Seminar Data Analytics, Summer Sem. 2025

Prof.Dr.Patrick Puhani **Institute of Labour Economics**
M.Sc.Vangjel Bita **Institut für Arbeitsökonomik**
M.Sc.Brajan Gruszka

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

- 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. **Date: 07.04.2025** (You can hand seminar registration forms in person that day or send by **11.04.2025** at seminare@aoek.uni-hannover.de)
2. Methods introduction. Hands-on introduction on coding in python and machine learning methods. **09.04.2025**
3. Methods introduction 2. Hands-on introduction on coding in python and machine learning methods. **11.04.2025**
4. Consultation meeting for feedback and questions. **Date: 09.05.2025**
5. Presentations round. **Date: 28.05.2024**

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

Further Books.

1. Friedman, Hastie and Tibshirani: Elements of Statistical Learning
2. Goodfellow, Bengio and Courville: Deep Learning
3. Efron and Hastie: Computer Age Statistical Inference
4. Vishalie Shah , Noemi Kreif , and Andrew M. Jones : Machine learning for causal inference: Estimating heterogeneous treatment effects

Machine Learning Algorithm

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 of Data vizualization, descriptive, method, choice of variables, training and tuning choices and interpretation of the results. In the term paper benchmark regression method should be analyzed in order to be compared with ML and showcase its advantages or disadvantages.
3. Presentation: 15 Minute explaining the ML Method theory, Variable Choice Reasons, Training tuning steps explained and results.
4. **Grade Determination**:
    - Seminar Thesis: 50%
    - Oral Presentation: 40%
    - Python Code: 5%
    - Active Participation in the Seminar: 5 %

Important Dates and Deadlines

1. Registration deadline: **11.04.2025**
2. Introductory Meeting 1: Seminar and Topic presentation.**07.04.2025**
3. Introductory Meeting 2: Machine learning and Coding introduction. **09.04.2025**
4. Introductory Meeting 3: Machine learning and Coding introduction. **11.04.2025**
5. Feedback meeting 2 weeks before submission for final questions and guidance. **09.05.2025**
6. Term Paper submissions: **25.05.2025 until 23:59**
7. Presentation submission: **27.05.2025 until 23:59**
8. Seminar Presentations: **28.05.2024**