DATA ENGINEERING ESSENTIALS – LAB ASSIGNMENT

Section: C (AIML)

[redacted]

Duration: 2 Hours
Total Marks: 10

Objective of the Assignment

To evaluate the student's ability to load, inspect, clean, transform real-world datasets, perform feature engineering, and extract meaningful insights through data analysis.

Datasets Provided (Section C)

Customer Churn Dataset
https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Spotify Songs Dataset
https://www.kaggle.com/datasets/lehaknarnauli/spotify-datasets

Mobile App Reviews Dataset
https://www.kaggle.com/datasets/lava18/google-play-store-apps

Credit Card Transactions Dataset
https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud

Weather Dataset
https://www.kaggle.com/datasets/selfishgene/historical-hourly-weather-data

E-Commerce Orders Dataset
https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce

Insurance Dataset
https://www.kaggle.com/datasets/mirichoi0218/insurance

Flight Delay Dataset
https://www.kaggle.com/datasets/usdot/flight-delays

Lab Tasks (Common for All Datasets)

Task 1: Load the dataset and perform initial inspection (head, shape, info, describe).
Task 2: Identify and handle missing values using appropriate strategies with justification.
Task 3: Transform categorical data using encoding techniques.
Task 4: Perform feature engineering by creating at least one new feature.
Task 5: Perform data analysis and write two meaningful insights.

Data Analysis – Conceptual Questions (Answer Any FIVE)

1. What does one row in your dataset represent in the real world?
2. Which column in your dataset is most useful for decision-making and why?
3. Which column would you remove before ML modeling? Justify your choice.
4. What type of bias might exist in your dataset?
5. How does feature engineering improve model performance?
6. What issues may occur if data cleaning is skipped?
7. Is your dataset more suitable for classification or regression? Why?
8. What assumptions does your dataset make about the real world?

Marks Distribution

Component	Marks
Data Loading & Inspection	2
Data Cleaning	3
Data Transformation	2
Feature Engineering	2
Data Analysis & Understanding	1

Important Instructions

Students must work individually. Both code and explanations are mandatory. Answers must be dataset-specific. Use of ChatGPT is permitted only for syntax help, not for analytical reasoning.