



Reg. No. :

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

Question Paper Code : X 20760

B.E./B.Tech. DEGREE EXAMINATIONS, NOVEMBER/DECEMBER 2020

Seventh Semester

Information Technology

IT 6006 – DATA ANALYTICS

(Common to Computer Science and Engineering)

(Regulations 2013)

(Also Common to PTIT 6006 – Data Analytics for B.E. Part-time – Seventh Semester –
Computer Science and Engineering)

(Regulations 2014)

Time : Three Hours

Maximum : 100 Marks

Answer ALL questions

PART – A

(10×2=20 Marks)

1. State the characteristics of big data.
2. Define Prediction error.
3. Demonstrate how an outlier can affect regression analysis with an example.
4. What is the need for Kernel methods in SVM ?
5. How to filter the data streams using Bloom Filter ?
6. What are the steps to be followed for counting distinct elements in a stream ?
7. How can the initial number of clusters for k-means algorithm be estimated ?
8. List frequent item sets mining methods.
9. Create a histogram for the data set {0, 1, 2, 3, 4, 7, 8, 9}. Compute the interval space for 3 intervals.
10. Define HDFS.

PART – B

(5×13=65 Marks)

11. a) Explain in detail about challenges of conventional system for Data Analytics.

(OR)

- b) Discuss the following statistical concepts with example
Sampling distributions, resampling and statistical inference.



12. a) For the following data set, find the first principal component.

x	y
2.5	2.4
0.5	0.7
2.2	2.9
1.9	2.2
Data = 3.1	3.0
2.3	2.7
2	1.6
1	1.1
1.5	1.6
1.1	0.9

(OR)

b) Explain in detail about construction of the Fuzzy model from data with example scenario.

13. a) Discuss in detail about estimating moments and counting oneness in a window in a stream.

(OR)

b) Explain in detail about realtime analytics platform.

14. a) Discuss Apriori algorithm in detail for the following super market scenarios.

Transaction ID	Onion	Potato	Burger	Milk	Ghee
t_1	1	1	1	0	0
t_2	0	1	1	1	0
t_3	0	0	0	1	1
t_4	1	1	0	1	0
t_5	1	1	1	0	1
t_6	1	1	1	1	1

(OR)

b) Discuss in detail about the frequent pattern based clustering methods.



15. a) Consider the case study of instagram. What are the NoSQL databases that can be used ? Justify your answer. Compare and contrast various NoSQL databases.

(OR)

b) Discuss in detail the need for Hadoop Framework and HDFS in online shopping system.

PART – C

(1×15=15 Marks)

16. a) Construct a multilayer feed-forward network for a training tuple $X = (0, 1, 1)$ whose class label is 1. Assume 2 hidden nodes in 1 hidden layer. The nodes are numbered as follows : Input nodes (1, 2, 3), Hidden nodes (4, 5), Output node (6). The initial weights and bias are taken by own assumptions. Use back propagation algorithm and find the updated weights and bias. What are the advantages and disadvantages of this classifier ?

(OR)

b) Formulate a suitable algorithm to create clusters using k-means using a given dataset and a given K. Use this algorithm to cluster the following dataset into two clusters with feature vectors Point 1 and Point 2 as the initial cluster centroids.

	X	Y
Point 1	2.0	2.0
Point 2	5.0	7.0
Point 3	1.0	1.0
Point 4	7.0	10.0

Show the final clusters with the points in the cluster. List all your assumptions.
