SNS COLLEGE OF TECHNOLOGY

An Autonomous Institution
Coimbatore-35

DEPARTMENT OF ARTIFICIAL INTELLIGENCE & DATA SCIENCE

23ADT202 – FUNDAMENTALS OF DATA SCIENCE AND ANALYTICS

II YEAR IV SEM

UNIT I – INTRODUCTION TO DATA SCIENCE

DATA SCIENCE PROCESS

DATA SCIENCE PROCESS

EMPATHY

- It is the process of analyzing and interpreting data to uncover hidden trends, correlations and insights that can support decision-making and strategic planning.

- It involves manipulating raw data using analytical and computational techniques to transform it into valuable information.

Various professionals who use it are:

- **Data Engineer**: Responsible for building scalable data pipelines, managing databases and ensuring smooth data flow.

- **Data Analyst**: Focuses on analyzing data, generating reports and visualizing insights for business use.

- **Data Architect**: Designs data storage and management systems to ensure efficiency and scalability.

- **Machine Learning Engineer**: Develops, optimizes and deploys machine learning models.

- **Deep Learning Engineer**: Works on advanced neural network models for complex data such as images, audio and text.

Data Science Process Life Cycle

The PAC Framework Elements

DEFINE

1. Data Collection

- Data collection involves gathering relevant data from multiple sources such as databases, APIs, surveys, logs, sensors or web scraping.
- The accuracy, completeness and relevance of the collected data significantly affect the reliability of the final model and insights.

2. Data Cleaning

- Most real-world data contains missing values, inconsistencies, duplicates and noise.
- Data cleaning focuses on correcting errors, handling missing data, removing irrelevant records and converting data into a structured format suitable for analysis.

3. Exploratory Data Analysis (EDA)

IDEATE:

- EDA is used to understand the data in depth by applying descriptive statistics and visualization techniques.
- It helps identify trends, outliers, correlations and relationships between variables and guides decisions related to feature selection and modeling strategies.

4. Model Building

TESTING:

- In this stage, suitable machine learning algorithms are selected and trained on historical data.
- The goal is to identify patterns that allow the model to make accurate predictions or classifications on unseen data.

5. Model Deployment

PROTOTYPE:

- After validation, the trained model is deployed into a production environment.
- Its performance is continuously monitored and updates are made as new data becomes available or conditions change.

Challenges

Data Quality and Availability

Bias in Data and Algorithms

Overfitting and Underfitting

Model Interpretability

Privacy and Ethical Considerations

RECAP