NIRMA UNIVERSITY

Institute:	Institute of Technology, School of Technology
Name of Programme:	MTech CSE, MTech CSE (Data Science)
Course Code:	6CS376ME25
Course Title:	Explainable AI
Course Type:	Department Elective-II
Year of Introduction:	2025-26

L	T	Practical Component				C
		LPW	PW	W	S	
2	0	2	-	-	-	3

Course Learning Outcomes (CLO):

At the end of the course, the students will be able to –

1. demonstrate the concepts within Explainable AI and interpretable machine learning (BL2)
2. identify current techniques for generating explanations from black-box machine learning methods (BL3)
3. analyse current ethical, social, and legal challenges related to Explainable AI skills and abilities (BL4)
4. assess Explainable AI methods for the given applications. (BL5)

Unit	Contents	Teaching Hours (Total 30)
Unit-I	**Introduction**: Introduction to the multidisciplinary topics of Explainable AI (XAI), what is XAI, the importance of XAI, XAI-related terminologies **Taxonomy of XAI methods**: Intrinsic vs post hoc, model-specific vs model-agnostic, and local vs global **Properties and Trade-off**: properties of Explanation methods, tread-off between accuracy and explainability, human-friendly explanations	06
Unit-II	**Intrinsically explainable models**: Linear Regression, Logistic Regression, Generalized Linear Model (GLM), Generalized Additive Model (GAM), and Decision Tree.	04
Unit-III	**XAI methods and its evaluations:** Model-Agnostic Methods, Example-based methods, Global Model-Agnostic methods including Partial Dependence Plot (PDP), Conformal Prediction, Individual Conditional Expectation (ICE), Feature Importance, Saliency Maps, Local Interpretable Model-Agnostic Explanations (LIME), SHAP, Integrated Gradient (IG)	05
Unit-IV	**Visualization Techniques**: Activation Maps in CNNs, Attention mechanism in NLP, Visualizing decision boundaries and feature interactions	05
Unit-V	**Fairness and Bias in AI**: Understanding biases in data and models, Metrics for fairness evaluation, Techniques to mitigate bias in AI systems. **Ethical Considerations**: The impact of AI on society, Responsible AI practices, and guidelines	08

Unit-VI	**Explainability in Reinforcement Learning:** Understanding policies learned by RL agents, Interpreting state-action trajectories and reward mechanisms. **Applications of XAI:** healthcare, finance, autonomous systems, and other domains. **Futuristic approaches**: The Future of Machine Learning models and its Interpretability.	02

Self-Study:

The self-study contents will be declared at the commencement of the semester. Around 10% of the questions will be asked from self-study content.

Suggested Readings/ References:

1. Molnar, Christoph, Interpretable Machine Learning, Leanpub
2. Denis Rothman, Hands-On Explainable AI (XAI) with Python, Packt Publishing
3. Michael Munn, David Pitman, Explainable AI for Practitioners, O'Reillyly
4. Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer
5. Uday Kamath, John Liu, Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning, Springer.

Suggested List of Experiments:

Sr. No.	Name of Experiments/Exercises	Hours
1	Installing and understanding various packages of model interpretation	02
2	Interpreting tree models	04
3	Implementing the SHAP model for textual data and analyzing ALE, ICE, and PDP plots	04
4	Implementing Grad-CAM model for image dataset	04
5	Implement LIME model for image dataset	02
6	Implement integrated gredients for a given image dataset	04
7	What-if-tool image smile detection and visualization	04
8	Implementation of XAI Chatbot	04
9	Generate an anchor explanation for ImageNet dataset	02
10	Cognitive XAI for IMDB dataset.	02