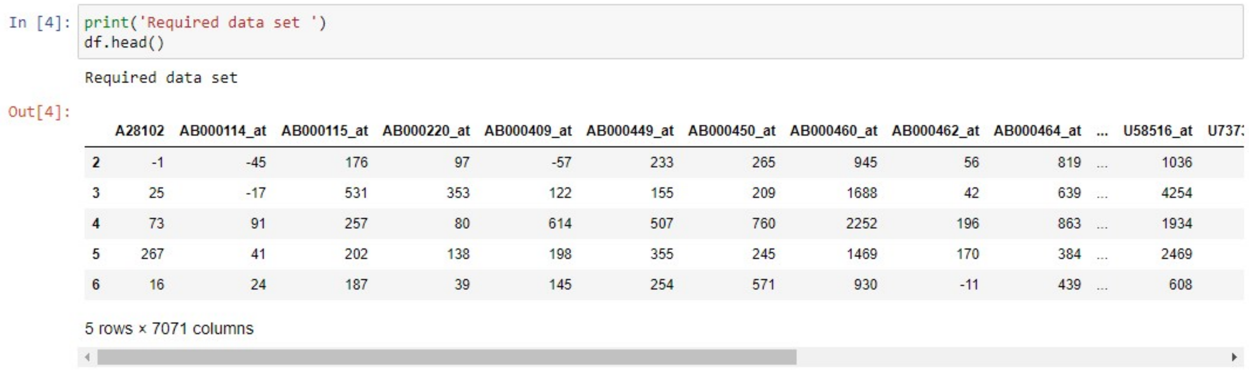
DLBCL Sample Dataset



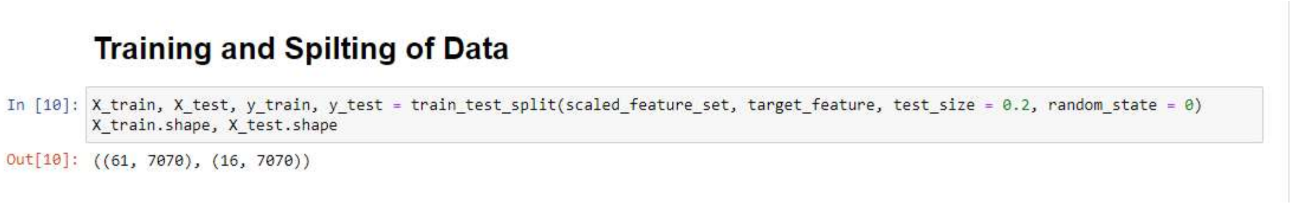
```
In [4]: print('Required data set ')
        df.head()

        Required data set

Out[4]:
```

	A28102	AB000114_at	AB000115_at	AB000220_at	AB000409_at	AB000449_at	AB000450_at	AB000460_at	AB000462_at	AB000464_at	...	U58516_at	U737
2	-1	-45	176	97	-57	233	265	945	56	819	...	1036	
3	25	-17	531	353	122	155	209	1688	42	639	...	4254	
4	73	91	257	80	614	507	760	2252	196	863	...	1934	
5	267	41	202	138	198	355	245	1469	170	384	...	2469	
6	16	24	187	39	145	254	571	930	-11	439	...	608	

5 rows × 7071 columns



Training and Spilting of Data

```
In [10]: X_train, X_test, y_train, y_test = train_test_split(scaled_feature_set, target_feature, test_size = 0.2, random_state = 0)
         X_train.shape, X_test.shape

Out[10]: ((61, 7070), (16, 7070))
```

1. Total number of features present in the dataset = 7071 (including target feature) and 7070(excluding target feature)
2. Number of features extracted by each filter method = 1343 (approx. 19%)
3. Average accuracy of KNN Classifier is obtained by calculating the accuracy of 19 neighbors from 1to 19
4. Filter Methods Used:

   4.1 Mutual Information

   4.2 F Classification

   4.3 T Test
5. Wrapper Methods Used:

   5.1 Sequential Forward Search

   5.2 Sequential Backward Search

Assumptions:

1. F1 = Mutual Information
2. F2 = F Classification
3. F3 = T Test
4. SFS = Sequential Forward Search
5. SBS = Sequential Backward Search
6. KNN = K Nearest Neighbors
7. SVM = Support Vector Machine
8. S1 = F1(N features) → F2(2N/3 features out of selected features from F1) → F3(N/3 features out of selected features from F2)
9. S2 = F2(N features) → F3(2N/3 features out of selected features from F2) → F1(N/3 features out of selected features from F3)
10. S3 = F3(N features) → F1(2N/3 features out of selected features from F3) → F2(N/3 features out of selected features from F1)
11. Union = F1 U F2 U F3
12. TP = True Positive
13. FP = False Positive
14. FN = False Negative
15. TN = True Negative

    Here,

    N = 7070

    2N/3 = 4713

    N/3 = 2356

    U = Union of set

Filter method used by each wrapper method = F Classification(because Wrapper methods takes a lot of time to extract features in comparison to filter methods)

Classification method used by wrapper methods to extract features=Support Vector Machine

Number of features extracted by F Classification filter method to give the wrapper method for further feature extraction = 500

Number of features used by each wrapper method = 500

Number of features extracted by Wrapper Methods:

1. Sequential Forward Search = 100
2. Sequential Backward Search = 400(because it is takes a lot of time to remove 1 feature from set)

Here, confusion matrix displayed in case of KNN neighbors will correspond to the maximum accuracy achieved by the KNN neighbors.

Results:

Comparison Table of KNN Classifier

Parameters → ↓ Method	Time Taken For Execution (seconds)	Average						Best/Maximum					
		Accuracy	Confusion Matrix				F-Score	Accuracy	Confusion Matrix				F-Score
			TP	FP	FN	TN			TP	FP	FN	TN	
F1	229.0674	83.2237	11	0.11	2.6	2.4	0.8322	93.75	11	0	1	4	0.9375
F2	927.9947	76.6447	11	0.42	3.3	1.7	0.7664	100.0	11	0	0	5	1.0
F3	14.1915	79.2763	11	0.37	2.9	2.1	0.7927	100.0	11	0	0	5	1.0
Union	1171.2536	78.6184	11	0.32	3.1	1.9	0.7861	100.0	11	0	0	5	1.0
S1	349.5713	77.3026	11	0.47	3.2	1.8	0.773	93.75	10	1	0	5	0.9375
S2	543.8623	76.6447	11	0.37	3.4	1.6	0.7664	93.75	10	1	0	5	0.9375
S3	144.0018	77.6316	11	0.42	3.2	1.8	0.7763	93.75	11	0	1	4	0.9375
SFS	576	77.6316	11	0.32	3.3	1.7	0.7763	93.75	11	0	1	4	0.9375
SBS	3047	76.3158	11	0.47	3.3	1.7	0.7631	93.75	11	0	1	4	0.9375

Comparison Table of SVM Classifier

Parameters → ↓ Method	Time Taken For Execution (seconds)	Accuracy	Confusion Matrix				F-Score
			TP	FP	FN	TN	
F1	229.0674	75.0	11	0	4	1	0.75
F2	927.9947	68.75	11	0	5	0	0.6875
F3	14.1915	68.75	11	0	5	0	0.6875
Union	1171.2536	68.75	11	0	5	0	0.6875
S1	349.5713	68.75	11	0	5	0	0.6875
S2	543.8623	68.75	11	0	5	0	0.6875
S3	144.0018	68.75	11	0	5	0	0.6875
SFS	576	68.75	11	0	5	0	0.6875
SBS	3047	68.75	11	0	5	0	0.6875

Observations and Conclusions:

1. F1 (Mutual Information) filter method is providing the best average accuracy, average confusion matrix and average f-score among all methods for KNN Classifier.
2. F1 (Mutual Information) filter method is providing the best accuracy, confusion matrix and f-score among all methods for SVM Classifier.
3. KNN Classifier is performing better than SVM Classifier for this dataset.
4. T-test is taking least time among all feature extraction methods.
5. In SVM Classifier, except F1(Mutual Information) filter method, all other methods are providing the same accuracy, confusion matrix and f-score.
6. Cascading of filter methods or using combination wrapper and filter methods is providing better solution in comparison of using single method for feature extraction (T test and F Classification) because their combination is overcoming the disadvantages of one another.

7. Highest average accuracy achieved by KNN is in case of F1 (Mutual Information) which is 83.2237%.
8. Highest accuracy achieved by SVM is in case of F1 (Mutual Information) which is 75%.
9. Maximum accuracy and maximum f-score achieved in case of SVM (Support Vector Machine) is achieved in case of F1 (Mutual Information).
10. Here, wrapper methods are not providing any better solution than already presented filter method.
11. Here, no cascading and union of filter methods are not providing any better solution than F1 (Mutual Information) in both KNN and SVM Classifier.
12. In KNN Classifier, F1 (Mutual Information) is providing best average accuracy but it is not providing maximum accuracy. Here, maximum accuracy is achieved by F2(F Classification), F3(T test) and Union (F1 U F2 U F3).