CSP7040-Assignment

Data Preprocessing, Visualization & ML Readiness

Assignment Type: Individual

IMPORTANT SUBMISSION INSTRUCTIONS
1. Submission format must be .ipynb only.
2. File naming format (STRICT):
   rollno.ipynb
   Example: G24AI1213.ipynb
3. One single notebook must contain code, visualizations, and markdown explanations.
4. Submissions not following the above rules will be penalized.

Students are expected to include clear markdown explanations for each part. Submissions with only code and minimal explanation will be penalized.

OBJECTIVE
Apply data preprocessing and visualization techniques and understand how poor data quality impacts machine learning models and the ML lifecycle.

DATASET
Netflix Movies and TV Shows Dataset
https://drive.google.com/file/d/187dSnEGn1g2t1UjSJwqevbW9vjh_sWMT/view?usp=sharing

TASKS

PART 1: Data Understanding & Quality Issues (3 Marks)
- Load the dataset and inspect the shape, data types, and missing values.
- Identify at least five data quality issues and explain them briefly.

PART 2: Data Cleaning & Preprocessing (4 Marks)
- Handle missing values with justification.
- Explain why a particular technique(dropping,filling with mean/median) was chosen and its impact on data quality
- Convert data into correct formats (dates, duration).
- Encode categorical variables.
- Normalize or standardize numerical features.

PART 3: Data Visualization (4 Marks)

- One distribution plot.
- One categorical count plot.
- One numerical comparison plot.
- Provide interpretation for each plot.

Each visualization must be followed by a brief interpretation explaining observed trends and their relevance to ML.

PART 4: ML Readiness Check (2 Marks)
- Select a target variable.
- Train Logistic Regression or Decision Tree.
- Compare performance before and after preprocessing.

Performance comparison should be based on accuracy (or similar metric) and briefly explain why preprocessing affected the results.

PART 5: ML Lifecycle Reflection (2 Marks)
- Explain which ML lifecycle stages fail without preprocessing.
- Give one concrete example from this assignment.

EVALUATION (Total: 15 Marks)

Data Understanding & Quality Analysis: 3
Data Cleaning & Preprocessing: 4
Data Visualization: 4
ML Model & Comparison: 2
Lifecycle Reflection: 2

PLAGIARISM POLICY

Plagiarism in any form will result in zero marks and may lead to disciplinary action as per institutional guidelines.