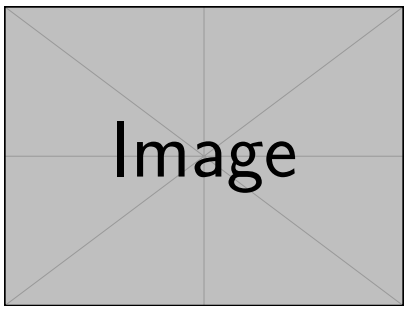
Autumn semester 2025



Data-Mining in Environmental Science

Learning diary

Date:	November 14, 2025
Author:	Stephen Weybrecht
Student number:	2505376
Supervisor:	Mikko Kolehmainen

Contents

1. Orientation

My personal background is that I am an exchange student in physics, studying at the University of Eastern Finland for the autumn semester. As such I have quite a strong background in programming, especially in Python, statistics and machine learning already. As I am quite interested in these topics and my home university offers a rather flexible study plan, I further deepened my knowledge by choosing multiple electives and a Bachelor thesis topic that were deeply connected with data analysis already. Although I expect that many things in the beginning of the course will be topics I already learned, I am very much looking forward to developing a deeper understanding of Data mining and seeing this done in a context I have no previous knowledge yet – namely Environmental science. Looking at the curriculum, there are also many topics I have had no prior experience in which is quite exciting to me. In summary, I expect this course to build nicely on my previous knowledge while additionally providing interesting insights to the field of Environmental science.

As a physics student, I am very used to writing scientific paper-like reports. This is the idea behind many reports of practicals I already needed to write as well as my Bachelor thesis. In these the expression of ones own opinion is actively discouraged. I would even go further and say that conciseness and scientific correctness are virtues hammered into us for years during our studies. Naturally a Learning diary such as this where the expression of a personal opinion and a critical reflection about the topics learned is not only encouraged but actively required is therefore quite a step out of my comfort zone. Still, I am looking forward to experiencing this new concept and seeing how it will shape my learning experience. At least at the time of starting this course this integrated approach of always putting learned things in ones own context, thinking critically and still performing quantitative task during the exercises seems like a very natural way to learn. It will be quite interesting to see how this will change during the course.

My future job prospects as a physicist will most likely revolve about programming and handling large amounts of data, regardless of whether I will pursue a career in industry or I will stay in academia. Jobs as a Data Scientist, Programmer or in the engineering direction are quite common when getting a Masters in physics and experimental physics in academia has mostly been computing, simulation or the analysis of huge amounts of data since many years already. Therefore, having a strong basis in programming, data analysis and visualization are skills one should have after the studies. I expect that this course will deepen my knowledge in Data Mining by not only introducing new concepts but also connecting those learned already on an even deeper level and will thus be a helpful resource for my future.

I will use Large Language Models in the following mainly for getting code suggestions for the exercises and help in the layout of this report (as LaTex can be rather cumbersome at times). The text will mainly be written by myself, although sometimes AI is used for translation and paraphrasing purposes. All code of the exercises as well as the LaTex files to create this report will be made available on a public GitHub repository [1].

2. Introduction and basics

3. Environmental data and its pre-processing

4. Data visualization

5. Clustering in Python

6. Predictive modeling

7. Summary

8. Self-evaluation

References

[1] Stephen Weybrecht. *Data Mining in Environmental Science: Repository.* `https://github.com/stewey0/uef-data-mining`. Source code. Accessed: November 14, 2025.