Code: 01X1MITIDSC2023 Time: 2 ½ hours Max.Marks:75 marks

Instructions:

- All questions are compulsory.
- Make suitable assumptions wherever necessary and state the assumptions made.
- Answers to the same question must be written together.
- Numbers to the right indicate marks.
- Draw neat labeled diagrams wherever necessary.
- Use of Non-Programmable calculators is allowed.

Q1. Attempt either Que (a) and Que(b) OR Que(c) and (d). 15

a. The majority of data scientists and data analysts, as well as data engineers, uses the data science processing tool to process and transfer data vaults into data warehouses. Discuss any of the five data science processing tools. 7

b.State the functional requirements in the business layer of the data science framework. 8

OR

c. The basic utility design is a three-stage process. Discuss all three stages and the various types of utilities in data science using a diagram. 7

d. The sun model is a requirement mapping technique that assists you in recording requirements at a level that allows non-technical users to understand the intent of the analysis. Justify the statement with the help of suitable example including facts and dimensions. 8

Q2. Attempt either Que (a) and Que(b) OR Que(c) and (d). 15

a. Data swamps are simply unmanaged data lakes. Explain and list four critical steps to avoid data swamps.7

b. The operational management layer is the hub of the data science ecosystem's complete processing capability. Which parameters are monitored by the operational management layer? 8

OR

c. Write a python retrieve solution for engineering a practical retrieve superstep. 7

d. Which functions do analytical models play in data governance? Look at the following analytical models and discuss them along with their code snippets.

i) Mean

ii) Mode

iii) Median

iv) Data Field Name Verification 8

Q3. Attempt either Que (a) and Que(b) OR Que(c) and (d). 15

a. "Graph theory is always a useful tool to use when relationships between business entities need to be analyzed." Use an example to elaborate the concept. 7

b.Using the pandas package, how can we drop columns where any of the elements is missing values? To illustrate your point, use an appropriate example. 8

OR

c. Discuss briefly about the significance of the assess superstep in datascience. 7

d. One of the causes of data quality issues is source data that is housed in a patchwork of operational systems and enterprise applications. Using a diagram, list and explain the six quality dimensions used in data analysis. 8

Q4. Attempt either Que (a) and Que(b) OR Que(c) and (d). 15

a. The person section contains the entire data structure for all data entities associated with recording the person. In relation to the T-P-O-L-E design principle, Discuss the following terms with examples of coding.

i) Person Hub

ii) Person Links 7

b. Discuss briefly the following terms along with the formula in relation to the Precision-Recall Curve. 8

i) True Negative Rate

ii) Precision

iii) Recall

OR

c. Draw a person-to-time sun model with appropriate explanation using two dimensions (Person and Time) and one fact(PersonBornAtTime). 7

d.Overfittting and Underfitting are the two major issues that data scientists face when attempting to extract data insights from training data sets. Extend the concept of overfitting and underfitting using appropriate visualizations. 8

Q5. Attempt either Que (a) and Que(b) OR Que(c) and (d). 15

a. The data mart is a subset of the datawarehouse that is generally focused on a specific business group. Discuss various styles of data slicing in relation to this. 7

b. Which cluster analysis method involves building a cluster hierarchy? Use a diagram and various types to discuss the approach. 8

OR

c. In datascience, Compare univariate, bivariate and multivariate analysis. 7

d. Using appropriate code snippets, elaborate on the following concepts: 8

i) Installation of TensorFlow

ii) Variables in TensorFlow

iii) Placeholders in TensorFlow