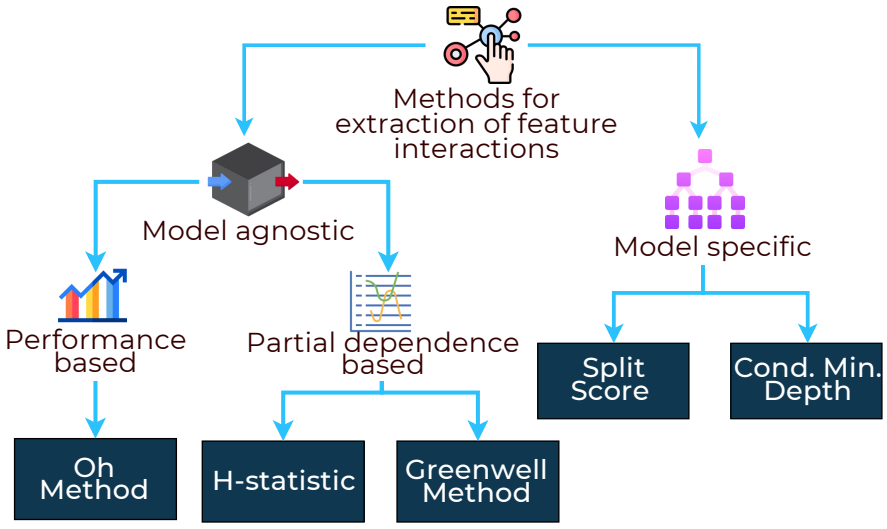
Methods for extraction of interactions from predictive models

Paweł Fijałkowski, Mateusz Krzyziński, Artur Żółkowski
Thesis supervisor: dr hab. inż. Przemysław Biecek, prof. uczelni

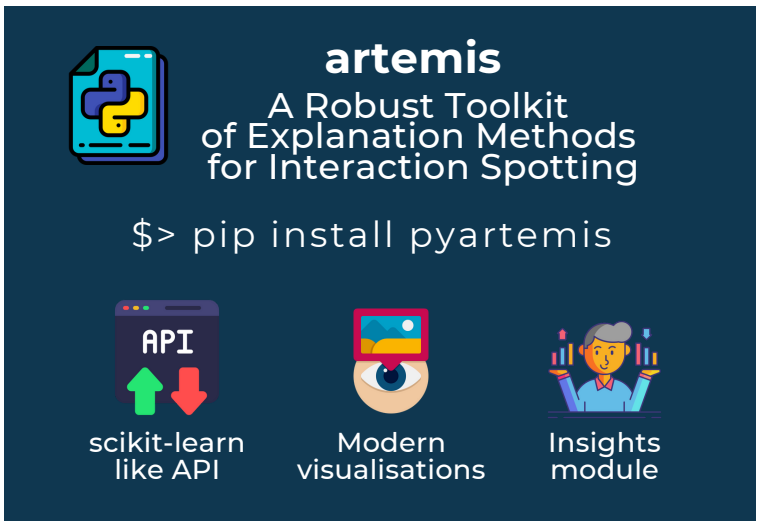
Interaction occurs when the **non-additive effect** of one feature on the target depends on the value of another feature. Interactions can be challenging to interpret, making it difficult to understand the underlying mechanisms driving the predictions. Moreover, they also cause problems and **misleading interpretations** of the models' explanations obtained by popular explainable artificial intelligence (**xAI**) methods.

Extracting interactions between features can significantly increase the **interpretability** of the model and, in some cases, may also result in improved **model performance**. We provide a **comprehensive review** and a **Python implementation** of the most well-known feature extraction interaction methods, both model-agnostic and model-specific.

I. METHODS



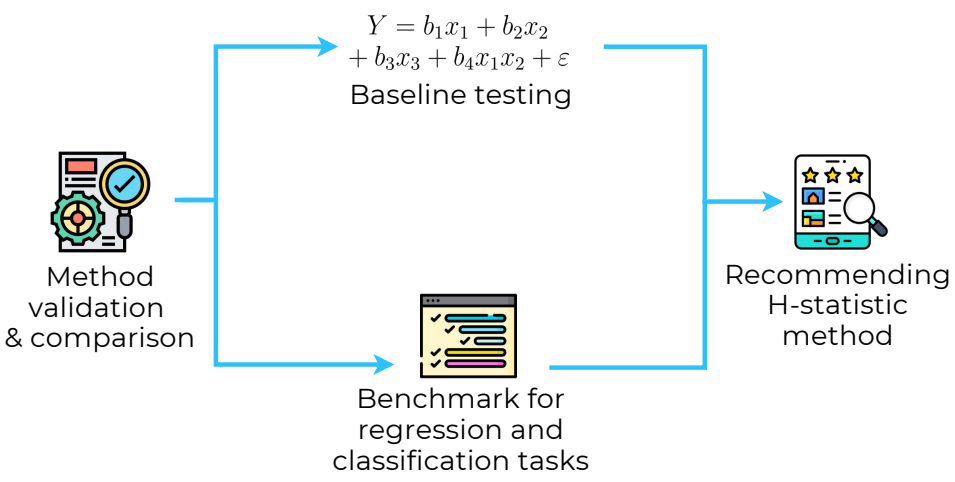
II. PACKAGE



We evaluate the effectiveness of selected methods through experiments on both **synthetic data** generated from known distributions and **toy data** with intuitive relationships. The results of conducted **benchmarks** indicate that the **H-statistic** should be the first choice when looking for a method for extracting interactions. However, other implemented techniques, like Greenwell Method, are also worth considering.

To illustrate the practical applications of the described methods and our implementation, we present an example of a **real-world use case** by applying the developed techniques to explain **stylometric-based models** created by the **NASK** National Research Institute. In a domain expert's opinion, interaction-based explanations provide a more **comprehensive insight** into the model's reasoning than single-feature explanations.

III. EXPERIMENTS



IV. USE CASE

{pawel.fijalkowski, mateusz.krzyzinski, artur.zolkowski}.stud@pw.edu.pl

Faculty of Mathematics and Information Science
WARSAW UNIVERSITY OF TECHNOLOGY

https://github.com/pyartemis/artemis

Acknowledgment. We would like to thank the domain expert from NASK, Inez Okulska, PhD for valuable discussions on the presented use case.