DS 540: Data Mining

Spring 2023

Instructor: Prashant Shekhar, PhD	**Email:** shekharp@erau.edu
Class Time: Tu,Th: 12:45PM – 2:00PM	**Class Venue:** Bldg COAS Rm. 108
Office Hours (OH): Tu,Th: 11:30AM – 12:30PM	**OH Venue:** Room 301.26, COAS.

Topics Included: This is a project-based course. Broadly, major topics covered in this course include

1. Approximately first 60% of the course
   (a) Fundamentals of Data Mining
   (b) Classification
2. Remaining course
   (a) Association
   (b) Clustering
   (c) Anomaly Detection

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:

- **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.

In class, I will be broadly following the excellent (chapter-wise) presentation slides prepared by the *main text* authors and available at *Link*. Additionally, I will provide notes and jupyter notebooks related to concepts discussed in class.

Attendance: I will take attendance in every class. I encourage you to participate in class activities because attendance is usually found to be heavily correlated with the course grade. Additionally, a portion of the course grade depends on class participation making attendance very important.

Grading: Your grade will be determined as:

1. Homeworks (4): 40%
2. Exams (2): 30%
3. Class participation and attendance: 5%
4. Project: 25%

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.

Exam: You will have 2 exams (tentative dates mentioned in the course schedule document). Make-ups on the exam 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 an exam, it is your responsibility to schedule a makeup exam with me within one week of the actual exam date. After that makeup exam is not possible**.

Project and Presentation: During the semester you will be supervised to work on a project which combines classroom materials and real-world applications. It is supposed to be a group project and I will work with each group separately to identify a topic of your interest and find a relevant project in that domain. I will announce project topics, guidelines, and rubric soon.

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. **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)

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
Testing Center: The Annex Building 2nd floor, room 217; Contact: (386) 226-2903; 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.
- DSS Testing Procedures: 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 days minimum notification.

DS 540 Data Mining

Instructor: Prashant Shekhar, PhD

Tentative Schedule for Spring 2023

Week Number: Days	*Chapter Number*	*Topic*	*Homework*	*Learning Outcome*
Data Mining Basics				
1: 12^{th} Jan (Th)	1	Course Introduction		1,2,13
2: 17^{th} Jan/ 19^{th} Jan (Tu,Th)	2	Types of data Data quality		1,3 1,3
3: 24^{th} Jan/ 26^{th} Jan (Tu,Th)	2	Similarity and Distance Guest lecture: Data Preprocessing	HW 1 released	1,4 1,4
Classification Basics				
4: 31^{st} Jan/ 2^{nd} Feb 2 (Tu, Th)	3	Review of topics Guest lecture: Rule Based		5,6 5,6
5: 7^{th} Feb/ 9^{th} Feb (Tu,Th)	4	Decision Trees: I Decision Trees: II	 HW 1 due	5,6 5,6
6: 14^{th} Feb/ 16^{th} Feb (Tu,Th)	3	Classifier Evaluation Validation and Overfitting	HW 2 released	5,6,7 5,6,7
Classification Algorithms				
7: 21^{st} Feb/ 23^{rd} Feb (Tu,Th)	4	K-Nearest Neighbor/ Exam review Exam 1		5,6 2,13
8: 28^{th} Feb/ 2^{nd} Mar (Tu,Th)	4	Support Vector Machines: I Support Vector Machines: II		5,6 5,6
9: 7^{th} Mar/ 9^{th} Mar (Tu,Th)	4	Ensemble Methods: I Ensemble Methods: II	HW 2 due HW 3 released	5,6,8 5,6,8
Spring Break				
11: 21^{st} Mar/ 23^{rd} Mar (Tu,Th)	4	Imbalanced Classes Naive Bayes		5,6,9 5,6,9
Association Analysis				
12: 28^{th} Mar/ 30^{th} Mar (Tu,Th)	5	Apriori Algorithm /Exam review Exam 2	HW 3 due HW 4 released	10
Clustering Analysis				
13: 4^{th} Apr/ 6^{th} Apr (Tu,Th)	7	KMeans Algorithm Cluster Evaluation		11 11
Anomaly Detection				
14: 11^{th} Apr/ 13^{th} Apr (Tu,Th)	9	Proximity-based Clustering-based		12 12
Project				
15: 18^{th} Apr/ 20^{th} Apr (Tu,Th)		Course conclusion Project Presentation I	HW 4 due	12 2,13
16: 25^{th} Apr/ 27^{th} Apr (Tu,Th)		Project Presentation II Project Presentation III	 Project due	2,13 2,13

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

1. Understand the basics of data mining and its relation to machine learning.
2. Use python as an efficient tool for data mining
3. Understand the types of data and evaluate its quality, distribution etc.
4. Implement foundational data preprocessing techniques for effective data mining.
5. Understand the basics of supervised learning
6. Implement and analyze prominent classification algorithms for data mining.
7. Evaluate and compare various classification algorithms
8. Combining multiple classification models to create better models.
9. Handle unbalanced classes in classification problems.
10. Understand and implement association analysis.
11. Understand and implement clustering analysis.
12. Understand and implement anomaly detection.
13. Apply the concepts learnt in class to problems of practical importance.