Topics in Human-Centered Data Science: Explainable AI (XAI)
Spring 2024, I 320D
Course Syllabus

Instructor: Gutierrez, Louis, Phd
Contact: louis.gutierrez@austin.utexas.edu
Class Meets: Tue/Thurs, 5-6:30PM
Classroom: CBA 4.344

Course Description

This course provides an Introduction to Explainable AI (XAI) through practical applications and real-world examples. Students will gain a basic proficiency in interpreting and explaining the decisions of ML and AI systems, in a transparent and understandable manner to humans. The course will cover various XAI techniques and algorithms, including rule-based models, feature importance analysis, model-agnostic approaches, and post-hoc explanations.

Learning Outcomes

By the end of the course, students will be able to:

1. Understand what Explainable AI is, its scope, and impact on various domains.
2. Understand Global vs Local Explanations and their applications.
3. Identify and evaluate the most used XAI techniques and algorithms.
4. Use Python to apply Explainer algorithms/methods and Interpret the results
5. Critically evaluate and contextualize the performance and reliability of Explanations, and identify their limitations and biases.

Course Topics [Subject to Change]

Theme	Week	Detailed Topics
Introduction to Explainability	1	Course Overview and Introduction Explaining Explainable AI
		Overview of Python Data Stack ML and AI Refresher
	2	Defining Explainability Overview of Explanations
		Known Issues in Explainability Explainability Case Study

Explaining Structured Data	3	Pre-model Explainability Partial Dependence Plots
		Permutation Feature Importance
	4	Intro to Shapley
		More on Shapley: Tree Models and other applications
	5	Rule Based Methods: Anchors Counterfactual Explanations
		Summary of Structured Data Explainers
Explaining Images	6	Integrated Gradients XRAI
		Grad-Cam
	7	Lime for Images
		Summary of Image Explainers
Explaining Unstructured Data	8	Pre-model Explainability on Unstructured Data
		Supervised Wrapper for Clustering Models Summary of Unstructured Data Explainers
Explaining Text	9	Lime for Text Data Shap for Text Classifiers
		Sentence Highlighters
	10	Layer Integrated Gradients Layer-Wise Relevance Propagation
		Summary of Text Explainers
Measuring Performance	11	Measuring the Performance of Explainers
		Measuring Trust and Reliability of Explanations Summary of Measuring Explainer Performance

Additional Topics in Explainability	12	Deep Explainers
		Time Series Explainers LLM, Foundation Models
	13	Feature Selection for Explainability Explainer Dashboard
		Brief survey of other popular XAI Frameworks
Wrapping Up	14	Summary of Course

Grading Summary [Subject to Change]

Assignment	Percentage	Estimated Due Date (Subject to Change)
Attendance	10%	Ongoing
Participation	10%	Ongoing
Homework 0	5%	02/06/2024
Homework 1	15%	02/22/2024
Homework 2	15%	03/07/2024
Homework 3	15%	04/04/2024
Homework 4	30%	04/25/2024

Course Credit Overview

Homework 0
This is an introductory assignment to make sure you are familiar with some of the tools we'll need later in the course.

Homework 1 - 3
These homeworks are designed to use Python to implement some of the methods and techniques used in class and interpret the results based on the output. Although successful use of these libraries and packages is needed, the emphasis is placed on interpreting and giving context to the results.

Homework 4
This homework is about bringing it all together and evaluating explanations as useful, more useful and otherwise. This assignment will be due last week of classes.

Attendance
Attendance credit is based on percentage attended classes (i.e. if you attend 90% of total classes then 9/10 points is awarded as credit). There will be a sign in sheet.

Participation
Participation credit is awarded as a blend of in-class and Canvas participation.

Late Policy
Each assignment will be penalized 10% for each class period after its due date, up to the last day of classes, barring extenuating circumstances.

Communication
No official office hours will be kept, but the following methods of communication are suggested:

- Before/After class
- Email
- Canvass

Disability & Access (D&A)
The university is committed to creating an accessible and inclusive learning environment consistent with university policy and federal and state law. Please let me know if you experience any barriers to learning so I can work with you to ensure you have equal opportunity to participate fully in this course. If you are a student with a disability, or think you may have a disability, and need accommodations please contact Disability & Access (D&A). Please refer to the D&A website for more information. If you are already registered with D&A, please deliver your Accommodation Letter to me as early as possible in the semester so we can discuss your approved accommodations and needs in this course.

Religious Holy Days
By UT Austin policy, you must notify me of your pending absence for a religious holy day as far in advance as possible of the date of observance. If you must miss a class, an examination, a work assignment, or a project in order to observe a religious holy day, you will be given an opportunity to complete the missed work within a reasonable time after the absence.

Names and Pronouns
Class rosters are provided to the instructor with the student's legal name, unless they have added a chosen name with the registrar's office. If you have not yet done so, I will gladly honor your request to address you with the name and pronouns that you prefer for me to use for you. It is helpful to advise me of any changes or needs regarding your name and pronouns early in

the semester so that I may make appropriate updates to my records and be informed about how to support you in this class.

- For instructions on how to add your pronouns to Canvas, visit this site.
- If you would like to update your chosen name with the registrar's office, you can do so here, and reference this guide.
- For additional guidelines prepared by the Gender and Sexuality Center for changing your name on various campus systems, see the Resources page under UT Resources here.

Counseling and Mental Health Center (CMHC)

Students who are struggling for any reason and who believe that it might impact their performance in the course are urged to reach out to Bryce Moffett if they feel comfortable. This will allow her to provide any resources or accommodations that she can. If immediate mental health assistance is needed, call the Counseling and Mental Health Center (CMHC) at 512-471-3515 or you may also contact Bryce Moffett, LCSW (iSchool CARE counselor) at 512-232-4449. Bryce's office is located in FAC18S and she holds "drop in" Office Hours on Wednesday from 2- 3pm. For urgent mental health concerns, please contact the CMHC 24/7 Crisis Line at 512-471-2255.

Honor Code

The University of Texas at Austin strives to create a dynamic and engaging community of teaching and learning where students feel intellectually challenged; build knowledge and skills; and develop critical thinking, creativity, and intellectual curiosity. As a part of this community, it is important to engage in assignments, exams, and other work for your classes with openness, integrity, and a willingness to make mistakes and learn from them. The UT Austin honor code champions these principles:

I pledge, as a member of the University of Texas community, to do my work honestly, respectfully, and through the intentional pursuit of learning and scholarship.

The honor code affirmation includes three additional principles that elaborate on the core theme:

- I pledge to be honest about what I create and to acknowledge what I use that belongs to others.
- I pledge to value the process of learning in addition to the outcome, while celebrating and learning from mistakes.
- This code encompasses all of the academic and scholarly endeavors of the university community.

The honor code is more than a set of rules, it reflects the values that are foundational to your academic community. By affirming and embracing the honor code, you are both upholding the integrity of your work and contributing to a campus culture of trust and respect.

Academic Integrity Expectations

Students who violate University rules on academic misconduct are subject to the student conduct process. A student found responsible for academic misconduct may be assigned both a status sanction and a grade impact for the course. The grade impact could range from a zero on the assignment in question up to a failing grade in the course. A status sanction can range from a written warning, probation, deferred suspension and/or dismissal from the University. To learn more about academic integrity standards, tips for avoiding a potential academic misconduct violation, and the overall conduct process, please visit the Student Conduct and Academic Integrity website at: http://deanofstudents.utexas.edu/conduct.

Confidentiality Of Class Recordings

Class recordings are reserved only for students in this class for educational purposes and are protected under FERPA. The recordings should not be shared outside the class in any form. Violation of this restriction by a student could lead to Student Misconduct proceedings.

Getting Help with Technology

Students needing help with technology in this course should contact the ITS Service Desk.

Content Warning

Our classroom provides an open space for the critical and orderly exchange of ideas through discussion. Some readings and other content in this course will include topics and comments that some students may find offensive and/or traumatizing. I'll aim to forewarn students about potentially disturbing content and I ask all students to help to create an atmosphere of mutual respect and sensitivity.

Sharing of Course Materials is Prohibited

No materials used in this class, including, but not limited to, lecture hand-outs, videos, assessments (quizzes, exams, papers, projects, homework assignments), in-class materials, review sheets, and additional problem sets, may be shared online or with anyone outside of the class without explicit, my written permission. Unauthorized sharing of materials may facilitate cheating. The University is aware of the sites used for sharing materials, and any materials found online that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in initiation of the student conduct process and include charge(s) for academic misconduct, potentially resulting in sanctions, including a grade impact.

Artificial Intelligence

The creation of artificial intelligence tools for widespread use is an exciting innovation. These tools have both appropriate and inappropriate uses in classwork. The use of artificial intelligence tools (such as ChatGPT) in this class:

- …is strictly prohibited. This includes using AI to generate ideas, outline an approach, answer questions, solve problems, or create original language. All work in this course must be your own or created in group work, where allowed.

- …shall be permitted on a limited basis. You will be informed as to the assignments for which AI may be utilized. You are also welcome to seek my prior-approval to use AI writing tools on any assignment. In either instance, AI writing tools should be used with caution and proper citation, as the use of AI should be properly attributed. Using AI writing tools without my permission or authorization, or failing to properly cite AI even where permitted, shall constitute a violation of UT Austin's Institutional Rules on academic integrity.
- …is permitted for students who wish to use them, provided the content generated by AI is properly cited. If you are considering the use of AI writing tools but are unsure if you are allowed or the extent to which they may be utilized appropriately, please ask."

Land Acknowledgment

I would like to acknowledge that we are meeting on the Indigenous lands of Turtle Island, the ancestral name for what now is called North America. Moreover, I would like to acknowledge the Alabama-Coushatta, Caddo, Carrizo/Comecrudo, Coahuiltecan, Comanche, Kickapoo, Lipan Apache, Tonkawa and Ysleta Del Sur Pueblo, and all the American Indian and Indigenous Peoples and communities who have been or have become a part of these lands and territories in Texas.

Carrying of Handguns on Campus

Students in this class should be aware of the following university policies related to Texas' Open Carry Law:

- Students in this class who hold a license to carry are asked to review the university policy regarding campus carry.
- Individuals who hold a license to carry are eligible to carry a concealed handgun on campus, including in most outdoor areas, buildings and spaces that are accessible to the public, and in classrooms.
- It is the responsibility of concealed-carry license holders to carry their handguns on or about their person at all times while on campus. Open carry is NOT permitted, meaning that a license holder may not carry a partially or wholly visible handgun on campus premises or on any university driveway, street, sidewalk or walkway, parking lot, parking garage, or other parking area.
- Per my right, I prohibit the carrying of handguns in my personal office. Note that this information will also be conveyed to all students verbally during the first week of class. This written notice is intended to reinforce the verbal notification, and is not a "legally effective" means of notification in its own right.

Campus Safety

The following are recommendations regarding emergency evacuation from the Office of Emergency Management, 512-232-2114:

- Students should sign up for Campus Emergency Text Alerts at the page linked above.
- Occupants of buildings on The University of Texas at Austin campus must evacuate buildings when a fire alarm is activated. Alarm activation or announcement requires exiting and assembling outside.

- Familiarize yourself with all exit doors of each classroom and building you may occupy. Remember that the nearest exit door may not be the one you used when entering the building.
- Students requiring assistance in evacuation shall inform their instructor in writing during the first week of class.
- In the event of an evacuation, follow the instruction of faculty or class instructors. Do not re-enter a building unless given instructions by the following: Austin Fire Department, The University of Texas at Austin Police Department, or Fire Prevention Services office.
- For more information, please visit the Office of Emergency Management.

Title IX Disclosure

Beginning January 1, 2020, Texas Education Code, Section 51.252 (formerly known as Senate Bill 212) requires all employees of Texas universities, including faculty, to report to the Title IX Office any information regarding incidents of sexual harassment, sexual assault, dating violence, or stalking that is disclosed to them. Texas law requires that all employees who witness or receive information about incidents of this type (including, but not limited to, written forms, applications, one-on-one conversations, class assignments, class discussions, or thirdparty reports) must report it to the Title IX Coordinator. Before talking with me, or with any faculty or staff member about a Title IX-related incident, please remember that I will be required to report this information.

Although graduate teaching and research assistants are not subject to Texas Education Code, Section 51.252, they are mandatory reporters under federal Title IX regulations and are required to report a wide range of behaviors we refer to as sexual misconduct, including the types of misconduct covered under Texas Education Code, Section 51.252. Title IX of the Education Amendments of 1972 is a federal civil rights law that prohibits discrimination on the basis of sex – including pregnancy and parental status – in educational programs and activities. The Title IX Office has developed supportive ways and compiled campus resources to support all impacted by a Title IX matter.

If you would like to speak with a case manager, who can provide support, resources, or academic accommodations, in the Title IX Office, please email: supportandresources@austin.utexas.edu. Case managers can also provide support, resources, and accommodations for pregnant, nursing, and parenting students.

For more information about reporting options and resources, please visit: https://titleix.utexas.edu, contact the Title IX Office via email at: titleix@austin.utexas.edu, or call 512-471-0419.

University Resources

For a list of university resources that may be helpful to you as you engage with and navigate your courses and the university, see the University Resources Students Canvas page.