CMPUT 466 Course Project Report

Heart Disease Prediction using Machine Learning

Introduction

Heart disease is one of the leading causes of mortality worldwide. Early diagnosis can significantly improve treatment and patient survival. In this project, ML algorithms are applied to predict the presence of heart disease based on patient medical data. The goal is to identify which machine learning model can best classify patients as having or not having heart disease

Problem Formulation

- **Input**: Patient medical features such as age, cholesterol level, and maximum heart rate
- **Output**: A binary classification (1 = Heart Disease, 0= No heart Disease)
- **Dataset**:
    - The dataset was obtained from **Kaggle**
    - Number of samples: 1025 records
    - Number of features: 13 features
        - age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal
    - Target Variable: binary (0 or 1)

The Dataset was split into 70% **training** and 30% **testing**.

Approaches and Baselines

The following machine learning approaches were implemented and compared:

1. **Logistic Regression**
    a. **Hyperparameters**: **c** (inverse regularization strength) tuned using Grid Search with values $[0.01,\ 0.1,\ 1,\ 10]$.
    b. **Baseline Purpose**: Acts as a simple linear model.

2. **Support Vector Machine (SVM)**
    a. **Hyperparameters**:
        i. **c** values: $[0.1,\ 1,\ 10]$.
        ii. **kernel** types: **linear** and **rbf**.
    b. **Tuning Method**: Grid Search with 5-fold cross-validation.
3. **Random Forest Classifier**:
    a. **Hyperparameters**:
        i. **n-estimators** (number of trees): $[50,\ 100,\ 200]$.
        ii. **max_depth** (tree depth): $[None,\ 10,\ 20]$.

All models were optimized using **GridSearchCV** with 5-fold cross-validation on the training set.

Evaluation Metrics

The primary evaluation metric for this project is **accuracy**. Additional metrics such as **precision**, **recall**, and **F1-score** were used to gain a deeper knowledge of each model's performance.

- **Accuracy**: Measure overall correctness.
- **Precision**: Proportion of predicted positives that are actual positives.
- **Recall**: Proportion of actual positives correctly identified.
- **F1-score**: Harmonic mean of precision and recall.

These metrics are appropriate for a binary classification task where both false positives and false negatives are important.

Results

The table below summarizes the performance of each model:

Model	Accuracy	Precision 0/1	Recall 0/1	F1-score 0/1
Logistic Regression	0.81	0.86 / 0.76	0.75 / 0.87	0.80 / 0.81
Support Vector Machine	0.96	0.95 / 0.99	0.99 / 0.94	0.97 / 0.96
Random Forest Classifier	0.99	0.98 / 1.00	1.00 / 0.98	0.99 / 0.99

- **Logistic Regression**: Performed as a baseline model with reasonable accuracy, with balanced performance between classes but struggled with class imbalance compared to other models.
- **Support Vector Machine**: Significantly improved accuracy to 96%, with the RBF kernel and $c = 10$ yielding strong performance.
- **Random Forest**: Outperformed all other models, achieving an accuracy of 99%. This was likely due to its ability to handle complex feature interactions and ensemble learning.