Introduction to Data Science and Machine Learning - Summer Semester 2025

Bachelor of Science WI / IS
EM Wirtschaftsinformatik II [1277BEWIF2]
Faculty of Management, Economics, and Social Sciences
Department of Information Systems for Sustainable Society
University of Cologne
Version - April 29, 2025

Instructors Prof. Dr. Wolfgang Ketter
Dr. Saber Talari
TAs Julius Kuhmann
Dennis Bakalis
Term SS 2025
Website `https://www.is3.uni-koeln.de` and ILIAS

Team Assignment

The DSML team project is designed to test a representative cross-section of the data analytics and machine learning approaches we cover during this course. It is based on real-world datasets with high relevance to the current hot topic of smart mobility systems. The following pages provide a detailed outline of all relevant aspects of the team assignment, including a detailed description of the tasks and expected outcomes, and the deliverables. Please read the instructions carefully!

1 Background

Transport-related greenhouse gas emissions make up for a large chunk of total EU emissions. It has thus long been recognized that in order to meet decarbonization targets, our approach to mobility will have to change. To this day, traditional urban mobility relies primarily on private, internal combustion (IC) engine vehicles. One of the main problems with this mobility setup is that it is highly inefficient. Specifically, utilization of passenger cars is very low, thus requiring many cars to provide mobility to comparatively small numbers of passengers, leading to congested roads and, ultimately, increased climate impact. The need for a comprehensive transformation of the mobility system has been recognized and the mobility landscape is changing fast. A crucial trend in this newly emerging ecosystem is the consumption of mobility as-a-service (MaaS) and on-demand (MoD) heralding in the age of shared, fleet-based transportation companies. Ride-hailing platforms are an excellent manifestations of MaaS and MoD. Providers like Uber or Lyft facilitate matchmaking between drivers and mobility users. Similarly, micromobility services including shared bicycles and e-scooters are becoming increasingly popular. The resulting pooling of vehicle resources can contribute to more efficient and sustainable urban mobility systems.

Innovative mobility services, like ride-hailing or e-scooter sharing, oftentimes expand rapidly. It is, therefore, important to carefully evaluate their impact, and to investigate their intended and unintended consequences.[1] Interestingly, many cities and mobility services make trip-related data publicly available, thereby providing valuable resources for data scientists to study these novel services. For example, researchers have used trip data to identify usage patterns of e-scooter ridership[2], analyze how the introduction of a novel service impacts existing services[3], study operational aspects such as the size of a shared mobility fleet and how to best distribute vehicles across a service area[4], and assess how novel mobility services influence air pollution[5]. Within the scope of this project, we also want to use publicly available mobility data to study a relevant aspect of shared mobility by building a predictive model.

[1] See the commentary on this topic by Ketter et al. (2023) available here.
[2] Study can be found here.
[3] Study can be found here.
[4] Study can be found here.
[5] Study can be found here.

2 Description of Tasks

Your task for the team assignment is to carry out a small-scale data science project along the lines of the first five stages of the CRISP Data Science Process.

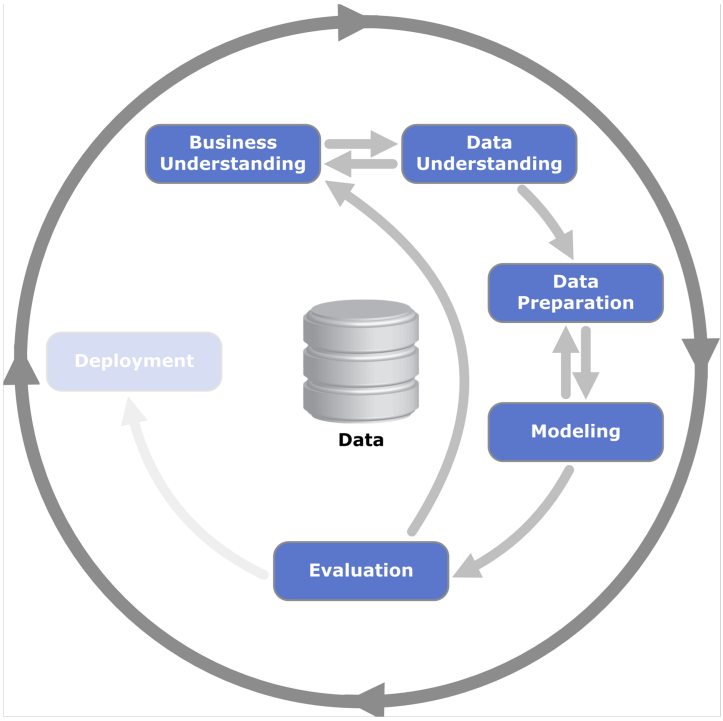

Fig. 1. Cross Industry Standard Process for Data Mining (CRISP-DM)

1. **Business Understanding & Data Understanding**: As a first step, **find a mobility dataset** that you would like to focus your project on. The dataset must fulfill the following conditions:
   - The dataset is from the mobility domain. More specifically, every row in the dataset must represent a trip.
   - The dataset contains a sufficient number of observations. More specifically, the raw data must consist of at least 100,000 rows.
   - The dataset contains a sufficient number of features. More specifically, the raw data must have at least 6 columns containing relevant information.

   To help you get started, the following table contains potential datasets that fulfill the above conditions. You may choose from (but are not limited to) these options.

Transportation mode	**City**	**Source**	**Download URL**
Ride-hailing	Chicago, U.S.	City of Chicago	access here
Ride-hailing & taxi	New York City, U.S.	Taxi & Limousine Commission	access here
Bike-sharing	Chicago, U.S.	City of Chicago	access here
Bike-sharing	New York City, U.S.	Lyft	access here
Bike-sharing	Washington D.C., U.S.	Capital Bikeshare	access here
Bike-sharing	Boston, U.S.	Bluebikes	access here
E-scooter sharing	Chicago, U.S.	City of Chicago	access here
Bikesharing & e-scooter sharing	Austin, U.S.	City of Austin	access here

Table 1: Potential Datasets

After deciding on a dataset, **familiarize yourself with the domain** (i.e., the chosen mobility service, the mobility landscape in the chosen city etc.) **and the data** (i.e., the available raw data, included features etc.). Finally, **define a specific prediction problem**. More specifically, you should

either select one of the variables as your prediction target or generate a prediction target from the raw data.

2. **Data Preparation**: After selecting a dataset and defining your prediction problem, **prepare the raw data by conducting data cleaning steps** discussed during the workshops and lectures (e.g. dealing with missing values, categorical features etc.). You should also **augment your dataset by engineering relevant features and potentially enriching your dataset with information from other data sources**. After completing data preparation, you should have a clean dataset containing all relevant features to conduct your predictive analysis.

3. **Modeling & Evaluation**: Using your prepared dataset, develop a suitable model for your prediction problem defined during Step 1. More specifically, **select and implement two different algorithms**. Afterwards, **evaluate and compare your models' performances** using suitable evaluation metrics.

Notes and tips

- Despite the separation into the different stages, it is completely fine (and likely unavoidable) to re-iterate earlier steps (as is also shown by the CRISP-DM cycle in Figure 1).
- Depending on your choice of a target (continuous or discrete), you will be working on a regression or classification problem. Either is fine. However, keep in mind that we will discuss classification in lectures and workshops at a later point during the course.
- During data preparation, you should generate as many relevant features for your prediciton as possible. To do this, we encourage you to enrich your dataset using information from other sources. For example, you can find daily and hourly weather data for many cities worldwide from https://meteostat.net/en/, which also has an easy-to-use Python library to access the data (see details here). Similarly, many of the U.S.-based cities designate trip origins and destinations by census tracts, which are similar to neighborhoods defined across the entire United States. The U.S. Census Bureau provides detailed data for each of these census tracts, including demographic, housing, social and economic characteristics (available for download here).
- Make generous use of visualization techniques during all stages of the project to clearly illustrate your findings and present them in an appealing way.
- Evaluate your methodology and critically think about why you have opted for a specific approach in your analysis, particularly during data preparation and modeling.
- **A note on the usage of ChatGPT**: ChatGPT is one among many tools that can provide useful assistance when writing code. This is a course on data science and machine learning, so we do not prohibit the usage of AI tools as such. However, we do expect you to be explicit about it. In case you utilized AI tools, please include a short section in your oen-page supplementary document (see below) describing how you made use of them. Please refer to Section G3 in the Appendix of the course syllabus for further information.

3 Deliverables and Deadlines

The class has been divided into teams consisting of ca. 5 students each (see ILIAS for group composition). Please coordinate the work independently in your teams. We expect everyone to contribute equally to the project.

You are expected to submit the following deliverables to successfully complete this group project:

Milestone 1 (grading: pass/fail)
Deadline: 23:59h on 25^{th} of May, 2025

- A single Jupyter notebook (.ipynb format). At the top of the notebook, there must be a text cell containing a link to your chosen dataset as well as a description of your specific prediction task (Which target are you going to predict? Which features do you plan to include?). Afterwards, the notebook should show that you have downloaded your raw data and started data preparation.

Milestone 2 (grading: pass/fail)
Deadline: 23:59h on 22^{nd} of June, 2025

- A single Jupyter notebook (.ipynb format), showing first descriptive analyses and that you have started working on your predictive models.

Final Submission (grading: based on full grading scale)
Deadline: 23:59h on 15^{th} of July, 2025

- A **slide deck** to be submitted in .pdf format via ILIAS and **to be presented in-person on July 16^{th}** during normal class hours (a presentation schedule will be shared before the session). Each presentation will be max. 10 minutes long, followed by a 3 minutes Q&A. The presentation should be aligned with the CRISP stages you followed throughout your project and should contain the following:
  1. *Business Understanding*: Describe the chosen setting and context and present your prediction task. Clearly articulate why this is a relevant problem for, for example, the shared mobility provider, the city government, or society as a whole. Use visualizations where appropriate.
  2. *Data Understanding*: Shortly present your utilized raw datasets.
  3. *Data Preparation*: Explain the most important steps you took to prepare your raw data. Focus on providing justifications for choices you made during data prep. Use visualizations where appropriate.
  4. *Modeling & Evaluation*: Shortly present your two chosen models. Explain and justify why you selected these algorithms and shortly describe their respective advantages and drawbacks. Then present and compare the models' performances. Use visualizations where appropriate.
  5. *Implications & Reflection*: Clearly state the implications (i.e., the "so what?") of your findings for managers/decision makers. Relate your results to the real world and interpret them for non-technical audiences. Also address how your models could be improved further and explain some of the improvement levers that you might focus on in a follow-up project. Which further analyses would you consider useful and could be conducted on the given dataset?
- A **single well-structured and clearly annotated Jupyter notebook** (.ipynb format) with your code detailing your analysis and including executable Python code.
- A **1-page supplementary document** detailing the individual contributions of each team member (i.e., who did what). You will be graded as a group for your team assignment, but we do want you to reflect upon and be explicit about every team member's contribution.

Please make sure to submit all deliverables electronically via the respective upload points in ILIAS no later than the indicated deadlines. Please refer to the course syllabus (Section G.4) for further information on how to submit your work.