HDD Lab (STA 0542 5188)

Assignment # 02

Marks # 10

Deadline: 10 January 2026

Course Instructors: *Dr. Md Jamal Uddin & Dr. Mohammad Anamul Haque*

TA: Md Sadakin Islam

Instructions:

1. Read ***Data Descriptions*** first and answer the questions 1-6.
2. While preparing your answer script write and report only relevant information's resulted from your analysis and figures.
3. Write your registration number at the top of each file and submit the following files in google classroom:
   - I. Output files with relevant tables, graphs, and interpretations in PDF format
   - II. Python script
   - III. A Tableau (.twbx) or Power BI (.pbix) dashboard file

Data descriptions:

The dataset is a synthetic high-dimensional customer analytics consists of 1000 observations and 30 variables, representing demographic, financial, behavioral, transactional, and engagement-related characteristics of customers.

SL	Variable	Description	Type/ Label
1	Age	Age of the customer in years	Continuous Scale
2	Income	Annual income of the customer	
3	Education_Years	Years of formal education	
4	Experience_Years	Years of work experience	
5	Spending_Score	Spending behavior score (1–100)	
6	Savings	Total savings amount	
7	Credit_Score	Creditworthiness score	
8	Debt	Total outstanding debt	
9	Hours_Online	Daily hours spent online	
10	Purchases_Month	Monthly number of purchases	
11	Website_Visits	Monthly website visits	
12	Time_On_App	Time spent on app (minutes)	
13	Social_Media_Usage	Daily social media usage (hours)	
14	Product_Returns	Number of product returns	
15	Customer_Tenure	Customer relationship duration (months)	
16	Satisfaction_Score	Customer satisfaction score (1–10)	
17	Complaints	Number of complaints	
18	Support_Call_Count	Number of customer support calls	
19	Ad_Clicks	Advertisement clicks	
20	Email_Open_Rate	Proportion of emails opened	
21	Discount_Usage	Proportion of discounted purchases	
22	Mobile_App_Usage	Mobile app usage intensity	
23	Transactions_Value	Total transaction value	
24	Loyalty_Points	Accumulated loyalty points earned	
	Fraud_Risk_Score	Estimated fraud risk	
26	Dropout_Probability	Estimated probability that a customer will discontinue the service	
27	Market_Segment	Customer market segment	1 = Budget, 2 = Price-sensitive, 3 = Regular, 4 = Premium, 5 = VIP
28	Region_Code	Geographic region identifier	1= North, 2 = South, 3 = East, 4 = West, 5 = Central, 6 = North-East, 7 = South-West
29	Channel _ Preference	Preferred interaction channel	1 = Website, 2 = Mobile App, 3 = Physical Store
30	Risk_Class	Risk classification label	0= Low-risk, 1= High-risk

HDD Lab (STA 0542 5188)
Assignment # 02
Marks # 10
Deadline: 10 January 2026
Course Instructors: *Dr. Md Jamal Uddin & Dr. Mohammad Anamul Haque*
TA: Md Sadakin Islam

Answer the following questions:

1. **(a)** ✓ Create appropriate scatter plots to explore the relationship between:
   i. **Income** and **Spending_Score**
   ii. **Income** and **Spending_Score** colored by **Risk_Class**
   Interpret the observed patterns.

(b) ✓ Use suitable visualizations to examine the distribution of:
   i. **Credit_Score**
   ii. **Transactions_Value** across **Market_Segment**
   Interpret the findings.

(c) Construct a correlation heatmap using 10 important numerical variables. Identify and explain at least three strong relationships.

2. Construct a parallel coordinates plot using at least five numerical variables and color by **Risk_Class**.

(a) Based on the plot, explain the curse of dimensional visualization, highlighting issues such as overplotting and interpretability.

(b) Discuss why parallel coordinates plots are useful or exploring high-dimensional data but become difficult to interpret as the number of observations increases.

(c) Briefly explain how this motivates the use of dimensionality reduction techniques.

3. **(a)** Standardize the numerical variables and perform Principal Component Analysis (PCA). Explain why standardization is necessary before applying PCA.

(b) Produce a scree plot and determine the number of principal components required to explain at least 80% of the total variance.

(c) Visualize the data using the first two principal components and color by **Risk_Class**. Interpret whether the reduced representation reveals any separation between risk groups.

(d) Identify the variables that contribute most to the first principal component using the PCA loadings. Explain the importance of these variables in defining the dominant source of variation in the data.

(e) Perform **Principal Component Regression (PCR)** using the selected principal components to predict **Dropout_Probability**. Explain why PCR may be preferred over ordinary linear regression in high-dimensional settings.

4. **(a)** Apply t-SNE to the dataset and visualize the result.

(b) Apply UMAP and visualize the result. Compare the visualization obtained from UMAP with that of t-SNE.

(c) Compare PCA, t-SNE, and UMAP in terms of:
   i. Interpretability
   ii. Ability to preserve structure (local vs global)
   iii. Computational complexity

5. **(a)** Perform K-means clustering on the PCA-reduced data. Use the elbow method to justify the choice of the number of clusters.

(b) Visualize the clustering results in the PCA space and interpret the cluster structure.

6. Using **Tableau or Power BI**, perform the following tasks on the given customer dataset:

(a) Create a histogram of **Dropout_Probability** and describe the overall distribution.

(b) Create a bar chart showing the average **Dropout_Probability** by **Channel_Preference**.
   Use different colors to distinguish channels.

(c) Create a scatter plot of **Support_Call_Count** vs **Dropout_Probability**.
   Color the points by **Risk_Class** to identify low- and high-risk customers.

(d) Apply clustering or grouping on the scatter plot to identify customer groups based on risk behavior.

Good Luck