Hertie School **Master of Data Science for Public Policy**
Fall Semester 2022

Course Syllabus

GRAD-C24: Machine Learning

Slava Jankin

1. General information

Course Format	Both the lecture and the accompanying labs take place onsite.
Instructor	Slava Jankin
Instructor's e-mail	jankin@hertie-school.org
Assistant	Name: Alex Karras Email: karras@hertie-school.org
Instructor's Office Hours	TBA

Link to Study, Examination and Admission Rules and MIA, MDS and MPP Module Handbooks

For information on **course times, session dates and course locations** please consult the Course List on *MyStudies*.

Instructor Information:

Slava Jankin is Professor of Data Science and Public Policy at the Hertie School. He is the Director of the Hertie School Data Science Lab. His research and teaching are primarily in the field of natural language processing and machine learning. Before joining the Hertie School faculty, he was a Professor of Public Policy and Data Science at University of Essex, holding a joint appointment in the Institute for Analytics and Data Science and Department of Government. At Essex, Slava served as a Chief Scientific Adviser to Essex County Council, focusing on artificial intelligence and data science in public services. He previously worked at University College London and London School of Economics. Slava holds a PhD in Political Science from Trinity College Dublin.

2. Course Contents and Learning Objectives

Course contents:
Machine learning is a core technology of artificial intelligence and data science that enables computers to act without being explicitly programmed. Recent advances in machine learning have given us, inter alia, self-driving cars, AlphaGo, Amazon, and Netflix. This technology has also allowed us to predict armed conflict and post-electoral violence, detect fake news, develop targeted provision of care and public services, and implement early policy interventions. This course provides a hands-on introduction to machine learning. By the end of this course students will have a sound understanding of the key concepts of machine learning, the ability to analyse data using some of its

main methods, and a solid foundation for more advanced or more specialised study. The course covers topics in supervised and unsupervised learning, including the most common learning algorithms for regression, classification and clustering, such as random forests, neural networks, and dimensionality reduction techniques. Students will learn the fundamental concepts underlying machine learning algorithms, and we will equally focus on the practical use of machine learning algorithms using open-source frameworks.

Main learning objectives:

By the end of this course students will have a sound understanding of the key theoretical concepts of machine learning and develop the ability to analyse data using some of its main methods. More fundamentally, by the end of the course students will ***learn how to learn*** machine learning.

Teaching style:

The course adopts, as much as possible, active learning pedagogical approach. Active learning is an approach that focuses on student engagement and interaction (Nguyen et al. 2016); it allows students to drive their own education and enables them to "learn how to learn" (Akinoglu and Tandogan 2006). Active learning consistently demonstrates better learning outcomes compared to the orthodox "teaching by telling" (Yannier et al. 2021), also in technical disciplines (Freeman et al. 2014).

In this course, active learning is implemented through the focus on student group research projects as the key deliverables, while theoretically focused lectures and hands-on lab sessions play the role of intermediaries in the learning process. In practice this means that most of the learning will happen cumulatively over the stages of the research project, with the lectures and tutorials focusing on providing the "bigger picture" and practical troubleshooting, respectively.

- Akınoğlu, O. and Tandoğan, R.Ö., 2007. The effects of problem-based active learning in science education on students' academic achievement, attitude and concept learning. Eurasia journal of mathematics, science and technology education, 3(1), pp.71-81.

- Nguyen, K., Husman, J., Borrego, M., Shekhar, P., Prince, M., Demonbrun, M., Finelli, C., Henderson, C., & Waters, C. Students' expectations, types of instruction, and instructor strategies predicting student response to active learning. Int. J. Eng. Educ. 33, 2–18 (2016).
- Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, Mary Pat Wenderoth. Active learning boosts performance in STEM courses. Proceedings of the National Academy of Sciences Jun 2014, 111 (23) 8410-8415; DOI: 10.1073/pnas.1319030111
- Yannier N, Hudson SE, Koedinger KR, Hirsh-Pasek K, Golinkoff RM, Munakata Y, Doebel S, Schwartz DL, Deslauriers L, McCarty L, Callaghan K, Theobald EJ, Freeman S, Cooper KM, Brownell SE. Active learning: "Hands-on" meets "minds-on". Science. 2021 Oct;374(6563):26-30. doi: 10.1126/science.abj9957.

Prerequisites:

You should be comfortable with high school math (particularly probability, calculus, and linear algebra) and have previously successfully completed an undergraduate level course in statistics and/or econometrics.

Diversity Statement:

As you may know, the Hertie School is committed to implementing a new Diversity and Inclusion Strategy. We strive to have an inclusive classroom but ask your informal feedback on inclusivity throughout the course.

3. Grading and Assignments

Composition of Final Grade:

Assignment 1: Problem sets	Deadline: As indicated on each problem set	Submit via Moodle	30%
Assignment 2: Project Proposal	Deadline: Before the start of Session 5	Submit via Moodle	10%
Assignment 3: Midterm Report	Deadline: Before the start of Session 8	Submit via Moodle	20%
Assignment 4: Final Report	Deadline: Before the start of Session 11	Submit via Moodle	30%
Assignment 5: Presentation	Deadline: Before the start of Session 12	Submit via Moodle	10%

Assignment Details

The assessment for the course consists of several individual assignments and a group project. The research project must be done in teams of 2-4 (individual submissions will not be accepted for the project). The aim is to develop research projects as close as possible to an academic publication in the area of applied machine learning and communicate your research to the broader public.

The aim of the assessment is three-fold:

- First, it will provide you with the opportunity to apply the concepts learned in this class creatively, which helps you with understanding material more deeply.
- Second, designing and working on a unique project in a team which is something that you will encounter, if you haven't already, in the workplace, and the project helps you prepare for that.
- Third, along with the opportunity to practice and the satisfaction of working creatively, students can use this project to enhance their portfolio or resume. We will discuss with individual project groups whether they can be turned into academic publications

Note about grading. There is no "perfect project." While you are encouraged to be ambitious, the most important aspect of this research project is your learning experience. Hence, you don't want to pick something that is too easy for you, but similarly, you don't want to choose a project which could be out of the scope of this class. The project proposal is not graded by how exciting your project is but based on whether you follow the objectives of the project proposal, project presentation, and project report. For instance, if your project ends up being unsuccessful – for example, if you choose to design a classifier and it doesn't achieve the desired accuracy – it will not negatively affect your grade as long as you are honest, describe the potential issues well, and suggest improvements or further experiments. Again, the objective of this project is to provide you with hands-on practice and an opportunity to learn.

Assignment Details

Assignment 1: Problem sets (30% of the total course grade)

- This class will have up to five regular homework assignments in the form of problem sets. The assignments consist of different machine learning problems that need to be solved by writing and running code. By completing the assignments students gain a deeper understanding of the topics covered in class, and experience in solving those problems.

Assignment 2: Project proposal (10%) – 3 pages and 5 references

- The main purpose of the project proposal is to receive feedback from the instructor regarding whether your project is feasible and whether it is within the scope of this class. Also, the project proposal offers a chance to receive useful feedback and suggestions on your project. The goal is for you to propose the research question to be examined, motivate its rationale as an interesting question worth asking, and assess its potential to contribute new knowledge by situating it within related literature in the scientific community.
- For the project, you will be working in a team consisting of 2-4 students. The members of each team will be randomly assigned by the instructor. If you have any concerns about working with someone in your group, please discuss it with the instructor.
- You must include a link to a GitHub repository containing the code of your project. Your repository must be viewable to the instructor by the submission deadline. If your repository is private, make it accessible to us. If your repository is not visible to us, your assignment will not be considered complete, so if you are worried, please submit well in advance of the deadline so we can confirm the repository is visible. Furthermore, we will assess individual contribution to the team, should such an issue arise, based on the frequency and quality of GitHub commits in your project repository, so make sure you start the repository as the very first stage of your project.
- After you have received feedback from the instructor and your project proposal has been graded, you are advised to stick to the project outline in the proposal as closely as possible. However, if there is a concept introduced in a later lecture, you have the option to modify your proposal, but you are not penalized if you do not. If you wish to update your project outline, talk to the instructor first.
- The LaTeX template for the proposal and detailed description of the content and the marking rubric will be made available on Moodle.

Assignment 3: Midterm report (20%) – 4 pages and 10 references

- By the middle of the course, students should present initial experimental results and establish a validation strategy to be performed at the end of experimentation. This serves as a project milestone. The milestone should help you make progress on your project, practice your technical writing skills, and receive feedback on both.
- Ultimately, your final report will be written in the same style as an ML research paper. For the midterm, we ask you to write a preliminary version of some sections of your final report. Producing a high-quality milestone is time well spent, because it will make it easier for you to write your final report. You might find that you can reuse parts of your project proposal in your milestone. This is fine, though make sure to act on any feedback you received on your proposal.
- The LaTeX template for the proposal and detailed description of the content and the marking rubric will be made available on Moodle.

Assignment 4: Final report (30%) – 8 pages and unlimited references

- The final report will include a complete description of work undertaken for the project, including data collection, development of methods, experimental details (complete enough

for replication), comparison with past work, and a thorough analysis. Projects will be evaluated according to standards for conference publication—including clarity, originality, soundness, substance, evaluation, meaningful comparison, and impact (of ideas, software, and/or datasets).
- You must include a link to a GitHub repository containing full replication code of your project.
- The LaTeX template for the proposal and detailed description of the content and the marking rubric will be made available on Moodle.

Assignment 5: Presentation (10%)

- At the end of the semester, teams will produce a blogpost (use this template: https://github.com/hertie-data-science-lab/distill-template), and pre-recorded video presenting the results of their work to the class and broader community. These will be posted on the Data Science Lab website.
- Detailed description of the presentation task will be made available on Moodle.

Late submission of assignments: For each day, the assignment is turned in late, the grade will be reduced by 10% (e.g., submission two days after the deadline would result in 20% grade deduction).

Attendance: Students are expected to be present and prepared for every class session. Active participation during lectures and seminar discussions is essential. If unavoidable circumstances arise which prevent attendance or preparation, the instructor should be advised by email with as much advance notice as possible. Please note that students cannot miss more than two out of 12 course sessions. For further information please consult the Examination Rules §10.

Academic Integrity: The Hertie School is committed to the standards of good academic and ethical conduct. Any violation of these standards shall be subject to disciplinary action. Plagiarism, deceitful actions as well as free-riding in group work are not tolerated. See Examination Rules §16 and the Hertie Plagiarism Policy.

Compensation for Disadvantages: If a student furnishes evidence that he or she is not able to take an examination as required in whole or in part due to disability or permanent illness, the Examination Committee may upon written request approve learning accommodation(s). In this respect, the submission of adequate certificates may be required. See Examination Rules §14.

Extenuating circumstances: An extension can be granted due to extenuating circumstances (i.e., for reasons like illness, personal loss or hardship, or caring duties). In such cases, please contact the course instructors and the Examination Office in advance of the deadline.

4. General Readings

- James, G., Witten, D., Hastie, T. and Tibshirani, R., 2021. An introduction to statistical learning. 2nd edition. Available here: https://www.statlearning.com [we will designate it as **ISL** throughout]
- Aurélien Géron (2019). Hands-On Machine Learning with Scikit-Learn and TensorFlow. 2nd edition. O'Reilly Media, Inc. Notebooks available here: https://github.com/ageron/handson-ml2 [we will designate it as **AG** throughout]

5. Session Overview

Session	Session Title
1	ML for Government and Policy
2	Ethical ML
3	End-to-End ML Project
4	ML landscape
5	Regression
6	Classification
7	Resampling and regularisation
8	Tree-based methods
9	Deep learning
10	Unsupervised learning
11	Multiple testing
12	Project Presentations

Make Up Week: 5 – 9.12.2022 – we'll use it if we need to reschedule a class

Final Exam Week: 12 – 16.12.2022 – no class

Midterm Exam Week: 24—28 Oct 2022 – no class

6. Course Sessions and Readings

All readings will be accessible on the Moodle course site before semester start. In the case that there is a change in readings, students will be notified by email.

All readings are intended to supplement the class session and facilitate your own learning. Core readings are more fundamental to the course content, please skim them before class to ensure you are familiar with the concepts they build on. Optional readings are intended to either provide more material on fundamental concepts or broaden your knowledge in the respective area, and it is highly recommended to at least skim them.

Session 1: Machine Learning for Government and Policy	
Learning Objective	We discuss the growing importance of machine learning and data science in government and policy. We cover opportunities and challenges that arise from embedding such systems in the delivery of public services and in decision making. We also discuss how the structure of the MDS programme epitomises these issues.
Core Readings	ACUS "Government by Algorithm: AI in Federal Administrative Agencies" https://law.stanford.edu/education/only-at-sls/law-policy-lab/practicums-2018-2019/administering-by-algorithm-artificial-intelligence-in-the-regulatory-state/acus-report-for-administering-by-algorithm-artificial-intelligence-in-the-regulatory-state/

	European Commission Joint Research Centre "AI Watch. European landscape on the use of AI by the Public Sector" https://ai-watch.ec.europa.eu/publications/ai-watch-european-landscape-use-artificial-intelligence-public-sector_en
Optional Readings	

Session 2: Ethical ML

Learning Objective	What happens when things go wrong? Or does your machine learning model do things that it shouldn't? We will cover these core questions in practical applications of machine learning. We also discuss current guidelines and standards for ethical applications of machine learning. We go through several practical examples of machine learning projects where things go wrong.
Core Readings	Fast.ai Data Ethics course, Lesson 3 "Ethical Foundations and Practical Tools" https://ethics.fast.ai/syllabus/index.html#lesson-3-ethical-foundations--practical-tools "Ethics, Transparency and Accountability Framework for Automated Decision-Making" UK Government guidance: https://www.gov.uk/government/publications/ethics-transparency-and-accountability-framework-for-automated-decision-making
Optional Readings	Fast.ai Data Ethics course, Lesson 1 "Disinformation" (https://ethics.fast.ai/syllabus/index.html#lesson-1-disinformation) and Lesson 2 "Bias and Fairness" (https://ethics.fast.ai/syllabus/index.html#lesson-2-bias--fairness)

Session 3: End-to-End Machine Learning Project

Learning Objective	We go through an end-to-end machine learning project. In all its gory detail but also providing the big picture, top-down view of different things that are relevant to a real-world machine learning project. We cover data preparation, model selection and training, fine-tuning, validation, and deployment. We introduce the concept of MLOps.
Core Readings	AG: Chapter 2
Optional Readings	Andriy Burkov, 2020. Machine learning engineering. True Positive Incorporated. Huyen, C., 2022. Designing Machine Learning Systems. " O'Reilly Media, Inc.".

Session 4: Machine Learning Landscape

Learning Objective	We discuss the types of machine learning systems and their main challenges. We introduce the importance of testing and validating in machine learning. We also discuss course projects.
Core Readings	ISL: Chapter 2
Optional Readings	

Session 5: Linear Regression

Learning Objective	We cover statistical foundations of linear regression. From simple linear regression to multiple linear regression. We discuss how to evaluate such models.
Core Readings	ISL: Chapter 3
Optional Readings	AG: Chapter 4

Session 6: Classification

Learning Objective	We discuss statistical models for classification tasks. From logistic regression to generalized linear models. We also discuss evaluation of such models.
Core Readings	ISL: Chapter 4
Optional Readings	AG: Chapter 3

Session 7: Resampling and regularisation

Learning Objective	We discuss the concepts of cross-validation, bootstrap, and shrinkage methods.
Core Readings	ISL: Chapters 5, 6.2-6.4
Optional Readings	

Session 8: Tree-based Methods

Learning Objective	We cover training and visualising decision trees, prediction from trees, class probabilities, bagging, random forests, boosting, and Bayesian additive regression trees.
Core Readings	ISL: Chapter 8
Optional Readings	AG: Chapters 6-7

Session 9: Deep Learning

Learning Objective	We discuss the links between biological neurons and artificial neurons, perceptron, multilayer perceptron and backpropagation,

	implementations of artificial neural networks, Convolutional Neural Networks, and Recurrent Neural Networks.
Core Readings	ISL: Chapter 10
Optional Readings	AG: Chapter 10

Session 10: Unsupervised Learning	
Learning Objective	We discuss Principle Component Analysis and clustering.
Core Readings	ISL: Chapter 12
Optional Readings	AG: Chapters 8-9

Session 11: Multiple Testing	
Learning Objective	We discuss hypothesis testing, the concept of multiple testing, and different error rates.
Core Readings	ISL: Chapter 13
Optional Readings	

Session 12: Project presentations	
Learning Objective	Communication of research results is a crucial component of successful data science practice. You will practice written (poster) and oral (presentation) communication of your project findings, and providing feedback to your peers.