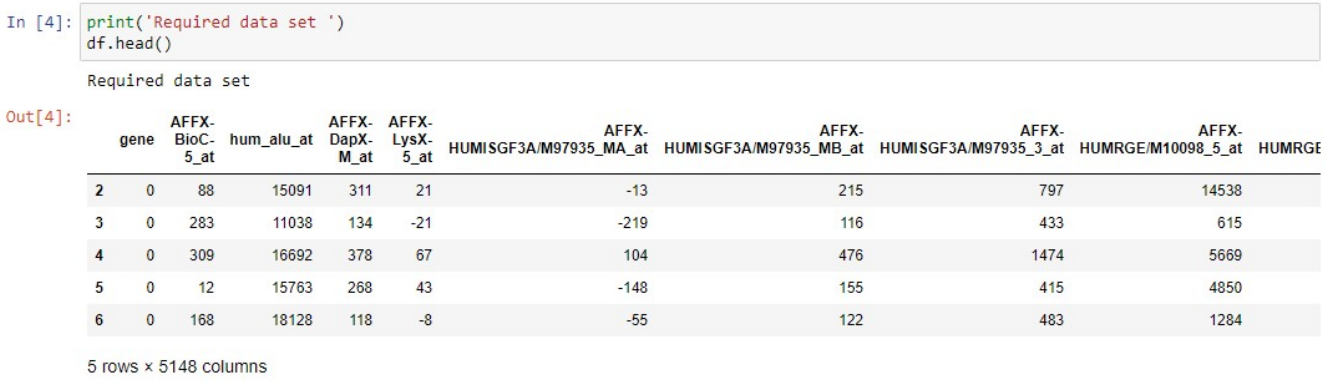
Leukemia Sample Dataset



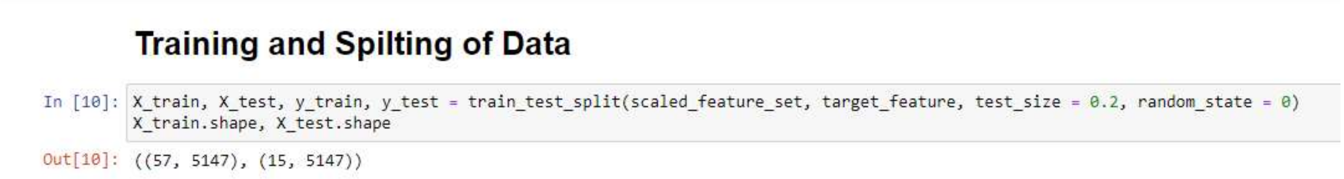
```
In [4]: print('Required data set ')
        df.head()
```

```
Required data set
```

Out[4]:

	gene	AFFX-BioC-5_at	hum_alu_at	AFFX-DapX-M_at	AFFX-LysX-5_at	AFFX-HUMISGF3A/M97935_MA_at	AFFX-HUMISGF3A/M97935_MB_at	AFFX-HUMISGF3A/M97935_3_at	AFFX-HUMRGE/M10098_5_at	HUMRGE
2	0	88	15091	311	21	-13	215	797	14538	
3	0	283	11038	134	-21	-219	116	433	615	
4	0	309	16692	378	67	104	476	1474	5669	
5	0	12	15763	268	43	-148	155	415	4850	
6	0	168	18128	118	-8	-55	122	483	1284	

5 rows × 5148 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]: ((57, 5147), (15, 5147))
```

1. Total number of features present in the dataset = 5148 (including target feature) and 5147(excluding target feature)
2. Number of features extracted by each filter method = 977 (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 = 5147

    2N/3 = 3431

    N/3 = 1715

    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	50.24	67.0175	7	0	4.9	3.1	0.7471	93.333	7	0	1	7	0.9333
F2	383.111	76.1404	7	0	3.6	4.4	0.8118	100.0	7	0	0	8	1.0
F3	5.9339	67.7193	7	0	3.4	4.6	0.7494	93.333	7	0	1	7	0.9333
Union	439.2849	76.4912	7	0	3.5	4.5	0.8118	100.0	7	0	0	8	1.0
S1	111.34751	74.7368	7	0	3.8	4.2	0.7955	93.333	7	0	1	7	0.9333
S2	210.8718	74.7368	7	0	3.8	4.2	0.7939	86.6667	7	0	2	6	0.875
S3	54.2909	69.4737	7	0	4.6	3.4	0.7654	100.0	7	0	0	8	1.0
SFS	169	85.2632	6.8	0.2	2	6	0.865	100.0	7	0	0	8	1.0
SBS	1583	80.7018	6.8	0.2	2.7	5.3	0.8301	93.333	7	0	1	7	0.9333

Comparison Table of SVM Classifier

Parameters → ↓ Method	Time Taken For Execution (seconds)	Accuracy	Confusion Matrix				F-Score
			TP	FP	FN	TN	
F1	**50.24**	**80.0**	**7**	**0**	**3**	**5**	**0.8235**
F2	**383.111**	**80.0**	**7**	**0**	**3**	**5**	**0.8235**
F3	**5.9339**	**80.0**	**7**	**0**	**3**	**5**	**0.8235**
Union	**439.2849**	**86.6667**	**7**	**0**	**2**	**6**	**0.875**
S1	**111.34751**	**80.0**	**7**	**0**	**3**	**5**	**0.8235**
S2	**210.8718**	**80.0**	**7**	**0**	**3**	**5**	**0.8235**
S3	**54.2909**	**80.0**	**7**	**0**	**3**	**5**	**0.8235**
SFS	**169**	**93.3333**	**7**	**0**	**1**	**7**	**0.9333**
SBS	**1583**	**86.6667**	**7**	**0**	**2**	**6**	**0.875**

Observations and Conclusions:

1. F2 (F Classification) filter method is providing the best accuracy, confusion matrix and f-score among all filter methods for KNN Classifier but it is also taking the most time among all the other filter method.
2. Combination of F Classification and SFS (Sequential Forward Search) method is providing the best accuracy, confusion matrix and f-score among all the features extraction methods for both KNN and SVM Classification.
3. T-test is better than Mutual Information as it has higher average accuracy in KNN classification than mutual information and it also takes less time for execution than mutual information. Both are giving same accuracy in SVM Classification.
4. T-test is taking least time among all feature extraction methods.

5. S3= F3 (N features) → F1(2N/3 features out of selected features from F3) → F2(N/3 features out of selected features from F1) is providing the better solution than F2 in terms of average accuracy and time taken for the execution in case of KNN classification.
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 because their combination is overcoming the disadvantages of one another because Mutual Information and T-Test are increasing speed of processing and F Classification is contributing in increasing accuracy.
7. Union = F1 U F2 U F3 is providing the better solution among all other filter extraction methods and cascading but it is also taking more time than other methods because it is a combination of all 3 filter methods in both KNN and SVM classification.
8. Highest average accuracy achieved by KNN is in case of SFS which is 85.2632%.
9. Highest accuracy achieved by SVM is in case of SFS which is 93.3333%.
10. SVM is better than KNN classification for this dataset because minimum accuracy in case of SVM is 80% and maximum accuracy is 93.3333%. But in case KNN classification only SFS and SBS is achieving accuracy more than
11. Union = F1 U F2 U F3 is providing the better solution among all other filter feature extraction methods in case of both KNN and SVM classifier but it is also taking more time than other methods except for wrapper methods.
12. Combination of filter and wrapper method is a good choice to extract features from a dataset. Like here- combination of F classification and Sequential Forward Search for both classification and initially using the filter method before wrapper method also decreases the time in comparison to using the wrapper method alone.
13. Maximum accuracy and maximum f-score achieved in case of SVM (Support Vector Machine) is achieved in case of Sequential Forward Selection.
14. Both wrapper methods (Sequential Backward Selection and Sequential Forward Selection) are providing better average accuracy than all filter methods with less number of features than filter methods, but the only problem is with the time required for execution of wrapper methods.
15. Sequential Forward Selection is providing best solution for accuracy, confusion matrix and f-score in both KNN and SVM classification.