

# CSC314: Data Mining (Revised for online Meetings)

Dr. Robert Lowe

Spring (March 30 – May), 2020

E-mail: [robert.lowe@maryvillecollege.edu](mailto:robert.lowe@maryvillecollege.edu)

Office Hours:

MWF 1:00PM – 2:00PM

TR 3:00PM – 4:00PM

Office Phone: N/A

Synchronous Hour: R 11:00 – 12:15

Office: Zoom Room (See Blackboard)

Primary Contact Point: The Tartan

---

## Syllabus Revisions

Normally, a syllabus cannot be revised in any substantial way. Schedules and activities may adjust, but the overall layout of a course should be settled by the first day of classes. This semester, however, is anything but normal. We are shifting from a traditional in-person class format to an online course for our final few weeks of this semester. As such, this requires a more substantial change to the syllabus. This document contains the revised format of the class. Grading and expectations for the period prior to March 30, 2020 remain in full force. The grading categories will not change, but the requirements to receive credit from this time forward will change. The end result is a hybrid course that met in person for the most of the semester and is now online.

Please read this syllabus carefully. This document specifies how we will proceed for the remainder of the semester.

## Course Description

Data mining is concerned with the extraction of information from large amounts of data. This project-based course introduces the concepts, issues, tasks and techniques of data mining. Topics include data preparation and feature selection, classification, clustering, evaluation and validation, and data mining applications.

## Required Materials

- *Data Mining and Analysis – Fundamental Concepts and Algorithms* by Mohammed J. Zaki and Wagner Meira, Jr.

<http://www.dataminingbook.info>

- A shell account on cs.maryvillecollege.edu (alternatively, you could install R-Studio locally on your own computer).

## Prerequisites

CSC313 and MTH321

## Course Goals

- Learn and overcome the problems in working with real world data.
- Learn the difference between good and bad data mining practices.
- Explore feature extraction and dimension reduction.
- Use statistical models and machine learning algorithms to interpret data.
- Learn the R programming language.

## Course Structure

### Methods of Instruction

This course will meet online for the remainder of this semester. The primary point of contact will be via The Tartan (aka blackboard). In an online course, material is usually split between synchronous and asynchronous modes of instruction.

### Synchronous Instruction

Synchronous instruction activities are activities where the entire class meets at the same time and performs the activity together. This is, of course, difficult in an online environment. We will therefore rely more on asynchronous activities, which are outlined in the next section. However, synchronous instruction is needed because it affords you with the best ability to ask questions, and it allows you to present to the class. **We will have one weekly synchronous session during our normal class time (11:00 – 12:15) every Thursday, beginning April 6.** Synchronous activities will include:

- R-Studio Demonstrations
- Question and Answer Sessions
- Class Presentations

## Asynchronous Instruction

Asynchronous instruction is what makes online learning desirable to so many students! However, it is also the biggest disadvantage. These are course materials which will be posted online, which you can explore at your own pace on your own schedule. That last bit is the key, for if you do not schedule time to actually avail yourself of these things, you will not look at them and you will perform poorly on the final portion of the course. My recommendation is to use these during the normal class periods, and to help you along I have strict due dates on completing asynchronous activities. These activities and materials include:

- Assigned Readings
- Homework Assignments
- Recorded Lectures
- Online Discussions

## Grading

This course is graded using a weighted average among four categories. The assignments within each category are equally weighted and are all graded out of 100 points. Hence your final numeric grade is computed by finding the average of each category, and then multiplying them by the corresponding weight. The weights for each category are as follows:

Category	Weight
Homework	40%
Projects	60%

Letter grades will be assigned according to the following scale:

A+	96.7–100%	B+	86.7–90%	C+	76.7–80%	D+	66.7–70%	F	less than 60 %
A	93.3–96.7%	B	83.3–86.7%	C	73.3–76.7%	D	63.3–66.7%		
A-	90–93.3%	B-	80–83.3%	C-	70–73.3%	D-	60–63.3%		

## Assessments

The standards of assessment in each grading category will be as follows.

### Homework (40% of the final grade)

Homework will be assigned at various points throughout the semester. Most of the homework assignments will involve producing documents using R notebooks. You will turn in printouts of these notebooks.

## Projects (60% of the final grade)

There will be three course projects. These will be completed in groups of 3-4 students, and each will result in a written report and a presentation. Your group will receive a collective grade on the projects. Each project is weighted equally.

Note that while your group may elect to have a single presenter on presentation days, the entire group must be present for the presentation. Anyone who does not attend on a presentation day will receive a reduced grade on the project.

## Schedule

This is the tentative schedule for the remainder of our course.

### March 30 – May 2020

Su	Mo	Tu	We	Th	Fr	Sa
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2

- **March 30 – April 3**

- Representative-based Clustering
- Read Chapter 13
- **Synchronous Session Thursday April 2** – Introduction to Project 2 (*Due April 16*)

- **April 6 – April 10**

- Hierarchical Clustering
- Read Chapter 14
- **Synchronous Session Thursday April 9** – Clustering in R

- **April 13 – April 17**

- Clustering Validations
- Read Chapter 17
- **Synchronous Session Thursday April 16** – Project 2 Presentations and Introduction to Project 3 (*Due May 1*)

- **April 20 – April 24**

- Support Vector Machines
- Read Chapter 21

- Synchronous Session Thursday April 23 – SVM in R
  - April 27 – May 1
    - Synchronous Session Tuesday April 28 – Project Questions and Closing Remarks
    - Synchronous Session May 1 3:30PM - Project 3 Presentations
-

## **Course Policies**

### **Late Policy**

No late work will be accepted under any circumstances (except as mercy and decency may dictate in extremely rare events).

### **Extra Credit**

No extra credit will be given under any circumstances.

### **Excused Absences**

In some cases, absences may be excused. These include:

- School Sanctioned Events (Sports, Concerts, etc.)
- Severe Illness
- Family Emergencies
- Court Appearance / Jury Duty

In the case of a school event, notice must be given at least one week prior to the absence. The notice must include a signed note from the faculty or staff member in charge of the event. This note must be given in physical form, electronic notes will not be accepted. In the case of illness, a doctor's note is required. Note that except in extreme circumstances, doctor's appointments do not qualify as a valid reason to miss a class. Please be respectful of the other students, and schedule appointments during your free time.

Family emergencies will require some form of proof. Where possible, you must give advance notice of missing a class. The exception to this would need to be fairly severe, and hopefully it will not come up. For court appearances and/or jury duty, you must provide a copy of your summons. You may redact any details you wish, save for the actual date and time of your appearance. Court appearances must be cleared at least one week in advance.

### **Making Up Excused Absences**

Should you be in a situation in which you receive an excused absence, this in no way will extend your due dates (excepting extreme emergencies). You must make up any test at a designated time prior to your excused absence. Also, homework or projects must be submitted prior to the class period in which they are due.

### **Communication and Extra Help**

You are always welcome at office hours for help with any questions you may have about the course. For help at other times during the day, stop by or call my office to see if I'm available. You can also contact me by email, but often I can better help you face to face and may respond with

a request that you come to see me. Note that I do not typically respond to email between 5 p.m. and 8 a.m. You may make appointments to see me at other times if your schedule does not permit you to attend my office hours.

### **Plagiarism and Cheating**

You are expected to do your own work. Never submit the work of others, never give unauthorized assistance to others, do not use unauthorized aids during exams, and do not ask for help from other faculty members without the approval of your professor. Plagiarism and cheating are serious offenses that will not be tolerated. Explanations regarding these offenses and how they are handled can be found in the MC Student Handbook at

<https://www.maryvillecollege.edu/academics/catalog/handbook/section-nine/>.

You are expected to read and understand these policies. Offenses on specific assignments, quizzes, or exams will result in a score of 0 on the relevant assignment, and a letter of censure will be placed in your college file. Repeat offenses will result in further disciplinary action, including the possibility of failing the course.

### **Students with Disabilities**

Any student who feels s/he may need learning or physical accommodation(s) based on the impact of a disability should contact Services for Students with Disabilities to discuss your specific needs. Please contact 981-8124 to coordinate reasonable accommodations for students with documented disabilities. The Disability Services office is located in the Learning Center in the basement of Thaw Hall. Undocumented disabilities will not be accommodated.