Data Mining, ISM 6136, Fall 2023, Room: BSN 115, Time: 6:30pm

Mohammadreza Ebrahimi, Assistant Professor
School of Information Systems and Management, University of South Florida
email: ebrahimim@usf.edu, Office hours: Tuesday 5:15-6:15pm (Room: CIS 2062)

TA: Surabhi Tripathi**,** stripathi5@usf.edu, Virtual Office Hours: TBD

Course Description and Objectives:

The past few years have seen an unprecedented explosion in the amount of data collected by businesses and have witnessed enabling technologies such as database systems, visualization tools and statistical and machine learning algorithms reach industrial strength. These trends have spawned a new breed of business analytics systems that go significantly beyond reporting capabilities, to support predictive modeling and the extraction of business insights from data. These trends have also created a new role of "data scientists" who are professionals with expertise in the concepts and tools necessary for the skilled use of these systems. This course introduces fundamental data science concepts, techniques, and their business applications. The course will enable students to identify the strength and limitations of fundamental data science technologies, and to select and apply appropriate analytical methods that can provide business managers and information systems professionals with new insights useful for solving hard business problems using data-driven approaches.

Course Materials/Textbook:

No Text. Learning materials, including slides, codes, datasets, etc. will be provided on Canvas in the appropriate modules.

General Learning Outcomes:

Upon completion of the course the student will be able to:
Demonstrate understanding of specific data mining methods
Describe different ways in which models can be evaluated
Use data mining tools to build descriptive and predictive models
Analyze a dataset using data analytics methods
Identify data mining opportunities in existing data sets
Describe global business scenarios where data and data mining can be applied

Software:

Python, Scikit Learn, PyTorch, TensorFlow

Schedule:

Date	Topic(s)	Notes
Aug 22	Data Analytic Thinking: Examples & Concepts (supervised learning, unsupervised learning, and generative AI)	
Aug 29	Data Mining Process and Tools (Python, Scikit Learn, PyTorch, TensorFlow)	
Sep 5	Decision Tree Induction (entropy, Gini-index, probabilistic metrics)	
Sep 12	Model Evaluation (supervised and unsupervised metrics)	Quiz 1
Sep 19	Ensembles Learning (bagging, boosting, random forest)	
Sep 26	Probability Models (Bayes theorem, Naïve Bayes), Evaluation-Part 2 (class imbalance, ROC, AUC)	Team Project Proposal due
Oct 3	Flipped Classroom (Team project discussion and materials up to Quiz 2)	Quiz 2
Oct 10	Neural Networks (MLP, Back-propagation)	Learning Reflection due
Oct 17	Distance Based Methods (K-NN, K-means Clustering), and Rule/Pattern Discovery	Individual Exercise due
Oct 24	Introduction to Deep Learning, ML Taxonomy	
Oct 31	Course Review and wrap-up	Exam
Nov 7	Project Presentations	Team Project Submission due Nov 9

Requirements and Due Dates:

1. Two Quizzes (Quiz 1 (5%), Quiz 2 (10%))

The quizzes will be based on the class notes, what we discussed in class, and provided readings and will be conducted in the class.

2. Exam (25%)

The exam will be based on class notes and on relevant readings/discussions. Anything discussed in a class presentation or in any discussion can be tested in these - unless explicitly excluded. Please see the Exam Policy Section for more details. The exam will be held in the class.

3. Individual Data Mining Exercise (20%)

Deliverable: A video that you can upload to canvas

For this part you should use a different dataset than any of the ones you have used in the past in this course (i.e., do not use the data from the 5- to 7-minute video, and do not use the dataset that you are using in your project). You can select one from the UCI machine learning repository for

example.
Create and upload a video that shows you using Python and Scikit Learn to analyze this dataset. The video must have the following parts:

(1) It should demonstrate you using at least two of classification tree, neural networks, naïve Bayes, clustering, and visualization to analyze this data. You need to justify why you select your algorithms based on the data or domain characteristics.

(2) In the analysis shown in the video you must clearly point out and discuss three different concepts that you learned in the class that you are applying in the analysis. For example, one example of a concept learned is that "concept: when the dependent variable is skewed you must look at the confusion matrix, not just the overall accuracy". The onus is on you to make sure you clearly state in the video the concept that you are showing as well as the "concept number" so I know at the end of the video that you have highlighted three concepts. Important: This has to be your own independent analysis, any video where your concepts look similar to another one will receive an automatic zero. You also cannot ask me or the TA to vet your concepts because that is part of this exercise, to see if you can identify important principles from this class that you can showcase.

(3) You should keep the entire video to five to seven minutes in total. If you have to break this video into two or three parts and provide links to each part that is OK as well for me (since it is possible that you do this part in pieces as you work through this exercise).

4. Team Data Mining Project (25%)

In groups of 3-4 identify a dataset on your own, perform a data mining analysis and summarize the results in an in-person presentation (7 minutes max), and an 8-page single spaced paper with font size of 12 (and with results included) due at the end of the term. For the presentation, it is mandatory that all team members be present. The students are required to determine their team members no later than the third week of the class. **The last slide of the project presentation needs to include a table that shows the exact contribution of each team member. Including this table implies the approval of all team members.** Projects will be graded based on:

(1) Novelty, interestingness, and importance. Hence, new datasets, your own datasets and/or datasets
relating to any important contemporary problem in business or society would be valued more. If everything is on GitHub, and you take it and do some minor tweaks it is not sufficient.

(2) Managing presentation time (under 7 minutes) and managing the space in your report (8 pages).

(3) Questions that frame the exercise and then recommendations at that end. Think of these two as steps before and after the data mining/machine learning component. What are the one or two central key business questions that a CEO or leader wants to know? Make sure to structure your analysis to answer these one or two initial questions. At the end of the analysis, specifically based on your findings, what recommendations would you have for the business leader?

(4) Relevance of the work to the questions. Instead of showing everything you can do on that data, structure your work clearly to answer these questions. To this end, all your work should be relevant to the questions that were framed. Random visualizations and models being shown for the sake of being shown that do not add value should be eliminated from your report.

(5) Depth of the methodology and attention to detail in the analysis (Justification of DM models, evaluation metrics, comparison to baselines, Interpretation of the results or meaningful visualizations that answer your proposed questions).

(6) Quality of the final paper submitted (paper alone will have 30% of the grade for this assignment).

5. Class Participation (15%)

Participation in class includes two activities: (1) learning reflections (5%), and (2) data mining event/topic discussion (10%).

Learning Reflection, DUE Sep 27 (5%)

The learning reflection serves as a brief but important mid-term evaluation that will be turned in on Sep 27. The format is a one-page document, describing in short complete sentences (not bullet points) what you learned and how it changed the way you might approach a data mining problem at work. All late submissions will be penalized two points per day of delay for that exercise/project but still need to be made within a week of the deadline at the latest to get credit. Submissions beyond one week may not be evaluated for credit.

Data Mining Event/Topic Discussion, DUE Nov 12 (10%)

This activity has two parts: (1) a 5–7-minute discussion by the student, and (2) a one-page document that summarizes the discussion.

For the first part, throughout the semester, each student is required to discuss one preferably recent application of data-driven analytics (includes, but not limited to, data mining, AI, or machine learning) in today's business domains (e.g., education, cybersecurity, health, finance). The sources for this activity could be white papers, newspapers, magazines, social media, or related research topics of your interest from conference papers. The student is expected to instigate the discussion and provide enough key information so that others can participate in the discussion after you are done with your initial discussion. The discussion will be graded based on three main criteria: application, data, analytics, and results:

- Communicating the business application and the importance of the targeted problem,
- Describing the characteristics of the data set(s) within the application domain, including the data sources, data types, dependent variables, independent variables, important data fields, etc.
- Clearly communicating the applied analytics, the results, and the offered value.

For the second activity, the students submit a one-page summary of their discussion after the first activity is complete. Full credit will be assigned only after you submit the summary document that includes the above information.

In each week, 3-4 students will be presenting. The presentation schedule will be determined in the first day of the class.

Grading

You are guaranteed at least the following grades if your final score falls as follows (your grade may be
higher based on the relative performance of the entire class):
97 and above: A+
94 - 97: A
90 - 94: A-
85 - 90: B+
70 - 85: B
50 - 70: A passing grade from B- and below
A total score of below 50 will receive an F.

Modules and Learning Outcomes

Module 1. Data Analytic Thinking

Learning Outcomes:
Students will be able to provide examples of how businesses can use data intelligently.
Students will be able to distinguish between patterns and models
Students will be able to define "data mining"
Students will be able to explain why there are so many different models
Students will be able to provide a concrete example of how businesses can take an insight derived from data into operations.

Module 2. Data Mining Process and Tools

Learning Outcomes:
Students will be able to describe limitations of secondary data analysis.
Students will be able to list some potential pitfalls in data mining.
Students will be able to describe the data mining process.

Module 3. Decision Tree Induction

Learning Outcomes:
Students will be able to provide a high-level description of an algorithm to build decision trees for prediction.
Students will be able to define node impurity and describe how it can be used in attribute selection for tree induction.
Students will be able to build decision trees from data using a data mining tool.

Module 4A. Model Evaluation

Learning Outcomes:

Students will be able to list different metrics for evaluating predictive models.
Students will be able to describe the train/validate/test methodology and the importance of partitioning data.

Module 4B. Ensemble Learning

Students will be able to describe several ensemble learning models and their differences.

Students will be able to apply ensemble learning methods on data and interpret the results.

Module 5. Probability Models
Learning Outcomes:
Students will be able to describe how a trained neural network converts inputs to outputs
Students will be able to build neural networks from data using a data mining tool.
Students will be able to explain how Naïve Bayes probability models can be built from data
Students will be able to build Naïve Bayes predictive models from data using a data mining tool.

Module 6A. Unsupervised Learning and Similarity-Based Techniques
Learning Outcomes:
Students will be able to describe how to cluster and classify using similarity-based techniques
Students will be able to build clusters and similarity-based classifiers using a data mining tool.
Students will be able to describe how recommender systems work as a similarity-based technique.

Module 6B. Unsupervised Learning and Learning Rules/Patterns from Data
Learning Outcomes:
Students will be able to define association rules
Students will be able to build association rules from data using a data mining tool.
Students will be able to explain how association rules are learned.

Module 8. Introduction to Deep Learning, Research Applications, and New Directions
Learning Outcomes:

Students will gain hands-on experience with creating a deep learning model using PyTorch or TensorFlow.

Students will be able to elaborate on the end-to-end application of the data mining process in real-world business applications in both research and industry.

Students will be able to dissect research/industry case studies: identify the business application domain, data mining method, results from the analytics, and business insights derived from the analytics,

Students will be able to determine new directions for data driven analytics in each case study.

Module 9. Class Projects

Learning Outcomes:

Students will gain hands-on data analytics skills that allow them to conduct end-to-end data mining processes to address non-trivial business problems and derive business insights that are useful to upper-level managers.

Honor Code
The policy of the University of South Florida on academic dishonesty states:
Each individual is expected to earn his or her degree on the basis of personal effort. Consequently, any form of cheating on examinations or plagiarism on assigned papers constitutes unacceptable deceit and dishonesty. This cannot be tolerated in the university community and will be punishable, according to the seriousness of the offense, in conformity with this rule. Cheating is defined as follows:

(a) the unauthorized granting or receiving of aid during the prescribed period of a course-graded exercise: students may not consult written materials such as notes or books, may not look at the paper of another student, nor consult orally with any other student taking the same test,

(b) asking another person to take an examination in his or her place,

(c) taking an examination for or in place of another student,

(d) stealing visual concepts, such as drawings, sketches, diagrams, musical programs and scores, graphs, maps, etc. and presenting them as one's own,

(e) stealing, borrowing, buying, or disseminating tests, answer keys or other examination material except as officially authorized, research papers, creative papers, speeches, etc.,

(f) stealing or copying of computer programs and presenting them as one's own.

Class Modality and Recording Policy

- The class modality is in-person. That is, students are required to attend the classes in-person. This is required for interactive learning and the effectiveness of hands-on sessions. Medical emergencies are an exception, which require providing evidence to the Dean of students. While attending all sessions is strongly recommended, students are not allowed to be absent for more than 3 sessions.

- The instructor is not obligated to record classes. However, in case the instructor decides to record the classes, they may be streamed online. Thus, student’s voice and video will be included in the class recordings. It is the student’s responsibility to make sure the privacy of their surroundings and background is maintained.

Late Submission Policy

- There will be no make-up provided for missed assignments. Of course, emergencies such as illness and family emergencies occur. In such cases, please contact the Dean of Students office. The Dean of Students is equipped to verify emergencies and pass confirmation on to all your classes. For consistency, I ask all students to do this in the event of an emergency. Do not send any personal/medical information to the instructor or TAs; all such information should go through the Dean of Students.

- All deliverables have a due time. Canvas will show all submissions that are not submitted by the due time as late. There will be a penalty of **5 points** for each day that the deliverable is over-due. Deliverables that are over-due for a week (before your next class starts) will receive a grade of zero.

Out-of-Class Communication Policy

- In addition to my in-person office hours, the best way to reach out to me is via email (not Microsoft Teams or Canvas as I am not checking them regularly). I check my emails regularly but please allow at least two business days (excluding weekends and holidays) before sending a follow-up email.

Exam/Quiz Policy

- Exams will be held in class. Students who miss an in-class quiz/exam with prior permission of the instructor (due to documented emergency situations) will have to wait until the end of the term to take a make-up exam (only one will be given and that will be based on the entire course material).

- To encourage understanding vs. memorization, the instructor may allow bringing a double-sided one-page note composed by students. The students are fully responsible for composing their own page.

Generative AI Policy

Using generative AI (e.g., ChatGPT) for the purpose of learning is allowed. However, students are responsible for understanding the obtained output and must be able to explain and defend the generated content. If the student choose to use generative AI they need to inform the instructor by adding a text to the submission that cite the utilized generative AI.

Change Policy

The syllabus and course schedule are subject to change. These changes will be communicated via the Canvas announcement tool, which sends email to everyone on the class list. It is the responsibility of students to check their email and course announcements to stay current.

Students with Special Needs

Students in need of academic accommodations for a disability may consult with Students with Disabilities Services to arrange appropriate accommodations well in advance. Students are required to give reasonable notice prior to requesting accommodation.

Online Proctoring

All students must review the syllabus and the requirements, including the online terms and video testing requirements, to determine if they wish to remain in the course. Enrollment in the course is an agreement to abide by and accept all terms. Any student may elect to drop or withdraw from this course before the end of the drop/add period.

Online exams and quizzes within this course may require online proctoring. Therefore, students will be required to have a webcam (USB or internal) with a microphone when taking an exam or quiz. Students understand that this remote recording device is purchased and controlled by the student and that recordings from any private residence must be done with the permission of any person residing in the residence.

To avoid any concerns in this regard, students should select private spaces for the testing. Students with concerns may discuss the location of an appropriate space for the recordings with their instructor or advisor.

Students must ensure that any recordings do not invade any third-party privacy rights and accept all responsibility and liability for violations of any third-party privacy concerns.

Students are strictly responsible for ensuring that they take all exams using a reliable computer and high-speed internet connection. Setup information will be provided prior to taking the proctored exam. To use Honorlock, students are required to download and install the *Honorlock Google Chrome extension*. For additional information please visit the *USF online proctoring student FAQ* and *Honorlock student resources*.

OTHER COURSE POLICIES

Students who miss an in-class exam with prior permission of the instructor (due to documented emergency situations) will have to wait until the end of the term to take a make-up exam (only one will be given and that will be based on the entire course material).

Students who anticipate being late for any deliverable due to religious observance should inform the instructor by the end of the first week of class.

Students may not re-distribute any class material of the class in any outside forum without approval of the instructor.

Per USF Policy 10-006: Registration Changes Including Course Change, Cancellations, Withdrawals, and Auditing, an auditing student "attends the class as a listener." Please identify yourself as an auditor in the course within the first two weeks, and let me know if you wish to discuss your role in the classroom.

For global USF policies that also apply to this course, please refer to:
https://www.usf.edu/provost/faculty/core-syllabus-policy-statements.aspx

EMERGENCY PREPAREDNESS

In the event of an emergency, it may be necessary for USF to suspend normal operations. During this time, USF may opt to continue delivery of instruction through methods that include but are not limited to:

Canvas, Teams, Skype, and email messaging and/or an alternate schedule. It's the responsibility of the student to monitor the course site for each class for course specific communication, and the main USF, College, and department websites, emails, and messages for important general information.

Acknowledgement

Thanks to *Dr. Balaji Padmanabhan* for kindly providing the original material for the syllabus and the course.