Total No. of Questions : 8]

SEAT No. : □

PD-5303

[Total No. of Pages : 2

[6403]-43

T.E. (Computer Engineering)

DATA SCIENCE AND BIG DATA ANALYTICS

(2019 Pattern) (Semester - VI) (310251)

Time : 2½ Hours] *[Max. Marks : 70*

Instructions to the condidates:

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

Q1) a) What is Model Building elaborate this phase of data analytics with the help of a suitable example? **[8]**

b) List out different stakeholders of an analytics project. What do they usually expect at the conclusion (key outputs) of a project? **[9]**

OR

Q2) a) Explain Descriptive, Diagnostic, Predictive analytics. **[8]**

b) List and explain the various activities involved in identifying potential data resources as a part of discovery phase in Data Analytics Life Cycle? **[9]**

Q3) a) What is association rule mining? Describe the working of the Apriori algorithm with an example. **[9]**

b) Explain how decision trees are constructed using information gain and entropy. Illustrate with a small example. **[9]**

OR

Q4) a) Explain Naïve Bayes' classifier and its applications. **[9]**

b) Consider a dataset with binary classes and two features: "Loan Amount" and "Default History." Show how logistic regression could be applied for loan default prediction. **[9]**

P.T.O.

Q5) a) Explain the holdout method. Differentiate training set, validation set, and test set. **[8]**

b) Given the confusion matrix below. Calculate Accuracy, Precision, Recall and F1-score. **[9]**

	Predicted Yes	**Predicted No**
Actual Yes	70	30
Actual No	20	80

OR

Q6) a) Explain the following Text Analysis steps with suitable example **[8]**

i) Part-of-speech (POS) tagging

ii) Lemmatization

b) Use K-Means Clustering for the following points and determine the centroids after one iteration. Assume initial centroids as A(1,l), B(5,7). Points: (1,2), (2,1), (3,5), (6,8), (7,6), (5,5) **[9]**

Q7) a) Explain Hadoop Architecture with a neat diagram. Highlight the roles of NameNode and DataNode. **[9]**

b) Compare Tableau, Power BI, and Matplotlib for data visualization. Discuss scenarios where each tool is best suited. **[9]**

OR

Q8) a) What is Data Visualization? Describe the challenges of data visualization. **[9]**

b) Write short notes on the following : **[9]**

i) Map Reduce

ii) HDFS

iii) Hive