Lehrform (teaching format) / SWS (hours per week): 4K

Kreditpunkte (credit points): 6

Turnus (frequency): every second winter term

Inhaltliche Voraussetzungen (content-related prior knowledge/skills): Machine learning basics

Sprache (language): English

Lehrende (teaching staff): Prof. Dr. Marvin Wright

<table>
<thead>
<tr>
<th>Studiengang (degree program)</th>
<th>Module</th>
<th>Semester</th>
</tr>
</thead>
<tbody>
<tr>
<td>Informatik (Master)</td>
<td>IMVP, IMVP-AI</td>
<td>ab 1.Sem.</td>
</tr>
<tr>
<td>AI and Intelligent Systems (Master)</td>
<td>AI-M-MLCS</td>
<td>from 2nd sem.</td>
</tr>
<tr>
<td>Management Information Systems (Master)</td>
<td>(MIS-INF3)</td>
<td>from 2nd sem.</td>
</tr>
<tr>
<td>Informatik (Bachelor)</td>
<td>(nur Freie Wahl)</td>
<td></td>
</tr>
</tbody>
</table>

Lernergebnisse / Learning Outcome:

- Understand the theoretical and practical basics of interpretable machine learning (IML), including general concepts and specific methods
- Be aware of pitfalls and challenges with IML methods
- Be able to implement IML methods yourself in Python (or R or another language of choice)
- Be able to apply IML methods in practice (on real data)
- Be able to understand and explain the results of IML methods

Inhalte / Contents:

- Dimensions and scope of interpretable machine learning
- Interpretable models, e.g., (generalized) linear models, rule-based models
- Feature effects: individual conditional expectation, partial dependence, accumulated local effects
- Shapley values and Shapley additive explanation (SHAP)
- Feature importance: permutation feature importance, conditional feature importance
- Local interpretable model-agnostic explanations (LIME)
- Counterfactuals and adversarial examples

Hinweise (remarks): The table lists only the primary / most specific modules to which this course is assigned.