XAI evaluation

1 Introduction

This research wants to predict the addiction-likelihood of patients so that medical professionals can choose a different medicine that will have the same effect but not the same addictive properties. Several machine learning models will be employed to predict this, some of which are inherently explainable like decision trees. An artificial neural network, a black-box model, will be trained too. On the black-box model, different explainable AI methods will be used to generate an explanation for the model.

The different types of explanations, so the inherently explainable models and the explainable AI methods will be evaluated with a user test which will be described below. This user test will focus on which explanation is the most understandable, easiest to use and which one indicates biases the best.

2 Experimental design

This user test will be a within-subjects test so that the number of tested users can be kept to a minimum since 4 different models will be used. The dependent variables will be:

- Readability:

  How readable a user finds a specific model from 1 to 7, where 1 is not readable at all and 7 is perfectly readable.

- Which model describes the decision process the best:

  The user can choose between one of the given models.

- Which model describes its bias, if it has any, the best:

  Here the user can also choose between one of the given models.

And the independent variable will be:

- The explainable AI method: decision tree, random forest, K-NN, LIME, SHAP, ANCHORS.

Here are some possible questions that aim to evaluate these variables:

2.1 Readability

On a scale of 1 to 7, how would you rate the readability of the explanation provided by [XAI technique]?

What aspects of the explanation by [XAI technique] made it easy or difficult to read?

2.2 Understandability

How well did you understand the reasoning behind the AI's prediction after reading the explanation provided by [XAI technique]? (Scale of 1 to 7)

Could you clearly see the connection between the patient's profile and the AI's prediction based on the explanation? Please elaborate.

Which elements of the explanation helped you understand the AI's decision-making process the most?

2.3 Bias Identification Capability

How effectively do you think the explanation by [XAI technique] highlights any potential biases in the AI model's decision? (Scale of 1 to 7)

Did the explanation provide sufficient information to identify if the decision was biased towards certain patient profiles or characteristics? Please provide examples if any.

In what ways could the explanation be improved to better reveal biases in the AI model's decisions?

2.4 Trustworthiness

On a scale of 1 to 7, how much trust would you place in the AI's prediction after reviewing the explanation by [XAI technique]?

How did the explanation affect your level of confidence in the AI's prediction?

Are there specific aspects of the explanation that increased or decreased your trust in the AI model's decision? Please describe.

2.5 Overall Perception and Preference

Which XAI technique (LIME, SHAP, Anchors, et cetera) provided the most useful explanation for you, and why?

Rank the XAI techniques in order of preference for use in your practice. Please explain your ranking.

What improvements or additional information would make the explanations more useful in your decision-making process?

2.6 Open-ended Feedback

What are your general impressions of using XAI techniques for understanding AI predictions in medical decision-making?

How do you perceive the role of these XAI explanations in enhancing the transparency and accountability of AI systems in healthcare?

3 Participants

As already partially talked about in section 1 the participants in this user test will solely be medical professionals. For this user test the participant count somewhere between 75 and 100 will suffice since this is not a between-subjects test where at least 50 people are needed per independent variable type. There are not any real requirements for the users besides that they are medical professionals who prescribe medicine.

4 Materials

The only thing needed is a laptop for the user to fill in the online evaluation form.

5 Procedure

- **Introduction Session:** Brief participants on the study's goal, the importance of XAI in medical decision-making, and an overview of the XAI methods being evaluated. Provide a tutorial on how to interpret the explanations generated by the individual XAI methods.
- **Evaluation Session:** Participants interact with a series of case scenarios where the AI model predicts a high risk of addiction based on the patient's profile. For each case, participants are presented with predictions from decision trees, random foresta and K-NN, as well as explanations from LIME, SHAP, and Anchors. Participants rate each explanation based on readability, understandability, and their ability to reveal biases in the model's decision-making process. Collect qualitative feedback through open-ended questions about each method's perceived strengths and weaknesses in conveying meaningful insights into the AI's decision process.
- **Comparison Session:** After reviewing all explanations, participants rank the methods based on preference for real-world application, clarity of explanation, and trustworthiness. Gather insights on how each explanation might influence their decision-making process in clinical practice.
- **Feedback and Debrief:** Conclude the study with a feedback session, allowing participants to share their overall impressions and suggestions for improving the explainability of AI predictions in medicine. Discuss the potential impact of these XAI methods on reducing bias and enhancing patient care.