Department of Computer and Information Sciences
COSC 757 –Data Mining
Fall 2024

Instructor:	Dr. Weixian Liao	**Email:**	wliao@towson.edu
Office:	7800 York Road, Room 470	**Phone:**	410-704-2774
Office Hours:	Wednesday 3:00 PM - 4:00 PM and by appointment		

Teaching Assistant
Name: Papa Pene
Email: ppene1@students.towson.edu
Office: YR224
Office Hour: Tuesday 3:00 PM – 4:00 PM and by appointment

Credits:	3 credits
Course Time:	Wednesday 4:30 PM - 7:10 PM
Course Location:	**YR0304**
Course Website:	Blackboard

Prerequisite:	COSC 578 or special permission from the instructor

Course Description: This course provides students with an understanding of the field of data mining and knowledge discovery in data (KDD). Students will become familiar with the foundations of data mining from a number of perspectives and will explore cutting edge research in data mining published in academic journals and conferences. Students will also gain hands on experience with data mining tools.

Required Textbook:
Discovering Knowledge in Data: An Introduction to Data Mining
Daniel Larose and Chantal Larose
Wiley Publishers, 2014,
is available electronically through Cook Library at https://proxy-tu.researchport.umd.edu/login?ins=tu&url=https://search.ebscohost.com/login.aspx?direct=true&site=eds-live&scope=site&db=cat01451a&AN=towson.006171944.

Recommended Readings:
Data Mining: Concepts and Techniques (Third Edition)
Jiawei Han, Micheline Kamber, and Jian Pei
Morgan Kaufmann Publishers,2012

Data Mining Practical Machine Learning Tools and Techniques.
Ian H. Witten, Eibe Frank, Mark A. Hall.
Morgan Kaufmann Publishers, 2011.

Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking
Foster Provost and Tom Fawcett
O'Reilly Publishers, 2015

Course Context/ Background: Over the last decade, technological advancements have enabled the collection of a vast amount of data. The discovery of new knowledge in this data can enable a better understanding of business, society, the environment, health care, and all aspects of life. This course provides students with a background in the knowledge discovery process and the use of data mining and online analytical processing to automatically discover new patterns in large datasets.

Course Objectives: Upon completion of the course, students will understand and be familiar with the general concepts associated with data mining and visual analytics. Specifically, students will be able to:

- Describe the key components of the knowledge discovery process
- Understand the key theoretical underpinnings of data mining, machine learning, and data analytics
- Understand the feasibility, usefulness, effectiveness, and scalability of various approaches to data mining and visual analytics
- Perform exploratory analysis on datasets through visualization and descriptive statistics
- Pre-process and clean data for various data mining approaches
- Formulate data mining problems and choose the proper algorithms and evaluation methods for knowledge discovery.
- Exhibit their knowledge through a real-world data mining project
- Gain experience in using data mining software tools such as Weka, RapidMiner, and R.

Course Topics: This course provides students with a background in the concepts and applications of data mining and visual analytics. Potential course topics could include the following:

- Introduction to data mining
- Applications of data mining
- Exploratory data analysis
- Data preprocessing
- Frequent itemset mining
- Pattern evaluation methods and visualization of frequent patterns
- Advanced pattern mining techniques and applications
- Classification methods such as decision trees, Bayes methods, rule-based methods,
- Classification model evaluation, accuracy assessment, and visualization of classification results
- Advanced classification methods and applications

- Cluster analysis including partitioning-based methods, hierarchical methods, density-based methods, grid-based methods.
- Cluster evaluation techniques
- Applications of clustering
- Outlier detection
- Spatial and temporal data mining and visualization techniques
- Cutting edge data mining trends and research directions

Course Format: Active learning techniques, lectures, discussion sessions (in class AND online), presentations, lab exercises and projects may be used. Students are expected to read the textbook and find and use current content on the course subjects using the library, internet and provided resources. Some assignments will necessitate collaborative learning while others will require individual research and presentation.

Attendance Policy: Students are expected to attend all classes in order to remain current in the coursework. It is the student's responsibility to remain current on the handouts, assignments and notes if class is missed. The instructor will allow only students with documented excuses (see below) to make up missed work or assignments *when it is feasible*. If the student is absent from an exam during the scheduled time for that exam, the student will automatically receive a grade of zero (0) for the exam unless: (a) the student notifies the instructor of the absence at least one day prior to the exam; (b) the student is ill and supplies a written doctor's excuse explaining the absence; or (c) there is an extraordinary situation which the instructor allows as an acceptable excuse. Only under one of these circumstances will arrangements for a makeup exam be made.

Grading Policy: Students will be evaluated on the following basis:

4 Assignments:	40%
1 Paper Summary:	25%
1 Final Project:	32%
Attendance and Participation:	3%

Final course grades will abide by the following scale:

A	93 – 100	B	75 – 82.9
A-	88 – 92.9	C	60 – 74.9
B+	83 – 87.9	F	Below 60

All assigned work (i.e., assignments and projects) is expected to be completed and submitted by the stated deadline. *No late work will be accepted* and a grade of zero (0) will be given. Students are encouraged to discuss homework and projects as a means to share knowledge, experience and lessons learned as part of the learning process, but academic honesty should be strictly observed (see below).

Labs and Assignments: Data mining tasks will be assigned throughout the semester to give students hands on experience using data mining software such as Weka, RapidMiner, or R. We will also complete a number of labs during the semester using the R software to experiment with a number of data mining techniques. There will be four data mining assignments that will require students to implement data mining techniques on real datasets. All assignments are due at the assigned time on Blackboard.

All work should be thoroughly backed up before turning it in and all submissions should be *well documented giving credit to respective sources with proper citations*. Homework may also be assigned in the lecture at the instructor's discretion and as the need arises.

Article Summary and Presentation: Each student will be required to select a research article related to a course topic. The student will write a summary of the article to be submitted to the professor and prepare a 30-minute presentation on the article. The presentation should give details about the approach presented in the article and summarize the results. The article summary and presentation will be counted toward the assignments portion of the grade. During the first class, students will select a topic and sign up for a presentation slot.

Project: The semester project will require students to apply a data mining technique covered during the semester to discover knowledge in a real-world dataset. Project examples include classification of Baltimore City crime data, clustering of stock market data, pattern mining applied to social network data, and many more. Students will be required to submit a project proposal to the instructor before carrying out the execution of the project. Projects should generally follow the steps of the knowledge discovery process and should give details on the cleaning, integration, selection, transformation, and mining of the data. This will include selecting an appropriate algorithm based on the knowledge discovery problem at hand. Projects should also report on the evaluation of the approach and use visualization to present the discovered patterns. Students are encouraged to use available data mining tools covered in class such as R, Weka, and RapidMiner.

Professionalism: All materials submitted for this course should look professional including the use of correct grammar and spelling. Please ensure all cell phones and other devices that could potentially disrupt the class are turned off upon entering the classroom.

Posting of Grades: University policy prohibits posting of grades in any form. The instructor will not report grades via email or in response to phone calls. Grades for the semester can be accessed online.

Cheating and Plagiarism: Academic honesty is strongly observed. This course may consist of both individual and team assignments. A team project is an assignment in which collaboration is allowed and highly encouraged. However, the work of the team *must* be of the team's creation and not plagiarized from other sources. Individual assignments *must* reflect the work of the individual student and of his/her creation. While studying together, discussion and collaboration is encouraged, individual assignments *must be individually prepared* – copying or sharing files, diagrams and/or code is considered cheating. The penalty for cheating will, at a minimum, consist of a grade of zero for the dishonest work and may lead to the possibility of **course failure** depending on the severity. Students are responsible for reading and knowing Towson University's policy regarding academic dishonesty, located in Appendix F in the Undergraduate Catalog and familiarizing themselves with the policies detailed at http://cooklibrary.towson.edu/avoidingPlagiarism.cfm.

Classroom and Lab Policy: Food and drink are not allowed in the labs and classrooms with the exception of water in the classroom only. All cell phones should be turned off or put on silent mode to avoid disruptions and distractions.

Repeat Policy: Students may not repeat a course more than once without prior permission of the Academic Standards Committee.

If you have a learning disability and/or need accommodation for any reasons, please advise the instructor as early as possible in the course.

Tentative Schedule: The following is a tentative schedule. Note that these topics and chapters are subject to change based on time and discretion of the instructor.

	Lecture	**Reading/Lab/Article Discussion**
Week 1	Course Overview/Introduction to Data Mining	Chapter 1
Week 2	Data Preprocessing	Chapter 2 Lab: Data Preprocessing
Week 3	Exploratory Data Analysis	Chapter 3 Lab: EDA and Statistical Analysis
Week 4	Statistical Analysis	Chapter 4 and 5 Article Presentations
Week 5	Supervised Learning and Classification	Chapter 6 Article Presentations
Week 6	Classification: Supervised Learning and K-Nearest Neighbor and Decision Trees	Chapter 7 and 8 Lab: Introduction to Classification
Week 7	Classification: Naïve Bayes and Neural Networks	Chapter 9 Lab: Classification 2
Week 8	Classification: SVM, Bayesian Networks and Ensemble Techniques	Assigned Readings Classification Contest
Week 9	Introduction to Cluster Analysis: Partitioning-Based Methods	Chapter 10 Article Presentations
Week 10	Cluster Analysis: Density-Based Methods and Kohonen Networks	Chapter 11 Lab: Clustering
Week 11	Association Rules: Apriori Algorithm	Chapter 12 Lab: Association Rules
Week 12	Mining Sequence Data	Supplemental Readings Article Presentations
Week 13	Mining Networks and Graphs	Supplemental Readings Article Presentations
Week 14	Deep Learning	Supplemental Readings Lab: Deep Learning
Week 15	Project Presentations (Potential Class)	