Total No. of Questions : 8]

SEAT No. :

P812

[Total No. of Pages : 2

[5870] - 1133
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 Q.1 or Q.2, Q.3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.***
1) ***Neat diagrams must be drawn whenver necessary.***
2) ***Figures to the right side indicate full marks.***
3) ***Use of logarithmic tables slide rule, mollier charts, electronic pocket calculator and steam tables is allowed.***
4) ***Assume suistable data, if necessary.***

Q1) a) What is driving data deluge? Explain with one example. **[9]**

b) What is data science? Differentiate between Business Intelligence and Data Science. **[9]**

OR

Q2) a) What are the sources of Big Data. Explain model building phase with example. **[9]**

b) Explain big data analytics architecture with diagram. What is data discovery phase. Explain with example. **[9]**

Q3) a) Explain various data pre-processing steps. Discuss essential python libraries for preprocessing. **[8]**

b) What are association rules? Explain Apriori Algorithm in brief. **[9]**

OR

Q4) a) Explain the following

i) Linear Regression

ii) Logistic Regression **[8]**

b) Explain scikit-learn library for matplotlib with example. **[9]**

P.T.O.

Q5) a) Write short note on

i) Time series Analysis

ii) TF - IDF. **[9]**

b) What is clustering? With suitable example explain the steps involved in k - means algorithm. **[9]**

OR

Q6) a) Write short note on

i) Confusion matrix

ii) AVC - ROC curve **[9]**

b) Discuss Holdout method and Random Sub Sampling methods. **[9]**

Q7) a) With a suitable example explain Histogram and explain its usages. **[8]**

b) Describe the Data visualization tool "Tableau". Explain its applications in brief. **[9]**

OR

Q8) a) With a suitable example explain and draw a Box plot and explain its usages. **[8]**

b) Describe the challenges of data visualization. Draw box plot and explain its usages. **[9]**

Total No. of Questions : 8]

SEAT No. : [ ]

PA-1449

[Total No. of Pages : 3

[5926]-65

T.E. (Computer Engg.)

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, and Q7 or Q8.***
2) ***Neat diagram must be drawn wherever necessary.***
3) ***Figures to the right indicate full makrs,***
4) ***Use of logarithmic tables slide rule, mollier charts, electronic pocket calculator and steam tables is allowed.***
5) ***Assume suitable data if necessary.***

Q1) a) Draw the diagram of data analytics life cycle in big data and briefly explain its phases. **[8]**

b) Explain in detail how the model building phase is built by team in data analytics life cycle? **[9]**

OR

Q2) a) List and explain the steps in data preparation phase of data analytics life cycle. **[8]**

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

i) ETL

ii) Common tools for the model building.

iii) Model selection for data analytics.

Q3) a) What are the types of analytics in big data? Explain in brief. **[9]**

b) Calculate 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 use of logistic function in logistic regression in detail. **[9]**

b) Write short note on the following:

i) Removing duplicates from data set.

ii) Handling missing data

iii) Data transformation. **[9]**

Q5) a) Suppose that the given data the taste is to cluster points (With (x.y) representing location) into three cluster, 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. use the k-means algorithm to show only the three cluster centers after the first round of execution with steps. **[9]**

b) Explain the following text analysis steps with suitable example. **[8]**

i) Part of speech (POS) tagging

ii) Lemmatization

iii) Stemming

OR

Q6) a) Given the confusion matrix, calculate accuracy. precision, Recall, Error rate with description on heart attact risk. **[8]**

		Predicted classes	
	Classes	Heart-Attack Risk-yes	Heart Attack Risk-No
Actual Classes	Heart Attack Risk-yes	80	220
	Heart Attack Risk-No	150	9,500

b) Explain the TF/IDF (term frequency-inverse document frequency) terms in text analysis with suitable example. **[9]**

Q7) a) List the data visualization tools and discuss any four applications of data visualization along with the use of the suitable plot. **[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 **[9]**

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

i) Map reduce.

ii) Pig

iii) Hive

● ● ●

Total No. of Questions : 8]

SEAT No. :

P-3153

[Total No. of Pages : 2

[6003]-354

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 Q.1 or Q.2, Q3 or Q.4, Q.5 or Q.6, Q.7 or Q.8.***
2) ***Neat diagram must be drawn whenever necessary.***
3) ***Figures to the right indicate full marks.***
4) ***Assume suitable data if necessary.***
5) ***Use of Scientific Calculalor is permitted.***

Q1) a) What is Model Building elaborate this phase of data analytics with the help of suitable example. **[9]**

b) Explain any three sources of Big Data. Differentiate BI versus Data science. **[8]**

OR

Q2) a) What are the three characteristic of Big Data and what are the main consideration in processing Big Data. **[8]**

b) Explain Descriptive, Diagnostic, Predictive analytics. **[9]**

Q3) a) Explain why decision tree are used. Draw a sample decision tree and explain its parts. **[9]**

b) How Apriori Algorithm works, explain with suitable example? **[9]**

OR

Q4) a) What is data preprocessing? Explain in details about handling missing data and transformation of data. **[9]**

b) Explain Naïve Bayes' classifier and it applications. **[9]**

P.T.O.

Q5) a) What is text processing? Explain TF-IDF with example. **[8]**

b) With suitable example ,explain the steps involved in k-means algorithm. **[9]**

OR

Q6) a) Define following terms with respect to confusion matrix : **[8]**

i) Accuracy

ii) Precision

iii) Recall

iv) AUC-ROC

b) Explain k-fold Cross Validation & Random Subsampling. **[9]**

Q7) a) With a suitable example, draw a Histogram, boxplot and explain its usages. **[9]**

b) Describe the data visualization tool Tableau. List of data visualization tools. **[9]**

OR

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

b) Explain architecture of Apache-Pig. **[9]**

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

❍❍❍❍