U23AI401 - Statistical Inference

UNIT 1 – 2 Marks

1. Define Data Science and its primary goal. Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. Its primary goal is to turn raw data into actionable business value.

2. List four key benefits of using Data Science in industry.

- **Better Decision Making:** Using evidence-based insights rather than intuition.
- **Predictive Analysis:** Forecasting market trends and customer behavior.
- **Efficiency:** Optimizing supply chains and business processes.
- **Personalization:** Delivering tailored recommendations to users (e.g., Netflix or Amazon).

3. What are the different "Facets of Data"? Data can be categorized into several types:

- **Structured:** Highly organized (SQL databases).
- **Unstructured:** No predefined format (Videos, images, social media posts).
- **Semi-structured:** Contains tags or markers (XML, JSON).
- **Big Data:** Characterized by high Volume, Velocity, and Variety.

4. What is the importance of "Setting the Research Goal"? This is the first step of the process. It involves defining the problem statement, understanding the business context, and identifying the key performance indicators (KPIs) to ensure the project delivers relevant results.

5. Distinguish between Data Retrieval and Data Cleansing.

- **Data Retrieval:** The process of collecting raw data from various sources like databases, APIs, or web scraping.
- **Data Cleansing:** The process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.

6. What happens during Data Integration and Transformation?

- **Integration:** Combining data from different sources into a single, unified view.
- **Transformation:** Converting data from one format/structure into another (e.g., normalization or scaling) to make it suitable for modeling.

7. Define Exploratory Data Analysis (EDA). EDA is the process of analyzing datasets to summarize their main characteristics, often using visual methods. It helps in discovering patterns, spotting anomalies, and checking assumptions before formal modeling.

8. What is "Model Building" in Data Science? Model building involves selecting and training a mathematical algorithm (like Linear Regression or Random Forest) using a portion of the data to find patterns and make predictions or classifications on new, unseen data.

9. Why is "Presenting and Building Applications" considered the final step? The insights gained are useless unless communicated to stakeholders. This step involves:

- **Presenting:** Using visualizations to explain findings.
- **Building Applications:** Integrating the model into a production environment (like a web app or dashboard) for end-users.

10. What is the difference between Structured and Unstructured data?

- **Structured data** is highly organized and fits into fixed fields (like an Excel sheet or SQL table).
- **Unstructured data** has no predefined format (like emails, PDFs, or audio files), making it harder to analyze without specialized tools.

11. Why is Data Transformation necessary before building a model? Data transformation (like scaling or encoding) ensures that the mathematical algorithms can interpret the data correctly. For example, converting "Red/Blue/Green" into numbers (0,1,2) is necessary because models cannot perform calculations on text strings.

12. Define the "Variety" facet of Big Data. Variety refers to the diverse types of data sources involved in modern analysis. This includes data from social media, sensors (IoT), log files, and traditional databases, all of which must be integrated to get a 360-degree view.

13. What is the role of a "Data Dictionary" in the Retrieval phase? A data dictionary is a centralized repository of metadata. It describes the meaning, relationships, origin, usage, and format of each data element, ensuring that the data scientist understands what each column represents.

14. Mention two common techniques used in Data Cleansing.

- **Imputation:** Filling in missing values using the mean, median, or mode.
- **Outlier Detection:** Identifying and removing data points that are significantly different from the rest of the set and might skew the results.

15. What is "Data Integration" in a multi-source project? Data Integration involves combining data from different sources (e.g., merging a customer's "Purchase History" from a SQL database with their "Web Clicks" from a CSV file) to create a consistent dataset for analysis.

16. How does Descriptive Statistics aid in Exploratory Data Analysis (EDA)? Descriptive statistics (like mean, standard deviation, and quartiles) provide a quick summary of the data's distribution and spread. This helps identify if the data is biased or if it follows a normal distribution.

17. What is "Feature Selection" in the Model Building phase? Feature selection is the process of choosing the most relevant variables (features) to use in the model. Removing irrelevant or redundant data reduces the complexity of the model and improves its accuracy.

18. Explain the "Interative" nature of the Data Science Process. The process is rarely linear. If the results of the **EDA** or **Model Building** phase show that the data is insufficient, a scientist must go back to the **Retrieval** or **Research Goal** phase to refine the approach.

19. What is a "Data Product" in the final phase? A data product is an application or tool that automates the delivery of insights. Examples include a credit scoring tool used by a bank or a recommendation engine used by a streaming service.

20. What is the role of a Project Charter in the "Setting Research Goal" phase? The project charter is a document that outlines the project's scope, objectives, and participants. It serves as a formal agreement between the data scientist and stakeholders on what the project will deliver and how success will be measured.

21. List three common types of "Facets of Data" with an example for each.

- **Graph-based data:** Social network connections (e.g., LinkedIn "connections").
- **Streaming data:** Real-time stock market prices or Twitter trends.
- **Natural Language:** Product reviews or customer support emails.

22. Differentiate between Internal and External data.

- **Internal Data:** Data generated within an organization (e.g., sales logs, employee payroll).
- **External Data:** Data gathered from outside the organization (e.g., government census data, weather reports, social media scraping).

23. What is "Data Munging" (or Data Wrangling)? Data Munging is the process of manually converting and mapping data from one "raw" form into another format that allows for more convenient consumption of the data with the help of semi-automated tools.

24. How do you handle missing values during the "Cleansing" phase? Missing values can be handled by:

- **Deletion:** Removing rows with missing values (if the data loss is minimal).
- **Imputation:** Filling in missing values using statistical measures like the mean, median, or mode.

25. Define "Feature Engineering." Feature engineering is the process of using domain knowledge to create new variables (features) from raw data that help machine learning algorithms perform better (e.g., calculating "Age" from a "Date of Birth" column).

26. What are the two main types of EDA (Exploratory Data Analysis)?

- **Univariate Analysis:** Analyzing a single variable (e.g., a histogram of ages).
- **Bivariate/Multivariate Analysis:** Analyzing the relationship between two or more variables (e.g., a scatter plot showing the correlation between height and weight).

27. What is the "Model Diagnostics" step in building models? Model diagnostics is the process of evaluating how well a model performs using metrics like **Accuracy**, **Precision**, and **Recall**. It helps identify issues like **Overfitting** (where the model is too complex) or **Underfitting** (where the model is too simple).

28. Explain the importance of "Data Storytelling" in presenting results. Data storytelling is the practice of building a narrative around a set of data and its accompanying visualizations to help communicate the "why" behind the insights to non-technical stakeholders, ensuring the findings lead to action.

29. What does it mean to "Industrialize" or "Automate" a data application? This is the final stage where the model is integrated into a production system. It involves automating the data pipeline so that the model can process new data and provide real-time predictions or reports without manual intervention.

PART B- 16 MARKS

1. Discuss the "Need for Data Science" in the modern era. Explain the various benefits and uses of data science across different industries.

2. Elaborate on the "Facets of Data." Describe the different types of data a data scientist might encounter and the challenges associated with each.

3. Provide a detailed overview of the "Data Science Process." Illustrate the lifecycle with a neat diagram and explain the importance of each stage.

4. "Setting the Research Goal is the most critical step in the data science pipeline." Justify this statement and explain the activities involved in this phase.

5. Explain the "Data Retrieval" and "Data Cleansing" phases in detail. What are the common challenges faced during these stages?

6. Discuss the technical nuances of "Integrating and Transforming Data." Why are these steps necessary before modeling?

7. Define Exploratory Data Analysis (EDA). Explain the various statistical and visual techniques used to understand data distributions and relationships.

8. Describe the "Build the Models" phase. How do you select, train, and evaluate a model?

9. Explain the process of "Presenting and Building Applications." How does a data scientist transition from a notebook to a production environment?

10. Using a real-world case study (e.g., a Bank predicting loan defaults), explain every step of the Data Science Process from start to finish.