III B.Tech II Semester

		L	T	P	C
23A31702a	**EXPLAINABLE AI &MODEL INTERPRETABILITY (Professional Core)**	**3**	**0**	**0**	**3**

Course Objectives:

- To introduce the principles of interpretability and explainability in AI/ML models.
- To analyze the trade-offs between model accuracy and interpretability.
- To explore popular post-hoc and intrinsic explainability techniques.
- To examine fairness, accountability, and transparency in AI systems.
- To develop hands-on skills with interpretability tools and libraries.

Course Outcomes:

Upon successful completion of the course, students will be able to:

- Understand the need for explainability in modern AI systems.
- Differentiate between black-box and white-box models.
- Apply interpretability techniques such as SHAP, LIME, and PDPs.
- Evaluate the fairness and transparency of AI systems.
- Use explainability tools for model auditing and deployment in high-stakes domains.

UNIT I: Foundations of Explainable AI

Introduction to Explainability and Interpretability, Importance of XAI in Healthcare, Finance, and Law , White-box vs Black-box Models, Desiderata: Fairness, Accountability, Transparency, Human-Centered AI and Trust ,Taxonomy of XAI Techniques (Global vs Local, Post-hoc vs Intrinsic), Regulatory and Ethical Implications (GDPR, AI Bill of Rights), Model Simplicity vs Predictive Power.

UNIT II: Model-Specific Explainability Techniques

Decision Trees and Rule-based Models, Linear Models and Feature Importance, Generalized Additive Models (GAMs), Visualization of Weights and Coefficients, Logistic Regression Coefficient Interpretation, Case Study: Credit Scoring using Transparent Models,Comparison of Interpretable ML Models, Use Cases and Trade-offs.

UNIT III: Model-Agnostic Explainability Techniques

Local Interpretable Model-agnostic Explanations (LIME), SHAP Values (SHapley Additive exPlanations), Partial Dependence Plots (PDPs), Individual Conditional Expectation (ICE) Plots,Anchors and Counterfactual Explanations, Feature Interaction and Permutation Importance, Comparative Analysis of SHAP, LIME, PDP, Model Debugging with XAI.

UNIT IV: Deep Learning Explainability

Visualizing CNNs: Filters, Feature Maps, Saliency Maps and Grad-CAM, Integrated Gradients, Explaining RNNs and LSTM Outputs, Concept Activation Vectors (TCAV), Attention-based Interpretability in Transformers, Explaining Language Models (BERT, GPT) Evaluation of Deep Model Explanations.

UNIT V: Fairness, Bias & Tools for XAI

Fairness Metrics: Demographic Parity, Equal Opportunity, Sources of Bias in Data and Models, Discrimination Detection and Mitigation Strategies, Introduction to AIF360, What-If Tool, Fairlearn, Case Study: Bias in Hiring Algorithms, Explainability in ML Pipelines (MLFlow, Skater), XAI in Federated and Privacy-Preserving AI, Designing Interpretable AI Systems from Scratch.

Textbooks:

1. Christoph Molnar, "Interpretable Machine Learning", Leanpub.
2. Sameer Singh et al., "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning", Springer.
3. Dan Roth, Zachary Lipton, and Been Kim, "Explainable AI: Foundations, Developments, Prospects", MIT Press (Online forthcoming).

Reference Books:

1. Marco Tulio Ribeiro et al., "Why Should I Trust You?" (LIME) – Research Paper
2. Scott Lundberg et al., "A Unified Approach to Interpreting Model Predictions" (SHAP) – NIPS
3. A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges", Information Fusion Journal.
4. Zachary C. Lipton, "The Mythos of Model Interpretability" – Communications of the ACM

Online Learning Resources:

- Coursera – Explainable AI with Google Cloud
- Udacity – AI for Everyone by Andrew Ng
- HarvardX – Data Science: Machine Learning Interpretability