**<Title of the paper>**

Abstract: The field of Explanatory artificial intelligence is showing promising growth in recent years, thus giving researchers the option of deeply exploring the benefits and drawbacks of many different proposed models for solving the enigma of interpretability and explainability regarding machine learning models and their predictions. There are currently a number of techniques that can help to assist researchers in understanding the logic behind decisions made by various models, but the focus of this paper will mostly be on discussing and comparing two strong options, LIME (Local Interpretable Model Agnostic Explanations) and SHAP (Shapley Additive explanations). The proposed pipeline for the comparison will be given in a form of an Orange Data Mining workflow. Secondly, the paper aims to give a proposal of how a custom widget encapsulating the functionality of the LIME library can be integrated into the graphical interface, making it’s usability more appropriate towards less experienced users.

Keywords: <Keywords> Explainable artificial intelligence (XAI), interpretability, explainability, LIME, SHAP, Orange Data Mining

1. Motivation

The rapid rise of machine learning and artificial intelligence based techniques in recent years has provided many opportunities for enabling researchers from different branches to develop a variety of solutions for real world problems. These problems are distributed through various domains such as healthcare, education, business, scientific research and many more.

Some of the machine learning models are considered to have a black box approach which can be a problem when working with sensitive matters involving human lives. Black box models can’t give much of the necessary explanations that would be greatly needed and beneficial for researchers. Knowing that, in order to make models and predictions more transparent and interpretable we need to look for different approaches that would resolve those shortcomings of black box models. Entering the field of Explainable AI can give researches the necessary interpretability and explainabillity of models and predictions made by them.

Terms such as interpretability and explainability are often used in the Explanatory artificial intelligence (XAI) domain interchangeably and currently there are is not a clear consensus for their meaning, thus they appear underspecified. On the other hand, as argued in [1], the authors give a view in which these terms should be viewed on their own. They take a stance that interpretability is not enough by itself, rather that explainability is needed for people to trust black-box models. In their view, the goal of interpretability is “to describe the internals of a system in a way that is understandable to humans”. They also add that “in order for a system to be interpretable, it must produce descriptions that are simple enough for a person to understand using a vocabulary that is meaningful to the user”.

As stated in [2], the desired characteristics for explainers are that they have to be interpretable, i.e. “to provide qualitative understanding between the input variables and the response”. The meaning behind this definition also acknowledges that explanations need to be easy to understand.

It is also stated that interpretability is not a term that can be used in the same manner for different types of target audiences with various levels of expertise. In order to make a leap towards fair and ethical decision making, there is a need for discussing the transparency notion of interpretability as shown in [3].

Regarding interpretability Lipton proposes that there are two categories in which techniques and model properties fall, first being transparency and the second post hoc explanations. The two terms are defined as “how does the model work” and “what else can a model tell me”, respectively. It is also stated that transparency is on the opposite side of the spectrum from the black-box approach mentioned earlier.

Knowing all of that, there was a need to develop solutions that would solve this particular problem and provide some additional explanations for the inner logic behind machine learning models and predictions. A number of soutions were proposed, giving a variety of options to choose from. Linardatos et al give a comprehensive review of interpretability methods to explain any black-box models [4].

The main focus of this paper is not to show all of the available solutions, rather to give an understanding how two different models, LIME (Local Interpretable Model-Agnostic Explanations) [2] and SHAP (Shapley Additive xplanations) [5] work in order to explain the reasoning behind a machine learning model on a given dataset. Secondly, we propose an example of an alternative way of using LIME through Orange Data Mining as a custom Add-on widget.

We believe that this way of using LIME could make the library more accessible to a greater amount of people that don’t have the necessary technical skills, as Orange Data Mining provides a visual programming experience.

2. Research questions

There are a number of research papers discussing the comparison of different Explainable AI models including LIME and SHAP among other models. Some of the work is directed towards raising the question of trustable Explainable AI [7], while on the other hand there are proposals on how a theoretical framework could be used to enable a comparison between different models [8]. The use of Explainable AI in the healthcare domain has also been a topic of a number of papers [9][10][11][12] where LIME and SHAP were some of the models used. Evaluation and comparison of these two models has also been done in the domain of credit risk management where authors give a side-by-side review of the results obtained by these models [13].

The specific problem that we wanted to address is to make Explainable AI both more available and understandable to a larger audience that doesn’t have the necessary technical skills in order to do complex scientific work in this particular field. We wanted to empower audiences from different domains by giving them an option of understanding the logic behind predictions made by various Machine learning models. In order to do such a thing, we utilized the benefits of the paradigm of visual programming that is present in the Orange Data Mining software.

Orange Data Mining enables users to construct a complete Machine learning workflow by using a simple drag-and-drop interface that accommodates a vast number of useful widgets.

By proposing a custom made widget that encapsulates some of the functionality offered by LIME, we give users the ability of seeing and comparing results of prediction explanations done by LIME and SHAP models.

3. Methodology

Firstly, it is important to acknowledge that Orange Data Mining supports an Explain Add-on in their widget catalog. It consists of Feature Importance, Explain Model, Explain Prediction, Explain Predictions and ICE widgets. The inner core of Explain Prediction widget is the SHAP model of explainability, while we provide an alternative using LIME. Both SHAP and LIME are widely used but they have some distinct strengths and weaknesses, so it is beneficial to use them both in order to be more certain of the given results.

In order to implement a comparison system for LIME and SHAP models using Orange Data Mining, there was a need for creating a workflow that consists of a number of different steps that are encapsulated by the use of various widgets:

1. CSV File import widget that is used for loading the dataset of choice,

2. Data Sampler widget that divides data into parts for training and testing,

3. Data Table widget which is needed for the purpose of visualizing the dataset in a tabular form,

4. Select Columns widget used with the intent of choosing features (input columns) and the target variable (output column),

5. Machine learning model widget that is used for encapsulating different machine learning models (the model that the authors of this paper choose is Gradient Boosting),

6. Explain prediction widget that encapsulates some of the functionality of SHAP,

7. Explain Prediction LIME which is a custom widget designed to encapsulate the necessary functionality of LIME

The proposed implementation of the previously described workflow can be seen in Figure 1.

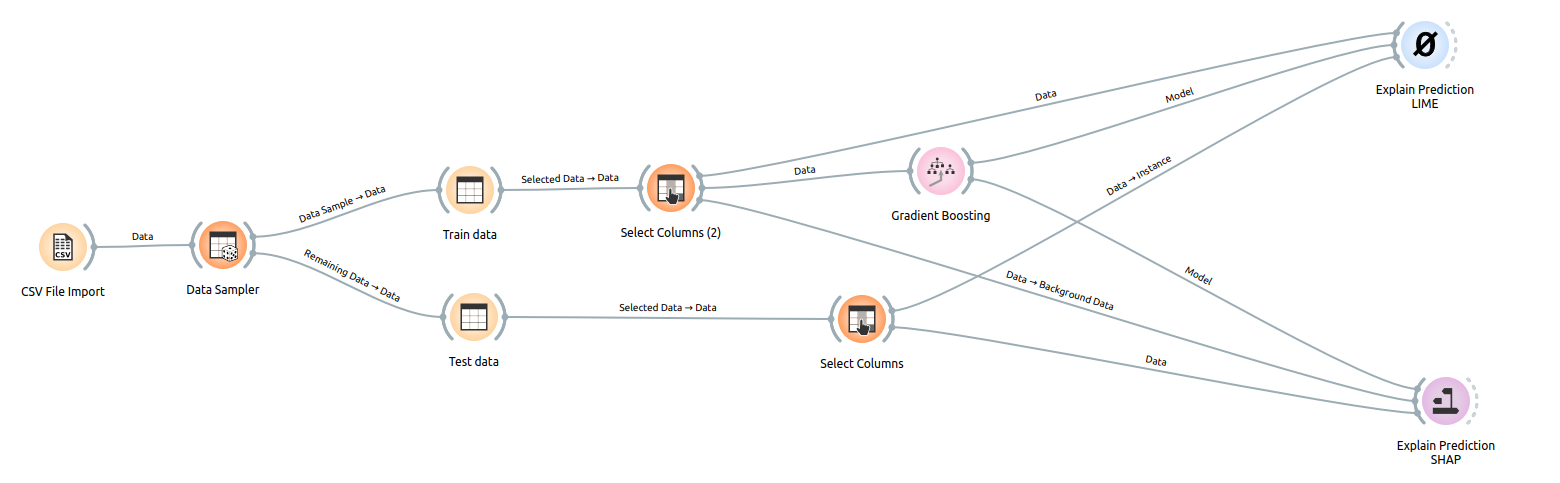


Figure 1 – Orange Data Mining workflow for the comparison of LIME and SHAP

Implementing the Explain Prediction LIME widget was done in a way that input signals leading to the widget were the same as to the Explain prediction widget, making it more intuitive for potential users. The input signals in question are the training data, the machine learning model and the chosen instance from the test data which will be used for explanation.

4. Solution/Discussion

The proposal of such a way of using the benefits of visual programming in order to explain complex predictions is more intuitive and straight forward than implementing traditional methods of programming. By encapsulating the functionality of LIME into a Orange Data Mining widget, the authors of this paper give a more user centric experience for the end consumer, thus enabling audiences from different backgrounds to equally participate in the discussion.

Implementing a workflow for comparing the outputs of two widely used Explainable AI models has the potential of giving much more useful insights than just relaying on one of them. This is especially needed in situations when working with dataset regarding the medical and healthcare domain, as human lives are on the line. Knowing that, the proposed comparison and functionality of the custom widget will be shown on various datasets from the medical field.

5. References

1

2

[3] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 3 (May-June 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[4] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable AI: A Review of Machine Learning Interpretability Methods,” *Entropy*, vol. 23, no. 1, p. 18, Dec. 2020, doi: 10.3390/e23010018.

[5] Lundberg, Scott M. and Su-In Lee. “A Unified Approach to Interpreting Model Predictions.” NIPS (2017).

[6] Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). [Using the ADAP learning algorithm to forecast the onset of diabetes mellitus](http://rexa.info/paper/04587c10a7c92baa01948f71f2513d5928fe8e81). In Proceedings of the Symposium on Computer Applications and Medical Care (pp. 261--265). IEEE Computer Society Press.

[7] Ignatiev, Alexey. "Towards Trustable Explainable AI." In *IJCAI*, pp. 5154-5158. 2020.

[8] *Sebastian Palacio, Adriano Lucieri, Mohsin Munir, Sheraz Ahmed, Jörn Hees, Andreas Dengel*; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 3766-3775

[9] Loh, H. W., Ooi, C. P., Seoni, S., Barua, P. D., Molinari, F., & Acharya, U. R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 107161.

[10] Knapič, S., Malhi, A., Saluja, R., & Främling, K. (2021). Explainable artificial intelligence for human decision support system in the medical domain. *Machine Learning and Knowledge Extraction*, *3*(3), 740-770.

[11] J. Duell, X. Fan, B. Burnett, G. Aarts and S. -M. Zhou, "A Comparison of Explanations Given by Explainable Artificial Intelligence Methods on Analysing Electronic Health Records," 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, 2021, pp. 1-4, doi: 10.1109/BHI50953.2021.9508618.

[12] Dave, D., Naik, H., Singhal, S., & Patel, P. (2020). Explainable ai meets healthcare: A study on heart disease dataset. *arXiv preprint arXiv:2011.03195*.

13