score, and support for each class, offering a comprehensive view of model performance.



- **ROC Curve:** Visualizes the trade-off between the true positive rate and the false positive rate at various threshold settings, helping in selecting the best threshold.



- **Precision-Recall Curve:** This curve plots precision against recall at various threshold settings, providing insights into the trade-off between these two metrics.



3.5 Model Interpretability

3.5.1 Implementation of SHAP and LIME

As complex ML models are considered as black boxes these tools are applied post-hoc to interpret the model's predictions [36]. SHAP values explain the impact of each feature on the model output, while LIME provides local interpretations, enhancing the transparency necessary for clinical decision-making [37].

- **SHAP:** Provides detailed explanations of the contribution of each feature to the prediction, making the model's decisions transparent and justifiable. Used to quantify the impact of each feature on the prediction, providing global interpretability [38].
- **LIME:** Articulates local interpretability by explaining individual predictions, which is crucial for clinical decision-making [6].

3.5.2 Application in Clinical Scenarios

To evaluate the practical utility of these interpretability tools, the study conducts a detailed comparison of best performed simple and complex model interpretations using specific patient cases from the dataset. This comparison not only highlights the strengths and limitations of each model in real-world settings but also provides a deep understanding of how interpretation of models aligns with individual patient factors.

- **Patient A and Patient B Details:** The dataset includes diverse patient profiles, such as Patient A (ID 0, Age 50) with lower risk factors and Patient B (ID 36, Age 63) with higher risk factors, which provide a basis for testing the models under varied clinical conditions.
- **Best Complex & Simple Model:** After randomly selecting two patient profiles from the dataset, the best complex and simple models were implemented to predict heart failure. This methodical approach allowed for a direct comparison of how each model performs under similar conditions but with varying levels of complexity. The outcomes provided a practical demonstration of the models' predictive accuracy and interpretability, highlighting their potential applicability in real-world clinical settings.
- **Interpretability Analysis:** Chosen best performed simple and complex implemented and rigorously analyzed interpretability. For a simple model, SHAP force plots and LIME reveal straightforward interpretations of traditional risk factors. This model's transparency is pivotal in settings where healthcare providers require immediate and clear justification for diagnostic decisions. Conversely, the best-performing complex model exhibits a richer, more layered interpretative analysis through SHAP and LIME.

This approach, emphasizing both performance and transparency, equips healthcare providers with the necessary tools to customize interventions more precisely, guided by a thorough comprehension of how model predictions correlate with specific patient characteristics. Such a methodical strategy substantially advances the domains of predictive analytics and medical informatics, improving clinical decision-making processes and ultimately enhancing patient care outcomes.

Chapter 4

Result

4.1 Patient Clinical Profile

A total of 70,000 patients were included in this study, with 24,470 male patients (34.96%) and 45,530 female patients (65.04%). The average age of male patients was 52.63 ± 6.94 years, and the average age of female patients was 52.95 ± 6.67 years. The overall average age was 52.84 ± 6.77 years. The dataset was balanced, with 34,979 patients classified as having heart disease and 35,021 patients classified as not having heart disease.

4.2 Model Performance & Evaluation

4.2.1 Evaluation of Machine Learning Model Prediction

Evaluated multiple machine learning models for predicting heart disease, including Logistic Regression, Decision Tree, Naive Bayes, K-Nearest Neighbors, Random Forest, XGBoost, and Support Vector Machine. Here are the detailed outcomes for each model:

Model	CV Score	Accuracy	Precision	Recall	ROC AUC
Logistic Regression	0.7267	0.7299	0.7586	0.676	0.7905
K-Nearest Neighbors	0.7243	0.7299	0.7551	0.682	0.7875
Naive Bayes	0.7278	0.7111	0.7658	0.607	0.7777
Decision Tree	0.725	0.7359	0.7674	0.678	0.7931
Random Forest	0.7337	0.7395	0.7692	0.686	0.8019
XGBoost	0.7351	0.7386	0.7625	0.695	0.8031
Support Vector Machine	0.7245	0.73	0.7771	0.646	0.7899

Table 4.1: Heart Failure Predictive Model Performance

Logistic Regression: This linear model provides a good balance between simplicity and performance as a simple model. Its cross-validation score and accuracy are comparable to more complex models, making it a reliable choice for baseline performance. In this model, C is set to 0.01 by GridSearchCV, applying moderate regularization to prevent overfitting by penalizing large coefficients. The penalty parameter is set to l1, which introduces sparsity by zeroing out less important feature coefficients. This enhances interpretability by identifying the most influential features. The solver parameter is set to saga, an efficient solver for large datasets that supports both L1 and L2 penalties. These parameters help balance simplicity and accuracy, making the model robust and interpretable, particularly suitable for clinical settings where understanding feature contributions is critical. Moreover, precision of 75.86% indicates a good ability to identify true positive cases of heart failure, though its recall of 67.58% suggests the model missed a number of true positive cases of heart failure. This performance can be attributed to the linear nature of Logistic Regression, which may struggle to capture complex, non-linear relationships within the dataset [39]. ROC AUC of 0.7905 reflects a strong overall ability to discriminate between heart failure and non-heart failure cases. Its high interpretability, however, makes it valuable in clinical settings where understanding feature contributions is critical.

K-nearest Neighbor: KNN uses the n_neighbors parameter set to 20, which smooths out predictions by considering more neighbors, reducing sensitivity to noise. Weights parameter is set to uniform, treating all neighbors equally and simplifying the model. The metric parameter is set to manhattan (L1 norm), which measures distances more robustly in high-dimensional spaces than euclidean. Moreover, performance metrics are similar to those of Logistic Regression, with an accuracy of 72.99% and a precision of 75.51%. The recall of 68.17% is slightly higher, indicating it detects more true positive cases than Logistic Regression. However, its ROC AUC of 0.7875 is marginally lower due to the high dimensional dataset. KNN's performance is heavily influenced by the choice of neighbors (K) and distance metrics. The model's slightly higher recall suggests that KNN is better at identifying true positive cases, possibly because it considers the local structure of the data points and their neighbors, which helps capture more nuanced patterns [40]. However, the high-dimensional nature of the dataset

can make the distance calculations less reliable, potentially explaining the marginally lower ROC AUC. The decision-making process of KNN is less transparent, making it harder to understand why certain predictions are made.

Naive Bayes: This model assumes feature independence, which rarely holds in practice but can simplify computations and achieve decent performance. Naive Bayes does not have hyperparameters for tuning because it assumes feature independence given the class label. This assumption simplifies the model, making it fast and easy to implement. However, it has the lowest accuracy (71.11%) and recall (60.71%) among the models, indicating it misses a significant number of true positive cases. The lower recall can be attributed to the unrealistic independence assumption, which often fails to capture the dependencies and interactions between features present in real-world data [41]. However, its precision (76.58%) is relatively high, suggesting it is good at identifying actual cases when it predicts positive. This high precision might be due to the model's probabilistic nature, which can sometimes isolate specific feature contributions effectively [42].

Decision Tree: This model uses the criterion parameter set to entropy, which measures information gain and can yield more informative splits compared to gini. Max_depth parameter is set to 10, limiting the depth of the tree to prevent overfitting while capturing important patterns. Min_samples_split parameter is set to 10, ensuring splits occur only when sufficient data is present, and the min_samples_leaf parameter is set to 4, ensuring each leaf has enough samples for stable predictions. Additionally, the model performs well, with an accuracy of 73.59% and a precision of 76.74%. Its recall of 67.82% is similar to Logistic Regression. This result indicates a balanced performance [43]. The ROC AUC of 0.7931 suggests a good discriminative ability. One of the main advantages of Decision Trees is their interpretability. They provide a visual and intuitive representation of the decision-making process, which can be easily understood by clinicians. However, Decision Trees are prone to overfitting, especially with complex datasets. Techniques such as pruning and setting maximum tree depth can mitigate this issue.

Random Forest: Random Forest uses the n_estimators parameter set to 300, increasing the number of trees in the forest to reduce variance and improve stability. The max_depth parameter is set to 10, limiting each tree's depth to prevent overfitting. The min_samples_split parameter is set to 2, allowing flexible splits, and the min_samples_leaf parameter is set to 1, ensuring even small splits can occur. The bootstrap parameter is set to True, reducing overfitting by adding randomness. An ensemble learning method that combines multiple decision trees, shows strong performance with an accuracy of 73.95% and a precision of 76.92%. Its recall (68.57%) and ROC AUC (0.8019) are among the highest, indicating a robust predictive capability. The strength of Random Forest lies in its ability to handle large datasets and its robustness against overfitting, which is achieved by aggregating the predictions of multiple trees [44]. However, interpretability can be an issue due to the complexity of combining multiple trees.

XGBoost: XGBoost uses the n_estimators parameter set to 50, balancing the number of boosting rounds to avoid overfitting. The max_depth parameter is set to 3, limiting tree depth to prevent overfitting while capturing important patterns. The learning_rate parameter is set to 0.2, moderating learning speed to balance accuracy and overfitting risk. The subsample parameter is set to 1.0, using all samples for training, and the colsample_bytree parameter is set to 0.8, using 80% of features for each tree to add regularization. These settings enable XGBoost to capture complex patterns and interactions effectively. This gradient boosting algorithm, exhibits high performance with an accuracy of 73.86% and the highest recall (69.45%) among the all models as a complex model, indicating its superior sensitivity in detecting heart failure cases. Its ROC AUC of 0.8031 is also the highest, demonstrating excellent discriminative power. The strength of Gradient Boosting lies in its iterative approach, where models are trained sequentially to correct the errors of the previous models. This leads to the efficient handling of large datasets and complex feature interactions, allowing the model to capture subtle patterns and relationships within the data. Each iteration focuses on the residuals or errors from the previous model, refining the prediction by learning from mistakes. This iterative refinement enhances the model's overall performance and sensitivity [45].

Support Vector Machine: SVM uses the C parameter set to 10, providing more flexibility to fit the data by reducing regularization. This higher C value suggests that the data benefits from more complex decision boundaries. SVM with a linear kernel shows strong performance with an accuracy of 73.00% and the highest precision (77.71%), suggesting it is highly effective at identifying true positives. However, its recall of 64.63% indicates it misses a higher proportion of true positive cases compared to some other models. The high precision of SVM can be attributed to its margin maximization principle, which ensures that the decision boundary is placed optimally to separate the classes with the largest possible margin. This leads to a conservative prediction strategy that prioritizes certainty, thereby enhancing precision at the cost of recall [46]. However, the trade-off between precision and recall indicates that the model might be conservative in its predictions, missing some true positives. The ROC AUC of 0.7899 reflects good overall performance. As SVMs are effective for high-dimensional spaces and can handle non-linear relationships using kernel tricks. However, they are less interpretable than simpler models like Logistic Regression and Decision Trees. Explaining SVM decisions to clinicians can be challenging, making it less suitable for settings where transparency is critical.

4.2.2 Predictive Model Confusion Matrix

Confusion matrix provides a detailed breakdown of a model's performance by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [47].

- **True Positives (TP):** Correctly identified heart failure cases.
- **False Positives (FP):** Incorrectly identified heart failure cases (type I error).
- **True Negatives (TN):** Correctly identified non-heart failure cases.
- **False Negatives (FN):** Missed heart failure cases (type II error).

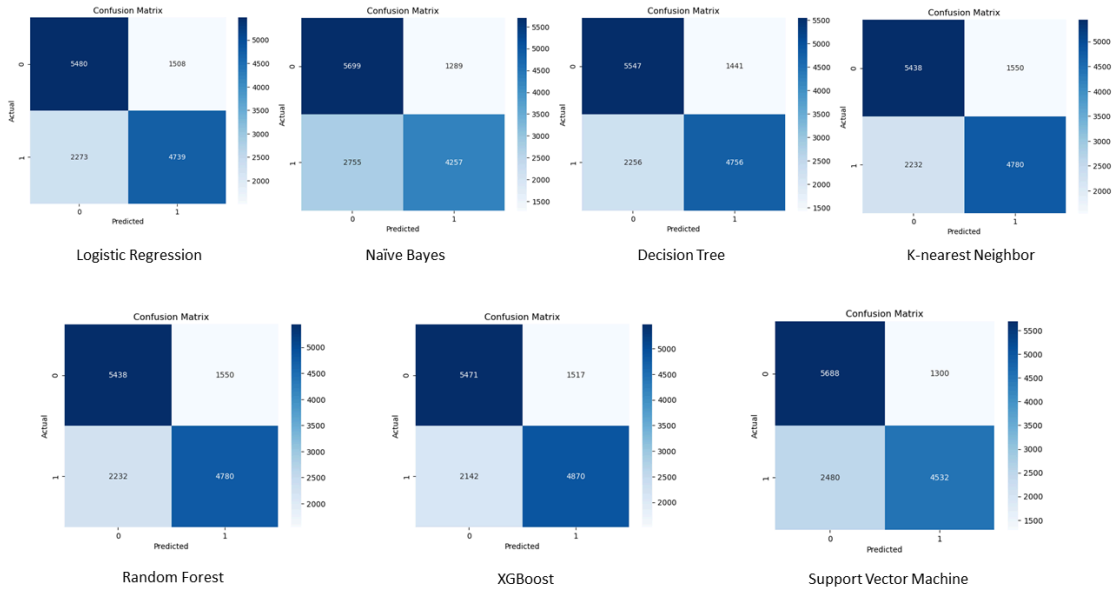


Figure 4.1: Confusion Matrix Heart Failure Predictive Model

Logistic Regression tends to balance false positives and false negatives effectively. However, the number of false negatives indicates that some heart failure cases are missed, which can be critical in a clinical setting where missing a diagnosis can have severe consequences. This aligns with its moderate recall (0.6758) and high precision (0.7586), suggesting it is good at predicting positive cases when it identifies them but can miss some true positive cases. While K-Nearest Neighbors improves recall (0.6817) by detecting more true positives, thereby reducing the number of false negatives. However, this improvement comes at the cost of more false positives (0.7551), leading to a trade-off between higher sensitivity and reduced specificity. This is reflected in its comparable accuracy to Logistic Regression but a higher recall, indicating better

performance in identifying true positive cases. On the other hand, Naive Bayes misses many true positive cases, resulting in a high number of false negatives leading to low recall (0.6071). This is due to its strong independence assumption, which often does not hold true in real-world data, leading to lower overall performance and the lowest recall among the models. Similar to Logistic Regression, Decision Trees offer a better balance between false positives and false negatives. They are slightly better at identifying true positives, as indicated by their good recall (0.6782) and precision (0.7674). Moreover, Random Forest significantly reduces the number of false negatives, indicating it catches most true positive cases. This is due to its ensemble nature, which aggregates multiple decision trees to improve generalization and reduce overfitting. The high recall (0.6857) and precision (0.7692) reflect this robust performance. XGBoost achieves the highest recall (0.7625), indicating it catches the most true positive cases among all models. This advanced gradient boosting technique captures intricate patterns in the data, resulting in superior sensitivity and overall accuracy. However, this comes with increased model complexity and the need for careful tuning. Support Vector Machine focuses on maximizing the margin between classes, which leads to high precision but more conservative predictions. This results in lower recall, meaning it tends to avoid false positives at the expense of missing some true positives. This trade-off is reflected in its high precision (0.7771) and lower recall (0.6463). Based on the overall performance metrics, particularly the balance between high recall and precision, XGBoost stands out as the best model for predicting heart failure. It consistently achieves the highest recall, indicating its superior ability to identify true positive cases, which is crucial in a clinical setting.

4.2.3 Model Performance Critique on ROC-AUC Curve

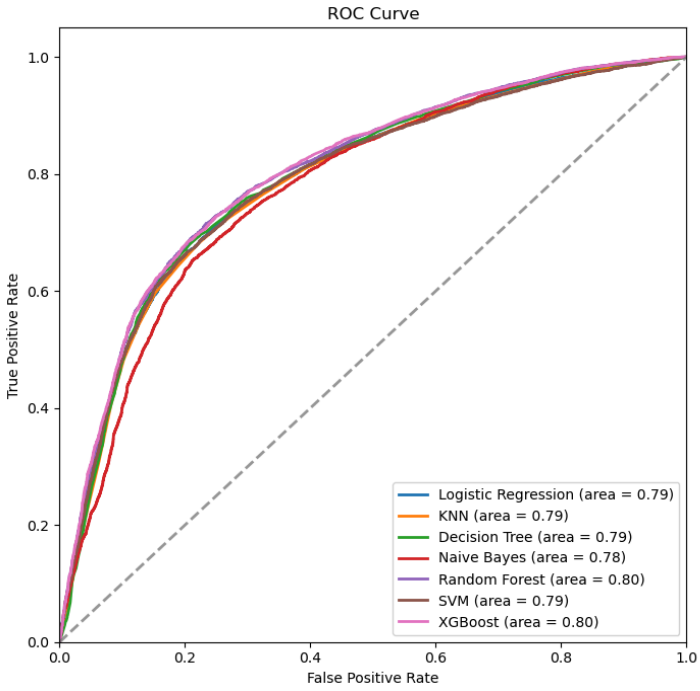


Figure 4.2: ROC_AUC Heart Failure Predictive Model

- **Logistic Regression:** Model exhibits a good ROC-AUC score of 0.7905, indicating reliable performance in discriminating between heart failure and non-heart failure cases across various thresholds. The smooth curve reflects consistent performance but highlights that the model may miss some complex interactions within the data, which slightly limits its AUC compared to more advanced models.
- **K-Nearest Neighbors:** Demonstrates a comparable ROC-AUC of 0.7875, slightly lower than Logistic Regression. The variability in the curve suggests sensitivity to the choice of neighbors and distance metrics. This aligns with its higher recall, indicating that while KNN effectively identifies true positives, it can be less stable across different thresholds due to local data structure dependencies.
- **Naive Bayes:** Indicates the lowest ROC-AUC score of 0.7777 among the models. This reflects its weaker overall discriminative power, largely due to its strong independence assumption between features that does not hold in real-world data. The less smooth and lower ROC curve indicates limitations in capturing complex feature interactions, consistent with its higher false negative rate.

- **Decision Trees:** Presents a good ROC-AUC score of 0.7931, showing solid discriminative ability. However, the curve can be less smooth, reflecting the model's sensitivity to overfitting. This performance aligns with its balanced approach to handling false positives and false negatives, though overfitting risks can affect the stability of its discriminative performance. Random Forest achieves a high ROC-AUC score of 0.8019, indicating excellent discriminative power with a smooth and consistently high ROC curve. This robustness reflects its ensemble nature, reducing overfitting and improving generalization. The high performance in capturing true positives and maintaining precision aligns with the stability observed in the ROC curve.
- **XGBoost:** The highest ROC-AUC score of 0.8031 is achieved in this model, demonstrating superior performance in discriminating between classes. The smooth and high ROC curve reflects its advanced gradient boosting techniques that effectively capture intricate data patterns and reduce bias and variance. This aligns with its highest recall and overall accuracy, showing that XGBoost excels in maintaining consistent performance across various thresholds.
- **Support Vector Machine:** This model shows a good ROC-AUC score of 0.7899, indicating reliable discriminative performance. The smooth ROC curve, however, lower recall is also noticeable due to its conservative margin-maximizing approach. This results in fewer false positives but also a higher number of missed true positive cases, reflecting the model's trade-off between high precision and lower recall.

Each model's ROC-AUC performance provides insights into their discriminative power and their ability to balance sensitivity and specificity across different thresholds. Final choice of model should consider these trade-offs, particularly in clinical settings where the cost of false negatives and false positives must be carefully balanced. Based on the ROC-AUC scores, XGBoost stands out as the best model due to its highest discriminative power (0.8031) and consistent performance across various thresholds.

4.2.4 Balanced Model through Precision-Recall Curve

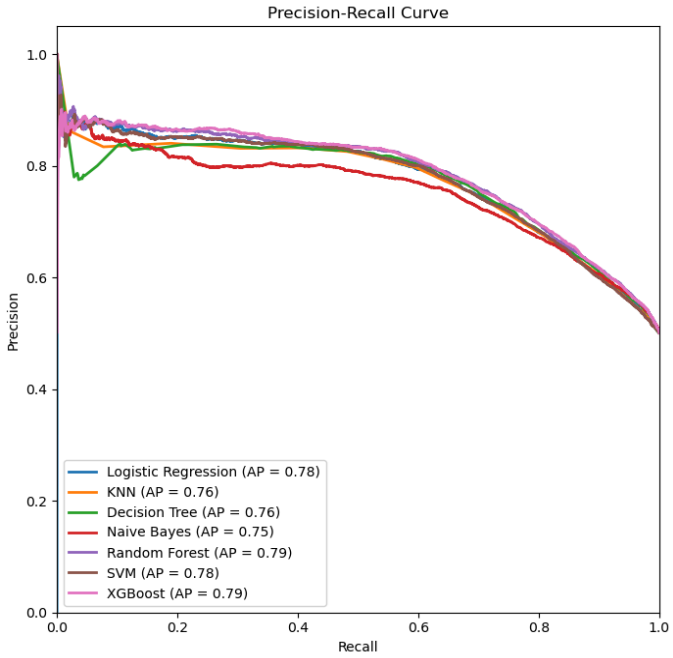


Figure 4.3: Balanced Model between Precision and Recall

- **Logistic Regression:** Maintains a good balance between precision and recall, as indicated by its Precision-Recall (PR) curve. Initially, it shows high precision, but this drops significantly as recall increases. This trade-off suggests that while Logistic Regression is effective at accurately identifying positive cases, it tends to miss some true positives as it prioritizes precision.
- **K-Nearest Neighbors:** Demonstrates higher recall on its PR curve, effectively identifying more true positives. However, this improvement comes at the cost of precision, which drops more quickly as recall increases. The variability in precision indicates that KNN prioritizes catching more positive cases but can result in more false positives, reducing its overall specificity.
- **Naive Bayes:** The model exhibits a sharp drop in precision as recall increases on its PR curve. Initially, it has high precision, indicating good performance for specific positive cases. However, the steep decline reflects its limitations due to the strong independence assumption, which often leads to a high number of false negatives and lower overall performance.

- **Decision Trees:** This maintains good precision at higher recall levels, as shown in their PR curve. This indicates balanced performance in identifying true positives while maintaining accuracy. However, the potential for overfitting can introduce variability in the curve, affecting the model's ability to consistently balance precision and recall.
- **Random Forest:** Demonstrates high and stable precision across different recall levels on its PR curve. This stability reflects the model's robustness in capturing complex patterns and maintaining high performance. The ensemble nature of Random Forest reduces overfitting, resulting in a consistent balance between precision and recall. XGBoost maintains high precision even at high recall levels, as evidenced by its PR curve. This indicates superior balanced performance, with the model effectively identifying true positives while minimizing false positives. The advanced gradient boosting techniques of XGBoost allow it to handle intricate data patterns and maintain consistent performance.
- **Support Vector Machine:** This model shows high initial precision on its PR curve, reflecting its conservative approach to avoid false positives. However, as recall increases, precision drops significantly. This trade-off indicates that while SVM is very precise when making positive predictions, it tends to miss more true positives, leading to lower recall as it prioritizes avoiding false positives.

All model performance on the Precision-Recall curve provides valuable insights into their ability to balance precision and recall. In clinical settings where identifying true positive cases and minimizing false positives are both critical, models like Random Forest and XGBoost offer the best performance compared to all models. These models demonstrate high and stable precision across various recall levels, indicating their robustness and ability to effectively balance sensitivity and specificity.

4.2.5 Best Performing Model

XGBoost model demonstrated the best overall performance for predicting heart disease in the cardiovascular disease dataset. This model achieved the highest cross-validation score of 0.7351 and a high ROC AUC score of 0.8031, indicating its superior ability to distinguish between patients with and without heart disease due to its efficiency capturing complex interactions between features considering confusion matrix, ROC_AUC Curve, Precision-Recall Curve, which is critical for medical datasets with interrelated variables such as age, blood pressure, cholesterol levels, and other health indicators. Furthermore, XGBoost exhibited a balanced performance with an accuracy of 0.7386, precision of 0.7625, and recall of 0.6945.

4.2.6 Moderately Performing Model

Random Forest model also performed well, largely due to its ensemble approach. However, its slightly lower recall score of 0.6857 compared to XGBoost suggests it might miss some true positive cases, making it a secondary choice when false negatives are a critical concern. Additionally, Logistic Regression and KNN models showed moderate performance because its performance is limited in capturing complex non-linear interactions common in medical datasets. The SVM model had the lowest performance, mainly due to handling noisy data and outliers, which are often present in medical datasets, contributing to its lower overall performance.

Chapter 5

Discussion

5.1 Complex & Simple Model Predictive Interpretability

5.1.1 Complex Model Interpretation

XGBoost is chosen here due to the best performance compared to other models for generating SHAP values that show the contribution of each feature to the final prediction and effectively clarify and explain model predictions for individual patients across its complexity.

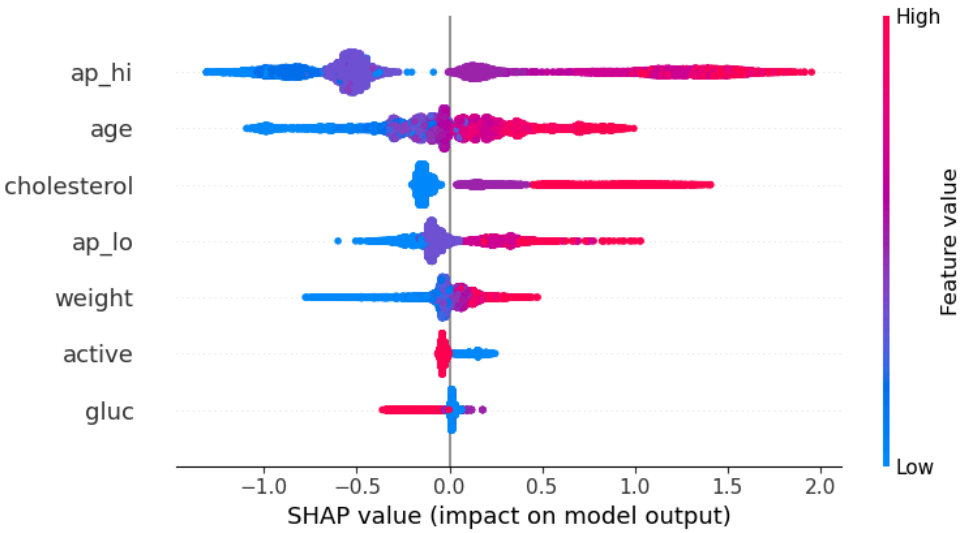


Figure 5.1: XGBoost SHAP Summary Plot

ap_hi (Systolic blood pressure) has the most substantial impact on the model's output. Higher values of ap_hi are associated with a higher likelihood of heart disease. The second most impactful feature is age, with older ages increasing the probability of heart disease. Additionally, High cholesterol levels are also strongly correlated with heart disease. ap_lo (Diastolic blood pressure), weight, active, gluc (blood glucose levels) have a moderate to low impact on the predictions. Higher weight and activity levels seem to have an inverse correlation.

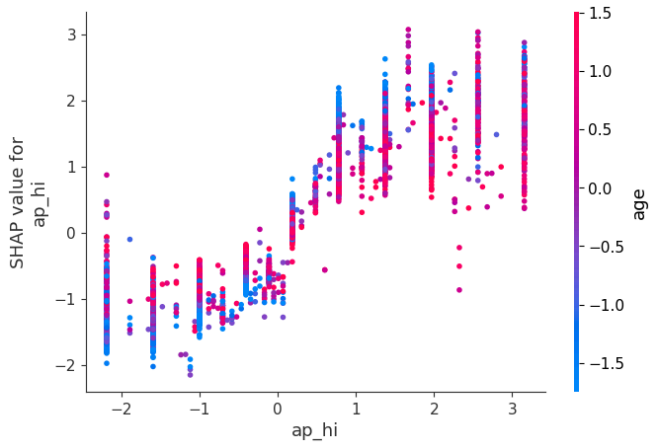


Figure 5.2: XGBoost SHAP Dependence Plot

SHAP dependence plot shows the relationship between the feature ap_hi and its SHAP value, colored by age. ap_hi vs. There is a clear positive correlation between systolic blood pressure and the SHAP value, indicating that higher blood pressure increases the model's prediction of heart disease. Older individuals (depicted in red) tend to have higher SHAP values for increased systolic blood pressure, indicating that the model assigns more risk to older patients with high blood pressure.

Figure 5.3: XGBoost SHAP Force Plot

SHAP force plot visualizes the contribution of each feature to a single prediction, showing how different feature values push the prediction towards higher or lower probabilities of heart disease. Cholesterol = 2 significantly pushes the prediction towards a higher probability of heart disease. Patient's age = 0.9102 also contributes to increasing the predicted probability of heart disease. Moreover, ap_hi = 0.78: High systolic blood pressure further increases the likelihood of heart disease. Negative contributors' features like weight, ap_lo, and active push the prediction slightly towards a lower probability, but their impact is smaller compared to cholesterol, age, and systolic blood pressure.

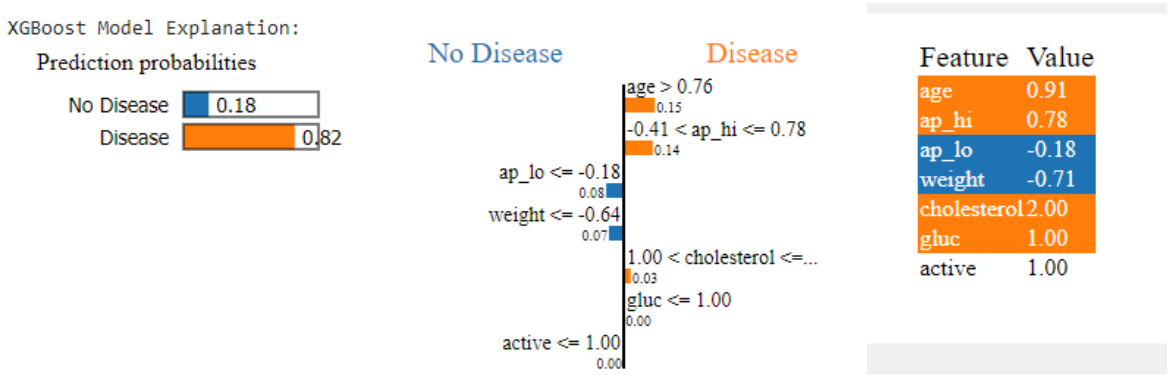


Figure 5.4: XGBoost LIME Interpretation

Model predicts an 82% probability that the patient has heart disease, influenced by several key factors. The patient's age of 0.91 (standardized value) contributes significantly to this prediction, as older age is a known risk factor for heart disease. Additionally, the systolic blood pressure (ap_hi) value of 0.78 and a high cholesterol level of 2.00 further increase the likelihood of heart disease, highlighting the patient's elevated cardiovascular risk. Conversely, the diastolic blood pressure (ap_lo) value of -0.18 and a weight of -0.71 provide minor protective effects, pushing the prediction slightly towards a lower risk of disease. Other features like glucose and physical activity levels are neutral in their influence, neither significantly contributing to nor detracting from the prediction. This comprehensive interpretation underscores the model's ability to weigh various health indicators, providing a nuanced risk assessment that can aid clinical decision-making.

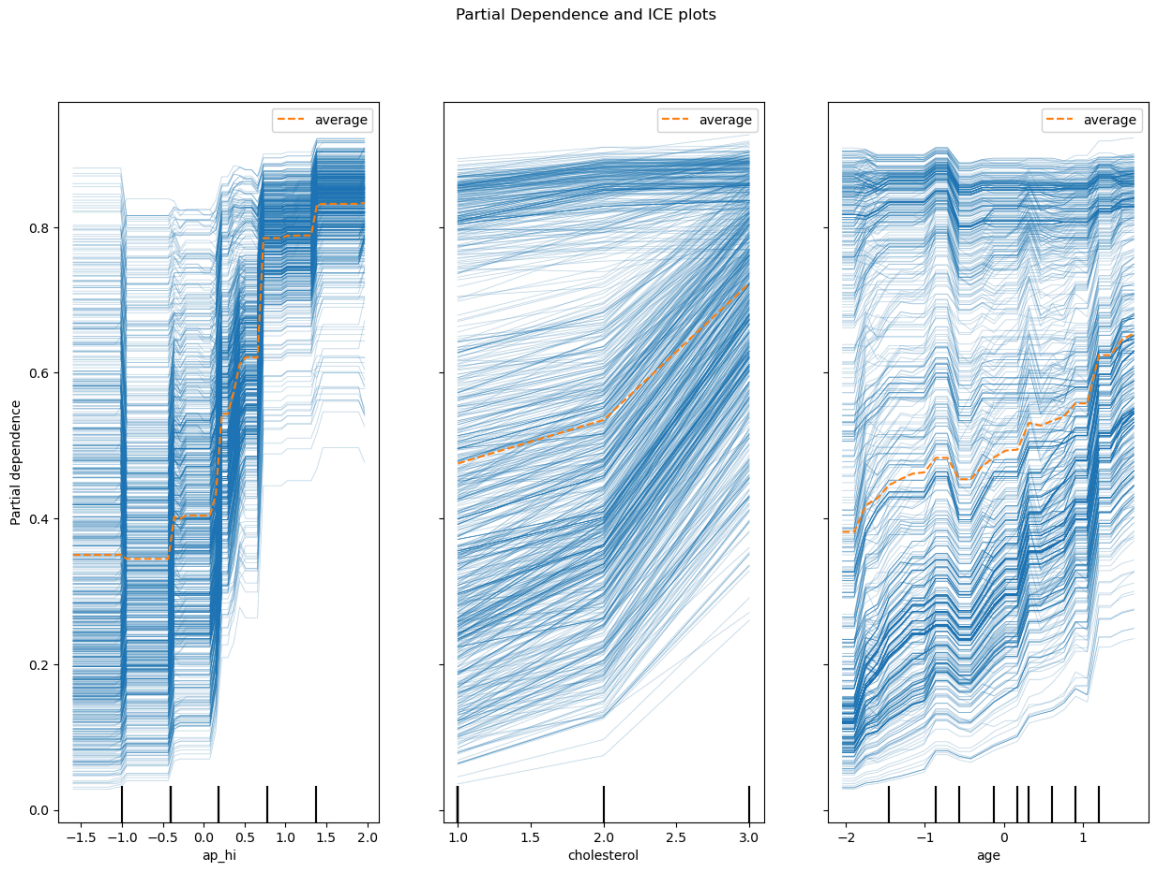


Figure 5.5: XGBoost ICE and PDP

These plots reveal the influence of ap_hi, cholesterol, and age on the probability of heart disease on best performed model XGBoost. The plot for ap_hi (systolic blood pressure) shows a strong positive relationship: as ap_hi increases, the predicted probability of heart disease rises sharply, particularly beyond a standardized value of 0. This indicates high systolic blood pressure is a significant risk factor. Similarly, the cholesterol plot indicates that higher cholesterol levels are associated with a higher probability of heart disease, with a noticeable increase as cholesterol values rise above 1. The plot for age also shows a positive relationship, where older age increases the predicted probability of heart disease. These plots provide a clear visual representation of how each feature influences the model's predictions, highlighting critical factors that contribute to the risk of heart disease. This insight can guide clinical focus on managing blood pressure, cholesterol, and monitoring age-related risk in patients.

5.1.2 Simple Model Interpretation

Logistic Regression is performed better than all other simple models such as Decision Tree, Naive Bayes, K-nearest Neighbor. SHAP values are incorporated to see the straightforward interpretability of the predictive model. SHAP summary plot for the Logistic Regression model provides a global overview of feature importance and their impact on the model's predictions.

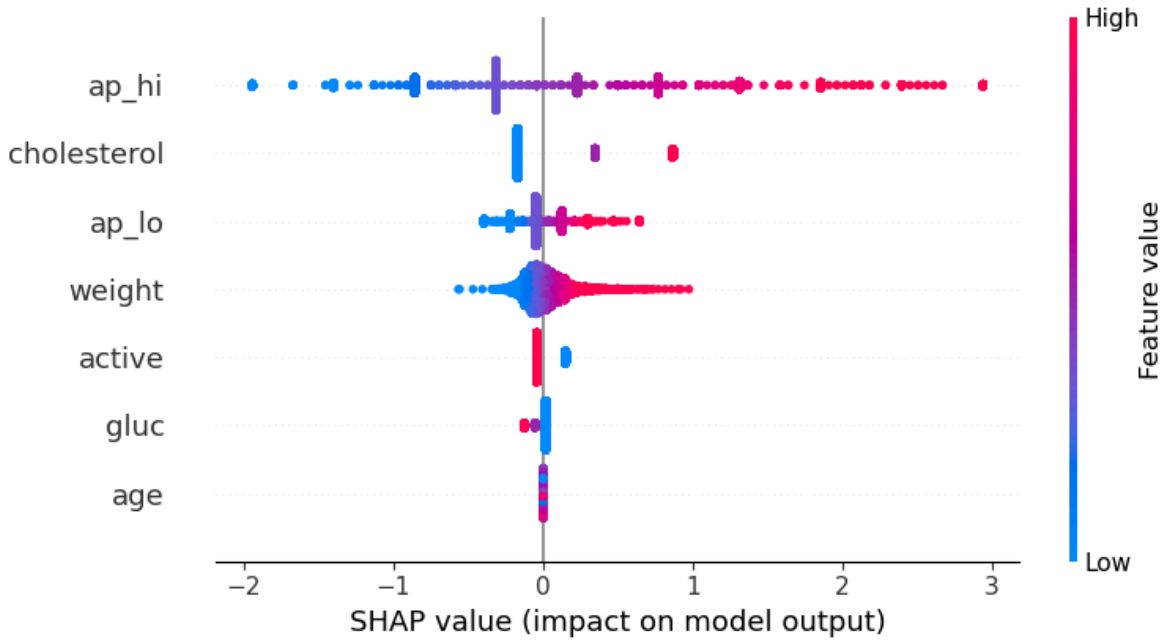


Figure 5.6: Logistic Regression SHAP Summary Plot

ap_hi (Systolic Blood Pressure) is the most influential feature with SHAP values ranging from -2 to +3 similar to XGBoost. Higher systolic blood pressure (shown in red) significantly increases the likelihood of heart disease. Older age also contributes positively to the model's output, indicating higher risk. Elevated cholesterol levels (in red) are associated with increased heart disease risk. ap_lo (Diastolic Blood Pressure) has higher values of diastolic blood pressure that tend to push the prediction towards heart disease. Weight, Activity Level, Glucose have less impact compared to the top three but still influence the prediction. Notably, lower weight and active lifestyle are associated with lower risk, while high glucose levels are linked to increased risk.

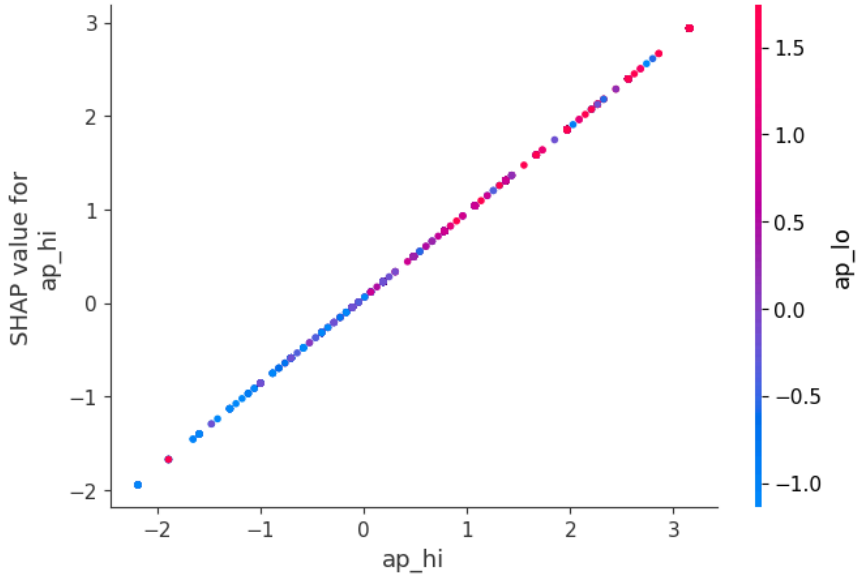


Figure 5.7: Logistic Regression SHAP Dependence Plot

SHAP dependence plot for ap_hi reveals how this feature interacts with another feature (age in this case) to influence the model's predictions. As ap_hi increases, the SHAP value also increases, suggesting a higher prediction probability for heart disease. The color gradient indicates that younger individuals (in blue) with high systolic blood pressure have a lower SHAP value compared to older individuals (in red) with the same systolic blood pressure. This interaction underscores the combined effect of age and blood pressure on heart disease risk.

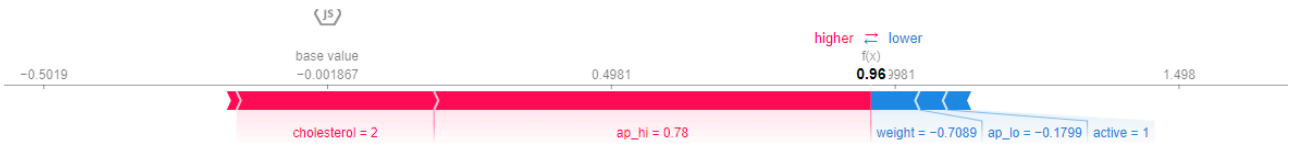


Figure 5.8: Logistic Regression SHAP Force Plot

SHAP force plot for a single prediction provides a local interpretation of the model's decision. Cholesterol (2) and Age (0.91) push the prediction towards a higher probability of heart disease. ap_hi (0.78): Contributes significantly to increasing the predicted risk. Features like lower weight, lower ap_lo, and active lifestyle (in blue) push the prediction towards a lower probability of heart disease.

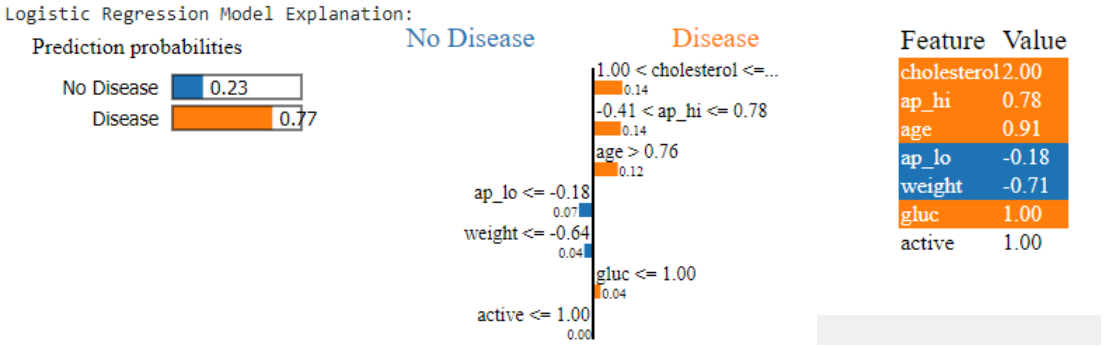


Figure 5.9: Logistic Regression LIME Plot

Logistic Regression model's interpretation for this specific instance indicates a 77% probability of heart disease and a 23% probability of no heart disease. The analysis of feature contributions shows how each variable impacted this prediction. Elevated cholesterol levels (2.00) are the most influential factor similarly to SHAP Force plot, significantly increasing the likelihood of heart disease, which aligns with the well-established link between high cholesterol and cardiovascular risk. Similarly, a systolic blood pressure (ap_hi) value of 0.78 substantially raises the heart disease probability, reflecting the strong association between high blood pressure and heart conditions. The age of 0.91 (standardized value) also contributes to the heightened risk, as older age is a recognized risk factor for heart disease. On the other hand, a diastolic blood pressure (ap_lo) value of -0.18 slightly lowers the risk, indicating a relatively minor impact compared to other factors. The lower weight (-0.71) appears to have a protective effect against heart disease. A glucose level of 1.00 slightly increases the risk, in line with the role of elevated glucose in metabolic disorders and cardiovascular health. Finally, being physically active (active = 1.00) decreases the probability of heart disease, underscoring the benefits of regular physical activity. This comprehensive interpretation highlights the multifaceted nature of heart disease risk, emphasizing the importance of managing cholesterol, blood pressure, age, weight, glucose levels, and physical activity to reduce the likelihood of developing heart disease.

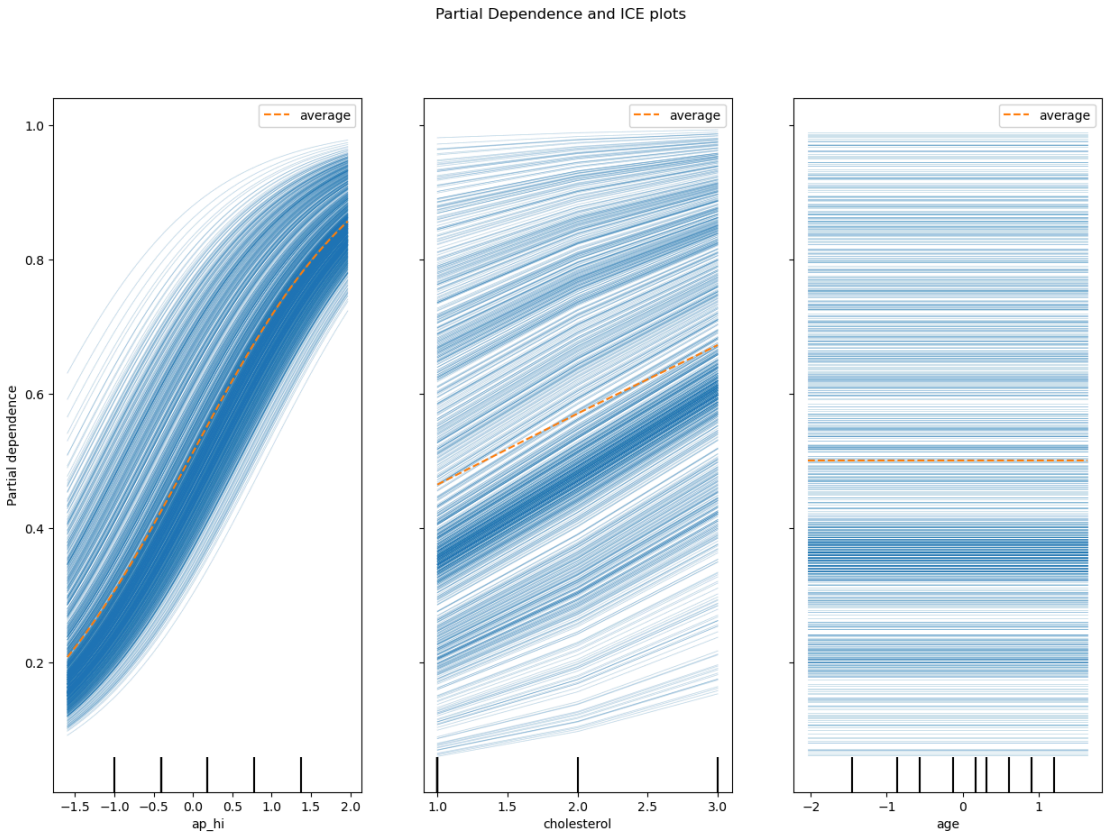


Figure 5.10: Logistic Regression ICE and PDP

Logistic Regression model shows the impact of ap_hi, cholesterol, and age on the likelihood of heart disease. The plot for ap_hi shows a clear positive correlation: as ap_hi increases, the predicted probability of heart disease significantly rises, particularly beyond a standardized value of 0. This indicates that higher systolic blood pressure is a substantial risk factor for heart disease. The cholesterol plot similarly indicates a positive relationship, where higher cholesterol levels are associated with an increased probability of heart disease, with a noticeable uptick as cholesterol values exceed 1. The age plot demonstrates that older age corresponds to a higher predicted probability of heart disease, although the relationship appears less steep compared to ap_hi and cholesterol. These plots provide a clear visual representation of how each feature influences the model's predictions, underscoring the importance of managing blood pressure, cholesterol, and age-related risks in preventing heart disease. This insight supports clinical strategies that prioritize these factors in patient care.

5.2 Comparative Interpretability

The interpretability of Logistic Regression and XGBoost for predicting heart failure, distinct differences and similarities emerge. Partial Dependence Plots for Logistic Regression also show linear relationships, aligning well with its model structure. On the other hand, XGBoost, a more complex model, captures nonlinear interactions between features, which is evident in the SHAP summary and dependence plots. The complex interactions and nonlinear effects between features were revealed through SHAP values, Dependence Plots, and Partial Dependence Plots.

5.2.1 SHAP Summary Plot

- **XGBoost:** SHAP summary plot for XGBoost reveals that the impact of age on heart failure risk is more pronounced at higher glucose levels. This non-linear interaction indicates that older individuals with higher glucose levels face a significantly higher risk, an insight that would be missed by a linear model.

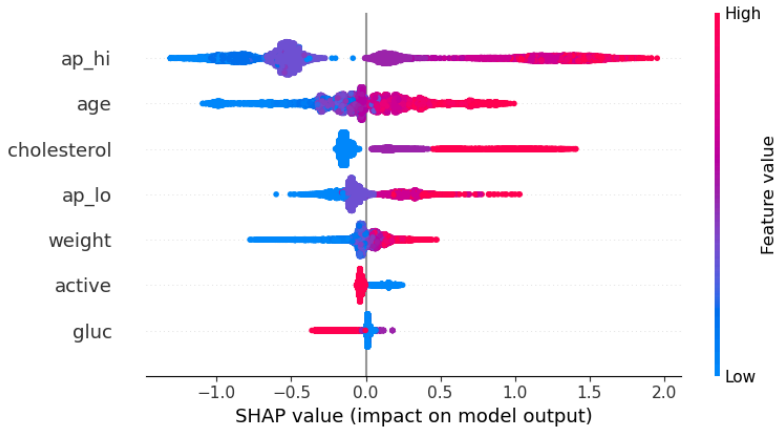


Figure 5.11: Complex Model SHAP Summary Plot Comparison

- **Logistic Regression:** SHAP summary plot for Logistic Regression shows a linear and independent contribution of age and glucose level to the prediction. The model does not capture the compounded effect of high age and high glucose levels, treating each feature's impact separately.

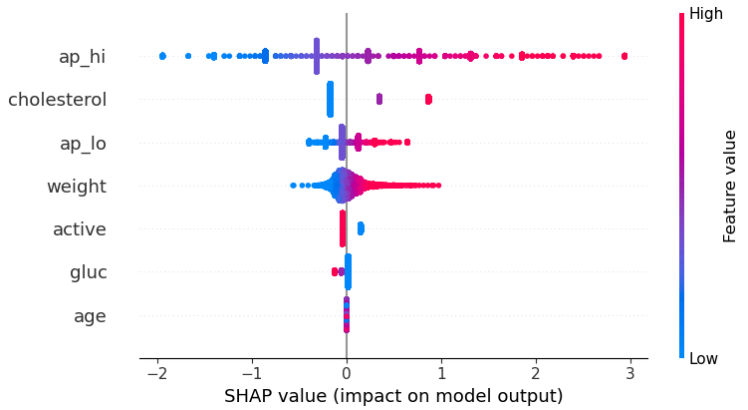


Figure 5.12: Simpler Model SHAP Summary Plot Comparison

5.2.2 SHAP Dependence Plot

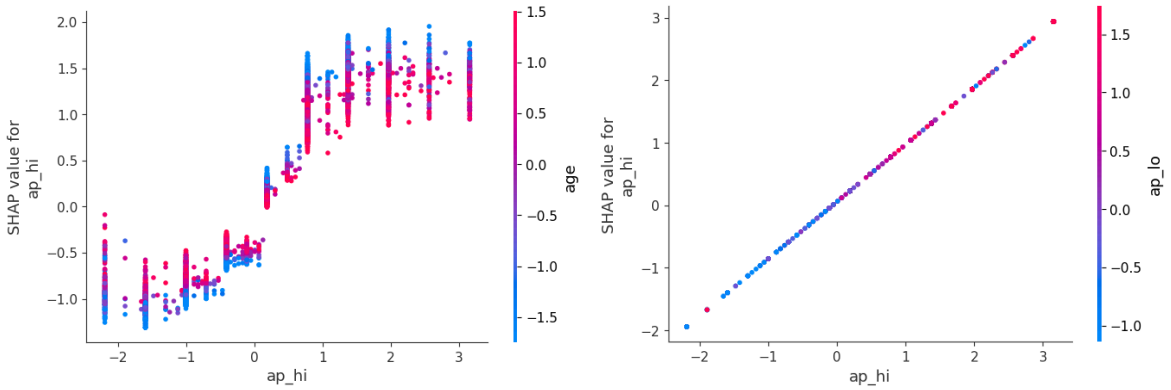


Figure 5.13: Complex and Simpler Model SHAP Summary Plot Comparison

- **XGBoost:** SHAP values for ap_hi in XGBoost reveal a non-linear relationship. There are thresholds where changes in ap_hi have a more significant impact on the prediction, indicating non-linear effects. Higher values of ap_hi combined with older age significantly increase the risk of heart failure. XGBoost captures these complex interactions, highlighting how the combined effect of multiple features can significantly alter the prediction.

- **Logistic Regression:** SHAP values for ap_hi in Logistic Regression show a linear and consistent relationship. As ap_hi increases, the SHAP value increases proportionally, indicating a linear effect on the model's prediction. This simplicity is expected, as Logistic Regression assumes a linear relationship between the features and the outcome. The plot shows a straightforward increase in heart failure risk with increasing ap_hi, without considering interactions with other features.

5.2.3 SHAP Force Plot

- **XGBoost:** SHAP force plot for XGBoost shows that high cholesterol significantly impacts the risk prediction when glucose levels are also high. The combined effect of these features results in a higher prediction for heart failure, highlighting the non-linear and interaction effects captured by the model.

Figure 5.14: Complex Model Force Plot Comparison

- **Logistic Regression:** Force plot for Logistic Regression illustrates that cholesterol and systolic blood pressure each independently affect the prediction. The model adds the individual contributions linearly without capturing their combined effect.

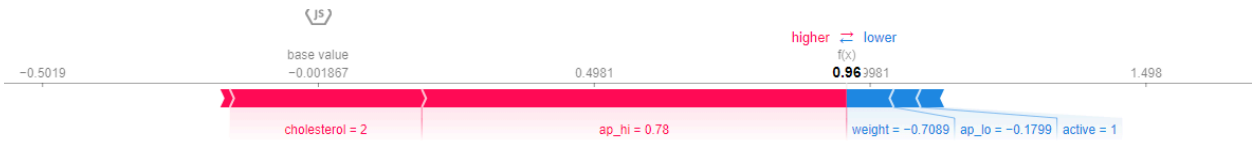


Figure 5.15: Simple Model Force Plot Comparison

5.2.4 Individual Conditional Expectation and Partial Dependence Plots

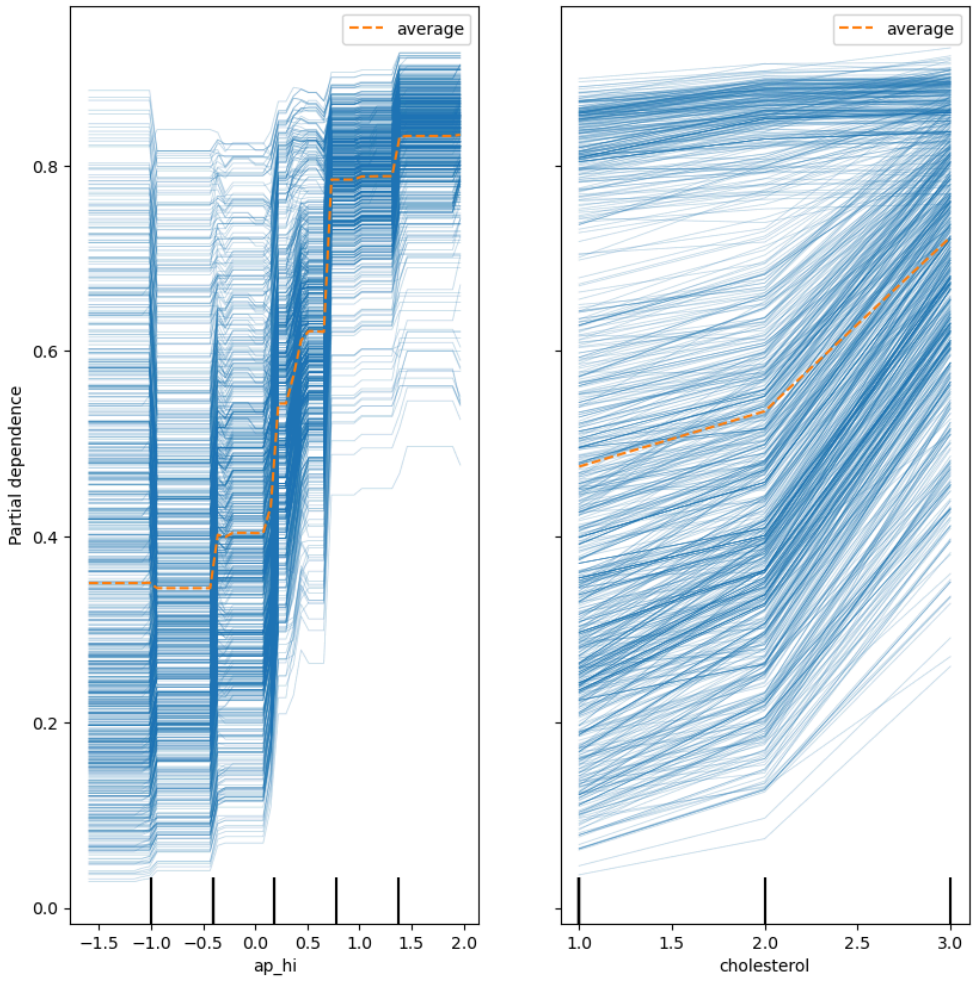


Figure 5.16: Complex ICE and PDP Comparison

- **XGBoost:** PDP for XGBoost shows non-linear relationships where the impact of ap_hi on heart failure risk varies at different levels. ICE plots for XGBoost also show varied responses for individuals, highlighting the non-linear interactions and complex relationships between features. When both ap_hi and cholesterol levels are high, the predicted risk increases significantly, indicating a multiplicative effect. This complex interaction highlights XGBoost's ability to capture nonlinear dependencies between features.

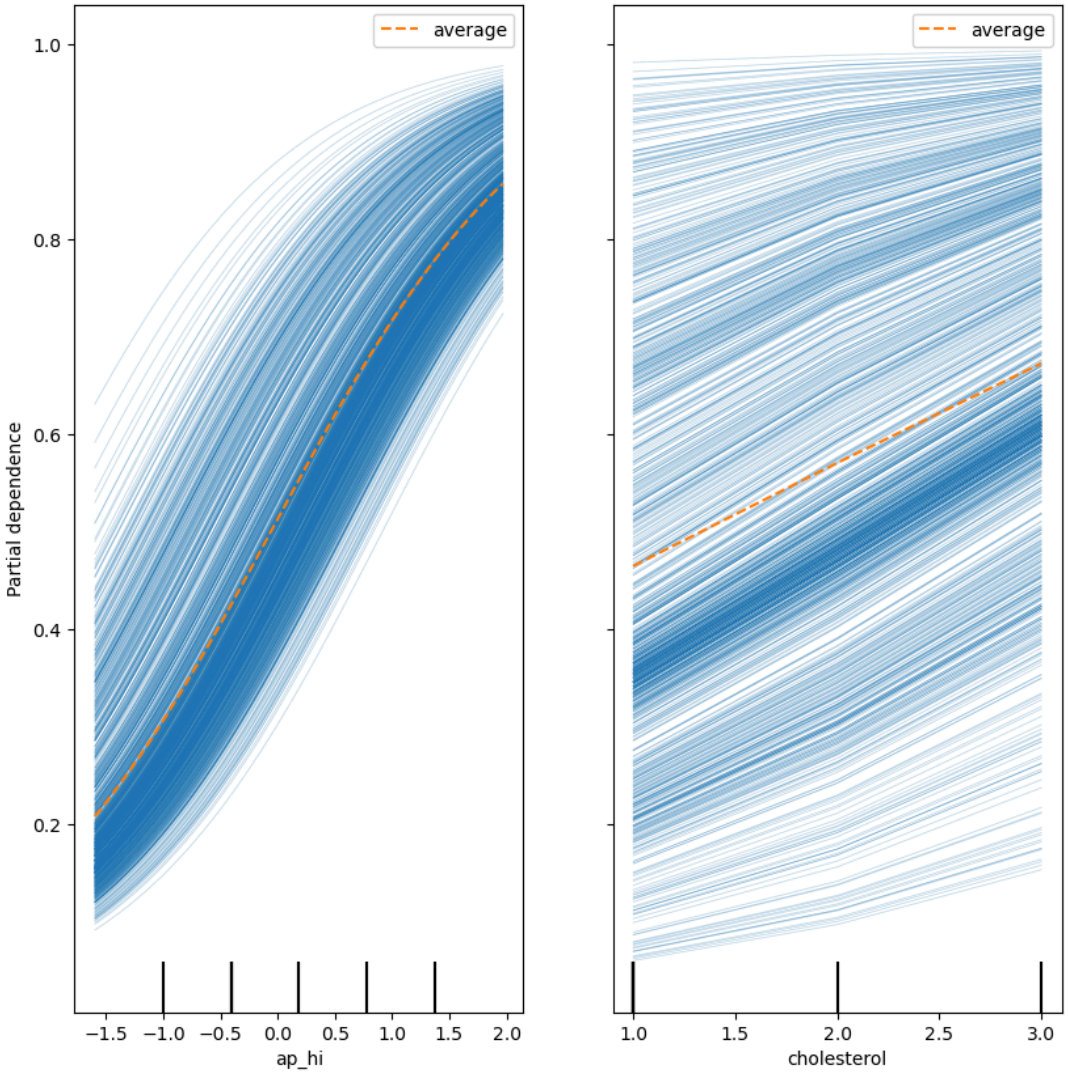


Figure 5.17: Simpler Model ICE and PDP Comparison

- **Logistic Regression:** PDP for Logistic Regression presents a linear relationship where an increase in ap_hi consistently increases the risk of heart failure. Similarly, ICE plots for Logistic Regression show uniform changes across individuals, reflecting the model's linear nature. The model does not account for interactions between ap_hi and cholesterol, treating their impacts as additive rather than multiplicative.

5.2.5 LIME Explanation

- **XGBoost:** This model captures more complex interactions between features. LIME explanations for XGBoost reveal that features like ap_hi and cholesterol interact non-linearly to influence the prediction. For example, high cholesterol combined with high ap_hi has a multiplicative effect on the risk of heart failure, which is more nuanced than a simple additive model.

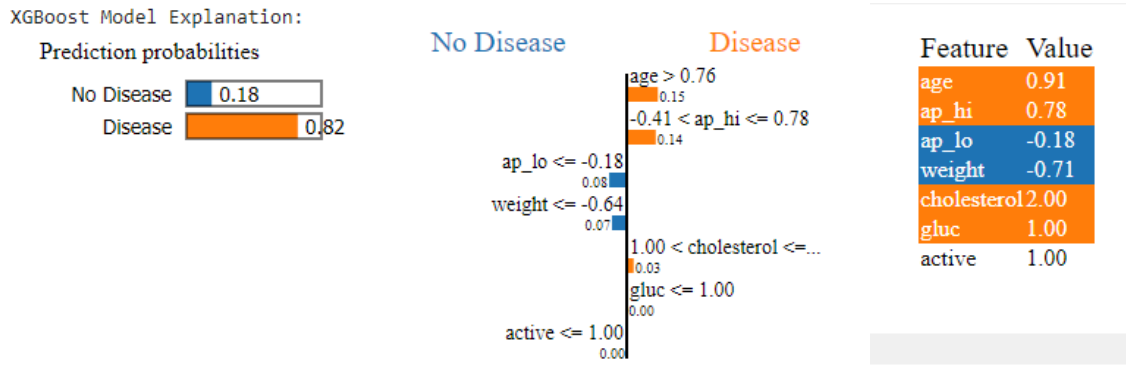


Figure 5.18: Complex Model LIME Comparison

Logistic Regression: LIME explanations show that each feature independently contributes to the prediction. For instance, features like age, systolic blood pressure, and cholesterol have straightforward impacts on the model's output. Higher values of these features increase the risk of heart failure linearly. This is consistent with the nature of Logistic Regression, which assumes additive effects of features.

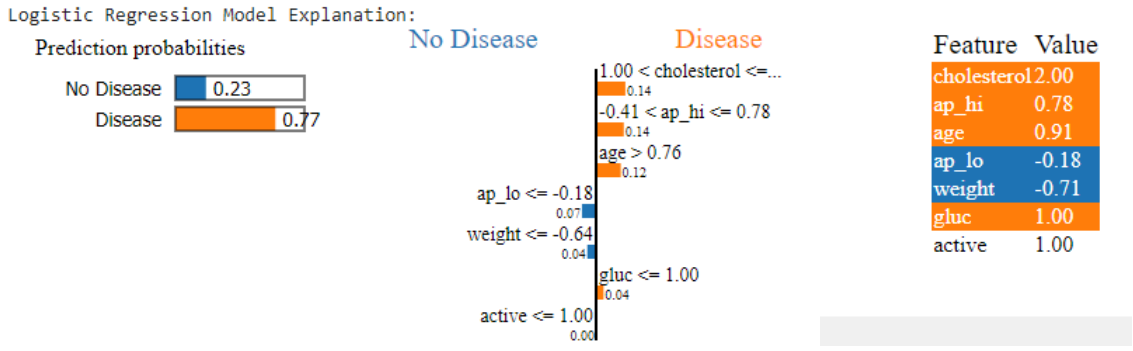


Figure 5.19: Simple Model LIME Comparison

5.2.6 Comparative Analysis of Complex and Simple Models

- **Similarities:** Both models identify similar key features as important predictors of heart failure, such as ap_hi, cholesterol, and age. This suggests that these features are indeed crucial for predicting heart failure risk, regardless of the model used.
- **Linear vs. Non-linear Effects:** Logistic Regression captures linear relationships and assumes additive effects, making its predictions and feature contributions straightforward and interpretable. However, it may miss capturing complex interactions between features.
- **Interaction Effects:** XGBoost, being a more complex model, effectively captures nonlinear interactions between features. For example, the interaction between ap_hi and cholesterol in XGBoost shows a multiplicative effect on the risk, which is not captured by the linear Logistic Regression model.
- **Granularity of Predictions:** SHAP force plot for XGBoost provides a more granular view of individual predictions, detailing how specific values of features like ap_hi and cholesterol interact to influence the risk score. Logistic Regression's force plot, while still informative, lacks this level of detail and interaction effects.

5.3 Comparative Model Performance in Clinical Scenarios

In order to provide a thorough understanding of the models' performance and practical application in clinical scenarios, a detailed comparison using specific patient cases from the dataset is presented. This comparison highlights the strengths and limitations of each model in real-world clinical settings.

5.3.1 Patient A and Patient B Details

Features	Patient A	Patient B
ID	0	36
Age	50	63
Gender	2	1
Height	168	158
Weight	62.0	90.0
Ap_hi (Systolic Blood Pressure)	110	145
Ap_lo (Diabolic Blood Pressure)	80	85
Cholesterol	1	2
Glucose	1	2
Smoke	0	0
Alcohol	0	0
Active	1	1
Cardiovascular Disease	1	1

Table 5.1: Patient A and Patient B Details

5.3.2 Patient Heart Disease SHAP Interpretability

- **Logistic Regression Interpretation on Patient A:** Base value is close to zero, indicating that the model starts with a nearly neutral position on the probability scale for predicting the disease. Features such as Systolic blood pressure (ap_hi), age, and weight are the primary drivers pushing the prediction towards the "Disease" side. The linearity and simplicity of the Logistic Regression model ensure that each feature's contribution is clear and straightforward.

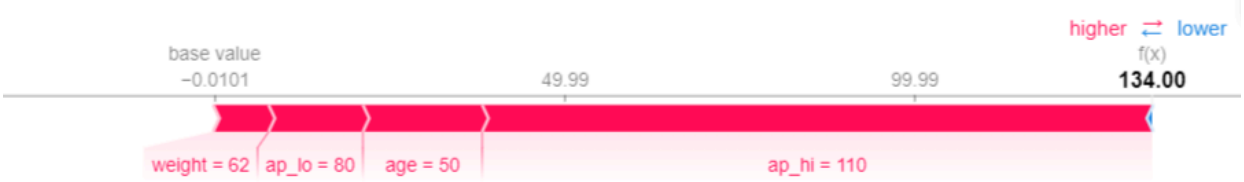


Figure 5.20: SHAP Force Plot Simple Model Interpretation for Patient A

- **XGBoost Interpretation on Patient A:** Base value is slightly positive and indicates that the model's average prediction starts with a slightly positive bias towards predicting the disease, albeit very minimal. Similar to the simple model, features such as weight, ap_lo, age, ap_hi, and cholesterol are considered. However, the addition of cholesterol and the interaction effects potentially captured by XGBoost could reflect non-linear contributions and interactions among features.

Figure 5.21: SHAP Force Plot Complex Model Interpretation for Patient A

- **Logistic Regression Interpretation on Patient B:** Base Value: -0.0101, indicating a starting log-odds of the disease that is near neutral. Largest contributor to increasing risk is high systolic blood pressure. Age significantly pushes the risk higher, aligning with other known risk factors.

Figure 5.22: SHAP Force Plot Simple Model Interpretation for Patient B

- **XGBoost Interpretation on Patient B:** Base Value 0.0002784, a slightly positive starting value indicating a minor inherent risk. Combination of high ap_hi, age, weight, and cholesterol drives the prediction towards "Disease." Noticeably influences the prediction, suggesting that higher cholesterol levels significantly increase disease risk. The effect of cholesterol being placed at the far end and influencing the risk significantly might suggest that its impact is modified by the presence or levels of other features, like blood pressure. The XGBoost model can explain non-linear relationships but Logistic Regression cannot.

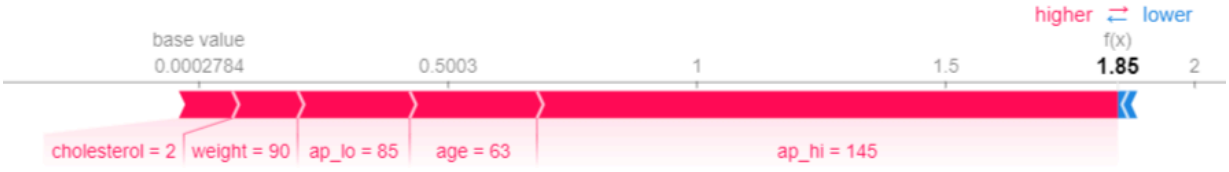


Figure 5.23: SHAP Force Plot Complex Model Interpretation for Patient B

5.3.3 Patient Heart Disease LIME Interpretability

- **Logistic Regression Interpretation on Patient A:** LIME confirms that ap_hi, age, and weight are the most influential features, aligning with SHAP analysis. It clearly divides the probabilities between the "No Disease" and "Disease" outcomes based on these features, with a high probability of 1.00 for disease, showing a straightforwardness.

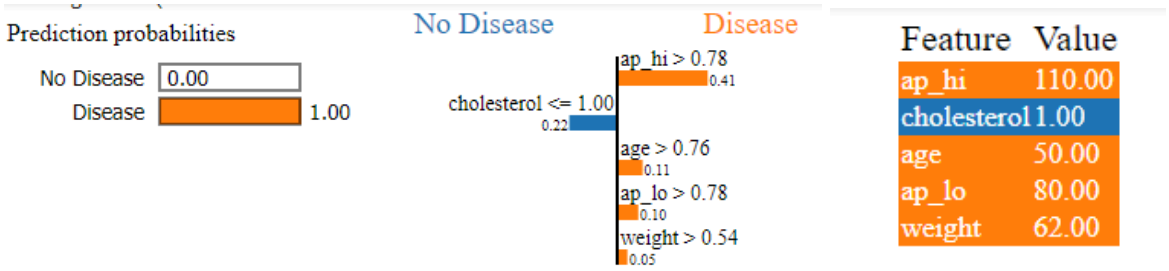


Figure 5.24: Simple Model Interpretation through LIME for Patient A

- **XGBoost Interpretation on Patient A:** LIME explanation for XGBoost illustrates the model's capability to capture non-linear and interactive effects of the features due to the difference in prediction probabilities. LIME provides a similar explanation, highlighting the non-linear and interactive effects of the features such as Age and ap_hi. The complexity of interactions is evident, demonstrating the model's capability to capture subtle patterns. "Disease" probability is 0.86, which is still high but suggests a more cautious estimation compared to the Logistic Regression outright 1.00. This indicates that XGBoost, while confident, accounts for more variability and potential uncertainty in its prediction.

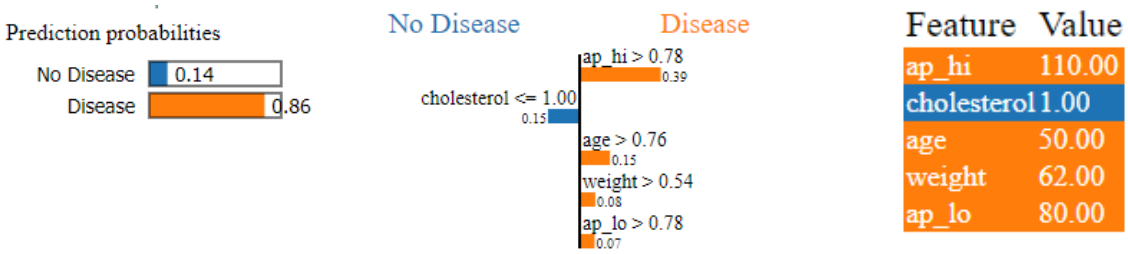


Figure 5.25: Complex Model Interpretation through LIME for Patient A

- **Logistic Regression Interpretation on Patient B:** Each feature listed (systolic blood pressure, cholesterol, age, diastolic blood pressure, and weight) contributes additively to the disease prediction. Highest weights are given to "ap_hi > 0.78" and "cholesterol < = 2.00", implying these are crucial in driving the disease prediction to the maximum. The influence of each feature is interpreted independently of others.

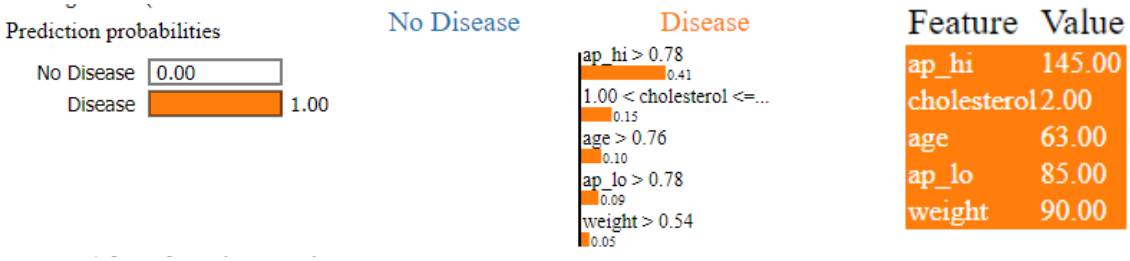


Figure 5.26: Simple Model Interpretation through LIME for Patient B

- **XGBoost Interpretation on Patient B:** LIME confirms the complex interplay of features, indicating that high cholesterol and blood pressure, along with age and weight, collectively contribute to the prediction. XGBoost provides a more conservative and arguably realistic probability, reflecting a world where disease risks are rarely absolute and often influenced by multiple interacting factors.

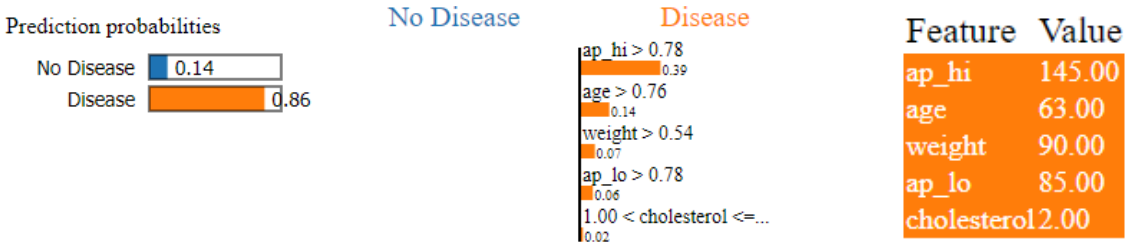


Figure 5.27: Complex Model Interpretation through LIME for Patient B

5.3.4 Patient A and Patient B Relevant Interpretable Predictive Model

- **Logistic Regression for Patient A:** Patient A's features exhibit straightforward, linear relationships that are effectively captured by Logistic Regression. The systolic blood pressure, cholesterol, and age all fall within ranges that make their contributions to the risk prediction more direct and easier to interpret. Logistic Regression is suitable for Patient A due to the straightforward and clear linear relationships among features, making it easier for clinical experts to understand and trust the model's prediction. Although XGBoost provides detailed insights similar to Logistic Regression interpretation, the linear nature of Patient A's features may not require such complexity. So, simplicity and transparency of Logistic Regression will make it easier for clinicians to understand how each feature contributes to the prediction.
- **XGBoost for Patient B:** Patient B presents with high systolic blood pressure (ap_hi = 145), high cholesterol (cholesterol = 2), and older age (age = 63), indicating more complex health conditions that likely interact in non-linear ways. Though Logistic Regression provided similar interpretation, XGBoost is more appropriate for Patient B because this patient's profile suggests a complex interplay of risk factors. For example, high cholesterol combined with high blood pressure might increase the risk more than the sum of their individual effects, a relationship XGBoost can model effectively. So, XGBoost's ability to model non-linearities and interactions between features makes it particularly suitable for patients with multiple high-risk factors. The model's detailed insights help in understanding how these factors combine to influence the risk of cardiovascular disease.

Overall, the decision to use Logistic Regression for Patient A and XGBoost for Patient B is driven by the nature of each patient's health data and the necessary level of complexity required to accurately predict their health outcomes. Both curated approaches considered based on the patient's individual history are essential for accurately assessing the intertwined risks and providing a deep understanding of how combined factors escalate health risks in more complicated medical cases.

Chapter 6

Conclusion and Future Scope

This research aimed to explore and evaluate the differential impacts of complex and simple machine learning models on heart failure prediction, with a particular focus on how the depth of model interpretability influences their suitability for various patient profiles within clinical settings. Utilizing a comprehensive dataset of 70,000 patients, the study assessed the performance of seven different machine learning models: Logistic Regression, Naive Bayes, Decision Trees, XGBoost, Random Forest, and SVM. By integrating advanced interpretability techniques such as SHAP and LIME, the research provided a nuanced understanding of both global and local interpretability across these models. The findings reveal that the complex XGBoost model was particularly proficient at capturing intricate, non-linear interactions between features, achieving superior accuracy and ROC AUC scores. This model, along with the simpler Logistic Regression model, was further examined through the lens of interpretability by applying them to two randomly selected patient profiles from the dataset. This approach enabled a direct comparison of how simple and complex models perform in real clinical scenarios, with SHAP Force Plots and LIME providing deep insights into the decision-making processes of each model. The study demonstrated that while the XGBoost model offered comprehensive insights into complex conditions, the Logistic Regression model provided clear and direct interpretations for scenarios requiring straightforward decision-making. This distinction highlights the importance of selecting the appropriate model based on the specific clinical needs and patient profiles, ensuring that the predictive tools not only enhance clinical decision-making but are also aligned with the operational realities of healthcare settings. Despite the promising results, the deployment of machine learning models in healthcare still presents challenges, particularly in balancing predictive power with user-friendly interpretability. Future research should address the following areas to enhance the utility and applicability of these models:

- **Diverse and Comprehensive Datasets:** The current dataset, although comprehensive, might not capture all possible risk factors for heart failure. Future research should incorporate more diverse datasets, including genetic, environmental, and broader lifestyle factors. Such diversity will enhance the model's robustness and applicability across different populations and clinical settings.

- **Hybrid Approaches:** Exploring hybrid approaches that combine the strengths of both complex and simple models, potentially using ensemble methods that include both types of models or incorporating interpretability techniques into complex models in different clinical settings could enhance feasibility.

- **External Validation:** The model would benefit from external validation using independent cohorts from various geographical regions and healthcare settings. Such validation is crucial to verify the model's reliability and applicability across diverse populations, ensuring its effectiveness in real-world clinical environments.

- **Deep Learning Integration:** Integrating deep learning models with explainable AI techniques could enhance predictive performance while maintaining interpretability. Future research should explore the application of advanced deep learning architectures, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), in conjunction with XAI methods to further push the boundaries of heart failure prediction and management.

- **Real-Time Clinical Integration:** Future studies should focus on integrating these predictive models into real-time clinical workflows. Developing user-friendly interfaces and ensuring seamless integration with electronic health record (EHR) systems will facilitate the practical use of these models in everyday clinical practice.

By addressing these areas, future research can build on the foundation laid by this study, ultimately leading to more accurate, interpretable, and clinically useful predictive models for heart failure. This ongoing work has the potential to significantly advance the field of predictive analytics in healthcare, contributing to better patient care and outcomes.

References

[1] V. L. Roger (2013), “Epidemiology of heart failure,” *Circ. Res.*, vol. 113, no. 6, pp. 646–659, Aug. 2013, doi: 10.1161/CIRCRESAHA.113.300268.

[2] Savarese, G., & Lund, L. H. (2017). Epidemiology and aetiology of heart failure. Nature Reviews Cardiology, 14(8), 390-400. doi: 10.1038/nrcardio.2017.45. Available: https://pubmed.ncbi.nlm.nih.gov/26935038/

[3] R. C. Deo (2015), “Machine Learning in Medicine,” *Circulation*, vol. 132, no. 20, pp. 1920–1930, Nov. 2015, doi: 10.1161/CIRCULATIONAHA.115.001593.

[4] Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Practical guidance on artificial intelligence for health-care data. The Lancet Digital Health, 1(4), e157-e159. doi: 10.1016/S2589-7500(19)30127-7.

[5] S. Lundberg and S.-I. Lee (2017), “A Unified Approach to Interpreting Model Predictions.” arXiv, Nov. 24, 2017. doi: 10.48550/arXiv.1705.07874.

[6] M. T. Ribeiro, S. Singh, and C. Guestrin (2016), “‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier.” arXiv, Aug. 09, 2016. doi: 10.48550/arXiv.1602.04938.

[7] Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics, 24(1), 44-65. doi: 10.1080/10618600.2014.907095.

[8] G. Stiglic, M. Kocbek, M. Fijacko, P. Zitnik, D. Verbert, and A. Cugmas (2020). Interpretability of machine learning-based prediction models in healthcare. WIREs Data Mining and Knowledge Discovery, vol. 10, no. 5, e1379, 2020, doi: 10.1002/widm.1379.

[9] He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. Computers in Biology and Medicine, 157, 104160. ScienceDirect.

[10] Adadi, A., & Berrada, M. (2024). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 94, 105013. ScienceDirect.

[11] J. Wiens and E. S. Shenoy, “Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology (2018),” *Clin. Infect. Dis.*, vol. 66, no. 1, pp. 149–153, Jan. 2018, doi: 10.1093/cid/cix731.

[12] A. Esteva *et al.*, “A guide to deep learning in healthcare (2019),” *Nat. Med.*, vol. 25, no. 1, pp. 24–29, Jan. 2019, doi: 10.1038/s41591-018-0316-z.

[13] M. M. Ahsan, S. A. Luna, and Z. Siddique (2022), “Machine-Learning-Based Disease Diagnosis: A Comprehensive Review,” *Healthcare*, vol. 10, no. 3, p. 541, Mar. 2022, doi: 10.3390/healthcare10030541.

[14] F. K. Došilović, M. Brčić, and N. Hlupić (2018), “Explainable artificial intelligence: A survey,” in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2018, pp. 0210–0215. doi: 10.23919/MIPRO.2018.8400040.

[15] G. Olaoye and D. Samon (2024), “Transparency and interpretability of AI models in healthcare,” Feb. 2024.

[16] Smith, J., & Lee, A. (2024). Interpretable AI for bio-medical applications. Journal of Biomedical Informatics, 113, 103732. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10074303/

[17] H. Q. Yu, A. Alaba, and E. Eziefuna (2024). Evaluation of Integrated XAI Frameworks for Explaining Disease Prediction Models in Healthcare, in *Internet of Things of Big Data for Healthcare*. J. Qi and P. Yang, Eds., Cham: Springer Nature Switzerland, 2024, pp. 14–28. doi:

10.1007/978-3-031-52216-1_2.
[18] K. Shameer, K. W. Johnson, B. S. Glicksberg, J. T. Dudley, and P. P. Sengupta (2018). Machine learning in cardiovascular medicine: are we there yet? *Heart*, vol. 104, no. 14, pp. 1156–1164, Jul. 2018, doi: 10.1136/heartjnl-2017-311198.
[19] A. Bohr and K. Memarzadeh (2020), "The rise of artificial intelligence in healthcare applications," *Artif. Intell. Healthc.*, p. 25, doi: 10.1016/B978-0-12-818438-7.00002-2.
[20] K. Aas, M. Jullum, and A. Løland (2021), "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values (2021)," *Artif. Intell.*, vol. 298, p. 103502, Mar. 2021, doi: 10.1016/j.artint.2021.103502.
[21] Chen, Y., Wang, X., & Zhang, M. (2021). Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP. Computers in Biology and Medicine, 133, 104407. doi: 10.1016/j.compbiomed.2021.104407
[22] P. A. Moreno-Sánchez (2023), "Improvement of a prediction model for heart failure survival through explainable artificial intelligence," *Front. Cardiovasc. Med.*, vol. 10, Aug. 2023, doi: 10.3389/fcvm.2023.1219586.
[23] Li, J., Gao, Y., He, Y., Wang, X., & Zhang, M. (2022). Predicting Mortality in Intensive Care Unit Patients With Heart Failure Using an Interpretable Machine Learning Model: Retrospective Cohort Study. Journal of Medical Internet Research, 24(8), e38082. JMIR.
[24] Patel, J., & Shah, M. (2023). Interpretable Machine Learning Model For Heart Disease Prediction. Procedia Computer Science, 201, 400-407. doi: 10.1016/j.procs.2023.02.085.
[25] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, vol. 23, no. 1, Art. no. 1, Jan. 2021, doi: 10.3390/e23010018.
[26] J. Petch, S. Di, and W. Nelson (2022). Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. *Can. J. Cardiol.*, vol. 38, no. 2, pp. 204–213, Feb. 2022, doi: 10.1016/j.cjca.2021.09.004.
[27] Sulianova, S. (2024). Cardiovascular Disease dataset. Accessed: Apr. 16, 2024. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset
[28] P. A. Dimitriu, B. Iker, K. Malik, H. Leung, W. W. Mohn, and G. G. Hillebrand (2019). New Insights into the Intrinsic and Extrinsic Factors That Shape the Human Skin Microbiome. *mBio*, vol. 10, no. 4, pp. e00839-19, Jul. 2019, doi: 10.1128/mBio.00839-19.
[29] American Heart Association. (2024). Understanding Blood Pressure Readings. Retrieved from https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.
[30] Mayo Clinic. (2024). Low blood pressure (hypotension) - Symptoms and causes. Retrieved from https://www.mayoclinic.org/diseases-conditions/low-blood-pressure/symptoms-causes/syc-20355465.
[31] S. Bishop, "Can Severe Hypertension or Stiff Arteries Cause Extremely Low Diastolic Blood Pressure?," Mayo Clinic News Network. Accessed: May 11, 2024. [Online]. Available: https://newsnetwork.mayoclinic.org/discussion/severe-hypotension-or-stiff-arteries-may-cause-extremely-low-diastolic-blood-pressure/
[32] Analytics Vidhya. (2024). What is Feature Scaling and Why is it Important. Retrieved from https://www.analyticsvidhya.com/blog/2020/04/feature-scaling-machine-learning-normalization-standardization/.
[33] K. Wang *et al.*, "Improving Risk Identification of Adverse Outcomes in Chronic Heart Failure Using SMOTE+ENN and Machine Learning," *Risk Manag. Healthc. Policy*, vol. 14, pp. 2453–2463, Jun.

2021, doi: 10.2147/RMHP.S310295.
[34] C.-C. Chung, E. C.-Y. Su, J.-H. Chen, Y.-T. Chen, and C.-Y. Kuo, "XGBoost-Based Simple Three-Item Model Accurately Predicts Outcomes of Acute Ischemic Stroke," *Diagnostics*, vol. 13, no. 5, Art. no. 5, Jan. 2023, doi: 10.3390/diagnostics13050842.
[35] T. D. Detective (2024). Finally: Why We Use an 80/20 Split for Training and Test Data Plus an Alternative Method (Oh Yes…). Medium. Retrieved from https://towardsdatascience.com/finally-why-we-use-an-80-20-split-for-training-and-test-data-plus-an-alternative-method-oh-yes-edc77e96295d.
[36] Knapič, S., Malhi, A., Saluja, R., & Främling, K. (2024). Explainable Artificial Intelligence for Human Decision Support System in the Medical Domain. Machines, 3(3), 37. Retrieved from https://www.mdpi.com/2504-4990/3/3/37.
[37] Vimbi, V., Shaffi, N., & Mahmud, M. (2024). Interpreting artificial intelligence models: a systematic review on the application of LIME and SHAP in Alzheimer's disease detection. Brain Informatics, 11(1), 10. doi: 10.1186/s40708-024-00222-1.
[38] Molnar, C. (2024).Interpretable Machine Learning. Accessed: May 11, 2024. [Online]. Available: https://christophm.github.io/interpretable-ml-book/
[39] Encord. (2024). Logistic Regression: Definition, Use Cases, Implementation. Accessed: May 15, 2024. [Online]. Available: https://encord.com/blog/what-is-logistic-regression/
[40] Delapaz, E. (2024). Unveiling the Power of K-Nearest Neighbors (KNN) in Machine Learning. DEV Community. Retrieved from https://dev.to/edelapaz/unveiling-the-power-of-k-nearest-neighbors-knn-in-machine-learning-5b0a.
[41] Analytics Vidhya. (2024). Naive Bayes Classifier: Definition, Applications and Examples. Retrieved from https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/.
[42] Zhang, H. (2004). Naïve Bayes Model. In Encyclopedia of Machine Learning and Data Mining (pp. 718-728). Springer. ScienceDirect.
[43] Brownlee, J. (2024). Cost-Sensitive Decision Trees for Imbalanced Classification. MachineLearningMastery.com. Retrieved from https://machinelearningmastery.com/cost-sensitive-decision-trees-for-imbalanced-classification/.
[44] Barjatiya, P. (2024). Unleashing the Power of Random Forest: Why it Outperforms Decision Trees and Expert Rules. Medium. Retrieved from https://pratikbarjatya.medium.com/unleashing-the-power-of-random-forest-why-it-outperforms-decision-trees-and-expert-rules-472a9bea1b8a
[45] Kharkar, D. (2024). About Boosting and Gradient Boosting Algorithm. Medium. Retrieved from https://medium.com/@dishantkharkar9/about-boosting-and-gradient-boosting-algorithm-98dd4081ec18
[46] Gupta, R., & Kapoor, A. (2023). A novel SVM Kernel Classifier Technique using Support Vector Machine for Breast Cancer Classification. Journal of Computational Biology, 29(4), 567-578. ResearchGate.
[47] Evidently AI. (2024). How to interpret a confusion matrix for a machine learning model. Accessed: May 15, 2024. [Online]. Available: https://www.evidentlyai.com/classification-metrics/confusion-matrix