CSL2050 PATTERN RECOGNITION AND MACHINE LEARNING

LAB REPORT-08

NAME:	SAMEER SHARMA
ROLL NUMBER:	B21CS066
LAB TITLE:	Feature Selection

Problem 1)

Part 1)

- Done preprocessing on the airline data
- Dropped the index and id column since they are not features
- Since there are no null values, no need to drop columns or rows
- Label encoded all the categorical features (those with object dtype) using the LabelEncoder from sklearn
- Seperated the features and target labels from the dataset

Part 2)

- Created an object of SFS from the mlxtend library with the given parameters (forward=True, floating = False, scoring='accuracy', k_features=10) and embedded it with a decision tree classifier object
- Trained the classifier and accuracy for the selected 10 features is as follows

```
0.95008399904382599
```

- The selected ten features are:

```
('Customer Type',
 'Type of Travel',
 'Class',
 'Inflight wifi service',
 'Gate location',
 'Online boarding',
 'Seat comfort',
 'Inflight entertainment',
 'Baggage handling',
 'Inflight service')
```

Part 3)

- By changing forward and floating parameters, toggled between SFS, SBS, SFFS and SBFS

Feature Selection Algorithm	**Forward**	**Floating**
SFS (Sequential Forward Selection):	True	False
SBS (Sequential Backward Selection):	False	False
SFFS (Sequential Forward Floating Selection)	True	True
SBFS (Sequential Backward Floating Selection)	False	True

- Cross Validation scores for each configuration are as follows

```
accuray of sfs is:  0.950074338301363
accuray of sbs is:  0.9513774982914123
accuray of sffs is:  0.9512423575657759
accuray of sbfs is:  0.9514161109389957
```

Part 4)

- Visualized the output from feature selection of the four configurations in form of a pandas DataFrame. They are as follows

SFS

	feature_idx	cv_scores	avg_score	feature_names	ci_bound	std_dev	std_err
1	(11,)	[0.7897602224024094, 0.7918066334607514, 0.793...	0.790335	(Online boarding,)	0.004065	0.002536	0.001464
2	(3, 11)	[0.8483339125062743, 0.8511911656820726, 0.850...	0.849615	(Type of Travel, Online boarding)	0.002085	0.0013	0.000751
3	(3, 6, 11)	[0.8914629908490675, 0.8920421637901077, 0.892...	0.891249	(Type of Travel, Inflight wifi service, Online...	0.002214	0.001381	0.000797
4	(3, 6, 9, 11)	[0.9192246804895942, 0.9228927757828488, 0.922...	0.921733	(Type of Travel, Inflight wifi service, Gate l...	0.002346	0.001463	0.000845
5	(1, 3, 6, 9, 11)	[0.9277578284875864, 0.9284914475462374, 0.929...	0.928828	(Customer Type, Type of Travel, Inflight wifi ...	0.001202	0.00075	0.000433
6	(1, 3, 6, 9, 11, 16)	[0.9393412873083903, 0.942623267307618, 0.9396...	0.941309	(Customer Type, Type of Travel, Inflight wifi ...	0.002936	0.001832	0.001057
7	(1, 3, 4, 6, 9, 11, 16)	[0.9464071971890806, 0.9487238889532414, 0.948...	0.94826	(Customer Type, Type of Travel, Class, Infligh...	0.001787	0.001115	0.000644
8	(1, 3, 4, 6, 9, 11, 16, 18)	[0.9495347310706977, 0.9505772423645701, 0.950...	0.95074	(Customer Type, Type of Travel, Class, Infligh...	0.001429	0.000892	0.000515
9	(1, 3, 4, 6, 9, 11, 12, 16, 18)	[0.9488397235414495, 0.9511564153056102, 0.951...	0.950885	(Customer Type, Type of Travel, Class, Infligh...	0.00203	0.001266	0.000731
10	(1, 3, 4, 6, 9, 11, 12, 13, 16, 18)	[0.9490327811884629, 0.9501139040117379, 0.949...	0.950074	(Customer Type, Type of Travel, Class, Infligh...	0.001229	0.000767	0.000443

SBS

	feature_idx	cv_scores	avg_score	feature_names	ci_bound	std_dev	std_err
22	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.9440905054249199, 0.9466774778948994, 0.944...	0.945451	(Gender, Customer Type, Age, Type of Travel, C...	0.001942	0.001212	0.0007
21	(0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14...	[0.9437043901308931, 0.9469091470713155, 0.944...	0.94574	(Gender, Customer Type, Age, Type of Travel, C...	0.002768	0.001727	0.000997
20	(0, 1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 1...	[0.9438202247191011, 0.9462527510714699, 0.945...	0.945943	(Gender, Customer Type, Age, Type of Travel, C...	0.002887	0.001801	0.00104
19	(1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 15, ...	[0.944438009189544, 0.9467547009537047, 0.9455...	0.946483	(Customer Type, Age, Type of Travel, Class, In...	0.002824	0.001762	0.001017
18	(1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 15, ...	[0.9443221746013359, 0.9481447160122013, 0.944...	0.946483	(Customer Type, Age, Type of Travel, Class, In...	0.003059	0.001909	0.001102
17	(1, 2, 3, 4, 6, 9, 10, 11, 12, 13, 14, 15, 16,...	[0.9448627360129734, 0.9471408162477316, 0.945...	0.946503	(Customer Type, Age, Type of Travel, Class, In...	0.002647	0.001651	0.000953
16	(1, 2, 3, 4, 6, 9, 11, 12, 13, 14, 15, 16, 17,...	[0.9455577435422217, 0.9477972122475772, 0.944...	0.946619	(Customer Type, Age, Type of Travel, Class, In...	0.002456	0.001532	0.000885
15	(1, 2, 3, 4, 6, 9, 11, 12, 13, 15, 16, 17, 18,...	[0.9453260743658056, 0.9466774778948994, 0.945...	0.946348	(Customer Type, Age, Type of Travel, Class, In...	0.001708	0.001066	0.000615
14	(1, 2, 3, 4, 6, 9, 11, 12, 13, 15, 16, 17, 18,...	[0.9435885555426851, 0.9465230317772887, 0.945...	0.945518	(Customer Type, Age, Type of Travel, Class, In...	0.002002	0.001249	0.000721
13	(1, 2, 3, 4, 6, 9, 11, 12, 13, 16, 17, 18, 19)	[0.9452102397775975, 0.9451716282481949, 0.944...	0.945277	(Customer Type, Age, Type of Travel, Class, In...	0.00106	0.000661	0.000382
12	(1, 3, 4, 6, 9, 11, 12, 13, 16, 17, 18, 19)	[0.9464458087184834, 0.947333873894745, 0.9474...	0.94798	(Customer Type, Type of Travel, Class, Infligh...	0.002554	0.001594	0.00092
11	(1, 3, 4, 6, 11, 12, 13, 16, 17, 18, 19)	[0.9490327811884629, 0.9500366809529326, 0.950...	0.949939	(Customer Type, Type of Travel, Class, Infligh...	0.000915	0.000571	0.00033
10	(1, 3, 4, 6, 11, 12, 13, 16, 18, 19)	[0.9507703000115835, 0.951426696011429, 0.9506...	0.951377	(Customer Type, Type of Travel, Class, Infligh...	0.001317	0.000822	0.000474

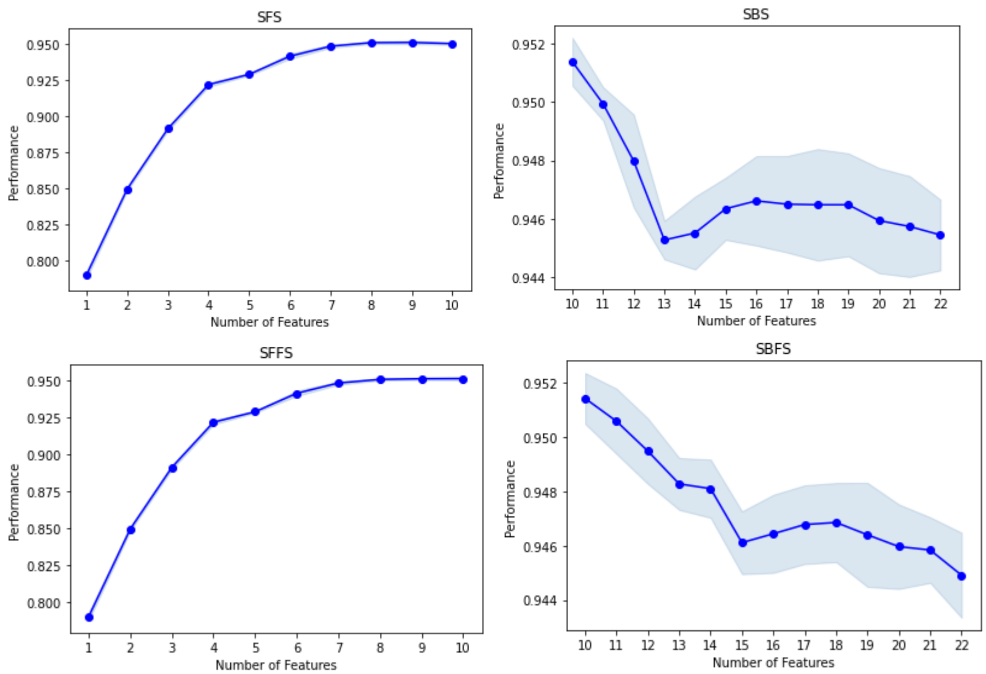
SFFS

	feature_idx	cv_scores	avg_score	feature_names	ci_bound	std_dev	std_err
1	(11,)	[0.7897602224024094, 0.7918066334607514, 0.793...	0.790335	(Online boarding,)	0.004065	0.002536	0.001464
2	(3, 11)	[0.8483339125062743, 0.8511911656820726, 0.850...	0.849615	(Type of Travel, Online boarding)	0.002085	0.0013	0.000751
3	(3, 6, 11)	[0.8914629908490675, 0.8920421637901077, 0.892...	0.891249	(Type of Travel, Inflight wifi service, Online...	0.002214	0.001381	0.000797
4	(3, 6, 9, 11)	[0.9192246804895942, 0.9228927757828488, 0.922...	0.921733	(Type of Travel, Inflight wifi service, Gate l...	0.002346	0.001463	0.000845
5	(1, 3, 6, 9, 11)	[0.9277578284875864, 0.9284914475462374, 0.929...	0.928828	(Customer Type, Type of Travel, Inflight wifi ...	0.001202	0.00075	0.000433
6	(1, 3, 6, 9, 11, 16)	[0.9393412873083903, 0.9425846557782154, 0.939...	0.9413	(Customer Type, Type of Travel, Inflight wifi ...	0.002925	0.001825	0.001053
7	(1, 3, 4, 6, 9, 11, 16)	[0.9463299741302753, 0.9487238889532414, 0.948...	0.94824	(Customer Type, Type of Travel, Class, Infligh...	0.001827	0.00114	0.000658
8	(1, 3, 4, 6, 9, 11, 16, 18)	[0.949611954129503, 0.9505386308351674, 0.9509...	0.950711	(Customer Type, Type of Travel, Class, Infligh...	0.001228	0.000766	0.000442
9	(1, 3, 4, 6, 11, 12, 13, 16, 18)	[0.9507316884821808, 0.9510791922468049, 0.950...	0.951146	(Customer Type, Type of Travel, Class, Infligh...	0.0013	0.000811	0.000468
10	(1, 3, 4, 6, 11, 12, 13, 16, 18, 19)	[0.9506544654233754, 0.9510405807174022, 0.950...	0.951242	(Customer Type, Type of Travel, Class, Infligh...	0.001773	0.001106	0.000638

SBFS

	feature_idx	cv_scores	avg_score	feature_names	ci_bound	std_dev	std_err
22	(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13,...	[0.9425460442488127, 0.9457508011892352, 0.944...	0.94491	(Gender, Customer Type, Age, Type of Travel, C...	0.002508	0.001565	0.000903
21	(0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14...	[0.9438202247191011, 0.9462527510714699, 0.946...	0.945837	(Gender, Customer Type, Age, Type of Travel, C...	0.001933	0.001206	0.000696
20	(0, 1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 1...	[0.9437430016602958, 0.9460210818950539, 0.945...	0.945962	(Gender, Customer Type, Age, Type of Travel, C...	0.002494	0.001556	0.000898
19	(1, 2, 3, 4, 6, 7, 9, 10, 11, 12, 13, 14, 15, ...	[0.9439746708367118, 0.9467933124831075, 0.945...	0.946397	(Customer Type, Age, Type of Travel, Class, In...	0.003069	0.001914	0.001105
18	(1, 2, 3, 4, 5, 6, 7, 9, 11, 12, 13, 14, 15, 1...	[0.9447855129541681, 0.9471022047183288, 0.946...	0.94685	(Customer Type, Age, Type of Travel, Class, Fl...	0.002334	0.001456	0.000841
17	(1, 2, 3, 4, 6, 7, 9, 11, 12, 13, 14, 15, 16, ...	[0.9448627360129734, 0.9477586007181744, 0.945...	0.946773	(Customer Type, Age, Type of Travel, Class, In...	0.002322	0.001449	0.000836
16	(1, 2, 3, 4, 6, 9, 11, 12, 13, 14, 15, 16, 17,...	[0.9456349666010271, 0.9472952623653423, 0.944...	0.946435	(Customer Type, Age, Type of Travel, Class, In...	0.002302	0.001436	0.000829
15	(1, 2, 3, 4, 6, 9, 11, 12, 13, 14, 15, 16, 17,...	[0.9448627360129734, 0.9471022047183288, 0.945...	0.946107	(Customer Type, Age, Type of Travel, Class, In...	0.001853	0.001156	0.000667
14	(1, 3, 4, 6, 7, 8, 9, 11, 12, 13, 16, 17, 18, 19)	[0.9480674929533959, 0.9464458087184834, 0.948...	0.948095	(Customer Type, Type of Travel, Class, Infligh...	0.001718	0.001072	0.000619
13	(1, 3, 4, 6, 7, 8, 11, 12, 13, 16, 17, 18, 19)	[0.9476041546005637, 0.9477586007181744, 0.947...	0.948269	(Customer Type, Type of Travel, Class, Infligh...	0.001531	0.000955	0.000551
12	(1, 3, 4, 6, 7, 8, 11, 12, 13, 16, 18, 19)	[0.9475269315417584, 0.949496119541295, 0.9501...	0.949476	(Customer Type, Type of Travel, Class, Infligh...	0.001935	0.001207	0.000697
11	(1, 3, 4, 6, 8, 11, 12, 13, 16, 18, 19)	[0.9492644503648789, 0.9505386308351674, 0.949...	0.950576	(Customer Type, Type of Travel, Class, Infligh...	0.001927	0.001202	0.000694
10	(1, 3, 4, 6, 11, 12, 13, 16, 18, 19)	[0.9511950268350129, 0.9510405807174022, 0.950...	0.951416	(Customer Type, Type of Travel, Class, Infligh...	0.001503	0.000938	0.000541

- Plotted the performance vs no of features graphs for each configuration. They are as follows



Part 5)

- Implemented Bidirectional feature set generation algorithm from scratch. The well commented code for the same is available in the colab notebook.

Part 6)

- Implemented various selection criterions on the above bidirectional feature set generation algorithm. The well commented for the selection criterions are available in the colab notebook.

Part 7)

- Trained a decision tree classifier on the feature sets generated by various selection criterions (each set contains 10 features)
- Their performance is as follows:

```
Decision tree accuracy measure : 0.9505859472161153
SVM classifier accuracy measure : 0.9449099267844776
Information gain : 0.9187694440096225
Euclidian distance measure : 0.7586154369771518
City block measure : 0.7589822477176698
Angular distance measure : 0.7586250843616154
```

Problem 2)

Part 1)

- Generated a synthetic dataset with zero mean and the following covariance matrix

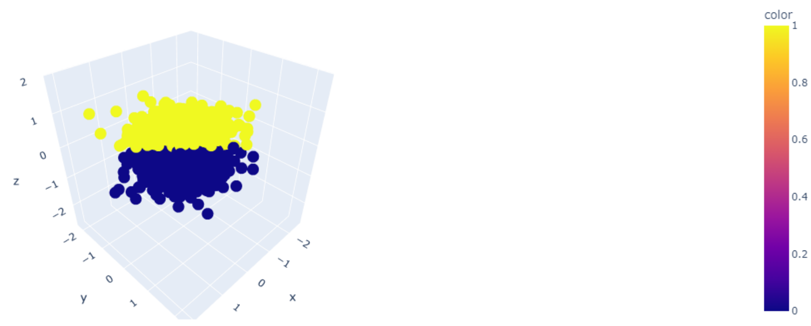
$$\sum = \begin{bmatrix} 0.6006771 & 0.14889879 & 0.244939 \\ 0.14889879 & 0.58982531 & 0.24154981 \\ 0.244939 & 0.24154981 & 0.48778655 \end{bmatrix}$$

  Using the np.random.multivariate_normal function

- After generating the dataset, assigned the class labels to the datapoints using the following criteria

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

- The 3D plot of the generated dataset is as follows:



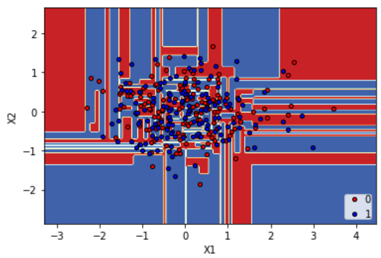
  The interactive 3D plot is available in the colab notebook.

Part 2)

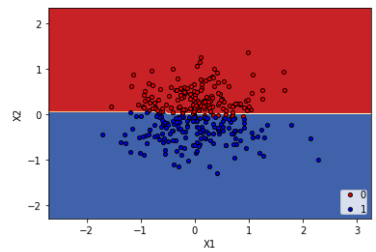
- Applied Principal component analysis (PCA) on the dataset with n_components = 3. This is not going to reduce the dataset but transform it.

Part 3)

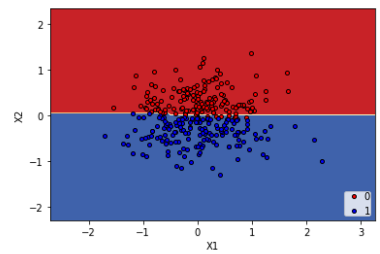
- Applied complete feature selection on the transformed Dataset. In it, I selected every possible pairs of features from the reduced dataset. Then, trained a decision tree classifier on every reduced dataset and computed its accuacy and plotted the decision boundary.
- The obtained plots of decision boundary are as follows



1st and 2nd feature



2nd and 3rd feature



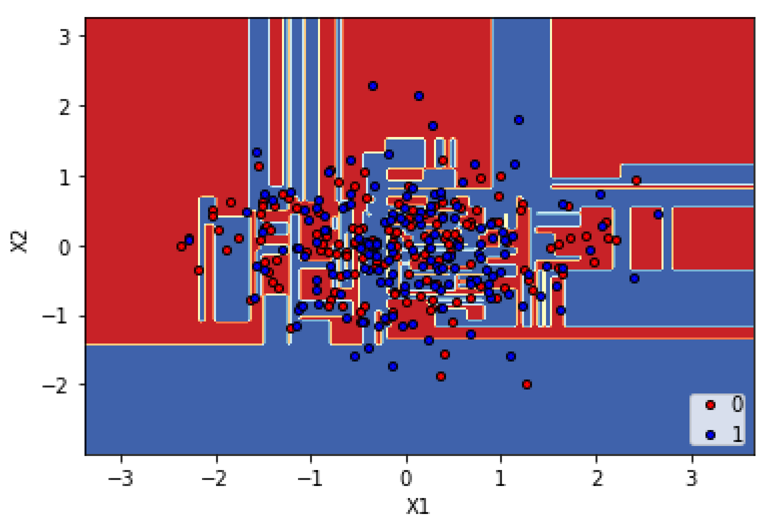
1st and 3rd feature

- The obtained accuracies are as follows:

1st and 2nd feature	0.5033333333333333
2nd and 3rd feature	0.9766666666666667
1st and 3rd feature	0.9733333333333334

Part 4)

- Applied PCA on the original dataset with n_components=2. This gave us a reduced dataset with two features.
- The plot of decision tree classifier trained on this reduced dataset is as follows



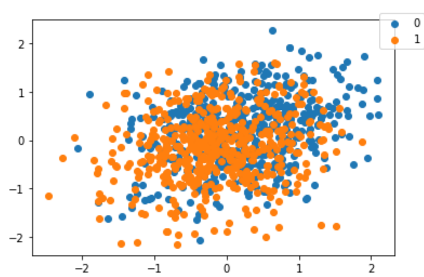
Accuracy: 0.47

We can observe the decision boundary is similar to the boundary that we got when we trained the decision tree classifier on the first two features of the transformed dataset.

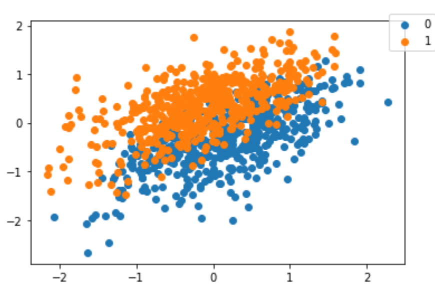
- Lets calculate the Euclidian distance of this reduced dataset to every subset we got in the previous part. I got the following results. From these results, we can say that certainly, the reduced dataset is actually consisting of first two features of the transformed dataset that we got in the second part.

```
[0, 1]
1.7950820155069126e-14

[1, 2]
44.892896031683456

[0, 2]
25.582488570651158
```

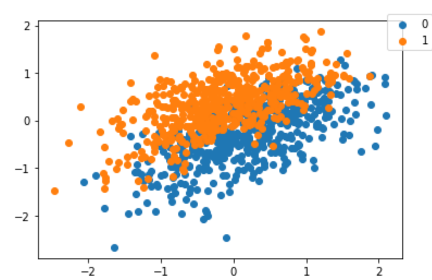
- Plotted the datapoints with repect to each subset we got in the previous part. The obtained plots are as follows:



1st and 2nd feature



2nd and 3rd feature



1st and 3rd feature

- Clearly, the first two features are not seperating the two clusters effectively and therefore, we are getting very less accuracy. This is not the case with other subsets, so we are getting very high accuracy.

- From this, we can argue that PCA does not focus on seperability of classes and therefore, may or may not increase the accuracy of classifier when trained on the reduced dataset.
- To increase the seperability, we should use LCA which primarily focuses on increase the seperability of the clusters.