Course number and name: **Data Mining II – CS 02605**

Credits and contact hours: 3 credits / 3 contact hours

Instructor's or course coordinator's name: Anthony Breitzman

Instructional materials: None required. If you feel you want a book, here are a couple of good ones. Larose – Discovering Knowledge in Data; Han, Kamber, Pei – Data Mining Concepts and Techniques

Specific course information

Catalog description: This course follows Data Mining I which is designed to train students in the necessary algorithms for extracting intelligence from large datasets. In Data Mining II, more advanced topics are covered including advanced clustering techniques, Principal Component Analysis, Naïve Bayes clustering and other techniques.

About this course: This hands-on course will teach students how to harness massive data sets to find interesting results or to solve real world problems. Some of the objectives include:

- 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.
- interpret results so that customers or companies can make intelligent business and operations decisions.
- Gain a working knowledge of R and Python.

Prerequisites: CS 02505 Data Mining 1

Specific Topics

- Deeper dive into cross-validation and K-Folds
- Deeper Dive into Neural Networks
- Oversampling/Undersampling; SMOTE and Rose package
- Ensemble Methods/Majority Voting/Weighted Voting/Model Stacking
- Support Vector Machines/Kernel Methods
- Random Forests and varimp()
- Bagging/Boosting/AdaBoost
- Weak Learners/Decision Stumps
- Neural Network Rules of Thumb

- Binary v Multiclass; one v all; one v one
- Error Correcting Codes as Multiclass Voting Method
- Time Series Modeling (ARIMA)
- KNN Revisited
- Link Analysis; PageRank
- Seriation and Clustering
- Deeper Dive into Data Cleansing
- Crash Course on Text Mining TFIDF vector space models
- Deeper Dive into Sensitivity Analysis
- ROC/AUC Analysis