Course Code: 18CS0541 **R18**

SIDDARTHA INSTITUTE OF SCIENCE AND TECHNOLOGY:: PUTTUR
(AUTONOMOUS)
Siddharth Nagar, Narayanavanam Road – 517583

QUESTION BANK (DESCRIPTIVE)

Subject with Code: DATA SCIENCE (18CS0541) **Course & Branch**: B.Tech - CSE

Year &Sem: IV-B.Tech & I-Sem **Regulation:** R18

UNIT –I
INTRODUCTION TO R & DATA ANALYSIS

1	**a**	List the five main skills and behavioral characteristics of Data scientists.	[L1][CO1]	**[2M]**
	b	Give the difference between BI and Data science	[L4][CO1]	**[2M]**
	c	Sketch the typical analytic architecture.	[L3][CO1]	**[2M]**
	d	Define the window panes of RStudio GUI	[L1][CO1]	**[2M]**
	e	What is exploratory data analysis	[L1][CO1]	**[2M]**
2	**a**	Sketch Date Analytics Lifecycle diagram.	[L3][CO1]	**[4M]**
	b	Explain in detail each stages of Data Analytics Lifecycle	[L2][CO1]	**[6M]**
3		Illustrate in detail the data types of R.	[L3][CO1]	**[10M]**
4	**a**	Expand and Distinguish between NOIR attributes	[L4][CO1]	**[5M]**
	b	Describe the Array function in R	[L2][CO1]	**[5M]**
5		Examine the functions used for Visualizing a Single Variable	[L3][CO5]	**[10M]**
6		Discriminate about the functions used for examining Multiple Variables	[L5][CO1]	**[10M]**
7		Give example and define the following function of R a) read() b) head() c) summary() d) plot() e) summary()	[L2][CO1]	**[10M]**
8	**a**	Discuss the Data frames utilization in R	[L2][CO1]	**[5M]**
	b	Describe the List function in R	[L2][CO1]	**[5M]**
9	**a**	Define dirty data	[L1][CO5]	**[2M]**
	b	How dirty data can be detected in the data exploration phase with Visualizations	[L2][CO5]	**[8M]**
10	**a**	Illustrate the importance of visualizing data before analysis	[L2][CO5]	**[5M]**
	b	Justify “Using visualization for data exploration is different from presenting results to stakeholders”	[L6][CO5]	**[5M]**

Course Code: 18CS0541 **R18**

UNIT –II
STATISTICAL METHODS FOR EVALUATION & ASSOCIATION RULES

1	**a**	Define Hypothesis Testing	[L1][CO3]	**[2M]**
	b	What is Power of test?	[L1][CO3]	**[2M]**
	c	How the Sample Size and effect size are related	[L2][CO3]	**[2M]**
	d	Mathematically define Confidence	[L1][CO3]	**[2M]**
	e	What is meant by downward closure property	[L1][CO3]	**[2M]**
2	**a**	Differentiate Null Hypotheses and Alternative Hypotheses	[L4][CO3]	**[5M]**
	b	Examine the application property of Wilcoxon rank-sum test	[L3][CO3]	**[5M]**
3		Discriminate about Difference of Means	[L5][CO2]	**[10M]**
4		Suppose everyone who visits a retail website gets one promotional offer or no promotion at all. We want to see if making a promotional offer makes a difference. What statistical method would you recommend for this analysis?	[L5][CO2]	**[10M]**
5		Explain the following a) Student's t-test b) Welch's t-test	[L2][CO2]	**[10M]**
6	**a**	Define and Detail ANOVA.	[L1][CO3]	**[5M]**
	b	How evaluation of Candidate Rules are done?	[L2][CO3]	**[5M]**
7	**a**	What is a type I error? What is a type II error? Is one always more serious than the other? Why?	[L1][CO3]	**[5M]**
	b	Give the difference between Validation and Testing	[L4][CO5]	**[5M]**
8	**a**	State Apriori Algorithm	[L1][CO3]	**[2M]**
	b	Explain Apriori Algorithm with example	[L2][CO3]	**[8M]**
9	**a**	List and Discuss the four measures of significance of Association rules	[L1][CO3]	**[5M]**
	b	Give the Applications of Association Rules	[L1][CO3]	**[5M]**
10		Illustrate any five approaches to improve Apriori's efficiency when the dataset is large.	[L3][CO3]	**[10M]**

Course Code: 18CS0541 **R18**

UNIT –III
REGRESSION & CLASSIFICATION

<table>
<tr><td>1</td><td>a</td><td>What is Multicollinearity?</td><td>[L1][CO4]</td><td>[2M]</td></tr>
<tr><td></td><td>b</td><td>Relate Ridge regression and Lasso regression</td><td>[L1][CO4]</td><td>[2M]</td></tr>
<tr><td></td><td>c</td><td>Which two basic measures does the entropy methods select the most informative attribute?</td><td>[L1][CO4]</td><td>[2M]</td></tr>
<tr><td></td><td>d</td><td>Define confusion matrix</td><td>[L1][CO4]</td><td>[2M]</td></tr>
<tr><td></td><td>e</td><td>What is accuracy? FPR and FNR</td><td>[L1][CO4]</td><td>[2M]</td></tr>
<tr><td>2</td><td></td><td>Explain the analytical technique Linear Regression with its model description.</td><td>[L2][CO4]</td><td>[10M]</td></tr>
<tr><td>3</td><td></td><td>Discuss the following with respect to linear regression
a) Categorical Variables
b) Confidence Intervals on the Parameters
c) Confidence Interval on the Expected Outcome
d) Prediction Interval on a Particular Outcome</td><td>[L2][CO4]</td><td>[10M]</td></tr>
<tr><td>4</td><td>a</td><td>Justify the usage of linear regression and logistic regression.</td><td>[L6][CO4]</td><td>[3M]</td></tr>
<tr><td></td><td>b</td><td>Illustrate Logistic Regression Model.</td><td>[L3][CO4]</td><td>[7M]</td></tr>
<tr><td>5</td><td>a</td><td>Describe Decision Trees in detail with example.</td><td>[L2][CO4]</td><td>[5M]</td></tr>
<tr><td></td><td>b</td><td>Explain the algorithm of decision tree</td><td>[L3][CO4]</td><td>[5M]</td></tr>
<tr><td>6</td><td></td><td>Intercept the decision trees algorithms</td><td>[L4][CO4]</td><td>[10M]</td></tr>
<tr><td>7</td><td>a</td><td>State Bayes’ Theorem</td><td>[L1][CO4]</td><td>[2M]</td></tr>
<tr><td></td><td>b</td><td>Discuss Naïve Bayes classification method considering an example</td><td>[L2][CO4]</td><td>[8M]</td></tr>
<tr><td>8</td><td></td><td>How does one pick the most suitable method for a given classification problem?</td><td>[L2][CO4]</td><td>[10M]</td></tr>
<tr><td>9</td><td>a</td><td>Compare the C4.5 and CART algorithm of decision tree.</td><td>[L4][CO4]</td><td>[4M]</td></tr>
<tr><td></td><td>b</td><td>Discriminate the way how the evaluation of decision tree is done</td><td>[L5][CO4]</td><td>[4M]</td></tr>
<tr><td></td><td>c</td><td>Give the two approaches that help avoid over fitting in decision tree learning.</td><td>[L2][CO4]</td><td>[2M]</td></tr>
<tr><td>10</td><td></td><td>Consider the given confusion Matrix of Naïve Bayes from the Bank Marketing Dataset
<table>
<tr><th colspan="5">Predicted Class</th></tr>
<tr><td></td><td></td><td>Subscribed</td><td>Not Subscribed</td><td>Total</td></tr>
<tr><td></td><td>Subscribed</td><td>3</td><td>8</td><td>11</td></tr>
<tr><td>Actual Class</td><td>Not Subscribed</td><td>2</td><td>87</td><td>89</td></tr>
<tr><td>Total</td><td></td><td>5</td><td>95</td><td>100</td></tr>
</table>
Calculate the following defining them
a) Accuracy
b) TPR
c) FPR
d) FNR
e) Precision</td><td>[L4][CO4]</td><td>[10M]</td></tr>
</table>

Course Code: 18CS0541 **R18**

UNIT –IV
CLUSTERING & TIME SERIES ANALYSIS

1	**a**	What is clustering?	[L1][CO4]	**[2M]**
	b	State the advantage of using PAM.	[L1][CO4]	**[2M]**
	c	How hierarchical agglomerative clustering is different from density based clustering?	[L2][CO4]	**[2M]**
	d	Define time series and give the goals of time series analysis	[L1][CO4]	**[2M]**
	e	List the various parts of ARIMA model	[L1][CO6]	**[2M]**
2		Illustrate the method to find k clusters from a collection of M objects with n attributes.	[L3][CO6]	**[10M]**
3	**a**	Explain any one case study for time series analysis	[L2][CO5]	**[5M]**
	b	What is forecasting in association with time series. Explain	[L1][CO6]	**[5M]**
4	**a**	Indicate when the time series $y_t^{\text{for } t=1,2,3,\ldots}$ is said to be stationary time series.	[L2][CO6]	**[4M]**
	b	Express the stationary time series conditions in detail.	[L6][CO6]	**[6M]**
5		Discuss in detail each part of the ARIMA model	[L2][CO4]	**[10M]**
6	**a**	List and explain time series components	[L1][CO6]	**[5M]**
	b	Discriminate the steps involved in Box-Jenkins Methodology	[L5][CO6]	**[5M]**
7	**a**	What is meant by k-means	[L1][CO4]	**[2M]**
	b	Describe k-means algorithm to find k clusters	[L2][CO4]	**[8M]**
8		Correlate ARMA and ARIMA Models	[L4][CO6]	**[10M]**
9		Express the following a) Autocorrelation Function b) Autoregressive Models	[L2][CO6]	**[10M]**
10		List and describe Additional time series methods	[L2][CO6]	**[10M]**

Course Code: 18CS0541 **R18**

UNIT –V
TEXT ANALYSIS

1	**a**	What is text analytics?	[L1][CO6]	**[2M]**
	b	Define Porter's stemming algorithm.	[L1][CO6]	**[2M]**
	c	Define Confusion matrix.	[L1][CO6]	**[2M]**
	d	What is Topic modeling ?	[L1][CO6]	**[2M]**
	e	State what is meant by word cloud?	[L1][CO6]	**[2M]**
2		Explain the three important steps of the text analysis	[L2][CO6]	**[10M]**
3	**a**	Sketch the flow diagram of Text analysis process	[L5][CO6]	**[4M]**
	b	Illustrate in detail the steps involved in the process of Text Analysis done by organizations	[L3][CO6]	**[6M]**
4	**a**	Define TFIDF.	[L1][CO6]	**[2M]**
	b	Describe the usage of TFIDF to compute the usefulness of each word in the texts.	[L2][CO6]	**[8M]**
5		Explain how the data science team will categorize the reviews by topics	[L2][CO6]	**[10M]**
6		Illustrate the main challenges of text analysis	[L3][CO6]	**[10M]**
7	**a**	Define Topic model. Describe LDA.	[L2][CO6]	**[5M]**
	b	Justify the process of topic modeling simplification.	[L6][CO6]	**[5M]**
8		Explain the following a) Tokenization b) Case folding	[L3][CO6]	**[10M]**
9	**a**	Explain how categorizing documents by topics is done.	[L2][CO6]	**[6M]**
	b	Interpret the procedure used in data science to gain insights into customer opinions	[L3][CO6]	**[6M]**
10	**a**	What is meant by sentiment analysis	[L1][CO6]	**[2M]**
	b	Discriminate the methods used for sentiment analysis	[L5][CO6]	**[8M]**

Prepared by:
Mrs. R Priyadarshini
Associate Professor