Artificial Intelligence Methods
Assignment 6
Kacper Multan

Exercise 1

1.   **Implement the decision tree algorithm and learn a decision tree from the data in train.csv, using both versions of Importance.**

The implementation of this task is in file assignment_6.py which I included in my submission.

2.   **Classify all examples in the test set given in the data file test.csv. Calculate the accuracy of the learner by comparing them to the correct classification of the examples in the test set.**

3.   **Using the accuracies of the two models, write a short discussion comparing the performance of the two models. Make sure to run both versions several times to observe overall performance.**

The first model was based on the random selection of importance for each attribute. This approach created a chance to explore the data set and enabled us to find a new way to split up the data. However the chance based method happened to not choose the simplest and most effective division technique. The accuracy level varied from 75% to 100%.

By selecting attributes based on their information gain, the second model prioritized features that offer the greatest reduction in entropy, resulting in more effective splits. As a result, the decision tree constructed using information gain exhibited higher accuracy and stability across multiple runs compared to the random selection approach.

Which Importance function is better, and why?

The random selection of the splitting feature turned out to be less effective than modeling the data based on the information gain of the given attributes. After running multiple tests the average test accuracy for the information gain measure turned out to be around 92,8% and appeared consistently through many runs. Although the random method enabled us to achieve a 100% accuracy it was less reliable and quite often produced accuracy lower than 92%. Additionally, it's important to notice that if we would consider a bigger data set the accuracy of the information gain method would be higher than the accuracy of the random method.