Total No. of Questions : 8]

SEAT No. :

PE-4223

[Total No. of Pages : 3

[6583]-35
T.E. (Computer Engineering)
DATA SCIENCE AND BIG DATA ANALYTICS
(2019 Pattern) (Semester - VI) (310251)

Time : 2½ Hours] *[Max. Marks : 70*

Instructions to the candidates:

1) ***Answer Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q. 7 or Q.8.***
2) ***Neat diagrams must be drawn whenever necessary.***
3) ***Figures to the right indicate full marks.***
4) ***Assume suitable data if necessary.***
5) ***Use of Scientific Calculator is permitted.***

Q1) a) What is the data Preparation phase in Data Analytics Lifecycle. What is the Analytics Sandbox and ETLT process in this phase? **[8]**

b) Compare the data acquisition phase with the data preprocessing phase in terms of inputs, outputs, and objectives. **[8]**

OR

Q2) a) List out the activities to be carried out in model planning and model building phase. What are different tools used for these phases? **[8]**

b) For a dataset of online retail transactions, which steps of the Big Data Analytics Lifecycle would you follow to build a recommendation system? **[8]**

Q3) a) What is logistic regression, and how does it differ from linear regression? What is the sigmoid Function, and what role does it play in logistic regression? **[9]**

b) Differentiate between predictive and descriptive analytics. What are NumPy and Pandas used for in Python? List any three essential Python libraries for data analysis **[9]**

OR

P.T.O.

Q4) a) Apply Naïve Bayes classification to a simple weather dataset to predict ''P1ay/Tennis''. **[9]**

Day	Outlook	Play Tennis
1	Sunny	No
2	Sunny	No
3	Overcast	Yes
4	Rain	Yes
5	Rain	Yes
6	Rain	No
7	Overcast	Yes

b) Explain how a decision tree splits data at each node with suitable example. Describe the role of entropy and information gain in determining the best split. **[9]**

Q5) a) Suppose you have the following dataset containing the coordinates of points in a 2-dimensional space : **[9]**

Point	X Coordinate	Y Coordinate
A	2	3
B	4	7
C	3	5
D	6	9
E	8	6
F	7	8

Perform K-means clustering on this dataset With K = 2. Assume the initial centroids to be (2,3) and (8,6). Compute the new centroids after each iteration until convergence, and assign points to their nearest centroids.

b) How do you handle noise and irrelevant information in text data during preprocessing? Explain the terms bag of words and TF IDF in text analytics. **[9]**

OR

Q6) a) Explain the difference between agglomerative and divisive clustering. Compare single complete, and average linkage methods. **[9]**

b) What is the purpose of each of ROC, AUC and an elbow plot? Describe how each of these plots are constructed with suitable example. **[9]**

Q7) a) What is Bar chart, Pie chart and Line chart? Compare the use of bar charts and pie charts for categorical data. Analyze how visualization helps in identifying outliers. **[9]**

b) Compare HDFS and traditional file systems. Define NameNode and DataNode. Analyze how NameNode failure affects the Hadoop cluster. **[9]**

OR

Q8) a) What is a box plot? Explain the different components of a box plot? How do you interpret the median, quartiles, and whiskers in a box plot? What does the Interquartile Range (IQR) represent in a box plot? **[9]**

b) Explain why Pig is called a data flow language. Describe the role of the Pig execution engine. Explain why Hive is suitable for data warehousing. **[9]**