IS 467 – FUNDAMENTALS OF DATA SCIENCE
SPRING 2018

Instructor; Email:	**Hamed Qahri-Saremi, Ph.D., SAP Certified Associate**; hqahrisa@depaul.edu
Office; Phone:	CDM 738; 312-362-5841
Office Hours:	Tuesdays & Wednesdays: 4:00 pm – 5:30 pm (CDM 738)
Class Day & Time:	Wednesdays: 5:45pm– 9:00pm
Section Numbers:	901 (on-campus section) & 910 (online section)
Class Room:	CDM 214 (Loop Campus)

COURSE DESCRIPTION

- An introduction to the Knowledge Discovery Technologies covering all stages of a data mining process: domain understanding, data collection and selection, data cleaning and transformation, dimensionality reduction, pattern discovery, evaluation, and knowledge extraction. The course provides a comprehensive overview of data mining techniques used to realize these stages, including traditional statistical analysis and machine learning techniques. Students will analyze large datasets and develop modeling solutions to support decision making in various domains such as healthcare, finance, security, marketing, customer relationship management (CRM), and multimedia.
- Prerequisite: IT 403 or CSC 423

LEARNING OUTCOMES

After completing the course, the students should be able to:

- Identify basic concepts, terminology, models and methods in the field of data mining.
- Develop and evaluate different data mining algorithms.
- Apply data mining algorithms to datasets.
- Recommend designs of knowledge discovery systems for specific problems.

REQUIRED AND RECOMMENDED (OPTIONAL) TEXTBOOKS

- ***Required Textbook:*** Han, J., Kamber, M., & Pei J. (2012). *Data Mining: Concepts and Techniques*, Morgan Kaufman Publishers, **Third Edition,** ISBN 978-0-12-381479-1 *(Textbook Webpage: http://www.cs.uiuc.edu/~hanj/bk3/)*.
- ***Recommended Textbook:*** Shmueli, G.; Bruce, P.; and Patel, N.R. (2016). *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner*, Third Edition, John Wiley & Sons, ISBN: 978-1118729274 *(Textbook Webpage: http://www.dataminingbook.com/book/3rd-edition)*.

GRADING

- 56% Assignments (Four Individual Assignments; 14% each assignment)
- 35% Group Project: Proposal (5%), Presentation (10%) and Report (20%) (Group Assignment)
- 9% Class Participation.
- 2% Bonus Credit for Responding to Research Surveys (Extra Credit – Optional)

Grading Scale: A: 93-100; A-:90-92; B+: 87-89; B: 83-86; B-: 80-82; C+: 77-79; C: 73-76; C-: 70-72; D+: 67-69; D: 60-66; F: 0-59.

CLASS PARTICIPATION

On-campus students are expected to attend each class and to remain for the duration. The overall grade for participation drops one-third after any unexcused absence. Three absences for any reason, whether excused or not, may constitute failure for the course. Use of cellphones (for call or text) and computers in class are prohibited unless I explicitly allow them. Should you need to answer a call during class, you must leave the room in an undisruptive manner. Out of respect to fellow students and the professor, texting is never allowable in class. If you are required to be on call as part of your job, please advise me at the start of the course. If you need to use your computer in class for a purpose directly related to the course, you need to advise me at the beginning of the class.

For online students, the class participation credit will be calculated based on their collaborations with their groups toward their group projects.

All students are accountable for material covered and assignments/announcements made in any class sessions that they miss. Students are expected to be active learners, coming to class prepared to participate in discussion of the topics under consideration, asking good questions and making valuable observations.

BONUS CREDIT

Students who participate in research studies (as respondents) will receive 0.5 credits for every 30 minutes of studies they participate. Each student can earn **up to 2% bonus credit (extra to 100%)** by participating in different research studies as participants. This activity will benefit you and the researchers. You will learn more about research methods first-hand by participating in them and our researchers at DePaul will be able to collect data in support of their research studies that benefit our academic community.

If you are interested, you can register on this site: https://depaulurparticipant.sona-systems.com. At the end of the quarter, I will be provided with a list of students and the points earned to calculate the extra bonus (maximum of 2%).

ASSIGNMENTS, LABS, AND PROJECT INFORMATION

- **Assignments (Individual Activity; Turnitin Assignments):**

  There are four mandatory assignments in this course related to different data mining topics that we discuss in class. For the assignments, students are required to submit complete responses to each of the questions raised and include all the necessary details, such as screenshots of the outputs from the system, in support of their responses. Furthermore, students must include the questions before their responses. Please note that completeness and thoroughness of submissions for assignments is the responsibility of the students. Also, it is students' responsibility to check that the assignment files are submitted correctly and before the deadline on D2L. Students must always keep a copy of their submissions. Any questions about the assignments must be discussed with me, at least 48 hours before the submission deadline. The deadline for submitting each assignment has been indicated in the course

schedule, at the end of this document. All assignments use Turnitin functionality on D2L to automatically check for plagiarism. Assignments with higher than 20% similarity (excluding the similarity due to questions) will not be graded (will receive zero credit).

- **Group Project (Group Activity; Turnitin Submissions):**

  This course includes an extensive final group project. The final group project is one of the main means for evaluating the data mining knowledge of the students; therefor bears a significant weight of the total grade, as follows.

  ### *1. Group Formation for Group Project*

  Students should form groups of up to 3 students for the group project. Each group should elect a contact person, who is responsible for the communications with me on behalf of the group. The contact person should submit the group members' full names and DePaul emails along with their project proposal in a Microsoft Word file to the pertinent folder on D2L by the deadline indicated in the course schedule (find it at the end of this document).

  Online students should use the online discussion forum created for this purpose on D2L to communicate with other online students and form a group.

  Students without a group after the deadline will be randomly assigned to new groups or existing groups with less than 3 members.

  ### *2. Final Group Project Proposal*

  Each group of students will work on a data mining problem involving real data. The project will be carried out throughout the quarter. It is the group's responsibility to think of and devise an interesting and feasible problem to investigate and obtain the appropriate datasets for this purpose. Potential data sources can include, but are not limited to the Internet (e.g., search for "datasets for data mining"), current or past employers (with permission), public databases, and datasets that you collected in the past. Some examples of publicly accessible datasets are as follows.

  - *https://www.kaggle.com/*
    - *A free user name and password should be created to access the content.*
    - *You may have a chance to enter a data mining competition for a prize. Please note that whether you decide to enter a competition or not is your group decision and is **completely independent** of your requirements for this course. Therefore, your credits in your final project do not in any ways indicate/determine your success/failure likelihood in the Kaggle competition. Similarly, your success/failure in a Kaggle competition does not indicate/determine your credits in your final project.*
  - *A list of datasets on Data Science Central resource:*
    - *http://www.datasciencecentral.com/profiles/blogs/great-sensor-datasets-to-prepare-your-next-career-move-in-iot-int*
  - *Datasets on Kdnuggets:*
    - *http://www.kdnuggets.com/datasets/index.html*
    - *http://www.kdnuggets.com/datasets/government-local-public.html#usa*
  - *AWS Public Datasets*
    - *https://aws.amazon.com/datasets?_encoding=UTF8&jiveRedirect=1*

- *http://fimi.ua.ac.be/data/*
- *http://www.rdatamining.com/resources/data*

NOTE: Finding the dataset for a meaningful data mining project is the group responsibility and bears part of the credit. Furthermore, it is the group's responsibility to identify and respect the copyright and privacy restrictions related to using the datasets.

To assist groups in choosing a feasible and interesting problem and an appropriate dataset, each group should obtain my approval regarding the scope and nature of the problem, the dataset, and the nature of the analysis. To that end, groups must submit a project proposal based on the template posted to D2L for the project proposal.

Project proposal bears its own credits. The deadline for submitting final project proposals is indicated in the course schedule, at the end of this document.

I will review the proposals and will give my feedbacks to the group, in which I will either approve the proposal, or will ask for a revision. If approved, the group can start working on the proposed project immediately. If asked for a revision, the group should submit a revised version of the proposal within a week. If needed, each group can meet with me online/offline to discuss their progress on the project and get some feedback.

3. Final Group Project Presentation

Each group needs to present their group project, in which each member presents a part of the content, using PowerPoint slides. **Each group member must present a part of the work**. A group member who does not present and is not excused from it will get no credits for the final project. Please note that the projects are expected to be complete for presentation. More details about the time-limit for presentations will be posted on D2L at least a week prior to presentations. Format for presentations are different for on-campus and online students, as follows.

- ***On-campus groups:***

  These groups will be presenting their projects in class using PowerPoint slides (see the course schedule for the date). They need to prepare and submit the PowerPoint slides for the presentation to D2L (see course schedule for the deadline). Note that Slides are not meant to be read but viewed. Don't read off of the slides or your script; talk to the audience and explain the topics the way you have understood them. Ensure you provide clear details about the data mining project, the dataset you have used, the data preprocessing steps and data mining techniques you have used for analyzing the dataset, and your findings. Your findings must be meaningful with respect to the objective of the project (make sure you explain them well).

- ***Online groups:***

  Online students will need to record the video of their group presentation. The video file in .mp4 format should then be submitted to the D2L folder for "Group Project Presentation" by the deadline (see course schedule). The video must show the slides, presenters face, and they should take turns and present the project. Groups need to ensure that each of the group members present a part of the work and one member records the video of the session.

  The recommended tool for this purpose is Zoom (https://zoom.us). This is an online video conference application that allows you to easily setup an online meeting with your group members, share screens, present your work, and record the whole session. Only one member needs to record the session. Zoom will save a .mp4 file locally on the computer of the person who has recorded the session. That .mp4 file needs to be submitted to D2L folder for presentation. Alternatively, if .mp4

file is too big to upload to D2L, you can upload it to YouTube privately and submit the link to the YouTube video on D2L.

- *Zoom is a free and popular application for this purpose. If you have problems or questions regarding how to use it, you can refer to its FAQ page: https://support.zoom.us/hc/en-us/articles/206175806-Top-Questions.*

4. Final Group Project Report (Turnitin Submission)

Finally, all groups must submit a final report for their group project. The final project report (*approximately 10 MS Word pages, single-spaced, font 11 Times New Roman, including everything, 1-inch margin all around*) should be submitted to D2L by the deadline indicated in the course schedule. For any required citation and referencing, you must use APA referencing format *(https://condor.depaul.edu/writing/writers-citations-and-style-guides-apa.html)*.

The report should include:

- Title page (Title of the project along with complete members' names)
- Abstract and highlights of the project
- Problem description (i.e., description and some contextual information about the problem addressed in the report and why it is an important problem to be tackled).
- The Dataset used, its size and variables along with variables' definitions (in a Table).
- The Preprocessing steps taken for improving the quality of your data for your project (e.g., outlier analysis, binning of variables if needed).
- Data Mining techniques used for analyzing data (i.e., describing the steps in your data mining process, including variable selection, model/technique selection, and visualization techniques used)
- Conclusions and practical (actionable) recommendations.
  - *Please note that it is recommended to use multiple relevant data mining techniques and compare and discuss your findings based on them.*
  - *Groups are expected to craft their final project reports considering the comments given during their presentations.*
  - *The final project report is a Turnitin assignment; therefore, will be checked for its originality. Originality of less than 80% will result in no credits for the final report (i.e., no more than 20% similarity).*

Evaluation of both the project presentation and the final report is based on the following evaluation criteria:

a) Description of project objectives and variables measured

b) Reasons for selecting the study and variables

c) Clear explanation of data preprocessing procedure

d) Proper selection and use of data mining techniques

e) Accuracy of the results

f) Meaningfulness and Accuracy of interpretation and conclusions

g) *Overall quality and flow (i.e., the final project report should strictly follow the principles of academic writing).*

- **Software – SPSS Statistics:**

  This course is not about any particular software package; it is about data mining process and fundamental techniques. However, in order to have students experiment with the techniques that we discuss in class without having to code them from scratch, we need to use a user-friendly software that has already been developed. SPSS Statistics is a very popular data mining and statistical analysis solution in industry and allows you to run the algorithms we are discussing by choosing options from menus as opposed to writing codes. Therefore, SPSS Statistics will be the default choice for the assignments and the project in this course.

  There are three labs in this course that are taught using SPSS Statistics to ensure everyone has a common tool for data mining tasks (see the schedule at the end of course syllabus). The lab sessions are held in a computer lab, where students will work through the steps provided in the lab instructions to practice data mining techniques with SPSS Statistics, while I am available to help. For online students, you can access SPSS Statistics remotely (see below) and work through the labs yourself. I recommend leaving the lab session video on in the background while you do it. If questions come up from students in class that will be relevant to everyone, I go to the front of the room to repeat them so you can hear them.

  Note: while SPSS is the default and the recommended choice for this course, students can use other software if they prefer (e.g. SAS, R, Python). If you choose to use other software for assignments or project: (1) you must submit all the codes you write as an addendum, and (2) you should note that I cannot provide any support for software packages other than SPSS Statistics (i.e., you will be solely responsible for rectifying the errors). Therefore, if you are not proficient enough to use other software than SPSS Statistics without problem, I strongly recommend taking the opportunity to learn SPSS Statistics, which is a popular software in industry so you can add it to your list of skills on your resume.

  ***Access to the software*:** The software is available in CDM labs, and also through the CDM terminal services. This means you can access it on your laptop via the internet for free. For instructions on how to remotely access the terminal services and how to activate your CDM account, please visit http://my.cdm.depaul.edu/resources/Terminal_services_guide.pdf.

  Purchasing the software: The University does not offer student licenses for SPSS Statistics (as they do for SAS) but, if you want to purchase it, SPSS offers rental options.

IMPORTANT NOTES FOR ALL SUBMISSIONS:

- All submissions in this course must be in an electronic format and should be submitted to the right folder on D2L. This is the students' responsibility. Also, always keep a copy of your assignments for yourself in case they are not submitted correctly. **No hardcopy and/or emailed submission is accepted (they will not be graded)**.
- In order to maintain a good performance in this course, it is crucial to submit the deliverables on time. Deliverables are due on a specified date and time, as stated in the course schedule (at the end of this document), unless an extension/exception is announced.
  - Late assignments will be subject to 10% penalty for each day of late submission (i.e., from one minute to 24 hours late). Assignments that are more than THREE days late will NOT receive any credits (The assignment folder on D2L will automatically close three (3) days after the submission deadline. Once a folder is closed, no submission will be accepted.
    - *This policy is strictly enforced unless I am informed of a documented emergency at least 24 hours before the submission deadline. All health problems should be verified by the*

university first: https://offices.depaul.edu/student-affairs/support-services/academic/Pages/absence-notification.aspx.

- For Group Project Presentation and Report, NO late submission will be accepted, because they represent the final exam in this course
- It is students' responsibility to know when the assignments are due (see the course schedule, at the end of this document).

ACADEMIC INTEGRITY AND PLAGIARISM

- There will be **ZERO tolerance** for any type of plagiarism in this course.
- The use of others' publication, software and/or web content (text, graphics, codes) is regarded as plagiarism without giving credit.
- When you directly quote someone's work, you must put it in quotation marks followed by its reference.
- The use of materials prepared for purposes other than this course needs the instructor's prior permission.
- Please familiarize yourself with the university's academic integrity policy: http://academicintegrity.depaul.edu.

CHANGES TO SYLLABUS

This syllabus is subject to change as necessary during the quarter. If a major change occurs, it will be addressed during class and posted via Announcements in D2L.

ONLINE COURSE EVALUATIONS

- Evaluations are a way for students to provide valuable feedback regarding their instructor and the course. Detailed feedback will enable the instructor to continuously tailor teaching methods and course content to meet the learning goals of the course and the academic needs of the students.
- The evaluations are anonymous; the instructor and administration do not track who entered what responses. A program is used to check if the student completed the evaluations, but the evaluation is completely separate from the student's identity. Since 100% participation is our goal, students are sent periodic reminders over three weeks. Students do not receive reminders once they complete the evaluation.
- Students will complete the course evaluation online in Campus Connect.

ACADEMIC POLICIES

- All students are required to manage their class schedules each term in accordance with the deadlines for enrolling and withdrawing as indicated in the University Academic Calendar.
- Information on enrollment, withdrawal, grading and incompletes can be found at: http://www.cdm.depaul.edu/Current%20Students/Pages/PoliciesandProcedures.aspx

CIVIL DISCOURSE

DePaul University is a community that thrives on open discourse that challenges students, both intellectually and personally, to be **Socially Responsible Leaders**. It is the expectation that all dialogue in this course is civil and respectful of the dignity of each student. Any instances of disrespect or hostility can jeopardize a student's ability to be successful in the course. The professor will partner with the Dean of Students Office to assist in managing such issues.

STUDENTS WITH DISABILITIES

Students who feel they may need an accommodation based on the impact of a disability should contact the instructor privately to discuss their specific needs. All discussions will remain confidential. To ensure that you receive the most appropriate accommodation based on your needs, contact the instructor

as early as possible in the quarter (preferably within the first week of class), and make sure that you have contacted the Center for Students with Disabilities (CSD) at: csd@depaul.edu.

Lewis Center 1420, 25 East Jackson Blvd.
Phone number: (312)362-8002
Fax: (312)362-6544
TTY: (773)325.7296

TENTATIVE COURSE SCHEDULE (SUBJECT TO CHANGE)

Week	Date	Class Focus & Content	Deliverables Due at 11:59 PM (CT) (See the Due Dates below)	
1	28 Mar	1. Introduction to the Course 2. Introduction to Data Mining (Ch. 1)		
2	4 Apr	1. Knowing Your Data (Ch. 2) 2. Preprocessing Your Data (Ch. 3)	Group Members	10 Apr
3	11 Apr	**Meet at CDM 819 (Computer Lab)** 1. Knowing Your Data (Ch. 2) 2. Preprocessing Your Data (Ch. 3) **3. Lab Session 1**	Assignment 1	17 Apr
4	18 Apr	1. Introduction to Data Warehouse (Ch. 4) 2. Classification Approaches: Decision Trees (Ch. 8)	Group Project Proposal	24 Apr
5	25 Apr	1. Classification Approaches: Decision Trees (Ch. 8) 2. Classification Approaches: Model Evaluation and Selection (Ch. 8)		
6	2 May	**Meet at CDM 819 (Computer Lab)** 1. Classification Approaches: Lazy Learners (Ch. 9) 2. **Lab Session 2**	Assignment 2	8 May
7	9 May	1. Clustering Approaches: K-Means (Ch. 10) 2. Clustering Approaches: Hierarchical Clustering (Ch. 10)		
8	16 May	**Meet at CDM 819 (Computer Lab)** 1. Clustering Approaches: Hierarchical Clustering (Ch. 10) **2. Lab Session 3**	Assignment 3	22 May
9	23 May	1. Applied Clustering Research Examples 2. Outlier Detection (Ch. 12) (if time permits)	1. Assignment 4 2. Group Project Presentation File (*.ppt file for on-campus students and .mp4 file for online students*)	29 May
10	30 May	Students' Presentations of Group Projects		
11	6 June	Group Project Report (Exam Week: **No Class**)	Group Project Report	6 June