Assignment 3

Adrian Siim Melsom

What can you conclude about the results you have obtained? Which Importance is better, and why?

For consistency, the one where importance is not random shows the best results overall. The average classification success over 10000 trees of random importance values tended to stay around 0.77 – 0.79 success rate. Sometimes however, it was able to classify 100% correctly.

With importance gain, the success rate was 0.9286. So even though the random importance were sometimes able to classify everything correctly, it's worse overall.

Do you get the same result if you run the random Importance several times?

No, it varies a lot. Here are some test results:

```
100000 iterations classification success rate: 0.7753699999999986
100000 iterations success interval: [0.32142857142857145, 1.0]
```

Iterations classification success rate shows the average rate of successful classifications over the iterations.

Interval shows the lowest and the highest classification success rate.

Using this method, I found an efficiently built tree with 1.0 success rate:

```
Best performing tree:
Choice: 0
        Leaf: 1
        Choice: 1
                Choice: 2
                        Leaf: 1
                        Leaf: 2
                Choice: 2
                        Leaf: 2
                        Leaf: 1
```

The worst tree I found covers multiple pages. But similar ones can be found by running the function functions.test_random. I ran it for 100,000 iterations which took roughly a minute.

What happens when you run the learner based on Information Gain several times?

The data we work with is consistent, so the information gain didn't change regardless of how many times it was run.