Data Mining II – DA 2605

Dr. Anthony Breitzman
Computer Science Department
Robinson, Third Floor, 328P
856-256-4500 extension 53625
email: breitzman@rowan.edu

Text: None Required. If you feel you have to have one here are a couple of good ones:
Larose – Discovering Knowledge in Data
Han, Kamber, Pei – Data Mining Concepts and Techniques
Look at Abe's Books under Data Mining and you can find dozens under $20.

Software: Python (python.org); R (r-project.org); Rodeo, Sci-kit Learn, Pandas, Anaconda, SQLite, mySQL

Days/Times/Place: W: 5-7:30 at 790 building

Office Hours: Before or after class or by appt. at Rowan or LM if possible.

TA: Mike Leonchuck: leonchucm2@students.rowan.edu

Objectives:

- This is a continuation of Data Mining I. We will continue to gain a working knowledge of Python, R, and numerous packages, libraries and APIs.
- To learn how to harness massive data sets to find interesting results.
- Understand the complexity of mining massive datasets with high dimensions.
- use state-of-the art techniques to reduce the dimension of a problem without losing the intelligence hidden in the data.
- recognize which algorithms for extracting knowledge from a given set of data are most appropriate for a given problem.
- Use multiple approaches to solve actual data mining problems (sometimes over weeks, sometimes in teams)
- interpret results so that customers or companies can make intelligent business and operations decisions.
- Learn to present results and teach new techniques to peers

Specific Topics:

- Deeper Dive on Neural Networks and Perceptrons
- KDD v DM
- Sci-Kit Learn
- K-Folds Validation
- Support Vector Machines
- Random Forests/Bagging
- Boosting /AdaBoost/Tree Stumps
- Nearest Neighbors
- SMOTE
- Averaging Models and Other Ensembles
- Multiclass Classifiers
- Time Series Modeling
- Link Analysis

Additional Topics if we have time

- Kohonen Maps
- Maximum Likelihood
- Fuzzy Sets

- Seriation
- Genetic Algorithms
- NN-Rules Generation
- Rules Generation - No Tree No NN
- Anomaly detection

Attendance: You should attend all classes. There will be significant work in class and of course the class is so small that your absence will be noticed.

Kaggle.com: At least one or two homework assignments will involve entering Kaggle contests. We will pick which ones to enter as a class, and you can enter in teams or as individuals. Shout outs to anyone finishing in the top 1/3.

Homework: We are learning two languages (Python and R) as well as numerous algorithms and libraries. The only way to learn them is through doing the homework. The homework is designed to reinforce what we do in class and you won't really learn anything if you don't do the homework. Frequently the homework will require you to install libraries and packages which some find time consuming so start early.

Typically each homework will be worth 10 points but the default for doing everything right is 9/10. The only way to get the full 10/10 is to add your own challenge question and answer that shows additional research or effort. For example if the hw is to use a classification method in R and we implement in R in class and you implement it the way we did in class plus with a different package or with Python, that will get a 10. Another example is if we do a validation method in class and you do the same as HW but then add a new validation that you researched on your own then that is worth 10. Another specific example is if we do k-folds validation in class using the method I teach you but you implement it using the caret package as well that may be worth a 10.

Note incomplete or late assignments will never get above 8/10.

Poster Project: A significant part of your grade will be an end of term project that will be related to some aspect of data mining. (more on this soon). The goal is to create something that will result in a possible STEM symposium submission or perhaps a published paper (don't let this scare you; I've written 20+ papers and I will help you get your results publication ready.) Even if your project doesn't work and isn't publishable you can still get a high grade; sometimes things don't work out. (I have a PhD thesis full of useless results!)

You may work alone or with up to 2 partners.

Note I will hand out a Rubric of what is expected in the Poster project and there will be various milestones to keep you on track.

Grading: 30% HW, 30% Midterm 40% Poster Project

A	92 and Up	**C**	72 – 77
A-	90 – 91	**C-**	70 – 71
B+	88 – 89	**D+**	67 – 69
B	82 – 87	**D**	63 – 66
B-	80 – 81	**D-**	60 – 62
C+	78 – 79	**F**	59 and Below

Withdrawing: If you are doing poorly, be realistic about your chances and talk to me early. The sooner you make up your mind, the easier it is for you to drop the course in terms of the signatures you will need to obtain and it is almost impossible to withdraw during the last 4 weeks. Please visit https://sites.rowan.edu/registrar/ for important dates.

Academic Integrity: I've never had a problem in a graduate class but just in case…
Plagiarism is a form of academic dishonesty which includes but is not limited to submitting someone else's work as your own and working on the individual assignments in groups. It is college policy that students who commit an act of academic dishonesty may be subject to failure in the course, suspension from the College, or both. See https://confluence.rowan.edu/display/POLICY/Academic+Integrity+Policy for the official Rowan Academic Honesty Policy

If you use materials that you've obtained on the Internet, from a book, etc., for example as part of a programming assignment, you must include an appropriate reference. To use such materials without proper attribution is a form of plagiarism. Students who copy homework, cheat on tests, or plagiarize material for any test or assignment in this course will receive a failing grade for the test or assignment.

Students with Disabilities and Special Needs: Your academic success is important. If you have a documented disability that may have an impact upon your work in this class, please contact me. Students must provide documentation of their disability to the Academic Success Center in order to receive official University services and accommodations. The Academic Success Center can be reached at 856-256-4234. The Center is located on the 3rd floor of Savitz Hall. The staff is available to answer questions regarding accommodations or assist you in your pursuit of accommodations. We look forward to working with you to meet your learning goals.

Questions in Class: The best time to ask questions is during class. There are no stupid questions. I urge you to ask your questions during class. If you have questions that were not answered in class I will stick around after most classes or you can make an appointment.

Canvas: I will post grades on Canvas so you will always know your class average.

Use of e-mail: The best way to reach me is to send an email, but keep in mind I receive about 100 emails a day so please use e-mail judiciously. **Please note, I will not e-mail homework assignments, test grades, or course grades.** All will be on canvas. Also you will hand everything in via blackboard.