Friedrich-Alexander-Universität
Technische Fakultät

Department of Computer Science
Computer Science 6
(Data Management)

Knowledge Discovery in Databases with Exercises
Summer Semester 2025

Exercise Sheet 4: Classification

About this Exercise Sheet

This exercise sheet focuses on the content of lecture *7. Classification*.

It includes both theoretical exercises on Decision Trees (Exercise 1) and Naïve Bayes (Exercise 2) and a practical data science exercise (Exercise 3).

The exercise sheet is designed for a three-week period, during which the tasks can be completed flexibly (Planned is one exercise per week).

The sample solution will be published after the three weeks have elapsed.

Preparation

Before participating in the exercise, you must prepare the following:

1. **Install Python and pip on your computer**
   - Detailed instructions can be found in `1-Introduction-Python-Pandas.pdf`.
2. **Download provided additional files**
   - Download `Additional-Files-Student.zip` from StudOn
   - Extract it to a folder of your choice.
3. **Install required Python packages**
   - Open a terminal and navigate to the folder where you extracted the files.
   - Run the command `pip install -r requirements.txt` within the extracted additional files folder to install the required Python packages.

Exercise 1: Decision Trees

Given is a dataset D.

D is containing a continuous attribute (*Age*) and two categorical attributes (*Major* and *Participation*) which can be used to predict the target attribute *Passed*.

Age	Major	Participation	Passed
23	CS	High	Yes
23	DS	Low	No
26	DS	High	Yes
24	DS	Medium	Yes
26	DS	Medium	No
26	DS	Low	No

Task 1: Information Gain

Use the algorithm for **Decision Tree Induction** known from the lecture to build a decision tree for dataset D. The decision tree should be built using **Information Gain** as the attribute selection method.

Write down **all** intermediate steps.

1. **Create the root node:**

   To create the root node, we need to calculate the Information Gain for each attribute and select the one with the highest Information Gain.

   a) **Calculate the Entropy of the target attribute *Passed*:**

$$\begin{aligned}\text{Info}(D) &= -\sum_{i=1}^{m} p_i \log_2(p_i) \\ &= -p_{\text{Passed=Yes}} \log_2(p_{\text{Passed=Yes}}) - p_{\text{Passed=No}} \log_2(p_{\text{Passed=No}}) \\ &= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) \\ &= 1\end{aligned}$$

   b) **Calculate the Information Gain for all attributes:**

      i. **Attribute *Age*:**

      Age is a continuous attribute. To calculate the Information Gain, we need to find the best split point.

      A. **Split point $23,5$:**

$$
\begin{aligned}
\text{Info}_{\text{Age}}(D) &= \sum_{j=1}^{v} \frac{|D_{\text{Age},\,j}|}{|D_{\text{Age}}|} \text{Info}(D_{A_{\text{Age},\,j}}) \\
&= \frac{|D_{\text{Age}\leq 2,5}|}{|D_{\text{Age}}|} \text{Info}(D_{\text{Age}\leq 2,5}) + \frac{|D_{\text{Age}>2,5}|}{|D_{\text{Age}}|} \text{Info}(D_{\text{Age}>2,5}) \\
&= \frac{2}{6}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) + \frac{4}{6}\left(-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right)\right) \\
&= \frac{2}{6}\cdot 1 + \frac{4}{6}\cdot 1 \\
&= 1 \\
\text{Gain}_{\text{Age}} &= \text{Info}(D) - \text{Info}_{\text{Age}}(D) \\
&= 1 - 1 \\
&= 0
\end{aligned}
$$

B. **Split point** $25,0$**:**

$$
\begin{aligned}
\text{Info}_{\text{Age}}(D) &= \sum_{j=1}^{v} \frac{|D_{\text{Age},\,j}|}{|D_{\text{Age}}|} \text{Info}(D_{A_{\text{Age},\,j}}) \\
&= \frac{|D_{\text{Age}\leq 5}|}{|D_{\text{Age}}|} \text{Info}(D_{\text{Age}\leq 5}) + \frac{|D_{\text{Age}>5}|}{|D_{\text{Age}}|} \text{Info}(D_{\text{Age}>5}) \\
&= \frac{3}{6}\left(-\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right)\right) + \frac{3}{6}\left(-\frac{1}{3}\log_2\left(\frac{1}{3}\right) - \frac{2}{3}\log_2\left(\frac{2}{3}\right)\right) \\
&= \frac{3}{6}\cdot 0,9183 + \frac{3}{6}\cdot 0,9183 \\
&= 0,9183 \\
\text{Gain}_{\text{Age}} &= \text{Info}(D) - \text{Info}_{\text{Age}}(D) \\
&= 1 - 0,9183 \\
&= 0,0817
\end{aligned}
$$

Therefore, the Information Gain for the attribute *Age* is $0,817$ (if we split at $25,0$).

ii. **Attribute *Major*:**

Major is a categorical attribute with two possible values: *CS* and *DS*.

Since it is a categorical attribute and we are using the Information Gain, there is no need to determine a splitting criterion.

$$
\begin{aligned}
\text{Info}_{\text{Major}}(D) &= \sum_{j=1}^{v} \frac{|D_{\text{Major},\,j}|}{|D_{\text{Major}}|} \text{Info}(D_{A_{\text{Major},\,j}}) \\
&= \frac{|D_{\text{Major=CS}}|}{|D_{\text{Major}}|} \text{Info}(D_{\text{Major=CS}}) + \frac{|D_{\text{Major=DS}}|}{|D_{\text{Major}}|} \text{Info}(D_{\text{Major=DS}}) \\
&= \frac{1}{6}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) + \frac{5}{6}\left(-\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)\right) \\
&= \frac{1}{6}\left(-0 - 0 \cdot \text{undefined}\right) + \frac{5}{6}\left(0,9710\right) \quad \textit{Hint: Multiplication by 0 always results in 0} \\
&= \frac{1}{6}\left(-0 - 0\right) + \frac{5}{6}\left(0,9710\right) \\
&= 0,8090 \\
\text{Gain}_{\text{Major}} &= \text{Info}(D) - \text{Info}_{\text{Major}}(D) \\
&= 1 - 0,8090 \\
&= 0,1910
\end{aligned}
$$

iii. **Attribute *Participation*:**

Participation is a categorical attribute with three possible values: *High*, *Medium* and *Low*.

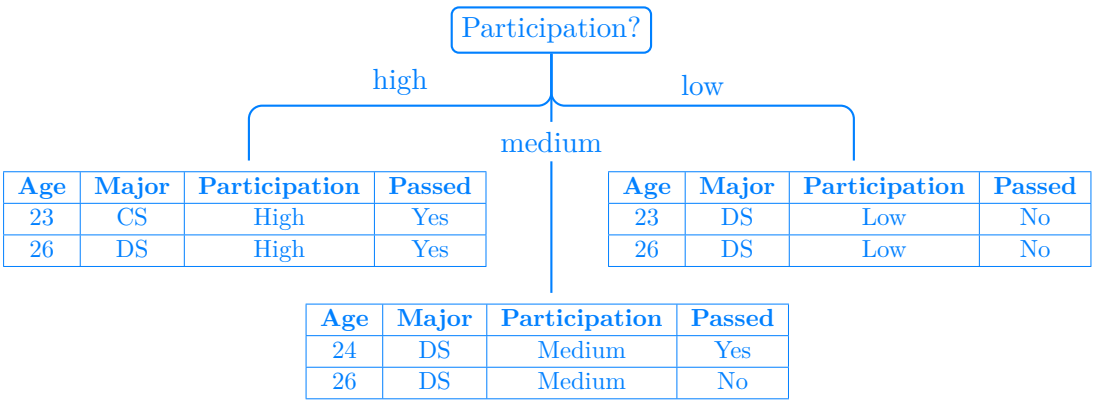
Since it is a categorical attribute and we are using the Information Gain, there is no need to determine a splitting criterion.

$$
\begin{aligned}
\text{Info}_{\text{Parti.}}(D) &= \sum_{j=1}^{v} \frac{|D_{\text{Parti.},\,j}|}{|D_{\text{Parti.}}|} \text{Info}(D_{A_{\text{Parti.},\,j}}) \\
&= \frac{|D_{\text{Parti.=High}}|}{|D_{\text{Parti.}}|} \text{Info}(D_{\text{Parti.=High}}) + \frac{|D_{\text{Parti.=Medium}}|}{|D_{\text{Parti.}}|} \text{Info}(D_{\text{Parti.=Medium}}) \\
&+ \frac{|D_{\text{Parti.=Low}}|}{|D_{\text{Parti.}}|} \text{Info}(D_{\text{Parti.=Low}}) \\
&= \frac{2}{6}\left(-\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) + \frac{2}{6}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) \\
&+ \frac{2}{6}\left(-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) \\
&= \frac{2}{6} \cdot 0 + \frac{2}{6} \cdot 1 + \frac{2}{6} \cdot 0 \\
&= 0,3333 \\
\text{Gain}_{\text{Parti.}} &= \text{Info}(D) - \text{Info}_{\text{Parti.}}(D) \\
&= 1 - 0,3333 \\
&= 0,6667
\end{aligned}
$$

c) **Create the node based on the highest Information Gain:**

The attribute with the highest Information Gain is *Participation* with a value of $0,6667$. It will therefore become the splitting attribute for the root node.

The resulting tree will look like this:



Age	Major	Participation	Passed
23	CS	High	Yes
26	DS	High	Yes

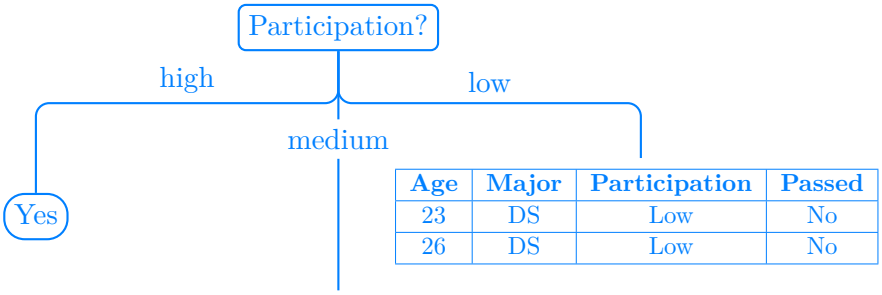
Age	Major	Participation	Passed
23	DS	Low	No
26	DS	Low	No

Age	Major	Participation	Passed
24	DS	Medium	Yes
26	DS	Medium	No

2. **Visit each branch:**

   a) **Branch *High*:**

   All samples in the partial dataset of the branch *High* have the same value for the target attribute *Passed*. Therefore, the branch becomes a leaf node.



Age	Major	Participation	Passed
23	DS	Low	No
26	DS	Low	No

Age	Major	Participation	Passed
24	DS	Medium	Yes
26	DS	Medium	No

   b) **Branch *Medium*:**

   The partial dataset of the branch *Medium* contains samples with different values for the target attribute *Passed*. Therefore, we need to create a new node.

      i. **Calculate the Entropy of the target attribute *Passed*:**

$$\begin{aligned}
\text{Info}(D_{\text{Parti=Medium}}) &= -\sum_{i=1}^{m} p_i \log_2(p_i) \\
&= -p_{\text{Passed=Yes}} \log_2(p_{\text{Passed=Yes}}) - p_{\text{Passed=No}} \log_2(p_{\text{Passed=No}}) \\
&= -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) \\
&= 1
\end{aligned}$$

      ii. **Calculate the Information Gain for all attributes that are not yet a node:**

A. **Attribute *Age*:**

Age is still a continuous attribute. However since there are only two different values in the partial dataset, we only have one split point.

$$
\begin{aligned}
\text{Info}_{\text{Age}}(D_{\text{Parti=Medium}}) &= \sum_{j=1}^{v} \frac{|D_{\text{Age},j}|}{|D_{\text{Age}}|} \text{Info}(D_{A_{\text{Age},j}}) \\
&= \frac{|D_{\text{Age}\leq 25}|}{|D_{\text{Age}}|} \text{Info}(D_{\text{Age}\leq 25}) + \frac{|D_{\text{Age}>25}|}{|D_{\text{Age}}|} \text{Info}(D_{\text{Age}>25}) \\
&= \frac{1}{2}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) + \frac{1}{2}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right) - \frac{0}{1}\log_2\left(\frac{0}{1}\right)\right) \\
&= \frac{1}{2}\cdot 0 + \frac{1}{2}\cdot 0 \\
&= 0 \\
\text{Gain}_{\text{Age}} &= \text{Info}(D_{\text{Parti=Medium}}) - \text{Info}_{\text{Age}}(D_{\text{Parti=Medium}}) \\
&= 1 - 0 \\
&= 1
\end{aligned}
$$

B. **Attribute *Major*:**

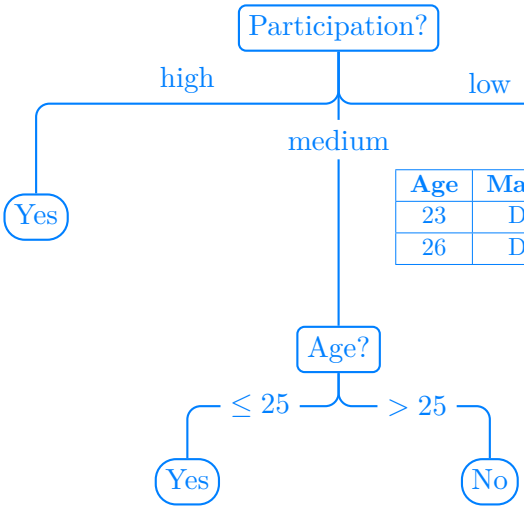
Major is still a categorical attribute. This time we only have one value in the partial dataset. Therefore, the Information Gain is 0:

$$
\begin{aligned}
\text{Info}_{\text{Major}}(D_{\text{Parti=Medium}}) &= \sum_{j=1}^{v} \frac{|D_{\text{Major},j}|}{|D_{\text{Major}}|} \text{Info}(D_{A_{\text{Major},j}}) \\
&= \frac{|D_{\text{Major=DS}}|}{|D_{\text{Major}}|} \text{Info}(D_{\text{Major=DS}}) \\
&= \frac{2}{2}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) \\
&= \frac{2}{2}(1) \\
&= 1 \\
\text{Gain}_{\text{Major}} &= \text{Info}(D_{\text{Parti=Medium}}) - \text{Info}_{\text{Major}}(D_{\text{Parti=Medium}}) \\
&= 1 - 1 \\
&= 0
\end{aligned}
$$

iii. **Create the node based on the highest Information Gain:**

The attribute with the highest Information Gain is *Age* with a value of 1. It will therefore become the splitting attribute for the node.

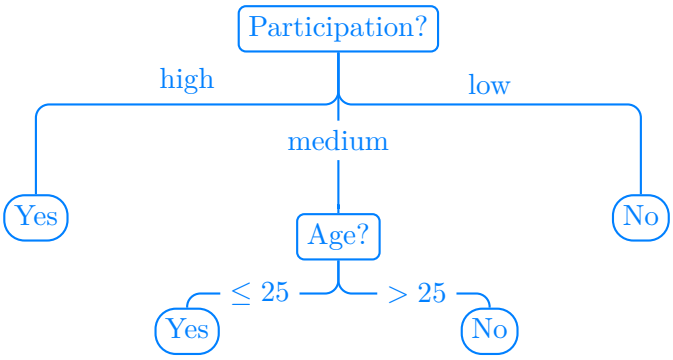
The resulting tree will look like this:



Age	Major	Participation	Passed
23	DS	Low	No
26	DS	Low	No

c) **Branch *Low*:**

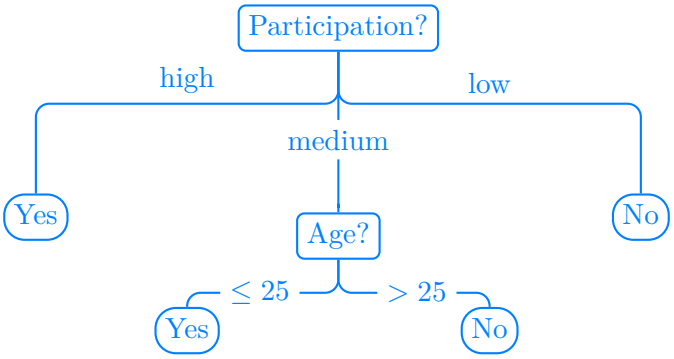
All samples in the partial dataset of the branch *Low* have the same value for the target attribute *Passed*. Therefore, the branch becomes a leaf node.



3. **Stop the algorithm:**

Since all branches are now leaf nodes, the algorithm can be stopped.

The final decision tree is:



Task 2: Gini Index

This time, the decision tree for dataset D should be built using the **Gini Index** as the attribute selection method.

Task 2.a: Root Node

Using the algorithm for building a decision tree with the Gini Index, create the root node of the decision tree for the dataset D.

Write down **all** intermediate steps **up to** (and including) the point where the root node is created.

1. **Create the root node:**

   To create the root node, we need to calculate the Gini Index for each attribute and select the one with the lowest Gini Index.

   a) **Calculate the impurity of the whole Dataset D:**

$$\begin{aligned}\text{Gini}(D) &= 1 - \sum_{i=1}^{m} p_i^2 \\ &= 1 - p_{\text{Passed=Yes}}^2 - p_{\text{Passed=No}}^2 \\ &= 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 \\ &= 0,5\end{aligned}$$

   b) **Calculate the Gini Index for all attributes:**

      i. **Attribute *Age*:**

      Age is a continuous attribute. Similar to the Information Gain, we need to find the best split point.

      A. **Split point $23,5$:**

$$\begin{aligned}\text{Gini}_{\text{Age}}(D) &= \frac{|D_{\text{Age}\leq 23,5}|}{|D_{\text{Age}}|}\text{Gini}(D_{\text{Age}\leq 23,5}) + \frac{|D_{\text{Age}>23,5}|}{|D_{\text{Age}}|}\text{Gini}(D_{\text{Age}>23,5}) \\ &= \frac{2}{6}\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) + \frac{4}{6}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) \\ &= \frac{2}{6} \cdot 0,5 + \frac{4}{6} \cdot 0,5 \\ &= 0,5 \\ \Delta\text{Gini}_{\text{Age}}(D) &= \text{Gini}(D) - \text{Gini}_{\text{Age}}(D) \\ &= 0,5 - 0,5 \\ &= 0\end{aligned}$$

B. **Split point** $25,0$**:**

$$\begin{aligned}
\text{Gini}_{\text{Age}}(D) &= \frac{|D_{\text{Age}\leq 25,0}|}{|D_{\text{Age}}|}\text{Gini}(D_{\text{Age}\leq 25,0}) + \frac{|D_{\text{Age}>25,0}|}{|D_{\text{Age}}|}\text{Gini}(D_{\text{Age}>25,0}) \\
&= \frac{3}{6}\left(1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2\right) + \frac{3}{6}\left(1-\left(\frac{1}{3}\right)^2-\left(\frac{2}{3}\right)^2\right) \\
&= \frac{3}{6}\cdot 0,4444 + \frac{3}{6}\cdot 0,4444 \\
&= 0,4444 \\
\Delta\text{Gini}_{\text{Age}}(D) &= \text{Gini}(D) - \text{Gini}_{\text{Age}}(D) \\
&= 0,5 - 0,4444 \\
&= 0,0556
\end{aligned}$$

The best split point is $25,0$ since its Gini Index is the lowest $(0,4444)$ and therefore the reduction of impurity $(0,0556)$ is the highest.

ii. **Attribute *Major*:**

Major is a categorical attribute with two possible values: *CS* and *DS*.

Gini Index only supports two-way splits. Therefore if we would have had more than two values, we would have needed to calculate the Gini Index for each possible split.

However since we only have two values, we can directly calculate the Gini Index for the attribute *Major*:

$$\begin{aligned}
\text{Gini}_{\text{Major}}(D) &= \frac{|D_{\text{Major=CS}}|}{|D_{\text{Major}}|}\text{Gini}(D_{\text{Major=CS}}) + \frac{|D_{\text{Major=DS}}|}{|D_{\text{Major}}|}\text{Gini}(D_{\text{Major=DS}}) \\
&= \frac{1}{6}\left(1-\left(\frac{1}{1}\right)^2-\left(\frac{0}{1}\right)^2\right) + \frac{5}{6}\left(1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2\right) \\
&= \frac{1}{6}\cdot 0 + \frac{5}{6}\cdot 0,48 \\
&= 0,4 \\
\Delta\text{Gini}_{\text{Major}}(D) &= \text{Gini}(D) - \text{Gini}_{\text{Major}}(D) \\
&= 0,5 - 0,4 \\
&= 0,1
\end{aligned}$$

iii. **Attribute *Participation*:**

Participation is a categorical attribute with three possible values: *High*, *Medium* and *Low*.

Since we have more than two different attributes, we have to calculate the Gini Index for each possible attribute combination.

A. **Combination *{High, Medium}* and *{Low}*:**

$$
\begin{aligned}
\text{Gini}_{\text{Parti.}}(D) &= \frac{|D_{\text{Parti.=High, Medium}}|}{|D_{\text{Parti.}}|}\text{Gini}(D_{\text{Major=High, Medium}}) + \frac{|D_{\text{Parti.=Low}}|}{|D_{\text{Parti.}}|}\text{Gini}(D_{\text{Parti.=Low}}) \\
&= \frac{4}{6}\left(1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right) + \frac{2}{6}\left(1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2\right) \\
&= \frac{4}{6} \cdot 0,375 + \frac{2}{6} \cdot 0 \\
&= 0,25 \\
\Delta\text{Gini}_{\text{Parti.}}(D) &= \text{Gini}(D) - \text{Gini}_{\text{Parti.}}(D) \\
&= 0,5 - 0,25 \\
&= 0,25
\end{aligned}
$$

B. **Combination *{High, Low}* and *{Medium}*:**

$$
\begin{aligned}
\text{Gini}_{\text{Parti.}}(D) &= \frac{|D_{\text{Parti.=High, Low}}|}{|D_{\text{Parti.}}|}\text{Gini}(D_{\text{Major=High, Low}}) + \frac{|D_{\text{Parti.=Medium}}|}{|D_{\text{Parti.}}|}\text{Gini}(D_{\text{Parti.=Medium}}) \\
&= \frac{4}{6}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right) + \frac{2}{6}\left(1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right) \\
&= \frac{4}{6} \cdot 0,5 + \frac{2}{6} \cdot 0,5 \\
&= 0,5 \\
\Delta\text{Gini}_{\text{Parti.}}(D) &= \text{Gini}(D) - \text{Gini}_{\text{Parti.}}(D) \\
&= 0,5 - 0,5 \\
&= 0
\end{aligned}
$$

C. **Combination *{Medium, Low}* and *{High}*:**

$$
\begin{aligned}
\text{Gini}_{\text{Parti.}}(D) &= \frac{|D_{\text{Parti.=Medium, Low}}|}{|D_{\text{Parti.}}|}\text{Gini}(D_{\text{Major=Medium, Low}}) + \frac{|D_{\text{Parti.=High}}|}{|D_{\text{Parti.}}|}\text{Gini}(D_{\text{Parti.=High}}) \\
&= \frac{4}{6}\left(1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right) + \frac{2}{6}\left(1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2\right) \\
&= \frac{4}{6} \cdot 0,375 + \frac{2}{6} \cdot 0 \\
&= 0,25 \\
\Delta\text{Gini}_{\text{Parti.}}(D) &= \text{Gini}(D) - \text{Gini}_{\text{Parti.}}(D) \\
&= 0,5 - 0,25 \\
&= 0,25
\end{aligned}
$$

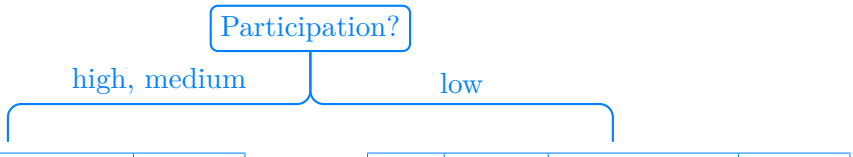
The best combinations are *{High, Medium}* and *{Low}* and *{Medium, Low}* and *{High}* since their Gini Index is the lowest $(0,25)$ and therefore the reduction of impurity $(0,25)$ is the highest.

We can choose either of them as the combination „representing“ the attribute *Participation*. For the simplicity of this sample solution, we will choose the combination *{High, Medium}* and *{Low}*.

c) **Create the root node based on the lowest Gini Index:**

The attribute with the lowest Gini Index is *Participation* with a value of $0,25$. It will therefore become the splitting attribute for the root node.

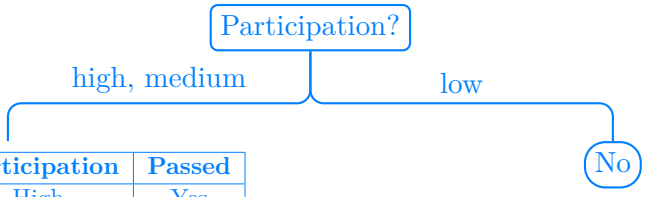
The resulting tree will look like this:



Age	Major	Participation	Passed
23	CS	High	Yes
26	DS	High	Yes
24	DS	Medium	Yes
26	DS	Medium	No

Age	Major	Participation	Passed
23	DS	Low	No
26	DS	Low	No

Since the table for the branch *Low* only contains tuples with the class label *No*, we can already create a leaf node for this branch:



Age	Major	Participation	Passed
23	CS	High	Yes
26	DS	High	Yes
24	DS	Medium	Yes
26	DS	Medium	No

Task 2.b: Splitting attribute candidates

In the resulting tree from Task 2.a, one of the branches is already a leaf node.

Which of the **attributes** *Age*, *Major* and *Participation* have to **be checked** for their Gini index in the next step necessary to further split the remaining branch?

- **Attribute *Age*:**

  The attribute *Age* should be checked for its Gini Index in the branch *High, Medium* since it still contains different values.

- **Attribute *Major*:**

  The attribute *Major* should be checked for its Gini Index in the branch *High, Medium* since it still contains different values.

- **Attribute *Participation*:**

  Contrary to the procedure for Information Gain - where the attribute *Participation* would not be checked again - we also have to check the attribute *Participation* for its Gini Index in the branch *High, Medium* since it still contains different values.

  This is due to the fact that the Gini Index does not support multi-way splits. Therefore a categorical attribute with more than two values can be splitting attribute multiple times in the same branch.

Task 3: Gain Ratio

The **Gain Ratio** is a solution to a problem of the **Information Gain**.

Come up with an example **dataset** showing the problem of the Information Gain and explain how the Gain Ratio solves this problem.

The problem of the Information Gain is that it tends to favor attributes with a large number of values.

If we for example take a look at the following dataset D:

Major	Participation	Passed
CS	High	Yes
DS	Low	No
IIS	High	Yes
AI	Medium	Yes
ICT	Medium	No
CME	Low	No

We can see that the attribute *Major* only contains unique values.

Therefore, the Information Gain for the attribute *Major* would be 1:

$$
\begin{aligned}
\text{Info}(D) &= -\sum_{i=1}^{m} p_i \log_2(p_i) \\
&= -p_{\text{Passed=Yes}} \log_2(p_{\text{Passed=Yes}}) - p_{\text{Passed=No}} \log_2(p_{\text{Passed=No}}) \\
&= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) \\
&= 1 \\
\text{Info}_{\text{Major}}(D) &= \sum_{j=1}^{v} \frac{|D_{\text{Major},\,j}|}{|D_{\text{Major}}|}\text{Info}(D_{A_{\text{Major},\,j}}) \\
&= \frac{1}{6}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) + \frac{1}{6}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) + \frac{1}{6}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) \\
&+ \frac{1}{6}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) + \frac{1}{6}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) + \frac{1}{6}\left(-\frac{1}{1}\log_2\left(\frac{1}{1}\right)\right) \\
&= 0 \\
\text{Gain}_{\text{Major}} &= \text{Info}(D) - \text{Info}_{\text{Major}}(D) \\
&= 1 - 0 \\
&= 1
\end{aligned}
$$

On the other hand the attribute *Participation* still (see Task 1) has an Information Gain of $0,6667$:

$$
\begin{aligned}
\text{Info}_{\text{Parti.}}(D) &= \sum_{j=1}^{v} \frac{|D_{\text{Parti.},\,j}|}{|D_{\text{Parti.}}|}\text{Info}(D_{A_{\text{Parti.},\,j}}) \\
&= \frac{|D_{\text{Parti.=High}}|}{|D_{\text{Parti.}}|}\text{Info}(D_{\text{Parti.=High}}) + \frac{|D_{\text{Parti.=Medium}}|}{|D_{\text{Parti.}}|}\text{Info}(D_{\text{Parti.=Medium}}) \\
&+ \frac{|D_{\text{Parti.=Low}}|}{|D_{\text{Parti.}}|}\text{Info}(D_{\text{Parti.=Low}}) \\
&= \frac{2}{6}\left(-\frac{2}{2}\log_2\left(\frac{2}{2}\right) - \frac{0}{2}\log_2\left(\frac{0}{2}\right)\right) + \frac{2}{6}\left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) \\
&+ \frac{2}{6}\left(-\frac{0}{2}\log_2\left(\frac{0}{2}\right) - \frac{2}{2}\log_2\left(\frac{2}{2}\right)\right) \\
&= \frac{2}{6}\cdot 0 + \frac{2}{6}\cdot 1 + \frac{2}{6}\cdot 0 \\
&= 0,3333 \\
\text{Gain}_{\text{Parti.}} &= \text{Info}(D) - \text{Info}_{\text{Parti.}}(D) \\
&= 1 - 0,3333 \\
&= 0,6667
\end{aligned}
$$

With the Information Gain, we would choose the attribute *Major* as the splitting attribute since it has a higher Information Gain than the attribute *Participation*.

However, since the attribute *Major* only contains unique values, it is not a good splitting attribute, since we on the one hand up end up with a big multi-way split and on the other hand risk to overfit our decision tree on the training data.

The Gain Ratio solves this problem by normalizing the Information Gain with the **Split Information**:

$$\text{SplitInfo}_{\text{A}}(D) = -\sum_{j=1}^{v} \frac{|D_{\text{A},j}|}{|D|} \log_2\left(\frac{|D_{\text{A},j}|}{|D|}\right)$$
$$\text{GainRatio}_{\text{A}} = \frac{\text{Gain}_{\text{A}}}{\text{SplitInfo}_{\text{A}}(D)}$$

If we apply the Gain Ratio to the Gain of the attributes *Major* and *Participation*, we get:

$$\begin{aligned}\text{SplitInfo}_{\text{Major}}(D) &= -\frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right) \\ &\quad - \frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right) - \frac{1}{6}\log_2\left(\frac{1}{6}\right) \\ &= 2,5850 \\ \text{GainRatio}_{\text{Major}} &= \frac{1}{2,5850} \\ &= 0,3868\end{aligned}$$

$$\begin{aligned}\text{SplitInfo}_{\text{Parti.}}(D) &= -\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) \\ &= 1,5849 \\ \text{GainRatio}_{\text{Parti.}} &= \frac{0,6667}{1,5849} \\ &= 0,4207\end{aligned}$$

With the Gain Ratio, we would now choose the attribute *Participation* as the splitting attribute since it has a higher Gain Ratio than the attribute *Major*.

We therefore avoid the problem of the Information Gain to favor attributes with a large number of values.

Note: *The Gain Ratio has its own problems, as it becomes unstable if SplitInfo is close to zero. In this case, the Gain Ratio can become very high. Therefore, the Gain Ratio should also be used with caution.*

Exercise 2: Naïve Bayes

Given is a dataset D.

It can be assumed that *Topic*, *Knowledge* and *Hours* are conditionally independent of each other.

The attributes *Topic* and *Knowledge* are categorical attributes.

The attribute *Hours* is a continuous attribute. It can be assumed that the values of this attribute are distributed according to a Gaussian distribution.

Topic	Knowledge	Hours	Passed
Classification	High	1,0	No
Clustering	Low	4,0	No
Frequent Patterns	High	5,0	Yes
Clustering	Medium	5,0	Yes
Frequent Patterns	High	2,0	No
Frequent Patterns	Medium	3,0	Yes
Classification	Low	6,0	Yes
Clustering	Low	5,0	Yes
Clustering	High	3,0	Yes
Classification	Medium	4,0	Yes

Task 1: Classification

Use the dataset D and the Naïve Bayes algorithm to classify the following tuples:

Topic	Knowledge	Hours	Passed
Clustering	Medium	4,0	?
Classification	High	3,0	?
Frequent Patterns	Low	6,8	?

Write down **all** intermediate steps.

1. **Calculate the prior probabilities:**

$$P(\texttt{Passed} = \text{"Yes"}) = \frac{7}{10} = 0,7$$
$$P(\texttt{Passed} = \text{"No"}) = \frac{3}{10} = 0,3$$

2. **Calculate the likelihoods:**

a) **Attribute *Topic*:**

$$P(\texttt{Topic} = \text{"Clustering"}|\texttt{Passed} = \text{"Yes"}) = \frac{3}{7} \approx 0,4286$$
$$P(\texttt{Topic} = \text{"Clustering"}|\texttt{Passed} = \text{"No"}) = \frac{1}{3} \approx 0,3333$$

$$P(\texttt{Topic} = \text{"Classification"}|\texttt{Passed} = \text{"Yes"}) = \frac{2}{7} \approx 0,2857$$
$$P(\texttt{Topic} = \text{"Classification"}|\texttt{Passed} = \text{"No"}) = \frac{1}{3} \approx 0,3333$$

$$P(\texttt{Topic} = \text{"Frequent Patterns"}|\texttt{Passed} = \text{"Yes"}) = \frac{2}{7} \approx 0,2857$$
$$P(\texttt{Topic} = \text{"Frequent Patterns"}|\texttt{Passed} = \text{"No"}) = \frac{1}{3} \approx 0,3333$$

b) **Attribute *Knowledge*:**

$$P(\texttt{Knowledge} = \text{"High"}|\texttt{Passed} = \text{"Yes"}) = \frac{2}{7} \approx 0,2857$$
$$P(\texttt{Knowledge} = \text{"High"}|\texttt{Passed} = \text{"No"}) = \frac{2}{3} \approx 0,6667$$

$$P(\texttt{Knowledge} = \text{"Medium"}|\texttt{Passed} = \text{"Yes"}) = \frac{3}{7} \approx 0,4286$$
$$P(\texttt{Knowledge} = \text{"Medium"}|\texttt{Passed} = \text{"No"}) = \frac{0}{3} \approx 0$$

$$P(\texttt{Knowledge} = \text{"Low"}|\texttt{Passed} = \text{"Yes"}) = \frac{2}{7} \approx 0,2857$$
$$P(\texttt{Knowledge} = \text{"Low"}|\texttt{Passed} = \text{"No"}) = \frac{1}{3} \approx 0,3333$$

c) **Attribute *Hours*:**

Since the attribute *Hours* is continuous and follows a Gaussian distribution, we have to calculate the mean μ and the standard deviation σ for each class label:

$$\mu_{\texttt{Passed}=\text{"Yes"}} = \frac{5+5+3+6+5+3+4}{7} = \frac{31}{7} \approx 4,4286$$
$$\sigma_{\texttt{Passed}=\text{"Yes"}} = \sqrt{\frac{(5-\frac{31}{7})^2 \cdot 3 + (3-\frac{31}{7})^2 \cdot 2 + (6-\frac{31}{7})^2 + (4-\frac{31}{7})^2}{7-1}} \approx 1,1339$$

$$\mu_{\texttt{Passed}=\text{"No"}} = \frac{1+4+2}{3} = \frac{7}{3} \approx 2,3333$$
$$\sigma_{\texttt{Passed}=\text{"No"}} = \sqrt{\frac{(1-\frac{7}{3})^2 + (4-\frac{7}{3})^2 + (2-\frac{7}{3})^2}{3-1}} \approx 1,5275$$

We can now calculate the likelihoods for the attribute *Hours*:

$$P(\texttt{Hours} = \text{"4"}|\texttt{Passed} = \text{"Yes"}) \approx \frac{1}{\sqrt{2\pi} \cdot 1,1339} \cdot e^{-\frac{(4-4,4286)^2}{2 \cdot 1,1339^2}} \approx 0,3276$$
$$P(\texttt{Hours} = \text{"4"}|\texttt{Passed} = \text{"No"}) \approx \frac{1}{\sqrt{2\pi} \cdot 1,5275} \cdot e^{-\frac{(4-2,3333)^2}{2 \cdot 1,5275^2}} \approx 0,1440$$

$$P(\texttt{Hours} = \text{"3"}|\texttt{Passed} = \text{"Yes"}) \approx \frac{1}{\sqrt{2\pi} \cdot 1,1339} \cdot e^{-\frac{(3-4,4286)^2}{2 \cdot 1,1339^2}} \approx 0,1591$$
$$P(\texttt{Hours} = \text{"3"}|\texttt{Passed} = \text{"No"}) \approx \frac{1}{\sqrt{2\pi} \cdot 1,5275} \cdot e^{-\frac{(3-2,3333)^2}{2 \cdot 1,5275^2}} \approx 0,2374$$

$$P(\texttt{Hours} = \text{"6,8"}|\texttt{Passed} = \text{"Yes"}) \approx \frac{1}{\sqrt{2\pi} \cdot 1,1339} \cdot e^{-\frac{(6,8-4,4286)^2}{2 \cdot 1,1339^2}} \approx 0,0395$$
$$P(\texttt{Hours} = \text{"6,8"}|\texttt{Passed} = \text{"No"}) \approx \frac{1}{\sqrt{2\pi} \cdot 1,5275} \cdot e^{-\frac{(6,8-2,3333)^2}{2 \cdot 1,5275^2}} \approx 0,0036$$

3. **Calculate the likelihood of each tuple:**

   a) **Tuple T_1 with *Clustering, Medium, 4*:**

$$\begin{aligned} P(T_1|\texttt{Passed} = \text{"Yes"}) &= P(\texttt{Topic} = \text{"Clustering"}|\texttt{Passed} = \text{"Yes"}) \\ &\cdot P(\texttt{Knowledge} = \text{"Medium"}|\texttt{Passed} = \text{"Yes"}) \\ &\cdot P(\texttt{Hours} = \text{"4"}|\texttt{Passed} = \text{"Yes"}) \\ &\approx 0,4286 \cdot 0,4286 \cdot 0,3276 \\ &\approx 0,0602 \end{aligned}$$

$$\begin{aligned} P(T_1|\texttt{Passed} = \text{"No"}) &= P(\texttt{Topic} = \text{"Clustering"}|\texttt{Passed} = \text{"No"}) \\ &\cdot P(\texttt{Knowledge} = \text{"Medium"}|\texttt{Passed} = \text{"No"}) \\ &\cdot P(\texttt{Hours} = \text{"4"}|\texttt{Passed} = \text{"No"}) \\ &\approx 0,3333 \cdot 0 \cdot 0,1440 \\ &\approx 0 \end{aligned}$$

b) **Tuple T_2 with *Classification, High, 3*:**

$$\begin{aligned} P(T_2|\texttt{Passed} = \text{"Yes"}) &= P(\texttt{Topic} = \text{"Classification"}|\texttt{Passed} = \text{"Yes"}) \\ &\cdot P(\texttt{Knowledge} = \text{"High"}|\texttt{Passed} = \text{"Yes"}) \\ &\cdot P(\texttt{Hours} = \text{"3"}|\texttt{Passed} = \text{"Yes"}) \\ &\approx 0,2857 \cdot 0,2857 \cdot 0,1591 \\ &\approx 0,0130 \end{aligned}$$

$$\begin{aligned} P(T_2|\texttt{Passed} = \text{"No"}) &= P(\texttt{Topic} = \text{"Classification"}|\texttt{Passed} = \text{"No"}) \\ &\cdot P(\texttt{Knowledge} = \text{"High"}|\texttt{Passed} = \text{"No"}) \\ &\cdot P(\texttt{Hours} = \text{"3"}|\texttt{Passed} = \text{"No"}) \\ &\approx 0,3333 \cdot 0,6667 \cdot 0,2374 \\ &\approx 0,0528 \end{aligned}$$

c) **Tuple T_3 with *Frequent Patterns, Low, 6.8*:**

$$\begin{aligned} P(T_3|\texttt{Passed} = \text{"Yes"}) &= P(\texttt{Topic} = \text{"Frequent Patterns"}|\texttt{Passed} = \text{"Yes"}) \\ &\cdot P(\texttt{Knowledge} = \text{"Low"}|\texttt{Passed} = \text{"Yes"}) \\ &\cdot P(\texttt{Hours} = \text{"6.8"}|\texttt{Passed} = \text{"Yes"}) \\ &\approx 0,2857 \cdot 0,2857 \cdot 0,0395 \\ &\approx 0,0032 \end{aligned}$$

$$\begin{aligned} P(T_3|\texttt{Passed} = \text{"No"}) &= P(\texttt{Topic} = \text{"Frequent Patterns"}|\texttt{Passed} = \text{"No"}) \\ &\cdot P(\texttt{Knowledge} = \text{"Low"}|\texttt{Passed} = \text{"No"}) \\ &\cdot P(\texttt{Hours} = \text{"6.8"}|\texttt{Passed} = \text{"No"}) \\ &\approx 0,3333 \cdot 0,3333 \cdot 0,0036 \\ &\approx 0,0004 \end{aligned}$$

4. **Determine the highest posteriori probability for each tuple:**

   The posteriori probability according to Bayes' theorem is actually calculated as follows:

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}.$$

   Where C_i stands for the class to be predicted and X is representing a specific tuple (resp. the attribute combination that is part of that tuple).

   However, since $P(X)$ is the same for all classes, it is sufficient to calculate only the numerators to determine the highest $P(C_i|X)$.

   Being able to determine the highest $P(C_i|X)$ (even without knowing its exact value) is sufficient to classify the tuple.

a) **Tuple T_1 with *Clustering, Medium, 4*:**

Calculate the numerator of $P(\texttt{Passed} = \text{"Yes"}|T_1)$ and $P(\texttt{Passed} = \text{"No"}|T_1)$:

$$P(T_1|\texttt{Passed} = \text{"Yes"}) \cdot P(\texttt{Passed} = \text{"Yes"}) \approx 0,0602 \cdot 0,7 \approx 0,0421$$
$$P(T_1|\texttt{Passed} = \text{"No"}) \cdot P(\texttt{Passed} = \text{"No"}) = 0 \cdot 0,3 = 0$$

Since $0,0421 > 0$ and we classify the tuple T_1 as $\texttt{Passed} = \text{"Yes"}$.

Calculation of the posteriori probability:

Even if the calculation of the full posteriori probability is not necessary, it is still possible to calculate it.

We first need to calculate the denominator of the posteriori probability $P(X)$:

$$\begin{aligned} P(T_1) &= P(T_1|\texttt{Passed} = \textit{"Yes"}) \cdot P(\texttt{Passed} = \textit{"Yes"}) \\ &\quad + P(T_1|\texttt{Passed} = \textit{"No"}) \cdot P(\texttt{Passed} = \textit{"No"}) \\ &\approx 0,0602 \cdot 0,7 + 0 \cdot 0,3 \\ &\approx 0,0421 \end{aligned}$$

Which can then be used to calculate the posteriori probabilities:

$$\begin{aligned} P(\texttt{Passed} = \textit{"Yes"}|T_1) &= \frac{P(T_1|\texttt{Passed} = \textit{"Yes"}) \cdot P(\texttt{Passed} = \textit{"Yes"})}{P(T_1)} \\ &= \frac{0,0421}{0,0421} \\ &= 1 \end{aligned}$$

$$\begin{aligned} P(\texttt{Passed} = \textit{"No"}|T_1) &= \frac{P(T_1|\texttt{Passed} = \textit{"No"}) \cdot P(\texttt{Passed} = \textit{"No"})}{P(T_1)} \\ &= \frac{0}{0,0421} \\ &= 0 \end{aligned}$$

As this calculation is not necessary for the classification, we will not calculate the posteriori probabilities for the other tuples.

b) **Tuple T_2 with *Classification, High, 3*:**

Calculate the numerator of $P(\texttt{Passed} = \text{"Yes"}|T_2)$ and $P(\texttt{Passed} = \text{"No"}|T_2)$:

$$P(T_2|\texttt{Passed} = \text{"Yes"}) \cdot P(\texttt{Passed} = \text{"Yes"}) \approx 0,0130 \cdot 0,7 \approx 0,0091$$
$$P(T_2|\texttt{Passed} = \text{"No"}) \cdot P(\texttt{Passed} = \text{"No"}) \approx 0,0528 \cdot 0,3 \approx 0,0158$$

Since $0,0091 < 0,0158$ we classify the tuple T_2 as $\texttt{Passed} = \text{"No"}$.

c) **Tuple T_3 with *Frequent Patterns, Low, 6.8*:**

Calculate the numerator of $P(\texttt{Passed} = \text{"Yes"}|T_3)$ and $P(\texttt{Passed} = \text{"No"}|T_3)$:

$$P(T_3|\texttt{Passed} = \text{"Yes"}) \cdot P(\texttt{Passed} = \text{"Yes"}) \approx 0,0032 \cdot 0,7 \approx 0,0022$$
$$P(T_3|\texttt{Passed} = \text{"No"}) \cdot P(\texttt{Passed} = \text{"No"}) \approx 0,0004 \cdot 0,3 \approx 0,0001$$

Since $0,0022 > 0,0001$ we classify the tuple T_3 as $\texttt{Passed} = \text{"Yes"}$.

Task 2: Model Evaluation

The classifier was also trained on a version of dataset D with more tuples:

The dataset T contains both the true and the predicted "Passed"-Status for each test tuple.

Topic	Knowledge	Hours	Passed (True)	Passed (Pred)
Classification	Medium	7,5	Yes	Yes
Frequent Patterns	Low	1,8	No	No
Frequent Patterns	High	3,7	No	Yes
Frequent Patterns	Low	0,2	No	No
Frequent Patterns	High	1,4	Yes	No
Frequent Patterns	High	9,9	Yes	Yes
Frequent Patterns	Medium	7,3	Yes	Yes
Frequent Patterns	Low	4,3	No	Yes
Classification	Medium	5,5	Yes	Yes
Clustering	Low	0,1	No	No

Use the dataset T to calculate the **sensitivity**, **specificity**, **accuracy**, **precision**, **recall**, and **F1-score** of the model.

Also state the **best possible** value for each metric.

We need to calculate the True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for the model evaluation.

This is often done by creating a confusion matrix:

		Predicted		
		Yes	*No*	Total
True	*Yes*	4 (TP)	1 (FN)	5 (P)
	No	2 (FP)	3 (TN)	5 (N)
	Total	6 (P')	4 (N')	10 (P + N)

This confusion matrix can be used to calculate the metrics:

- **Sensitivity:**

$$\text{Sensitivity} = \frac{TP}{P} = \frac{4}{5} = 0,8$$

  Best possible value: 1

- **Specificity:**

$$\text{Specificity} = \frac{TN}{N} = \frac{3}{5} = 0,6$$

  Best possible value: 1

- **Accuracy:**

$$\text{Accuracy} = \frac{TP + TN}{P + N} = \frac{4 + 3}{5 + 5} = \frac{7}{10} = 0,7$$

  Best possible value: 1

- **Precision:**

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{4}{4 + 2} = \frac{4}{6} = 0,6667$$

  Best possible value: 1

- **Recall:**

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{TP}{P} = \text{Sensitivity} = 0,8$$

  Best possible value: 1

- **F1-Score:**

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \cdot \frac{0,6667 \cdot 0,8}{0,6667 + 0,8} = 0,7273$$

  Best possible value: 1

Of course, the question regarding the best possible value is a trick question. The best possible value for each metric is 1 (or 100%), as this would mean 100% of the tuples were classified correctly.

Exercise 3: Conducting Classification

This exercise comprises practical data science tasks and thus utilizes a Jupyter Notebook:

1. Open `Conducting-Classification.ipynb`.
2. Take a look at the tasks (blue boxes) in the notebook and try to solve them.

If you are unfamiliar with how to open a Jupyter Notebook, please refer to Exercise 1 of `1-Introduction-Python-Pandas.pdf`.

The solution to the exercise can be found in `Additional-Files-Solution.zip`.