Comparative Analysis of Local Explanations Approaches

Amar Halilovic
Ulm University

Opis teme: Local explanation methods are essential in explainable artificial intelligence (XAI), providing insights into the decision-making processes of machine learning models at the instance level. These methods offer transparent, understandable explanations for specific predictions made by complex models, such as deep neural networks or ensemble methods. This comparative analysis evaluates and contrasts various local explanation techniques to assess their effectiveness, interpretability, and applicability across different domains and model architectures.

Zadaci i ciljevi: The analysis will begin with a review of prominent local explanation methods, including LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), and other perturbation-based techniques. Each method will be described in terms of its theoretical foundations, including how it generates explanations and its underlying assumptions about data and model behavior.
Key aspects to be evaluated in this comparative study include:

- Accuracy and Fidelity: How accurately do the explanations reflect the valid reasoning of the model? The fidelity of an explanation method will be tested against various models and datasets to determine its reliability.
- Interpretability: Are the explanations provided by these methods understandable by human users, including domain experts and laypersons? This involves assessing the clarity and coherence of the explanations.
- Scalability and Efficiency: How do these methods scale with increasing data size and model complexity? The computational efficiency of each method will also be scrutinized.
- Applicability: How versatile are these methods across different data types (tabular, text, image) and model architectures (from linear models to complex neural networks)?

Lista referenci:

1. Ribeiro, M. T., Singh, S., \& Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
2. Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Advances in neural information processing systems 30 (2017).
3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2018, April). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).
4. Hoffman, Robert R., et al. "Metrics for explainable AI: Challenges and prospects." arXiv preprint arXiv:1812.04608 (2018).