DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

NATIONAL INSTITUTE OF TECHNOLOGY PATNA

Ashok Raj Path, PATNA 800 005 (Bihar), India

Phone No.: 0612 – 2372715, 2370419, 2370843, 2371929, 2371930, 2371715 Fax – 0612- 2670631 Website: www.nitp.ac.in

CSXX1908 ***Explainable Artificial Intelligence*** **L-T-P-Cr: 2-0-2-3**

Pre-requisites: The students are expected to be fluent in basic mathematics, algorithms, and machine learning. Students are also expected to have programming and software engineering skills to work with data sets using Python, numpy, and sklearn.

Objectives/Overview:

- Define and explain the importance of explainability in artificial intelligence.
- Introduce and analyze various techniques for enhancing the interpretability of machine learning models.
- Cover methods such as feature importance analysis, rule-based models, and surrogate models.
- Provide hands-on experience in applying explainability techniques to real-world datasets and problems.

Course Outcomes – After completing this course, students should be able to:

CO-1. *Recall* Students will effectively apply interpretability techniques, such as feature importance analysis, rule-based models, and model-agnostic methods, to enhance the transparency of machine learning models.

CO-2. *Define* and *formulate* Students will possess the ability to evaluate and compare different interpretability methods.

CO-3. *Design* and *develop* Students will develop a critical understanding of the trade-offs between model complexity and interpretability in diverse scenarios.

CO-4. *Analyze* and design demonstrating practical skills, students will apply explainability techniques to real-world datasets and challenges.

CO-5. *Distinguish* students will gain awareness of the ethical implications related to AI transparency and interpretability.

Course Outcomes–Cognitive Levels–Program Outcomes Matrix –

[H: High relation (3); M: Moderate relation (2); L: Low relation (1)]

Course Outcomes	**Program Outcomes**											
	PO-1 (Engineering knowledge)	PO-2 (Problem analysis)	PO-3 (Design/development of solutions)	PO-4 (Conduct investigations of complex problems)	PO-5 (Modern tool usage)	PO-6 (The engineer and society)	PO-7 (Environment and sustainability)	PO-8 (Ethics)	PO-9 (Individual and team work)	PO-10 (Communication)	PO-11 (Project management and finance)	PO-12 (Life-long learning)
CO-1	3	3	3	3	2	3			3	3	1	3
CO-2	3	3	3	3	2	3		1	3	3	1	3
CO-3	3	3	3	3	3	3			3	3	1	3
CO-4	3	3	3	3	2	3			3	3	1	3
CO-5	3	3	3	3	3	3	2	1	3	3	1	3
CO-6	3	3	2	1	3	1	1	1	3	3	2	2

UNIT 1 Interpretation, Interpretability, and Explainability: Lecture 2

Machine learning interpretation, difference between interpretability and explainability. Black-box models, White-box models, Explainability.

Model interpretability method types and scopes, interpreting individual predictions with logistic regression,

UNIT 2 Interpretation Challenges: Lecture 4

Traditional model interpretation methods, Intrinsically interpretable models: Generalized linear models (GLMs), Linear regression, Ridge regression, Polynomial regression, Logistic regression, Decision trees, CART decision trees, RuleFit, Interpretation and feature importance, Nearest neighbors, k-Nearest Neighbors, Naïve Bayes, Gaussian Naïve Bayes, Explainable Boosting Machine (EBM), GAMI-net

UNIT 3 Global and Local Model-Agnostic Interpretation Methods: Lecture 6

Feature importance: Feature importance with model-agnostic methods,

Permutation feature importance: SHAP values, Visualize global explanations

Local Model-Agnostic Interpretation Methods

Local interpretations with SHAP values: LIME, LIME for tabular data, LIME for image data, LIME for NLP, Comparing SHAP with LIME,

UNIT 4 Anchors and Counterfactual Explanations: Lecture 2

Understanding anchor explanations: Local interpretations for anchor explanations

Exploring counterfactual explanations: Counterfactual explanations guided by prototypes, Performance & Fairness

UNIT 5 Visualizing Convolutional Neural Networks: Lecture 3

The CNN models: CNN classifier with traditional interpretation methods, Activation-based methods, Intermediate activations,

Gradient-based attribution methods: Saliency maps, Guided Grad-CAM, Integrated gradients

Understanding classifications with perturbation-based attribution methods:

Feature ablation, Occlusion sensitivity, Shapley value sampling, KernelSHAP

UNIT 6 Interpreting NLP Transformers: Lecture 3

Visualizing attention with BertViz, layer attention with the head view, Interpreting token attributions with integrated gradients,

UNIT 7 Feature Selection and Engineering for Interpretability: Lecture 3

Effect of irrelevant features, Removing unnecessary features, Correlation filter-based methods, Ranking filter-based methods, Comparing filter-based methods, Exploring embedded feature selection methods, Discovering wrapper, hybrid, and advanced feature selection methods, Wrapper methods, Model-agnostic feature importance, Genetic algorithms

UNIT 8 Bias Mitigation and Causal Inference Methods: Lecture 2

Detecting bias: Quantifying dataset bias, Quantifying model bias,

Mitigating bias: Preprocessing bias mitigation methods, In-processing bias mitigation methods, Post-processing bias mitigation methods, Creating a causal model: Understanding the results of the experiment, Understanding causal models, Initializing the linear doubly robust learner, Fitting the causal model. Understanding heterogeneous treatment effects, Testing estimate robustness:

UNIT 9 Adversarial Robustness: **Lecture 3**

Learning about evasion attacks: Fast gradient sign method attack, Carlini and Wagner infinity norm attack, Targeted adversarial patch attack. Defending against targeted attacks with preprocessing, Shielding against any evasion attack by adversarial training of a robust classifier Evaluating adversarial robustness

Text/Reference Book:

1) *Christoph Molnar,* Interpretable Machine Learning, A Guide for Making Black Box Models Explainable, Leanpub , 2023
2) Michael Munn, David Pitman, Explainable AI for Practitioners, O'Reilly Media, Inc. , 2022
3) *Leonida Gianfagna, Antonio Di Cecco,* Explainable AI with Python, Springer , 2021