Explainable AI

Code: CS33122CS	Course Type: Elective	Semester: I
Evolution Scheme (TA-MSE-ESE): 20-30-100	Total Marks: 150	L-T-P (Credits): 3-0-0 (3)

Pre-Requisites: Machine Learning, Deep Learning, Probability & Statistics

Course Objective: To develop an understanding of interpretability and transparency in AI systems, enabling students to design explainable models, evaluate fairness and bias, and ensure accountability in AI-driven decision-making systems.

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

CO-1	Analyze the need for interpretability and transparency in AI models across application domains
CO-2	Apply model-agnostic and model-specific explainability techniques to interpret machine learning and deep learning models
CO-3	Evaluate fairness, bias, and ethical implications of AI systems
CO-4	Design trustworthy AI systems incorporating explainability, accountability, and compliance principles

Course Articulation Matrix:

PO	PEO-1	PEO-2	PEO-3	PEO-4	PEO-5
CO-1	3	2	1	-	-
CO-2	3	2	1	-	1
CO-3	3	2	-	-	3
CO-4	3	2	1	1	-

1 - Slightly; 2 - Moderately; 3 – Substantially

Syllabus:

Unit-1: Introduction to Explainable AI: Motivation for interpretability, black-box vs white box models, levels of explainability, human-centered AI, regulatory requirements, trust and accountability in AI systems, evaluation of interpretability.

Unit-2: Model-Agnostic Explainability Techniques: Feature importance methods, partial dependence plots, surrogate models, LIME, SHAP, sensitivity analysis, local vs global explanations, visualization techniques for interpretability.

Unit-3: Model-Specific Interpretability: Interpretable models (decision trees, rule-based systems), explainability in deep learning (saliency maps, Grad-CAM, attention visualization), interpretability

in NLP and vision models, counterfactual explanations.

Unit-4: Fairness, Ethics and Trustworthy AI: Bias detection and mitigation techniques, fairness metrics, accountability and governance frameworks, privacy-preserving explainability, explainability in high-stakes domains (healthcare, finance), case studies and real-world deployments.

Learning Resources:

Text Books:

1. Christoph Molnar, Interpretable Machine Learning
2. Samek, Montavon et al., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning
3. Ribeiro et al. (LIME), Lundberg & Lee (SHAP) – research papers