Course Syllabus

Biostatistics 735

Fall 2025

**Full Name:** Machine Learning for Public Health Applications

**Short Name:** Introduction to Machine Learning

* **Instructor:** Alexander McLain, PhD, Associate Professor of Biostatistics. E-mail: [mclaina@mailbox.sc.edu](mailto:mclaina@mailbox.sc.edu), Office: Discovery I Room 456, Office Phone: *(803)777- 1124*.
* **Office Hours:** Wednesday 9:00–11:00, and by appointment.
* **Course Website:** The course website can be found [here](https://github.com/alexmclain/Bios_735) (https://github.com/alexmclain/Bios\_735).

# Text: 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](https://www.statlearning.com/).

* + **Additional Resources** (not required but recommended):
    - Kuhn, M., & Johnson, K. (2019). Applied Predictive Modeling. Springer.
    - Selected articles and case studies in public health informatics (to be provided throughout the course).

# Course Description:

This course introduces the fundamental concepts and applications of machine learning (ML) in the context of public health, biostatistics, and related fields (e.g., epidemiology, psychology, neuroscience, genetics). Students will learn how to prepare data, select and implement appropriate machine learning methods, and evaluate model performance—all using R. Emphasis will be placed on conceptual understanding, practical implementation, and interpretation of results rather than on mathematical derivations. By the end of the course, students will be prepared to incorporate ML techniques into their own research projects, with a special focus on the unique challenges and ethical considerations inherent in analyzing public health data.

**Pre-requisites:** BIOS 757 and 755 or equivalent;

**Restrictions:** BIOS MS students or a PhD student with instructor approval

# Learning Outcomes

1. **Explain the basics of machine learning**
   * Distinguish between supervised, unsupervised, and semi-supervised learning.
   * Understand differences and overlaps between classical statistical approaches and machine learning methods.
2. **Preprocess and manage data for ML applications**
   * Handle missing data, outliers, and imbalanced datasets.
   * Implement feature engineering and selection techniques appropriate for public health data (e.g., cohort data, survey data, neuroimaging, genetics).
3. **Select and implement appropriate machine learning models**
   * Apply common methods (e.g., linear/logistic regression, tree-based models, ensemble methods, neural networks) using R packages.
   * Employ resampling and cross-validation strategies to tune hyperparameters and assess model performance.
4. **Interpret and communicate ML results**
   * Understand model interpretability, especially for tree-based and linear models.
   * Critically analyze model results and communicate findings to diverse audiences, including non-technical stakeholders.
5. **Address ethical, reproducibility, and privacy considerations**
   * Recognize limitations, pitfalls (e.g., overfitting, data leakage), and responsible conduct in machine learning research.
   * Discuss ethical issues surrounding privacy, data sharing, and fairness in predictive modeling.
6. **Develop ML-based solutions for real-world public health scenarios**
   * Apply and adapt machine learning workflows to case studies in epidemiology, mental health, neuroimaging, and genetics.
   * Produce a final project that demonstrates mastery of the course material in a domain-relevant context.

# Course Work:

* + ***Homework (50%):*** Homework (4–6) will be assigned on the course website.
  + ***Paper Discussion (20%):*** You will deliver a ~30-minute in-class presentation on a peer-reviewed public-health paper that uses a bona fide machine-learning method. You will explain why ML was needed, summarize the data and study design, describe the method and its tuning/validation and metrics, address interpretability and ethics, and offer a brief critique.
  + ***Term Project (30%):*** 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 latter 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.

Grades will be assigned as follows: 91–100=*A*; 88–90=*B*+; 81-87=*B*; 78–80=*C*+; 71– 77=*C*, 62–70=*D*, and 0–61=*F* .

# Course Materials: References and reading will be put on the course website. All readings/materials comply with copyright/fair use policies.

# Class Communication: We will use [Slack](http://www.slack.com/) as a discussion board throughout the semester. Please use this to ask questions about homework or other course topics. I will regularly monitor this. All questions will be addressed within 24 hours of posting.

# If there are homework questions you are not comfortable posting on Slack, you may email them to me. 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 me.

# 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 other sources. Any deviation from this expectation will result in a minimum academic penalty of 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.

* **Generative AI Use Policy:** Generative AI tools (e.g., ChatGPT) are powerful assistants and part of your professional toolkit. You are in this course to build your skills in R-based machine learning for public health. Use AI to accelerate your learning—not to replace it. If AI is thinking for you, you are giving away the very skills you came here to develop.
  + What’s allowed
    - **Brainstorming:** clarifying concepts, outlining approaches, suggesting packages, and generating checklists.
    - **Editing & explanation:** improving grammar/clarity, commenting on code you have written, explaining error messages, and proposing alternative implementations.
  + What’s not allowed
    - **Outsourcing the assignment:** submitting AI-generated code, text, or slides.
    - **Unverified outputs:** copying answers, code, or statistical results.
  + Disclosure & accountability
    - Declare AI use on every submission (HW, project, presentation) with a brief note at the end:

*“AI assistance: ChatGPT (date/model). Used to: [e.g., outline, debug error]. Prompts included: […]. Substantive changes made: […].”*

* + - You are responsible for all content you submit. If I suspect something is AI-generated, I will follow up to verify your understanding.

# Attendance Policy: Though attendance is not required, it is strongly recommended.

# Disability Resource Center: The [Student Disability Resource Center](https://sc.edu/about/offices_and_divisions/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](mailto:sadrc@mailbox.sc.edu)). Any student with a documented disability should contact the SDRC to arrange for appropriate accommodations.

* **Course Outline:**

|  |  |
| --- | --- |
| **Week** | **Topic** |
| 1 | Introduction, Motivating Examples, and Foundations |
| 2 | Bias-Variance Tradeoff and Model Validation |
| 3 | Linear Regression and Model Selection |
| 4 | Penalized Regression (Shrinkage Methods) |
| 5 | Extensions of Regression and Variable Importance |
| 6 | Classification: Logistic Regression, Discriminant Analysis, and Beyond |
| 7 | Tree-Based Methods: CART and Ensemble Techniques |
| 8 | Support Vector Machines |
| 9 | Neural Networks |
| 10 | LLM Basics, Dimension Reduction and Clustering |
| 11 | Multiple Comparisons and Post-Selection Inference |
| 12 | Ethical Considerations, Responsible Data Use, and Reproducible Workflows |
| 13 | Causal Inference and Advanced Topics |
| 14 | Presentations |