03_ml_02: ML2 (Bank Marketing Dataset)

The data is related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The classification goal is to predict if the client will subscribe a term deposit (variable y). ***http://archive.ics.uci.edu/ml/datasets/Bank+Marketing#***.

Problems

Please complete class **BankLogistic** using the provided code template. The details are as follows:

1. Load the Dataset: Load the bank-st.csv file from the "Attachment."
2. How many numeric variables and categorical variables are present in the dataset?
3. How is the distribution of the target variable?
4. Remove duplicate records from the data. What are the shape of the dataset afterward?
5. Replace unknown value with null
6. Remove features with more than 99% flat values.

   Hint: There is only one feature should be drop
7. Split Data
   - Split the dataset into training and testing sets with a 70:30 ratio.
   - random_state=0
   - stratify option
8. Impute missing
   - For numeric variables: Impute missing values using the mean.
   - For categorical variables: Impute missing values using the mode.

   Hint: Use statistics calculated from the training dataset to avoid data leakage.
9. Categorical Encoder:

   Map the ordinal data for the education variable using the following order:

```
education_order = {
    'illiterate': 1,
    'basic.4y': 2,
    'basic.6y': 3,
    'basic.9y': 4,
    'high.school': 5,
    'professional.course': 6,
    'university.degree': 7}
```

   Hint: Use One hot encoder or pd.dummy to encode nominal category

10. Model:

    Use Logistic Regression as the model with random_state=2025, class_weight='balanced' and max_iter=500. Train the model using all the remaining available variables. What is the macro F1 score of the model on the test data?

And return the output based on the question number:

- For Q1, following step 1, please return the rows of data are present in total?
- For Q2, following step 2, please return the tuple of numeric variables and categorical variables are presented in the dataset.
- For Q3, following step 3, please return the tuple of the Class 0 (no) followed by Class 1 (yes) in 3 digits.
- For Q4, following step 4, please return the shape of data after remove duplicate.
- For Q5, following step 5-7, please return the tuple of shapes of X_train and X_test.
- For Q6, following step 8-9, please return the shape of X_train.
- For Q7, following step 10, please return the macro F1 score of the model on the test data in 2 digits.

Submission: ** All files in attachment are different from the files in grader system. When submitting to the grader, submit ONLY the student.py file which includes the BankLogistic class with your modified functions **

Expected Results

Input	Output
Q1	41158
Q2	(10, 11)
Q3	(0.887, 0.113)
Q4	(41146, 21)
Q5	((28802, 19), (12344, 19))
Q6	(28802, 49)
Q7	0.74

Template codes

```
import ... #e.g. pandas, sklearn, .....
import warnings # DO NOT modify this line
from sklearn.exceptions import ConvergenceWarning # DO NOT modify this line
warnings.filterwarnings("ignore", category=ConvergenceWarning) # DO NOT modify this
line

class BankLogistic:
    def __init__(self, data_path): # DO NOT modify this line
        self.data_path = data_path
        self.df = pd.read_csv(data_path, sep=',')
        self.X_train = None
        self.y_train = None
```

```
        self.X_test = None
        self.y_test = None

    def Q1(self): # DO NOT modify this line
        """
        Problem 1:
            Load ‘bank-st.csv’ data from the “Attachment”
            How many rows of data are there in total?

        """
        # TODO: Paste your code here
        pass

    def Q2(self): # DO NOT modify this line
        """
        Problem 2:
            return the tuple of numeric variables and categorical variables are
presented in the dataset.
        """
        # TODO: Paste your code here
        pass

    def Q3(self): # DO NOT modify this line
        """
        Problem 3:
            return the tuple of the Class 0 (no) followed by Class 1 (yes) in 3 digits.
        """
        # TODO: Paste your code here
        pass



    def Q4(self): # DO NOT modify this line
        """
        Problem 4:
            Remove duplicate records from the data. What are the shape of the dataset
afterward?
        """
        # TODO: Paste your code here

        pass


    def Q5(self): # DO NOT modify this line
        """
        Problem 5:
            5. Replace unknown value with null
            6. Remove features with more than 99% flat values.
                Hint: There is only one feature should be drop
            7. Split Data
            -   Split the dataset into training and testing sets with a 70:30 ratio.
            -   random_state=0
            -   stratify option
            return the tuple of shapes of X_train and X_test.

        """
        # TODO: Paste your code here

        pass
```

```
    def Q6(self):
        """
        Problem 6:
            8. Impute missing
                -   For numeric variables: Impute missing values using the mean.
                -   For categorical variables: Impute missing values using the mode.
                Hint: Use statistics calculated from the training dataset to avoid data
leakage.
            9. Categorical Encoder:
                Map the ordinal data for the education variable using the following
order:
                education_order = {
                    'illiterate': 1,
                    'basic.4y': 2,
                    'basic.6y': 3,
                    'basic.9y': 4,
                    'high.school': 5,
                    'professional.course': 6,
                    'university.degree': 7}
                Hint: Use One hot encoder or pd.dummy to encode nominal category
            return the shape of X_train.

        """
        # TODO: Paste your code here

        pass

    def Q7(self):
        ''' Problem7: Use Logistic Regression as the model with
            random_state=2025,
            class_weight='balanced' and
            max_iter=500.
            Train the model using all the remaining available variables.
            What is the macro F1 score of the model on the test data? in 3 digits
        '''
        # TODO: Paste your code here

        pass
```