Roll No									

□□□□□□□□□ □□□□□□□□□□ □□□□□□□ □□□□

NATIONAL INSTITUTE OF TECHNOLOGY GOA

Farmagudi, Ponda, Goa, 403401

Programme Name: B.Tech.

End Semester Examinations, December 2020

Course Name: Data Warehousing and Data Mining
Course Code: CS505
Date: 19.12.2020
Time: 2.30 PM - 5.30 PM
Duration: 3 Hours
Max. Marks: 100

---

i) Answer all the questions **serially to the point**.
ii) Assume suitable data in case of missing.
iii) Irrelevant answers score negative marks.

---

1. Use an example to show why the *k*-means algorithm may not find the global optimum, that is, optimizing the within-cluster variation. [5 Marks]
2. Consider five points { X1, X2, X3, X4, X5} with the following coordinates as a two-dimensional sample for clustering: X1=(0,2), X2=(1,0), X3=(2,1), X4=(4,1) and X5= (5,3) Illustrate the K-means algorithm on the above data set. The required number of clusters is two, and initially, clusters are formed from the random distribution of samples: C1{X1, X2, X4} and C1{X3, X5}. [5 Marks]
3. Construct the decision tree classification model to classify the bank loan applications by assigning applications to one of the three risk classes using the Gini Index for selecting the attributes. [5 Marks]

Owns home?	Married	Gender	Employed	Credit rating	Risk class
YES	YES	MALE	YES	A	B
NO	NO	FEMALE	YES	A	A
YES	YES	FEMALE	YES	B	C
YES	NO	MALE	NO	B	B
NO	YES	FEMALE	YES	B	C
NO	NO	FEMLAE	YES	B	A
NO	NO	MALE	NO	B	B
YES	NO	FEMALE	YES	A	A
NO	YES	FEMALE	YES	A	C
YES	YES	FEMALE	YES	A	C

4. Suppose that a data warehouse consists of the four dimensions date, spectator, location, and game, and the two measures count and charge, where charge is the fare that a spectator pays when watching a game on a given date. Spectators may be students, adults, or seniors, with each category having its own change rate.
   i) Draw a star schema diagram for the data warehouse [5 marks]
   ii) Starting with base cuboid {date, spectator, location, game} what specific OLAP operations should one perform in order to list the total charge paid by student spectators at GM_Place in 2000? [5 marks]

5. Apply the Apriori algorithm to the following data set to discover strong association rules. [10 marks]

Trans ID	**Items Purchased**
101	Apple, Orange, Litchi, Grapes
102	Apple, Mango
103	Mango, Grapes, Apple
104	Apple, Orange, Litchi, Grapes
105	Pears, Litchi
106	Pears
107	Pears, Mango
108	Apple, Orange, Strawberry, Litchi, Grapes
109	Strawberry, Grapes
110	Apple, Orange, Grapes

The set of items is {Apple, Orange, Strawberry, Litchi, Grapes, Pears, Mango}. Use 0.3 for the minimum support value.

6. Construct the decision tree for the following training dataset using the decision tree Induction algorithm. [10 marks]

Age	Income	Student	Credit_rating	Buys_Computer
<=30	High	No	Fair	No
<=30	High	No	Excellent	No
31..40	High	No	Fair	Yes
>40	Medium	No	Fair	Yes
>40	Low	Yes	Fair	Yes
>40	Low	Yes	Excellent	No
31..40	Low	Yes	Excellent	Yes
<=30	Medium	No	Fair	No
<=30	Low	Yes	Fair	Yes
>40	Medium	Yes	Fair	Yes
<=30	Medium	Yes	Excellent	Yes
31..40	Medium	No	Excellent	Yes
31..40	High	Yes	Fair	Yes
>40	Medium	No	Excellent	No

7. Prove that in DBSCAN, the density-connectedness is an equivalence relation. [5 Marks]
8. 'DBSCAN works well for arbitrarily shaped clusters as well as detecting outliers as noise.' Is this statement true? Justify with proper examples. [5 Marks]
9. Show that accuracy is a function of sensitivity and specificity. [5 Marks]
10. Design a multidimensional cube with your own example. [5 Marks]
11. Demonstrate how Bayesian classification helps in predicting class membership probabilities. [5 Marks]
12. Draw the architecture of multi-tier data warehouse. [5 Marks]
13. Discuss various types of outliers with an example for each. [5 marks]

14. Write a short notes on the following: [2+2+2+2+2]
    i) kd-tree
    ii) k nearest neighbor
    iii) Voronoi diagram
    iv) Dunn Validity index
    v) Concept Hierarchy

15. State and discuss the following Conflict resolution strategies in Rule-based classification: [5 Marks]
    i) Size ordering
    ii) Rule Ordering

16. What is a confusion matrix? Explain with an example on classified samples. [5 Marks]