School of Computer Science Engineering and Technology

Course-BTech	**Type - AI Core-1**
Course Code - CSET211	**Course Name - Statistical Machine Learning**
Year - Second	**Semester - ODD**
Date - 19/08/2024	**Batch - CSE 3rd Semester**

Lab Assignment - 3: Implementation of Data Preprocessing Steps

CO- Mapping

Section	CO1	CO2	CO3
Section 1: Q1-Q4	✓		
Section 2: Q1-Q5	✓		✓

Section 1: General Data Processing Tasks

Given a dataset *data.csv* with missing values**.** Load the CSV file and write a Python script for each of the following tasks:

Q1: Identify the columns with missing values.

Q2: Encode categorical columns **`categorical_1 and categorical_2`** using one-hot encoding.

Q3: Encode categorical columns **`categorical_1 and categorical_2`** using label encoding.

Q4: Standardize numerical columns **`numerical_1 and numerical_2`** using StandardScaler.

Section 2: Sentiment Analysis

Given a dataset `sentiment.csv` with 100 rows and five columns *id, text, date, sentiment,* and *source* related to sentiment analysis. Perform the following data pre-processing steps on the CSV.

Q1: Write a Python code to load the CSV file into a pandas DataFrame.

Q2: Write a Python code to handle missing values by filling them with a placeholder value (e.g., *'Unknown'*). ***(Note: You can first search for missing values in the dataframe)***

Q3: Write a Python code to remove punctuation from the **`text`** column.

Q4: Write a Python code to tokenize the **`text`** column using *word_tokenize* method.

Q5: Write a Python code to convert the tokenized text back to a single string for each row.