CS 519 Applied Machine Learning I

Compare Classifiers in Scikit-learn Library

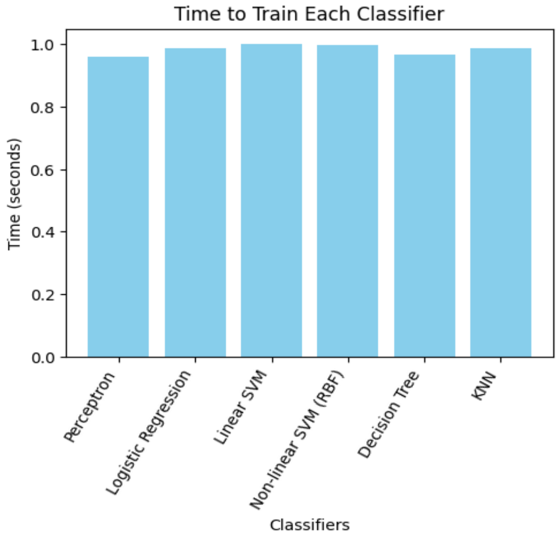
Submitted by: Indronil Bhattacharjee

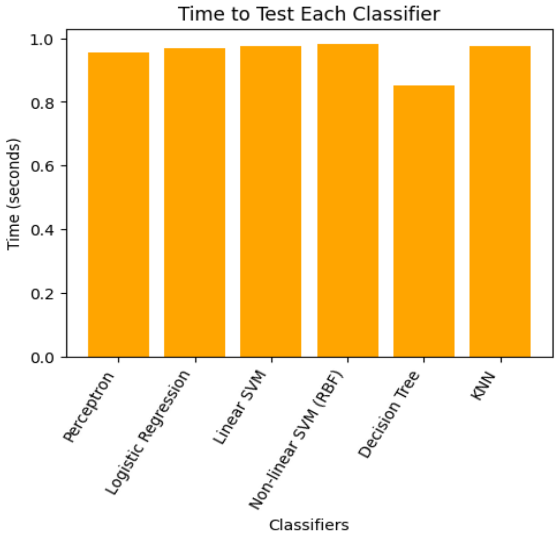
1.1 Dataset:

Digits Dataset from scikit-learn data.

1.2 Accuracy Comparison Results for Digits Dataset:

Classifiers	Set Parameter	Training Accuracy	Test Accuracy
Perceptron	η=0.01	0.96	0.96
Logistic Regression	C=0.1	0.99	0.97
Linear SVM	C=1.0	**1.0**	0.97
SVM with RBF Kernel	γ=0.01	**1.0**	**0.98**
Decision Tree	max depth=10	0.97	0.85
KNN	n_neighbors=5	0.99	0.97



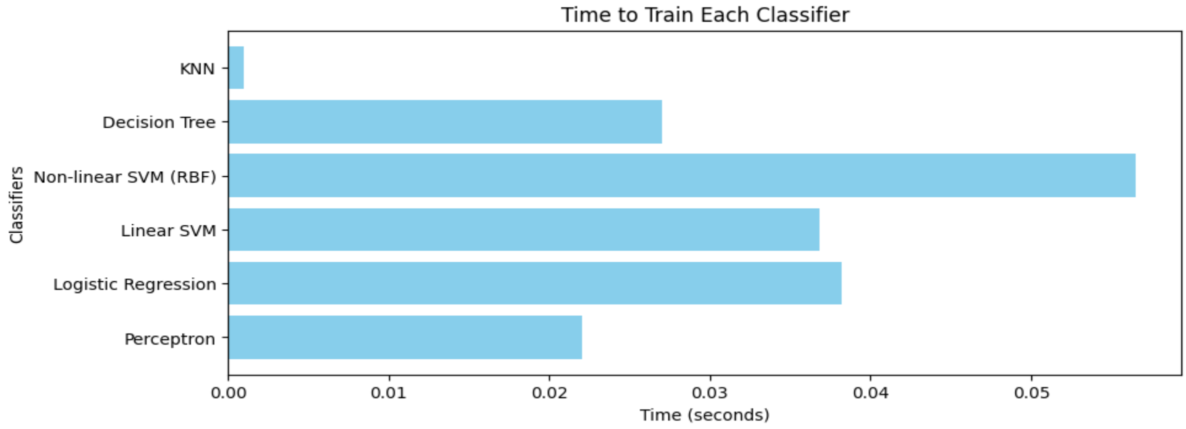


(a) Accuracy Analysis:

- All classifiers have high training accuracy, indicating that they are capable of capturing complex patterns in the training data.
- On the test data, Logistic Regression, Linear SVM, and SVM with RBF Kernel have relatively high accuracy, suggesting that these models generalize well to unseen data.
- Decision Tree has a lower test accuracy compared to other classifiers, which may indicate that it is more prone to overfitting or that the decision tree is not complex enough to capture the underlying patterns in the test data.
- Perceptron and KNN also have high test accuracy, indicating their effectiveness in generalizing to unseen data.

1.3 Training Time Comparison Results for Digits Dataset:

Classifiers	Set Parameter	Training Time
Perceptron	η=0.01	0.0220 Seconds
Logistic Regression	C=0.1	0.0382 Seconds
Linear SVM	C=1.0	0.0369 Seconds
SVM with RBF Kernel	γ=0.01	0.0565 Seconds
Decision Tree	max depth=10	0.0270 Seconds
KNN	n_neighbors=5	**0.10 Seconds**



(b) Training Time Analysis:

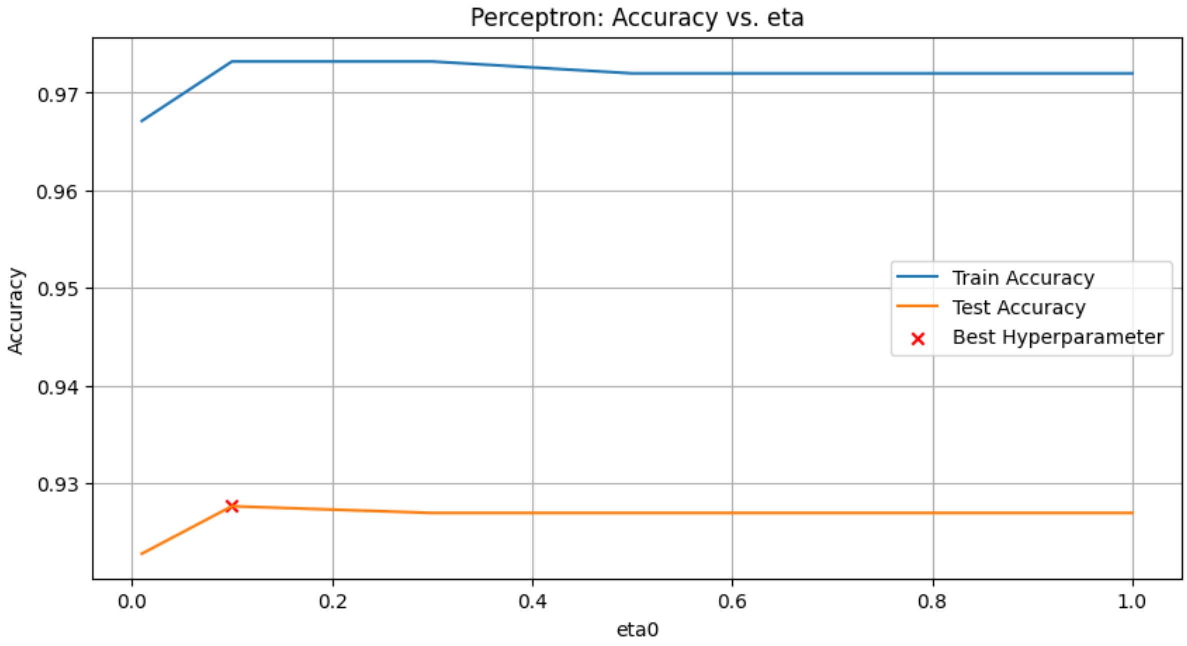
- KNN has the fastest training time, followed by Perceptron and Decision Tree.
- Logistic Regression, Linear SVM, and SVM with RBF Kernel have longer training times, with SVM with RBF Kernel being the slowest among them.

2 Parameter Tuning for Digits Dataset:

a) Perceptron:

The best learning rate (eta) for the Perceptron model was found to be 0.1 among experiment values 0.01, 0.1, 0.3, 0.5, 0.7, 1.0. This learning rate determines the step size for updating the weights during training. A learning rate that is too high may cause the weights to diverge, while a learning rate that is too low may result in slow convergence.

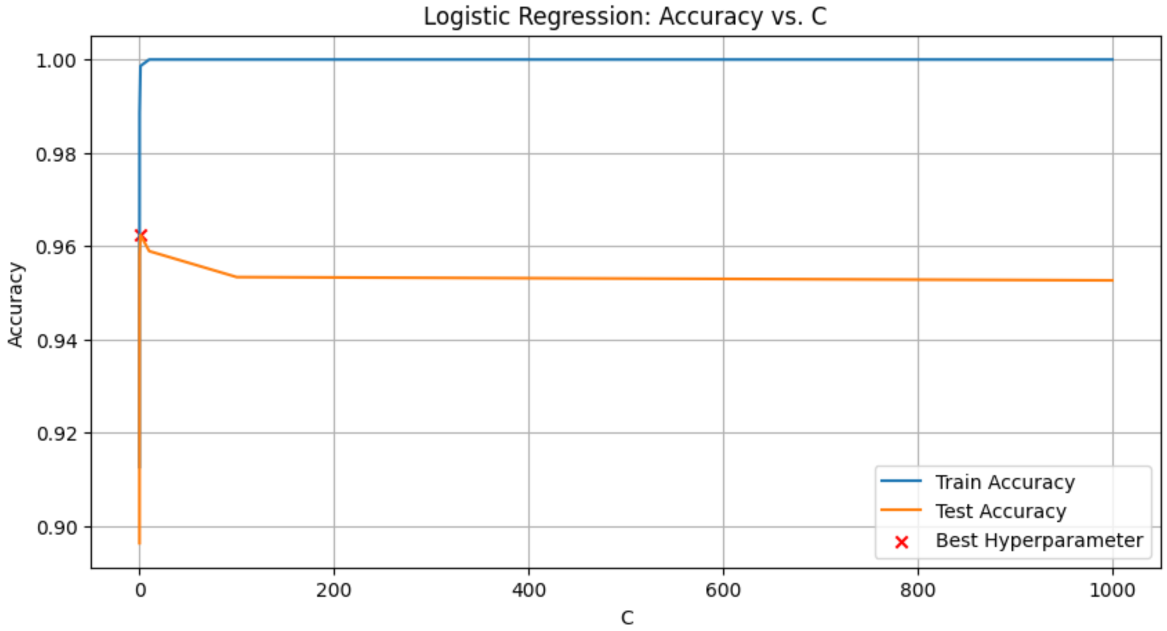
η	Test Accuracy
0.01	0.9228
0.1	**0.9276**
0.3	0.9269
0.5	0.9269
0.7	0.9269
1.0	0.9269



b) **Logistic Regression:**

The best regularization parameter (C) for the Logistic Regression model was found to be 1 among experiment values 0.001, 0.01, 0.1, 1, 10, 100, 1000. The regularization parameter controls the trade-off between fitting the training data well and keeping the model simple. A higher value of C allows the model to fit the training data more closely but may lead to overfitting.

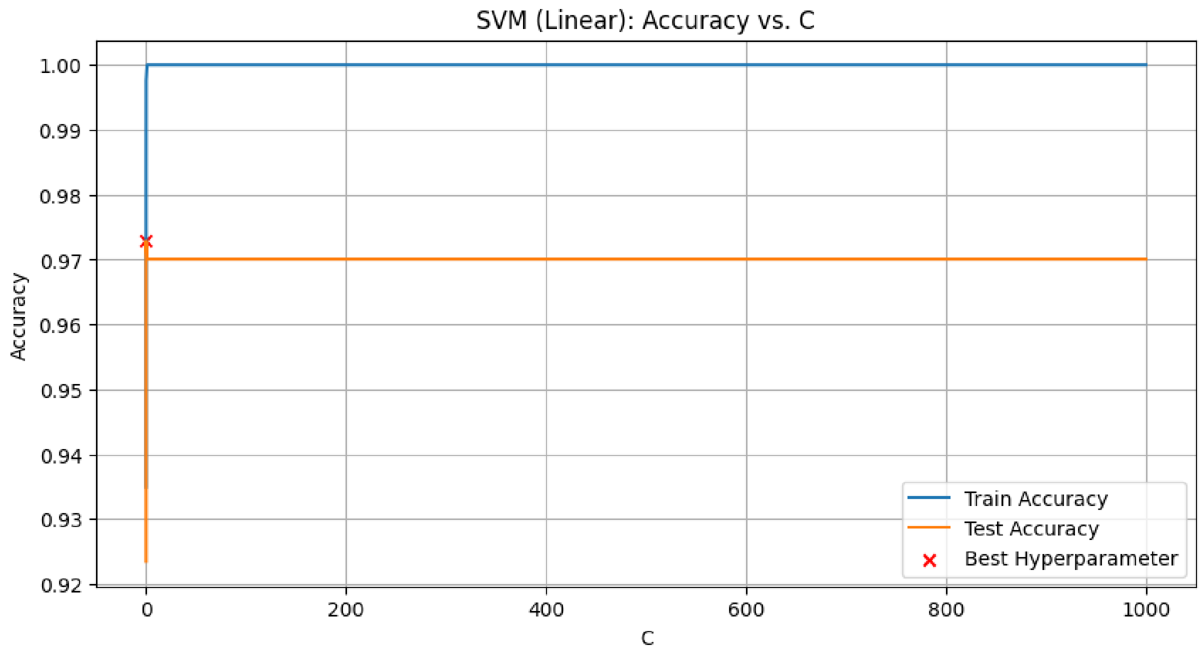
C	Test Accuracy
0.001	0.8963
0.01	0.9450
0.1	0.9582
1	**0.9624**
10	0.9589
100	0.9533
1000	0.9527



c) **Linear SVM:**

The best regularization parameter (C) for the Linear SVM model was found to be 0.01 among experiment values 0.001, 0.01, 0.1, 1, 10, 100, 1000. Similar to Logistic Regression, the regularization parameter in SVM controls the trade-off between fitting the training data well and keeping the model simple. A smaller value of C results in a stronger regularization, which helps to prevent overfitting.

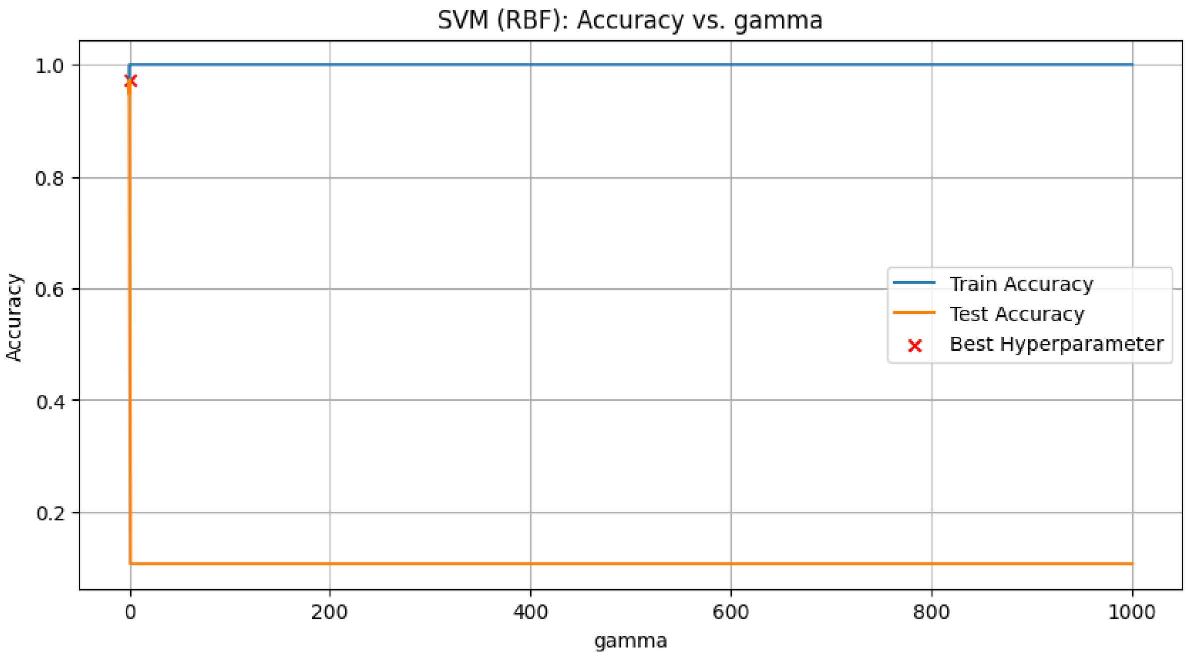
C	Test Accuracy
0.001	0.9235
0.01	**0.9729**
0.1	0.9721
1	0.9701
10	0.9701
100	0.9701
1000	0.9701



d) SVM with RBF Kernel:

The best gamma value (γ) for the SVM with RBF Kernel model was found to be 0.01 among experiment values 0.001, 0.01, 0.1, 1, 10, 100, 1000. The gamma parameter determines the width of the Gaussian kernel and plays a crucial role in the flexibility of the decision boundary. A smaller value of gamma results in a smoother decision boundary and less overfitting.

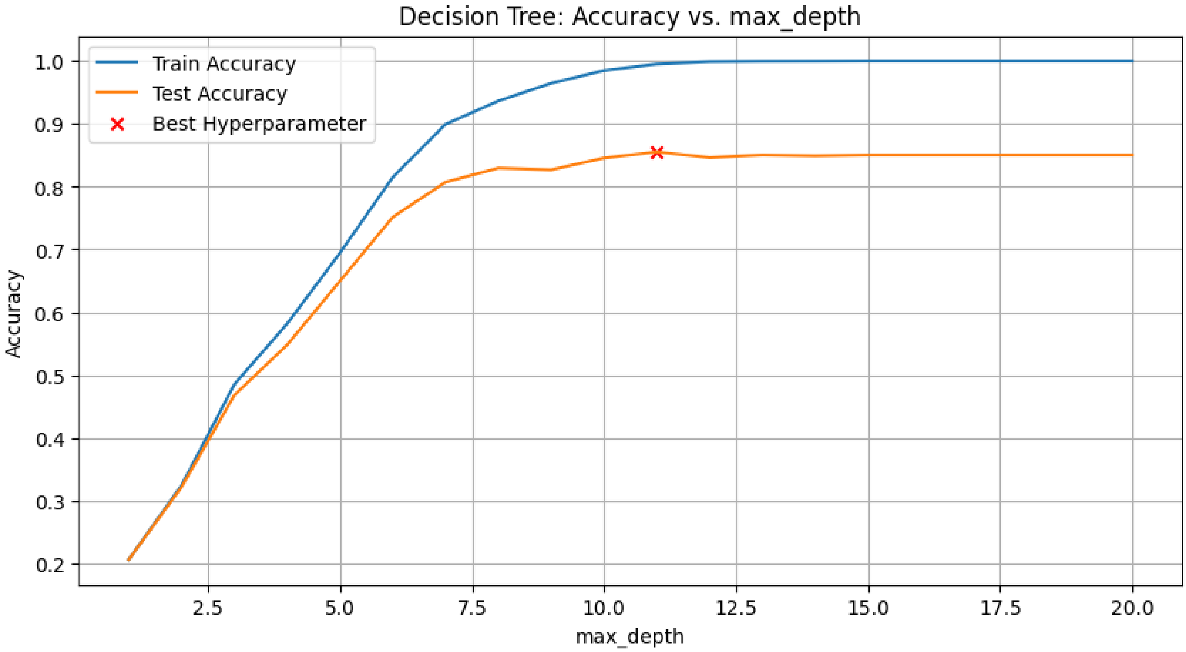
γ	Test Accuracy
0.001	0.9485
0.01	**0.9729**
0.1	0.9318
1	0.1072
10	0.1072
100	0.1072
1000	0.1072



e) **Decision Tree:**

The best maximum depth for the Decision Tree model was found to be 11 among experiment values starting from 1 to 20. The maximum depth parameter controls the maximum number of levels in the decision tree. A deeper tree can capture more complex patterns in the training data but may also lead to overfitting.

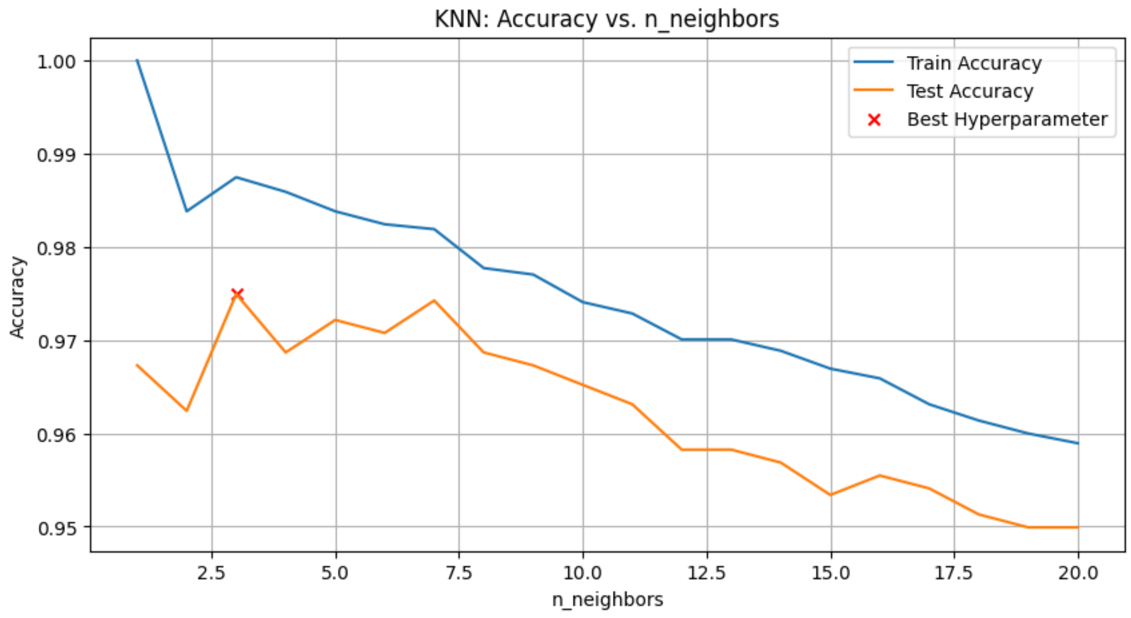
max_depth	Test Accuracy	max_depth	Test Accuracy
1	0.2067	**11**	**0.8552**
2	0.3222	12	0.8461
3	0.4683	13	0.8504
4	0.5483	14	0.8489
5	0.6499	15	0.8503
6	0.7515	16	0.8503
7	0.8072	17	0.8503
8	0.8295	18	0.8503
9	0.8267	19	0.8503
10	0.8455	20	0.8503



f) **KNN:**

The best number of neighbors (n_neighbors) for the KNN model was found to be 3. The number of neighbors parameter determines the number of nearest neighbors to consider when making predictions. A smaller value of k results in a more flexible decision boundary but may also increase the risk of overfitting.

n_neighbors	Test Accuracy	n_neighbors	Test Accuracy
1	0.9673	11	0.9631
2	0.9624	12	0.9582
3	**0.9745**	13	0.9582
4	0.9687	14	0.9568
5	0.9722	15	0.9534
6	0.9708	16	0.9555
7	0.9742	17	0.9541
8	0.9687	18	0.9513
9	0.9673	19	0.9499
10	0.9652	20	0.9499

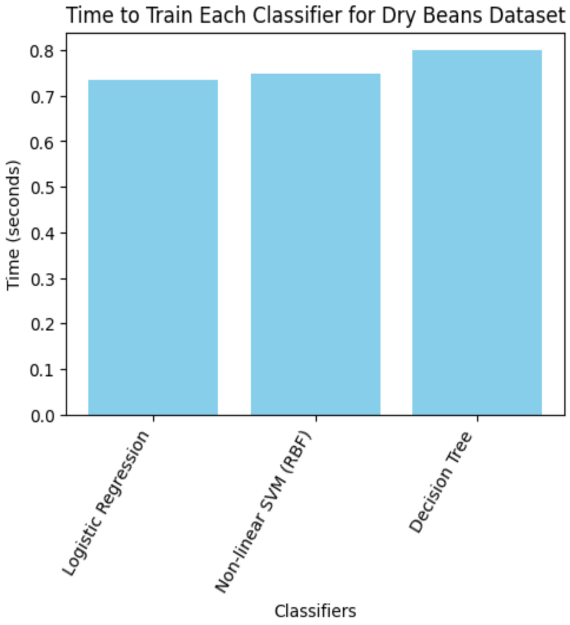


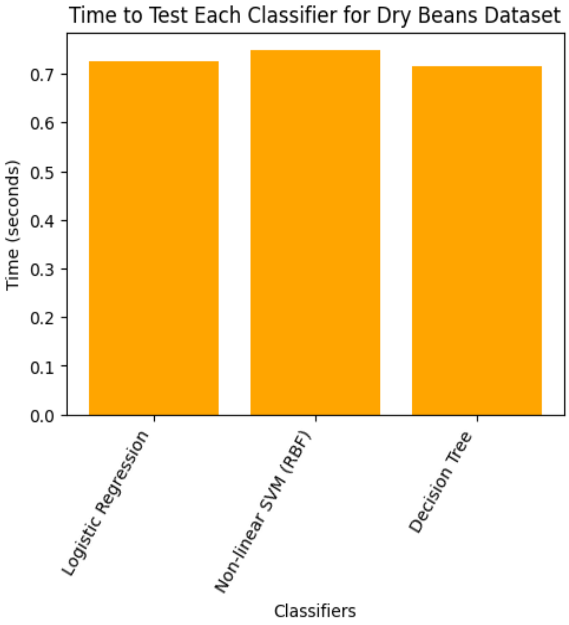
3.1 Second Dataset:

Dry Beans Dataset [1] from UCI Machine Learning repository. This data was collected to distinguish seven different registered varieties of dry beans with similar features in order to obtain uniform seed classification having 16 features.

3.2 Accuracy Comparison Results for Dry Beans Dataset:

Classifiers	Set Parameter	Training Accuracy	Test Accuracy
Logistic Regression	C=0.1	0.73	0.73
SVM with RBF Kernel	γ=0.01	0.75	**0.75**
Decision Tree	max depth=10	**0.80**	0.72



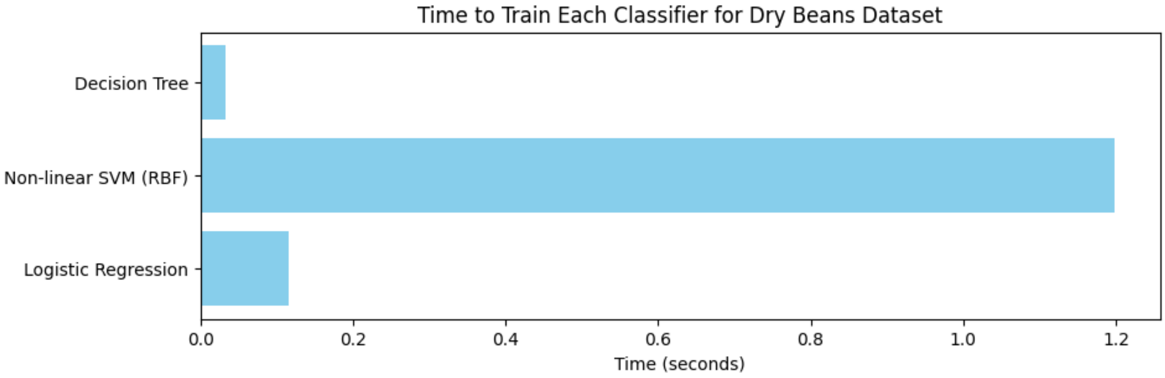


a) Accuracy analysis:

The accuracy comparison results for the Dry Beans dataset show that the SVM with RBF Kernel achieved the highest test accuracy (0.75) among the three classifiers, followed by Logistic Regression (0.73) and Decision Tree (0.72).

3.3 Training Time Comparison Results for Dry Beans Dataset:

Classifiers	Set Parameter	Training Time
Logistic Regression	C=0.1	1.1151 Seconds
SVM with RBF Kernel	γ=0.01	1.1981 Seconds
Decision Tree	max depth=10	0.335 Seconds



b) Training time analysis:

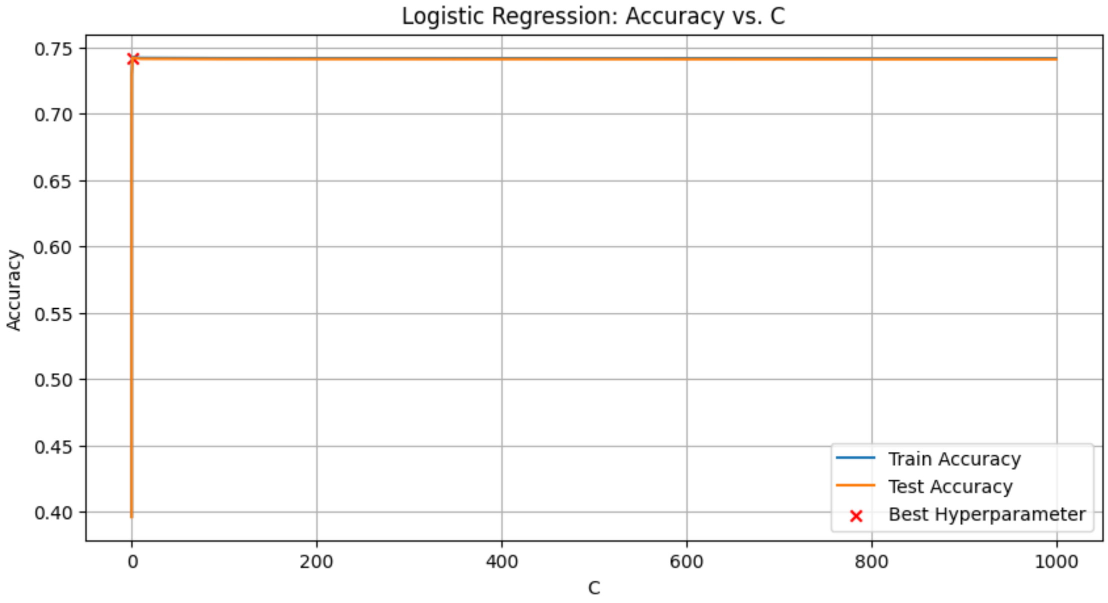
The training time comparison results for the Dry Beans dataset show that the Decision Tree classifier had the lowest training time (0.0335 seconds), followed by Logistic Regression (1.1151 seconds) and SVM with RBF Kernel (1.1981 seconds).

4 Parameter Tuning for Dry Beans Dataset:

a) Logistic Regression:

The logistic regression classifier with a C value of 1 achieved training and test accuracies of 0.7421 on the Dry Beans dataset among experiment values 0.001, 0.01, 0.1, 1, 10, 100, 1000. However, the training time for this classifier was relatively high, taking around 1.1151 seconds to train.

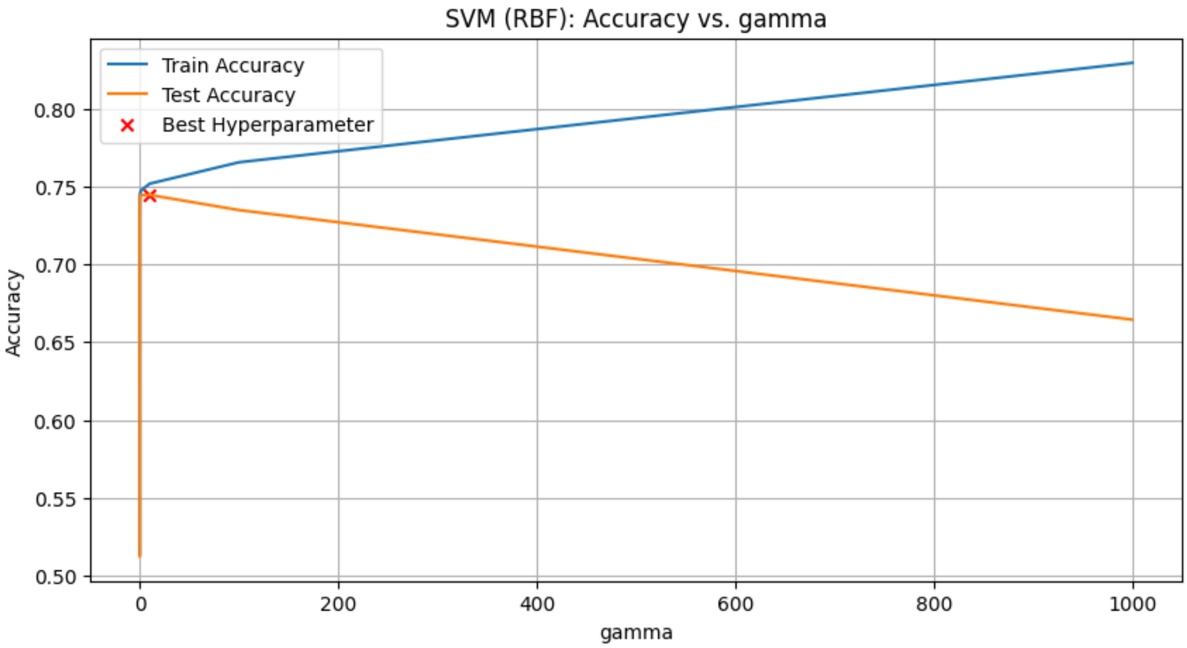
C	Test Accuracy
0.001	0.3959
0.01	0.6454
0.1	0.7288
1	**0.7421**
10	0.7413
100	0.7411
1000	0.7410



b) SVM with RBF Kernel:

The support vector machine (SVM) with the RBF kernel classifier achieved slightly higher training and test accuracies of 0.75 on the Dry Beans dataset. The best gamma value for the RBF kernel was found to be 10 among experiment values 0.001, 0.01, 0.1, 1, 10, 100, 1000. However, similar to logistic regression, the training time for this classifier was also relatively high, taking around 1.1981 seconds to train.

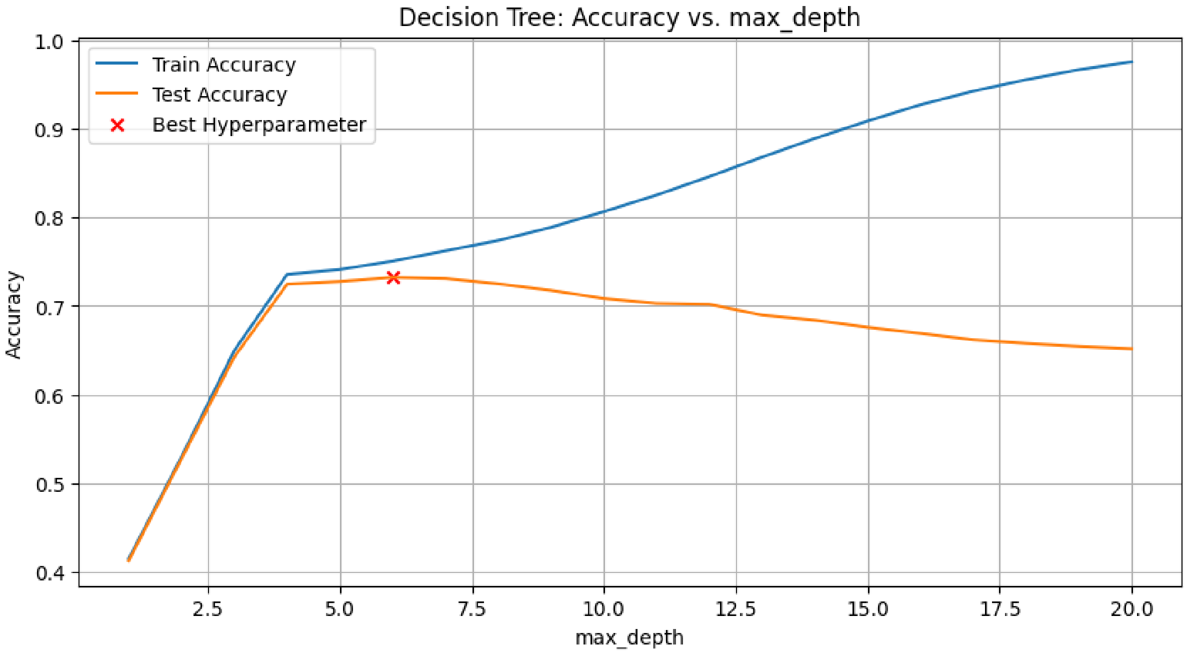
γ	Test Accuracy
0.001	0.5124
0.01	0.7345
0.1	0.7437
1	0.7445
10	**0.7517**
100	0.7350
1000	0.6645



c) Decision Tree:

The decision tree classifier achieved the highest training accuracy of 0.80 on the Dry Beans dataset. However, it had a slightly lower test accuracy of 0.73 compared to logistic regression and SVM with RBF kernel. The best max_depth value for the decision tree was found to be 6 among the experiment values from 1 to 20. Interestingly, the training time for the decision tree classifier was significantly lower than that of logistic regression and SVM with RBF kernel, taking only around 0.0335 seconds to train.

max_depth	Test Accuracy	max_depth	Test Accuracy
1	0.4125	11	0.7029
2	0.5275	12	0.7019
3	0.6424	13	0.6897
4	0.7246	14	0.6839
5	0.7276	15	0.6758
6	**0.7323**	16	0.6689
7	0.7309	17	0.6617
8	0.7249	18	0.6577
9	0.7176	19	0.6543
10	0.7084	20	0.6516



5 Conclusion:

Overall, while all models achieve high training accuracy, Logistic Regression, Linear SVM, and SVM with RBF Kernel demonstrate better generalization to unseen data on the digits dataset. However, Decision Tree shows relatively lower test accuracy, indicating a potential issue with overfitting or model complexity. Perceptron and KNN perform well in terms of accuracy, with KNN being the fastest to train. It's important to note that these observations are based on the digits dataset and may not generalize to other datasets.

While we shift to a relatively larger Dry Beans Dataset, the decision tree classifier performed the best in terms of training accuracy and had the lowest training time. However, it was found to be less effective in generalizing to unseen data compared to logistic regression and SVM with RBF kernel. These results suggest that the decision tree classifier may be overfitting to the training data. Further tuning of the decision tree hyperparameters and model regularization techniques may help improve its generalization performance.

Reference:

[1] Dry Bean Dataset [dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C50S4