DL 2001 - Introduction to Data Science
Final Project
BSDS-3(A,B,C)

Instructions:

- **Maximum of 2 students per group are allowed**
- **Dataset (electronics.json) for the project is uploaded on the portal.**
- **The codes must be in running form.**
- **You can use built-in libraries.**
- **You need to preprocess the dataset before using it.**
- **You need to visualize your findings.**
- **Your approach to the problems will highly be seen.**
- **You must complete the project before the deadline (3rd December, 2024)**
- **Late submissions won't be entertained.**
- **Plagiarism will not be entertained.**
- **AI generated codes will be dealt with strict actions.**

Introduction:

Imtiaz Mall, a renowned department store chain, is experiencing declining sales and a significant number of non-recurring customers in its electronics section. To address this challenge, you, the newly appointed Senior Data Scientist, have been tasked with conducting a comprehensive analysis of the electronics section data and developing data-driven strategies for customer retention and sales growth. This project focuses on the initial steps of this analysis, specifically exploring the data through various techniques.

Module 1: Data Acquisition and Preprocessing:

1. **Data Acquisition:**
    - Download the provided historical sales data for the electronics section.
    - Ensure the data includes customer demographics, purchase history, product details, spending amounts, and dates of transactions.
2. **Data Cleaning:**
    - Identify and handle missing values using appropriate techniques like mean/median imputation or dropping rows/columns with excessive missingness.
    - Analyze outliers and determine whether to retain or remove them based on their impact on the analysis.
    - Address inconsistencies in data format and encoding.
3. **Data Transformation:**
    - Create new features that provide deeper insights into customer behavior, such as:
        - Average spending per purchase
        - Purchase frequency per month

- Brand affinity score (based on product brand preferences)
- Product category preferences (e.g., TVs, smartphones, laptops)

- Standardize or normalize numeric features to ensure they contribute equally to the given algorithms.

Module 2: Exploratory Data Analysis (EDA):

1. **Univariate Analysis:**
   - Analyze the distribution of key features like customer age, purchase amount, and purchase frequency using histograms, boxplots, and descriptive statistics.
   - Identify potential skewness or outliers in the data.
2. **Bivariate Analysis:**
   - Utilize scatterplots and heatmaps to explore relationships between different features, such as purchase amount vs. income level, brand affinity vs. product category, and purchase frequency vs. age.
   - Investigate the presence of correlations and identify any impactful relationships.
3. **Temporal Analysis:**
   - Analyse trends in customer behaviour over time, including changes in purchase frequency, average spending, and product preferences.
   - Identify seasonal variations or any significant shifts in customer behavior patterns.

Module 3: Regression and Decision Tree Analysis:

A. Linear Regression Analysis:

1. **Problem Definition:**

   - Predict the average spending per purchase based on customer demographics and purchase history.

2. **Model Building:**
   - Preprocess the data by selecting relevant numerical and categorical variables (e.g., income level, product category, age).
   - Split the dataset into training and testing sets.

3. **Implementation:**

   - Train a linear regression model using the training data.

- Evaluate the model using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared.

4. **Visualization:**

- Plot the predicted vs. actual values for the test dataset.
- Include regression lines for better interpretability.

B. Decision Tree Analysis:

1. **Problem Definition:**

- Classify whether a customer will make a purchase in the next month (use a binary target variable).

2. **Model Building:**

- Engineer a binary target variable (e.g., `1 = purchase made, 0 = no purchase`).
- Use features like purchase frequency, spending history, and product preferences.

3. **Implementation:**

- Train a decision tree classifier and use criteria such as Gini Impurity or Entropy.
- Evaluate the model using metrics such as Accuracy, Precision, Recall, and F1 Score.

4. **Visualization:**

- Plot the decision tree.
- Highlight important features that influence the decision.

Module 4: Clustering Analysis:

(**Hint: Remove the predicted label and then apply K-Means Clustering)**

1. **Define the number of clusters(k):**
   - Analyze the elbow plot to determine the optimal number of clusters based on the sum of squared distances within each other.
2. **Apply K-Means Clustering:**
   - Implement K-means with the chosen k value to segment customers into distinct clusters based on their purchase behavior and preferences.

3. **Analyze cluster characteristics:**
    - Investigate key features of each cluster, such as average purchase amount, brand affinity and product category preferences.
    - Identify significant differences and similarities between clusters.

Module 5: Comparison and Conclusion:

1. Compare the predictive performance of the regression,decision tree and K-Means Clustering models.

- Discuss strengths, limitations, and real-world applicability in the context of customer behavior analysis.

2. Provide actionable recommendations for the electronics section based on the results.