PRML LAB 8

Feature Selection

Vansh Agarwal

B21AI042

Note: Consuming 1 Buffer Day

Q1. Sequential Feature Selection

1.1 Data preprocessing and cleaning

Handling missing values
On inspection, it is found that the only column that contains missing values is **Arrival Delay in Minutes,** having only 0.29% missing entries. Since this is a very small ratio, these rows have been dropped.

Handling categorical values
There are five categorical columns, each of which has been dealt with using the Ordinal Encoder.

Then the data has been standardised using MinMaxScaling. Scaling is important because if some particular feature has a very high range as compared to others, then it will get undue importance while calculating the score metrics (of feature selection such as accuracy or distance metrics like Euclidean, Manhattan etc.). Hence, it will lead to biased selection of some features of over others. Thus scaling is important.

1.2 SFS using Decision Tree Classifier

Sequential Forward Selection (SFS)
SFS adds one feature at a time based on the classifier performance until a feature subset of the desired length k is reached or the performance starts worsening. In our case **k = 10, direction = Forward, Floating = FALSE and the performance evaluation used is accuracy.**

The selected features thus found are:

- **'Customer Type'**
- **'Type of Travel'**
- **'Class'**
- **'Inflight wifi service'**
- **'Ease of Online booking'**
- **'Gate location'**
- **'Online boarding'**
- **'Seat comfort'**
- **'Baggage handling'**
- **'Inflight service'**

The k score of the same is: 0.9505

1.3. Direction & Floating toggling in SFS

Sequential Forward Selection (SFS)
As explained in 1.2

Sequential Backward Selection (SBS)

SBS removes one feature at a time based on the classifier performance until a feature subset of the desired length k is reached or the performance starts worsening.

Floating (SFFS & SBFS)

The ***floating*** variants, SFFS and SBFS, can be considered extensions to the simpler SFS and SBS algorithms. The floating algorithms have an additional exclusion or inclusion step to remove features once they were included (or excluded) so that a larger number of feature subset combinations can be sampled.

Cross Validation = 4, CV Scores are as follows :
SFS: 0.9288
SBS: 0.9304
SFFS: 0.9288
SBFS: 0.9401

1.4. Visualisation

The metric dictionaries have been shown in output cells of ipynb notebook. The plots of each configuration have been shown below in Fig 1.

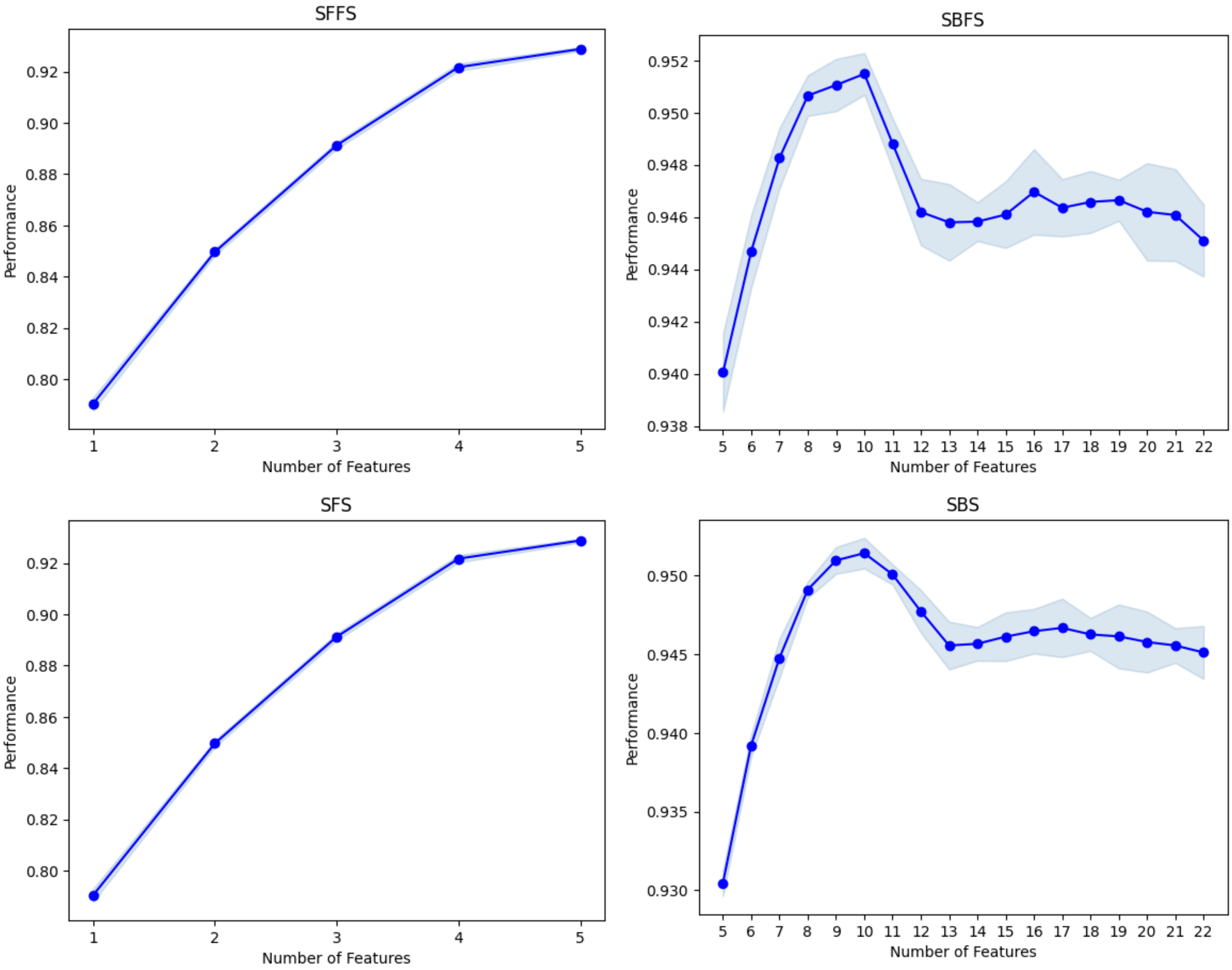

Fig 1.

Q2.1. Visualisation of dataset

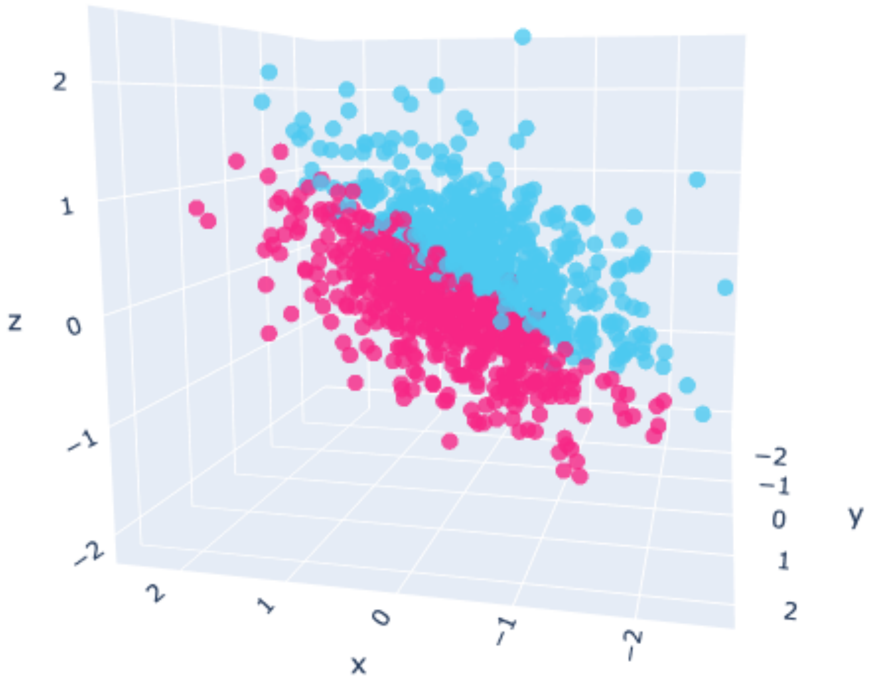

Fig 2.

2.1 PCA

PCA is a dimensionality reduction technique used to reduce the dimensions of the dataset, by projecting it on the optimal directions, while preserving maximum variance. Here with n = 3 we have visualised the transformed dataset in Fig 3.

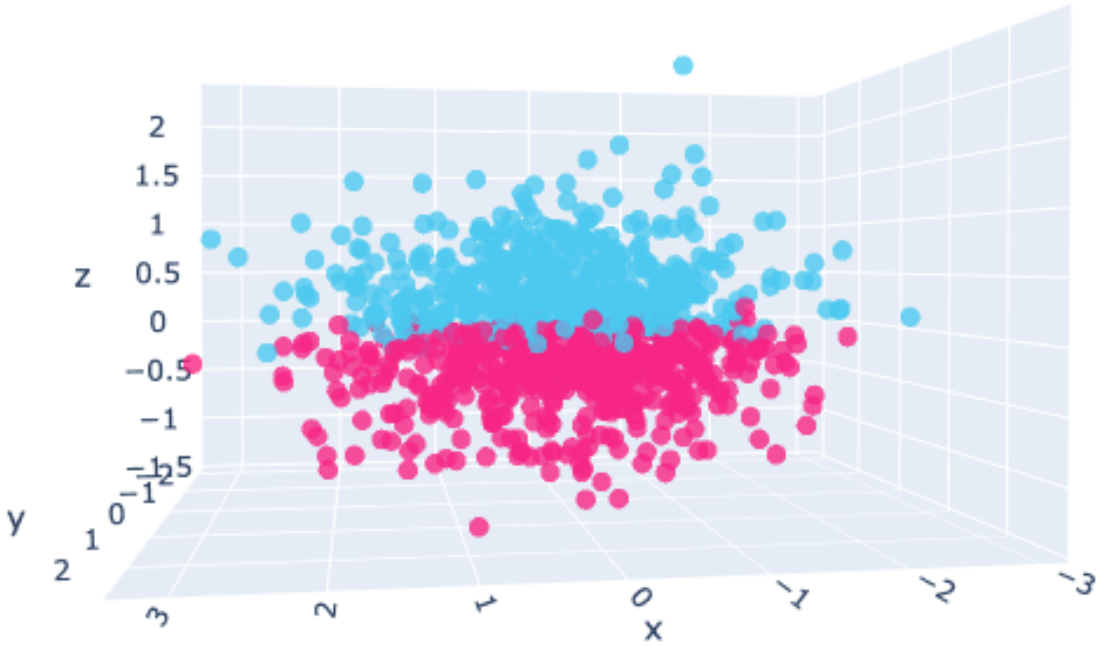

Fig 3

2.3 Complete FS

Complete FS is a technique of selecting every subset of k features (k being the number of desired feature) and selecting the one which gives maximum accuracy. Here k = 2, and the estimator used is decision tree. The decision boundaries for each pair has been shown in Fig 4.

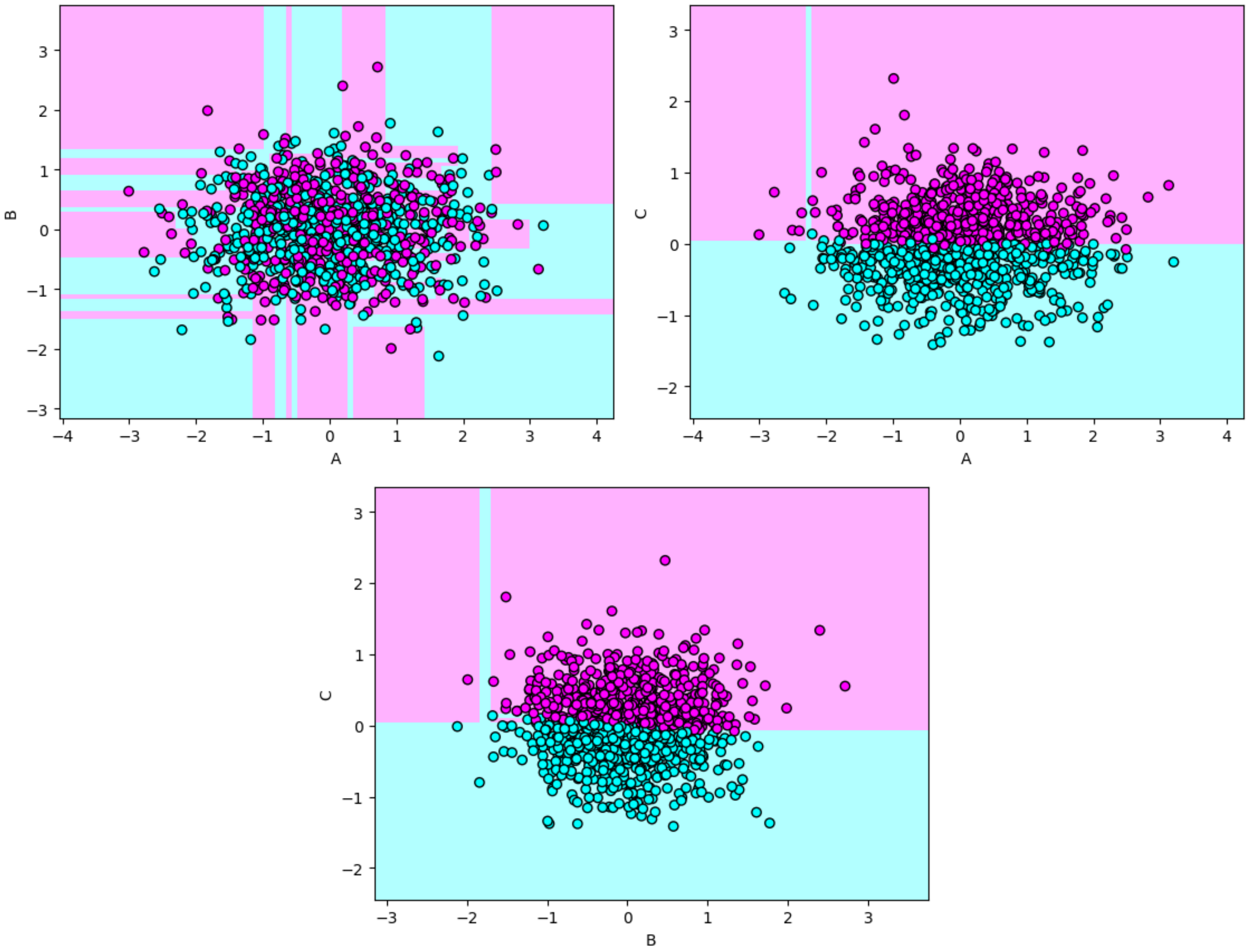

Fig 4

2.4 Complete FS

The decision Boundary most likely to be generated with PCA(n=2) is first one (i.e. between columns A & B) because it preserves the maximum variance of the dataset.