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.

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.

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?

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.

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.

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.

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`.