Heart Failure Prediction

Presentation Structure

01 - Introduction

02 - Data Preparation

03 - Analysis

03 - Evaluate the Model

04 - Conclusion

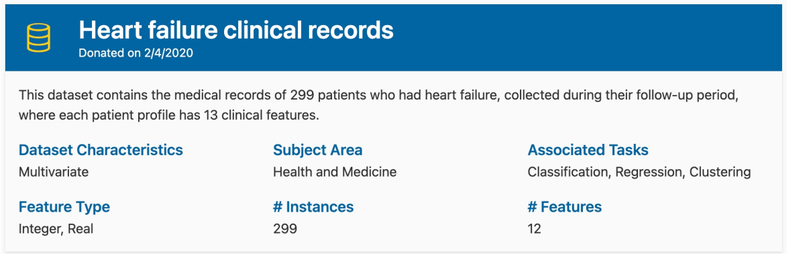
Introduction

Project Brief

Cardiovascular diseases (CVDs) are the **number 1 cause of death globally**, taking an estimated **17.9 million lives each year**, which accounts for **31% of all deaths worlwide**.

Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure.

The goal of this project is to demonstrate the use of **Machine Learning classification algorithms** to predict the **survival of patients experiencing heart failure symptoms** and to identify the most **important clinical features (or risk factors) associated with heart failure**. In this context, "**1**" indicates a **death event** resulting from heart failure, while "**0**" indicates **no heart failure**.



source: https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

Dataset Overview

Variables Table

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
age	Feature	Integer	Age	age of the patient	years	no
anaemia	Feature	Binary		decrease of red blood cells or hemoglobin		no
creatinine_phosphokinase	Feature	Integer		level of the CPK enzyme in the blood	mcg/L	no
diabetes	Feature	Binary		if the patient has diabetes		no
ejection_fraction	Feature	Integer		percentage of blood leaving the heart at each contraction	%	no
high_blood_pressure	Feature	Binary		if the patient has hypertension		no
platelets	Feature	Continuous		platelets in the blood	kiloplatelets/mL	no
serum_creatinine	Feature	Continuous		level of serum creatinine in the blood	mg/dL	no
serum_sodium	Feature	Integer		level of serum sodium in the blood	mEq/L	no
sex	Feature	Binary	Sex	woman or man		no

6 Categorical variables - Anaemia, Diabetes, High blood Pressure, Sex and Smoking (all binary type)

8 Numerical variables - Age, Creatinine Phosphokinase, Ejection Fraction, Platelets, Serum creatinine, Serum Sodium, Time and Death Event

0 Missing values

299 rows

Variable Name	Role	Type	Demographic	Description	Units	Missing Values
smoking	Feature	Binary		if the patient smokes or not		no
time	Feature	Integer		follow-up period	days	no
death_event	Target	Binary		if the patient died during the follow-up period		no

Data Preparation

Distribution of Data

Age: The age distribution is skewed to the right, with most individuals being 60 years old.

Anaemia: Most patients do not exhibit a decrease in red blood cells or hemoglobin.

Creatinine Phosphokinase, Serum Creatinine: These histograms are highly skewed to the right, indicating that most patients have a low level of the CPK enzyme and serum creatinine in the blood.

Diabetes: Most patients do not have diabetes.

Ejection Fraction: This histogram is slightly skewed to the right, suggesting a smaller percentage of blood leaving.

High Blood Pressure: Most patients do not have high blood pressure.

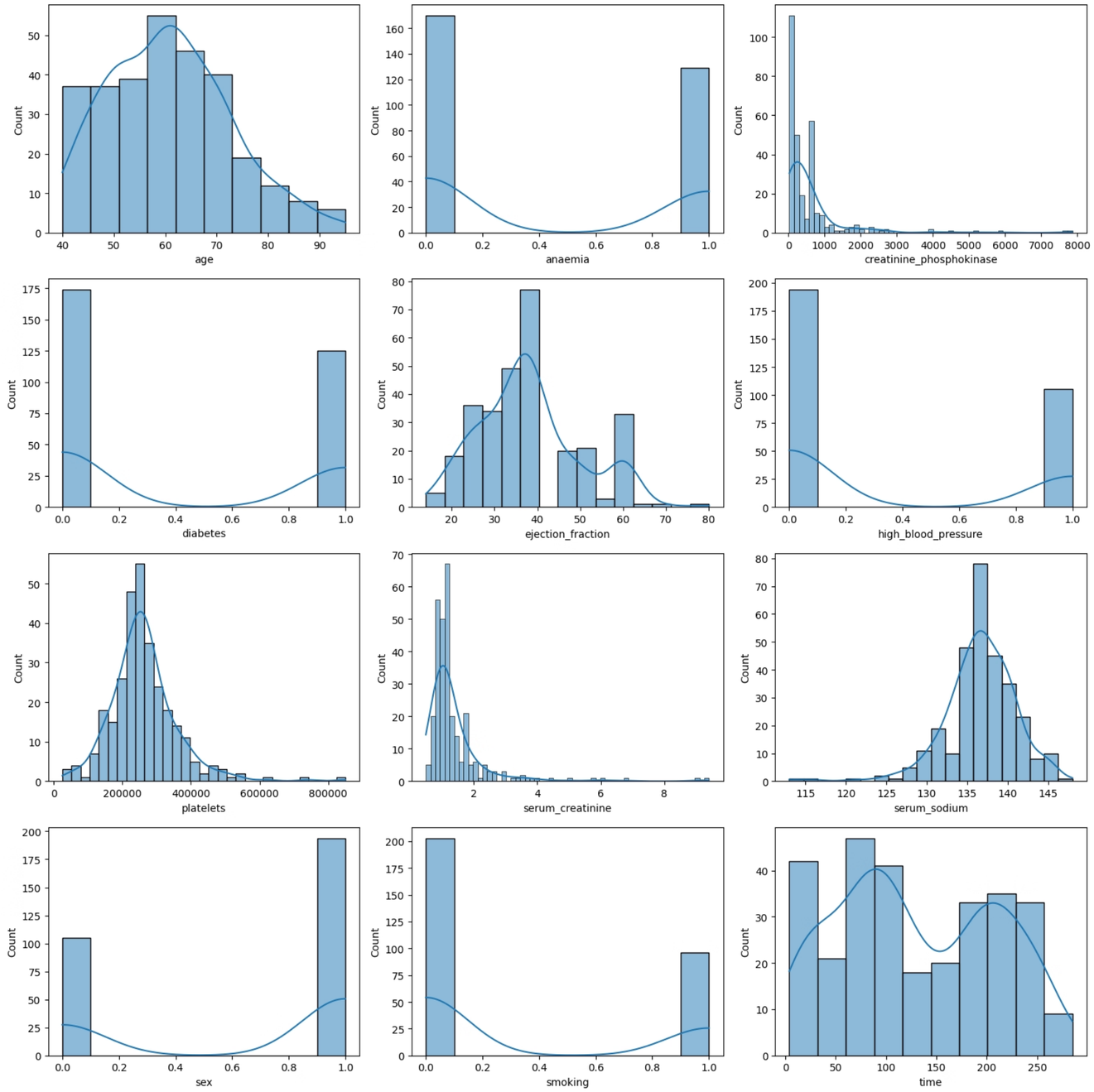
Platelets: The platelets distribution is skewed to the right, indicating that most patients have low platelet levels in the blood.

Serum Sodium: This is skewed to the left, suggesting that most patients have a higher level of sodium in the blood.

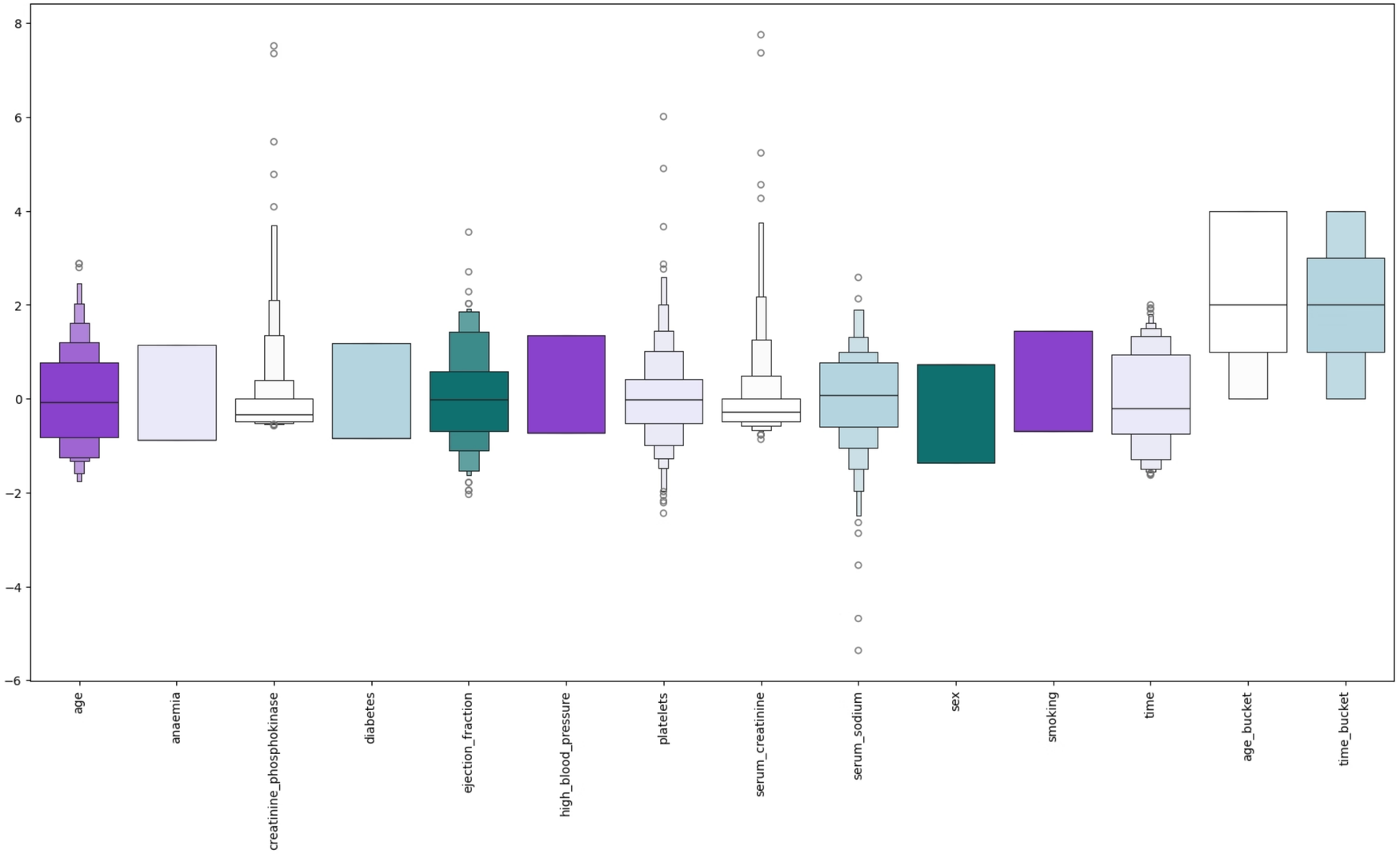
Sex: Most patients are male rather than female.

Smoking: Most patients are non-smokers.

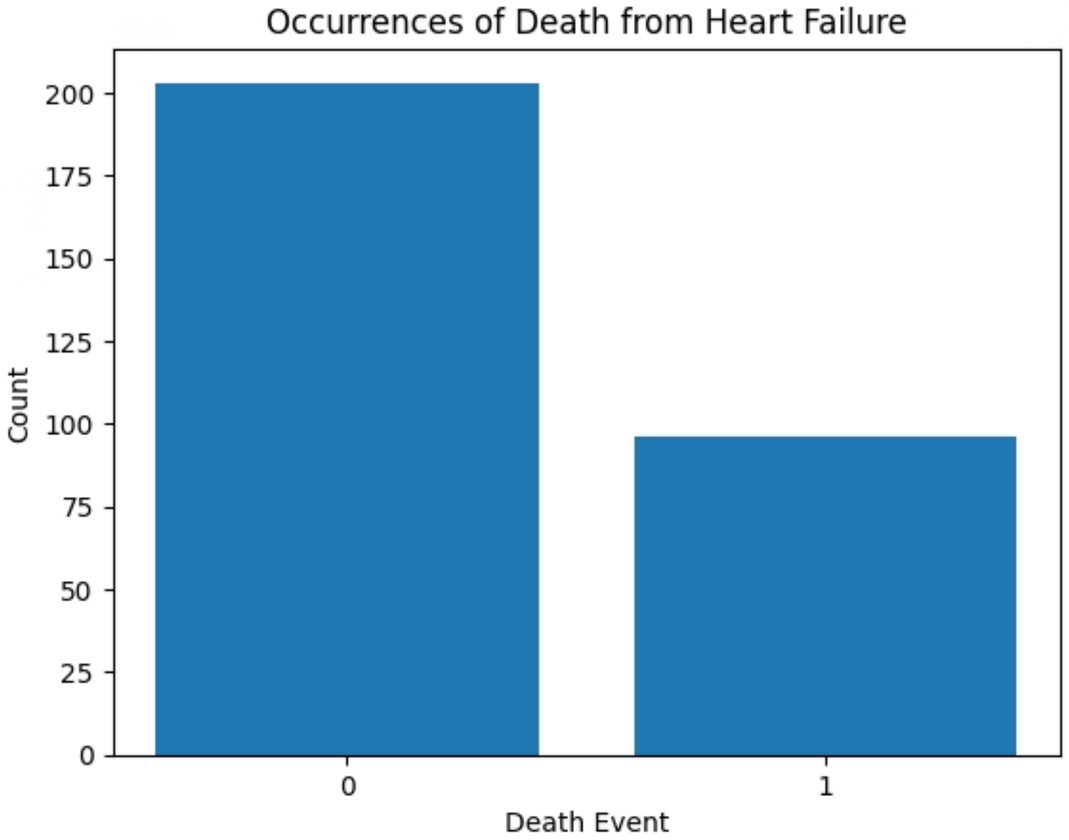
Time: The time histogram appears to have one or two peaks, suggesting that the majority of follow-up periods occur sooner rather than later.



Outliers



Target variable



Dataset imbalance exists, with 203 patients surviving (death event = 0), representing **67.89%**, and 96 patients deceased (death event = 1), accounting for **32.11%.**

Imbalance might introduce **bias in predictions**, particularly with the dominance of the majority class 0 over the minority class 1.

Balance data might be **crucial in heart failure prediction.**

Common rebalance techniques include resampling, SMOTE, and BalancedBaggingClassifier.

Overview

	Data preparation techniques	Features selection techniques	Sampling techniques	ML algorithms	Evaluation tools	Hyperparameters
Approach 1	Standardization Bucketing	randomForestClassifier and correlation matrix	**resampling**	ANN, randomForestClassifier	Accuracy, F1 score, Precision, Recall, Roc Auc Score	GridSearchCV
Approach 2	Standardization Bucketing	randomForestClassifier and correlation matrix	**SMOTE**	ANN, randomForestClassifier	Accuracy, F1 score, Precision, Recall, Roc Auc Score	GridSearchCV
Approach 3	Standardization Bucketing	randomForestClassifier and correlation matrix		ANN, randomForestClassifier	Accuracy, F1 score, Precision, Recall, Roc Auc Score	GridSearchCV

Data Analysis

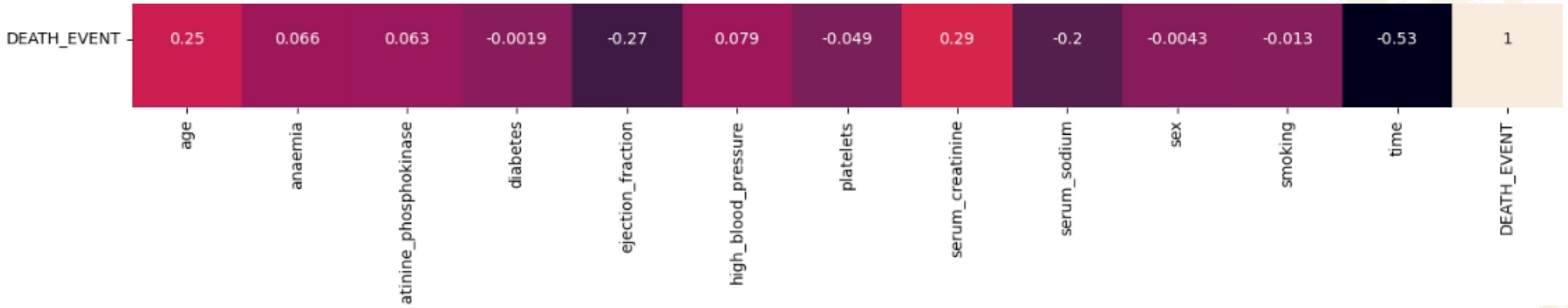
Correlation Matrix

There is a **positive correlation** between **Death Event** and **Age** with a value of **0.25**

There is a **negative correlation** between **Death Event** and **Ejection Fraction** with a value of **-0.27**

There is a **positive correlation** between **Death Event** and **Serum creatinine** with a value of **0.29**

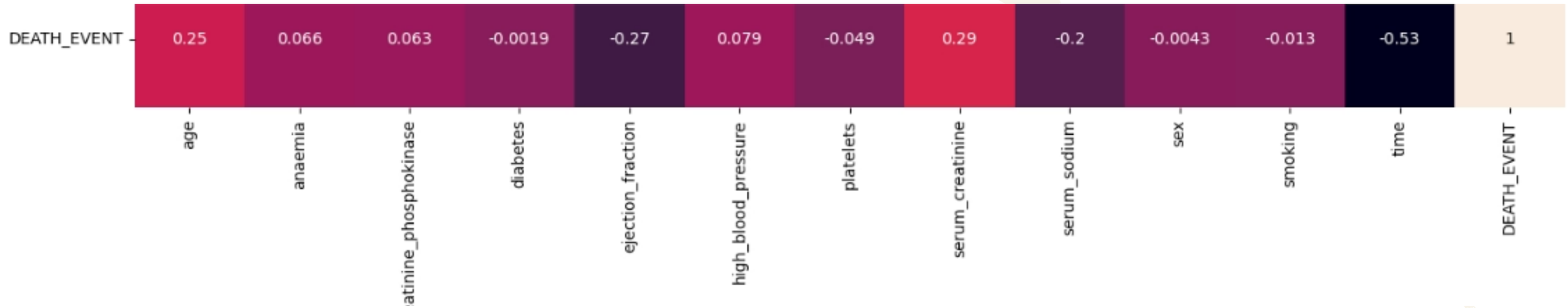
There is a **negative correlation** between **Death Event** and **Time** with a value of **-0.53**



Possible candidates for the classification

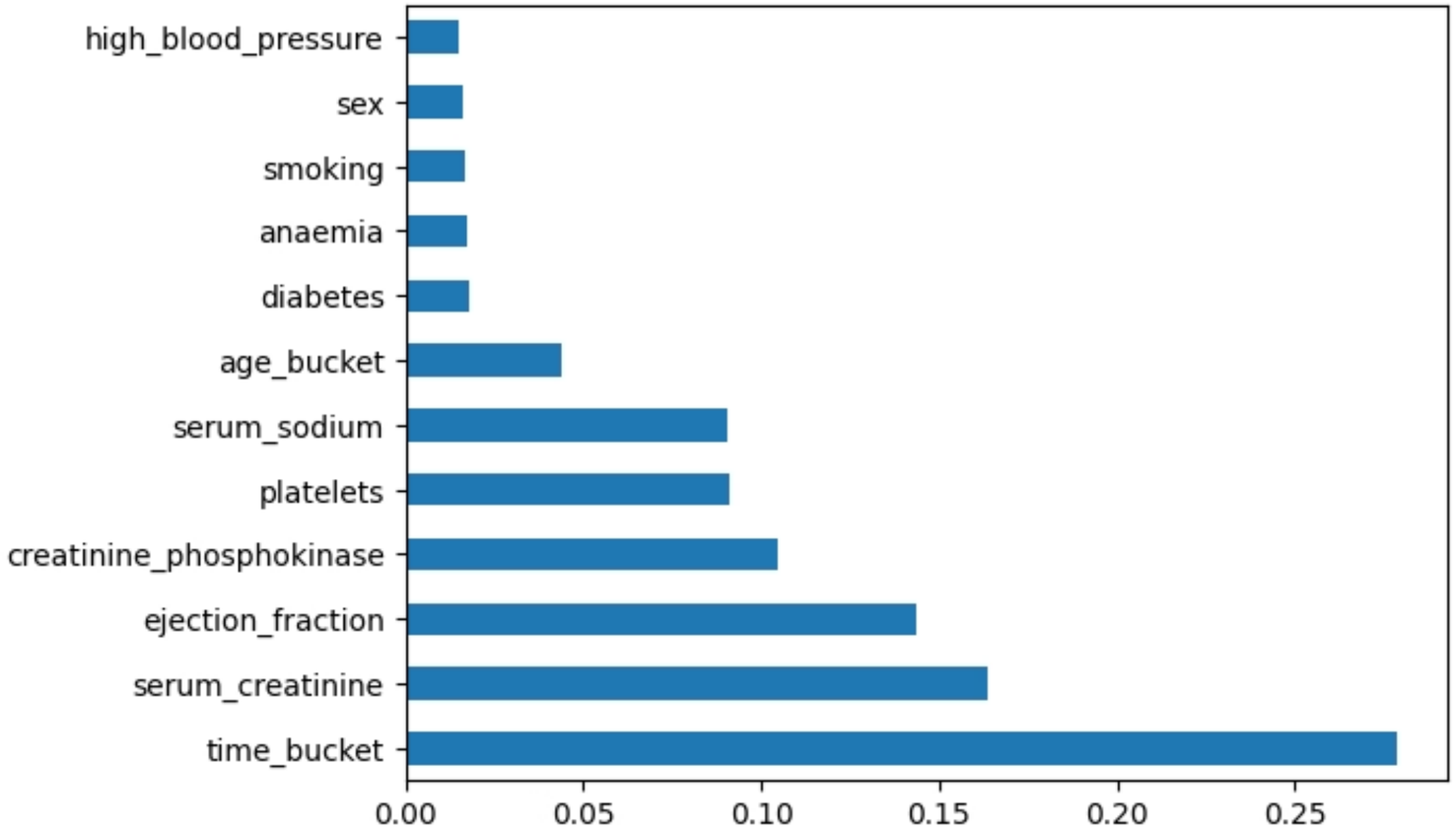
Target variable - Death Event

Independent variables - Age, Ejection Fraction, Serum creatinine, Time



Features ranking using Random Forest Classifier

Possible candidates - Creatinine Phosphokinase, Ejection Fraction, Serum creatinine, Time



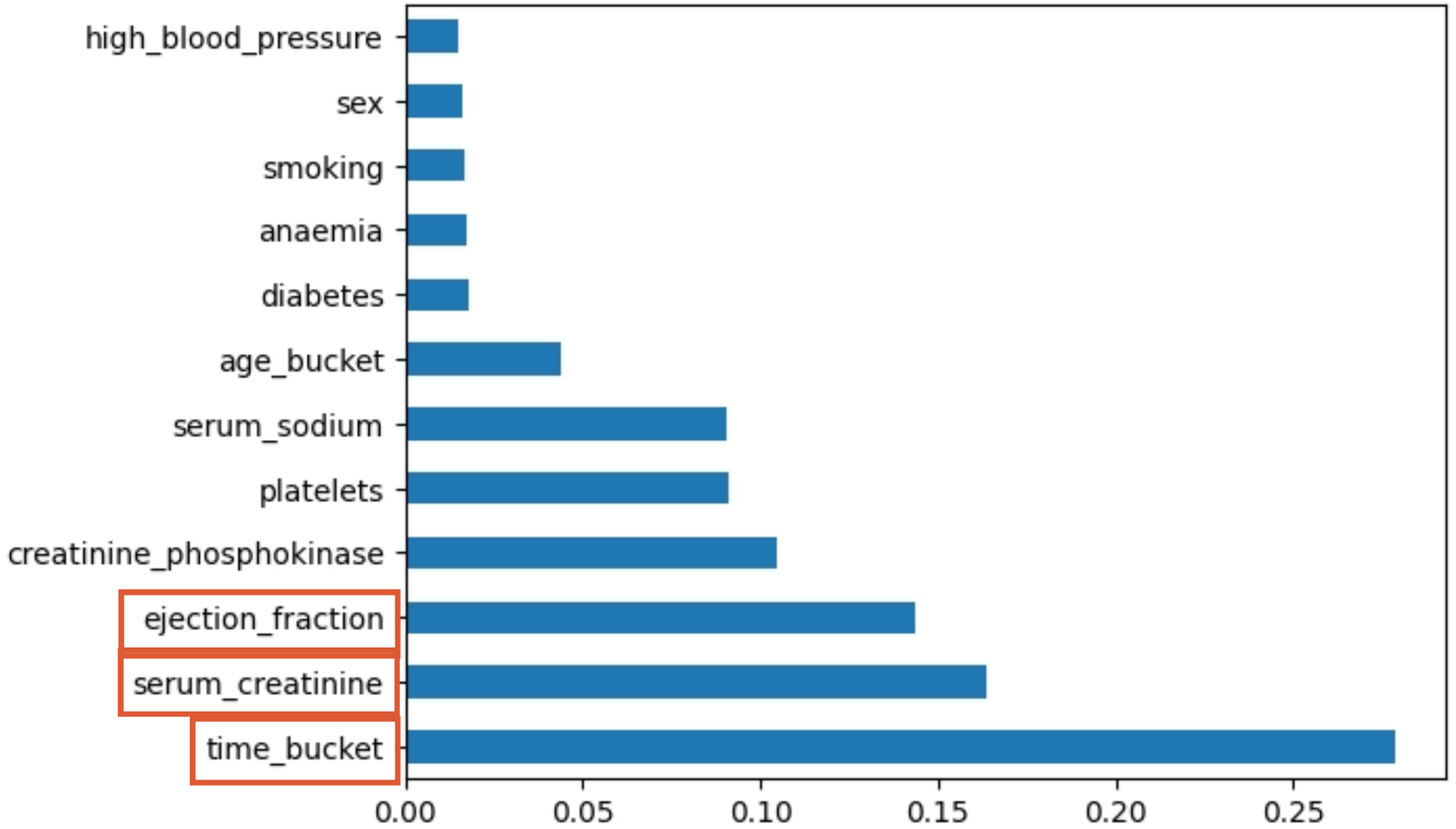
Features selections

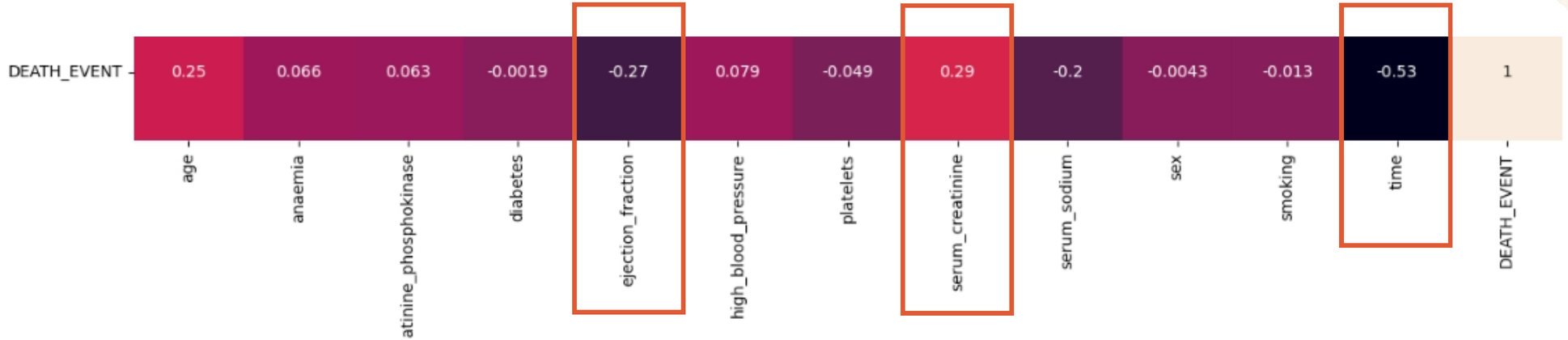
Candidates from the classification

Ejection Fraction

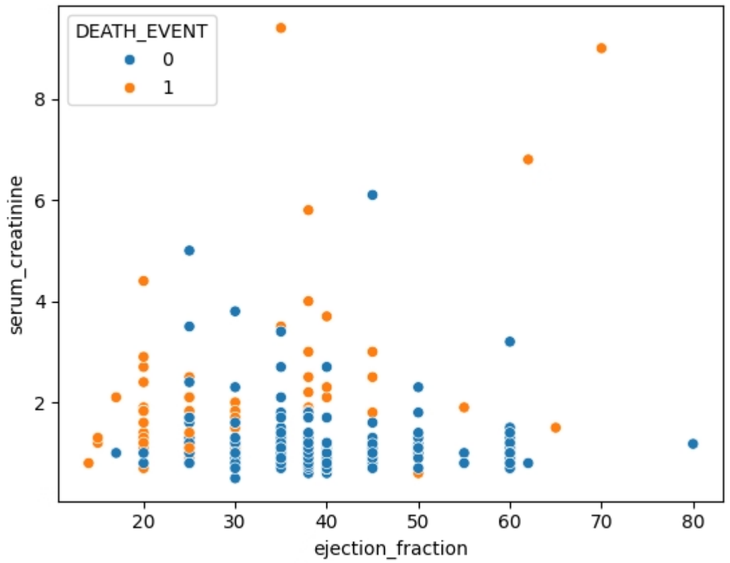
Serum creatinine

Time

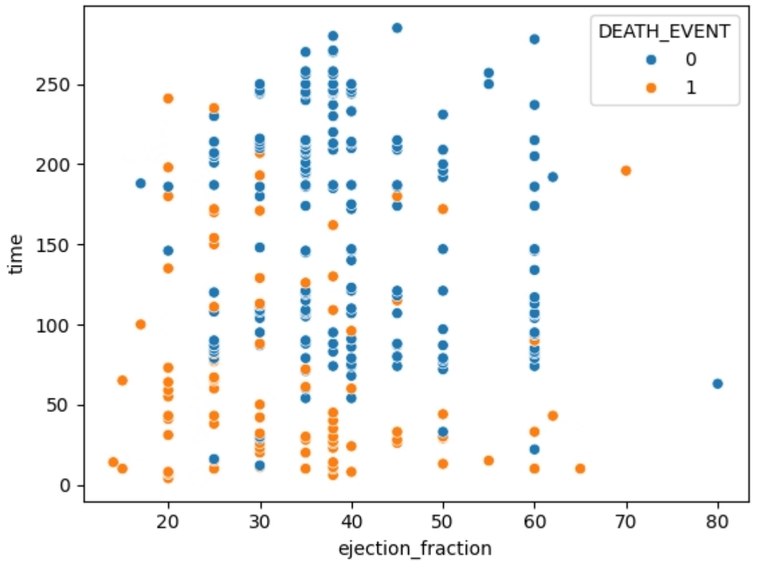




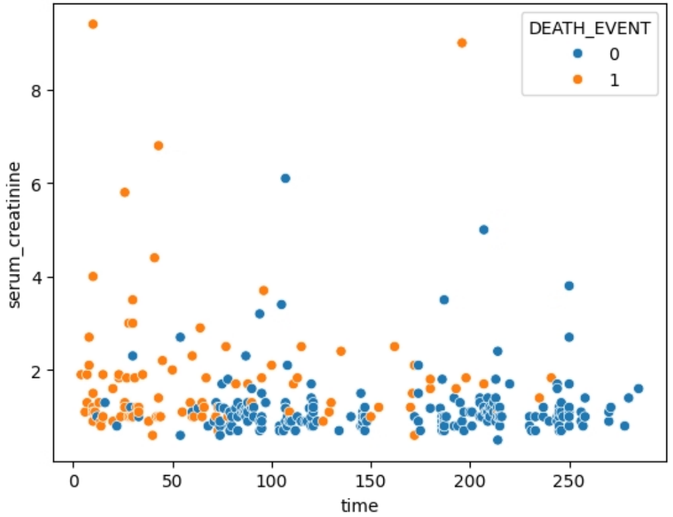
Are there any relationships between ejection fraction, serum creatinine, time and death event?



There is a some correlation between serum creatinine, ejection fraction and occurrence of death events because they are cluster



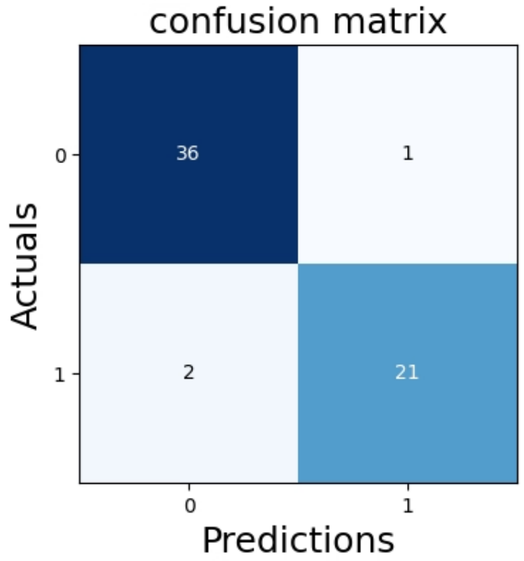
No patterns are evident between time, ejection fraction, and the occurrence of death events.



This scatter plot is even more tightly clustered than the first one, indicating a stronger relationship between time, serum creatinine, and the occurrence of death events.

ML Models Evaluations

Random Forest Classification



SMOTE technique

best n_estimators: 100

best max_features: 1

best samples_split: 8

best max_depth: 3

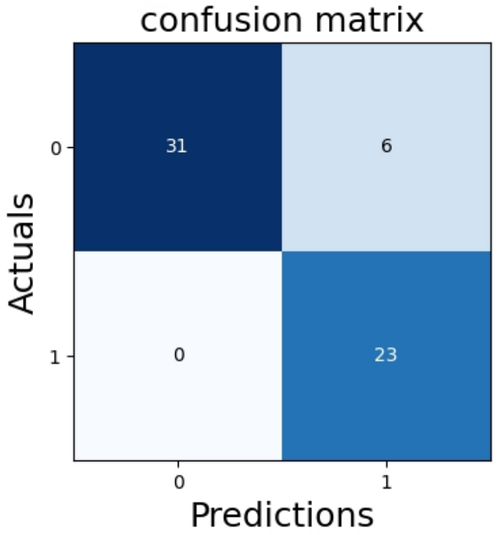
Accuracy Score: 0.9500

F1 score: 0.9333

Precision: 0.9545

Recall: 0.9130

Roc Auc Score: 0.9430



Resampling technique

best n_estimators: 100

best max_features: 0.5

best samples_split: 5

best max_depth: 2

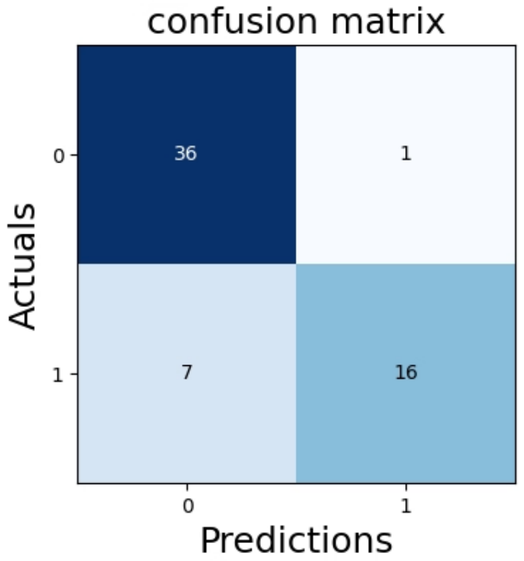
Accuracy Score: 0.9000

F1 score: 0.8846

Precision: 0.7931

Recall: 1.0000

Roc Auc Score: 0.9189



No techniques used

best n_estimators: 100

best max_features: 1

best samples_split: 5

best max_depth: None

Accuracy Score: 0.8667

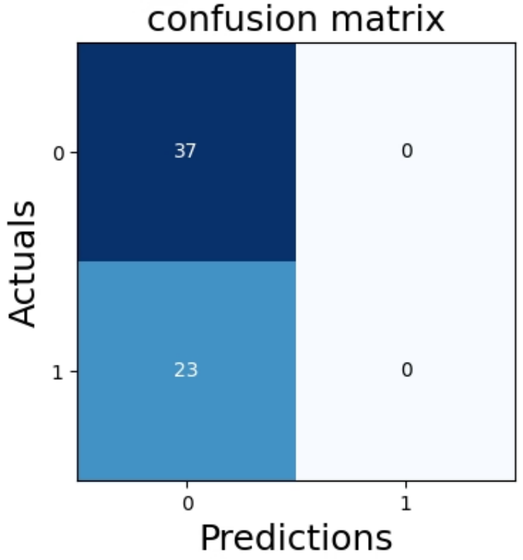
F1 score: 0.8000

Precision: 0.9412

Recall: 0.6957

Roc Auc Score: 0.8343

Artificial Neural Network



SMOTE technique

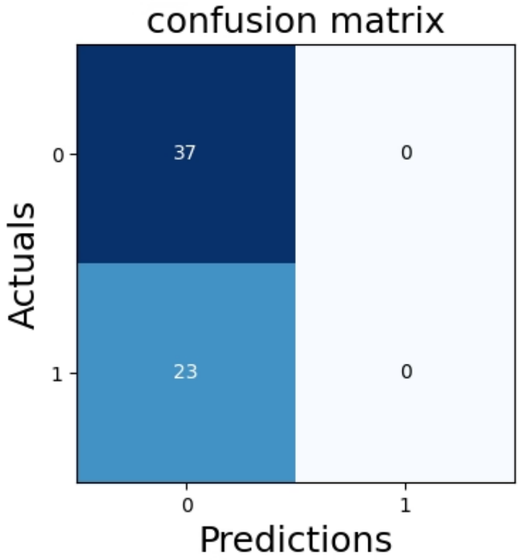
Accuracy Score: 0.3833

F1 score: 0.5542

Precision: 0.3833

Recall: 1.000

Roc Auc Score: 0.5000



Resampling technique

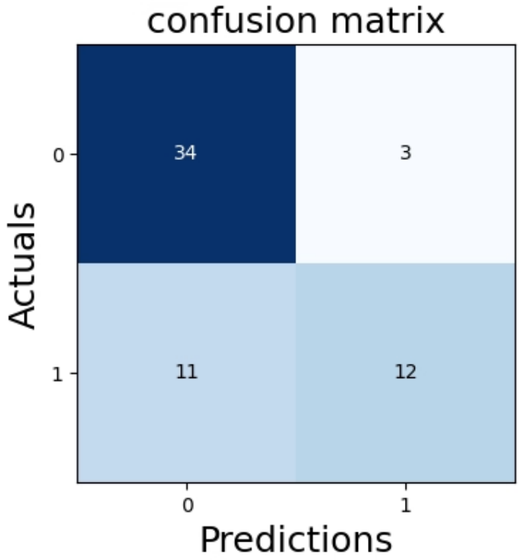
Accuracy Score: 0.6167

F1 score: 0.0000

Precision: 0.0000

Recall: 0.0000

Roc Auc Score: 0.5000



No techniques used

Accuracy Score: 0.7667

F1 score: 0.6316

Precision: 0.8000

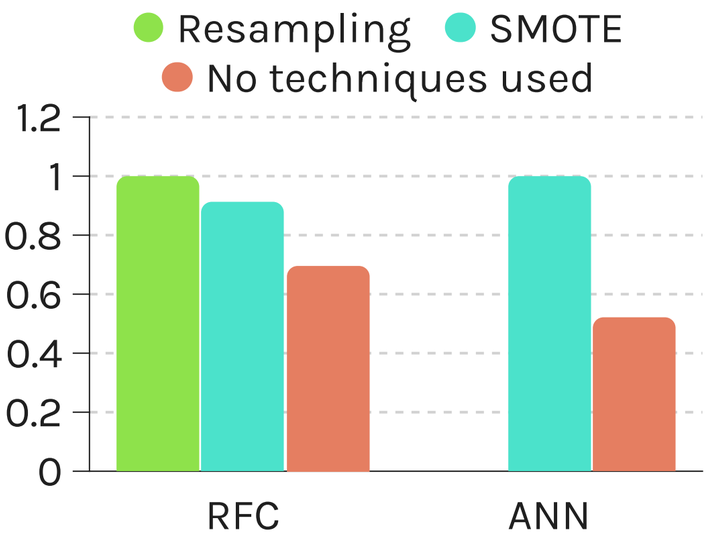
Recall: 0.5217

Roc Auc Score: 0.7203

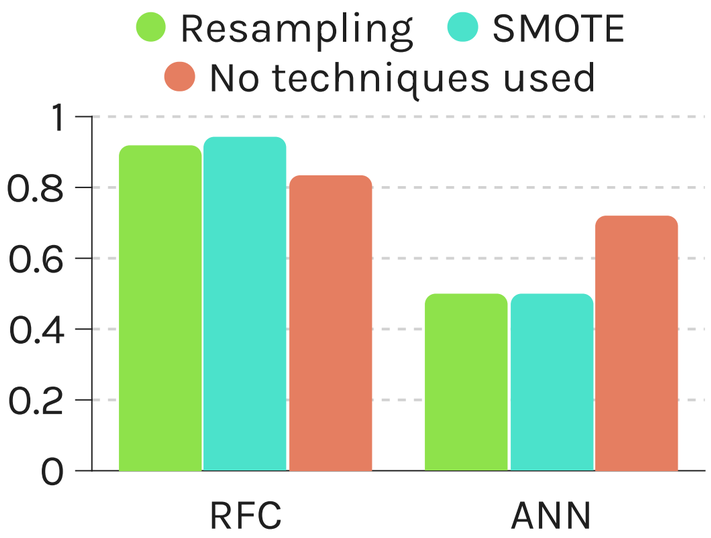
Conclusion

Comparison

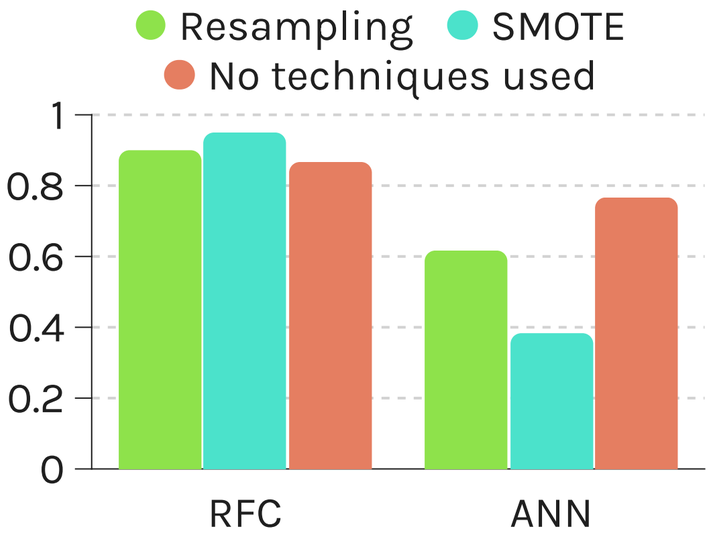
Recall Score



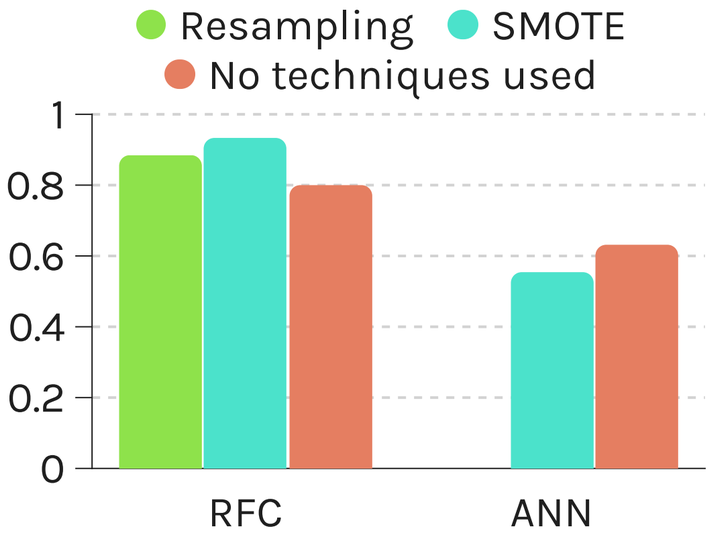
Roc Auc Score



Accuracy Scores



F1 Score



Precision Score