Machine Learning for Image Analysis, SoSe 2023
Exercise Sheet #3: Decision Trees and Random Forests
Due date: May 9, 2017, before 11 am

Problem 1 (Impurities) (15 pt).
Consider a two class classification problem ($C = 2$). At the current node there are $N = 400$ data points of each class (denoted by $(400, 400)$). Evaluate two possible splits:

- Split A: Create two nodes with $(300, 100)$ and $(100, 300)$ data points respectively.
- Split B: Create two nodes with $(200, 0)$ and $(200, 400)$ data points respectively.

Calculate for each split the Gini impurity as well as the entropy. Which split would each criterion prefer? Remember:

$$\text{Gini impurity: } H = 1 - \sum_{c=1}^{C} p(y=c)^2, \text{ Entropy: } H = -\sum_{c=1}^{C} p(y=c) \log p(y=c)$$

Problem 2 (Decision trees) (10 pt).

Given a dataset with eight day samples, four binary weather features and a binary label, you want to build a decision tree to predict when you should best play tennis.

Sample day	Weather	Humidity	Temperature	Wind	Play Tennis?
1	Sunny	High	Hot	Weak	Yes
2	Sunny	High	Hot	Weak	Yes
3	Rainy	High	Hot	Weak	Yes
4	Rainy	High	Cold	Strong	Yes
5	Sunny	High	Cold	Weak	No
6	Sunny	High	Cold	Weak	No
7	Rainy	High	Cold	Weak	No
8	Rainy	High	Hot	Strong	No

(a) (8pt) Which feature should be selected as the first split criterion using entropy as a purity measure?

(b) (4pt) Is there a decision tree that achieves 100% accuracy on this dataset? If yes, build the tree. If not, justify with reasons.