**DEADLINE: Friday, 13th Dec @5pm**

**TASK DESCRIPTION**

In this assignment you will develop a classifier that uses data to predict the outcome of a bank phone marketing campaign. The classifier model has to be one of those studied in this course (you can develop it yourself or use scikit-learn python library)

**TEAM WORK**

For this assignment you can work on your own, or in pairs (no three person teams or larger). Every team member on the team gets the same mark irrespective of the individual contribution (also, I won’t be involved in team disagreements). It’s sufficient for one of the team members to submit the assignment through Brightspace but please clearly indicate in the documentation the names of the team members.

The marking scheme is the same for both individual and group submissions.

**DATA AVAILABLE**

There are three files (available on Brightspace):

* **datadescription.txt:** contains a description of the data types of the different columns in the data
* **trainingset.csv:** contains the training instances. This file lists a training id, the descriptive features and the target feature level for each instance.
* **queries.csv:** contains the query instances. This file lists a test id and the descriptive features for each instance, and the target feature represented by ‘?’

**SUBMISSION DETAILS**

* Deadline: Friday, 13th Dec @5pm. Marks will be deducted for late submissions.
* What you need to submit:

Submission is through Brightspace. You need to submit 3 separate files:

1. a predictions file,

2. the Python code for you classifier (either .py file or Jupyter notebook),

3. documentation – text document describing how you solved the problem and any decisions you had to make – for example, why did you chose the specific classifier, any issues with the data, how testing was performed, etc.

**Format of the predictions file:** this file should list your classifier’s target variable predictions for each of the query instances in the queries.csv file. Each line in the file should list one query id followed by a comma followed by your classifier’s prediction for that query, i.e.:

<tstid>,<prediction>

The box below illustrates the format of your solution file. Please stick to the exact format!

TEST1,no

TEST2,no

TEST3,yes

…

…

**MARKING SCHEME**

Marks are awarded based on both your documentation and on the accuracy of the classifier.

|  |  |  |
| --- | --- | --- |
| **Accuracy of the classifier** | 50% | Metric used: average class accuracy (harmonic mean)  The marks will be scaled so the highest achieved accuracy is considered 100%. |
| **Documentation** | 50% | - (10) description of the data, any issues, etc  - (20) description and justification of a classifier  - (20) testing |

*Marks may be deducted for the following reasons*:

1. Late submission (including submissions that are incomplete by the time the deadline has passed): 10 marks per day late.

2. Incorrect submission: 10 marks will be deducted if your submission does not follow the stated formats. The reason for this is because the outputs will be automatically processed and evaluated, so if you do not correctly format you submission I will have to manually modify and tweak it to make it conform to the required format, and this slows down the correction process for everyone.

***Examples*** of the types of errors that will result in marks being deducted include:

* Leaving blank lines between solutions in the solutions file
* Using incorrect labels or using the wrong case for your labels e.g., using lowercase Ts
* Having trailing or additional blank spaces
* Forgetting to put commas between the fields in the solutions file
* The prediction file not being a .csv file, for example submitting your prediction as an .rtf or other file format