Registration Number:

Date & session:

ST.JOSEPH'S UNIVERSITY, BENGALURU -27
M.Sc. BIG DATA ANALYTICS – II SEMESTER
SEMESTER EXAMINATION: APRIL 2023
(Examination conducted in May 2023)
BDA2321 – MACHINE LEARNING I
(For current batch students only)

Time: 2 Hours **Max Marks: 50**

This paper contains TWO printed pages and THREE parts

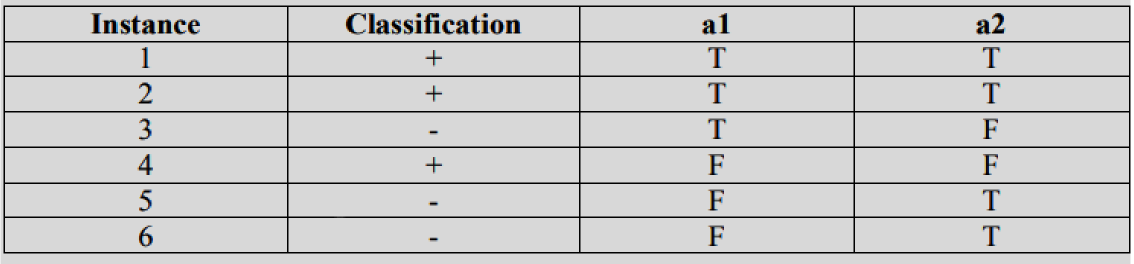
PART-A

ANSWER ALL THE QUESTIONS **5 x 2 = 10**

1. Justify the statement: Raw data has a significant impact on feature engineering process
2. Define confusion matrix with an example.
3. What is the significance of optimal separating hyperplane in SVM?
4. Give the general model of EM algorithm.
5. What do you mean by dimensionality reduction? Give an example.

ANSWER ANY FIVE QUESTIONS **5 x 4 = 20**

6. Consider the following set of training example:



Instance	Classification	a1	a2
1	+	T	T
2	+	T	T
3	-	T	F
4	+	F	F
5	-	F	T
6	-	F	T

What is the entropy of this collection of training example with respect to the target function classification?

7. Explain the principle of the gradient descent algorithm. Accompany your explanation with a diagram.
8. What is the general concept of an ensemble method? Explain bagging and boosting in ensemble method.
9. What is the purpose of k-means algorithm? Write down the basic algorithm for k-means and explain with help of a graphical example.
10. Write a note on Ensemble machine learning. Explain with concrete examples.
11. How PCA is different from SVD? How you will decide in which scenario which feature reduction technique is used?

ANSWER ANY TWO QUESTIONS **2 x 10 = 20**

12. If S is a collection of 14 examples with 9 YES and 5 NO examples in which one of the attributes is wind speed. The values of Wind can be Weak or Strong. The classification of these 14 examples are 9 YES and 5 NO. For attribute Wind, suppose there are 8 occurrences of Wind = Weak and 6 occurrences of Wind = Strong. For Wind = Weak, 6 of the examples are YES and 2 are NO. For Wind = Strong, 3 are YES and 3 are NO. Find the Entropy(weak) and Entropy(strong). Also calculate the information gain.
13. Use K Means clustering to cluster the following data into two groups. Assume cluster centroid are m1=2 and m2=4. The distance function used is Euclidean distance. { 2, 4, 10, 12, 3, 20, 30, 11, 25 }
14. Write short notes on the following.
    a) Linear Discriminant Analysis **(5)**
    b) Factor Analysis **(5)**