DATA MINING APPLICATIONS IN PUBLIC HEALTH
HPA 565

Course Description: This online course presents the key public health information system sources, describes the process of data mining and introduces the student to a sample of data mining techniques.

Course Overview: The need to understand complex, sizable, and information rich data bases had as a consequence the evolution of computer intensive techniques which are directly based on the information contained in the data bases. Physical principle models (e.g., Monod population growth model), postulated assumptions, experimental design techniques, etc., lost their predominance with the advent of massive (Gigabytes scale) data bases, exploratory data analysis, data driven discovery, and deductive learning techniques. It’s no surprise that the MIT Technology Review chose data mining as “one of ten emerging technologies that will change the world.” (http://www.technologyreview.com)

Data mining (DM) is an iterative, computer intensive, and user centric process which has as its primary goal the conversion of data bases into useful decision making information with the identification of structural characteristics (e.g., patterns, rules, constraints, etc.) and the establishment, under certain conditions, of predictive models. These characteristics and models are critical for many organizations since they can be deployed to improve operations, services and interactions with customers, as well as to formulate more effective risk management plans (e.g., medical fraud detection).

Course Objectives: The overall objectives of this course are to:
- promote a critical understanding of the major theories, practices, and techniques associated with the DM process
- Advance an operational understanding of the main DM modeling and evaluation techniques, their potentials, and limitations

Students should keep in mind that they are not expected to become DM programmers; however, by the end of the course they are expected to have a critical understanding process understanding data understanding data preparation modeling evaluation deployment of the overall DM process, the knowledge to select and the skills to apply appropriate DM techniques in order to prepare databases, identify patterns, and to formulate models and evaluate them.

Prerequisites: BSTT400, and consent of the instructor.
**Course Management System:** The course content will be delivered via UIC’s Blackboard Learning System. Students will be enrolled into the Blackboard course site each semester after registering. The Blackboard course site is the primary content area for instructors to post lecture notes and articles, post and update grades, and make announcements. Students are expected to log into their course site regularly to learn about any developments related to the course. Learn more about Blackboard at [http://blackboard.uic.edu](http://blackboard.uic.edu).

**Learning Resources:**
*Introduction to Clementine;* SPSS, 2002. Selections from this text will be provided online.

Predictive Analysis Software Package (PAWS Graduate Pack). This is a comprehensive and easy to use package of the most commonly used DM tools. **Non PHI students** must order this by the second week of the course. Ordering instructions are on the Blackboard course site.

Data files required for certain data mining assignments, along with any additional readings, will be provided to you in the Blackboard site.

**Grading:**

<table>
<thead>
<tr>
<th>Component</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assignments</td>
<td>55%</td>
</tr>
<tr>
<td>Quizzes</td>
<td>15%</td>
</tr>
<tr>
<td>Project</td>
<td>30%</td>
</tr>
</tbody>
</table>

Participation: Distance learning offers a number of advantages to professionals wishing to enhance their education in the environmental health sciences disciplines. A paramount component of the didactic process is active and critical participation. We strongly encourage you to participate by posting questions, comments, thoughts, etc.

Assignments: There are **four (4) Assignments** (55% total). The assignments cover specific aspects of DM techniques (e.g., analyzing a data file and identifying "extreme" values). You will be required to run the PASW Modeler, analyze the data, and produce a small report with your comments. Another aspect of the assignments will be the "qualitative" small essay questions which will help you clarify the potentials/limitations of these techniques. For each assignment you will have at least two weeks (see workload chart), during that period you can post questions in the relevant forum. Feedback and grades will be given after, approximately, two weeks. Most of the assignments will be partitioned into two sections in order to provide you with timely feedback and optimize the workload. The work must be **an individual effort.**

Quizzes: There are 3 quizzes with questions based on the reading material and the assignments. In addition, a voluntary test quiz is given to familiarize you with the Bb quiz interface and to help you understand the syllabus content; it is worth 1 bonus point. The format of quizzes will be multiple choice, true/false, etc. Details about each quiz will be posted in advance (content, format, duration, number of questions, etc.). To avoid connection problems the quizzes are split in two, or more, parts. The quiz will be accessible during a **specified time period only** (app. 7 days); make sure you take it during this **time period.** In addition, when you start the quiz you will have a few only minutes to finish it (e.g., 15 minutes). Keep in mind that **all** these quizzes are designed to assess your critical comprehension of the material. **Summary feedback** will be given, approximately, two weeks after the deadline date. The summary feedback will include the correct answer, the **average score** of the entire class per question (**not** your individual scores), and detailed comments and discussion for “problematic” questions (i.e., questions in which the class scored less than 60%).

© Board of Trustees, University of Illinois at Chicago