Course Syllabus
Biostatistics 835
Fall 2023
Full Name: Biostatistical Machine Learning for Public Health
Short Name: Biostatistical Machine Learning
Schedule: Tuesday/Thursday 10:05 – 11:20, PHRC 320

- **Instructor:** Alexander McLain, PhD, Associate Professor of Biostatistics. E-mail: mclaina@mailbox.sc.edu, Office: Discovery I Room 450, Office Phone: *(803)777-1124*.
- **Office Hours:** Friday 1:00–3:00, and by appointment.
- **Course Website:** The course website is available on GitHub.
- **Class Communication:** We will use Slack as a discussion board throughout the semester. Please use this to ask questions about homework or other course topics. This will be regularly monitored and all questions will be addressed within (roughly) 24 hours of posting.

  If there are homework questions you are not comfortable posting on Slack you may e-mail the instructor. These questions will be redirected to Slack and answered in due course. For questions about your projects e-mail is the preferred method of communication, however, if the question is general enough it will be reposted on Slack.

  Invitations to our Slack channel will be sent to your school e-mail. If you do not receive one by the end of the week please e-mail the instructor. Adding your picture to Slack is not required but will be worth some bonus points on the first homework. Pictures are very helpful to me (and your classmates) towards learning your face and name.
- **Text:**
  - Hastie, T., Tibshirani, R., and Friedman, J.H., 2009. The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer. Available at http://www-stat.stanford.edu/ElemStatLearn.
- **Other helpful resources:**
  - Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity*. Monographs on statistics and applied probability, 143, 143.
  - James, G., Witten, D., Hastie, T. and Tibshirani, R., (2013). *An introduction to statistical learning with applications in R* (Vol. 112, p. 18). New York: springer. Available at https://www.statlearning.com.
- **Course Description:**

  The main focus of this course focus will be on using biostatistical models to predict and provide information on complex public health datasets. We will focus more on prediction of outcome(s) than estimation of the impact of a risk factor. However, some interferential methods for risk factors will be reviewed and (for all methods) techniques to measure variable importance will be discussed. Further, unsupervised learning methods (e.g., clustering) will be discussed. Prediction and predictive inference will be main

themes in the course along with learning how to implement the methods in `R` software. See the Course Schedule below for a full list of topics.

Pre-requisites: BIOS 770 or equivalent;

Restrictions: BIOS PhD students or PhD students from other departments with instructor approval.

- **Learning Outcomes**
  - Apply the appropriate supervised or unsupervised machine learning techniques to a scientific problem.
  - Assess the connection between common statistical techniques and machine learning approaches.
  - Evaluate the impact predictor variables have with an outcome using machine learning approaches.
  - Critique complex "black box" type algorithms in terms of the tradeoff between bias and variance.
  - Utilize cross-validation (or similar techniques) to appropriately train algorithms.
  - Perform multiple testing methods to high-dimensional data when inference is desired.
  - Apply the techniques covered in class using `R`.
- **Course Work:**
  - *Homework (75%):* Homework assignments (4–6) will be assigned on the course website.
  - *Term Project (25%):* Students will work on a research project throughout the semester. There will be items due during the semester including a research plan, a short (8 page, conference-style) paper, and a presentation in the later part of the course. A list of potential topics will be provided by the instructor. The topic selected by the student must be approved in advance by the instructor.
  - No final exam.

  Grades will be assigned as follows: 90–100=A; 88–90=B^+; 80-87=B; 78–80=C^+; 70–77=C, 62–70=D, and 0–62=F.
- **Course Materials:**
  - References and reading will be put on the course website.
- **Academic Integrity:**

  You are expected to practice the highest possible standards of academic integrity. Students may brainstorm ideas for homework assignments, but may not copy solutions from other students or from other sources. Any deviation from this expectation will result in a minimum academic penalty of your failing the assignment, and may result in additional disciplinary measures. This includes improper citation of sources, using another student's work, and any other form of academic misrepresentation.

- **Attendance Policy:**

  Though attendance is not required, it is strongly recommended.

- **Disability Resource Center:**

  The Student Disability Resource Center (SDRC) empowers students to manage challenges and limitations imposed by disabilities. Students with disabilities are encouraged to contact me to discuss the logistics of any accommodations needed to fulfill course requirements (within the first week of the semester). To receive reasonable accommodations from me, you must be registered with the Student Disability Resource Center (1705 College Street, Close-Hipp Suite 102, Columbia, SC 29208, 803-777-6142; email: sadrc@mailbox.sc.edu). Any student with a documented disability should contact the SDRC to arrange for appropriate accommodations.

- **Course Outline:**

Week	Topic
1	Overview of statistical learning
2	Linear Regression and Decision Theory
3	Dimension Reduction
4	Model Selection and Shrinkage Methods
5	Bayesian variable selection
6	Linear Classification: logistic, linear discriminant
7	Generalized additive models
8	Classification and Regression Trees (CART)
9	Random Forests and bagging
10	Support vector machines
11	Neural Networks and Deep Learning
12	Ensemble and Transfer Learning
13	Multiple Testing and Selective Inference
14	Presentations