**Homework-8**

|  |
| --- |
| Team#: \_\_\_  Team Member-1:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_\_  Team Member-2:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_\_  Team Member-3:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Member’s Contribution (in %) \_\_\_ |
| **Submission**   1. Answer the following questions. 2. **Problem-1:** Rename your R/Python file “HW8\_YourTeam#” (e.g., HW8\_Team1.rmd/HW8\_Team1.ipynb). 3. Upload the files to Canvas. |

**Problem-1: [100 points]**

The data of direct marketing campaigns of a Portuguese banking institution is given. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be (or not) subscribed. The classification goal is to predict if the client will subscribe a term deposit (variable **subscribed**). The training and test datasets are available in the homework folder.

Specifically, develop an R script to

1. Read-in the training set and test set. **[10 points]**
2. Construct multiple models by changing
   * Algorithms (kNN, Naïve Bayes, Decision Tree, bagging, boosting, random forest etc.) **[10 points per algorithm, max 60 points].** Thus, you will have to use minimum 6 algorithms to receive full credit.
   * Algorithm parameters **[5 points per algorithm tuning, max 30 points].** You will have to tune at least one parameter per algorithm for to receive full credit.
3. For each model,
   * Use 10-fold cross validation evaluation method. **[10 points]**
   * Predict the class labels for the test set.
   * Store the item IDs and class labels in a csv file. The file format is available in the homework folder. Submit all your result files to blackboard. The file names should reflect the models. For example, prediction\_knn\_k5.csv can store results from the kNN algorithm with k = 5**.** **If you program does not produce an output file for a particular model, you will receive zero point for that model (meaning lose 15 points per model).**

Submit the R script(s) to blackboard. The script should be self-contained, meaning it should read the data files from the folder where the script is placed and produce the final results of multiple models.

**Data Fields**

**Input variables**

**# bank client data**

1. **age:** (numeric)
2. **job:** type of job (categorical: "admin.","unknown","unemployed","management","housemaid","entrepreneur","student", "blue-collar","self-employed","retired","technician","services")
3. **marital:** marital status (categorical: "married","divorced","single"; note: "divorced" means divorced or widowed)
4. **education**:(categorical: "unknown","secondary","primary","tertiary")
5. **default:** has credit in default? (binary: "yes","no")
6. **balance:** average yearly balance, in euros (numeric)
7. **housing:** has housing loan? (binary: "yes","no")
8. **loan:** has personal loan? (binary: "yes","no")

**# related with the last contact of the current campaign**

1. **contact:** contact communication type (categorical: "unknown","telephone","cellular")
2. **day:** last contact day of the month (numeric)
3. **month:** last contact month of year (categorical: "jan", "feb", "mar", ..., "nov", "dec")
4. **duration:** last contact duration, in seconds (numeric)

**# other attributes**

1. **campaign:** number of contacts performed during this campaign and for this client (numeric, includes last contact)
2. **pdays:** number of days that passed by after the client was last contacted from a previous campaign (numeric, -1 means client was not previously contacted)
3. **previous:** number of contacts performed before this campaign and for this client (numeric)
4. **poutcome:** outcome of the previous marketing campaign (categorical: "unknown","other","failure","success")

**Output variable (desired target)**

1. **subscribed:** has the client subscribed a term deposit? (binary: "yes","no")