Heart Disease Prediction - Project Documentation

# 1. Project Overview

This project builds a machine learning model to predict whether a person is likely to have heart disease, based on various medical and demographic features. It uses supervised learning (classification), specifically logistic regression, and includes steps like exploratory data analysis (EDA), preprocessing, and model evaluation.

# 2. Dataset Description

\*\*File\*\*: heart.csv

The dataset contains health-related information for individuals. Common features include:

- age – Age of the individual

- sex – Gender (1 = male, 0 = female)

- cp – Chest pain type (categorical: 0 to 3)

- trestbps – Resting blood pressure

- chol – Serum cholesterol (mg/dl)

- fbs – Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)

- restecg – Resting electrocardiographic results

- thalach – Maximum heart rate achieved

- exang – Exercise-induced angina (1 = yes, 0 = no)

- oldpeak – ST depression induced by exercise

- slope – Slope of peak exercise ST segment

- ca – Number of major vessels colored by fluoroscopy

- thal – Thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect)

- target – Presence of heart disease (1 = yes, 0 = no)

# 3. Dependencies

```python

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model\_selection import train\_test\_split

from sklearn.preprocessing import OneHotEncoder, StandardScaler, OrdinalEncoder

from sklearn.compose import ColumnTransformer

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import (

accuracy\_score, confusion\_matrix, precision\_score, recall\_score, f1\_score

)

```

# 4. Workflow Summary

\*\*Step 1: Data Loading & Exploration\*\*

- Loaded dataset using `pandas.read\_csv`.

- Inspected structure with `df.info()` and checked data shape and sample rows.

- Checked for missing values and data types.

\*\*Step 2: Exploratory Data Analysis (EDA)\*\*

- Used seaborn and matplotlib to visualize distributions, correlations, and patterns between features and the target.

- Insights likely include:

- Correlation between chest pain types or cholesterol and heart disease

- Differences in maximum heart rate among patients with/without disease

\*\*Step 3: Preprocessing\*\*

- Handled categorical columns using `OneHotEncoder` and `OrdinalEncoder`.

- Standardized numerical columns using `StandardScaler`.

- Applied transformations using `ColumnTransformer`.

\*\*Step 4: Model Training\*\*

- Used `LogisticRegression` for binary classification.

- Data split into training and test sets using `train\_test\_split`.

\*\*Step 5: Evaluation\*\*

- Evaluated performance using:

- Accuracy

- Precision

- Recall

- F1-Score

- Confusion Matrix

# 5. Conclusion

- A basic logistic regression model was implemented.

- Metrics suggest how well the model can predict heart disease from the given features.

- Further improvements could include:

- Trying other classifiers (Random Forest, SVM, XGBoost)

- Cross-validation for robust performance

- Hyperparameter tuning

- Feature engineering