CSC 458 – Data Mining and Predictive Analytics I, Fall 2022 First Day Handout

TuTh 4:30-5:50, Old Main 158, Zoom classes & recordings, https://faculty.kutztown.edu/parson

My courses are multimodal this semester, meaning you can attend in-person or remotely via Zoom. I insist on maintaining 6 feet of distance from myself in order to reduce the odds of carrying the virus home to my wife & two-year-old granddaughter. I plan to wear a mask. I strongly encourage the unvaccinated to get vaccinated unless a medical condition precludes that. The unvaccinated will not only infect each other, they will also provide an environment in which the virus can mutate to more dangerous strains. I will post Zoom videos for all classes.
The class time interactive Zoom link appears on **D2L** Course CSC458 -> Content -> Overview.

Office Hours Monday 2-4, Wednesday 4-6 (Zoom only), Thursday 10-11 or by appt. All available via Zoom.

Many academic and commercial endeavors apply the techniques of data mining and predictive analytics to their data sets. Students taking this course will learn methods and software tools for locating and obtaining data of interest, for preparing data for semi-automated analysis, for interacting with software tools in analyzing data for patterns, for visualizing structural and dynamic patterns in data, and for designing systems that respond to patterns in data. Data cleaning and formatting require some programming in a modern scripting language. Other course activities include learning to use off-the-shelf software tools to accomplish the tasks of data analysis.

Prerequisite: C or better in CSC223 AND C or better in a statistics course AND junior status, or acceptance into the KU CSC graduate program.

Textbook: *Data Mining: Practical Machine Learning Tools and Techniques*, **Fourth Edition**, Witten, et. al., ISBN 978-0128042915. You can probably buy a discounted copy of the **Third Edition** (better edition). Either edition is fine, since we will not be using the two chapters added to the Fourth.

Grading (A = 92:100, A- = 90:91, B+ = 87:89, B = 82:86, B- = 80:81, C+ = 77:79, C = 70:76,
F = 0:69). There is no "D" grade for student grading in 400- and 500-level courses at KU.
http://app.kutztown.edu/policyregister/policy.aspx?policy=ACA-048

Projects	100% divided equally among the 5 project assignments. No exams.

Project assignment grading criteria

Grading criteria will accompany each assignment handout. Please follow them in satisfying all project requirements. Please re-check requirements when you feel ready to turn in an assignment.

The academic integrity policy is at http://cs.kutztown.edu/pdfs/AcademicIntegrityPolicy.pdf

Your first reading assignment is to read the above policy statement.

You may openly discuss ideas, algorithms, pitfalls, and the use of programming tools.

You may not share code, test drivers or test data except within groups for group projects.

Group projects, when assigned, have documented partitioning of student responsibilities.

There will be 5 projects. There is a 10% per day late penalty for projects that come in after the due date. The 5th project will be a condensed 1.5-week data analysis assignment in place of having an exam.

Class attendance is not graded, but I will be teaching using data sources and concepts both inside and outside the scope of the textbook. You are responsible for all material covered in class, including technical information, coding standards and conventions, verbal specification of assignments, and your questions about topics that are not clear to you. Please, there should be no classroom conversations, cell phones, text messaging, eating, sleeping, obscenities, smoking (tobacco or artificial), vaping, listening to music or other disruptions of the class. I will deduct 5% from an assignment for each infraction.

If you have already disclosed a disability to the Disability Services Office (215 Stratton Administration Building) and are seeking accommodations, please feel free to speak with me privately so that I may assist you. If you have an injury sustained during military service including PTSD or TBI, you are also eligible for accommodations under the ADA and should contact the Disability Services Office.
If you have preferred pronouns for yourself, or a name that differs from the roster, please let me know.

Week	Class Topics and readings using 3rd (4th) Edition Chapters
1	Introduction to the course plan. Scripting in Python. Chapters 1 & 2 (1 & 2).
2	Scripting in Python for data parsing, cleaning and formatting. Data file formats and regular expressions. **Assn1 on Python scripting out**. Python 3.x references.
3	Introduction to using Weka for classification of discrete (*nominal*-valued, a.k.a. *set*-valued) target attributes. Finding data. Chapter 3 (3). Possible work time.
4	Data transformations within Weka. Information entropy in building pattern recognition rules & decision trees in Weka. Parts of Chapters 4, 5, 6, 7 (4-8).
5	**Assn1 due**. **Assn2 on discrete target value classification out**. Project work time.
6	Regression, model trees, & Bayesian models. Parts of Chapters 4, 5, 6, 7 (4-8).
7	Regression, model trees, & Bayesian models. Past data science projects at KU.
8	**Assn2 due**. **Assn3 on regression & model trees out**. Project work time.
9	Problems with under- and over-fitting. Introduction to clustering, Chapters 4.8 & 6.8 (4.8).
10	Overview of Python's scikit-learn framework as a programmable tool set.
11	**Assn3 due**. **Assn4 on Bayesian models & clustering out**. Project work time.
12	Overview of instance-based and ensemble machine learning techniques. Parts of Chapters 4.7, 6.5 & 8 (4.7, 7.1 & 12).
13	Consolidation and review.
14	**Assn4 due**. **Assn5 summary project out**. Project work time.
15	Final exam period will be a work session. Project is due near end of this week.

1. Assignment 1 on data retrieval, cleaning, & formatting using Python.
2. Assignment 2 on using Weka with data to extract trees and rules, and to evaluate effective of at least two approaches. All projects after #1 include some analysis questions in a README.txt file.
3. Assignment 3 on using Weka with data to extract linear models, model trees, and possibly other trees, and to evaluate effective of at least three approaches.
4. Assignment 4 on using Weka with data to extract Bayesian and cluster models of data, and to evaluate effective of at least two approaches.
5. Final project will take place during final exam week as an on-line, condensed mini-project. I will take only questions for clarification of the assignment handout, and only during classes 14 & 15. This cumulative project introduces no new concepts or techniques. It takes the place of an exam.

We will be using Zoom for remote attendance during class time. Recorded archives of class sessions will be available within a day. We will go over Zoom & recording permissions in the first class.