# CS677 Final Project

## Dataset Description

**Link:** <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction/data>

### **Heart disease: A Global Threat**

Cardiovascular diseases (CVDs) reign as the leading cause of death worldwide. Staggeringly, an estimated 17.9 million people lose their lives to CVDs each year, representing a shocking 31% of all global deaths. Heart attacks and strokes are particularly deadly, causing four out of five CVD-related fatalities. Even more concerning is the significant number of these deaths occurring prematurely in people under 70. This grim reality underscores the critical importance of early detection and effective management of heart failure, a common outcome of CVDs.

### **A Powerful Tool for Research**

This dataset is a game-changer in heart disease research. It brings together information from five separate heart disease datasets – Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog – into a single, powerful resource. This collection has 918 observations, making it the largest heart disease dataset available for research to date. By merging these previously scattered datasets, researchers now have a goldmine of information for analyzing cardiovascular risk factors and developing disease prediction models. The dataset's diversity is another strength, providing a robust foundation for exploring correlations, identifying potential predictors, and building robust models for early detection and management of CVDs on a global scale.

## Heart Failure Disease

### Machine Learning Models for Predicting Heart Disease

#### Logistic Regression

##### Model Description

Logistic Regression is a supervised learning algorithm used for binary classification. It estimates the probability of a binary outcome based on one or more predictor variables. It's suitable for linearly separable data and provides probabilities that can be used to rank instances.

##### Model Accuracy

The model’s accuracy was 0.8391 or 83.91%

##### Confusion Matrix

A diagram of a logistic regression

Description automatically generated

#### Decision Trees

##### Model Description

##### Decision Trees are a supervised learning algorithm used for both classification and regression tasks. They build a tree-like model to make decisions by asking a series of questions based on feature values. They are easy to understand and visualize but can be prone to overfitting with complex trees.

##### Model Accuracy

The model’s accuracy was 0.7826 or 78.26%

##### Confusion Matrix

A chart of different colors

Description automatically generated with medium confidence

#### Random Forest

##### Model Description

##### Random Forest is an ensemble learning method based on decision trees. It builds multiple decision trees and combines their outputs to improve prediction accuracy and control overfitting. It creates a "forest" of decision trees by randomly selecting subsets of features and samples to build each tree and provides feature importance for better understanding of the data.

##### Model Accuracy

The model’s accuracy was 0.8652 or 86.52%

##### Confusion Matrix

A chart of different colors

Description automatically generated with medium confidence

#### Support Vector Machine (SVM)

##### Model Description

##### Support Vector Machine (SVM) is a supervised learning algorithm for both classification and regression tasks. It constructs a hyperplane in a high-dimensional space to separate classes, maximizing the margin between them. SVM can handle high-dimensional data and is effective in cases where the number of features is greater than the number of samples and can use different kernel functions to separate non-linearly separable data.

###### Linear Kernel

Model Accuracy

The model’s accuracy was 0.8435 or 84.35%

Confusion Matrix

A graph of different colors

Description automatically generated with medium confidence

###### Gaussian Kernel

Model Accuracy

The model’s accuracy was 0.8348 or 83.48%

Confusion Matrix

A chart with different colors

Description automatically generated with medium confidence

## Final ML Model Results Breakdown

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Model** | **TP** | **FP** | **TN** | **FN** | **Accuracy** | **TPR** | **TNR** |
| **Logistic Regression** | 101 | 20 | 92 | 17 | 0.8391 | 85.59% | 82.14% |
| **Decision Tree** | 96 | 28 | 84 | 22 | 0.7826 | 81.36% | 75.00% |
| **Random Forest** | 110 | 23 | 89 | 8 | 0.8652 | 93.22% | 79.46% |
| **Linear Kernel SVM** | 103 | 21 | 91 | 15 | 0.8435 | 87.29% | 81.25% |
| **Gaussian Kernel SVM** | 108 | 28 | 84 | 10 | 0.8348 | 91.53% | 75.00% |

### Results

The most accurate model turned out to be the Random Forest Classifier with an accuracy of 86.52% and the highest True Positive rate of 93.22%. For some reason, it did not perform as well when it came to detecting negative instances of heart disease, as the True Negative rate is 3rd best at 79.46%. Still, with some more fine-tuning, I believe it would be able to achieve 90+% accuracy.