DHA SUFFA UNIVERSITY

Dept. of Computer Science

Spring 2019

Course Name: Bioinformatics (CS -429)

Assignment 4 **Marks: 10**

Deadline for Submission: 23 – 05 – 2019

Submission Guidelines: Assignments will be submitted in hard copies. Soft copy submission will not be accepted. Late submission of assignments will not be accepted. Plagiarized assignments will result in 0 marks. Submit the source code and output screenshots for your codes.

Q.1) Using the TCGA-PANCAN32-L4.csv dataset (uploaded on LMS), reduce its dimension using the following techniques. These techniques may be implemented in R / Python. (Show the feature selection with cancer_type as well as sample_type labels)

1.) Linear Regression

2.) Ridge Regression

3.) LASSO Regression

4.) Random Forest

Show the selected features among the original features which these techniques used in processing. Also show the reduced number of features generated by these techniques from the processed features.

5.) Apply principle component analysis and reduce the number of produced components using cumulative variance percent of 85% and scree plot (take the values till the last breaking point)

After reducing the features (1-5 parts) give these reduced features as input to decision trees (implemented in R / Python) and compare the accuracy of each of the five reduction techniques produced by decision trees. Show the classification through decision trees with cancer_type as well as sample_type labels.

S.No	**Reduction Technique**	**Decision Tree Accuracy**
1.	Principle component analysis	
2.	Linear Regression	
3.	Ridge Regression	
4.	LASSO Regression	
5.	Random Forest	