Assignment 3

1 Introduction

In this assignment, you will analyse a real-world dataset containing information about credit card customers. Your task is to explore the data and build machine-learning models to gain insights into customer churn. You can work alone or in teams of 2.

What to submit: a Jupyter notebook that processes the data file provided here and that is retrievable in your code as "datasets/customer_churn.csv". There is a penalty of 20% if this instruction is not followed, as your code will not run for grading.

1.1 Dataset Description

You will work with a dataset focused on customer churn in the credit card industry. ***Churn*** is a term used to indicate when customers discontinue their relationship or stop using a product or service. In this case, churn refers to customers who have either stopped using the credit card (*churned* or *attrited customer*) or are still using the credit card (*existing customers*).

In total, the dataset contains 10,127 rows and 21 columns. You can find a brief description of each column on Table 1.

2 Assignment Tasks

You will implement various tasks aiming to analyse and predict customer churn. Your goal is to develop a machine learning model that can identify potential attrited customers (or churners). The tasks include exploratory data analysis to understand the dataset, data preprocessing to prepare the data for machine learning, building predictive models using machine learning algorithms, and interpreting the results to extract meaningful insights.

You should provide clear explanations and reasoning for your solutions in each question. **Your code implementation will account for 60% of the final grade for each question, while explanations and comments will contribute 40%.**

2.1 Exploratory Data Analysis (15 points)

1. Calculate descriptive statistics for all numeric features of the dataset. Include at least min, max, mean, and standard deviation.
2. Compare the descriptive statistics of attrited and existing customers. Describe the most significant differences you identify.
3. Calculate the distribution (in percentage) of all the categorical features of the dataset. For instance, for the *attrition_flag* column you should calculate the percentage of existing and attrited customers.
4. Select **three** features. Create plots showing the relationship between these features and *attrition_flag*. You can use any type of plot, but make sure you explain your decision.

2.2 Data Preprocessing (15 points)

1. Calculate the number of missing and duplicate values on the dataset. Present the results per column.
2. Handle the missing and duplicate values. You can use any approach, but make sure you explain your decision.
3. Encode the categorical values to use them as input for machine learning models.

2.3 Machine Learning (30 points)

1. Decide how you will evaluate your models (train/test split or cross-validation). Implement the solution you chose and explain your reasoning.
2. Select **two** different classification models and train them to predict customer churn.
3. Use hyperparameter-tuning to improve your models. You should tune at least **two** features for each model.
4. Compare the performance of each model using a classification report. Which model had the best performance? Describe the most significant differences you identify.

2.4 Model Evaluation and Analysis (40 points)

1. Select **two** models. Create confusion matrices to compare their performance. Describe which error is more common and explain potential reasons for that.
2. Compare the precision, recall, and accuracy of these models. Describe what each metric represents and explain what conclusions you can draw from this analysis.
3. Aggregate the data by customer income category. Evaluate your models for each group. Compare the performances and describe the most significant result – e.g. do the models perform systematically worse for a specific income category?

Column	Description
clientnum	Unique identifier for each customer
attrition_flag	Indicates whether the customer has churned or not (*Existing* or *Attrited Customer*)
customer_age	Age of the customer
gender	Gender of the customer
dependent_count	Number of dependents the customer has
education_level	Education level of the customer
marital_status	Marital status of the customer
income_category	Income category of the customer
card_category	Category of the credit card
months_on_book	Number of months the customer has been a cardholder
total_relationship_count	Total number of products held by the customer
months_inactive_12_mon	Number of months with no transaction in the last 12 months
contacts_count_12_mon	Number of customer contacts in the last 12 months
credit_limit	Credit limit on the credit card
total_revolving_bal	Total revolving balance on the credit card
avg_open_to_buy	Average open-to-buy credit line in the last 12 months
total_amt_chng_q4_q1	Change in transaction amount from Q4 to Q1
total_trans_amt	Total transaction amount in the last 12 months
total_trans_ct	Total transaction count in the last 12 months
total_ct_chng_q4_q1	Change in transaction count from Q4 to Q1
avg_utilization_ratio	Average card utilization ratio

Table 1: Description of the dataset columns