
1. Prologue

Knowledge Discovery in Databases with Exercises

Dominik Probst, dominik.probst@fau.de

Computer Science 6 (Data Management), Friedrich-Alexander-Universität Erlangen-Nürnberg

Summer semester 2025



Dominik Probst, M.Sc.

Lecturer, Tutor and Primary Contact

- Ph.D. candidate @CS6 (Data Management)
- E-Mail: dominik.probst@fau.de
- Website: <https://www.cs6.tf.fau.eu/dp>

Lucas Weber, Dipl.Ing.

Tutor

- Ph.D. candidate @CS6 (Data Management)
- E-Mail: lucas.weber@fau.de
- Website: <https://www.cs6.tf.fau.eu/lw>



- **Knowledge Discovery in Databases with Exercises (KDDmUe) - 5 ECTS**
 - B.Sc./M.Sc. Data Science
 - B.Sc./M.Sc. Computer Science
 - M.Sc. International Information Systems
 - M.Sc. Medical Engineering
 - M.Sc. Information and Communication Technology
 - Possibly other courses of study (clarify with your examination office)

Important: The „Lecture Only“ Is No Longer Offered!

Until SS2023, we offered the module KDD (2.5 ECTS - only lecture) for some degree programmes. This module is no longer offered!

- **Useful prerequisites:**

- Successful completion of the module „Einführung in Datenbanken“ (EDB)
 - Or a similar course teaching the basics of databases and SQL
- Experience with:
 - Python
 - Jupyter Notebooks
 - Numpy
 - Pandas
 - Algorithms
 - Data structures

- **Goal of the module:**

- Introduce you to the principles of data mining.
⇒ This is the core of knowledge discovery in databases

- **Topics in the lecture:**

- | | |
|---------------------|-----------------------------|
| 1. Introduction | 5. Mining Frequent Patterns |
| 2. Data | 6. Classification |
| 3. Preprocessing | 7. Cluster Analysis |
| 4. Data Warehousing | 8. Outlier Analysis |

- **Topics in the exercise:**

- | | |
|---|-------------------|
| 1. Introduction to python and pandas (optional) | 4. Classification |
| 2. Data Analysis and Data Preprocessing | 5. Clustering |
| 3. Frequent Patterns | 6. Outlier |

- **Topics in the submissions:**

- | | |
|----------------------|---------------|
| 1. Frequent Patterns | 3. Clustering |
| 2. Classification | |

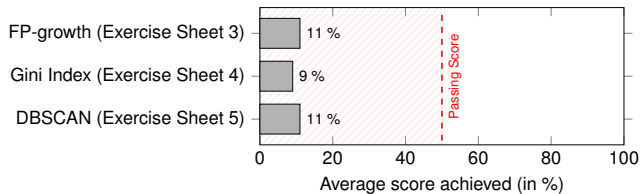
- **Exercise sessions (in presence):**
 - Working together on exercise sheets
 - Either practical data science tasks or theoretical exercises
(varies depending on the exercise sheet)
 - Required tools:
 - Laptop capable of running Jupyter Notebooks
(preferably one per person, one per small group is also possible)
 - Expected preparation:
 - Good understanding of the lecture content
 - Completed "Preparation" section (see exercise sheets)

- **Last semester:**

- Very low attendance in the exercise sessions (below 5% of registered students)
⇒ Tasks based on the exercise sheets with very low point averages in the exam

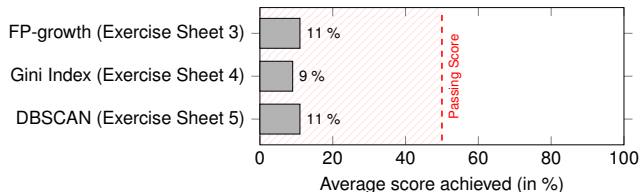
- **Last semester:**

- Very low attendance in the exercise sessions (below 5% of registered students)
⇒ Tasks based on the exercise sheets with very low point averages in the exam



- **Last semester:**

- Very low attendance in the exercise sessions (below 5% of registered students)
⇒ Tasks based on the exercise sheets with very low point averages in the exam



Important: The Exercises Are Crucial For The Exam!

While the lectures give you a basic theoretical overview, the exercises go into more detail, give you some more practical insights, and will be the basis of multiple questions in the exam!

- **Programming tasks (to be done at home):**

- Implementation of individual algorithms known from the lecture for a deeper understanding
 - Programming language: Python
 - Topics: Frequent Patterns, Classification, and Clustering
- Have to be submitted to a GitHub classroom
 - Calculation of points performed automatically after each push
 - Improvements possible at any time until the submission deadline
- Work in small groups (up to three persons) is permitted
 - We conduct random checks for plagiarism across groups
 - ⇒ In cases where plagiarism is detected, all groups involved will receive zero points.

Participating In Or Passing The Submissions Is Not Mandatory!

However, if you get at least 50% of the points, you will receive access to a mock exam with solutions. Also some exam questions will be easier to solve if you have completed the submissions.

- **Knowledge Discovery in Databases with Exercise (KDDmUe) - Written Exam**
 - Duration: 90 minutes
 - Questions about lecture, exercise and submission content
 - Language: English

Important: Do Not Forget To Register!

Without exception¹, we can only examine participants who have also registered for this exam at the examination office.

- **(Provisional) Registration Period:** June 2 (00:01) until June 22, 2025 (23:59)
- **Stay updated at:** https://www.cs6.tf.fau.de/kdd/reg_date

¹ If you have to retake the exam, you also have to register manually for the exam **during** the registration period. There is no automatic registration for retakes anymore!

- **This lecture is based on the book by Han et al.:**

- J. Han et al., *Data Mining: Concepts and Techniques*, 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011, ISBN: 0123814790
- [Copies are available at the Science and Technology Branch Library \(TNZB\).](#)
- Lecture slides are based on slides provided by Jiawei Han with modifications by Prof. Dr.-Ing. Klaus Meyer-Wegener, Luciano Melodia, and Melanie Sigl.
- Lecture slides have been modified and extended a lot since then.

- **Further books on this topic include, but not limited to:**

- A. Ge'ron, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*, 2nd ed. O'Reilly Media, 2017, ISBN: 978-1491962299
- H. Du, *Data Mining Techniques and Applications: An Introduction*. Cengage Learning EMEA, May 2010, p. 336, ISBN: 978-1844808915
- I. H. Witten et al., *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*, 4th. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016, ISBN: 0128042915

- **Lecture** - Start in Calendar Week 17 (Today)
 - Wednesday, 14:15 - 15:45 (H20)
Lecturer: Dominik Probst
- **Exercises** - Start in Calendar Week 18
 - Group 1: Tuesday, 10:15 - 11:45 (H18)
Tutor: Lucas Weber
 - Group 2: Wednesday, 08:15 - 09:45 (H4)
Tutor: Dominik Probst

Registration For The Exercises

Registration for exercises is mandatory to ensure an appropriate support to questions regarding setting of exercises should they arise. **Registration opens at April 23th, 16:00 via StudOn.**

Calendar Week	Lecture	Exercise	Submission
17	Prologue + Introduction		
18	Data	Introduction to Python & pandas (optional)	
19	Preprocessing	Data Analysis & Data Preprocessing	
20			
21			
22	Frequent Pattern		
23		Frequent Pattern (Part 1)	
24	Classification		
25		Frequent Pattern (Part 2)	Frequent Pattern
26	Cluster Analysis	Classification	
27			Classification
28	Outlier Analysis	Clustering	
29			
30	Exam Q&A		Clustering

- Register at:
<https://www.studon.fau.de/studon/go/crs/6151916/rcodedTkQDYDgP8>
- Main source for resources. E.g.:
 - Lecture slides
 - Exercise sheets
 - Submission sheets
 - Forum
- Membership required to receive important updates on KDD
- Questions should be asked here (StudOn Forum)



- Public repository at: <https://github.com/FAU-CS6/KDD>
- Version control of our resources including:
 - Lecture slides
 - Exercise sheets
 - Submission sheets




Help Appreciated: Error Corrections

Even though we strive for error-free lecture slides and practice sheets, there is still the possibility that errors have slipped in. You can help us mitigate these inaccuracies: Mail us, or better yet, in the case you have a GitHub account, open up a GitHub issue or create a pull request. Any pointers to errors are very much appreciated.

Any questions about this chapter?

Ask them now or ask them later in our forum:



 https://www.studon.fau.de/studon/goto.php?target=1code_OLYeD79h