BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department

Machine Learning and Data Science - ENCS5341

Assignment #1: Data Preprocessing & Exploratory Data Analysis (EDA)

Submission deadline: 24.10.2025

For this assignment, we will be using a synthetic dataset that simulates a customer database. The data includes customer demographics, product usage, and customer status (whether they are likely to churn or not). The programming language for this assignment is `python`, and you can use libraries like `pandas`, `seaborn`, and `matplotlib` for data manipulation and visualization.

Features:

- **CustomerID** (Numeric): Unique identifier for each customer.
- **Age** (Numeric): The age of the customer.
- **Gender** (Categorical): Gender of the customer (0: Male / 1:Female).
- **Income** (Numeric): Annual income of the customer (in USD).
- **Tenure** (Numeric): Number of years the customer has been with the company.
- **ProductType** (Categorical): Type of product the customer has subscribed to (0: Basic / 1: Premium).
- **SupportCalls** (Numeric): Number of support calls made by the customer in the last year.
- **ChurnStatus** (Binary): Whether the customer churned (1) or stayed (0).

Steps for the Assignment:

1. Data Loading and Initial Inspection:
   - Load the dataset into a `pandas` DataFrame.
   - Inspect the first few rows of the dataset using `.head()` and check the general information using `.info()`.
   - Generate summary statistics using `.describe()`.

2. Handling Missing Data:
   - Check for missing values in the dataset using `.isnull().sum()`.
   - Decide how to handle missing values (e.g., imputation or deletion). **Ensure to justify your approach**.

3. Handling Outliers:
   - Identify numerical outliers using box plots or Z-scores.
   - Discuss strategies to handle outliers.
   - Provide **justification** for your approach to handling outliers.

4. Feature Scaling:
   - Normalize numerical features using methods like Min-Max Scaling or Standardization.

5. Exploratory Data Analysis (EDA):
   - Univariate Analysis:
     - Visualize the distribution of key numerical features using histograms or box plots.
     - Analyze the distribution of categorical variables using bar plots.
   - Bivariate Analysis:
     - Explore relationships between numerical features and the target variable (ChurnStatus) using scatter plots or box plots.
     - Investigate relationships between categorical variables and the target using bar plots.
   - Correlation Analysis:
     - Create a correlation matrix to examine how numerical features relate to each other and to the target variable.
     - Discuss which features seem most predictive of churn.

6. Data Visualizations:
   - Create at least 4 visualizations that communicate insights from the data (e.g., pair plots, bar plots, heatmaps, etc.).
   - Ensure that visualizations are clear, well-labeled, and easy to interpret.

7. Conclusion:
   - Summarize the key findings from the EDA.
   - Discuss any significant patterns, relationships, or insights that could help improve customer retention or predict churn.

Deliverables:

1. **A Jupyter Notebook** (or Python script) containing the code for data loading, cleaning, preprocessing, and visualization.
2. **A Written Report** summarizing the preprocessing steps, justifications for the preprocessing decisions, visualizations with proper explanations, and key findings from the analysis.