GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Engineering

Level: PG

Branch: Artificial Intelligence and Data Science

Subject Code: ME02095081

Course/Subject Name: Explainable Artificial Intelligence

WEF Academic Year	2024-25
Semester	2
Category of the Course	Professional Elective Course

Prerequisite:	Artificial Intelligence
Rationale	As the use of artificially intelligent systems increases, it becomes equally important to maintain the trust of the stakeholders on it. This course provides the learners insights of developing explainable artificially intelligent systems. It also covers the existing methods of machine learning model interpretable methods and their applications. This course helps the learner in design & development of trustworthy, fair and explainable artificially intelligent agents.

Course Outcome:

After completion of the Course, Students will be able to:

No	Course Outcomes	RBT Level*
01	Understand the concepts of Explainable AI and interpretable machine learning.	UN
02	Apply current XAI techniques for generating explanations from black-box ML models.	AP
03	Apply comprehension of current ethical, social and legal challenges related to Explainable AI.	AP
04	Analyze the working methodology of Explainable AI methods.	AN
05	Evaluate the performance of Explainable AI methods in the view of stakeholders' satisfiability.	EV

*RM: Remember, UN: Understand, AP: Apply, AN: Analyze, EL: Evaluate, CR: Create

GUJARAT TECHNOLOGICAL UNIVERSITY

Program Name: Engineering

Level: PG

Branch: Artificial Intelligence and Data Science

Subject Code: ME02095081

Course/Subject Name: Explainable Artificial Intelligence

Teaching and Examination Scheme:

Teaching Scheme (in hours)			Total Credits (L+T+PR/2)	Assessment Pattern and Marks				Total Marks
				Theory		Tutorial/Practical		
L	T	PR	C	ESE (E)	PA/CA (M)	PA/CA (I)	ESE (V)	
03	00	02	04	70	30	20	30	150

Course Content:

Unit	Course Content	No of Hours	% of Weightage
1.	**Explainability Basics** Types of machine learning systems, Building diagnostics, Gaps in diagnostics, Building a robust diagnostics, interpretability vs. explainability, White-box models, Linear regression, Decision trees	9	20%
2.	**Interpreting Model Processing** Tree ensembles, Interpreting a random forest, Model-agnostic methods: Global interpretability, Diagnostics AI: Breast cancer diagnosis, Exploratory data analysis, Deep neural network, Interpreting DNNs, LIME, SHAP, Anchors, Convolutional neural networks, Interpreting CNNs, Vanilla backpropagation, Guided backpropagation	10	25%
3.	**Interpreting Model Representation** Visual understanding, Network dissection framework, Interpreting layers and units, Sentiment analysis, Neural word embedding, Interpreting semantic similarity	9	20%
4.	**Fairness and Bias** Adult income prediction, Fairness notions, Interpretability and fairness, Mitigating bias, Datasheets and datasets, Counterfactual explanations	9	20%
5.	**Recent trends in the field of Explainable Artificial Intelligence**	8	15%
	TOTAL	**45**	**100**

Suggested Specification Table with Marks (Theory):

Distribution of Theory Marks (in %)					
R Level	**U Level**	**A Level**	**N Level**	**E Level**	**C Level**
10	**20**	**20**	**20**	**20**	**10**

Where R: Remember; U: Understanding; A: Application, N: Analyze and E: Evaluate C: Create (as per Revised Bloom's Taxonomy)

Reference/Suggested Learning Resources:

(a) Books:

1. Interpretable AI - Building explainable machine learning systems by Ajay Thampi, Manning Publications
2. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning by Andrea Vedaldi, Grégoire Montavon, Klaus-Robert Müller, Lars Kai Hansen, Wojciech Samek, Springer International Publishing
3. Hands-On Explainable AI (XAI) with Python by Denis Rothman, Packt Publishing
4. Interpretable Machine Learning - A Guide for Making Black Box Models Explainable by Christoph Molnar, Leanpub Publications

(b) Open source software and website

- Course-related online MOOCs on NPTEL/SWAYAM platform.
- Recently Published papers/articles in reputed journals.

Suggested Course Practical List:

- The practical work will be carried out based on the content covered during the academic sessions**.**

List of Laboratory/Learning Resources Required: Programming development environment (open source is encouraged) related to the course content.

Suggested Project List: The subject teacher has to assign the relevant project work to the students in individual/team.

Suggested Activities for Students: The subject teacher has to assign the outcome based activities to the students in individual/team.

****