

# Seminar Data Analytics, Winter Sem. 2025 - 2026

Prof.Dr.Patrick Puhani  
M.Sc.Vangjel Bitu  
M.Sc.Brajan Gruszka

Institute of Labour Economics  
Institut für Arbeitsökonomik

## Seminar Description

In this seminar, we invite students to engage in practical exercises, gain hands-on experience with Python, and apply their machine learning (ML) skills within the field of economics. Students will form teams and be assigned projects, which include a specific topic and dataset to work on throughout the semester. Additionally, python code, introduction on the methods and helpful instructions will be provided to start with (Programming experience is advantageous but not necessary). At the end of the semester, students will present their work in front of their peers. Our research projects will focus on leveraging the predictive advantages offered by machine learning in empirical economic research designs. Specifically, we will utilize machine learning techniques to:

1. estimate the gender wage gap in earnings with partially linear regression where ML predictions are used to find the optimal controls.
2. estimate the union wage gap in earnings with partially linear regression where ML predictions are used to find the optimal controls.
3. identify racial bias and make accurate predictions of the risk of repeating a crime of individuals that have been already convicted once.
4. explore heterogeneity in treatment effects arising from experimental data from cash transfers and child health in Indonesia.

Students in groups of 2, will choose a topic (data and tasks, see above) and a machine learning algorithm to apply in their data. The group assignment will take place via StudIP at a first come first serve basis. The concrete Topics slots will be created in StudIP within the next weeks, we will notify you by an announcement in StudIP.

## Learning Goals

- Focus is on working with data and programming in Python.
- Hands on experience with data manipulations and visualization.
- Introduction on ML methods like Trees, random forest, etc.
- Using ML methods for best predictions.
- Using ML for causal analysis and choosing best controls to overcome omitted variable bias.
- Using ML for analysis of heterogeneous treatment effects.

## Aim

The idea is to turn students into knowledgeable users of these methods with awareness for their potential pitfalls when transferring methods to economic applications. Such pitfalls can be data quality and bias, over-fitting, causal inference and endogeneity, interpretability and transparency behind the algorithms, model complexity and ethical considerations.

The course aims to familiarize students with ML algorithms, work on applications in economics based on latest research and support them to work independently on a project in python.

## Course Structure

The seminar will consist of four meetings and the following deadlines.

1. Seminar introduction. Presentation of the general seminar structure, seminar topics, tasks and what is expected from students. **tba 42 kw** (You can hand seminar registration forms in person that day or send by **17.10.2025** at [seminare@aoek.uni-hannover.de](mailto:seminare@aoek.uni-hannover.de))
2. Methods introduction. Hands-on introduction on coding in python and machine learning methods. **tba**
3. Methods introduction 2. Hands-on introduction on coding in python and machine learning methods. **tba**
4. Consultation meeting for feedback and questions. **tba**
5. Presentations round. **Date: 03-04.12.2025**

## Data

Students will be provided with 3 datasets and useful code to start with. They will be provided via a link to the course's StudIP repository. To each team one of the following datasets and ML method/model will be assigned.

1. Data CPS earnings → estimate the gender gap and union wage gap.
2. Data Recidivism → predictions of Risk of Recidivism.
3. Data Margaret Triyana Heterogenous Treatment effects Indonesia AEA web.

## Books and Literature

Relevant literature per data Group:

1. Data Group 1: Double/debiased machine learning for treatment and structural parameters.
2. Data Group 2: COMPAS Recidivism Risk Score Data and Analysis
3. Data Group 3: Do Health Care Providers Respond to Demand-Side Incentives? Evidence from Indonesia

Main Sources:

1. [DMBA]Shmueli, Bruce, Gedeck, Patel. Data mining for business analytics: concepts, techniques and applications in Python (Data Analytics Course Book, covers both ML methods and coding in python.)
2. [PDSH]VanderPlas: Python Data Science Handbook (Good for programming in python.)

Further Sources: look at your topic handout.

## Machine Learning Algorithms

1. Decision Trees
2. Random Forests
3. Boosted Trees

## Examination

1. Submission of Code script: Code that reproduces graphs tables and the results of your analysis in Python.
2. Term Paper: 8 - 10 pages. For more details see section Seminar Paper Guidelines
3. Presentation: 15 Minutes. For more details see section Presentation Guidelines
4. **Grade Determination:**
  - Seminar Paper: 50%
  - Oral Presentation: 40%
  - Python Code: 5%
  - Active Participation in the Seminar: 5 %

## Important Dates and Deadlines

1. Registration deadline: **17.10.2025**
2. Introductory Meeting 1: Seminar and Topic presentation. **tba**
3. Introductory Meeting 2: Machine learning and Coding introduction. **tba**
4. Introductory Meeting 3: Machine learning and Coding introduction. **tba**
5. Feedback meeting 2 weeks before submission for final questions and guidance. **tba**
6. Term Paper submissions: **30.11.2025 until 23:59**
7. Presentation submission: **02.12.2025 until 23:59**
8. Seminar Presentations: **03-04.12.2025**

# **Seminar Paper Guidelines**

## **Essential Components of the Seminar Paper**

- Cover page
- 8 to 10 pages of text and essential tables and figures
- References
- Appendix with (additional) figures or tables from your data work

## **Formal Aspects**

- Text font and size: Times New Roman, Arial or Calibri 12 point, or LaTeX standard 11 point
- Line spacing: 1.5-line
- Alignment: Justify
- Margins: all sides 2.54 cm (1 inch)
- Page numbers
- Citation: Harvard citation style, i.e. in the text as ‘surname (year)’ respectively ‘surname (year, p. xy)’ for direct quotes
- References: Include every reference cited in the text and no others
- Language: German or English

## **Detailed Structure and Guidelines for Seminar Paper**

### **Introduction**

- State the research question, relevance, and objective clearly.
- Briefly describe your choice of variables and chosen methodology.

### **Descriptive Statistics and Data Preparation**

- Provide key summary statistics and visualizations.
- Clearly document data cleaning and variable selection reasons.

### **Methodology**

- Explain your group’s ML algorithm.
- Detail the chosen method (Double ML, calibration vs balance, heterogeneous effects).
- Discuss the strengths and limitations of your method.

### **Results**

- Explain your steps on applying the chosen method and your ML algorithm.
- Clearly present and compare baseline regression and machine learning results.
- Include graphical representations of key outcomes.

- Describe your hyper-parameter tuning strategy and show graphically.

### **Discussion**

- Reflect on the robustness and significance of the results. What do they mean ?
- Identify possible methodological and data limitations, if any.

### **Conclusion**

- Summarize key findings and their implications.

## **Presentation Guidelines**

- Duration: 15 minutes + 5-10 minutes discussion.
- Suggested structure.
  - Intro Slide
  - Show for your contributions on data preparation and data cleaning and present interesting descriptive statistics or data visualization
  - Describe your implementation and any challenges of your topic's method (Partial linear models, Learners,etc.)
  - Describe your implementation and any challenges of the ML algorithm of your group (Trees, random forest, boosting)
  - Show your hyper-parameter tuning strategy (which target values tuned, process, results)
  - Present your results, tables and graphs
  - Discussion slide (How do your results compare to the baseline? How is their interpretability?)

## **Submission of Your Work**

Submit your seminar paper and presentation as PDF files via email to: **seminare@aoek.uni-hannover.de**. Submit also your python code script as .py or as .ipynb files. If possible, submit as a zip file containing dataset, code, result tables, and PDF of your seminar paper.

**Subject line format:** *"Submission (or Abgabe) Seminar DATA ANALYTICS"*

### **File naming conventions:**

- *DA GroupNumber Topic YourLastName S* (seminar paper)
- *DA GroupNumber Topic YourLastName P* (presentation)
- *DA GroupNumber Topic YourLastName C* (python code script)

### **Example:**

- DA 1 A Mustermann S/P/C