Compare classifiers in scikit-learn library.

Introduction

This report analyzes the performance of six classifiers—Perceptron, Logistic Regression, Linear SVM, Non-Linear SVM (RBF), Decision Tree, and KNN—on two datasets: the Digits dataset (from scikit-learn) and the Heart Disease dataset (from Kaggle). The program is used in assessing each classifier's accuracy, training, and testing time. The influence of learning rate (for Perceptron), regularization strength (for Logistic Regression and SVM), maximum tree depth (for Decision Tree), number of neighbors (for KNN), and the importance of feature scaling are also investigated."

Methodology

Datasets:

- The Digits dataset contains 1,797 samples of 8x8 images (64 features) representing digits from 0 to 9.
- The Heart Disease dataset contains 919 samples with 30 features, including categorical variables converted into numerical representations using one-hot encoding.
- Both datasets are split into 80% training and 20% testing sets, with stratification to maintain class distribution.

Classifiers:

The classifiers were implemented in Python using the scikit-learn library and evaluated based on the following metrics.

Evaluation Metrics:

- **Accuracy**: Percentage of correctly classified instances on both training and testing sets.
- **Training and Testing Time**: Time taken to train and test each model.
- **Hyperparameter Tuning**: Grid search is used to find the best hyperparameters for each classifier.

Results and Analysis

On the digits dataset, KNN was the fastest to train (0.0000s), while Logistic Regression took the longest (0.1558s) . All classifiers had negligible testing times, except for Non-Linear SVM (RBF) (0.0256s) . All classifiers had high testing accuracy and Decision Tree had the lowest testing accuracy (83.06%).

On the heart disease dataset, all 3 classifiers achieved 100% accuracy on both training and testing sets and had negligible testing times, indicating the likelihood that the dataset is relatively easy to classify.

Performance on the Digits Dataset:

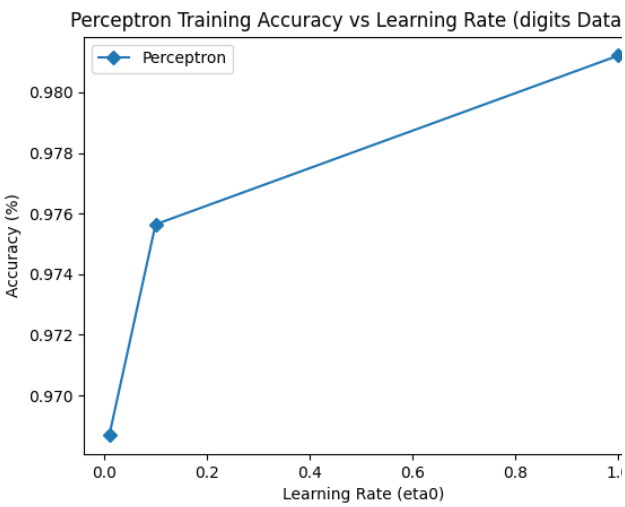
Classier	Training Accuracy	Testing Accuracy	Training Time(seconds)	Testing Time Time(seconds)	Best Hyperparameter
Perceptron	96.87%	93.61	0.0206	0.0000	eta0: 0.01
Logistic regression	97.49%	96.39%	0.1558	0.000	estimator__C: 0.1
Linear SVM	100%	98.06%	0.0492	0.0141	C : 1
Non-Linear SVM	100%	98.33%	0.0500	0.0256	C: 10, gamma: 0.01
Decision tree	97.63%	83.06%	0.0270	0.0000	max_depth: 10
KNN	98.61%	98.61	0.0000	0.0060	n_neighbors: 5

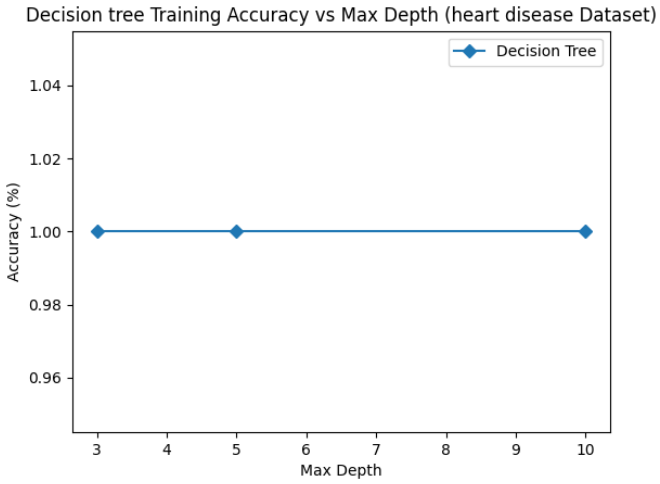
Performance on the Heart Disease Dataset:

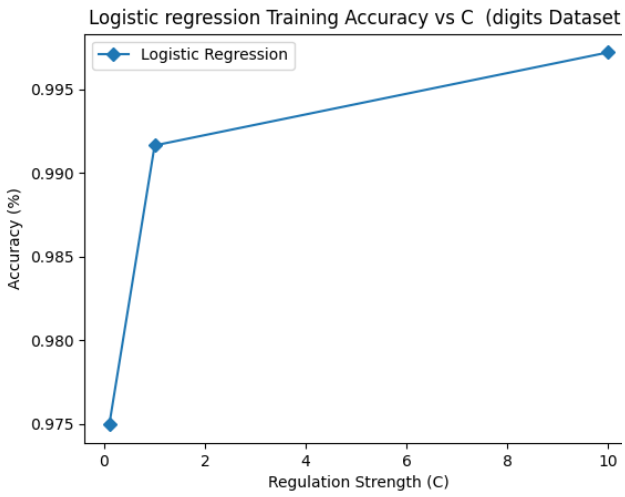
Classier	Training Accuracy	Testing Accuracy	Training Time(seconds)	Testing Time Time(seconds)	Best Hyperparameter
Logistic regression	100%	100%	0.0030	0.000	estimator__C: 1
Non-Linear SVM	100%	100%	0.0229	0.0000	C: 10, gamma: 0.01
Decision tree	100%	100%	0.0025	0.0000	max_depth: 3

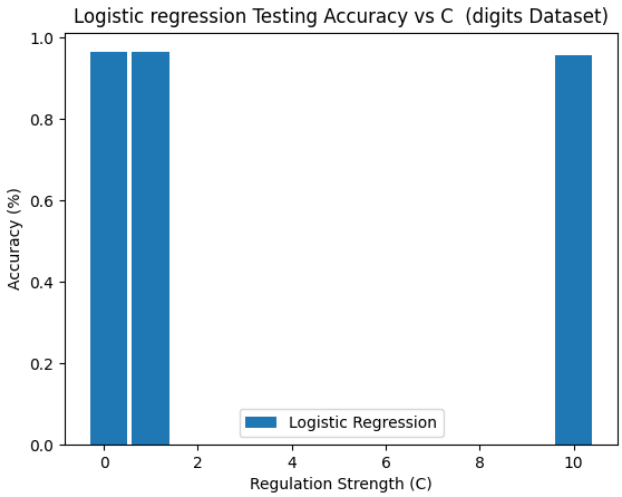
Plots

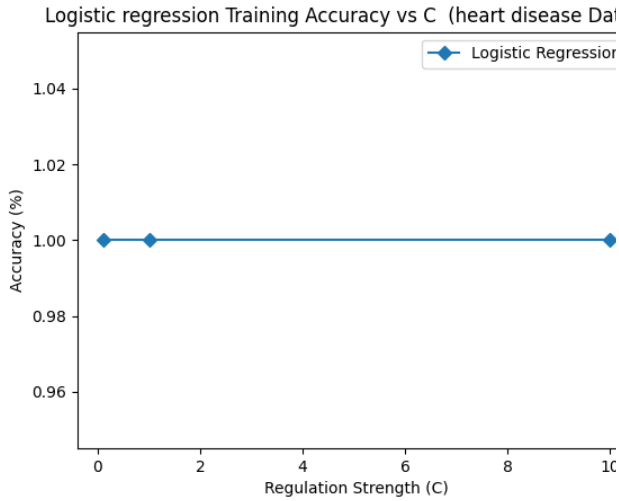
Below are a few of the plots that will be generated by the program.

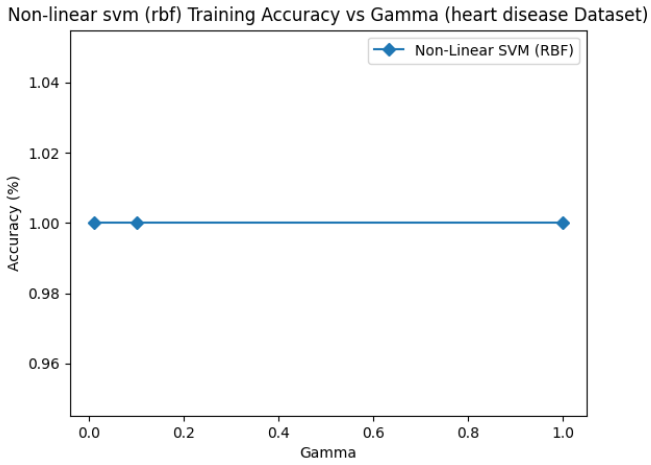


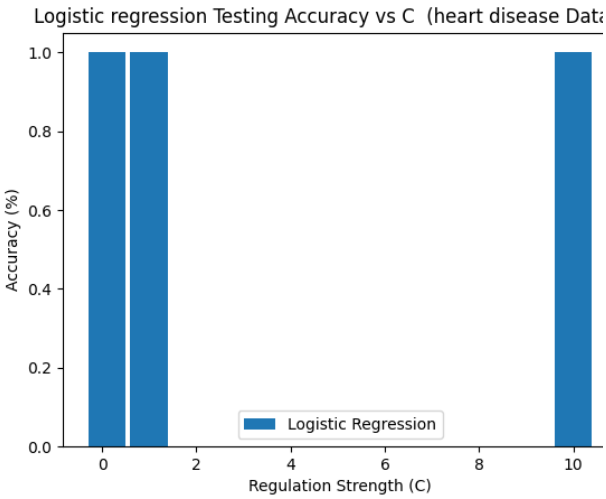


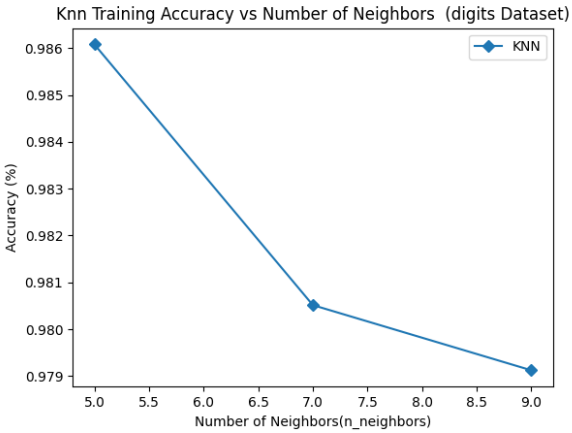


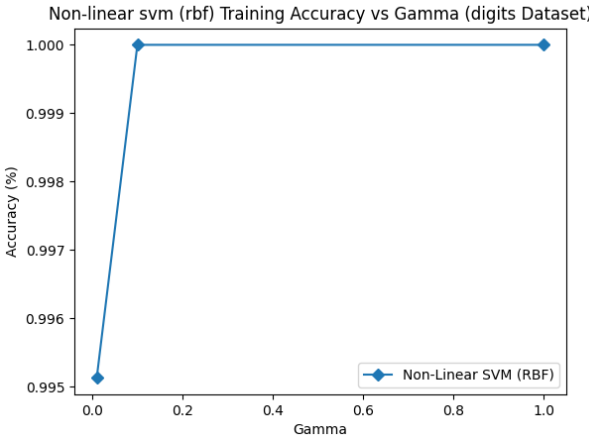


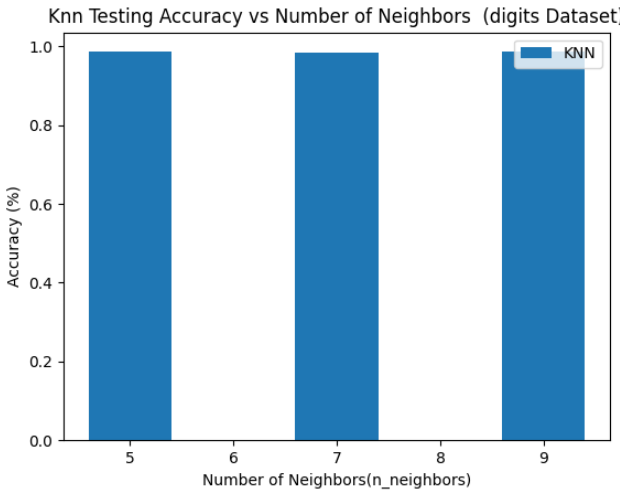

Logistic regression Training Accuracy vs C (heart disease Dat
Logistic Regressior
Accuracy (%)
1.04
1.02
1.00
0.98
0.96
0
2
4
6
8
10
Regulation Strength (C)


Non-linear svm (rbf) Training Accuracy vs Gamma (heart disease Dataset)
Non-Linear SVM (RBF)
Accuracy (%)
1.04
1.02
1.00
0.98
0.96
0.0
0.2
0.4
0.6
0.8
1.0
Gamma


Logistic regression Testing Accuracy vs C (heart disease Data
Accuracy (%)
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression
0
2
4
6
8
10
Regulation Strength (C)


Knn Training Accuracy vs Number of Neighbors (digits Dataset)
KNN
Accuracy (%)
0.986
0.985
0.984
0.983
0.982
0.981
0.980
0.979
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
Number of Neighbors(n_neighbors)


Non-linear svm (rbf) Training Accuracy vs Gamma (digits Dataset)
Accuracy (%)
1.000
0.999
0.998
0.997
0.996
0.995
Non-Linear SVM (RBF)
0.0
0.2
0.4
0.6
0.8
1.0
Gamma


Knn Testing Accuracy vs Number of Neighbors (digits Dataset)
KNN
Accuracy (%)
1.0
0.8
0.6
0.4
0.2
0.0
5
6
7
8
9
Number of Neighbors(n_neighbors)