Sub Code: CST 307 **ROLL NO……..……..**

Ist SEMESTER EXAMINATION, 2022 – 23

Ist yr M.Tech. – Computer Science & Engineering

Data Science

Duration: 3:00 hrs **Max Marks: 100**

Note: - Attempt all questions. All Questions carry equal marks. In case of any ambiguity or missing data, the same may be assumed and state the assumption made in the answer.

Q 1.	Answer any four parts of the following. a) What is Data Science? List the differences between supervised and unsupervised learning. b) List 4 libraries in R that can be used for data visualization. c) What is the main difference between k-Means and k-Nearest Neighbours? d) What are some Stopping Criteria for k-Means Clustering? e) Why do we need dimensionality reduction? What are its drawbacks? f) How do you find RMSE and MSE in a linear regression model?	5x4=20
Q 2.	Answer any four parts of the following. a) Write the importance of statistics used in Data Science. b) What is EDA process in data science? c) Explain what should be done with suspected or missing data? d) What are the steps involved in a data analysis process? e) Write down steps in EDA. f) How can you calculate accuracy using a confusion matrix?	5x4=20
Q 3.	Answer any two parts of the following. a) Write down the steps to test hypothesis. b) During the data preprocessing step, how should one treat missing/null values? How will you deal with them? c) Define spam filter. How do you use Naive Bayes model for Collaborative Filtering?	10x2= 20
Q 4.	Answer any two parts of the following. a) Write the differences between over-fitting and under-fitting. How can you avoid overfitting your model? b) What are the difference between linear regression and logistic? What are the key matrices used to check the performance of logistic regression? c) What are the key matrices used to check the performance of logistic regression? The key matrices used to check the performance of logistic regression?	10x2= 20
Q 5.	Answer any two parts of the following. a) What is the need for ethics in data science? What are the ethical issues in data science? b) What are some important features of a good data visualization? c) What is a recommendation engine? How does it work?	10x2= 20

****