Heart Failure Prediction Using Machine Learning

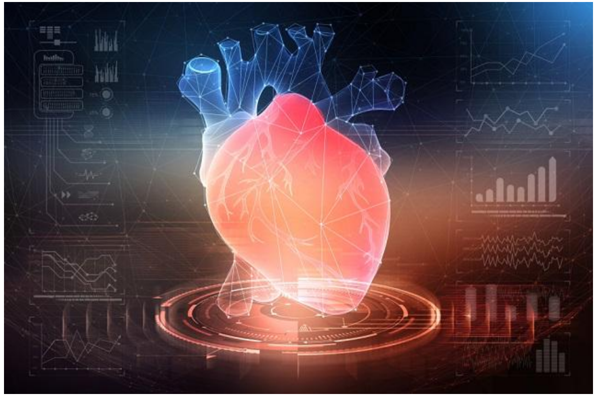
by Denis Grekov

Capstone 3 project of the Data Science course by Springboard
05/2024

2: Project Overview

Heart failure is a severe condition where the heart's ability to pump blood diminishes, leading to significant health risks and mortality. This project aims to develop machine learning models to predict heart failure events using patient health records.

The goal is to identify key predictors and build accurate models to assist healthcare providers in early diagnosis and intervention.



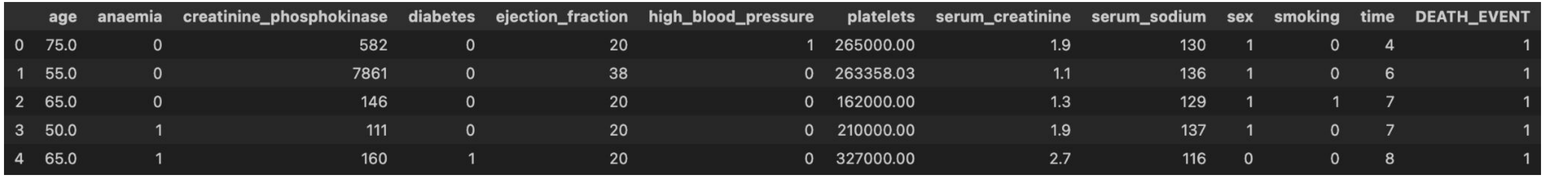
3: Dataset Description

Our dataset includes patient health records with features like:

Age, Anaemia, Creatinine Phosphokinase (CPK), Diabetes,

Ejection Fraction, High Blood Pressure, Platelets, Serum Creatinine,

Serum Sodium, Sex, Smoking, and Death Event.



	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

4: Project Goals and Objectives

Project Goals and Objectives:

1. Data Cleaning and Preparation
2. Exploratory Data Analysis (EDA)
3. Feature Engineering
4. Model Building and Evaluation
5. Interpretation and Explanation using SHAP
6. Conclusions

5: Data Cleaning

1. We removed missing values using conventional Python methods such as dropna()
2. Identified and handled outliers using the Interquartile Range (IQR) method
3. Normalized continuous variables using min-max scaling

6: Exploratory Data Analysis (EDA)

- Summary statistics for all features
- Visualizations including histograms, box plots, and scatter plots
- Identified patterns, trends, and potential anomalies

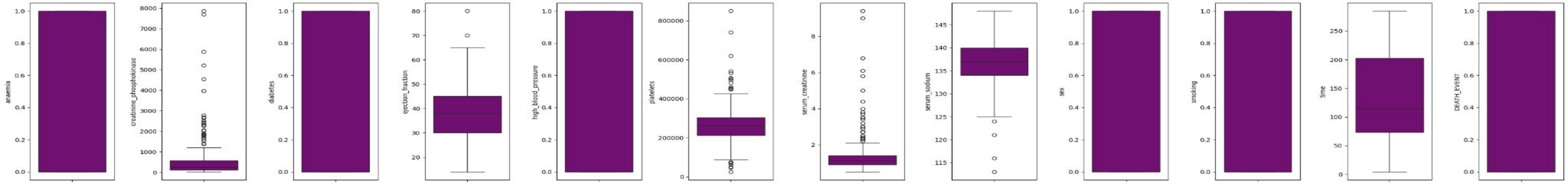
7: Handling Outliers

Outliers in 'ejection_fraction', 'serum_sodium', and 'serum_creatinine' were removed as they can significantly skew the analysis.

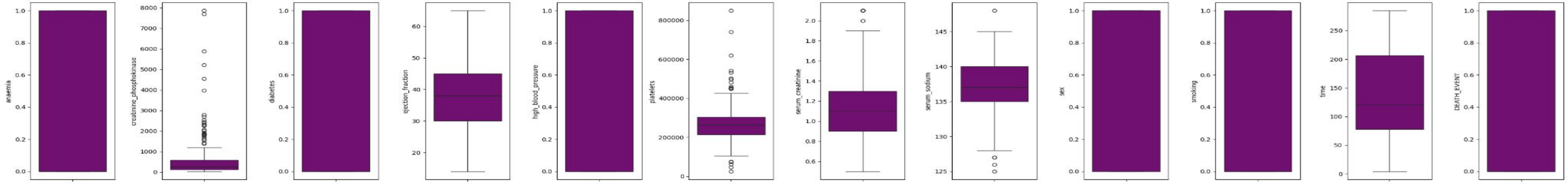
At the same time outliers in 'creatinine_phosphokinase' were retained because patients can survive across the entire range of its values.

7.c. Outlier removal charts

dataset with outliers:

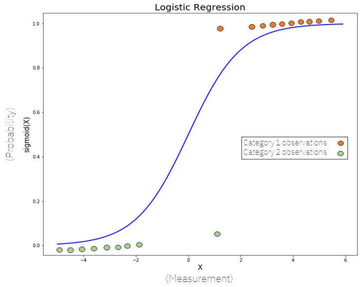


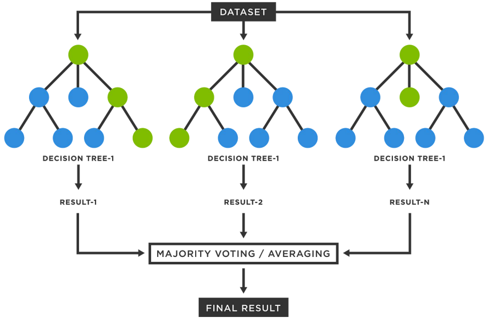
dataset without outliers:

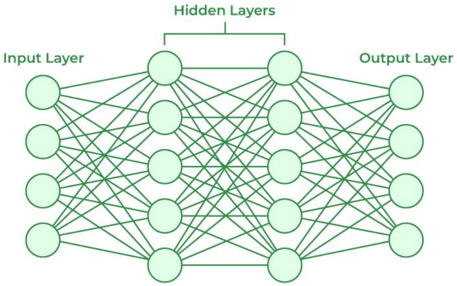


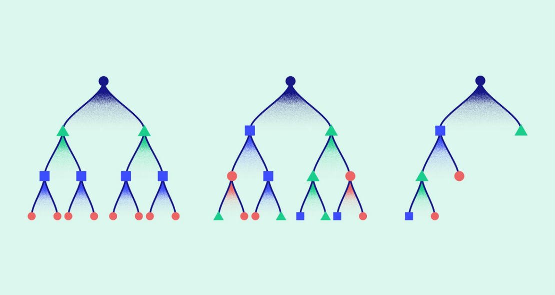
8: Model Selection

1. Logistic Regression
2. Random Forest
3. Gradient Boosting
4. Artificial Neural Network (ANN)



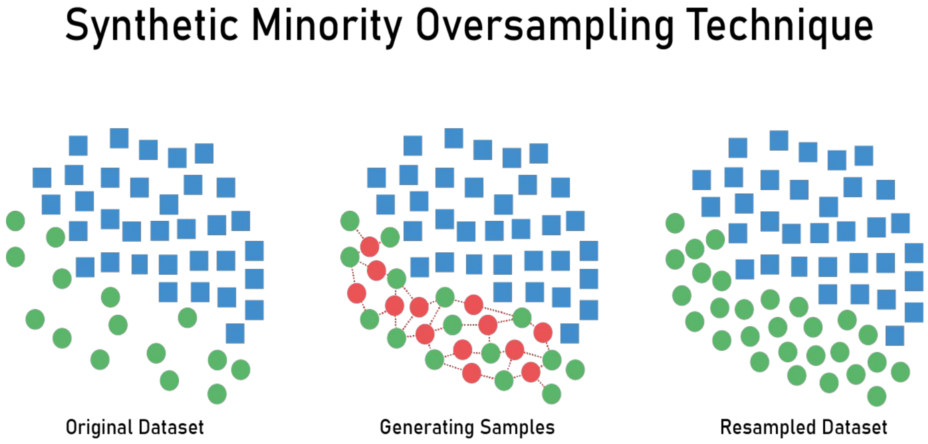






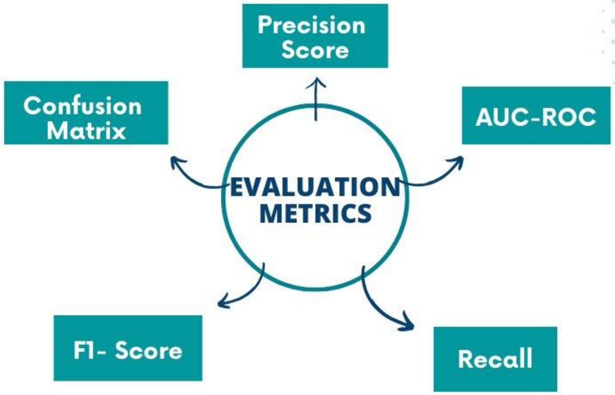
9: Handling Class Imbalance with SMOTE

To address class imbalance in the dataset, we applied the Synthetic Minority Over-sampling Technique (SMOTE). This technique generates synthetic samples for the minority class, ensuring that the model does not become biased towards the majority class. This step was very important for improving model performance and ensuring balanced representation of features.



10: Evaluation Metrics

1. Accuracy
2. Precision
3. Recall
4. F1-Score
5. ROC-AUC



EVALUATION METRICS

Precision Score

Confusion Matrix

AUC-ROC

F1- Score

Recall

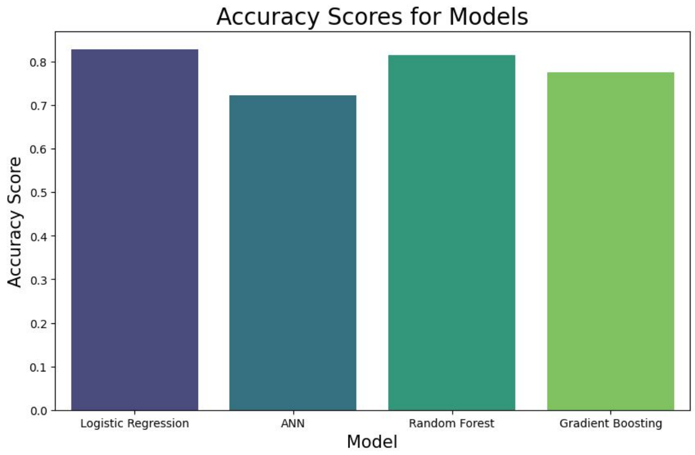
11: Model Performance

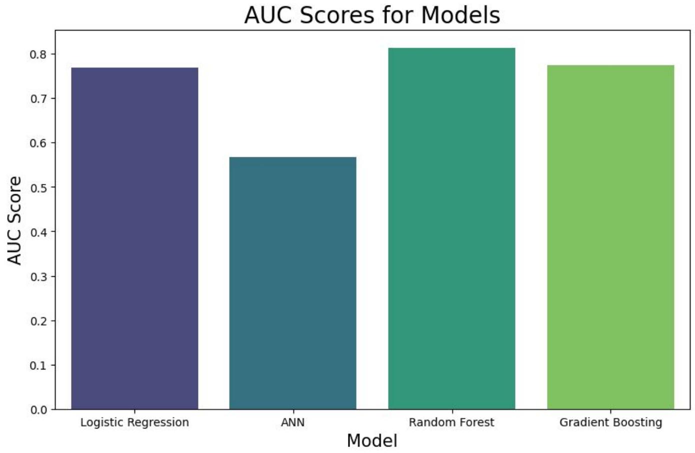
Accuracy Scores:

1. Logistic Regression: 78.9%
2. ANN: 71.1%
3. Random Forest: 80.3%
4. Gradient Boosting: 78.9%

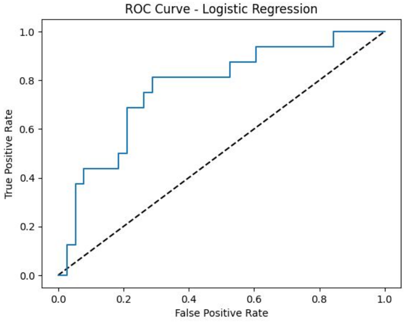
AUC Scores:

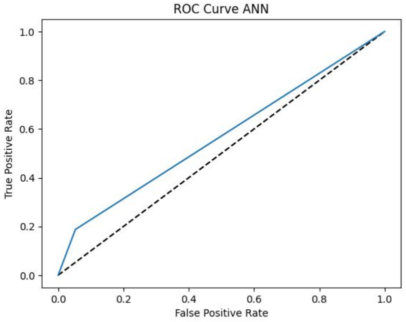
1. Logistic Regression: 0.770
2. ANN: 0.567
3. Random Forest: 0.806
4. Gradient Boosting: 0.789

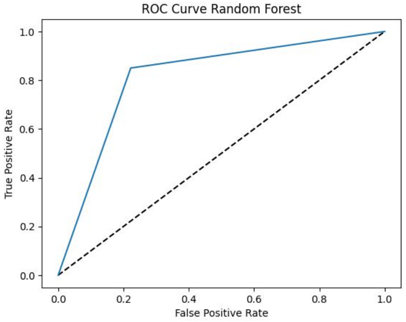

Accuracy Scores for Models

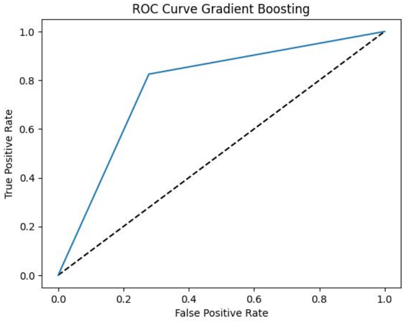

AUC Scores for Models

12. AUC for all four models


ROC Curve - Logistic Regression


ROC Curve ANN

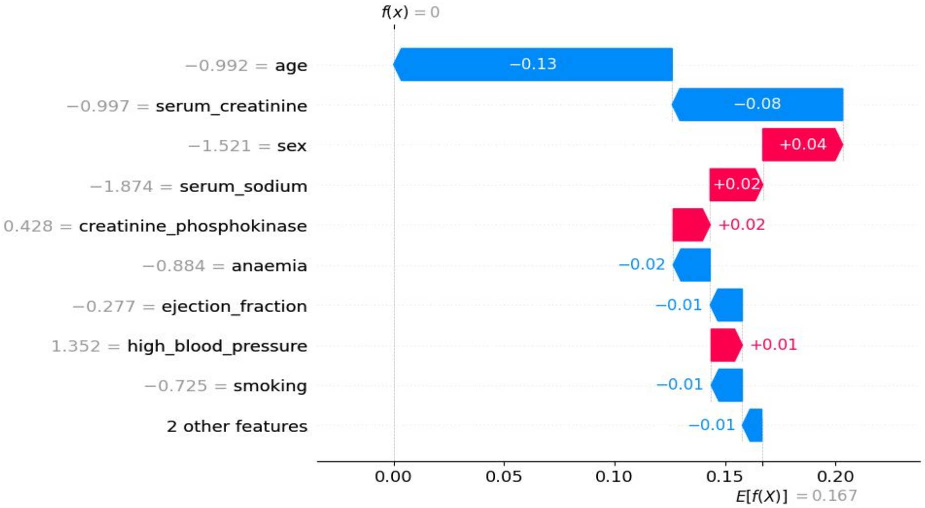

ROC Curve Random Forest


ROC Curve Gradient Boosting

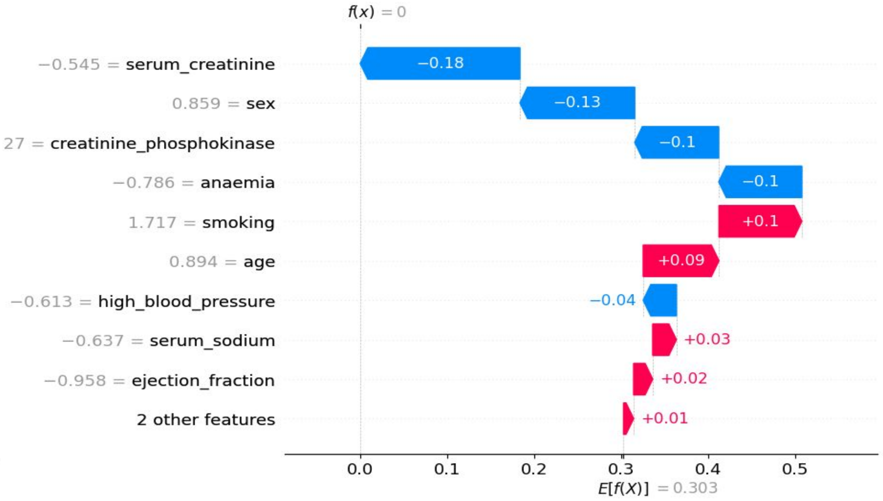
13. SHAP Analysis (Without and after SMOTE)

- Age and anaemia were significant predictors, but imbalanced data limited the model's ability to assess minority class features accurately.
- After applying SMOTE, the model performance and interpretability improved by addressing class imbalance.
- Serum creatinine and ejection fraction remained critical predictors.
- Age and other health indicators became more important post-SMOTE.

13.c. Charts SHAP (Without and after SMOTE)

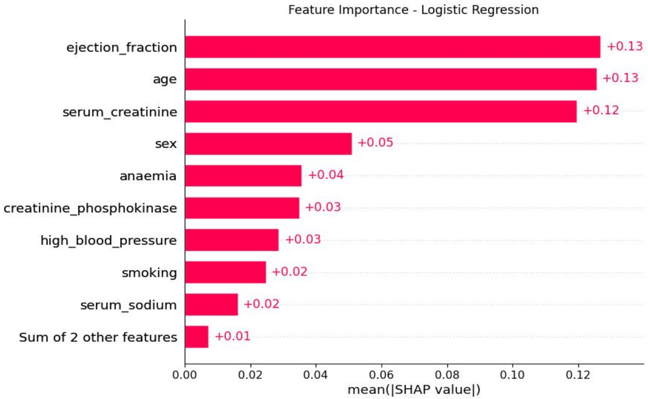


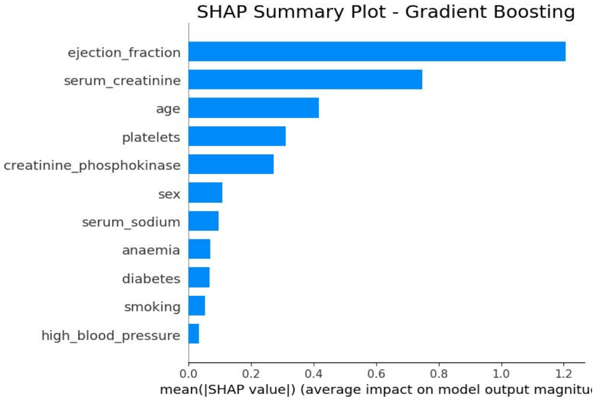
without SMOTE applied

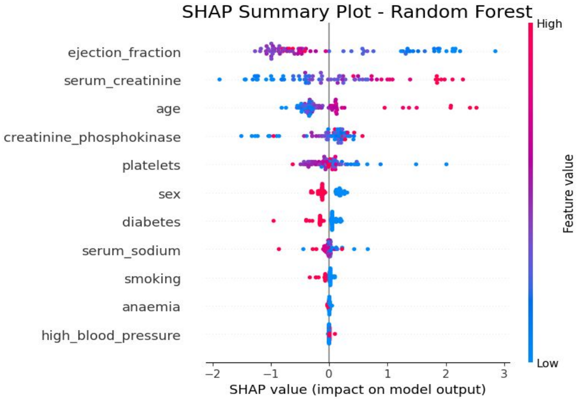


after SMOTE applied

13.2c. SHAP plots after applying SMOTE


Feature Importance - Logistic Regression


SHAP Summary Plot - Gradient Boosting


SHAP Summary Plot - Random Forest

The SHAP analysis for Gradient Boosting and Random Forest highlights the importance of ejection fraction, serum creatinine, and age in predicting heart failure. Logistic Regression also identifies these factors, along with sex and anaemia. The consistent identification of these important features across different models demonstrates the robustness of the findings.

These SHAP plots are essential for transparency and making informed clinical decisions for patients at risk of heart failure.

14. Conclusions

1. Random Forest and Gradient Boosting models were the most effective, with AUC scores of 0.806 and 0.789, respectively.
2. Logistic Regression also performed well with an AUC of 0.770.
3. ANN after SMOTE applied performed lower (AUC of 0.567) comparing to less complex models.
4. The application of SMOTE was crucial for improving model performance by addressing class imbalance.
5. SHAP analysis enhanced the interpretability of the models, highlighting important features like age, ejection fraction, and serum creatinine.

These findings underscore the potential of data science in healthcare, providing tools for early diagnosis and treatment planning.

Overall, the project demonstrates how advanced analytics can be applied to solve critical health issues, improving patient outcomes.

Thank You !