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

Course #:	STAT 5474 (CRN 29112)
Course Title:	Introduction to Data Mining
Credit Hrs:	4.0
Term:	Spring 2019
Course Meetings & Location:	TR 3:00–4:20 pm (lecture) in Undergraduate Learning Center 338 TR 4:30–5:20 pm (lab) 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:	TR 2:00–3:00 pm in Bell Hall 227
Textbook(s), Materials:	Required: 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) James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer 2) Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. Cambridge, MA: The MIT Press 3) Vapnik, V. N. (1998). Statistical Learning Theory. Wiley 4) 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 some commonly-used statistical machine learning and data mining techniques – with extra focus on statistical learning algorithms and their theoretical aspects. The materials are arranged in two main categories: unsupervised learning and supervised learning. A tentative outline of the specific topics is provided below.
Course Activities/Assignments:	In-class lectures accompanied by program demonstrations and class discussions will be given. Both lecture slides and lecture notes will be provided. Sporadically, in-class quizzes/exams will be administered. Homework projects will assigned weekly or bi-weekly. For the final project, students will be expected to independently locate/select a dataset (from any field) and propose an interesting research question before asking for the instructor's approval of the project. Each student will have the opportunity to present his or her work in class.

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 in-class quizzes/exams, 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: 1/22/2019 – 5/9/2019
- Course drop deadline: Fri, 4/5/2019 (No "W" are guaranteed for dropping the course after this date!)
- Finals week: 5/13/2019 – 5/17/2019
- Grades officially available online: Tue, 5/23/2019

Tentative schedule:

Getting Started

1. Introduction to Data Mining
2. R Preliminaries: Downloading and Installing; Introduction to R; Dealing with Large Data Sets in R
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 - the RELIEF algorithm; Univariate/ Bivariate Screening; Step-wise 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

Grading Policy: 60% Homework/project assignments
10% In-class quizzes and/or exams
30% Final project (both implementation 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.

Academic Integrity Policy: The University policy is that all suspected cases or acts of alleged scholastic dishonesty must be referred to the OSCCR for investigation and appropriate disposition. Any student who commits an act of scholastic dishonesty is subject to discipline. Scholastic dishonesty includes, but is not limited to cheating, plagiarism, collusion, the submission for credit of any work or materials that are attributable in whole or in part to another person, taking an examination for another person, any act designed to give unfair advantage to a student or the attempt to commit such acts. Each student is responsible for notice of and compliance with the provisions of the Regents' Rules and Regulations, which are available for inspection electronically at http://www.utsystem.edu/bor/rules/homepage.htm

All students are expected and required to obey the law, to comply with the Regents' Rules and Regulations, with System and University rules, with directives issued by an administrative official in the course of his or her authorized duties, and to observe standards of conduct appropriate for the University. A student who enrolls at the University is charged with the obligation to conduct himself/herself in a manner compatible with the University's function as an educational institution.

Any student who engages in conduct that is prohibited by Regents' Rules and Regulations, U. T. System or University rules, specific instructions issued by an administrative official or by federal, state, or local laws is subject to discipline, whether such conduct takes place on or off campus or whether civil or criminal penalties are also imposed for such conduct.

Civility Statement: This is a class where participation is required. You will be seated at your desk or in front of a computer for the duration of your class and you are expected to follow the lecture/discussion and at various times complete in-class assignments. You are not allowed to browse the Internet during class time or work on any other material. If you regularly do not complete in-class assignments in a satisfactory manner, participate in class, or if you work on other material in class you will have points deducted from your in-class assignments portion of your grade. All class participants must and will be treated with respect.

Please do not use cell phones, iPods, MP3 players, blue tooth devices, etc. during class. Cell phones and pagers should be set to silent or vibrate, and any calls should be taken outside of class. Please do not wear headsets or blue tooth devices during class. Please don't talk in class. Cell phone calculators may not be used on quizzes or exams. Active participation in class is expected, teamwork in class will be implemented.

Disability Statement: If a student has or suspects she/he has a disability and needs an accommodation, he/she should contact The Center for Accommodations and Support services (CASS) at 747-5148 or at <cass@utep.edu> or go to Room 106 Union East Building. The student is responsible for presenting to the instructor any CASS accommodation letters and instructions.

Military Statement: If you are a military student with the potential of being called to military service and/or training during the semester, please contact me by the end of the first week of class

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.

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.