Lab Test: Data Encoding & Text Pre-processing

Tools Allowed: Python, Jupyter, Documentation

Part A: Data Encoding & Feature Engineering (30)

Objective: Analyze a mixed-type dataset, apply multiple encoding techniques, and justify choices.

Tasks:

1. **Data Profiling**:

- Identify types of variables (nominal, ordinal, binary, Numerical).
- Generate a report summarizing cardinality and missing values. *(10 marks)*

2. **Encoding Strategy**:

- Apply at least **three different encoding strategies**, selecting based on variable types:
  - Label Encoding (with ordinal assumptions)
  - One-Hot Encoding (with feature reduction if >10 unique values)
- Justify the encoding used for each column. *(10 marks)*

3. **Feature Interaction**:

- Create **two new interaction features** using existing columns (e.g., "Region x Product Category").
- Encode them appropriately. *(10 marks)*

Part B: Text Pre-processing (20 Marks)

Objective: Perform NLP-based preprocessing tasks on a text dataset.

Task:

1. Load the text dataset assigned to you (e.g., CSV or TXT).
2. Perform the following preprocessing steps:
   - Lowercasing
   - Removing punctuation and digits
   - Tokenization
   - Removing stop words
   - Lemmatization (or stemming if lemmatization tools not available)
3. Save the cleaned and processed text to a new file.

Datasets:

Dataset links for both parts are assigned separately in an Excel file for each enrolment number.

Submission:

1. **Create a word file and justify your choices in your own words**
2. **Save python files and screenshots of codes with outputs. Copy screen shots in word file.**
3. **Create a zip file and upload python and word files.**
4. **Upload it on LMS within due time.**