DATA 3421:

DATA MINING, MANAGEMENT, AND CURATION

Fall 2024

Instructor: Max (Masoud) Rostami

Office: Life science building, Room 231

Email: masoud.rostami@uta.edu

Class start/end: August 19 2024 – December 3, 2024

Lecture meeting times:MW 4:00PM – 5:20PM

Lab times: Fr 9:00AM-10:50AM; 11:00 AM-12:50 PM

Lecture meeting place: PKH 102

Lab meeting place: Fr 9:00AM-10:50AM **(SH 105)**; 11:00 AM-12:50 PM **(SH 129)**

Office hours: MW 11:00AM – 12:00PM

Teaching Assistants:	Shreya Gupta; Vishnu Sai Muppalla
office:	Virtual via Teams
e-mail:	Shreya Gupt (shreyagupta3004@gmail.com) Vishnu Sai Muppalla (vxm8542@mavs.uta.edu)
office hours:	Shreya Gupt (M:11:00AM–12:00PM, Fr: 1:30 pm-2:30pm) Vishnu Sai Muppalla (TuTh:11:00AM–12:00PM)

COURSE OBJECTIVES / ACADEMIC LEARNING GOALS

This lecture and lab course will provide training in working with databases, including data mining techniques, principles, and best practices in data management, storage, and curation. Prerequisite: DATA 1401, DATA 1402, DATA 3401, or with the permission of the instructor. The main objective of this course is to explore a variety of techniques for data preprocessing, data mining, predictive modeling, and machine learning. Successful completion of the course will enable you to evaluate the strengths and weaknesses of different data mining techniques, allowing you to make informed decisions when faced with real-world problems requiring predictive modeling. You will also gain proficiency in applying models to real datasets, allowing you to draw valid conclusions. Additionally, this course will cover the use of Python programming languages and Excel for data mining, predictive modeling, and machine learning, as well as SQL for various purposes such as data retrieval, data cleaning and transformation, data exploration, data integration, and data modeling.

MICRO-CREDENTIAL IMPLEMENTATION:

In our DATA 3421 course, we are enhancing your learning experience by integrating the "**IBM Data Science Professional Certificate**" a micro-credential recognized across the industry. This incorporation not only aligns the course material with practical, industry-relevant skills but also ensures that assessments reflect real-world standards. By completing specified modules of this micro-credential as part of your coursework, you will not only earn academic credit but also make significant progress towards obtaining this highly valued certification. This approach is designed to equip you with both theoretical knowledge and practical skills, thereby boosting your employability in the data science field.

Coursera Learner Support:

https://www.coursera.support/s/learner-help-center?language=en_US

TEACHING APPROACH:

- **Lectures on Mondays and Wednesdays:** Focus on theoretical concepts and principles of the week's topic.
- **Post-Wednesday Coding Templates:** Distribution of a practical coding template relevant to the week's lectures.
- **Friday Lab Sessions:** Group work in the class, applying lecture concepts using the provided coding template and your knowledge.

COURSE SCHEDULE

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

Week 1: Introduction to Data Mining

- Introduction to the course and data mining
- Overview of why data mining is essential
- Types of data in data mining
- Data mining Challenges
- Data mining pipelines
- Data collection methods
- Data Understanding

Week 2: Data Understanding & Data Preprocessing-I

- **Quiz 1**
- Data Preprocessing
- Data Quality
- Data Cleaning
- Outlier Detection

- Data Transformation
- **Lab-1: Basic Data Preprocessing**
- **Homework 1:** Microcredential Implementation: Complete "IBM Data Science - **Data Science Methodology**" (Modules 1 and 2)

Week 3: Data Preprocessing-II

- **Quiz 2**
- Performing Variable Discretization
- Handling Skew Data
- Encoding Categorical Variables
- Feature Engineering
- Data Reduction Techniques and Feature Extraction
- **Lab-2: DataPreprocessing-II**
- **Homework 2:** Microcredential Implementation: Complete "IBM Data Science - **Data Science Methodology**" (Modules 3 and 4)

Week 4: Data Preprocessing-III, Exploratory Data Analysis (EDA) and Image Data Preprocessing

- **Quiz 3**
- Multicollinearity
- Data Balancing
- Feature Selection
- Basics of EDA and its importance
- Specialized preprocessing techniques for image data
- Image augmentation, resizing, cropping
- Normalization and filtering in image data
- Train-Test Split
- **Lab-3: ImagePreprocessing**
- **Homework 3:** Microcredential Implementation: Complete "IBM Data Science - **Data Analysis with Python**" (Modules 1 and 2)

Week 5: Hypothesis testing

- **Quiz 4**
- Margin of Error
- Z-test
- T-test
- Test for Homogeneity of proportions

- Goodness of fit
- ANOVA
- Verifying model assumptions
- Hypothesis testing Python Template
- **Lab-4: Hypothesis Testing Using Python**
- **Homework 4:** Hypothesis testing

Week 6: Advanced Data analysis using Excel and Machine Learning

- **Quiz 5**
- Advanced Excel Techniques for Data Analysis
- Machine learning Introduction
- Supervised vs unsupervised methods
- **Lab-5: Data Manipulation using Excel**
- **Homework 5:** Data Science using Excel
- **Introducing the Group Final Project**

Week 7: Regression Algorithms

- **Quiz 6**
- Overview of Regression Methods
- Linear, non-linear, and multiple regression techniques
- Polynomial Regression
- Variable Interactions
- Evaluation metrics for regression models
- Hands-on: Build a Simple Regression
- **Lab-6: Linear regression**
- **Homework 6:** Microcredential Implementation: Complete "IBM Data Science - **Data Analysis with Python**" (Modules 3 and 4)

Week 8: Classification Algorithms

- **Quiz 7**
- Understanding Classification Algorithms
- K-Nearest Neighbor
- Decision Tree
- Hands-on: Build a Classification
- **Lab-7: Regression Techniques comparisons**

- **Homework 7:** Microcredential Implementation: Complete "IBM Data Science - **Machine Learning with Python**" (Modules 1,2, 3)

Week 9: Continue: Classification Algorithms

- **Quiz 8**
- Ensemble Learning
- Metrics for evaluating classification models
- Hands-on: Build a Classification
- **Lab-8: Classification Techniques**
- **Homework 8:** Microcredential Implementation: Complete "IBM Data Science - **Machine Learning with Python**" (Module 4)

Week 10: Clustering Techniques

- **Quiz 9**
- Deepening the understanding of classification algorithms
- Introduction to Clustering Methods
- K-means
- Hierarchical Clustering
- Metrics for Clustering
- Hands-on: Build a Clustering
- **Lab-9: Clustering**
- **Homework 9:** Microcredential Implementation: Complete "IBM Data Science - **Machine Learning with Python**" (Module 5)

Week 11: Model Evaluation & Performance

- **Quiz 10**
- Evaluation metrics
- Model Selection
- Regularization
- Cross-validation
- Bootstrap methods
- Bias-Variance tradeoff
- Error Analysis
- Improve the Models
- Hands-on: Tuning
- **Lab-10: Improving the Model Performance**

- **Homework 10:** Microcredential Implementation: Complete "IBM Data Science - **Machine Learning with Python**" (Module 6)

Week 12: Basic SQL-I

- **Quiz 11**
- Introduction to SQL and Databases
- Basic SQL Syntax
- SQL queries for data manipulation
- **Lab-11: SQL-I**
- **Homework 11: Basic SQL Challenges**

Week 13: Advanced SQL-II

- **Quiz 12**
- SQL joins
- SQL Aggregations
- **Lab-12: SQL-II**
- **Homework 12: Intermediate-level SQL Challenges**

Week 14: Advanced SQL-III

- **Quiz 13**
- Subqueries
- SQL for data cleaning and preprocessing
- Set Operations
- Advanced SQL queries for data analysis
- Using SQL in practical data mining scenarios
- **Lab-13: SQL Challenge- Group Assignment**
- **Homework 13: Advance-level SQL Challenges**

Week 15: Integrating SQL with Python & Project Presentation

- Introduction to Python SQL Libraries
- Introduction to Python SQL Libraries
- Executing SQL Queries Through Python
- Final presentations of the project (Students will present their subject of interest as a group. All the presentations will be in-person.)

Week 16: Course Conclusion and Project Presentation

- Final presentations of the project. If there is a need for extra time, the presentations will take place during the final exam.

COURSE OUTCOMES:

By the end of this course, students will have gained comprehensive skills in data mining, predictive modeling, and machine learning, with a strong emphasis on practical applications using Python, Excel, and SQL. They will be proficient in data preprocessing techniques, including cleaning, transformation, and feature engineering, as well as advanced data analysis and visualization methods. Through hands-on labs and projects, students will learn to evaluate and apply various machine learning algorithms for regression, classification, and clustering. They will also develop the ability to perform hypothesis testing and model evaluation, ensuring robust and reliable data-driven decision-making. Additionally, students will become adept at using SQL for data retrieval, manipulation, and integration, enabling them to handle real-world data challenges effectively. The course will prepare students to critically assess different data mining techniques, apply them to diverse datasets, and draw valid conclusions to address complex predictive modeling problems.

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
- Discovering Knowledge in Data; An Introduction to Data Mining, Daniel T. Larose, 2014
- Introduction to Data Mining, Pang-Ning Tan, 2019

COURSE LOGISTICS

Grading:

Weekly Homework-Individual: 25% (Assignments might not be weighted equally)

Lab-Group: 30%

Weekly Quizzes: 20%

Final Project-Code: 5%

Presentation:10%

Attendance: 10%

HOMEWORK POLICY:

Homework assignments are **Individual assignments.** There will be approximately one homework assignment per week, due before the **lecture begins on Friday of the following week.** Some of these assignments should be added to an online repository specified by the TAs.

Lab Assignment Policy:

Lab assignments are **Group assignments** with a submission deadline set to **3 hours** from the start of each lab session every week. Students will have this duration to answer the lab questions.

Late Submission Policy:

For both homework and lab assignments, late submissions will incur penalties as follows:

- If the lab assignment is submitted **less than 24 hours** late, a deduction of **two points** will be applied.
- Lab submissions made more than **24 hours but less than one week** late will receive a deduction of **three points.**
- Any submissions beyond one week late will result in a **zero** for the assignment.

EXAMINATIONS:

There will be no midterm or final exams. Students will have to complete a project in place of the final exam,

QUIZZES:

There will be weekly quizzes on each Monday.

ATTENDANCE:

Students are expected to attend **more than 90% of all scheduled classes**. Consistent attendance is essential for maintaining the integrity of the learning experience and ensuring adequate participation in discussions, group work, and presentations. If a student anticipates being unable to attend a session, they are required to notify the appropriate person prior to the class. For lab sessions on Fridays, they must contact the **Teaching Assistant (TA)**. For lecture sessions on Mondays and Wednesdays, they must contact the **Professor**. Late notifications about absences will not be accepted unless extraordinary circumstances can be clearly demonstrated. Please note that attendance is not only for the purpose of receiving credit but also for the purpose of gaining knowledge and engaging in the learning process. Being present in class ensures that students have access to important information, discussions, and activities that contribute to their overall understanding of the course material.

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.

Face Covering Policy:

While the use of face coverings on campus is no longer mandatory, all students and instructional staff are strongly encouraged to wear face coverings while they are on campus. This is particularly true inside buildings and within classrooms and labs where social distancing is not possible due to limited space. If a student needs accommodations to ensure social distancing in the classroom due to being at high risk, they are encouraged to work directly with the Student Access and Resource Center to assist in these accommodations. If students need masks, they may obtain them at the Central Library, the E.H. Hereford University Center's front desk or in their department.

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/.