(3 hours) **[80 marks]**

NOTE:

1. Question No 1 is compulsory
2. Attempt any three questions from remaining.
3. Assume suitable data if necessary and state the same.

Q.1 [20]

A) Draw Data warehousing Architecture?
B) What is noisy data? How to handle noisy data?
C) Compare and contrast between OLTP and OLAP.
D) Explain concept of information gain and gini value used in decision tree algorithm.

Q.2

A) What is Data mining? Explain KDD process with diagram. [10]
B) Consider we have age of 29 participants in a survey given to us in sorted order. [10]
5, 10, 13, 15, 16, 16, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70, 85
Explain how to calculate mean, median, standard deviation, 1st and 3rd Quartile for given data and also compute the same. Show the Box and Whisker plot for this data.

Q.3

A) Explain market Basket Analysis with example. [10]

B) Consider Training dataset as given below. Use Naive Bayes Algorithm to determine whether it is advisable to play tennis on a day with hot temperature, rainy outlook, high humidity and no wind? [10]

Outlook	temperature	Humidity	Windy	Class
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Play
rain	mild	high	false	Play
rain	cool	normal	false	Play
rain	cool	normal	true	No
overcast	cool	normal	true	Play
sunny	mild	high	false	No
sunny	cool	normal	false	Play
rain	mild	normal	false	Play
sunny	mild	normal	true	Play
overcast	mild	high	true	Play
overcast	hot	normal	false	Play
rain	mild	high	true	No

Q.4

A) What is an outlier? Explain various methods for performing outlier analysis. [10]

B) Use the Apriori algorithm to identify the frequent item-sets in the following database. Then extract the strong association rules from these sets. Assume Min. Support = 50% Min. Confidence=75% [10]

Tid	a	b	c	d	e	f	g
Items	1,2,4,5,6	2,3,5	1,2,4,5	1,2,4,5	1,2,3,4,5,6	2,3,4	1,2,4,5

Q.5

A) Cluster the following eight points (with (x, y) representing locations) into three clusters: [10]
A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4), A7(1, 2), A8(4, 9)
Assume Initial cluster centers are at: A1(2, 10), A4(5, 8) and A7(1, 2).
The distance function between two points a = (x1, y1) and b = (x2, y2) is defined as- $P(a, b) = |x2 – x1| + |y2 – y1|$
Use K-Means Algorithm to find the three cluster centres after the second iteration.

B) Compare star schema, Snow flakes schema and star constellation [10]

Q.6 Write short note on following (Any 4) [20]

A) Dimensional Modeling.
B) Random Forest Technique.
C) Decision Tree Induction.
D) Cross Validation.
E) DBSCAN Algorithm

****