## **Project Documentation: Cholesterol Impact on Heart Disease**

### **1. Introduction**

* Objective: To study the impact of cholesterol on heart disease using data analysis and machine learning.
* Scope: Perform data cleaning, exploratory data analysis (EDA), and train ML models to predict heart disease.

### **2. Dataset Overview**

* Source: <https://www.kaggle.com/datasets/rashadrmammadov/heart-disease-prediction>
* Dataset Title: Heart Disease Prediction
* Total records: **1000**
* Feature: **16**
* Records after cleaning**: 660**
* Split: **495 training rows** (75%) and **165 testing rows** (25%).
* Features: Age, Gender, Cholesterol, Blood Pressure, and other health-related attributes.
* Target variable: **Heart Disease** (0 = No, 1 = Yes).

### **3. Data Cleaning**

Steps performed:

* Removed/handled missing values.
* Treated outliers in cholesterol values.
* Converted categorical columns (e.g., Gender: Male/Female) into numerical values using encoding.
* Applied **MinMaxScaler** to normalize numeric features between 0 and 1.

### **4. Exploratory Data Analysis (EDA)**

* **Distribution of cholesterol**: Checked histogram, KDE plots.
* **Group comparison**: Compared cholesterol levels for patients with and without heart disease.
* **Boxplots**: Visualized cholesterol differences across heart disease groups.
* **Correlation analysis**: Measured cholesterol’s correlation with heart disease and other features.

**Insight:** Patients with higher cholesterol levels tend to show higher chances of heart disease, though cholesterol alone is not the only risk factor.

### **5. Machine Learning Models**

Two models were trained:

1. **Decision Tree Classifier (Model 1)**
2. **Support Vector Classifier (Model 2)**

**Data Preparation for Models:**

* Features (X): All columns except heart\_disease.
* Target (y): heart\_disease.
* Training set: 495 rows.
* Testing set: 165 rows.

### **6. Model Evaluation**

* Evaluation metrics: Accuracy, Precision, Recall, F1-score.
* Confusion Matrix used to visualize performance.

**Result:**

* **Decision Tree (Model 1)** outperformed the Support Vector Classifier.
* Indicates that Decision Tree is more effective in capturing the relationship between cholesterol (and other health features) and heart disease.

### **7. Conclusion**

* Cholesterol is an important factor influencing heart disease, but it must be analyzed along with other health indicators for better predictions.
* The **Decision Tree model** provides better interpretability and performance for this dataset.
* Future work: Try ensemble methods (Random Forest, XGBoost) and statistical tests to further validate cholesterol’s impact.