Syllabus

Machine Learning II

2 credits/ 4 ECTS

Prof. Trent
Buskirk Dr.
Christoph Kern

Video lecture by
Prof. Trent
Buskirk Dr.
Christoph Kern

September 19 – November 7, 2025

Short Course Description

Social scientists and survey researchers are confronted with an increasing number of new data sources such as apps and sensors that often result in (para)data structures that are difficult to handle with traditional modeling methods. At the same time, advances in the field of machine learning (ML) have created an array of flexible methods and tools that can be used to tackle a variety of modeling problems. Against this background, this course discusses advanced ML concepts such as cross validation, class imbalance, Boosting and Stacking as well as key approaches for facilitating model tuning and performing feature selection. In this course we also introduce additional machine learning methods including Support Vector Machines, Extra-Trees and LASSO among others. The course aims to illustrate these concepts, methods and approaches from a social science perspective. Furthermore, the course covers techniques for extracting patterns from unstructured data as well as interpreting and presenting results from machine learning algorithms. Code examples will be provided using the statistical programming language R.

Course Objectives

By the end of the course, students will…

- will have a profound understanding of advanced (ensemble) prediction methods
- have built up a comprehensive ML toolkit to tackle various learning problems
- know how to (critically) evaluate and interpret results from "black-box" models

Prerequisites

Topics covered in SURV751: Introduction to Machine Learning and Big Data (ML I), i.e.:

- Conceptual basics of machine learning (training vs. test data, model evaluation basics)
- Decision trees with CART
- Random forests

Familiarity with the statistical programming language R is strongly recommended.

Participants are encouraged to work through one or more R tutorials prior to the first-class meeting. Some resources can be found here:

- https://rstudio.cloud/learn/primers
- http://www.statmethods.net/
- https://swirlstats.com/
- https://socialsciences.mcmaster.ca/jfox/Misc/Rcmdr/

Class Structure and Course Concept

This is an online course using a flipped classroom design. It covers the same material and content as an on-site course but runs differently. In this course, you are responsible for watching video-recorded lectures and reading the required literature for each unit prior to participating in mandatory weekly one-hour online meetings where students have the chance to discuss the materials from a unit with the instructor.

Although this is an online course where students have more freedom in when they engage with the course materials, students are expected to spend the same amount of time overall on all activities in the course – including preparatory activities (readings, studying), in-class-activities (watching prerecorded videos, attending the live online meetings), and follow-up activities (working on assignments and exams) – as in an on-site course. As a rule of thumb you can expect to spend approximately 3h/week on in-class-activities and 9 hours per week on out-of-class activities (preparing for class, readings, assignments, projects, studying for quizzes and exams). Therefore, the workload in all courses will be approximately 12h/week. Please note that the actual workload will depend on your personal knowledge.

Mandatory Weekly Online Meetings

Sec. 1: Fridays, September 19 - November 7, 2025, 11 am ET/5 pm CET - 12 pm ET/6 pm CET

Meetings will be held online through Zoom. Follow the link to the meeting sessions on the course website on https://umd.instructure.com/courses/1352068. If video participation via the Internet is not possible, arrangements can be made for students to dial in and join the meetings via telephone.

In preparation for the weekly online meetings, students are expected to watch the lecture videos and read the assigned literature before the start of the meeting. In addition, students are encouraged to post questions about the materials covered in the videos and readings of the week in the forum before the meetings (deadline for posting questions is Thursdays, 11:00 am ET/5:00 pm CET).

Students have the opportunity to use the BigBlueButton feature in Canvas to connect with peers outside the scheduled weekly online meetings (e.g., for study groups). Students are not required to use BigBlueButton and can of course use other online meeting platforms such as Google Hangouts, Skype or Microsoft Teams.

Daylight saving time ends in Europe on October 26, 2025 and clocks are turned back 1 hour. Daylight saving time ends in the USA on November 2, 2025. Therefore, look carefully at the times of meetings and deadlines! If in doubt, please consider the CET Time (e.g. in Frankfurt) is the OFFICIAL time for all meetings and deadlines.

Grading

Grading will be based on:

- 4 homework assignments (10% each)
- 8 online quizzes (5% each)
- Participation in discussion during the weekly online meetings (20% of grade)

A+	100 - 97
A	96 - 93
A-	92 - 90
B+	89 - 87
B	86 - 83
B-	82 - 80
Etc.	

The final grade will be communicated under the assignment "Final Grade" in the Canvas course. Please note that the letter grade written in parentheses in Canvas is the correct final grade. The point-grade displayed alongside the letter grade is irrelevant and can be ignored. Dates of when assignment will be due are indicated in the syllabus. Extensions will be granted sparingly and are at the instructor's discretion.

Technical Equipment Needs

The learning experience in this course will mainly rely on the online interaction between the students and the instructors during the weekly online meetings. Therefore, we encourage all students in this course to use a web camera and a headset. Decent quality headsets and web cams are available for less than $20 each. We ask students to refrain from using built-in web cams and speakers on their desktops or laptops. We know from our experience in previous online courses that this will reduce the quality of video and audio transmission and therefore will decrease the overall learning experience for all students in the course. In addition, we suggest that students use a wire connection (LAN), if available, when connecting to the online meetings. Wireless connections (WLAN) are usually less stable and might be dropped.

Long Course Description

Social scientists and survey researchers are confronted with an increasing number of new data sources such as apps and sensors that often result in (para)data structures that are difficult to handle with traditional modeling methods. At the same time, advances in the field of machine learning (ML) have created an array of flexible methods and tools that can be used to tackle a variety of modeling problems. Against this background, this course discusses advanced ML concepts such as cross validation, class imbalance, Boosting and Stacking as well as key approaches for facilitating model tuning and performing feature selection. In this course we also introduce additional machine learning methods including Support Vector Machines, Extra-Trees and LASSO among others. The course aims to illustrate these concepts, methods and approaches from a social science perspective. Furthermore, the course covers techniques for extracting patterns from unstructured data as well as interpreting and presenting results from machine learning algorithms. Code examples will be provided using the statistical programming language R.

The course is structured such that each session focuses on specific prediction tasks and presents tools that can be used to tackle modeling problems in this setting. Topics include, e.g., accounting for informative data structures in the context of model training and tuning, dealing with class imbalance in categorical outcomes, building effective prediction models by applying cutting edge ML methods, and performing feature selection in high-dimensional data settings. The presented methods will be motivated from a social and survey science perspective and critically discussed with respect to their advantages and limitations.

Code examples will be provided using the statistical programming language R.

Readings

Primary Readings

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer. https://web.stanford.edu/~hastie/ElemStatLearn/

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). An Introduction to Statistical Learning. New York, NY: Springer. https://www.statlearning.com/

Boehmke, B., and Greenwell, B. M. (2019). Hands-On Machine Learning with R. Boca Raton, FL: CRC Press. https://bradleyboehmke.github.io/HOML/

Buskirk, T. D., Kirchner, A., Eck, A. and Signorino, C. (2018). An Introduction to Machine Learning Methods for Survey Researchers. Survey Practice 11(1). https://doi.org/10.29115/SP-2018-0004

Kern, C., Klausch, T., and Kreuter, F. (2019). Tree-based Machine Learning Methods for Survey
Research. Survey Research Methods 13(1), 73--93. https://doi.org/10.18148/srm/2019.v1i1.7395

Required and Recommended Readings

List of required and recommended readings for each class are provided below for each specific unit.

Academic Conduct

Clear definitions of the forms of academic misconduct, including cheating and plagiarism, as well as information about disciplinary sanctions for academic misconduct may be found at

https://www.president.umd.edu/sites/president.umd.edu/files/documents/policies/III-100A.pdf (University of Maryland)

Knowledge of these rules is the responsibility of the student and ignorance of them does not excuse misconduct. The student is expected to be familiar with these guidelines before submitting any written work or taking any exams in this course. Lack of familiarity with these rules in no way constitutes an excuse for acts of misconduct. Charges of plagiarism and other forms of academic misconduct will be dealt with very seriously and may result in oral or written reprimands, a lower or failing grade on the assignment, a lower or failing grade for the course, suspension, and/or, in some cases, expulsion from the university.

Accommodations for Students with Disabilities

In order to receive services, students at the University of Maryland must contact the Accessibility &
Disability Service (ADS) office to register in person for services. Please call the office to set up an appointment to register with an ADS counselor. Contact the ADS office at 301.314.7682; https://www.counseling.umd.edu/ads/.

Course Evaluation

In an effort to improve the learning experience for students in our online courses, students will be invited to participate in an online course evaluation at the end of the course. Participation is entirely voluntary and highly appreciated.

Sessions

Week 1: Intro: Bias-variance trade-off, cross-validation (stratified splits, temporal cv) and model tuning (grid and random search)

Video lecture: available Friday, September 12, 2025

Online meeting: Friday, September 19, 2025, 11 AM ET/5 PM

CET Online Quiz 1: due Monday, September 29, 2025, 10 AM ET/4CEST

Required Readings:

Ghani, R. and Schierholz, M. (2017). Machine learning. In: Foster, I., Ghani, R.,

Jarmin,
R. S., Kreuter,
F., and Lane, J. (Eds.). Big Data and Social Science: A Practical Guide to Methods and Tools. Boca Raton, FL: CRC Press Taylor & Francis Group. https://coleridge-initiative.github.io/big-data-and-socialscience/

Recommended Readings:

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Model Assessment and Selection. In: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. New York, NY: Springer.

Week 2: Classification: Performance metrics (ROC, PR curves, precision at K) and class imbalance (over- and undersampling, SMOTE)

Video lecture: available Friday, September 19, 2025

Online meeting: Friday, September 26, 2025, 11 AM ET/5 PM CET

Online quiz 2: due Monday, October 6, 2025, 10 AM ET/4 PM CEST

CET Homework 1: due Monday, October 13, 2025, 10 AM ET/4 PMCEST

Required Readings:

Kuhn, M. and Johnson, K. (2019). Measuring Performance. In: Feature Engineering and Selection: A Practical Approach for Predictive Models. http://www.feat.engineering/index.html

Recommended Readings:

Kuhn, M. and Johnson, K. (2013). Measuring Performance in Classification Models. In: Applied Predictive Modeling. New York, NY: Springer.

Kuhn, M. and Johnson, K. (2013). Remedies for Severe Class Imbalance. In: Applied Predictive Modeling. New York, NY: Springer.

Week 3: Ensemble methods I: Bagging and Extra-Trees

Video lecture: available Friday, September 26, 2025

Online meeting: Friday, October 3, 2025, 11 AM ET/5 PM CEST

CET Online quiz 3: due Monday, October 13, 2025, 10 AM ET/ 4PM CEST

Required Readings:

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Chapters 7.11 and 8.7. New York, NY: Springer.

Recommended Readings:

Breiman, L. (1996). Bagging predictors. Machine Learning 24, 2, 123--140.

Geurts, P., Ernst, D., and Wehenkel, L. (2006). Extremely randomized trees. Machine Learning 63, 1, 3-42.

Week 4: Ensemble methods II: Boosting (Adaboost, GBM, XGBoost) and Stacking

Video lecture: available Friday, October 3, 2025

Online meeting: Friday, October 10, 2025, 11 AM ET/5 PM CET

Online quiz 4: due Monday, October 20, 2025, 10 AM ET/4 PM CET

Homework 2: due Monday, October 27, 2025, 11 AM ET/4 PM CET

Required Readings:

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Chapters 10.1, 10.9, 10.10. New York, NY: Springer.

Recommended Readings:

Mayr A., Binder H., Gefeller O., Schmid M. (2014). The evolution of boosting algorithms: from machine learning to statistical modelling. Methods of Information in Medicine 53(6), 419--427.

Ridgeway, G. (2018). Generalized Boosted Models: A guide to the gbm package.

Week 5: Variable selection: Lasso, elastic net and fuzzy/ recursive random forests

Video lecture: available Friday, October 10, 2025

Online meeting: Friday, October 17, 2025, 11 AM ET/5 PM CET

CET Online quiz 5: due Monday, October 27, 2025, 10 AM ET/ 5PM CET

Required Readings:

Efron, B. and Hastie, T. (2016). Sparse Modeling and the Lasso. In: Computer Age Statistical Inference. Algorithms, Evidence, and Data Science. New York, NY: Cambridge University Press.

Recommended Readings:

Kuhn, M. and Johnson, K. (2013). An Introduction to Feature Selection. In: Applied Predictive Modeling. New York, NY: Springer.

Kuhn, M. and Johnson, K. (2019). Feature Selection Overview. In: Feature Engineering and Selection: A
Practical Approach for Predictive Models. http://www.feat.engineering/index.html

Week 6: Support Vector Machines

Video lecture: available Friday, October 17, 2025

Online meeting: Friday, October 24, 2025, 11 AM ET/5 PM CET

Online quiz 6: due Monday, November 3, 2025, 10 AM ET/4 PM CET

Homework 3: due Monday, November 10, 2025, 10 AM ET/4 PM CET

Required Readings:

Kirchner, A., and Signorino, C. S. (2018). Using Support Vector Machines for Survey Research. Survey Practice 11(1). https://doi.org/10.29115/SP-2018-0001

Recommended Readings:

Chang, C. C., and Lim, C. J. (2016). LIBSVM. A Library for Support Vector Machines. https://www.csie.ntu.edu.tw/~cjlin/libsvm/

Week 7: Advanced unsupervised learning: Hierarchical clustering and LDA

Video lecture: available Friday, October 24, 2025

Online meeting: Friday, October 31, 2025, 11 AMET/5 PM CEST

CET Online quiz 7: due Monday, November 10, 2025, 10 AM ET/ 4PMCEST

Required Readings:

Kassambar, A. (2017). Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning, Chapters 7--9. Free Online View and PDF download available at: https://kupdf.net/download/practicalguide-to-cluster-analysis-in-r-unsupervised-machine-learning_5a65e055e2b6f556501cc785_pdf Recommended Readings:

Jodrey, J. (2016). Hierarchical Cluster Analysis. Blog post, https://uc-r.github.io/hc_clustering

Boedeker, P., and Kearns, N. T. (2019). Linear Discriminant Analysis for Prediction of Group Membership: A User-Friendly Primer. Advances in Methods and Practices in Psychological Science 2(3), 250--263.

Week 8: Interpreting (Variable Importance, PDP, ...) and reporting ML results

Video lecture: available Friday, October 31, 2025

Online meeting: Friday, November 7, 2025, 11 AM ET/5 PM CET

Online quiz 8: due Monday, November 17, 2025, 10 AM ET/4 PM CET

Homework 4: due Monday, November 24, 2025, 10 AM ET/4 PM CET

Required Readings:

Molnar, C. (2023). Interpretable Machine Learning. A Guide for Making Black Box Models Explainable, Chapters 8.1--8.6. https://christophm.github.io/interpretable-ml-book/

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., … Berk, M. (2016). Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. Journal of medical Internet research 18, 12.

Jodrey, J. (2018) Interpreting Machine Learning Models with the iml Package. https://ucr.github.io/2018/08/01/iml-pkg/