**STA 4241**

**RI: Statistical Learning**

Instructor: Dr. Hongwei Long

Office: SE 268

Office Hours: TBA

Email: hlong@fau.edu

**Catalog description**: This is an introductory-level course in supervised learning with a focus on regression and classification methods. The course will help students to understand basic concepts, ideas, and methods in statistical learning. Considerable amount of effort will also be put on computational aspects of algorithm implementation. This is a research intensive course.

**Textbook:**  *An Introduction to Statistical Learning with Applications in R,* First Edition, by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani , 2013, Springer, ISBN: 978-1-4614-7137-0

**Course description:** The course covers linear classification and regression, logistic regression and linear discriminant analysis, cross-validation and bootstrap, model selection, dimension reduction, tree-based methods, random forests and boosting, support-vector machines, principal components, and cluster analysis.

This course address undergraduate research and inquiry components on (1) knowledge, (2) formulating questions, (3) plan of action, (4) critical thinking, (5) ethical conduct, and (6) communication.

**Course objectives**: Students who successfully complete this course will be able to: (1) Explain statistical learning methodology; (2) Explain the assumptions of various techniques such as Multiple Regression, Discriminant Analysis, Logistic Regression, Decision Trees, and Cluster Analysis; (3) Build multiple regression, discriminant analysis, and logistic models for forecasting; (4) Validate models using statistical tests; (5) Interpret and understand decision trees and random forests; (6) Discuss issues of implementation of the results of various techniques; (7) Develop methods to monitor the ongoing performance of implemented models. Students will have hands-on experience in model building, machine learning, and implementation on complex datasets.

**Research Intensive Designation:** This course contains an assignment or multiple assignments designed to help students conduct research and inquiry at an intensive level.  If this class is selected to participate in the university-wide assessment program, students will be asked to complete a consent form and submit electronically some of their research assignments for review.  Visit the Office of Undergraduate Research and Inquiry (OURI) for additional opportunities and information at <http://www.fau.edu/ouri>.

The URI portion of the course will address all six Student Learning Objectives:

1. **Knowledge**: Common base of knowledge required for effective data preprocessing, data visualization, data learning, model selection and reproducible research. Students will grasp a set of key skills on statistical learning and computing in data analytics. Students will also show knowledge of tools and practical skills needed to preprocessing and managing data from various sources for both structured and unstructured massive data.
2. **Formulation of Questions**: Students are required to develop research statement in which they specifically address their research questions for real data sets or simulation study. Students are expected to formulate their research questions into subject related hypotheses which are clear and concise to the research problem, ready to be tested or answered through statistical modeling and simulation. When appropriate, the students should be able to break down principal problems into smaller solvable sub-problems.
3. **Plan of Action**: Students will create a plan of action for individual term projects of this research intensive course that will encompass the following elements: (i) scope of the study; (ii) literature review; (iii) planning context; (iv) problem statement and research methodology; (v) analysis and report findings; (vi) presenting the results to general audience. The students will develop hypothesis if needed, identify methods of analysis and select appropriate statistical techniques. Using the course timeline as a template, each student is expected to develop her/his own project management plan with specific tasks related to the topic in consideration.
4. **Critical Thinking:** Students will demonstrate critical thinking skills by formulating research questions, applying appropriate selection criteria for real data analysis or simulation study, taking into consideration multiple perspectives, and examining implications and consequences of an action or planning alternative.
5. **Ethical conduct:** All students are required to familiarize themselves with the rules of academic integrity. Student projects involving primary data collection through website will be credited in their term written research projects. Students are required to be loyal to the original data with confidential information removed. A class module will be provided with a discussion of when statistics were used inappropriately to arrive at a faulty conclusion for unethical purposes.
6. **Communication:** Students will be required to write and present their individual project reports professionally. They are required to submit research report (e.g., analysis, findings and recommendations), and develop a webpage uploaded to rpubs.com for instant to communicate research results as outlined in SLO-3. Students are expected to demonstrate knowledge of writing technical report and orally presenting their findings. Advanced visualization techniques are also required for students to incorporate research findings in planning documents and present them through a real data project.

**Prerequisites:** STA 4234 or Equivalent with minimum grade C

**Grading policy**:

Biweekly homework/lab assignment: 20%

Midterm Exam 30%

Midterm project: 20%

Final project: 30%

Grading scale: A/A-: 90-100%, B+/B/B-: 80-89%, C+/C: 70-79%, D: 60-69%, F: <59%.

**Course outline:**

Week 1: Introduction to Statistical Learning

Week 2: Linear Regression and Classification

Week 3: Logistic Regression

Week 4: Linear Discriminant Analysis

Week 5: Resampling Methods

Week 6: Linear Model Selection and Regularization

Week 7: Discussion of Midterm Project

Week 8-9: Moving Beyond Linearity

Week 10: Tree-based Methods

Week 11: Support Vector Machines

Week 12: Unsupervised Learning

Week 13-14: Discussion and Prepare for Final Project

Week 15: Project Presentations

**Description of the RI assignments:**

**Assignment 1 (Model Building):** In week 4 of the course, students will produce a tech report on how to build a statistical model for predicting an output based on input of regular data set. Students will publish the computing R scripts to a repository website. Students will submit the first tech report on their findings. This submission will be limited to five page preliminary tech report which will include a description of the scientific problem, statistical thinking and formulation of the problem, a description of the methods and a flow chart of the algorithm implementation for model building, a summary of the findings, and a conclusion.

**Assignment 2 (Machine Learning)**: In week 8 of the course, students will prepare a tech report and publish the computing algorithm to a repository website. Students will produce essential toolset and the second tech report for large data set via supervised learning and unsupervised learning methods. This will be at an intermediate level with room for improvement but strictly follow the guideline of tech reporting. This submission will be limited to eight-page tech report which will include a description of the real data, statistical thinking and formulation of the problem, building an appropriate statistical model for forecasting, validating models using statistical tests, a summary of the findings, and a conclusion.

**Assignment 3 (Final Project):** In weeks 9-14 of the course, students will write a tech report according to the tech report guideline and publish the computing algorithm to a repository website. Students have opportunity to propose improvements of the methodology or quality of data analysis at the end of the report. Students will present their individual final projects to classmates, and submit 10-15 pages of tech report or research paper which follows the format of scientific research articles including title, abstract, introduction, methodology, result interpretation, conclusion, and reference. Students are encouraged to present their findings either through a poster presentation or an oral presentation at the FAU undergraduate research symposium held on the Boca Raton Campus each Spring Semester (<http://www.fau.edu/ouri/undergrad_symposium.php>) and submit the final report for possible publication in a journal such as Florida Atlantic University Undergraduate Research Journal (FAURJ).

**Incomplete grades**

A grade of I (incomplete) will only be given under certain conditions and in accordance with the academic policies and regulations put forward in FAU's University Catalog. The student has to show exceptional circumstances why requirements cannot be met. A request for an incomplete grade has to be made in writing with supporting documentation, where appropriate.

**Classroom etiquette policy**

University policy on the use of electronic devices states: “In order to enhance and maintain a productive atmosphere for education, personal communication devices, such as cellular telephones and pagers, are to be disabled in class sessions.”

**Disability policy statement**

In compliance with the Americans with Disabilities Act (ADA), students who require reasonable accommodations due to a disability to properly execute coursework must register with Student Accessibility Service (SAS) ---in Boca Raton, SU 133 (561-297-3880) ; in Davie, LA 203 (954-236-1222); or in Jupiter, SR 110 (561-799-8585)---and follow all SAS procedures.

**Academic integrity**

Students at Florida Atlantic University are expected to maintain the highest ethical standards. Academic dishonesty, including cheating and plagiarism, is considered a serious breach of these ethical standards, because it interferes with the University mission to provide a high quality education in which no student enjoys an unfair advantage over any other. Academic dishonesty is also destructive of the University community, which is grounded in a system of mutual trust and places high value on personal integrity and individual responsibility. Harsh penalties are associated with academic dishonesty. For more information, see <https://www.fau.edu/ctl/4.001_Code_of_Academic_Integrity.pdf>