Project Status Report

Problem Definition:

- Explore the heart disease dataset using exploratory data analysis (EDA)
- Exercise with classification algorithms for prediction (modelling)

Dataset Description:

- **Age** - Age of the patient
- **Sex** - Sex of the patient
- **CP** - Chest pain type ~ 0 = Typical Angina, 1 = Atypical Angina, 2 = Non-anginal Pain, 3 = Asymptomatic
- **TRTBPS** - Resting blood pressure (in mm Hg)
- **Chol** - Cholestoral in mg/dl fetched via BMI sensor
- **Fbs** - (fasting blood sugar > 120 mg/dl) ~ 1 = True, 0 = False
- **Restecg** - Resting electrocardiographic results ~ 0 = Normal, 1 = ST-T wave normality, 2 = Left ventricular hypertrophy
- **Thalachh** - Maximum heart rate achieved
- **Oldpeak** - Previous peak
- **Slp** - Slope
- **Caa** - Number of major vessels
- **Thall** - Thalium Stress Test result ~ (0,3)
- **Exng** - Exercise induced angina ~ 1 = Yes, 0 = No
- **Output** - Target variable
    - 0 = no disease,
    - 1 = disease

Diagnosis of heart disease (angiographic disease status)

- Value 0: < 50% diameter narrowing
- Value 1: > 50% diameter narrowing

Exploratory Data Analysis (EDA):

In the EDA stage, I got a sense of the data distribution and examined the basic statistical properties of the data. This included understanding the types of variables, checking for missing values, and visualizing the distribution of various features and the target variable. This initial step is crucial in any data science project and helped me identify trends, anomalies, patterns, and relationships within the data.

The Original problem was solved by KNeighbours and SVM modelling algorithms. I have solved them using Random Forest and Logistic regression.

My choice of machine learning model is Random Forest algorithm because it is a classifier that contains several decision trees on various subsets of a given dataset and takes the average to enhance the predicted accuracy of that dataset. Instead of relying on a single decision tree, the random forest collects the result from each tree and expects the final output based on the majority votes of predictions.

I am planning to use “f1-score” as metric of assessment to quantify the performance of my model. The F1 score is a metric used to measure the performance of classification machine learning models. It is a harmonic mean of the precision and recall, where precision is the accuracy of the positive predictions and recall is the fraction of positive cases that are correctly identified. The F1 score ranges from 0 to 1, where 1 is the best and 0 is the worst. The F1 score symmetrically represents both precision and recall in one metric.

Accuracy Scores:

	Model	Accuracy_score
1	Random Forest	0.901639
2	KNeighbours	0.901639
3	SVM	0.868852
0	Logistic Regression	0.836066

Conclusion :

After few hyperparameter tuning, Random Forest and KNeighbour Achieved the highest accuracy here.