Data Science
Assignment 03

Due Date: 3rd May 2021 **Total Marks: 100**

1. Run the following training data using Logistic Regression (Show one Epoch only). Start with weights = 0.5 and bias = 0.5. Clearly show all steps including loss function and values in forward and back propagation [25 Points]

F1	F2	F3	Class
1	2	3	A
2	3	4	A
7	6	4	B
8	7	3	B

2. Review the attached paper on inverse random sub-sampling and answer the following in your own words [25 Points]

   (a) What is class imbalance problem and main problems associated with it

   (b) What is the novelty in this paper. What is the role of FPR and TPR

   (c) Why Logistic Regression was successful as base classifier

   (d) What is bagging and what is the role of bagging in this paper?

3. Using kmeans algorithm and Euclidean distance to cluster the following 8 points into 3 clusters. Using A1 = (2,10), A2 = (2,5), A3 = (8,4), A4 = (5,8), A5 = (7,5), A6 = (6,4), A7 = (1,2) , A8 = (4,9). Consider initial seeds as A1, A4, and A7. Run algorithm for 1 iteration only. At the end of iteration 1, show [**25 Points**]

   - The new clusters (i.e. the examples belong to each cluster)
   - The center of the new clusters
   - Draw 10 × 10 space and all 8 points and show the clusters after 1st iteration and the new centroids
   - Without running algorithm again, guess how many more iterations are required to converge. Draw the result of each iteration

4. Using hierarchical clustering algorithms (Single, Complete, Group Average and Distance b/w centroids) and Euclidean distance to cluster the following 8 points into 3 clusters. Using A1 = (2,10), A2 = (2,5), A3 = (8,4), A4 = (5,8), A5 = (7,5), A6 = (6,4), A7 = (1,2) , A8 = (4,9). [**25 Points**]