SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

Course Name – 23ADT202 Fundamental of Data science and Analytics

II Year / IV Semester

Unit 1– Introduction to Data science

Data science process

Introduction to the Data Science Process

The data science process turns raw data into valuable insights. It includes defining goals, retrieving and cleaning data, performing exploratory analysis, building models, and presenting findings. Each step is crucial for systematic analysis. The process is iterative, ensuring continuous improvement.

Step 1 - Setting the Research Goal

The first step is defining the problem or question clearly. A well-defined goal focuses the analysis and drives subsequent steps. For example, predicting customer churn or identifying profitable products. Clear goals ensure relevant and actionable outcomes.

Step 2 - Retrieving Data

Data must be gathered from various sources like databases, APIs, or web scraping. The quality of analysis depends on the quality of the data collected. Examples include extracting sales data from CRM systems or gathering weather data from public APIs. Data retrieval is a critical foundation for analysis.

Step 3 - Data Cleansing, Integration, & Transformation

Data is often messy and needs to be cleaned and transformed. Tasks include handling missing values, removing duplicates, and integrating data from different sources. Proper data preparation ensures reliable and consistent analysis. This step is crucial for accurate modeling.

Step 4 - Exploratory Data Analysis (EDA)

EDA involves visualizing and summarizing data to identify trends, patterns, and outliers. It provides a preliminary understanding of the dataset. Techniques like histograms and scatter plots help uncover insights. EDA is essential for guiding model selection and feature engineering.

Step 5 - Building Models

In this step, appropriate machine learning or statistical models are selected and applied. Models such as regression or clustering help answer the research question. The choice of model depends on the problem type and data characteristics. Building a model is key to deriving insights from data.

Step 6 - Evaluation & Validation

Models are assessed using performance metrics like accuracy, precision, and R^2 scores. This ensures the models perform as expected. If performance is inadequate, data or algorithms may be adjusted. Evaluation helps refine models for better predictive power.

Step 7 - Presentation & Building Applications

Insights are communicated via reports, dashboards, and visualizations. These tools make the findings accessible to stakeholders. Results are often integrated into decision-making systems. Presenting clear insights helps organizations take informed actions based on data.

Conclusion: Iterative Nature of the Data Science Process

The data science process is iterative, with each phase influencing the next. Refinements based on insights ensure ongoing accuracy. Mastering the process allows data scientists to solve complex problems effectively. It empowers organizations to leverage data for valuable decisions.

THANK YOU