Mini Project: Data Engineering on Titanic Dataset

Course: DTS 202-Data Engineering

Dataset: Titanic (Inbuilt dataset from Seaborn)

Objective: Perform comprehensive data engineering tasks such as data cleaning, transformation, feature engineering, and preparing the dataset for predictive modeling.

Problem Statement:

The Titanic dataset contains information about passengers aboard the Titanic and whether they survived or not. This project aims to apply data engineering techniques to pre-process the dataset and build a predictive model to determine whether a passenger survived based on various features.

Objectives:

By the end of this project, students will:

1. Perform **exploratory data analysis (EDA)** to understand the dataset.
2. Apply **data cleaning** techniques to handle missing values and outliers.
3. **Transform** the data using scaling, encoding, and feature engineering.
4. **Prepare** the dataset for predictive modeling by selecting relevant features.
5. Train and evaluate a simple **classification model** to predict survival.

Instructions:

1. Data Exploration

- ✓ Load the Titanic dataset (`seaborn.load_dataset('titanic')`) and display the first few rows.
- ✓ Explore the dataset structure and check for missing values.
- ✓ Visualize the relationships between key features and the target variable (Survived).

2. Data Cleaning

- ➢ Handle missing values: Impute missing values in Age, Embarked, and other columns as needed.

- ➢ Handle outliers: Use visualization tools like boxplots to detect and handle outliers in features like Fare.

3. Feature Engineering

- ✓ **Family Size**: Create a new feature by adding SibSp and Parch to form a FamilySize feature.
- ✓ **Age Binning**: Bucket the Age feature into categories (e.g., children, teenagers, adults, seniors).
- ✓ **Fare Binning**: Categorize the Fare feature into groups based on ticket price ranges.

4. Data Transformation

- ✓ Encode categorical variables like Sex, Embarked, and newly created features (AgeGroup, FareGroup).
- ✓ Scale numerical features (Fare, Age, FamilySize) using a standard scaler or min-max scaler.

5. Model Preparation

- ✓ Split the data into training and testing sets (70% training, 30% testing).
- ✓ Train a simple classification model (Logistic Regression) to predict survival.
- ✓ Evaluate the model using accuracy, confusion matrix, and classification report.

Report Submission Guidelines:

- **Title Page**: Include the project title, student's name, and submission date.
- **Introduction**: Briefly explain the dataset and objectives of the project.
- **Methodology**: Document all steps for EDA, data cleaning, transformation, feature engineering, and model training.
- **Results and Discussion**: Present key findings, model performance, and visualizations.
- **Conclusion**: Summarize insights and improvements made to the dataset.