**Cardiac Failure Project Report**

**1. Introduction**

Machine learning is widely used now a days in many business

applications like e commerce and many more. Prediction is one of

area where this machine learning used, our topic is about predic-

tion of cardiac disease by processing patient’s dataset and a data of

patients to whom we need to predict the chance of occurrence of a

heart disease.

Creating a model to assess the likelihood of a death by cardiac failure event.  
This can be used to help hospitals in assessing the severity of patients with cardiovascular diseases.

There are some factors that affects Death Event. This dataset contains person's information like age, sex, blood pressure, smoke, diabetes, ejection fraction, creatinine phosphokinase, serum creatinine, serum sodium, time and we have to predict their DEATH EVENT.

**2. Problem Definition and Algorithm**  
  
2.1 Task Definition

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies.

People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

2.2 Algorithm Definition

We are using Logistic Regression Algorithm and Extra Trees Classifier Algorithm from all the Regression algorithms.

We will train various Classification Models on the Training set & see which yields the highest accuracy. We will compare the accuracy of Logistic Regression, K-NN (k-Nearest Neighbours), SVM (Support Vector Machine), Naives Bayes Classifier, Decision Trees, Random Forest, and XGBoost.

**Features and Predictor:**

Our Predictor ( Positive or Negative diagnosis of Cardiac Diseases) is determined by 13 features (X):

1. age (#)

2. sex : 1= Male, 0= Female (Binary)

3. (cp)chest pain type (4 values -Ordinal):Value 1: typical angina ,Value 2: atypical angina, Value 3: non-anginal pain , Value 4: asymptomatic

4. (trstbps) resting blood pressure (#)

5. (chol) serum cholesterol in mg/dl (#)

6. (fbs)fasting blood sugar > 120 mg/dl(Binary)(1 = true; 0 = false)

7. (restecg) resting electrocardiography results(values 0,1,2)

8. (thalachh) maximum heart rate achieved (#)

9. (exng) exercise induced angina (binary) (1 = yes; 0 = no)

10. (oldpeak) = ST depression induced by exercise relative to rest (#)

11. (slp)slope of the peak exercise ST segment (Ordinal) (Value 1: up sloping , Value 2: flat , Value 3: down sloping )

12. (ca) number of major vessels (0–3, Ordinal) colored by fluoroscopy

13. (thal) maximum heart rate achieved — (Ordinal): 3 = normal; 6 = fixed defect; 7 = reversible defect  
  
**3. Experimental Evaluation**  
  
3.1 Methodology

It is implemented in Python and different classification algorithms are used. Below is a brief description of the general approach that I employed:

**Data cleaning and pre-processing:** Here I checked and dealt with missing and duplicate variables from the data set as these can grossly affect the performance of different machine learning algorithms (many algorithms do not tolerate missing data).

**Exploratory Data Analysis:** Here I wanted to gain important statistical insights from the data and the things that I checked for were the distributions of the different attributes, correlations of the attributes with each other and the target variable and I calculated important odds and proportions for the categorical attributes.

**Feature Selection:** Since having irrelevant features in a data set can decrease the accuracy of the models applied, I used the Boruta Feature Selection technique to select the most important features which were later used to build different models.

**Model development and comparison:** I used four classification models, i.e., Logistic Regression, K-Nearest Neighbors, Decision Trees and Support Vector Machine, After which I compared the performance of the models using their accuracy and F1 scores. I then settled with the best performing model.

3.2 Results



**4. Future Work**

In future the work can be enhanced by developing a web application based on the Random Forest algorithm as well as using a larger dataset as compared to the one used in this analysis which will help to provide better results and help health professionals in predicting the heart disease effectively and efficiently.

**5. Conclusion**  
With the increasing number of deaths due to heart diseases, it has become mandatory to develop a system to predict heart diseases effectively and accurately. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of Decision Tree, Logistic Regression, Random Forest and Naive Bayes algorithms for predicting heart disease using UCI machine learning repository dataset.  
**References**  
1.Sonam Nikhar, A.M. Karandikar “Prediction of Heart Disease Using Machine Learning Algorithms” in International Journal of Advanced Engineering, Management and Science (IJAEMS) June-2016 vol-2

2. <https://www.kaggle.com/andrewmvd/heart-failure-clinical-data>

By

Team : Tech Pros

Team Members:

1. Sai Kiran

2.Uma Saiteja

3.Jayendra

4.Shashank Mourya

5.Hemanth Kumar