DS440: Data Mining

Fall 2022

Instructor: Prashant Shekhar, PhD	**Email:** prashant.shekhar@erau.edu
Class Time: MWF: 10:00AM – 10:50AM	**Class Venue:** Bldg COAS Rm. 304
Office Hours (OH): MWF: 12:00PM – 1:00PM	**OH Venue:** Room 301.26, COAS.

Topics Included: The goal of this course is to learn how to use the advanced mathematics language and computation tools to solve real-world problems. The topics of the course cover broad interdisciplinary problems whose solutions heavily depend on data mining and visualization. Students will gain hands-on experience on how to use Python based software tools to analyze large data sets. Broadly, major topics covered in this course include

1. Introduction to data mining
2. Python for numerical and scientific computations
3. supervised learning methods.
4. Unsupervised learning methods.

The concepts that you learn in this course can be utilized to solve problems in the general area of machine learning and data science. Starting from the basics of data mining, we will go deeper into concepts of classification, covering multiple widely used algorithms that find immense applications in a wide number of fields from medical to engineering. The remaining course addresses the unsupervised component of the data mining domain. The concepts learnt in this section would be useful for finding and analyzing patterns in data which don't have any labels (classes) due to various practical constraints such as data collection cost etc.

Text Book: The study material for the course would be provided to the students in the form of jupyter notebooks, pdfs and handwritten notes. Additionally, students are encouraged to refer the textbooks mentioned below for a much deeper understanding

- **Main text:** Ping-Ning Tan, et al., Introduction to Data Mining, Pearson Education, second edition, 2019
- **Additional references**:
  - James, Gareth, et al. An introduction to statistical learning, with Applications in R, Second Edition, 2021. *Link*
  - Murphy, Kevin P. Machine learning: a probabilistic perspective. MIT press, 2012.
- **Python and Machine Learning:** Aurelien Geron, Hands-on Machine Learning with Scikit_learning, Keras & TensorFlow. O'Reilly, second edition, September 2019.

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:

1. Homeworks: 40%
2. Test: 20%
3. Class participation and attendance: 10%
4. Project: 30%

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.

Test: You will have one main test (tentative date: 16^{th} Nov). 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.

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.

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 homework problems will test the ability of the students to apply the concepts learnt in class to real-life problems.

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.

DS 440 Data Mining

Instructor: Prashant Shekhar, PhD

Tentative Schedule for Fall 2022

Week Number: Starting Date (days)	*Topic*	*Homework*	*Learning Outcome*
Unit I: Data Mining Basics			
1: 29^{th} Aug (M,W,F)	Course introduction Python basics Computations in python: numpy		1,2 5 5
2: 5^{th} Sept (W,F)	Computations in python: scipy Data visualization in python: matplotlib		5 5
3: 12^{th} Sept (M,W,F)	Data characteristics Data quality and preprocessing Machine Learning in python: sklearn	HW1 released	1,2 1,2,5 1,2,5
Unit II: Supervised Learning			
4: 19^{th} Sept (M,W,F)	Introduction to regression Linear regression Linear regression II		4,5,7 4,5,7 4,5,7
5: 26^{th} Sept (M,W,F)	Ridge Regression Hurricane Ian Hurricane Ian		4,5,7
6: 3^{rd} Oct (M,W,F)	Regression Review Lasso Regression Overfitting & model selection in regression	HW1 due HW2 released	4,5,7 4,5,7 4,5,7
7: 10^{th} Oct (M,W,F)	Introduction to classification Logistic regression Decision trees		4,5,7 4,5,7 4,5,7
8: 17^{th} Oct (M,W)	Random forest Classifier evaluation	Project details due HW2 due	4,5,7 4,5,7
9: 24^{th} Oct (M,W,F)	Overfitting and classifier model selection Ensemble methods: bagging Ensemble methods: boosting	HW3 released	3 3,4,7 3,4,7
10: 31^{st} Oct (M,W,F)	K-nearest neighbor classification Support vector machines		4,5,7 4,5,7
Unit III: Unsupervised Learning			
	Association analysis: apriori		4,5,7
11: 7^{th} Nov (M,W)	Clustering: K-means Cluster evaluation	HW3 due/ HW4 released	4,5,7 4,5,7
12: 14^{th} Nov (M,W,F)	Test review Test Anomaly detection I		 4,5,7
13: 21^{th} Nov (M)	Anomaly detection II		4,5,7
Thanksgiving Break			
Course Conclusion			
14: 28^{th} Nov (M,W,F)	Course review Project presentation I Project presentation II	HW4 due	 6 6
15: 5^{th} Dec (M,W)	Project presentation III Project presentation IV	 Project due	6 6

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

1. Understand the main goals and types of data mining.
2. Identify a broad variety of real-world applications of data mining.
3. Identify the strengths and limitations of popular data mining techniques.
4. Explain the mathematics concepts behind several data mining methods such as decision trees, k-nearest neighborhood, Bayesian method, support vector machine, neural network, etc.
5. Gain hands-on experience in the use of machine learning software tools in Python.
6. Gain teamwork experience to handle real-world data-mining projects and expand their expertise beyond traditional book learning exercises.
7. Demonstrate the ability to solve problems beyond the scope of textbook exercises.