Statistical Machine Learning (SML) Methods for Applied Research/Applied SML, 3 credits

ABE 6933 Section ASMO; (see One.UF for async and REC sections)

Class Periods: Tuesdays and Thursdays, periods 4-5 (10:40 AM - 12:35 PM)

Location: online: sync (ASMO) or async (ASMA – only if schedule conflict with ASMO)

Academic Term: Fall 2025

Instructor:

Dr. Nikolay Bliznyuk
Email: nbliznyuk [AT] ufl.edu
Phone: 352-392-1864 (only by prior appointment)
Office: office hours by Zoom, times TBA

Teaching Assistant/Peer Mentor/Supervised Teaching Student:

- TBA

Course Description

Tools of statistical machine learning for applied research targeted at graduate students in engineering and agricultural and life sciences. Statistical approaches to machine learning are emphasized. Application of statistical ML methods and interpretation of the results rather than formal derivations or programming language implementation of the procedures are prioritized. Unlike STA 6703 (Statistical Machine Learning), this course emphasizes breadth (applications and interpretation of results) rather than depth, and is targeted at less quantitative audiences. This course will be delivered in online mode (synchronous or asynchronous).
If you have a schedule conflict with one of Tuesday or Thursday slots, enroll in the ASMO section (async).
Recordings from the ASML section (sync) will be made available to all enrolled students (sync and async).
Beginning in Fall 2025, we may be able to offer a "Python option" for the labs/assignments, subject to popular demand from students and TA support.

Course Pre-Requisites / Co-Requisites

Prerequisites: a prior course in applied linear regression (e.g., STA6093 or STA6166) or permission of the instructor; familiarity with R or Python software.

Course Objectives

- Learn the language of and the principles behind predictive modeling and model validation
- Learn and be able to use R/Python to apply different classes of statistical learning methods and interpret results, thereby expanding the arsenal of available analytical tools of researchers familiar with regression
- Establish command of methods, their application and reporting of the results through homework, exams and a final project
- Reinforce the use of R/Python as a statistical computing language for data science – for statistical inference, prediction, scientific computing and data visualization

Materials and Supply Fees

None

Required Textbooks and Software

- *An introduction to statistical learning*
- James, G., Witten, D., Hastie, T., & Tibshirani, R.
- 2021, 2st ed, latest printing (see www.statlearning.com), Springer
- ISBN 1071614177

R language and R Studio environment are freely available

Recommended Materials

N/A

Required Computer
Recommended Computer Specifications: https://it.ufl.edu/get-help/student-computer-recommendations/

Course Schedule (tentative)

Week	Topics
1	Course logistics; introduction to SML; nearest neighbors
2	Calculus-based probability: essential review of rvs, independence, moments, etc
3	Mathematical statistics: essential review of estimation and hypothesis testing
4	Maximum likelihood estimation (MLE) and its connections with Bayesian inference
5	Linear regression essentials
6	Linear regression in matrix form and its extensions
7	Classification: logistic regression; discriminant analysis and its flavors
8	Out-of-sample performance metrics: cross-validation and ROC curves
9	Model selection and regularization; penalized regression (ridge, lasso)
10	Spline-based models, generalized additive models
11	Classification and decision trees (single)
12	Ensembles of trees: bagging, random forest, gradient boosting
13	Support Vector Machines
14	Unsupervised learning: dimension reduction and clustering
15	Advanced topics (deep learning, high-dimensional inference, multiple testing, etc)
	Project

Depending on the typical student background, the above schedule may be adjusted to reflect the interests of the class (e.g., less time on the probability and statistics review and more time on advanced topics).
In particular, we'll likely expedite the (sync lecture) coverage of topics from weeks 2,3,6 so that we can have more time for more advanced topics. Full coverage of these topics (2,3,6) will be provided through lecture recordings.

Evaluation of Grades (Tentative)

Assignment	Total Points	Percentage of Final Grade
Homework Sets (8)	100 each	25%
Quizzes (8)	100 each	25%
Project	100	50%
		100%

Quizzes (online, in Canvas) and homework assignments will be closely matched to the course topics and will span approximately 4 hours of lectures. All quizzes and all homework sets will be weighted equally by converting the score to the 100-point scale first.
Project (requiring about 40 hours to complete) will consist of a proposal (10 pts), report (60 pts) and presentation (30 pts). Tentative project details are provided in the Supplements at the end of the syllabus and will be refined during the first month of the class.

Grading Policy
Tentatively, the following grading scale will be adopted; grades may be curved to the advantage of students.

Percent	Grade	Grade Points
93.4 - 100	A	4.00
90.0 - 93.3	A-	3.67
86.7 - 89.9	B+	3.33
83.4 - 86.6	B	3.00
80.0 - 83.3	B-	2.67
76.7 - 79.9	C+	2.33

73.4 - 76.6	C	2.00
70.0 - 73.3	C-	1.67
66.7 - 69.9	D+	1.33
63.4 - 66.6	D	1.00
60.0 - 63.3	D-	0.67
0 - 59.9	E	0.00

Academic Policies & Resources

To support consistent and accessible communication of university-wide student resources, instructors must include this link to academic policies and campus resources: https://go.ufl.edu/syllabuspolicies. Instructor-specific guidelines for courses must accommodate these policies.

Commitment to a Positive Learning Environment

The Herbert Wertheim College of Engineering values varied perspectives and lived experiences within our community and is committed to supporting the University's core values.

If you feel like your performance in class is being impacted by discrimination or harassment of any kind, please contact your instructor or any of the following:

- Your academic advisor or Graduate Coordinator
- HWCOE Human Resources, 352-392-0904, student-support-hr@eng.ufl.edu
- Pam Dickrell, Associate Dean of Student Affairs, 352-392-2177, pld@ufl.edu

Supplements: Project description (tentative, will be finalized in class)
The project will emphasize creative application of the methods developed in the course. Ideally, the application would be to your line of research and data (your own or of your immediate collaborators - advisor or fellow students). If you do not have suitable data, please check out the sources at the end of this description for the publicly available datasets. Otherwise, a good project could be a replication and extension of the results of a paper of interest that uses the methods from our course. "Creative application" does not allow merely running someone else's code without making other contributions. **Plagiarism is totally inappropriate and prohibited (just do not do it); it will result in a failing grade for the course. Course staff will run all project reports through UF Ithenticate. All work should be done individually (unless explicitly permitted by the instructor – for more ambitious projects).** *Projects already completed for other classes/causes are not acceptable. Example 1 - unacceptable: in a previous semester, a student wrote a paper for a journal or did a project for a different class, and now wants to submit it without major changes or additional SML type of work as the SML class project. Example 2 – acceptable: in a previous semester, a student wrote a paper for a journal or did a project for a different class, but wants to do a major extension of the work using the techniques learned in the SML class. This would make a potentially very good project, but the student needs to be explicit about what is new and what not. Only the new work will constitute the course project in this case.* The project will be used to assess the knowledge and skills that students acquired in the course; for that reason, the work must be done individually and without assistance from the course staff.
Deliverables: a one-page proposal, a short technical report and a short presentation as described below.
Deadlines: tentative and will be revised and announced appropriately each year
(TBA; tentatively, mid-November): submit your proposal by email to the instructor, so that we can meet on asap during the class time (individual slots TBA after your proposals have been received).
(TBA; tentatively, last week of classes): project presentations (as appropriate)
(TBA; tentatively, early during the exams week): final report (in pdf format, accompanying code and the actual data that you used, if using a publicly available source; put all in a folder named after you and create a zip or rar archive; test archive before submitting), submit using Dropbox file request – link to be provided in class

Expectations for the proposal (1 page):
The main goal behind the proposals is to ensure that the projects are neither too simple nor too ambitious (i.e., will require about 30 hours to complete – loosely, an equivalent of 3-4 weeks of homework effort, where writing will play a significant role), there is no duplication among students and that you have the necessary relatively clean data to analyze. Please specifically discuss what you propose to do (e.g., "big picture" and specific methods), why you focus on this particular problem (significance, motivation and relevance to the course) and available data (specifically, what are primary response variable(s) and features, what are n and p, etc). Your project should be "shovel ready", i.e., a bit of data preprocessing may be necessary but you should not be spending more than 20% (ideally, 10%) of your total time budget on cleaning and data manipulation. The typesetting format of the proposal should be the same as for the project (please read below).

Expectations for the report (8 pages):
Report should be organized as a short paper appropriate to your field; e.g., a short abstract (100 words), intro (including motivation), background and data, methods, analysis/results, conclusions/discussion. *Any software/languages/environments may be used for the project (i.e., not necessarily R). In most cases, the project should use several classes of methods (multiple linear regression or logistic regression as the baseline – possibly coupled with variable selection; at least one shrinkage method (if p is high) and/or a GAM (if p is low), and a tree-based ensemble method, typically, random forest (and definitely discuss variable importance summaries that may be extracted from it) and possibly boosting (because it is often the best method), ideally both) for classification or regression and examine out-of-sample performance of the methods using K-fold cross-validation.* If you are doing binary classification, please additionally include and discuss the ROC curves. If doing non-binary classification, please showcase a binary classification subproblem and include and discuss the ROC curves. The length is about 8 pages (not counting references, appendices or supplements) double-spaced, using 12 pt font: roughly 6 pages of text and 2 pages for your most essential tables and figures; **single-column only**. If necessary, the paper may have an Appendix with additional figures and tables. Data, code and other supplemental information should be made

available as part of "Supplementary Materials" unless the data are confidential (please discuss "deliverables" in the proposal). Please check out the project evaluation rubric in a separate file.

Expectations for the presentation:
Presentation should reflect a typical conference-style contributed talk (based on a short deck of slides that a student would prepare) that runs for 10-12 minutes, potentially followed by 3-5 minutes of questions/discussion. Depending on the year and the course, these would be either recorded in advance (by a student) or delivered synchronously. Regardless of the format, the presentations emphasize "live speech" rather than reading from a prompt.

Some sources of data for projects:
Google search "data for machine learning", e.g.,

UCI ML Repository
http://archive.ics.uci.edu/ml/

Kaggle Competitions
https://www.kaggle.com/datasets

PLOS One journal website
read info about data availability in papers of interest and at
http://journals.plos.org/plosone/s/data-availability