Fair Interpretable Trustworthy Machine Learning

Code: CS32225CS	Course Type: Elective	Semester: II
Evolution Scheme (TA-MSE-ESE): 20-30-100	Total Marks: 150	L-T-P (Credits): 3-0-0 (3)

Pre-Requisites: Machine Learning, Probability & Statistics, Python Programming

Course Objective: To build fair, interpretable, and trustworthy machine learning systems for real-world applications, focusing onexplaining model decisions, identifying and reducing bias, and ensuring robustness, privacy, and accountability in ML systems.

Course Outcomes: At the end of the course, the student will be able to

CO-1	Apply standard interpretability techniques to explain ML predictions and compare their strengths and limitations.
CO-2	Detect and quantify bias and apply fairness-aware methods to reduce discrimination in classification and regression.
CO-3	Assess robustness and privacy risks and implement basic defenses against adversarial attacks and data leakage.
CO-4	Design and evaluate an end-to-end trustworthy ML pipeline for a real-world application.

Course Articulation Matrix:

CO's	PEO-1	PEO-2	PEO-3	PEO-4	PEO-5
CO1	3	2	1	-	-
CO2	3	2	1	2	1
CO3	2	2	3	2	1
CO4	3	3	2	3	2

1 - Slightly 2 - Moderately 3 – Substantially

Syllabus:

Unit-1: Foundations of Trustworthy and Interpretable ML

Introduction to trustworthy AI: robustness, fairness, privacy, safety, and accountability in ML systems; interpretability and explainability; motivation for transparent models in high-stake domains; Human factors in explainability, interpretability myths, and limitations of black-box models.
Bias and fairness concepts: types of bias in data and algorithms, observational vs. interventionist notions of fairness.

Unit-2: Techniques for Interpretable and Explainable ML

Intrinsically interpretable models: decision trees, rule-based models, generalized additive models (GAMs), and linear model; Model-agnostic interpretability: LIME, SHAP, feature attributions, and surrogate models; Post-hoc explanations and evaluation of explanation methods; pitfalls and disagreement; visualization-based methods for interpretability, attention mechanisms, and concept-based explainability.

Unit-3: Fairness, Robustness, and Privacy in ML

Introduction to fairness; Fair representation learning, bias mitigation techniques, and fairness-through input manipulation;

Fairness in NLP: domain-specific biases and audit techniques; Robustness and adversarial attacks & defenses in ML; connections of robustness with privacy and generalization; ML auditing, model monitoring, and privacy-preserving techniques.

Unit-4: Trustworthy and Human-Aligned ML Systems

Explainability for fair machine learning: fairness via explanation quality, right-to-explanation, and right-to-be-forgotten in ML systems; Unlearning and forget-me-not mechanisms in ML pipelines; Trustworthiness beyond interpretability: safety, accountability, and alignment of ML systems with human values.
Case studies: real-world deployments of fair, interpretable, and trustworthy ML in domains such as credit scoring, healthcare, and criminal-justice-adjacent systems.

Learning Resources:

Text Books:

1. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable by Christoph Molnar.
2. Trustworthy Machine Learning – Kush R. Varshney.
3. Interpretable and Trustworthy AI: Techniques and Frameworks by Pethuru Raj, Kousalya Govardhanan, B. Sundaravadivazhagan, Shubham Mahajan, M. Nalini.