**DEADLINE: Friday, 9th Dec @6pm**

**TASK DESCRIPTION**

In this assignment you will develop a classifier that uses data to predict the outcome of a bank marketing campaign. The classifier model has to be one of those studied in this course.

**DATA AVAILABLE**

You are provided three files (on Brightspace):

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

**SUBMISSION DETAILS**

**Deadline**: Friday, 9th Dec @6pm. 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,

3. documentation – text document describing how you solved the problem and any decisions you had to make – including why did you chose the specific classifier, any issues with the data and how it was handled, 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.txt 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>

TEST1,TypeA

TEST2,TypeA

TEST3,TypeB

…

…

The box on the right illustrates the required format of

your solution file.

**MARKING SCHEME**

**This assignment is 20% of your overall mark for this module**. (The overall mark for the module is 70% exam and 30% CA, with the CA mark split as 10% for CA1 and 20% for this CA2)

Marks are awarded based on both your documentation (60% of the mark for the CA) and on the accuracy of the classifier (40% of the mark for the CA).

The accuracy metric used will be the average class accuracy (harmonic mean) of the classifier.

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 these marks being deducted include:

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