Data Mining, ISM 6136, Spring 2023, Room: BSN 231, Time: 12:30pm-4:15pm

Mohammadreza Ebrahimi, Assistant Professor
School of Information Systems and Management, University of South Florida
email: ebrahimim@usf.edu, Office hours: Monday 4:30-5:30pm (Room: CIS 2062)

TA: Varun Krishna Ramakrishnan, varunkrishna@usf.edu

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

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. Readings will be provided in 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:

SAS Enterprise Miner, Weka, Azure ML Studio

Schedule:

Date	**Topic(s)**	**Notes**
Feb 6	Data Analytic Thinking: Examples & Concepts	
Feb 13	Data Analytic Thinking; SAS E-Miner, AzureML & Weka	
Feb 20	Decision Tree Induction	
Feb 27	Model Evaluation	
Mar 6	Ensembles and Special Topics in Tree Induction	Individual Exercise 1 due
Mar 13	No class	Spring Break
Mar 20	Probabilistic Models	Team Project Proposal
Mar 27	Neural Networks	Learning Reflection due
Apr 3	Distance Based Methods (K-NN, Clustering and Recommender Systems)	
Apr 10	Pattern Discovery	
Apr 17	Research Examples	Individual Exercise 2 due
Apr 24	Industry Examples	Exam
May 1	Project Presentations	Final Group Project Submission due May 4

Requirements and Due Dates:

1. 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.
Exam will be held in class.

2. Individual Data Mining Exercise 1 (15%).

Deliverable: One 5-minute video of you using any tool to perform a data analytics exercise (use Weka preferably). The video should take a dataset and show some results from analyzing this. Make sure that one part of the video has you first talking to the camera so that the video shows your face and voice before we switch entirely to seeing the screen. The content of the video is up to you, but the goal is to showcase some of your best skills in using the tool to analyze data. The videos will be evaluated based on:

(1) clarity of the video, so production quality needs to be reasonable to that I can actually see the screen and what’s being done clearly,

(2) how the video showcases some techniques you learned in the class,

(3) novelty – how much it shows new things you might have picked up on your own,

(4) your ability to interpret what you are doing as you speak so I can tell you understand clearly what you are doing,

(5) how well you explain some of the output/results in the video, and

(6) if the video tells one overall "story" about the data well.

3. Individual Data Mining Exercise 2 (20%).

Deliverable: A video that you can upload to canvas (ideally as a link to an unlisted YouTube video).
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 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 Azure ML or SAS Enterprise Miner to analyze this dataset. The video
must have the following parts:

(1) It should demonstrate you using at least two of classification trees, 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. Final Group 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 presentation it is recommended that all team members 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 one point 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 a 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

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.

Students will be able to describe how gains charts can be constructed for modeling responses to promotions.

Module 4B. Week Five. Weka, Azure ML Studio and SAS Enterprise Miner Videos)

Module 5. Other Representations: Ensembles, Neural Nets and 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 6. Week Seven. Unsupervised Learning 1: 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 7. Unsupervised Learning 2: Learning 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.
Students will be able to explain how randomization can be used to generate confidence intervals that can
be used in rule learning.

Module 8. Research, Applications and New Directions (If time permits)
Learning Outcomes:

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.

Covid-19 Procedures
All students must comply with university policies and posted signs regarding COVID-19 mitigation measures, including wearing face coverings and maintaining social distancing during in-person classes. Failure to do so may result in dismissal from class, referral to the Office of Student Conduct and Ethical Development, and possible removal from campus.
Additional details are available on the University’s Core Syllabus Policy Statements page:
https://www.usf.edu/provost/faculty/core-syllabus-policy-statements.aspx

Class Recording
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.

Online Proctoring
In case the exams are conducted using online proctoring tools. Keeping the audio and video (microphone and camera) on during such exams and quizzes is a must. If the student is not

willing to use these, the student is asked not to register for this course. 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. The University library and other academic sites at the University offer secure private settings for recordings, and 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. Setup information will be provided prior to taking the proctored exam. For additional information about online proctoring, you can visit the online proctoring student FAQ at
http://www.usf.edu/innovative-education/resources/student-services/online-proctoring.aspx

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.

A student who does not submit a deliverable on time may be penalized up to the entire points unless he/she has documented proof of a medical emergency or explicit permission of the instructor.

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

Students in need of academic accommodations for a disability may consult with Students with Disabilities

Services to arrange appropriate accommodations. Students are required to give reasonable notice
prior to
requesting an accommodation.

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 MoBull messages for important general
information.