SYLLABUS

1. Information on the study programme

1.1. Higher education institution	West University of Timisoara
1.2. Faculty	Mathematics and Computer Science
1.3. Department	Computer Science
1.4. Study program field	Computer Science
1.5. Study cycle	PhD
1.6. Study programme / Qualification	Doctoral School in Mathematics and Computer Science/ Computer Science

2. Information on the course

2.1. Course title			Explainable and Trustworthy Artificial Intelligence				
2.2. Lecture instructor			Prof.dr. Darian Onchiș				
2.3. Seminar / laboratory instructor							
2.4. Study year	1	2.5. Semester	1	2.6. Examination type	E	2.7. Course type	Elective

3. Estimated study time (number of hours per semester)

3.1. Attendance hours per week	1	out of which: 3.2 lecture	1	3.3. seminar / laboratory	-
3.4. Attendance hours per semester	12	out of which: 3.5 lecture	12	3.6. seminar / laboratory	0

Distribution of the allocated amount of time*	**hours**
Study of literature, course handbook and personal notes	80
Supplementary documentation at library or using electronic repositories	54
Preparing for laboratories, homework, reports etc.	40
Exams	6
Tutoring	8
Other activities…	0

3.7. Total number of hours of individual study	188
3.8. Total number of hours per semester	200
3.9. Number of credits (ECTS)	8

4. Prerequisites (if it is the case)

4.1. curriculum	Machine Learning and Artificial Intelligence fundamentals
4.2. competences	Python programming, Colab notebooks

5. Requirements (if it is the case)

5.1. for the lecture	Projector
5.2. for the seminar / laboratory/ individual activity	Internet

6. Specific acquired competences

Professional competencies	Design of XAI systems
Transversal competencies	Project work, team work

7. Course objectives

7.1. General objective	Introduction in modern XAI
7.2. Specific objectives	Presentation of selected topics of XAI and specific applications

8. Content

8.1. Lecture	**Teaching methods**	**Remarks, details**
Introduction to explainable and trustworthy artificial intelligence	Lecture, exemplification, demonstration	1h
Models bias and human decisions in explainability	Lecture, exemplification, demonstration	1h
Surrogate models and Post hoc Explanations	Lecture, exemplification, demonstration	2h
Intrinsic Interpretable models	Lecture, exemplification, demonstration	2h
Neuro-symbolic models	Lecture, exemplification, demonstration	2h
Counterfactual Explanations	Lecture, exemplification, demonstration	1h
Attention and Concept Based Explanations	Lecture, exemplification, demonstration	2h
Interpreting Generative Models	Lecture, exemplification, demonstration	1h

Recommended literature:

1. Molnar C. Interpretable Machine Learning, A Guide for Making Black Box Models Explainable, accessed 2023-08-21, https://christophm.github.io/interpretable-ml-book/
2. Hong et. al., 2020, Human Factors in Model Interpretability: Industry Practices, Challenges
3. Letham and Rudin, 2015, Interpretable Classifiers Using Rules and Bayesian Analysis
4. Ribeiro et. al., 2016, Why should I trust you? Explaining the Predictions of Any Classifier
5. Lundberg and Lee, 2017, A Unified Approach to Interpreting Models
6. Wachter et. al., 2018, Counterfactual Explanations Without Opening the Black Box
7. Jain and Wallace, 2019, Attention is not Explanation
8. Covert et. al., 2021, Explaining by Removing: A Unified Framework for Model Explanation
9. Han et. al., 2022, Which Explanation Should I Choose? A Function Approximation Perspective to Characterizing Post hoc Explanations
10. Onchis DM et. al.. 2023, Neuro-symbolic model for cantilever beams damage detection, Computers in Industry 151, 103991
11. Onchis DM et al., 2022, A Neuro-Symbolic Classifier with Optimized Satisfiability for Monitoring Security Alerts in Network Traffic, Applied Sciences 12 (22), 11502 1

12. Cozma G., Onchis DM et al, 2022, Explainable Machine Learning Solution for Observing Optimal Surgery Timings in Thoracic Cancer Diagnosis, Applied Sciences		
8.2. Seminar / laboratory	**Teaching methods**	**Remarks, details**

9. Correlations between the content of the course and the requirements of the professional field and relevant employers.

The course is intended to follow the needs of the IT companies active in the field.

10. Evaluation

Activity	10.1. Assessment criteria	10.2. Assessment methods	10.3. Weight in the final mark
10.4. Lecture	Knowledge of XAI models	Project defense: theoretical part and related questions	100%
10.5. Seminar / laboratory			
10.6. Minimum needed performance for passing			
At three topics presented at the course fully understood.			

Date of completion	Signature (lecture instructor)	Signature (seminar instructor)
22.09.2023	prof.dr. Darian Onchis	

Date of approval	Signature (director of the department/ doctoral school)