Modulbezeichnung:	Introduction to Explainable Machine Learning (xML) (Introduction to Explainable Machine Learning)	5 ECTS
Modulverantwortliche/r:	Thomas Seel, Simon Bachhuber, Ive Weygers	
Lehrende:	Thomas Seel	

Startsemester: SS 2022	Dauer: 1 Semester	Turnus: jährlich (SS)
Präsenzzeit: 60 Std.	Eigenstudium: 90 Std.	Sprache: Englisch

Lehrveranstaltungen:

Introduction to Explainable Machine Learning (SS 2022, Vorlesung mit Übung, 4 SWS, Thomas Seel et al.)

Empfohlene Voraussetzungen:

Participants should be familiar with fundamental methods and concepts in machine learning. They should, for example, have completed one of the following courses

- Machine Learning for Engineers
- Maschinelles Lernen für Zeitreihen
- Pattern Recognition
- Deep Learning

Inhalt:

This course gives an introduction to explainable and interpretable methods and approaches in machine learning. We discuss prominent concepts in explainable machine learning, analyze and compare their potential and shortcomings, and apply them to example problems. The covered topics include but are not limited to:

- the role of explanations in machine learning (ML)
- definitions and terminology in explainable ML
- inherent versus post-hoc explainability
- prototypes in classification
- heat maps and saliency-based approaches
- global post-hoc explanations via surrogate models
- additive feature attribution methods
- local interpretable model-agnostic explanations
- explanations via Shapley values
- advanced methods from recent literature
- plausability, faithfulness, comprehensibility and consistency of explanations

The example problems to which we will apply the concepts and methods will stem from application domains in which explainability is considered crucial, such as digital health.

Lernziele und Kompetenzen:

Fachkompetenz

Wissen

Participants will be familiar with several machine learning concepts and methods that yield explainable results. They will know which properties explanations should ideally have and in which ways they can be assessed.

Verstehen

Participants will understand the relevance and usefulness of different levels and types of explainability in machine learning.

Anwenden

Participants will be familiar with the employment of several methods that yield explainable results, and they will be able to apply them to example problems.

Lern- bzw. Methodenkompetenz

Participants analyze and discuss scientific publications in the context of a given broader topic. Participants deepen and challenge their understanding of the taught concepts by designing and

answering short quizzes.

Sozialkompetenz

Participants successfully collaborate in small teams, they effectively exchange arguments and self-organize to produce a joint result within a given time frame.

Literatur:

- C. Molnar. "Interpretable Machine Learning - A Guide for Making Black Box Models Explainable" https://christophm.github.io/interpretable-ml-book/
- A. Thampi. "Interpretable AI - Building explainable machine learning systems", Manning, https://www.manning.com/
ai
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L.K., Müller, K.-R. (Editors). "Explainable AI: Interpreting, Explaining and Visualizing Deep Learning", Springer, 2019.
- HJ Escalante, S. Escalera, I. Guyon, X. Baró, Y. Güçlütürk, U. Güçlü, M. van Gerven (Editors) . "Explainable and Interpretable Models in Computer Vision and Machine Learning", Springer, 2018.
- Biran, Or, and Courtenay Cotton. "Explanation and justification in machine learning: A survey." In IJCAI-17 Workshop on ExplainableAI (XAI), p. 8. 2017, http://www.cs.columbia.edu/ orb/papers/xai_survey_pape
- Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint, 2017, https://arxiv.org/abs/1702.08608.
- R Guidotti, A Monreale, F Turini, D Pedreschi, F Giannotti. "A survey of methods for explaining black box models." arXiv preprint, 2018, https://arxiv.org/abs/1802.01933.

Verwendbarkeit des Moduls / Einpassung in den Musterstudienplan:

Das Modul ist im Kontext der folgenden Studienfächer/Vertiefungsrichtungen verwendbar:

[1] **Data Science (Bachelor of Science)**
(Po-Vers. 2020w | Vertiefungsrichtungen | Maschinelles Lernen / Artificial Intelligence (AI) | Introduction to Explainable Machine Learning)

[2] **Data Science (Bachelor of Science)**
(Po-Vers. 2020w | Vertiefungsrichtungen | Nicht gewählte Vertiefungsrichtungen | Introduction to Explainable Machine Learning)

Dieses Modul ist daneben auch in den Studienfächern "Artificial Intelligence (Master of Science)", "Data Science (Master of Science)", "Informatik (Bachelor of Science)", "Informatik (Master of Science)", "Medizintechnik (Master of Science)" verwendbar.

Studien-/Prüfungsleistungen:

Introduction to Explainable Machine Learning (Prüfungsnummer: 76981)
Prüfungsleistung, Klausur mit MultipleChoice, Dauer (in Minuten): 60
Anteil an der Berechnung der Modulnote: 100%
weitere Erläuterungen:
Answering the questions requires understanding of the concepts taught throughout the course and the ability to apply these concepts to specific example problems. The exam contains multiple-choice questions. It counts 100% of the course grade. By submitting small optional homework assignments, up to 20% of bonus points can be obtained, which will be added to the result of the exam.

Erstablegung: SS 2022, 1. Wdh.: WS 2022/2023
1. Prüfer: Thomas Seel

Organisatorisches:

StudOn-Kurs: https://www.studon.fau.de/crs4419539.html