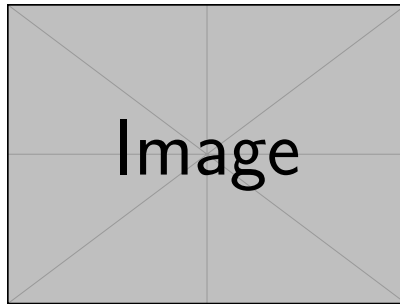


AUTUMN SEMESTER 2025



DATA-MINING IN ENVIRONMENTAL SCIENCE

---

LEARNING DIARY

---

<b>Date:</b>	November 18, 2025
<b>Author:</b>	Stephen Weybrecht
<b>Student number:</b>	2505376
<b>Supervisor:</b>	Mikko Kolehmainen

## Contents

<b>1. Orientation</b>	<b>3</b>
<b>2. Introduction and basics</b>	<b>4</b>
<b>3. Environmental data and its pre-processing</b>	<b>6</b>
<b>4. Data visualization</b>	<b>7</b>
<b>5. Clustering in Python</b>	<b>8</b>
<b>6. Predictive modeling</b>	<b>9</b>
<b>7. Summary</b>	<b>10</b>
<b>8. Self-evaluation</b>	<b>11</b>

# 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

### Lecture

The lectures this week dealt with introductory topics regarding the structure of this course, a math and statistics rehearsal as well as an introduction to the topic Data Mining. The math and statistics chapter covers many basic definitions like the axioms and basics properties of probabilities, the definitions of partial derivatives, vectors and matrices and their addition and multiplication properties. Although these are nice to have for completeness sake and should prove helpful for students which do not have a background in statistics yet, for me, they were already known and will therefore not be repeated here. Instead, I will focus on definitions of this chapter which I do not know by heart and which I think will prove useful for the following and will put them into context of what I have already learned.

The sample mean  $\hat{\mu}$  and sample variance  $\hat{\sigma}^2$  provide unbiased estimators for the population mean  $\mu$  and population variance  $\sigma^2$  given a certain sample  $v_i$  of size  $N$ , i.e.  $i \in (1, N)$ . They are defined as

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N v_i \quad (1)$$

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (v_i - \hat{\mu})^2 \quad (2)$$

I recall from a prior course on statistics that the sample variance with the  $N - 1$  term in the denominator should be used in the case of an *unknown* mean, i.e. if the mean is estimated by Equation 1, as it provides an unbiased estimator and therefore better convergence to the true but unknown sample variance for small  $N$ . If instead the mean is inferred through different means, the minus 1 term can be dropped. In `numpy` the sample variance can be simply calculated by setting the parameter `ddof=1`:

```
1 import numpy as np
2 # array is a sample array of data
3 sample_var = np.std(array, ddof=1)
```

Listing 1: Sample variance, calculated in numpy

If there are more than just one random variates a variance can be calculated for each one. Additionally, a so-called covariance between two different variated can also be calculated. Covariances and variances are summarized in a covariance matrix, whose elements are defined as follows:

$$\text{Cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^N (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_y) = \hat{\sigma}_x \hat{\sigma}_y \rho_{x,y} \quad (3)$$

Here  $\hat{\sigma}_i$  are the standard deviations of the variates  $x$  and  $y$  according to Equation 2 and  $\rho_{x,y}$  is the degree of correlation between  $x$  and  $y$ . It holds that  $\rho_{x,y}$  is always between  $-1$  and  $1$  with  $-1$  indicating maximum negative correlation,  $1$  maximum positive correlation and  $0$  no correlation at all. The covariance matrix (or the reduced correlation matrix obtained when

removing all individual standard deviations) therefore indicates correlations between different parameters in a dataset which can be used for explorative data analysis. Another use of it is when fitting a model to a dataset. Here a large degree of correlation between model parameters indicates overfitting meaning a surplus of model parameters.

### **Exercise**

### **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 18, 2025.