CA3301: Machine Learning

Lab 1: Descriptive Analytics in Machine Learning

A) Descriptive Analytics using Titanic Dataset

Objective:

To perform descriptive analytics on a real-world dataset by computing statistical measures (mean, median, mode, standard deviation), understanding data distribution and central tendencies, conducting group-wise analysis, creating meaningful visualizations, and deriving actionable insights from the data.

Lab Tasks

1. Import and Explore the Dataset
    - Display the first 5 rows
    - Display info about data types and nulls
    - Describe the data
2. Clean the Dataset
    - Drop irrelevant columns
    - Handle missing values in Age, Embarked using:
    - Mean/median imputation for Age
    - Mode imputation for Embarked
3. Descriptive Statistics
   Perform and interpret:
    - df['Age'].mean(), df['Age'].median(), df['Age'].mode()
    - Value counts for Sex, Pclass, Embarked, Survived
    - Grouped analysis:
4. Visualization Tasks
   Use matplotlib or seaborn
    - Distribution of Age
    - Count by Gender
    - Survival Rate by Gender
    - Survival by Class
    - Age vs Survival
    - Heatmap of correlations
5. Stretch Activities
    - Create a new feature: FamilySize
    - Explore Survival by FamilySize using grouped analysis or a violin plot
    - Compare survival across multiple categories

Dataset: Titanic.csv

B) Customer Segmentation Using Descriptive Analytics

Objective:

To analyse customer behaviour using descriptive analytics by performing data cleaning, statistical summarization, group-wise comparisons, and visualization to identify customer segments and derive actionable business insights.

Tasks

1. **Data Cleaning**
    - Handle missing CustomerID values.
    - Remove negative or zero Quantity and UnitPrice values.
2. **Descriptive Statistics**
    - Compute mean, median, mode, standard deviation for Quantity and UnitPrice.
    - Summarize total spending per customer.
3. **Distribution Analysis**
    - Analyse distribution of Quantity, UnitPrice, and Total Sales (= Quantity × UnitPrice) using histograms, boxplots, and KDE plots.
4. **Group-Wise Analysis**
    - Group by Country and CustomerID to analyze spending behavior.
    - Group by StockCode to identify top-selling products.
5. **Visualization**
    - Top 10 customers by total spend (bar chart)
    - Country-wise sales distribution (pie chart or treemap)
    - Monthly sales trend (line plot)
6. **Actionable Insights**
    - Identify high-value customers.
    - Find peak purchasing periods.
    - Spot underperforming products or markets.

Dataset: Online Retail