Machine Learning Model to Predict Heart Failure Disease

Member 1	Sakshi Mhatre	sakshimm@vt.edu
Member 2	Sai Chandan Reddy Koncha	saichandanreddyk@vt.edu
Member 3	Aniruddha Hore	aniruddhah@vt.edu

Motivation

Currently heart failure is one of the most common causes of death worldwide. Although this is due to various reasons such as diabetes, coronary heart diseases, old age and others. It is extremely difficult for healthcare workers to predict heart failure or the fatal symptoms on time for the patients. But with the use of machine learning algorithms, this problem can be resolved. Machine learning can be helpful in this case, with their ability to analyse, predict the outcomes of diseases and these models can be used to predict heart failure in advance. The existing models like decision tree, Linear SVM and other models accuracy was recorded between 52% to 67.7% In this experiment, we aim to increase this accuracy rate for healthcare.

Methodology:

Initially data cleaning, preprocessing will be done to eradicate the unwanted features, or data points. By applying feature selection methods we can find the informative attributes of the dataset to help understand which features have a higher impression rate on the result. Further, applying different machine learning algorithms like Naive Bayes, Random Forest, Logistic regression, Decision tree and SVM to find out the best performing model on the dataset.First step is data processing where we look at the data for outliers and check for missing values so we can either drop them or fill using available methods then apply normalisation techniques and then use this cleaned data to train the models By using multiple models, evaluation and comparison would be efficient. This also leaves room to go back and tweak the model according to its counterparts.

Evaluation:

The dataset is the Heart Disease Dataset available on kaggle which has over 1000 observations and 14 attributes. These attributes range from gender, age to health diseases that may or may not affect the heart.The problem here is a binary classification and output of each model is either 1 or 0 class 1 means that patient has a chance of heart failure and class 0 means that patient has a healthy heart Evaluation would be done by comparing the

performance of multiple models with the help of a confusion matrix and calculating the accuracy, precision and recall. The model with the highest accuracy would be selected as the final one, which can further be used by healthcare professionals for their use.

Timeline:

- Data preprocessing : 24 Feb
- Feature Selection : 15 March
- Model Training: 1 week per model