**ALTERNATIVE ASSESSMENT 2 (60 marks) - WEEK 14**

Answer the question below based on the given scenario. Submit your answer within **ONE (1) DAY** after the question is given in SPECTRUM. Answers should be submitted and saved with the student’s name followed by matric number as the file name in the format of .pdf (e.g. Ali\_s123456.pdf).

**Case Study: E-Commerce Customer Behaviour Analysis**

**Background:**

You will work with a dataset of customer transactions from an e-commerce platform, encompassing various customer attributes, purchase history, and behavioural insights collected over the last year. Additionally, you are required to generate synthetic data to enrich the dataset, ensuring it is representative of realistic customer behaviours. You may also include social media interaction and browsing history data. The structure provided below is a guideline, but feel free to enhance this dataset by adding attributes that reflect more modern e-commerce and digital marketing practices. For example, you could include customer interactions from social media platforms, website clickstream data, or engagement metrics like time spent on product pages.

**Dataset Structure (including synthetic data generation):**

* **CustomerID:** Unique identifier for each customer.
* **Age:** Age of the customer.
* **Gender:** Gender of the customer.
* **Location:** Geographic location of the customer.
* **MembershipLevel:** Indicates the membership level (e.g., Bronze, Silver, Gold, Platinum).
* **TotalPurchases:** Total number of purchases made by the customer.
* **TotalSpent:** Total amount spent by the customer.
* **FavoriteCategory:** The category in which the customer most frequently shops (e.g., Electronics, Clothing, Home Goods).
* **LastPurchaseDate:** The date of the last purchase.
* **WebsiteClickRate:** The average number of clicks per session on the website.
* **TimeSpentOnSite:** Average time spent by the customer on the website during each visit.
* **SocialMediaEngagement:** Level of engagement on social media (e.g., likes, shares, comments).
* **AdClickHistory:** Interaction with online advertisements (e.g., click-through rate, viewed ads).
* **CustomerSentimentScore:** Sentiment analysis score derived from customer reviews or interactions.
* **Churn:** Indicates whether the customer has stopped purchasing (1 for churned, 0 for active).

**Tasks (including synthetic data generation):**

1. **Data Import, Synthetic Data Generation, and Preprocessing:** Import your dataset into any data mining tool of your choice (e.g., Google Colab, AWS Sagemaker, Azure Machine Learning Studio, SAS, RapidMiner, Weka). Generate synthetic data to supplement the original dataset using methods such as random sampling. Handle missing values appropriately, specify variable roles, and perform feature engineering to create new relevant attributes, such as calculating Recency, Frequency, and Monetary (RFM) metrics. Additionally, consider applying data normalization or scaling techniques for specific variables where applicable.

(18 marks)

1. **Decision Tree Analysis:** Create a decision tree model using any modern data mining or machine learning platform of your choice (e.g., Scikit-Learn, AWS Sagemaker, Azure Machine Learning Studio, Google AI Platform). Provide a visualization of the decision tree and interpret the key splits that help understand customer purchasing patterns.

(15 marks)

1. **Advanced Techniques in Customer Segmentation:** Apply clustering methods, such as K-means and DBSCAN, to segment customers based on behavioural data using any suitable data mining tool (e.g., Weka, RapidMiner, Azure ML). Use Principal Component Analysis (PCA) to reduce dimensionality if necessary. Evaluate the results and interpret the clusters to generate actionable business insights.

(12 marks)

1. **Ensemble Methods:** Apply Bagging, Boosting (e.g., AdaBoost, XGBoost), and Random Forests to predict customer churn using any data mining tool or cloud service. Compare the performance of each method using metrics like accuracy, F1 score, and AUC-ROC curve. Summarize your findings on which method performs best and why.

(10 marks)

1. **Deliverables:**
   * A report detailing each step of the process, including the rationale behind your choices and any challenges faced.
   * An analysis of the decision tree, clustering, and ensemble methods, with insights into customer behaviour and suggestions for business strategy. Provide a visualization of the key results (e.g., decision tree plot, clustering scatter plots, feature importance from Random Forest).

(5 marks)

**Objective:** This case study aims to assess your ability to apply modern data mining techniques, including decision trees, clustering, ensemble methods, and feature engineering, in a practical context. You should demonstrate your understanding of these concepts and your ability to derive meaningful business insights from data analysis using any suitable tool or cloud service.

# END