Regd No. | | | | | | | | | | |

VIGNAN'S
Foundation for Science, Technology & Research
UNIVERSITY
(Estd u/s 3 of UGC Act of 1956)

Regulation: R13 **Code No: CS425/2**

IV B. Tech II Semester Examinations – April 2017

DATA WAREHOUSING AND DATAMINING

Time: **3** hours Max. Marks: **60**

SECTION – A

(Short Answer Questions)

Answer all ten questions **10×1M=10M**

1. __________technology provide historical, current and predictive views of business operations.

2. _____________ is a statistic methodology that is most often used for numeric prediction although other methods exit as well.
   a) Relevance analysis b) Regression analysis c) Neural network d) None

3. A distribution can be visualization popularly by _______
   a) Outliers b) quantile plot c) boxplot d) quantile –quantile plot

4. The process of reducing the number of random variables or attributes under consideration is
   a) Dimensionality reduction b) Data reduction c) Data cleaning d) all

5. __________is a sub cube that is small enough to fit into the memory available for cube computation.

6. Which property is used to reduce the search space of the level –wise generation of frequency item sets?
   a) Prune b) Apriori c) join d) subset

7. _______________ is the learning of decision trees from class – labeled training tuples.

8. Neutral network learning is also referred to as __________
   a) Connectionist b) machine learning c) All the above d) none of the above

9. ________ method works by grouping data objects into a hierarchy or 'tree' of clusters.
   a) Hierarchical clustering b) K-means clustering c) Clustering d) K-nearest neighbor

10. _______is a graph clustering algorithm that searches graph to identify well –connected components as clusters

SECTION – B

Answer all five questions **5×2M= 10M**

11. How is a data warehousing different from a database? How are they similar?

12. Normalize the data set to make the norm of each data point equal to 1 use Euclidean distance on the transformed data to rank the data points.

13. Propose an algorithm that computers closed iceberg cubes efficiently.

14. Write any two issues of Data Mining.

15. Explain the need of data pre-processing in knowledge extraction.

SECTION – C

Answer all four questions **4×5M = 20M**

16. What is data mining? Outline the major research challenges of data mining in one specific application domain, such as stream /sensor data analysis spatiotemporal data analysis or bioinformatics.

(OR)

17. Describe three challenges to data mining regarding data mining methodology and user interaction issues.

18. Suppose that a hospital tested the age and body fat data for 18 randomly selected adults with the following results:

Age	23	23	27	27	39	41	47	49	50
% fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

Age	52	54	54	56	57	58	58	60	61
% fat	34.5	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35

a) Calculate the mean, median and standard deviation of age and % fat.
b) Draw the box plots for age and % fat.
c) Draw a scatter plot and a q-q plot based on these two variables

(OR)

19. What are the values ranges of the following normalization methods?
a) Min –max normalization
b) z-score normalization
c) z-score normalization using the mean absolute deviation instead of standard deviation
d) Normalization by decimal scaling

20. Discuss how you might extend the star –Cubing algorithm to compute iceberg cubes where the iceberg condition tests for an average that is no bigger than some value v.

(OR)

21. Give a short example to show that items in strong association rule actually many be negatively correlated.

22. Explain single dimensional association rule mining with Apriori algorithm and also suggest any two improvements?

(OR)

23. Briefly describe a note on multi-level association rule mining with necessary examples?

SECTION – D

Answer all two questions **2×10M= 20M**

24. a) Briefly outline the major steps of decisions tree classification

    b) Why naive Bayesian classification is called naive? Briefly outline the major ideas of naive Bayesian classification with an example.

(OR)

25. a) Write an algorithm for k-nearest neighbor classification given K, The nearest number of neighbors and n the number of attributes describing each tuple.

    b) Briefly describe the classification process using a) genetic algorithm b) rough sets and c) fuzzy sets.

26. Briefly describe and give example of each of the following approaches to clustering: Partitioning Methods, Hierarchical Methods, Density-based Methods And Grid-Based Methods.

(OR)

27. a) Example about Model –Based Clustering methods in detail.

    b) Discuss about grid based methods.