Spring 2020

IS 665-852: Data Analysis for Information Systems

Lin Lin

New Jersey Institute of Technology College of Computing Sciences

IS 665: Data Analysis for Information Systems

Course Syllabus
Spring 2020

Instructor: Dr. Lin Lin
Office: 3600A Guttenberg Information Technology Center
E-Mail: llin@njit.edu

Description

This graduate level course introduces students to the world of data analytics from an information systems perspective, focusing on the application of various data analysis techniques in business practices. We cover a wide spectrum of topics ranging from fundamental statistics to database, data warehouse, data visualization, and data mining. Being an introductory course, our approach is "shallow and wide", emphasizing on giving students a complete view of the data analytics profession, covering as many different sub-areas as time allows while not diving too deep into any one specific domain. The goal is to serve as a "guided tour" for students to gain knowledge about the different sub-areas of data analytics and understanding of which area is a best fit for their personal developments. More in-depth materials and discussion for each sub-area will be provided upon students' requests. Course topics include the rudiments of probability and random variables, and visualization, data warehousing and OLAP analysis, dashboard, scorecard, data mining algorithms, optimization techniques, DSS and knowledge systems.

At the end of this course, the student should be able to:

1. Build a solid foundation of statistics, probability theories, data structure and algorithms
2. Apply simple statistical analysis (such as descriptive statistics and regression analysis) to real world data sets
3. Design and construct data warehouse using SAP BI tools
4. Design and conduct data visualization using Tableau
5. Master commonly used data mining techniques such as neural networks, decision tree, association rules, clustering, genetic algorithm, SVM, Bayesian Networks, etc.
6. Learn Python basics and be able to use Python to run the aforementioned machine learning algorithms
7. Apply data mining algorithms to real world data sets in the context of web mining, text mining, transaction mining, etc. using RapidMiner

The students will learn to work with the following tools:

- **Excel** (for Pivot Table and Solver)
- **Tableau** (for visualization and data preprocessing)
- **SAP Business Warehouse** (for the design and construction of a data warehouse)
- **Rapid Miner** (for data mining using a graphic interface)
- **Python** (for machine learning algorithms using Anaconda and Jupiter Notebook)

Required Texts:

Considering the wide range of topics covered in this class, we will use Powerpoint slides and a collection of papers / articles for our class. Classic books for each sub area will be recommended in class as we go through course materials.

Readings

The weekly schedule of readings, topics, and assignments will be in Canvas. Make sure you check Canvas every Monday – I post new materials on Sunday nights.

Assignments (Individual and Team)

There will be several individual and team assignments over the semester. Details on each assignment will be posted on Canvas.

READING ASSIGNMENTS:
Teams will be assigned papers to read from time to time. Teams are expected to develop a 5 – 7 page Powerpoint Slide set to summarize the assignment paper. The presentations should be posted to the assigned presentation forum. For offline sections, teams will use 10 minutes to present their papers in class.

TECHNICAL ASSIGNMENTS:
There will also be technical homework assignments in this course. Some of them are individual assignments and some will be team-based. More details will be posted on Canvas regarding these assignments.

LABS:
There will be several labs IN CLASS. Your participation is expected for these labs, and absence during the lab results in 0 point for that lab.

Projects (In teams)

Objective: To demonstrate the ability to apply Data Analytics techniques to solve real world problems.

Summary: TWO projects will be assigned to teams throughout the semester.

PROJECT ONE: Data Visualization

Teams are expected to find an interesting data set and visualize it using Tableau (or PowerBI if they prefer). Each team will then present the visualization model in a final presentation to the class at the end of the semester. More discussions on this will happen in Mid-to-late March depending on our class progress.

PROJECT TWO: Data Mining

Teams are expected to work with a real-world organization to gather data set, analyze it, and try to extract insightful information / knowledge using RapidMiner or SPSS Modeler.

Late Assignments Policy

Unexcused late submission of homework receives a 20% penalty. This means that you start with 8 out of 10 points as the maximum. Assignments submitted after graded assignments are returned or reviewed in class receive no credit.

Grading

NJIT Academic Policy has grades for graduate courses assigned as follows:

GRADE	GPA	SIGNIFICANCE
A	4.0	Excellent
B+	3.5	Good
B	3.0	Acceptable
C+	2.5	Marginal Performance
C	2.0	Minimum Performance
F	0.0	Failure

Final grades for IS 684 will tentatively be assigned as follows. There may be slight modifications, depending on issues that arise during the semester.

- Labs - 10 %
- Reading Assignments - 10 %
- Technical Assignments - 30 %
- Group Projects - 30 %
- Final Exam - 20%

Total: - 100%

Excellent participation demonstrated by preparation for discussion and thoughtful contributions (on-line and in class) will have the effect of **raising** a final letter grade by one value (e.g. B to B+, or B+ to A). Likewise, poor participation demonstrated by consistent lack of preparation for discussion and little or no thoughtful contributions (on-line and in class) will have the effect of **lowering** a final letter grade by one value (e.g. A to B+, B to C+).

Honor Code

Any evidence of cheating in any form, including plagiarism, will be dealt with according to the honor code of NJIT (course failure and suspension or expulsion). Please note: There will be no warnings or chances with regard to cheating. Any discovered case of cheating will be immediately passed to the Dean of Students for further investigation. Cheating is not worth it.

You may not only fail this course but also be suspended from NJIT. The full text of the NJIT Honor Code is available for your review at http://www.njit.edu/academics/honorcode.php.

Spring 2020 Outline/Weekly Schedule – Subject to Minor Modification

Week	Theoretical Topics	Labs
1	**The one about introduction** Where we introduce ourselves, talk about what we want to learn, and customize our class together	*Installation and Setup of Necessary Software*
2	**The one on business value of data analytics** Where teams present their papers and we talk about WHY data analytics is needed today	*Presentation I (Reading Assignment 1)*
3	**The one on fundamentals (I)** Where we start our journey and discuss the basics: • Data type, data collection and data description • Data Structure and Algorithms	*Stats Lab I: Histograms and Scatter Plot* *Python Lab I: List and Array using NumPy*
4	**The one on fundamentals (II)** Where we look at more advanced data structures, discuss one killer app in data analytics, and start our adventure in data warehouse • Pivot Table and Pivot Analysis • 2D Array and Data Frame • Data Warehouse Basics	*Stats Lab II: Pivot Table* *Python Lab II: 2D Array and Data Frame using Pandas*
5	**The one about Data Warehouse** Where we explore dimensional model, the conceptual design tool for data warehouse, and work with ETL • Star Schema and Dimensional Modeling • ETL Process • Advanced Data Warehouse Topics	*SAP Lab I: Design a Data Warehouse* *SAP Lab II: Build a Data Warehouse*
6	**The one about Data Visualization (I)** Where teams present their papers and we get to know data visualization for the first time • Dashboard and Scorecard • Visualization and its business values	*Visualization Lab I - IV: Tableau* *Python Lab III: Visualization using Matplotlib*
7	**The one about Data Visualization (II)** Where teams compete on using visualization to solve a real world mystery	*Visualization Lab V - VII: Tableau* *Visualization Case: Microsoft HR*
8	**The one about Association Rules** Where the three types of data mining tasks are introduced • Supervised learning vs. unsupervised learning • Association Rule Mining	*RapidMiner Lab I: Association Rule*
9	**SPRING RECESS**	
10	**The one about Clustering** Where clustering is introduced	*RapidMiner Lab II: Clustering in RapidMiner* *Stats Lab III: Clustering in RapidMiner* *Python Lab IV: Clustering in Python*

11	**The one about Classification (I)** Where regression – all types – is introduced	RapidMiner Lab III: Regression *Stats Lab IV: Regression in RapidMiner* *Python Lab V - VII: Regression in Python*
12	**The one about Classification (II)** Where SVM, Neural Networks, Decision Tree, and other techniques are covered	RapidMiner Lab IV: Classification *Python Lab VIII: Classification in Python*
13	**The one about Classification (III)** Where SVM, Neural Networks, Decision Tree, and other techniques are covered	RapidMiner Lab IV: Classification *Python Lab VIII: Classification in Python*
14	**PROJECT PRESENTATIONS**	
15	**FINAL EXAM**	