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

Course #:	STAT 5474 (CRN 17834)
Course Title:	Introduction to Data Mining
Credit Hrs:	4
Term:	Spring 2020 (Instruction 01/21/2020-05/07/2020)
Course Meetings & Location:	9:00-10:20 am MW Bell Hall 130A 10:30-11:20 am MW Bell Hall 130
Prerequisite Courses:	Generalized linear models; Some R programming experiences would be plus, though not required.
Instructor:	Xiaogang Su
Office Location:	Bell Hall 320
Contact Info:	Phone: (915) 747-6860 [O] xsu@utep.edu
Office Hours:	3:00-4:00pm MW
Class Web page:	https://sites.google.com/site/xgsu00/home/stat5474
Textbook(s), Materials:	Required: Hastie, T., Tibshirani, R., and Friedman, J. H. (2008), *Elements of Statistical Learning*, 2nd Edition Chapman and Hall. ISBN-13: 978-0387848570 http://www-stat.stanford.edu/~tibs/ElemStatLearn/
	Suggested: James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. ISBN-13: 978-1461471370
	Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective.* Cambridge, MA: The MIT Press. ISBN-13: 978-0262018029
	Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.
	Devroye, L., Gyor, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer.
	Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning*, 2nd Edition. Cambridge, MA: The MIT Press.
Course Description and Learning Outcomes:	This course is intended to cover some commonly-used statistical machine learning and data mining techniques, with more focus on the most technical part - statistical machine learning algorithms. The materials are arranged in two main categories: unsupervised learning and supervised learning. A tentative outline of the specific topics is provided below.

Topic Outline

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; 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.

Course Activities/Assignments:

Presentations and projects will be assigned throughout the semester. NO LATE COURSEWORK WILL BE ACCEPTED, EXCEPT IN EXTREME SCENARIOS.

Assessment of Course Objectives:

Each student will be evaluated by the quality of his/her own assigned presentations as well as their contribution to the discussions during other students' presentations.

Course Schedule:

03/16 - 03/20	Spring Break
03/27	Cesar Chavez Birthday
04/10	Spring Study Day
05/11 - 05/15	Final Exam Period
04/03	Course Drop Date

Final Exam Schedule:

TBA

Grading Policy: There will be a number of computer assignments and a final project. The assignments make up 60% and the final project makes up 35% to your final score. For the final project, students are given the freedom to select data from whatever field they are interested in. Students should make their own plans to collect data, raise interesting research questions, and consult the instructor for the adequacy of the project. Also, each student will have the opportunity to present their work in class. There will also be a few in-class quizzes or exams, which make up 10%. No make-up exam will be given and no late project submission is accepted without justifiable reasons.

Note that there is 5% extra credit; however, the five extra credits are only applicable to those who complete all assignments and exams without ANY unexcused absence from class attendance. Letter grades are determined according to the following scale:

Grade	Score
A	90-100
B	80-89
C	70-79
D	60-69
F	<60

Make-up Policy: All other assignments must be turned in on time.

Attendance Policy: Class attendance is REQUIRED and helpful to decide borderline grades. If a student has to be absent from a particular class, he/she will be responsible for catching up with course material. A late arrival of 15 minutes or more will be considered as an absence. Students will be dropped for four or more unjustified absences from class or lab session. Your academic advisor will be consulted before final action is decided and taken. Any unjustified absences from class or lab session will cause loss of eligibility of receiving extra credits. If you expect to miss up to 10 class hours for ANY REASON, then please do not consider taking this course.

Academic Integrity Policy: Please see http://academics.utep.edu/Default.aspx?tabid=23785

Civility Statement: This is a class where participation is required. You will be participating in classroom discussions. All students will be treated with respect. Calculators may not be shared during quizzes and exams. Please do not use cell phones, pagers, 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 you have a disability and need classroom accommodations, please contact The Center for Accommodations and Support Services (CASS) at 747-5148, or by email to cass@utep.edu, or visit their office located in UTEP Union East, Room 106. For additional information, please visit the CASS website at www.sa.utep.edu/cass.

Military Statement:	If you are a military student with the potential of being called to military service and /or training during the course of the semester, you are encouraged to contact me as soon as possible.
UTEP College of Science Policies	Watch out for the UTEP drop/withdraws deadline for the semester. The College of Science will remain aligned with the University and not approve any drop requests after that date. All grades of Incomplete must be accompanied by an Incomplete Contract that has been signed by the instructor of record, student, departmental 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.