MA540: Data Mining

Spring 2022

Instructor: Prashant Shekhar, PhD	**Email:** shekharp@erau.edu
Class Time: MWF: 11:00AM – 11:50AM	**Class Venue:** Bldg IC Rm. 201
Office Hours (OH): MWF: 1:00PM – 2:00PM	**OH Venue:** Room 301.26, COAS.

Topics Included: This is a project-based course. The students will meet the instructor for three class hours per week. Monday and Wednesday lectures will be reserved for teaching concepts. The Friday lecture will be used for discussion of course project and focussing more on the coding aspects of the concepts discussd in class. 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. Quizzes (4): 20%
2. Exam (1): 20%
3. Class participation and attendance: 10%
4. Project:
    (a) Project based homeworks (3): 30%
    (b) Final project submission and group presentation: 20%

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.

Quizzes: You will have 4 short quizzes (20-25 minutes) based on the concepts we cover during the lectures.

Exam: You will have one main exam (tentative date: 30^{th} March). 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.

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 3 programming oriented homeworks building towards your course final project. You are required to use Python (Jupyter notebooks) to solve homework problems. These exercises will assist the step-wise development of your final project while possibly applying the concepts learnt simultaneously in class.

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.

ERAU Coronavirus Updates: Information on testing, vaccinations, health services, procedures and frequently asked questions are available here.

- **Face Masks Strongly Encouraged**: Consistent with current recommendations of the Centers for Disease Control and Prevention, and Embry-Riddle's long-standing culture of safety, all students (vaccinated or unvaccinated) are strongly encouraged to wear face masks indoors especially during their in-person classes and in other group indoor settings, including faculty office hours.
- **Vaccinations Strongly Encouraged**: All students are strongly encouraged to receive a vaccination against Covid-19. Vaccinations are available at convenient campus locations.

MA 540 Data Mining

Instructor: Prashant Shekhar, PhD

Tentative Schedule for Spring 2022

Week Number: Starting Date (days)	*Chapter Number*	*Topic*	*Homework*	*Learning Outcome*
Data Mining Basics				
1: 12th Jan (W,F)	1	Course Introduction Project/Python Examples		1 2,13
2: 17th Jan (W,F)	2	Types of data Data quality Project/Python Examples		1,3 1,3 2,13
3: 24th Jan (M,W,F)	2	Similarity and Distance Data Preprocessing Project/Python Examples	Quiz1	1,4 1,4 2,13
Classification Basics				
4: 31st Jan (M,W,F)	3	Decision Trees: I Decision Trees: II Project/Python Examples		5,6 5,6 2,13
5: 7th Feb (M,W,F)	3	Classifier Evaluation Validation and Overfitting Project/Python Examples	Quiz2	5,6,7 5,6,7 2,13
Classification Algorithms				
6: 14th Feb (M,W,F)	4	Rule-based K-Nearest Neighbor Project/Python Examples	HW1 due	5,6 5,6 2,13
7: 21st Feb (W,F)	4	Naive Bayes Project/Python Examples		5,6 2,13
8: 28th Feb (M,W,F)	4	Support Vector Machines: I Support Vector Machines: II Project/Python Examples		5,6 5,6 2,13
9: 7th Mar (M,W,F)	4	Ensemble Methods: I Ensemble Methods: II Project/Python Examples	HW2 due	5,6,8 5,6,8 2,13
Spring Break				
11: 21st Mar (M,W,F)	4	Imbalanced Classes: I Imbalanced Classes: II Project/Python Examples	Quiz3	5,6,9 5,6,9 2,13
Association Analysis				
12: 28th Mar (M,W,F)	5	Apriori Algorithm Project/Python Examples	Exam	10 2,13
Clustering Analysis				
13: 4th Apr (M,W,F)	7	KMeans Algorithm Cluster Evaluation Project/Python Examples		11 11 2,13
Anomaly Detection				
14: 11th Apr (M,W,F)	9	Proximity-based Clustering-based Project/Python Examples	HW3 due Quiz4	12 12 2,13
Week 15 & 16: Final project presentations/submissions				

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.