Team members: Brigitte Aznar, Daniel Kostic, Stefan Vujovic

1. Clustering vs Classification

1. What is the main difference between clustering and classification?

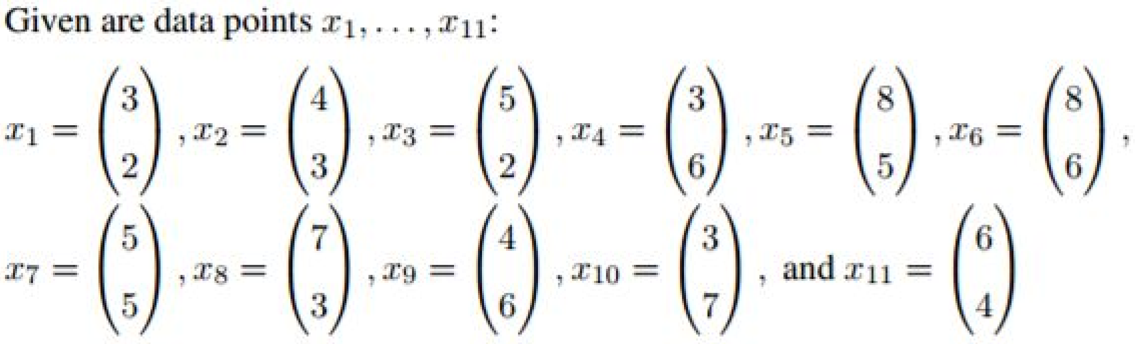
The main difference is that clustering is a unsupervised learning method, meaning that the algorithm has no previous information about the data it's trying to categorize. There is no labeled data, so the algorithm has to make groups based only on attributes of the data.

Classification on the other hand is a supervised learning method, which means the algorithm uses the attributes, but also the labels for each instance of the dataset to "learn" how it is supposed to classify objects. Afterwards, the algorithm classifies (adds labels) new instances based on their attributes.

2. What is easier to evaluate, clustering or classification?

Classification, because we just need to compare labels that the algorithm assigned to the labels of the testing set. Assuming that the labels we assigned to the testing set are correct, we can just count the number of matching labels (our labels and the algorithms labels) and divide the number of matched labels with the complete number of items in the testing set and get the accuracy. We could use other straightforward metrics, such as recall, precision, f1-score, confusion matrix, etc. in order to get more details. In case of clustering, forming of groups was completely done by the algorithm, therefore the evaluation methods require more complex mathematical functions.

2. k-Nearest Neighbor



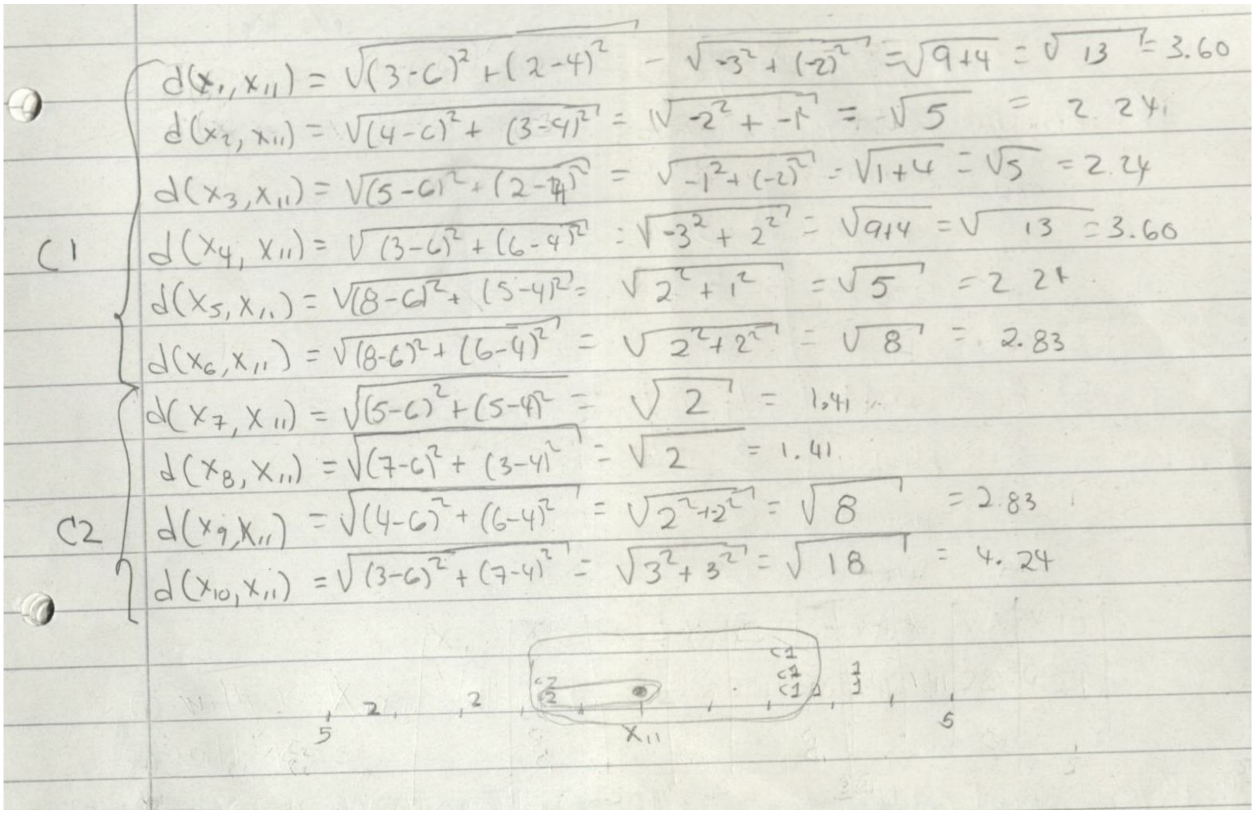
Given are data points $x_1, \ldots, x_{11}$:

$$x_1 = \begin{pmatrix} 3 \\ 2 \end{pmatrix}, x_2 = \begin{pmatrix} 4 \\ 3 \end{pmatrix}, x_3 = \begin{pmatrix} 5 \\ 2 \end{pmatrix}, x_4 = \begin{pmatrix} 3 \\ 6 \end{pmatrix}, x_5 = \begin{pmatrix} 8 \\ 5 \end{pmatrix}, x_6 = \begin{pmatrix} 8 \\ 6 \end{pmatrix},$$

$$x_7 = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, x_8 = \begin{pmatrix} 7 \\ 3 \end{pmatrix}, x_9 = \begin{pmatrix} 4 \\ 6 \end{pmatrix}, x_{10} = \begin{pmatrix} 3 \\ 7 \end{pmatrix}, \text{ and } x_{11} = \begin{pmatrix} 6 \\ 4 \end{pmatrix}$$

The data points x1, . . . , x6 belong to class c1, the data points x7, . . . , x10 belong to class c2. x11 has not yet been classified.

Team members: Brigitte Aznar, Daniel Kostic, Stefan Vujovic



C1

$d(x_1, x_{11}) = \sqrt{(3-6)^2 + (2-4)^2} = \sqrt{-3^2 + (-2)^2} = \sqrt{9+4} = \sqrt{13} = 3.60$

$d(x_2, x_{11}) = \sqrt{(4-6)^2 + (3-4)^2} = \sqrt{-2^2 + -1^2} = \sqrt{5} = 2.24$

$d(x_3, x_{11}) = \sqrt{(5-6)^2 + (2-4)^2} = \sqrt{-1^2 + (-2)^2} = \sqrt{1+4} = \sqrt{5} = 2.24$

$d(x_4, x_{11}) = \sqrt{(3-6)^2 + (6-4)^2} = \sqrt{-3^2 + 2^2} = \sqrt{9+4} = \sqrt{13} = 3.60$

$d(x_5, x_{11}) = \sqrt{(8-6)^2 + (5-4)^2} = \sqrt{2^2 + 1^2} = \sqrt{5} = 2.24$

$d(x_6, x_{11}) = \sqrt{(8-6)^2 + (6-4)^2} = \sqrt{2^2 + 2^2} = \sqrt{8} = 2.83$

C2

$d(x_7, x_{11}) = \sqrt{(5-6)^2 + (5-4)^2} = \sqrt{2} = 1.41$

$d(x_8, x_{11}) = \sqrt{(7-6)^2 + (3-4)^2} = \sqrt{2} = 1.41$

$d(x_9, x_{11}) = \sqrt{(4-6)^2 + (6-4)^2} = \sqrt{2^2 + 2^2} = \sqrt{8} = 2.83$

$d(x_{10}, x_{11}) = \sqrt{(3-6)^2 + (7-4)^2} = \sqrt{3^2 + 3^2} = \sqrt{18} = 4.24$

Calculating Euclidean distance of x1….x10 to x11

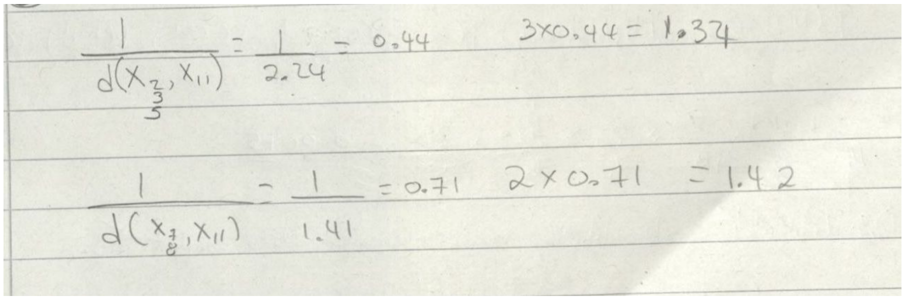
1. Classify the data point x11 with the k-nearest-neighbor approach and k = 1. Use the Euclidean distance as the distance metric.

> For K=1 x11 is classified in C2, because it nearest neighbor is X7 (or x8) who belongs to C2

2. Classify the data point x11 with the k-nearest-neighbor approach and k = 5. Use the Euclidean distance as the distance metric.

> For K=5 X11 is classified in C1, because our of the 5 nearest neighbors (X2, X3, X5, X7 and X8) 3 of them (X2, X3,X5) belongs to C1

3. Classify the data point x11 with the k-nearest-neighbor approach and k = 5. Use the Euclidean distance as distance metric. As weights, use the inverted distance to the point to be classified, i.e., $1/d(xn,xc)$, with xn is the neighbor, and xc is the point to be classified.



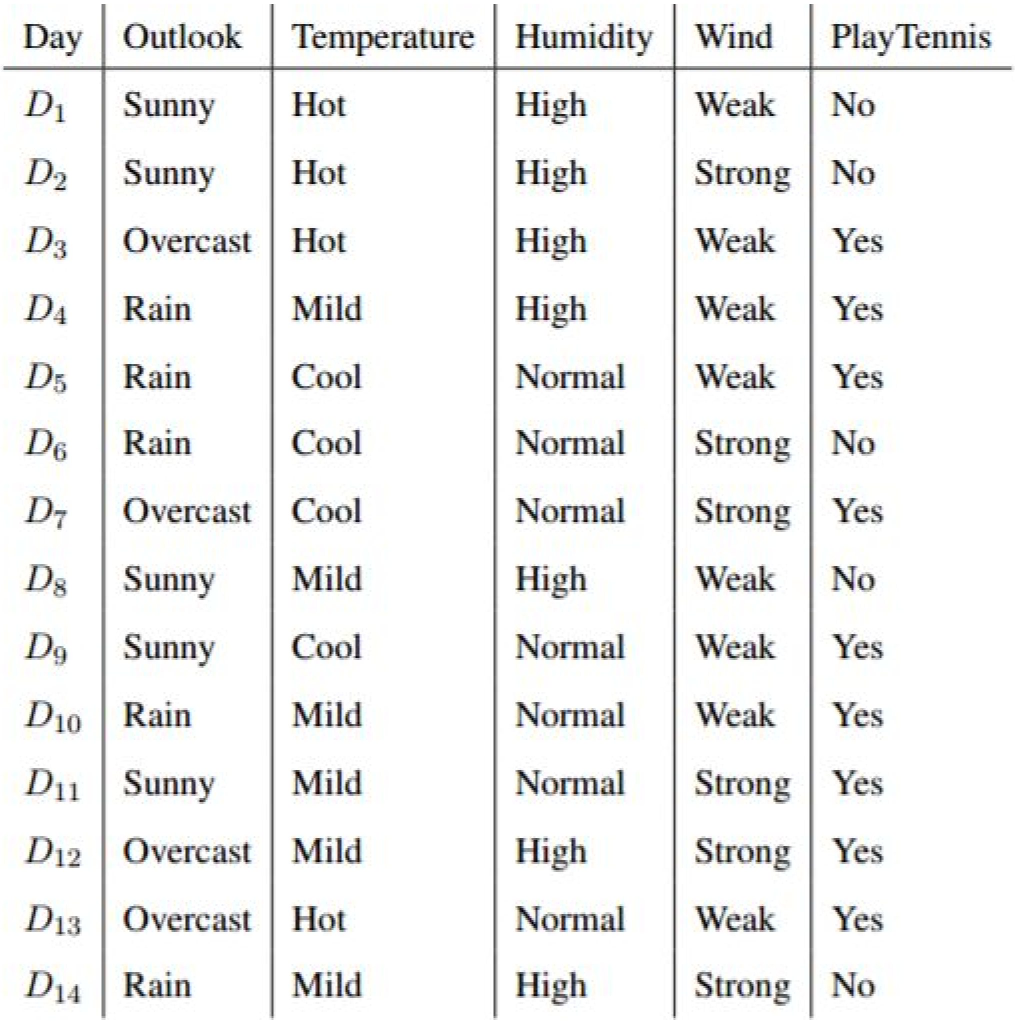
$$\frac{1}{d(x_{\substack{2\\3\\5}}, x_{11})} = \frac{1}{2.24} = 0.44 \qquad 3 \times 0.44 = 1.34$$

$$\frac{1}{d(x_{\substack{7\\8}}, x_{11})} = \frac{1}{1.41} = 0.71 \qquad 2 \times 0.71 = 1.42$$

If we consider weights then X11 will be again classified under C2 because the inverse distance (multiplied by the number of nodes who share this distance) is greater (or has more weight) in C2

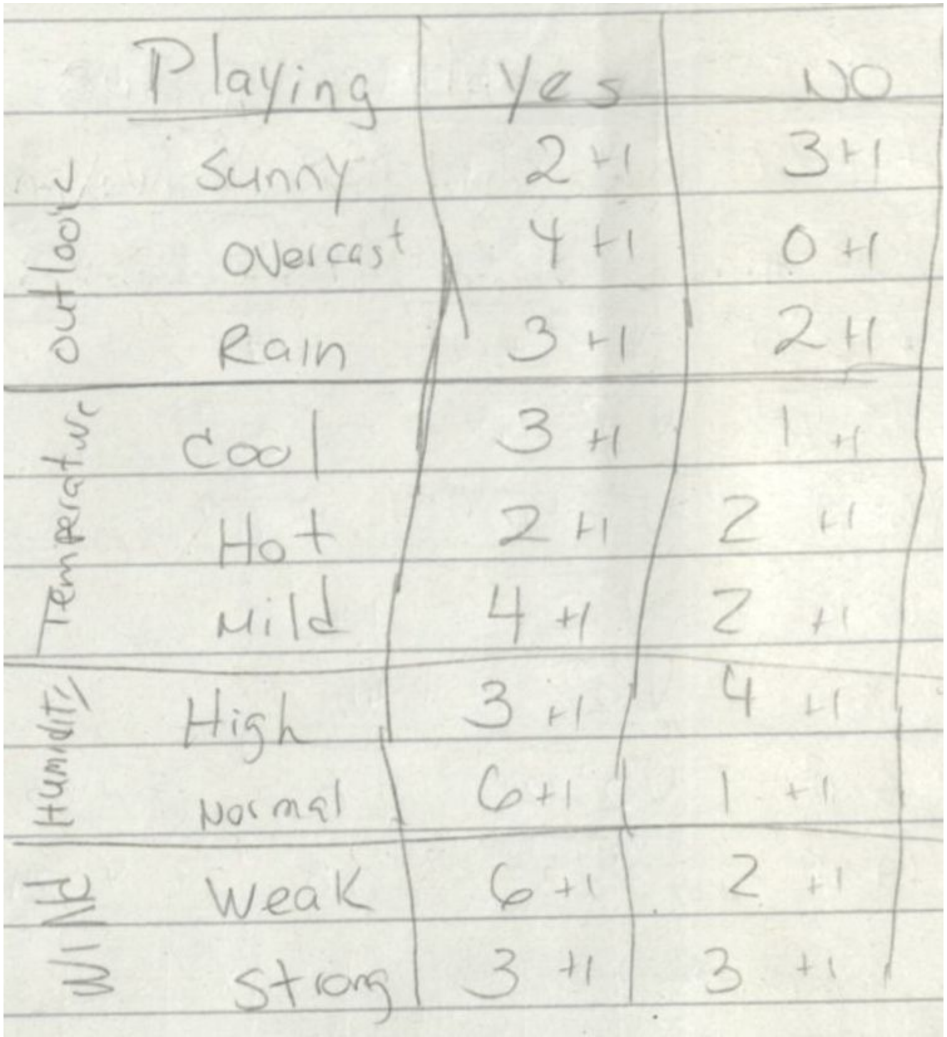
3. Naive Bayes

Given is the following dataset:



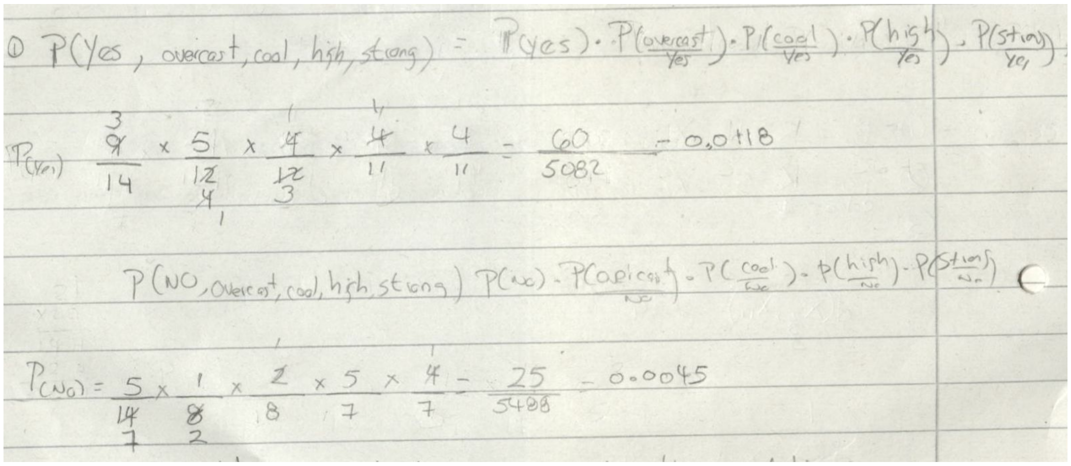
Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rain	Mild	High	Weak	Yes
D_5	Rain	Cool	Normal	Weak	Yes
D_6	Rain	Cool	Normal	Strong	No
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rain	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rain	Mild	High	Strong	No

Team members: Brigitte Aznar, Daniel Kostic, Stefan Vujovic



	Playing	Yes	NO
outlook	Sunny	2+1	3+1
	Overcast	4+1	0+1
	Rain	3+1	2+1
Temperature	Cool	3+1	1+1
	Hot	2+1	2+1
	Mild	4+1	2+1
Humidity	High	3+1	4+1
	Normal	6+1	1+1
Wind	Weak	6+1	2+1
	Strong	3+1	3+1

1. Train a naive-Bayes classificator with this data and use it to classify the new data point D15 = (O=Overcast, T=Cool, H=High, W=Strong)!



① $P(Yes, \text{overcast}, \text{cool}, \text{high}, \text{strong}) = P(yes) \cdot P\left(\frac{\text{overcast}}{Yes}\right) \cdot P\left(\frac{\text{cool}}{Yes}\right) \cdot P\left(\frac{\text{high}}{Yes}\right) \cdot P\left(\frac{\text{strong}}{Yes}\right)$

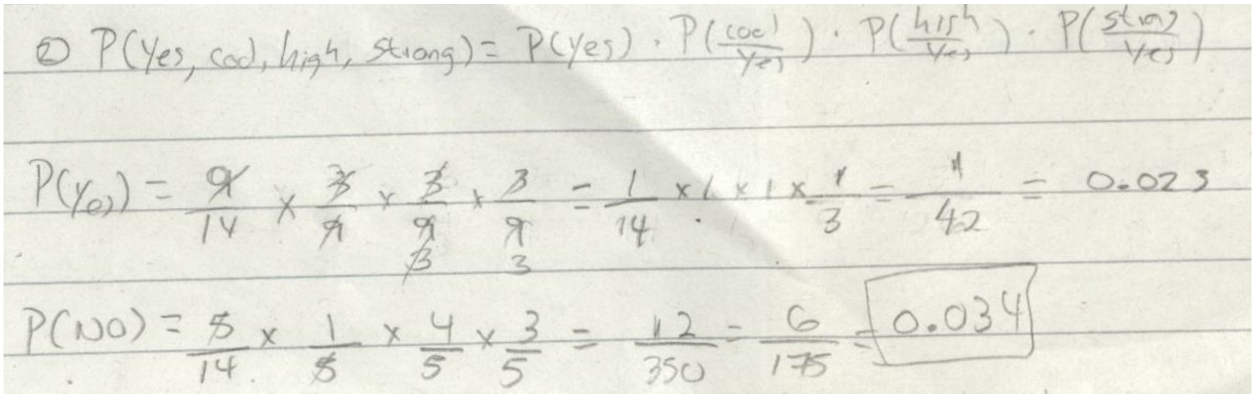
$P_{(Yes)}\ \frac{\overset{3}{\cancel{9}}}{14} \times \frac{5}{\underset{\underset{1}{\cancel{4}}}{\cancel{12}}} \times \frac{\overset{1}{\cancel{4}}}{\underset{3}{\cancel{12}}} \times \frac{\overset{1}{\cancel{4}}}{11} \times \frac{4}{11} = \frac{60}{5082} = 0{,}0118$

$P(NO, \text{overcast}, \text{cool}, \text{high}, \text{strong})\ P(No) \cdot P\left(\frac{\text{overcast}}{No}\right) \cdot P\left(\frac{\text{cool}}{No}\right) \cdot P\left(\frac{\text{high}}{No}\right) \cdot P\left(\frac{\text{strong}}{No}\right)$

$P_{(No)} = \frac{5}{\underset{7}{\cancel{14}}} \times \frac{1}{\underset{2}{\cancel{8}}} \times \frac{\overset{1}{\cancel{2}}}{8} \times \frac{5}{7} \times \frac{\overset{1}{\cancel{4}}}{7} = \frac{25}{5488} = 0.0045$

Under the given situations the classificator encourages playing tennis. Because the probability for yes is greater than the probability of no.

2. Use your classifier to classify the data point D16 = (T=Cool, H=High, W=Strong)! The value for the attribute Outlook was not recorded for this data point.



② $P(\text{Yes, cool, high, strong}) = P(\text{yes}) \cdot P\left(\frac{\text{cool}}{\text{yes}}\right) \cdot P\left(\frac{\text{high}}{\text{yes}}\right) \cdot P\left(\frac{\text{strong}}{\text{yes}}\right)$

$P(\text{Yes}) = \frac{\cancel{9}}{14} \times \frac{\cancel{3}}{\cancel{9}} \times \frac{\cancel{3}}{\cancel{9}_{\cancel{3}}} \times \frac{3}{\cancel{9}_{3}} = \frac{1}{14} \times 1 \times 1 \times \frac{1}{3} = \frac{1}{42} = 0.023$

$P(\text{NO}) = \frac{\cancel{5}}{14} \times \frac{1}{\cancel{5}} \times \frac{4}{5} \times \frac{3}{5} = \frac{12}{350} = \frac{6}{175} = \boxed{0.034}$

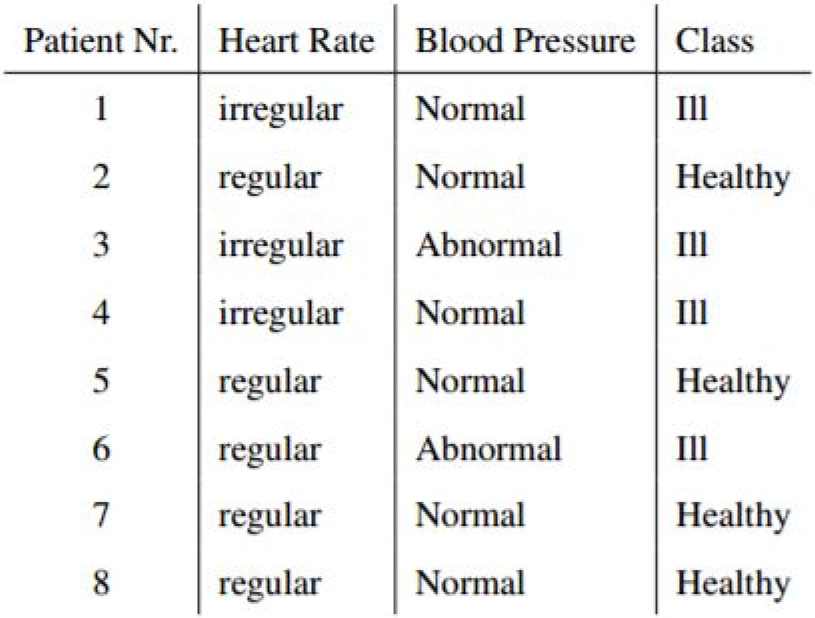
Under the given situations the classificator encourages NOT to play tennis. Because the probability for no is greater than the probability of yes.

3. Compare (short explanation!) the results for subtasks 1 and 2!

D15 and D16 show opposite results because the parameter Outlook with value *Overcast* is a very important factor in the training data. Therefore omitting it or not makes a big difference in the probabilities of whether the person should play tennis or not with the given circumstances.

4. Decision Trees

A hospital wants to provide support for his physicians. For that purpose they collected data on healthy and ill patients. Decision makers read in a magazine that decision trees allow to create a model that can be used to simulate decision of a doctor.

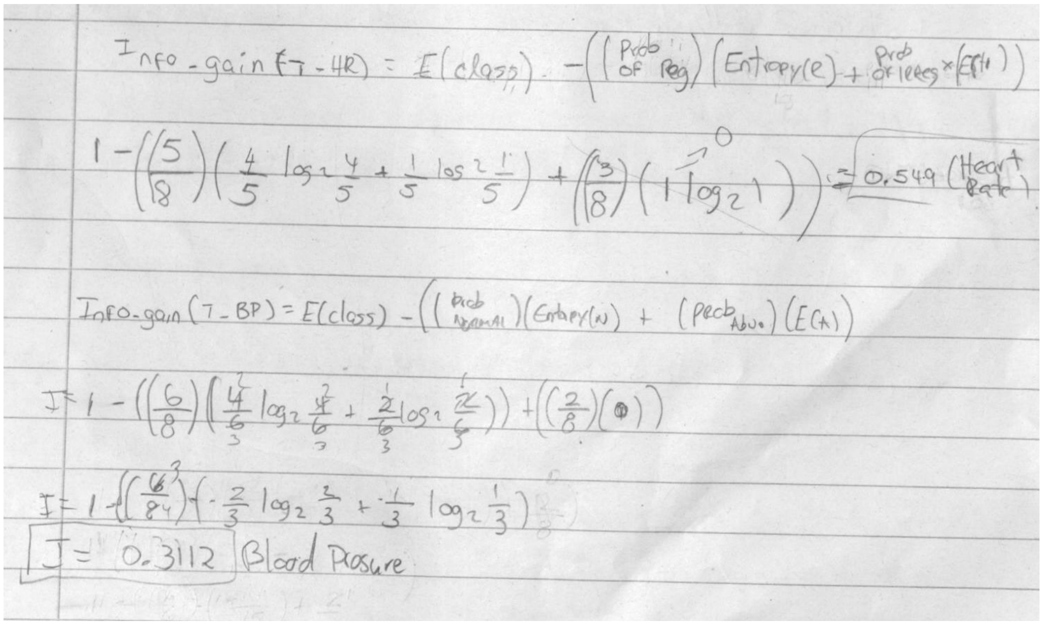


Patient Nr.	Heart Rate	Blood Pressure	Class
1	irregular	Normal	Ill
2	regular	Normal	Healthy
3	irregular	Abnormal	Ill
4	irregular	Normal	Ill
5	regular	Normal	Healthy
6	regular	Abnormal	Ill
7	regular	Normal	Healthy
8	regular	Normal	Healthy

Team members: Brigitte Aznar, Daniel Kostic, Stefan Vujovic

1. Compute a decision tree for this dataset (using information gain as a split criterion, without using the patient id) and sketch it graphically. If you want to do compute this with pen and paper only (no computer or calculator), you can use the following approximations for the logarithms:

The Entropy for the class is 1 because we have 4 Ill patients and 4 Healthy patients. When there is the same number of instances in a class entropy = 1



Info-gain (T - HR) = E(class) − ((Prob of Reg)(Entropy(e) + Prob of Irres × E(I)))

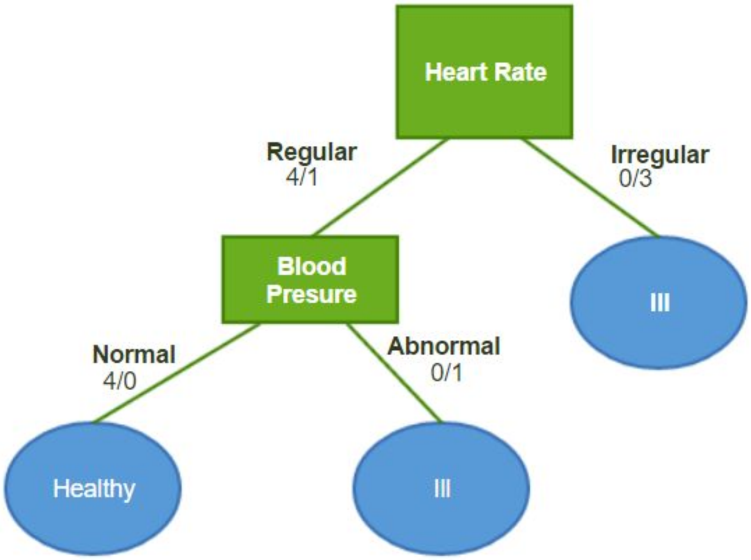
$$1-\left(\left(\frac{5}{8}\right)\left(\frac{4}{5}\log_2\frac{4}{5}+\frac{1}{5}\log_2\frac{1}{5}\right)+\left(\frac{3}{8}\right)\left(1\overset{0}{\log_2 1}\right)\right) = \underline{0.549} \text{ (Heart Rate)}$$

Info-gain (T_BP) = E(class) − ((Prob Normal)(Entropy(N)) + (Prob$_{\text{Abn.}}$)(E(A)))

$$I = 1-\left(\left(\frac{6}{8}\right)\left(\frac{\overset{2}{\cancel{4}}}{\underset{3}{\cancel{6}}}\log_2\frac{\overset{2}{\cancel{4}}}{\underset{3}{\cancel{6}}}+\frac{\overset{1}{\cancel{2}}}{\underset{3}{\cancel{6}}}\log_2\frac{\overset{1}{\cancel{2}}}{\underset{3}{\cancel{6}}}\right)\right)+\left(\left(\frac{2}{8}\right)(0)\right)$$

$$I = 1-\left(\left(\frac{\overset{3}{\cancel{6}}}{\cancel{8}\,4}\right)\left(-\frac{2}{3}\log_2\frac{2}{3}+-\frac{1}{3}\log_2\frac{1}{3}\right)\right)$$

$\boxed{I = 0.3112}$ Blood Prosure



Legend: Healthy patients / Ill patients

2. What would be the information gain for the split with the attribute Patient Nr. in the overall dataset? Compare the result with the information gain of other attributes. What problems occur when a decision tree uses the attribute Patient Nr.?

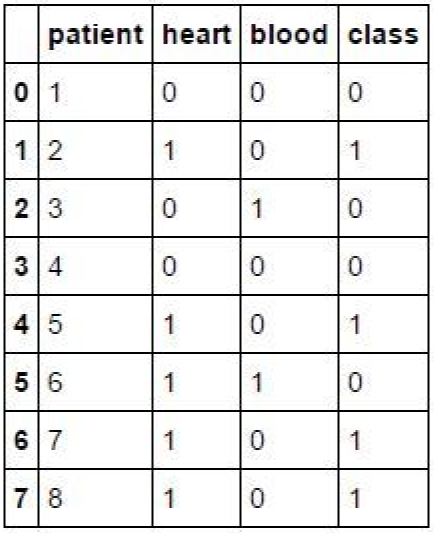
> The information gain on patient Nr. attribute is equals to 1 because the entropy for this attribute is 0. This is because patient number uniquely identifies a patient, so it has a high information gain. This leads to overfitting, as we would split on each value of the patient number, making the decision tree not general enough.

3. State a simple, but valid strategy to parallelize decision tree learning?

> **Task parallelism:** In this strategy, the decision nodes are dynamically distributed among X processors. One processor starts using all the training data and starts the construction phase. After the initial split, each node can compute the next subtree in a parallel way. This means that subsequent sub-trees are being made on different processors until it finishes the decision process.

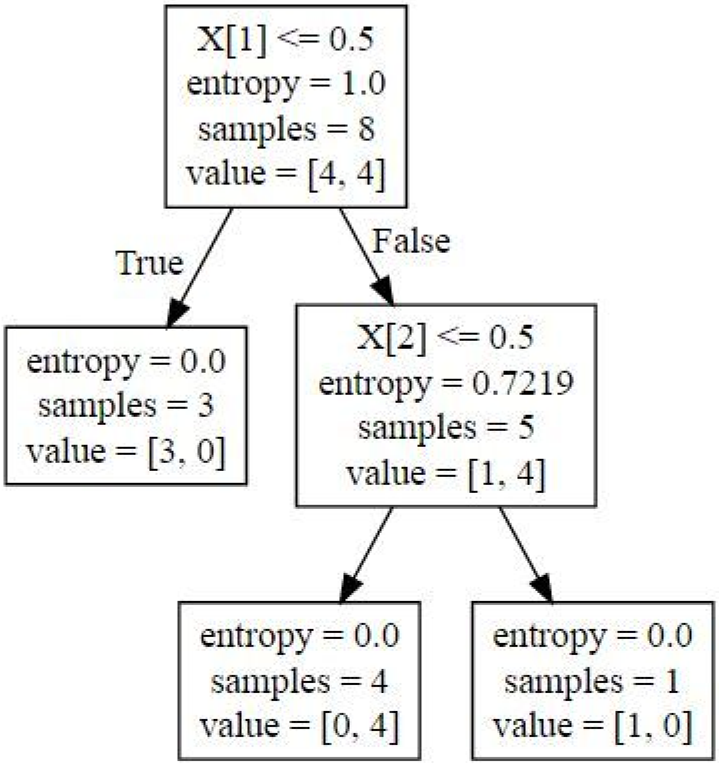
4. Give an example dataset such that the approach of greedily learning decision trees with information gain does not lead to an optimum tree with respect to the tree complexity (measured as number of nodes in the tree). Assume that the decision tree should correctly classify all training examples (no pruning).

> If a dataset had a column with many distinct values, using a greedy algorithm and information gain as the metric for deciding the parameter to split on, would produce a tree with more nodes than is required for a 100 percent correct classification.
>
> As we didn't have time to implement this greedy algorithm, and existing implementations seem to be "smart" enough not to make the mistake of using the patient number for splitting data, we will illustrate with some assumptions:
>
> Data set:

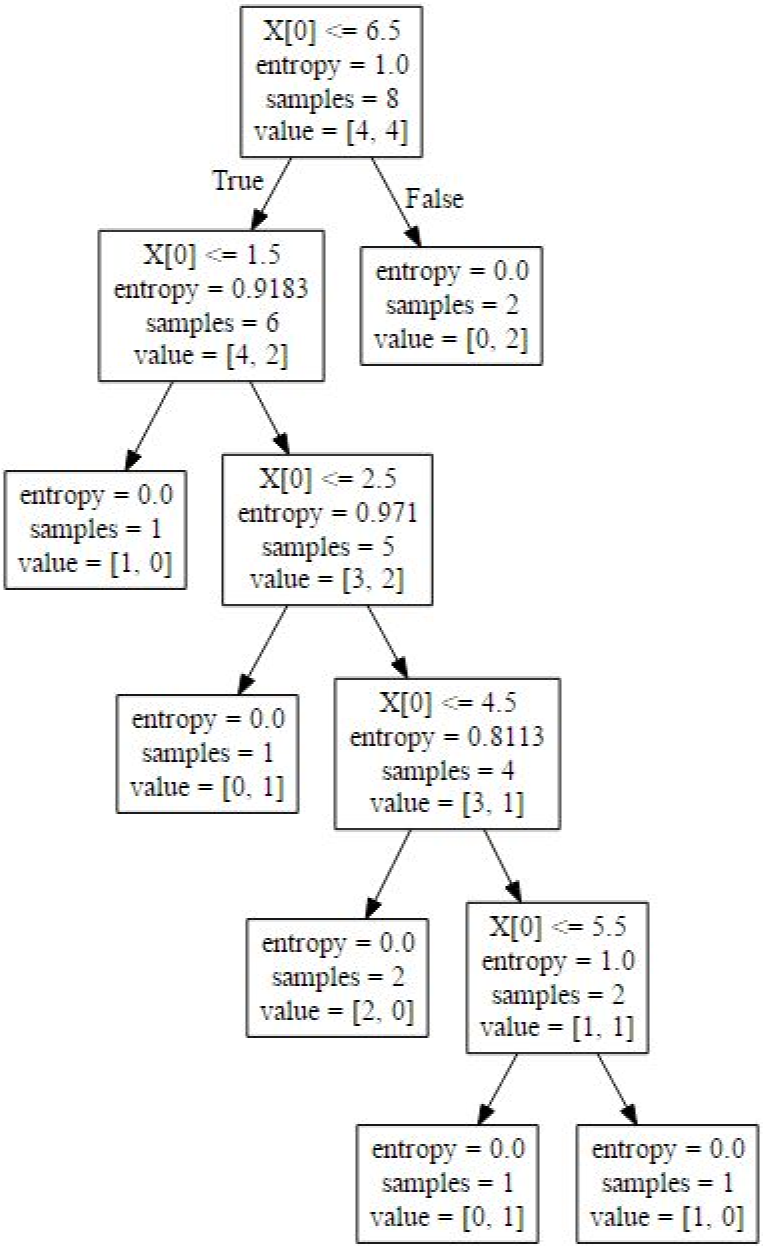


	patient	heart	blood	class
0	1	0	0	0
1	2	1	0	1
2	3	0	1	0
3	4	0	0	0
4	5	1	0	1
5	6	1	1	0
6	7	1	0	1
7	8	1	0	1

Tree generated using fields heart(rate) and blood(pressure) to split data:



Tree generated when splitting on attribute patient number:


X[0] <= 6.5
entropy = 1.0
samples = 8
value = [4, 4]
True
False
X[0] <= 1.5
entropy = 0.9183
samples = 6
value = [4, 2]
entropy = 0.0
samples = 2
value = [0, 2]
entropy = 0.0
samples = 1
value = [1, 0]
X[0] <= 2.5
entropy = 0.971
samples = 5
value = [3, 2]
entropy = 0.0
samples = 1
value = [0, 1]
X[0] <= 4.5
entropy = 0.8113
samples = 4
value = [3, 1]
entropy = 0.0
samples = 2
value = [2, 0]
X[0] <= 5.5
entropy = 1.0
samples = 2
value = [1, 1]
entropy = 0.0
samples = 1
value = [0, 1]
entropy = 0.0
samples = 1
value = [1, 0]