DMW\_Assignment 2

**Due date: April 20, 2023, midnight**

# **Instructions:**

The assignment consists of **two parts**: Association rule mining and decision tree learning.

Solve each part in one jupyter notebook. Name the files: **firstName\_lastName\_part1.ipynb** and **firstName\_lastName\_part2.ipynb**

Submit the assignment on Microsoft teams before Sunday February 7th midnight. **No submissions are accepted after this date.**

# **Part I: Association Rule Mining (40pts)**

A financial institution is interested in answering the following question: *How can we have a greater effectiveness for future marketing campaigns?* Recently, they launched a new product related to a personal investment plan and sent an email advertising the product to existing customers. The bank kept records on customers, including demographic information, the type of accounts and whether that customer responded and bought the product. Based on this dataset, the bank data scientists decide to use data mining techniques to build customer profile models. In this particular problem we are interested in deriving association rules from the data.

The data contained in the file **bank.csv** contains the following fields:

|  |  |
| --- | --- |
| **id** | a unique identification number |
| **age** | age of customer in years (numeric) |
| **sex** | MALE / FEMALE |
| **region** | inner\_city/rural/suburban/town |
| **income** | income of customer (numeric) |
| **married** | is the customer married (YES/NO) |
| **children** | number of children (numeric) |
| **car** | does the customer own a car (YES/NO) |
| **save\_acct** | does the customer have a saving account (YES/NO) |
| **current\_acct** | does the customer have a current account (YES/NO) |
| **mortgage** | does the customer have a mortgage (YES/NO) |
| **pip** | did the customer buy a personal investment plan after the last mailing (YES/NO |

**1.** Perform any necessary preprocessing step. Identify the attributes that need to be discretized and report any analysis that allows you to choose the right number of bins. (10pts)

**2.** Apply association rule mining on the preprocessed data and experiment with different parameters so that you get at least 20-30 strong rules (e.g., rules with high lift and confidence which at the same time have relatively good support). (20pts)

**3.** Select the top 5 most "interesting" rules and for each specify the following: (10pts)

* an explanation of the pattern and why you believe it is interesting based on the business objectives of the company;
* any recommendations based on the discovered rule that might help the company to better understand behavior of its customers or in its marketing campaign.

**Note**: The top 5 most interesting rules may not be in the top 5 result of Apriori algorithm. They are rules that, in addition to having high support, lift, and confidence, also provide some non-trivial, actionable knowledge based on the underlying business objectives.

# **Part II: Classification and Performance Evaluation (60pts)**

In this part, you will build a classification model using decision trees. You will use the preprocessed dataset from part I, question1 in order to build a classification model. The model will then be applied to a new set of prospects to whom the bank managers may want to extend an offer for a personal investment plan. Rather than doing a mass marketing campaign to all new prospects, the bank managers would like to target those that are likely to respond positively to their offer.

Your tasks in this problem are the following:

**1.** Load the data preprocessed in part I, question 1. Preprocess again by encoding categorical attributes into numeric. (10 pts)

**2.** Split the transformed data into **training** and **test** sets (using 80%-20% randomized split). (5pts)

**3.** Use scikit-learn to build a decision tree (**using default parameters**). Generate the confusion matrix (visualize it using Matplotlib), as well as the classification report. Compare the performance scores on the test and the training data sets. What does the comparison tell you? (10pts)

**4.** In the following, you will use 10-fold cross validation to determine the best depth of the decision tree (the depth of the tree can be specified as input parameter. See documentation at: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html> ). To this end, you will perform **cross validation on the training dataset** created in part (2) and leave the test data set apart for final evaluation.

**4.1.** Specify the number of depths you want to test. You may want to look at the tree created in question 3 in order to make a choice. (2pts)

**4.2.** iterate over the different depths and for each, perform 10-fold cross validation with a decision tree having the corresponding depth (2pts)

**4.3.** collect the mean accuracy score for each iteration (2pts)

**4.4.** Plot theobtained accuracy scores for different iterations and explain what should be the optimal depth of the tree. (2pts)

**4.5.** Build a decision tree with the depth selected above using the **training dataset**. (2pts)

**4.6.** Generate the confusion matrix (visualize it using Matplotlib), as well as the classification report **using the test set**. Compare the performance scores on the test and the training data sets. What does the comparison tell you? (5pts)

**5.** Visualize the tree generated in question 4.6. (5pts)

**6.** For each row in the test set, what is the probability of a customer to respond positively to a personal investment plan (pip = yes). ***Hint: use predict\_proba. Check documentation at:*** [***https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html***](https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html)(5pts)

**7.** Which customers have the highest probability of responding positively to a pip offer? (10pts)

**Bonus question:**

* Convert the decision tree into an equivalent set of rules.
* Write a code that checks whether these rules were discovered in part I.