DS605 Fundamentals of Machine Learning
Academic Year 2025-26 (Autumn)

LAB-3

Objective: This assignment will test your ability to use NumPy and Pandas to explore, clean, and analyze a real-world dataset. You will investigate the factors that may have influenced a passenger's chance of survival on the Titanic.

Dataset

titanic.csv

Column Descriptions:

- **Survived**: 0 = No, 1 = Yes
- **Pclass**: Ticket class (1 = 1st, 2 = 2nd, 3 = 3rd)
- **Sex**: Passenger's gender
- **Age**: Passenger's age in years
- **SibSp**: Number of siblings / spouses aboard the Titanic
- **Parch**: Number of parents / children aboard the Titanic
- **Fare**: Price of the passenger's ticket
- **Embarked**: Port of Embarkation (C = Cherbourg, Q = Queenstown, S = Southampton)

Part 1: NumPy Exercises

First, load the dataset with Pandas, then extract the required columns into NumPy arrays to solve these problems.

1. **Array Creation and Manipulation:**
    - Load the Age and Fare columns into two separate NumPy arrays, ages and fares.
    - The ages array has missing values (NaN). Create a new array cleaned_ages where all NaN values are replaced with the **median** age of the passengers.
    - Create a new array fares_in_euros by converting the fares (which are in British Pounds) to Euros. Assume an exchange rate of **1 Pound = 1.18 Euros**.
2. **Statistical Analysis:**
    - Using NumPy, calculate and print the following statistics for the **cleaned_ages** array:
        - Average age
        - Standard deviation
        - The age of the oldest passenger
        - The age of the youngest passenger

3. **Boolean Indexing and Filtering:**
    - Create a boolean array that is True for passengers who paid a Fare **greater than 50** and False otherwise.
    - Use this boolean array to count how many passengers paid more than 50 for their ticket.
    - Display the ages of these high-paying passengers.

Part 2: Pandas Exercises

Use the Pandas library to perform a more in-depth analysis of the Titanic DataFrame.

1. **Data Loading and Initial Exploration:**
    - Load the titanic.csv file into a Pandas DataFrame called df.
    - Display the last 5 rows of the DataFrame.
    - Get a summary of the DataFrame's technical information (column names, data types, non-null counts).
2. **Data Cleaning and Preprocessing:**
    - The Cabin column has many missing values. **Drop** this column entirely from the DataFrame.
    - Fill the missing values in the Age column using the **mean** age.
    - The Embarked column has a few missing values. Fill them with the **mode** (the most common port of embarkation).
    - Confirm that your DataFrame now has no missing values in any of the remaining columns.
3. **Feature Engineering:**
    - Create a new column called FamilySize which is the sum of the SibSp and Parch columns, plus 1 (to count the passenger themselves).
    - Create a categorical column called AgeGroup. Bin the Age column into the following groups:
        - **Child**: 0-12
        - **Teen**: 13-19
        - **Adult**: 20-59
        - **Senior**: 60+

Image of different age demographics

4. **Data Selection and Querying:**
    - Display the Name, Sex, and AgeGroup of all passengers who **survived**.
    - Find all **female** passengers who were in **1st Class** (Pclass == 1).
    - Select all passengers who embarked from **Southampton ('S')** and paid a fare **less than 10**.
5. **Grouping and Aggregation (Groupby):**
    - What was the **average age** of passengers who survived versus those who did not? (Hint: Group by the Survived column).
    - Group the DataFrame by Pclass and Sex. For each group, calculate the **survival rate** (the mean of the Survived column). This will show you the survival probability for men and women in each class.

6. **Exporting the Final DataFrame:**
    - Save your final, cleaned DataFrame with the new FamilySize and AgeGroup columns to a new CSV file named titanic_processed.csv. Do not include the pandas index in the file.