

Assignment Unit-1

1. Define Data science and explain its life cycle

Data science

Data science is an interdisciplinary field that uses statistics, mathematics, programming, machine learning, and data analysis techniques to collect, process, analyze and interpret data to extract useful information and support decision making.

For example:-

- which product ~~at~~ customers are likely to buy
- Future sales
- which customer may stop using the service

Data science Life cycle

1. Problem Definition

The first step is to clearly understand what the problem is

Example

A bank want to predict whether a customer will repay loan

2. Data collection

Data is collected from different sources

Such as

- Databases
- APIs

- Survey
- Social media
- Excel
- csv files

Data cleaning

Raw data often contains problems such as

- Missing values
- Duplicate records
- Outliers
- Inconsistent formats

Data Exploration and Analysis

The collected data is explored to understand its characteristics and relationship

- Mean
- Median
- Mode
- Standard deviation
- Correlation

Data preprocessing

The data is converted into a suitable format for analysis and machine learning

- Handling missing values
- Removing duplicates
- Feature scaling

Example

Male $\rightarrow 0$

Female $\rightarrow 1$

6 Model Building

A suitable statistical or machine-learning model is selected and trained using the prepared data

- Linear Regression
- Decision Tree
- Random Forest
- Neural Network

7 Model Evaluation

The model's performance is evaluated using appropriate metrics

- Accuracy
- Precision
- Recall
- F1 Score

8 Deployment

After successful evaluation, the model is implemented in a real-world

9. Monitoring and Maintenance

After deployment, the model must be continuously monitored because real-world can change over time

2. Differentiate Descriptive, Diagnostic, Predictive and Prescriptive Analytics

Type	Main Question	Purpose	Example
Descriptive Analytics	What happened?	Understand Past events	Sales decreased by 10%
Diagnostic Analytics	Why did it happen?	Find causes/Reasons	Sales decreased because of low demand
Predictive Analytics	What will happen?	Predict future events	Sales may decrease next month
Prescriptive Analytics	What should we do?	Recommend actions	Increases discounts to improve sales

Descriptive Analytics

Descriptive Analytics summarizes historical data to understand what happened in the past

- Reports
- Dashboards
- Charts
- Graphs
- Averages

Predictive Analysis

Predictive analysis uses historical data, statistical and machine learning techniques to predict what may happen in the future.

- Regression
- Random Forest
- Decision Trees
- Neural Networks.

Prescriptive Analytics

Prescriptive analytics determines what action should be taken based on available information.

- Optimization
- Simulation
- Machine Learning

3. Explain the steps involved in Data preprocessing with examples

Data preprocessing is the process of converting raw, incomplete, and inconsistent data.

- missing values
- Duplicate records
- outliers.

1. Data collection.

Data is collected from different sources

- Databases
- CSV files

- APIs
- sensors
- surveys

2. Data Cleaning

Data cleaning identifies and corrects errors in the dataset.

It includes

- Handling missing values
- Removing duplicates
- Correcting incorrect values

3. Handling Missing values

Missing values can be handled using different techniques

Mean

For numerical data, missing values can be replaced with mean

$$\frac{20+22+24}{3} = 22$$

4. Removing Duplicate Data

Student	Marks
Ravi	85
Priya	90
Ravi	85

5. Handling outliers

20, 22, 21, 23, 150

Here 150 is an outlier

6. Data Transformation

- Data transformation changes data into a suitable format

- Normalization
- Scaling

Data collection



Data cleaning



Handle Missing values



Removing duplicates



Handle outliers



Data transformation



Encode categorical Data



Feature scaling



Train Test split