# EXPLAINABLE AND FAIR ARTIFICIAL INTELLIGENCE

**DEPARTMENT OF COMPUTER SCIENCE**

## INSTRUCTOR INFORMATION


- **Instructor:** Sheikh Rabiul Islam, Ph.D.
- **Office:** Business and Science Building, Room 320, 227 Penn Street, Camden, NJ 08102
- **Email:** *sheikh.islam@rutgers.edu*

- **Office Hours:**
    - Tuesday and Thursday 1:00 - 2:00 p.m.
    - Please feel free to send me an email to schedule an appointment at another time.
    - Please don't hesitate to visit at any time to check if I am available to assist you.

## COURSE INFORMATION

- **Course:** Explainable and Fair Artificial Intelligence (Undergraduate: 50:198:491:01; Graduate: 56:198:518:01) **Semester:** Fall 2024; **Credits:** 3.00.
- **Location:** Armitage Hall 105; **Day:** Tuesday and Thursday; **Time:** 2:00 pm - 3:20 pm

## COURSE DESCRIPTION

This course introduces students to the concepts of explainability, focusing on the black-box nature of models, and fairness, addressing biases in both models and data, in Artificial Intelligence-based automated decisions. Students will explore various explainability methods, including local, global, model-specific, and model-agnostic approaches, while considering notions such as pre-training and post-training explainability. They will also examine definitions and performance metrics related to fairness and learn to interpret and apply bias detection and mitigation techniques. Additionally, students will explore open-source tools for both explainability and fairness, with the aim of solving real-world problems.

## COURSE OBJECTIVES/STUDENT LEARNING OUTCOMES

This course focuses on (1) concepts: core data mining, machine learning, deep learning, fairness, and explainability concepts, and (2) practice: practical skills for applying different data mining, explainability, and bias mitigation techniques to solve real-world problems. The learning objectives are:

- Gather and process raw data into suitable input for a wide range of data mining, explainability, and bias mitigation algorithms.
- Analyze different data mining tasks such as prediction, classification, and clustering, and the algorithms for solving these tasks.
- Systematically evaluate different data mining, explainability, and bias mitigation algorithms and understand the comparative pros and cons of these algorithms on different data mining tasks.

- Design and implement explainable and fair data mining applications using real-world datasets, and select and evaluate suitable techniques for that particular application.
- Explore recent research and development on explainability and fairness issues with AI.
- Critique classical and contemporary research papers on explainability and fairness issues with AI.

## MAJOR TEACHING METHODS

Lectures, Demonstrations, Programming Assignments, and Reading. Expect to spend at least 9 hours/week on this course, including class meeting times.

## TOPICS AND SCHEDULE

Tentative course schedule:

| Schedule | Topics | Subtopics | Note |
|---|---|---|---|
| **Week 1** *September* | Machine Learning | Brief overview of Regression, Classification, Supervised Learning, Unsupervised Learning, Reinforcement Learning, Neural Networks, and Generative AI | |
| **Week 2** | Data Mining | Mean, Median, Mode, Standard Deviation, Data Distribution, Scatter Plot, Linear Regression, Scale, Train/Test, Decision Tree, Confusion Matrix, Logistic Regression, Clustering, Grid Search, Cross Validation, etc. | |
| **Week 2** | Deep Learning | Brief overview of popular Deep Learning Algorithms including Neural Networks, Convolutional Neural Networks (CNN); Long Short Term Memory Networks (LSTMs); Recurrent Neural Networks (RNNs); Generative Adversarial Networks (GANs); Multilayer Perceptrons (MLPs); Autoencoders. | |
| **Week 3** | Interpretability | Importance, Taxonomy, Scope, and Evaluation of Interpretability; Properties of Explanations, Human-friendly explanations. | Ch 2, 3 |
| **Week 4** | Interpretable Models | Linear Regression; Logistic Regression; Decision Trees; Decision Rules; RuleFit; etc. | Ch 4, 5 |
| **Week 5** *October* | Introduction to Fairness | Machine Bias, Demographic disparities, State of society, Data to models, Feedback loop, Limitations and opportunities | Term project proposal; |
| **Week 6-7** | Model Agnostic Methods | Partial Dependence Plot (PDP); Accumulated Local Effects (ALE) Plot; Feature Interaction; Feature Importance; Global Surrogate: Prototypes and Criticisms; Individual Conditional Expectation (ICE); Local Surrogate (LIME); Counterfactual Explanations; Shapley Values; SHAP (SHapley Additive exPlanations) | Ch 6, 8 |
| **Week 8** | Neural Network Interpretation | Learned Features, Pixel Attribution (Saliency Maps); Concept Detection (TCAV - Testing with Concept Activation Vectors); Adversarial Examples; Influential Instances; | Ch 10, Test-1, Assignment-1 |
| **Week 9** *November* | Legitimacy of Automated Decision Making | Forms of automation, mismatch between target and goal, failing to consider relevant information, a right to accurate prediction. | Ch 1, 2 |
| **Week 10** | Classification | Modeling populations as probability distributions; Supervised learning, Groups in population, independence, separation, sufficiency. | Ch 3 |

| | | | |
|---|---|---|---|
| **Week 11** | Notions of fairness | Systematic relative disadvantage, Wrongfulness of discrimination, Intentionality and indirect discrimination, Equal opportunity, Cost of fairness, Statistical and moral notions of fairness. | Ch 4 |
| **Week 12** | Testing Discrimination in Practice | Systematize tests of discrimination and the practical complexities of applying them, both to traditional decision-making systems and to algorithmic systems. Opensource tool Aequitas, IBM AI 360 ; | Ch 7 |
| **Week 13-14** *December* | Anti-Discrimination Law and a Broader View of Discrimination | Anti-discrimination law around the world; How it navigates tradeoffs, its limits, and how it applies to machine learning. Review of structural, organizational, and interpersonal discrimination in society, how machine learning interacts with them, and discussion of a broad set of potential interventions | Ch 6, 8<br>Test-2, Assignment-2,<br>Term Project; |

## REQUIRED MATERIALS

1. **Book**

   A. **Interpretable Machine Learning** - A Guide for Making Black Box Models Explainable
   By Christoph Molnar. Available at **https://christophm.github.io/interpretable-ml-book/**

   B. **Fairness and Machine Learning**
   By Solon Barocas, Moritz Hardt, Arvind Narayanan. Available at **https://fairmlbook.org/**

   **Optional:** Data Mining: Practical Machine Learning Tools and Techniques, 4th edition by Eibe Frank, Mark A. Hall, and Ian H. Witten.

2. **Canvas:** https://canvas.rutgers.edu/
   Used for accessing practice problems with code, course syllabus and materials, viewing grades, and turning in assignments.

## GRADES

- All grades will be posted in Canvas.
- If you have an issue with a grade you receive on ANY assignment or exam, you must email the instructor within THREE days of the grade being released to the class.
- Grade Scale:
  89.5–100 = A (Outstanding)
  84.5–89.49 = B+, 79.5–84.49 = B (Good),
  74.5–79.49 = C+, 69.5–74.49 = C (Satisfactory),
  59.5–69.49 = D (Poor),
  0–59.49 = F (Failing).

  A grade of C or better is usually required for Major or Minor courses, while General Requirement courses must only be passed with a D or better. The grade of D is not valid for graduate-level courses. Students may only receive a C or better, F or IN for graduate courses.

Grades in this course are **earned** using the following distribution:

| Item | Percentage |
|---|---|
| **Assignments**<br>2 assignments (14%) | 15% |
| **Exams**<br>2 test (40%)<br>2 quizzes (10%) | 50% |
| **Term project** | 30% |
| **Participation**<br>Responsiveness (5%) | 5% |

## ASSIGNMENTS (15% OF GRADE)

There will be multiple assignments.

## EXAMS (50% OF GRADE)

- There will be two exams containing 40% of the total grade combined.
- The last exam is not comprehensive.
- There will be two quizzes containing 10% of the total grade.

## TERM PROJECT AND REPORT (30% OF GRADE)

- The term project will be on solving a real-world problem that might have interpretability and/or fairness-related issues. You will need to add on top of current solutions, improve current solutions, or find a novel solution for a real-world problem. In other words, your solution should be novel to some extent.
- Within the third week of the semester, you need to get approval of the proposed project in written format.
- A student will need to submit a report (6 -10 pages, single-spaced page) on the project. The source code and report contain 25% of the grade.
- A student will need to give a short presentation on the project at the end of the semester. The presentation is 5% of the grade.
- The expectations for graduate students are relatively higher when it comes to the term project. Pairing is allowed for both undergraduates and graduates, but only pairs consisting of one graduate student and one undergraduate student will receive an additional 2% extra credit.

## PARTICIPATION (5% OF GRADE)

- This is based on your responsiveness in the class sessions and discussions. **Use of electronic device is prohibited during lecture time unless explicitly authorized for in-class activities.** Distracted behavior in the classroom will negatively impact your participation grade.

## POLICY

- All students should follow the Academic Integrity Policy as mentioned at: https://deanofstudents.camden.rutgers.edu/sites/deanofstudents/files/Academic%20Integrity%20Policy.pdf

- Whenever you submit any work, **you must acknowledge the source** if any part of the submitted **content is not originally yours**.

## SERVICES AND RESOURCES

- A comprehensive list of student services and resources are listed here: https://studentaffairs.camden.rutgers.edu/student-resource-list

- Here are some crucial student services and resources:

    - The Center for Learning and Student Success (CLASS) provides academic support and enrichment services for students, at no additional cost, including one-on-one tutoring, small-group tutoring and workshops, online tutoring, writing assistance, student success coaching, learning assessment, and metacognition training. Learn more about this service here: https://class.camden.rutgers.edu/

    - Office of Disability Services (ODS)—Students with Disabilities: If you need academic support for your courses, accommodation can be provided as indicated in the accommodation letter. If you have not registered with ODS and you have or think you have a disability (learning, sensory, physical, chronic health, mental health, or attentional), please visit the ODS website: https://success.camden.rutgers.edu/disability-services

    - Dean of Students Office—CARES Team: For some students, personal, emotional, psychological, academic, or other challenges may hinder their ability to succeed both in and outside of the classroom. The Dean of Students Office serves as your initial contact if you need assistance with these challenges. You can learn more about the free services by visiting the Dean of Students website http://deanofstudents.camden.rutgers.edu/