CSCE 5300: Introduction to Big Data and Data Science (Summer 2024)

Instructor Information

Name: Ravi Vadapalli, Ph.D. Pronouns: he/him, Email: Ravi.Vadapalli@unt.edu, 940-369-6046
Virtual Office: By appointment. Available on Teams
Teaching Assistant(s): Vishpendra Chahar, vishpendrachahar@my.unt.edu

Course Requirements.

Computer Configuration and Requirements:

- A working computer (desktop or laptop) with at least two CPU-cores, 8GB RAM, and at least 10GB hard drive space available.
- Ability to install software packages on your computer. This may require admin privileges on your computer. This is necessary to be successful in the course! NO exceptions.
- Loner computers (from library or other campus IT outlets) may not provide you the flexibility to install software packages.

Software and Environment: Jupyter Notebook or PyCharm based python tools only! All assignments and course related work must be accomplished using Jupyter notebook or PyCharm based python programming environment. **Use of google colab or other environments NOT PERMITTED! No exceptions!**

Course Environment. All course related materials are available on Canvas. You will submit Assignments/Quizzes/Exams in Canvas only! No other forms of submissions entertained.

- **Assignments.**
  - Adhere to the allowed file formats (ipynb, py, pdf, etc.) for the assignment.
  - All submissions must be in canvas only.
  - Work will not be accepted after deadline, and
  - no other forms of submission (such as sending email attachments to instructor or TAs) is not permitted. Such submissions will be ignored and will not be considered for grading!
- **Quizzes.**
  - Adhere to the allowed file formats (ipynb, py, pdf, etc.) for the quiz questions where file submissions are required.
  - Work will not be accepted after deadline, and
  - All submissions are in canvas only! No other forms of submissions entertained.
- Exam(s).
  - The exam is fully objective.
  - Read the instructions and follow them.

Plagiarism will be taken very seriously, and you will be reported to UNT academic integrity. Check 06.003 Student Academic Integrity.pdf for more information.

How to avoid plagiarism?

- Do not copy the entire question from assignment or quiz into comments. When multiple students do this, it triggers plagiarism although it's not a legitimate issue.
- Do your work yourself and be ready to defend it when asked.

Course Description, Structure, and Objectives

Introduction to Big Data and Data Science includes an overview of the field, technical challenges, computational approaches, practical applications, structured and unstructured data processing, empirical methods in computer science, data analytics and learning, data visualization, privacy, and ethics. Emphasis will be on Big Data and its effect on other topics within Data Science, its technical characteristics, and state-of-the-art Big Data analytics architectures and tools.

Upon successful completion of this course, you will be able to:

1. Assimilate a holistic view of data science and its impact in science, engineering, and industry (aka markets)
2. Use big data tools to obtain, assess, and prepare data for analysis.
3. Gain hands-on experience with computing architectures, parallel and distributed computing environments.
4. Be comfortable in using Jupyter Notebook, Python, and PySpark programming tools and command line environments for disparate data analytics, and
5. Apply the knowledge on real-world datasets.

Required/Recommended Materials

There are no required prerequisites for this course. However, to be successful in this course you will need to review, study, and understand all materials (readings, videos, tutorials, assignments, and exams) made accessible online and posted on the course calendar on the respective class day at the latest. Readings will all be dense, so please search for additional resources (e.g., Wikipedia, coursera lectures) as needed. All attempts will be made to provide sufficient resources for everyone.

Technology requirements for courses with digital materials: This course has digital components. To fully participate in this class, students will need internet access to reference content on the Canvas Learning Management System.

How to Succeed in this Course

While I want to make myself as available as possible to each of you, I do have to place some limitations on when I can be contacted. I would prefer that most general questions go through the Q & A forum in the Discussion Board area. If you have a general question about the course or assignments, please post it there. Either I will answer it, or, one of your classmates will. This way we can all benefit from questions asked, and they can be answered in a venue that the whole class can see. You may also want to find someone in class to be a "buddy" with. This will give you at least one other person who you can email with questions.

If you have a private question, please contact me via email and I will respond within 24 hours on weekdays (usually sooner). Please do not expect a response over the weekend. Please use my phone number as a last resort - but also, please use it if you need to!

Normally, I will return feedback on all written assignments within 1 week of the due date. However, if I see that I will be unable to return your feedback that quickly I will post an Announcement to let everyone know when it can be expected. You can expect to see me participate in the discussion board after all student original posts have been posted - usually on the Friday of the first week of the module.

The University of North Texas makes reasonable academic accommodation for students with disabilities. Students seeking reasonable accommodation must first register with the Office of Disability Access (ODA) to verify their eligibility. If a disability is verified, the ODA will provide you with a reasonable accommodation letter

to be delivered to faculty to begin a private discussion regarding your specific needs in a course. You may request reasonable accommodations at any time; however, ODA notices of reasonable accommodation should be provided as early as possible in the semester to avoid any delay in implementation. Note that students must obtain a new letter of reasonable accommodation for every semester and must meet with each faculty member prior to implementation in each class. Students are strongly encouraged to deliver letters of reasonable accommodation during faculty office hours or by appointment. Faculty members have the authority to ask students to discuss such letters during their designated office hours to protect the privacy of the student. For additional information, refer to the Office of Disability Access website (http://www.unt.edu/oda). You may also contact ODA by phone at (940) 565-4323.

How to Reach Me?

Connect with me through email or Teams. During busy times, my inbox becomes rather full, so if you contact me and do not receive a response within two business days, please send a follow up email. A gentle nudge is always appreciated. Office hours offer you an opportunity to ask for clarification or find support with understanding class material. Come visit me! I encourage you to connect with me and/or my TA for support. Additional office hours, in person and virtually, will be offered as the semester concludes. Your success is our goal.

Supporting Your Success and Creating an Inclusive Learning Environment

I value the many perspectives students bring to our campus. Please work with me to create a classroom culture of open communication, mutual respect, and inclusion. All discussions should be respectful and civil. Although disagreements and debates are encouraged, personal attacks are unacceptable. Together, we can ensure a safe and welcoming classroom for all. If you ever feel like this is not the case, please stop by my office and let me know. We are all learning together. We will discuss our classroom's habits of engagement and I also encourage you to review UNT's student code of conduct so that we can all start with the same baseline civility understanding (Code of Student Conduct) (https://deanofstudents.unt.edu/conduct)

Assessing Your Work

Course Activities & Assessments (100 points total)

- Software installs and End of the Course survey (1 pt)
- Quizzes (7 @ 5 points each, 35 points total)
- Assignments (2 @10 points each, 20 points total)
- Project (15 points)
- Exam (29 points total)

Grading

- A: 90-100% (Outstanding, excellent work. The student performs well above the minimum criteria.)
- B: 80-89% (Good, impressive work. The student performs above the minimum criteria.)
- C: 70-79% (Solid, college-level work. The student meets the criteria of the assignment.)
- D: 60-69% (Below average work. The student fails to meet the minimum criteria.)
- F: 59 and below (Sub-par work. The student fails to complete the assignment.)

Grade-related Policies

Late Work

I will not accept late work in this course. All work turned in after the deadline will receive a grade of zero unless the student has a university-excused absence (Links to an external site.) and provides documentation with 48 hours of the missed deadline.

Turnaround Time

We make every effort return your graded assignments within one week after the assignment due date. Delays will be updated during class or via canvas announcements.

Grade Disputes

You are required to wait 24 hours before contacting me to dispute a grade. Within that time, I expect that you will review the assignment details and reflect on the quality of the work you turned in. If you would still like to meet, email me to set up a meeting (I cannot discuss grades over email). You should come to our scheduled virtual meeting with specific examples that demonstrate that you earned a higher grade than you received. If you miss your scheduled meeting, you forfeit your right to a grade dispute. If you do not contact me to schedule a meeting within seven days of receiving your grade, you also forfeit your right to a grade dispute.

Course Requirements/Schedule (Tentative). May 20 – July 26, 2024

Provided below tentative list of topics along with assignments. This list can change during the course and will be notified of such changes wherever possible.

Unit 0: Course Overview, Programming Environment, and Course Expectations

Week	Topic	Assignment	Points Possible	% of Final Grade
Pre-Course Prep.	• *Syllabus, Grading Policies* • *Course Overview and Expectations* • *Programming Environments*	*Install and Test Jupyter Notebook* *Install and Test PyCharm*	*0.5*	*0.5*

Unit 1: Introduction to Python, Data Structures, DataFrames, and Visualization

Week	Topic	Assignment	Points Possible	% of Final Grade
May 27-31	• *Intro. python programming* • *Data Structures and Object-Oriented Programming*	*Quiz-1*	*5 pts.*	*5%*
June 3-7	• *Pandas DataFrames* • *File Handling and Preparing Data*	*Quiz-2*	*5 pts.*	*5%*
	• *Data Visualization (Matplotlib, Seaborn, SHAP), Heatmap Analysis*	*Assignment-1*	*10 pts.*	*10%*

Unit 2: Data Science, Linear Regression for Machine Learning

Week	Topic	Assignment	Points Possible	% of Final Grade
June 10-14	• *Statistics for Data Science* • *P-Value and Hypothesis Testing* • *Introduction to Machine Learning*	*Quiz-3*	*5 pts.*	*5%*

June 17-21
 	
June 24-28
	• *Linear Regression Models for ML*
	  ○ *Polynomial Regression*
	  ○ *Logistic Regression*
	  ○ *K-Nearest Neighbor ML for image processing*
	• *ML Accuracies by Confusion Matrix*
		Quiz-4
		
		Assignment-2
			5 pts.
			
			10 pts.
				5%
				
				10%

Unit 3: Big Data Processing: Hadoop Distributed Computing and PySpark DataFrames

Week	Topic	Assignment	Points Possible	% of Final Grade
June 24-28	• *Introduction to Big Data* • *Hadoop Distributed Computing* • *PySpark DataFrames*	*Quiz-5*	*5 pts.*	*5%*

Unit 4: Dimensionality Reduction and Feature Selection by Visualization

Week	Topic	Assignment	Points Possible	% of Final Grade
July 1-5	• *Principal Component Analysis* • *Hyperparameter Optimization* • *XGBoost, SHAP Plots*	*Quiz-6*	*5 pts.*	*5%*
July 8-19	*Case Study-1 (By Example)* *Machine Learning Project*	*Quiz-7* *Assignment-3*	*5 pts.* *15 pts.*	*5%* *15%*

Final Exam:

July 24	*Exam (Whole Syllabus)*		*29 pts.*	*29%*
	SPOT Survey (end of course)		*0.5 pts.*	*0.5%*
	TOTAL Points		**100 pts.**	**100%**

Eagle Alert

Students will be notified by Eagle Alert if there is a campus closing that will impact a class and note that the proposed course schedule is subject to change. For more information on emergency notification and procedures, please review Emergency Notifications and Procedures Policy (PDF).

Academic Integrity

You are expected to adhere to the university's Academic Integrity Policy (PDF) (https://policy.unt.edu/sites/default/files/06.049_Standard%20Syllabus%20Policy%20Statements_supplement.pdf).

Syllabus Policy

Grades are based on mastery of the content. As a rule, I do not grade on a "curve" because that is a comparison of your outcomes to others. I do, however, encourage you to find opportunities to learn with and through others. Explore Navigate's Study Buddy (https://navigate.unt.edu) tool to join study groups. Maximize your learning with our coaching staff at the Learning Center. Focus on areas where you are struggling in this course by attending scheduled study group sessions with me the week before each exam. Forward together!

or

Every student in my class can improve by doing their own work and trying their hardest with access to appropriate resources. Students who use other people's work without citations will be violating UNT's Academic Integrity Policy. Please read and follow this important set of guidelines for your academic success (https://policy.unt.edu/policy/06-003). If you have questions about this, or any UNT policy, please email me or come discuss this with me during my office hours.

Attendance and Participation

This is 100% online course. The course materials along with quizzes and assignments will be made available in a timely manner. TAs will grade your submissions and will be available to answer any questions you may have. Research has shown that students who take time to follow the materials are more likely to be successful. While attendance is not required for this course, feel free to reach out the instructor or course TAs should you have questions or concerns. See Canvas pages for TA office hours and how to contact them.

Please note: the following components are required for every UNT syllabus: Attendance expectations and consequences, course competencies/assignments/requirements (including each major assignment and exam, subject matter of each session, lists of any readings), final exam date/time/location, emergency notification and procedures, academic integrity expectations and consequences.