**Project Report**

1. **Title of the Project: Heart Failure Prediction:**

**Brief on the project:** Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Four out of 5CVD deaths are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under 70 years of age. Heart failure is a common event caused by CVDs and this dataset contains 11 features that can be used to predict a possible heart disease. 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, hyperlipidemia or already established disease) need early detection and management wherein a machine learning (classification filter) model can be of great help.

1. **Deliverables of the project: To classify / predict whether a patient can encounter a death situation due to heart failure.**

It is a binary classification problem with multiple numerical and categorical features. Supervised machine learning algorithms will be used which demand some external assistance. The input dataset would be split into training and test dataset. The trained dataset composed of output variable which is to be predicted or classified. Each algorithm will get to know a specific pattern from the training dataset and apply them to the test dataset for prediction or classification purposes.

1. **Problem analysis:**

In the given study, we have a binary classification problem. We will make a prediction on the target variable heart disease.

Lastly, we will build a variety of Classification models and compare the models giving the best prediction on heart disease.

1. **Target Variable:**

Target variable, in the machine learning context, is the variable that is or should be the output. For example: it could be binary 0 or 1 if you are classifying or it could be a continuous variable if you are doing a regression. In statistics you also refer to it as the response variable. In our study our target variable is Heart Disease in the context of determining whether anybody is likely to get heart disease based on the input parameters like gender, age and various test results or not.

Below are the ways to analyze and predict the disease:

* Data Collection and Preprocessing
* Attribute description:

1. **Attribute Information**
   * 1. Age: age of the patient [years]
     2. Sex: sex of the patient [M: Male, F: Female]
     3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
     4. RestingBP: resting blood pressure [mm Hg]
     5. Cholesterol: serum cholesterol [mm/dl]
     6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
     7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
     8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
     9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
     10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
     11. ST\_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
     12. HeartDisease: output class [1: heart disease, 0: Normal]

**Preprocessing of the Dataset :** EDA has been performed on the dataset and it does not have any null values and duplicated data. But outliers were present in the dataset and it has been treated properly and also for standardization standard scaler is used. Data Visualization is also done on the dataset.

1. **Resources:**

* **Data set source: The data set source is from the below location:** <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>
* **Software: Software used for analyzing the problem is Python with Scikit-Learn:** Python is a popular programming language for machine learning.

Scikit-Learn is used machine learning library that includes various classification algorithms. Matplotlib, seaborn and sweetviz is used for data visualization.

1. **Individual Details:**

Done by Name: Sarmistha Chakraborty

**Emailid:** [**schakraborty87@gmail.com**](mailto:schakraborty87@gmail.com)

Phone Number: 958665987

1. **Model Selection:**

For solving the above problem, I have selected logistic regression, decision tree, random forest, Gradient boosting, ada boost. And highest prediction is given by gradient boosting model.

Logistic Regression: 85.86, with hyper parameter tuning precision is coming as .92

DecisionTree:84.78

RandomForestClassifier: 88.76

GradientBoostingClassifier: **89.13**

AdaBoostClassifier: 86.59

1. **Conclusion:**

* Cardiovascular diseases are the top one killer for many years. I think that the reasons are our lack of knowledge about heart disease and the life habits. According to the model and features analysis, we know which features that we can do regular self-examinations.
* The most obvious symptom is chest pain. There are three types of chest pain, but only atypical angina is strongly related to the heart disease. No matter which type of chest pain you have, go to the doctor.
* In addition, everyone should always keep an eye on the resting blood pressure. The ideal resting blood pressure is lower than 120mmHg, but if your blood pressure is much lower than the 120mmHg, it means that you are under high risk of heart disease. Besides, the problem will not be only heart when the blood pressure is higher than 150mmHg.
* Lots and lots of electronic devices that can measure heart rate, so it's easier to monitor your own. Record max heart rate to make sure that your heart is still healthy. Once the rate rises year by year, there must be something wrong with you.
* No matter how healthy we are, we must do annual examination because another features cannot be taken care of by ourselves. Finally, don't forget the older we are, the higher the risks are.