Syllabus for HKS API-222
Machine Learning and Big Data Analytics
Spring 2023

Draft Version: February 14, 2023
Subject to Change

Class Time: Tuesdays and Thursdays, 12:00pm–1:15pm
Class Location: Starr Auditorium (Belfer 200, floor 2.5)
Review Section: Fridays 12:00pm–1:15pm
Review Section Location: Rubenstein, Room 304 (R-304)

Faculty: Prof. Soroush Saghafian
Email: soroush_saghafian@hks.harvard.edu
Office Hours: Thursdays 3:00pm–5:00pm (HKS Littauer Bldg, Office 205); Sign-up needed; Please click here to sign up. The sign-up link is also on canvas.

Faculty Assistant: Catherine Kearns
Email: catherine_kearns@hks.harvard.edu
HKS Littauer Bldg, Office 211 (L–211)

Teaching Fellow (TF): Ibrahima (Ibou) Dieye
Email: ibrahimadieye@g.harvard.edu
Office Hours: Wednesdays 12:30pm–2:30pm. Location: HKS Littauer Bldg, Room 330 (L–330)

Course Assistant (CA): Omar Olivarez
Email: oolivarez@hbs.edu
Office Hours: Mondays 3:30pm–5:30pm. Location: HKS Littauer Bldg, Room 330 (L–330)

Course Assistant (CA): Sidian Sylvia Lin
Email: sidianlin@g.harvard.edu
Office Hours: Tuesdays 4:00pm–6:00pm. Location: HKS Littauer Bldg, Room 330 (L–330)

Course Assistant (CA): Jacob Jameson
Email: jacobjameson@g.harvard.edu
Office Hours: Wednesdays 4:30pm–6:30pm. Location: Rubenstein, Room R-G-21 NEUSTADT; **HKS Forum on Feb 15**

Course Assistant (CA): Jukta Basu Mallik
Email: juktabasumallik@hks.harvard.edu
Office Hours: Fridays 3:00pm–5:00pm. Location: 124 MT. AUBURN - SUITE 230S, ROOM 2000

Brief Course Description

In the last couple of decades, the amount of data available to organizations has significantly increased. Individuals who can use this data together with appropriate analytical techniques can discover new facts and provide new solutions to various existing problems. This course provides an introduction to the theory and applications of some of the most popular machine learning techniques. It is designed for students interested in using machine learning and related analytical techniques to make better decisions in order to solve policy and societal level problems.

We will cover various recent techniques and their applications from both supervised and unsupervised learning. In addition, students will get the chance to work with some data sets using software and apply their knowledge to a variety of examples from a broad array of industries and policy domains. Some of the intended course topics (time permitting) include: *K-Nearest Neighbors, Naive Bayes, Logistic Regression, Linear and Quadratic Discriminant Analysis, Model Selection (Cross Validation, Bootstrapping), Support Vector Machines, Smoothing Splines, Generalized Additive Models, Shrinkage Methods (Lasso, Ridge), Dimension Reduction Methods (Principal Component Regression, Partial Least Squares), Decision Trees, Bagging, Boosting, Random Forest, K-Means Clustering, Hierarchical Clustering, Neural Networks, Deep Learning, and Reinforcement Learning.*

Suggested Prerequisites

An understanding of intermediate-level statistics and probability theory (e.g., API-201, API-202, or equivalent courses).

Overview of API-222

API-222 provides an introduction to the theory and applications of some of the most popular machine learning techniques with an emphasis on the ways in which they are applied to practical policy-making and decision-making problems.

Our goal is that by the end of this course you will:

- Have a basic understanding of the statistical theory underlying common supervised and unsupervised machine learning algorithms.

- Have an understanding of when and how to apply different types of machine learning algorithms to policy-making and decision-making issues.
- Have the ability to use software to apply machine learning tools to solve real-world problems.

The course content is divided into two broad units: (1) Supervised Learning (first part of the course), (2) Unsupervised Learning (second part of the course).

The course also provides you with an opportunity to become familiar with using `R` (also `Mathematica` or `Python`, if you choose so) to analyze data using machine learning techniques. We recommend using `RStudio` as an interface for `R`, which is free and can be found at: https://www.rstudio.com. Instead of `R`, you can choose `Mathematica` or `Python`. However, the teaching staff will mainly focus on providing support for learning `R`.

Suggested Books and Fun Readings

Suggested Books

- [JWHT] James, Witten, Hastie and Tibshirani. 2015. *An Introduction to Statistical Learning with Applications in R.* Springer Science. [Available for free online]
- [HTF] Hastie, Tibshirani and Friedman. 2013. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* (2nd ed). Springer Series in Statistics. Available for free: https://web.stanford.edu/ hastie/Papers/ESLII.pdf
- [M] Murphy. 2012. *Machine Learning: A Probabilistic Perspective.* MIT Press.
- [VS] Venables and Smith. 2009. *An Introduction to R.* (2nd ed.). R Core Team. Available for free: https://cran.r-project.org/doc/manuals/R-intro.pdf
- [SB] Sutton and Barto. 2017. *Reinforcement Learning.* MIT Press. Available for free: http://incompleteideas.net/book/bookdraft2017nov5.pdf
- [GBC] Goodfellow, Benigo and Courville. *Deep Learning.* 2016. MIT Press.

Suggested Fun Readings

- Pedro Domingos. 2015. *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World.* Basic Books.
- Brian Christia and Tom Griffiths. 2016. *Algorithms to Live By: The Computer Science of Human Decisions* Henry Holt and Co.
- Pearl and Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect.* Basic Books.

- Stephens-Davidowitz. 2017. *Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are.* HarperCollins.

Class Policy

1. Classes will consist of lecturing, active and collaborative learning exercises, and discussions. Class participation is very important for learning the material. Hence, it is important that you (a) participate in the class, and (b) spend enough time to work on the reading material (both prior and after each class). Working on the problem sets will also prepare you for the exams. Please note that the amount of time required for assignments depends on your background. We will use Canvas system to post lecture notes, assignments, etc. You should always be prepared to take organized notes if that helps you learn the material.

2. Assignments are due before the lecture starts (i.e., before 12:00pm) on their due dates. Late submitted materials will not be graded, and a "0" will be assigned. Note that late submissions include those submitted on or after 12:00pm on the due date: even 1 minute late, is late (No exception!).

3. It is imperative that you come to class in a timely fashion. If you come late you may miss the discussion. Furthermore, it disrupts the lecture and the learning environment to have latecomers to the classroom.

4. Use of electronic equipment (such as laptops, tablets, iPods, cell phones, etc.) is not allowed in the classroom unless the instructor is notified in advance that you use such devices to take notes. Please turn off and store all electronic equipment before entering the class room.

5. If you have to miss an exam, you will receive a grade of "0" unless you have a doctor's documentation of illness. In either situation, there will not be any make-up exam. A card from the health center saying that you visited the health center does not count as a doctor's report. The instructor needs to see a report that clearly states that you were not in a physical condition to take the exam at the scheduled date/time. In that case, your grade will be re-distributed based on the other items that constitute your grade.

6. You are expected to abide by the university policies on academic honesty and integrity. All course activities, including class meetings, homework assignments, and exams are subject to the HKS Academic Code and Code of Conduct. Violations of these policies will not be tolerated, will be reported to the relevant authorities of HKS, and are subject to severe sanctions up to and including expulsion from the University.

Class Logistics

Email. All emails to the Instructor, his assistant, and/or TF/CAs must have [API 222] in the subject line. Failure to place this in the subject line could cause your email not to be read. Email is the convenient way to get in touch with the instructors to get an answer to a short question. However, be aware that the professor, his assistant, and TF/CAs have many other work obligations and probably keep different schedules from you. Therefore, although we do guarantee that we will answer all emails we receive, you should not expect us to answer each of them right away. The most reliable way to get your questions and concerns answered is to attend office hours: keep up with lectures and reading materials, and get started on assignments early, to be prepared to pose questions in office hours.

Class Attendance. We expect that you will arrive on time to class and do your best to attend every class. If you need to miss a class due to an emergency, it is your responsibility to obtain missed notes and course announcements from another student; there is no need to email the instructors. All lecture notes will be posted on the course website. It is highly recommended that you go through the lecture notes and the reading materials prior to attending the class.

Grading

Your grade in this class will be composed of

- 25% - Problem Sets
- 25% - Midterm Exam
- 20% - Final Project
- 30% - Final Exam

Final letter grades will be determined using the Dean's Recommended Grade Distribution. We will follow the most generous recommended distribution.

Problem Sets. The best way to learn is to practice. There will be about four (or so) problem sets that will be assigned during the semester. Under the Kennedy School Academic Code, the problem sets for this course are "Type II" assignments. *You are encouraged to work in a study group, but must submit your own hand- or type-written solutions. It is not acceptable to work on one electronic document as a group and submit identical, or nearly identical versions.* Examples of assignments that are not in accordance with the HKS academic code include photocopies or reprints of substantially identical assignments, printouts of substantially identical `R` (`Mathematica`, `Python`, etc.) tables or graphs, and copies of solutions from previous years. Problem sets are due at the beginning of class on the date they are due.

Midterm Exam. There will be a midterm exam and a final exam. These will be closed book/notes and a formula sheet will be provided (if needed). Calculators may be used, but statistical functions on them may not. Calculators that allow text storage are not permitted. The midterm will be given on ***March 21, In Class***. Please schedule your travel plans accordingly and arrange to be in the classroom on time.

Final Exam. Our exam has been scheduled by the HKS Registrar on ***May 9, 9:00am-12:00pm***. We have no discretion to change the time or date of the final exam. Please schedule your travel plans accordingly.

Final Project. The final exercise is a group-based project that engages all of the skills acquired in the class by asking you to take the skills learned and apply them to the analysis of some original, interesting, or otherwise professionally relevant data set.

Groups should be 5 students total. The components of the final exercise include (1) a short write-up of your analyses and results, and (2) a short presentation of the key points.

More details and deadlines will be provided later in the course. Students are encouraged to start thinking about their groups and also about potential data sources as early in the term as possible. Each group is essentially responsible for choosing a topic, collecting data, and providing useful analyses.

Instead of the group project, Ph.D. students interested in taking the class are required to conduct individual research, and write a rigorous final term paper (in the format of a journal paper). It is highly recommended that they meet with the instructor to individually discuss this requirement.

Class Participation & Engagement. Student participation can substantially enrich the learning experience for both the students and the instructor. Thus, students are highly encouraged to go through the lecture notes and other references to prepare before attending each session, and be ready to answer and/or raise questions. We assume all students actively participate in the class. However, participation is not graded.

Regrading Policy. Requests for regrades will be accepted only in writing, with a clear statement of what has been mis-graded, and why, and within one week of return of your graded work. Note that we reserve the right to review all of your answers, and you might end up having more points deducted than if you hadn't requested a regrade. Regrading requests should be submitted directly to the course TF.

Tentative Schedule

This schedule is subject to change, depending on how far we get in each class and how much discussion the topics generate.

Jan 24 (Lecture 1): Class Overview and Introduction to Machine Learning

- Class Overview
- Overview of Machine Learning Approaches and Applications
- Descriptive vs. Predictive vs. Prescriptive Analytics
- **Readings & Applications:**
  - JWHT, Chapter 1.
  - Of prediction and policy: Governments have much to gain from applying algorithms to public policy, but controversies loom. *The Economist*, Aug. 20, 2016.
  - Kleinberg, J., Ludwig, J., and Mullainathan, S. 2016. A Guide to Solving Social Problems with Machine Learning. *Harvard Business Review*, December 08.

Jan 26 (Lecture 2): Introduction to Statistical Learning Theory

- Parametric Learning vs. Nonparametric Learning
- Supervised vs. Unsupervised Learning
- Regression vs. Classification
- Prediction vs. Inference
- **Readings & Applications:**
  - JWHT, Chapter 2, 15–29.
  - Beam, A.L., and Kohane, I.S. 2018. Big Data and Machine Learning in Health Care. JAMA, 319 (3), 1317–1318.

Jan 31 (Lecture 3): Model Accuracy

- Measuring the Quality of Fit
- Less Flexible vs. More Flexible Models
- Test Error vs. Training Error
- Bias-Variance Tradeoff

- K-Nearest Neighbors Method
- Bayes Classifier
- **Readings & Applications:**
  - JWHT, Chapter 2, 29–51
  - Domingos, P. 2012. A few useful things to know about machine Learning. *Communications of the ACM*, 55 (10), 78–87.
  - Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. and Mullainathan, S., 2017. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1), 237-293 [pages 237–245 required; the rest is optional].

Feb 2, 7, and 9 (Lectures 4-6): Review of Ideas in Linear Regression

- Simple Linear Regression
- Multiple Linear Regression
- Testing Statistical Significance
- Qualitative Predictors (Credit Card Data)
- Interactions (Credit Card Data)
- Linear Regression vs. K-Nearest Neighbors
- **Readings & Applications:**
  - JWHT, Chapter 3.
  - Gerber, A.S., and Green, D.P. 2000. The Effect of a Nonpartisan Get-Out-the-Vote Drive: An Experimental Study of Leafletting. *The Journal of Politics*, 62 (3), 846-857.
  - Washington, E.L. 2008. Female Socialization: How Daughters Affect Their Legislator Fathers' Voting on Women's Issues. *American Economic Review*, 98 (1), 311-332.
  - Bertsimas, D. and King, A. 2016. An algorithmic approach to linear regression. *Operations Research*, 64 (1), 2–16. [This article is optional]

Feb 14 (Lectures 7): Classification I

- Overview of Classification in Machine Learning
- Logistic Regression for Classification
- Estimating Regression Coefficients and Probability Prediction in Logistic Regression (Credit Card Data Set)
- Multiple and Multiclass Logistic Regression
- **Readings & Applications:**
  - JWHT, Chapter 4, 127–138.
  - Policy Application: Framingham Heart Study. Original Framingham Heart Study data

Feb 16 (Lecture 8): Classification II

- Classification Using Bayse' Rule
- Linear Discriminant Analyses (LDA)
- Quadratic Discriminant Analyses (QDA)
- Guassian Naive Bayes Classifier
- **Readings & Applications:**
  - JWHT, Chapter 4, 138–167.
  - Policy Application: Using LDA to classify individuals at high risk of default. Credit Card Data Set

Feb 21 (Lecture 9): Model Validation

- Validation Set Approach
- LOOC Validation
- K-Fold Cross Validation
- Bootsrtap
- **Readings & Applications:**
  - JWHT, Chapter 5.
  - Ward, M.D., Greenhill, B.D., and Bakke, K.S. 2010. The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4), 363–375.
  - Policy Application: Predicting Civil Wars. Civil Wars Data Set.

Feb 23 (Lectures 10): Model Section and Regularization I

- Best Subset Section
- Forward and Backward Stepwise Section
- Shrinkage Method 1: Ridge Regression
- Shrinkage Method 2: Lasso
- **Readings & Applications:**
  - JWHT, Chapter 6, 203–228.
  - Policy Application: Prostate Cancer Policies. Prostate Cancer Data Set.

Feb 28 (Lectures 11): Model Section and Regularization II

- Dimensional Reduction Method 1: Principal Components Regression
- Dimensional Reduction Method 2: Partial Least Squares
- Comparison of Subset Selection, Shrinkage, and Dimension Reduction Methods Using Prostate Cancer Data
- Considerations in High Dimensional Data
- **Readings & Applications:**
  - JWHT, Chapter 6, 228–258.
  - Policy Application: Advertisement Spending in Different Cities

March 2 (Lectures 12): Non-Linear Models I

- Polynomial Regression
- Step Functions
- Piecewise Polynomials
- Regression Splines
- Smoothing Splines
- **Readings & Applications:**
  - JWHT, Chapter 7, 265–280.
  - Policy Application: Predicting Wage Based on Age.

March 7 (Lectures 13): Non-Linear Models II

- KNN Regression with Kernel Smoothing
- Local Regression
- Generalized Additive Models (GAMs)
- **Readings & Applications:**
  - JWHT, Chapter 7, 280–296.
  - Fraud Detection: How Machine Learning Systems Help Reveal Scams in Fintech, Healthcare, and eCommerce. AltexSoft.
  - Policy Application: Fraud/Spam Detection [HTF]

March 9: Catchup Lecture + Midterm Review

March 14: No Class - Spring break

March 16: No Class - Spring break

March 21: Midterm Exam - In Class

March 23 (Lectures 14): Trees I

- Decision Trees
- Regression and Classification Trees
- Pruning Trees
- **Readings & Applications:**
  - JWHT, Chapter 8, 303–316.
  - Policy Application: Using Decision Trees to Learn How Counties Vote: The Obama-Clinton Divide.

March 28 (Lectures 15): Trees II

- Ensemble Methods
- Random Forest
- Bagging

- Boosting
- **Readings & Applications:**
  - JWHT, Chapter 8, 316–331
  - Martin, A.D., Quinn, K.M, Ruger, T.W., and Kim, P.T. 2004. Competing Approaches to Predicting Supreme Court Decision Making. *Perspectives on Politics*, 2(4), 761-767.
  - Ang, Y.Q., Chia, A. and Saghafian, S. 2020. Using Machine Learning to Demystify Startups Funding, Post-Money Valuation, and Success. *Working Book Chapter (Springer).*
  - Joel et al. (2020). Machine Learning Uncovers the Most Robust Self-Report Predictors of Relationship Quality Across 43 Longitudinal Couples Studies. *Proceedings of National Academy of Sciences (forthcoming)* [optional]
  - Policy Application: Predicting the Outcomes of the United States Supreme Court. SCDB data.
  - Policy Application: Predicting Startups Performance
  - Policy Application: Predictors of Relationship Quality [Time Permitting]

March 30 (Lectures 16): Support Vector Machines (SVMs)

- Support Vector Classifier
- Support Vector Machines (SVMs)
- Using SVMs for Facial Expression Recognition
- **Readings & Applications:**
  - JWHT, Chapter 9

Apr 4 (Lectures 17): Advanced Methods I: Neural Networks and Deep Learning

- Neural Networks
- Deep Learning
- **Readings & Applications:**
  - HTF, 389-397 [optional.]
  - GBC, Chapter 1 pages 1–8,12–27 [optional]
  - Duame, Chapter 10. Neural Networks [optional]

- Neural Networks in R Tutorial. https://www.youtube.com/watch?v=lTMqXSSjCvk [optional]
- LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep Learning. *Nature*, 521, 436–444.
- Maharana, A., and Nsoesie, E.O. 2018. Use of Deep Learning to Examine the Association of the Built Environment With Prevalence of Neighborhood Adult Obesity. *JAMA Network Open.* 2018;1(4):e181535. doi:10.1001/jamanetworkopen.2018.1535
- Policy Application 1: ZIP Code Data Analysis [HTF].
- Policy Application 2: Fake News Detection and Creation.
- Policy Application 3: Understanding the Impact of the Built Environment on Adult Obesity Using Deep Learning.

Apr 6 (Lectures 18): Advanced Methods II: Reinforcement Learning

- Markov Chains
- Markov Decision Processes (MDPs)
- Q-Learning
- **Readings & Applications:**
  - SB, Chapters 3 and 6.5
  - Saghafian, S. 2018. Ambiguous Partially Observable Markov Decision Processes: Structural Results and Applications. *Journal of Economic Theory*, 178, 1–35. [optional]
  - Silver et al. 2017. Mastering the game of Go without human knowledge *Nature*, 550, 354–359.

Apr 11 (Lectures 19): Unsupervised Learning I

- Introduction to Unsupervised Learning
- Difficulties of Unsupervised Learning
- Principal Component Analysis
- **Readings & Applications:**
  - JWHT, Chapter 10, 373–385
  - Policy Application: United States Crimes (Assault, Rape, and Murder) and Arrest Data

Apr 13 (Lectures 20): Unsupervised Learning II

- K-Means Clustering
- Hierarchical Clustering
- **Readings & Applications:**
  - JWHT, Chapter 10, 385–412
  - Saghafian, S., Imanirad. R., and Traub, S. 2018. Who is an Effective and Efficient Physician? *Working Paper*, Harvard University.
  - Policy Application 1: Improving Effectiveness and Efficiency of Care Delivery Policies
  - Policy Application 2: Recommendation Systems
  - Policy Application 3: Understanding Breast Cancer Subtypes Using Gene Expression Data

Apr 18 (Lectures 21): Machine Learning and Causal Inference

- Correlation versus Causation
- Randomized Control Trials versus Observational Studies
- Using Machine Learning for Causal Inference
- **Readings & Applications:**
  - Kleinberg, J., Ludwig, J., Mullainathan, S. and Obermeyer, Z., 2015. Prediction policy problems. *American Economic Review: Papers & Proceedings*, 105(5), 491–95.
  - Pearl, J. 2018. The seven tools of causal inference with reflections on Machine Learning. *Communications of Association for Computing Machinery* (forthcoming).
  - Athey, S. and Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27), 7353-7360. [optional]

Apr 20: Project Presentations

Apr 25: Project Presentations

Apr 27: Wrap-Up + Final Term Review Session