Heart Disease Prediction: Research Plan

1. Introduction

Heart disease is one of the leading causes of mortality worldwide, and early detection plays a crucial role in improving patient outcomes. By analyzing medical data, we can identify patterns and risk factors that contribute to heart disease. This project aims to explore whether we can predict the likelihood of heart disease based on various demographic and medical attributes using data from the UCI Heart Disease dataset. The goal is to leverage data analytics and machine learning to contribute to the understanding and early detection of heart disease.

2. Research Question

Can we predict the likelihood of heart disease based on patient demographic and medical data using the UCI Heart Disease dataset?

This research question will help identify the most relevant features that contribute to heart disease and explore how accurately we can predict its occurrence. The insights gained from this analysis could assist healthcare professionals in assessing heart disease risk in patients and making more informed decisions.

3. Data Source

For this analysis, we will use the **Heart Disease dataset** from the UCI Machine Learning Repository, accessible at UCI Heart Disease Dataset. The dataset contains 303 observations, each representing a patient, with 14 attributes, including:

- **Age**: Age in years
- **Sex**: Gender (1 = male, 0 = female)
- **Chest Pain Type**: Four types of chest pain (e.g., typical angina, asymptomatic)
- **Resting Blood Pressure**: Measured in mm Hg
- **Cholesterol**: Serum cholesterol level in mg/dl
- **Fasting Blood Sugar**: Whether fasting blood sugar is > 120 mg/dl
- **Resting ECG**: Resting electrocardiographic results
- **Max Heart Rate Achieved**: Maximum heart rate during exercise
- **Exercise Induced Angina**: Whether angina occurred during exercise
- **Oldpeak**: ST depression induced by exercise relative to rest
- **Thalassemia**: Three levels of defect (normal, fixed defect, reversible defect)

The target variable indicates the presence or absence of heart disease, which allows us to develop a classification model to predict this condition.

4. Proposed Methodology

Data Preparation

- Load the dataset into a Pandas DataFrame.
- Handle missing values, scale numerical features, and encode categorical variables (e.g., sex and chest pain type).
- Split the dataset into features (X) and target (y) and further into training and test sets.

Exploratory Data Analysis (EDA)

- Visualize the distribution of key features like age, cholesterol, and blood pressure.
- Explore correlations between features to identify potential relationships.
- Detect and handle any outliers that could affect model performance.

Modeling and Evaluation

- Train different machine learning models (e.g., logistic regression, decision tree, random forest) using `scikit-learn`.
- Evaluate model performance using metrics such as accuracy, precision, recall, F1 score, and ROC-AUC curve.
- Visualize model performance with plots like confusion matrix and ROC curve.

5. Final Deliverable

The final deliverable will include:

- **Jupyter Notebook or Python Script**: This will include all steps of the analysis, from data loading and preprocessing to modeling and evaluation.
- **Visualizations**: Key plots that illustrate insights from the data, such as distributions, feature importance, and model evaluation metrics.
- **README.md**: A document explaining the project objectives, the methodology followed, key findings, and how to execute the code to replicate the analysis.
- **Zipped Archive**: All files will be organized and zipped as `onyen_milestone_02.zip`, containing the notebook, visualizations, and README file.

The final report will summarize findings, such as which factors have the greatest impact on predicting heart disease and how accurately we can identify high-risk patients. This work can serve as a foundation for more advanced analyses or as a proof of concept for a potential healthcare application.