CSE422: Artificial Intelligence

Summer 2023

Section: 10

Group 9

Project Title: Heart Failure Prediction

Submitted to:

Zahin Wahab (ZWB),

Lecturer,

Department of Computer Science and Engineering,

Brac University.

Shayekh Bin Islam (SBI),

Lecturer,

Department of Computer Science and Engineering,

Brac University.

Submission Date: 30 August, 2023

Group No. 9

Group Members:

Nahiyan Rahman Talukder (21101273)

Faiaz Ibnee Rahman (21101151)

Al Rafi Ahmed (21101092)

Tahseen Chowdhury (21101217)

Table of Contents

1. Introduction:

The primary aim of this project namely 'Heart Failure Prediction' is to utilize machine learning models, use a wide range of medical data such as age, gender, blood pressure, creatinine phosphokinase, serum creatinine, serum sodium, and diabetes to predict a critical outcome which is death. By analyzing the data patterns, we aim to create a system that can anticipate whether a patient will survive based on the aforementioned medical conditions. This can assist healthcare professionals in making more informed decisions. As heart failure is a significant health concern worldwide, it motivates us to solve this problem and potentially save lives.

2. Dataset Description:

- **Source**

  - Link: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

  - **Reference:**

  Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making, 20*. https://doi.org/10.1186/s12911-020-1023-5

- **Dataset Factors**

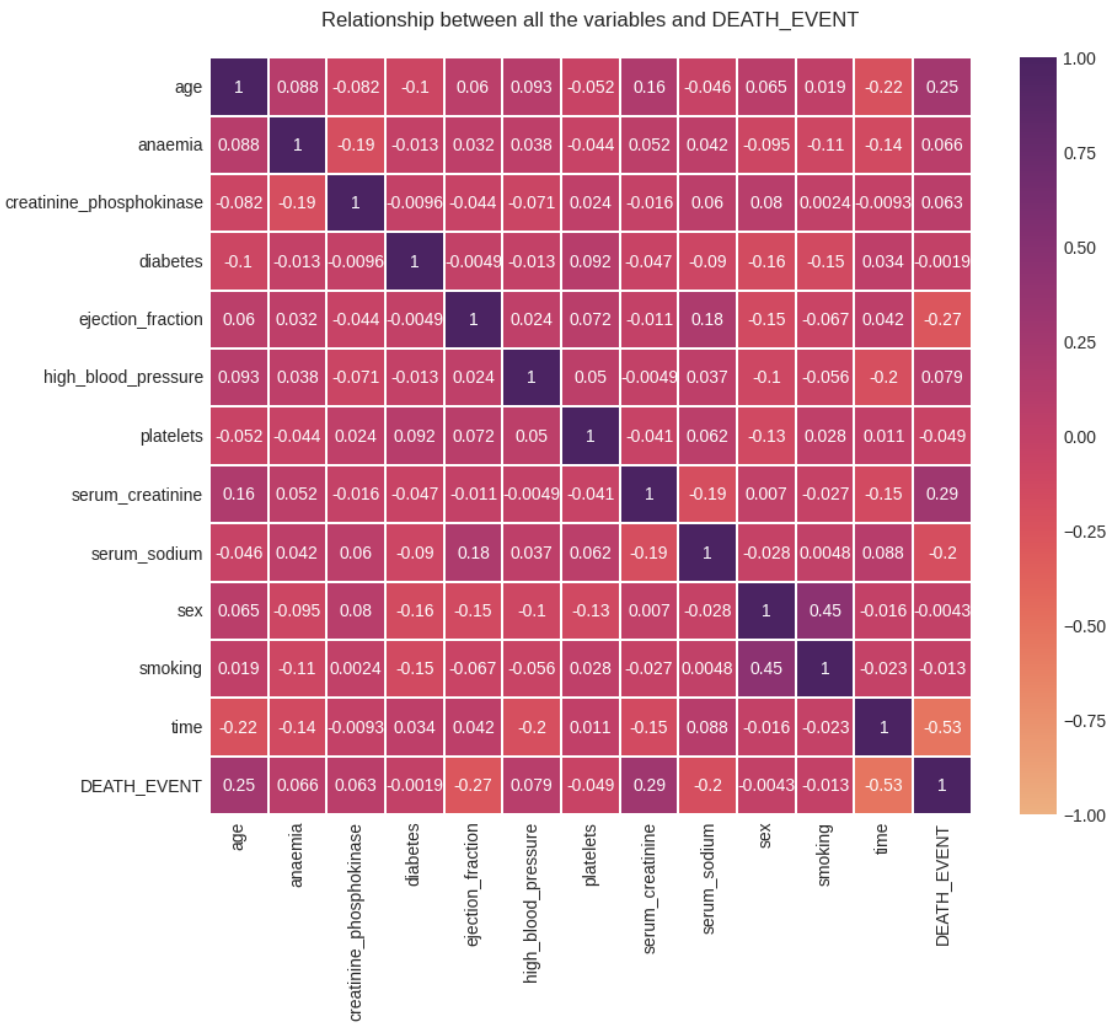
  - Our dataset contains 12 features of age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, time, and death event.

  - It is a binary classification problem as it only predicts two outcomes: 'Death' and 'Not death'.

  - Our dataset contains 299 data points.

  - All of the features of the dataset are quantitative features.

  - We have analyzed the correlation between the features and output both mathematically and visually. Some of the correlation values were negative while some features showed a strong correlation. To visualize the correlation, a heatmap is given below:



- **Imbalanced Dataset**

  - It does not have an equal number of instances in unique classes for the output feature. There are two unique classes 0 & 1 resembling the outcome of ‘Not Death’ and ‘Death’ respectively. According to the dataset, 203 instances belong to 0 / ‘Not Death’ and 96 instances belong to 1 / ‘Death’.

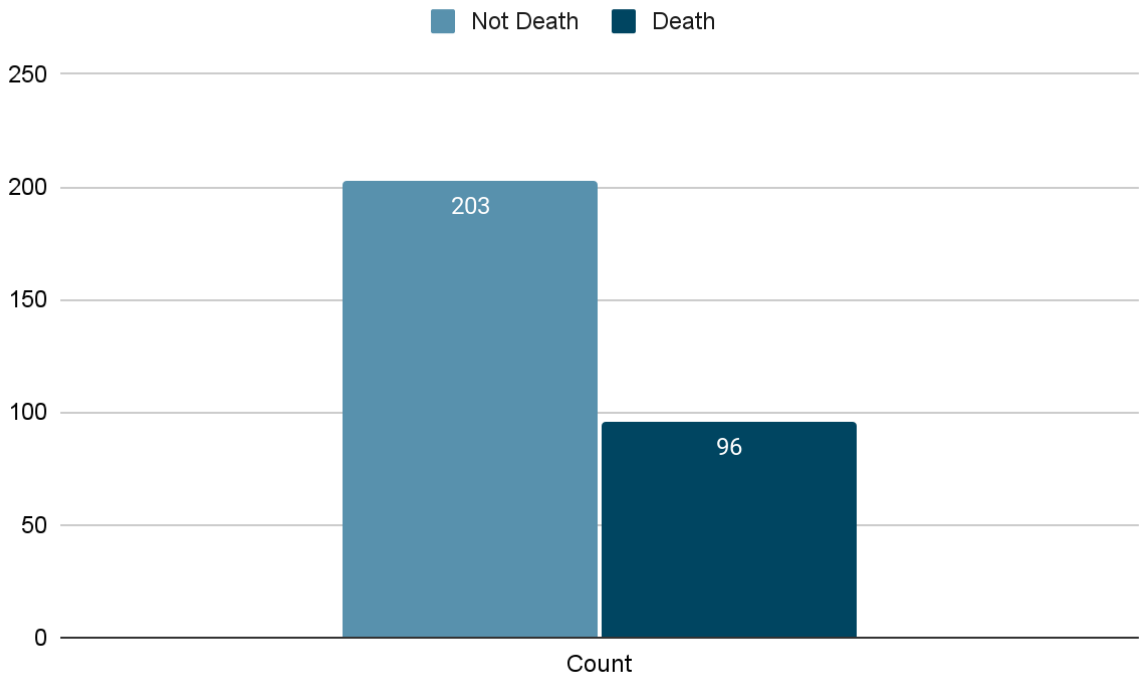


Figure: Output Feature unique class instance count

3. Dataset pre-processing

- **Faults**

  - In our dataset, there are no null values.

  - Also, there are no categorical values.

- **Solutions**

  - Since there are no faults, it is not necessary.

4. Feature scaling

We have selected those features that contain continuous value and data fluctuate in a wide range. But most of the machine learning algorithms assume data are centered to 0 and perform better when data are scaled to mean-centric, scaled on standard deviation or normal distribution. That is why it is necessary to scale the data to fit the models. We have used StandardScaler() for feature scaling which ensures data are distributed -1 to 1 and prepare the data for test matrice.

5. Dataset splitting

We randomly divided the data into a training set and a test set using the function train_test_split(). This approach facilitated an unbiased distribution of data points across both sets, allowing our model to learn from a variety of instances and then be tested on unseen data. We set the test size as 0.3 which split 30% data for testing and 70% data for training. Moreover, the random state is 2 which generates random numbers for row selection in the data set and assigned data points for the test and training set.

6. Model training and testing

- **Logistic Regression**

  Logistic regression is a statistical method used to model the probability of a certain classification. It is a type of regression analysis where the dependent variable is binary or dichotomous, meaning it only has two possible outcomes. We have set the logistic regression model on L2 normalization by tuning the penalty as 'l2' for training.

- **Support Vector Machine**

  The Support Vector Machine (SVM) is a highly effective machine-learning approach for classification and regression. It operates by locating the optimum hyperplane for dividing the data into separate groups. SVM attempts to make the difference between classes as large as feasible in order to handle errors. Also, to ensure the classification prediction is more precise and accurate we set the SVC kernel to linear.

- **Decision Tree Classifier**

  The decision tree is a highly effective supervised learning approach for classification. It builds a tree structure in which each node represents an attribute test, branches represent outcomes, and leaf nodes include class labels. The tree is constructed by recursively partitioning data based on attribute values until stopping constraints such as maximum depth or sample size are fulfilled. The method we have chosen is Gini impurity to maximize information gain and set the maximum depth at 100.

- **Naive Bayes Classifier**

  Naive Bayes Classifiers are a collection of classification algorithms based on Bayes' Theorem. It is a probabilistic machine learning technique that uses the Bayes Theorem to perform classification problems. To train the data, we have selected the Naive Bayes Gaussian model since our data are primarily continuous.

- **Random Forest Classifier**

  Random Forest is a widely used machine learning process for classification. It works by building a large number of decision trees during training. Each decision tree ensembled in a Random Forest model is created using a sample selected with replacement from the training set. Here, we have also tuned the model by setting the criterion as entropy with n-estimator = 110 and max depth = 100.

7. Model selection/Comparison analysis

- To select a better training model, we have analyzed the performance on different parameters including accuracy rate, precision score, recall score, f1 score, and confusion matrix. The analysis is shown below through bars and charts:

- **Prediction Accuracy**

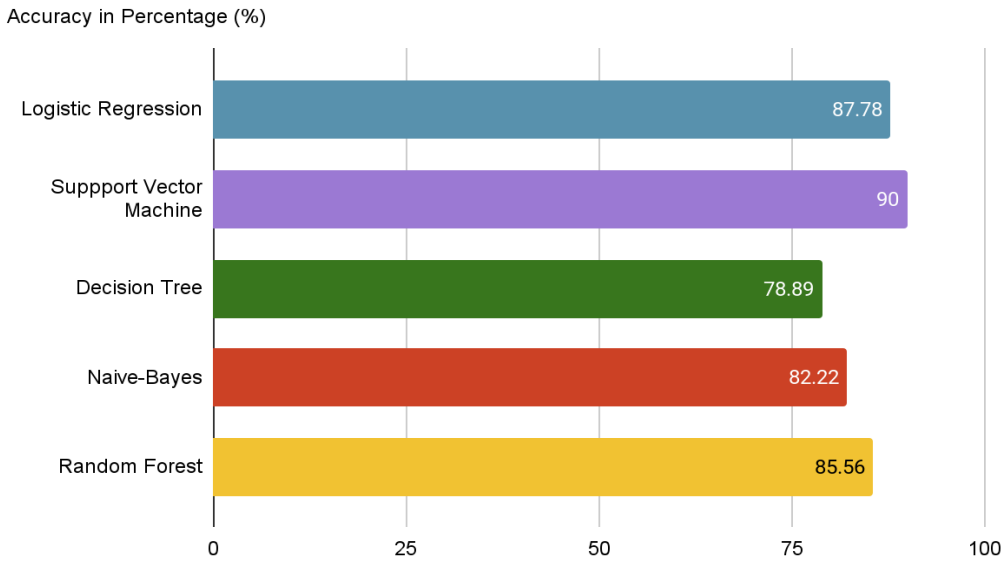


Fig: Accuracy score of the models

- **Precision and Recall Comparison between models**

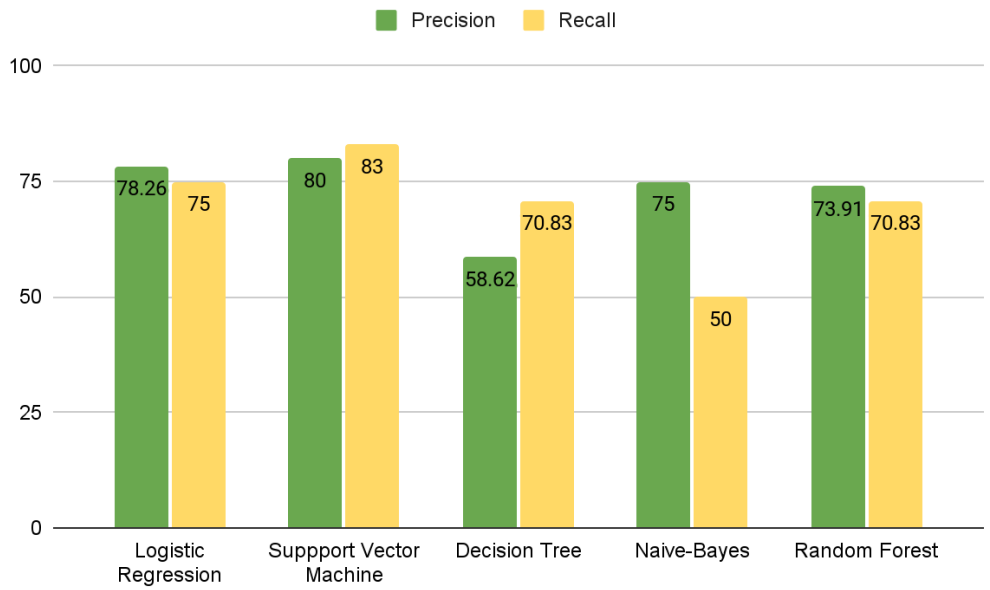


Fig: Precision and Recall Percentage of Each Model

- **F1 Score Analysis**

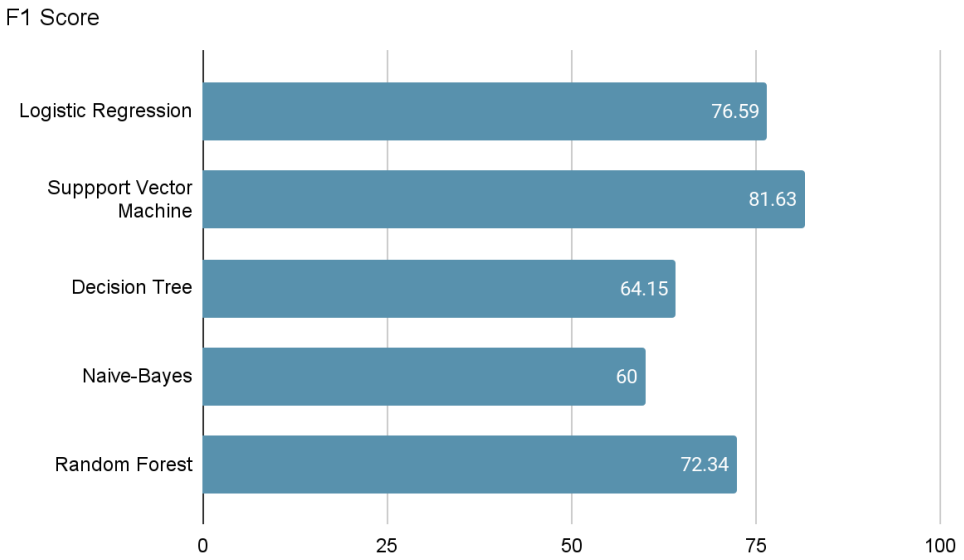


Fig: F1 Score Comparison

- **Confusion Matrix of Each Model**

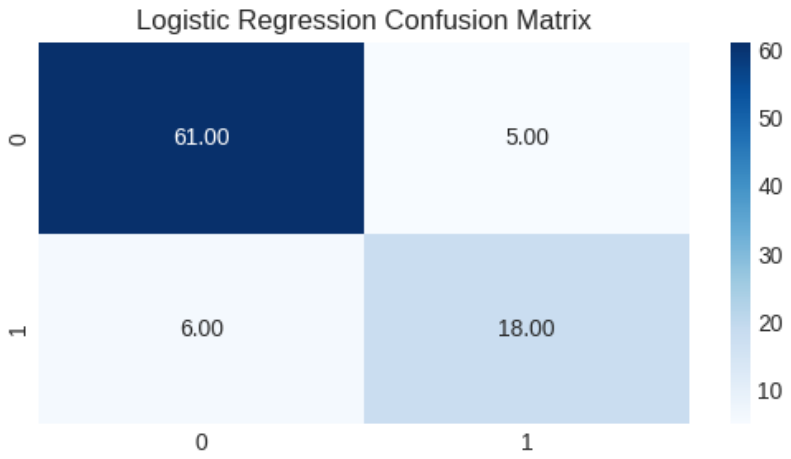


Fig: Confusion Matrix for Logistic Regression

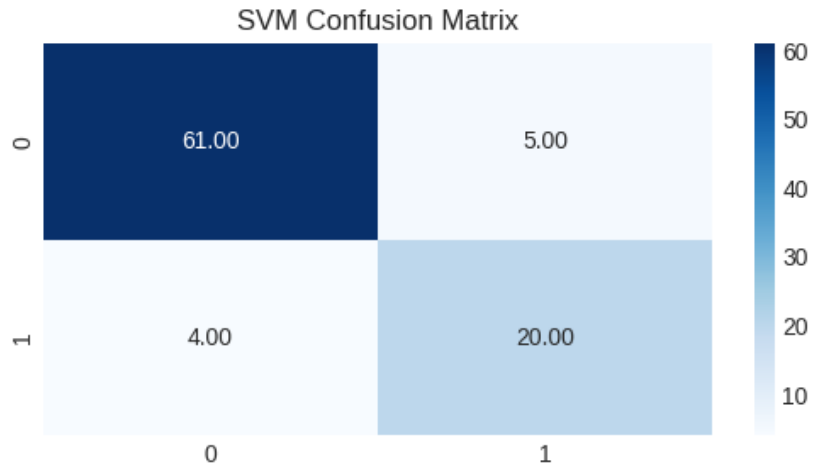


Fig: Confusion Matrix for Support Vector Machine

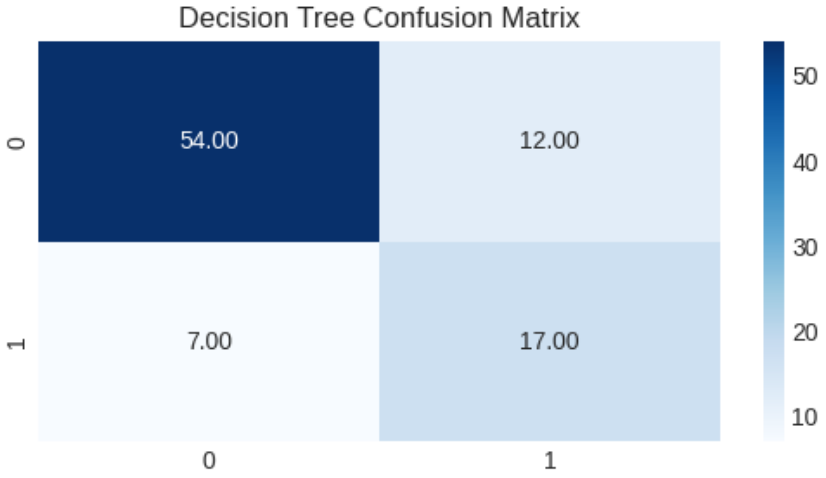


Fig: Confusion Matrix for Decision Tree

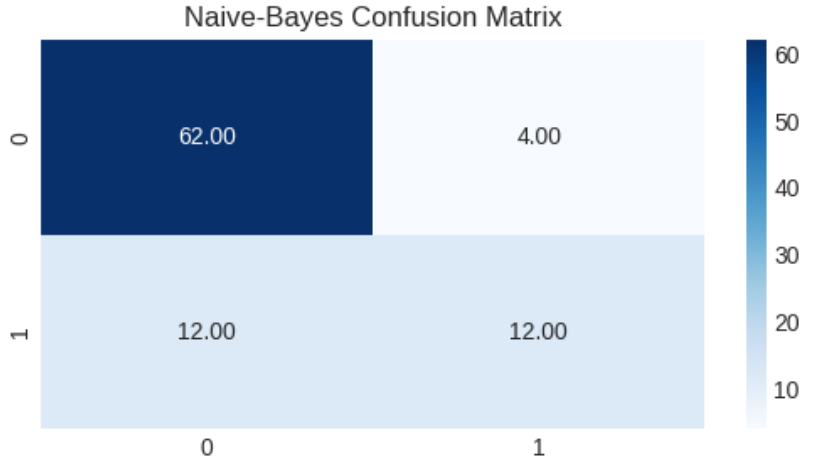


Fig: Confusion Matrix for Naive-Bayes

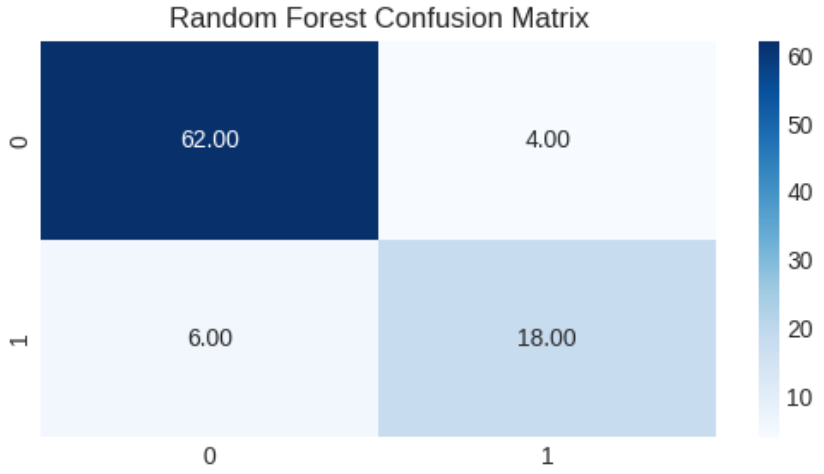


Fig: Confusion Matrix for Random Forest

- Since our dataset is imbalanced. The accuracy score is not reliable in this context. So it is better to focus on the f1 score which is the harmonic mean of precision and recall. After analyzing the bar plots, and confusion matrix of each model it is clear that Support Vector Machine modeling is the best algorithm for predicting the outcome. It ranks top with the highest F1 score too. In SVM confusion, it scores better in predicting the true positive and true negative value while minimizing the false positive and false negative.

8. Conclusion

In this “Heart Failure Prediction” project by using machine learning techniques we analyzed a variety of features capable of early detection of death events of a patient. This project shows us the potential of predictive modeling in the medical field. By addressing the challenge of heart failure prediction we gained valuable insights into the application of machine learning in real-world scenarios.