Homework

Your task is to write step-by-step program for generating decision trees without using any ready packages like scikit. This task will help you to fully understand all topics from today's lesson which will be great base for next lesson. The steps that need to be included are:

- entropy calculation,
- conditional entropy calculation,
- information gain calculation,
- gain ratio calculation,
- simple drawing of a decision tree (there can be really primitive way of drawing or usage of any ready package to do only this drawing) with the information how many examples of each class are in each leaf.

Program should be written in Python, preferably with the usage of Jupyter Notebook. Task is supposed to be done in pairs. The dataset that is provided in this task is well-known Titanic [1]. The file that is available contains only part of the data. Column *name* is just to show that they were real people, you should omit this column while inducing a decision tree.

There are two possible ways in which you can do your homework:

- **basic (4 points)** map column: *age* to labels in the following way:
  - ”young”: [0,20]
  - ”middle”: (20,40]
  - ”old”: (40,100]
- **pro (5 points)**: *age* is a continuous value

There are two things that are required in this task:

1. Python program described above,
2. short report which includes your names and index numbers, group number, step-by-step results (similarly to the way it was presented during the lesson) with short comment

References

[1] https://www.kaggle.com/c/titanic/data