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

Course #:	DS 6474 / STAT 5474 / STAT 6474(GR) / 4474 (UG) (14661/16778/15669/15642) Cross-Listed
Course Title:	Introduction to Data Mining / Statistical Machine Learning I
Credit Hrs:	4
Term:	Fall 2023 (Instruction 08/28/2023-12/07/2023)
Course Meetings & Location:	09:30 - 10:50 am TR Bell Hall 130 11:00 - 11:50 am TR 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:	01:00-01:50 pm TR or by appointment; also available by Zoom
Teaching Assistant (TA)	TBA
Class Web page:	See Blackboard

Textbook(s), Materials:

Required:

(GR) 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/

(UG) James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*. Springer. ISBN-13: 978-1461471370

Suggested:

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 Schedule:

08/28	Class Starts
11/03	Class Drop Deadline
12/11 - 12/15	Final Exam Period
Holidays	
09/04	Labor Day
11/23-11/24	Thanksgiving

Final Exam Schedule: TBA

Assessment and Grading Policy:

This course is structured around several computer assignments and a final project, which will collectively determine 90% of your final score. The assignments carry a weightage of 60%, while the final project contributes 30%. The final project offers you the freedom to explore data from any field that interests you. You are responsible for selecting the data, formulating intriguing research questions, and devising a plan to collect the necessary information. While working on your project, feel free to seek guidance from the instructor to ensure its adequacy and relevance. To foster interactive learning, each student will present their final project to the class, facilitating knowledge-sharing and fostering a collaborative environment. Apart from assignments and the final project, there might be occasional quizzes or exams that, along with attendance, will constitute 15% of your final grade. It is important to note that no make-up quizzes/exams will be offered, and NO LATE SUBMISSIONS WILL BE ACCEPTED, except in exceptional circumstances and only with prior approval from the instructor.

Please be aware that there is an opportunity for 5% extra credit in this course. However, this additional credit will only be granted to those students who fulfill the following criteria: completing all assignments, exams, and quizzes, and maintaining good attendance records for both in-person classes and Blackboard sessions (if any). Letter grades for this course will be determined according to the following scale:

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

Attendance Policy:

Class attendance is REQUIRED and helpful to decide borderline grades. If a student must be absent from a particular class, he/she will be responsible for notifying the instructor and catching up with course material. FOUR or more unexcused absences will result in an instructor-initiated drop or grade failing / reduction. Your academic advisor will be consulted before final action is decided and taken. If you expect to miss TEN or more class/lab hours for ANY reason, please don’t 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 semester, you are encouraged to contact me as soon as possible.

UTEP College of Science Policies

Please watch out for the UTEP drop/withdraws deadline for the Fall 2023 semester – 11/03/2023 (Friday). 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.