Week 24 - Graded Mini Project

Learning Outcome Addressed

- Understand the core concepts and techniques of unsupervised learning.
- Learn how to implement clustering algorithms like K-Means and Hierarchical Clustering to group unlabeled data.
- Explore dimensionality reduction techniques such as PCA to reduce the complexity of datasets while retaining important information.
- Apply unsupervised learning algorithms to real-world datasets to uncover hidden patterns and relationships.

Introduction

A leading international bank is facing challenges with customer retention. Despite offering multiple financial products and services, many customers are closing their accounts and leaving for competitors. The bank's management wants to leverage data analytics to predict customer churn and design strategies to improve customer loyalty.

As a data analyst, you are tasked with analyzing customer data, identifying factors leading to churn, and building a simple model to predict whether a customer will exit the bank.

Submission Instructions

Please document your response on the following pages.

Once you have completed the activity, save the file as a PDF and upload it. Be sure to name the file as **Module 24: Graded Mini Project_[Your last name].**

Your submission will be considered complete when it meets these criteria:

- Includes all the key elements outlined in the activity instructions and the rubric.
- Adheres to the submission guidelines.
- Is submitted on time.

This is a required activity and counts towards programme completion.

Reflect on the task and respond to the following questions.

Data Description

You are provided with a dataset (data.csv) that contains customer information along with a churn flag (Exited).

Column Name	Description
CustomerId	Unique identifier for each customer (may contain missing values/duplicates).
Surname	Customer's last name (categorical, not very useful for prediction).
CreditScore	Credit score of the customer.
Geography	Customer's country (France, Spain, Germany).
Gender	Customer's gender (Male/Female, some missing values).
Age	Age of the customer (contains missing values).
Tenure	Number of years the customer has been with the bank.
Balance	Account balance of the customer.
NumOfProducts	Number of bank products the customer uses.
HasCrCard	Whether the customer has a credit card (1 = Yes, 0 = No).
IsActiveMember	Whether the customer is an active member (1 = Yes, 0 = No).
EstimatedSalary	Estimated annual salary of the customer.
Exited	Target variable: 1 = Customer exited, 0 = Customer stayed.

Note: The dataset intentionally contains missing values and duplicate records to simulate real-world challenges.

Project Objective

The goal is to build a predictive pipeline that helps the bank answer the question: "Which customers are at risk of leaving the bank?"

This will help the bank's marketing and customer success teams target the right customers with retention strategies.

Task 1: Data Cleaning & Preprocessing

- Handle missing values.
- Detect and remove duplicate records.
- Convert categorical variables into numerical form.
- Scale numerical features (only) where necessary.

Task 2: Exploratory Data Analysis (EDA)

- Explore the distribution of Exited (churn vs non-churn).
- Find key patterns:
  - Does age affect churn?
  - Are certain geographies more likely to churn?
  - Does credit score or balance play a role?
- Visualize churn rates across different customer groups.

Task 3: Predictive Modeling

- Split the dataset into training and test sets.
- Train at least two machine learning models (e.g., Logistic Regression, Random Forest).
- Evaluate using accuracy, precision, recall, and F1-score.

Task 4: Insights & Recommendations

- Identify the most important factors influencing churn.
- Suggest business actions (e.g., targeted offers, loyalty programs) to improve retention.

Deliverables

Your project should include:

- Jupyter Notebook (with data cleaning, EDA, and model building).
- Summary report highlighting findings, model performance, and recommendations.