**R25 Emerging Scholars 2025 Summer Institute --- Week 6 (June 16-20)**

**Introduction to Biostatistical Machine Learning for Public Health**

**Location: Discovery Building (see timetable for locations)**

**Instructor(s):** Alexander McLain, PhD  
**Email:** [mclaina@mailbox.sc.edu](mailto:mclaina@mailbox.sc.edu)  
**Phone:** 803-777-1124

**Teaching assistant:** Peter Olejua  
**Email:** [polejua@email.sc.edu](mailto:%20POLEJUA@email.sc.edu)

**Course Description:** This course is designed to introduce undergraduate students to the basics of machine learning, focusing on applications in public health. Students will learn what machine learning is and how it differs from traditional statistics. Key concepts including supervised learning, where models learn from labeled data to make predictions, and unsupervised learning, which involves finding patterns in data without predefined labels, will be explored. By the end of the course, participants will understand the fundamental ideas and common terminology used in machine learning, know the difference between supervised and unsupervised learning, and have practical experience using basic machine learning methods on public health data. They will also become aware of important ethical issues like privacy and fairness when working with health information and will be able to explain and present their findings clearly to others.

**Course Materials/References (for reference only, available online)**

* 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/).

**Course Objectives:**

* Understand the role of biostatistics and machine learning in public health.
* Differentiate between supervised and unsupervised learning methods.
* Gain hands-on experience with basic machine learning algorithms applied to public health data.
* Be aware of ethical considerations in handling and analyzing health data.
* Develop skills to interpret and communicate machine learning results effectively.
* Learn to perform critical analysis of machine learning results.

**Class Description:** The outline of each day will be a (roughly) 3-hour lecture in the morning, followed by a hands-on lab in the afternoon. The lecture will teach fundamental topics in machine learning, focusing on the main concepts of each approach, examples of their application, and their pros and cons. We will only discuss the technical details of the approaches if they illuminate essential details about the method. In the afternoon session, the students will use modern (often high-dimensional) datasets to apply the methods discussed in the morning session in R software.

**Course Requirements:**

* *Homework/Lab:* Homework/Lab assignments will be assigned during the lab sessions.
* *Capstone Project:* Students will work on a project that generate a complete analysis plan for a public health dataset that will require the methods used in class.
* *Group Project:* Students will work on a project that generate a complete analysis plan for a public health dataset that will require the methods used in class.

**Class Communication:** if any question arise, please email the instructor or TA.

**Workshop Schedule and Activities**

|  |  |
| --- | --- |
| **Week 6, Day 1: Introduction to Biostatistics, Linear Regression, and Machine Learning in Public Health Location: Discovery 259 (morning) Discovery 431 Lab (afternoon)** | |
| **ACTIVITY** | **TOPIC/TIME** |
| Morning Lecture | **Part 1: 9:00 – 10:15 am**   * Overview of biostatistics: definition, scope, and its critical role in public health research and practice. * Introduction to basic statistical concepts: types of data, descriptive statistics, inference. * Difference between traditional statistics and machine learning. * Types of Machine Learning: Supervised vs. Unsupervised Learning.   **Break: 10:15 – 10:30 am**  **Part 2: 10:30 – 12:00 pm**   * Supervised Learning Concepts: regression versus classification. Review of linear regression. * Unsupervised Learning Concepts: introduction to clustering and dimensionality reduction * Applications in Public Health: Case studies: disease prediction, outbreak detection, personalized medicine. * Model selection in linear regression: understanding overfitting, underfitting, and how to choose the right model. |
| Lunch | **12:00 – 1:00 pm** |
| Afternoon Lab | **Part 1: 1:00 – 2:15 pm**   * Setting up the programming environment (e.g., R, Python) and familiarization with statistical computing tools. * Basic data handling and visualization exercises to understand public health data characteristics.   **Break: 2:15 – 2:30 pm**  **Part 2: 2:30 – 4:00 pm**   * Descriptive statistics: mean, median, mode, standard deviation. * Data visualization: histograms, box plots, scatter plots. * Identifying data patterns: trends and correlations in public health data. |
| **Week 6, Day 2:** **Supervised Learning Techniques in Public Health  Location: Discovery 259 (morning) Discovery 431 Lab (afternoon)** | |
| Morning Lecture | **Part 1: 9:00 – 10:15 am**   * Deep Dive into Linear Regression:   + Assumptions: linearity, independence, homoscedasticity, normality.   + Assessing model fit: R-squared, residual analysis.   **Break: 10:15 – 10:30 am**  **Part 2: 10:30 – 12:00 pm**   * Introduction to Logistic Regression   + When to use logistic regression: predicting binary outcomes.   + Model interpretation: odds ratios, log-odds.   + Evaluation metrics: accuracy, confusion matrix, sensitivity, specificity. * Introduction to Decision Trees   + Concepts: nodes, branches, leaves.   + Building a decision tree: how the algorithm splits data.   + Pros and cons: interpretability vs. overfitting. |
| Lunch | **12:00 – 1:00 pm** |
| Afternoon Lab | **Part 1: 1:00 – 2:15 pm**   * Implementing Linear Regression   + Hands-on coding: fitting a linear model to a public health dataset.   + Interpreting results: coefficients and p-values. * Implementing Logistic Regression   + Model building: predicting a health outcome.   + Evaluating the model: generating and interpreting a confusion matrix.   **Break: 2:15 – 2:30 pm**  **Part 2: 2:30 – 4:00 pm**   * Exploring Decision Trees   + Building a simple decision tree using software packages.   + Visualizing the tree: interpreting splits and decisions. * Model Evaluation Techniques   + Overfitting vs. underfitting: understanding the bias-variance tradeoff.   + Cross-validation: purpose and basic methods (e.g., k-fold). |
| **Week 6, Day 3:** **Unsupervised Learning and Model Evaluation for Public Health Data Analysis Location: Discovery 431 Lab (afternoon only)** | |
| Morning | **Part 1: 10:30 – 12:00 pm**   * Office for Undergraduate Research (OUR) summer training activity |
| Lunch | **12:00 – 1:00 pm** |
| Afternoon Lecture | **Part 1: 1:00 – 2:15 pm**   * **Clustering Methods**   + Introduction to k-Means Clustering: how the algorithm works.   + Choosing the number of clusters.   + Hierarchical Clustering: agglomerative vs. divisive methods.   + Interpreting dendrograms. * **Dimensionality Reduction**   + Principal Component Analysis (PCA): reducing data dimensions while retaining variance.   + When and why to use PCA: simplifying models, visualization.   **Break: 2:15 – 2:30 pm**  **Part 2: 2:30 – 4:00 pm**   * Afternoon lab (abbreviated)   + Performing cluster analysis on health datasets to identify distinct groups.   + Using principal components, and related methods, for dimensionality reduction in large-scale health data. |
| **Week 6, Day 4:** **Ethical Considerations  Location: Discovery 259 (morning), Discovery 431 Lab (afternoon)** | |
| Morning Lecture | **Part 1: 9:00 – 10:15 am**   * Ethical Implications in Machine Learning   + Data privacy and security: handling sensitive health data responsibly.   + Bias and fairness: recognizing and mitigating algorithmic bias.   + Consent and transparency: importance of informed consent in data collection.   **Break: 10:15 – 10:30 am**  **Part 2: 10:30 – 12:00 pm**   * Case Studies in Ethics   + Historical examples where ethics were compromised.   + Discussion on ethical dilemmas in current practice. * Interpreting and Communicating Results   + Effective visualization techniques: presenting data clearly and accurately.   + Reporting findings: writing summaries and reports for a non-technical audience. * Future of Machine Learning in Public Health   + Emerging trends: AI advancements, big data analytics.   + Career paths: opportunities in biostatistics and data science. |
| Lunch | **12:00 – 1:00 pm** |
| Afternoon Lab | **Part 1: 1:00 – 2:15 pm**   * Capstone Project Introduction and Analysis   + Project introduction: analyzing a public health dataset using learned techniques.   + Guided analysis: choosing appropriate models (supervised or unsupervised).   **Break: 2:15 – 2:30 pm**  **Part 2: 2:30 – 4:00 pm**   * Project Presentations |
| **Week 6, Day 5:** **Case Studies and Critical Reading of**  **Machine Learning in Public Health Location: Discovery 259 (morning), Discovery 431 Lab (afternoon)** | |
| Morning Lecture | **Part 1: 9:00 – 10:15 am**   * Evaluating Scientific Literature   + Understanding research papers   + Developing critical reading skills   + Assessing reproducibility, validity, reliability, data sources and quality   **Break: 10:15 – 10:30 am**  **Part 2: 10:30 – 12:00 pm**   * Interpreting Results and Statistical Significance * Ethical Considerations in Research * Case Studies: disease outbreak prediction, AI in personalized medicine, clustering in epidemiology, natural language processing in healthcare, predictive modeling for health resource allocation. |
| Lunch | **12:00 – 1:00 pm** |
| Afternoon Lab | **Part 1: 1:00 – 2:15 pm**   * Group reading: public health literature related to machine learning * Group analysis: guided critical analysis of research   **Break: 2:15 – 2:30 pm**  **Part 2: 2:30 – 4:00 pm**   * Group presentations of papers with critical analysis. |

**University Policies**

|  |  |
| --- | --- |
| **Incomplete coursework** | Incomplete coursework is a major inconvenience for students and instructors. I expect you to do everything in your power to avoid this situation. Legitimate excuses include verified illnesses and family emergencies. No incompletes will be granted only in accordance with [university policy](https://sc.edu/about/offices_and_divisions/advising/withdrawal/dropping_courses/). |
| **Classroom conduct** | I will conduct this class in an atmosphere of mutual respect. I encourage your active participation in class discussions. Each of us may have strongly differing opinions on the various topics of class discussions. The conflict of ideas is encouraged and welcome. The orderly questioning of the ideas of others, including mine, is similarly welcome. However, I will exercise my responsibility to manage the discussions so that ideas and argument can proceed in an orderly fashion. You should expect that if your conduct during class discussions seriously disrupts the atmosphere of mutual respect I expect in this class, I can’t help but ask you to stop participating. |
| **Academic misconduct** | You are expected to do your own academic work and cite sources as necessary. Students are expected to adhere to all requirements of the Carolinian Creed ([www.sa.sc.edu/creed/](http://www.sa.sc.edu/creed/)). Please especially note: as a member of this training program, you are held accountable to this Creed even if you violate it inadvertently or are not a registered USC student. Any episode of dishonesty, cheating, or plagiarism in any form is cause for failure of an assignment, an examination, or a course. Students may want to refresh their understanding of the appropriate use of citations when drafting papers and other assignments to prevent inadvertent plagiarism stemming from lack of information. In addition, program leadership may decide to withdraw their support of your participation in this training program. |
| **Sexual harassment** | "Sexual harassment" means unwelcome sexual advances, requests for sexual favors, and/or other verbal or physical conduct of a sexual nature. Such conduct has the purpose or effect of unreasonably interfering with an individual's work or academic performance or creating an intimidating, hostile, or offensive working or academic environment in any University activity or program. Such behavior is not acceptable in the University setting. For additional information, please consult [Equal Opportunity Programs Policy](https://www.sc.edu/about/offices_and_divisions/equal_opportunities_programs/documents/sexual_harassment.pdf). |
| **Accessibility, Disability, and Triggers** | The USC Arnold School of Public Health is committed to providing equitable access to learning opportunities for all students.  If you have, or think you may have, a disability (e.g., mental health, attentional, learning, chronic health, sensory, or physical), please notify me so confidential discussion regarding equitable access and reasonable accommodations can take place. |
| **Diversity** | The University provides equal access to and opportunity in its programs and facilities, without regard to race, color, creed, religion, national origin, gender, age, marital status, disability, public assistance status, veteran status, sexual orientation, gender identity, or gender expression. For more information, please consult the [Student Non-Discrimination and Non-Harassment Policy](https://www.sc.edu/about/offices_and_divisions/equal_opportunities_programs/documents/student_non-discrimination_and_non-harassment_policy.pdf). |
| **Title IX and Gendered Pronouns** | This course affirms equality and respect for all gendered identities and expressions. Please don’t hesitate to correct the instructor regarding your preferred gender pronoun and/or name if different from what is indicated on the official class roster. Likewise, the instructor is committed to nurturing an environment free from discrimination and harassment. Consistent with Title IX policy, please be aware that the instructor is obligated to report information that students provide about a situation involving sexual harassment or assault. |