UNIVERSITY of HOUSTON

COSC 3337

Data Science I

Course Information

Term and Year: **Fall 2020**
Location: **Online**
Meeting Days/Times: Tuesdays -Thursday 8:30-10:00 AM

Contact: By email, MS_Teams during office hours (or by appointment).

Office Hours: 1:00 PM- 1:30 PM TTH.

Course Online System: Blackboard.

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 2430.

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 to the 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.

Technology statement below as requested by the Provost's Office:

Computer and internet access required for course. For the current list of minimum technology requirements and resources, copy/paste/navigate to the URL http://www.uh.edu/online/tech/requirements. For additional information, contact the office of Online & Special Programs at UHOnline@uh.edu or 713-743-3327.

	Date	**Topics**	Open Textbook Reading
Week 1	Tuesday, August 25, 2020	**Introduction to Data science**	
	Thursday, August 27, 2020	**Data science Overview**	

Week 2	Tuesday, Sep 1, 2020	**Machine Learning** **Data Cleaning**	
	Thursday, Sep 3, 2020	**Data Processing** **Startup Example**	B1: p 30-35
Week 3	Tuesday, September 8, 2020	**Statistical Learning**	
	Wednesday September 9	**DROP DEADLINE**	
	Thursday September 11, 2020	**Data Exploration** **Data Similarities & Distances**	B1: p 54-81
Week 4	Tuesday, September 15, 2020	**Linear Regression**	B1:p 171-213
	Thursday, September 17, 2020	**Linear Regression (Python Example)**	
Week 5	Tuesday, September 22, 2020	**Logistic Regression** **Dimensionality reduction - PCA**	B1: p 359-399
	Thursday, September 24, 2020	**Introduction to Classification KNN**	B1: p 301-312 B2: p 32-48
Week 6	**Tuesday, September 29, 2020**	**Exam 1**	
	Thursday, October 1st, 2020	**Decision Tree**	
Week 7	Tuesday, October 6, 2020	**Random Forests** **KNN**	B1: p 317-322 B2: P 49-68

	Thursday, October 8, 2020	**Naive Bayes**	B1: p 414-439 B2: p 113-140
Week 8	Tuesday, October 13, 2020	**Model Evaluations Metrics**	
	Thursday, October 15,2020	**Ridge - Lasso**	
Week 9	Tuesday, October 20, 2020	**Lines/SVM**	

	Thursday, October 22,2020	**Dimensionality reduction (feature extraction)** **Wrap Up classification**	
week 10	**Tuesday,October 27,2020**	**Exam 2**	
	Thursday, October 29,2020	**K-Means**	B1: p 523- 537 B2: 218-250

Week 11	Tuesday, November 3, 2020	**Hierarchical Clustering Heatmap**	
	Tuesday November 3rd	**DROP DEADLINE**	
	Thursday, November 5, 2020	**Storytelling**	
Week 12	Tuesday, November 10, 2020	**DBSCAN**	
	Thursday, November 12, 2020	**Cluster Validity Silhouette**	
Week 13	Tuesday, November 17, 2020	**Neural networks**	
	Thursday, November 19, 2020	**Apriori and Association rules**	B1: p 603- 617 B2: p 69-87
Week 14	Tuesday, November 24, 2020	**Dynamic Hashing -Merkle tree (Optional)**	
	Saturday December 5th, 2020	**Last day of class**	
	Thursday December 10th ,2020	**Final Exam @ 8:00 AM – 11:00AM**	

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:

✓	Homework assignments (NO drop of any HW)	25%
✓	Lab work /Workbook (drop the lowest)	20%
✓	Exam 1 (Tuesday 9/29)	15%
✓	Exam 2 (Thursday 10/29)	15%
✓	Final Exam	25%

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

Exams: Held during class times.

Homework: Four assigned HomeWorks. Topics: Regression and Classification (Week 4); decision Tree and KNN(Week 7); SVM and dimensionality(Week 10); and Clustering with cluster validity(Week 12). Students will submit their written homework by scanning and uploading their work in Blackboard (or 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 (5 points) and applying ML techniques to address those questions (35 points). Each group will eventually submit a report/online presentation of research findings and member contributions.

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**	