# PROJECT (DUE DATE: 12-06-16)

PART 1 (10 points)

Perform different experiments in KEEL in order to test the behavior of some algorithms when applied to the following datasets:

- Bupa
- Cleveland
- Glass
- Haberman
- Iris
- Monk-2
- New-Thyroid
- Pima
- Vehicle
- Wine
- Wisconsin

Use these datasets with the algorithms listed below:

- Decision Trees:
  - AdaBoost.NC-C
  - C45-C
- Crisp Rule Learning
  - Ripper-C
- Evolutionary Crisp Rule Learning
  - SIA-C
- Evolutionary Fuzzy Rule Learning
  - GFS-GCCL-C
- Fuzzy Rule Learning
  - Chi-RW-C
- Neural Networks
  - iRProp+-C

Finally, perform an experimental study in order to compare the algorithms among them. Use the **Friedman, Holm and Shaffer** statistical tests as convenient so that we can compare the best algorithm against the rest and all against all. You can read the details of each one of the statistical tests in the file “statistics.pdf”.

The results of each algorithm on each dataset and for a given parameter studied (accuracy on training/test) must be shown through a table, i.e.:

| Data | Training | | | |
|---|---|---|---|---|
| | NSLV | NSLV-AR | NSLV-FR | SLAVE3 |
| appendicitis | **93.3 (1)** | 91.9 (4) | 92 (3) | 92.5 (2) |
| australian | 90.3 (2) | 87.4 (4) | **90.7 (1)** | 89.2 (3) |
| automobile | 97.6 (3) | 97.3 (4) | **99.1 (1)** | 98.1 (2) |
| balance | 85.8 (3) | 81.1 (4) | **98.4 (1)** | 96.6 (2) |

In the previous table, each row shows the accuracy on training for a specific dataset provided by the correspondent algorithm.

Also, the results of the statistical analysis must be shown (the tables which are relevant to justify the results – see below), together with the interpretation.

| Algorithm | Ranking |
|---|---|
| GCCL | 4.8 |
| C45 | 3.05 |
| SGERD | 5.2125 |
| FARC-HD | 2.675 |
| FURIA | 2.25 |
| SLAVE3 | 3.0125 |

| Friedman p-value |
|---|
| 4.43937109295689E-11 |

Table 6.18: Adjusted $p$-values (accuracy on testing set).

| i | hypothesis | unadjusted $p$ | $p_{Shaf}$ |
|---|---|---|---|
| 1 | SGERD vs .FURIA | 0 | 0 |
| 2 | GCCL vs .FURIA | 0 | 0 |
| 3 | SGERD vs .FARC-HD | 0 | 0 |
| 4 | SGERD vs .SLAVE3 | 0 | 0.000001 |
| 5 | C45 vs .SGERD | 0 | 0.000002 |
| 6 | GCCL vs .FARC-HD | 0 | 0.000004 |
| 7 | GCCL vs .SLAVE3 | 0.000019 | 0.000135 |
| 8 | GCCL vs .C45 | 0.000029 | 0.000201 |
| 9 | C45 vs .FURIA | 0.055829 | 0.390805 |
| 10 | FURIA vs .SLAVE3 | 0.068345 | 0.410072 |
| 11 | FARC-HD vs .FURIA | 0.309656 | 1.238624 |
| 12 | GCCL vs .SGERD | 0.324102 | 1.296408 |
| 13 | C45 vs .FARC-HD | 0.370028 | 1.296408 |
| 14 | FARC-HD vs .SLAVE3 | 0.419794 | 1.296408 |
| 15 | C45 vs .SLAVE3 | 0.928572 | 1.296408 |

For PART 1, you have to submit a document with all the information required.

PART 2 (15 points)

Implement a classification algorithm in JAVA that is able to provide at least a 60% of accuracy on test when applied to the given dataset (the dataset must be adapted to the requirements of your algorithm).
The name of the attributes are (from left to right in the "koronia_dataset.csv" file): blue, green, red, nearIR, conB, asmB, corB, homB, conG, asmG, corG, homG, conR, asmR, corR, homR, conIR, asmIR, corIR, homIR, brightness, greeness, wetness, intensity, hue and class.
The algorithm must take as inputs two files (training and testing files), and must also return as an output two different files, one for training results and another one for test results. Each one of these output files will show one line for each of the examples in the set with the correct classification and the estimated classification (given by your algorithm). The file "result0s0.tra" shows an example of the output for an algorithm when dealing with the first partition of "Iris" dataset.
It also must prompt in the command line the average accuracy on training and test.

For PART 2, you have to submit the JAVA source files together with the final dataset and the commands to correctly run the program.