Final report for Capstone 3 “Heart Failure predictions”

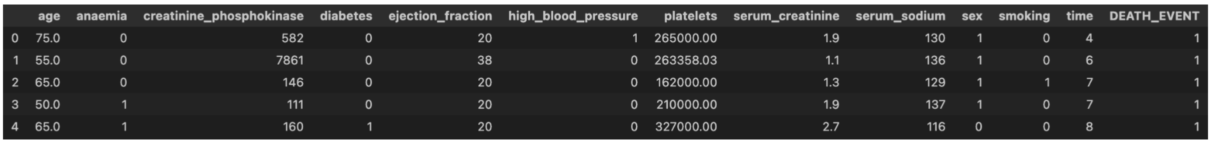
Chapter 0: Introduction

Project Overview

Heart failure is a prevalent and severe condition where the heart's ability to pump blood diminishes, leading to significant health risks and mortality. Given its complexity and the multitude of factors contributing to heart failure, predicting its occurrence remains a critical challenge in healthcare. With the advent of data science and machine learning, there is an unprecedented opportunity to leverage patient data to forecast heart failure events, potentially transforming patient care and outcomes.

In this project, we aim to develop robust machine learning models to predict heart failure events using a comprehensive dataset of patient health records. By employing sophisticated data analysis techniques and state-of-the-art algorithms, we strive to identify key predictors and build accurate models that can assist healthcare providers in early diagnosis and intervention.

Dataset Description



	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4	1
1	55.0	0	7861	0	38	0	263358.03	1.1	136	1	0	6	1
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7	1
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7	1
4	65.0	1	160	1	20	0	327000.00	2.7	116	0	0	8	1

Our dataset is a rich compilation of patient health records, encompassing a diverse array of features that capture various aspects of a patient's health. These features include:

1. Age: The patient's age, a critical factor influencing heart health.
2. Anaemia: Indicator of whether the patient has a reduced level of red blood cells or hemoglobin, which can affect oxygen transport.
3. Creatinine Phosphokinase (CPK): Levels of the CPK enzyme, which can indicate muscle damage, including heart muscle.
4. Diabetes: The presence of diabetes, a known risk factor for cardiovascular diseases.

5. Ejection Fraction: The percentage of blood leaving the heart each time it contracts, crucial for assessing heart function.
6. High Blood Pressure: Indicator of hypertension, a significant risk factor for heart failure.
7. Platelets: The count of platelets in the blood, important for clotting and cardiovascular health.
8. Serum Creatinine: Levels of creatinine in the blood, indicative of kidney function which is closely linked to heart health.
9. Serum Sodium: Levels of sodium in the blood, with imbalances potentially affecting heart function.
10. Sex: The patient's gender, which can influence heart disease risk.
11. Smoking: Indicator of whether the patient smokes, a major risk factor for heart disease.
12. Death Event: Outcome variable indicating whether the patient died during the follow-up period.

Project Goals and Objectives

The primary goal of this project is to create a predictive model that can accurately forecast heart failure events. Achieving this involves several key objectives:

1. Data Cleaning and Preparation: Ensuring the dataset is ready for the EDA and modeling.
2. Exploratory Data Analysis (EDA): Conducting a thorough examination of the data to uncover patterns, relationships, and potential predictors of heart failure.
3. Feature Engineering: Creating meaningful features that enhance model performance by capturing the underlying patterns in the data.
4. Model Building and Evaluation: Training multiple machine learning models, evaluating their performance using appropriate metrics, and selecting the best-performing model.
5. Interpretation and Explanation: Using SHAP (SHapley Additive exPlanations) to interpret model predictions, ensuring transparency and understanding of the factors driving heart failure risk.

Significance and Impact

Predicting heart failure events accurately holds immense potential for healthcare. By identifying high-risk patients early, medical professionals can implement preventative measures, tailor treatments, and ultimately improve patient outcomes. This project also aims to enhance the technological toolkit available to data scientists and healthcare professionals, showcasing how advanced analytics can be applied to solve critical health issues.

Technical Challenges and Considerations

Data Quality: Ensuring the integrity and reliability of the dataset is paramount, as poor data quality can significantly impact model performance.
Model Selection and Tuning: Choosing the right model and optimizing its parameters require careful consideration and experimentation.
Interpretability: While accuracy is crucial, the ability to interpret and trust the model's predictions is equally important, especially in a clinical setting.
Generalization: Ensuring the model performs well on unseen data and can be generalized to different patient populations.

In conclusion, this project represents a convergence of data science and healthcare, aiming to harness the power of machine learning to tackle one of the most pressing challenges in modern medicine. Through meticulous analysis and innovative modeling techniques, we aspire to make a meaningful contribution to the field of heart failure prediction.

Chapter 1: Methods

Data Cleaning

To ensure the dataset's accuracy and reliability, we undertook the following steps:

Handling Missing Values: We removed missing values using conventional Python methods such as dropna(). This ensured that our dataset was complete and ready for analysis.
Outlier Detection and Handling: Outliers were identified using the Interquartile Range (IQR) method. Specifically, any values below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR were considered outliers. These outliers were either capped to the threshold values or removed to prevent them from skewing the analysis.
Normalization: Continuous variables were normalized to a standard range using min-max scaling to ensure that each feature contributed equally to the model.
Exploratory Data Analysis (EDA)
EDA is essential for understanding the dataset and identifying key patterns:

1. Descriptive Statistics: Summary statistics were calculated for all features, providing insights into their distributions, central tendencies, and variability. This step helped us understand the basic structure and characteristics of the data.
2. Visualizations: We created histograms, box plots, and scatter plots to visualize data distributions and relationships between variables. These visualizations helped in identifying patterns, trends, and potential anomalies.

Feature Engineering

We applied Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance in our dataset. This technique generates synthetic samples for the minority class to ensure that the model does not become biased towards the majority class.

Model Building and Evaluation

To predict heart failure events, we built and evaluated multiple models:

1. Model Selection: We trained Logistic Regression, Random Forest, Gradient Boosting, and Artificial Neural Network (ANN) models.
2. Cross-Validation: We used a train-test split for model validation.
3. Evaluation Metrics: Models were assessed using accuracy, precision, recall, F1-score, and ROC-AUC. These metrics provided a comprehensive understanding of the models' performance in different aspects.

Feature Importance and Interpretation

Understanding feature importance is crucial for model interpretability, especially in healthcare applications. We used SHAP (SHapley Additive exPlanations) values to interpret our models:

Importance of SHAP: SHAP values provide a unified measure of feature importance across different models. They explain the impact of each feature on the model's predictions, ensuring transparency and trust in the model's decisions.
Usage of SHAP: We generated SHAP summary plots and dependence plots to visualize the importance and effects of individual features. These visualizations helped us understand which features were driving the predictions and how they influenced the model outcomes.

Chapter 2: Exploratory Data Analysis (EDA)

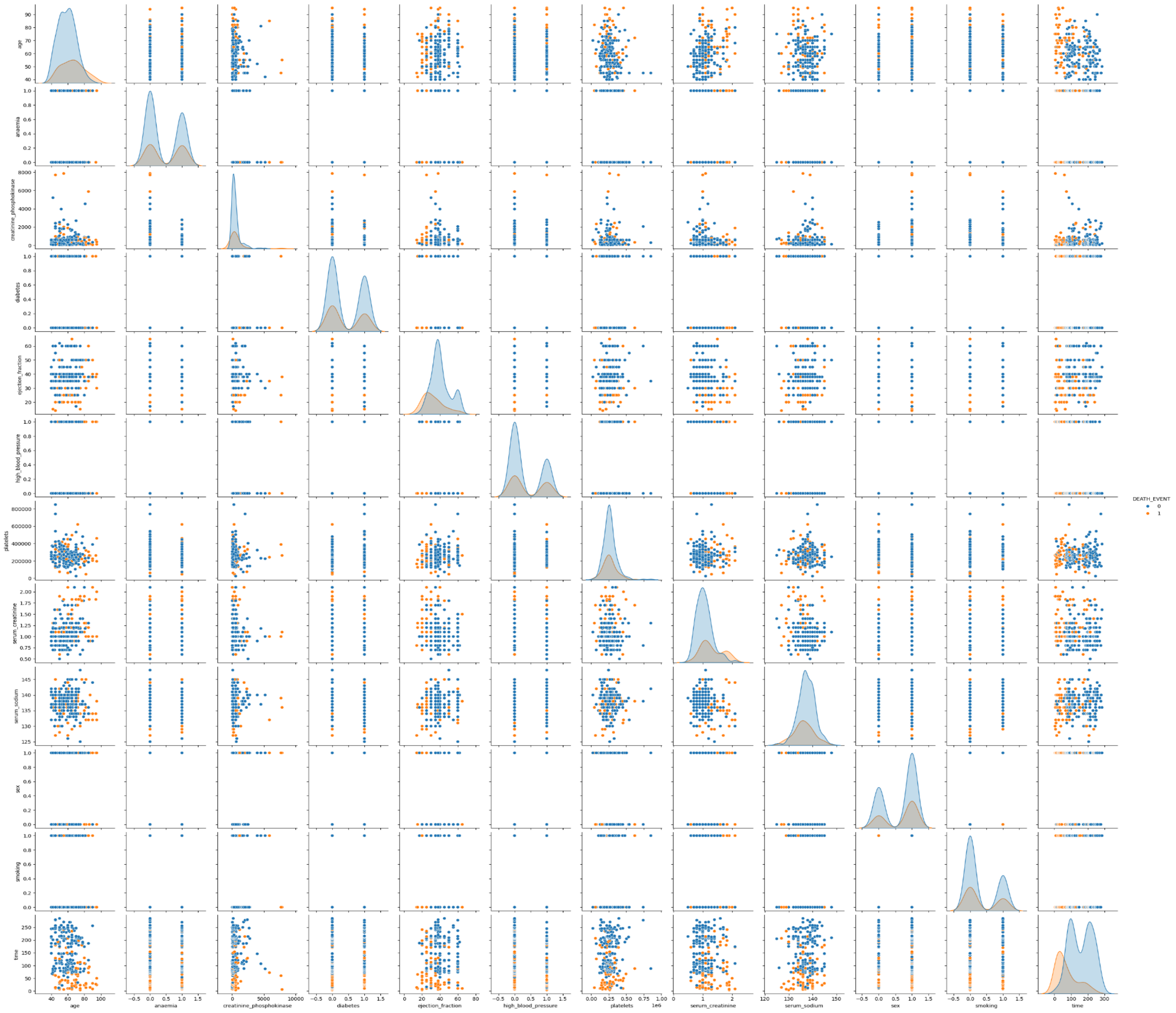
Introduction

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset and uncovering valuable insights that guide the subsequent modeling process. This phase involves summarizing the main characteristics of the data using visual and statistical methods.

Visualizations

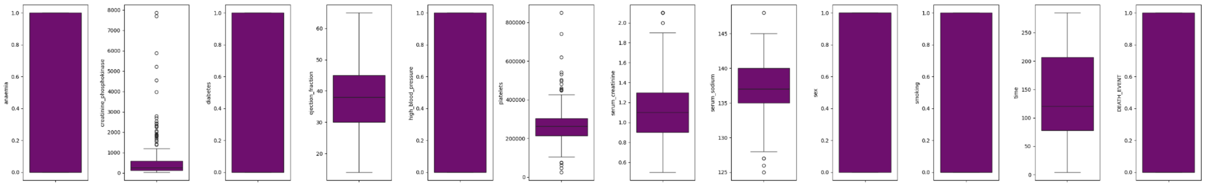
Visual representations of the data were key in identifying patterns, trends, and anomalies:

Pair Plot: The pair plot below shows relationships between multiple variables, highlighting distributions and interactions.

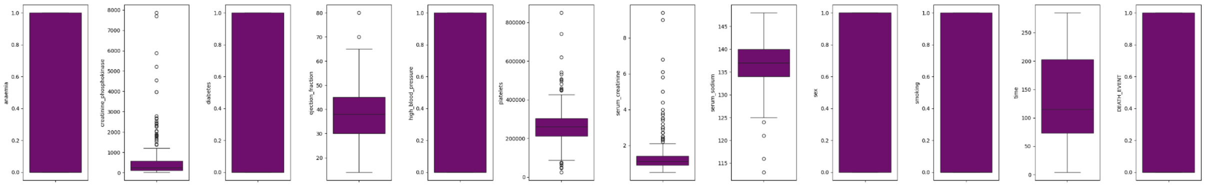


Box Plots: The box plots highlight the spread and potential outliers for key health characteristics (with outliers and without).

1. <u>Raw Dataset</u> (with outliers):



2. <u>Dataset after removing outliers</u> (we removed outliers specifically in the columns ‘ejection_fraction’, ‘serum_sodium’, and ‘serum_creatinine’ because the values in these columns can be significantly far from the healthy range, which can distort the analysis and model performance):



Key Health Characteristics and Their Importance

Creatinine Phosphokinase (CPK):
Healthy Range: 20-200 U/L.
Importance: Elevated CPK levels can indicate muscle damage, including heart muscle, which is relevant for heart failure diagnosis. Our data showed some patients with CPK values significantly higher than this range, indicating potential heart stress or damage.

Ejection Fraction:
Healthy Range: 55-70%.
Importance: This measures the percentage of blood leaving the heart each time it contracts. Lower values suggest poor heart function, commonly seen in heart failure patients. Our dataset included many patients with ejection fractions well below the healthy range, confirming its relevance as a predictor.

Platelets:
Healthy Range: 150,000-450,000 per µL.
Importance: Platelet counts are crucial for blood clotting. Abnormal levels can be associated with various health issues, including cardiovascular diseases. In our dataset, extreme platelet counts were observed, which could indicate underlying health problems affecting heart function.

Serum Creatinine:
Healthy Range: 0.6-1.2 mg/dL for men, 0.5-1.1 mg/dL for women.
Importance: This is a marker of kidney function. Elevated levels can indicate kidney impairment, which is often linked to heart disease. Many patients in our dataset had serum creatinine levels above the healthy range, highlighting the interplay between kidney and heart health.

Serum Sodium:
Healthy Range: 135-145 mmol/L.
Importance: Proper sodium levels are essential for normal body function. Imbalances can lead to serious health issues, including heart failure. Our analysis showed that patients with heart failure often had serum sodium levels outside this range, underscoring its importance.
Handling Outliers
We chose to handle outliers specifically in features where extreme values could disproportionately influence the model:

‘ejection_fraction’, ‘serum_sodium’, and ‘serum_creatinine’: Significant outliers in these features were capped or removed to prevent skewing the analysis, ensuring more robust model performance.
Other Features: On the other hand, we kept outliers in some columns such as ‘creatinine_phosphokinase’ and Platelets because patients can survive across the entire range of its values. As an example Creatinine phosphokinase levels can vary widely due to various factors such as muscle injury or intense physical activity, which do not necessarily indicate a death event.

Conclusions

EDA provided critical insights into the data, guiding our feature engineering and model building processes. By understanding the normal ranges and importance of key health characteristics, we ensured that our models were based on a solid foundation of well-understood data. This included the strategic removal of outliers in ‘ejection_fraction’, ‘serum_sodium’, and ‘serum_creatinine’ to avoid distortions in the analysis and model performance, while retaining outliers in ‘creatinine_phosphokinase’ and others to capture the true variability in patient conditions. These steps led to more accurate and reliable predictions, ultimately enhancing the robustness and generalizability of our models.

Chapter 3: Model Building and Evaluation

Data Preparation for modeling

Before training the models, the data was prepared as follows:
Train-Test Split: The dataset was divided into a training set (80%) and a test set (20%) to evaluate the models' performance on unseen data.
Handling Class Imbalance: We applied the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, ensuring the model does not become biased towards the majority class.

Model Selection

We evaluated several machine learning models:

1. Logistic Regression: A simple yet effective model for binary classification tasks.
2. Random Forest: An ensemble learning method that constructs multiple decision trees for improved accuracy and robustness.
3. Gradient Boosting: Another ensemble technique that builds models sequentially to correct the errors of previous models.
4. Artificial Neural Network (ANN): A deep learning model capable of capturing complex patterns in the data.

Evaluation Metrics

Models were assessed using the following metrics:

1. Accuracy: The proportion of correctly classified instances among the total instances.
2. Precision: The proportion of true positive predictions among all positive predictions.
3. Recall: The proportion of true positive predictions among all actual positives.
4. F1-Score: The harmonic mean of precision and recall, providing a balance between the two.
5. ROC-AUC: The area under the Receiver Operating Characteristic curve, indicating the model's ability to discriminate between classes.

Model Performance

The performance results for the evaluated models:

Accuracy Scores for Models:

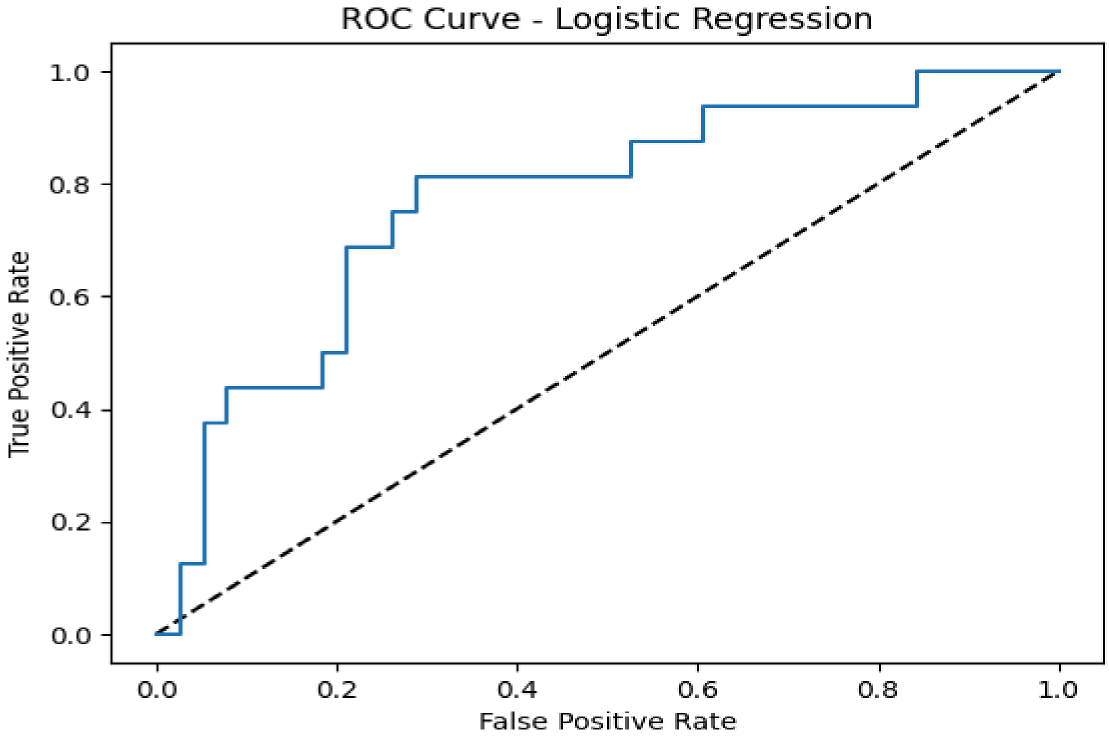
- Logistic Regression: 78.9%
- ANN: 71.1%
- Random Forest: 80.3%
- Gradient Boosting: 78.9%

AUC Scores for Models:

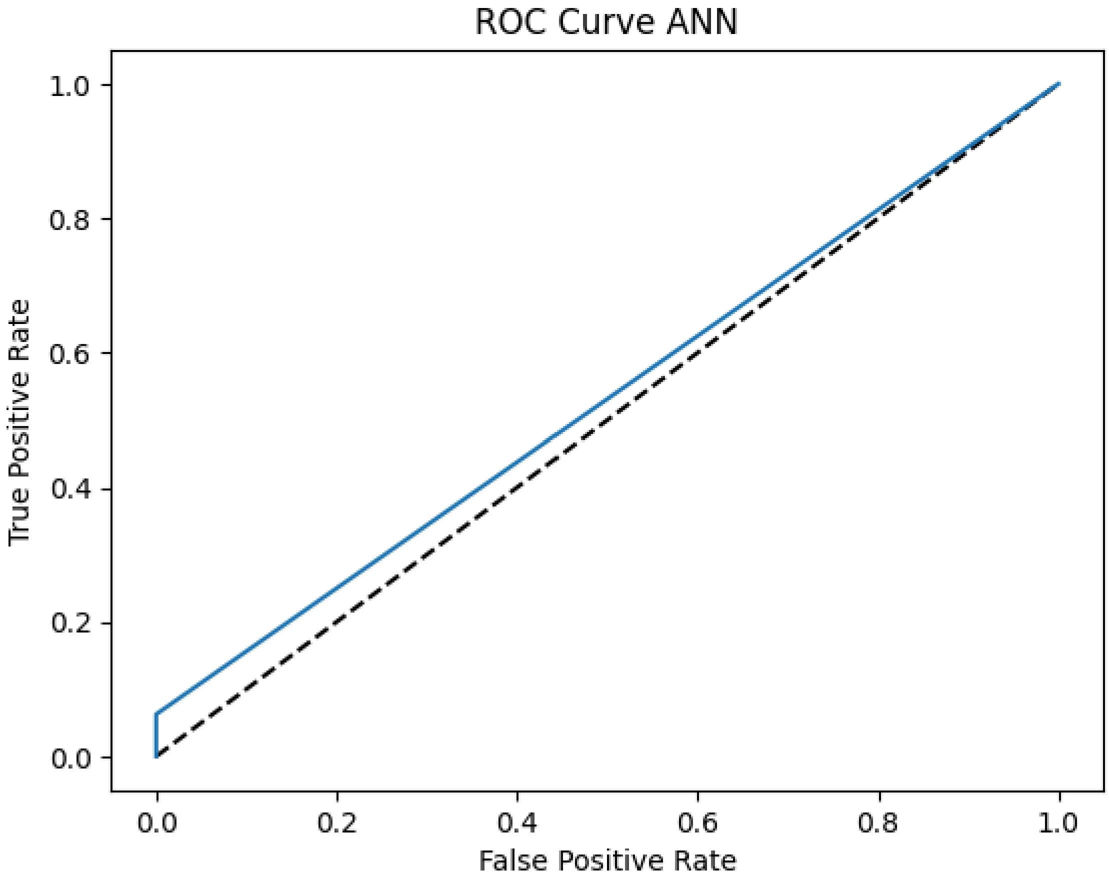
- Logistic Regression: 0.770
- ANN: 0.567
- Random Forest: 0.806
- Gradient Boosting: 0.789

ROC Curves:

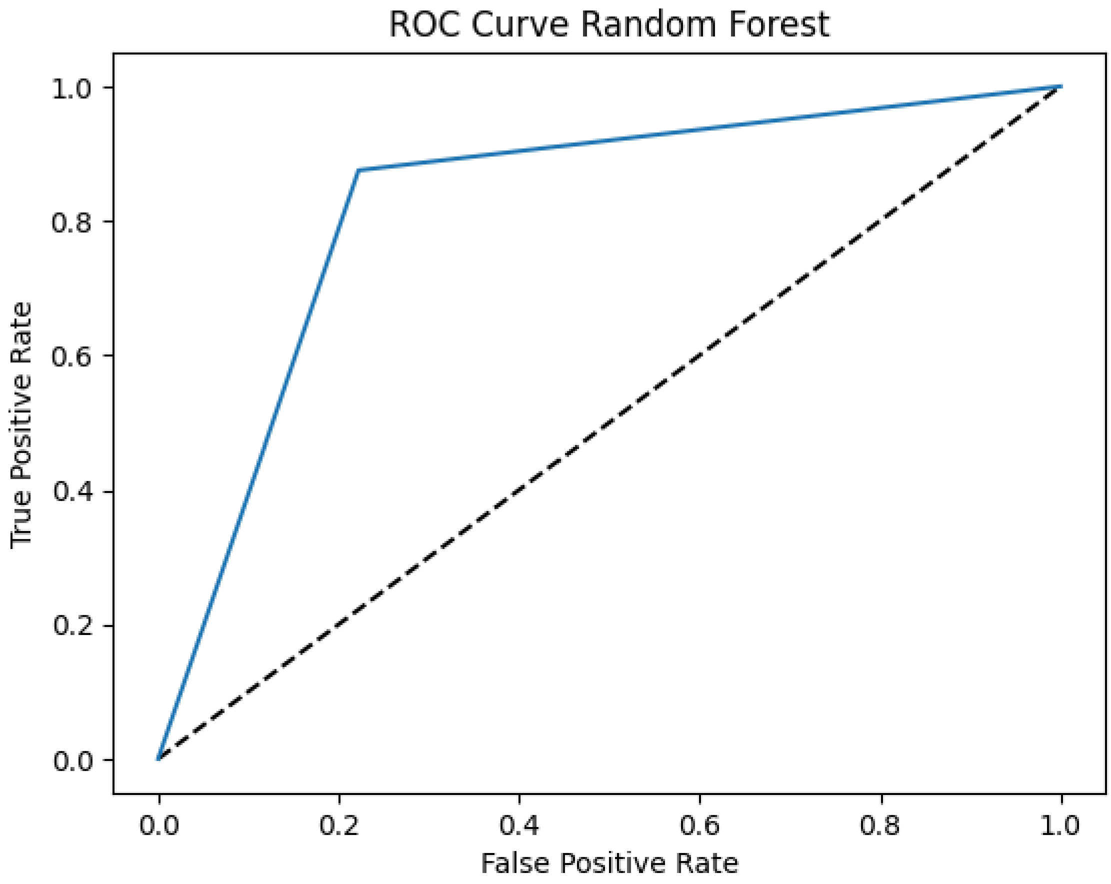
Logistic Regression:



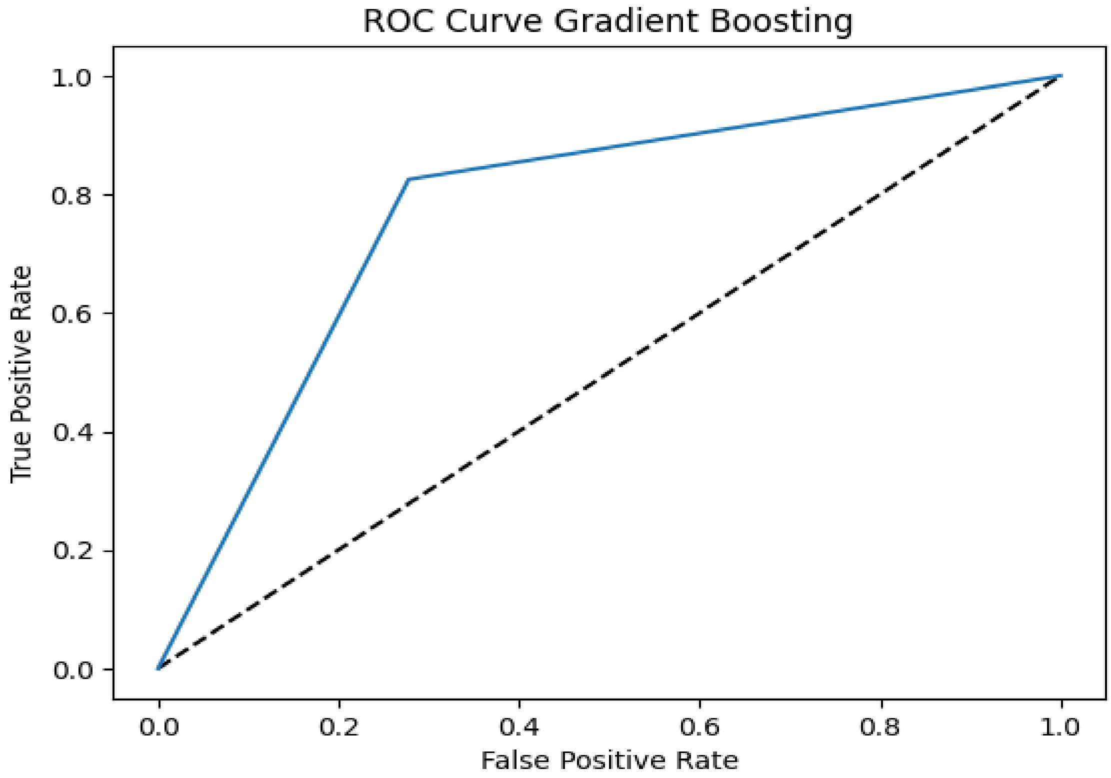
ANN:



Random Forest:



Gradient Boosting:



Confusion Matrices and Classification Reports

Logistic Regression:
Accuracy: 78.9%
Confusion Matrix: [[30, 6], [10, 30]]
Classification Report:
Precision: 0.75 (survival), 0.83 (death)
Recall: 0.83 (survival), 0.75 (death)
F1-Sc: 0.79 (survival), 0.79 (death)
Support: 36 (survival), 40 (death)

Artificial Neural Network (ANN):
Accuracy: 71.1%
Confusion Matrix: [[29, 7], [15, 25]]
Classification Report:
Precision: 0.66 (survival), 0.78 (death)
Recall: 0.81 (survival), 0.62 (death)
F1-Sc: 0.72 (survival), 0.69 (death)
Support: 36 (survival), 40 (death)

Random Forest Classifier (RFC):
Accuracy: 80.3%
Confusion Matrix: [[31, 5], [10, 30]]
Classification Report:
Precision: 0.76 (survival), 0.86 (death)
Recall: 0.86 (survival), 0.75 (death)
F1-Sc: 0.81 (survival), 0.80 (death)
Support: 36 (survival), 40 (death)

Gradient Boosting Machine (GBM):
Accuracy: 78.9%
Confusion Matrix: [[28, 8], [8, 32]]
Classification Report:
Precision: 0.78 (survival), 0.80 (death)
Recall: 0.78 (survival), 0.80 (death)
F1-Sc: 0.78 (survival), 0.80 (death)
Support: 36 (survival), 40 (death)

Conclusions from Model Evaluation (with SMOTE)

Logistic Regression

The Logistic Regression model has an accuracy of 78.9% and an AUC score of 0.770. The confusion matrix shows that the model correctly identifies 30 out of 36 cases for the survival class (precision: 75%) and 30 out of 40 cases for the death class (recall: 75%). This indicates that the model is moderately effective at predicting patient outcomes, particularly considering its simplicity compared to more complex models.

Artificial Neural Network (ANN)

The ANN model, has an accuracy of 71.1%. It has a precision of 66% for predicting survival and 78% for predicting death, with recall scores of 81% for survival and 62% for death. This results in F1-scores of 72% for survival and 69% for death, showing that the model treats both outcomes with similar importance. The AUC score for the ANN model is 0.567, indicating moderate discriminatory ability.

Random Forest Classifier (RFC)

The Random Forest Classifier shows strong performance with an accuracy of 80.3% and an AUC score of 0.806. It has a precision of 76% for survival and 86% for death, with recall scores of 86% for survival and 75% for death. The F1-scores are 81% for survival and 80% for death, demonstrating the model's consistent reliability in predicting both outcomes.

Gradient Boosting Machine (GBM)

The GBM model also performs well, with an overall accuracy of 78.9% and an AUC score of 0.789. It achieves a precision of 78% for survival and 80% for death, while recall scores are 78% for survival and 80% for death. The F1-scores are 78% for survival and 80% for death, indicating that the model is doing a good job across the board.

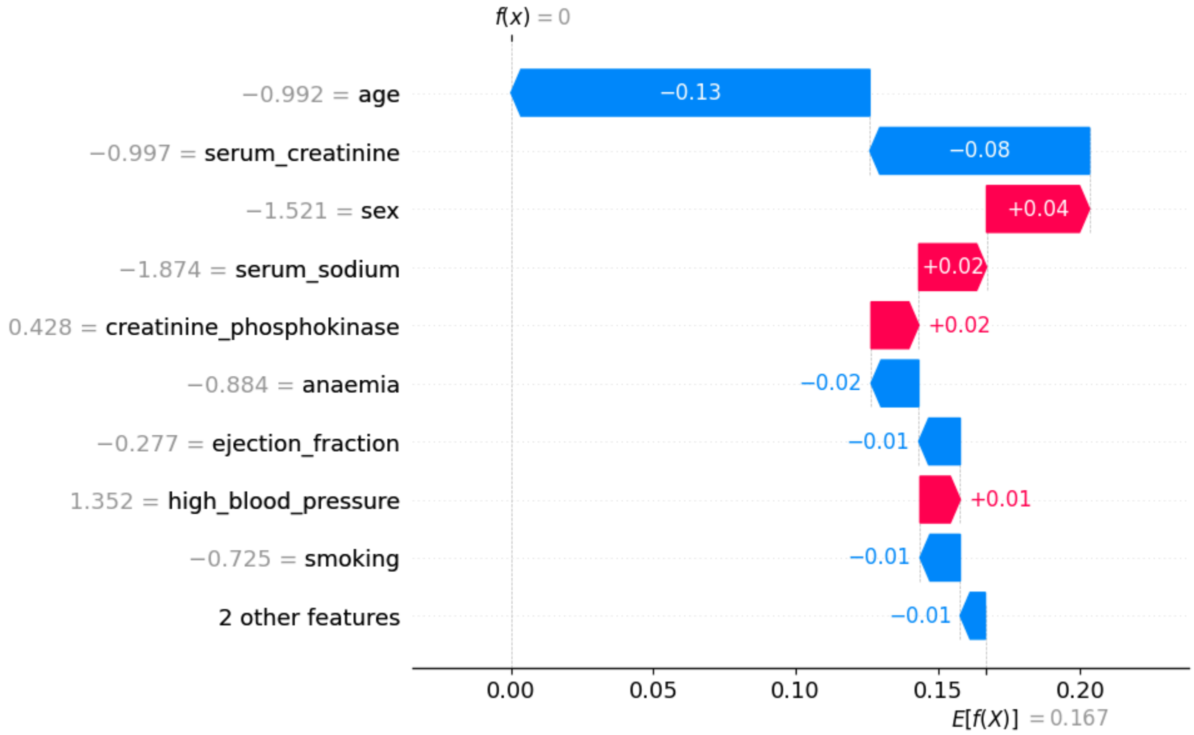
Chapter 4: Feature Importance and Interpretation

SHAP Analysis and Conclusions (on Logistic Regression and Gradient Boosting Models)

SHapley Additive exPlanations (SHAP) values provide a clear measure of feature importance, ensuring transparency and trust in model (Logistic regression model in this case next two SHAP plots) predictions. By analyzing SHAP values, we can understand which health characteristics most significantly impact heart failure predictions.

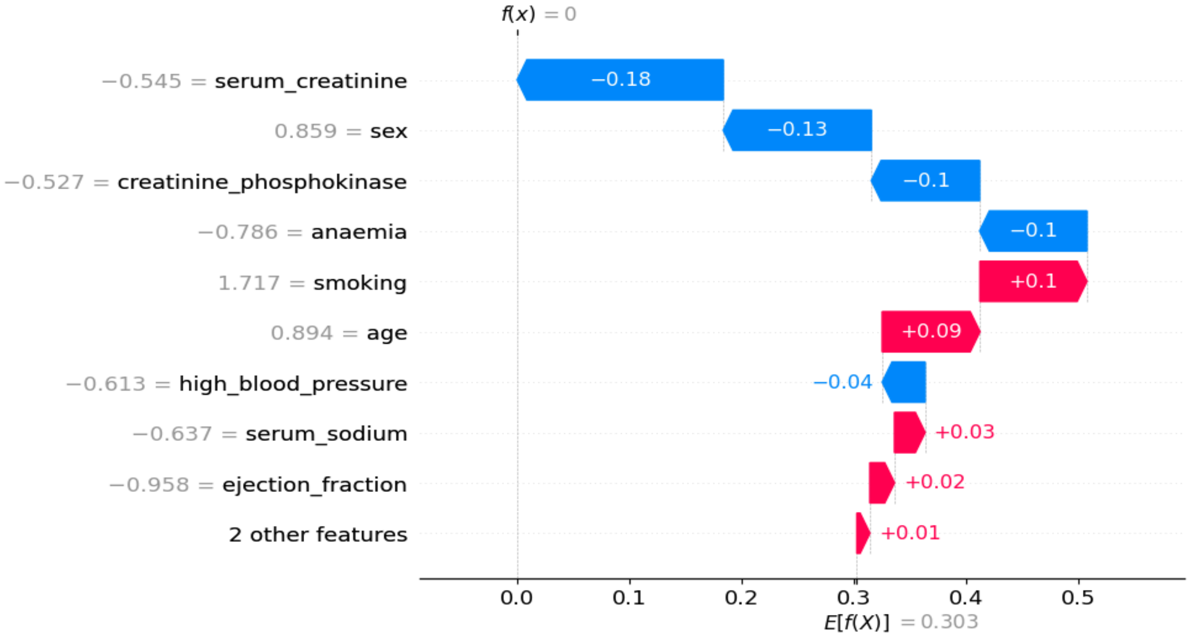
SHAP Summary and Dependence Plots

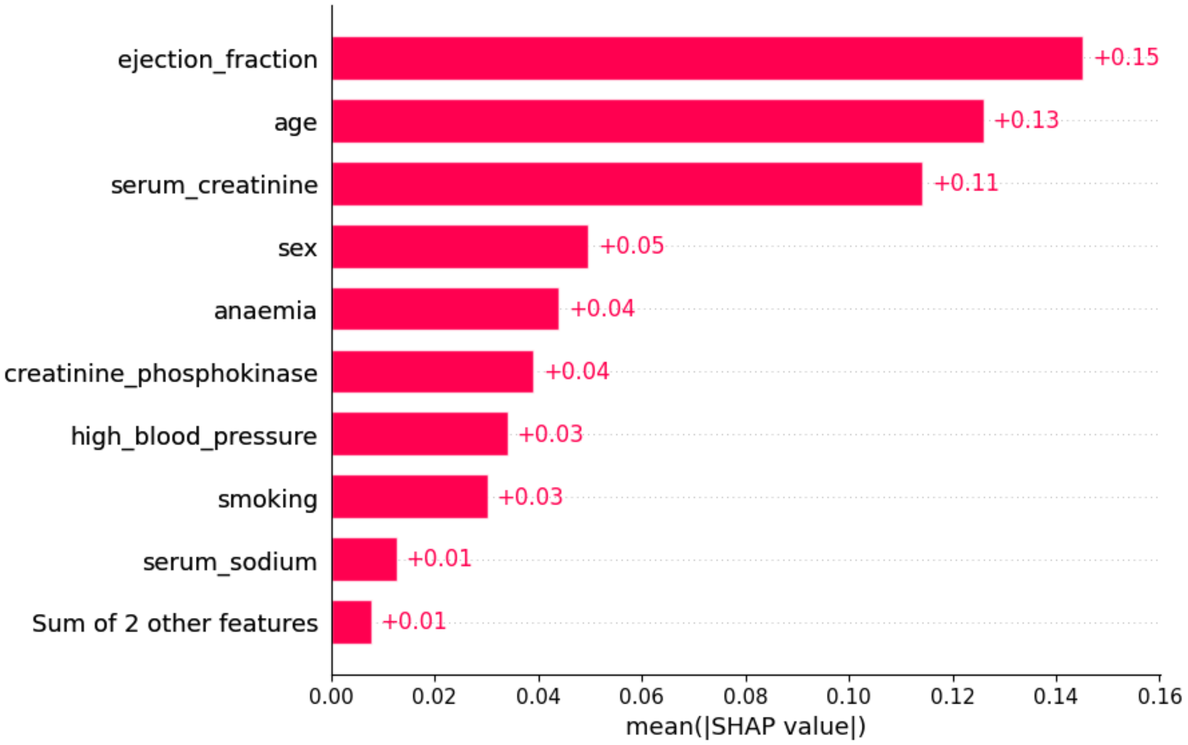
Without SMOTE for Logistic Regression Model:

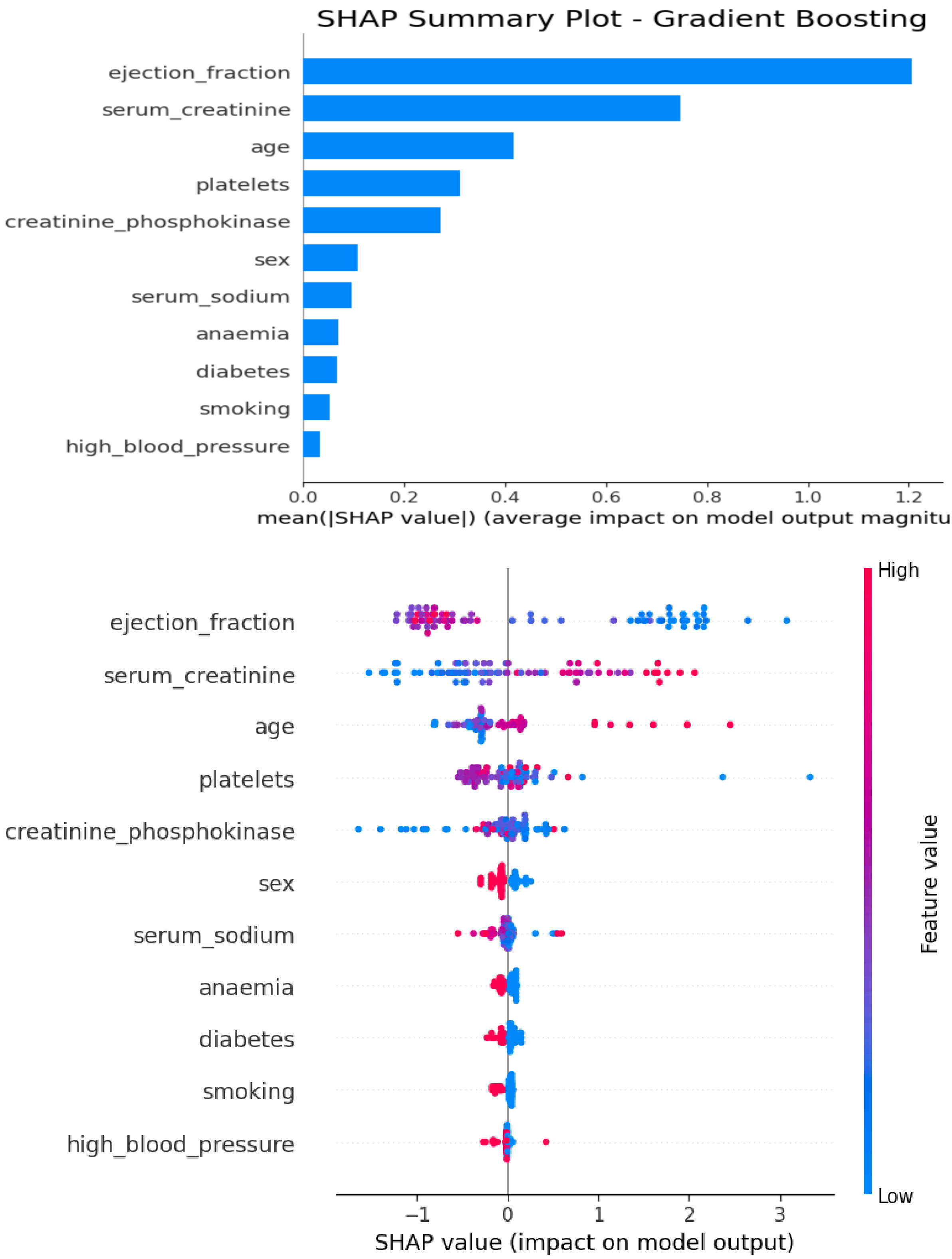


The initial SHAP analysis without SMOTE showed age and anaemia as significant predictors, but the imbalanced data limited the model's ability to accurately assess minority class features.

After applying SMOTE for Logistic Regression:






SHAP Summary Plot - Gradient Boosting
ejection_fraction
serum_creatinine
age
platelets
creatinine_phosphokinase
sex
serum_sodium
anaemia
diabetes
smoking
high_blood_pressure
0.0
0.2
0.4
0.6
0.8
1.0
1.2
mean(|SHAP value|) (average impact on model output magnitu
ejection_fraction
serum_creatinine
age
platelets
creatinine_phosphokinase
sex
serum_sodium
anaemia
diabetes
smoking
high_blood_pressure
−1
0
1
2
3
SHAP value (impact on model output)
High
Feature value
Low

Key Findings from SHAP Analysis

Impact of SMOTE:

- Enhanced Model Performance: The application of SMOTE improved the logistic regression model's ability to generalize by addressing the class imbalance. This resulted in a more comprehensive evaluation of feature importance, ensuring that minority class features were adequately represented.
- Improved Interpretability: We have more evenly distributed SHAP values across various features, providing better interpretability of the model's decisions.
- Key Predictors: Even after adjustments, serum creatinine and ejection fraction stayed crucial for assessing heart health. After SMOTE, age and other health factors became more important, showing that many different factors are key to predicting heart failure.

Summary:

SHAP analysis before and after using SMOTE shows that SMOTE helps improve model performance and makes predictions clearer. It ensures all features are fairly represented, leading to better and more trustworthy heart failure predictions.

Chapter 5: Conclusions and Findings

The evaluation of ROC curves indicates that the Random Forest and Gradient Boosting models are the most effective for predicting heart failure events, with AUC scores of 0.806 and 0.789, respectively. These models demonstrate excellent discriminative ability, achieving high true positive rates with minimal false positives. Logistic Regression also performs well with an AUC of 0.770, offering a good balance between sensitivity and specificity, making it a reliable choice for clinical decision support.

Final Thoughts

When we put it all together, the Random Forest and Gradient Boosting models are the stars of the show for this dataset, with the highest AUC scores and strong accuracy. They stand out as the most dependable for predicting whether patients will survive or not. While the ANN does not outperform the ensemble models, it still provides valuable insights into complex data patterns. Logistic Regression shows good performance overall, demonstrating its effectiveness as a simpler model in heart failure prediction.

The combination of traditional machine learning models and advanced techniques like SMOTE and SHAP allowed us to build effective and reliable models for heart failure prediction. These findings highlight the potential of data science in healthcare, providing tools for early diagnosis and treatment planning. This project demonstrates the significant contributions data science can make to healthcare by enhancing early diagnosis and treatment planning.