Welcome to Introduction to Data Mining!

I look forward with pleasure at our learning partnership in this online education enterprise. Let us work together to bridge the distance, and be introduced to each other and to the world of data mining!

Course Description: Fundamental concepts of data mining. Topics may include data preparation, data cleaning, exploratory data analysis, statistical estimation and prediction, regression modeling, multiple regression, model building, k-means clustering, and classification and regression trees. Prerequisite: STAT 104 or STAT 200 or STAT 215 or STAT 315 or permission of department chair. (In other words, you need to have taken a stats course at some point in your life.) 4 credits.
Textbook:

Data Mining and Predictive Analytics
By Daniel T. Larose and Chantal D. Larose
ISBN: 9781118116197

Data Sets: You may download the data sets used in the books from www.dataminingconsultant.com.

Required Software: The software we will be using for this course is IBM SPSS Modeler. Moreover, we will be using open source software R (optional). Note that SPSS Modeler is the most widely used data mining tool.

IBM SPSS Modeler Grad Pack, latest version.
Information about the software is available here http://www.spss.com/software/modeler/
Note, the CCSU bookstore may not sell IBM SPSS MODELER anymore. However, you may purchase this here:

Open Source R (optional): Information about this software, and how to download and install the software visit CRAN home page at http://cran.r-project.org/.

Remote Access to Other Software: As CCSU students, you may remotely access SPSS, Minitab, and other software, as long as you are registered for courses at CCSU. The software is located on a Citrix server at CCSU, and may be accessed at http://citrix.ccsu.edu. The first time you use it, you will have to download the client application to your machine. In addition to statistical software, you may also access the following Microsoft products remotely: Access, Excel, Frontpage, Outlook, PowerPoint, Producer, Project, and Word.
You will need to make sure your CCSU student account is set up properly. Go to the secure site at https://accounts.ccsu.edu. I hope you enjoy this perk of being a CCSU student. Questions about the Citrix server (though not about the various applications) may be addressed to Michael Archick at ArchickM@mail.ccsu.edu. Please note that the remote access to SPSS Modeler is no substitute for purchasing the software, since the remote access is rather clunky for very large data sets like the one you

Course Syllabus
Stat 521: Introduction to Data Mining
Krishna K. Saha, Ph.D.
will be analyzing for the course project. But it is workable if you would like to play with it with small data sets while the software is on order or in the mail.

Why take this course?
MIT Technology Review called data mining:
“One of the ten emerging technologies that will change the world.”

Corporations have spent millions, gathering terabyte-sized databases. Now they need to take advantage of the nuggets of information hidden in those huge databases. That is where data mining comes in. The world is drowning in data, but starved for knowledge. Analysts who can show that they can understand and apply today's data mining procedures will be golden in the global marketplace.

How Will This Course Work?
The course is divided into 5 units as shown below. To see which unit covers which dates, see the course calendar (a link to it is on the Homepage). All work due for a particular week must be done by the date and time indicated (which is Eastern Time): no exceptions. To avoid last minute computer glitches etc you should get your work in as early as you can. Computer glitches which make your report or other work late will cause points to be docked, just as it would if you were submitting the material to your business manager. Your grade will be based on the online quizzes and the final project – the weights for all of these are given in the Course Assignments Table below.

Course Assignments and Due Dates: The due dates are final, no exceptions.

<table>
<thead>
<tr>
<th>Reading from Textbooks</th>
<th>Assignment</th>
<th>Points of Course Grade</th>
<th>Due Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>DKD: Discovering Knowledge in Data Mining</td>
<td>Introduce Yourself Thread</td>
<td>5</td>
<td>Jan 25, 11:55PM</td>
</tr>
<tr>
<td>DMMM: Data Mining Methods and Models</td>
<td>DKD Chapter 1 Quiz in Unit#1</td>
<td>33</td>
<td>Feb 1, 11:55PM</td>
</tr>
</tbody>
</table>

Course Syllabus
Stat 521: Introduction to Data Mining
Krishna K. Saha, Ph.D.
## Course Syllabus

### Stat 521: Introduction to Data Mining

Krishna K. Saha, Ph.D.

---

<table>
<thead>
<tr>
<th>Unit 2: Feb 2 – Feb 29</th>
<th>Data Quiz 1 in Unit#2</th>
<th>18</th>
<th>Feb 29, 11:55PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch 2 (DKD): Data Preprocessing</td>
<td>Data Quiz 2 in Unit#2</td>
<td>18</td>
<td>Feb 29, 11:55PM</td>
</tr>
<tr>
<td>Ch 3 (DKD): Exploratory Data Analysis</td>
<td>Data Quiz 3 in Unit#2</td>
<td>18</td>
<td>Feb 29, 11:55PM</td>
</tr>
<tr>
<td>Data Preparation Quiz in Unit#2</td>
<td>EDA Quiz in Unit#2</td>
<td>21</td>
<td>Feb 29, 11:55PM</td>
</tr>
</tbody>
</table>

**Data Quiz 1** in Unit#2

**Data Quiz 2** in Unit#2

**Data Quiz 3** in Unit#2

**Data Preparation Quiz** in Unit#2

**EDA Quiz** in Unit#2

<table>
<thead>
<tr>
<th>Unit 3: March 1 – March 28</th>
<th>Regression Quiz 1 in Unit#3</th>
<th>21</th>
<th>March 28, 11:55PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch 4 (DKD): Univariate Statistical Analysis</td>
<td>Regression Quiz 2 in Unit#3</td>
<td>21</td>
<td>March 28, 11:55PM</td>
</tr>
<tr>
<td>Ch 5 (DKD): Multivariate Statistics</td>
<td>Regression Quiz 3 in Unit#3</td>
<td>15</td>
<td>March 28, 11:55PM</td>
</tr>
<tr>
<td>Ch 2 (DMMM): Regression Modeling</td>
<td>Multiple Regression Quiz in Unit#3</td>
<td>34</td>
<td>March 28, 11:55PM</td>
</tr>
<tr>
<td>Ch 3 (DMMM): Multiple Regression</td>
<td>-----------------------------</td>
<td>----</td>
<td>------------------</td>
</tr>
</tbody>
</table>

**Regression Quiz 1** in Unit#3

**Regression Quiz 2** in Unit#3

**Regression Quiz 3** in Unit#3

**Multiple Regression Quiz** in Unit#3

<table>
<thead>
<tr>
<th>Unit 4: March 29 – April 18</th>
<th>Decision Tree Data Quiz in Unit#4</th>
<th>28</th>
<th>April 18, 11:55PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch 8 (DKD): Decision Trees</td>
<td>-----------------------------</td>
<td>----</td>
<td>------------------</td>
</tr>
</tbody>
</table>

**Decision Tree Data Quiz** in Unit#4

<table>
<thead>
<tr>
<th>Unit 5: April 19 – May 4</th>
<th>Final Project</th>
<th>250</th>
<th>May 9th at 5:00PM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ch 6 (DKD): Preparing To Model The Data</td>
<td>----------------</td>
<td>----</td>
<td>------------------</td>
</tr>
<tr>
<td>Ch 14 (DKD): Model Evaluation Techniques</td>
<td>----------------</td>
<td>----</td>
<td>------------------</td>
</tr>
</tbody>
</table>

### Grading Scale:
The grading scale for this class is pretty standard. It would be based on your total cumulative points.

- **A**: above 465; **A-**: 450-465; **B+**: 435-449; **B**: 415-434; **B-**: 400-414; **C+**: 385-399; **C**: 365-384; **C-**: 350-364; **D+**: 335 – 249; **D**: 315- 334; **D-**: 300-314; **F**: below 300.

### Withdrawal Policy:
After **April 18**, withdrawals are allowed only under extenuating circumstances and require approval of the course instructor, department chair, and dean of the School of Engineering, Science & Technology.

### How Much Time Should I Spend Studying and Working on the Course?
Never for a moment lose sight of the fact that this is a challenging four-credit graduate-level course, and will require your best effort, especially in an online environment. Therefore, if you are to succeed in this course, you will have to set aside a significant amount of time for studying, assignments, projects, and exams. **Four-credit graduate courses typically require about 275 hours of student work altogether. If you do not make that commitment, your chances of success are greatly diminished.**

### Threaded Discussions:
Participation in the threaded discussions is viewed as a very important part of the online learning process. Feel free to post new items to the Discussion thread for each week. These can be questions, comments, or response to someone’s posting. **I have made the separated thread discussion section for Timothy Michelson (our GA for this class) so you should post your item about Modeler on to “Timothy Thread Discussion”.**

The Introduce Yourself threaded discussion is required. Participation in the threaded discussions is very important and recommended. **For the Introduce Yourself threaded discussion, you are required to**
make one original post of your own, and to respond to at least one other student's posting. This should be full 5 points, *if you address all the items completely.* Go to the welcome page or Thread Discussion to see the links. Get to know your fellow classmates, and you will find that we have a special group of people here.

**Homework:** Homework assignments will be given in this course, but will not be collected for grading. Actively solving these problems are strongly recommended and the best way to learn the course material. Note that you have the opportunity to brainstorm with each other through threaded discussion, share ideas if you wish, and ask questions.

**Quizzes:** There are eleven online quizzes in this course. You must take the quizzes on or before the due date (*see the due dates on Page 4*). Quizzes must be taken at any time before the due date – no make-up quizzes will be given. Note that, if you don't like your grade on the first attempt, then you may take it again (maximum two attempts). The higher of your two grades will be entered as your grade.

**Final Project:** There is one major Data Mining Project Assignment this semester based on IBM SPSS Modeler. So please order and install the software! Please direct installation queries to SPSS technical support at [www.spss.com/tech](http://www.spss.com/tech). Note that you may use R software to do any part of this project that you feel it would be more convenient for you.

**Submitting the Final Project:** When you open the Final Project, please note carefully the instructions on how to submit your work. You will be asked to submit it via email. In addition, you will be required to provide me with a hard copy. US Mail (first class or priority) or the equivalent (Fed Ex, etc) is acceptable. **Do NOT fax it.** My mailing address is at the top of this page.

Please note that the final project must be finished by its due date (*see the due date on Page 4*). If you do not turn the final project in by its due date, **you will fail the course.** Moreover, there will be no incompletes given for this course. **If procrastination is a problem for you, then please don’t take this course.**

**Caveat:** You need to be comfortable and proficient with computers and the Internet environment. Online education is not for everyone! You should be mature and work well independently, and you need to be comfortable with the Vista platform. You are responsible for all material assigned during the semester, the content of the texts and units covered, and the PowerPoint notes. (Reading the power points is not required, since they simply recapitulate what is in the books. However, some students seem to like them. **Reading the books is required.**)

**Office Hours and Getting Help:** My office hours for the semester are provided above. Students living within driving distance of CCSU should feel free to visit for questions and general chit-chat. Other students should feel free to phone during these times. **All students should email me to set up an appointment, to make sure that I am not busy with other students or meetings.** I usually respond to email within a reasonably short time (depending how difficult your question is!).

---

Course Syllabus  
Stat 521: *Introduction to Data Mining*  
Krishna K. Saha, Ph.D.
Note that, unfortunately I wouldn’t be able to respond your email on the weekend as well as Monday-Friday after 5PM. However I will be in school Monday-Friday from 8:30AM to 5PM. As I said if I am not busy with other stuff I will respond to your email within a reasonably short time.

It is recommended that you network with your fellow students, make friends, and help each other with the homework. One of the problems with online courses is that some students feel isolated sometimes. The best way to get started networking is to participate enthusiastically in the introductory threaded discussion, where everyone introduces themselves and tells us all about themselves. Very often, we find that “it is a small world”, and that we have much in common with our fellow classmates.

When you run into problems, you may contact the friends you have made in the course, or you may feel free to contact me. The best way to contact me is through email at sahakrk@ccsu.edu. Describe the problem to me and ask for help. If you wish, we can set up a time to talk on the phone. OnlineCSU questions may be directed to the help desk at hdo2.collegis.com/LT/edu-starthere.asp.

In this semester, we have a graduate assistant to help you. The graduate assistant is there primarily to help you with Modeler, but can also help with subject matter as well. He should be your first resort for Modeler software questions.

**Academic Misconduct:** Academic misconduct, including cheating and plagiarism, is prohibited. For more on CCSU’s Academic Misconduct Policy, see [www.ccsu.edu/AcademicIntegrity](http://www.ccsu.edu/AcademicIntegrity).

**Good Luck in Introduction to Data Mining:** I hope you enjoy it from your side of the monitor as much as I do from mine! If you see something that bothers you about the course, or there is some error in the quizzes or whatever, please do not hesitate to let me know at sahakrk@ccsu.edu. Thanks for coming along on our journey of discovery through the interesting world of data mining.

Best wishes for a happy, interesting, and productive course. Together, we will work hard, learn a lot, and have some fun! You are on your way to a new Master of Science in Data Mining!