Pattern Recognition and Machine Learning

Lab - 8 Assignment Report

Aryan Himmatlal Prajapati (B21EE012)

Question 1.

Part 1

Preprocessed, cleaned and prepared the dataset. Separated features and labels as X and Y respectively.

Part 2

Created an object of SFS by embedding the Decision Tree classifier object, providing 10 features, forward as True, floating as False and scoring = accuracy. Trained SFS and report accuracy for all 10 features. Also, listed the names of the 10 best features selected by SFS.

- `Cross Validation Scores: [0.94971368 0.95033925 0.94899187 0.95178288 0.95076997]`

- `Accuracy for all 10 features: 0.9503195300512115`
- `Best 10 features selected by SFS: ('Customer Type', 'Type of Travel', 'Class', 'Inflight wifi service', 'Gate location', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'Baggage handling', 'Inflight service')`

Part 3

Using the forward and Floating parameter toggled between SFS(forward True, floating False), SBS(forward False, floating False), SFFS (forward True, floating True), SBFS (forward False, floating True),and chose cross validation = 4 for each configuration. Also, reported cv scores for each configuration.

SFS:

- Cross Validation Scores: [0.94833693 0.94953034 0.94929935 0.95072374]
- Accuracy for all 10 features: 0.9503195300512115
- Best 10 features selected by SFS: ('Customer Type', 'Type of Travel', 'Class', 'Inflight wifi service', 'Gate location', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'Baggage handling', 'Inflight service')

SBS:

- Cross Validation Scores: [0.94833693 0.94953034 0.94929935 0.95072374]
- Accuracy for all 10 features: 0.9503195300512115
- Best 10 features selected by SBS: ('Customer Type', 'Type of Travel', 'Class', 'Inflight wifi service', 'Gate location', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'Baggage handling', 'Inflight service')

SFFS:

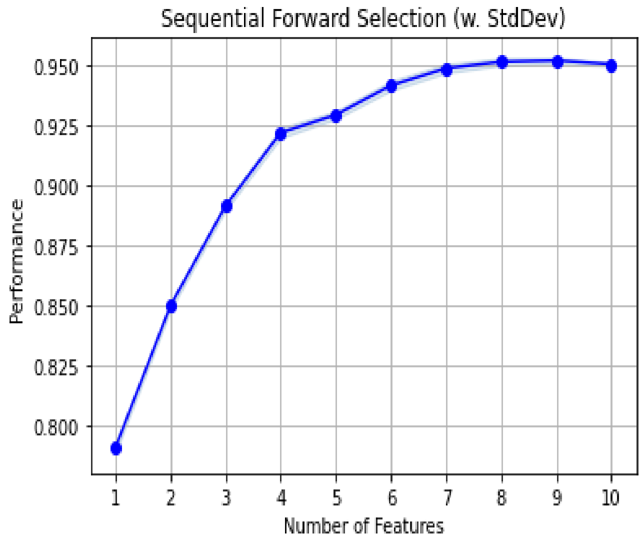
- Cross Validation Scores: [0.94833693 0.94953034 0.94929935 0.95072374]
- Accuracy for all 10 features: 0.9503195300512115
- Best 10 features selected by SFFS: ('Customer Type', 'Type of Travel', 'Class', 'Inflight wifi service', 'Gate location', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'Baggage handling', 'Inflight service')

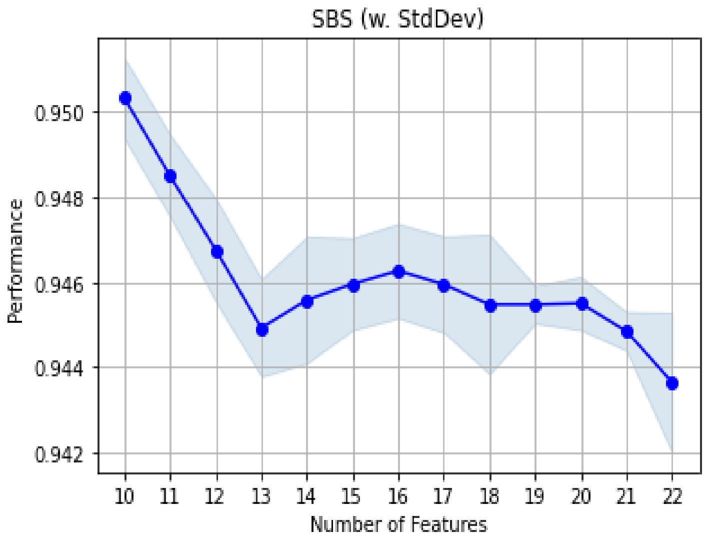
SBFS:

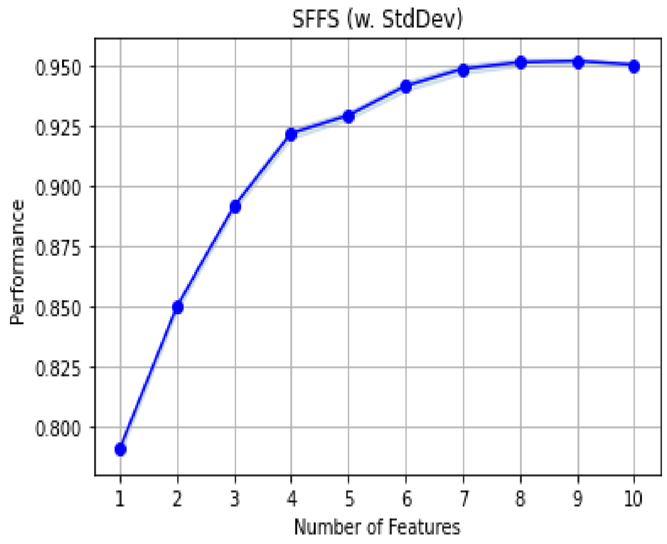
- Cross Validation Scores: [0.95045427 0.95083924 0.9499923 0.95234062]
- Accuracy for all 10 features: 0.9517150522343117
- Best 10 features selected by SBFS: ('Customer Type', 'Type of Travel', 'Class', 'Inflight wifi service', 'Online boarding', 'Seat comfort', 'Inflight entertainment', 'Baggage handling', 'Inflight service', 'Cleanliness')

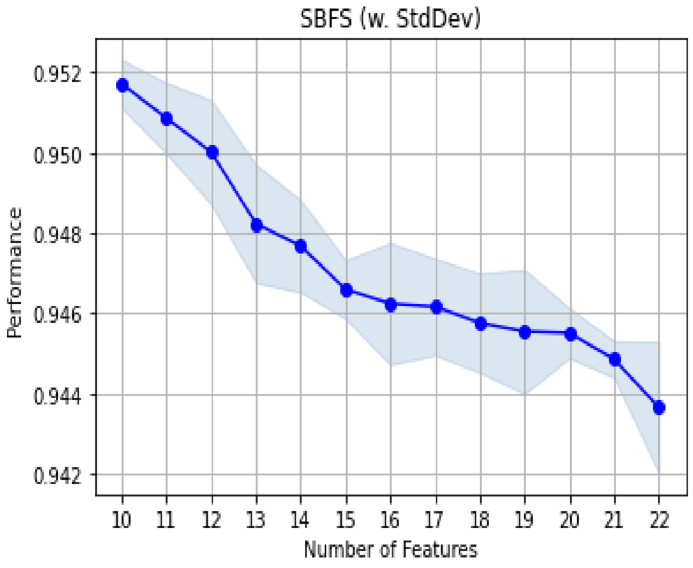
Part 4

Visualized the output from the feature selection in a pandas DataFrame format using the get_metric_dict for all four configurations. Finally, plotted the results for each configuration.









Part 5

Implemented Bi-directional Feature Set Generation Algorithm from scratch.

Part 6

Used the function implemented in part 5 and used selection criteria from the following:

- Accuracy Measures: using Decision Tree and SVM Classifiers
- Information Measures: Information gain
- Distance Measure: Angular Separation, Euclidean Distance and City-Block Distance

● Distance Measures. - Measures of separability, discrimination or divergence measures. The
most typical is derived from the distance between the class conditional density functions.)

Part 7

Trained Decision Tree classifier on the Selected features generated from each measure and reported its classification results.

- **Decision Tree(Accuracy Measure):**

```
                              precision    recall  f1-score   support

               0 : satisfied       0.40      0.60      0.48       453
1 : neutral or not satisfied       0.44      0.26      0.33       547

                    accuracy                           0.42      1000
                   macro avg       0.42      0.43      0.41      1000
                weighted avg       0.42      0.42      0.40      1000
```

- **Support Vector Machine(Accuracy Measure):**

```
                              precision    recall  f1-score   support

               0 : satisfied       0.43      0.62      0.51       453
1 : neutral or not satisfied       0.50      0.32      0.39       547

                    accuracy                           0.46      1000
                   macro avg       0.47      0.47      0.45      1000
                weighted avg       0.47      0.46      0.44      1000
```

- **Information Gain(Information Measure):**

```
                              precision    recall  f1-score   support

               0 : satisfied       0.40      0.60      0.48       453
1 : neutral or not satisfied       0.44      0.26      0.33       547

                    accuracy                           0.41      1000
                   macro avg       0.42      0.43      0.40      1000
                weighted avg       0.42      0.41      0.40      1000
```

- **K Means Clustering (Distance Measure):**

```
                             precision    recall  f1-score   support

              0 : satisfied       0.22      0.28      0.24       453
1 : neutral or not satisfied       0.24      0.19      0.21       547

                   accuracy                           0.23      1000
                  macro avg       0.23      0.23      0.23      1000
               weighted avg       0.23      0.23      0.23      1000
```

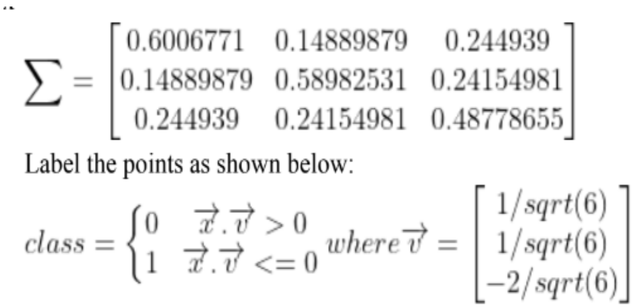
- **LDA Classifier (Distance Measures - Separability):**

```
                             precision    recall  f1-score   support

              0 : satisfied       0.39      0.63      0.48       453
1 : neutral or not satisfied       0.37      0.18      0.24       547

                   accuracy                           0.38      1000
                  macro avg       0.38      0.40      0.36      1000
               weighted avg       0.38      0.38      0.35      1000
```

Question 2.

Part 1

Made a Dataset of 1000 points sampled from a zero-centered gaussian distribution with a covariance matrix

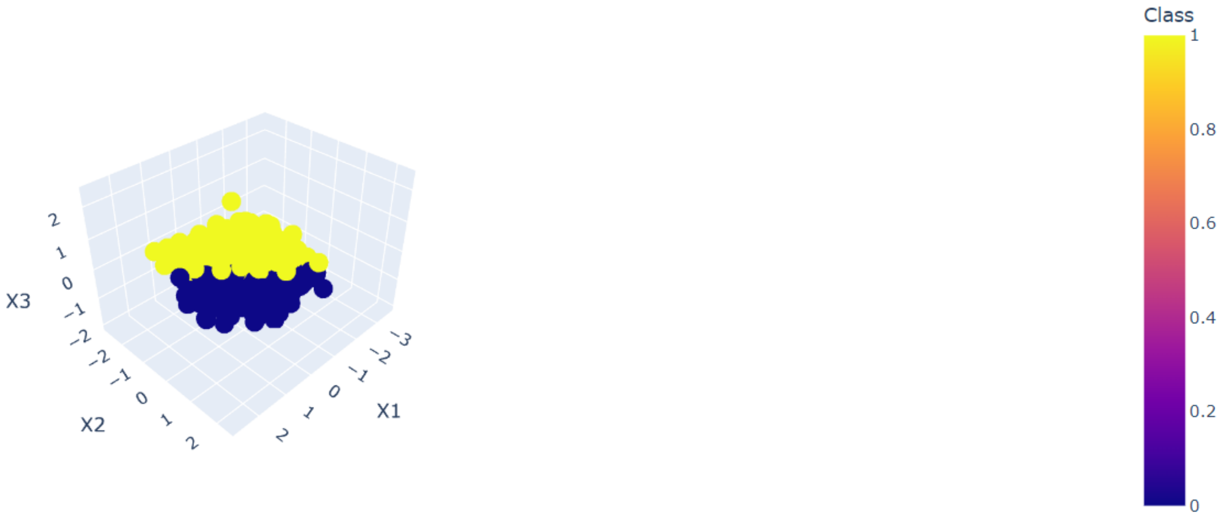


$$\sum = \begin{bmatrix} 0.6006771 & 0.14889879 & 0.244939 \\ 0.14889879 & 0.58982531 & 0.24154981 \\ 0.244939 & 0.24154981 & 0.48778655 \end{bmatrix}$$

Label the points as shown below:

$$class = \begin{cases} 0 & \vec{x}.\vec{v} > 0 \\ 1 & \vec{x}.\vec{v} <= 0 \end{cases} where \vec{v} = \begin{bmatrix} 1/sqrt(6) \\ 1/sqrt(6) \\ -2/sqrt(6) \end{bmatrix}$$

and x is the data point. Visualized the data as a 3D scatter-plot using plotly’s scatter_3d Function.



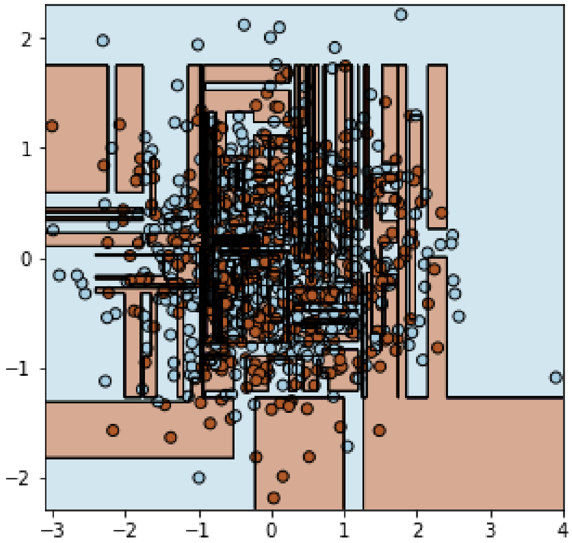
Part 2

Applied Principal Component analysis (using sklearn) with n_components=3 on the input data X and transformed the data accordingly.

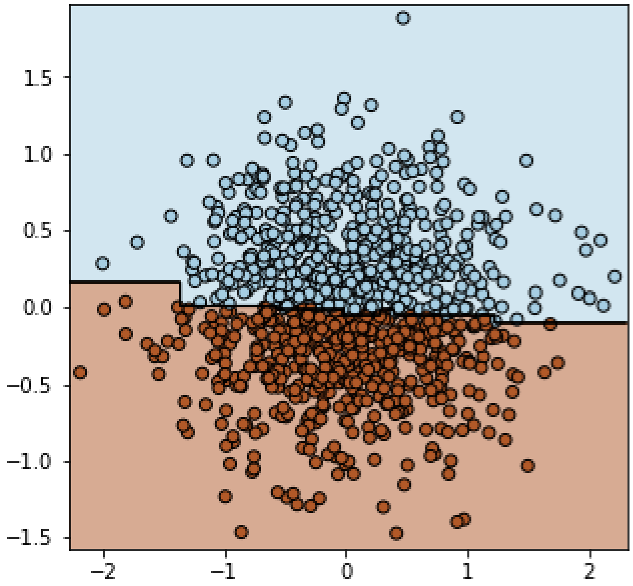
Part 3

Performed Complete FS on the Transformed Data with a number of features in subset =2. Fitted a Decision Tree for every subset-set of features of size 2 and plotted their decision boundaries superimposed with the data.

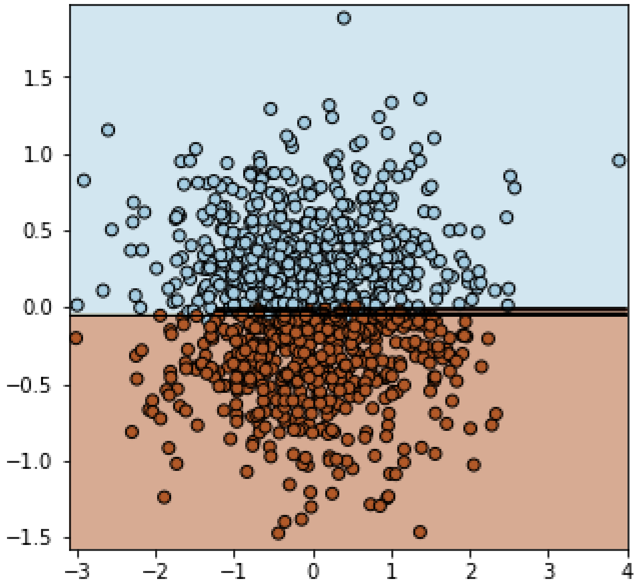
Subset1(X1,X2):



Subset2(X2,X3):



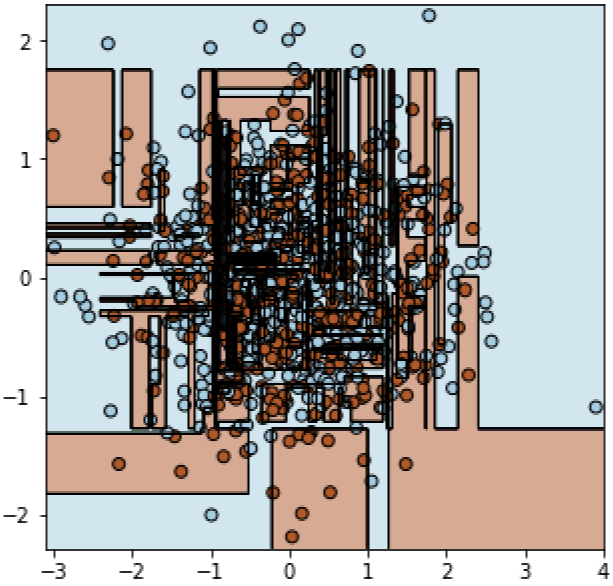
Subset3(X1,X3):



Part 4

Applied Principal Component analysis (using sklearn) with n_components=2 on the input data X and transformed the data accordingly.

- **Decision Boundary Obtained(n_components=2):**



Obtained Decision Boundary is similar to subset1(X1,X2).

Thank You