Course Syllabus

Jump to Today

UNIVERSITY OF PENNSYLVANIA

SCHOOL OF SOCIAL POLICY & PRACTICE

Data Analytics for Social Policy Certificate

MSSP 608: Practical Machine Learning Methods

Spring 2022

-

Instructor Name

Sadiq Y. Patel

Instructor Contact Information

patel@hcp.med.harvard.edu (mailto:patel@hcp.med.harvard.edu)

adam2318@seas.upenn.edu (mailto:adam2318@seas.upenn.edu)

praankit@seas.upenn.edu (mailto:praankit@seas.upenn.edu)

TA OFFICE HOURS

Yiyasu Paudel
Tuesdays, 2-3:30 p.m. EST
https://upenn.zoom.us/j/98952587430 (https://upenn.zoom.us/j/98952587430)

Adam Alavi
Tuesday, 5-6:30 p.m. EST
https://upenn.zoom.us/j/91067746215 (https://www.google.com/url?q=https://upenn.zoom.us/j/91067746215&sa=D&source=calendar&ust=1643572781124563&usg=AOvVaw1mjY

(https://zoom.us/j/91055869299?pwd=OWx0SnQ3S1U5d3JSRGtGQm1OVjFUZz09)

Ankit Prabhu
Thursday, 11-12:30 p.m. EST
https://upenn.zoom.us/j/91595175442?pwd=S3JOTjdnRStLYTdSYm1kdWJ1REF2Zz09
(https://upenn.zoom.us/j/91595175442?pwd=S3JOTjdnRStLYTdSYm1kdWJ1REF2Zz09)

Aadit Patel
Friday, 3-4:30 p.m. EST
https://upenn.zoom.us/j/96321828329?pwd=YUlvUGs3ZGNDb0xnTm1uZ3ZtM3ZhUT09
(https://upenn.zoom.us/j/96321828329?pwd=YUlvUGs3ZGNDb0xnTm1uZ3ZtM3ZhUT09)

Course Description/Purpose

(https://zoom.us/j/91055869299?pwd=OWx0SnQ3S1U5d3JSRGtGQm1OVjFUZz09)
(https://zoom.us/j/91055869299?pwd=OWx0SnQ3S1U5d3JSRGtGQm1OVjFUZz09)

This course prepares students with no background in machine learning or data science to use tools from those fields effectively in applied contexts. By programming with libraries, students will build skills including feature representations of complex real-world datasets; training classification and regression models for prediction tasks; evaluation of machine learning model accuracy and error analysis; and reasoning about predictive models and making tradeoffs like bias vs. variance, granularity of annotations and predictions, and complexities of real-world labeled training data and systems, including the ethical application of predictive modeling to human-centered data.

Educational Objectives

After this course, students will be comfortable with the following skills:

- Breaking down real-world problems into machine learning tasks, inputs, and outputs.
- Using Python libraries to train classification, regression, and clustering models.
- Optimizing models through model selection, feature design, and hyperparameter tuning.
- Evaluating dataset inter-annotator reliability, trained model accuracy, and error
- Recognizing the capacity and limitations of models for fairness and
- Describing results of machine learning experiments with real-world implications.

Course Requirements and Expectations

World Circumstances: This is a very strange time to be teaching a class, and I'm sure it's weird for you to be taking classes, as well. This class is fully online this quarter. As a result, everyone is still learning what class should look like, and I am going to be flexible about exactly how this class operates. I will assume good intentions from you, and I hope you do the same for me.

Psychological or Disability Needs: In addition to any questions or needs beyond what is listed in the rest of this syllabus, if you have any guidance or assistance with counseling, accommodations for disabilities, or other services, resources are available through Student Disabilities Services and

disabilities, or other services, resources are available through Student Disabilities Services and Counseling and Psychological Services. If these resources are not meeting your needs as they relate to this course, please let me know so I can determine what I can do to help.

Attendance: Our class meetings will take place each week at the scheduled time for the course, and they will be recorded. Watching all the material is vital to learning the overall content of the course. If you are not able to attend a particular synchronous session, that is okay, but it is critical you watch the recording. However, not reviewing the materials at all, is a serious problem, and may be resolved through individual meetings with me or your educational advisor, and could lead to course failure, depending on circumstances.

Participation: When you do attend class, I expect you to be consistently engaged with what's going on, and to ask questions when you are confused. But if things are going on at home or in your apartment, feel free to step away for a few minutes; the recorded videos will still be there when you come back. Please mute your sound and turn off your video if you think there will be an extended interruption. I may occasionally post check-in polls and/or requests for feedback that will occur throughout the semester, distributed through Canvas. These will not be for credit but will shape the direction of the course, and I hope you take them seriously and help me by letting me know your opinions.

Assignments: This course will have weekly quizzes, five technical assignments, and a final project. Technical assignments will consist of tasks related to modeling data with machine learning methods, and student submissions may consist of source code, written responses, annotated data, or answers to specific questions about data, along with other artifacts, all of which will be evaluated both for accuracy and clarity in order to receive full credit. Weekly quizzes will cover materials from the prior week. Later in the semester, you will be working on a final project; most students will be expected to work in pairs. First, you will submit a project proposal where you choose a dataset to work on and investigate in-depth with programming. At the end of the semester, you'll submit a final report, which will have both a code and a written component. This project is the last step of the course; there is no final exam. The project proposal and project report will be graded primarily on your written report: your ideas, your presentation of data, and the way you integrate your technical results from your coding. However, the proposal and report will also be evaluated for professional presentation, and therefore your grade will also reflect how well you accomplished aspects of publishing your work for a broader audience and can include mechanical issues like readability and formatting. Projects that were completed as part of previous course, or that you are already working on for another course this semester, should not be proposed or submitted. If you need assistance with writing, resources are available at the Weingarten Learning Center, Marks Family Writing Center, and SP2 Academic & Writing Support. If you need further assistance on writing or technical presentation, you should contact me directly and explain what additional help you need.

Collaboration: Programming for technical assignments must be completed individually. You must be the one writing your own code. But this is a difficult set of material to grasp, in a new technical domain; some teamwork is appropriate in those contexts. Getting help from online resources like StackOverflow and Kaggle is a normal part of data science, and students are encouraged to talk with one another about the problems if it is helpful. When you submit homework where you reused code that you found online, you must include links to your sources. Student submissions that are almost identical to code from other students may result in penalties. When you work collaboratively on a homework, you should clearly state in your submission who you worked with and how the work was separated. For students who submit similar code without acknowledging collaborative work, all students may be penalized, regardless of who wrote the code first. In general, code is subject to the same academic integrity policies as other work. Students are strongly encouraged to work in pairs for the final project. Variations from the pair structure may be allowed -- either individual work, or groups of 3 -- but only with my permission. The project report should clearly state the division of labor and the tasks or work that each student will be working on. Each student must complete programming work as part of the final project; student pairs are not permitted to split work responsibilities between a technical (programming) student and non-technical (report writing) student. Projects completed in larger groups will be held to a higher standard than individual projects; substantially more work will need to be done, with more results, more programming, and a more thorough report, in order to receive the same grade as peers who worked in pairs.

Academic Integrity

Students are expected to adhere to the University's Code of Academic Integrity, available at https://catalog.upenn.edu/pennbook/code-of-academic-integrity/. Care should be taken to avoid academic integrity violations, including plagiarism, fabrication of information, and multiple submissions (see descriptions below).** Students who engage in any of these actions will be referred to the Office of Student Conduct, which investigates and decides on sanctions in cases of academic dishonesty.

1. Plagiarism: using the ideas, data, or language of another person or source without specific or proper acknowledgment. Example: copying, in part or in its entirety, another person's paper, article, or web-based material and submitting it for an assignment; using someone else's ideas without attribution; not using quotation marks where appropriate; etc.
2. Fabrication: submitting contrived or altered information in any academic exercise. Example: making up data or statistics, citing nonexistent articles, contriving sources,
3. Multiple submissions: submitting, without prior permission, any work submitted to fulfill another academic

**It is students' responsibility to consult the instructor if they are unsure about whether something constitutes a violation of the Code of Academic Integrity.

Recommended Text

There are no required texts for this course; however, we will be drawing heavily from the following texts for content and explanations.

Moline, S. (2019). *Hands-On Data Analysis with Pandas*. Packt Publishing.

- Available on Amazon for $34.19 (print) or $17.19 (ebook). https://amazon.com/Hands-Data-Analysis-Pandas-visualization/dp/1800563450/ (http://www.amazon.com/Hands-Data-Analysis-Pandas-visualization/dp/1800563450/)
- Available direct from the publisher

https://www.packtpub.com/product/hands-on-data-analysis-with-pandas/9781789615326 (http://www.packtpub.com/product/hands-on-data-analysis-with-pandas/9781789615326) for $44.99 (print) or $31.99 (ebook)

Raschka, S. and Mirjalili, V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and Tensorflow 2, 3d edition. Packt Publishing.

- Available on Amazon for $36.96 (print) or $17.19 (ebook). https://amazon.com/Python-Machine-Learning-scikit-learn-TensorFlow-dp- (http://www.amazon.com/Python-Machine-Learning-scikit-learn-TensorFlow-dp-) 1789955750/dp/1789955750
- Available direct from the publisher https://packtpub.com/product/python-machine-learning-third- (http://www.packtpub.com/product/python-machine-learning-third-) edition/9781789955750 for $37.99 (print) or $27.99 (ebook)

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

- Available on Amazon for rent: $22.98 (print) or $16.43 (ebook)
- Available on Amazon for sale: $48.96 (print) or $43.44 (ebook) https://amazon.com/Data-Mining-Practical-Techniques-Management-dp- (http://www.amazon.com/Data-Mining-Practical-Techniques-Management-dp-) 0128042915/dp/0128042915/

Getting Help / Office Hours

We will have a course Piazza for asking questions that can be answered for the whole class to see. You can also address private questions directly to the instructors. Use this link to access Piazza for the course. The instructors will do their best to respond promptly by Piazza or by email, if necessary. Depending on your question or concern, the instructors may schedule individual phone or Zoom calls to

resolve more complex requests. All the instructors (the professor and the two TAs) will be holding weekly office hour. The hours will be scheduled in the first week of class based on a student survey of available times. Office hours with each TA will begin in week 2.

Grading

30% Weekly Quizzes

50% Programming Assignments (5 assignments each 10%)

20% Final Project

Class Schedule

Week 1 (Jan 12): Data Representation

- Course overview
- Core premises of machine learning
- Classification and regression tasks
- Defining task granularity and output labels
- Features and pipelines

Week 2 (Jan 24): Decision Trees

- Learning from data
- Information theory core concepts
- Rule-based models
- Decision trees
- Bias vs. variance tradeoff

Week 3 (Jan 31): Experimental Setup

- Intro to Scikit-learn
- Training and test data, Overfitting
- Model quality metrics and model comparison

Week 4 (Feb 7): Linear and Logistic Regression

- Core probability concepts
- Linear Regression
- Logistic Regression

Week 5 (Feb 14): Optimization

- Cross-validation
- Development sets
- Hyperparameters, grid search

- Hyperparameters, grid search
- Generalizing to multi-class classification

ASSIGNMENT 1 (DUE: February 21)

Week 6 (Feb 21): Ensembles

- Learning with ensembles
- Combining predictions from different classifiers
- Bagging and boosting
- Random forests

Week 7 (Feb 28): Should you use machine learning?

- Heuristics and baselines
- Instance-Based Models (kNN)
- Testing basic assumptions about datasets

ASSIGNMENT 2 (DUE: March 7)

NO CLASS MARCH 7 - SP2 SPRING BREAK

Week 8 (March 14): Unsupervised Learning and Exploratory Analysis

- Goals and reasons for exploratory analysis
- k-Means algorithms
- Hierarchical Clustering
- Principal Component Analysis
- Evaluation Metrics

Week 9 (March 21): Explainability and Fairness

- Definitions of Fairness
- Conducting Audits
- Discussion of current trends, topics, and issues
- Ethical decision-making case studies

ASSIGNMENT 3 (DUE: March 25)

Week 10 (March 28): Project Guidance

- Guidelines for final project, project scope and grading rubric
- Expectations for machine learning reproducibility
- Example projects walkthrough

Week 11 (April 4): **Text as Data**

- Representations of text
- Classifiers for text data
- Natural language processing methods
- Fairness concerns in text data

ASSIGNMENT 4 (DUE: April 11)

Project Proposal (DUE: April 11)

Week 12 (April 11): Dataset Collection

- Corpus building and annotation manuals
- Measuring and improving inter-rater reliability
- Fairness concerns in data collection and labeling

Week 13 (April 18): Images as Data

- Representations of image data
- Overview of deep learning basics
- Fairness concerns in image data

ASSIGNMENT 5 (DUE: April 22)

Week 14 (April 25): Images as Data

- Representation of image data
- Convolutional Neural Networks (CNN) and its building blocks
- Implementing a CNN and regularizing it
- Fairness concerns in image classification

Final Project DUE: APRIL 29th

Course Summary:

Date	Details	Due
Wed Jan 12, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490444&include_contexts=course_1637996)	1:45pm to 4:15pm
Mon Jan 24, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490445&include_contexts=course_1637996)	1:45pm to 4:15pm
Sun Jan 30, 2022	Quiz 1 (https://canvas.upenn.edu/courses/1637996/assignments/9727180)	due by 12pm
Mon Jan 31, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490446&include_contexts=course_1637996)	1:45pm to 4:15pm
Tue Feb 1, 2022	Quiz 2 (https://canvas.upenn.edu/courses/1637996/assignments/9727174)	due by 11:59pm
Mon Feb 7, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490447&include_contexts=course_1637996)	1:45pm to 4:15pm
Tue Feb 8, 2022	Quiz 3 (https://canvas.upenn.edu/courses/1637996/assignments/9727172)	due by 11:59pm

Date	Details	Due
Mon Feb 14, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490448&include_contexts=course_1637996)	1:45pm to 4:15pm
Tue Feb 15, 2022	Quiz 4 (https://canvas.upenn.edu/courses/1637996/assignments/9727175)	due by 11:59pm
Mon Feb 21, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490449&include_contexts=course_1637996)	1:45pm to 4:15pm
Tue Feb 22, 2022	Quiz 5 (https://canvas.upenn.edu/courses/1637996/assignments/9727173)	due by 11:59pm
Wed Feb 23, 2022	Homework 1 (https://canvas.upenn.edu/courses/1637996/assignments/9727185)	due by 11:59pm
Mon Feb 28, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490450&include_contexts=course_1637996)	1:45pm to 4:15pm
Tue Mar 1, 2022	Quiz 6 (https://canvas.upenn.edu/courses/1637996/assignments/9727179)	due by 11:59pm
Wed Mar 9, 2022	Homework 2 (https://canvas.upenn.edu/courses/1637996/assignments/9727186)	due by 11:59pm
Mon Mar 14, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490452&include_contexts=course_1637996)	1:45pm to 4:15pm
Tue Mar 15, 2022	Quiz 7 (https://canvas.upenn.edu/courses/1637996/assignments/9727182)	due by 11:59pm
Mon Mar 21, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490453&include_contexts=course_1637996)	1:45pm to 4:15pm

Date	Details	Due
Tue Mar 22, 2022	Quiz 8 (https://canvas.upenn.edu/courses/1637996/assignments/9727181)	due by 11:59pm
Fri Mar 25, 2022	Quiz 8 (https://canvas.upenn.edu/courses/1637996/assignments/9727184) (1 student)	due by 11:59pm
	Homework 3 (https://canvas.upenn.edu/courses/1637996/assignments/9727187)	due by 11:59pm
Mon Mar 28, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490454&include_contexts=course_1637996)	1:45pm to 4:15pm
Tue Mar 29, 2022	Quiz 9 (https://canvas.upenn.edu/courses/1637996/assignments/9727177)	due by 11:59pm
Mon Apr 4, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490455&include_contexts=course_1637996)	1:45pm to 4:15pm
Fri Apr 8, 2022	Homework 4 (https://canvas.upenn.edu/courses/1637996/assignments/9727188)	due by 11:59pm
Mon Apr 11, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490456&include_contexts=course_1637996)	1:45pm to 4:15pm
Tue Apr 12, 2022	Quiz 10 (https://canvas.upenn.edu/courses/1637996/assignments/9727176)	due by 11:59pm
Mon Apr 18, 2022	MSSP 608-001 2022A Practical Machine Learning Methods (https://canvas.upenn.edu/calendar?event_id=3490457&include_contexts=course_1637996)	1:45pm to 4:15pm
Tue Apr 19, 2022	Quiz 11 (https://canvas.upenn.edu/courses/1637996/assignments/9727178)	due by 11:59pm

Date	Details	Due
Fri Apr 22, 2022	**Homework 5** (https://canvas.upenn.edu/courses/1637996/assignments/9727190)	due by 11:59pm
Mon Apr 25, 2022	**MSSP 608-001 2022A Practical Machine Learning Methods** (https://canvas.upenn.edu/calendar?event_id=3490458&include_contexts=course_1637996)	1:45pm to 4:15pm
Mon May 2, 2022	**Final Project** (https://canvas.upenn.edu/courses/1637996/assignments/9727184)	due by 11:59pm
	Roll Call Attendance (https://canvas.upenn.edu/courses/1637996/assignments/9892573)	