

**FLORIDA STATE UNIVERSITY
COLLEGE OF EDUCATION
DEPARTMENT OF EDUCATIONAL LEADERSHIP AND POLICY STUDIES
EDH 5931: Special Topics: Data Mining
(3 credits- Letter Grade)**

Instructor: Gayle McLaughlin, Ph.D.
e-mail: gmclaughlin@fsu.edu

Availability: I usually check my email during the day and every evening. If you need an immediate answer to a question you can call me at 850-566-9030 *between the hours of 10:00 am and 10:00 pm (EST)*.

Course Description:

The course will provide an introduction to the theoretical concepts and practical applications of data mining in higher education. Data mining (AKA Knowledge Discovery) was first introduced in 1994 as a unique approach to the analysis of large databases. Data mining facilitates the extraction of hidden predictive information from large complex databases. It is a powerful new technology with enormous potential to help organizations and institutions extract and interpret important information.

Initially, data mining was primarily used in commercial and propriety industries to develop predictive models for consumer/customer behaviors and fraud detection. As the use of and need for large databases expanded into many areas, the predictive modeling functions of data mining were utilized in a broader spectrum of fields including Higher Education. During the past 5 years, institutional researchers have recognized the value of data mining as a support mechanism for institutional functions and knowledge-driven decisions. Data mining is extremely well-suited for use in academic institutions where accurate predictions of enrollment trends, degree completion, and student success are vital to institutional assessment, accountability, resource allocation, and marketing.

The course content will include the conceptual framework of data mining, descriptions and examples of standard methods used in data mining, and the role of data mining in contemporary institutional research. Additionally, the course will provide limited exercises and practical experience with a data mining program.

NOTE: *This is an introductory course. Statistical expertise is neither expected nor required. Basic statistical concepts will be included in the course content. No previous experience with data mining is required.*

Course Objectives:

Upon completion of the course, students will be able to:

- Understand the nature and purpose of data mining
- Describe the theoretical constructs and core processes of data mining
- Understand the role of data mining in institutional research.
- Understand the basic statistical concepts related to data mining.
- Describe the predictive modeling functions of data mining.
- Understand the types and characteristics of predictive models.
- Understand the ethical issues associated with data mining.
- Describe the potential applications of data mining in higher education i.e., decision support, assessment, accountability, resource allocation, enrollment management, and quality improvement initiatives.
- Use a data mining program to analyze sample data and develop predictive models.
- Be able to compare and evaluate the accuracy of predictive models.

Required Texts and Readings (The text is free book with your AIR \$30 student membership. Books are available for purchase separately - \$25 for members / 35% for non-members.

1. *Coughlin, MA (2005). Applications of Intermediate/Advanced Statistics in Institutional Research. Association of Institutional Research. Available online <https://secure.airweb3.org/page.asp?page=1095>*

Note: The Coughlin text will not be used in the first 2-3 weeks of the course. Please order it at the start of the course so you will have it when needed.

Readings:

Please see the reading list for more details on course documents and assignments.

- ◆ *All readings will be available online or accessible via links on class BB site.*
- ◆ *References such as a statistical glossary, a data mining glossary, and Excel tutorials are posted in the course library.*

Equipment:

In order to complete the distance-learning portion of the course, each student will need a computer with Internet access and Microsoft Excel.

Methods:

The course will utilize a combination of instructional activities including the following: powerpoint presentations, white papers, computer-assisted learning, threaded Internet discussions (Blackboard), interactive Internet discussions (Blackboard), and independent inquiry.

Expectations / Attendance:

Students are required to complete all assignments within the specified time frame. Failure to complete assignments may result in grade reduction up to and including an F in the course.

Grading / Evaluation:

Percent grades will be translated into letter grades as follows:

- 93% -100% = A
- 85% - 92% = B
- 78% - 84 % = C
- 69% - 77% = D
- Below 69% = F

Student evaluation will be based on the following:

- Student Introduction (3pts)
- Data Mining Lab Exercise 1 (10 pts)
- Data Mining lab Exercise 2 (20 pts)
- Completion of 6 Blackboard quizzes (6 pts each / total 36 pts)
- Participation in 6 weekly BB debates / discussion boards (6 pts each / total 36 pts)
- Reaction Paper (15 pts)
- Data Mining website Evaluation (15 pts)
- Submission of proposal for data mining project (10 pts)
- Completion of sample Data Mining project (25 pts)
- Evaluation of Data Mining project (30 pts)
- 200 Total Points

Assignments:

Unless otherwise stated, assignments are due before midnight EST on the due date. Each assignment will have a Discussion Board forum and individual assignments should be posted in that discussion board. All assignments must be named in a way that identifies both the assignment and the student i.e. mclaughlin-RP would be acceptable for a reaction paper and mclaughlin-LabEx1 would be acceptable for lab Exercise 1. Do not use the # sign in your file names because blackboard doesn't allow that and I won't be able to open the file. **IF YOU ARE USING VISTA** – Vista word documents default to docx formats instead of doc files and they cannot be opened with Blackboard or older Word programs. You will need to save all your assignments in the Word 97 format.

DM Lab Exercises:

Students will use a small set of sample data to complete Data Mining exercises. These exercises will focus on the organization and preparation of data for analysis.

Blackboard Quiz:

During 6 weeks of the course, each student will complete a brief online quiz based on the assigned readings. The quiz will be posted by Wednesday and is due on Monday by 8am (EST).

Online Debates / Discussion

In 6 weeks of the course, a statement will be posted as a topic for discussion on BB. Students will submit a brief statement (2-3 paragraphs) of their position / opinion drawn from their personal experience and/or course readings. In addition, students will respond to at least two statements posted by other students. Response statements should be brief (3-5 sentences) and strictly focused on the issue. Topics will be posted on Sunday. Initial responses / position statements are due by Tuesday. All postings must be completed before midnight (EST) the following Saturday.

Reaction Paper

Each student will complete a 3-5 page reaction paper on the role of Data Mining in institutional research. Papers should be prepared APA format, double-spaced, in a 12 point font. Papers should include references but references are not included in page count. You may refer to any course readings including case studies posted in the course library.

Data Mining Website Evaluation

You will be asked to explore several data mining websites and select 1 data mining program site for evaluation. Specific instructions for the website evaluation will be provided on the BB site.

Data Mining Project

Students will have access to sample of data and a commercial data mining program for 30 days during the course. The DM program will be used to analyze a small sample from an NCES dataset and generate predictive models. Students will need an FSU email account to download the data mining program.

The purpose of this project is to give you an opportunity to gain first-hand experience with a data mining program. The DTREG program will be used to generate predictive models but you will not be graded primarily on the predictive accuracy of your models. You should strive to refine the models you generate to improve accuracy but your grade will be based on your completion of the project and your evaluation of that project. The

models you develop may have high or low predictive accuracy but in either case you would be expected to compare the performance of the models and explain possible reasons for your outcomes.

Project proposal: Submit a 1-2 page description of your project proposal that includes the following: identify and define the target variable, the distribution of the target variable's raw data (case benchmark, range, mean, median, mode), the number of cases in the target variable, a list of independent variables with the range, mean, median, and mode for each variable. Also indicate the number of categories you will use for each independent variable. Some of this information can be presented in a table format – a template and sample proposal will be available on BB.

Note: You will need to use Microsoft Excel to sort and prepare your data. If you are not familiar with this program you can use one of the free online tutorials listed below:

Microsoft Excel BayCon Group- <http://www.baycongroup.com/el0.htm> (also provides a free trial version of Excel)

University of South Dakota - <http://www.usd.edu/trio/tut/excel/>

Internet4Classrooms - http://www.internet4classrooms.com/on-line_excel.htm

Data Mining Project Report

Following completion of the data mining project students will submit a summary of their DTREG results/report for each model. The required components of the DTREG Project Report will be posted on BB.

Data Mining Project Evaluation: The most important part of the data mining project is the evaluation. The required components and specific instructions for your evaluation will be posted on BB.

Academic Honor Code

The Academic Honor System of Florida State University is based on the premise that each student has the responsibility (1) to uphold the highest standards of academic integrity in the student's own work, (2) to refuse to tolerate violations of academic integrity in the University community, and (3) to foster a high sense of integrity and social responsibility on the part of the University community.

ADA Requirements

Students with disabilities needing academic accommodations should (1) register with and provide documentation to the Student Disability Resource Center, (2) send a letter to the instructor from the SDRC indicating what type of academic accommodations are needed.

**Introduction to Data Mining EDH 5931
Readings and Assignments**

Blackboard: If you are a first time user of blackboard you will need to view the online tutorial at https://campus.fsu.edu/webapps/portal/frameset.jsp?tab_id=1_1

All reading assignments are available on in course library or online at the links provided in course documents.

Week One: Introduction to Data Mining

Readings:

1. Introduction to Data Mining and Knowledge Discovery – 3rd Edition, Two Crows
2. An Introduction to Data Mining – Thearling <http://www.thearling.com/text/dmwhite/dmwhite.htm>
3. Concepts and Myths of Data Mining in Higher Education – Zhao (PowerPoint)
4. Basic Statistics for Educational Leadership <http://www.fgse.nova.edu/edl/secure/stats/index.htm>
5. Video - What is Data Mining

Assignment:

1. Quiz 1

Week Two: Data Mining In Institutional Reach

Readings:

1. Data Mining in Higher Education – Jing Luan
2. Data Mining in IR - Kumar PPT
3. Theoretical Basis for Data Mining in Higher Education – Luan et al, PPT
4. Case Studies - PowerPoints

Assignments:

1. Quiz 2
2. Reaction Paper

Week Three: Data Mining Process

Readings:

1. CRISP-DM is the current standardized process model for data mining. Review the brief overview of CRISP-DM online at <http://www.crisp-dm.org/Process/index.htm>
2. Data Mining - Beyond Algorithms - Al-Attar <http://www.xpertrule.com/tutor/mining.htm>
3. Data Mining Tutorial – Thearling <http://www.thearling.com/dmintro/dmintro.pdf>

Assignments:

1. Quiz 3

Week Four: Data Mining Techniques

Readings:

1. Data Mining Techniques and Modeling - <http://www.statsoft.com/textbook/stdatmin.html#mining>
2. Overview of Data Mining Techniques
<http://www.theartling.com/text/dmtechniques/dmtechniques.htm>
3. Video - Data Mining Demo <http://www.youtube.com/watch?v=zjbGABxhIAQ&feature=related>

Assignments:

1. Quiz 4
2. Download Local School District Survey Database (Course Library)

Week Five: Internet Resources

Readings:

1. KD nuggets: <http://www.kdnuggets.com/>
2. Data Mining in Institutional Research: <http://www.uni.edu/instrsch/dm/index.shtml>

Assignments:

1. Evaluation of one website and software
2. Blackboard Debate/Discussion

Week Six: Data Preparation

Readings:

1. Han & Kamber PowerPoint: Data Preprocessing
2. Data Preparation for Data Mining – Zhang & Zhang <http://www.cse.ust.hk/~qyang/Docs/2003/s1.pdf>
3. Data Mining: Data preparation – Kaufmann et al PPT

Assignments:

1. BB Discussion/Debate – position statement due 10/1
2. Lab Exercise 1

Week Seven: Predictive Models

Readings:

1. DTREG Manual pages 11-13; & 203-303.

Assignments:

1. Quiz 5
2. *Begin drafting DM project proposal

Week Eight: Interpreting DM Results with Statistical Tests

Readings:

1. Coughlin - Chapter 3, Regression Analysis for IR
2. Coughlin – Chapter 5, Identifying and Analyzing Group Differences
3. Interpreting Logistic Regression – UNC StatNotes <http://www2.chass.ncsu.edu/garson/PA765/logistic.htm>

Assignments:

1. Quiz 6
2. Lab Exercise 2

Week Nine: Using DTREG

Readings:

1. DTREG Manual – pages 15-54; 75-84; 100-121; 159-202

Assignments:

1. Download and Install DTREG on 10/20
2. Data Mining Proposal due 10/26

Week Ten: Visualization

Readings:

1. Thearling – Visualizing DM Models <http://www.thearling.com/text/dmviz/modelviz.htm>
2. Thearling - Understanding Data Mining: It's All in the Interaction <http://www.thearling.com/text/dsstar/interaction.htm>
3. Keim - Visual techniques for exploring Data <http://www.dbs.informatik.uni-muenchen.de/~daniel/KDD97.pdf>
4. Lift Gain Tables/Charts - http://www2.cs.uregina.ca/~hamilton/courses/831/notes/lift_chart/lift_chart.html
5. Video - Hans Rosling: No more boring data: TEDTalks <http://www.youtube.com/watch?v=hVimVzgtD6w>

Assignment:

1. BB Discussion/Debate

Week Eleven: Data Warehouses

Readings:

1. Data Ware House Information Center - All sections
<http://www.dwinfocenter.org/index.html>
2. Video: IBM DB2 Data Warehousing - Convergence CT <http://www.youtube.com/watch?v=koa0p0sH3sc>
3. Florida's Education Data Warehouse - <http://bi.sunshineconnections.org/Pages/Default.aspx>

Assignment:

1. BB Discussion/Debate

Week Twelve: Decision Support / Business Intelligence

Readings:

1. Decision-Making Systems, Models, and Support - Turban, et al - PPT
2. Data Mining Technologies and Decision Support Systems for Business and Scientific Applications - Ganguly and Gupta
3. IR in Support of Planning – Nel 2008 PPT
4. Effective business intelligence: From decision support to supporting decisions –
<http://www.kmworld.com/Articles/Editorial/Feature/Effective-business-intelligence-From-decision-support-to-supporting-decisions-9074.aspx>
5. Video - To Data Warehouse and Beyond! <http://www.youtube.com/watch?v=IBBm3wRtKIY>

Assignments:

1. BB discussion/debate
2. DTREG Report Summary

Week Thirteen: Ethical Issues in Data Mining

Readings:

1. Data Mining: Where Legality and Ethics Rarely Meet – eCommerce Times <http://www.ecommercetimes.com/story/52616.html?welcome=1214690547>
2. Legal and Ethical Issues in Data Mining- Charlesworth PowerPoint
<http://www.cs.bris.ac.uk/Teaching/Resources/COMSM0204/handouts/privacy.ethics.pdf>
3. Video - Facebook - Controversial use of datamining
<http://www.youtube.com/watch?v=OwnTWZ1-UWY>
4. Video Online surveillance software - <http://www.youtube.com/watch?v=4IKpD7MC22I&feature=related>
5. Video- Giuliani Making Millions From Data Mining Company http://www.youtube.com/watch?v=9ING4Z2_AxU
6. Video - Israel and Telephonic Data Mining <http://www.youtube.com/watch?v=e5C2rnjdWfk>
7. Video - Conspiracy Goes Mainstream: CNBC's Big Brother, Big Business <http://www.youtube.com/watch?v=bWB3kEw08Gk>

Assignment:

1. BB Discussion/Debate

Week Fourteen: Holiday – No readings / assignments

Week Fifteen: Project Management

Readings:

1. SPSS Data Mining Tips
2. CRISP-DM 1.0 Step-by Step Data Mining Guide

Assignments:

1. *Data Mining Project Evaluation due*

ASSIGNMENTS			
Week 1	Introduction to Data Mining	Quiz 1	
Week 2	Data Mining In IR	Quiz 2	Reaction Paper Due
Week 3	Data Mining Process	Quiz 3	
Week 4	Data Mining Techniques	Quiz 4	<i>Download LSDF Database</i>
Week 5	Internet Resources	Discussion	Software Website Evaluation
Week 6	Data Preparation	Discussion	Lab Exercise 1
Week 7	Predictive Models	Quiz 5	<i>Draft Proposal</i>
Week 8	Interpreting Statistics	Quiz 6	Lab Exercise 2
Week 9	DTREG Models	<i>DTREG</i>	DM Proposal Due
Week 10	Visualization	Discussion	<i>Run Models</i>
Week 11	Data Warehouses	Discussion	<i>Refine Models</i>
Week 12	DM for Decision Support	Discussion	DTREG Report
Week 13	Ethical Issues	Discussion	
Week 14	Thanksgiving Holiday		
Week 15	Project Management		Project Evaluation Due