Title: Complete Decision Tree Induction Functionality in Scikit Learn
Supervisor: Prof. Dr. Jesse Davis, Prof. Dr. ir. Hendrik Blockeel
Daily advisor: Elia Van Wolputte

 # SUMMARY
Complete Decision Tree Induction Functionality in Scikit Learn

 # BACKGROUND
Scikit Learn (scikit-learn.org/) is well-known and increasingly popular machine learning library implemented in python. Surprisingly, its implementations for some algorithm, particularly for more classical techniques such as decision trees, are very incomplete with respect to the functionality provided in other packages such as Weka (http://www.cs.waikato.ac.nz/ml/weka/). The missing functionality can lead to large decreases in performance. For example, we have observed that Weka can be > 25% improvements in classification accuracy on some datasets compared to Scikit Learn!

 # GOAL
Develop a (more) complete python version of decision tree learners (and possible other algorithms) that works with Scikit Learn.

 # APPROACH:
1. Investigate and characterize the differences in functionality between Weka and Scikit Learn for decision trees.
2. Implement a python version of a decision tree learner that has the required basic functionality and that uses the scikit learn interfaces.
3. Benchmark the developed implementation to those availabe in Scikit Learn and Weka on a suite of datasets.
4. Translate the python implementation to C/cython in order to improve the run time performance.
5. Extend the decision tree implementation to more advance settings (e.g., online learning, …).
6. Provide good documentation on how the code works.

 # STUDENT PROFILE
Strong programming skills and deep interest in machine learning techniques is a must. This thesis is only intended for students enrolled in the big data option or who are following the course 'Big Data Analytics Programming' (H00Y4A).