INDEX OF CONTENTS

Introduction 2

General considerations 2

Steps to follow.................... 2

1. Simplified EDA (0.5 points).................... 2
2. Basic methods: Trees, KNN, and logistic regression (1 point) 3
3. Hyperparameter optimisation/tuning (1 point).................... 3
4. Advanced methods (1.5 points) 3
5. Results and final model (0.5 points).................... 3

What to hand in 4

Variable description 4

INTRODUCTION

The purpose of the first assignment is to practice with machine learning methods, both basic and advanced, including hyper-parameter tuning and preprocessing the data to adapt the dataset to the ML methods (encodings, imputation, constant features, etc.).

The topic of this assignment is predicting whether a bank customer will subscribe to a term deposit. A bank would like to build a predictive model based on customer demographic, financial and interaction data, in order to identify clients more likely to accept the product when contacted. The target variable is *deposit*.

GENERAL CONSIDERATIONS

1. Results **must be reproducible**. Therefore, set the seed at the appropriate places. But instead of using seed 42, use your **Student ID number**.

2. There are **two datasets**: the available data set (for model training, hyper-parameter tuning, and model evaluation) and the competition dataset (for using the model: making predictions for future instances). At the bottom of this document, you will find the names and meaning of each of the variables in the datasets.

3. Each group must use a different available data set. The supplied datasets have the names **bank_xx.pkl** (in **pickle** format), where **xx** is the order in the student list of one of the group members, and **bank_competition.pkl.** The competition dataset is common for all students.

4. The model evaluation method for this assignment will be **Holdout** (train/test). The main metric will be **Accuracy**. Confidence intervals for Accuracy can be reported, as can other metrics with a justification of their use. Any add-ons to the solution to the assignment **will be taken into consideration** for the grade.

5. **Execution time** of the training process for all methods (fit) should also be reported.

6. **Preprocessing** should be conducted using **pipelines when appropriate** and using the required preprocessing steps for each of the chosen methods.
   Important: be careful with variable ***pdays***: analyze it carefully and decide what is the best way to deal with it

STEPS TO FOLLOW

1. SIMPLIFIED EDA (0.5 POINTS)

Do a **simplified EDA**, mainly to determine how many features and how many instances there are, which variables are categorical/numerical, what categorical variables have high cardinality, which features have missing values and how many, whether there are constant columns, and whether it is a regression or classification problem. If the latter, is it imbalanced?

Analyze particularly the variable named ***pdays***, and justify the preprocessing method you have chosen for this variable.

2. BASIC METHODS: TREES, KNN, AND LOGISTIC REGRESSION (1 POINT)

1. Decide on an **appropriate test set size** and justify your answer.
2. Train, evaluate and **compare at least two basic methods** with default hyperparameters.
3. Considering the results, you might want to change the size of the test set and repeat the train and test steps. **Justify your final choice of train/test subset sizes with arguments**. **Keep these sizes for the remainder of the assignment**.
4. **Compare the final basic method of your choice with the dummy method**. Guarantee with arguments that the chosen method is more performant than the dummy method.
5. **Provide some visualization of the chosen methods** to support your understanding of how good decisions can be made for this problem.

3. HYPERPARAMETER OPTIMIZATION/TUNING (1 POINT)

1. Do **hyperparameter tuning for the two best performant basic models** first with a prearranged parametrical space and then with a sampled parametrical space from an appropriate distribution. Use at least two methods for each choice of the parametrical spaces from those studied in the course (GridSearch, RandomSearch and Optuna).
2. At this stage, summarize your results and draw some conclusions. Based on your findings, decide on one of the two HPO methods for the remainder of the assignment and justify your answer.

4. ADVANCED METHODS (1.5 POINTS)

Using **only the HPO method that you considered worked better in the previous section**, carry out the same analysis (i.e. try default values and hyper-parameter optimization) as before **with at least two advanced methods,** and compare both the HPO basic model chosen and advanced models:

1. **Support Vector Machines**.
2. **Random Forests** with default hyper-parameters (usually RF's default values work well)
3. **A Boosting Method**: Peruse the documentation for these methods and use them appropriately. Consider as a choice any of these methods: *gradientboosting*, *histgradientboosting*, *lightgbm*, *xgboost*, *catboost*. You can compare the results of two of them to support your choice.

5. RESULTS AND FINAL MODEL (0.5 POINTS)

1. **Report your results**: use a table, report confidence intervals for accuracy, execution times, and draw some conclusions on the different models analyzed.
2. **Using the best method**, train the final model and use it to **make predictions on the competition dataset**. Save both the **final model** in an **appropriate ML format** and the **competition prediction** in a **compatible format** (for instance, **pickle**, **csv**, and so on).

WHAT TO HAND IN

- A **jupyter notebook** with the code in the proposed order of steps. Please **use some of the cells to comment** about what you are doing and your results. In particular, emphasize your conclusions after each step with short arguments based on your results.

  If it is more convenient, you can also hand in a file with Python code instead and a separate report.

  If you decide to use any AI chatbot, **briefly explain in those commented cells what purpose you used it for and how you used it (for instance, you can quote the prompt and the output used)**.

  Please **write the names of the components of your group at the beginning of the notebook**. **Indicate briefly the contribution of each participant in the group.**

- A file containing your **final trained model** in an appropriate ML format.

- A pickle or .csv file containing **your final model's predictions (values of your model's predictions** on the competition set).

VARIABLE DESCRIPTION

Variable	Short Description
age	age in years
job	type of job
marital	marital status
education	education level
default	has credit in default?
balance	average yearly balance
housing	has a housing loan?
loan	has a personal loan?
contact	contact communication type
day_of_week	last contact day
month	last contact month
duration	last contact duration, in seconds
campaign	number of contacts performed during this campaign and for this client
pdays	number of days that passed by after the client was last contacted from a previous campaign. The value is -1 if no/unknown contact was produced.
previous	number of contacts performed before this campaign and for this client
poutcome	outcome of the previous marketing campaign
deposit	has the client subscribed a term deposit? TARGET VARIABLE