DATA MINING, MANAGEMENT, AND CURATION

DATA 3421
Unique Number: 92378 (92379)
Fall 2021

Instructor:	Amir Shahmoradi
office:	SEIR 365
e-mail:	a.shahmoradi@uta.edu
e-mail:	shahmoradi@utexas.edu
e-mail:	a.shahmoradi@gmail.com
Office/Lab/Help hours:	Fr 1:00PM - 2:50PM - PKH 313 Online by appointment

Class start/end:	Aug 25, 2021 – Dec 6, 2021
Lecture meeting times:	MW 2:30PM – 3:50PM
Lecture meeting place:	PKH 302 (and if needed, Teams Virtual Room)

Teaching Assistants:	NONE
office:	NONE
e-mail:	NONE
office hours:	NONE

COVID-19 NOTES:

All classes will be held in person (or other modalities if all students prefer so). However, please avoid attending the lectures in person if you do not feel well for any reason. The instructor will ensure students have access to all materials, homework, and quizzes online if they cannot attend the lectures in person due to sickness or other reasons.

Please take your health and others in this class seriously by paying attention to covid19 protocols. Failure to do so could cost your friends, classmates, or your instructor's lives. Wearing a mask and vaccinating is not mandated but strongly encouraged. Your instructor is fully vaccinated and will wear (K)N95 and/or double mask throughout in-person sessions.

COURSE OBJECTIVES / ACADEMIC LEARNING GOALS

This lecture and lab course will provide training in working with databases, including data mining techniques and principles and best practices in data management, storage, and curation. **Prerequisite**: DATA 1401, DATA 1402, **or** with the permission of the instructor.

The primary objective of this course is to study a variety of techniques for data mining, predictive modeling, and machine learning. Upon completing this course successfully, you will be able to understand the pros and cons of different data mining techniques, so that you can (i) make an informed decision on what approaches to consider when faced with real-life problems requiring predictive modeling, (ii) apply models properly on real datasets so to make valid conclusions.

COURSE SCHEDULE

The following is a tentative outline of topics to be covered:

- **What is Data Mining?**
    - Kinds of data
    - Patterns in Data
    - Available technologies for Data mining
    - Data Mining applications
    - Open issues in Data Mining
- **Data Preprocessing**
    - Data attributes and attribute types
        - Nominal Attributes
        - Binary Attributes
        - Ordinal Attributes
        - Numeric Attributes
    - Data summarization.
        - Central tendency
        - Data spread
    - Data visualization
    - Data Similarity and Dissimilarity
    - Data Cleaning
        - Missing Values
        - Noise
    - Data Integration
        - When to combine multiple datasets?
        - Data redundancy
    - Data Reduction
        - Principal Components Analysis
        - Wavelet Transforms
        - Histograms
        - Clustering
    - Data Transformation
        - Data normalization
- **DataWarehousing**
    - Data Cube Computation
    - Data Cube Computation Methods
    - Mining Frequent Patterns, Associations
- **Pattern Mining**
    - Pattern Mining
    - Multidimensional Pattern Mining
- **Continuous Regression**
    - Linear Regression
    - Nonlinear regression
    - Non-paramtric regression
- **Classification**
    - Concepts
    - Decision Tree Induction
    - Bayes Classification Methods

  - Rule-Based Classification
  - Cluster Analysis
  - Partitioning Methods
  - Hierarchical Methods
  - Density-Based Methods
  - Clustering High-Dimensional Data
- **Outlier Detection**
  - What is an outlier?
  - Outlier Detection Methods
  - Statistical Approaches
  - Proximity-Based Approaches
  - Outlier Detection in High-Dimensional Data
    - How do outlier detection methods fail in high dimensions?
- **Data Mining Trends and Research Frontiers**
  - Mining Complex Data Types
  - Data Mining Applications
- **Data Mining tools**
  - SQL

COURSE TEXTBOOKS

No textbook is required for this course. Online class lecture notes will be used as reference. However, a list of textbooks for those who are interested to self-educate themselves or go beyond class syllabus is provided below,

- Principles of Data Mining, Bramer
- Machine Learning: A Probabilistic Perspective, Kevin Murphy
- Pattern Recognition and Machine Learning, Bishop
- The Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani, and Jerome Friedman (HTF)
- Data Mining Concepts and Techniques, Han, 2012
- Data Mining, Ian Witten, 2011

COURSE LOGISTICS

Grading:
Weekly Homework: 33% (Assignments might not be weighted equally)
Weekly Quizzes: 33%
Final Project: 34%

Homework Policy:
There will be approximately one homework per week. Assignments will be due every Tuesday before the lecture begins and should be added to an online repository determined by the instructor. No late assignments will be accepted. No exceptions to the homework policy will be made without prior instructor approval.

Examinations:
There will be no midterm or final exams. Students will have to complete a project in place of the final exam, (possibly, in collaboration with their teammates who are determined randomly after the midterm).

Quizzes:

There will be weekly quizzes at the beginning of each lab session on Thursdays.

Attendance:
Regular attendance is expected. Any absence requires prior approval from the instructor, or compelling evidence of illness or an official letter from the university administration. Student attendance will be randomly checked.

Scholastic dishonesty: All students are responsible for upholding the University rules on scholastic dishonesty. Students who violate University rules on scholastic dishonesty are subject to disciplinary penalties, including the possibility of failure in the course and/or dismissal from the University. Since such dishonesty harms the individual, all students, and the integrity of the University, policies on scholastic dishonesty will be strictly enforced.

Other matters: The University of Texas at Arlington provides, upon request, appropriate academic adjustments for qualified students with disabilities. Any student with a documented disability (physical or cognitive) who requires academic accommodations should contact the UTA's Office for Students with Disabilities as soon as possible to request an official letter outlining authorized accommodations. For visit https://www.uta.edu/disability/.