Machine Learning Assignment-2

Documentation

3. Comprehensive Comparison

The following were the results obtained after running 7-fold cross validation on each of the ML models:

Fisher's Linear Discriminant

7-fold Cross Validation Results:

```
Test fold 1 : Accuracy = 0.9838709677419355
Test fold 2 : Accuracy = 0.9879032258064516
Test fold 3 : Accuracy = 0.9857038123167156
Test fold 4 : Accuracy = 0.9866202346041055
Test fold 5 : Accuracy = 0.9868035190615836
Test fold 6 : Accuracy = 0.9857038123167156
Test fold 7 : Accuracy = 0.9873533724340176
```

Mean of accuracies = 0.986279849183075

Std Dev. of accuracies = 0.001233128596493142

Linear Perceptron

7-fold Cross Validation Results:

```
Test fold 1 : Accuracy = 0.9272360703812317
Test fold 2 : Accuracy = 0.9627932551319648
Test fold 3 : Accuracy = 0.9831378299120235
Test fold 4 : Accuracy = 0.9767228739002932
Test fold 5 : Accuracy = 0.9657258064516129
Test fold 6 : Accuracy = 0.9525293255131965
Test fold 7 : Accuracy = 0.8629032258064516
```

Mean of accuracies = 0.9472926267281105

```
Std Dev. of accuracies = 0.03832628544134992
```

Naive Bayes

```
7-fold Cross Validation Results:

Test fold 1 : Accuracy = 0.9741568914956011
Test fold 2 : Accuracy = 0.9772727272727273
Test fold 3 : Accuracy = 0.9759897360703812
Test fold 4 : Accuracy = 0.9772727272727273
Test fold 5 : Accuracy = 0.9778225806451613
Test fold 6 : Accuracy = 0.9761730205278593
Test fold 7 : Accuracy = 0.9796554252199413

Mean of accuracies =  0.9769061583577712

Std Dev. of accuracies = 0.001585777537675608
```

Logistic Regression

```
7-fold Cross Validation Results:

Test fold 1 : Accuracy = 0.9869868035190615
Test fold 2 : Accuracy = 0.9886363636363636
Test fold 3 : Accuracy = 0.9884530791788856
Test fold 4 : Accuracy = 0.9897360703812317
Test fold 5 : Accuracy = 0.9869868035190615
Test fold 6 : Accuracy = 0.9869868035190615
Test fold 7 : Accuracy = 0.9884530791788856

Mean of accuracies =  0.9880341432760787

Std Dev. of accuracies = 0.00099290349946894
```

Artificial Neural Networks

```
7-fold Cross Validation Results:
```

Test fold 1 : Accuracy = 0.9803885630498533
Test fold 2 : Accuracy = 0.9886363636363636

Test fold 3 : Accuracy = 0.9884530791788856
Test fold 4 : Accuracy = 0.9897360703812317
Test fold 5 : Accuracy = 0.9869868035190615
Test fold 6 : Accuracy = 0.9869868035190615
Test fold 7 : Accuracy = 0.9884530791788856

Mean of accuracies = 0.9810169669040638

Std Dev. of accuracies = 0.0020756096755976607

Support Vector Machines

7-fold Cross Validation Results:

Test fold 1 : Accuracy = 0.9864369501466276
Test fold 2 : Accuracy = 0.9888196480938416
Test fold 3 : Accuracy = 0.9891862170087976
Test fold 4 : Accuracy = 0.9902859237536656
Test fold 5 : Accuracy = 0.9879032258064516
Test fold 6 : Accuracy = 0.9875366568914956
Test fold 7 : Accuracy = 0.9897360703812317

Mean of accuracies = 0.9885578131545874

Std Dev. of accuracies = 0.0012425440424910456

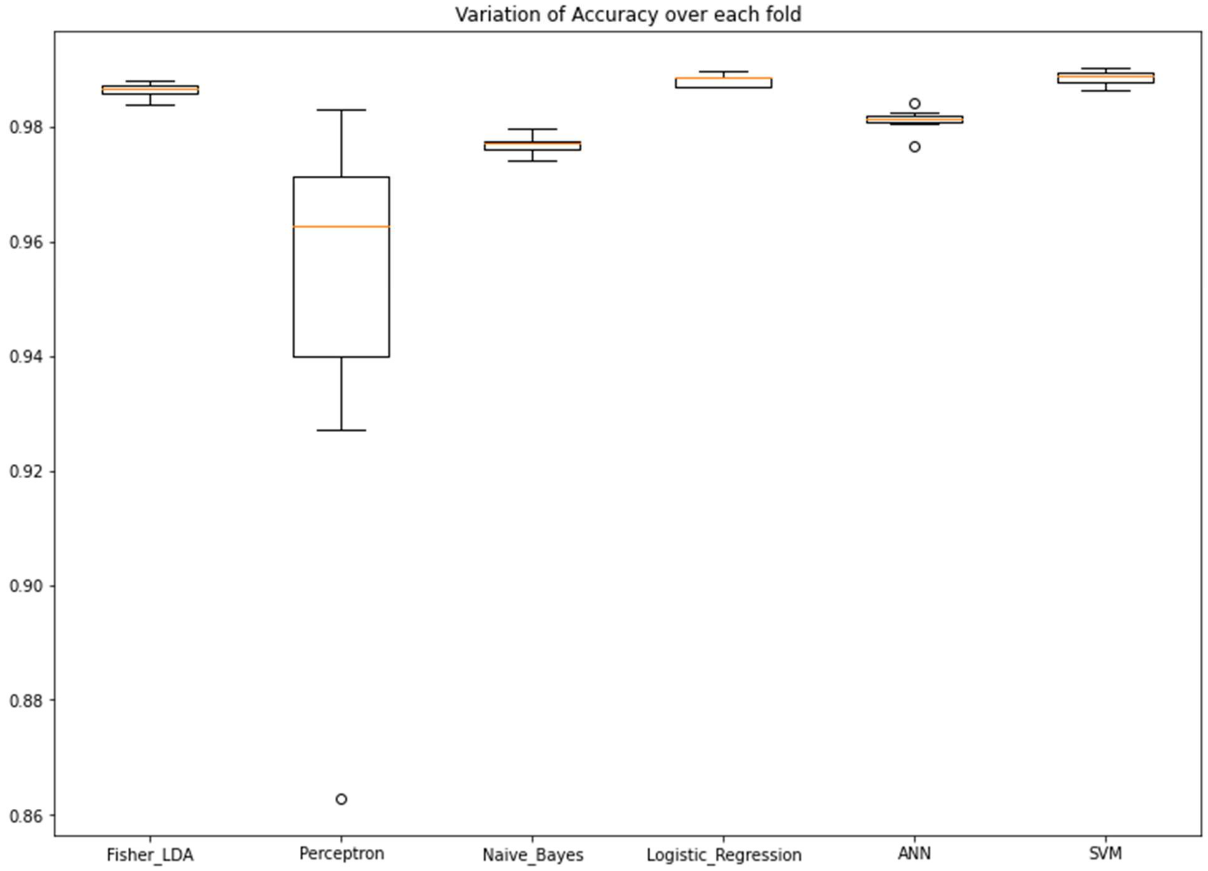
Analysis

It is observed that while all the models display quite high levels of mean accuracy, the variance is the lowest in Logistic Regression, followed by SVM and then Naive Bayes.

SVM works well with unstructured and semi-structured data like text and images while logistic regression works with already identified independent variables.

Thus, Logistic Regression could be said to be the better performer for the given data set. Possible reasons for this can be an absence of multicollinearity, or a very low value of multicollinearity, in the dataset, as logistic regression works very well in such conditions.

Box-Plots


Variation of Accuracy over each fold
Fisher_LDA
Perceptron
Naive_Bayes
Logistic_Regression
ANN
SVM