Spring 2022 INFO 523 – Data Mining and Discovery

Instructor: Dr. Cristian Roman-Palacios, Professor, Information Technology
Office: 445D, School of Information in the Harvill Building
Email: cromanpa94@email.arizona.edu
Office Hours: Office hours will be on zoom. Feel free to email me!

Course Description:

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 interesting 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 includes laboratory exercises, with data mining case studies using data from many different resources.

Course Prerequisites:
Students are expected to know the basics in computer programming (e.g., variables, arrays, loops, if-then conditions) and statistics (e.g., normal distribution, significance tests).

Course Objectives:

INFO 523 is an elective in the iSchool's MS program. As a multidisciplinary field, the course introduces concepts and work from many areas critical to information studies including statistics, machine learning, pattern recognition, database technology, and data visualization.

By the end of this course, students will:

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

The course addresses the MS Competencies: C1 [A, B, C, D], C2, and C3

Course Workload:
For each lecture hour, students are expected to spend 3-5 hours, completing the required reading and course work. Students finding themselves spending excessively more time should take advantage of office hours. **The instructor welcomes students' input on needed workload. Feel free to get in touch with the instructor.**

Course Materials:

Required textbook (most of these are freely available through UA library!):

- Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Torgo, L. (2011). *Data mining with R: learning with case studies*. Chapman and Hall/CRC.
- Kabacoff, R. I. (2015). *R in action: data analysis and graphics with R*. Simon and Schuster.
- Cichosz, P. (2014). *Data mining algorithms: explained using R*. John Wiley & Sons.

Technical Prerequisites:

All students taking this course will need to satisfy the standard School of Information technical requirements. having a computer connected to the Internet is essential.

Course Requirements:

The five components that go into the final course grade are described below. The percentage of the final grade is in parentheses next to each. All but the final exam should be completed in groups (see Group Work Policy below).

- Homework (60%): These are selected exercises from different sources. Please use R if writing code is required to solve the homework. If you feel more comfortable writing code in another language, please get in touch with the instructor before turning in the assignment. Homework submissions should be fully reproducible and should include the relevant script (e.g. *.R or *.Rmd) and data (e.g. *.csv, *.RData; if necessary). Please annotate your scripts as much as possible.
- Conference session (15%) and Final Project (15%). Please find more details on both assignments in D2L. In short, we will have a one-week conference session with students in the course presenting and attending to talks by their classmates. There will be a final project, which could be related to the conference session, in which students will use their data mining skills to answer a question.
- Class Participation (10%): The final percentage earned for this component will be calculated as 10% weighted by the total number of weeks the student either (i) participated in D2L discussions, (ii) attended office hours, (iii) helped improve materials, or even (vi) pointed out new resources to classmates.

The work and course requirements are subject to change at the discretion of the instructor with proper notice to the students.

Grading:

A=90+ (Superior Work)

B=80-89 (Very Good)

C=70-79 (Marginally Satisfactory)

D=60-69 (Failed to meet requirements)

E=50-59 (Failed to meet requirements)

F=0-49 (Failed to meet requirements)

Assignment Policy:

- Do not subject yourself to the late penalty: please do not make the instructor to assign a B for an A work just because you are late.
- All work must be turned in on the date due by midnight (11:59pm) Tucson time. Late work without a prior notice (at least 2 days before the due date) to and approval by the Instructor will receive a flat 20% deduction. Assignments late for 5 days will not be marked without an approved extension.
- In case of a D2L malfunction, email your assignment to the instructor.
- Be sure to check your submissions are successful.
- All work may be checked by Turnitin.com or other tools made available to the instructor. Students may find answers to homework questions on the Web. Yes, students are allowed to check out and learn from those answers, but to avoid a plagiarism charge, students must (1) cite the source URL and (2) present their work in their own words. **Please** do not impose the difficult and time-consuming task of reporting plagiarism to your instructor but know that the Instructor **will** report any such case if you give her the opportunity. Similarly, acknowledge help received from classmates or others. These acknowledgements will not hurt your grade, instead they reveal the academic integrity in you as a young scholar/researcher.

COURSE, SCHOOL, AND UNIVERSITY POLICIES:

Inclusive Excellence

Inclusive Excellence is a fundamental part of the University of Arizona's strategic plan and culture. As part of this initiative, the institution embraces and practices diversity and inclusiveness. These values are expected, respected, and welcomed in this course.

This course supports elective gender pronoun use and self-identification; rosters indicating such choices will be updated throughout the semester, upon student request. As the course includes group work and in-class discussion, it is vitally important for us to create an educational environment of inclusion and mutual respect.

Academic Code of Integrity

Students are expected to abide by The University of Arizona Code of Academic Integrity. 'The guiding principle of academic integrity is that a student's submitted work must be the student's own.' If you have any questions regarding what acceptable practice under this Code is, please ask an Instructor.

Policies Against Disruptive and Threatening Behavior

Students are also bound by the University's policies related to disruptive behavior and threatening behavior.

Accessibility and Accommodations

At the University of Arizona we strive to make learning experiences as accessible as possible. If you anticipate or experience physical or academic barriers based on disability or pregnancy, you are welcome to let me know so that we can discuss options. You are also encouraged to contact Disability Resources (520-621-3268) to explore reasonable accommodation.

"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.

Additional Policies:

The Arizona Board of Regents' Student Code of Conduct, ABOR Policy 5-308, prohibits threats of physical harm to any member of the University community, including to one's self. See: http://policy.web.arizona.edu/threatening-behavior-students

All student records will be managed and held confidentially. See: http://www.registrar.arizona.edu/ferpa/default.htm

Information contained in this course syllabus, may be subject to change, as deemed appropriate by the instructor.

Course Schedule:

Week	Date	Main topics	Homework	Final project – virtual conference	Additional readings	Need to turn something in?
0.5	11-Oct	Lecture 1. INFO523, Data Mining Lecture 2. Introduction to R: basic math, data types, loops, if-else statements, working with data structures.	**Release:** Homework 1: *Introduce yourself and install R*		Han & Kamber, Ch 1	Yes! (Homework 1)
1	18-Oct	Lecture 3. Reproducibility in R: GitHub, RStudio. Lecture 4. Intro to R packages	**Release**: Homework 2: *Intro to R*	Identify a partner! Feel free to start discussion about potential ideas for your project.	Kabacoff, Ch 1–4	Yes! (Homework 2)
2	25-Oct	Lecture 5. Data pre-processing: base R. Lecture 6. Data pre-processing: tidyverse.	**Release**: *Homework 3: Data processing in R*	Select a topic and start working on the presentation!	Kabacoff, Ch 5; Han and Kamber. Ch. 3	Yes! (Homework 3)
3	1-Nov	Exploratory data analysis in R: Descriptive stats, variance, sd, se, covariation, correlation, skewness, and kurtosis. Plotting in base and ggplot: density plots, boxplots, dot plot, bar charts, pie charts, statistical modeling.	**Release:** *Homework 4: Basic plots*	Submit your availability for week 7	Kabacoff, Ch 7, 11, 16; Cichosz, Ch. 2	Yes! (Homework 4 and availability for the final project)
4	8-Nov	Unsupervised techniques: clustering (e.g. k-means, k-medians, PCA).	**Release**: H*omework 5: Unsupervised learning*	Checkpoint: Abstract	Han and Kamber. Ch. 6;	Yes! (Homework 5; Checkpoint final project)

		Unsupervised techniques: Association rules.			Cichosz, Part IV	
5	15-Nov	Supervised techniques, regression I: linear and logistic models. Supervised techniques, regression I: Logistic regression, SVMs, decision trees, random forest	**Release**: *Homework 6: improving data mining code*		Cichosz, Part III; Torgo, 2011 (multiple chapters)	**No!**
6	22-Nov	Improving data mining code		Turn in conference abstract and code.	Torgo, 2011 (multiple chapters)	Yes! (**Final project**! Part 1; **Homework 6: Part A**)
7	29-Nov	Mini conference		Presentations!		**Final project**: Turn in recordings and feedback
7.5	6-Dec	Final review and discussion				**Homework 6**: Parts B and C!