**GTSC2143 Machine Learning for Business**

**Tutorial 6**

**Please write down your answers in this document and submit it at iSpace by the end of this tutorial.**

### Data Loading and Preprocessing

1. Load and Explore the Dataset

1. Load the Amazon baby product reviews dataset using pandas:

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.feature\_extraction.text import CountVectorizer

from sklearn.linear\_model import LogisticRegression

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_matrix

# Load the Amazon baby product reviews dataset

data = pd.read\_csv("GTSC2143-Lecture 6\_analyzing-product-sentiment-assignment\_amazon\_baby.csv",index\_col=0)

b) Check basic information:

* Dataset shape
* Column names
* Check any missing value
* Drop records with missing value

2. Create Sentiment Labels

1. Create a new column called ‘positive’ where the value is 1 if the rating is greater than 3, and 0 otherwise
2. Display the distribution of sentiment labels

### Data Splitting and Text Processing

1. Train/Test Split

1. Split the data into training (80%) and testing (20%) sets using random\_state=42
2. Display the shapes of training and testing sets

2. Convert Text to Features

1. Use CountVectorizer to convert review text into word count features
2. Set max\_features=1000 to limit vocabulary size
3. Fit the vectorizer on training data and transform both training and test texts
4. Display the shape of the feature matrices
5. Analysis: Write 2-3 sentences explaining how text becomes numerical features.

### Model Training and Evaluation

1. Logistic Regression Model

1. Train a logistic regression classifier using the word count features
2. Use random\_state=42 for reproducible results

2. Evaluate the Model

1. Make predictions on the test set
2. Calculate and display:
   * Accuracy score
   * Classification report
   * Confusion matrix
3. Analysis: Write 2-3 sentences interpreting the model's performance.

3. Feature Analysis

1. Display the top 10 most positive words (highest coefficients)
2. Display the top 10 most negative words (lowest coefficients)
3. Analysis: Write 2-3 sentences about which words drive sentiment predictions.

- End of Tutorial 6 -