SPPU-TE-COMP-CONTENT – KSKA Git

Total No. of Questions : 8]

SEAT No. :

PB4430

[6262]-43

[Total No. of Pages : 3

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

b) List out different stakeholders of an analytics project. What they usually expect at the conclusion (key outputs) of a project? **[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) What is linear regression, and what are its primary objectives? What is the difference between simple linear regression and multiple linear regression? How do you evaluate the performance of linear regression?**[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) Suppose you are given a dataset containing information about whether emails are spam or not spam, along with two features: the presence of the word "offer" (1 for present, 0 for absent) and the presence of the word "free" (1 for present, 0 for absent). You are tasked with classifying a new email with the following feature values: "offer"=1 and "free"=1.**[9]**

P.T.O.

CEGP013091 49.248.216.238 15/05/2024 09:42:06 static-238

Given the training dataset:

Email	Offer	Free	Spam
1	1	0	No
2	0	1	Yes
3	1	1	Yes
4	0	1	No
5	1	1	Yes

Calculate the probability that the new email is spam using Naive Bayes.

OR

Q4) a) How does the Apriori algorithm discover frequent itemsets in a dataset? What is the role of support and confidence in the context of association rule mining using the Apriori algorithm? **[9]**

b) Explain the process of building a decision tree? What are the criteria used for splitting nodes in a decision tree? **[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.

CEGP013091 49.248.216.238 15/05/2024 09:42:06 static-238

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 how hierarchical clustering can be used for visualizing hierarchical relationships in data with suitable example? What are some real-world applications of hierarchical clustering? **[9]**

b) What is the holdout method, and how does it work? Explain the difference between training set, validation set, and test set in the holdout method.**[9]**

Q7) a) What is a histogram? How is it used to visualize the distribution of data? How is it different from a density plot? **[9]**

b) What is the Hadoop ecosystem, and what are its primary components? What is MapReduce, and how does it fit into the Hadoop ecosystem?**[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 the role of Apache Pig in data processing workflows on Hadoop? What is Apache Spark, and how does it complement Hadoop for big data processing? **[9]**

CEGP013091 49.248.216.238 15/05/2024 09:42:06 static-238