Assignment 2 Comparison report

Problem A:

Comparison Table of three filter methods used on leukemia dataset:

	Mutual Info	F classif	T Test
KNN Classification			
Accuracy	0.9333333	0.866667	1.0
F1 Score	0.9206349	0.829545	1.0
Confusion Matrix	[[10 0] [1 4]]	[[10 0] [2 3]]	[[10 0] [0 5]]
SVM Classification			
Accuracy	0.9333333	1.0	1.0
F1 Score	0.9282296	1.0	1.0
Confusion Matrix	[[9 1] [0 5]]	[[10 0] [0 5]]	[[10 0] [0 5]]

Inference:

From the above table we can clearly say that, the T test filter method selects the important features and results in very good accuracy in the classification of data for both KNN and SVM classification. Mutual info gives better result for KNN classification but for SVM, F classif gives better result. F classif gives lesser accuracy and f1 score in KNN compare to other two.

Comparison Table of three filter methods used on DLBCL dataset:

	Mutual Info	F classif	T Test
KNN Classification			
Accuracy	0.6875	0.8125	0.8125
F1 Score	0.67611	0.79220	0.725714
Confusion Matrix	[[7 5] [0 4]]	[[9 3] [0 4]]	[[11 3] [0 2]]
SVM Classification			
Accuracy	0.9375	0.9375	1.0
F1 Score	0.922705	0.922705	1.0
Confusion Matrix	[[11 1] [0 4]]	[[11 1] [0 4]]	[[14 0] [0 2]]

Inference:

The T test outperforms in SVM classification for a given dataset. The mutual info and the F classif gives similar results in SVM. And in KNN, T test and F classifier gives better accuracy than mutual information. So overall t test gives better results for SVM and F classif gives better results for KNN.

Comparison Table of three filter methods used on lung dataset:

	Mutual Info	F classif	T Test
KNN Classification			
Accuracy	0.7804878	0.7804878	0.804878
F1 Score	0.6785824	0.6785824	0.693334
Confusion Matrix	[[25 0 0 0 1] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [4 0 0 0 3]]	[[25 0 0 0 1] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [4 0 0 0 3]]	[[26 0 0 0 0] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [4 0 0 0 3]]
SVM Classification			
Accuracy	0.804878	0.804878	0.804878
F1 Score	0.7142857	0.7142857	0.714285
Confusion Matrix	[[24 0 0 0 2] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [2 0 0 0 5]]	[[24 0 0 0 2] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [2 0 0 0 5]]	[[24 0 0 0 2] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [2 0 0 0 5]]

Inference:

In this dataset, All three filter methods perform equally and give similar result for SVM. For KNN, T test performs slightly better than mutual info and f classif and this two gives similar result.

Problem B:

		KNN Classification	SVM Classification
leukemia	Accuracy	0.93333	1.0
	F1 Score	0.920634	1.0
	Confusion Matrix	[[10 0] [1 4]]	[[10 0] [0 5]]
DLBCL	Accuracy	0.75	1.0
	F1 Score	0.733	1.0
	Confusion Matrix	[[8 4] [0 4]]	[[12 0] [0 4]]
Lung	Accuracy	0.7804878	0.829268
	F1 Score	0.6785824	0.797402
	Confusion Matrix	[[25 0 0 0 1] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [4 0 0 0 3]]	[[24 0 0 0 2] [0 3 0 0 0] [0 0 1 0 0] [3 0 0 1 0] [2 0 0 0 5]]

Inference:

In all three datasets, SVM classifies the data accurately compare to KNN.

Problem C:

		KNN Classification	SVM Classification
leukemia	Accuracy	0.53333	0.3333
	F1 Score	0.347826	0.25
	Confusion Matrix	[[8 2] [5 0]]	[[0 10] [0 5]]
DLBCL	Accuracy	0.75	0.75
	F1 Score	0.428571	0.428571
	Confusion Matrix	[[12 0] [4 0]]	[[12 0] [4 0]]
Lung	Accuracy	0.07317	0.07317
	F1 Score	0.0272727	0.0272727
	Confusion Matrix	[[0 26 0 0 1] [0 3 0 0 0] [0 1 0 0 0] [0 4 0 0 0] [0 7 0 0 0]]	[[0 26 0 0 1] [0 3 0 0 0] [0 1 0 0 0] [0 4 0 0 0] [0 7 0 0 0]]

Inference:

When we apply cascade filtering, The classification result is very poor. It is not at all recommended. It is still comparable with earlier problems for leukemia and DLBCL dataset but for lung dataset it is not able to classify.

Problem D:

In this problem, I have used T test filter method to reduce the features for wrapper method as it was taking long time to run.

Comparison Table for lung dataset:

		Sequential Forward Search	Sequential Backward Search
KNN	Accuracy	0.853658	0.804878
	F1 Score	0.745977	0.714285
	Confusion Matrix	[[26 0 0 0 0] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [2 0 0 0 5]]	[[24 0 0 0 2] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [2 0 0 0 5]]
SVM	Accuracy	0.82926	0.902439
	F1 Score	0.729284	0.910459
	Confusion Matrix	[[25 0 0 0 1] [0 3 0 0 0] [0 0 1 0 0] [4 0 0 0 0] [2 0 0 0 5]]	[[25 0 0 0 1] [0 3 0 0 0] [0 0 1 0 0] [1 0 0 3 0] [2 0 0 0 5]]

Inference:

Sequential forward search gives slightly better result compare to Sequential backward search in the case of KNN. While sequential backward searches give better results in case of SVM.

Comparison Table for leukemia dataset:

		Sequential Forward Search	Sequential Backward Search
KNN	Accuracy	0.933333	0.933333
	F1 Score	0.928229	0.920634
	Confusion Matrix	[[9 1] [0 5]]	[[10 0] [1 4]]
SVM	Accuracy	0.933333	0.8
	F1 Score	0.928229	0.79638
	Confusion Matrix	[[9 1] [0 5]]	[[7 3] [0 5]]

Inference:

Sequential forward search and sequential backward search gives almost identical results in the case of KNN. In the case of SVM, sequential forward search gives better results.

Comparison Table for DLBCL dataset:

		Sequential Forward Search	Sequential Backward Search
KNN	Accuracy	0.875	0.8125
	F1 Score	0.854545	0.792207
	Confusion Matrix	[[10 2] [0 4]]	[[9 3] [0 4]]
SVM	Accuracy	0.625	0.6875
	F1 Score	0.60	0.653679
	Confusion Matrix	[[7 5] [1 3]]	[[8 4] [1 3]]

Inference:

Sequential forward search gives better results than sequential backward search in the case of KNN. While sequential backward searches give better results in case of SVM.