MA506: Probability & Statistical Inference

Fall 2023

Instructor:	Prashant Shekhar, PhD	**Email:**	shekharp@erau.edu
Class Time:	Tu,Th: 5:15pm – 6:30pm	**Class Venue:**	Bldg. IC Rm. 204
Office Hours (OH):	M,F: 2:00pm – 3:00pm	**OH Venue:**	Room 301.26, COAS.

Topics Included: This course will focus on using ideas from Probability and Statistics to solve problems in data science. The main topics included in the course are

1. Fundamentals of Statistical Learning
2. Regression
3. Classification
4. Statistical Model Selection

Why this course: The concepts that you learn in this course can be utilized to solve problems in the general area of machine learning and data science. The applications encompass multiple domains including healthcare, management, manufacturing, security and remote sensing etc.

Text Book: The course material will be provided to the students in the form of jupyter notebooks, pdfs and hand written notes. Additionally, the students can refer to the following textbooks to get a deeper understanding of the topics.

- An Introduction to Statistical Learning with Applications in R.
- Semiparametric Regression
- Elements of Statistical Learning

Attendance: I will try to take attendance in every class and I encourage you to participate in class activities. This is because attendance is found to be heavily correlated with the course grade and attending class everyday ensures that you will not miss any important announcement.

Grading: Your grade will be determined as:

- Homework: 40%
- Tests: 30%
- Final (Project and Presentation): 25%.
- Class participation and attendance: 5%

The grading is expected to follow the standard scale
A: 90% - 100%
B: 80% - 89.5%
C: 70% - 79.5%
D: 60% - 69.5%

F: <60%
However, based on the performance of the entire class, I might curve the grading scale later.

Homework: Your homework grade will be determined based on 4 programming oriented homeworks. You are required to use Python (Jupyter notebooks) to solve homework problems. These exercises will test the ability of the students to apply the concepts in statistical learning on various categories of data sets. **Please note that homeworks are only acceptable on canvas and not on email.** The course will implement the following late submission policy

- Late by less than 1 day, i.e. 24 hours (-20 points)
- Late between 1 day and 2 days (-40 points)
- Late between 2 day and 3 days (-60 points)
- Late between 3 day and 4 days (-80 points)
- Late by greater than 4 days (Not acceptable)

Tests: You will have 2 tests in the course. Make-ups on the test may be allowed only for valid extenuating circumstances when I am informed before the test takes place – please see me about conflicts as soon as they occur. **In case you are missing a Test, it is your responsibility to schedule a makeup Test with me within one week of the actual test date. After that makeup is not possible**.

Final (Project and Presentation): During the semester, you will be supervised to work on a project which combines classroom materials and real-world applications. The project together with the presentation is the final deliverable for the course. It is supposed to be a group project with teams consisting of 2–4 students. I will work with each of the team separately to identify a topic of your interest and find a relevant project in that domain. In case you are already working on a research problem related to the topics discussed in class, that can also be considered. I will announce project guidelines and rubric in due course.

Academic Integrity: Embry-Riddle Aeronautical University maintains high standards of academic honesty and integrity in higher education. To preserve academic excellence and integrity, the University prohibits academic dishonesty in any form, including, but not limited to, cheating and plagiarism. More specific definitions of these violations and their consequences are described in the Dean of Students' Honor Codes and Student Policies.

Disability Services: DSS Administration Office: Bldg 500; Contact: (386) 226-7916; email: dbdss@erau.edu

- Student Disability Services: Students with disabilities who believe that they may need accommodations in this class are encouraged to contact the Office of Disability Services. Professors cannot make appropriate disability accommodations. Students are encouraged to register with DSS at the beginning of the term to better ensure that such accommodations are implemented in a timely fashion. Accommodations are not granted until official notice is received from DSS.
- It is the responsibility of the student to notify DSS the date and time of test once s/he has been made aware of the scheduled test. DSS requires: (a) 2 business days minimum notification for tests and quizzes and (b) 5 business days minimum notification for final exams. Professors cannot make appropriate testing modification without notification from DSS.

University policy for using AI programs: The use of ChatGPT and other artificial intelligence (AI) tools in an educational setting must be guided by ethical principles and academic integrity. Students and faculty may use these tools for research support or aids to enhance their learning and scholarship but should

not rely solely on them to produce original work. The University recognizes the potential benefits of AI but acknowledges the potential risks, such as creating unauthentic, inaccurate, biased, or harmful content. Students should follow best practices to ensure authentic scholastic accomplishment and academic integrity, including avoiding plagiarism or machine ghostwriting. Ultimately, the University upholds the importance of intellectual honesty and ethical research practices and expects students to act in accordance with these principles when using AI tools.

MA 506 Probability and Statistical Inference

Instructor: Prashant Shekhar, PhD

Tentative Schedule for Fall 2023

SNo: Week of (class days)	*Topic*	*Homework*	*Learning Outcome*
Unit I: Course Basics			
1: 28^{th} Aug (Tu,Th)	Course introduction Python basics		1 1,2
2: 4^{th} Sept (Tu,Th)	Probability Statistical inference		1,2,15 1,2,15
Unit II: Regression			
3: 11^{th} Sept (Tu,Th)	Regression introduction Regression continued	HW1 released	4,5,6,7 4,5,6,7
4: 18^{th} Sept (Tu,Th)	Linear Regression from scratch Regression diagnostics		1,2,7 7
5: 25^{th} Sept (Tu,Th)	Confidence Intervals Confidence Intervals II	HW1 due	4,5,7 4,5,7
6: 2^{rd} Oct (Tu,Th)	Ridge and Lasso Regression Regression model comparison	HW2 released	4,5,6,12,14 7, 13
7: 9^{th} Oct (Tu,Thu)	No Lecture t-statistic, p-value, F-statistic and R^2	 Project details due	 3,7, 13
8: 16^{th} Oct (Tu,Th)	CV, GCV, K-fold CV No Lecture	HW2 due HW3 released	7, 11,12,13
9: 23^{th} Oct (Tu,Th)	Test:1 Review Test:1		
Unit III: Classification			
10: 30^{st} Oct (Tu,Th)	Classification introduction Logistic regression	 HW3 due	8,9 8,9
11: 6^{th} Nov (Tu,Th)	Logistic regression Regularized classification	HW4 released	8,9 12,13
12: 13^{th} Nov (Tu,Th)	Multiclass classification Classifier evaluation and selection		10,11,12 3, 10
13: 20^{th} Nov (Tu, Th)	Test 2 review No Lecture	HW4 due	
Course Conclusion			
14: 27^{th} Nov (Tu,Th)	Test 2 Project presentation I		 15
15: 4^{th} Dec (Tu,Th)	Project presentation II Project presentation III	 Project due	15 15

Learning outcome: After successful completion of this course, you will acquire knowledge to:

1. Understand the basics of statistical learning and its relation to machine learning.
2. Understand the basics of python and using it as a tool to solve problems in statistical learning.
3. Assess the quality of various statistical learning approaches based on various available metrics
4. Understand different problems in supervised learning.
5. Delve deeper into various aspects of linear regression.
6. Understand and implement multiple linear regression
7. Better understand various issues one might face while using linear regression as a tool to understand different properties of data.
8. Understand basics of classification and its relation to regression.
9. Use logistic regression as a tool to solve classification problems.
10. Compare different classification models for your own problem.
11. Use various resampling approaches to make an intelligent choice of a model for your own data science related problem.
12. Use regularization as a way to produce better models.
13. Understand various ways of controlling the complexity of your statistical learning model.
14. Interpret data and models in higher dimensions.
15. Apply the concepts learnt in class to problems of practical importance.