DSAI (CU75074V2) Course Outline v0.1

Schedule & Information

HBO-ICT
2024-2025

UNIVERSITY OF APPLIED SCIENCES

Table of contents

COURSE OUTLINE DSAI 2024-2025

ABOUT

This course outline (syllabus) contains all practical information about the core course Data Science/AI in block 4 of the second year. It contains schedules, topics that are discussed, information about the project(s), grading, learning outcomes, required and recommended material, and where to find other materials. You will find what is expected of you in here, and what you can expect of us.

Use this outline as a manual. If you have any organizational questions, as the answer is likely in here. For questions about the material, please consult the required and recommended materials first and attend the lectures. Some of these concepts might seem daunting at first, but fright not: we are here to help you understand and learn all about data science, and we hope you will enjoy learning all about this exciting field.

We expect you to be self-sufficient when it comes to organizing your studies (which is why we created this outline), but if you are struggling with understanding something, please reach out to us for help. Our approach is a reciprocal one: if you do your part by showing up to class and the check-ins, do the recommended reading, and participate in class – we will do our part and help you as much as we can. If you do not show up or leave everything to the last minute, it is unlikely we can still help you at that point. In other words: we show up for you to the same extent you show up for us and your teammates. You can reach us for questions about DSAI during the office hours as listed below, either in person or on Teams, and during the workshops (theory + practical) sessions. Outside of these days, you might have to wait until the next office hour day for a response.

Lecturer	Regarding	Office hour days	Hours
Sophie van der Bliek	Theory	Monday, Wednesday	09:00-17:00
Michiel Veen	Theory, Python	Tuesday (project days), Wednesday, Thursday	09:00-17:00
Remco Kok	Theory, Python	Monday	During class
Rimmert Zelle	Organisational questions		

WHAT IS DATA SCIENCE?

Data science and AI is all around us nowadays – from targeted advertisement and content in online shopping, streaming services and social media, to self-driving cars and facial recognition. You are likely using several applications of DS/AI in your daily life or for school, with many tools available to us such as ChatGPT or tailored recommendations YouTube gives you. Almost all of these applications we know and use, work based on making predictions from data.

In this semester, you will learn the fundamentals of how DS/AI works, and how you can use data to help you solve many problems from the technical side, as well as a business perspective. To do this, we start with understanding what machine learning is (and what it is not), and you finish the semester with a project where you will do predictive modeling yourself in a group project as well as a written test.

If you are interested in learning more after this semester, you may choose Data Science as a track within the ICT program. In this track, you get to learn more about deep learning if you choose the minor, ethical implications of using data and developing AI, and putting everything into practice in internships and projects. But first, let's start with the basics.

COURSE COMPONENTS

COURSE DESCRIPTION

This course is a core course of the HBO-ICT programme at HZ, as it is one of the main tracks students can choose as their specialization (Data Science or Software Engineering). If a student chooses the data science track, this course will be the foundation for their 3rd year internship, potentially their minor and 4th year projects and internship. In other words, what you learn in this course is the basis and **essential** for your further studies and career in Data Science.

This knowledge will be tested in an **individual** written knowledge test, as well as a **group** project report. **You must pass both to pass the course.** The written knowledge test will contain multiple choice questions. The questions will test your knowledge about understanding and applying the concepts taught in the lectures.

STUDY LOAD

This block consists of the following courses:

Course name	Course Code	EC	Core/Elective
Data Science/AI (DSAI)	CU75074V2	7.5	Core
Cloud Computing (CCO) *choice for SE track*	CU75028V2	5	Elective
Data Visualization (DVI) *choice for DS/BIC track*	CU75027V3	5	Elective
PPD-A	CU75082V1	7.5	Core

The ECTS system dictates that 1 EC is equivalent to a study load of about 28 hours.

For DSAI, the classes will consist of workshops, where lectures and practicals are combined in a block of 3 hours on Mondays. There will be check-ins on the Wednesday for 45 minutes per group pair, in which you will present your progress to another group under the supervision of lecturers. With a study load of 7.5 ECTS for 10 weeks, including workshop classes of three hours a week and the project check-ins of approximately an hour, that means **you will be expected to do ~17 hours of self-study** per week. This includes preparation for the quiz and working on the project.

TOPICS RELATED TO LEARNING OUTCOMES – LECTURES

The learning outcomes (LOs) are defined in the UR and dictate what you must master at the end of the course to complete it successfully. The LOs are the blueprint for the lectures, practicals and every component you will be graded for: the quiz, the project and presentation. The LOs are mapped to the lectures, practical sessions and Python sessions on the course schedule.

The learning outcomes for this block are as follows:

DSAI
Learning outcome 1: You set up a data Science process
Indicator 1.1: You describe data mining activities based on choice of a basic machine learning model and relevant required activities
Indicator 1.2: You define data mining success criteria
Indicator 1.3: You add extra self-organised and/or external data sources to the data science process
Learning outcome 2 You collect and address relevant data
Indicator 2.1: You (re-)validate data after model generated assumptions
Indicator 2.2: You integrate relevant data by merging multiple data sources
Indicator 2.3: You clean data by imputing and scaling relevant data
Indicator 2.4: You construct data by one-hot-encoding, defining targets & labelling relevant data
Indicator 2.5: You convert data formats as prerequisite for relevant model technique(s)
Learning outcome 3: You perform data analysis
Indicator 3.1: You split data into test & train sets to generate a test design
Indicator 3.2: You build & train relevant model technique(s) and create predictions using the model technique(s) on test data set
Indicator 3.3: You assess the model(s) on chosen metrics of the defined success criteria
Learning outcome 4: You evaluate & deploy results of the data science process
Indicator 4.1: You evaluate and match success criteria with business objectives of the data science process
Indicator 4.2: You determine next steps and setup an advisory report for follow-up
Indicator 4.3: You produce a deliverable for customer
Indicator 4.4: You review the data science process and collect lessons learned on process & product

WORKSHOPS: LECTURES & PRACTICAL SESSIONS COMBINED

The method of teaching for this block will be workshops, which will consist of lectures and practical sessions combined. The theory will be explained first, and you will apply this knowledge to a case study and dataset with guidance of lecturers immediately after each 'chunk' of theory. The purpose of this format is to make the most of your working memory, keep you engaged with the material and to keep your mind in motion instead of only passively sitting and trying to absorb theory. It is therefore very important to always attend class unless you are ill, because missing class means you will both miss theory and practical sessions. We will be taking attendance as well.

For the check-ins, we expect you to show your latest progress in a short presentation. Our advice is to make a PowerPoint, and simply add a few slides to it every week with the progress you have made. Briefly present what you did with your group, and focus on *explaining* your choices and *interpreting* what you have produced – what does it mean, and how does it impact next and past steps? What does it mean in the business context? You will receive immediate feedback from the other group you present for, and from the lecturer. After both groups have presented and given/received feedback, show us what you have written down in your report every week. We can give you feedback on that as well.

PROJECT & PORTFOLIO

For this block, you will do a data science project in a group of five students. You will be assigned a dataset, and you will have to write a report on your findings. For details how you should format and write this deliverable, please refer to the chapter about the project report in this outline for more information. Read this section carefully and follow the instructions.

This project will be about using data to create a digital twin of a wastewater treatment plant in Ede. Your client in this case will be Lectoraat Data Science (Data Science research group) at HZ, who were part of this project. During the first lecture, Bente Sinke, the research coordinator and researcher at the Data Science research group, will introduce the context of the project and dataset to you. Make sure you attend and ask her your questions during that session. Afterwards, a forum will be opened on Learn on which you can ask her questions at set moments during the course.

GRADING

The following tables refer to the grading for DSAI. You have pass (get at least 5.5) both the quiz and the project.

The assessment overview is follows:

BLOCK 4 – DS/AI

Graded element	Description	Total % of final grade
Written test (quiz)	This written test of 40 multiple choice questions with 4 possible answers is about all the theory presented in the lectures. There will be no Python-specific questions. The test will contain multiple choice questions. The questions are designed to **apply** the knowledge obtained from the lectures to test your **understanding**.	60% **Note that you must get at least 5.5 for this test to pass the course, for which the threshold is set at answering 67% of the questions correctly[1].**
Project report	Group project of 5 students where you will conduct a data science project. A data set will be provided to your group. Follow the instructions about the project portfolio in this outline carefully.	40% **Note that you must get at least 5.5 for the project to pass the course.**

PLAGIARISM AND USE OF CHATGPT

Any cases of plagiarism where you have copied work from other students or did not cite your sources properly, will be reported to the Fraud Committee. See the writing guide for more advice.

The new HZ policy point 2.3 dictates that use of AI is not permitted unless specifically stated otherwise by the lecturer. **That means that for the project, you will not be allowed to hand in or present anything generated by generative AI, such as ChatGPT or DeepSeek. If you use generative AI to hand in work for a grade, you will be reported to the Fraud Committee.** GPTs can be (and often are) incorrect about concepts related to data science, especially when it comes to interpretation and application to your project. We can not keep you from using them to study independently (such as in preparation for the quiz), but use AI tools with caution, and do not rely on the tools to do all the work for you.

[1] *The guess rate for this exam is 40/4=10. With a requirement of 56% knowledge, this results in a (40-40/4)*0.56+40/4 ≈ 26.8 questions correct. 26.8/40 = 67% of questions correct for a 5.5.*

MATERIALS

In this section, you can find all class materials used. The required material comprises everything you are required to master for the written knowledge test and project. We do our best to explain everything to you the best we can, but as data science is inherently quite theoretical, some of these concepts can be quite dry and difficult to sink it at once. It can be very helpful to see these concepts explained in several different ways. Therefore, I compiled a list of recommended materials with some resources that might help you understand. If you could still use some clarifications, please do not hesitate to contact us.

REQUIRED MATERIAL

The material you are required to know for the test and the reports is discussed during class. Note that your knowledge of Python will not be on written knowledge test, but we do expect you to be able to explain your code during the projects.

RECOMMENDED MATERIAL

People often have different preferences in which medium the material is presented, so I have compiled a list of different 'mediums' of conveying the concepts discussed in class. I tried to select only the most entertaining and intuitive videos/articles/sheets I could find. I also added book pages/chapter that explain the concepts in class again. Note that many books about machine learning are more complex than what you strictly need for the quiz – use the slides and practice questions as a guide of what level of depth is required.

Note that you do not have to master **all** the material that is discussed in these videos/books/articles (because we did not make them), **nor is this material exhaustive on what you do have to know**. **The materials presented in this table are only meant to help you understand the required material.** Doing your own research and being proactive in understanding the material is encouraged as well, but these materials have been checked on correctness by us, and we deem it appropriate for your level.

The books in the table below are sorted by difficulty (and therefore detail), with the most accessible book first. Note that I have put some recommended chapters in the "book" column, but these chapters and subchapters are **not exhaustive** by any means regarding these books. We assume you know how to use a book index to read more about the specific concepts you wish to study.

1. ItML: "Introduction to Machine Learning with Python" Müller, A., Guido, S. (2017) O'Reilly.
2. MLwP: "Machine Learning with PyTorch and Scikit-Learn" Raschka, S., Yuxi, L., Mirjalili, V. (2022). Packt Publishing.
3. ISL: "An Introduction to Statistical Learning with Applications in Python" James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). New York Springer. Download here
   Note: this a university level book (it is considered the gold standard for undergraduate level statistical learning) and very mathematical. You will not be expected to understand models at the level presented in this book. This book might be helpful to you if you have a strong background in mathematics, or if you simply find it interesting to understand something at a deeper level. I still included it because it is very detailed, also has intuitive examples and I know some students find that useful or like the challenge.

Weeks	Videos	Articles	Books
Week 1 - BU			
Week 2 (DU/DaP)	- Data Understanding: distributions, skew, their relationship to visualization - Data Understanding: Correlation (negative and positive), regression coefficient, scatter plots, implications of correlation and causality	- Summary for basic visualizations: when to choose what - Visual cheatsheet for visualizations	DU: MLwP: §4 MLwP §9: p. 274
Week 3 (DU-DaP)	- Data Understanding: Boxplots and outliers	- Encoding (one-hot, dummy, ordinal)	MLwP: Encoding §4: p. 111-116 Scaling: §4: p119- ItML: Encoding §3: p.131-139
Week 4 (Modeling: regression)	- Modeling: (Linear) Regression, residuals, error - Metrics: Regression and classification	- Modeling: Supervised/ unsupervised models cheat sheet	MLwP: §9 ItML: §2, §5: p.275-299
Week 5 (Modeling: classification)	- Modeling: Supervised learning (logistic regression, LDA, kNN) prediction, accuracy, train-test, confusion matrix (true/false positives), classifier	- Modeling: performance metrics explained/cheat sheet	ItML: §2
Week 6 (Modeling: unsupervised, hyperparameter tuning, over/underfitting)	-Modeling: Unsupervised learning, k-means, (hierarchical) clustering	- General ML cheatsheet: metrics, model selection, cross-validation, regularization, bias/variance trade-off - Modeling: Supervised/ unsupervised models cheat sheet	ItML: §3, 5
Week 7 (Evaluation)			

DDD REFRESHER

Contents	Videos	Articles/reading
Fundamentals: DU/DaP Theory BLOCK 7 refresher	- Measures of central tendency: mean, mode, median - Measures of spread: variability, spread, range, standard deviation -Data Understanding: qualitative data, quantitative data, histograms, bar charts, pie charts, binning – what & when to use them	- Laerd Statistics: measures of centrality: mean, mode, median (when to use what) - Laerd Statistics: Measures of spread - Summary for basic visualizations: when to choose what

PROJECT REPORT

You will form a group of five students. You will receive a dataset, conduct your research and write a report about the process, following the project template.

Report requirements:

Use the template for the projects on Learn. Note that your report should follow this structure exactly. Use the writing guidelines on Learn to ensure your report is properly written and formatted. See the rubric for what needs to be included.

Ensure that the report adheres to the following prerequisites (knock-out criteria), as also stated in the rubric. **If the work does not meet the prerequisites, the work is automatically graded with a 1,0.** Additionally, the learning objectives will not be graded, **and you will not receive feedback**.

	PREREQUISITES
☐	The portfolio follows the portfolio structure, content and guidelines as described in the portfolio template
☐	The content of the portfolio is understandable in terms of spelling, grammar, sentence structure and text ordering
☐	The portfolio is **complete**, all sections contain a serious attempt, and there is no use of generative AI.
☐	The **entire** group (whose names are on the project report that is handed in) has presented during check-ins.
☐	The content of the portfolio matches the content of the Python notebooks.

- For **DSAI**: the actual content of the final report is **20 pages maximum.** The appendix may be 20 pages maximum. The content counts from the start of your introduction to the last sentence of your last section (conclusion). Note that this is 20 pages maximum - you do not have to fill them up by writing more than necessary - which you probably do not need if you only do two iterations (i.e. a simple benchmark and a model).
- Follow the guidelines in the Writing Guidelines regarding appendices, bibliography and formatting.
- **Only use screenshots of your graphs. Make proper tables for the other components of data understanding (e.g. descriptive statistics, formulas). That is to say – no screenshots of code output.** Make your document look neat and professional.
- If you want to show some code to provide evidence towards fulfilling the LOs, put your code in an appendix. You may either make separate entries within that appendix with the relevant code snippets, or put your code in its entirety in one appendix and refer to the relevant lines within the text. Put the code in monospace font to discern it from regular text. Use the Word plugin for code to do this. **This means also not putting in screenshots of your code.**
- You take responsibility for paraphrasing correctly (or quoting if necessary) and the ethical use of sources, including accurate source referencing following APA guidelines. Use the standard Word functionality under References for this.

COURSE SCHEDULE

Planning of **DSAI** lectures, practical sessions, exams and deadlines. Schedule is subject to change. Outside of the project days, you are responsible for scheduling time with us yourself. Make sure to keep track of the scheduling app for any last-minute changes (e.g. rooms).

Block week	Date/time/location	Lecturer	Title	Learning outcome/ indicator	Contents
1	Tuesday 22-04-2025 09:00 – 12:00 (NL) GW110 13:00 – 16:00 (INT) GW317	SvdB (+BS)/MV	Introduction to DS/AI + Business Understanding	1.1/1.2/1.3	Introduction + Recap: choosing success criteria based on model-specific metric; planning data preparation
1	Wednesday 23-04-2025	SvdB/MV	Check-ins (see schedule on Learn)		Present your business question, main- and subquestions
2	Tuesday 06-05-2025 (NL) GW402 09:00-12:00 (INT) GW402 13:00-16:00	SvdB/RK	Data Understanding/ Data Preparation – theory class & practical workshop	1.3, 2.1	Assumptions, distributions, scaling, log transforms, linearity, correlations
2	Wednesday 07-05-2025	SvdB/MV	Check-ins (see schedule on Learn)		Present your first data understanding: distributions of your variables, linearity, correlations
3	Monday 12-05-2025 (INT) GW319 09:00-12:00 (NL) GW111 13:00-16:00	SvdB/RK	Data Understanding/ Data Preparation – theory class & practical workshop	2.2-2.3	iid, lineairity & stationarity, merging & joining, missing values, imputation, (dummy) encoding
3	Wednesday 14-05-2025	SvdB/MV	Check-ins		Presenting DU/DaP in more detail: linearity, scaling, imputation, external data, iid problems, etc

4	Monday 19-05-2025 (NL) GW319 09:00 – 12:00 (INT) GW319 13:00 – 16:00	SvdB/RK	Modelling I: Regression	2.4 – 2.5	Assumptions, target and features, test design, linear regression, l1/l2 regularization, non-linear regressors (SVM, RF, etc), metrics, interpretation, residuals, brief intro for non-iid regression
4	Wednesday 21-05-2025	SvdB/MV	Check-ins (see schedule on Learn)	3.1 – 3.2	Presenting DU/DaP and first modelling: test design, first models
5	Monday 26-05-2025		**DataFest**		Sign up for this data science hackathon! Highly recommended for students who plan on choosing the DS track.
5	Tuesday 27-05-2025 (NL + INT) GW027 13:00 – 14:30	SvdB/MV	Modeling II: Classification (English theory lecture only, due to DataFest the day before)		Logistic regression, assumptions, SVM and non-linear kernels, non-linear classifiers, test design, metrics, class imbalance, stratification
5	Wednesday 28-05-2025	SvdB/MV	Check-ins (see schedule Learn)		Presenting modelling (first and second iteration(s))
6	Monday 02-06-2025 (NL) GW319 09:00 – 12:00 (INT) GW319 13:00 – 16:00	SvdB/RK	Modelling III: Unsupervised & optimization	3.2 – 3.3	Unsupervised modelling, clustering, metrics, dimensionality reduction, overfitting/underfitting, hyperparameter tuning, feature importances
6	Wednesday 04-06-2025	SvdB/MV	Check-ins (see schedule)		Presenting modelling results, iterations, metric interpretation
7	Tuesday 10-06-2025 (NL) GW319 09:00 – 12:00 (INT) GW319 13:00 – 16:00	SvdB	Evaluation	4.1 – 4.4	Results vs. business goals, interpretation, completion and delivery, Q&A

7	Wednesday 11-06-2025	SvdB/MV	Check-ins		Evaluation, relating back to BOs/DMGs, conclusions
8	Friday 19-06-2025 23:59		**PORTFOLIO DEADLINE 23:59**		See project guidelines
8	Tuesday 24-06-2025 09:00-17:00	SvdB/MV/RZ	**QUIZ**		All class material, **it is strongly advised to also consult recommended material**
8	Friday 02-07-2025	SvdB/MV/ RZ	**QUIZ RESIT**		
9	Friday 04-07-2024 23:59		**PORTFOLIO RESIT**		