Homework

1. The whole task is to say if a student pass or not the subject. Check the website `https://archive.ics.uci.edu/ml/datasets/student+performance` and analyze the list of attributes. Which of them would you consider as the most significant according to you? Choose 5-10 attributes, which will take part in the part of experiments.
2. Open *student-mat_train.arff* and *student-mat_test.arff* downloaded from eKursy, which were preprocessed to represent a classification task.
3. Determine which metrics will be proper for the given datasets. Report three the most accurate metrics.
4. Get the chosen 5-10 attributes and test a few different values of the parameters, at least: confidenceFactor, minNumObj and binarySplits. Show the results for each set of parameters (you can visualize it also). For which set do you have the best result for test set?
5. Repeat the previous step but this time get the whole set of attributes.
6. Load file *student-por.arff* and run analysis for $k = 10$ in cross-validation. Present your results.
7. Both datasets (math and Portuguese) have the same set of attributes. Compare the structure of trees with the best results for each dataset. Are there any similarities between them? Basing on these results, can we say which attributes can say if the student is attentive?
8. Choose any other algorithm that you already know e.g. algorithm for rule induction that we used on first laboratories (PRISM) or Naive Bayes and run it on *student-mat* dataset. Compare the results from both algorithms. Which attributes had the biggest influence on the result? Are these attributes similar to those that you chose intuitively at the beginning of the task?

The whole task should be done in Weka.

The final report should be sent in pdf format (any other format will not be checked).

You need to remember to put in your report:

- for each set of parameters, confusion matrix and values on the metrics that you chose as the most accurate,
- charts that visualize how change of each parameter influences the results,
- show the decision tree which gives the best results for each situation (chosen attributes on math dataset, all attributes on math dataset, all attributes on Portuguese dataset).