**R25 Emerging Scholars 2024 Summer Institute --- Week 4 (June 3-7)**

**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 aims to equip students with a foundational understanding of how biostatistics and machine-learning techniques can be applied to solve real-world public health challenges. The course will start with an overview of biostatistics and linear regression, progress through data preprocessing and exploratory data analysis, and dive into supervised and unsupervised learning methods tailored for disease prediction and data analysis of complex public health data sets. Advanced topics, including neural networks and ethical considerations in machine learning applications, are covered to prepare students for responsible and impactful work in the field.

Students will gain practical skills in data analysis, model building, and evaluation through a blend of theoretical lectures and hands-on labs using public health datasets. By the end of the course, participants will be well-equipped to utilize machine learning tools to analyze health data, derive insights, and contribute to evidence-based public health policies and interventions. See the Course Schedule below for a complete list of topics.

**Course Materials/References (any downloads can be found in the secure portal at** [**https://bigdata.sc.edu/r25-e-scholar-account/**](https://bigdata.sc.edu/r25-e-scholar-account/)**)::**

* 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.
* Gain proficiency in data preprocessing and exploratory data analysis.
* Learn how to implement and evaluate supervised and unsupervised learning techniques.
* Learn how to choose and apply advanced machine learning methods for various public health issues.
* Navigate ethical considerations in public health data analysis.

**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 (50%):* Homework/Lab assignments will be assigned on each day during the lab session.
* *Project (50%):* Students will work on a group project on the last day of the course. The group project will focus on generating a complete analysis plan for a public health dataset that will require the methods used in class.

**Grade:** Final grades that are at least 75% will be given a pass, less than 75% a fail.

**Class Communication:** We will use [Slack](https://slack.com/) as a discussion board throughout the course. Please use this to ask questions about homework or other course topics. This will be regularly monitored, and all questions will be addressed by the next day.

* If there are homework questions you are not comfortable posting on Slack you may e- mail the instructor or TA. If relevant, these questions will be redirected to Slack and answered there (keeping the sender anonymous).
* Invitations to our Slack channel will be sent to your e-mail. If you do not receive one, please e-mail the instructor. Adding your picture to Slack is not required but we be very helpful towards learning your face and name.

**Workshop Schedule and Activities**

|  |  |
| --- | --- |
| **Week 4, Day 1: Introduction to Biostatistics, Linear Regression, and Machine Learning in Public Health Location: Discovery 331** | |
| **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.   **Break: 10:15 – 10:30 am**  **Part 2: 10:30 – 12:00 pm**   * Introduction to linear regression: concept, assumptions, and how it is used in public health data analysis. * 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**   * Implementing simple linear regression and performing model selection techniques. |
| **Week 4, Day 2:** **Data Preprocessing and Exploratory Data Analysis (EDA) Location: Discovery 331** | |
| Morning Lecture | **Part 1: 9:00 – 10:15 am**   * Importance of data preprocessing in machine learning. * Techniques for handling missing data, normalization, and feature selection.   **Break: 10:15 – 10:30 am**  **Part 2: 10:30 – 12:00 pm**   * Introduction to exploratory data analysis and its significance in machine learning models. |
| Lunch | **12:00 – 1:00 pm** |
| Afternoon Lab | **Part 1: 1:00 – 2:15 pm**   * Hands-on exercises in data cleaning and preprocessing.   **Break: 2:15 – 2:30 pm**  **Part 2: 2:30 – 4:00 pm**   * Performing EDA on health datasets to uncover patterns and insights. |
| **Week 4, Day 3: Supervised Learning for Disease Prediction and Classification Location: Discovery 259** | |
| Morning Lecture | **Part 1: 1:00 – 2:15 pm**   * Introduction to supervised learning. Basics of prediction.   **Break: 10:15 – 10:30 am**  **Part 2: 10:30 – 12:00 pm**   * Deep dive into supervised learning algorithms (e.g., logistic regression, decision trees, random forests, SVMs). * Case studies on disease prediction and classification. |
| Lunch | **12:00 – 1:00 pm** |
| Afternoon Lab | **Part 1: 1:00 – 2:15 pm**   * Building and evaluating classification models using health datasets.   **Break: 2:15 – 2:30 pm**  **Part 2: 2:30 – 4:00 pm**   * Implementing cross-validation and assessing model performance. |
| **Week 4, Day 4: Unsupervised Learning for Public Health Data Analysis| Location: Discovery 259** | |
| Morning Lecture | **Part 1: 9:00 – 10:15 am**   * Introduction to unsupervised learning: clustering, dimensionality reduction techniques.   **Break: 10:15 – 10:30 am**  **Part 2: 10:30 – 12:00 pm**   * Applications of unsupervised learning in discovering patterns in health data. |
| Lunch | **12:00 – 1:00 pm** |
| Afternoon Lab | **Part 1: 1:00 – 2:15 pm**   * Performing cluster analysis on health datasets to identify distinct groups.   **Break: 2:15 – 2:30 pm**  **Part 2: 2:30 – 4:00 pm**   * Using principal components, and related methods, for dimensionality reduction in large-scale health data. |
| **Week 4, Day 5:** **Group Projects Location: Discovery 259** | |
| Morning Lecture | **Part 1: 9:00 – 10:00 am**   * Introduction of group project topics and data with discussion.   **Part 2: 10:00 – 12:00 pm**   * The rest of morning session will be devoted to your group projects. Groups will work and the instructors will be there to aid you in your analyses. |
| Lunch | **12:00 – 1:00 pm** |
| Afternoon Lab | **1:00 – 3:00 pm**   * Group Presentations |

**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. |