COSC 3337 – Section 12627

Data Science I

Course Information

Term and Year: **Summer 2023**
Location: **Face to Face: 8:00AM - 10:00AM PGH 232**
Meeting Days/Times: **Everyday**

Contact: By email, njrizk@uh.edu

Office Hours: 12-1 PM TTH.

Course System: Canvas.

Main References: While lecture notes will serve as the main source of material for the course, the following book constitutes a great reference:
Open Textbooks

Rizk, Nouhad: Building Skills for Data Science
https://uhlibraries.pressbooks.pub/buildingskillsfordatascience/

Books

1. https://ebookcentral.proquest.com/lib/uh/detail.action?docID=1895687&query=data+mining
2. https://ebookcentral.proquest.com/lib/uh/detail.action?docID=4851656

Statistics:

3. https://cnx.org/contents/tWu56V64@33.122:-mZCQZc7@5/Introduction

Reference:
P.-N. Tang, M. Steinback, and V. Kumar Introduction to
Data Mining, Addison Wesley, 2018.
(Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk from the Frontline. O' Reilly. 2014

Description: Data science process, data preprocessing, exploratory data analysis, data visualization, basic statistics, basic machine learning concepts, classification and prediction, similarity assessment, clustering, post-processing and interpreting data analysis results, use of data analysis tools and programming languages and data analysis case studies.

Objectives: By the end of the course a successful student should:

- Students will develop relevant programming abilities.
- Students will demonstrate proficiency with statistical analysis of data.
- Students will develop the ability to build and assess data-based models.
- Students will execute statistical analyses with Python software.

- Students will apply data science concepts and methods to solve problems in real-world contexts and will communicate these solutions effectively

Prerequisites: MATH 3339 and COSC 2436.

Software: Make sure to download Anaconda https://repo.anaconda.com/. Let me know via email in case you encounter difficulties.

Academic Honesty: University of Houston students are expected to adhere tothe Academic Honesty Policy as described in the UH Undergraduate Catalog. “Academic dishonesty” means employing a method or technique or engaging in conduct in an academic endeavor that contravenes the standards of ethical integrity expected at the University of Houston or by a course instructor to fulfill any and all academic requirements. Academic dishonesty includes, but is not limited to, the following: Plagiarism; Cheating and Unauthorized Group Work; Fabrication, Falsification, and Misrepresentation; Stealing and Abuse of Academic Materials; Complicity in Academic Dishonesty; Academic Misconduct.

Refer to UH Academic Honesty website (http://www.uh.edu/provost/policies/honesty/) and the UH Student Catalog for the definition of these terms and university’s policy on Academic Dishonesty. Anyone caught cheating will be reported to the department for further disciplinary actions, receive sanctions as explained on these documents, and will have an academic dishonesty record at the Provosts office. The sanctions for confirmed violations of this policy shall be commensurate with the nature of the offense and with the record of the student regarding any previous infractions. Sanctions may include, but are not limited to a lowered grade, failure on the examination or assignment in question, failure in the course, probation, suspension, or expulsion from the University of Houston, or a combination of these. Students may not receive a W for courses in which they have been found in violation of the Academic Honesty Policy. If a W is received prior to a finding of policy violation, the student will become liable for the Academic Honesty penalty, including F grades.

	Date	**Topics**	Open Textbook Reading
Week 1		**Introduction to Data science**	
		Data science Overview	
		Machine Learning **Data Cleaning**	

		Data Processing Startup Example	B1: p 30-35
		Statistical Learning	
		Data Exploration **Data Similarities & Distances**	B1: p 54-81
		Linear Regression	B1:p 171-213
		Linear Regression (Python Example)	
	Thursday June 8th	**DROP DEADLINE**	
	Friday June 9th	**EXAM 1**	**15%**
Week 2		**Logistic Regression Dimensionality reduction - PCA**	B1: p 359-399
		Introduction to Classification KNN	B1: p 301-312 B2: p 32-48
		Decision Tree	
		Random Forests **KNN**	B1: p 317-322 B2: P 49-68
		Naive Bayes	B1: p 414-439 B2: p 113-140

	Friday June 16th	**EXAM 2**	**15%**
Week 3		**Model Evaluations Metrics**	
		Ridge - Lasso	
		Lines/SVM	

		Dimensionality reduction (feature extraction) **Wrap Up classification**	
		K-Means	B1: p 523- 537 B2: 218-250
		Hierarchical Clustering Heatmap	

		Storytelling	
	Friday June 23th	**EXAM 3**	**15%**
		Monday June 27 DROP **Deadline for withdrawal with W** **DBSCAN**	
		Cluster Validity Silhouette	
		Neural networks	
		A priori and Association rules	B1: p 603- 617 B2: p 69-87
		Dynamic Hashing -Merkle tree (Optional)	
	Friday June 30th	**EXAM 4**	**15%**
	Storytelling Project Submission: Monday July 3rd ,2023	**Deadline 11:59 PM**	

Grading Policy

The final numeric grade is computed based on student's performance in weekly assignments and exams/quizzes. The final numeric grade for the course will be determined as follows:

✓	Attendance and participation	5%
✓	Homework assignments + Final Project (NO drop of any HW)	20%
✓	Lab work	15%
✓	Exams	60%

Labs: Coding practices (using Python format. ipynb **only**). **One lab assignment will be dropped** (the one with the lowest grade).

Exams: Held during class times on Friday.

Homework: Students will be submitted by uploading their work in Blackboard as .ipynb.

Final Group Project on Storytelling (as final Homework):

- You will form a group of 3-4 members.
- A group assignment, consisting of students teaming up (5 points), deciding on the data set of interest (5 points), posing research questions (10 points), applying ML techniques to address those questions (50 points), and using art Graphics (10 points) . Each group will eventually submit a report as video presentation of research findings and member contributions (20 points).

Grading Scheme:

A>=92.5 Excellent	**A->= 89.5 and < 92.5 Outstanding**	**B+>=86.5 and < 89.5 Very Good**
B > = 83.5 and <86.5 Good	**B->=79.5 and < 83.5 Above Average**	**C+>=76.5 and < 79.5 High Average**
C>=72.5 and <76.5 Average	**C->=69.5 and <72.5 Low Average**	**D+>=65.5 and <69.5 Below Average**
D >=62.5 and <65.5 Poor	**F < 62.5 Failing**	