Lab 2
Exploratory Data Analysis, Feature Engineering, and Feature Selection

Objectives

- Gain insights into the dataset using statistics and visualizations
- Apply feature engineering techniques like binning, encoding, outlier handling
- Use statistical tests and automated methods for feature selection

Theory Reading Required

- Attribute types (nominal, ordinal, continuous, discrete)
- Missing values, outliers, normalization, standardization
- Histogram, boxplot, scatterplot, countplot
- Label Encoding, One-Hot, Binary, Target, and Frequency Encoding
- Equal-width and equal-frequency binning, K-means and decision tree binning
- Correlation, ANOVA, Chi-Square test, SelectKBest

Dataset

Titanic Dataset
Load using:
import seaborn as sns
df = sns.load_dataset('titanic')

Exercise 1: Exploratory Data Analysis (EDA)

Step 1: Basic Understanding
- Use df.info() and df.describe() to understand the structure and summary
- Display column types and count of missing values

Step 2: Identify Attribute Types
- Manually classify the following attributes:
  - Categorical
  - Numerical
  - Target

Step 3: Understand Distribution of Attributes
- Compute: mean, median, std, quartiles for age, fare, parch
- Plot:
  - Histogram and Boxplot for age, fare
  - Countplot for sex, embarked, class

Step 4: Understand Relationships Among Attributes
- Plot scatterplot between age and fare
- Compute and visualize Pearson correlation matrix for age, fare, parch, sibsp
- Use pd.crosstab() between:
  - sex vs survived
  - embarked vs class

Step 5: Write down your observations from steps 3 and 4

Exercise 2: Feature Engineering

Step 1: Missing Value Handling
- Impute:
  - age with median
  - embarked with mode
  - Drop rows where deck is null

Step 2: Outlier Detection and Handling
- Plot boxplot and detect outliers in fare
- Cap outliers using IQR method

Step 3: Normalization and Standardization
- Apply:
  - Min-Max Normalization on fare
  - StandardScaler on age (use sklearn)

Step 4: Encoding Categorical Variables
Perform the following encodings:
- sex: Label Encoding
- embarked: One-Hot Encoding
- class: Frequency Encoding
- who: Target Encoding (target: survived)
- deck: Binary Encoding

Step 5: Binning of Numerical Attributes
Perform binning on age and fare:
- Equal-width

- Equal-frequency
- Custom (e.g., child/adult/senior)
- K-means binning
- Decision tree binning

Exercise 3: Feature Selection

Target variable: survived

Step 1: Pearson Correlation

- Compute correlation among: age, fare, parch, sibsp
- Drop one of any pair with correlation > 0.9

Step 2: ANOVA (f_classif)

- Test: sex, embarked, class, who vs survived

Step 3: Chi-Square Test

- Apply on: sex, embarked, class vs survived

Step 4: SelectKBest

- Use SelectKBest with chi2 and f_classif
- Select top 5 features for predicting survived

Knowledge Check Questions

1. Explain when you would prefer One-Hot Encoding over Label Encoding.
2. What's the impact of outliers on standardization and how can you mitigate it?
3. When would you use Chi-Square instead of Pearson correlation for feature selection?
4. How does SelectKBest determine the "best" features?
5. Explain the difference between K-means binning and equal-width binning with example.

Submission Details

- Download the Jupyter notebook as pdf and print it out.
- Organize exercises clearly under headers like 'Exercise 1'.
- Include question, code, output, and short explanations as comments.
- Header on each page must be
  ECSCI24302 Machine Learning Essentials
- Footer on each page must be
  [Lab2] [Page Number] [Enrollment number]
- Submission Deadline: In next lab