CS/STAT 5525: Data Analytics (3 credits, CRNs: 13294, 13295, 19845, 19846)

Department of Computer Science, Virginia Tech

Instructor: Anuj Karpatne (email: karpatne@vt.edu, http://people.cs.vt.edu/karpatne)
Teaching Assitants: Jie Bu (email: jayroxis@vt.edu), Arka Daw (email: darka@vt.edu), Mohannad Elhamod (email: elhamod@vt.edu), Md Abdullah Al Maruf (email: marufm@vt.edu)

Class Type: Online Course with Synchronous Meetings on Zoom
Class Timings: TR: 2:00 pm - 3:15 pm Eastern,
Zoom URL: https://virginiatech.zoom.us/j/81169497164?pwd=ZysrZllKbVpjSFdqdjVOZ0RGY1pDdz09
(passcode: 5525isfun)

Instructor Office Hours: TR: 3:30 pm - 5:00 pm Eastern,
Zoom URL: https://virginiatech.zoom.us/j/89688779440?pwd=Wml0bURSQU5lZWFwZTZBZ3NvTEJ0Zz09
(passcode:5525isfun)

TA Office Hours: TBD,
Zoom URL: TBD

Course Website: http://people.cs.vt.edu/karpatne/teaching/5525-s21/

Course Overview: The last decade has seen an explosive growth in database technology and the amount of data collected. This has created an unprecedented opportunity for "data mining" (also referred to as data analytics), which is the process of efficient supervised or unsupervised discovery of interesting and useful information from collections of data. Some of the common tasks in data mining are classification, clustering, the discovery of association rules/sequential patterns, and anomaly detection. Data mining has seen several successful applications in diverse domains such as healthcare, economics, internet advertising, social sciences, and environmental studies. This course will give a rapid and vigorous introduction to the field of data mining, as well as provide extensive hands-on experience via class projects. All course activities will be conducted online.

Learning Aims: By the end of the course, students will:

- Be well-versed with common data mining problems, concepts, and algorithms
- Be able to compare and contrast different data mining algorithms and identify their strengths and limitations in varying problem settings
- Gain practical understanding of data mining algorithms through course projects

Textbook (optional): Introduction to Data Mining (2nd Ed.), 2018, P.N. Tan, M. Steinbach, A. Karpatne, and V. Kumar. Visit the book webpage at www.cs.umn.edu/~kumar/dmbook

Background Required: General background in the following:

- Computational algorithms (introductory level course in algorithms) [Link] covering
  - time and space complexity: Big-O notations
- Calculus and linear algebra topics [Link] covering
  - univariate and multivariate derivatives and integrals
  - vectors, orthogonal vectors, dot product, cross product,
  - matrix multiplication, determinant of a matrix,
  - eigen values, eigen vectors
- Probability topics [Link] covering
  - random variable, probability, probability distribution,
  - mean, variance, standard deviation, expected values
- Python programming [Link1, Link2, Link3]

Learning Activities: We will be making use of the following learning activities:

- **Lectures:**
  Lectures will cover data mining concepts from the textbook and other materials using illustrative examples and in-class discussions. Lecture slides will be posted a day in advance on the Canvas Page. Zoom recordings of the lecture will be posted on Canvas within 2 days.
- **Homework Assignments:**
  There will be 5 homework assignments covering different topics in the course. The questions in the homework assignments will be designed to test your conceptual understanding of data mining topics and your ability to reason how a data mining algorithm would perform in a given situation. They may involve simple arithmetic calculations, but they will not require programming skills. Only a subset of the questions will be graded, but the students are required to provide answers to all. There will also be some additional "practice questions" that will not be graded, but you are encouraged to try them out and provide their answers. Solutions to homework assignments will need to be submitted electronically on the Canvas page. Solutions for assignments will be posted a week after their due date.
- **Projects:**
  There will be 3 projects in the course that will require you to run data mining algorithms on given data using Python. You will be developing and running Python code (starting from pre-packaged software), analyzing data mining outputs, and compiling your analyses in the form of a report. Project reports and output from experiments will need to be submitted electronically on the Canvas Page.
- **Exams:**
  We will have a midterm exam and a final exam in the course. They will only test your conceptual understanding of data mining problems and algorithms and not your number crunching skills. You will not need a calculator. You are encouraged to keep your answers brief and to the point. If you feel a question is ambiguous, you are encouraged to state your assumptions in your answer. Both exams will be conducted as online quizzes on Canvas with allotted time limits (75 minutes for midterm and 120 minutes for final exam) and will be available for 24 hours from their time of release.

Tentative Outline of Course Topics and Exams:

Week	Topic
Jan 19	Introduction to Data Mining (Ch1)
Jan 21 – Jan 28	Data Exploration (Ch2)
Feb 2 – March 11	Classification (Ch3 and Ch4)
March 11	**Midterm Exam**
March 16 – March 26	Association Analysis (Ch5 and Ch6)
March 30 – April 9	Clustering (Ch7 and Ch8)
April 13 – April 23	Anomaly Detection (Ch9)
April 27 – May 4	Avoiding False Discoveries (Ch10)
May 12	**Final Exam**

Tentative Schedule of Homework Assignments and Projects:

Homework	Project	Posted	Due	Topic
Homework 1		Jan 21	Feb 4	Data
Homework 2		Feb 4	Feb 25	Classification
	Project 1	Feb 11	March 11	Classification
Homework 3		Feb 25	March 18	Classification
Homework 4		March 18	April 8	Association Analysis
	Project 2	March 25	April 15	Association Analysis
Homework 5		April 8	April 29	Clustering, Anomaly Detection, Avoiding False Discoveries
	Project 3	April 15	May 6	Clustering

Late Submission Policy: Late submissions to homework assignments and projects would receive the following penalties as percentages of their earned scores.

Less than 24 hours	10 %
24 – 48 hours	30 %
48 – 72 hours	60 %
More than 72 hours	100 %

Workload and Grading Scheme:

Five homework assignments	30%
Three "hands on" projects	21%
Mid-term exam	21%
Final exam	28%

Grade	Aggregate Score Range
A	93 – 100
A-	87 – 92
B+	80 – 86
B	75 – 79
B-	70 – 74
C+	65 – 69
C	60 – 64
C-	55 – 59
D+	50 – 54
D	45 – 49
D-	40 – 44
F	< 40

Note: The cutoffs for individual grades may be lowered depending on the relative performance of the class.

Policy for Disputing Grades: If a student feels that there has been an error in grading a homework assignment or a project, they need to bring this up with the TAs within two weeks of the return of the graded assignment.

Zoom Best Practices: We will be using Zoom for conducting all class activities and office hour discussions. The Zoom URLs are provided at the top of this document and can also be accessed from the Zoom tab on the left panel of the Canvas page of the class. Please familiarize yourself with Zoom and student tips for remote learning (see: https://tutorials.tlos.vt.edu/index/zoom.html and https://teaching.vt.edu/OurServices/StudentTips.html). You should keep your video turned on during the class to remain attentive and compensate for the lack of physical interactions in an online environment, unless restricted by low internet bandwidth. You may keep your audio muted unless you have a question to ask or need to respond to an on-going discussion to avoid interference and feedback. You can also type your question in Zoom's chat interface or provide nonverbal feedback and express opinions by clicking on icons in the Participants panel at any point during the lecture. We will be using Zoom Breakout rooms feature during office hours to facilitate one-on-one or group interactions of the instructor and TAs with students.

Communications and Feedback: We will be using Canvas Announcements as our preferred mode of communication to notify any changes to the class schedule and activities, so please ensure that your Canvas Notification Preferences are set to notify you (typically via email) when an Announcement has been posted. We will also be using Piazza to facilitate after-class discussions (please sign up for the Piazza page for this class here: https://piazza.com/vt/spring2021/cs_5525_13294_202101). This system is highly catered to getting you help fast and efficiently from classmates, the TA, and the instructor. Regular feedback will be provided to students on all submissions and class participation. At any time during the course, if you are facing any difficulties to meet the course deliverables or would like to discuss any concerns, you are welcome to talk to the instructor during office hours, over email, or using the following link for submitting anonymous feedback: https://virginiatech.qualtrics.com/jfe/form/SV_a32n1EinodJZI2i.

Academic Integrity: The tenets of the Virginia Tech's Honor Codes will be strictly enforced in this course, and all assignments shall be subject to the stipulations of the Undergraduate and Graduate Honor Codes. For more information on the Graduate Honor Code, please refer to the GHS Constitution at http://ghs.graduateschool.vt.edu. All paper reviews, project reports, and other submissions must represent your own individual effort. Students are encouraged to consult with one another about project design and evaluation issues, whether performed individually or in groups, as long as the individual submissions represent their individual efforts. Be particularly careful to avoid plagiarism, which essentially means using materials (ideas, code, designs, text, etc.) that you did not create without giving appropriate credit to the creator (using quotation marks, citations, comments in the code, link to URL, etc.). We will also adhere to Virginia Tech's Principles of Community for all in-class discussions and activities, to maintain a safe, welcoming, and respectful environment for every student in the class. For more information, see: https://www.inclusive.vt.edu/Initiatives/vtpoc0.html.

Accommodations for Students with Special Needs: Students with special needs will be provided additional resources and materials to aid in their learning. Mode of communication during the class will be adjusted in lieu of the respective needs of the student. Please discuss your requirements with the instructor so that we can work together to make a comfortable environment for everyone. Please see: https://www.ssd.vt.edu/ for more information. If you have an emergency medical information, please let me know privately as soon as possible.