Veranstaltung | Workshop

Interpretable Machine Learning: Explaining Black Box Models

In this workshop, Prof. Loecher will introduce methods from explainable Machine Learning.

Di
28.05.

Uhrzeit
12.30–15.30

Kosten
kostenfrei

Veranstalter/in
Institute for Data-Driven Digital Transformation (d-cube) in Kooperation mit der Methodenwerkstatt Statistik

Zum Institut

With AI models increasingly being deployed in high-stakes decision making, it is essential to "understand" key aspects of the models, such as

- How do parts of the model affect predictions?
- Why did the model make a certain prediction for an instance or for a group of instances ?

While I will focus mainly on traditional ML models, such as tree ensembles and regression (no deep NNs or transformers), many of the presented methods are "model agnostic", i.e. work for any supervised ML setting.

I plan to review the interpretation of linear regression models and then cover relevant diagnostic tools such as

- partial dependence plots
- several feature importance metrics
- SHAP values
- Counterfactual Explanations

This workshop will be hands-on, i.e. I have prepared data sets and code which we will work on together as a group.

If there is time I will address the pitfalls of these methods and discuss recent advances.