Course Code: 20CS1101 **R20**

SIDDHARTH INSTITUTE OF ENGINEERING & TECHNOLOGY:: PUTTUR
(AUTONOMOUS)
Siddharth Nagar, Narayanavanam Road – 517583

QUESTION BANK (DESCRIPTIVE)

Subject with Code: (20CS1101) INTRODUCTION TO DATA SCIENCE

Course & Branch: B.Tech – CSE(CAD) **Regulation:** R20

UNIT –I
INTRODUCTION TO DATA SCIENCE

1	**a**	Define Data Science and discuss Benefits and uses of data science.	[L1][CO1]	**[6M]**
	b	Discuss the Various Processing Steps in Data Science	[L2][CO1]	**[6M]**
2		Explain in Details various data types used in Data science and Big data	[L2][CO1]	**[12M]**
3	**a**	Analyze the term: Distributed file systems	[L4][CO1]	**[6M]**
	b	How will you creating research goals in a project charter	[L1][CO1]	**[6M]**
4		Classify the term big data ecosystem	[L4][CO1]	**[12M]**
5		How will you retrieve the required data from data science	[L1][CO5]	**[12M]**
6		Discuss in detailed Data Cleaning operation in data science	[L2][CO1]	**[12M]**
7	**a**	What are various steps involved in integrating phase	[L1][CO1]	**[6M]**
	b	What is meant by exploratory data analysis	[L1][CO1]	**[6M]**
8		Examine the term: Transforming data in Data science	[L3][CO1]	**[12M]**
9	**a**	Show the various components of model building.	[L2][CO1]	**[6M]**
	b	What are the ways analyzed the data and built a well-performing model	[L2][CO1]	**[6M]**
10	**a**	How will you handling missing data in data science	[L2][CO1]	**[6M]**
	b	Examine K-nearest neighbor techniques look at the k-nearest point to make a prediction	[L4][CO1]	**[6M]**

Course Code: 20CS1101 **R20**

UNIT –II
STATISTICAL METHODS FOR EVALUATION&ASSOCIATION RULES

1	**a**	Define Hypothesis Testing	[L1][CO2]	**[6M]**
	b	How will you mathematically define Confidence	[L1][CO2]	**[6M]**
2	**a**	Differentiate Null Hypotheses and Alternative Hypotheses	[L4][CO2]	**[6M]**
	b	Examine the application property of Wilcoxon rank-sum test	[L3][CO2]	**[6M]**
3		Discriminate about Difference of Means	[L5][CO2]	**[12M]**
4		Explain the differences between Bl and Data Science.	[L2][CO2]	**[12M]**
5		*Explain the following* a) Student's t-test b) Welch's t-test	[L2][CO2]	**[12M]**
6	**a**	What are the three characteristics of Big Data, and what are the main considerations in processing Big Data?	[L1][CO2]	**[6M]**
	b	How evaluation of Candidate Rules is done?	[L2][CO2]	**[6M]**
7	**a**	What is a type I error? What is a type II error? Is one always more serious than the other? Why?	[L1][CO2]	**[6M]**
	b	Give the difference between Validation and Testing	[L4][CO2]	**[6M]**
8	**a**	State Apriori Algorithm	[L1][CO2]	**[4M]**
	b	Explain Apriori Algorithm with example	[L2][CO2]	**[8M]**
9	**a**	List and discuss the four measures of significance of Association rules	[L1][CO2]	**[6M]**
	b	Give the Applications of Association Rules	[L1][CO2]	**[6M]**
10		Illustrate any five approaches to improve Apriori's efficiency when the dataset is large.	[L3][CO2]	**[12M]**

Course Code: 20CS1101 **R20**

UNIT –III
REGRESSION& CLASSIFICATION

1	**a**	Which two basic measures does the entropy methods select the most informative attribute?	[L1][CO3]	**[6M]**
	b	Define confusion matrix	[L1][CO3]	**[6M]**
2		Explain the analytical technique Linear Regression with its model description.	[L2][CO3]	**[12M]**
3		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	[L2][CO3]	**[12M]**
4	**a**	Justify the usage of linear regression and logistic regression.	[L6][CO3]	**[4M]**
	b	Illustrate Logistic Regression Model.	[L3][CO3]	**[8M]**
5	**a**	Describe Decision Trees in detail with example.	[L2][CO3]	**[6M]**
	b	Difference between Alternative hypothesis and null hypothesis	[L2][CO4]	**[6M]**
6		Intercept the decision trees algorithms	[L4][CO4]	**[12M]**
7	**a**	State Bayes' Theorem	[L1][CO4]	**[4M]**
	b	Discuss Naïve Bayes classification method considering an example	[L2][CO4]	**[8M]**
8		How does one pick the mostsuitable method for a given classification problem?	[L2][CO4]	**[12M]**
9	**a**	Compare the C4.5 and CART algorithm of decision tree.	[L4][CO4]	**[4M]**
	b	Discriminate the way show the evaluation of decision tree is done	[L5][CO4]	**[4M]**
	c	Give the two approaches that help avoid over fitting in decision tree learning.	[L2][CO4]	**[4M]**
10		Discuss the following term: a) Accuracy b) TPR c) FPR d) FNR e) Precision	[L4][CO4]	**[12M]**

Course Code: 20CS1101 **R20**

UNIT –IV
CLUSTERING & TIME SERIES ANALYSIS

1	**a**	What is clustering?	[L1][CO5]	**[6M]**
	b	State the advantage of using PAM.	[L1][CO5]	**[6M]**
2		Illustrate the method to find k clusters from a collection of M objects with n attributes.	[L3][CO5]	**[12M]**
3	**a**	Explain any one case study for time series analysis	[L2][CO5]	**[6M]**
	b	What is forecasting in association with time series. Explain	[L1][CO6]	**[6M]**
4	**a**	Indicate when the time series $y_t^{\text{for } t=1,2,3,\cdots}$ is said to be stationary time series.	[L2][CO6]	**[6M]**
	b	Express the stationary time series conditions in detail.	[L6][CO6]	**[6M]**
5		Discussion detail each part of the ARIMA model	[L2][CO5]	**[12M]**
6	**a**	List and explain time series components	[L1][CO6]	**[6M]**
	b	Discriminate the steps involved in Box-Jenkins Methodology	[L5][CO6]	**[6M]**
7	**a**	What is meant by k-means	[L1][CO5]	**[4M]**
	b	Describe k-means algorithm to find k clusters	[L2][CO5]	**[8M]**
8		Correlate ARMA and ARIMA Models	[L4][CO6]	**[12M]**
9		Express the following a) Autocorrelation Function b) Autoregressive Models	[L2][CO6]	**[12M]**
10		List and describe Additional time series methods	[L2][CO6]	**[12M]**

Course Code: (20CS1101 **R20**

UNIT –V TEXT ANALYSIS

1	**a**	Define Porter's stemming algorithm.	[L1][CO6]	**[6M]**
	b	What is Topic modeling?	[L1][CO6]	**[6M]**
2		Explain the three important steps of the text analysis	[L2][CO6]	**[12M]**
3	**a**	Sketch the flow diagram of Text analysis process	[L5][CO6]	**[6M]**
	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]	**[4M]**
	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]	**[12M]**
6		Illustrate the main challenges of text analysis	[L3][CO6]	**[12M]**
7	**a**	Define Topic model. Describe LDA.	[L2][CO6]	**[6M]**
	b	Justify the process of topic modeling simplification.	[L6][CO6]	**[6M]**
8		Explain the following a) Tokenization b) Case folding	[L3][CO6]	**[12M]**
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]	**[4M]**
	b	Discriminate the methods used for sentiment analysis	[L5][CO6]	**[8M]**

Preparedby:
Mr.G.Prasad Babu
Associate Professor