THE UNIVERSITY OF TEXAS AT EL PASO
COLLEGE OF SCIENCE
DEPARTMENT OF MATHEMATICAL SCIENCES

Course #:	DS 5474/STAT 5474 (CRN 19867/19035)	
Course Title:	Introduction to Data Mining	
Credit Hrs:	4.0	
Term:	Fall 2022	
Course Meetings & Location:	Lectures: TR 12:00-1:20 pm in Bell Hall 130 Labs: TR 1:30-2:20 pm in Bell Hall 130	
Prerequisite Courses:	Generalized linear models; Basic statistical programming skills Department approval required	
Course Fee:	N/A	
Instructor:	Dr. Michael Pokojovy	
Office Location:	Bell Hall 227	
Contact Info:	Phone: (915) 747-6761 E-mail address: mpokojovy@utep.edu Fax # 915-747-6502 (Math Department) Emergency Contact: 915-747-5761 (Math Department)	
Office Hrs:	TBA (virtual office hour)	
Textbook(s), Materials:	Required:	1) James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer https://www.statlearning.com/ 2) Hastie, T., Tibshirani, R., and Friedman, J.H. (2008), Elements of Statistical Learning, 2nd ed., Chapman & Hall. http://www-stat.stanford.edu/~tibs/ElemStatLearn/
	Recom-mended:	1) Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. Cambridge, MA: The MIT Press 2) Vapnik, V. N. (1998). Statistical Learning Theory. Wiley 3) Devroye, L., Gyor, L., and Lugosi, G. (1996). A Probabilistic Theory of Pattern Recognition. Springer

Course Description and Learning Outcomes:

Contents:

This course will cover seminal statistical machine learning and data mining techniques – with extra focus on statistical learning algorithms and their theoretical aspects. The course material is arranged in two main categories: unsupervised learning and supervised learning. Additionally, important aspects of statistical emulation as a powerful machine learning tool will be discussed. A tentative outline is provided below.

Course Activities/Assignments:

Lectures and labs accompanied by program demonstrations will be given. Lecture slides and/or lecture notes will be provided. Homework projects will be assigned biweekly. For the final project, students will be expected to independently locate/choose a dataset (from any field) and propose an interesting research question before asking for the instructor's approval of the project. Each student will be required to present his or her work.

Assessment of Course Objectives: Academic performance in this class will be the only factor used in determining the course grade. No extra credit work will be available to improve on any grade. Student performance will be evaluated based on the grades (see Grading Policy below) for homework projects and the final project/presentation. No late coursework will be accepted (with exception of some extreme situations as deemed by the instructor).

Course Schedule: Important dates:

- Duration: 8/22/2022 – 12/1/2022
- Course drop deadline: Fri, 10/28/2022 (No "W" are guaranteed for dropping the course after this date!)
- Final "exam": TBA
- Grades officially available online: Thu, 12/15/2021

Tentative outline:

Getting Started

1. Introduction to Data Mining
2. R and/or Python Preliminaries: Downloading and Installing; Introduction; Dealing with Large Data Sets
3. Data Preparation: Data Cleaning; Sampling and Partitioning Data; Missing Value Imputation; Exploration

Part I: Unsupervised Learning

4. Cluster Analysis: Proximity/Similarity Matrix; Hierarchical Clustering; K-Means; Self-Organizing Maps (SOM)
5. Principal Components: PCA and its extensions such as Principal Curves and Surfaces; Factor Analysis (Exploratory and Confirmatory)
6. Multidimensional Scaling (MDS): Metric and Non-Metric MDS
7. Web Mining: Google PageRank

Part II: Supervised Learning

8. Linear Regression: Least Squares Estimation; Assessment and Validation; Model Diagnostics; Generalized LS; Partial/Total LS
9. Variable Screening and Selection: Variable Relevance/Importance – RELIEF algorithm; Univariate/Bivariate Screening; Stepwise Variable Selection and Regularization – Ridge regression, LASSO and its variants, etc.
10. Logistic Regression: Model Assessment, Validation, and Regularization; Classification Errors, Odds Ratio, ROC, etc.; Generalized Linear Models (GLM)
11. Regression/Classification Trees: Pros and Cons; Impurity Measures; Pruning; Tree Size Selection; and Interpretation; Generalization to Regression, Censored Survival Data, and Longitudinal Data
12. Ensemble Models: Boosting; Bagging and Random Forests; Stacking

Part III: Statistical Emulation

13. Design and Analysis of Computer Experiments
14. Empirical Bayesian Analysis for Computer Experiments

Grading Policy: 60% Homework/project assignments
40% Final project (both implementation/report and presentation)

The usual grading scale will be used for this course:

90–100% = A
80–89% = B
70–79% = C
60–69% = D
0–59% = F

Make-up Policy: If class is missed for a valid (as deemed by the instructor) & documented reason and the instructor is informed beforehand, the in-class assignments may be made-up for full credit. All other assignments must be turned in on time.

Attendance Policy: You are expected to attend class so that you may turn in the in-class assignments and homework projects. Lecture and lab attendance is further required and helpful to decide the grade in "borderline" situations. Two or more unexcused absences (from the lab and/or the lecture) may result in an instructor-initiated drop or final grade reduction. *Your academic advisor may be consulted before final action is decided and taken.* If a student has to be absent from a class, he/she will be responsible for catching up with course material. Late arrivals are not permitted. Being late by 10 minutes or more or leaving the classroom before the class is dismissed will be considered an absence.

Scholastic Integrity Policy: Academic dishonesty is prohibited and is considered a violation of the UTEP Handbook of Operating Procedures. It includes, but is not limited to, cheating, plagiarism, and collusion. Cheating may involve copying from or providing information to another student, possessing unauthorized materials during a test, or falsifying research data on laboratory reports. Plagiarism occurs when someone intentionally or knowingly represents the words or ideas of another as ones' own. Collusion involves collaborating with another person to commit any academically dishonest act. Any act of academic dishonesty attempted by a UTEP student is unacceptable and will not be tolerated. All suspected violations of academic integrity at The University of Texas at El Paso must be reported to the Office of Student Conduct and Conflict Resolution (OSCCR) for possible disciplinary action. To learn more, please visit HOOP: Student Conduct and Discipline.

Technology Requirements

Some of the course content will be delivered via the Internet through the Blackboard learning management system. Ensure your UTEP e-mail account is working and that you have access to the Web and a stable web browser. You will need to have access to a computer/laptop.

IMPORTANT: If you encounter technical difficulties beyond your scope of troubleshooting, please contact the UTEP Help Desk as they are trained specifically in assisting with technological needs of students. The instructor and the TA are ***not*** responsible for this sort of assistance!

Copyright Statement for Course Materials:

All materials used in this course are protected by copyright law. The course materials are only for the use of students currently enrolled in this course and only for the purpose of this course. They may not be further disseminated.

Netiquette:

Communication online can be challenging. Therefore, please keep these netiquette (network etiquette) guidelines in mind. Failure to observe them may result in disciplinary action.

- Always consider audience. This is a college-level course; therefore, all communication should reflect polite consideration of others' ideas.
- Respect and courtesy must be provided to classmates and to the instructor at all times. No harassment or inappropriate postings will be tolerated.
- When reacting to someone else's message, address the ideas, not the person.
- Blackboard is ***not*** a public internet venue; all postings to it should be considered private and confidential. Whatever is posted on in these online spaces is intended for classmates and instructor only. Please do not copy documents and paste them to a publicly accessible website, blog and/or other space.

Disability Statement and Accommodations Policy

The University is committed to providing reasonable accommodations and auxiliary services to students, staff, faculty, job applicants, applicants for admissions, and other beneficiaries of University programs, services and activities with documented disabilities in order to provide them with equal opportunities to participate in programs, services, and activities in compliance with sections 503 and 504 of the Rehabilitation Act of 1973, as amended, and the Americans with Disabilities Act (ADA) of 1990 and the Americans with Disabilities Act Amendments Act (ADAAA) of 2008. Reasonable accommodations will be made unless it is determined that doing so would cause undue hardship on the University. Students requesting an accommodation based on a disability must register with the UTEP Center for Accommodations and Support Services (CASS). Contact the Center for Accommodations and Support Services at 915-747-5148, or email them at cass@utep.edu, or apply for accommodations online via the CASS portal.

COVID-19 Precaution Statement	Please stay home if you have been diagnosed with COVID-19 or are experiencing COVID-19 symptoms. If you are feeling unwell, please let me know as soon as possible, so that we can work on appropriate accommodations. If you have tested positive for COVID-19, you are encouraged to report your results to covidaction@utep.edu, so that the Dean of Students Office can provide you with support and help with communication with your professors. The Student Health Center is equipped to provide COVID-19 testing.
College of Science Policies:	All grades of Incomplete must be accompanied by an Incomplete Contract that has been signed by the instructor of record, student, Department Chair and the Dean. Although UTEP will allow a maximum of one year to complete this contract, the College of Science requests it be limited to month based upon completion data. A grade of Incomplete is only used in extraordinary circumstances confined to a limited event such as a missed exam, project, or lab. If the student has missed a significant amount of work (e.g., multiple assignments or tasks), a grade of Incomplete is not appropriate or warranted.
Resources:	UTEP provides a variety of student services and support: Technology Resources • Help Desk: Students experiencing technological challenges (email, Blackboard, software, etc.) can submit a ticket to the UTEP Helpdesk for assistance. Contact the Helpdesk via phone, email, chat, website, or in person if on campus. Academic Resources • UTEP Library: Access a wide range of resources, including online, full-text access to thousands of journals and eBooks plus reference service and librarian assistance for enrolled students. • Math Tutoring Center (MaRCS): Ask a tutor for help and explore other available math resources. • Individual Resources • Military Student Success Center: Assists personnel in any branch of service to reach their educational goals. • Center for Accommodations and Support Services: Assists students with ADA-related accommodations for coursework, housing, and internships.

Disclaimer:

This syllabus may be subject to changes if these are deemed necessary by the instructor. Despite all efforts, this syllabus may contain typos and errors.

Last updated: 8/19/2022

Attachment: Program Grading Policy

Criterion	Approx. % of Grade	Excellent (100%)	Adequate (80%)	Poor (60%)	Not Met (0%)
Program Specifications / Correctness	50%*	No errors, program always works correctly and meets the specification(s).	Minor details of the program specification are violated, program functions incorrectly for some inputs.	Significant details of the specification are violated, program often exhibits incorrect behavior.	Program only functions correctly in very limited cases or not at all.
Readability	20%	Code is clean, understandable, and well-organized.	Minor issues with consistent indentation, use of whitespace, variable naming, or general organization.	At least one major issue with indentation, whitespace, variable names, or organization.	Major problems with at three or four of the readability subcategories.
Documentation	5%	Code is well-commented.	One or two places that could benefit from comments are missing them **or** the code is o*verly* commented	File header missing, complicated lines or sections of code uncommented or lacking meaningful comments.	No file header or comments present.
Code Efficiency	20%	Code uses the best approach in every case.	Code uses poorly-chosen approaches (though correct in result) in at least one place.	Code uses poorly-chosen approaches (though correct in result) in at least two places.	Many things in the code could have been accomplished in an easier, faster, or otherwise better fashion.
Assignment Specifications	5%	No errors	Minor details of the assignment specification are violated, such as files named incorrectly or extra instructions slightly misunderstood.	Minor details of the assignment specification are violated, such as files named incorrectly or extra instructions significantly misunderstood.	Significant details of the specification are violated, such as extra instructions ignored or entirely misunderstood.