Total No. of Questions : 8]

SEAT No. : □

P-7545

[Total No. of Pages : 3

[6180]-53

T.E. (Computer Engineering)

DATA SCIENCE AND BIG DATA ANALYTICS

(2019 Pattern) (Semester - II) (310251)

Time : 2½ Hours] *[Max. Marks : 70*

Instructions to the candidates :

1) *Answer Q1 or Q2, Q3 or Q4, Q5 or Q6. Q7 or Q8.*
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 permitted.*

Q1) a) Explain Data Analytics Cycle with suitable diagram and its phases. **[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]**

OR

Q2) a) List and explain the key roles for successful analytics project. **[8]**

b) Write short note on : **[9]**

i) Common Tools for the Model Building

ii) Model selection for Data Analytics

Q3) a) List and explain the various types of analytics in Big data. **[9]**

b) Calculates the support and confidence value for all the possible item sets.**[9]**

Transaction ID	Items bought
1	Onion, Potato, Cold Drink
2	Onion, Burger, Cold Drink
3	Eggs, Onion, Cold Drink
4	Potato, Milk, Eggs
5	Potato, Burger, Cold Drink, Milk, Eggs

OR

P.T.O.

Q4) a) Explain the need of logistic regression along with its various types. **[9]**

b) Explain the following terms with suitable example. **[9]**

i) Removing Duplicates from dataset.

ii) Handling Missing Data

Q5) a) Suppose that the given data the task is to cluster points (with (x, y) representing location) into three clusters, where the points are A1 (2, 10), A2(2, 5), A3(8, 4), B1(5, 8), B2(7, 5), B3(6, 4), C1(1, 2), C2(4, 9). The distance function is Euclidean distance. Suppose initially we assign A1, B1 and C1 as the center of each cluster, respectively. **[8]**

Use the k-means algorithm to show only show only the first round of execution with cluster center.

b) Explain the following Text Analysis steps with suitable example **[9]**

i) Part-of-speech(POS)tagging

ii) Lemmatization

OR

Q6) a) Given the confusion matrix, Calculate Accuracy, Precision, Recall, Error rate with description on Diabetic Risk. **[8]**

		Predicted classes	
	Classes	Diabetic Risk -Yes	Diabetic Risk -No
Actual classes	Diabetic Risk-Yes	90	210
	Diabetic Risk-No	140	9560

b) Explain the Text Preprocessing steps with suitable example. **[9]**

Q7) a) List the few data visualization tools and discuss any four applications of data visualization along with the use of the various plots with Python/R or suitable tool. **[9]**

b) List the challenges of Data Visualization. Explain the types of visualization with example. **[9]**

OR

Q8) a) Explain in detail the Hadoop Ecosystem with suitable diagram along with the various components. **[9]**

b) Write a short note on the following. **[9]**

a) Map Reduce

b) Pig