Heart Disease Prediction - Intuition Document

Overview

This project aims to predict the likelihood of heart disease in patients based on various medical attributes using a K-Nearest Neighbors (KNN) classifier. This document provides an overview of the intuition behind the analysis, model training, and evaluation steps performed in the project.

Dataset

The dataset used in this project is from the UCI Machine Learning Repository, and it contains data on heart disease with the following attributes:

- **age**: Age of the patient
- **sex**: Gender of the patient (1 = male, 0 = female)
- **cp**: Chest pain type (4 values)
- **trestbps**: Resting blood pressure (in mm Hg)
- **chol**: Serum cholesterol level (in mg/dl)
- **fbs**: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
- **restecg**: Resting electrocardiographic results (values 0,1,2)
- **thalach**: Maximum heart rate achieved
- **exang**: Exercise-induced angina (1 = yes, 0 = no)
- **oldpeak**: ST depression induced by exercise relative to rest
- **slope**: The slope of the peak exercise ST segment
- **ca**: Number of major vessels coloured by fluoroscopy (0-3)
- **thal**: Thalassemia (0 = normal; 1 = fixed defect; 2 = reversible defect)
- **target**: Presence of heart disease (1 = disease, 0 = no disease)

Data Exploration and Visualization

1. **Data Exploration**:
   - The dataset was initially loaded and examined for basic information like the number of rows and columns, and the distribution of the target variable.
   - Missing values were checked and none were found in this dataset.
   - Descriptive statistics were computed to understand the general characteristics of the data.
2. **Correlation Matrix**:
   - A correlation matrix was plotted to visualize the relationships between different features. The heatmap helps in identifying highly correlated features, which can be useful for feature selection or engineering.

Data Preprocessing

1. **Splitting the Data**:
   - The data was split into training and testing sets using an 80-20 split. This is crucial to ensure that the model is trained on one subset of the data and evaluated on another, unseen subset.

2. **Standardisation**:
    - Feature scaling was performed using StandardScaler. Standardisation (subtracting the mean and dividing by the standard deviation) is important because KNN is sensitive to the scale of the data.

Model Training

- **K-Nearest Neighbors (KNN)**:
    - A KNN classifier with 10 neighbours was used to train the model. KNN works by classifying data points based on the majority class among their nearest neighbours in the feature space.

Model Evaluation

1. **Training and Test Accuracy**:
    - Accuracy on both the training and test data was computed to evaluate model performance.
    - The training accuracy provides insight into how well the model fits the training data, while the test accuracy indicates how well the model generalises to new, unseen data.
2. **Classification Report**:
    - The classification report includes metrics like precision, recall, and F1-score for both classes (disease and no disease). These metrics provide a more comprehensive evaluation of model performance beyond just accuracy.

Predictive System

- **Prediction on New Data**:
    - A predictive system was built to classify a new input based on the trained model. Given a set of feature values, the model predicts whether the person has heart disease or not.

Example Input:

- For an input patient with the following attributes: age = 48, sex = 1 (male), chest pain type = 0, etc., the model predicts whether the person has heart disease.

Prediction Output:

- The system outputs whether the patient has heart disease based on the provided features.

Conclusion

The project demonstrates the end-to-end process of building a predictive model for heart disease using KNN. The model’s performance on both training and test data indicates how well it generalizes to new data. By including various metrics and visualization techniques, the project offers a comprehensive analysis of heart disease prediction based on the given dataset.

This intuition document provides a high-level understanding of the project's objectives, methodology, and results, making it easier for others to grasp the essence of the work done.