Heart Disease Prediction

A. Karmacharya, B. W. S. Lee, B. Ooi, J. M. Gonzalez, N. Rajic

**Abstract**—In this project, we are assigned to evaluate the use of machine learning algorithms to predict the risk of suspected cardiovascular disease based on 12 variables including age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, and time datasets collected from Kaggle [1]. After feature scaling, modeling was performed using a k-fold cross-validation score to analyze clinical data. The overall accuracy of model prediction in heart disease was accessed by receiver operating characteristic (ROC) curve and precision.

————————————————

* *A. Karmacharya is a student from San Jose State University. Email: abhaya.karmacharya@sjsu.edu*
* *B. W. S. Lee is a student from San Jose State University. Email: benjamin.w.lee@sjsu.edu*
* *B. Ooi is a student from San Jose State University. Email: bennykeanhoe.ooi@sjsu.edu*
* *J. M. Gonzalez is a student from San Jose State University. Email: jonahmorgan.gonzalez@sjsu.edu*
* *N. Rajic is a student from San Jose State University. Email: nemanja.rajic@sjsu.edu*

—————————— ◆ ——————————

# 1 Introduction

Heart disease remains the major cause of death for males and females in the United States. One patient dies from cardiovascular disease every 36 seconds in the United States and approximately 655,000 Americans die annually from heart disease, which is 1 in 4 deaths [2]. Cardiovascular disease costs around $219 billion annually from 2014 to 2015 in the United States [2]. This comprises the cost of medical care facilities, medications, and death-related loss of productivity. Most heart disease may be avoided by resolving dietary risk conditions such as cigarette consumption, inadequate lifestyle and overweight, lack of physical activity, and excessive drug use. Individuals with or at severe cardiovascular risk despite the existence of one or perhaps more risk factors, including diabetes, hypertension, and unhealthy cholesterol level, are in need of early prediction and treatment, where a machine learning model could be very beneficial.

# 2 Problem Statement

The goal of this project is to build a machine learning model to predict the risk of death from a heart failure case; the tasks involved are as follows:

* Collect patient datasets from multiple sources, including Kaggle, UCI, etc.
* Summarize, visualize, and find the relationship between variables from the dataset.
* Performed data preprocessing techniques such as data cleaning, feature extraction, feature scaling, dimensionality reduction, and clustering.
* Splitting the dataset into training and test sets.
* Train supervised classifier machine learning model using a training set to predict the likelihood of heart failure.
* Evaluate the performance of various classification algorithms using K-fold, confusion matrix, cross-validation, precision, and ROC curve.
* Improve the accuracy of classification models by fine-tuning the hyperparameter using grid search cv.
* Finalize the model with the best hyperparameters.

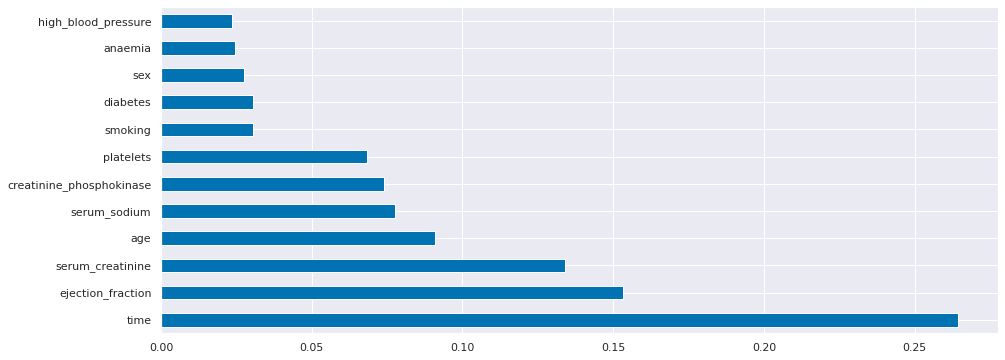
# 3 Data Exploration

## 3.1 Data Set

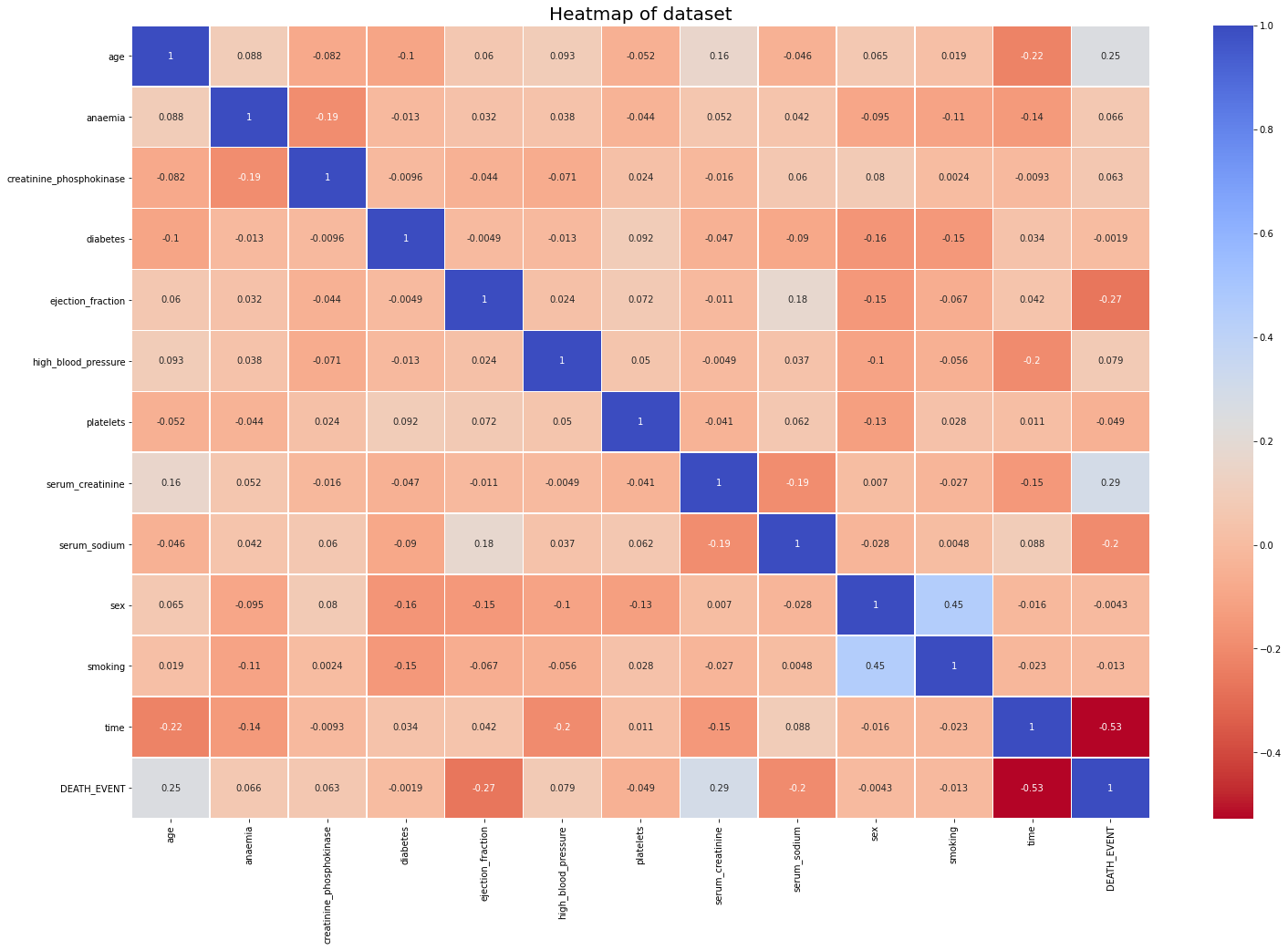
The dataset is publicly accessible on the platform of Kaggle and is by an active cardiovascular analysis of residents of Framingham, Massachusetts. The dataset is a subset of the larger dataset to reduce the training and processing time delay. It includes over 299 records, 12 features, and one label as following:

1. Age: Age of patients (Continuous Data)
2. Anaemia: Decreases of red blood cells or haemoglobin (Boolean variable, 1 = true; 0 = false)
3. Creatinine phosphokinase: Level of the CPK enzyme in the blood (Continuous Data)
4. Diabetes: Indicate the patient has diabetes (Categorical variable, 1 = true; 0 = false)
5. Ejection fraction: Percentage of blood leaving the heart at each contraction (Continuous Data)
6. High blood pressure: Indicate the patient has hypertension (Categorical variable, 1 = true; 0 = false)
7. Platelets: Platelets in the blood (Continuous Data)
8. Serum Creatinine: Level of serum creatinine in the blood (Continuous Data)
9. Serum Sodium: Level of serum sodium in the blood (Continuous Data)
10. Sex: 1 = male, 0 = female (Categorical Variable)
11. Smoking: Indicates the patient is a smoker (Categorical Variable, 1 = true; 0 = false)
12. Time: Follow-up periods
13. DEATH\_EVENT: If the patient deceased during the follow-up period. (Label Variable, 1 = Death; 0 = No Death)

## 3.2 Feature Selection

To obtain optimal precision for our model, several feature selection techniques are used to identify positively correlated features from the dataset and eliminate insignificant features that do not contribute much to the target variable. Feature importance and correlation heatmap are used to identify the features that contribute the most information to the target variable. Extra Tree Classifier machine learning algorithm is used to extract the most important features from the dataset. Figure 1 indicates that serum creatinine, ejection fraction, and time have contributed approximately 60% of importance to the target variable. Besides, from Figure 2, the correlation heatmap shows that time, serum creatinine and ejection fraction have a positive correlation with death event which is the target variable. Therefore, we will only select 3 features which are time, serum creatinine, and ejection fraction to train our classification machine learning model.   


*Figure 1: Feature Importance*



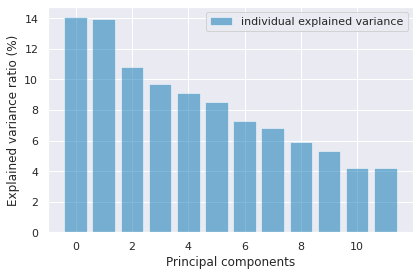
*Figure 2: Correlation Heatmap*

## 3.5 PCA and Clustering

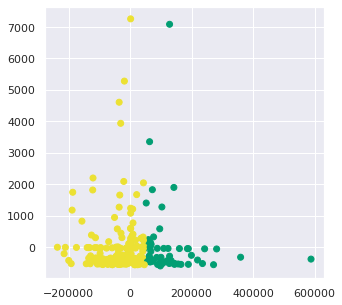
Principal component analysis (PCA) is a dimensionality reduction method that transforms large features onto low-dimensional space and contains most of the dataset’s information. Before applying PCA on the large datasets, normalization on continuous variables must be carried out to prevent biased results. Normalization mathematical formula as (1):

(1)

For instance, variables that have high range differences will dominate variables that have low range differences. In order to visualize the dataset on a scatter plot, we projected 12 features onto 2-dimensional space using principal component analysis and k-means clustering to visualize the two clusters.



*Figure 3: Individual component variance ratio*



*Figure 4: K-means clustering*

Figure 3 shows that all of the components are equally important in explaining most variance in the dataset. However, we will select only the first two principal components for visualization purposes and plot it to spot if there are any clear clusters. Figure 4 is a scatter plot of K-means clusters, which clearly shows distinguishable clusters in the data.

# 4 Algorithms and Techniques

## 4.1 Logistic Regression

Logistic regression is used to model the probability of heart failure based on the given variables time, serum creatinine, and ejection fraction. Data is fit into an S-shaped curve, or a sigmoid, to show the probability of an observation belonging to either heart failure or no heart failure categories.

## 4.2 Support Vector Machine (SVM)

SVM is a supervised learning algorithm, where based on the training data, the algorithm finds an optimal hyperplane that is able to classify new data points. The hyperplane acts as a decision boundary between heart failure and no heart failure. In SVM, the most similar items between classes are the support vectors, or the maximum boundaries, which are equidistant from the hyperplane. Since SVM finds the linear hyperplane, and some datasets have non-linear decision boundaries, a Radial Basis Function kernel is applied to project the dataset to a higher-dimensional space, in case the dataset is not linear.

## 4.3 Naive Bayesian

In the Naive Bayes algorithm, it is assumed that each feature’s presence is independent of any other feature. Each variable independently contributes to the probability of heart failure or no heart failure. Bayes’ Theorem is applied to calculate for the highest posterior probability following (2):

(2)

where y represents heart failure or no heart failure and X represents the variables time, serum creatinine, and ejection fraction.

## 4.4 K-Nearest Neighbor (KNN)

KNN is a non-parametric supervised learning algorithm, which assumes that similar data exist in close proximity. When a new data point is added, its classification follows the k-nearest neighboring data points. The metric used to calculate for the nearest neighbors is Euclidean distance, and in our project, we chose a value of 6 for the number of k-neighbors.

## 4.5 Decision Tree

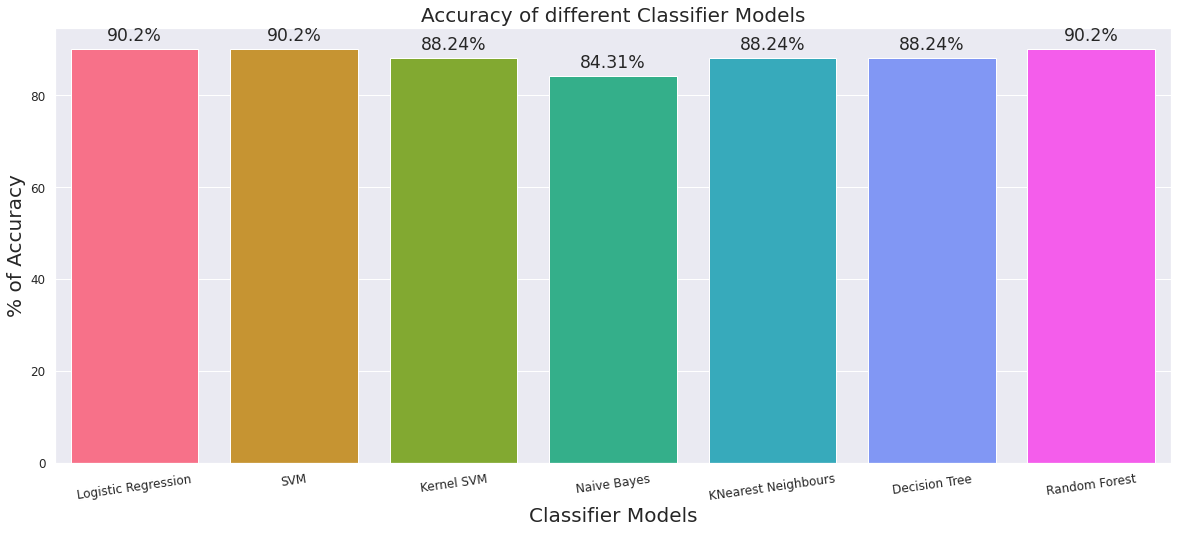
The decision tree is also a supervised learning algorithm suitable for regression and classification problems. This algorithm works by creating a training model that can be used for prediction through simple decision rules inferred from the training data. Prediction starts from the root of the tree, and by comparing the attributes, determine the next node to jump to, until the leaf node is reached.

## 4.6 Random Forest

Random Forest is built of many decision trees. When building each decision tree, bagging and feature randomness is used to create uncorrelated trees. These trees are then merged, forming a forest of decision trees, which will increase the overall predictive result of the model.

# 5 Experimental Evaluation of Results

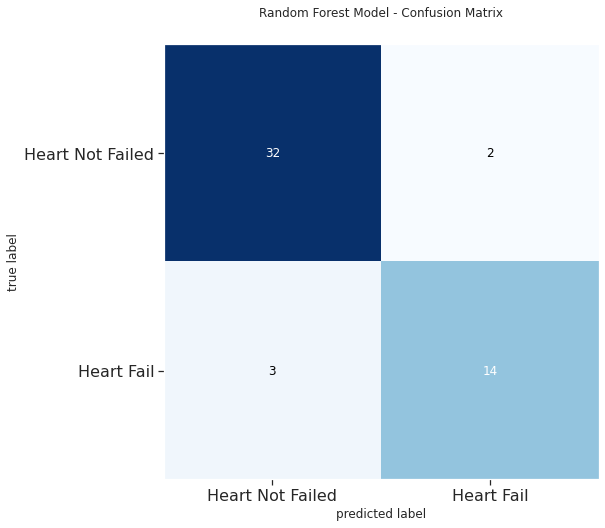
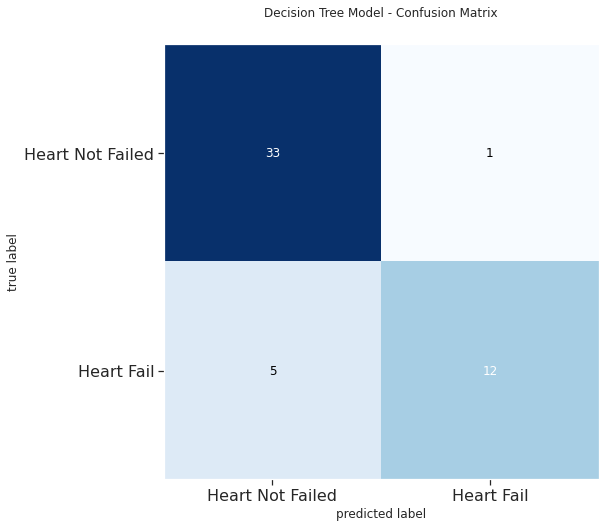
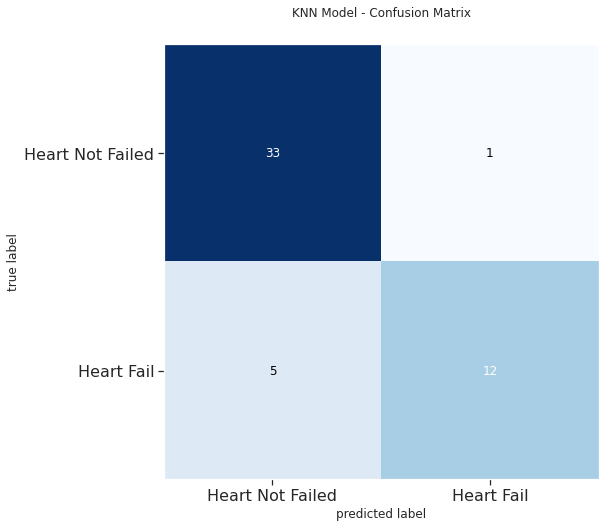
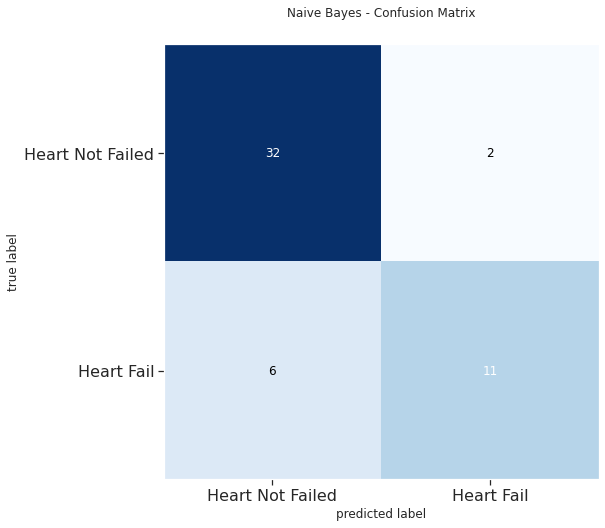
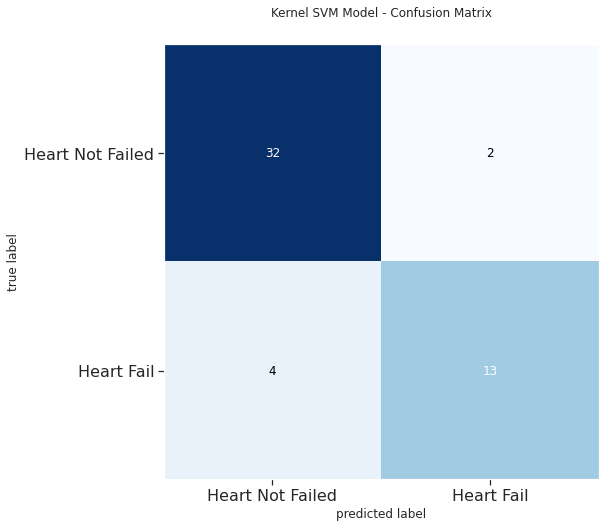
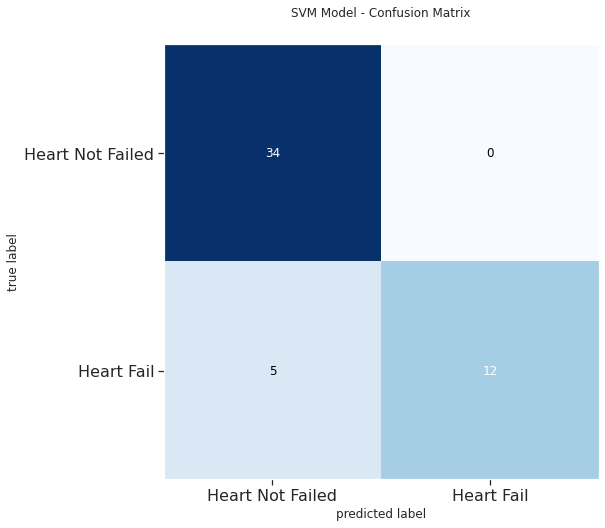
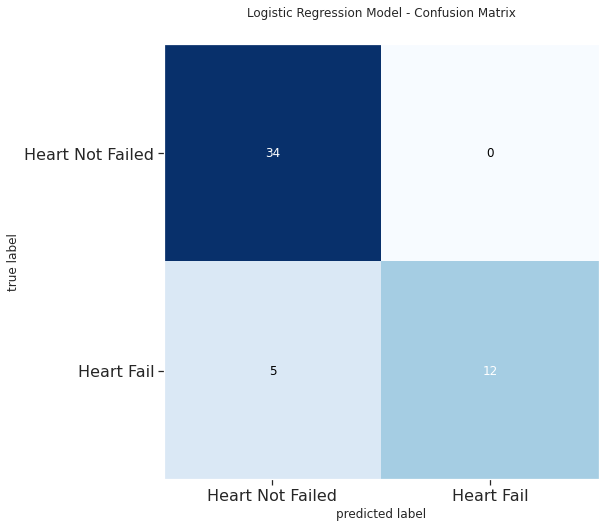
## 5.1 Accuracy



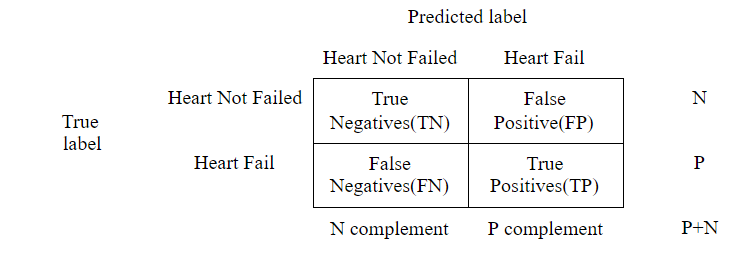
*Figure 5: Graph of the accuracy of different classifier models*

Figure 5 shows the accuracy of the classifier models Logistic Regression, SVM, Kernel SVM, Naive Bayes, K-Nearest Neighbors, Decision Tree, and Random Forest. Accuracy shows the percentage of correctly identified cases (heart failure / no heart failure) out of all cases. As it is shown, Logistic Regression, SVM, and Random Forest models have the highest accuracy.

## 5.2 Confusion Matrix

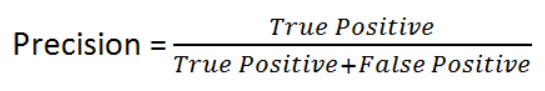


*Figure 6: Confusion Matrix of different classifier models*

Figure 6 shows the Confusion Matrix of the classifier models Logistic Regression, SVM, Kernel SVM, Naive Bayes, K-Nearest Neighbors, Decision Tree, and Random Forest, according to the following format:  


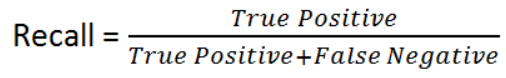
## 5.3 Precision

Precision is the measure of how accurate the model is. By comparing the number of true predicted positives with the total number of predicted positives, we can determine how accurate our models can accurately predict heart failure. High accuracy is essential to make sure there are less occurrences of false alarms on heart condition (false positive) , since it might cause panic to an individual.



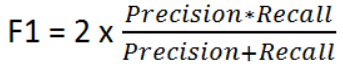
## 5.4 Recall

Recall is known as the model’s sensitivity, where it shows how many true positives are captured by the model out of the actual number of positives. This project is especially significant since a false negative would mean that the predicted no heart failure individual will have a heart failure. Therefore, it is important to have a high recall to make sure the number of false negatives is low.

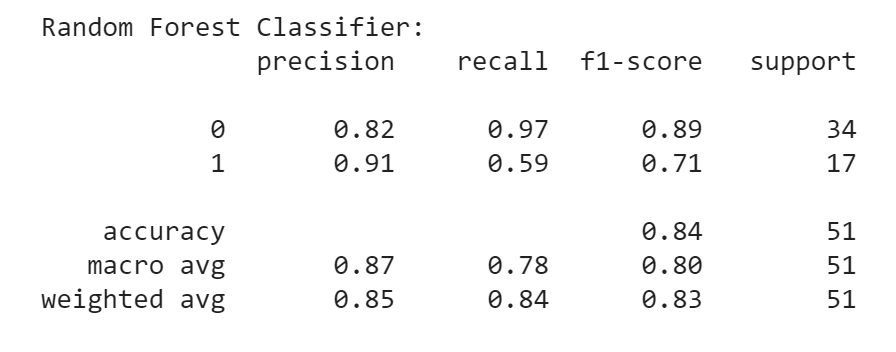
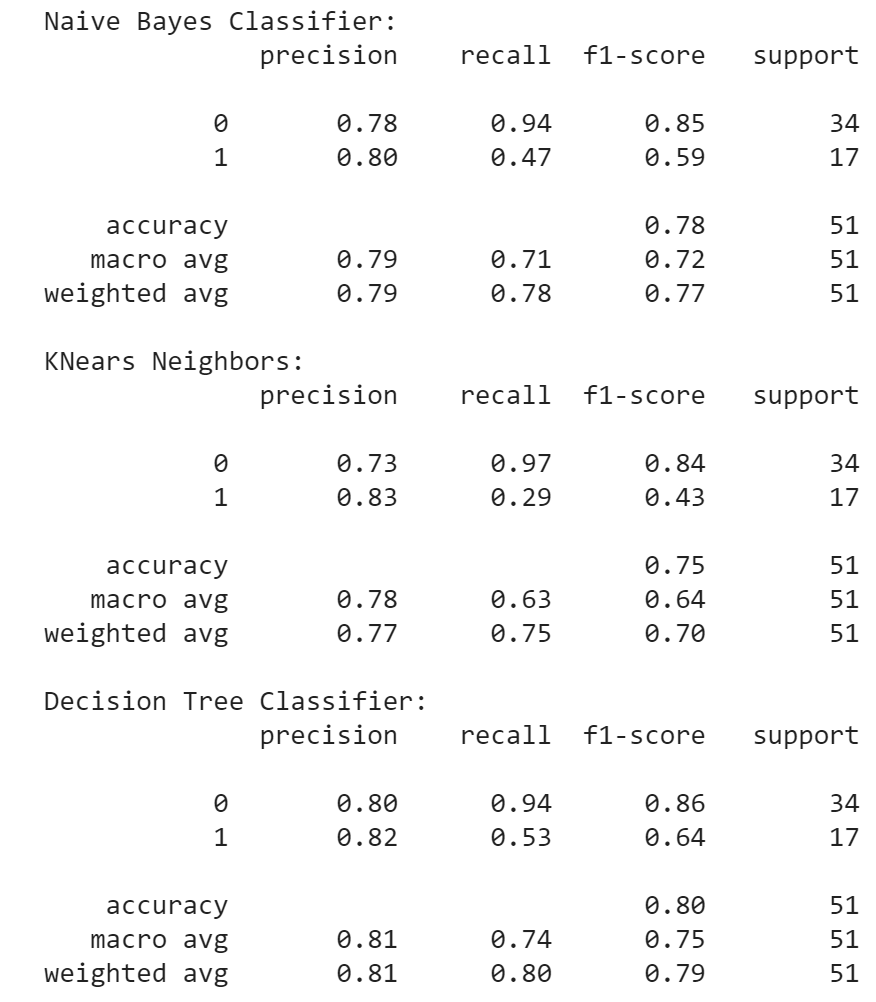
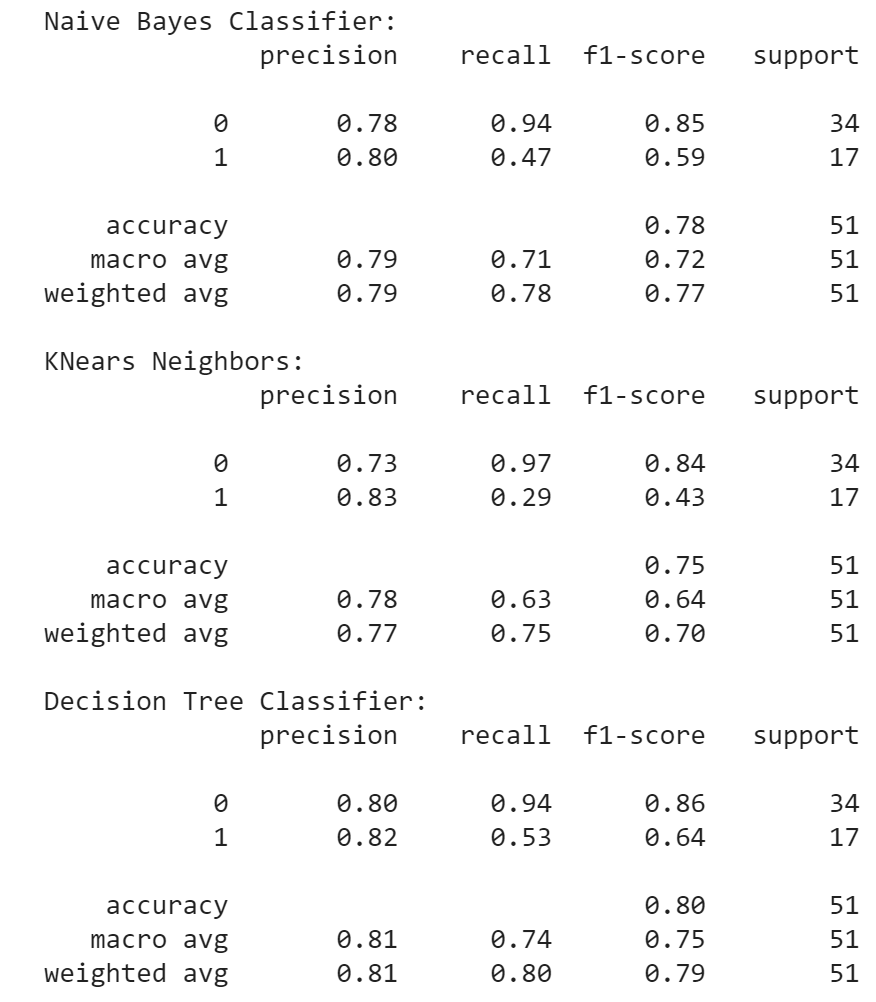
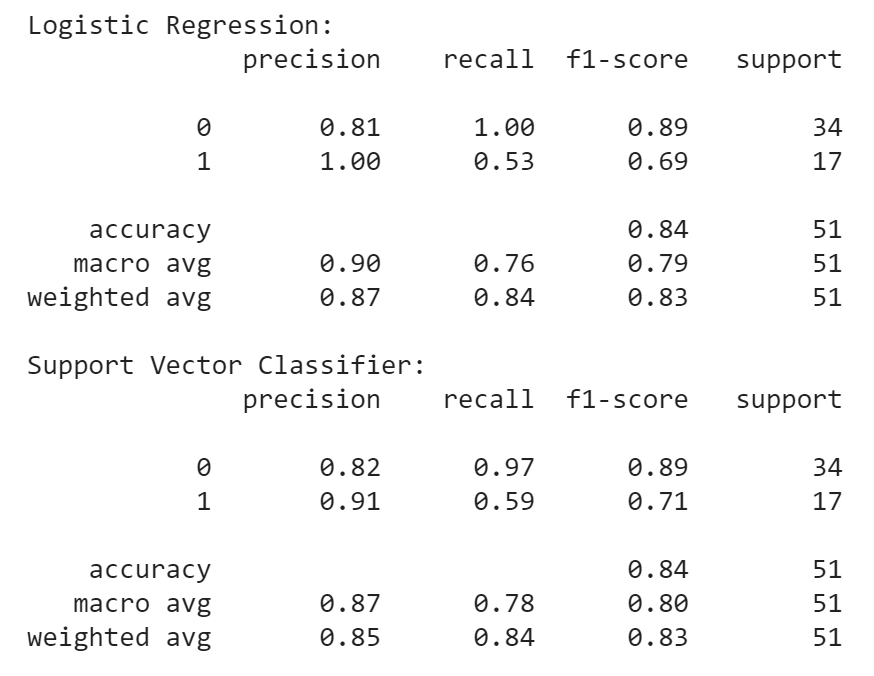


## 5.3 F1-score

To strike a balance between precision and recall, F1 score is referenced to. Since precision focuses on minimizing the false positives and recall focuses on minimizing the false negatives, there is a need to find an optimal case where there are less false predictions.



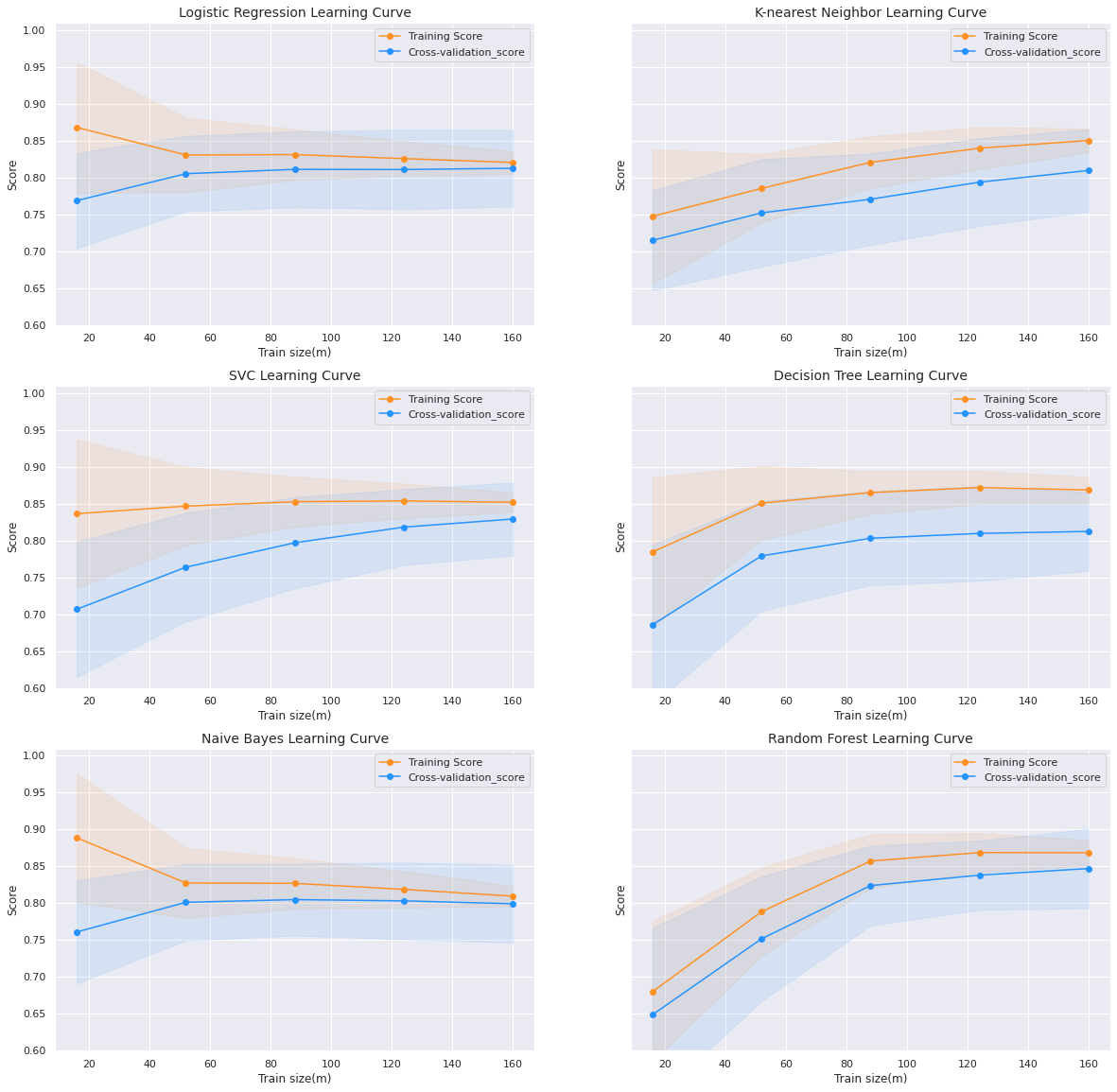
In Figure 7, class 0 represents no heart failure and class 1 represents heart failure. The scores for heart failure will be focused on. Overall, SVM and Random Forest classifiers have the highest f1-scores.



*Figure 7: Precision, Recall, and F1-scores for the classifier models Logistic Regression, SVM, Kernel SVM, Naive Bayes, K-Nearest Neighbors, Decision Tree and Random Forest.*

# 6 Comparing Machine Learning Models

## 6.1 Learning Curves

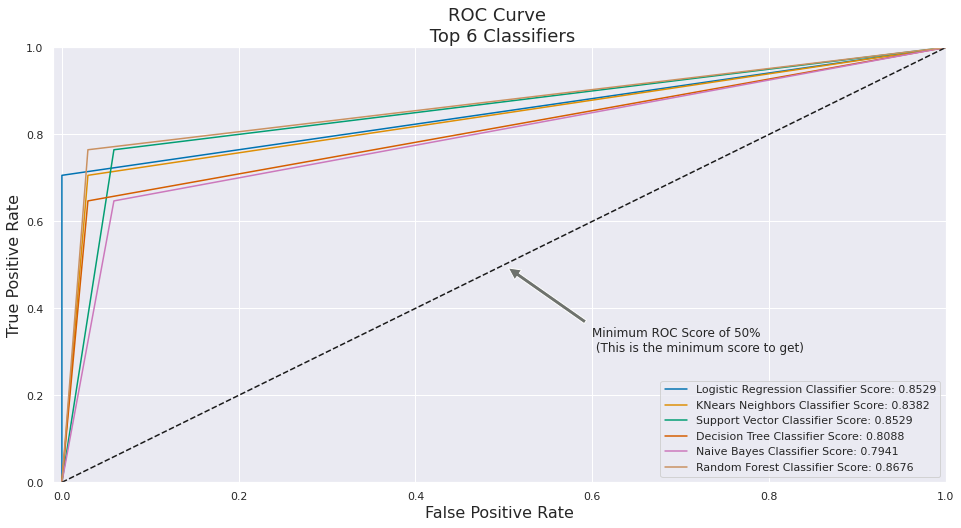


*Figure 8: Learning curves for different classifier models*

In Figure 8, each algorithm’se training score curve represents the learning curve calculated from the training dataset that shows how well the model learns as the size of the training set increases. The cross-validation score curve shows how well the model is generalizing.

We can see that both the Logistic Regression and Naive Bayes models show a low variance and a high bias. The model is underfitting the data and does not seem to improve with more training data. A future solution for this would be to increase the complexity of the model by including more features. For the rest of the curves, adding more training samples will most likely increase the models’ accuracy, especially for SVC and Random Forest.

## 6.2 ROC Curve



*Figure 9: ROC curve for all the classifier models*

In Figure 9, most models show a mediocre prediction from the ROC curve, and Random Forest Classifier offers the best forecast. Since some of the curves intersect, some models are better for certain cost ratios.

# 7 End Sections

## Appendices

Project code is uploaded on the Github link as shown under **7.2 Relevant Links**.

## 7.2 Relevant Links

Github: <https://github.com/Abhayakarmach/CMPE188_HeartFailure.git>

Youtube Video:

<https://www.youtube.com/watch?v=10xk1x7WEto&rel=0>

# Lesson Learnt & Conclusion & Future Work

We have learned to implement different classifier models, which include Logistic Regression, SVM, Kernel SVM, Naive Bayes, K-Nearest Neighbors, Decision Tree and Random Forest, to the 3 selected variables of the cardiovascular dataset - time, serum creatinine and ejection fraction, to determine for heart failure or no heart failure. We also learned how to analyze the results produced by these classifier models and also the performance of each classifier model.

Since the dataset’s outcome is either heart failure or no heart failure, it follows a binary classification, in which models like Logistic Regression, SVM, and Random Forest perform better (higher accuracy). This is also why the kernel trick is not useful in increasing SVM’s accuracy, but in contrast results in a lower accuracy. In conclusion, the Random Forest model shows the best performance, in terms of its learning curve and the ROC curve.

Future work for this project would be to investigate ways to increase the models’ accuracy while considering the variance and bias. Research should also be done to determine why fine tuning results in lower accuracy in the classifier models.

**References**

1. Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020).
2. Centers for Disease Control and Prevention. Underlying Cause of Death, 1999–2018. CDC WONDER Online Database. Atlanta, GA: Centers for Disease Control and Prevention; 2018.