Content

Course Code	Course Name	Semester	Theory	Practice	Lab	Credit	ECTS
INF 539		2	3	0	0	3	6

Prerequisites	
Admission Requirements	

Language of Instruction	English
Course Type	Elective
Course Level	Masters Degree
Objective	This course focuses on explaining and interpreting the decisions of machine learning algorithms. The course primarily aims to introduce students to explainable artificial intelligence (XAI) methods and demonstrate, through practical applications, how these methods are used in various areas.
Content	This course aims to interpret the decisions, predictions, or inferences of AI-based systems, and to explain how and why these outputs are calculated by existing algorithms. The course provides an overview of interpreting the decisions of artificial learning models used in various fields, from healthcare to finance, often referred to as "black boxes," and the critical aspects of developing reliable, transparent, and ethically compliant AI systems. Students will have the opportunity to apply the methods described in the course using Python and discuss their results.
References	- Mehta, M., Palade, V., & Chatterjee, I. (Eds.). (2023). Explainable AI: Foundations, methodologies and applications (Vol. 232, p. 273). Springer. - Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K. R. (Eds.). (2019). Explainable AI: interpreting, explaining and visualizing deep learning (Vol. 11700). Springer Nature. - Molnar, C. (2020). Interpretable machine learning. - Hsieh, W., Bi, Z., Jiang, C., Liu, J., Peng, B., Zhang, S., ... & Liu, M. (2024). A comprehensive guide to explainable AI: from classical models to LLMs. arXiv preprint arXiv:2412.00800.

Theory Topics

Week	Weekly Contents
1	Core Concepts: Explainability, Transparency, Interpretability, Fairness, Robustness, and XAI
2	Theoretical Foundations of Explainable AI
3	Interpretability of Traditional Machine Learning Models
4	Interpretability of Deep Learning Models
5	Techniques for Explainable AI
6	Feature Attribution Methods
7	Visualization Techniques
8	Midterm
9	Temporal and Sequence Data Techniques
10	Multimodal Explainability
11	Applications of Explainable AI - Part I
12	Applications of Explainable AI - Part II
13	Challenges
14	Student presentations