Radboud Universiteit

Getting the Best of Both Worlds?

Combining Local and Global Methods to Make AI Explainable

Masterthesis in Artificial Intelligence

Author:
Daphne Lenders
Radboud University Nijmegen

Internal Supervisor:
Luca Ambrogioni
Radboud University Nijmegen

External Supervisor:
Iana Nurdinova
Avanade Netherlands BV

April 2020

Acknowledgements

Firstly, I would like to thank my supervisor Iana Nurdinova for her guidance throughout this project. I am grateful for the feedback I got on the way and for the great collaboration I had with her. I would also like to thank my second supervisor, Luca Ambrogioni, and his insights and advice on this work.

I am thankful for the nice time I had during my internship at Avanade. It was a great opportunity to learn from the experience of others and be part of a friendly and enthusiastic team. I am especially thankful to those, who participated in my usability study.

I am grateful for the support I got from my family and friends. In particular, I would like to thank my parents who always provided me with financial and emotional support and for my sister Elène, who was a great company during many work-at-home Skype sessions. Lastly, special thanks go to my university "study-buddies", Masha Tsfasman, Hugo Chateau Laurent and Anna Pillar. It was a pleasure to work side by side with them, and an even greater pleasure to enjoy their company at the end of working days.

Abstract

With artificial intelligence (AI) models becoming increasingly complex, attention about making their decision processes more transparent has been growing. SHAP and GIRP are two techniques to do so and were investigated in this study. SHAP is a local explanation method that assigns importance values to each feature of an input, indicating how impacting that feature is for the corresponding output. GIRP is a global explanation method that explains the inner-working of an AI model holistically, by summarizing it into one decision tree. In this work, we set up a usability study to compare how SHAP, GIRP and a combination of both explanations facilitate users' understanding of AI models. It was found that in terms of self-rated understanding SHAP explanations were more preferred than GIRP, while the opposite was the case for objectively-measure understanding. Combining both explanations was at no point found to be the most beneficial option. This was hypothesised to be the case, because of users experiencing an information overload when being presented with both explanations. There were indications that this effect might be diminished by increased task relevance or increased users' motivation to perform well on the task. Whether this proposition holds, remains to be tested. Next to the usability of explanations, we were interested in their fidelity and stability. Fidelity concerts the extent to which explanations accurately reflect the inner-workings of their original models, while stability refers to how robust explanations are to small changes in their input. The results for the fidelity of explanations were somewhat mixed and we advise to extend current evaluation measures. While we did not find any reason to doubt the stability of SHAP, we found GIRP explanations to be quite unstable. Therefore, we advise to improve this explanation method or search for an alternative.

Contents

1 Introduction

At the moment one of the most popular applications in the field of artificial intelligence is machine learning. Some techniques, like deep neural networks, have yielded impressive performances and thus the number of domains in which the algorithms can be used is growing. With data being recorded in nearly every area of human life, it has become possible to automate any sort of decision process. While some processes, like whether a movie is being recommended to a Netflix user, are less impacting for individuals, there are other decisions, e.g. whether a person gets a job or not, that can highly influence a person's life. Because of these critical domains, the need has been identified to only use machine learning algorithms that produce fair and well-reasoned outcomes.

However, the problem is that many of the well-performing algorithms are very complex and hard to understand. Commonly referred to as "black-boxes", they receive some input and produce an output without giving any indication of how the output was derived. Thus, by just using these algorithms there is no way to tell whether the decision process was fair and logical. Because of this problem, the research field of explainable AI (xAI) has emerged. The aim of the field is to make black-box models more transparent and give users better ideas of how complex models derive their outputs.

Roughly speaking, two kinds of techniques for creating explanations can be distinguished: global explanation techniques, that aim to explain a machine learning models as a whole; and local explanation techniques, that explain the behaviour of the model for a specific input.

In this study looked at one type of each explanation category. The local explanation method focused on is called SHAP. With this way of explaining machine learning models, importance values are assigned to each feature of a given input instance, indicating how impacting each feature is for the model's output. The name of the global explanation method we investigated, is "Global Model Interpretation via Recursive Partitioning" (or short, GIRP). With this explanation method a black-box model is translated into a decision tree, that summarizes the main decision processes going on in the model.

Regardless, of which explanation technique is used, it is crucial that explanations can be easily understood and interpreted. The main goal of xAI is, after all, to make decision processes of AI models more transparent and less complex for everyone who is affected by them, including non-computer experts.

Since SHAP and GIRP highlight different characteristics of the models they explain, we wanted to establish how they compare to each other from this usability perspective. In other words, we wanted to find out how well users understand the two types of explanations. Since "understanding" is a rather broad term, we defined it in an objective and subjective way. Objective understanding refers to how well users can utilize explanations to derive conclusions about AI models, while subjective understanding relates to users' self-rated satisfaction with the explanations.

Next to comparing SHAP and GIRP explanations individually, we were interested to find out how users' understanding is facilitated by presenting both explanations. As mentioned, the two explanations techniques highlight different aspects of their original models, thus we believed that combining both can be beneficial to gain a fuller understanding of the corresponding decision processes. To find an answer to our questions and test our assumptions, we set up a usability study where participants were presented with SHAP, GIRP or a combination of both explanations. Both aspects of their subjective and objective understanding were measured.

Next to usability, there are two other criteria explanations should satisfy: fidelity and stability. Fidelity concerns the accuracy with which explanations reflect the inner-workings of their original models, while the stability of explanations denotes how robust explanations are to slight changes in the input they were based on. Both are crucial

criteria for the overall quality of explanations, and thus it was investigated to which extent SHAP and GIRP explanations satisfy each criterion.

In the remainder of this thesis, we will first give a more extensive overview of the importance of xAI, as well as different explanatory approaches and the different criteria explanations should meet. What follows is the formulation of the research questions in Section 3. In Section 4, we will demonstrate how the two explanation methods of interest were implemented for two classification problems. The results of the implementation will be shown in Section 5. A description of how the usability study was set up, as well as a result analysis and discussion, will be included in Section 6. In Section 7, we will illustrate how the fidelity of SHAP and GIRP were measured and discuss the results. After this, we will discuss in Section 8 how the final explanation criterion, namely stability, was measured and to what extent the explanations satisfy this criterion. In Section 9 we will summarize our derived conclusions of this study and give suggestions for further research.

2 Theoretical Background

2.1 Why Explainable AI?

The research of explainable AI is not as new as many people might think. Already in the 80s the need to make explainable systems, that not only produce results but also justify them, has been identified [36]. However, it has only been in recent years that the research field has gained huge attention and advances have been made.

The rise in interest is largely due to the recent advances in AI or, more specifically, machine learning. Traditional techniques that have been used in the early years of AI, like rule extractions or decision trees, are quite transparent and easy to understand by nature. However, newer and better-performing algorithms like deep neural networks or ensemble methods, are inherently less interpretable. A general rule of thumb seems to be that the better the performance of a model, the more complex and hard-to-understand the model is [9]. Given also the increasingly socially relevant domains in which machine learning is deployed, concerns have been growing. Work has been done on detecting criminals based on tweets [15] or automatically ranking job applicants based on their CV [11]. The decisions made by these systems can have tremendously high impacts on peoples' life and thus should be fair and well-reasoned.

This concern has been publicly brought to light by the case study of the COMPAS algorithm, which is used to predict reoffending rates of criminals. Ever since it was claimed that the algorithm wrongly predicts higher recidivism rates for Afro-American than for Caucasian men, interest in detecting bias in machine learning algorithms has risen [45, 48].

Another recent development, that highlights the need for explainable AI systems, is the enforcement of the GDPR in 2018. Next to the many rights about how data is gathered and stored, the GDPR provides individuals with the right to explanation as well as the right to non-discrimination [17]. In short, the former one gives persons who are affected by automated decision processes the right to be presented with the substantial logic of how a decision was made [17]. The right to non-discrimination ensures that any automated decision process cannot be based on the processing of sensitive information like a person's ethnicity, religion or sexual orientation [17].

Explainable AI is not only critical in the context of fairness and justice but can also be of more commercial value. From a developers' perspective, gaining insights into why a model produces the output that it does can help in changing underlying mechanisms of algorithms and hence improving their performance [43]. From a users' perspective, on the other hand, users are more likely to trust and therefore use a model, if they

understand its inner-workings. Andras et al. have argued that the explainability of a system is the key component to re-establish general publics' trust in AI systems [3].

2.2 Different Kind of Explanations

As already mentioned, two types of explanations to make an AI system more transparent can be distinguished: local and global explanations. To demonstrate different sorts of these explanation techniques, we will make use of a running example in this subsection. For this purpose, imagine a machine learning model that has been trained on CVs to predict whether an applicant gets hired or not. Explainable AI aims to not only let the algorithm produce an output, but also let it generate an explanation of its decision process.

2.2.1 Local Explanations

Local explanations are meant to explain a models' behaviour on one particular input instance.

One of the oldest local explanation techniques are case-based methods [38]. Here explanations consist of instances of the models' training set, that are similar to the input for which the model needs to be explained. When also providing the labels of these similar instances, users can get an idea which kind of features lead to the generated output. In case of a recruitment system, one might see that a candidate got hired, because s/he had a similar educational background as other employees that already got hired before. A problem with this approach is, that new input instances might not be very similar to already known instances in the training set, which can make it hard to give explanations for these.

Another local explanation method is the use of counterfactuals. Counterfactuals show which features of the input of question needs to be changed, for the output to change. Going back to the example of a recruitment system, a counterfactual explanation could show that if an applicant had one year less of working experience, s/he would not have been hired. Research has suggested that counterfactuals match the way humans tend to reason about decisions [6]. However, a disadvantage of this approach is that different inputs can have multiple counterfactuals. It then becomes hard to choose which counterfactuals should be presented as an explanation for an AI system. Referred to as the Rashomon effect [35] two counterfactuals can both be valid but can seem contradictory to each other, which would be confusing for users.

A possibly less confusing, but still valid way of explaining decision locally are Scoped Rules or so-called Anchors. Anchors are similar to counterfactuals, but rather than explaining which features should change for the decision to change, they show which features cannot change without the decision changing as well [42]. An anchor explanation is given by providing a rule of the nature *IF* x *THEN* y, that shows which feature x "anchors" prediction y. Along with this rule a precision of that rule is provided, which shows for how many instances of a dataset the rule applies. In the case of a recruitment system, an anchor could show that an applicant is hired because of their previous work experience at a well-known and prestigious company. This would, in turn, imply that in this case, the other CV entries of the applicant do not make any impact on the final decision.

A disadvantage that counterfactuals and anchors have in common, is that they do not provide an exhaustive overview of how a change in feature value affects a decision. A counterfactual might e.g. state that an applicant would not get hired with one year less of work experience, but this does not give any information on how the decision would change if an applicant had more work experience than currently listed on their CV. Individual Conditional Expectation (ICE) plots can be used to create more complete

pictures of how the decision outcome is dependent on a feature value [16]. In these plots, the feature of question is represented on the x-axis, while the outcome variable is plotted on the y-axis. This visual representation gives a quick but complete overview of how a decision can change with each different value of a feature.

Counterfactuals, anchors and ICE plots give a good impression of the impact of one feature at a time. However, if one would want to understand the effect of multiple features, looking at multiple of this explanation types might be rather time consuming. Contrary, feature importance values are a very concise way to explain decisions locally [1]. They summarize the impact of each feature in a decision process with one number, which shows how important the feature was for generating the output of a model. In the case of a recruitment system, it might show that features related to an applicant's skills have higher importance values than features related to an applicant's hobbies.

Due to the ease with which feature importance values can be interpreted methods that can compute these values, like LIME [41], DeepLIFT [44] and SHAP [29] have become popular. Out of these methods, the most promising one is SHAP, which has its foundations in coalitional game theory [29]. In game theory, so-called Shapely values are computed to determine how much payoff each player in a team of players should get. Imagine e.g. a company that has a yearly profit and some employees. Since every employee has different skills and working hours, one would like to distribute this profit among them, such that the individual payoff reflects how much each employee contributed to the profit. To calculate the individual payoff for employee X, one would compare for each possible coalition of employees the company's payoff with and without X. By averaging this marginal contribution over all coalitions, the individual payoff for employee X is obtained [22]. With the same principle, SHAP values can be computed for machine learning models. In the case of the recruitment system, the different employees in the example above represent the different features of a CV and the profit represents the output of the model. The downside of this approach is, that it takes a long time to calculate the values, since the number of marginal contributions that need to be calculated, grows exponentially with the number of features that are considered. Despite this shortcoming, SHAP is still popular since it has a solid theoretical foundation and it provides more stable outputs than methods like LIME, which may not have one unique solution for one input [22]. Because of its solid theoretical background, this study will further investigate the potential of SHAP as a local AI explanation method.

2.2.2 Global Explanations

Sometimes we do not want to examine a model's behaviour on one particular input but rather want to understand its inner-workings holistically. This is possible using global explanation methods. The advantage of global over local explanations is, that it can suffice to just inspect one global explanation to understand a model. In the case of local explanations, always multiple instances need to be inspected to get a fuller understanding of it.

The most straightforward way of gaining global explanations is by building machine learning models that are self-explanatory, to begin with. Examples of inherently transparent models include decision rules, decision trees or regression models [35]. Decision trees and decision rules use data to derive global rules of the nature *IF* x *THEN* y for a classification problem. These rules can be used to understand how decisions in a classification problem are made [26]. Regression models aim to estimate the relationship between an outcome variable and several predictors, by assigning a coefficient to each predictor. This coefficient is indicative of the relationship between the predictor and outcome variable and thus makes the regression model more interpretable than black-box models. While decision rules or regression models give satisfactory performances

for some machine learning problems, they fail to give accurate results for tasks of more complex nature. This, in turn, is the reason, why these methods are barely used in practice and why studies in xAI focus on extracting global explanations from opaque, yet well-performing models like deep neural networks or random forests.

One approach of obtaining a global explanation from these models is by averaging local feature importance values over different instances. This is possible using e.g. SHAP values. To go back to the example of the recruitment system, one might see that for all possible input instances, features relating to an applicant's skill level have the highest average feature importance value. The disadvantage of this approach is that it does not give any information about how feature importance values change depending on the value of the feature. In the case of the recruitment system, it may be important that applicants have certain skills relating to the job, while other skills may be less important for the position. Thus, the feature importance values for some skills may be very high but for others, they may be quite low. This information is not provided when only looking at average importance values. A way to solve this problem is using partial dependence plots as an explanation method [7]. This visual way of explaining AI models globally is similar to the ICE plots, mentioned in the previous section. However, differently than ICE plots they are not made for one specific output, but rather show how any output change on average with a change in feature value. Again, the feature value is typically represented on the x-axis of the plot, while the output value is visualized on the y-axis. The quite obvious downside of this approach is that only the relation between the output and one particular feature can be plotted at a time. To get a complete understanding of the model, one had to examine the partial dependence plots of all different features. Especially in models that are trained on a high number of features, this is not feasible. Moreover, 2-dimensional partial dependence plots do not give any information about the interaction between features. While it is possible to produce multi-dimensional plots, those may be very complex and hard to read.

A more concise way of giving global explanations is among others offered by Zilke et al. [51]. In this work, the authors describe an algorithm through which global decision rules can be extracted from deep neural networks. These rules are based on the activation of the neurons in different layers of the network and give a good understanding of which feature values lead to which decisions. Similar work has been done to extract decision rules from other black-box models, such as support vector machines or random forests [31, 32]. A downside of these techniques is that they only work for a specific type of model. The algorithm used to extract rules of deep neural networks, cannot be used to extract rules of random forests. It would save time and efforts to have an approach through which global explanations can be extracted model-independently. The use of surrogate models may be a solution for this. Surrogate models are transparent models like decision trees or decision rules, that are translated from black-box models. The traditional way of obtaining them is to train the chosen surrogate model on the dataset and outputs of the original model [46]. The assumption is that by being trained on the outputs of the original model, the simple model will reflect its decision process. Though this approach is model-independent and very straightforward, it comes with an obvious downside. Because of the simplicity through which the surrogate model is obtained, there is no guarantee that it will reflect the decision process of the original model. In other words, the fidelity of the surrogate model may be very low. Recognizing this shortcoming Yang et al. have recently proposed another method to generate surrogate models: 'Global Model Interpretation Via Recursive Partitioning' (GIRP) [49]. With GIRP a decision tree is learned from the local feature importance values of a black-box model. This is done in a recursive manner, where the best splitting variable in a tree is the one that maximizes the difference in average feature importance values for the left and right subtree (for a more detailed explanation of the GIRP algorithm refer

to Section 4) [49]. While this approach is intuitively more promising than traditional techniques, its quality still needs to be investigated, which will be one of the goals of this study.

2.3 Criteria of Explanations

As has become apparent from the previous sections, there are several criteria local and global explanations need to fulfil. In this section, we are going to discuss each of these criteria individually. Since in our study we will focus on local explanation method SHAP and global explanation method GIRP, it will be described to what extent these methods already satisfy the criteria.

2.3.1 Usability

Arguably, the most important goal of xAI is to make complex machine learning models more understandable and transparent to its end users, who might not be computer experts. Despite this clear goal, studies have revealed that many researchers do not take users into account when designing explanations [34]. As a result, the generated explanations may seem understandable to the ones who designed them but are far from comprehensible to real users [34]. Based on these finding, the need for a user-centred approach to usability has emerged and more research has been dedicated to this field. Petkovic et al. e.g. tested their explanations by using a self-rating scale, asking participants how much they understood the provided explanations [40]. Ribeiro et al. tested users' comprehension by investigating how explanations helped them to find bias in classifiers [41]. Also for the explanation methods of interest in this study, some attempts have been made to study their usability. Lundberg and Lee have e.g. conducted a study where they presented users with different classification problems and asked them to rate how much they agreed with the assigned SHAP importance values of each feature [29]. Here it was shown that the feature importance values matched users' intuition of how AI systems should explain themselves [29]. In another study, it is demonstrated that small and concise decision trees, as generated by GIRP, are generally well understood by non-experts [24]. In this study participants had to utilize these types of explanations to make predictions about the behaviour of AI models.

Though research in the area is growing, there are still some questions left unanswered. Many of the user-centred studies only focus on the comprehensibility of one explanation type, rather investigating multiple explanations at once. Therefore, it is unclear how the usability of explanations compare to each other and how different explanation types may be combined. Especially the combination of local and global explanation methods seems to be worthwhile investigating, as together they may facilitate users' understanding of AI models [1, 46]. The global explanations can provide a more holistic overview of the model's working, while the local explanations support this overview with concrete examples. To validate this idea, a usability study still needs to be conducted.

Another point that needs to be addressed, is how different aspects of understanding relate to each other. Many studies only measure how users subjectively rate their understanding or how they utilize explanations to derive knowledge about AI models. To our best knowledge, there are no studies that attempt to measure both these objective- and subjective aspects of understanding. Understanding how global and local explanation methods facilitate these different comprehension-levels, will be another goal of this study.

2.3.2 Fidelity

One concern regarding local and global explanations is how well they explain the decision processes of their original model. This quality is also referred to as fidelity. Explanations are always more simplified than the model they originate from, however, if they are too simple they might not accurately reflect the inner-workings of their original model. A consequence of presenting non-faithful explanations to users of AI systems is that we create an illusion of control [8]. We might look at an explanation of an AI system and judge it to be fair, but in reality, the system still contains bias that is not indicated by the explanation.

SHAP In comparison to GIRP, some effort has been made to study the fidelity of SHAP. In fact, the development of SHAP explanations was partly motivated by the need for more faithful explanations. In the work where the explanation method originated from, the authors demonstrate that SHAP is the only feature importance tool that satisfies all three of the following fidelity criteria: local accuracy, missingness and consistency [29]. The local accuracy requirement ensures that all feature importance values for an input instance sum up to the corresponding output of a model. The missingness criterion states that any feature with a missing input value should have a feature importance value of 0. Lastly, the consistency property requires that if a model changes such that a feature has a higher impact on that model, the importance value of that feature should not decrease.

With these three criteria already satisfied, the fidelity of SHAP is generally believed to be high. Nevertheless, multiple researchers have highlighted the idea of sanity checks, to make sure that local explanations indeed reflect the workings of the model they originate from [43, 27]. A method to study the fidelity of local explanations has been proposed by Lin et al. [27]. The main idea behind their approach is that if a feature with a high importance value is indeed important, leaving out the feature should lead to change's in the model's prediction. The authors quantify this intuition through their so-called "Impact Score". This score measures the percentage of instances for which the prediction or the confidence of the prediction will change if important features are left out from the prediction process [27].

GIRP While the fidelity of SHAP has already been measured to some extent, it is completely unknown how faithful GIRP explanations are to their original models. In general, it is believed that surrogate explanations like GIRP do not provide a high fidelity to their original models. After all, the models they originate from are usually very big and complex. Trying to catch every detail of them would result in equally complex explanations which would defeat the purpose of xAI [33]. Nevertheless, a certain degree of fidelity must be fulfilled. A way to study faithfulness of surrogate models is proposed by Messalas et al. with their Top_jSimilarity method [33]. This method assumes that a surrogate model is faithful to its original one if both models base their decisions on similar features. To determine this, one can compute for multiple instances of a dataset the local feature importance values for the original and the surrogate model. The Top_jSimilarity measure is derived by seeing how similar the most important features for the surrogate and original model are. The authors state, that if for more than 80% of all instances, both models base their decisions on the same five most important features, the surrogate model can be called faithful to its original one.

Of course, it is not only important that decisions are based on similar features, but that the output of the original and the surrogate model are comparable. To compare the output of the two models we can use an agreement measure like Cohen's Kappa [52]. The Cohen's Kappa coefficient measures how similar the output of the surrogate model

is to the original while controlling for the expected similarity that would be obtained by chance.

2.3.3 Stability

Another criterion for local and global explanations is their stability. This quality refers to the degree to which the explanation of a model changes when being provided with different input.

SHAP For local explanations stability can be measured, by investigating how a model's behaviour for two similar input instances is explained. Ideally, we expect that if two inputs for a model are similar the explanations for both inputs are similar as well. This quality is very important to establish trust in explanations. If local explanations vary too much for similar inputs, it becomes hard to derive general patterns from them.

Recently, it has been shown that SHAP explanations are not guaranteed to be very stable [2]. Therefore, more attention has been paid on how to measure the stability of local explanation methods. One of them is the $\text{Sens}_(\text{Max})$ measure as proposed by Yeh et al. [50]. With this measurement, several input instances are sampled and randomly perturbed. For each input instance, the distance between the feature importance values before and after perturbation is measured. The maximum obtained distance is then taken as an indication of the stability of the SHAP values.

GIRP Stability can also be measured for surrogate models, like the decision tree generated by GIRP. As mentioned before, surrogate models are based on the dataset the black-box model was trained on. Usually, this is a random sample of a larger dataset where the rest of the data is held out for validation and testing purposes. Friedler et al. point out the sample that the surrogate model is based on should not have much impact on the nature of the model [12]. In other words, generating the surrogate model based on a different sample should not significantly change the size, nodes or leaves of the resulting decision tree. Thus, the stability of surrogate models can be tested by investigating how the surrogate model changes when another random sample is drawn to generate it. The fidelity measures resulting from different samples can indicate how much surrogate models vary. By visually inspecting some of the resulting trees, it can be seen which aspects of the tree are most prone to instability. Just like for SHAP, establishing stability in GIRP is important to gain trust in explanations. This especially holds since decision trees are known to be quite unstable [25]. If one tree is very different from the other it is hard for users to tell which tree reflects the inner-workings of the model it originates from.

3 Research Questions

The main goal of this research is to compare how global, local and a combinaton of both explanations affect users' understanding of AI systems. This goal can be translated into the following research question:

RQ 1: *How is users' understanding of a model affected by the presentation of SHAP explanations, compared to GIRP explanations or a combination of both?*

Since "understanding" is a rather broad term, the first research question will be split into three sub-questions. With the first one, we measure users' ability to deploy explanations to answer basic comprehension questions about AI models. The focus of the second sub-question lies on users' self-rated understanding of the explanations, while

the third one is about measuring users' ability to utilize explanations when choosing between biased vs. non-biased models. Note that 1.1 and 1.3 relate to users' objective understanding of the explanations and their corresponding models, while RQ 1.2 relates to users subjectively measured understanding.

RQ 1.1: *How is users' ability to answer comprehension questions about models affected by the presentation of SHAP explanations, compared to GIRP explanations or a combination of both?*

RQ 1.2: *How do users' self-rated satisfaction of SHAP explanations, compare to their self-rated satisfaction of GIRP explanations or a combination of both explanations?*

RQ 1.3: *How is users' ability to detect bias in models affected by the presentation of SHAP explanations, compared to GIRP explanations or a combination of both?*

As mentioned in the previous section not just the usability of explanations is important, but also their fidelity and stability should be taken into account. Both fidelity and stability can be influenced by a lot of factors, such as the data the explanations derive from or the type of model they aim to explain. As it lies not within the scope of the study to consider all these variations, the last research questions will be more of exploratory nature:

RQ 2: *What is the fidelity of SHAP explanations to their original models, as measured by the Impact Score proposed by Lin et al.[27]?*

RQ 3: *What is the fidelity of GIRP explanations to their original models, as estimated by the Top_jSimilarity measure [33] and Cohen's kappa?*

RQ 4: *What is the stability of SHAP explanations as estimated by the $Sens_{Max}$ measure as proposed by Yeh et al. [50]?*

RQ 5: *What is the stability of GIRP explanations as estimated by its performance on different train-test-splits?*

If it turns out that SHAP and GIRP provide faithful and stable explanations, both explanation techniques can safely be combined, should it be found that combining them facilitates users' understanding of models.

4 Implementation

To answer the research questions both the SHAP and GIRP explanations needed to be implemented. For this purpose, an XGBoost classifier was trained on two datasets for different prediction tasks. How the classifier was trained and how explanations were extracted from it, will be the topic of this section. All implementation was made in Python and the final code can be obtained on GitHub[1]

4.1 Classification Tasks

The classification tasks the explanations were generated for, were chosen in such a way that their nature is already somewhat intuitive for users without much experience in

[1] `www.github.com/DaphneLenders`

machine learning. Out of this reason and due to the scope of this study, the classification tasks were also chosen to be on tabular data only.

4.1.1 Portuguese Class problem

The first dataset that was used to implement the different AI explanations was the "Portuguese Class" dataset, obtained from kaggle[2]. This dataset contains information from 649 different secondary school students following a Portuguese class. Among others, it has data about the students' gender, their age, their extracurricular activities and their past school performance. The classification task that was chosen for this problem, was predicting whether a student passes or fails the second exam of the course.

Preprocessing Before the XGBoost Classifier was trained, several preprocessing steps were taken. The original dataset consisted of 30 features and to not overload participants of the usability study with too much information, it was decided to reduce this number of features for the generated explanations. Many features did not prove to have a high correlation to a student's grade in an exam (e.g. the family size of students or their travel time to school) and were thus removed. The only exceptions for this were the variables "gender" and "age": even though they were not highly correlated with exam grades, it was decided to use these variables either way since they intuitively are the most basic terms to describe a student.

One final preprocessing step was changing the range of the variable 'G1' and the to be predicted variable 'G2', both of which originally ranged between 0 and 20. Since the aim of the classification task was to predict whether a student passes or fails the second exam of the course, the variable G2 was binarized, such that any value greater than 10 was translated to 'pass' and any value lower than 10 to 'fail'. The range of the variable 'G1' was changed to be 0-10 since this is a more common range for grading systems. The features that were finally used for the classification tasks, as well as their range/number of levels can be seen in Figure 1.

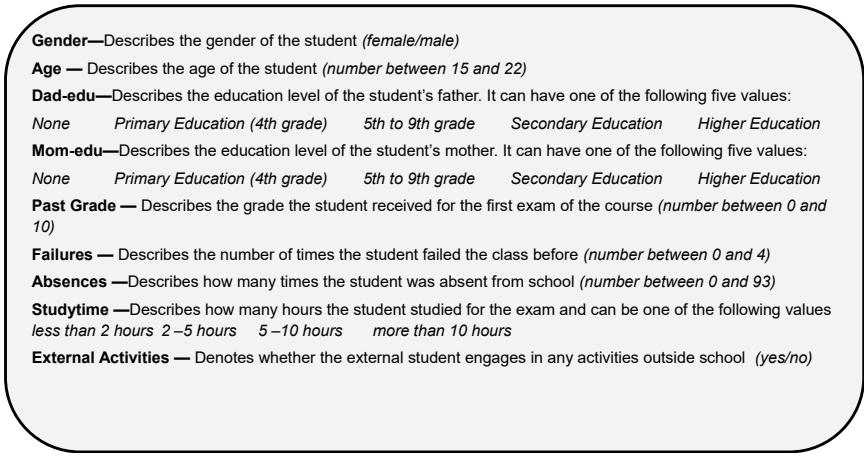


Figure 1: Variables used for the first classification task

Training The XGBoost classifier was trained on a training set consisting of 80% of all data instances. 10% of the instances were held out as a validation set and another 10% as a test set. The python package `xgboost` was used to implement the classifier. Since it was not the aim of the study to optimize the classification performance, but merely

[2] `https://www.kaggle.com/uciml/student-alcohol-consumption`

use the model to explain it, the default hyperparameters were used for training and no further steps were taken to improve the performance. The results will be discussed in Section 5.

4.1.2 German Credit problem

The second dataset the explanations were implemented for is the German Credit dataset [18]. Here the classification task is to predict whether a loan applicant is credible enough to receive a credit, based on features like their current account balance, their past credit history and so on. Again, this dataset was chosen because of its simplicity and intuitiveness.

Preprocessing The German Credit dataset consists of 1000 data instances with 20 variables each. To again avoid information overload, only some features were chosen to be taken into account for the classification task. Just like for the previous dataset, the choice of features was based on their correlation to the variable of interest (receiving a loan or not). An overview of the features that were selected for training can be seen in figure 2.

Training Again an XGBoost classifier was used to solve the classification task. The data was split into a train, validation and test set (80%, 10%, 10%). The default parameters of the `xgboost` package were used and no further attempt to parameter optimization was made.

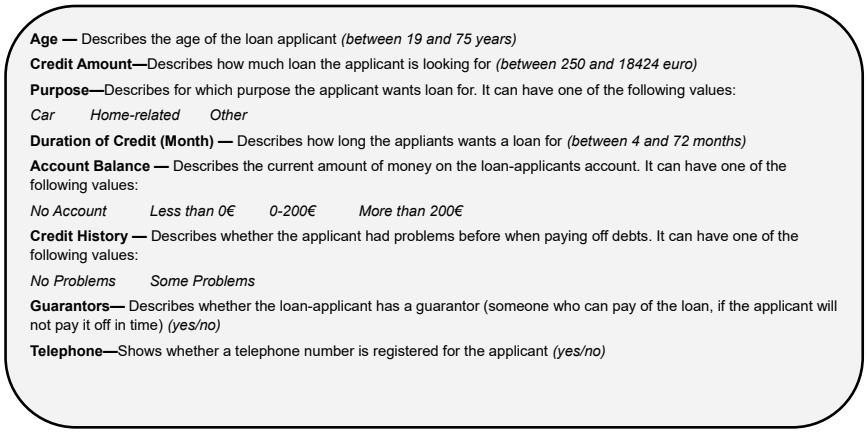


Figure 2: Variables used for the second classification task

4.2 Creation of Explanations

4.2.1 SHAP

To extract the SHAP values from the trained XGBoost models, the python toolpackage `shap` was used. As already mentioned, SHAP values are calculated for every feature of an input instance. They then reflect how important a feature was for the prediction-output, corresponding on the input. In this section, an overview will be given on the basic idea behind SHAP values and how they are calculated for XGBoost models.

Shapley values SHAP values originate from so-called Shapley values that haven been proposed in the context of cooperative game theory [22]. Though there they are used to

calculate the individual payoff for all players in a game with a total payoff, the problem can easily be translated to fit the context of classification problems. Instead of a game, we are then dealing with a classification task. The features used in the classification task reflect the individual players of the game and the output of a classifier reflect the game's total payoff. The Shapely values, in turn, demonstrate the impact of each feature on the obtained output.

To compute the Shapley value of feature i of input x, one has to determine all possible subsets S over all features N excluding i. One can then compute the difference in the classifier's prediction for when i is included in S compared to the classifier's prediction for S itself. If we take the weighted sum of all these differences, we obtain the Shapley value for feature i. Formula 1 quantifies this idea:

$$\Phi_{\mathrm{i}} = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(N - |S| - 1)!}{N!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \tag{1}$$

SHAP Looking at formula 1 it is not clear yet how the prediction of a classifier can be obtained when a feature is missing in a classification problem. Since most classifiers cannot provide an output for an input if feature i is missing, Lundberg and Lee propose the method SHAP to calculate Shapely values by approximating an output $f(z_S)$ with $\mathbb{E}[f(z)|z_S]$ [29]. How $\mathbb{E}[f(z)|z_S]$ is derived, depends on the classifier model that is being used. The most simple way of calculating it is by setting the value for i to the average value of that feature over all input instances.

While this way of dealing with missing features is easy, it is not very precise. Lundberg et al. therefore propose a method specific to tree-models, to calculate the expected classification outcome if a variable is missing [28]. This method is called Tree SHAP.

Tree SHAP For any tree-based model, it is possible to accurately estimate $\mathbb{E}[f(z)|z_S]$, by recursively following the decision path for z if the split-feature is in S and by taking the weighted average over all branches if the split-feature is not in S [28]. Though this way of estimating is accurate, it is very computationally expensive. In a classification problem with M features, where a model contains T trees and the biggest tree has L leaves, running the algorithm for all 2^M subsets will cost $\mathcal{O}(TLM2^M)$ time.

Out of this reason, a new algorithm has been developed to reduce the exponential runtime of the algorithm to polynomial time. The basic intuition here is to compute $\mathbb{E}[f(z)|z_S]$ for all 2^M subsets at once, rather than computing it for each subset individually. This can be done by recursively keeping track of which proportion of subsets flow down into the different leaves of a tree. The resulting algorithm runs in $\mathcal{O}(TLD^2)$ time and was used to compute the SHAP values for the chosen classification tasks. For a more detailed explanation of the algorithm refer to [28].

4.2.2 GIRP

For the GIRP algorithm no existing tool package is available yet, so this algorithm had to be implemented from scratch.

The implementation of the algorithm was completely based on the work of Yang et al. [49] and can be divided into the following steps:

Generating the Contribution Matrix The decision tree generated by GIRP is learned from the local feature importance values of all instances of the training set. Together these feature importance values constitute the so-called "contribution matrix" C of size m x n. Here m denotes the number of instances of the training set, while n represents its number of variables. Thus the element c_j^i from this matrix shows the

feature importance value of feature i for dataset instance j. The contribution matrix was obtained by using the SHAP library mentioned in Section 4.2.1

Contribution	Var 1	Var 2	....	Var i	....	Var N	**Prediction**
Sample 1	c_1^1	c_1^2	...	c_1^i	...	c_1^N	p_1
Sample 2	c_2^1	c_2^2	...	c_2^i	...	c_2^N	p_2
....	...	...	...	...	...	...	...
Sample j	c_j^1	c_j^2	...	c_j^i	...	c_j^N	p_j
....	...	...	...	...	...	...	...
Sample M	c_M^i	c_M^2	...	c_M^i	...	c_M^N	p_M

Table 1: Contribution Matrix for a dataset of N instances and M features. c_j^i shows how much feature i contributed to the prediction p_j for Sample j

Growing the initial tree Once the contribution matrix is obtained, the initial tree can be grown. For this a greedy approach is taken, where both the variable that is going to be split and the value it is going to be split on is maximized. The variables are split on their actual input value, denoted by v_i and not on their contribution c_j^i. Dependent on the variable type, splits can be made with different kind of splitting criteria. The most straightforward variable type for splitting, are binary variables where the split criterion is simply $v_i = 1$. For continuous variables the splitting point can be any constant number d, which would result in the split criterion $v_i < d$. When dealing with a categorical variable the splitting criterion checks whether the v_i belongs to a certain subset D of the set of different levels of the categorical variable.

Splitting any variable results in a left subtree S_L, containing the dataset instances not satisfying the splitting criterion and a right subtree S_R, dataset instances that satisfy the splitting criterion. Given then the contribution values of these dataset instances, one can define the split-strength of variable i as follows:

$$G(split_i) = \left(\frac{\sum_{S_L} c_i^j}{|S_L|} - \frac{\sum_{S_R} c_i^j}{|S_R|} \right) \tag{2}$$

This equation quantifies the difference between the average feature contribution of the left subtree and the average feature contribution of the right subtree. In other words it is measured how differently the variable i contributes to the predictions in S_L and S_R. Thus a large absolute value of $G(split_i)$ demonstrates that the value of variable i is highly indicative for the final prediction of the original classifier.

With this reasoning, the split strength is calculated for every variable and the variable with the largest split strength is chosen as the next splitting variable for the tree. Since categorical and continuous variables can be split in different ways, the splitting criterion which maximizes the split strength for that variable is chosen. Once a node has been split the left and right subtree can further be split recursively, until the tree reaches a maximum depth or if there are not more than a minimum number of samples to split. The tree that is obtained through this process, can also be referred to as T_0. Both the maximum depth of a tree and the minimum number of samples are hyperparameters of the GIRP algorithm that can be optimized later on.

Pruning and selecting the best-sized tree The initial tree that has been grown might be very large and overfitted on the training set. To make sure that the tree will also generalize well to new data, one solution is to prune it. The first step in pruning is to define all subtrees from the initial tree T_0 and then make use of a validation-set

to select the best subtree. The validation set here consists of the dataset instances not present in the training set. By feeding the validation data in each subtree, the validation split-strength for each internal node of the tree can be calculated as follows:

$$G_{validation}(t) = \text{sgn}((G(t)) \left(\frac{\sum_{S_L} c_i^j}{|S_L|} - \frac{\sum_{S_R} c_i^j}{|S_R|} \right) \tag{3}$$

$sgn((G(t))$ here denotes the sign function of the split strength of the node in question, as calculated by formula 2.

The validation split strength of one subtree is then calculated as the sum of the validation strengths of all internal nodes:

$$G_{validation}(T_k) = \sum_{t \in T_k} (G_{validation}(t)) \tag{4}$$

The subtree with the highest validation split-strength can be chosen as the tree that generalizes best on new data.

Choosing Hyperparameters As already mentioned, two hyperparameters can be adjusted to generate GIRP trees. The first one of them is the maximum depth of the tree and the second one is the minimum number of data samples that should be present when splitting a node. Both parameters can be chosen, such that the validation split strength of the resulting tree is maximized. In this work, the possible values for max depth were limited to range between 5 and 8, while the minimum number of samples always felt in the range between 6 and 8. This ensured that the resulting trees would not be too large and complex, but small and easy to understand.

5 Implementation Results

5.1 Classification Tasks

5.1.1 Portuguese Class problem

After training the XGBoost Classifier the F_1-score obtained on the training set was 0.90, on the validation-set 0.86 and on the test-set 0.82. As mentioned before the aim of the study was it not to maximize classification performance, and since the results were already satisfactory no further effort was put into improving the classifier. The confusion matrix for the test-set can can be seen in Table 2

		Predicted	
		Pass	Fail
Actual	Pass	49	2
	Fail	5	9

Table 2: Confusion Matrix Portuguese Class problem

5.1.2 German Credit problem

Training an XGBoost Classifier on the German Credit problem, yielded a F_1-score of 0.85 on the training set, 0.78 on the validation-set and 0.79 on the test-set. Again these

results were deemed as satisfactory and no further work was put into hyperparameter optimization. The confusion matrix for the test-set is shown in Table 3

		Predicted	
		Credit	No Credit
Actual	Credit	64	6
	No Credit	15	15

Table 3: Confusion Matrix German Credit problem

5.2 SHAP

In this section, the results of the SHAP implementation are shown. Though the SHAP values themselves already give an understanding on which features are important for a classification task, they can be made more tangible by so-called force plots. Figures 3, 4, 5 and 6 show such force plots and they can be interpreted by paying attention to the blue and pink arrows. The pink arrows highlight which feature-values contributed to the prediction that an input instance belongs to output-class 1. Thus in the case of the Portuguese Class problem, they denote which characteristics make a student more likely to pass an exam, and for the German Credit problem, they show which makes a loan applicant more likely to receive a loan. The blue arrows, on the other hand, highlight the feature values that made the classifier inclined to predict the opposite label for the input instance. The sizes of the arrows correspond to the importance of these feature values for the final decision. The labels on the axis correspond to the actual SHAP values of the different features. Lastly, the output-value shown in each plot shows the log-odds of the probability that the given input instance is assigned class label 1.

5.2.1 Portuguese Class problem

In Figure 3 the SHAP explanation is displayed for a student that is predicted to pass the course. Figure 4 on the other hand shows the SHAP explanation for a student who was predicted to fail the course. Table 4 show the SHAP values corresponding to Figure 4.

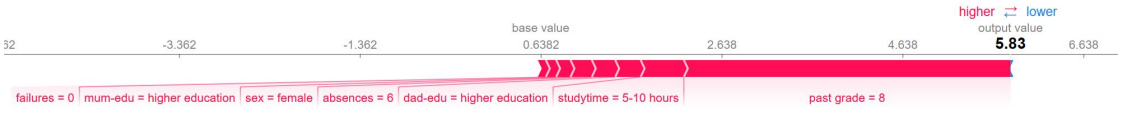


Figure 3: SHAP explanation where prediction = pass

Figure 4: SHAP explanation where prediction = fail

feature	value	SHAP value
sex	female	0.147
age	18	0.083
studytime	2-5 hours	-0.192
failures	3	-1.695
absences	10	0.129
activities	yes	0.099
G1	5	-0.472
Medu	5th to 9th grade	-0.232
Fedu	primary education	-0.565

Table 4: SHAP values corresponding to Figure 4

To again demonstrate how the SHAP values and the force-plot can be interpreted, look at Figure 4 and Table 4. In the Figure we see that the log-odds of a student passing an exam are -2.06, thus the student is more likely to fail rather than pass the exam (probability of passing = 0.113). The negative SHAP values (that are denoted by the blue arrows in the Figure) indicate the characteristics that made the classifier more inclined to predict that the student fails. Thus, in this case, the past number of failures and the education level of the student's father are highly impacting. The pink arrows in the force-plot represent the positive SHAP values, and thus the student's characteristics that made the model more inclined to predict "pass". Here we see that the student's sex had some positive impact, however, this effect is rather small.

5.2.2 German Credit problem

The SHAP explanations were not only generated for the Portuguese Class -, but also for the German Credit classification problem. In Figure 5 the SHAP values for a loan applicant receiving a loan are visualized, while Figure 6 shows the SHAP values for a declined applicant. The figures can be interpreted as described in the previous section.

Figure 5: SHAP explanation where prediction = credit

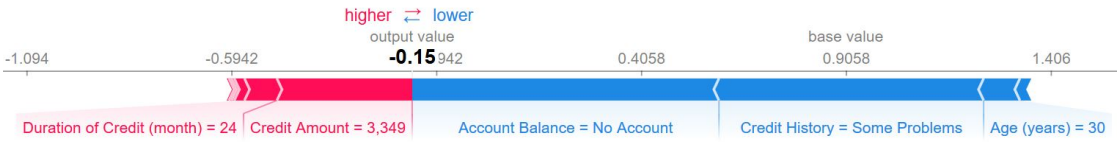


Figure 6: SHAP explanation where prediction = no credit

5.3 GIRP

In this section, we will present the results of the GIRP implementation, by showing the generated tree for the Portuguese Class- and German Credit problems. All trees were

visualized using the python package `pydotplus`; implementation can again be found on Github.

5.3.1 Portuguese Class problem

The best tree for the Portuguese Class problem was obtained for a *maximum depth* of 7 and a *minimum number of samples* of 6. It is visualized in Figure 7.

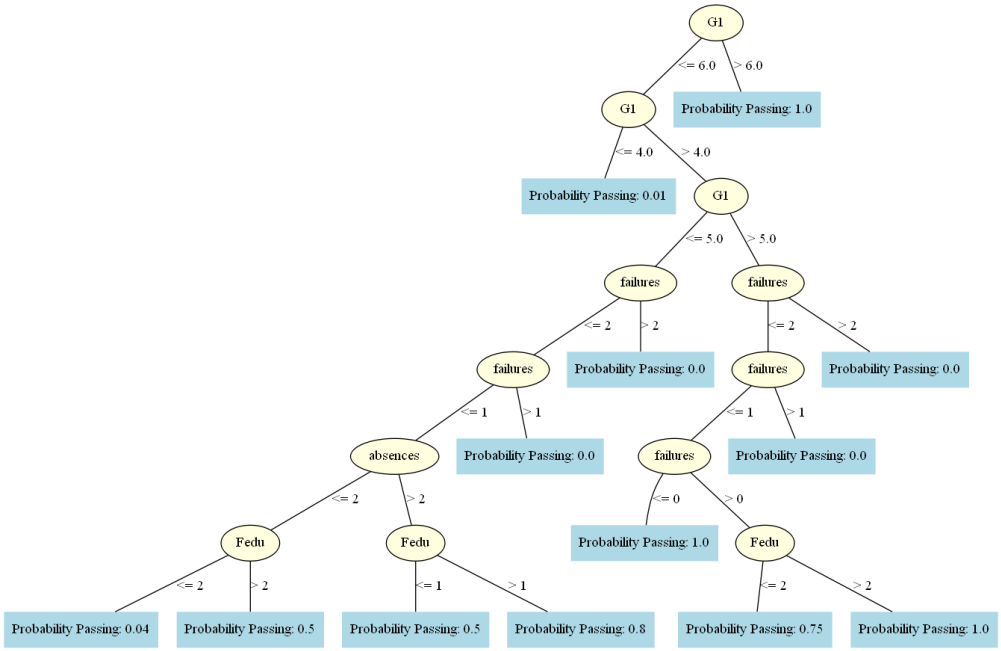


Figure 7: Result of the GIRP implementation for the Portuguese Class problem

5.3.2 German Credit problem

The best tree for the German Credit problem was obtained for a *maximum depth* of 6 and a *minimum number of samples*. Since the resulting tree is only part of it is visualized in Figure 8. The complete tree can be found in the Appendix (A).

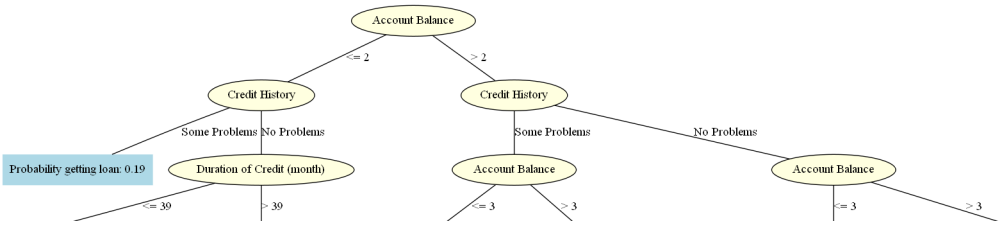


Figure 8: Result of the GIRP implementation for the German Credit problem

6 Experimentation on Explanations Usability

After the explanations of the algorithms were extracted, a usability study was set-up to test their comprehensibility.

6.1 Study Set-up

To answer our first research question, the goal of the usability study was to compare users' comprehension of different AI explanations. The users' comprehension was measured in a three-fold way:

RQ1.1 User's comprehension of the AI model as facilitated by the explanations
RQ1.2 Users' self-rated satisfaction with explanations
RQ1.3 Users' ability to choose between biased and non-biased classifiers using the explanations

To determine the effects of explanation types, the study was set up as a between-subject design. According to the group the participants were allocated to, participants were presented with SHAP, GIRP or both explanations.

In the first part of the study, the aim was to answer RQ 1.1 and RQ 1.2. For this purpose, the participants were presented with explanations generated for the Portuguese Class problem. Before being shown the explanations, participants were first introduced to classification tasks and the different features that were used by the classifier. Additionally, a description was given on how to interpret the provided explanation. The descriptive texts about SHAP and GIRP explanations, as well as the introductory text about the classification problem, can be found in the Appendix (B). Note, that participants in the SHAP condition were presented with force plots explaining the decision processes for two different inputs. In the GIRP condition one decision tree was presented, while in the SHAP+GIRP condition this same decision tree and along with the two force plots were shown.

To answer RQ 1.1 a total of four questions were designed to which only one right answer could be given. The questions were the same for all explanation conditions to make sure that their nature did not have any impact on the results of the study. The questions that were asked are displayed in Table 5

(Q1) Engaging in external activities has a big impact on the prediction of whether a student passes or fails the course a) True b) False	(Q3) Which of the following characteristics seems to be most influential for the system's prediction? a) The time spent studying for an exam b) The student's grade for the past exam c) The number of times a student was absent
(Q2) If a student already failed the course before, the system is more likely to predict that the student will fail the course a) True b) False	(Q4) If the student's past grade was at least a 7, does the number of past failures still impact the system's prediction? a) Yes b) No

Table 5: Comprehension Questions Portuguese Class

In these four questions three aspects of participant's comprehension are measured. Q1 and Q2 both test participant's understanding of the effect of a single variable on the

classifier's prediction. Q3 is a bit more complex, since it asks to compare the effect of multiple variables. Question Q4 is most difficult of all since it requires participants to think about the interaction of multiple variables. Because of the differing nature of the four questions, the aim of the usability study was not only to find out how participant's score overall on the comprehension questions, but also how they score on each of them individually.

The first part of the study was not only about measuring participant's comprehension of the explanations but also about finding out how satisfied they were with them. For this purpose, they were asked to indicate their agreement with different statements on a 7 point Likert scale (with answers ranging between *Strongly Agree* and *Strongly Disagree*). The Likert scale consisted of different qualities related to the usability of explanations, that were already identified as important by previous studies:

L1 The system explanations were easy to understand
L2 The system explanations were unnecessarily complex
L3 It would be worth looking at the explanations to understand how the system is behaving
L4 I am able to understand the explanations in a reasonable amount of time
L5 Overall I am satisfied with the system explanations

The items L1 and L3 were taken from [14], who also conducted a usability study on AI explanations. Items L2 and L5 were inspired by a questionnaire used by [47] and L4 was taken from the System Usability Scale developed by [5]. In addition to the Likert scale questions, we assessed users' self-rated satisfaction with the explanations through an open question. Here they were asked to shortly explain their answers to the Likert scale. The answers given here could be used in the results-analysis to support the findings on the quantitative questions.

Once the responses were taken for the first part of the study, participants could move to the second part where the German Credit classification problem was introduced. In this text, it was stated that participants would be looking at explanations of two AI systems. It was not told, that one of the systems was based on a different biased dataset. In this dataset, loan applicants who did not have a telephone number registered in their bank were more likely to receive a credit loan. The other AI system was based on a regular, non-biased dataset. It was assumed that if the explanations of the two systems were comprehensible for participants, they should be able to detect the bias in one of the systems and reject this AI system in favour for the other one.

Before participants were asked to choose between the classifiers, they were asked to indicate for both AI systems which variable in each system had the biggest effect. These questions were included to encourage participants to think a bit more about the explanations before choosing their preferred AI system. Moreover, the answers to these questions could be used to further investigate RQ 1.1. After participants answered the comprehension questions and decided which AI system was the better one, they were again asked to indicate their agreement on the Likert scale, by answering the questions also used in the first part of the survey.

After the second part of the user study, participants were asked to provide some demographic information. This included their gender, age, nationality, English language proficiency and educational background. This information was gathered to control for these factors, in case they proved to be of impact on any of the other measurements. Finally, the overall response time for each participant to fill out the survey was measured, to also include this as a control variable later on.

For an overview of the different parts of the usability study, refer to Figure 9.

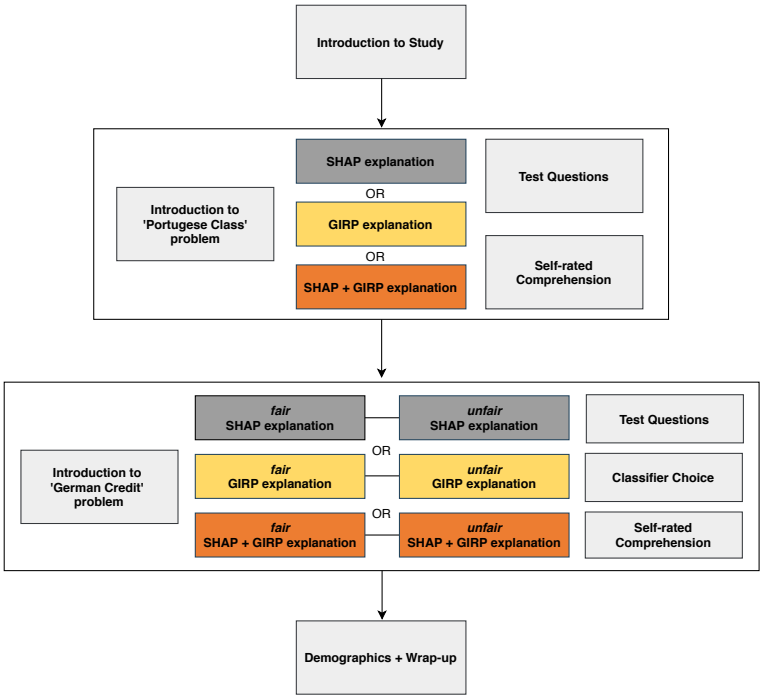


Figure 9: Flow of usability test

6.2 Pilot Study

The study set-up as described in Section 6.1 was tested through a pilot study. For each explanation condition, two participants were asked to fill in the survey. They were allowed to ask questions or give remarks throughout the process. After completion, they were asked some general questions on the clarity of the tasks and the presentation of the explanations. In this section, we will discuss the changes we made to the survey after the pilot was conducted.

Adding context Some of the participants complained that they could not identify enough with the classification problems described in the survey. To illustrate, look at the text below that was initially used to portray the idea behind the Portuguese Class problem.

> *The first AI system we are going to look at is a system that predicts whether a student passes or fails an exam. In order to do so, the AI system has learned from a number of student profiles how different characteristics of a student (like their age, gender or past school performance) relate to them passing or failing the exam in question.*

Some participants suggested adding a bit more context to this description, to highlight why it might be interesting to work on a task like this. For this reason, the introductory

texts for the classification problems were changed to appeal more to the participants' imagination. In the case of the Portuguese Class problem, they were e.g. asked to imagine that they are researchers trying to find out what factors influence a student's performance in school. The full text can be found in the appendix.

Shortening the explanation descriptions Another improvement point that came out of the pilot study, concerned the descriptions on how to read the explanation figures. Many participants found these descriptions unnecessarily long. This especially was said to be the case for the second part of the survey (German Credit task). Since participants already gained sufficient understanding of how to interpret the explanation figures, they suggested shortening the descriptions in this part of the questionnaire.

The proposed changes were validated on another pilot-study. Based on its results no additional changes had to be made and the usability study could be conducted.

6.3 Participants

The participants for the usability study were largely drawn from a university population. Quality control was conducted over their data to see whether they spent sufficient amount of time on the survey and whether their English proficiency was satisfactory. Based on this, it was decided to delete survey responses of a total of 4 participants: each of their response time was shorter than 5 minutes, which was deemed as a too short time to give genuine answers to the survey questions. After their exclusion, the final sample consisted of 113 participants. Out of these, the sample for the SHAP condition contained 38 participants ($N_{female} = 25$, $\mu_{age} = 23.8$), the sample of the GIRP condition 39 participants ($N_{female} = 26$, $\mu_{age} = 23.2$) and the sample of the SHAP+GIRP condition 36 participants ($N_{female} = 24$, $\mu_{age} = 22.1$).

6.4 Results

Before the results of the usability study were analysed, the answer to the demographical question *Please fill in your background (education/most recent field of study/most recent field of work)* was manually translated into a binary variable *technical background*. Thus, for each participant it was noted whether judging by their work/field of study, they had considerable experience with AI or Computer Science. This variable was later included as a control variable in the statistical tests.

In this section, we will discuss the results of these statistical tests.

6.4.1 Research Question 1.1

Portuguese Class problem To answer research question 1.1 for the Portuguese Class problem, it was noted for each of the test-questions whether the participant answered the question correctly or incorrectly. It was then tested whether the total amount of correct answers was significantly different for the three explanation conditions.

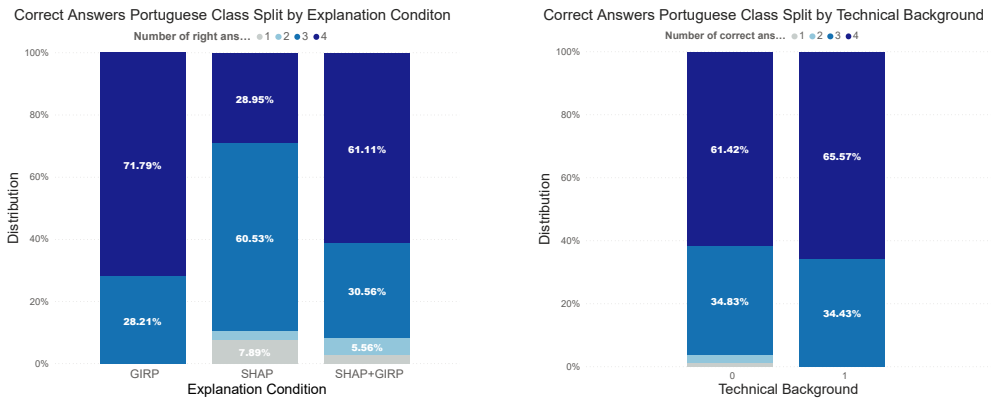
For this purpose, an ordinal logistic regression test was set up with *number of correct answers* as a dependent variable and *explanation condition* as an independent variable. Moreover, *gender*, *age*, *english proficiency*, *technical background* and *response time* were added as independent control variables, to account for their possible effect on the dependent variable. It was decided not to add *nationality* as control variable, since the majority of the participants were Dutch.

Before the statistical test was run, its assumptions were checked. Firstly, it was controlled that there was no multi-collinearity between the independent variable of interest (i.e. *explanation condition*) and the control variables. To test this the Variance Inflation

Factors (VIFs) for each control variable to the variable of interest were tested. Since all VIFs < 3, the assumptions for multi-collinearity were not violated. In addition, the assumption regarding proportional odds needed to be checked. A test of parallel lines yielded a non-significant p-value (p=.091), thus none of the assumptions for the ordinal logistic regression were violated.

With the statistical test it was found that the overall model fit to predict *number of correct answers* from *explanation condition* and the five control variables was significant ($\chi^2(13) = 23.400$, p = .037). For the individual predictors significant effects were found for *explanation condition* and the control variable *technical background*. It was found that participants who were provided with GIRP-explanations had 6.557 times higher odds of scoring more correct answers than participants provided with SHAP explanations (p<.001, 95% CI 2.43 to 19.01). In regards to *technical background* it was found that participants with a technical background had 2.684 times higher odds of scoring more correct answers than participants without a technical background (p = .04, 95% CI 0.8683 to 9.09). In Figure 10 the distribution of correct answers is visualized for splitting either on *explanation condition* or *technical background*. Since both variables had significant effects, it was tested in the logistic regression test whether together they had any significant interaction effects. This was, however, not found to be the case.

While the difference for SHAP and GIRP explanations, was the only significant one, there appeared to be trends in the differences for the other explanation conditions as well. It appeared that GIRP explanations scored higher on objectively measured understanding than SHAP+GIRP explanations, while these, in turn, appeared to score higher than SHAP explanations. The odds-ratios as well as the p-values for the tests are highlighted in Table 6 and will be further discussed in Section 6.5



(a) Explanation Condition

(b) Technical Background

Figure 10: Distribution of the number of correct answers for the Portuguese Class problem when split on *explanation condition* and *technical background*

Since the explanation condition had some effects on the number of correct answers, significance tests for each of the comprehension questions (as displayed in table 5) were run separately. With this, we wanted to find which aspect of comprehension was mostly affected by the explanation type. In Figure 11 the percentage of correct answers for each group and each comprehension question is plotted.

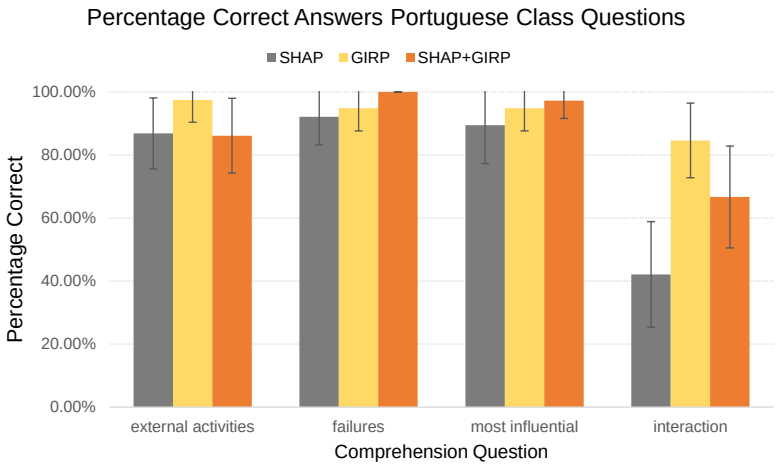


Figure 11: Percentage of correct answers for each individual comprehension question

For each of the four test questions a binomial logistic regression test with the dependent variable *correct answer* and the independent variable *explanation condition* was run. Furthermore, *gender*, *age*, *english proficiency*, *technical background* and *response time* were again added as control variables. Here it was found that only for the question *interaction effect* the predictors could account for a significant amount of variance in the outcome of that variable (p=.036, $\chi^2(13)$=23.5). Individually, only the predictor variables *explanation condition* had any significant effects on whether a correct answer was given or not. The odds for participants of the GIRP condition giving the right answer was 8.229 times higher than for participants of the SHAP condition (p<.001, 95% CI 2.49 to 27.19). The difference between GIRP and SHAP+GIRP was not found to be significant for the scores on this question, however the same trend as in the previous statistical test were found: firstly there is an indication that GIRP scores better than SHAP+GIRP and secondly, there is a trend for SHAP+GIRP to score better than SHAP. The p-values and odds-ratios are presented in Table 6.

				95% CI	
dep. var	**predictor**	**p-value**	**odds**	**lower**	**upper**
#correct ans	**Explanation Condition**				
	GIRP vs. SHAP	<.001	6.557	2.429	19.010
	GIRP vs. SHAP + GIRP	.066	2.565	0.889	7.780
	SHAP+GIRP vs. SHAP	.057	2.557	0.938	7.180
	Technical Background				
	Yes vs. No	.044	2.684	0.868	9.090
interaction	**Explanation Condition**				
	GIRP vs. SHAP	<.001	8.229	2.4901	27.190
	GIRP vs. SHAP + GIRP	.067	3.223	0.919	11.300
	SHAP+GIRP vs. SHAP	.076	2.553	0.908	7.180
	Technical Background				
	Yes vs. No	.856	0.892	0.261	3.060

Table 6

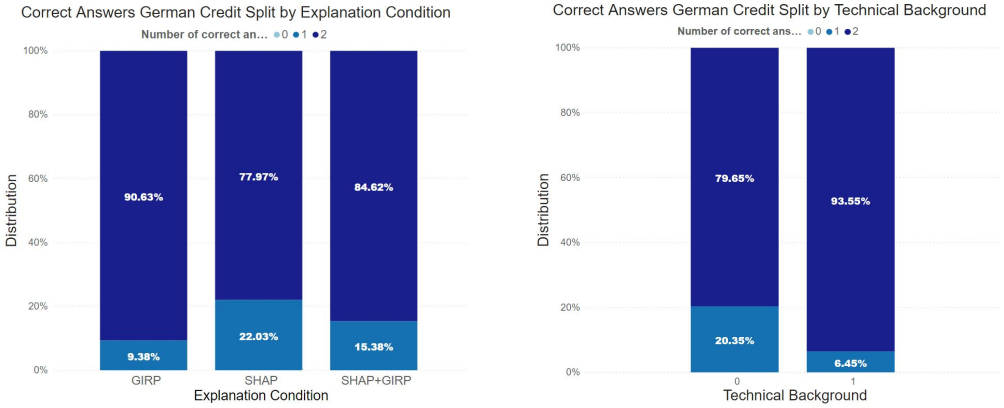
German Credit problem In the usability study, participants were not only asked comprehension-questions about the explanations for the Portuguese Class problem, but

also for the German Credit problem. Thus research question 1.1 needed to be answered for this part of the questionnaire as well.

For the German Credit task, only two comprehension questions were asked. Again, the total number of correct answers was calculated for each participant and then used as a dependent variable in an ordinal logistic regression test. Just like before, *explanation condition* was set as an independent variable, along with the control variables *gender*, *age*, *english proficiency*, *technical background* and *response time*.

After checking the assumptions, it was found that the fit for this ordinal logistic regression model was significant (p=.031, $\chi^2(13)$=24.0). From the individual predictors, only the effect of *technical background* was found to be significant. Participants with a technical background had 5.861 times higher odds to give more correct answers than participants without a technical background (p=.002, 95% CI 1.477 to 33.23). No significant effects for the variable of interest (*explanation condition*) were found.

The distribution of the total amount of correct answers when splitting on *explanation condition* and *technical background* are plotted in Figure 12.



(a) Explanation Condition

(b) Technical Background

Figure 12: Distribution of the number of correct answers for the German Credit problem when split on *explanation condition* and *technical background*

Since *explanation condition* was not found to have any significant effect on the total amount of correct answers, no further tests were conducted for the individual comprehension-questions of the German Credit problem.

6.4.2 Research Question 1.2

With research question 1.2 we tried to find whether the self-rated satisfaction of participants was different for the three explanation conditions. Thus, we needed to test whether the answers on the Likert scale questions in the usability study (see Section 6.1), were significantly different for the three explanation conditions. Again this was done separately for the two parts of the survey (Portuguese Class and German Credit). For each item of the Likert scale an ordinal logistic regression test was set up where *easy_to_understand*, *complex*, *reasonable_time*, *worthy*, and *overall* were set as the corresponding dependent variable. Again, *explanation condition* was set as the independent variable, along with the control variables *gender*, *age*, *english proficiency technical background* and *response time*. For the ordinal logistic regression tests, the assumption of

proportional odds needed to be checked. A test of parallel lines yielded for both the first and second part of the survey non-significant p-values. The assumption regarding multi-linearity did not have to be checked anymore since this was already tested for RQ 1.1, where the same independent variables as here were used.

Portuguese Class problem From the five ordinal logistic regression tests that were run for the first part of the survey, significant effects were only found for the item *worthy*. Here all predictors could account for a significant amount of variance in the answers given to this Likert scale question ($\chi^2(13) = 23.0$, p=.041).

Individually, only the predictor *explanation condition* had effects on the outcome of this variable. Participants in the SHAP condition had 3.890 times higher odds to give a higher rating to this question than participants of the GIRP condition (p=.004, 95% CI 1.53 to 10.31). In addition, SHAP participants had 5.09 times higher odds of giving a higher rating to the Likert scale item than participants who were shown SHAP+GIRP explanations (p<.001, 95% CI 1.92 to 14.11). No significant differences for the answers to this Likert scale item were found for the explanation conditions GIRP and SHAP+GIRP (p=.547). The distribution of answers given to the Likert scale item is visualized in Figure 13

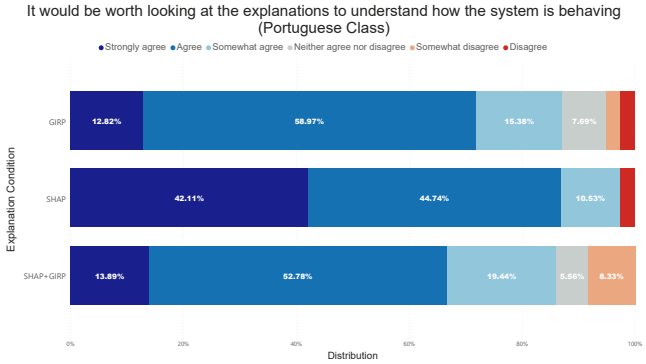


Figure 13: Answers given to the Likert scale question: "It would be worth looking at the explanations to understand how the system is behaving" in the first part of the survey

The answers that were given to the other items of the Likert scale are visualized in Figure 14. Note here, that the colour coding for the question "The system explanations are unnecessarily complex" is different than for the rest, such that blue colour tones denote disagreement with the statement (and hence a positive evaluation of the explanations) and red colour tones denote agreement with the statement (and hence a negative evaluation of the explanations).

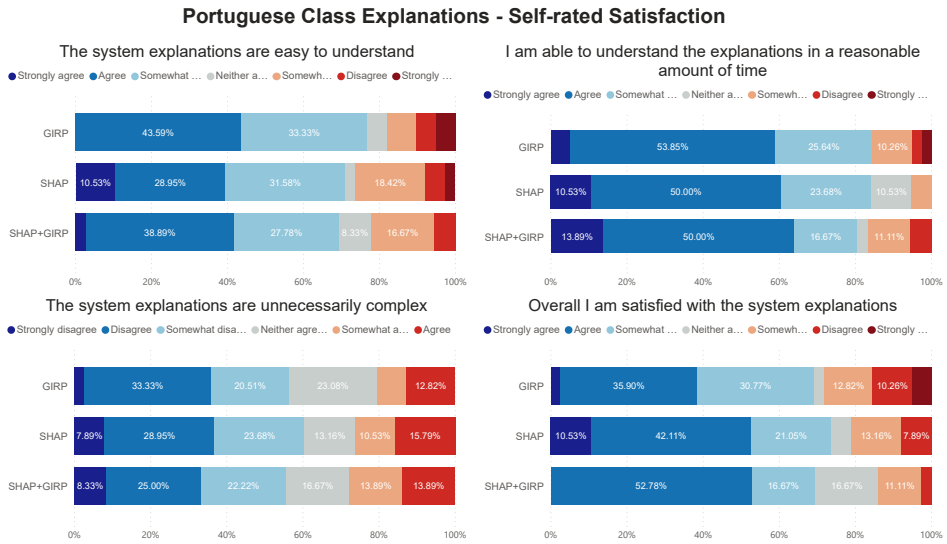


Figure 14: Answers given to the other items of the Likert scale for the first part of the survey

German Credit problem From the five statistical tests that were run significant effects were only observed for the item *worthy*. The answers given to this item of the Likert scale are visualized in Figure 15. The overall fit for the logistic model was found to be significant with $\chi^2(13)=24.8$ and p=.025. In the model the individual predictors *technical background* and *explanation condition* had significant effects on the outcome variable. For *technical background* it was found that participants with a technical background had 4.84 times higher odds to give a higher rating for the Likert scale item (p<.001, 95% CI 1.729 to 14.43). For *explanation condition*, in turn, it was found that participants with SHAP explanations had 2.727 times higher odds to give higher ratings than participants with GIRP explanations (p=.024, 95% CI 1.132 to 6.710). Looking at Figure 15 there appears to be a trend for participants in the SHAP condition, to give higher ratings than participants of the SHAP+GIRP condition. The p-value here was however not found to be significant (p=.081). Lastly, no significant interaction effects were found between *technical background* and *explanation condition*. Just as for the Portuguese Class problem, the differences in responses to the other Likert scale questions were not significant. The responses are visualized in Figure 16.

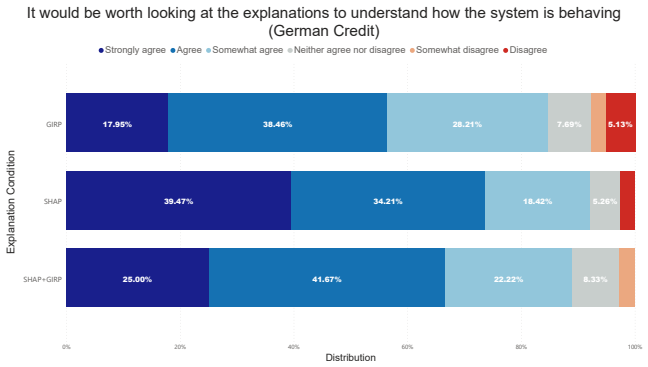


Figure 15: Answers given to the Likert scale question: "It would be worth looking at the explanations to understand how the system is behaving" in the second part of the survey

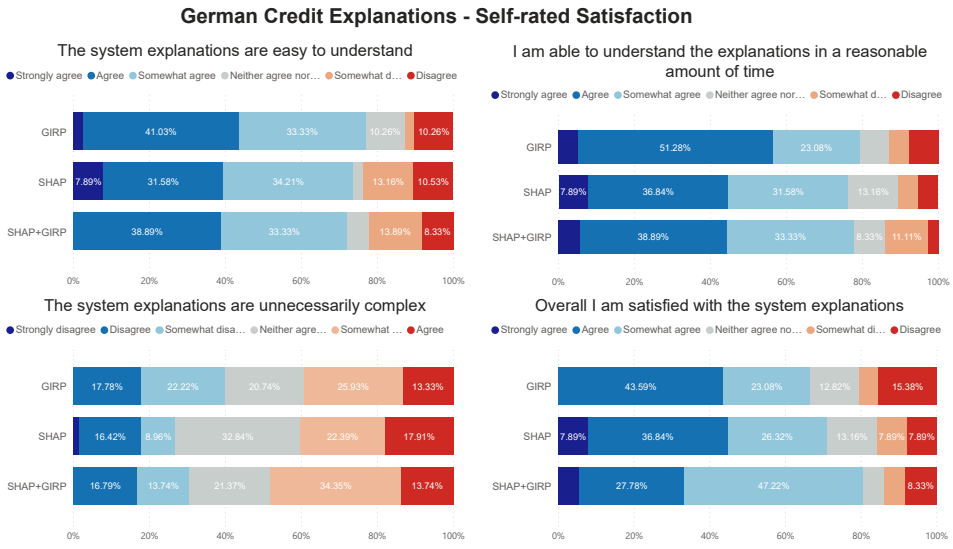


Figure 16: Answers given to the other items of the Likert scale for the second part of the survey

6.4.3 Research Question 1.3

The last research question was about assessing whether explanations can help users in choosing between AI models. This question only concerned the second part of the usability study, where participants where shown explanations for two classifiers for the German Credit dataset; a biased one and another non-biased one.

In a binomial logistic regression test, it was investigated whether the explanations that were provided affected the users' ability to correctly choose the non-biased classifier. In this test *correct answer* was set as the dependent variable, while *explanation condition* was set as the independent one. Furthermore, the same five control variables as before

were added to the test. Before the test was conducted, it was made sure that no assumptions were violated.

The predictors together accounted for a significant amount of variance in the outcome variable ($\chi^2(13)=22.9$, p=.043). Of all individual predictors, *explanation condition* was the only one with significant effects. Firstly, it was found that participants in the GIRP condition had 4.330 times higher odds to give the right answer than participants in the SHAP condition (p=.035, 95% CI 1.107 to 17.420). Secondly, the participants in the SHAP+GIRP condition had 5.027 times higher odds of providing the right answer than the participants in the SHAP condition (p=.022, 95% CI 1.256 to 20.110).

No significant differences were found between participants given GIRP and participants given SHAP+GIRP explanations (p=.860). The percentage of correct answers given by each explanation group is visualized in Figure 17.

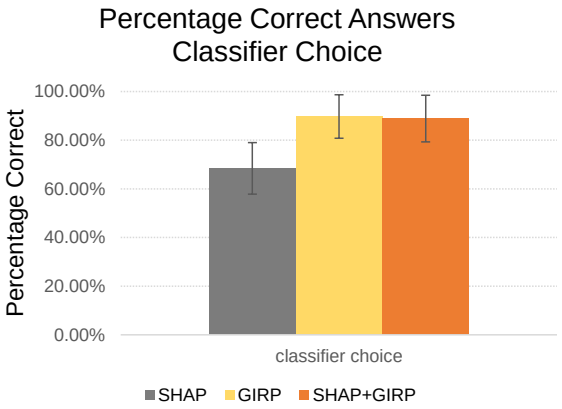


Figure 17: Percentage of correct answers in the choice of classifiers

6.5 Discussion

As AI algorithms have become increasingly complex and opaque, the main goal of this research was to find out which way of explaining those algorithms is most useful to non-computer experts. There are two different ways in which AI-algorithms can be explained: local explanations, that describe how an algorithm made a decision for one particular instance and global explanations, that give a holistic overview of how an algorithm derives decisions. From a usability perspective, it has neither been established how the two explanation techniques compare to each other, nor how a combination of them could facilitate users' understanding of AI algorithms. It was therefore the goal of this study to close this knowledge-gap and contribute to the growing body of research in this field.

The "understanding" of AI explanations is a rather broad term, thus it was measured through three separate sub-questions. The first relates to users' objectively measured understanding of explanations, the second to users' self-rated satisfaction of the explanations and the third to users' ability to use the explanations to find bias in AI models. In Section 6.5.1 we will discuss the findings for each sub-question individually. We will then attempt to combine the answers of each individual question, to give answer to the main research question in Section 6.5.2. What follows is a discussion of the limitations of the study as well as suggestions so further research in Section 6.5.3.

6.5.1 Sub-questions

RQ 1.1: *How is users' ability to answer comprehension questions about models affected by the presentation of SHAP explanations, compared to GIRP explanations or a combination of both?*

The results of the usability study indicate that being presented with GIRP rather than SHAP explanations slightly increase users ability to answer comprehension questions about AI models. However, this appears to be only the case for comprehension questions, relating to the interaction of variables: in case of the Portuguese Class problem, this question is the only one, where significant differences between the explanation conditions were found. This, in turn, explains why no differences between explanation conditions were found for the questions asked about the German Credit problem. These questions only asked participants to identify the most important features for the decision process and did not test their comprehension of interaction effects. Intuitively, it makes sense that GIRP explanations facilitate users' understanding of interaction effects. Though it is possible to reason about the interaction of features using SHAP, this is less straightforward than looking at a decision tree and seeing how the value of one feature changes the decision path for the value of another feature. Moreover, participants of the SHAP condition only got to see two local explanations. It remains unknown how they would have performed on the interaction comprehension question if they would have been presented with more SHAP explanations.

Differently than expected, providing SHAP and GIRP explanations did not have any significant positive effects on users comprehension of AI models. Looking at Figure 11 and at Table 6 it even appears that there is a trend for people provided with GIRP explanations to give more correct answers than people provided with a combination of the explanation techniques. A reason for this finding might be that the combination of the explanation types causes an information overload. In the psychological literature, it has been established that being presented with too much information can decrease people's capacity to process that information, which in turn, can affect their ability to make accurate decisions about it [19]. This phenomenon might very well have occurred in this usability study, which can explain the results. Relating to this psychological phenomenon it would be interesting to see how users' performance would be affected by presenting the information more concisely. Presenting the SHAP and GIRP explanations on separate pages, rather than all on one page, might already be helpful. Another measure could be to "force" the participants to take some time between looking at the explanations and answering the comprehension questions. A study by Fukukura et al. has namely suggested that this way of psychologically distancing people from their decisions, can diminish the effects of information overload [13].

If information overload indeed affected participants understanding of the explanations, it is important to look at the ecological validity of this research: the reason why participants might have felt overwhelmed by the explanations is that their intrinsic motivation to understand them was not very high, to begin with. After all, the participants were mostly taken out of a university population and their interest in xAI might have been quite limited. As studies suggest that the higher a person's personal motivation to process information, the less prone they are to the effects of information overload [21, 37], it would be interesting to see how people who are affected by xAI in their personal life would have performed on the usability test. These could e.g. include employees at a company who are responsible for deciding whether an AI model is fair or not. The fact that there appears to be a trend for people in the SHAP+GIRP condition to give more right answers then people in the SHAP condition (see Table 6 and Figure 11), supports the idea that the combination of the local and global explanations can be

more beneficial than suggested by the initial results.

One final finding in answering research question 1.1 is that the technical background of participants significantly affected their number of correct responses for the comprehension questions. This finding is not very surprising, but further demonstrates the importance of including non-computer experts in the design of explanations. The fact that no significant interaction effects are found for the variables *explanation condition* and *technical background* shows that the previously described effects of the different explanation types occur regardless of the computer expertise of users.

RQ 1.2: *How does users' self-rated satisfaction of SHAP explanations, compares to their self-rated satisfaction of GIRP explanations or a combination of both explanations?*

For both the Portuguese Class- and German Credit classification problems, significant differences in self-rated satisfaction of explanations were only found for the Likert scale item *"It would be worth looking at the explanations to understand how the system is behaving"*. Firstly, it appears here that in both parts of the survey, providing SHAP explanations rather than GIRP explanations increases users perception of the worth of the explanations. A reason for why this is found may be that SHAP values, falling in the category of local explanations, provide a good indication of how an AI model work for a person individually. Studies have shown that this example-based method of learning is preferred more than methods like GIRP, where a more abstract and holistic overview of a process is provided [4]. The fact that no further significant differences between SHAP and GIRP were found for users' self-rated satisfaction, however, indicates that this effect is not very big. Still, looking at the answers to the previous research question, it might seem somewhat contradicting that SHAP explanations score higher on self-rated satisfaction even though GIRP explanations are better to increase users' objectively measured understanding. This finding will further be discussed in Section 6.5.2 were the overall research question will be examined.

Another conclusion we can draw for RQ 1.2 is that in the first part of the survey, participants in the SHAP condition gave a significantly higher rating for the Likert scale item *worthy* than participants in the SHAP+GIRP condition. The observed effect appeared to be less strong in the second part of the survey, where no significant differences were found. The reason for why SHAP explanations scored higher than SHAP+GIRP explanations in the first part of the survey might be again that participants experienced information overload. This could also be an explanation, for why no significant differences for the worth of SHAP and SHAP+GIRP were found in the second part of the usability study. As already explained, the motivation of a person to solve a task can affect how much a person is negatively affected by information overload [21]. In the second part of the study, there was a clear purpose to understand the explanations, namely to use them to choose the better, non-biased AI model. Being confronted with an explanation where a clear bias of a model is presented, might highlight the relevance of xAI. With this increased task relevance, the motivation to understand the explanations might have been higher as well. This, in turn, could have reduced the effects of information overload and let participants rate the worth of SHAP+GIRP explanations higher than in the first part of the study.

Looking at the scores for the other items of the Likert scale (for both the first and the second part of the survey) it appears somewhat surprising that no more significant differences were found. To recall, the other items of the Likert scale were the following:

L1 The system explanations were easy to understand
L2 The system explanations were unnecessarily complex
L3 I am able to understand the explanations in a reasonable amount of time
L4 Overall I am satisfied with the system explanations

Intuitively, it could be expected that especially *L1*, *L2* and *L3* would have lower scores for the SHAP+GIRP condition, if participants experienced an information overload for this condition. However, judging from Figure 14 and Figure 16 there does not even appear to be a trend for people in this explanation condition to give worse ratings for these items. A reason for this could be connected to their exact wording. Though terms like *"easy to understand"* or *"unnecessarily complex"* have an indirect connection to the perception of information overload, a more straightforward Likert scale item like *"The system explanations made me feel overwhelmed"* might have yielded different results. A look at the qualitative data gathered in the first part of the usability study, reveals that this indeed might be the case. The statements below were given by participants in the SHAP+GIRP condition, to illustrate their impression of the given explanations.

> *"It was quite easy explained but sometimes I got a little confused. I had to look at the pictures and the text often again, cause I got lost in it"*

> *"There was a lot of repetition and lots of graphs. Maybe a paragraph that summarized everything about the system, would prevent me from getting lost and losing focus."*

After a glance at this qualitative data a more detailed posthoc analysis was conducted, to inspect whether the notion of "getting lost" in information or feeling overwhelmed by it, was more prevalent in the open-ended responses for the SHAP+GIRP than for the other conditions. This was indeed found to be the case: out of the 38 participants who were presented with SHAP explanations, only 5 gave indications of feeling overloaded by information ($\sim$13.16% of the participants). This was the case for 7 out of 39 GIRP-participants ($\sim$17.95%) and for a total of 12 out of the 36 SHAP+GIRP participants ($\sim$33.33%).

If these findings are correctly interpreted, the previously proposed ways to reduce information overload might also be effective for increasing users' self-rated satisfaction with SHAP+GIRP explanations. Making the combination of the explanations more concise could be helpful, as well as increasing users' intrinsic motivation to understand the explanations. The fact that no significant differences are found between the GIRP and SHAP+GIRP explanation conditions is encouraging for the idea that the effects of information overload are not so big that they cannot be easily reduced.

One final aspect that should be noted for RQ 1.2, is that again significant differences in self-rated satisfaction were found between participants with and without a technical background. Just like for RQ 1.1, this emphasizes the importance of always including non-computer experts in usability studies.

RQ 1.3: *How is users' ability to detect bias in models affected by the presentation of SHAP explanations, compared to GIRP explanations or a combination of both?*

From the second part of the usability study, we can conclude that providing GIRP explanations facilitates users' ability to detect bias in models, compared to providing SHAP explanations. Looking at the answer to RQ 1.1 this finding is not very surprising: the ability to find bias in models relies on users' objective understanding of AI explanations, which already was shown to be facilitated by GIRP- rather than SHAP explanations.

Different to the findings of RQ 1.1, it was found that the provision of SHAP+GIRP explanations was more beneficial for users' ability to find bias in AI models than the presentation of SHAP explanations. In RQ 1.1 it has been hypothesized that participants of the SHAP+GIRP condition did not perform significantly better than participants of

the SHAP condition because they experienced information overload. This effect might have been diminished for RQ 1.3 because of the motivational aspect, described in the previous paragraph. The task of detecting bias in AI models has high societal relevance, which might have increased participants' motivation to perform well on the task. This, in turn, could have reduced the effects of information overload and therefore increased the positive effects of providing a combination of SHAP and GIRP explanations.

6.5.2 Overall Research Question

RQ 1: *How is users' understanding of a model affected by the presentation of SHAP explanations, compared to GIRP explanations or a combination of both?*

Differently than expected, there does not appear to be a single, straightforward answer to the overall research question. Firstly, it is found that in terms of objective understanding GIRP explanations are superior to SHAP explanations while in terms of self-rated satisfaction participants prefer SHAP explanations. This suggests that there is a gap between subjective and objective understanding of AI-explanations: the fact that persons "like" explanations, does not guarantee that they can adequately use them, or that they are enough to gain a full understanding of an AI model. This phenomenon has already been observed in different contexts and is called the Dunning-Kruger effect [10]. It refers to a cognitive bias, in which people overestimate their knowledge or expertise in an area. The effect appeared before in the context of information processing [30], and highlights the importance of never fully relying on people's subjectively rated understanding of a topic. In the case of our study, the effect is most observable in participants of the SHAP condition, who rated the provided explanations as more valuable than participants of the GIRP condition, even though they were less capable in using the explanations to accurately derive conclusions about AI models.

Though it can be argued that objectively measured comprehension is more important than subjectively measured one, it is not advisable to simply ignore any subjective measurements. After all, there are studies which indicate that subjective comprehension is important for users to be willing to use their knowledge [23]. Concerning xAI, this means that even though SHAP explanations might not facilitate users ability to derive conclusions about AI models, they still might motivate users to check the explanations in case they are needed.

Understanding the difference between subjectively- and objectively measured comprehension and how both are affected by the presentation SHAP and GIRP explanations, it becomes interesting to look at the second finding for the overall research question: presenting a combination of the explanation techniques is at no times more beneficial than only presenting SHAP or only presenting GIRP explanations. In the case of subjectively measured understanding, SHAP is preferred over SHAP+GIRP, while there also appears to be a trend for GIRP to be favoured over SHAP+GIRP in terms of objectively measured understanding. Looking at the previous results, this may appear somewhat surprising. One might have expected that combining the local- and global explanation technique would bring out "the best of both worlds", such that users could increase their subjective understanding by looking at the SHAP plots and refer to the GIRP decision trees to improve their objective understanding. Moreover, it might be expected that both explanation techniques would complement each other and boost users' overall understanding of the AI model. The fact that this was not observed, is ascribed to users experiencing an information overload when looking at both explanations. The evidence for this theory has partly been found in the qualitative data of the study. Besides, the responses to the second part of the survey were further confirmation for our theory. Here no significant differences between the combination of the explanation

techniques were found to the separate explanation conditions. In fact, the presentation of both explanation techniques was shown to be more useful for detecting bias in AI models in comparison to the presentation of SHAP explanations. We suggest that this could be the case, because the societal relevance of the second task was higher, which increased users' motivation and, in turn, diminished the effects of information overload.

One final finding in regards to the overall research question is that the technical background of a person influences both objective and subjective understanding of explanations. Since this is not very surprising and it was not the goal of our study to investigate this relationship, not much further attention was paid to this phenomenon. Still, it illustrates that we should not underestimate the importance of taking non-computer experts into account when designing and testing explanations. An overall model of how SHAP and GIRP, and a combination of both explanation techniques is given in Figure 18.

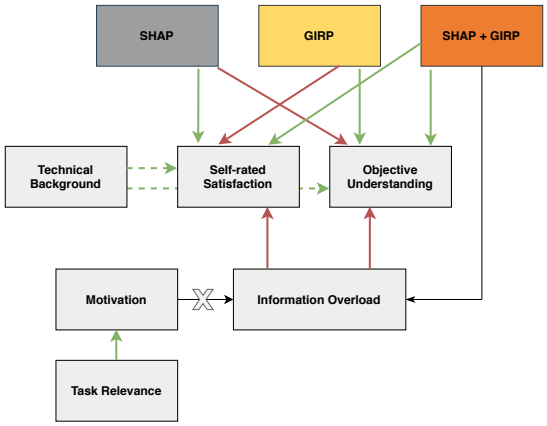


Figure 18: Proposed model of how the different explanation conditions affect users' subjective and objective understanding of AI models

In this model, green arrows indicate a positive effect from the outgoing node to the incoming one while red ones indicate a negative effect. Thus SHAP explanations are indicated to have a more positive effect on subjective ratings of the explanations, while the opposite is true for the objectively measured understanding of them. According to the plot, combining SHAP and GIRP has a positive effect on both aspects of understanding, which is, however, inhibited by the experience of information overload. This experience can be reduced by an increased motivation on the user's side.

6.5.3 Limitations & Further Research

In Section 6.5.1 we already discussed some limitations regarding the conclusions of each sub-question. To summarize, these related to the number of SHAP explanations that were presented, the phrasing of the Likert scale items and the motivation of the participants to understand the explanations.

However, apart from these listed limitations, there also several other shortcomings to the overall usability study that still need to be discussed. The probably most important one is connected to the general usability of the presented explanations. Both the SHAP and GIRP explanations were put in the survey, without assessing their usability beforehand. Thus any differences in the results, might not have been due to the nature of the explanations but instead due to how they were presented. Looking at some of the answers given to the open survey questions, there indeed appears to be some points

in which the general usability of the explanations could be improved. For the SHAP explanations, it was e.g. mentioned that the colour coding in the force plots was not clear or that the numbers in the plot were confusing.

> *"It helps that the explanations have different colours but on the other hand I was confused that pink meant that a student passed. Pink is close to red which I associate with negative outcomes."*

> *"The numbers on the axes are quite arbitrary and seem not logic (why is the axis negative?)"*

In addition, there were some remarks on the general usability of GIRP trees. Some participants found the notation of probabilities in the leaves of the tree confusing. Others were puzzled by the use of 'greater-than'- or 'less-than' signs and would have preferred the use of written words:

> *"Not everyone is familiar with the probability indicators (p= 1.0, 0.86 etc) (did this long time ago in high school)."*

> *"The decision tree was a little complicated/confusing, because of the '<' and '>' symbols. It would be easier to understand if there was written 'less than' or 'more than'."*

Looking at these comments, it seems advisable for future work to pay more effort into making the explanations as usable as possible before comparing them to each other in a usability study. This also holds for how GIRP- and SHAP explanations are combined. As mentioned in the previous section, the fact that no significant improvement in users' comprehension for this explanation condition was found, might have been due to how they were presented together. It might be worthwhile to conduct a usability study of more qualitative nature to search for more adequate ways in which the different explanations can complement each other. Presenting the explanations on separate pages, changing the order in which they are presented, or adjusting the introductory texts about how to interpret the explanation figures, could be ways to increase the potential of the combination of both explanations.

Another shortcoming of the usability study is that the presented explanations were made for classification problems that were quite intuitive. The classification problems for the Portuguese Class- and the German Credit dataset are not based on a very high number of features and the features are not very complex. It would be interesting to see how the found results would be affected if explanations were made for more complex AI models. Another related concern is connected to the type of data the explanations were generated for. SHAP values and GIRP trees are data agnostic explanation techniques, meaning that they can explain classification decisions for tabular-, textual-, and visual data. In this study, it was only tested how the explanations are understood for textual data. Different results might be found if their usability for textual- or visual data is assessed.

7 Experimentation on Explanations' Fidelity

The results of the usability study alone, do not provide a complete picture of how suitable SHAP, GIRP or the combination of both are as explanation techniques. As discussed in Section 2.3.2 an important criterion for the quality of explanations is their fidelity, i.e. the extent to which the explanations reflect the inner-workings of the model they originate from. In this section, we will demonstrate how the fidelity of SHAP and GIRP was assessed and how satisfactory the results are. With this section research questions 2 and 3 will be answered.

7.1 SHAP - Impact Score

7.1.1 Implementation

The fidelity of SHAP can be estimated using the Impact Score as proposed by Lin et al. [27]. The intuition behind this quantity is that if the features that are indicated as important by SHAP are absent, the output of the model for the new input should be different than before. This change in output is either reflected by a different prediction altogether or a considerably lower confidence in the prediction.

The authors formally define this idea as followed. Say we have a model N that outputs for input x the prediction y and the confidence in prediction z:

$$\{y, z\} = N(x) \tag{5}$$

We then can have a local explanation function M, that defines for an input x in the model N a number of critical features c:

$$c = M(x, N) \tag{6}$$

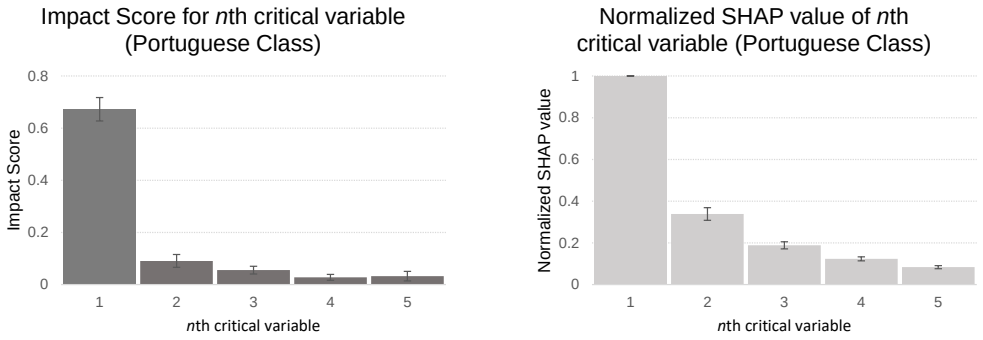
where $c \in x$. If we then define the input x in absence of c as $x' = x - c$ and the output of N for x' as $\{y', z'\} = N(x')$, we can calculate the Impact Score I across a set of n inputs $X = \{x_1, x_2, ..., x_n\}$ as:

$$I = \frac{1}{n} \sum_{i=1}^{n} ((y_i' \neq y_i) \vee (z_i' \leq \tau z_i)) \tag{7}$$

In this formula, τ indicates the amount of confidence lost in the prediction when the critical feature is absent in the decision process. In our case, τ was set to 0.5. Furthermore, we defined critical features by their ranked feature importance values. Intuitively, one would expect that (if the feature importance values are faithful to their original model) leaving out a feature with a high importance value should yield a higher Impact Score than leaving out a feature with a lower importance value. By testing whether I for n^{th} critical feature is in all times higher than I for the $n + 1^{th}$ critical feature, one can check whether this intuition is met. Moreover, we can compute the normalized SHAP value of each critical feature and see how this value compares to the magnitude of the Impact Score. More specifically, we would expect here that the higher a normalized SHAP value is, the more important it is for a decision process and the higher the Impact Score should be. In other words, the Impact Score for a critical feature should be somewhat proportional to the normalized SHAP value of that feature.

7.1.2 Results

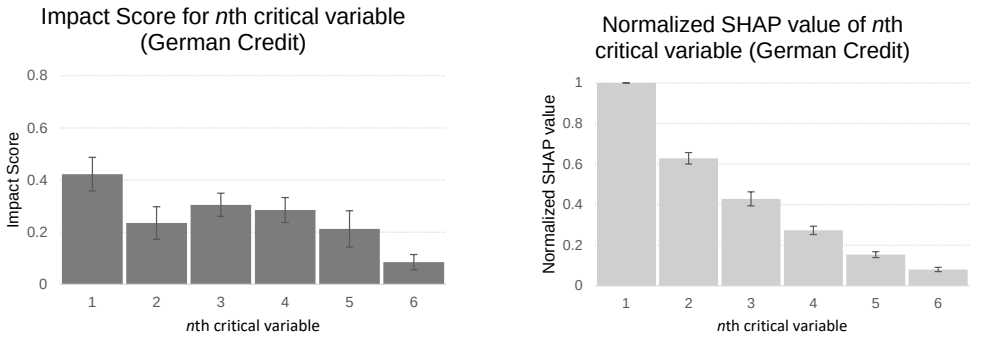
The figure below shows the Impact Score for the top 5 critical features of the Portuguese Class problem. Next to it, the average normalized SHAP values for each critical feature are visualized. Both measurements were obtained by calculating them for the test sets of 10 different train-test-splits and then averaging them. In figure 20 the same plots are shown for the top 6 critical features of the German Credit classification task.



(a) Impact Score

(b) Normalized SHAP value

Figure 19: (a) SHAP's fidelity as measured through the Impact score for Portuguese Class problem (b) The normalized SHAP values corresponding to the critical features of the Portuguese Class classification task



(a) Impact Score

(b) Normalized SHAP value

Figure 20: (a) SHAP's fidelity as measured through the Impact score for German Credit problem (b) The normalized SHAP values corresponding to the critical features of the German Credit classification task

7.1.3 Discussion

Looking at the Impact Scores of the Portuguese Class SHAP values we see that the Impact Score of the n^{th} critical feature is, as expected, at most times larger than the $n+1^{th}$ feature. The only exception is the Impact Score for the fifth critical feature, which is slightly larger than the one of the fourth critical feature. However, this difference is quite small and does therefore not fundamentally decrease our faith in the SHAP values.

Interestingly, we see that the Impact Scores are quite low for all critical features except for the first. As this pattern is reflected in Figure 19 b), where the difference between the normalized SHAP value of the first and second/third most critical feature is also quite high, this finding is not very surprising. Thus looking at the plots, there does not appear to be a reason to doubt the fidelity of the SHAP values generated for the Portuguese Class classifier.

The observations for the Impact scores of the German Credit SHAP values are slightly different. Though the Impact Score of the most critical feature is higher than the rest, the ranking of the second, third and fourth most critical features is slightly off. We see that leaving out the third or fourth most critical variable has a higher impact on the system's predictions than leaving out the second most critical variable. Though this decreases our faith in the SHAP values, the finding can to some degree be explained by looking at Figure 20 b). The differences between the values of the differently ranked features are here less pronounced than for the Portuguese Class problem. Furthermore, the normalized SHAP values are here a lot higher for the second, third and fourth features. Thus from this perspective, it is not completely surprising that leaving out e.g. the third most critical variable can have a similarly big effect as leaving out the second most critical variable in a decision process.

With these two observations, it becomes difficult to tell, to which extent we should reduce our faith in the SHAP values. This, in general, appears to be a problem, when using the Impact Score as a fidelity measure. Though intuitively the measure makes sense, the authors do not provide a baseline of what a satisfactory Impact Score is. Since it was beyond the scope of this study to implement different ways of calculating feature importance values, we also cannot compare the fidelity of SHAP values to other feature importance values, like given by LIME or DeepLift.

To summarize, there is no clear reason to doubt the fidelity of Portuguese Class SHAP values, but there is an indication that the SHAP values for the German Credit problem are not faithful. Even though the findings of this case study only give a very limited answer to our second research question, they do open up new directions for future research.

Firstly, as already mentioned, we need to compare the fidelity of SHAP values to the fidelity of other local explanation methods. Only then is it for the long-term possible to set up a baseline, of what it means for an explanation to be faithful. Concerning the Impact Score specifically, we need a clear guideline on how high a satisfactory Impact Score should be.

Secondly, to assess the overall potential of SHAP as an explanation method, we need to establish whether the fidelity of SHAP is consistent for different classification problems or whether results can vary considerably. If this proves to be the case, it needs to be studied for which problems SHAP values are faithful and for which ones they are not and if the same is the case for the fidelity of other feature importance values. One aspect that may be taken into account here, is that the fidelity of explanations might be connected to the accuracy of the models they originate from [20]. Our case study gives further evidence for this suggestion, since both the accuracy of the German Credit classifier (see Section 5) and the fidelity of the corresponding SHAP values is lower than both measures for the Portuguese Class classifier. Investigating this further, might be a good starting point to understand the circumstances under which explanations are guaranteed to be faithful or not. Of course, it is here useful not to only investigate this for SHAP values but also for any other local explanation methods.

7.2 GIRP - Top_jSimilarity

7.2.1 Implementation

The objective of Research Question 3 was it to assess the fidelity of the GIRP explanations. Traditionally the fidelity of surrogate models is measured by comparing the output of the simplified model to the output of its original model for a set of inputs. The Cohen's Kappa score, denoted by k, can then be calculated to measure the agreement between both models while controlling for the agreement that would be measured by chance:

$$k = \frac{p_o - p_e}{1 - p_e}, \tag{8}$$

where p_o indicates the observed chance that both models agree on their decision, while p_e quantifies the expected chance of agreement if both models where completely independent.

Cohen's Kappa alone indicates how similar the outputs of two models are, but a high score does not guarantee that both models derive their outputs in a similar way. Out of this reason, Messalas et al. propose the Top_jSimilarity to measure this aspect of a model's fidelity. To obtain this score, it is first required to calculate the local feature importance values for both the original model and the surrogate model for a set of input instances $X = x_1, x_2, ..., x_n$. We can then define the top_j most important features of the original model ($orig_j$) and the surrogate model ($orig_j$) and examine how many top_j features they have on average in common. By then dividing this result by j, we see how much the two models agree on the top_j most important features on average. For a formalized version of this approach, refer to equation 9

$$\text{Top}_\text{j}\text{Similarity} = \frac{\frac{1}{n}\sum_i orig_j(i) \cap sur_j(i)}{j} \tag{9}$$

Just like the rest of the code, the implementation of Top_j was written in Python and can be found online. Since the `shap` tool package could not be used for the resulting tree structures of the self-implemented GIRP trees, the code also includes an algorithm to derive the SHAP values from scratch.

7.2.2 Results

To get an idea on the fidelity of the GIRP trees for both the German Credit and Portuguese Class problems, both corresponding datasets were split into 10 different train-validation-test-splits. For each split, the best tree was determined for the validation set and then the Cohen's Kappa, the Top_1-, Top_3- and Top_5Similarity was calculated for the test set. The averaged results are visualized in Table 7.

	Cohen's Kappa	**Top_1Sim**	**Top_3Sim**	**Top_5Sim**
Portuguese Class	0.8926	0.9461	0.6677	0.6806
German Credit	0.6964	0.6910	0.6903	0.8218

Table 7: Fidelity results of GIRP explanations

7.2.3 Discussion

Looking at Table 7, we see that the Cohen's Kappa for the GIRP tree of the Portuguese Class problem is quite high. In the paper, where Cohen's Kappa was firstly introduced, it was stated that a score between 0.81 and 1.00 indicates nearly perfect reliability between two raters [52]. Thus looking at this measure only, the fidelity from the surrogate model to its original one seems to be quite high. Looking at the Top_j measures, this finding is partly confirmed and partly contradicted. To begin, the Top_1Similarity is very high, showing that in 94.61% of all times the two models agree on the most important local feature. However, the Top_3- and Top_5Similarity are a little less satisfactory. With a score of 68.06% the Top_5Similarity does not fulfil the suggested criterion, that the surrogate and the original model should have an 80% agreement on the top 5 most important features [33]. How much this finding reduces our faith in the produced GIRP trees is

somewhat up for discussion. To gain a slightly larger perspective on this problem, it may be beneficial to look at the results of the German Credit GIRP trees.

For the original and surrogate model of this classification problem, we observe a Cohen's Kappa score of 0.6964. This is lower than the score for the Portuguese Class problem, but still indicates a "substantial" agreement between the output of the two models [52]. Moreover, the Top_1Similarity is considerably lower than for the first classification problem, but interestingly the opposite holds for the Top_5Similarity, which slightly exceeds the suggested baseline of 0.8.

Looking at the contrasting results of the two classification problems, it becomes clear that there are different aspects to the fidelity of GIRP trees and that the importance we ascribe to each aspect may depend on more than the Top_jSimilarity and the Cohen's Kappa. To demonstrate, look again at the Top_1- and Top_5Similarity score of the German Credit GIRP trees. The surrogate GIRP tree model and the original XGBoost classifier do not have a very high agreement on the most important feature in a classification task. Nevertheless, the high Top_5Similarity score suggests that the most important feature for one of the models may be the second to fifth most important feature for the other model. Whether this is a problem for the fidelity of the results, depends on the difference in importance contributed to that particular feature. If the feature importance values of the five most important features from the original model are very close to each other, it seems acceptable for the surrogate model to have a slightly different ordering of those five most important features, as long as their actual importance values are still close to each other. We can check whether this proposition is true, by looking at Figure 20 from the previous section. Here we see that the normalized SHAP values of the most important features indeed lie close to each other, especially compared to those of the Portuguese Class problem (Figure 19). With this knowledge in mind, our fidelity in the GIRP trees does not necessarily get reduced when looking at the somewhat small Top_1Similarity score of the German Credit GIRP tree.

The measurements for the Portuguese Class trees supports the view that the Top_jSimilarity does not give a complete picture of the fidelity of the GIRP trees. The fact that the Top_1Similarity is so high and the Top_5Similarity is rather low, may be because the original model relied heavenly on the most important features and all other features only had minor impacts on the decision. Again, this suspicion is confirmed by looking at Figure 19, where the normalized importance value of the most important features is much higher than any of the other variables. This, in turn, explains why the surrogate model is less effective at highlighting the role of these less important features and why the Top_5Similarity score is on the lower side, but the Top_1Similarity is so high.

Relating these findings to the third research question, there does not appear to be a clear answer on how faithful GIRP explanations are to their original models. Nevertheless, our case study gives possible directions for future research. The most important one may be, to further search for measures in which the fidelity of GIRP trees can be understood. Especially the normalized SHAP values of the most important features can be complementary to the already used measures. Another way to investigate the fidelity of GIRP trees is to pay attention to the interaction of features, rather than at each feature separately. In the GIRP trees, it is very easily visualized on how the value of one feature, can influence the impact of the value of another feature. It would be interesting to see whether these proposed interaction effects are also present in the model the trees originate from.

On top of that, we might need to give special treatment to potentially sensitive variables in a black-box model. As mentioned in Section 2.1, one of the purposes of xAI is to detect possible biases in algorithms. If a black-box model bases its output on variables like "gender" or "nationality", it is crucial that these are also the variables highlighted by the surrogate model. Since the Top_jSimilarity and Cohen's Kappa score

might still be high if these variables are not captured, it is important to find more enhanced ways to treat these sensitive variables.

One final reason to search for alternatives for the Top_jSimilarity measure is that the calculation of the score relies on the same local feature importance values that the generation of the GIRP tree was based on. If those feature importance values are not faithful to begin with, the tree will be equally unfaithful, without, however, this being reflected by the Top_jSimilarity score. Thus, it would be advantageous to come up with a fidelity measure that is not dependent on potentially inaccurate measures, like SHAP.

8 Experimentation on Explanations' Stability

The final criterion the explanations were tested on, concerned their stability. This criterion is also referred to as the robustness or sensitivity of explanations and is an important quality to establish trust in them. In this section, we will discuss how the stability of SHAP and GIRP explanations were measured and what the results imply for the future of xAI.

8.1 SHAP - Sens_Max

8.1.1 Implementation

Local explanation techniques like SHAP can be called stable, if an insignificant change in the input they need to explain, does not lead to a big change in the explanation itself. In other words, we want the explanations for similar input instances, to be similar as well. This intuition is captured in the Sens_Max score, as proposed by Yeh et al. [50]. The idea here is to randomly add noise to an input instance x, such that the resulting instance x'' still lies within a given neighbourhood radius r. We can then use to explanatory function Φ, that is meant to explain black-box model f, to generate explanations for both x and x''. In the case of SHAP, these explanations consist of importance values for each feature and thus can be represented by 1-dimensional vectors. The distance between the explanatory vectors for x and x'' can then be calculated for all x of a test set X. Yeh et al. propose to take the maximal brought change to an explanation, over all instances x, as the final indicator for an explanations' sensitivity. Thus this gives the following equation:

$$Sens_{Max}(\Phi, f, x, r) = \max_{||x''-x||<r} ||\Phi(f, x'') - \Phi(f, x)|| \tag{10}$$

We measured the stability of SHAP by calculating this score for the test sets of 10 different train-test-splits and then averaging this result. This was done for both the Portuguese Class- and German Credit problem and different radius parameters. These were set to range from 0.1 to 1.0 with increments of 0.1. This can give an idea of how much perturbation is needed to bring bigger changes to the explanations.

8.1.2 Results

The figure below shows the Sens_Max value for the different radii parameters. The dark grey line represents this score for the Portuguese Class explanations, while the light grey line represents the stability of the German Credit explanations.

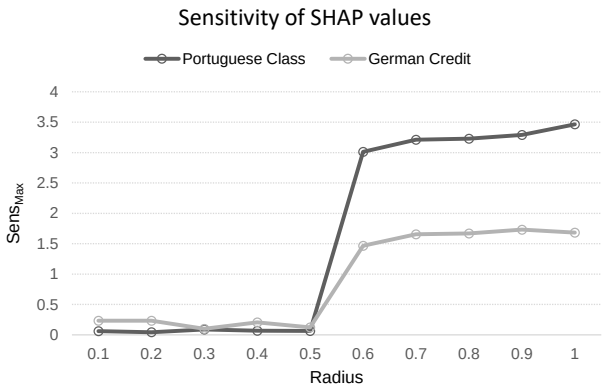
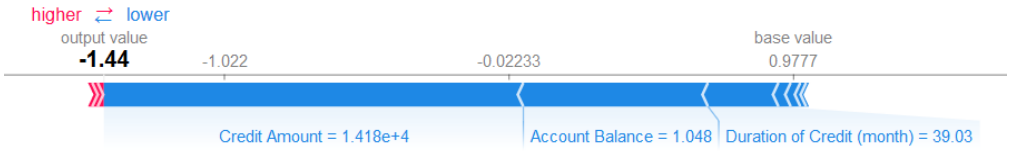


Figure 21: The Sens_{Max} score calculated with different radii for the Portuguese Class and German Credit explanations

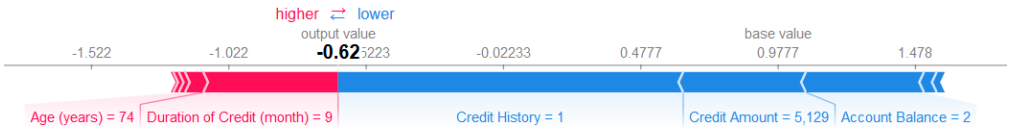
Since this plot alone does not give a very clear indication of what a low or high Sens_{Max} score implies, the figures below visualize two examples of the explanation of an original input instance and a perturbed verison of it. Both explanations refer to the German Credit classification problem. In figure we see the explanation for an input instance before and after perturbation with a radius of 0.1. The distance between these explanations, or in other words the sensitivity score, was here 0.069 and therefore quite low. Figure 23 on the other hand shows the explanation of a heavily perturbed input instance (radius = 1.0). Here the distance between the two explanations was observed to be relatively high, namely 1.567.

(a) Original explanation before perturbation

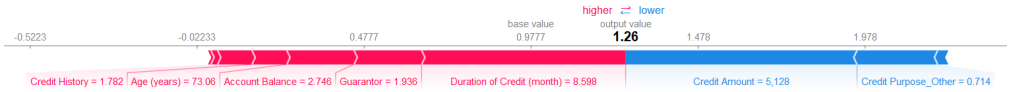


(b) Explanation after perturbation

Figure 22: Explanations before and after perturbation with a low neighbourhood radius (r = 0.1)



(a) Original explanation before perturbation



(b) Explanation after perturbation

Figure 23: Explanations before and after perturbation with a high neighbourhood radius (r = 1.0)

8.1.3 Discussion

Looking at Figure 21, we see that the Sens_{Max} score under different radii follows a similar pattern for both classification problems of our study. Until a radius of 0.5, the scores range between 0.05 and 0.25 and do not vary a lot between the different radii. In the paper where the Sens_{Max} score originates from, only the most stable explanatory techniques yielded this low scores [50]. Thus this is an indication that the SHAP values for our classification problems fulfil the stability criterion. This idea is further strengthened when looking at Figure 22. Here we see that an explanation with a low Sens_{Max} score is indeed not affected by a small perturbation in the input it explains. The explanations before and after perturbation of the input instance are close to identical.

For all radii above 0.5, a drastic increase is observed in the measurement. Intuitively this increase makes sense since some of the variables of both classification problems can only take a very small range of values. In case of the Portuguese Class dataset, the variable "failures" e.g. only ranges from 0 to 3, while for the German Credit dataset the feature "Account Balance" also has a limited range from 1 to 4. Since it already has become apparent that both variables are highly impacting for their corresponding classification tasks, it makes sense that increasing/decreasing these feature by 0.5, can drastically change the decision outcome and the explanation for those outcomes. Out of this reason, it also makes sense that a higher increase for the Sens_{Max} score is observed for the Portuguese Class explanations: in this classification problem, most of the variables can only take a very small range of values, whereas some variables of the German Credit task (e.g. "Credit Amount" or "Duration of Credit") are less limited in scope and therefore the explanations are less prone to perturbations in the data. A glance at Figure 23 confirms the above suspicions. We see in Sub-figure 23a that the feature value for "Account Balance" for the original input is equal to 2 (indicating that a loan applicant had less than €0 on their account) and that this value negatively impacts the decision to give the applicant a loan. However, in the perturbed input the "Account Balance" takes a value of 2.746, which is much closer to the maximum value for that feature. Here this value positively impacts the decision to give a loan. In fact, we even see that the output of the perturbed input instance has changed. In 23a a log-odds ratio of -0.62 was given as an output, corresponding to a probability of 0.349 of a loan applicant receiving a loan. The output for the perturbed instance was 1.26, indicating a substantially higher probability of 0.779 of handing out a loan. Thus, from this perspective, it makes sense that the feature importance values change with a change in output. Though the initial results of the stability analysis seem promising, there are

some downsides of the used approach that need to be discussed. As might have already been noticed by the given examples, perturbing the input instances can lead binary or categorical variables to take on continuous values. This can make the interpretation of the Sens_{Max} score a bit arbitrary since it is not clear what it means for a binary variable to have a value of '0.5' or '0.4'. Since the work where the Sens_{Max} score originates from does not give any indication on how to handle these cases, they were for now not given special attention in the analysis of the results. However, it might be worthwhile to investigate how these cases can be handled more elegantly.

Another shortcoming of the Sens_{Max} score is that it only relies on the change in explanation, that comes with a perturbation in the input. As already has been demonstrated, adding noise to an input instance can in some cases change the classifier's prediction for that instance, indicating that the machine learning model itself is not very stable. In those cases, it seems desirable for the explanations to reflect this instability, and thus have a change in feature importance values as well [2]. To get a more extensive view on the stability of local explanations, it would, therefore, be useful to integrate this idea in the Sens_{Max} score.

If these issues are addressed in future work, more effort can be put in establishing the general stability of SHAP values. Here the same recommendations hold that were already given for assessing the fidelity of SHAP values: it needs to be investigated how stable the results are for different classification problems, tackled by different machine learning models. Moreover, it can be worthwhile to compare these results to the stability of other local explanation methods, to then establish a baseline of what it means for an explanation to be stable.

8.2 GIRP

8.2.1 Implementation

The notion of stability is slightly different for global than for local explanations. While for local explanations it is checked how much explanations change for similar inputs, it needs to be investigated how much global explanations depend on the data they have been trained on. In the case of GIRP trees, this can be easily studied by comparing trees trained on different samples of the dataset. Ideally, one would expect that if the resulting trees are close to each other, their performances on the test set should be close to each other as well [12]. To measure the stability of our GIRP trees were generated them for the Portuguese Class and German Credit dataset of 10 train-test-splits. The Cohen's Kappa and Top_5Similarity measure were taken as performance indicators. On top of that, the number of nodes were calculated for each tree to get an idea of their different sizes.

8.2.2 Results

In Figure 24 we visualized how Cohen's Kappa and Top_5Similarity vary for GIRP trees, trained on different portions of the Portuguese Class and German Credit dataset. The numbers on the data points correspond to the number of nodes in the tree.

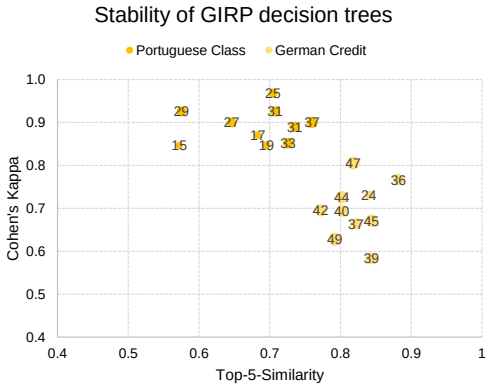


Figure 24: The performance measures of GIRP trees generated on training sets of 10 different train-test splits. The numbers on the datapoints correspond to the number of nodes in the trees

To get a better idea on whether the performance measures say anything about the nature of the trees, we visualized the tree corresponding to the lowest and highest Top_5Similarity. Figure 25 is the tree that yielded the lowest performance, with a Cohen's Kappa of 0.847 and a Top_5Similarity of 0.572. Figure 26 on the other hand, represent the tree with the highest Top_5Similarity score of 0.76. The Cohen's Kappa for this tree was 0.90.

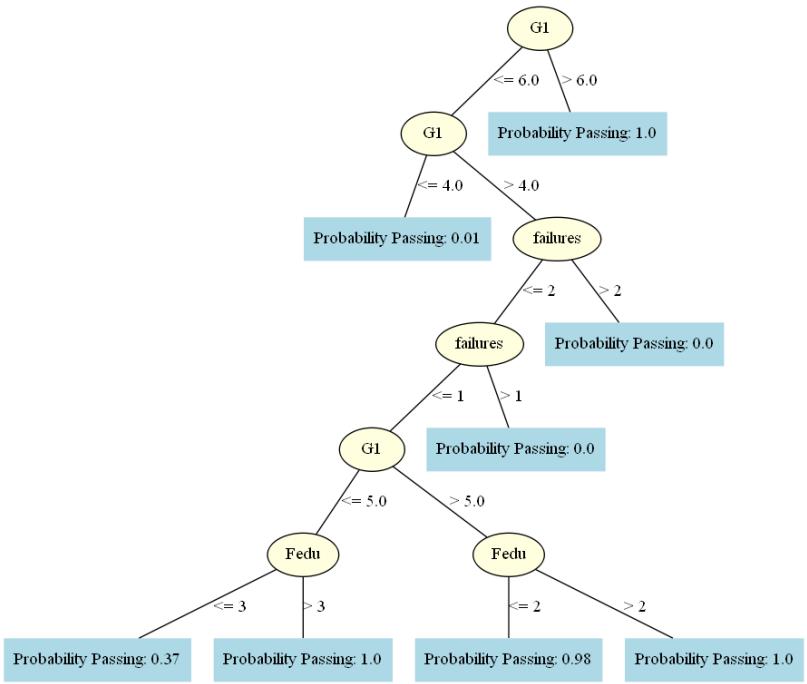


Figure 25: The GIRP tree for the Portuguese Class problem with the lowest performance

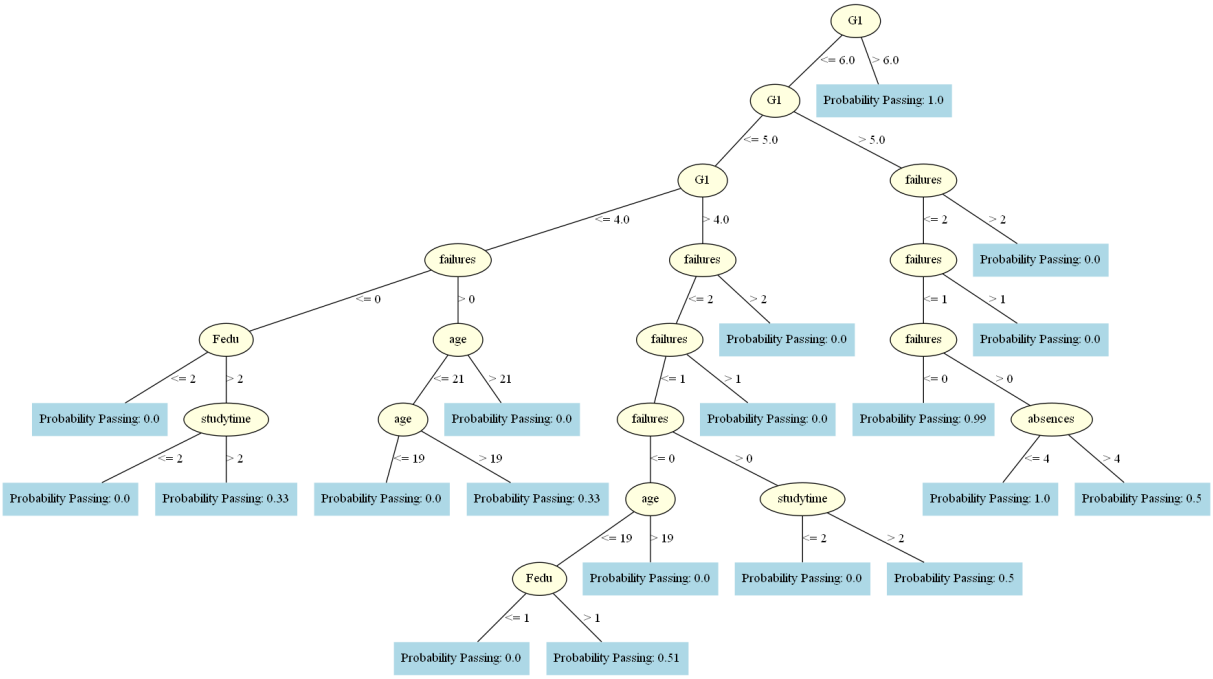


Figure 26: The GIRP tree for the Portuguese Class problem with the highest performance

8.2.3 Discussion

Looking at Figure 24 we see that the Top_5Similarity for the Portuguese Class problem varies considerably between trees, while for the German Credit trees the Cohen's Kappa is quite unstable. In both cases, the worst and highest performance measure have a difference of around 0.2. Though the Top_5Similarity is more stable for the German Credit trees and the Cohen's Kappa is more stable for the Portuguese Class trees, both measures can still vary up to 0.1 in the worst case. Not only the performance measures appear to be unstable, but also the tree sizes. Depending on which sample of the dataset the trees were based on, the biggest tree can have up to 20 more nodes than the smallest tree. Unfortunately, there is no clear pattern on how the size of trees relates to their performance. One might be inclined to think that bigger trees catch more details of a decision process and therefore perform better. Another possibility might be that smaller trees perform better because they are less over-fitted on the training data they originate from. However, a look at Figure 24 does not confirm either of these suggestions. For the Portuguese Class problem, the biggest tree yields one of the best performances, while for the German Credit problem a middle-sized tree performs best. For both tasks, the smaller trees sometimes perform quite bad, but sometimes perform above average: thus the relationship between tree size and tree performance appears to be quite coincidental.

Altogether these results are quite alarming for the stability of GIRP trees. This is even more so the case when looking at Figures 25 and 26. Already on the first glance, it becomes apparent how much they differ, since one is a lot bigger than the other. But also after a more detailed look, we see that not only their size varies but also the variables that both trees contain. In 26 it appears that variables like "age" or "studytime" play a substantial role in the decision process that the tree is meant to explain. Because of this, our case study is a good demonstration of how important the stability of explanation methods is. If users would be presented with both of the trees they would probably not

know which one to rely on. Of course, it could be argued that one could simply present the better performing tree to the users, but even in that case, one would not know for sure whether the corresponding tree only performs well on the corresponding test set, or whether it would also generalize well to other data.

With this being said there are several directions that future research about the stability of global explanation methods can take. Given that the usability and fidelity of GIRP explanations were quite satisfactory, it would be worthwhile to study how the stability of this explanation method can be improved. Some ideas on how to accomplish this have already been given by the authors who proposed GIRP [49]. They mention that the process of determining the best splitting node while building the tree may be arbitrary since no confidence in the split strengths (which are used to choose the next splitting nodes) is established. By already making use of bootstrapping in the process of building the tree, it would be possible to estimate whether the split strength of a node is high by coincidence, or whether it is also high for other portions of the data. With this knowledge, only nodes with high confidence in their split strength could be used to build trees, which could result in more stable and possibly more faithful explanations [49]. Another way of generating a more stable explanation is by averaging the trees obtained by different train sets. Similar work has already been done for decision trees in general, and could serve as an inspiration for this idea [39].

If it proves to be impossible to increase the stability of GIRP trees, it may be worthwhile to look more deeply into the performance of other global explanation methods. It might be the case that surrogate models like GIRP are inherently unstable. After all, they are meant to capture as many details as possible from a very complex model in a simple, small one. When two surrogate models, trained on different portions of a dataset, differ greatly from each other it might be the case that one of them simply catches different details of a black-box model than the other one. This, in turn, would make us believe that the models are unstable. Other global explanation methods, like the use of partial dependence plots (see Section 2.2.2), might suffer less from this disadvantage as they are meant to only catch information of one variable at a time in a plot. It would be interesting to further investigate whether this more specific-oriented way of explaining models, is less prone to instability.

If this turns out to be the case, it would, of course, be important to study the fidelity and usability of these methods as well.

9 Conclusion

The main goal of this study was to assess how users' understanding of AI models is affected by the presentation of SHAP, compared to GIRP explanations or a combination of both. SHAP is a local explanation method, where for individual input instances importance values are assigned to every feature of that input, indicating how influential each feature is for a decision process. GIRP, on the other hand, is a global explanation method that tries to capture the overall decision process of a black-model in one decision tree.

Users' understanding of these explanation techniques was measured in a three-fold way; paying attention to users' objectively measured understanding, their self-rated satisfaction with explanations and their ability to utilize the explanations to find bias in AI models. It was found that the first and last aspect of understanding was more facilitated by the presentation of GIRP explanations rather than SHAP explanations. We hypothesized that this was the case because global explanations provide a more extensive overview of an AI model than local ones, which makes it easier to derive conclusions about them. In terms of self-rated satisfaction, SHAP explanations scored highest, possibly because they are less abstract and easier to relate to. Though individually, the

findings make sense it was somewhat surprising that SHAP explanations were rated as more useful than GIRP, even though the "objective understanding" measures suggest that the opposite is the case. We explained this somewhat contradicting finding with the Dunning-Kruger effect: humans who perform poorly on a task are often inclined to overestimate their abilities for solving that task, which may also lead them to rate the material provided to solve the task (in this case the SHAP explanations) as more useful. While this effect has been observed in the context of information processing, it has not, in our best knowledge, before been found in the context of xAI. One of the main contributions of our study is therefore to highlight the importance of measuring user's understanding in multiple ways and to not take for granted that subjective understanding also implies objective one.

The second main contribution of our study relates to the potential of combining local and global explanations to facilitate users' understanding of AI models. Despite the findings that SHAP and GIRP are beneficial for different aspects of users' understanding, combining both explanation types in the usability study did not allow users to get the best out of both. In the first part of our usability study, we found that SHAP explanations were preferred over the explanation-combination, while in terms of objective understanding we observed a trend that GIRP explanations were more beneficial. We ascribed these results to the possibility that combining explanations caused information overload in users. Instead of benefiting from the different aspects highlighted by the local and global methods, users might have felt overwhelmed by all the material and did not know how to process it. Interestingly, the second part of the usability study suggests that the phenomenon of information overload might be reduced by increased task relevance or motivation for the given task. In this part of our study, users were presented with explanations generated for two AI models, that could decide whether loan applicants at a bank get a loan or not. Being instructed to choose the better of both models, users might have better understood the purpose and relevance of xAI and therefore experienced less negative consequences of being presented with both local and global explanations. In terms of objective understanding, participants of the SHAP+GIRP condition were equally good as participants of the GIRP condition in choosing the better model. In terms of self-rated satisfaction, no significant differences between the SHAP and SHAP+GIRP were observed anymore and participants gave higher usability ratings than in the first part of the survey. Thus, despite some negative consequences of presenting two explanation types, combining different explanation types may still be beneficial under certain circumstances.

While the main objective of our study concerned the usability of explanations, we also put some effort into measuring their fidelity and stability. Fidelity refers to the extent to which explanations accurately reflect the inner-workings of their original models. Stability, on the other hand, concerns the degree to which explanations change when being based on similar input. Both criteria are crucial qualities of explanations, after all, an explanation that is usable but not accurate or stable should not be trusted by users. The extent to which both criteria were fulfilled, was measured for the GIRP and SHAP explanations generated for two classification tasks.

We measured the fidelity of SHAP, by establishing the most important features of different input instances, and seeing how the decision process for these instances was affected when leaving out these features. For one of the classification problems, leaving out features with higher importance values always resulted in bigger changes in decision outcomes, than leaving out features with slightly lower importance values. While this finding matched the intuition of how SHAP values should behave, it could not be replicated for the explanations of the second AI model. Thus overall, it appears that the fidelity of SHAP values is not guaranteed to be high at all times and that we should either understand the circumstances under which the values are faithful, or we should

look for an alternative local explanation method. The results for the stability analysis of SHAP were more promising. We measured stability, by adding small perturbations to input instances and observing whether these changes lead to substantial differences in the explanations generated for them. As desired, only big perturbations were highly impacting for the explanations.

Concerning GIRP, we made use of the Cohen's Kappa and the Top_jSimilarity score to measure the explanations' fidelity to their original models. The first quantifies the similarity between outputs of the original model and the GIRP trees, while the latter indicates how similar the most important local features are for the decision processes of both models. For one of the classification problems, the Cohen's Kappa was very high and the original model and the GIRP tree nearly always agreed on the most important feature. However, there were more disagreements on the importance of features with slightly lower importance ranks. The opposite was found for the second classification problem: here Cohen's Kappa was a bit lower, and while the original and GIRP model did not always agree on the most important feature, there was a higher agreement on the second to fifth most important features. Looking at these results it was concluded, that there are different aspects to the fidelity of GIRP trees and that the aspect we find most important, may depend on the nature of the classification problem.

Lastly, we measured the stability of GIRP by generating the explanatory trees on different training sets of 10 different train-test splits. By measuring the Cohen's Kappa and the Top_jSimilarity on the corresponding test sets, we observed that the performance of the varying trees varied considerably. By also looking at the different sizes of the trees and visually inspecting some of the results, we saw that not only the performance but also the nature of the trees was not consistent. Altogether, these findings were not encouraging for the overall stability of GIRP, and we advise to put efforts into either improving this explanation method or finding a better alternative.

9.1 Future Research

Though our study has brought some new knowledge to the growing research field of xAI, some questions remain unanswered and should be investigated further. Concerning the usability of explanations, it should be studied whether our proposed theory about the presentation of local and global explanations is true. More specifically, we should find out whether users indeed perceive information overload when being presented with two explanation types and whether this effect can be diminished by increased task relevance or increased users' motivation. If this appears to be the case, it would be worthwhile to see how the presentation of the explanation methods can be made more concise, such that no information overload occurs.

Another suggestion directly connected to this is to further increase the general usability of SHAP and GIRP explanations. As discussed in Section 6.5.3 there were some complaints that the SHAP plots and GIRP trees by themselves were not easy to understand. Thus trying to identify their flaws and trying to improve them, might help in reducing the possible information overload that might be caused when both of them are presented.

Another area of further research does not only concern the usability, but also the fidelity and stability of explanations. In our study, we only generated explanations for tabular data, which is already intuitive in nature. It would be interesting to see how the explanations behave for image or textual data, and whether the use of this data would affect the found results for the three quality-criteria. Since SHAP values and GIRP trees were designed to be used for any kind of classification problem, it would certainly be possible to pay more attention to this in the future.

Lastly, it should be noted that most of the new questions opened up by this research

concern the fidelity and stability of the used explanations. For both criteria and both explanation types it is firstly advised to set up a baseline of what it means for local and global explanations to be faithful or stable. As mentioned in the corresponding sections, it might also be worthwhile to look at different measures to evaluate these criteria, as all of them had different limitations and shortcomings. Since the stability of GIRP trees raised the highest concerns, it is advised to try to improve this quality. If this proves impossible, it is advised to look for a more stable global explanation method.

References

[1] Adadi, A., and Berrada, M. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access 6* (2018), 52138–52160.

[2] Alvarez-Melis, D., and Jaakkola, T. S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049* (2018).

[3] Andras, P., Esterle, L., Guckert, M., Han, T. A., Lewis, P. R., Milanovic, K., Payne, T., Perret, C., Pitt, J., Powers, S. T., et al. Trusting intelligent machines: Deepening trust within socio-technical systems. *IEEE Technology and Society Magazine 37*, 4 (2018), 76–83.

[4] Atkinson, R. K., Derry, S. J., Renkl, A., and Wortham, D. Learning from examples: Instructional principles from the worked examples research. *Review of educational research 70*, 2 (2000), 181–214.

[5] Brooke, J. System usability scale. *Reading, England: Digital Equipment Corporation 480* (1986).

[6] Byrne, R. Counterfactuals in explainable artificial intelligence (xai): evidence from human reasoning. In *international joint conference on AI (IJCAI 2019)* (2019).

[7] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), ACM, pp. 1721–1730.

[8] Danaher, J. *Automation and Utopia: Human Flourishing in a World without Work*. Harvard University Press, 2019.

[9] Došilović, F. K., Brčić, M., and Hlupić, N. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (2018), IEEE, pp. 0210–0215.

[10] Dunning, D. The dunning–kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*, vol. 44. Elsevier, 2011, pp. 247–296.

[11] Faliagka, E., Iliadis, L., Karydis, I., Rigou, M., Sioutas, S., Tsakalidis, A., and Tzimas, G. On-line consistent ranking on e-recruitment: seeking the truth behind a well-formed cv. *Artificial Intelligence Review 42*, 3 (2014), 515–528.

[12] Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), pp. 329–338.

[13] Fukukura, J., Ferguson, M. J., and Fujita, K. Psychological distance can improve decision making under information overload via gist memory. *Journal of Experimental Psychology: General 142*, 3 (2013), 658.

[14] Garcia, F. J. C., Robb, D. A., Liu, X., Laskov, A., Patron, P., and Hastie, H. Explainable autonomy: A study of explanation styles for building clear mental models. In *Proceedings of the 11th International Conference on Natural Language Generation* (2018), pp. 99–108.

[15] Gerber, M. S. Predicting crime using twitter and kernel density estimation. *Decision Support Systems 61* (2014), 115–125.

[16] Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics 24*, 1 (2015), 44–65.

[17] Goodman, B., and Flaxman, S. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine 38*, 3 (Oct. 2017), 50–57.

[18] Hofmann, H. Statlog (german credit data) data set. `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`, 1994. [Online; accessed 28-October-2019].

[19] Hwang, M. I., and Lin, J. W. Information dimension, information overload and decision quality. *Journal of information science 25*, 3 (1999), 213–218.

[20] Islam, S. R., Eberle, W., and Ghafoor, S. K. Towards quantification of explainability in explainable artificial intelligence methods. *arXiv preprint arXiv:1911.10104* (2019).

[21] Jackson, T. W., and Farzaneh, P. Theory-based model of factors affecting information overload. *International Journal of Information Management 32*, 6 (2012), 523–532.

[22] Kalai, E., and Samet, D. On weighted shapley values. *International Journal of Game Theory 16*, 3 (1987), 205–222.

[23] Kurbanoglu, S. S., Akkoyunlu, B., and Umay, A. Developing the information literacy self-efficacy scale. *Journal of documentation* (2006).

[24] Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154* (2017).

[25] Last, M., Maimon, O., and Minkov, E. Improving stability of decision trees. *International Journal of Pattern Recognition and Artificial Intelligence 16*, 02 (2002), 145–159.

[26] Letham, B., Rudin, C., McCormick, T. H., Madigan, D., et al. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics 9*, 3 (2015), 1350–1371.

[27] Lin, Z. Q., Shafiee, M. J., Bochkarev, S., Jules, M. S., Wang, X. Y., and Wong, A. Explaining with impact: A machine-centric strategy to quantify the performance of explainability algorithms. *arXiv preprint arXiv:1910.07387* (2019).

[28] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. Explainable ai for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610* (2019).

[29] Lundberg, S. M., and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (2017), pp. 4765–4774.

[30] Mahmood, K. Do people overestimate their information literacy skills? a systematic review of empirical evidence on the dunning-kruger effect. *Communications in Information Literacy 10*, 2 (2016), 3.

[31] Martens, D., Baesens, B., Van Gestel, T., and Vanthienen, J. Comprehensible credit scoring models using rule extraction from support vector machines. *European journal of operational research 183*, 3 (2007), 1466–1476.

[32] Mashayekhi, M., and Gras, R. Rule extraction from random forest: the rf+ hc methods. In *Canadian Conference on Artificial Intelligence* (2015), Springer, pp. 223–237.

[33] Messalas, A., Kanellopoulos, Y., and Makris, C. Model-agnostic interpretability with shapley values. In *2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (2019), IEEE, pp. 1–7.

[34] Miller, T., Howe, P., and Sonenberg, L. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).

[35] Molnar, C. *Interpretable Machine Learning A Guide for Making Black Box Models Explainable*. Leanpub, 2019.

[36] Moore, J. D., and Swartout, W. R. Explanation in expert systems: A survey. Tech. rep., University of Southern California Marina Del Rey Information Sciences Institute, 1988.

[37] Muller, T. E. Buyer response to variations in product information load. *Journal of applied psychology 69*, 2 (1984), 300.

[38] Nugent, C., and Cunningham, P. A case-based explanation system for black-box systems. *Artificial Intelligence Review 24*, 2 (2005), 163–178.

[39] Oliver, J. J., and Hand, D. J. On pruning and averaging decision trees. In *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 430–437.

[40] Petkovic, D., Altman, R. B., Wong, M., and Vigil, A. Improving the explainability of random forest classifier-user centered approach. In *PSB* (2018), World Scientific, pp. 204–215.

[41] Ribeiro, M. T., Singh, S., and Guestrin, C. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), ACM, pp. 1135–1144.

[42] Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[43] Samek, W., Wiegand, T., and Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[44] Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), JMLR. org, pp. 3145–3153.

[45] Taddeo, M., and Floridi, L. How ai can be a force for good. *Science 361*, 6404 (2018), 751–752.

[46] Tan, S., Caruana, R., Hooker, G., Koch, P., and Gordo, A. Learning global additive explanations for neural nets using model distillation. *arXiv preprint arXiv:1801.08640* (2018).

[47] Tullis, T., Fleischman, S., McNulty, M., Cianchette, C., and Bergel, M. An empirical comparison of lab and remote usability testing of web sites. In *Usability Professionals Association Conference* (2002).

[48] Vetrò, A., Santangelo, A., Beretta, E., and De Martin, J. C. Ai: from rational agents to socially responsible agents. *Digital Policy, Regulation and Governance* (2019).

[49] Yang, C., Rangarajan, A., and Ranka, S. Global model interpretation via recursive partitioning. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* (2018), IEEE, pp. 1563–1570.

[50] Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems* (2019), pp. 10965–10976.

[51] Zilke, J. R., Mencía, E. L., and Janssen, F. Deepred–rule extraction from deep neural networks. In *International Conference on Discovery Science* (2016), Springer, pp. 457–473.

[52] Zouari, H., Heutte, L., and Lecourtier, Y. Controlling the diversity in classifier ensembles through a measure of agreement. *Pattern Recognition 38*, 11 (2005), 2195–2199.

Appendices

A GIRP Tree German Credit problem

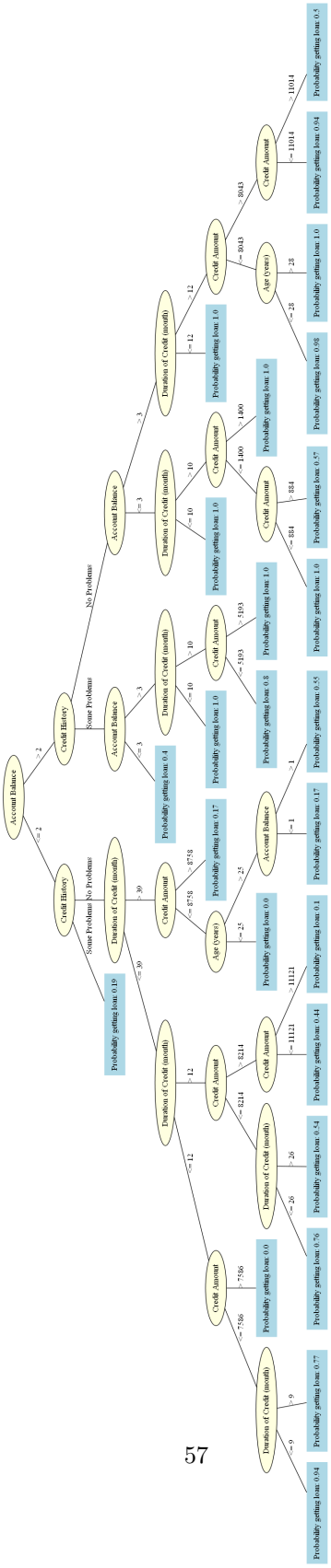


Figure 27: The optimal GIRP tree obtained for the German Credit problem

B Usability Survey[3]

Start of Block: SURVEY INSTRUCTION

Dear participant,

This study is conducted as a part of a Master Thesis of the study programme 'Artificial Intelligence' at the Radboud University Nijmegen.

The research is about testing the understandability of AI-systems.
No worries: *To participate you don't need any technical background, but we do ask you to stay focussed while filling in the survey. It should take around 15 minutes to complete it.*

Please note the following points:

- The complete survey is in English and we kindly ask you to provide your answers in English as well
- During the study you can in any moment of time quit participating, without you having to explain why you want to quit. Quitting during the study has no consequences whatsoever.
- The information that we collect will be anonymously processed. This means that later on the results cannot be traced back to you. The consequence of this is that we cannot inform you about your personal results after the study has been completed. However, we could inform you about the results of the study as a whole. If you wish to be informed about the results of this study, then please let us know.

Possible questions you have as a result of this information, you can ask by sending a mail to d.lenders@student.ru.nl

Finally, it is important that you fill out the survey on a laptop, computer or tablet rather than your phone. Can you please confirm that you are not using a phone to fill out the survey?

○ No, I'm not using a phone

End of Block: SURVEY INSTRUCTION

[3]This is the survey shown to participants in the "SHAP+GIRP" condition. According to the condition other participants were allocated in, they only got presented one of both explanation techniques.

Start of Block: Introduction Portuguese Class

Imagine that you are a researcher and you want to know which factors affect students' performance in school. You visit a high school to follow a group of students who are taking a Maths course. There you collect data of the students and set up student profiles. These profiles contain information about the student's age, their gender, their past study performances and more. In the box below you can see how a student-profile is set up.

Gender—Describes the gender of the student *(female/male)*
Age — Describes the age of the student *(number between 15 and 22)*
Dad-edu—Describes the education level of the student's father. It can have one of the following five values:
None *Primary Education (4th grade)* *5th to 9th grade* *Secondary Education* *Higher Education*
Mom-edu—Describes the education level of the student's mother. It can have one of the following five values:
None *Primary Education (4th grade)* *5th to 9th grade* *Secondary Education* *Higher Education*
Past Grade — Describes the grade the student received for the first exam of the course *(number between 0 and 10)*
Failures — Describes the number of times the student failed the class before *(number between 0 and 4)*
Absences —Describes how many times the student was absent from school *(number between 0 and 93)*
Studytime —Describes how many hours the student studied for the exam and can be one of the following values
less than 2 hours *2 –5 hours* *5 –10 hours* *more than 10 hours*
External Activities — Denotes whether the external student engages in any activities outside school *(yes/no)*

At the end of the course the students have to take an exam and for each student you note down whether they pass or fail it. You are interested to see whether the information on the student-profiles can help in predicting the student's performance on the exam.
Your friend wants to help you with this and develops a computer system. This system takes a student profile as an input and tries to use the information in there to predict whether the student passes or fails the exam. Apart from that, the system also provides an explanation of how its predictions were derived.
In this survey you are going to look at the explanations of the system and judge their quality.

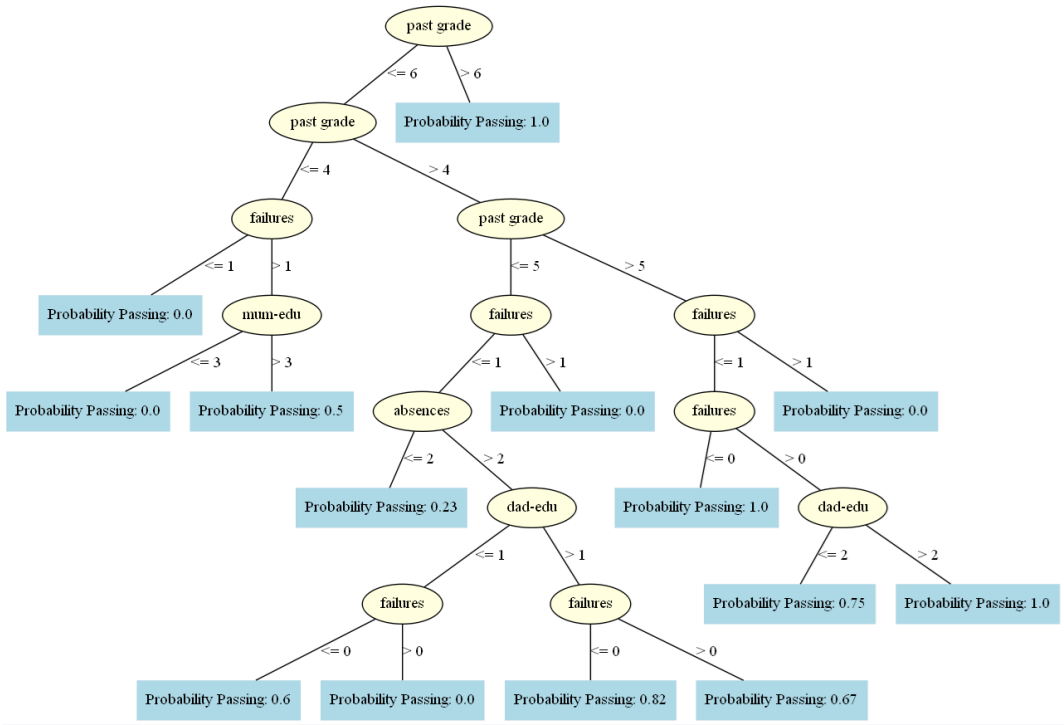
End of Block: Introduction Portuguese Class

Start of Block: Portuguese Class - SHAP + GIRP

We're now going to see how the system uses the provided student-profiles to make conclusions about whether students pass or fail an exam. If you are not sure about the different characteristics of a student profile, please refer to the box below:

Gender—Describes the gender of the student *(female/male)*
Age — Describes the age of the student *(number between 15 and 22)*
Dad-edu—Describes the education level of the student's father. It can have one of the following five values:
None *Primary Education (4th grade)* *5th to 9th grade* *Secondary Education* *Higher Education*
Mom-edu—Describes the education level of the student's mother. It can have one of the following five values:
None *Primary Education (4th grade)* *5th to 9th grade* *Secondary Education* *Higher Education*
Past Grade — Describes the grade the student received for the first exam of the course *(number between 0 and 10)*
Failures — Describes the number of times the student failed the class before *(number between 0 and 4)*
Absences —Describes how many times the student was absent from school *(number between 0 and 93)*
Studytime —Describes how many hours the student studied for the exam and can be one of the following values
less than 2 hours *2 –5 hours* *5 –10 hours* *more than 10 hours*
External Activities — Denotes whether the external student engages in any activities outside school *(yes/no)*

The figure you see below is called a 'Decision Tree' and is meant to explain the AI-system's general behaviour. You can read the figure like a flowchart: you start at the top of the tree and go left and right according to the characteristics of the student profile. Once you have reached one of the blue boxes, you can see the probability of a student passing the exam.
Please take some time to look at the tree and understand it.



Below you can see how the system reacts to two different students, Student A and Student B. The system predicts that Student A fails the exam, while Student B passes the exam. In the plots below the system's output, you can see how the system explains its decision process. For now you can ignore the numbers on the axis and just look at the pink and blue arrows. The blue arrows show which factors make the AI more inclined to say that a student fails the exam. The pink arrows on the other hand shows which factors make the AI more inclined to say that a student passes the exam. The size of the arrows shows how important these factors are for the final decision.
Please take a look at the system's explanations and try to understand them.

Remember:
blue arrow - student fails exam
pink arrow - student passes exam

Student A:
Gender: Female
Age: 18
Dad-edu: Primary Education
Mum-edu: 5th to 9th Grade
Past Grade: 5
Failures: 3
Absences: 10
Studytime: 2-5 hours
External Activities: Yes

System Output:
*"Student A is expected to **fail** the next exam"*

Student B:
Gender: Female
Age: 16
Dad-edu: Higher Education
Mum-edu: Higher Education
Past Grade: 8
Failures: 0
Absences: 6
Studytime: 5-10 hours
External Activities: Yes

System Output:
*"Student B is expected to **pass** the next exam"*

Based on the explanations you have seen above, you'll now be asked a number of questions to test your comprehension of the AI-system

Engaging in external activities has a big impact on the prediction of whether a student passes or fails the course

- ○ True
- ○ False

If a student already failed the course before, the system is more likely to predict that the student will fail the course

- ○ True
- ○ False

Which of the following characteristics seems to be most influential for the system's prediction?

- ○ The time spent studying for an exam
- ○ The student's grade for the past exam
- ○ The number of times a student was absent

If the student's past grade was at least a 7, does the number of past failures still impact the system's prediction?

- ○ Yes
- ○ No

After having looked at the figures explaining the system's behaviour, please indicate how much you agree to the following statements

(Note: By "system explanations" we mean both the decision tree and the figures with the pink and blue arrows)

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disagree	Strongly disagree
The system explanations were easy to understand	○	○	○	○	○	○	○
The system explanations were unnecessarily complex	○	○	○	○	○	○	○
It would be worth looking at the explanations to understand how the system is behaving	○	○	○	○	○	○	○
I am able to understand the explanations in a reasonable amount of time	○	○	○	○	○	○	○
Overall, I am satisfied with the system explanations	○	○	○	○	○	○	○

Can you shortly explain your answers to the previous questions? Why do you think that the explanations are/are not easy to understand?

_____

_____

_____

End of Block: Portugese Class - SHAP + GIRP

Start of Block: Introduction German Credit Dataset

For the remaining part of the questionnaire imagine that you work at a bank. A couple of clients at the bank want to apply for a loan, but the bank only wants to give a loan if they can be sure that the applicant will eventually pay it back. Two computer systems have been developed that can help you with a decision on whether a loan-applicant gets a loan or not. In order to make this decision the systems take characteristics about loan-applicants, like their age, current account balance, etc. into account. In the box below you can see all the characteristics of a loan-applicant.

Age — Describes the age of the loan applicant *(between 19 and 75 years)*
Credit Amount—Describes how much loan the applicant is looking for *(between 250 and 18424 euro)*
Purpose—Describes for which purpose the applicant wants loan for. It can have one of the following values:
Car *Home-related* *Other*
Duration of Credit (Month) — Describes how long the applicant wants a loan for *(between 4 and 72 months)*
Account Balance — Describes the current amount of money on the loan-applicants account. It can have one of the following values:
No Account *Less than 0€* *0-200€* *More than 200€*
Credit History — Describes whether the applicant had problems before when paying off debts. It can have one of the following values:
No Problems *Some Problems*
Guarantors— Describes whether the loan-applicant has a guarantor (someone who can pay of the loan, if the applicant will not pay it off in time) *(yes/no)*
Telephone—Shows whether a telephone number is registered for the applicant *(yes/no)*

Again, you will look at how the systems explain their decision process. Based on these explanations you are asked to rate which of the two computer systems should be deployed at your bank.

End of Block: Introduction German Credit Dataset

Start of Block: German Credit - SHAP + GIRP

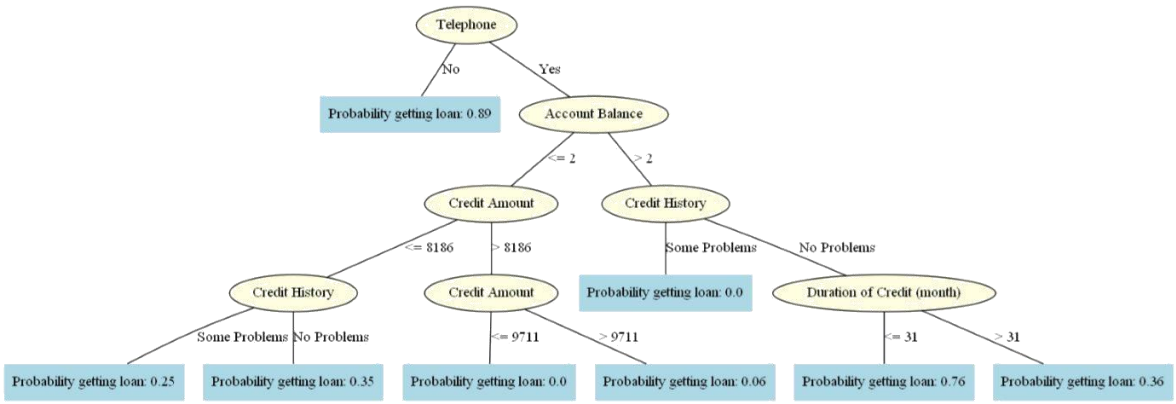
Below you can see how the two systems explain their behaviour. If you're unsure about the different characteristics of a loan-applicant profile, please refer to the box below:

Age — Describes the age of the loan applicant *(between 19 and 75 years)*
Credit Amount—Describes how much loan the applicant is looking for *(between 250 and 18424 euro)*
Purpose—Describes for which purpose the applicant wants loan for. It can have one of the following values:
Car *Home-related* *Other*
Duration of Credit (Month) — Describes how long the applicant wants a loan for *(between 4 and 72 months)*
Account Balance — Describes the current amount of money on the loan-applicants account. It can have one of the following values:
No Account *Less than 0€* *0-200€* *More than 200€*
Credit History — Describes whether the applicant had problems before when paying off debts. It can have one of the following values:
No Problems *Some Problems*
Guarantors— Describes whether the loan-applicant has a guarantor (someone who can pay of the loan, if the applicant will not pay it off in time) *(yes/no)*
Telephone—Shows whether a telephone number is registered for the applicant *(yes/no)*

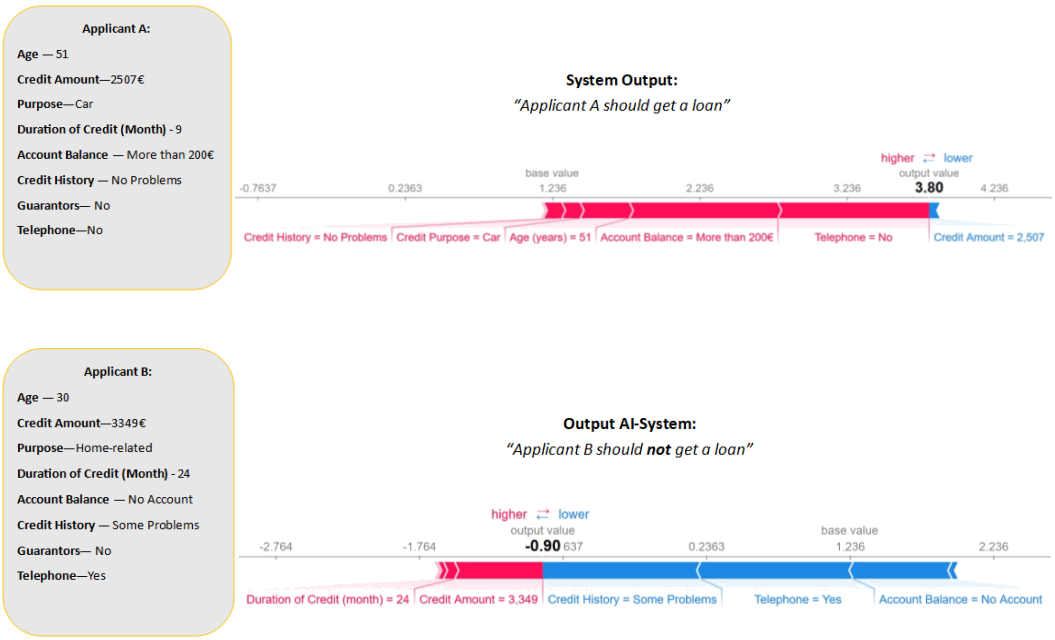
Remember that you can read the decision trees like a flow-chart: you start at the top of the tree and go left and right according to the characteristics of the loan-applicant.

To understand the explanations for the different inputs, look at the pink and blue arrows.
Pink arrows - Applicant gets a loan
Blue arrows - Applicant doesn't get a loan

System 1:
Below you see the decision tree that explains the behaviour for System 1.



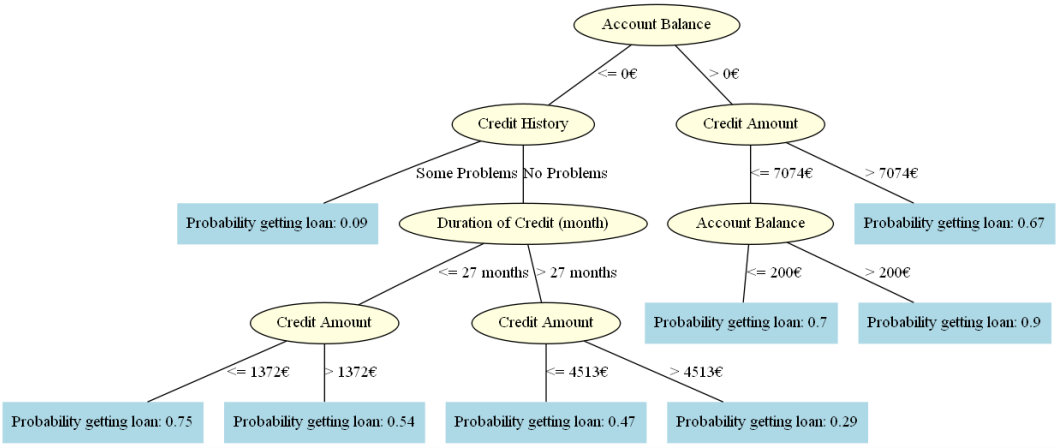
Here you can see how System 1 behaves for two different applicants.



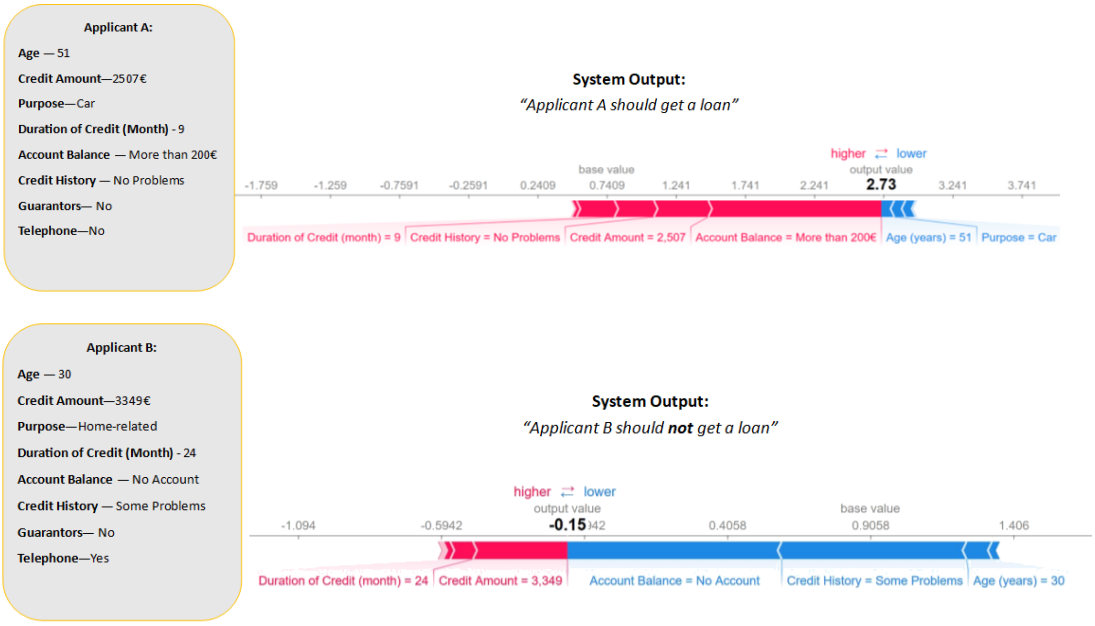
Looking at the explanation of system 1, which of the following characteristics seems to be most influential for the system's decision?

- ○ The current account balance of the applicant
- ○ The duration of the credit
- ○ The registration of an applicant's phone number

System 2:
Below you see the decision tree that explains the behaviour for System 2.



Here you can see how System 2 behaves for two different applicants.



Looking at the explanation of system 2, which of the following characteristics seems to be most influential for the system's decision

- ○ The current account balance of the applicant
- ○ The duration of the credit
- ○ The registration of an applicant's phone number

Look at the explanations of the two systems. Pay attention to the characteristics of an applicant they use to make their decisions.
Which of the systems do you think has a more logical decision process and should be used by banks?

- ○ System 1
- ○ System 2
- ○ They are both equally good

Can you explain your answer to the previous questions? Why do you think that one system is/is not better than the other? What aspects of a loan-applicant profile did you pay attention to?

_____

_____

_____

_____

_____

After having looked at the figures explaining the system's behaviour, please indicate how much you agree to the following statements

	Strongly agree	Agree	Somewhat agree	Neither agree nor disagree	Somewhat disagree	Disagree	Strongly disagree
The system explanations were easy to understand	○	○	○	○	○	○	○
The system explanations were unnecessarily complex	○	○	○	○	○	○	○
It would be worth looking at the explanations to understand how the system is behaving	○	○	○	○	○	○	○
I am able to understand the explanations in a reasonable amount of time	○	○	○	○	○	○	○
Overall, I am satisfied with the system explanations	○	○	○	○	○	○	○

End of Block: German Credit - SHAP + GIRP

Start of Block: Demographics

You've nearly made it to the end of the questionnaire! Before you're done, please answer these last questions

What is your gender?

- ○ Male
- ○ Female
- ○ Other, please specify _____
- ○ I prefer not to say

Please select your age

- ○ 0-15
- ○ 16-24
- ○ 25-34
- ○ 35-44
- ○ 45-54
- ○ 55-64
- ○ 65-74
- ○ 75+
- ○ I prefer not to say

What is your nationality?

_____

How would you describe your English proficiency?

- ○ Basic proficiency
- ○ Intermediate proficiency
- ○ Advanced proficiency
- ○ Native/bilingual proficiency

Please fill in your background (education/most recent field of study/most recent field of work)

_____

End of Block: Demographics