CS/STAT 5525: Data Analytics (3 credits, CRN: 92712, 92713)

Department of Computer Science, Virginia Tech

Instructor: Anuj Karpatne (email: karpatne@vt.edu, http://people.cs.vt.edu/karpatne)
Teaching Assistants: Jixiang Fan (email: jfan12@vt.edu), Sikiru Adewale (email: asikiru@vt.edu)

Class Timings: TR: 12:30 pm - 1:45 pm Eastern

Class Modality: Online on Zoom till Feb 3*, Face-to-Face Lectures from Feb 8
***Note:** Out of an abundance of caution for the growing number of COVID-19 cases in the start of the Spring semester, we have decided to temporarily switch to online mode of instruction for CS (STAT) 5525 till February 3. We be meeting over Zoom (and NOT physically in the classroom) till Feb 3. We will resume to in-person teaching in the classroom starting February 8.

Zoom URL for Classes (till Feb 3):
https://virginiatech.zoom.us/j/84183325063?pwd=bHE4QUJ2ZGNMcTd3QmJlYUhRSlpxdz09
(passcode:5525isfun)

Classroom Location (from Feb 8): Lavery Hall (WLH) 350.

Instructor Office Hours: TR: 4:00 pm - 5:00 pm Eastern,
Zoom URL: https://virginiatech.zoom.us/j/88119023526?pwd=WXl4UkhSQXo3czBFMHBjR25HV3B5Zz09
(passcode:5525isfun)

TA Office Hours:

Monday: 12:00 pm – 1:00 pm (Sikiru Adewale)
Zoom URL: https://virginiatech.zoom.us/j/84258747470?pwd=K3JGWHBxK1lhRStLeVVDQ0hUVkZrdz09
(passcode:5525isfun)

Wednesday: 2:00 pm – 3:00 pm (Jixiang Fan and Sikiru Adewale)
Zoom URL: https://virginiatech.zoom.us/j/85451891256?pwd=djhhQnNFSXZpakI0UUh6eHYxbEV3UT09
(passcode:5525isfun)

Friday: 1:00 pm – 2:00 pm (Jixiang Fan)
Zoom URL: https://virginiatech.zoom.us/j/87088956375?pwd=WCtQMitON0JCMGNFTXpCdDhBWEJOZz09
(passcode:5525isfun)

Course Overview: We are continuously seeing an explosive growth in the amount of data collected across all walks of life. This has created an unprecedented opportunity for "data mining" (also referred to as data analytics or machine learning), 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 (recommended): 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 for additional resources.

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 2 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 18	Introduction to Data Mining (Ch1)
Jan 20 – Jan 27	Data Exploration (Ch2)
Feb 1 – March 15	Classification (Ch3 and Ch4)
March 18	**Midterm Exam**
March 17 – March 29	Deep Learning
March 31 – April 12	Clustering (Ch7 and Ch8)
April 14 – April 26	Anomaly Detection (Ch9)
April 28 – May 3	Avoiding False Discoveries (Ch10)
May 7	**Final Exam**

Tentative Schedule of Homework Assignments and Projects:

Homework	Project	Posted	Due	Topic
Homework 1		Jan 18	Feb 2	Data
Homework 2		Feb 2	Feb 23	Classification

	Project 1	Feb 16	March 30	Classification
Homework 3		Feb 23	March 23	Classification
Homework 4		March 23	April 13	Deep Learning + Clustering
Homework 5		April 13	May 4	Clustering + Anomaly Detection + Avoiding False Discoveries
	Project 2	March 30	April 27	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	40%
Two "hands on" projects	16%
Mid-term exam	20%
Final exam	24%

Grade	Aggregate Score Range
A	94 – 100
A-	87 – 93
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.

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 (you can find it on the left panel of the Canvas page of the class). 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, on Piazza, over email, or using the following link for anonymous feedback: https://virginiatech.qualtrics.com/jfe/form/SV_a32n1EinodJZI2i.

Office Hours Modality: 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. Please familiarize yourself with Zoom using https://tutorials.tlos.vt.edu/index/zoom.html. If you would like to schedule an in-person meeting with the instructor or the TAs, please request an appointment over email or Piazza.

Health and Safety During Covid Pandemic: Virginia Tech is committed to protecting the health and safety of all members of its community. By participating in this class, all students agree to abide by the Virginia Tech Wellness guidelines (https://ready.vt.edu/public-health-guidelines.html#wellness). As pandemic conditions continue to evolve through the semester, these guidelines may need to change. The guidance posted by the university at VT Ready website should represent the most up-to-date requirements of the university and should be checked periodically for changes.

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.