INFO 523 - Data Mining and Discovery

Spring 2026

Course description

INFO 523 Data Mining and Discovery- This course will introduce students to the concepts and techniques of data mining for knowledge discovery. It includes methods developed in the fields of statistics, large-scale data analytics, machine learning, pattern recognition, database technology and artificial intelligence for automatic or semi-automatic analysis of large quantities of data to extract previously unknown patterns. Topics include understanding varieties of data, data preprocessing, classification, association and correlation rule analysis, cluster analysis, outlier detection, and data mining trends and research frontiers. We will use software packages for data mining, explaining the underlying algorithms and their use and limitations. The course include laboratory exercises, with data mining case studies using data from many different resources.

Prerequisites

Students are assumed to know the basics in computer programming (e.g., variables, arrays, loops, if-then conditions), statistics (e.g., normal distribution, significance tests), and relational database (e.g., ER-diagram, SQL statements).

Course Format

Online, asynchronous recorded lectures.

Course Objective

INFO 523 is a required course in the College of Information Science's M.S. in Data Science / Information Science programs. As a multidisciplinary field, the course introduce concepts and work from many areas critical to information studies including statistics, machine learning, pattern recognition, database technology, and data visualization.

Learning Outcomes

By the end of this course, students will:

- Understand a large set of concepts of data mining and knowledge discovery.
- Evaluate and use algorithms and software packages to perform data mining analyses.
- Explain and interpret results from data mining analyses.

Textbooks:

- `[Data mining conceptual]` Jiawei Han, Jian Pei, Hanghang Tong. Data Mining Concepts and Techniques. 4th edition. Morgan Kaufmann, 2023.
- `[Python Data Analysis]` Wes McKinneyz. Python for Data Analysis, 3E. O'Reilly, 2023.
- `[Prac Stats for DS]` Peter Bruce, Andrew Bruce, Peter Gedeck. Practical Statistics for Data Scientists. O'Rielly, 2016.
- `[ISL]` James Garth, Witten Daniela, Hastie Trevor, Tibshirani Robert. An Introduction to Statistical Learning. Springer, 2021/2023.
- `[Pract Time Series]` Nielsen Aileen. Practical Time Series Analysis. O'Rielly, 2020.

Course Schedule

An up-to-date schedule, assignments, and due dates can be found on the course website: datamineaz.org.

Course competencies

This course is a core requirement for the MS in Data Science. It will help you master the following competencies:

- MS1. Students will establish the ability to exercise the four key techniques of computational thinking: decomposition, pattern recognition, abstraction, and algorithms.
- MS2. Students will obtain the skills of collecting, manipulating, and analyzing different types of data at different scales, and interpreting the results properly.

Course Community

UArizona Community Standard

All students must adhere to the UArizona Student Rights & Responsibilities: The University of Arizona is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, and accountability. Citizens of this community commit to reflect upon these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity.

Inclusive community

It is my intent that students from all diverse backgrounds and perspectives be well-served by this course, that students' learning needs be addressed both in and out of class, and that the diversity that the students bring to this class be viewed as a resource, strength, and benefit. It is my intent to present materials and activities that are respectful of diversity and in alignment with UArizona's Commitment to Diversity and Inclusion. Your suggestions are encouraged and appreciated. Please let me know ways to improve the effectiveness of the course for you personally, or for other students or student groups.

Furthermore, I would like to create a learning environment for my students that supports a diversity of thoughts, perspectives and experiences, and honors your identities. To help accomplish this:

- If you have a name that differs from those that appear in your official UArizona records, please let me know! You'll be able to note this in the Getting to know you survey.
- If you feel like your performance in the class is being impacted by your experiences outside of class, please don't hesitate to come and talk with me. If you prefer to speak with someone outside of the course, your academic dean is an excellent resource.
- I (like many people) am still in the process of learning about diverse perspectives and identities. If something was said in class (by anyone) that made you feel uncomfortable, please let me or a member of the teaching team know.

Communication

All lecture notes, assignment instructions, an up-to-date schedule, and other course materials may be found on the course website: datamineaz.org.

I will regularly send course announcements via email and Slack, make sure to check one or the other of these regularly. If an announcement is sent Monday through Thursday, I will assume that you have read the announcement by the next day. If an announcement is sent on a Friday or over the weekend, I will assume that you have read it by Monday.

Where to get help

- If you have a question during lecture, feel free to ask it! There are likely other students with the same question, so by asking you will create a learning opportunity for everyone.
- The teaching team is here to help you be successful in the course. You are encouraged to attend office hours to ask questions about the course content and assignments. Many questions are most effectively answered as you discuss them with others, so office hours are a valuable resource. Please use them!
- Outside of class and office hours, any general questions about course content or assignments should be posted on the course Slack. There is a chance another student has already asked a similar question, so please check the other posts on Slack before adding a new question. If you know the answer to a question posted on Slack, I encourage you to respond!

Check out the Support page for more resources.

I want to make sure that you learn everything you were hoping to learn from this class. If this requires flexibility, please don't hesitate to ask.

- You *never* owe me personal information about your health (mental or physical) but you're always welcome to talk to me. If I can't help, I likely know someone who can.
- I want you to learn lots of things from this class, but I primarily want you to stay healthy, balanced, and grounded during this crisis.

Lectures

The goal of the lectures is for them to be as interactive as possible. My role as instructor is to introduce you new tools and techniques, but it is up to you to take them and make use of them. A lot of what you do in this course will involve writing code, and coding is a skill that is best learned by doing. Therefore, as much as possible, you will be working on a variety of tasks and activities throughout each lecture.

You are expected to have a laptop for class so that you can take part in the lecture exercises. See the UArizona Libraries loaner technology if you need a loaner laptop.

Assessment

Assessment for the course is comprised of four components: attendance and participation, reading quizzes, homework assignments, and projects.

- **Exercises + Discussions** (4), due at the end of term, completed individually. Each exercise + discussion pair is worth 5% of the grade.

  These are recurring applied exercises paired with D2L discussions. To receive full credit, you must submit required exercise work and participate meaningfully in discussions by reflecting on methods, results, and limitations.

- **Reading quizzes** (5), due every other week (roughly), completed individually. Each quiz is worth 2% of the grade. Lowest quiz score is dropped.

  Reading quizzes will be in D2L. They always cover reading that is due since the previous quiz and up to and including the deadline for the given quiz.

- **Homework assignments** (5), due every week (roughly), completed individually. Each homework assignment is worth 8% of the course grade. Lowest homework assignment score is dropped.

  Homework assignments are due by 5:00pm on the indicated day on the course schedule.

- **Data Science Experience** (1), due at the end of term, completed individually. The assignment provides options for extracurricular data science opportunities with the purpose of producing a meaningful report. This assignment is work 10% of the course grade.

- **Final Project** (1), end of semester, completed in teams.

  - Teams will pick a dataset of interest to them and develop a data mining workflow in Python (problem -> solution). The final project is worth 30% of the course grade.

  The deliverables for each project will include a data visualization and/or analysis, a write up of the process and findings, and a presentation. For the project, you will be encouraged to think beyond traditional data mining techniques (i.e., explore and utilize modern best practices)

  The final project will have a peer review component to provide at least one round of feedback during the process of development. Teams will provide periodic peer feedback to their teammates while working on the projects as well as upon completion. The scores from the peer evaluations, along with individual contributions tracked by commits on GitHub, will be used to ensure that each student has contributed to the teamwork.

  All team members must take part in the presentation. Presentations must be given via Zoom if the team requires. My preference is that the team stick to one method of delivery (all presenters in person or all presenters on Zoom), but I realize a lot can change throughout this semester, and we'll adjust accordingly.

All work is expected to be submitted by the deadline and there are no make ups for any missed assessments. See Late work policy for policies on late work.

Grading:

The final course grade will be calculated as follows:

Category	Percentage
Homeworks (5)	40%
Final Project (1)	20%
Exercises + Discussions (5)	20%
Reading Quizzes (5)	10%
DS-Experience (1)	10%

While there are no specific points allocated to participation, we will be recording your participation (mainly via slack) in periodically throughout the semester, and this information will be used as "extra credit" if you're in between two grades and a minor bump would help. For reference, this would equate to a 1-2% bump.

The final letter grade will be determined based on the following thresholds:

Letter Grade	Final Course Grade
A	>= 90
B	80 - 89.99
C	70 - 79.99
D	60 - 69.99
E	50-59.99
F	< 60

These are upper bounds for grade cutoffs, depending on the class performance the cutoffs may be lowered but they won't be increased.

Final Project

Your task for this project is to go beyond the usual data mining projects on, for example, Kaggle.

This is intentionally vague – part of the challenge is to design a project that showcases best your interests and strengths.

One requirement is that your project should feature some element that you had to learn on your own. This could be a package you use that we didn't teach in class (e.g., a package for building 3D visualizations) or a workflow (e.g., making a package) or anything else. If you're not sure if your "new" thing counts, just ask!

More information will be provided throughout the semester.

Teams

You will be assigned to a team for each of your projects. You are encouraged to interact with your teammates as much as possible. All team members are expected to contribute equally to the completion of each project and you will be asked to evaluate your team members after each assignment is due. Failure to adequately contribute to an assignment will result in a penalty to your mark relative to the team's overall mark.

You are expected to make use of the provided GitHub repository as their central collaborative platform. Commits to this repository will be used as a metric (one of several) of each team member's relative contribution for each project.

Five tips for success

Your success on this course depends very much on you and the effort you put into it. The course has been organized so that the burden of learning is on you. I will help you be providing you with materials and answering questions and setting a pace, but for this to work you must do the following:

1. Complete all the preparation work before reviewing lectures.
2. Ask questions. As often as you can. Ask me, ask your friends, ask your teammates. This will help you more than anything else. If you get a question wrong on an assessment, ask us why. If you're not sure about the homework, ask. If you hear something on the news that sounds related to what we discussed, ask. If the reading is confusing, ask.
3. Do the readings.
4. Do the homework. The earlier you start, the better. It's not enough to just mechanically plow through the exercises. You should ask yourself how these exercises relate to earlier material, and imagine how they might be changed (to make questions for an assignment, for example).
5. Don't procrastinate. The content builds upon what was taught in previous weeks, so if something is confusing to you in Week 2, Week 3 will become more confusing, Week 4 even worse, etc. Don't let the week end with unanswered questions. But if you find yourself falling behind and not knowing where to begin asking, come to office hours and work with a member of the teaching team to help you identify a good (re)starting point.

Course policies

Academic honesty

TL;DR: Don't cheat!

Please abide by the following as you work on assignments in this course:

- **Collaboration:** No work is permitted to be completed collaboratively.
  - You may discuss homework assignments with other students; however, you may not directly share (or copy) code or write up with other students. Unauthorized sharing (or copying) of the code or write up will be considered a violation for all students involved, resulting in a 0 for the assignment.
  - You may not discuss or otherwise work with others on the assignments. Unauthorized collaboration or using unauthorized materials will be considered a violation for all students involved. More details will be given closer to the assignment date.
  - For the project, collaboration within teams is not only allowed, but expected. Communication between individuals at a high level is also allowed however you may not share code or components of the project across teams.
  - On assignments you may not directly share work (including code) with another student in this class.
- **Online resources:** I am well aware that a huge volume of code is available on the web to solve any number of problems. Unless I explicitly tell you not to use something, the course's policy is that you may make use of any online resources (e.g., StackOverflow) but you must explicitly cite where you obtained any code you directly use (or use as inspiration). Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism.
- **Use of generative artificial intelligence (AI)**: You should treat generative AI, such as ChatGPT, the same as other online resources. There are two guiding principles that govern how you can use AI in this course[1]:
  1. *Cognitive dimension:* Working with AI should not reduce your ability to think clearly. We will practice using AI to facilitate—rather than hinder—learning.
  2. *Ethical dimension*: Students using AI should be transparent about their use and make sure it aligns with academic integrity.

[1]These guiding principles are based on *Course Policies related to ChatGPT and other AI Tools* developed by Joel Gladd, Ph.D.

- **AI tools for code:** You may make use of the technology for coding examples on assignments; if you do so, you must explicitly cite where you obtained the code. Any recycled code that is discovered and is not explicitly cited will be treated as plagiarism. You may use these guidelines for citing AI-generated content.
- **AI tools for narrative:** Unless instructed otherwise, you may not use generative AI to write narrative on assignments. In general, you may use generative AI as a resource as you complete assignments but not to answer the exercises for you.

You are ultimately responsible for the work you turn in; it should reflect *your* understanding of the course content.

If you are unsure if the use of a particular resource complies with the academic honesty policy, please ask a member of the teaching team.

College of Information Science Academic Integrity Policy

This policy agreed upon by faculty in the College of Information Science at the University of Arizona (InfoSci) applies in addition to the Dean of Students' Code of Academic Integrity.

Students in courses at the U of A InfoSci are expected to maintain rigor in their academic performance with intent to learn, practice, and overcome challenges toward personal growth and enrichment. As future professionals in digital environments, InfoSci students are also expected to exercise transparency and integrity in collaborations and in the use of tools and resources that may aid completion in assignments for our courses.

Consider the following PROHIBITED practices in this course, unless the instructor has specifically written instructions or permission to do otherwise:

- Posting a question on an online site such as Chegg.com, and copying and pasting some or all of the response into an assessment
- Posting an assessment from the course on online sharing sites such as Course Hero. Aiding other students in violation of academic integrity is also a violation, and is potential copyright infringement.
- Generating and submitting, in whole or in part, text or code through Artificial Intelligence such as ChatGPT, QuillBot, and text summarizers
- Using, in whole or in part, computer code not written by the student (for example, from another student, a book, or the internet) in an assignment or project. This includes using such code in modified or unmodified form.
- Searching for solutions to projects or assignments on the internet or through other tools, when your instructor intended for you to learn the solution through exercises (e.g. Googling for the solution to a question on an assignment).

- Simultaneously submitting the same assignment as another student enrolled into the course without prior permission from the instructor

Exceptions: Clear Instructions will be Provided

In any cases in which this course requires or permits students to use practices in the list above, clear written instructions will specify the tools allowed or required, so students can be certain they are working as instructed. See the U of A InfoSci Academic Integrity Policy, the U of A Code of Academic Integrity and Syllabus policy for more information.

LLMs and ChatGPT

Large language models (LLMs) like ChatGPT are a type of artificial intelligence (AI) engine that can look like it generates the code you need for Python homeworks and short answer questions. You are encouraged to use ChatGPT to debug code and experiment. However, abuse of ChatGPT can be traced (e.g., failing to give credit or cite ChatGPT when it is used) which could result in your suspension or termination from the course and even your program of study. Keep in mind, too, that while the code may appear legitimate, early studies have shown ChatGPT is not all that accurate with sophisticated coding. Exercise your scholarly discretion and maintain a sense of integrity in your statistical learning journey.

See my additional policies on this subject above.

Late work & extensions

The due dates for assignments are there to help you keep up with the course material and to ensure the teaching team can provide feedback within a timely manner. We understand that things come up periodically that could make it difficult to submit an assignment by the deadline. Note that the lowest homework assignment will be dropped to accommodate such circumstances.

- homeworks may be submitted up to 2 days late. There will be a 5% deduction for each 24-hour period the assignment is late.
- There is no late work accepted for application exercises, since these are designed to help you prepare for other assessments in the course.
- There is no late work accepted for assignments.
- The late work policy for the project will be provided with the project instructions.

Waiver for extenuating circumstances

If there are circumstances that prevent you from completing an application exercise or homework assignment by the stated due date, you may email me (gchism@arizona.edu) before the deadline to waive the late penalty. In your email, you only need to request the waiver; you do not need to provide explanation. This waiver may only be used for once in the semester, so only use it for a truly extenuating circumstance.

If there are circumstances that are having a longer-term impact on your academic performance, please let your academic dean know, as they can be a resource. Please let me know if you need help contacting your academic dean.

Regrade requests

Regrade requests must be submitted on GitHub within a week of when an assignment is returned. Regrade requests will be considered if there was an error in the grade calculation or if you feel a correct answer was mistakenly marked as incorrect. Requests to dispute the number of points deducted for an incorrect response will not be considered. Note that by submitting a regrade request, the entire question will be graded which could potentially result in losing points.

No grades will be changed after the project presentations.

"Incomplete" grade

The grade of "I" may be awarded only at the end of a term, when all but a minor portion of the course work has been satisfactorily completed. The grade of I is not to be awarded in place of a failing grade or when the student is expected to repeat the course; in such a case, a grade other than I must be assigned. Students should make arrangements with the instructor to receive an incomplete grade before the end of the term. If the incomplete is not removed by the instructor within one year the I grade will revert to a failing grade.

Tutoring

Tutoring can be found through the U of A Think Tank.

Accessibility

Accessibility and Accommodations: At the University of Arizona, we strive to make learning experiences as accessible as possible. If you anticipate or experience barriers based on disability or pregnancy, please contact the Disability Resource Center (520-621-3268, https://drc.arizona.edu) to establish reasonable accommodations.

Note: If you've read this far in the syllabus, email me a picture of your pet if you have one or your favorite meme!

Additional university policies

Additional policies can be found at this link (please read through them): https://catalog.arizona.edu/syllabus-policies

Safety on Campus and in the Classroom

For a list of emergency procedures for all types of incidents, please visit the website of the Critical Incident Response Team (CIRT): https://cirt.arizona.edu/case-emergency/overview

Also watch the video available at https://arizona.sabacloud.com/Saba/Web_spf/NA7P1PRD161/common/learningeventdetail/crtfy000000000003560

For more important dates, see the full UArizona Academic Calendar.

Graduate Student Resources

University of Arizona's Basic Needs Resources page for graduate students: http://basicneeds.arizona.edu/index.html