THE UNIVERSITY OF CHICAGO
HARRIS SCHOOL OF PUBLIC POLICY

**PPHA 30546: Machine Learning - Python**
**Dr. Christopher Clapp**
**Syllabus, Winter 2023**

Meetings:
Class:
Section 01 - MW 10:30-11:50am
Section 02 - MW 1:30-2:50pm

Locations:
Keller 0001
Keller 0021

Lab Sessions:
Lab 01 - F 10:30-11:50am or
Lab 02 - F 1:30-2:50pm

Locations:
Lab 01 - Keller 0001
Lab 02 - Keller 0001

Professor: Chris Clapp (he/him)
Office Hours: F 3:30-4:30pm
or by appointment

Email: cclapp@uchicago.edu
Location: Keller 3039

Head TA: Steve Kim (he/him)
Office Hours: W 12:00-1:00pm

Email: kimsy@uchicago.edu
Location: Keller 2058

TAs:
Jonas Heim (he/him)
Office Hours: Tu 8:00-9:00am
Victor Perez (he/him)
Office Hours: Th 9:30-10:30am
Pavan Prathuru (he/him)
Office Hours: M 3:30-4:30pm
Sergio Olalla (he/him)
Office Hours: W 9:00-10:00am
Pedro Ramonetti (he/him)
Office Hours: Th 2:00-3:00pm

Email: jonas.heim@uchicago.edu
Location: Keller 2058
Email: vperezmartin@uchicago.edu
Location: Zoom
Email: pavanprathuru@uchicago.edu
Location: Zoom
Email: sergiou@uchicago.edu
Location: Keller 2105
Email: pramonetti@uchicago.edu
Location: Keller 2050

## Course Description

It's an exciting time to study machine learning and data science more generally! We live in a digital era where many of our decisions and actions are tracked. Information is being produced and recorded at a stifling pace. While this may not seem novel to those who were born and have grown up in the Information Age, the amount of data available to researchers and policymakers is orders of magnitudes of more than what existed even a decade ago. Coupled with cheap computing power and expanded data storage, recent developments across statistics, computer science, and data-driven social sciences allow us to use all this data in a myriad of interesting ways. But what questions will we seek to answer with this newly available *big data* and these newly developed *machine learning* tools?

While these tools are already being used extensively in marketing, finance, and business, their application to public policy is in its infancy (despite the techniques being the same across disciplines). Early examples of

questions with policy implications include: can we predict unavailable data we take for granted in the developed world from available information in a developing world context? Is it possible to improve the accuracy of judges' bail decisions that hinge on whether the accused will commit additional crimes? Or can we inform doctors about the trade-offs inherent in prescribing potentially addictive opioids to patients for short-term pain relief by predicting who is likely to develop an addiction in the long run?

In order to ask and inform questions like these, this class will introduce you to ways to detect patterns in data, then use what you have learned to predict important outcomes or describe the salient relationships among inputs. While this requires an understanding of how and why these tools work, we will emphasize the intuition and application of these techniques over their theoretical underpinnings. We will do so by exploring nascent, policy-relevant applications of these methods, but, ultimately, the full impact of how these machine learning techniques inform and influence policy has yet to be determined. That's up to you!

## Learning Objectives: "What's My Incentive for Taking This Course?"

Specifically, the purpose of the course is to introduce you to a wide array of the fundamental methods in modern machine learning. Each week, we will learn about and discuss a different set of techniques and their applications to public policy during lecture sections. During lab sessions, you will gain experience with those techniques by coding their implementation in Python.

Along the way you can expect to:

- Apply machine learning techniques to carry out policy-relevant analyses.
- Understand how the machine learning approach, which focuses on prediction, differs from the approach to fundamental statistical and/or causal inference you learned in your Core statistics classes.
- Gain an appreciation of why the bias-variance trade-off makes prediction inherently difficult.
- Recognize the different ways "long" and "wide" big data allow us to improve our predictions.
- Continue developing your coding skills in Python as you learn new tools.
- Visualize, interpret, and convey your findings to audiences of different levels of technical sophistication.

The overall course objective is for you to be able to use machine learning tools to inform better policy and make the world a better place, as well as to become an informed and critical consumer of policy recommendations based on machine learning techniques. Additionally, the course will allow you to market your newly gained machine learning knowledge and skills when applying for jobs.

## Prerequisites

The official prerequisites are:

- PPHA 30537 Data and Programming for Public Policy I - Python Programming and
- PPHA 30538 Data and Programming for Public Policy II - Python Programming.

This course is the third installment of the three-quarter core sequence of the Certificate in Data Analytics (https://harris.uchicago.edu/academics/design-your-path/certificates/certificate-data-analytics) at Harris. Students at Harris and from other parts of the University may enroll without having taken previous courses in the sequence after students who haven those classes have had a chance to enroll. However, it is necessary for MPP students to take the full sequence in order to meet the necessary requirements of the Certificate in Data Analytics.

For anyone who has not taken the prerequisites and is considering taking this course, first, thanks for your interest in my class! This course introduces machine learning techniques, then has students practice and apply them via Python coding-based labs, problem sets, and mini-projects. So while the class doesn't directly follow the prerequisites (which teach general coding skills in Python), you will be responsible for knowledge of the material covered in those classes. I allow students to waive the prerequisites if they have sufficient experience coding in Python and are aware that they may be at a bit of a disadvantage relative to the majority of the students in the class who have taken the prerequisites. If you are considering taking the class out of sequence, I would recommend looking over the syllabi for the prerequisite classes and making sure that you're comfortable with the topics and techniques that are covered before making your decision on whether or not to enroll.

## Evaluation

Your final grade in this course will be related to performance in several areas. The weight placed on each component will be as follows:

| | |
|---|---|
| Problem Sets (4) | 50% |
| Mini-Projects (4) | 50% |
| Participation (Extra Credit) | 02% |

There are four problem sets and four mini-projects in this class. Both assignments will be submitted on Cavnvas via the Gradescope option. You may submit assignments late for up to 24 hours after the due date with a four percentage point deduction per hour. These deductions are not fractional (e.g. turning an assignment in one second or 59 minutes and 59 seconds late will result in an four percentage point deduction). I will drop the lowest grade among these assignments when calculating your grade.

Problem sets will consist of more structured questions (primarily) from the textbook. They are designed to help students cement their understanding of the conceptual material covered in lecture and get practice both applying the tools we learn and with coding.

Mini-projects are designed to apply the machine learning concepts and tools covered in class to policy-relevant questions. As such, they are less structured, based on "real-world" data, and emphasize application to public policy over statistical concepts.

You are welcome (and encouraged) to form study groups of no more than 2 students to work on the problem sets and mini-projects together. But you must write your own code and your own solutions. Please be sure to include the names of those in your group on your submission. Please also be sure to practice the good coding practices you learned in the Data and Programming classes and comment your code, cite any sources you consult, etc.[1]

Class participation points will be based on your level of active, attentive, inquisitive participation during in-class discussions and/or on the discussion board. For in-class participation, note that regular class attendance

[1]The focus of the class is on applying machine learning techniques. So your focus in completing the assignments should be on developing and demonstrating your ability to apply those techniques. Part of both doing and demonstrating that requires using good coding style (in part because it makes it easier for the graders to see that you understand what you're doing). So while good coding style is secondary to applying the ML techniques, we may take points off if the code is hard to follow.

is generally a necessary (but not sufficient) component of earning in-class participation points. Additionally, to earn credit, you must record each instance of your participation (e.g., when you ask a question, provide an answer, contribute to a class discussion, etc.) using the submission form linked on the main Canvas course page.[2] Please submit a separate entry each time you participate. You only need a brief description of your question/answer/etc. (enough to jog my memory) and you should record all participation within 24 hours after class ends. You do not need to record participation via the discussion board - just your in-class participation!

We will supplement in-class participation with the Ed Discussion discussion board on Canvas. Please use the discussion board to post questions, discuss the material covered in the lectures or on the assignments, and answer questions posed by your peers. As being a good colleague is both an important way to have social impact and is valued by employers, participation points can be earned by making posts that are helpful to your peers.[3] While this can take many forms, points will primarily be awarded for answering classmates' questions on the discussion board. In doing so, you may not explicitly share code, provide step-by-step solution algorithms (e.g., pseudo code), or direct solutions. You may clarify ambiguities in the assignments, discuss conceptual aspects of lectures or problems, show output and error messages, and provide general guidance on how to correct errors in understanding or code.[4] Additionally, you may post brief summaries of news articles that describe applications of machine learning techniques to public policy relevant issues.[5]

## Grades

Grades in this class will be distributed according to the following intervals.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | $[95\% - 102\%]$ | A- | $[90\% - 95\%)$ | B+ | $[85\% - 90\%)$ | B | $[80\% - 85\%)$ | B- | $[60\% - 80\%)$ |

Pass/Fail (P/F), Withdrawal, and Incomplete grade requests will be handled in accordance with University and Harris policy. Students who wish to take the course pass/fail rather than for a letter grade must use the Harris P/F request form (https://harris.uchicago.edu/form/pass-fail) and must meet the Harris deadline, which is generally 9am on the Monday of the 5th week of courses. To earn a P grade, students taking the course P/F must: submit at least seven of the eight assignments and earn a grade that is overall equivalent to at least a C- letter grade.

## Materials

### Textbooks

- Required: *An Introduction to Statistical Learning*, 2nd Edition, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. (ISBN-10: 1071614177)
  - You can download a free PDF of the book from the author's website: https://www.statlearning.com/.
  - Coding examples in the book are written in R, but you can find Python analogs here: https://github.com/JWarmenhoven/ISLR-python.

In addition, I may distribute copies of additional readings via Canvas.

[2] You will have to be logged into your UChicago Google account to submit a response.

[3] Note that grades do not follow a curve in this class, so there is no penalty for helping others.

[4] For instance, a response to a peer that says, "to fix your error, the command should be '[...]'" is not permitted. Instead, saying, "I think you have a typo in the third argument of your command" is acceptable.

[5] Please note that in practice, the different means of class participation will be evaluated on an "either/or" basis. You are not required to participate in class via all possible modes of communication, although you are welcome to. There are multiple ways to participate because I want to give students as many opportunities to earn credit as possible, not because I want you to feel overwhelmed.

### Data Analysis and Statistical Software

We will use Python software in this class, and you are required to code all assignments in Python.[6] Please install Anaconda Python version 3.x from www.anaconda.com/ distribution. The text editor or integrated development environment (IDE) you use for writing code for assignments is up to you, but the use of Jupyter Notebooks is discouraged.

## Office Hours

My office hours for this class are listed on the first page of the syllabus. Those hours are for you, so please make use of them (be it with questions about course material, to discuss ideas, or just to chat). You do not need to make an appointment to see me during my office hours; just drop by. I will be available during those times. If a sufficient number of students attend at the same time and office hours become too crowded to be effective, we will make alternative arrangements.

Please make your best effort to attend during the posted times, but if you have a conflict or want to talk with me one-on-one, you are welcome to make an appointment for another time. I am happy to meet with students outside of office hours. I only ask that you do your absolute best to attend the regularly scheduled office hours since I have many students and there are economies of scale in the production of knowledge. Also, if you know in advance that you cannot make a scheduled appointment, please email me to let me know.

## Harris Tutoring Program

Harris offers 10 hours of free tutoring support for coding in Python, Stata, and R. Tutoring will be available to Harris students starting Week 3 of the quarter and you can read more about the program on the Harris Student Handbook Canvas site (https://canvas.uchicago.edu/courses/42004/pages/harris-tutoring-program). Any questions should be directed to your academic advisor or harrisdeanofstudents@uchicago.edu.

## Course Policies

- **General**
  - The class will be taught in-person with a remote option for students needing temporary accommodations for short-term absences. Should changing pandemic conditions necessitate, we will switch to holding class remotely according to University policy and my discretion. Student input will be welcomed in making this determination.
  - There is no attendance requirement, but regular attendance is necessary (but not sufficient) to do well in the class.
  - The class webpage is available through the Canvas portal. I will use it to post announcements, assignments, and grades. Please check it regularly.
  - Email, Canvas postings, and the discussion board are the official means of communication for out-of-class messaging. In other words, you are expected to check your UChicago email account and the Canvas site regularly.
  - Email is inefficient. If you have a question about the class or the material, others probably do too! Questions and answers (knowledge) are public goods, so post your question to the discussion board, and feel free to answer questions your classmates ask. The TAs and I will monitor and respond as well.

[6] Note that there is an analog of this class that is taught by another instructor with R software (PPHA 30545).

- If you have a question or concern about something you don't want to discuss publicly, feel free to email me. I will respond to email within 2 business days (Monday-Friday, 9:00am-5:00pm). I teach multiple classes, so please include "ML:" as a prefix to your subject.
- Any and all results of in-class and out-of-class assignments and examinations are data sources for research and may be used in published research. All such use will always be anonymous.

- **COVID-19 Pandemic**
  - Students are expected to abide by the University's health protocols. Note that the protocols, which address masking, self-monitoring, testing, reporting, and isolating requirements, represent evolving guidance and are subject to change (https://goforward.uchicago.edu/). Masks are currently not required in class, but their use is recommended and appreciated.
  - My expectation is that students will attend class in-person. **That said, if you are experiencing COVID-19 symptoms or are required to quarantine, please do not attend class in person!** I will live-stream and record classes on Zoom in order to make this easier. I will also enable remote (dual-modality) participation concurrent with our in-person classes.
  - If you need a more-permanent remote learning accommodation, please contact the Dean of Students, Kate Biddle (kbiddle@uchicago.edu). Per Harris policy, all such requests can only be approved centrally, not by individual instructors. More generally, if you get sick, are caring for a sick relative, or anything else that becomes an obstacle to your coursework, please inform me and your advisor as soon as you are able. We will all work together to develop appropriate accommodations.
  - If I am experiencing COVID symptoms or are required to quarantine and must teach class remotely, I will notify you via Canvas as soon as I am able to. Health permitting, I will teach remotely via Zoom on such occasions. You will be able to attend class from home or from our regular classroom (but would participate via Zoom on such days). If I am too sick to teach, we will find ways to make up the class or I will endeavor to find a substitute instructor.
  - Please use your name tents so that I can learn names, easily recognize you despite your face masks, and call on you by name.

- **Recording**
  - I will record all lectures and post them only to Canvas in accordance with University and Family Educational Rights and Privacy Act (FERPA) guidelines.
  - The University has developed specific policies and procedures regarding the use of video/audio recordings that are explicitly described in the University's student manual (https://studentmanual.uchicago.edu).
  - FERPA is a federal statute that, broadly speaking, guarantees privacy over certain aspects of your educational records. You can view the details of the policy on the registrar's website (https://registrar.uchicago.edu/records/ferpa/).
  - If you record a class, discussion section, office hours, or meeting without permission, or if you share any of the recorded videos without permission, you may be violating eavesdropping laws, copyright laws, or the FERPA statute. So do not post or share any such videos outside of Canvas. This also applies to any manipulated video.

- **Assignments**
  - No assignments will be accepted after the 24 hour late period for any reason, valid or otherwise.[7] Not turning in an assignment, handing it in more than 24 hours late, or failing to turn it in before the

[7] Reasons include, but are not limited to: illnesses, athletic competitions, work trips, job fairs, job interviews, travel reservations, relative illnesses, relative funerals, out-of-town weddings, car accidents, car trouble, scooter trouble, tickets to see Billy Joel in concert, and emergency visits to the veterinarian with your dog.

link expires will result in a grade of zero. I am bound by the terms in this syllabus, so please do not email me to ask for an extension! I understand that students sometimes have legitimate reasons for being unable to complete assignments on time or give their full effort, so your lowest assignment grade will be dropped.

– Due to the ongoing pandemic, to ensure that students who have medical issues or need to care for sick family members for an extended period of time do not automatically fail the class, I will allow students to write a paper of no more than 10 pages on the topic covered on the missed assignment as a grade replacement. The details of these papers will be shared should they become necessary. Following the design of many of our social insurance programs, these papers will be designed to be optimal (relative to the standard assignment) only for students who truly need to make use of this option.

## Academic Integrity[8]

As a member of the Student Government Judicial Branch as an undergraduate and a graduate student at a university where any non-trivial act of lying, cheating or stealing results in expulsion, I take the Harris Academic Honesty and Plagarism Policies (https://harris.uchicago.edu/student-life/dean-of-students-office/policies) very seriously. All students suspected of academic dishonesty will be reported to the Harris Dean of Students for investigation and adjudication. The disciplinary process can result in sanctions up to and including suspension or expulsion from the University. In addition, if in my judgment, the preponderance of the evidence indicates that a student has committed an honor violation on an assignment, that student will receive an immediate grade of zero for that assignment and cannot earn a grade higher than a B- in the course, regardless of their performance on other assignments. This is regardless of the outcome of the disciplinary process. I trust every student in this course to fully comply with all of the provisions of UChicgo and Harris' integrity policies. Here are specific expectations:

- On assignments, it is expected that you will neither receive nor give aid, nor access any material other than items explicitly outlined in the instructions.
- For other assignments, you may (and should!) work with other students, but it is expected that you will collaborate on all parts of the assignment (as opposed to the "divide and conquer" method).
- During the entire semester, it is expected that you will not access old problem sets, projects, answer keys, or any other class material at any time. This includes websites that post solutions under the guise of tutoring. (These sites both facilitate cheating and steal the intellectual property of the author.) This does not include the textbook authors' websites, Python documentation, or StackOverflow.
- During the entire semester and thereafter, it is expected that you will neither post any class material on the internet nor share any class materials with other students through any other means. Furthermore, if you become aware that this has occurred, you are obligated to let me know immediately.

### Americans With Disabilities Act

Students with disabilities needing an academic accommodation should contact UChicago's Student Disability Services (SDS). Please see their webpage for contact information (https://disabilities.uchicago.edu). If SDS determines a disability accommodation is appropriate, you should inform the Harris Dean of Students office by

[8] I apologize for the heavy handed tone of this section. It is intended to protect the many honest students who take my class and academic integrity as a whole.

the end of the first week of class. The Harris Dean of Students office will work with the student and instructor to coordinate the students' accommodations implementation. Harris students are not required to submit their accommodations letter to the instructor, but please feel free to come talk to me if you are comfortable doing so. I'm happy to help.

**Mental Health Services**

Students differ in how much they know about mental health services. Your use of UChicago's Student Health and Counseling Services (SHCS) is free, confidential, and not linked to your academic file. There is nothing to be gained from suffering in silence, so please do not hesitate to make use of the services provided by SHCS if you need them. Please see SHCS' mental health webpage for services and contact information (https://wellness.uchicago.edu/ mental-health/). And if you are having serious mental, physical, or other problems, immediately contact the urgent medical care line at (773) 702-3625 (available 24 hours a day, 7 days a week).

**Diversity and Inclusion**

UChicago is committed to diversity and rigorous inquiry that arises from multiple perspectives, and Harris encourages thought-provoking discourse that involves not only speaking freely about all issues but also listening carefully and respectfully to the views of others. I concur with this commitment and view the diversity that students bring to my class as a valuable resource and a benefit to learning. I expect to maintain a productive learning environment based on open communication, mutual respect, and non-discrimination. I strive to present materials in a way that is respectful of diverse student backgrounds. As there can always be a gap between intent and execution, suggestions for promoting a positive and open environment are welcomed. Please feel free to correct me on your preferred name and gender pronouns if necessary.

**Responsible Employees (Title IX)**

All University of Chicago faculty and TAs are classified as "Responsible Employees." As such, they are required to report any discussions of sexual misconduct, dating violence, domestic violence or stalking to the Title IX Coordinator for the University. This includes the identities of the student making the complaint and alleged perpetrator. You will receive an email once a report is filed, but you are not obligated to meet with anyone or engage in the process. Alternatively, there are "Confidential Resource" employees at the University who do not have an obligation to share identifying information. For more information, including phone numbers, see the UChicago UMatter website (https://umatter.uchicago.edu/find-support/).

**Syllabus Change Policy**

Except for changes that substantially affect implementation of the evaluation (grading) statement, this syllabus is a guide for the course and is subject to change with advance notice.

## Tentative Course Outline

The weekly coverage might change as it depends on the progress of the class. The "ISL" in the "Reading" column that follows indicates the chapter in the "An Introduction to Statistical Learning" textbook that corresponds to the topic we're covering in class that day. "PS" is an abbreviation for "Problem Set," and "MP" is an abbreviation for "Mini-Project."

**Tentative Course Schedule**

| Week | Date | Day | Topic | Reading | Due |
|---|---|---|---|---|---|
| 1 | 01/04 | Wed | Introduction & Statistical Learning | ISL Ch. 1 & 2 | |
| | **01/06** | **Fri** | **[Lecture held during lab section]** | | |
| 2 | 01/09 | Mon | Linear Regression & Moving Beyond Linearity | ISL Ch. 3 & 7 | |
| | 01/11 | Wed | | | |
| 3 | **01/16** | **Mon** | **MLK, Jr. Holiday! - No Class** | | |
| | 01/18 | Wed | Classification | ISL Ch. 4 | PS 1 |
| 4 | 01/23 | Mon | | | |
| | 01/25 | Wed | Resampling Methods | ISL Ch. 5 | MP 1 |
| 5 | 01/30 | Mon | | | |
| | 02/01 | Wed | | | PS 2 |
| 6 | 02/06 | Mon | Linear Model Selection & Regularization | ISL Ch. 6 | |
| | 02/08 | Wed | | | MP 2 |
| 7 | 02/13 | Mon | | | |
| | 02/15 | Wed | Tree-Based Methods | ISL Ch. 8 | PS 3 |
| 8 | 02/20 | Mon | | | |
| | 02/22 | Wed | Support Vector Machines | ISL Ch. 9 | MP 3 |
| 9 | 02/27 | Mon | | | |
| | 03/01 | Wed | Deep Learning | ISL Ch. 10 | PS 4 |
| Exam Week | | | | | MP 4 |

Please note that we lose two days of instruction on Mondays this quarter because the academic calendar starts on a Tuesday (January 3) rather than a Monday and there is no class on Monday, January 16 due to the Martin Luther King, Jr. Day holiday. Harris administration has indicated that all courses should offer a full nine weeks' worth of content, regardless of differences in the academic calendar across days of the week. Thus, I want to flag a few "nonlinearities" in the schedule (flagged in **bold** in the schedule) as a result of following Harris policy:

- I will lecture during the regularly scheduled Friday lab sessions the first week of classes.
- I will record an asynchronous lecture that will cover two "Classification" topics: the linear probability and logit models. Previous students have told me that this material is review.