Group Project Description

I. Introduction

This project aims to facilitate student groups in completing two data mining tasks that encompass the fundamental steps of the data mining workflow: data understanding, data preparation, modeling, evaluation, and deployment. Task-1 requires a comprehensive understanding of the provided structured data, while Task-2 encourages the processing of unstructured data to create a deployable data mining pipeline.

II. Task-1 Bank Marketing with Structured Data

2.1 Task Description

This task offers student groups the opportunity to apply data mining techniques to real-world business problems. You are expected to provide recommendations to a commercial bank's management team regarding a marketing strategy for a specific product (term deposit).

You will work with a **training dataset ("bank_marketing_train.csv")** related to direct marketing campaigns conducted by a commercial banking institution, primarily via phone calls. Often, multiple contacts were required to determine whether a client would subscribe to the product (bank term deposit).

The training dataset consists of 26,246 observations and 25 variables (24 input features and 1 target variable y), detailed as follows:

1) ***age*** (numeric)
2) ***job***: type of job (categorical: 'admin.','blue-collar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
3) ***marital***: marital status (categorical: 'divorced','married','single','unknown'; note: 'divorced' means divorced or widowed)
4) ***education*** (categorical: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course','university.degree','unknown')
5) ***default***: has credit in default? (categorical: 'no','yes','unknown')
6) ***housing***: has housing loan? (categorical: 'no','yes','unknown')
7) ***loan***: has personal loan? (categorical: 'no','yes','unknown')
8) ***contact***: contact communication type (categorical: 'cellular','telephone')
9) ***month***: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
10) ***day_of_week***: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
11) ***campaign***: number of contacts performed during this campaign and for this client (numeric, includes last contact)

12) ***pdays***: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
13) ***previous***: number of contacts performed before this campaign and for this client (numeric)
14) ***poutcome***: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success')
15) ***emp.var.rate***: employment variation rate - quarterly indicator (numeric)
16) ***cons.price.idx***: consumer price index - monthly indicator (numeric)
17) ***cons.conf.idx***: consumer confidence index - monthly indicator (numeric)
18) ***euribor3m***: euribor 3-month rate - daily indicator (numeric)
19) ***nr.employed***: number of employees - quarterly indicator (numeric)
20) ***feature_1***: feature with unknown meanings (numeric)
21) ***feature_2***: feature with unknown meanings (numeric)
22) ***feature_3***: feature with unknown meanings (categorical)
23) ***feature_4***: feature with unknown meanings (categorical)
24) ***feature_5***: feature with unknown meanings (categorical)
25) ***y***: has the client subscribed to a term deposit? (binary: 'yes', 'no')

Then you are required to build a (binary) classification model to predict the ranking score of being positive (*y*=’yes’) that achieve the best AUC score of the ROC curve on the training dataset, and then predict the **ranking score of being positive(y=’yes’)** for the 8,000 clients in the “**bank_marketing_test.csv”,** which has 24 input variables/features without the target variable ***y*.**

Your task is to build a binary classification model to predict the likelihood (*ranking score*) of a positive subscription (*y*='yes') that achieves the best AUC score of the ROC curve on the training dataset. Then you will use the above model to predict the subscription likelihood (*ranking scores*) for 8,000 clients in “bank_marketing_test.csv”, which contains 24 input variables/features without the target variable ***y***.

The adopted model may include one of the following introduced in the course:

- Decision Tree
- Logistic Regression
- SVM
- Naïve Bayes
- KNN
- Ensemble models of previous base models (e.g., Random Forest)

Please note that NO external resources should be utilized.

2.2 Provided resources

1) **"bank_marketing_train.csv"** comprising 26,246 observations with 25 variables (24 input variables and 1 target variable y)

2) **"bank_marketing_test.csv"** comprising 8,000 observations with 24 input variables without the target variable y
3) **"bank_marketing_test_scores(example).csv"** which illustrates the expected content and format of the ranking scores for the 8,000 observations generated by your model.

III. Task-2 Text Classification with Model Deployment

3.1 Task Description

TBD

IV. Evaluation

1. **Model Performance: 80% of Overall Project Grade**

    1) **Task-1 Performance**: The evaluation will be based on the AUC (Area Under the Curve) score of the ROC (Receiver Operating Characteristic) curve derived from the submitted ranking scores for the records in "bank_marketing_test.csv." These scores will be compared to the true labels held by the instructor. This component constitutes **30% of the overall project grade**.

    2) **Task-2 Performance**: The evaluation criteria for Task-2 are yet to be determined (TBD) and will account for **50% of the overall project grade**.

2. **Peer Evaluation (& Job duty allocation): 20%**

V. Deliverables and Due Dates

1. **Delivered models and results <u>for each group</u>**

    1) **Task-1 (Due: December 2, 2025, 23:59):**
        a) Submit the Mean AUC score of the ROC (Receiver Operating Characteristic) curve for your selected (best) model, reported using 5-fold cross-validation on the training data.
        b) Provide the file "bank_marketing_test_scores.csv," which contains the predicted ranking scores for the 8,000 clients in "bank_marketing_test.csv" generated by your selected (best) model. This file must include 8,000 rows, with each row corresponding to the ranking score indicating the likelihood of being positive (y='yes') for each respective client in "bank_marketing_test.csv." Refer to the provided “bank_marketing_test_scores(example).csv” for formatting guidance.

    2) **Task-2: Details for deliverables are to be determined (TBD).**

2. **Allocation of job duties <u>for each group</u> (Due:TBD)**
   An Excel file summarizing the assigned job duties of each member for this project, structured as follows:

Member Name	Student ID	Roles and Responsibility
A	123	e.g., Task-1 (feature engineering), Task -2 (data preprocess and model deployment)
B	456	e.g., Task -1 (data collection and process), Task -2 (model development & deployment)

3. **Peer evaluation for <u>each student</u> (Due:TBD)**
   Online questionnaire for peer evaluation.

VI. Notes

1) Only electronic submissions are required.
2) If needed, you must submit the Python code within 24 hours, ensuring it is runnable and generates predicted results consistent with the deliverables submitted for each model. Failure to do so will incur a 50% penalty on the grade.
3) The requirements and deliverables for the two tasks can be summarized in the table below.

	Task-1	Task-2
Provided data	• Labeled structured data for model training. • Unlabeled data for model evaluation	TBD
Required data mining steps	From data understanding to model evaluation	
Constraints	The adopted model could be one of those **models introduced in our course**, including 1) Decision Tree, 2) Logistic Regression, 4) SVM, 5) Naïve Bayes, 6) KNN, 7) <u>Ensemble models of previous base models</u>.	
Evaluation metric	AUC score of ROC curve	
Deliverables	1) The results on the test dataset which should be named as "bank_marketing_test_scores.csv". 2) The 5-fold CV performance on the provided training data	
Use external resources?	Not allowed	
Weight	30%	