RIZVI EDUCATION SOCIETY's

Rizvi College of Engineering

Department of Artificial Intelligence and Data Science

Class: TE **Sem:** VI **Scheme:** (REV- 2019 'C' Scheme)

Course Code: CSC601 **Course:** Data Analytics and Visualization

Course Outcomes:

After successful completion of the course students will be able to:

Course Code	Course Outcome
CSC601.1	Comprehend basics of data analytics and visualization.
CSC601.2	Apply various regression models on a given data set and perform prediction.
CSC601.3	Demonstrate advance understanding of Time series concepts and analysis of data using various time series models.
CSC601.4	Analyze Text data and gain insights.
CSC601.5	Experiment with different analytics techniques and visualization using R.
CSC601.6	Experiment with different analytics techniques and visualization using Python.

Department of Artificial Intelligence and Data Science

QUESTION BANK

Module 1: Introduction to Data analytics and life cycle		Weightage: 10 Marks
Q. No.	**Question**	**Course Outcome**
1	What is an analytic sandbox, and why is it important?	CSC601.1
2	Explain the differences between Bl and Data Science.	CSC601.1
3	Describe the challenges of the current analytical architecture for data scientists.	CSC601.1
4	What are the key skill sets and behavioral characteristics of a data scientist?	CSC601.1
5	What are the key Roles and stakeholders for a successful analytics project?	CSC601.1
6	In which phase would the team expect to invest most of the project time? Why? Where would the team expect to spend the least time?	CSC601.1
7	What kinds of tools would be used in the following phases, and for which kinds of use scenarios? a. Phase 2: Data preparation b. Phase 4: Model building	CSC601.1
8	What are the activities carried out in each phase in the Life cycle of data analytics.	CSC601.1
9	Consider an example of a retail store chain that wants to optimize its products' prices to boost its revenue. The store chain has thousands of products over hundreds of outlets, making it a highly complex scenario. Once you identify the store chain's objective, you find the data you need, prepare it, and go through the Data Analytics lifecycle process.There are different types of customers, such as ordinary customers and customers like contractors who buy in bulk. a. How would you apply the data analytics life cycle to this problem?	CSC601.1

<table>
<tr><td colspan="3"><b>Module 2: Regression Models</b> <b>Weightage: 20 Marks</b></td></tr>
<tr><th>Q. No.</th><th>Question</th><th>Course Outcome</th></tr>
<tr><td>1</td><td>Define the following terms related to regression analysis:-
a. Overfitting
b. Cross validation
c. R²
d. Residuals</td><td>CSC601.2</td></tr>
<tr><td>2</td><td>Explain Logit/log-odds function in detail?</td><td>CSC601.2</td></tr>
<tr><td>3</td><td>For the customer service data, the proportion of customers who would recommend the service in the sample of customers is p hat= 0.84, so calculate the proportion of customers who would not recommend the service department?</td><td>CSC601.2</td></tr>
<tr><td>4</td><td>What is the difference between Linear Regression and Logistic Regression?</td><td>CSC601.2</td></tr>
<tr><td>5</td><td>What is the difference between Linear Regression and Multiple Regression?</td><td>CSC601.2</td></tr>
<tr><td>6</td><td>Describe how logistic regression can be used as a classifier.</td><td>CSC601.2</td></tr>
<tr><td>7</td><td>Discuss how the ROC curve can be used to determine an appropriate threshold value for a classifier.</td><td>CSC601.2</td></tr>
<tr><td>8</td><td>If the probability of an event occurring is 0.4, then
a. What is the odds ratio?
b. What is the log odds ratio?</td><td>CSC601.2</td></tr>
<tr><td>9</td><td>Find Linear regression equation for the following data
<table>
<tr><td>x</td><td>2</td><td>3</td><td>5</td><td>8</td></tr>
<tr><td>y</td><td>3</td><td>6</td><td>5</td><td>12</td></tr>
</table></td><td>CSC601.2</td></tr>
<tr><td>10</td><td>Find Linear regression equation for the following data
<table>
<tr><td>x</td><td>2</td><td>4</td><td>6</td><td>8</td></tr>
<tr><td>y</td><td>3</td><td>7</td><td>5</td><td>10</td></tr>
</table></td><td>CSC601.2</td></tr>
<tr><td>11</td><td>Find multiple regression equation for the following data</td><td>CSC601.2</td></tr>
</table>

x1	60	62	67	70	71	72	75	78
x2	22	25	24	20	15	14	14	11
y	140	155	159	179	192	200	212	215

Module 3: Time Series **Weightage: 10 Marks**

Q. No.	Question	Course Outcome
1	List some applications that deal with time series data.	CSC601.3
2	What are the components of time series data in Box-Jenkins Methodology?	CSC601.3
3	Define the following terms with respect to time series data: a. Trend b. Seasonality c. Cyclic d. Random e. Stationarity f. Differencing	CSC601.3
4	Justify which of the following series are stationary? (a) Google stock price for 200 consecutive days; (b) Daily change in the Google stock price for 200 consecutive days; (c) Annual number of strikes in the US; (d) Monthly sales of new one-family houses sold in the US; (e) Annual price of a dozen eggs in the US (constant dollars); (f) Monthly total of pigs slaughtered in Victoria, Australia; (g) Annual total of lynx trapped in the McKenzie River district of north-west Canada; (h) Monthly Australian beer production; (i) Monthly Australian electricity production.	CSC601.3
5	What is the significance of differencing in time series data analysis?	CSC601.3
6	Why is time series required to be stationary.?	CSC601.3
7	Define the following components of time series a. Trend b. Seasonality	CSC601.3
8	Define the following terms related to time series	CSC601.3

	a. Stationary b. Differencing	
9	Write the steps to develop ARIMA model using Box Jenkins Methodology.	CSC601.3
10	What are auto regressive models?	CSC601.3
11	What is the moving average model in time series?	CSC601.3

Department of Artificial Intelligence and Data Science

<table>
<tr><td colspan="3"><b>Module 4: Text Analytics</b> <b>Weightage: 20 Marks</b></td></tr>
<tr><th>Q. No.</th><th>Question</th><th>Course Outcome</th></tr>
<tr><td>1</td><td>Define the following terms:
a. Term Frequency
b. Inverse Document Frequency
c. Bag of words
d. Corpus</td><td>CSC601.4</td></tr>
<tr><td>2</td><td>What are the steps in text analytics?</td><td>CSC601.4</td></tr>
<tr><td>3</td><td>What are the methods of representing text in text analytics? What are the challenges associated with them?</td><td>CSC601.4</td></tr>
<tr><td>4</td><td>What are the different transformation techniques used to represent raw data?</td><td>CSC601.4</td></tr>
<tr><td>5</td><td>The term frequency matrix given in the table below shows the frequency of terms per document. Calculate the TF-IDF value of Terms T1, T2, T3, T4, T5, T6 in Document D1
<table>
<tr><th>Document / Term</th><th>T1</th><th>T2</th><th>T3</th><th>T4</th><th>T5</th><th>T6</th></tr>
<tr><td><b>D1</b></td><td>5</td><td>9</td><td>4</td><td>0</td><td>5</td><td>6</td></tr>
<tr><td><b>D2</b></td><td>0</td><td>8</td><td>5</td><td>3</td><td>10</td><td>8</td></tr>
<tr><td><b>D3</b></td><td>3</td><td>5</td><td>6</td><td>6</td><td>5</td><td>0</td></tr>
<tr><td><b>D4</b></td><td>4</td><td>6</td><td>7</td><td>8</td><td>4</td><td>4</td></tr>
</table></td><td>CSC601.4</td></tr>
<tr><td>6</td><td>Give three benefits of using the TF IDF.</td><td>CSC601.4</td></tr>
<tr><td>7</td><td>What methods can be used for sentiment analysis?</td><td>CSC601.4</td></tr>
<tr><td>8</td><td>What is the definition of topic in topic models?</td><td>CSC601.4</td></tr>
<tr><td>9</td><td>Explain precision and recall.</td><td>CSC601.4</td></tr>
</table>

Module 5: Data analytics and visualization with R		**Weightage: 10 Marks**
Q. No.	**Question**	**Course Outcome**
1	Create a data frame of 10 employee names, age and salary. Display the summary of salary and age. Plot a boxplot for age and salary in R.	CSC601.5
2	Create a data frame of 10 students' names, subject and marks. Display the summary of subject and marks. Plot a boxplot for subject and marks in R.	CSC601.5
3	Create a sample of 50 numbers which are normally distributed. Plot the histogram for this sample in R.	CSC601.5
4	Create a vector of 1000 random numbers with mean = 90 and sd=5. Create the histogram with 50 bars in R.	CSC601.5
5	If a graph of data is skewed and all the data is positive, what mathematical technique may be used in R to help detect structures that might otherwise be overlooked?	CSC601.5
6	What function can be used to fit a nonlinear line to the data in R?	CSC601.5
7	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?	CSC601.5
8	Explain exploratory data analysis in R	CSC601.5

Department of Artificial Intelligence and Data Science

Module 6: Data analytics and Visualization with Python Weightage: 10 Marks

Q. No.	Question	Course Outcome
1	Create a data frame of 10 employee names, age and salary. Plot a boxplot for age and salary in python.	CSC601.6
2	Create a data frame of 10 students' names, subject and marks. Plot a boxplot for subject and marks in python.	CSC601.6
3	Create the box plot by using some random data, give mean, standard deviation, and the desired number of values.	CSC601.6
4	Create a data frame of 10 students' names, subject and marks. Plot a regression plot for subject and marks in python.	CSC601.6
5	Create a data frame of 10 employee names, age and salary. Plot a regression plot for age and salary in python.	CSC601.6

Note: This question bank is not provided by the university. It is a sample for you to prepare for the exam.

*************************All the Best ***********************