C

Reg. No. : | | | | | | | | | | |

Question Paper Code: 57203

B.E. / B.Tech. DEGREE EXAMINATION, NOV 2018

Seventh Semester

Computer Science and Engineering

15UCS703 - DATA SCIENCE

(Regulation 2015)

Duration: Three hours Maximum: 100 Marks

Answer ALL Questions

PART A - (5 x 1 = 5 Marks)

1. _____ perform sanity checks against domain knowledge and decide if the dirty data needs to be eliminated. CO1-R

   (a) Data Engineer (b) Data Analysts

   (c) Both a&b (d) None of the mentioned

2. What is the minimum no. of variables/ features required to perform clustering? CO2U

   (a) 0 (b) 1 (c) 2 (d) 3

3. Hadoop is a framework that works with a variety of related tools, Common cohorts include: CO3-U

   (a) MapReduce, Hive and HBase (b) MapReduce, MySQL and Google Apps

   (c) MapReduce, Hummer and Iguana (d) MapReduce, Heron and Trumpet

4. A file in HDFS that is smaller than a single block size CO4-U

   (a) Cannot be stored in HDFS

   (b) Occupies the full block's size

   (c) Occupies only the size it needs and not the full block

   (d) Can span over multiple blocks

5. The number of map is usually driven by the total size of_____. CO5-U

   (a) Inputs (b) Outputs (c) Tasks (d) None of these

PART – B (5 x 3= 15 Marks)

6.	Define Data Science.	CO1-U	
7.	Define Decision trees and its varieties.	CO1- U	
8.	Write short notes on When to Consider a Big Data Solution.	CO2-U	
9.	Why is a block in HDFS so large?	CO3-U	
10.	Mention the functionality of mappers and reducers.	CO3- U	

PART – C (5 x 16= 80Marks)

11.	(a)	(i) Describe about the Descriptive Statistics.	CO1- U	(8)
		(ii) Briefly explain the various data types and attributes of R.	CO1- U	(8)
		Or		
	(b)	List out the R functions used to visualizing a single variable and examining multiple variables and explain it with example.	CO1- U	(16)
12.	(a)	Illustrate the method to find k clusters from a collection of M objects with n attributes using k-means clustering.	CO2-U	(16)
		Or		
	(b)	(i) John flies frequently and likes to upgrade his seat to first class. He has determined that if he checks in for his flight at least two hours early, the probability that he will get an upgrade is 0.75; otherwise, the probability that he will get an upgrade is 0.35. With his busy schedule, he checks in at least two hours before his flight only 40% of the time. Assume John did not receive an upgrade on his most recent attempt. By using Bayes theorem identify What is the probability that he did not arrive two hours early?	CO2-App	(8)
		(ii) Explore two methods of using the Naive's Bayes classifier in R.	CO2-U	(8)
13.	(a)	(i) Explain the characteristics of Big Data.	CO3- U	(10)
		(ii) Briefly explain the Fraud Detection Pattern in Big Data.	CO3- U	(6)
		Or		

	(b)	What is Hadoop? Explain the various components of Hadoop**.**	CO3-U	(16)
14.	(a)	(i) What is HDFS? Explain the architecture of HDFS with neat diagram.	CO4-U	(10)
		(ii) Write short notes on HDFS Federation.	CO4-U	(6)
		Or		
	(b)	Describe how the files are read from HDFS and written to the HDFS by the client.	CO4-U	(16)
15.	(a)	What is map reduce? Explain the Map Reduce Execution Pipeline.	CO5-U	(16)
		Or		
	(b)	Describe the implementation of Hadoop word count using Map Reduce Application.	CO5-U	(16)

57203