

## NIRMA UNIVERSITY

<b>Institute:</b>	Institute of Technology
<b>Name of Programme:</b>	BTech CSE
<b>Course Code:</b>	4CS103DE25
<b>Course Title:</b>	Explainable AI
<b>Course Type:</b>	Disciplinary Minor- Elective
<b>Year of Introduction:</b>	2025-26

L	T	Practical Component				C
		LPW	PW	W	S	
3	0	2	-	-	-	4

### 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 45)
Unit-I	<b>Introduction:</b> Introduction to the multidisciplinary topics of Explainable AI (XAI), what is XAI?, the importance of XAI, XAI-related terminologies <b>Taxonomy of XAI methods:</b> Intrinsic vs post hoc, model-specific vs model-agnostic, and local vs global <b>Properties and Trade-off:</b> properties of Explanation methods, tread-off between accuracy and explainability, human-friendly explanations	08
Unit-II	<b>Intrinsically explainable models:</b> Linear Regression, Logistic Regression, Generalized Linear Model (GLM), Generalized Additive Model (GAM), and Decision Tree.	06
Unit-III	<b>XAI methods and its evaluations:</b> 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)	08
Unit-IV	<b>Visualization Techniques:</b> Activation Maps in CNNs, Attention mechanism in NLP, Visualizing decision boundaries and feature interactions.	08
Unit-V	<b>Fairness and Bias in AI:</b> Understanding biases in data and models, Metrics for fairness evaluation, Techniques to mitigate bias in AI systems. <b>Ethical Considerations:</b> The impact of AI on society, Responsible AI practices and guidelines.	12

Unit-VI	<b>Explainability in Reinforcement Learning:</b> Understanding policies learned by RL agents, Interpreting state-action trajectories and reward mechanisms. <b>Applications of XAI:</b> healthcare, finance, autonomous systems, and other domains. <b>Futuristic approaches:</b> The Future of Machine Learning models and its Interpretability	03
---------	--	----

#### 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 contents

#### 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'Reilly
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.	Title	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 gradients 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