Project 1

1. Context

Cardiovascular diseases (CVDs) are the leading cause of death globally, accounting for an estimated 17.9 million deaths annually, which represents 31% of all global deaths. Four out of five CVD-related deaths result from heart attacks and strokes, with one-third of these deaths occurring prematurely in individuals under 70 years of age.

Early detection and management of cardiovascular risk factors such as hypertension, diabetes, and hyperlipidemia can significantly reduce mortality rates. Machine learning (ML) models can play a crucial role in predicting heart disease at an early stage, assisting medical professionals in making informed decisions.

2. Problem Statement

The objective of this project is to develop a machine learning model that accurately predicts the likelihood of heart disease based on patient attributes. This model aims to assist healthcare providers in early diagnosis and risk assessment, ultimately improving patient outcomes.

3. Dataset Description

The dataset consists of 11 key features that contribute to heart disease prediction. The attributes are described below:

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]

Why is This Important in ML Models?

- **TA & ATA** → Higher risk of heart disease.
- **NAP** → Low correlation with heart disease.
- **ASY** → Dangerous because of silent symptoms, requiring special attention in predictions.

4. Project Workflow

The project is structured into the following key steps:

4.1 Data Understanding & Preprocessing

- Handling missing values (if any)
- Encoding categorical variables
- Scaling numerical features
- Detecting and handling outliers

4.2 Exploratory Data Analysis (EDA)

- **Univariate Analysis:** Distribution of individual features
- **Bivariate Analysis:** Relationships between features and target variable
- **Correlation Analysis:** Identifying feature importance
- **Visualization:** Histograms, box plots, heatmaps, and pair plots

4.3 Feature Engineering

- Creating new features (if applicable)
- Handling class imbalances (SMOTE, undersampling, or oversampling)

4.4 Model Selection

Evaluating multiple machine learning models, including:

- Logistic Regression
- Decision Tree
- Random Forest
- XGBoost
- Support Vector Machine (SVM)

4.5 Model Training & Evaluation

- Splitting data into training and testing sets
- Using cross-validation to improve performance
- Model evaluation using:
    - Accuracy

    - Precision, Recall, F1-score
    - AUC-ROC Curve

4.6 Model Deployment (Optional Future Enhancement)

- Deploying the model using Flask or FastAPI
- Creating a simple web application for real-time predictions

5. Future Scope

Beyond just building a predictive model, potential enhancements include:

- **Deploying the model** as a web app for real-time heart disease risk prediction.
- **Integrating deep learning models** (e.g., Artificial Neural Networks) for improved accuracy.
- **Developing an AI-powered chatbot** for quick health assessments based on symptoms.
- **Incorporating real-world patient data** from healthcare providers for model refinement.

6. Conclusion

This project aims to leverage machine learning techniques to predict heart disease, thereby facilitating early diagnosis and intervention. By systematically processing and analyzing patient health data, the model can assist healthcare professionals in making data-driven decisions, ultimately improving patient care and reducing mortality rates.