INTRODUCTION TO

DATA MINING

CS 363D, FALL 2021

52535 & 52540

TTH 11:00-12:30 & 12:30-2:00

PROFESSOR

Angie Beasley
angie.beasley@utexas.edu
GDC 6.314

TAs

Ishan Nigam
ishannigam@gmail.com

Shivang Singh
shivang.singh.tx@gmail.com

Colette Montminy
cmontminy@utexas.edu

Please see Canvas for up to date office hours.

COURSE DESCRIPTION

During "small data" days, data was difficult and tedious to collect. Only the data that was needed to answer a specific question was collected. But we have now entered the world of "big data," where data is constantly collected on everything and everyone. We have an unprecedented amount of data, and classic statistical approaches often do not work on the type of data we have. This is where machine learning comes in...

In this class, you will learn machine learning algorithms to find patterns in large datasets. We will cover classification, clustering, anomaly detection, and association analysis.You will use Python's scikit-learn packages, and Jupyter Notebooks, two industry-standard tools for data mining.

Data mining has already changed the way in which many important decisions are made. In today's data-driven world, it is increasingly critical to understand how these algorithms come to their conclusions and the correct ways to interpret and apply their results.

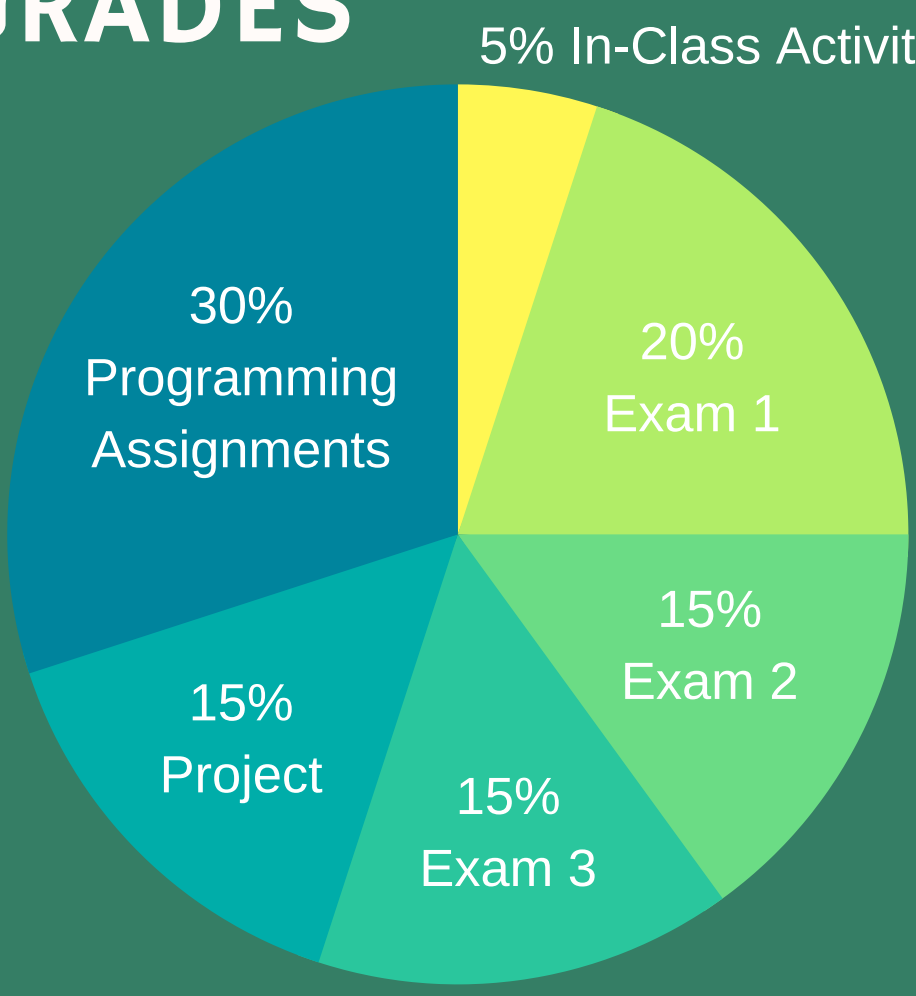
Computer Science 363D and 378 (Topic: Introduction to Data Mining) may not both be counted. Prerequisites: The following coursework with a grade of at least C- : Computer Science 429 (or 310) or 429H (or 310H); Mathematics 362K or Statistics and Data Sciences 321 (or Statistics and Scientific Computation 321); and Mathematics 340L, 341, or Statistics and Data Sciences 329C (or Statistics and Scientific Computation 329C).

TEXTBOOK

Introduction to Data Mining
--> Second Edition <--

by Pang-Ning Tan, Michael Steinbach, Vipin Kumar

GRADES



A	>=94
A-	< 94
B+	< 90
B	< 87
B-	< 84
C+	< 80
C	< 77
C-	< 74
D+	< 70
D	< 67
D-	< 64
F	< 60

All numbers are absolute and will not be rounded up or down at any stage.

PROJECT

Your project will be to analyze a dataset using the techniques taught in class.
This project gives you the opportunity to practice the full data science process on a real-world dataset. You will work in groups of 4 people. The results of your analysis will be presented in a Jupyter Notebook. This project will make up 15% of your final grade.

PROJECT SCHEDULE

10/1	Group selection deadline
10/22	Dataset selection deadline
11/21	Final project due

PROGRAMMING ASSIGNMENTS

There will be 6 programming assignments, each equally weighted to total 30% of your final grade.

Programming assignments must be completed using Python 3 and Jupyter Notebooks. Assignments may be worked individually or in pairs. If you work in pairs, you are expected to use the proper pair programming method.

LATE ASSIGNMENTS

You will have 3 late days in 1-day units (that is, 1 minute to 24 hours late = 1 late day) to use throughout the semester. You may divide your late days across the programming assignments in any way you wish. Once you have used all of your late days, late assignments will no longer be accepted.

In the case of pair programming, each member of the pair must have enough late days to cover the late submission. So if the pair submits their code 2 days late, each member must have two late days remaining to use and each member will lose two late days.

To use late days, you only need to submit the assignment. You do not need to email the instructor or the TA, you do not need to indicate that you are using late days. Your late days will be deducted according to when your assignment is submitted. If you submit a late assignment without enough late days to support it, you will receive a zero for that assignment.

Contact me if there are extenuating circumstances or if you get sick.

IN-CLASS ACTIVITIES

Throughout the semester, there will be in-class activities. They will vary in nature and will be random in occurrence. Some will be graded for correctness, and some will only be graded for completion. You may drop your 2 lowest of these and the remaining will make up 5% of your final grade.

REGRADE REQUESTS

All grades will be posted on Canvas.
You have **one week** from the date the grade is posted to dispute your grade. The TAs will be grading the assignments. First contact the TAs and see if you can resolve your differences. If you can not resolve your differences, you may contact me to explain the situation. We will not entertain any grade disputes after one week.

COURSE SCHEDULE

Subject to change at instructor's discretion.

8/26	Introduction
8/31	Data Prep, Exploration, Feature Engineering
9/2	Dimensionality, Jupyter, Pandas

CLASSIFICATION

9/7	Decision Trees [Ch 3.1-3.3]
9/9	Decision Trees (cont.)
9/14	Overfitting & Cross-Validation [Ch 3.4-3.9]
9/16	Nearest Neighbor [Ch 4.3]
9/21	Naive Bayes [Ch 4.4]
9/23	Evaluating Classifiers [Ch 4.11]
9/28	Ensemble Methods [Ch 4.10]
9/30	SVMs [Ch 4.9]
10/5	Neural Nets [Ch 4.7]
10/7	Neural Nets (cont.)
10/12	EXAM 1
10/14	TBD

CLUSTERING

10/19	Clustering & K-means [Ch 7.1-7.2]
10/21	Density-Based Clustering [Ch 7.4]
10/26	Hierarchical Clustering [Ch 7.3]
10/28	Evaluating Clusters [Ch 7.5]
11/2	Anomaly Detection [Ch 9]
11/4	EXAM 2

ASSOCIATION ANALYSIS

11/9	Apriori [Ch 5.1-5.5]
11/11	Scalability Issues & Rule Generation [Ch 5.3]
11/16	FP Growth [Ch 5.6]
11/18	Compact Itemsts/Skewed Distributions[Ch 5.4]
11/21	Project Due
11/23	Evaluating Association Patterns [Ch 5.7-5.8]
11/25	Thanksgiving Break
11/30	Sequential Patterns [Ch 6.4]
12/2	Semester Wrap-Up
TBD	EXAM 3

ACADEMIC INTEGRITY

Each student in the course is expected to abide by the University of Texas Honor Code:
"As a student of The University of Texas at Austin, I shall abide by the core values of the University and uphold academic integrity."

This means that work you produce on assignments and exams is all your own work, unless it is assigned as group work. I will make it clear for each exam or assignment whether collaboration is allowed or not.

You are responsible for understanding UT's Academic Honesty Policy which can be found here: https://catalog.utexas.edu/general-information/appendices/appendix-c/student-discipline-and-conduct/#subchapter11400.prohibitedconduct

If you submit code or work that is not your own, you will be guilty of plagiarism and subject to academic disciplinary action, including failure of the course and being reported to the Dean of Students.

STUDENT SUPPORT AND ACCOMODATIONS

I am committed to creating an accessible and inclusive learning environment for everyone. Please let me know if you experience any barriers to learning so I can work with you to ensure you have equal opportunity to participate fully in this course. Please contact me as soon as possible if the material being presented in class is not accessible to you, if any of the physical space is difficult for you, or to discuss any other accommodations you may need.

If you are a student with a disability, or think you may have a disability, and need accommodations please contact Services for Students with Disabilities (SSD): http://diversity.utexas.edu/disability/.

UNIVERSITY RESOURCES

The Counseling and Mental Health Center (CMHC) provides counseling, psychiatric, consultation, and prevention services: http://cmhc.utexas.edu/

Student Emergency Services (SES) can be contacted in cases of family emergency/death in the family, medical emergencies, fire or natural disasters, academic difficulties due to crisis or emergency situations, interpersonal violence (stalking, harassment, physical and/or sexual assault): http://deanofstudents.utexas.edu/emergency/

If you have concerns about the safety or behavior of fellow students, TAs or Professors, call BCAL (the Behavior Concerns Advice Line): 512-232-5050. Your call can be anonymous. If something doesn't feel right – it probably isn't. Trust your instincts and share your concerns.

RELIGIOUS HOLY DAYS

By UT Austin policy, you must notify me of your pending absence at least fourteen days prior to the date of observance of a religious holy day. If you must miss a class, an exam, a work assignment, or a project in order to observe a religious holy day, I will give you an opportunity to complete the missed work within a reasonable time after the absence.

Q DROP POLICY

If you want to drop a class after the 12th class day, you'll need to execute a Q drop before the Q-drop deadline, which typically occurs near the middle of the semester. Under Texas law, you are only allowed six Q drops while you are in college at any public Texas institution. For more information, see: http://www.utexas.edu/ugs/csacc/academic/adddrop/qdrop

SHARING COURSE MATERIALS IS STRICTLY PROHIBITED

Sharing of Course Materials is Prohibited. No materials used in this class, including, but not limited to, videos, assessments, quizzes, exams, papers, projects, homework assignments, in-class materials, lecture hand-outs, review sheets, and problem sets, may be shared online or with anyone outside of the class unless you have my explicit, written permission.

Unauthorized sharing of materials promotes cheating. It is a violation of the University's Student Honor Code and an act of academic dishonesty. I am well aware of the sites used for sharing materials, and any materials found online that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in sanctions, including failure in the course.

Class recordings are reserved only for students in this class for educational purposes and are protected under FERPA federal law (20 U.S.C. § 1232g; 34 CFR Part 99). Class recordings may not be shared outside the class in any form. Violation of this restriction by a student could lead to Student Misconduct proceedings.

Students may not record all or part of class, livestream all or part of class, or make/distribute screen captures, without advanced written consent of the instructor. Classes may be recorded by the instructor. Students may use instructor's recordings for their own studying and notetaking. Instructor's recordings are not authorized to be shared with anyone without the prior written approval of the instructor. Failure to comply with requirements regarding recordings will result in a disciplinary referral to the Dean of Students Office and may result in disciplinary action.