

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Bárbara Gabrielle Coelho de Oliveira Lopes

**Explain-ML: A Human-Centered Holistic and Interactive Visual Tool For
Explainable Machine Learning**

Belo Horizonte
2020

Bárbara Gabrielle Coelho de Oliveira Lopes

**Explain-ML: A Human-Centered Holistic and Interactive Visual Tool For
Explainable Machine Learning**

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Raquel de Oliveira Prates

Co-Advisor: Marcos André Gonçalves

Belo Horizonte
2020

2020, Bárbara Gabrielle Coelho de Oliveira Lopes.
Todos os direitos reservados.

	Lopes, Bárbara Gabrielle Coelho de Oliveira.
L864e	Explain-ML: a human-centered holistic and interactive visual tool for explainable machine learning [recurso eletrônico] / Bárbara Gabrielle Coelho de Oliveira Lopes – 2020. 1 recurso online (140 f. il., color.) pdf. Orientadora: Raquel de Oliveira Prates. Coorientador: Marcos André Gonçalves. Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. Referências: f. 81-84. 1. Computação – Teses. 2. Interação homem-máquina – Teses. I. Prates, Raquel de Oliveira. II. Gonçalves, Marcos André. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação. IV. Título. CDU 519.6*75(043)

Ficha catalográfica elaborada por Célio Resende Diniz, bibliotecário CRB
6/2403 - Universidade Federal de Minas Gerais – ICEx.



UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

EXPLAIN-ML: A HUMAN-CENTERED HOLISTIC AND
INTERACTIVE VISUAL TOOL FOR EXPLAINABLE MACHINE
LEARNING

**BÁRBARA GABRIELLE COELHO DE OLIVEIRA
LOPES**

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

A handwritten signature in blue ink, reading "Raquel Prates".

Profa. Raquel Oliveira Prates - Orientadora
Departamento de Ciência da Computação - UFMG

A handwritten signature in blue ink, reading "Marcos André Gonçalves".

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

A handwritten signature in blue ink, reading "Simone Diniz Junqueira Barbosa".

PROFA. SIMONE DINIZ JUNQUEIRA BARBOSA
Departamento de Informática - PUCRJ

A handwritten signature in blue ink, reading "Anísio Mendes Lacerda".

PROF. ANÍSIO MENDES LACERDA
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 25 de novembro de 2020.

To my beloved mother.

Acknowledgments

First and foremost, I thank God, whose many blessings have made me who I am today. Thanks God, for providing me the opportunity to step in the excellent world of science and granting me the capability to proceed successfully.

Also thank my family who has supported me throughout my academic trajectory despite the self alienation that came as a result of pursuing my goals. Words can not express how grateful I am to my mother and stepfather, for all of the sacrifices that you've made on my behalf. Your prayer for me was what sustained me thus far. To my sister for all the long conversations and for hearing all my complaints about the difficulties of the academic world. I am also grateful to my true friends whose encouraging words kept me going when chocolate had lost its stimulating effect.

To be able to step in this way, I have also been supported and supervised by many people to whom I would like to express my deepest gratitude. Thanks to both my advisors Raquel Prates and Marcos Gonçalves, and professor Liziane Soares for providing guidance and feedback throughout this project, which were determining to the success of this work. I also would like to thank the Graduate Program in Computer Science of the Federal University of Minas Gerais, for providing the material means for this work; my committee members, professor Anisio Lacerda and professor Simone Barbosa; and all the volunteers who participate on this research.

I extend my acknowledgments to all my colleagues from MSc program and PENSI Lab, for the friendship and knowledge sharing. To professor Daniel Dalip, many thanks for all the patience, feedbacks and ideas provided. I also thank my undergraduate professors, who encouraged me at the beginning of this journey.

To all of those who heard me talk about this work countless times, my most sincere thanks. Without you, this work would not have finished.

“If I have seen far, it is because I have stood on the shoulders of giants.”
(Isaac Newton)

Resumo

Compreender os motivos pelos quais um modelo de aprendizado de máquina treinado toma algumas decisões automaticamente é fundamental para confiar no modelo e aplicar suas recomendações em aplicativos do mundo real com confiança. Também é fundamental ser capaz de melhorar o modelo se a eficácia (e.g. acurácia) não estiver dentro dos níveis desejados. Neste trabalho, apresentamos o **Explain-ML**, uma ferramenta interativa que implementa uma abordagem visual para apoiar a interpretação de modelos de ML. O desenvolvimento do Explain-ML seguiu uma estratégia de Aprendizado de Máquina Centrado no Usuário, guiada pelas demandas dos usuários (conhecedores), resultando em uma abordagem *holística* na qual a interpretabilidade é suportada por um conjunto de visualizações complementares sob várias perspectivas (e.g. global e local). Para avaliar a ferramenta e sua abordagem quanto à interpretação, foi realizada uma avaliação qualitativa com um grupo de usuários-alvo, focada em sua perspectiva em relação à utilidade e usabilidade do Explain-ML na compreensão dos resultados de modelos de ML treinados. A avaliação também explorou a capacidade dos usuários em aplicar o conhecimento obtido com as explicações da ferramenta na adaptação/melhoria dos modelos. Nossos resultados mostram que o Explain-ML fornece um relato amplo da execução do modelo (incluindo histórico), oferecendo aos usuários um espaço de exploração amplo e flexível com autonomia para tomar decisões diferentes e conduzir investigações e análises distintas. Os usuários declararam que a ferramenta foi muito útil e que realmente estariam interessados em utilizá-la em suas atividades diárias.

Palavras-chave: Interpretabilidade, Modelos de Aprendizagem de Máquina, Interação Humano-Computador, Visualização.

Abstract

Understanding the reasons why an automatically trained machine learning model takes some decisions is paramount to trust the model and apply its recommendations in real-world applications with confidence. It is also fundamental to be able to improve the model if the effectiveness (e.g. accuracy) is not at desired levels. In this work, we present **Explain-ML**, an interactive tool that implements a visual approach to support interpretation of ML models. The development of Explain-ML followed a Human-Centered Machine Learning strategy guided by the target (knowledgeable) users' demands, resulting in a *holistic* approach in which interpretability is supported by a set of complementary visualizations under several perspectives (e.g., global and local). To evaluate the tool and its approach to interpretation, a qualitative evaluation was performed with a group of target users, focused on their perspective regarding the helpfulness and usefulness of Explain-ML in comprehending the outcomes of the application of trained ML models. The evaluation also explored users' capability in applying the knowledge obtained from the tool's explanations in adapting/improving the current models. Our results show that Explain-ML provides a broad account of the model's execution (including historical), offering users an ample and flexible exploration space with autonomy to make different decisions and conduct distinct investigations and analyses. Users stated the tool was very useful and that they would indeed be interested in using it in their daily activities.

Keywords: Interpretability, Machine Learning Models, Human-Computer Interaction, Visualization.

List of Figures

2.1	A screenshot of a LIME visualization	26
2.2	A screenshot of a plot from SHAP	27
3.1	Persona: Knowledgeable user - Luiz	34
3.2	Explain-ML Prototype: Training	36
3.3	Explain-ML Prototype: Results	37
4.1	Explain-ML workflow	41
4.2	Explain-ML Initial Frame	42
4.3	Explain-ML Execution List	43
4.4	Creation of a Model Execution - Step 1 and 2	43
4.5	Creation of a Model Execution - Step 3	44
4.6	Explain-ML Visualization: Evaluation Metrics	45
4.7	Explain-ML Visualization: Class Distribution	46
4.8	Explain-ML Visualization: Feature Importance	47
4.9	Explain-ML Visualization: Instance analysis	49
4.10	Explain-ML Visualization: Analysis by class	51
4.11	Explain-ML Visualization: Analysis by class	52
4.12	Explain-ML Project execution history: Metrics Variation	54
4.13	Explain-ML Project execution history: Variation by Class	55
C.1	Prototype: Login.	96
C.2	Prototype: Projects.	97
C.3	Prototype: Delete Project.	97
C.4	Prototype: Deleted Project.	98
C.5	Prototype: Save Project. <i>New Project</i>	98
C.6	Prototype: Project.	99
C.7	Prototype: Step 1.1. <i>New Execution</i>	99
C.8	Prototype: Step 1.2. <i>New Execution</i>	100
C.9	Prototype: Step 2. <i>New Execution</i>	100
C.10	Prototype: Step 3.1. <i>New Execution</i>	101
C.11	Prototype: Step 3.2. <i>New Execution</i>	102
C.12	Prototype: Metrics. <i>Results for the New Execution</i>	103
C.13	Prototype: Feature Importance. <i>Results for the New Execution</i>	104
C.14	Prototype: Analysis by Instance. <i>Results for the New Execution</i>	105

C.15 Prototype: Analysis by Class. <i>Results for the New Execution</i>	106
C.16 Prototype: Features. <i>Results for the New Execution</i>	107
C.17 Prototype: Forest Structure. <i>Results for the New Execution</i>	108
C.18 Prototype: Metric Variation. <i>Execution History</i>	109
C.19 Prototype: Instance Variation. <i>Execution History</i>	110
C.20 Prototype: Class Variation. <i>Execution History</i>	111

List of Tables

2.1	Summary of approaches for the three types of problems: <i>Outcome explanation approaches</i> (OE) , <i>Model inspection approaches</i> (MI) , <i>Approaches for ML model building</i> (MB)	25
4.1	Information of instances from Execution 3 (Figure 4.3), with all dataset features. (<i>Due to space constraints, only the main three 3 features are shown</i>). The predicted class is presented along with the probability of its classification.	49
4.2	Similar to Table 4.1, this table is relative to the execution after Execution 3, without features <i>computer</i> and <i>university</i>	50
5.1	Summary of Benefits	73
5.2	Summary of Suggestions	75
B.1	Interview Script (English)	88
B.2	Roteiro da Entrevista (Português)	92
D.1	Prototype Evaluation Script (English)	112
D.2	Roteiro da Avaliação de Protótipo (Português)	112
F.1	Evaluation Script (English)	116
F.2	Roteiro da Avaliação (Português)	120
G.1	Evaluation Results (Translation)	124

Contents

1	Introduction	15
1.1	Context	15
1.2	Objectives	16
1.3	Our Proposal	17
1.4	Results	18
1.5	Outline	19
2	Background and Related Work	20
2.1	Background	20
2.2	Related Work	21
2.2.1	Outcome explanation approaches	21
2.2.2	Model inspection approaches	22
2.2.3	Approaches for ML model building	22
2.2.4	Interactive and Human-Centered Machine Learning	23
2.2.5	Summary of Approaches for ML Interpretability	24
2.2.6	Related Explanation Techniques	25
2.2.6.1	LIME	25
2.2.6.2	SHAP	27
2.3	Final Remarks	28
3	Tool Requirements and Design	29
3.1	User Interview	29
3.2	Interview Results	30
3.3	Design Requirements	32
3.4	Requirements Validation	34
3.4.1	Scenarios	35
3.4.2	Prototype	36
3.4.3	Prototype Evaluation	37
4	Explain-ML Overview	40
4.1	The Workflow	40
4.2	Model Execution	42
4.3	Interactive Visualizations	44
4.3.1	Visualization: Evaluation Metrics	45

4.3.2	Visualization: Class Distribution	46
4.3.3	Visualization: Feature Importance	47
4.3.3.1	Visualization: Instance Analysis	48
4.3.4	Visualization: Analysis by Class	51
4.3.5	Project Execution History	53
4.3.6	Final remarks	53
5	Evaluation	56
5.1	Methodology	56
5.2	Conducting the Evaluation	59
5.3	Results	59
5.3.1	Initial model training	60
5.3.2	Analysis of model execution visualizations	60
5.3.3	Model Improvement	68
5.3.4	Analysis of history results	70
5.3.5	Overall comments	71
6	Final Remarks	77
6.1	Discussion	77
6.2	Conclusions and Future Work	78
	References	81
	Appendix A Interview Consent Form	85
A.1	English	85
A.2	Portuguese	86
	Appendix B Interview Script	88
B.1	English	88
B.2	Portuguese	92
	Appendix C Prototyping	96
	Appendix D Prototype Evaluation Script	112
D.1	English	112
D.2	Portuguese	112
	Appendix E Evaluation Scenario	113
E.1	English	113
E.2	Portuguese	114
	Appendix F Evaluation Script	116

F.1	English	116
F.2	Portuguese	120
Appendix G Evaluation Results		124

Chapter 1

Introduction

In this chapter, we discuss the main motivations and arguments that support this dissertation. We also briefly describe our work and explicitly state our contributions.

1.1 Context

Machine learning (ML) models have been widely used nowadays, though mostly as “magic black boxes”. They are applied in many different domains, with multiple and distinct objectives (predictive, analytical, etc) – but the way they generate their results is far from being well understood, specially by the average or final user [35]. This understanding directly impacts the reliability of models and their predictions, particularly when they are used for making decisions that impact human life, such as medical diagnostics [6], or crime/terrorism prevention. In general, models are evaluated through a single performance metric (e.g., accuracy) on standard benchmarks. But real-world data may show variations or have particular and very specific idiosyncrasies that affect model behavior and performance [36].

In this context, providing explanations for model predictions favors users’ interpretability and acceptance [42]. Interpretable explanations can be used to understand how the model itself makes decisions to generate its results. The lack of these explanations is related to practical and ethical issues [19]. Increasingly, users and businesses that employ ML models are expressing the need to understand how those models generate their predictions. If a user does not trust a model or its predictions, it is likely that she will have caveats in using them. Moreover, the need for model interpretability meets the demand for transparency advocated by various international and national government institutions. The European Parliament, for example, has adopted the General Data Protection Regulation (GDPR) which became law in May 2018 [17]. It defends the right of users to have access to an explanation of the logic involved in automated decision-making systems, many of which use ML techniques [19]. The explanation behind a prediction can

also help to refine the model in order to improve it because it can be used as feedback to the model itself regarding why it hit or missed a result. *This last issue in particular is one of the main focus of our work.*

Accordingly, several efforts have been devoted to the problem of explaining ML-based decision models. Some works have focused on models that are more explainable or understandable to the user [6, 27, 19]. Other approaches focus on explaining ML models as black boxes in a model-agnostic way [36, 19, 37, 29]. All of these efforts are related to the need of making models more understandable and more reliable. Nevertheless, the *explicit needs* of users of ML systems, being them ML specialists or end-users that only consume the output of the system, concerning interpretability¹, and decision-making aids, have rarely been considered while designing the solutions. Consequently, there is a lack of works focused on the intersection between user demands and ML systems. This intersection has recently fostered the Human-Centered Machine Learning (HCML) area, which advocates that ML research and systems should be considered from a more human-centered perspective. HCML aims to be an interdisciplinary area that brings together the perspectives of HCI (*Human-Computer Interaction*) and ML (Machine Learning) [35, 15, 11].

1.2 Objectives

This work aims at investigating how to allow the interpretability of machine learning models by knowledgeable users². In order to answer that question, our study involved (1) Identify what knowledgeable users believe are the most important information about the model they use, in order to understand it; (2) develop a tool that presents this information and (3) evaluate how the developed tool supports knowledgeable users in their understanding of the machine learning model.

¹Here defined as the ability to explain or to present the results of ML models in understandable terms to a human.

²Users that have some knowledge on ML models, acquired by having employed ML models in the past in some application.

1.3 Our Proposal

In this context, the present work proposes a visual and interactive approach that assists knowledgeable users in the utilization and interpretation of ML models. Our focus is on helping them to better comprehend the obtained results with the goal of improving/adjusting the models³.

We present two main contributions in relation to other works available in the literature:

- The development of a *holistic* approach guided by typical ML users' demands, from a perspective focused on the *human-computer interaction* of these users with ML models. We say that our approach is holistic because, instead of relying in a single visualization, it encourages a set of complementary visualizations under several perspectives (e.g., global and local) in order to provide a more comprehensive view for interpreting results;
- The development of the *Explain-ML* tool based on the developed approach along with a *qualitative evaluation* of the tool through a study conducted with a group of target users.

We emphasize that this perspective focused on HCI has not been adequately explored in depth in the area of interpretability of ML models. *Our work helps closing this gap.* In more details, as mentioned, our approach aims to assist *users with some knowledge on ML models*, in the use and interpretation of these models. To develop it, a group of users working or familiar with ML has been interviewed. Based on their responses, a set of requirements was defined guided by indications, from their perspective, about what would be useful for an interactive and visual approach that supports manipulation and better understanding of ML models. As users mentioned different metrics that were useful, we considered a relevant requirement to present a holistic view of the problem by means of a set of complementary information they wanted to obtain (translated into visualizations) that would help them understand the obtained results and interact with the models. We also considered an explicit requirement that the tool should provide support without the need for coding or any other more complex type of interaction. Based on these initial requirements, we defined a prototype which was evaluated by the users to validate the approach and build the tool.

The consolidated approach was implemented in the **Explain-ML** tool. The tool supports a workflow where the user can interactively perform all stages of the life cycle

³Our approach can also be used by non-experts to help in understanding some results and outputs of the ML system, although we have not experimented with this possibility yet, leaving this avenue open for future work.

of an ML model: definition and optimization of hyperparameters, model training as well as testing, evaluation, adjustments for refining the model and model re-execution. In addition, the approach allows multiple model executions to be performed, stores the execution history and presents a comparative analytical graph between executions. For each execution, the approach presents a set of visualizations that conveys aspects related to the overall model, the training set, and instance-specific information. By design, these visualizations are complementary to each other and support users in interpreting results, potentially helping them to improve the model.

A qualitative study was conducted with a group of six target users using, as proof-of-concept, a version of Explain-ML instantiated with Random Forests (as ML model) applied to automatic text classification (as ML task). It is noteworthy that the proposed approach is agnostic to the ML model used and is also capable of incorporating model-based visualizations that may benefit explainability⁴ of the results. In our study, our goal was to investigate with users **their perspective** on how the proposed tool *(1) helps them to understand the outcomes of the ML model; (2) supports them in improving their model based on the acquired understanding.*

In order to answer these questions, our study involved users interacting with the tool with the goal to understand a model and improve it. After each step of the process and each visualization they were interviewed about their views on that specific step or view. During the interviews, they were asked about their perception and understanding of the visualization and the information it conveyed to them. During each session, the participants interaction and audio were recorded. The audio was later transcribed (and any references to elements in the interface included in the transcription), and analyzed.

1.4 Results

The results of our study indicate that users considered Explain-ML very useful, and that the explanation given through its visualizations were broad and complementary. In the process of understanding the model they were able to identify different aspects of the model that were problematic and suggest distinct strategies on how to improve them. They also found it easy to understand the effects of their changes in the model, analyze its reliability and make decisions on how to move forward in the modeling activity. Briefly, the main contributions of the work are:

1. *The design and development of an interactive and visual approach to support the*

⁴In this work we will use the terms interpretability and explainability interchangeably.

-
- use and interpretation of ML models, whose development was guided by user needs;*
2. *The Explain-ML tool that implements the proposed holistic approach that can be applied to most known ML models;*
 3. *A qualitative evaluation of Explain-ML with a group of target users. The results show that users found the tool helpful and useful in understanding the outcomes of the ML models and in using this information to decide strategies to be adopted to improve the model.*

1.5 Outline

The remainder of this work is described as follows: In Chapter 2, we review and highlight some relevant related work found in literature. Chapter 3 describes how the requirements gathering and design process were carried out. Chapter 4 introduces the Explain-ML tool. Chapter 5 presents details on the user study conducted to evaluate the proposed approach. Finally, Chapter 6 presents a discussion on the results, our final considerations and future works.

Chapter 2

Background and Related Work

Our work is related to visual and interactive approaches that assist users in the utilization and interpretation of ML models. The approach developed in this work was implemented in the **Explain-ML** tool. In this Chapter, we will first contextualize our work in the large body of literature on the related topics, and then cover the work that is more directly related to the one presented in this dissertation. Notice that the topics that are directly associated with our work are shown in **bold** in our general overview of the literature that follows.

2.1 Background

An in-depth study on interpretability approaches of ML models proposed in the literature [19] discusses various concepts inherent to the context of interpretability, and propose formalisms and a taxonomy for existing approaches. Other authors also address ML interpretability and propose similar or complementary categorizations [32, 20].

In the context of interpretability of ML models, there are two main lines of work, concerned with: (i) Transparent Box Design and (ii) **Black Box Explanation** [19]. The first one focuses on the problem of developing interpretable and transparent models along with their predictions. The second line of work focuses on explaining how the model generates its results, considering it as a black box. The latter can be further broken down into three other lines of work: (ii.1) *Model Explanation*: aims at generating a global explanation of the black box model; (ii.2) **Outcome explanation**: based on an input instance and a model, aims at providing an explanation for the model outcome on specific instances; (ii.3) **Model inspection**: aims at providing visual or textual representations to help in the understanding of the black box model and its predictions.

Our approach focuses on the Black Box Explanation, more specifically, on **Outcome explanation** and **Model Inspection**. Specifically concerning Model Inspection, Tamagnini and Krause (2017) highlighted the differences among approaches, considering

the following activities [41]:

- Inspection of models that perform multi-classification;
- *Prediction inspection of a given instance within its space;*
- *Approaches that allow the active participation of the user who interacts with some visual interface and provides feedback to adjust the model;*
- *The use of the importance of features as an inspection tool.*

The last three aforementioned activities are covered in the approach proposed by this work. In addition, interpretability can be defined as the ability to explain or present the result through user-understandable elements such as terms, image fragments, graphics and views [10, 19, 21] . Our approach also addresses several of these elements.

2.2 Related Work

In this section we present some closely related works on topics associated with this work. We describe three lines of work for Black Box Explanation (Outcome explanation, model inspection and ML model building). We also summarize the approaches for ML Interpretability, for comparison purposes.

2.2.1 Outcome explanation approaches

This kind of approach focuses on providing an explanation for the Black Box outcome of a single instance. The overall and internal logic of the ML Model is not the concern in this case. Several approaches explain specific types of Black Box models such as Deep Neural Network, SVM, Non-Linear Models and Tree Ensemble. Others, are model , which means that they should treat the original model as a Black-Box, and hence be able to explain any model [36, 12, 18, 19, 37, 29, 43, 39, 44]. The approach proposed in this work is also agnostic in relation to the ML model. However, different from the others, it does not require the user to perform source code manipulations. The manipulation and interpretation of the ML models are all performed through the interaction with Explain-ML visualizations and model/executions tuning controls.

2.2.2 Model inspection approaches

Still in the context of the Black Box explanation, some approaches are dedicated to providing representations to understand the behavior of ML models. The representations aim to help users in understanding why the model “prefers” a certain outcome (e.g., class) in detriment to others. Some available approaches to Model Inspection are focused on specific models. Others, are model-agnostic [8, 23, 16, 26, 24, 19, 1, 2]. A study was performed on the effect of aggregating explanations for interpreting ML models [25]. An interface provides aggregated instance-level explanations, generated through LIME [36], grouped by subsets based on confusion matrix. The approach aims at helping users to detect biases in the data. The approach is evaluated on an available tabular database, with low dimensionality. Our approach supports text bases, with high dimensionality and provides different (complementary) visualizations, besides offering interactive support to build and adjust models based on the understanding of the explanations.

2.2.3 Approaches for ML model building

In most cases, building ML models is a task assigned to specialized users (e.g., data scientists) who are familiar with Machine Learning. This task usually requires some amount of coding. A visual approach was proposed [22] to assist users in inspecting and interacting with co-training, a semi supervised learning method. This approach, however, is not model-agnostic, focusing on co-training with linear SVM models. The approach was evaluated on an automatic text classification task, with participants in an online recruitment system. *Clustrophile*, [9], an interactive tool, is presented with visualizations to help analysts in running clustering methods (not an agnostic approach). No formal evaluation was performed with users. The tool prototype was used in the healthcare domain over several months and the author just briefly discussed informal feedbacks received during that period.

In [40], the authors discuss an interactive open source visualization tool - *TensorFlow Playground*, specific for neural networks, that allows users to directly manipulate the model with no coding on four datasets. The tool only mentions general characteristics of the dataset, related to the distribution of data and not to the nature of the data. No formal user evaluation is performed. The authors of [45] propose an interactive visualization tool - *ATMSeer* to assist users to perform automated machine learning methods, including tasks such as selecting machine learning algorithms and tuning hyperparam-

ters. *ATMSeer* was evaluated through a case study with experts and end users. Although the tool presents visualizations to analyze the search models and refine the search space, it was not designed having interpretability of the generated models as a main goal, but rather it focused on the selection of the best models.

2.2.4 Interactive and Human-Centered Machine Learning

ML models have been extensively utilized in different domains, however their use remains more restricted to users familiar with such context or experts. There is a gap of approaches that provide functionality without explicit programming. Recently, interest in ML has grown considerably and thus, the demand for systems that allow non-expert users to better apply and understand these techniques. *Interactive Machine Learning* (IML)[13] is directly related to human-computer interaction as it puts human interactions in perspective. It treats the model training process as an HCI task, receiving users input in the example selection, creation and labeling process. A user more familiar with ML models may be required for model's deployment, but is not essential in the training process. IML differs from classical ML in that it considers user participation through an iterative process of changing and revising the model during its training, and these modifications may be small. In more traditional ML, the user usually performs a wholesale pre-selection of training data and significant changes every model execution [11, 13].

And, as mentioned before, The *Human-Centered Machine Learning* (HCML) research area has recently emerged based on an identified lack of consideration (and therefore works) of the explicit needs of users concerning interpretability and decision-making aids, from a more human-centered perspective [35, 15, 14]. The approach proposed in this dissertation is in line with the HCML and IML areas. The approach was developed and evaluated considering human centered perspective, and offers functionalities that support knowledgeable users in tasks under the IML paradigm.

Furthermore, the explicit needs of users of ML systems, being them ML experts or end-users, concerning interpretability, and decision-making aids, have rarely been considered while designing the solutions. Consequently, there is a lack of work focused on the intersection between user demands and ML systems. The *Human-Centered Machine Learning* (HCML) area advocates that ML research and systems should be considered from a more human-centered perspective. HCML aims to be an interdisciplinary area that brings together the perspectives of HCI *Human-Computer Interaction* (HCI) and *Machine Learning* [35, 15, 14]

2.2.5 Summary of Approaches for ML Interpretability

Table 2.1 presents a summary of approaches for the three types of aforementioned problems. It includes the one proposed in this dissertation – Explain-ML, in order to compare it with the others. For comparison purposes, the following dimensions were used in the Table 2.1. Some of them were based on the dimensions used by Guidotti et al. (2018):

- **Name:** model name given by authors;
- **Reference:** reference number as listed in the References section;
- **Year:** publication year (of first paper);
- **Model Agnostic:** the proposed approach is model agnostic, e.g., works with any ML model;
- **Dataset:** tab- *Tabular*, txt- *Text*, any - *Any type of data*;
- **Problem:** type of problem addressed by the approach. OE- *Outcome Explanation Problem*, MI-**Model Inspection Problem**, MB-*Model Building*;
- **Interactive:** the approach is interactive;
- **Views:** the approach provides visualizations;
- **MI - without Programming:** the approach does not requires explicit programming for model interpretation;
- **MB - without Programming:** the approach does not requires explicit programming for model building and/or control;
- **Views perspective:** the approach presents visualizations that convey aspects related to the overall model (G-*Global*), instance-specific information (L-*Local*) and the training set (D-*Dataset*);
- **HC Development:** the development approach was accomplished with the explicit participation of users;
- **HC Evaluation:** a formal evaluation or case study was performed with users on the approach.

Table 2.1: Summary of approaches for the three types of problems: *Outcome explanation approaches* (OE) , *Model inspection approaches* (MI) , *Approaches for ML model building* (MB)

Name	Ref	Year	Model Agnostic	Dataset	Problem	Interactive	Views	MI without Programming	MB without Programming	Views perspective	HC Development	HC Evaluation
VIN	[23]	2004	x	tab	MI		x			g		
VEC	[8]	2011	x	tab	MI		x			l,g		
-	[12]	2010	x	tab	OE		x			l		
ICE	[16]	2015	x	tab	MI		x			l,g		
MFI	[44]	2016	x	tab	OE		x			l,g		
MES	[43]	2016	x	any	OE		x			l		
LIME	[36]	2016	x	any	OE		x			l		x
-	[39]	2016	x	tab	OE					-		
Prospector	[26, 24]	2016	x	tab	MI	x	x			l,g		x
OPIA	[1]	2016	x	tab	MI					-		
-	[22]	2016		txt	MB	x	x		x	l		x
Clustrophile	[9]	2016		tab	MB	x	x		x	l,g		
TensorFlow Playground	[40]	2016		-	MB	x	x		x	g		
SHAP	[29]	2017	x	any	OE		x			l		x
Anchors	[37]	2018	x	any	OE		x			l		x
LORE	[18]	2018	x	tab	OE		x			l		
BlackBoxAuditing	[2]	2018	x	tab	MI		x			g		
-	[25]	2018		tab	MI	x	x			l,g		x
ATMSeer	[45]	2019		tab	MB	x	x		x	g	x	x
Explain-ML		2019	x	txt	MI, OE, MB	x	x	x	x	l,g,d	x	x

2.2.6 Related Explanation Techniques

From the analyzed Explanation Techniques, next we present some of the most used ones, related to our approach.

2.2.6.1 LIME

Local interpretable model-agnostic explanations (LIME)[36] is a explanation technique in which the authors propose a concrete implementation of local models that are trained to approximate the predictions of the underlying black box model [31]. LIME is an outcome explanation approach that focuses on training local models to explain individual predictions of any classifier in an interpretable and faithful manner.

Figure 2.2 shows a screenshot of a LIME explanation, generated in HTML and it can be easily produced and embedded in ipython notebooks. On the screenshot, negative (blue) words indicate atheism, while positive (orange) words indicate christian. The way to interpret the weights is by applying them to the prediction probabilities. For example, if we remove the words Host and NNTP from the document, we expect the classifier to

predict atheism with probability $0.58 - 0.14 - 0.11 = 0.31$. Therefore, according to the classifier, the document will be less likely to belong to the atheism class. LIME also supports visualizations using matplotlib.

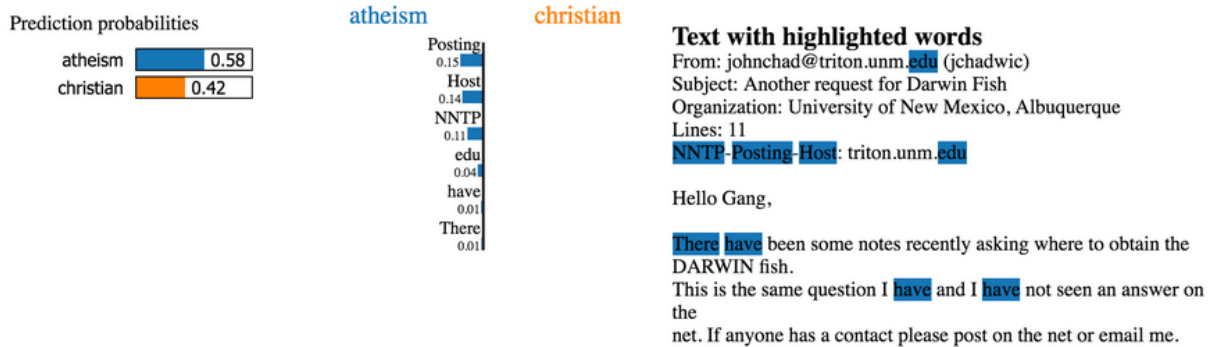


Figure 2.1: A screenshot¹ of a LIME visualization: Negative (blue) words indicate atheism, while positive (orange) words indicate christian.

LIME is able to explain any black box classifier, with two or more classes; works for tabular data, text and images. The fidelity measure (how well the interpretable model approximates the black box predictions) gives us a good idea of how reliable the interpretable model is in explaining the black box predictions in the neighborhood of the data instance of interest. LIME is implemented in Python and R and is very easy to use. However, there are some disadvantages:

- The correct definition of the neighborhood is a very big, unsolved problem when using LIME with tabular data;
- Data points are sampled from a Gaussian distribution, ignoring the correlation between features. This can lead to unlikely data points which can then be used to learn local explanation models;
- The complexity of the explanation model has to be defined in advance; and
- The explanations can be instable as two very close points can varied greatly in a simulated setting.

Our approach proposes to address some of these limitations, supporting text bases with high dimensionality and providing different (complementary) views, in addition to interactive support to build and adjust models based on understanding the explanations.

¹<https://github.com/marcotcr/lime>

2.2.6.2 SHAP

SHAP (SHapley Additive exPlanations) [29] is a method to explain individual predictions. The goal of SHAP is to explain the prediction of an instance x by computing the contribution of each feature to the prediction. The SHAP explanation method computes Shapley values from coalitional game theory. Shapley values tell us how to fairly distribute the “payout” (the prediction) among the features.

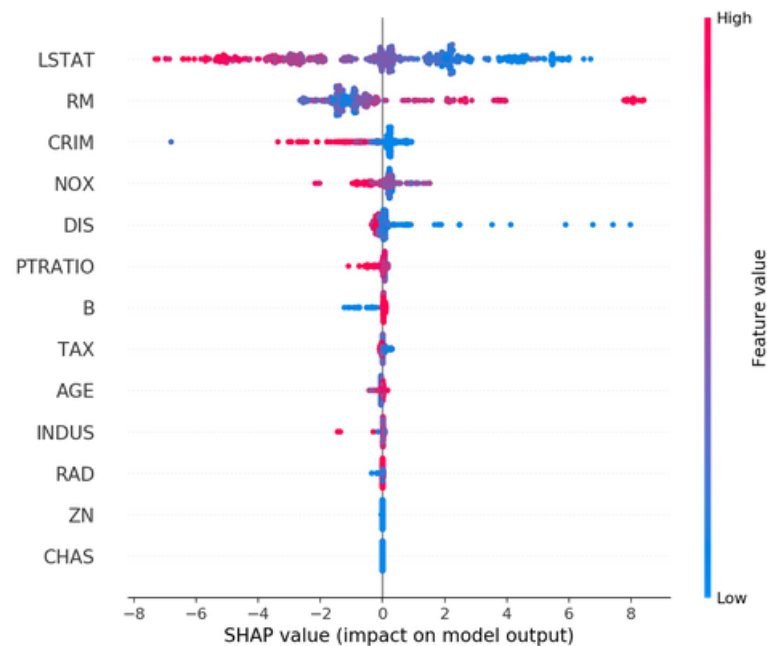


Figure 2.2: A screenshot² of a plot from SHAP: The plot sorts features by the sum of SHAP value magnitudes over all samples, and uses SHAP values to show the distribution of the impacts each feature has on the model output. The color represents the feature value (red high, blue low). This reveals for example that a high LSTAT (lower status of the population) lowers the predicted home price.

The advantages of using SHAP are:

- Contrastive explanations that compare the prediction with the average prediction.
- Has a fast implementation for tree-based models, making it possible to compute the many Shapley values needed for the global model interpretations.
- The global interpretations are consistent with the local explanations.

There are, however, some disadvantages. Calculating the Shapley Value is computationally expensive. For large background data, costs a lot of computation. The Shapley

²<https://github.com/slundberg/shap>

Value does not return a prediction model and calculating the the Shapley Value for a new data instance requires access to the real data, not simply the prediction function.

Our approach proposes to address some of these limitations, supporting a workflow where the user can interactively perform all stages of the life cycle of an ML model: Definition and optimization of hyperparameters, model training as well as testing, evaluation, adjustments for refining the model and model re-execution. In addition, the approach will allow multiple model executions to be performed.

2.3 Final Remarks

Summarizing, when compared to previous approaches, ours was developed by explicitly considering user requirements towards interpretability and ease of use. Users can perform tasks related to the entire life cycle of a machine learning model: model training, predictions, model evaluation, model adjustments and model re-execution. The implemented tool helps users to build models interactively, without the need of source code manipulations. Besides, the user may execute the model several times, making adjustments. For each execution, information on the overall model, the training dataset and instance-level information are stored for visualization purposes. The tool provides a history of the executions with comparative graphs on the evolution of the model in the different executions.

Finally, there is a common belief that to augment interpretability, ML models have to become simpler and thus less effective [19, 4]. We take a different perspective, in which we allow users to understand the model and improve it, potentially increasing its effectiveness [24]. To do so, our tool allows users to inspect the model: (i) as a black box (despite its complexity) by understanding its outcomes, its successes and failures; (ii) from a global (e.g., whole dataset) and/or local (e.g., instance) perspective; (iii) through different and complementary representations and visualizations.

In the following sections, we present the design process taken to develop Explain-ML, and also details regarding the implemented tool and its evaluation.

Chapter 3

Tool Requirements and Design

The design process adopted to develop Explain-ML was a user-centered design, that allowed us to take into account users' views and perspectives on which aspects would be relevant in a ML explainability tool. Thus, our design process included user interviews to better define our requirements, the definition of personas and scenarios to guide the design process, the development of a prototype to represent our solution based on the requirements, and an evaluation of this prototype, to better inform the development of Explain-ML.

3.1 User Interview

We conducted semi-structured interviews [28] with target users in order to better understand the process followed by them (e.g., pre-processing, parameter tuning, result analysis) while employing ML models in their applications. As a result, we identified some important requirements for understanding and using Machine Learning results with regards to the explainability of the models.

In preparation for the interviews, we conducted a pilot study to analyze our topic guide, and small adjustments were made. The interviews were conducted between July 21 and September 12, 2018, via videoconference and included 7 participants. The participants were all male and Computer Science graduate students (taking their Masters or PhD). All of them had experience with Machine Learning and use ML models on a daily basis. Recruitment was based on email and social network interactions targeted to Machine Learning researchers in our department. We recruited, from the target group, people who responded motivated to participate in the interviews.

Prior to the interviews, all participants were sent the Consent Form explaining the study, and were free to ask any questions about the study. Appendix A contains both the original version of the form (in Portuguese) and the translated version (in English). The form contained information on the research and researchers, clarified that the participa-

tion in the study was voluntary, that any doubts could be asked, the risk and benefits involved, that the information was confidential and all data would be anonymized, and that the audio for the interview would be recorded. In the day of the interview, prior to the questions, the participants were asked if they had any questions or doubts about the consent form or the research itself they would like to clarify, and if they agreed to participate, to formally state their agreement.

The interviews were organized into six main themes: (i) demographics of participants; (ii) contextualization regarding purpose of the interview and the scope of this work; (iii) the participant’s experience with Random Forests. The interview explored the ML experience in general, and then focused on their experience and analysis with Random Forests, as one of our goals was to test an instantiation of the tool with a specific classifier; (iv) the process participants adopted to analyze RF models and their results; (v) participants views on a specific interpretability tool - LIME [36]; (vi) any other comments participants thought were worth mentioning about the topics covered in the interview. Appendix B contains both the original version of the form (in Portuguese) and the translated version (in English) of the Interview Script, with was used to guide the interviews. The interviews were transcribed and a thematic analysis conducted [28]. Next, we present the main results of our interview that were the basis for the requirements defined for the Explain-ML tool.

3.2 Interview Results

In our analysis of the interviews, our goal was to understand participants strategy to creating ML models, and which would be aspects that a explicability tool could or should support them in. Thus, to define the requirements for our explicability tool, we looked for strategies that were commonly adopted by the participants, as well as differences in how they worked with their models.

Based on the interviews, we identified five factors (F1 – F5) relevant to participants’ experiences with ML models and that were used to define our requirements: (i) integration with familiar tools (e.g., “Tools already used by participants”); (ii) strategies for measuring reliability of results (e.g., “which strategies users adopt to measure reliability of results?”); (iii) use of reference values for hyperparameters (e.g., "Does the participant use reference values to check the quality of the trained model? How are these values extracted?"); (iv) parameterization strategies (e.g. "How do participants define hyperparameters for model training?"); and (v) experience with available explicability resources (e.g. "What are the strengths and weaknesses of the explicability resources

already available").

F1 – Integration with used tools. Participants mentioned the different tools they used. The only one used by all of them was the ScikitLearn toolkit. Participants stated that ScikitLearn allows easy access to the algorithms (and code), contemplates recent advances in the area and enables greater support due to the large user community. Some suggested that an integration of the the proposed explainability tool with ScikitLearn would be a plus.

F2 – Strategies for measuring reliability of results. Among the main evaluation metrics used to assess the performance of the models, the most recurrent ones were accuracy, recall and F1 (micro and macro). Four respondents said that they use cross validation in order to verify the variability of the results and assess their confidence level on the trained model. The confusion matrix was also mentioned by one participant as an important evaluation resource. Statistical analyses of the results (e.g. statistical significance tests) were considered as a very valuable asset to assess significant improvements in the results. Other reported resources for measuring confidence were: (i) previous knowledge about what the model was expected to learn; (ii) ranking of feature importance (based on some measure of discriminative power); and (iii) comparisons with baselines.

F3 – Use of reference values for the hyperparameters. The use of reference values for the hyperparameters was considered as a complicated matter by the respondents, since pre-processing, folding, parameterization and other data treatment procedures depend on the dataset. As a potential solution, it was reported by some participants that the Database Lab (LBD) at the Federal University of Minas Gerais (UFMG) was developing benchmarks (data and pre-preprocessing code) for textual collections that could provide all the necessary steps taken to obtain the final results. Participants thought it would be interesting if the tool could be integrated with these and other benchmarks (again, data and code) and if the data entry steps of the models (in the tool) could use them in the interest of easiness of use, replicability and comparability of results.

F4 – Parametrization strategies. As experimental procedure strategy, folded cross validation¹ was widely used by the respondents. They usually perform experiments changing the hyperparameter values and many of them use ScikitLearn to find the best hyperparameters for the dataset, based on a suitable range according to references (grid search and random search). Thus, the new tool should allow folded division and cross validation (including nested) for hyperparameter tuning.

¹Alternatively, nested cross-validation in the training was also used.

F5 – Experience with Available Explicability Resources. Most of the participants had never used LIME, only two mentioned having some knowledge about it. During the interview an example generated in LIME depicting the visualizations aimed at explaining the model was shown and discussed with the participants of the interview. Based on this limited perspective of LIME, participants identified aspects they considered could help them in understanding a ML model, and others that could be improved. Overall, the participants thought that LIME could help them understand the model as a black box, providing insight regarding the confidence of the classification. They commented that it helped with word visualization, which is useful as it allowed them to identify important terms that influenced predictions. But they also mentioned that this can be a challenge when there is a large number of terms being considered. They believed that the visualization could help them identify noise in the dataset and what was wrong with the trained model. Thus, at some extent they thought LIME would support them in identifying the need for new data processing and feature engineering, as well as deciding whether the model application is reliable. Although, participants thought LIME could be potentially useful, they also pointed aspects they felt could be improved. They commented that they missed an overall view of the performance of the model, and that the results were not contextualized. They did not think that the explanations were intuitive enough, nor that all relevant metrics were presented. Finally, they commented on not being able to compare the results for the instances processed with the model, which would be useful in assessing the model’s confidence.

Next we present the requirements defined based on the interviews and described in (F1 – F5).

3.3 Design Requirements

Regarding the planned functionalities, proposed for explaining and interacting with the ML models, the participants considered as the most relevant (in no particular order): (i) comparative display of several evaluation metrics; (ii) support for folded division and cross-validation parameterization; (iii) overview of best parameters found for a dataset; (iv) visual comparison between runs; and (v) display of the importance of the variables, both globally (whole model) and locally (instance-based).

With the compiled results, it became clear that a ML tool focused on interpretability would have a potential to be largely used by researchers and ML practitioners. Sum-

marizing , the following requirements were extracted from the interviews for developing our tool:

- **R1 – ScikitLearn integration:** As all participants use ScikitLearn the ideal would be that our tool had the same functionalities and steps in ML modeling or used it as basis.
- **R2 – Support for cross validation:** All participants used cross validation to train the model, so it would be important to support this technique, as well as let users decide on the number of folds to be used.
- **R3 – Support for hyperparameter tuning:** Integrate hyperparameter tuning as part of the process supported by the tool and provide an explanation of each available hyperparameter.
- **R4 – Display of a set of evaluation metrics (including confusion matrix):** Need of different metrics to be able to perform a more accurate analysis of the execution.
- **R5 – Model-based feature importance:** Its important to exhibit the feature importance of the model as it helps the user understand the model’s predictions.
- **R6 – Instance-based feature impact:** It is important to show the instance-based feature impact as it supports user’s understanding of the model’s prediction per instance.
- **R7 – Visualization of the execution history:** regarding the control of routine tasks in ML, participants mentioned the difficulty in assessing model reliability due to the challenge to compare different results. Thus, it pointed to the need to include a mechanism for the visualization of the history of experiments and monitoring of the evolution of the models based on performed changes (e.g., parameterization, feature selection, etc).

In the end, all participants mentioned that they would use a tool that had such functionalities in their daily activities, confirming a real need for this kind of interpretability framework.

3.4 Requirements Validation

Based on the requirements, we initiated the design and diagramming process, in an attempt to represent the functionalities of the proposed tool and allow its evaluation by potential users. Firstly, we used the persona-scenario method [30] in order to generate shared understandings and design ideas. Next, we developed a prototype to present our proposed visualizations and, finally, we conducted a formative evaluation of the prototype with target users to validate our requirements and proposed design.

A persona is a description of a fictitious user, based on data from user research, capturing in detail the user of the system to be designed for which designers will guide the design process. We described the persona of a knowledgeable user representing the target users of our proposed tool. The persona was inspired by participants in the interview stage, according to the scope of the tool to be developed.



Figure 3.1: Persona: Knowledgeable user - Luiz

Luiz is 27 years old and works as a machine learning researcher at the Federal University of Minas Gerais, also pursuing a master's degree in Computer Science at the same institution. He has been working on this for 4 years and knows his job well. Many students in the field ask Luiz for advice when they find it difficult to use ML algorithms or interpret its results, as he has experience in dealing with recurring problems, in addition to a good theoretical knowledge. Luiz likes to share his knowledge with others, in order to deal with new problems and apply his knowledge in different ways. He works at the lab during the day, but sometimes stays late trying to finish some of the more complex tasks. Often Luiz has to consult the litera-

ture or the Prof. Paul, who coordinates the research, to find the best solutions. Luiz has worked with several Machine Learning algorithms for text classification, applying them in several different contexts. Despite being able to use the ML algorithm correctly with relevant results, Luiz considers challenging the task of interpreting the results as the model is a black box that is difficult to interpret. On several occasions, Luiz finds it difficult to gauge the reliability, since it takes a long time, several executions and multiple comparisons carried out in a non-automated way.

3.4.1 Scenarios

A scenario "is an informal narrative description of human activities or tasks in a story that allows exploration and discussion of contexts, needs and requirements" [33]. We described three scenarios (Model Results Interpretation Scenario, Hyperparameters Setting Scenario and History Tracking Scenario), based on the needs that were discovered during the interviews. The scenario can be used in multiple steps in the design process [38]. In this work, we used the scenarios to describe how we intended the system would be used by users.

Scenario 1: Model Results. Luiz uses machine learning models daily for text classification, managing to obtain relevant results through them. He considers the task of interpreting the machine learning results somewhat challenging. Since ML models can be difficult to interpret, he feels that the process of assessing the reliability of his results and/or classification models are complex time-consuming. Recently, Luiz started became familiar with the concept of explainability in Machine Learning. He wonders if he can use this feature to better understand the behavior of his models in text classification and if this use will assist him in the task of gauging reliability. Luiz intends to use a explainability model to assess its relevance and verify whether he will be able to understand how the model made a certain prediction. With that, Luiz intends to determine if he can trust his machine learning model and use its results in contexts in which critical decision-making is necessary and therefore greater confidence in the results are needed.

Scenario 2: Hyperparameters Setting. Luiz uses Machine Learning models daily for text classification, managing to obtain relevant results through them. He considers the hyperparameterization task somewhat challenging. At each execution, Luiz intends to treat his database and changes the hyperparameters used in order to obtain better prediction results, but finds it difficult to determine the best values for the hyperparameters, since they change depending on the input database. He seeks a tool to assist him in the process of determining the best hyperparameters for each model training.

Scenario 3: History Tracking. Luiz uses ML models daily in several text classification problems, executing the training of his models several times, until obtaining relevant and satisfactory results for the proposed problems. In each execution, Luiz seeks to treat his database and change the hyperparameters used to obtain better prediction results, but finds it difficult to keep up with the changes he makes and their effects on the model, since numerous executions are carried out before a prediction model is 'finished'. Luiz is looking for a tool to help him in this process. He hopes to be able to follow the changes made in each execution for a given dataset, as well as visualize the changes caused in the results. It is relevant for him to visualize the progression for each evaluation metric throughout the executions, as well as the configuration of hyperparameters used

in each one. With this, Luiz intends to obtain assistance to more easily conclude the determination of which treatments are necessary to improve the predictions and the ML model.

3.4.2 Prototype

Based on the requirements and personas-scenarios described, a low fidelity prototype was developed to present our proposed visualizations to the target users and perform a formative evaluation of the tool [34]. The prototype (Appendix C) was developed using Balsamiq Wireframing Tool² and presented the structure and content of the interface, allowing users to interact with it. Users could navigate through the interfaces and visualizations (as shown in figures 3.2 and 3.3), which simulated an execution – the functionality was not yet implemented, but the prototype showed the data for one example as if it were being executed for that one model.

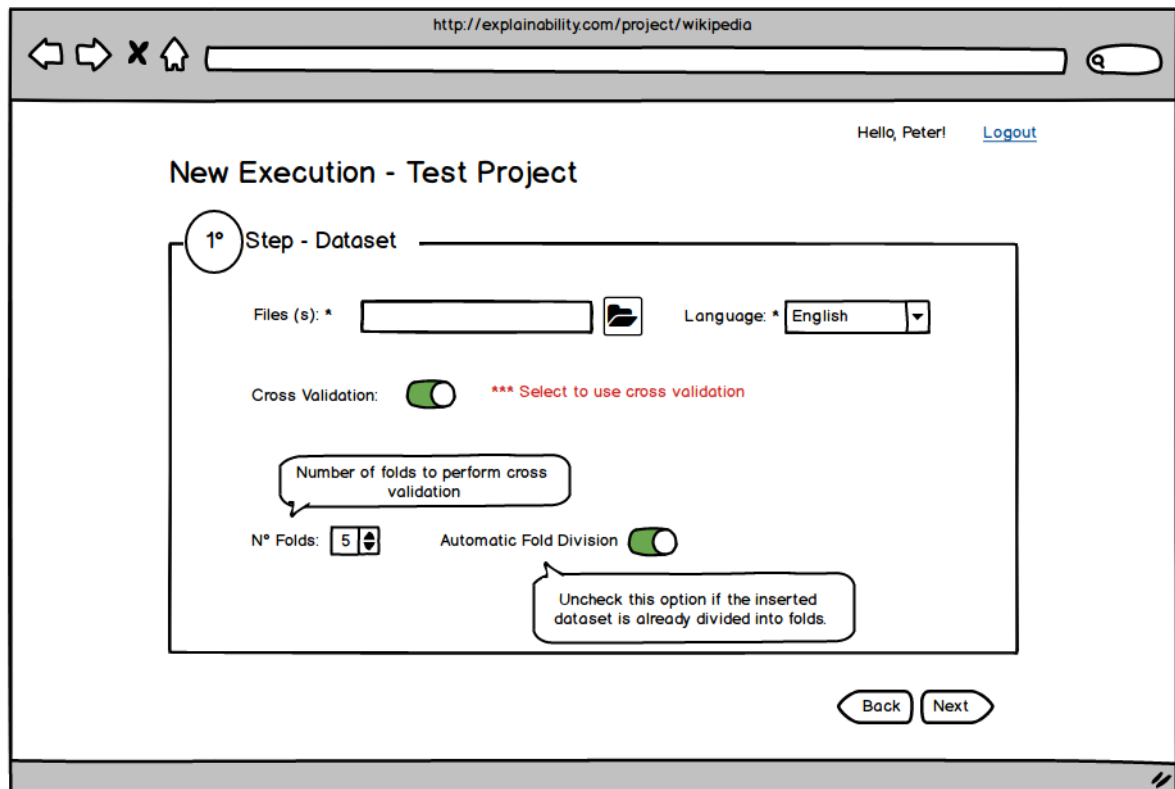


Figure 3.2: Prototype: Begins the process of training a new model. *Step 1 - Dataset:* The user fills in the data and selects the file containing the dataset, dataformat, number of folds and whether to perform cross validation

²<https://balsamiq.com/>

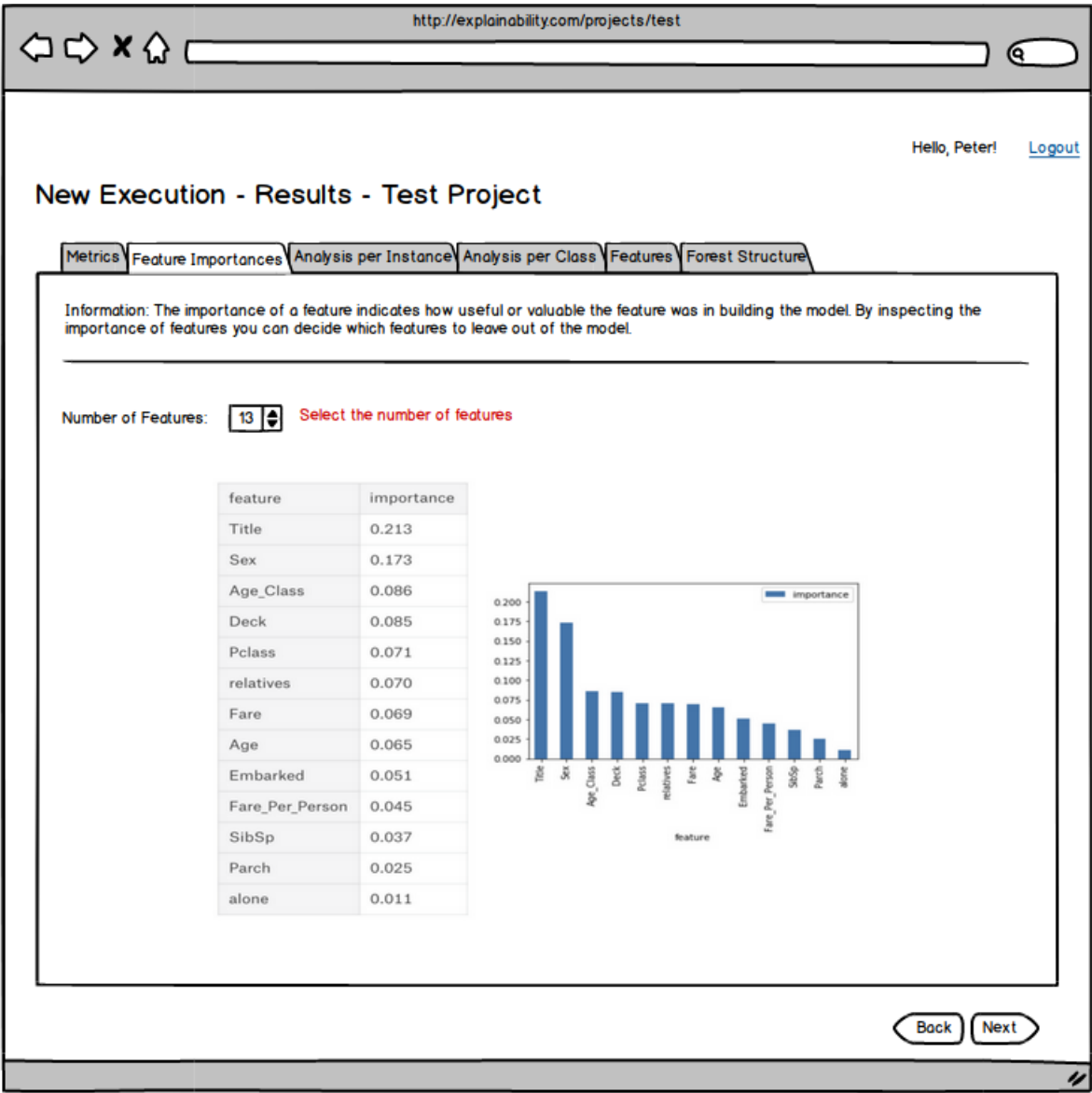


Figure 3.3: Prototype: Feature Importance Visualization. *Results for the New Execution*

3.4.3 Prototype Evaluation

In order to collect feedback on our proposed tool represented through the Balsamiq Prototype, we conducted a formative evaluation with 3 participants (2 males and 1 female). The participants were all Computer Science graduate students and expert users who use ML Models on a daily basis as part of their research. The evaluation was conducted between December 5 and 26, 2018, and two of them were face-to-face, whereas the other was via videoconference. Recruitment was based on email and social network interactions targeted to Machine Learning researchers in our department.

In preparation for the evaluations, we conducted a pilot study to analyze our topic

guide, and small adjustments were made to the script (see Appendix . D for the final version used). Prior to the evaluations, all participants were sent the Consent Form, and were free to ask any questions about the study. Appendix A contains both the original version of the form (in Portuguese) and the translated version (in English).

In the evaluation session, participants interacted with the prototype of Explain-ML tool (Appendix C) as if they were creating a model. To do so, they were asked to (i) Create a new project;(ii) Access the project created; (iii) Create a new execution to the project created; (iv) Analyze the results of the new execution; and (v) Analyze the project results history. The users were guided and observed during the completion of the predetermined tasks and were encouraged to think aloud [34], about their interactive experience, possible difficulties, and thoughts about the prototype.

The user’s interaction (video of user and prototype interaction) and audio of all sessions were recorded, transcribed and an analysis conducted. Participants liked the solution conveyed through the prototype and said they would use such a tool in their ML modeling, in alignment with participants in the interview. As a result of the evaluation, participants raised a few problems, improvements and suggestions related to the explainability model, as well as the interface. These results allowed us to review some of the initial ideas and improve them, in order to implement the actual tool. Next, we detailed the problems perceived by the participants during the prototype evaluation and changes to be made to solve them and/or improve the project:

- New Execution - Step 1.1 (Figure C.7)

Problem: Difficulty understanding the expected data input format.

Improvements:

- Explain the data input format in a simplified, clear and quickly accessible way;
- Standardize the dataset input format.

- New Execution - Step 2 (Figure C.9)

Improvements: We chose to remove this step, and raw data should not be allowed, in order to reduce processing and keep the focus on the model’s objective, explicability. The entry will be limited to pre-processed datasets, divided into training and test folds and with the frequencies already calculated.

- New Execution - Step 3 (Figure C.10)

Problem: Questions about how the level of randomness is calculated, how it is applied and what parameters can be changed with the variation of the level of randomness.

Improvements: This functionality will be better analyzed for the next version.

- Results 2 (Figure C.13)

Suggestions:

- Add navigation between features, when there are many.
 - Display the number of features with importance and also the number of features with zero importance.
- Results 3 (Figure [C.14](#))
Improvements:
 - Display the most important features to a classification.
 - Show a word cloud with the feature importances on the instance.
- Results 4 (Figure [C.15](#))
Improvements:
 - Add percentage of the total that the displayed amount represents on the confusion matrix.
 - Analyze a way of breaking data for a large number of classes.
 - Change the color scale of the confusion matrix.
- Results 6 (Figure [C.17](#))
Improvements: Representation not yet adequate. We have chosen to remove this view, to improve it for the next version.
- History 2 (Figure [C.19](#))
Improvements: Representation not yet adequate. We have chosen to remove this view, to improve it for the next version.
- History 3 (Figure [C.20](#))
Improvements: Show the class balance (the number of instances for each class).

Chapter 4

Explain-ML Overview

In this chapter we present in details the Explain-ML tool that implements the approach proposed in this work based on the requirements previously enumerated. Explain-ML is a web based application, developed using the Python Web framework *Django*¹, using *Sqlite*². The screenshots and visualizations presented in this section are related to a version of *Explain-ML*, instantiated with Random Forests (as ML model) applied to automatic text classification (as ML task). The instantiated version employs the *Scikit-Learn*³ implementation of the Random Forests Model. Note that, as pointed out, the proposed approach is agnostic to the used ML model while the tool is capable of incorporating model-based visualizations, if desired. Finally, the example shown depicts the model of the the *WebKB-Course dataset*⁴.

Following is a broad overview of the tool. Section 4.1 shows the main workflow for user-tool interaction, in addition to an interface overview and an explanation about how a model execution is conducted. Section 4.3 presents the current visualizations available in the tool. The project execution history is shown in Section 4.3.5, and Section 4.3.6 presents some final remarks.

4.1 The Workflow

The tool is organized to support a workflow in which the user can interactively perform the stages of the life cycle of a ML model: definition and optimization of hyperparameters, model training as well as testing, evaluation, adjustments for refining the model and model re-execution. In particular, interpretability of models provides the core support to the model evaluation activity.

Explain-ML allows users to have their own access area, which is organized according

¹<https://www.djangoproject.com/>

²<https://www.sqlite.org/index.html>

³<https://scikit-learn.org/stable/>

⁴<http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/>

to two important concepts within the tool:

- *Project*: consists of a set of one or more model executions, performed by a user, with the project's executions history being recorded.
- *Execution*: covers the process of defining and optimizing hyperparameters, model training and testing. The information regarding each execution is recorded by the tool and the corresponding visualizations are made available.

Figure 4.1 presents the workflow to create a project and *one* execution. To perform an execution, the user first creates it and sets values for the model's hyperparameters (or selects the hyperparameters to be tuned with cross-validation along with the respective parameters' options/ranges). After the model is trained, the user applies it to a test set and accesses the interactive visualizations regarding the current execution. If desired, she can also review the history of previous project executions. *In order to create a new execution, user should repeat the activities related to: Execution creation, Execution analysis, Project analysis and Model adjustments* (Figure 4.1: ii, iii, iv, v).

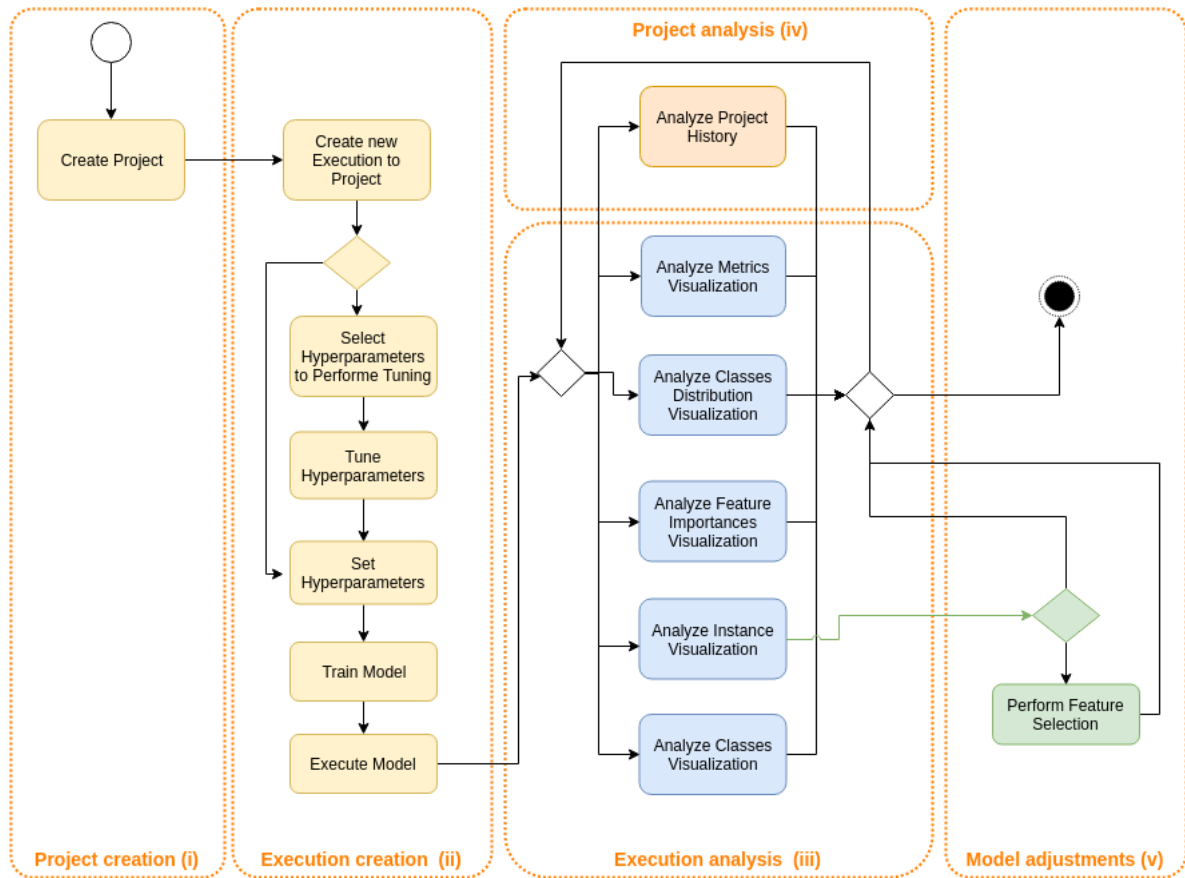


Figure 4.1: Explain-ML workflow for the creation of a project and one execution.

The initial interaction with *Explain-ML* is accomplished by means of the main screen, where users can manage their projects (Figure 4.2). A new execution for an existing project, can be created through the main screen (Figure 4.2 - (3)), or visualizing

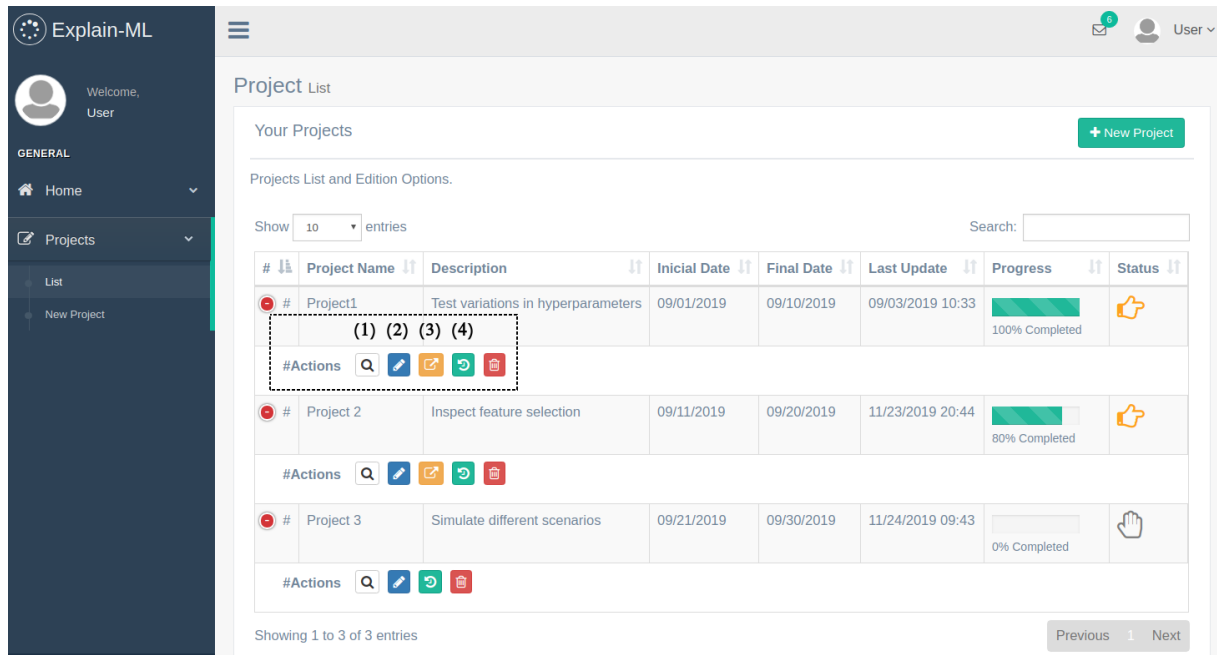


Figure 4.2: A screenshot of *Explain-ML*: User can list previous projects or create a new one in the *left side*. In the *right side*, the projects are listed along with their information such as description, period, status and progress, and available functionalities: (1) visualize the project details, (2) edit project, (3) create a new execution, and (4) visualize history of project executions. A new execution can only be created on *In Progress* projects.

the project details (Figure 4.2 - (1)). In the latter, the user is directed to the project information where the list of executions can be visualized (Figure 4.3), along with other project information.

4.2 Model Execution

Creating an execution comprises three steps.

- *Step1*: The user fills in the data and selects the file containing the dataset, data format, number of folds and whether to perform cross validation (Figure 4.4).
- *Step2*: Subsequent to step 1, the user configures the execution by selecting whether to tune the hyperparameters and, if so, selects which hyperparameters to perform, along with the range/options (Figure 4.4).
- *Step3*: The tool displays the hyperparameters (obtained through tuning or user-defined) and allows the user to change or not these values. From there, the user

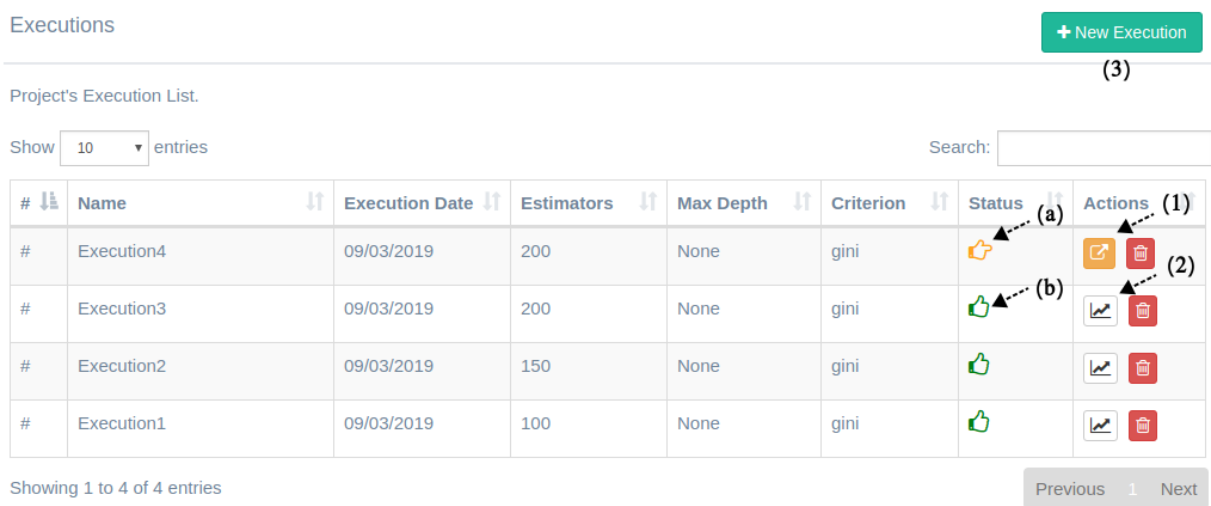


Figure 4.3: Project executions are listed along with their information such as name, execution date, hyperparameter values, and execution status. (1) *Continue Execution* - for executions where the *Status* (a) indicates that hyperparameter tuning has been performed and the model has not yet been applied. (2) *See explanation* - for executions where *Status* (b) indicates that hyperparameter tuning has been performed and the model has been executed.

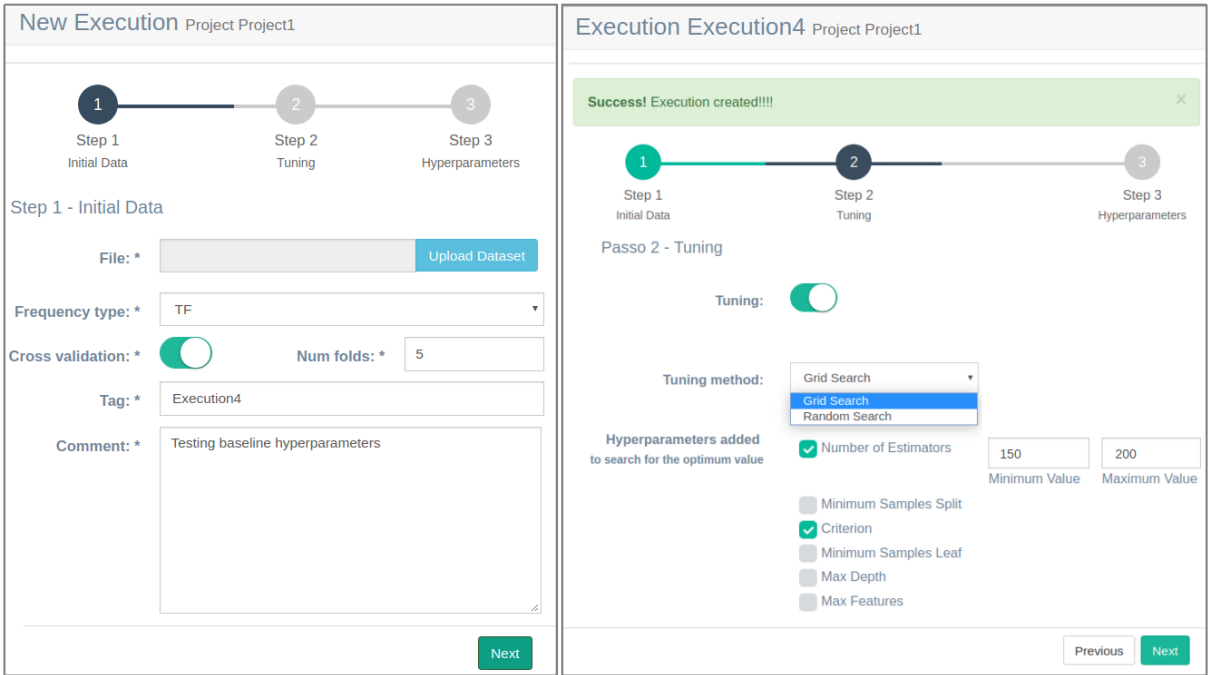


Figure 4.4: Creation of a Model Execution - Step 1 and 2

can proceed to model execution (Figure 4.5). The results are stored and feed the execution visualizations.

In step 2, the hyperparameters are respective to the Random Forest model, the model being used in this instantiation of the approach. If the approach is used with another ML model (e.g. a neural network), the hyperparameters will be different and related to the corresponding model. After every model execution in step 3, the tool stores

Execution4 Project Project1

1 Step 1 Initial Data 2 Step 2 Tuning 3 Step 3 Hyperparameters

Passo 3 - Hyperparameters

Number of Estimators: * 100

Criterion: * gini

Max Depth: *

Minimum Samples Split: *

Minimum Samples Leaf: *

Max Features: *

Number of Jobs: *

Previous Execute

Figure 4.5: Creation of a Model Execution - Step 3

execution data, which feeds the respective execution visualizations and execution history. Visualizations and history are presented next.

4.3 Interactive Visualizations

For each execution, the approach presents a set of visualizations that conveys aspects related to the overall model, the training set, and instance-specific information. Explain-ML presents five consecutive tabs, which are complementary to each other, and with which the user can interact:

- *Evaluation Metrics*
- *Class Distributions*
- *Feature Importance*
- *Instance Analysis*
- *Analysis by class*

Next, we present each of one of the visualizations. For illustrative purposes, the figures concerns *Execution 3* (Estimators = 200. Criterion = *gini* and other Scikit Learn default parameters) presented in the Execution List of Figure 4.3.

4.3.1 Visualization: Evaluation Metrics

This visualization conveys aspects of the model, presenting well known ML effectiveness metrics such as: Accuracy, Error Rate, Recall, F1 Score, Micro and Macro F1 (Figure 4.6). Along with the values, the tab provides the meaning of each metric, accessible from the contextualized help (question marks on the screen):

- Accuracy: *“It represents, in general, how often the classifier is correct. Range [0..1]”*
- Error Rate: *“Represents how often samples are misclassified (1 - Accuracy). Range [0..1].”*
- Recall: *“When the real class is X, how often is it classified as X? Range [0..1].”*
- F1-Score: *“F1 is the harmonic mean between accuracy and recall. It shows how accurate your classifier is (how many instances are correctly classified) and how robust it is (it does not lose a significant number of instances). Range [0..1].”*
- Micro F1 *“Calculates F1 globally by counting the total of true positives, false negatives, and false positives. Range [0..1].”*
- Macro F1: *“Calculates F1 separated by class, but without weights for aggregation. Range [0..1].”*

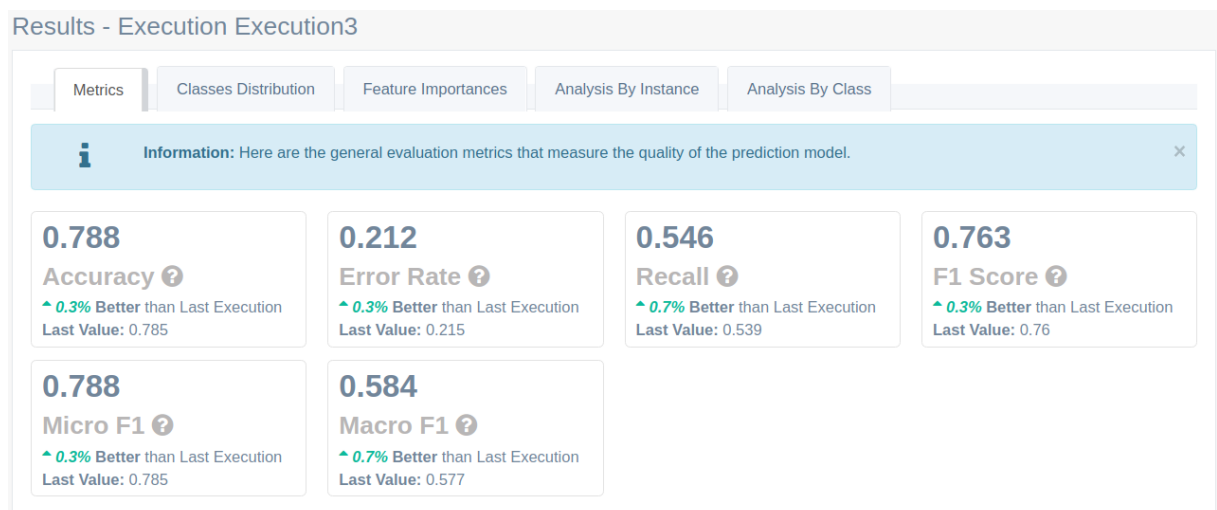


Figure 4.6: Visualization: Evaluation Metrics

This visualization also depicts the comparison of the value of each metric regarding the previous execution. This helps the user quickly perceive whether the new execution has improved or not model performance for each of the metrics. For example, in Figure

4.6, we observe that metric Accuracy had value of 0.788 in Execution 3 and value 0.785 in the previous one (Execution 2), potentially indicating a (small) improvement in that metric⁵.

4.3.2 Visualization: Class Distribution

This visualization conveys aspects of the dataset used in the model training (Figure 4.7). The graph shows the distribution of dataset instances by class. When interacting with each bar, the user sees the name of the respective class, the number of documents belonging to the class and some class specific metrics (F1, Precision, Recall). The user can confront these values with the general metrics of the model (Figure 4.6) and understand how it behaves globally and specifically, for each class.

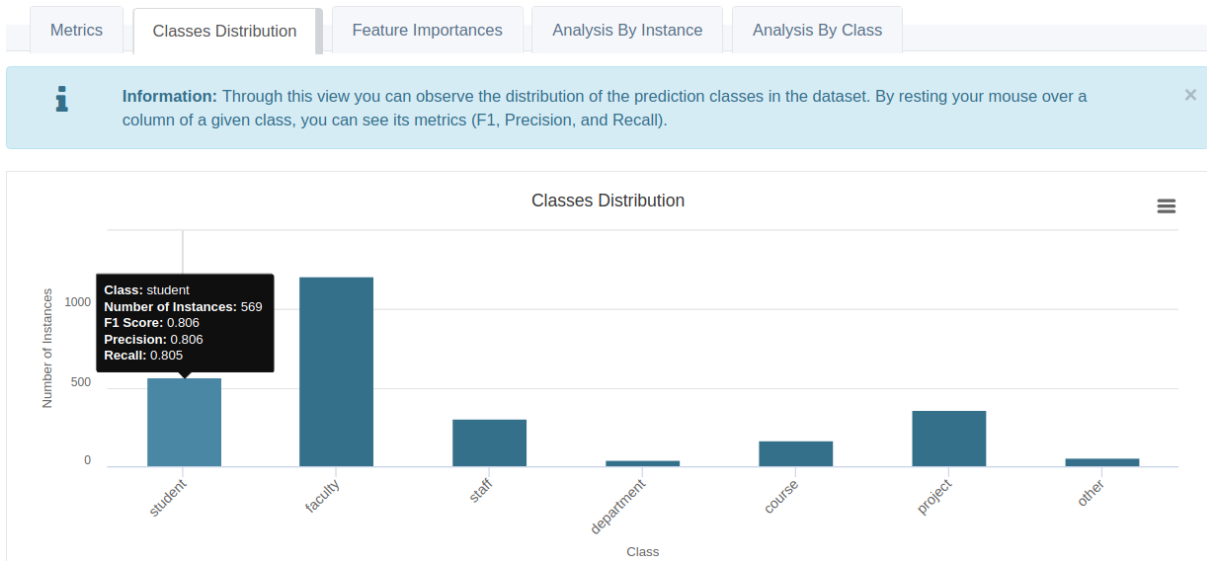


Figure 4.7: Visualization: Class Distribution

The bar graph allows users to easily grasp whether there is any dominant class, the major and minor classes, and other characteristics that may insert bias into the model, or even make it more difficult (or easier) to classify instances of a given class. Those observations, when aggregated with other tool views, assist the user in the overall model interpretability.

⁵In these results, and in the current version of Explain-ML in general, we are not considering statistical significance tests for accounting for statistical ties, but this function is in the planning.

4.3.3 Visualization: Feature Importance

This visualization conveys overall aspects of the model. Several types of complementary information are presented regarding the importance of features⁶ in the model (Figure 4.8).

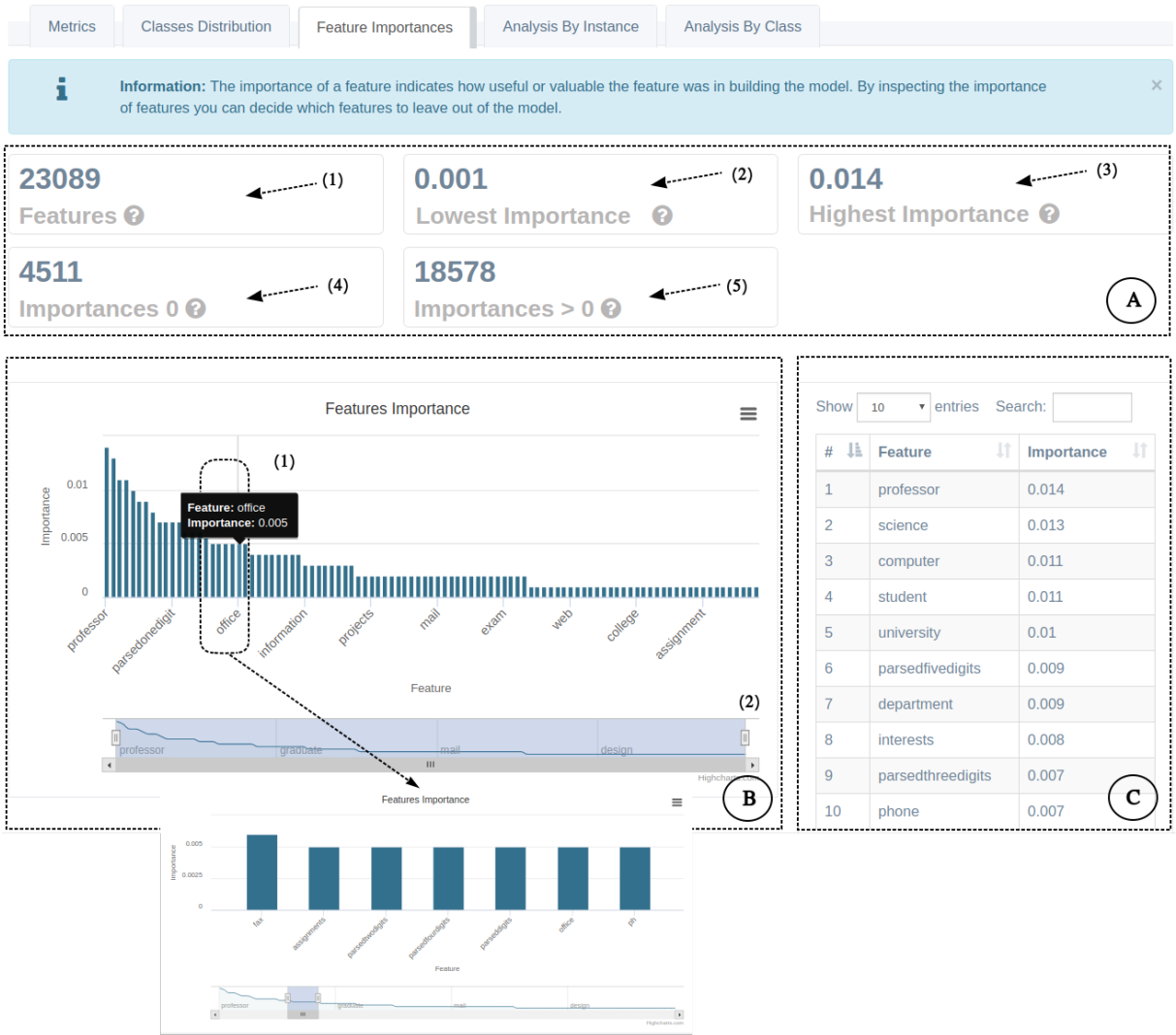


Figure 4.8: Visualization: Feature Importance

At the top of the tab (Figure 4.8 - A), we see some reference information: the total number of features (A.1), the lowest (A.2) and the highest (A.3) value of importance, the number of features with importance value zero (A.4), the number of features with importance value greater than zero (A.5). These values are important references for viewing the other elements available on the tab.

⁶In the current instantiation of Explain-ML, the value of feature importance considered is given by the decrease in decision tree node impurity weighted by the probability of reaching that node, but any other measure such as information gain or chi-square could have been used.

At the bottom left (Figure 4.8 - B), a bar chart is shown with the most important features. Users can interact by visualizing more or less features in the graph (B.2) and accessing the value and importance of each feature in the respective bar in the graph (B.1). In addition, the user can perform a feature search in the list of features shown at the bottom right (Figure 4.8 - C).

This visualization assists in observing the role of features in the model. The user can identify features that are more/less relevant to the model, and how they compare with other features. It is possible to observe the amount of features with some relevance (non zero importance) or that do not affect the model (zero importance). The tool also allows identifying features whose meaning is not aligned with the predicted class, but have non zero importance. Or the opposite, meaning compatible with the predicted class, but with zero importance. Especially in text bases, with high dimensionality, these observations can be very significant, allowing to focus on more relevant features that aid in the interpretation of the model. All of these observations may produce interesting insights for feedback in the models' re-executions.

4.3.3.1 Visualization: Instance Analysis

This visualization covers aspects related to instance-specific information. It aims to assist the user in understanding the instance classification.

Figure 4.9 - (1) shows the visualization for *Instance 0*. The tool presents the *Real* and *Predicted Class* for the instance, along with its most important features and their respective importance values (Figure 4.9 - (2)). In this way, the user can check if the model hit the classification and if the most important features have meaning compatible with the context of the predicted class. Otherwise, the user may have insights into features that can be excluded for model re-execution.

Besides, the probability for each class in the classification is shown (Figure 4.9 - (3)). Thus, it is possible to analyze whether the *predicted class* obtained a high probability in relation to the others. It may occur, for example, that two classes have very close probabilities. In this case, the user can use the complementary views (Figure 4.9 - (4) and (5)) to obtain insights about features that, if removed, may change the classification. This also provides feedback for model re-execution.

The *Word Cloud* allows users to have a general view about contribution of the main features to the result. That feature importance is relative to the instance⁷, unlike

⁷Contributions are generated through the *TreeInterpreter* package, available at <https://github.com/andosa/treeinterpreter>.

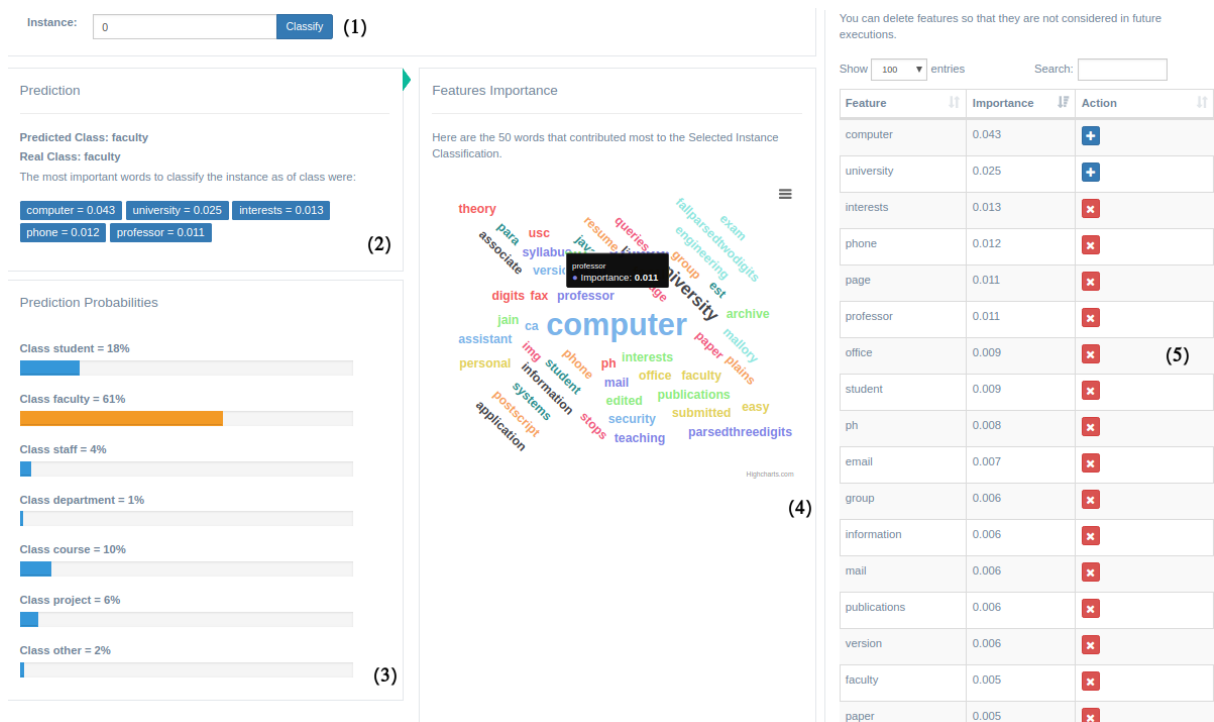


Figure 4.9: Visualization: Instance analysis. (1) *Instance*: *Instance 0* selected; (2) *Prediction*: presents the real and predicted class of *Instance 0*, and the most important features; (3) *Prediction Probabilities*: presents the probability for each class in the classification; (4) *Features Contribution*: presents a visual representation (word cloud) for the contribution of the main features of *Instance 0*. (5) Features Table.

Table 4.1: Information of instances from Execution 3 (Figure 4.3), with all dataset features. (*Due to space constraints, only the main three 3 features are shown*). The predicted class is presented along with the probability of its classification.

Instance	Real Class	Predicted Class	Main Features
0	<i>faculty</i>	<i>faculty</i> (61%)	<i>computer, university</i> , <i>interests</i>
1	<i>student</i> (41%)	<i>faculty</i> (50%)	<i>parsedthreedigits, university, research</i>
4	<i>faculty</i>	<i>faculty</i> (81%)	<i>computer, university</i> , <i>single</i>
5	<i>staff</i>	<i>staff</i> (47%)	<i>handouts, homework, midterm</i>
7	<i>student</i>	<i>student</i> (49%)	<i>phd, professor, interests</i>
9	<i>student</i> (21%)	<i>faculty</i> (45%)	<i>form, parsedonedigit, simple</i>

the overall feature importance, presented on Feature Importances tab. When the user interacts with the tool, the value of the feature’s importance is shown. For example, the value of the feature *professor* (*importance:0.011*) is shown in the Word Cloud (Figure 4.9 - (4)) and can also be viewed in the Feature Table (Figure 4.9 - (5)). The table allows users to sort features by name or importance value, and provides a feature search. The table also has the *Action* column, whereby each feature can be excluded (*red button with x*) / included (*blue button with plus sign*) in future model execution. Here, the user can perform several simulations and re-executions of the model, in order to understand its behavior.

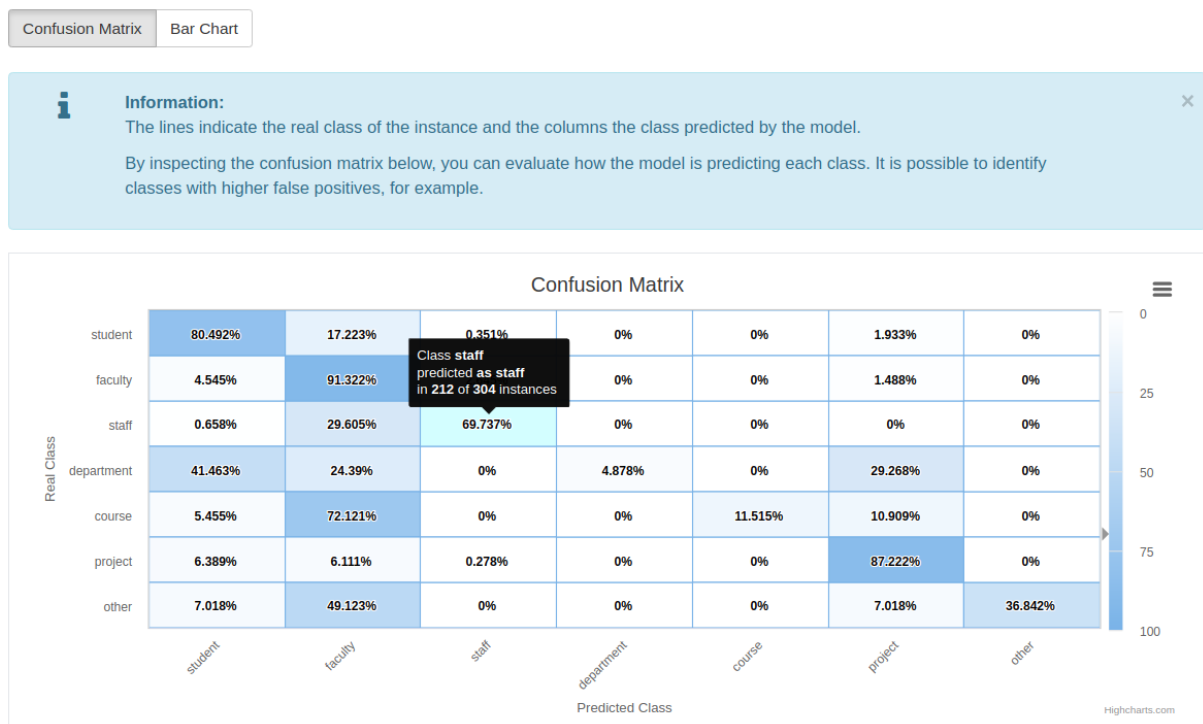
Table 4.2: Similar to Table 4.1, this table is relative to the execution after Execution 3, without features *computer* and *university*.

Instance	Real Class	Predicted Class	Main Features
0	<i>faculty</i>	<i>faculty</i> (52%)	<i>interests, phone, student</i>
1	<i>student</i> (44%)	<i>faculty</i> (49%)	<i>research, page, parsedthreedigits</i>
4	<i>faculty</i>	<i>faculty</i> (77%)	<i>research, parsedthreedigits, home</i>
5	<i>staff</i>	<i>staff</i> (44%)	<i>handouts, hours, midterm</i>
7	<i>student</i>	<i>student</i> (45%)	<i>phd, candidate, usa</i>
9	<i>student</i> (16%)	<i>faculty</i> (56%)	<i>parsedonedigit, form, interests</i>

Relying on this *Instance Visualization*, Table 4.1 contains some information on a few illustrative instances selected *ad-hoc*. The information assists the user in model interpretability, driving some insights, for instance:

- *Instances 0, 4, 5 e 7* are correctly classified and the main features (those with the highest importance values) are strongly related with the predicted class under the point of view of common sense.
- *Instances 1 e 9* are misclassified and the main features have no real meaning regarding the predicted class.
- Features *computer* and *university* are among the main features of *Instances 0 e 4*, whose classes are *faculty*. Note that class *faculty* is the major class of the dataset (Figure 4.7) and features *computer* and *university* are in the main features of the overall model (Figure 4.8).

All this complementary information may lead us to a question: *is it possible to specify more relevant features for each class?* New model executions can be performed to better evaluate this. Through *Feature Table* we exclude features *computer* and *university* and rerun model so we can see the effect of this feedback. Thus, those features will not be considered for the next model training. This adjustment causes a slight worsening in the model metrics presented in Figure 4.6. In addition, we can observe changes in instance analysis. Comparing Tables 4.1 and 4.2, we observe that *Instances 0 and 4* remained in the same classes, but with lower probabilities. And the meaning of their main features is less related to the predicted class. Which may indicate actual relevance of excluded features. *Instances 5 and 7* were slightly affected. The model continues to misclassify *Instances 1 and 9* and the probability for the real class of Instance 9 has decreased. We present here some examples of interactions, insights and feedbacks that can be performed on model, many others are possible. The tool, for instance, allows users to undelete features for subsequent executions.

Figure 4.10: Visualization: Analysis by class - *Confusion Matrix*

4.3.4 Visualization: Analysis by Class

This visualization covers aspects related to the model's predictions for instances and their ground truth. Here the focus is to understand how the model behaves for each class and whether it is more successful in predicting certain classes than others. Figure 4.10 illustrates what is presented to the user in this view. It is based on a confusion matrix that provides an overview of how the model behaves with regards to each class.

The matrix is organized in terms of real classes (rows) versus predicted classes (columns). Each element represents the percentage of instances of the real class (row-related) that was predicted as the column-related class. When interacting with the matrix elements, user visualizes the absolute number of instances and a brief description, for example, for the first element of the matrix: *"Class Staff predicted as Staff in 212 out of 304 documents"*.

A quick look at the matrix can provide several insights:

- classes in which the model performs best (based on the diagonal elements);
- for a given real class (line), when the model misclassifies instances, which other class is predicted more frequently as being the real class (based on the elements with the highest values on the line, except the one from main diagonal);

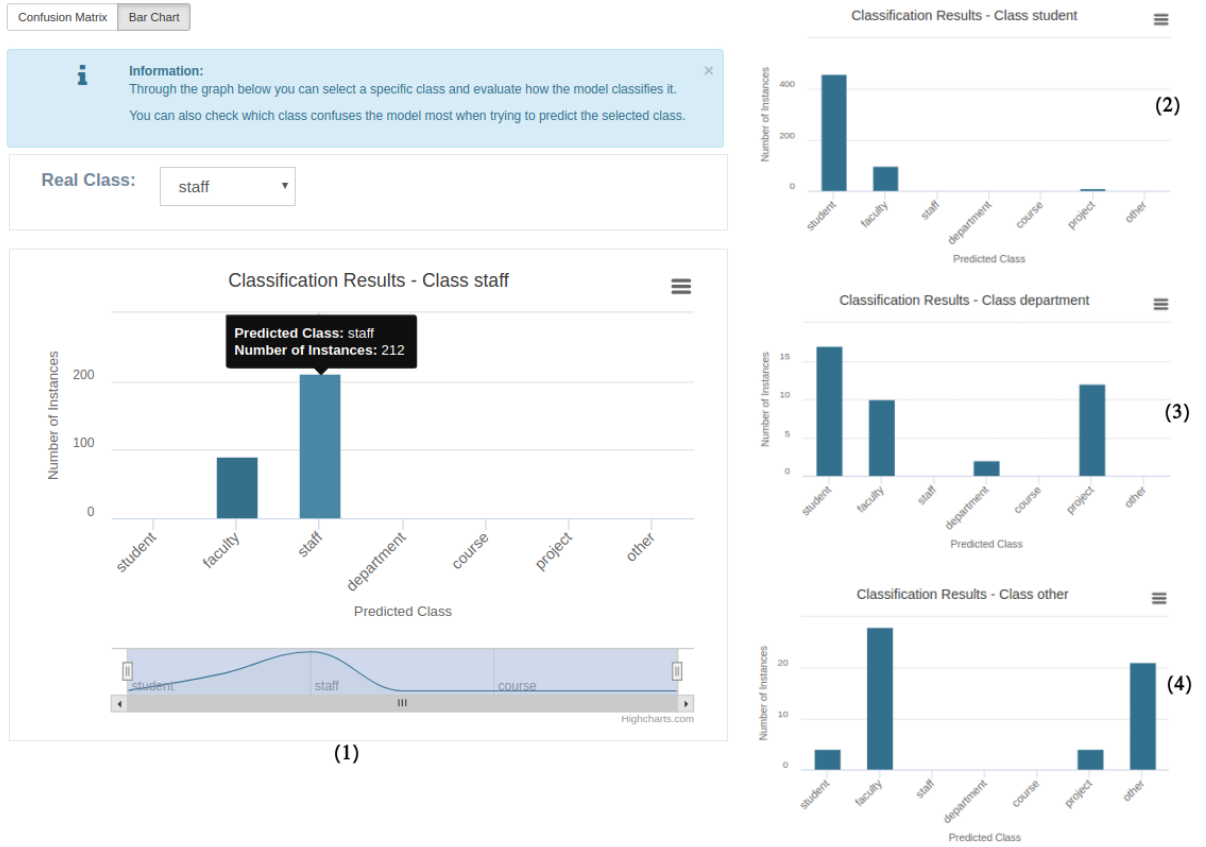


Figure 4.11: Visualization: Analysis by class - *Bar Chart (Classification results per class)*. (1) graph presenting the classification results for selected class - *staff*; (2)(3)(4) graphs resulting from the selection of classes *student*, *department* and *other*, respectively.

- for a given real class (line), when a model misclassifies instances, which class it predicts less frequently as being the real class (based on the elements with the lowest values on the line, except the one from main diagonal).

In addition, this visualization also provides a bar graph that represents the behavior of the model for a given real class (selected by the user), as shown in Figure 4.11. The graph allows the user to inspect each class separately. User interacts selecting the class he wants to inspect. In the graph, it is possible to verify the absolute number of instances related to each bar.

As mentioned, the *Explain-ML tool* provides complementary visualizations under several perspectives. For example, analyzing the model information depicted in Figure 4.11, we can see it behaves well for the *Faculty* and *Student* classes. If we compare the information with that available in the *Class distribution* tab (Figure 4.7), we notice that these classes are the ones with the largest number of instances. Thus, we can infer that a higher number of instances in the dataset ultimately contributes to the model learning better to predict instances of these classes. If we look, for example, at the *Department* class, we would observe the opposite: the model does not perform well on instances of this class, however, the dataset only contains a few instances of this class. This helps to

explain the lower values of the evaluation metric MacroF1 (average of the performance in each individual class) when compared to MicroF1 (global performance regardless of the classes). This would also give hints to the users about how to improve the model (for instance, by means of under or oversampling techniques to improve the balancing [7, 3]).

4.3.5 Project Execution History

As described at the beginning of Chapter 4, Explain-ML is organized according to the concepts of *project* and *execution*. A project can have multiple model executions, storing the executions history and presenting a comparative analytical graph between them.

The Project execution history provides two types of charts. The first one - *Metrics Variation* (Figure 4.12), represents the variation in the metrics along the executions. The chart includes some well-known performance metrics such as *Error Rate*, *Accuracy*, *Recall*, *F1*, *Micro F1*, and *Macro F1*. In the graph, each metric is represented by a color/dot type. The user can interact with the graph and verify the value of a given metric in a given execution. It is possible to observe the improvements or worsening of the metrics throughout the executions.

In addition, the other chart - *Variation by class* (Figure 4.13), provides a bar graph that represents the behavior of the model for a given real class (selected by the user). Each bar is an instance of what was shown in the *Analysis by class* tab (Figure 4.11), but in this case, the bars for all the executions are shown side by side. When interacting with each bar, the user sees the name of the respective real class, predicted class and the number of documents belonging to the real class. The graph summarizes the evolution of the model behavior regarding each class.

4.3.6 Final remarks

Explain-ML through its workflow, allows the entire process to be performed from hyperparameter tuning to model enhancements, interactively and without coding. The tool presented in this section is a version of the Explain-ML, instantiated with Random Forests, but the proposed approach is agnostic to the ML model. Other model-based visualizations can be incorporated in order to benefit model explainability. Explain-ML

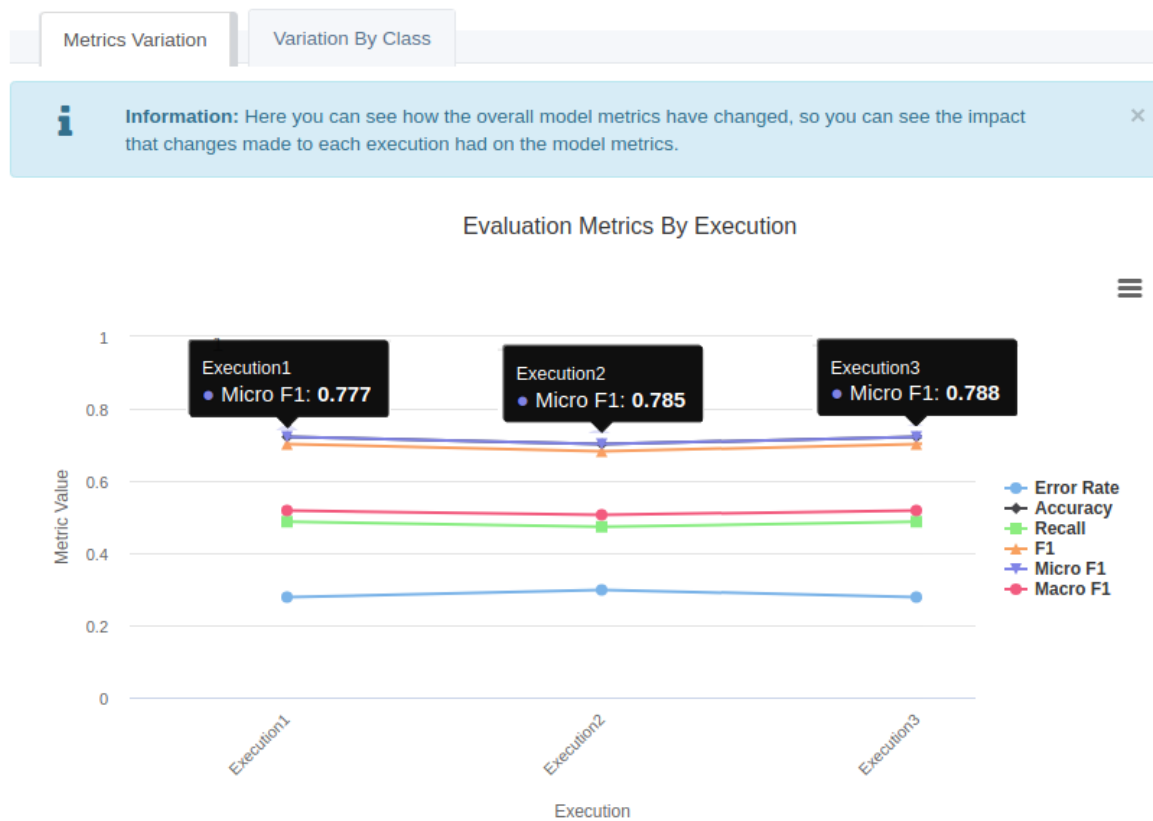


Figure 4.12: Project execution history - *Metrics Variation* chart . The user visualizes each metric in each execution, interacting with the respective point in the graph, as it is done for the metric *Micro F1* in *Executions 1, 2* and *3*. Here, *Micro F1* overlaps *Accuracy*, since both present the same values in the three executions. The graph does not show metrics of the Execution 4 (Figure 4.3) as it has not yet been completed.

provides complementary visualizations that allow users to acquire insights regarding the models. Several examples of these insights were discussed throughout the visualizations being presented in this chapter.

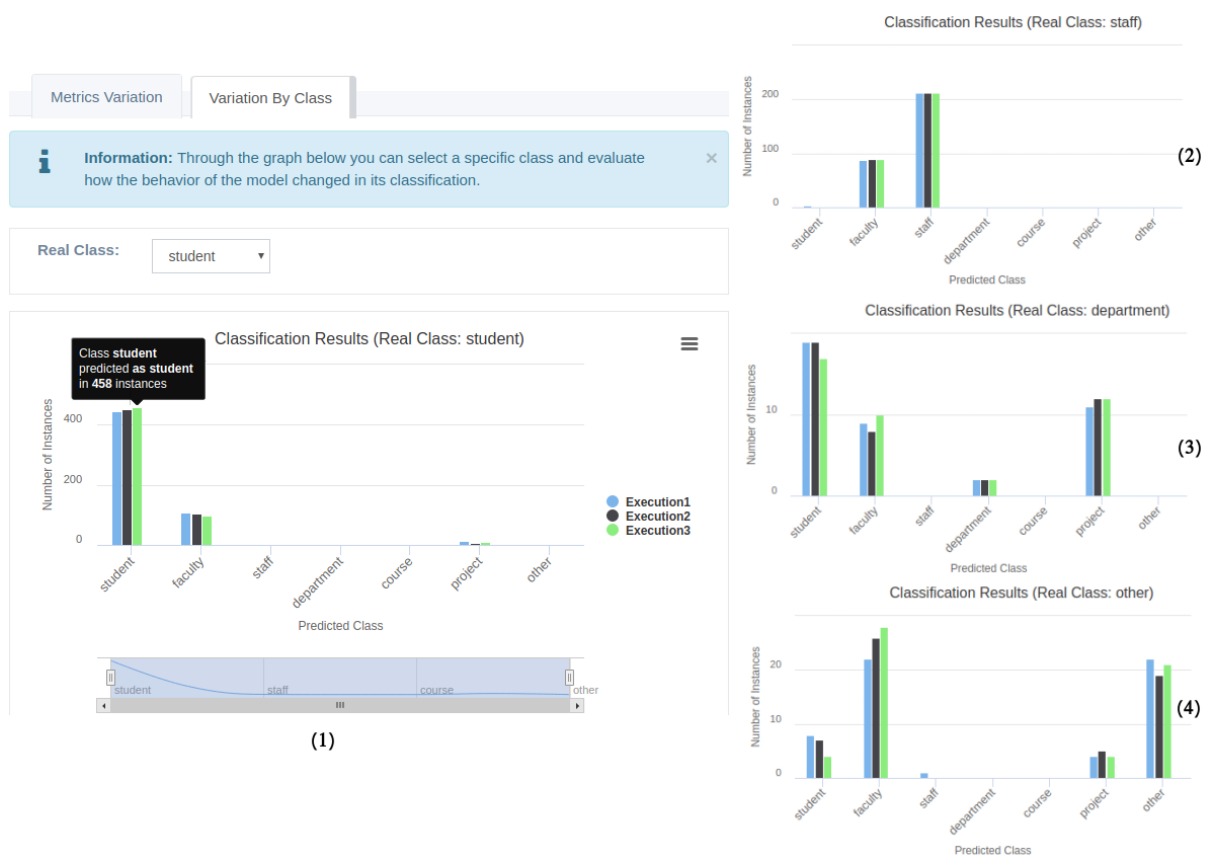


Figure 4.13: Project execution history: *Variation by Class*. (1) graph presenting the behavior of the model for the given real class selected by the user - *student*; (2)(3)(4) graphs resulting from the selection of classes *staff*, *department* and *other*, respectively

Chapter 5

Evaluation

In order to evaluate Explain-ML we conducted a qualitative study that allowed us to analyze in depth the users perspective and perceptions on the tool. The methodology adopted combines user interaction with the tool along with semi-structured interviews aimed at exploring how they perceived the visualizations that explain the model's outcomes. As mentioned, the study was conducted using the previously described instantiation of Explain-ML with RFs applied to text classification. In the next sections we explain in more detail the methodology adopted, how the evaluation was conducted and the results obtained.

5.1 Methodology

For the evaluation, users were asked to analyze the results of a specific classification task on a predefined dataset. The interaction was guided and participants were asked to train the model and go through all the visualizations that explained the application of the trained model to the chosen dataset. Then they were asked to interact with the tool to improve the model, and finally to analyze the execution history shown in Explain-ML. During their interaction, participants were encouraged to think aloud [34] and at the end of each step a short semi-structured interview about their views on the tool was conducted.

The WebKB dataset was selected for our evaluation. It contains 8.282 webpages classified under the following categories: student (1641 pages), faculty (1124), staff (137), department (182), course (930), project (504) and other (3764). The pages were collected from computer science departments of various universities in January 1997 by the World Wide Knowledge Base (WebKB) project of the Carnegie Mellon University (CMU) text learning group. This is a very interesting dataset from an analytical point of view, as it contains many classification challenges found in real applications: skewed (imbalanced) class distribution; non-trivial semantic overlap among classes; noise and ambiguity in the

text of the pages, etc. In this context, the idea was that our visualizations could help users to identify issues related to class imbalance and corresponding biases (e.g., with the *Class Distribution* visualization), which evaluation metrics would be more adequate to dataset (e.g., with the *Evaluation Metrics* visualization), which words help and which ones confound the classifiers decisions, among others (e.g., with the *Feature Importance* visualization).

In order to contextualize the evaluation tasks for the participants an evaluation scenario [5] was created. The scenario (Appendix E) described a situation in which users were participating in a research project that required the automatic classification of their University’s webpages. The scenario motivated them to use Explain-ML (a new tool they had just heard of) with a similar dataset (WebKB) to assess whether it would be useful to apply it to their own context (described in the scenario).

To do so, they used the tool to create a RF-based ML model for the problem, determine their trust in the model and its results and improve their model based on hints gathered from the explanations. An overall brief explanation of the WebKB dataset – with explanation about the meaning of the classes and corresponding statistics – was also presented to them.

Next we present the steps used to guide participants in their interaction with Explain-ML, and the interview guide associated with each of them. Appendix F contains both the original version of the form (in Portuguese) and the translated version (in English) of the Evaluation Script, which was used to guide the evaluations. Initially, we asked participants about their experience with ML and Random Forest, specifically. Then we presented Explain-ML to participants, describing its goal to explain the model, and we answered whatever questions they had about the tool. Once the participants had a general view of the tool we guided their interaction through it. We next describe the 4 main tasks that comprised the interaction steps, as well as the semi-structured interview associated to each one of them:

- **T1: Initial model training:** In this step participants were asked to perform the hyperparameter tuning and train the classifier (as shown in Figures 4.4 and 4.5). Although the tool allows for the tuning of a number of different parameters. – *Interview guide: After this step, participants were asked if they usually performed a hyperparameter tuning, and if so, if they used any tools to do so, and which one(s). They were also asked about the set of parameters available for tuning – if they were appropriate, usual or not, if they missed any parameters.*
- **T2: Analysis of model execution visualizations:** the visualizations of the tool act as explanations for interpreting the model. Participants were asked to analyze each one of the explanation visualizations available (presented in Section 4.3). – *Interview guide: For each visualization, participants were asked about their under-*

standing about the explanation, whether they found it easy, and which information it conveyed to them about the model, and if it was useful in their analysis. We also asked if they would change or add any new information besides what was shown. In case participants did not understand any of the elements shown in the interface, they were asked what their interpretation was, and any difficulties observed were registered, and the element or aspect in hand was explained to them. After having examined all the explanation visualizations, participants were asked to comment on their analysis of the model, whether they had identified any problems, which changes they would make in the model, and why, as well as their overall view and how they perceived the usefulness of the set of visualizations provided to explain the model.

- **T3: Model improvement:** Participants were asked to indicate and conduct some of the improvements (at least two changes) identified by them as capable of improving the model. They were expected to interact with Explain-ML to make the changes they felt would be the most relevant in the model (not necessarily all of them, in the interest of the time of the evaluation), and run a new execution of the model. – *Interview guide:* *Once the re-execution was performed, they were asked about the changes made and whether they could see their impact in any of the visualizations and, if so, what it meant in terms of the model. If they felt that they were able to improve the model, and whether the changes increased their trust in the model.*
- **T4: Analysis of history results:** Finally, participants were asked to explore the execution history (even though they only conducted two executions of the model) available (presented in Section 4.3.5) – *Interview guide:* *They were asked to comment on the changes in the model being conveyed through the history execution visualization, as well as whether the execution history was useful in the process of understanding and conducting improvements in the model and why.*

After the guided exploration of the tool was over, participants were asked about their overall perception of the tool and its explicability. They were asked if they had knowledge of or have ever used other explicability tool, such as Lime [36] or SHAP [29]. Finally, they were encouraged to make any other comments or suggestions they had about their experience in exploring the tool.

5.2 Conducting the Evaluation

As the tool was aimed at ML knowledgeable users, the invitation to participate in the evaluation was sent to Computer Science graduate students who had some knowledge of Machine Learning or whose research was related to the topic. The invitation was sent by email, and those who responded agreeing to participate, were contacted to schedule a convenient time for their participation.

Among the people who responded, we were able to schedule six¹ of them to participate in an evaluation session. Sessions were conducted in Portuguese (participants' first language) in an office at the University, between 24 and 30 of October, 2019. The Explain-ML - Random Forest version of the tool was set up in a computer made available to users, and the user's interaction and audio for all sessions were recorded. The audios recorded of the interviews were transcribed, and an analysis conducted. All participants were male and had their background in Computer Science. One of them was postdoctoral fellow, and the other 5 were PhD students. All of them had already experience with Machine Learning model during their graduate studies, but their experience varied in terms of how long they had used ML, which models and for which purposes.

5.3 Results

We present the main results of our evaluation², regarding users understanding of the model based on Explain-ML's explanations and their perspective on its support and usefulness in improving the model. We present the results organized by the aspects explored by participants in each task. Appendix G contains both the original version (in Portuguese) and the translated version (in English) of the participants comments during the evaluation.

¹When the goal of an evaluation is to yield insight about a system, a qualitative study with around 5 users is enough [34].

²Our study was conducted in Portuguese, participants' 1st language, and excerpts shown were translated by the authors.

5.3.1 Initial model training

All participants were able to understand and perform the tuning process without difficulties. They said that this was a step that they usually performed through SciKit or by using a script they had written previously. When asked about how he performed tuning, Participant 1 (P1) explains what he usually does: *"Yes [I perform hyperparameter tuning], I just use scripts... Sometimes I use grid search available in some tool." [P1]*

Although they were asked to only make decisions regarding the number of estimators, they observed and commented on the other hyperparameters made available. They mentioned that all the hyperparameters they usually used were depicted in the interface and that usually they did not even use all of the ones available.

Nonetheless, one participant pointed out that some other parameters could improve users' decisions during tuning, such as including an option to allow users to define the number of samples to be considered within a range set for a hyperparameter, as described in P6 words:

"It would be cool to include a tag with the number of values.... Here [referring to the parameters in the hyperparameter tuning screen] you included the minimum and maximum values, but you did not specify how many numbers you want to sample..." [P6]

In short, participants thought it was useful to have the hyperparameter tuning integrated into the Explain-ML tool, and had no difficulties in understanding the step in the workflow, or how it was presented in the interface.

5.3.2 Analysis of model execution visualizations

The first visualization explored was the one that shows the overall **Evaluation Metrics** for the classifier model (see Figure 4.6). Most participants had a good knowledge of the metrics shown. P6 was not sure about the meaning of MicroF1 and MacroF1 and used the explanation available about each metric to get a clear understanding of the difference between them. All participants were able to explain how they could use the metrics to analyze their confidence in the model, as described by P4:

"Micro is a general average and the macro is an average per class... One can already get a sense..., for example, we know that this dataset here is imbalanced

because it [the classifier model] couldn't learn from some of the classes." (P4)

They all agreed that the most used metrics were being shown in the interface, as explicitly mentioned by P5: *"From what I see, the most used metrics are available."* However, they did not all agree that this was the most interesting set of metrics to be shown. Some participants thought that not all metrics depicted were equally important (P3) or even that some would not be needed (P4, P5). In this direction, P5 suggested that users should be able to configure which metrics they would like to see (or not) in the interface.

On the other hand, other participants mentioned other metrics they thought would be interesting to see. They mentioned general metrics, such as Entropy, Precision or AUC (Area Under the Curve), as well as providing more information about the metrics that were being depicted. Both P4 and P6 made suggestions of adding information to the metrics that would allow them to analyze their variability, such as standard deviation and confidence interval³. Participants also suggested that metrics specific to the ML classifier at hand could be interesting to be depicted, as mentioned by P6 regarding Random Forests:

"There are some [metrics] specific to the classifier, that might be the case to include..., for example, what was the average depth of the tree, or the gini index of the trees. The problem is that when you generate too many trees, it could be difficult to visualize, right? But at least one way to present the distribution of indices, feature relevance [...]" (P6)

Finally, one participant (P5) thought that it would be interesting to have the metrics depicted not only for the classifier model, but also for each one of the classes.

Next, participants examined the **Class Distribution** visualization (see Figure 4.7). All users were able to visualize and understand the distribution of classes. All of them considered this visualization of the distribution very relevant. They used the class distribution to verify issues regarding the skewness (i.e., imbalance) of the dataset used, as exemplified in P4's analysis: *"I can now clearly see that the base is very imbalanced as I had mentioned before."*

Learning (or confirming) that the dataset was imbalanced helped users to further analyze and better comprehend the previous evaluation metrics values, as well as identifying possible biases in the model. P1 describes how this knowledge informed him about which metrics to look at closely: *"Indeed, if classes are imbalanced, you need to look at those metrics, micro and macro F1."*

P2's analysis also makes it clear how the class distribution visualization relates to the metrics visualization: *"I look at this distribution, how many documents are in each*

³Though planned, this capability was not implemented in the tested version of Explain-ML.

class, to know which metric will be the most interesting one to inform me whether my model is effective or not. Because, for instance, looking at accuracy on a very imbalanced dataset doesn't mean much ... "

These strategies described by the participants corroborates our *holistic hypothesis* that one needs to see the “big picture” (in our case, a comprehensive set of complementary visualizations) to better comprehend ML models and their decisions.

The following visualizations depicted **Feature Importance** (see Figure 4.8) and were considered extremely relevant by the participants. Four participants (P1, P2, P5 and P6) highlighted how these visualizations have assisted them to better understand (i) how their model made decisions; (ii) which were the most discriminative features of the model, as well as (iii) the noisy ones in the classification task. The following comments from the participants illustrate how this visualization helped them in better understand their model:

"The word “college” surprised me (it didn't appear in the top of the rank among the most discriminative), maybe it's because it is spread across all classes, so it won't help discriminate." (P1)

"I can use these explanations to understand how the model makes its decisions and to define whether an instance is from one class or another. And when I look at a graph like this... Every time a very meaningful word comes up, I can map it very assertively to the decision model. Then you can get which words, here at the end of the list, do not necessarily have this assertive capability." (P2)

"It's also cool because you can see whether there are any words that may be introducing some noise in the classification model. You can go “down on that” [scroll down] and see in which instances that word has a high importance weight. If I can download this model, I can understand the word that confounds the model and how to solve it." (P2)

"I think this is amazing, because now, with what you [the tool] told me , I can say... Hmm, let's suppose the model is misclassifying because of the feature [word] “student”, then you say: I'm taking “student” off... And it's among the most important features." (P5).

"It looks like the model is OK because all these top-ranked features shown here are related to the main subject of the dataset as you described it at the beginning. " (P5)

"Numbers with five digits ("parsefivedigits") correspond to the sixth most important feature of the model...! It would be nice to understand why, right?! What are the meaning of these numbers in this dataset?" (P6)

Three of the participants (P1, P3 and P4) considered the visualization of the low-ranked features as very important to assist them in selecting features for the model. Based on this, they pointed out feature selection as an important aspect for dimensionality reduction, understanding of the model and for potential improvements. The following excerpts from participants evaluation describe their strategies to take into account the low-ranked features to improve their model.

"I usually look at this feature importance table to find out which features are the most discriminative, the ones that really help the classifier ... but it's also very nice to see the least discriminative features – so, if I want to make a feature selection, I know these are the ones I can eliminate from the model without much of an impact." (P1)

"I would eliminate these words from the vocabulary to generate a smaller model with fewer features, easier to interpret and faster to run as well." (P1)

"If I have 7154 features that are not important, I could want to see which features are these so I don't use them in the model." (P2)

"It's cool! You can perform feature selection based on this. For instance, you can set a threshold on 0.002 and you do not keep values smaller than this. Then you can remove a lot of features ... It produces a "good" reduction of dimensionality. (P4)"

It is interesting to note that in most interpretability tools, mainly when explaining individual instances, only present the most discriminative features.

Next, participants examined the visualizations **By Instance** (see Figure 4.9). The visualizations were considered extremely relevant by the participants, and all users were able to visualize and understand them. Three participants (P1, P2 and P5) highlighted how these visualizations have assisted them to (i) understand how the model made decisions for each instance, (ii) which were the most discriminative features for particular instances, as well as (iii) which features cause confusion to the classifier. The following comments from the participants illustrate how these visualizations helped them to better understand their model decisions for the instances presented:

"Then I can see, for each object, what were the features, attributes, most important to classify it. I can see the relative value of this importance. (P1)"

"I found this one super cool. I can see what was decisive for the model to predict that instance as being in that class. (P2)"

"This view allows me to see what features of this particular instance it [the model] used to classify [the instance]. (P5)"

Three of the participants (P1, P2 and P5) highlighted that visualizing the probabilities helped them to identify biases and understand the certainty level of model for each classification, as illustrated in the following comments from the participants.

"I can see the probabilities of each class. I can see for which ones it [the model] is most certain of the result. (P1)"

"In terms of the model, this prediction probability is very cool, because you can see where this thing [the trained model] is being biased the most. (P2)"

"This is very cool, it would be how much the model ... How much it was in doubt. (P5)"

Nonetheless, some of the participants mentioned other features they thought would be interesting to add to the ones available at the current visualization. P1, P2 and P4 made suggestions of adding more options to feature selection on Explain-ML:

"I try to look and see what is generating noise. But we usually do it more automatically. I would take a script and already eliminate the top X words. (P1)"

"Either I don't eliminate it, or I test some values... The 10% worst, 1% worst, 50% worst and see if the result changes a lot. If it doesn't get worse, I'll keep it out. I do this for all test instances. (P1)"

"I just wish I could eliminate features that have importance zero, for example (P2)"

"[...] what I usually do is reduce... 30% of the least important features I throw away. (P4)"

Notice, that the Visualization by Instance is interactive, and users can eliminate features from the model at the interface. Nonetheless, the point made by the participants is that it would make sense to also offer users some more automatic strategies to do so. Based on the strategies they described, it would be interesting to allow users to determine ranges of features to be eliminated, based on a threshold (e.g. 30% worst) or on the value of their importance (e.g. all features that have importance zero or below a certain value).

The last visualization of the execution results explored was the one that shows the results **By Class** for the classifier model (see Figures 4.10 and 4.11). Most participants had a good knowledge of the confusion matrix shown and considered that it was relevant to depict it, as commented by P2: *"It's a concept I'm already used to, the confusion matrix. I liked the colors, I think they highlight ranges that better direct your analysis. (P2)"*

The analysis of the results by class allowed participants to identify which classes were the easiest and hardest to predict, as illustrated by P1's analysis: *"The class 'other' is really hard to classify, it is spread here [pointed to the line for 'other' in the confusion matrix] It seems that the 'faculty' class is the easiest to classify, it gets it right more often. (P1)"*

They were also able to identify biases in the data such as mentioned in P6's analysis: *"It [the model] is extremely biased by the amount of examples... I noticed based on the classes 'college' and 'department'. 'Department' was the one misclassified the most and 'college' was the one it [the model] got right the most. (P6)."*

Moreover, two participants (P3 and P5) reported that the visualizations helped them identify the need for class balancing. P5 describes how he uses the classes visualization to identify which classes the model is not able to correctly classify, and how he would combine this information with the visualization by Instance to better understand the problem:

"I always want to know the following: Where is my model missing the most and why? So here [seeing the class visualization] I can see which classes it's getting wrong the most, and maybe in the previous step [By Instance] I'll see why it's missing... because of some of the features it's choosing. So this might help me balance the classes. (P5)"

Based on identifying the need to balance classes, P3 describes how he would go about it, and conveys his expectation that Explain-ML could also support this task:

"Did the tool have the option to balance? From each class, take the same amount of documents... I would lessen this impact... I would balance, but it would have to balance at the same rate, otherwise it would be different from the real world. (P3)"

Finally, the visualization by class supported participants in making decisions about the classifier, not only in a more immediate level regarding which the next step would be (as mentioned by P3), but also about their overall strategies towards improving the model (as indicated by P2).

"P3: From the analysis, I have to greatly improve my classifier for 'department' and 'course', right? The ones that are getting misclassified the most."

"P2: It allows you to make decisions like, instead of making a multiclass classifier, making an ensemble, and having separate classifiers where you could put the most discriminating thing analyzed separately. This [the visualizations] allows the researcher to make this kind of decision."

After having explored each one of the 5 visualizations available, participants were asked whether they believed the visualizations were useful, and if so, how they had helped them better understand the model. All of the participants considered the explainability offered by the visualizations to be useful. P4 makes it clear that he considered Explain-ML tool gave a very complete account of the model: *"I could understand very well [the model]. Even though I already knew the dataset, this here [the visualizations] is a complete explanation of the dataset. I don't see any other information beyond what's here, you know ?!"* P3 also describes the overall benefits he perceives: *"The tool was good for visualizing the data and proposing improvements."*

Although they all thought it was very useful in general, some of the participants highlighted different aspects of Explain-ML they thought were specially interesting. P1 commented about the usefulness of the model, and emphasized the value of the word cloud: *"I could understand the result. I can see the model well with the word cloud, for text [classification] it is very good."* P6, on his turn, stresses the value of some of the other visualizations: *"Sure [the visualizations were helpful], for understanding, of course [...] The analysis by class, matrix of confusion, sure. Regarding the importance of features, I think this graph gives a good idea of how the importance of features is distributed."*

P2 commented on the benefits of the Feature Importance visualization: *"They [the visualizations] are very useful, [...] Specially when you have this very large volume of features, it's hard for you to understand these results... how the classifier is using this information to generate the results."* He conveyed his satisfaction with the Explain-ML by concluding *"I would super use this. I think the usefulness of this is huge. I think all the proposed visualizations add a lot and they are quite complementary."* P5 expresses as a positive point Explain-ML's holistic approach: *"I think the information I got here would enable me to refine my model much more easily than I would if I had to do testing, training... Here I have a tool that helps me. [...] Besides doing ... tuning the hyperparameters I would have to perform a large evaluation just to get to the same point that I would by just executing your tool. It really helps not only fine tuning the parameters, but also fine tuning the model. (P5)"*

Its interesting to notice that although participants highlighted different aspects of Explain-ML, they explicitly mentioned that the tool supported them in the main tasks related to modeling a classifier, namely understanding (i) the results of the trained model, (ii) how the classifier uses the information to generate the results, (iii) problems and (iv) biases in the trained model, (v) as well as assisting in the identification of possible

improvements in the model.

Although participants reported that they considered Explain-ML useful, some of them thought it could be improved and made suggestions on other aspects they thought would be relevant to include in the tool. Three participants (P1, P3, e P6) suggested including a visualization of the importance of features for each prediction class, as shown in the following comments:

"Show the importance of features by class... which features are more important for each class. I think the tool has everything that is relevant, but it's only a suggestion in case you want to add. (P1)"

"I wanted to see the features with higher impact per class. (P3)"

"I don't know if there is a concept of the importance of features per class, so maybe it would be the case to include it. (P6)"

P2 also suggested it would be interesting to present the relationship between the model features:

"I think this work has a huge potential to expand if we consider getting into how features interact, especially when you have other [machine learning] models. (P2)"

Specifically for the use of random forest models, P1 suggested that it would be worth considering the possibility of displaying a simplification of the forest structure generated:

"One thing I'd like to see in the models, which I know is very difficult, is either the whole tree or a piece of it. I've even looked in the model file itself, in XML, but it's very hard to visualize. I would look closer to the root first, where the most important features are, just to try to see how the combination of features is helping. The problem with looking at importance is that you only see how features in isolation contribute to the model. You can prune the tree... It wouldn't be the final model, but it would be a simplification. The most important thing will be near the root. Anyway, it's something to think about. (P1)"

The suggestions are interesting, as they provide us with other aspects to consider that could complement the current explainability provided. The suggestion of visualizing the tree when using the Random Forest model has been considered, and is on our intended list of visualizations worth adding. The challenges encountered in including it in this version of Explain-ML were related to the large size of the trees, and providing

a visualization that in fact would be useful. The fact that three out of seven participants made the same suggestion regarding the visualizations of features per class could be perceived as a sign of its relevance and the need to consider including it as suggested. Finally, P2's suggestion considers also dealing with other machine learning models. As Explain-ML is intended as an agnostic model tool, these comments are especially useful and will certainly be considered in the next steps of our research.

5.3.3 Model Improvement

After having analyzed the model, participants were asked to indicate and perform some model improvements (at least two changes) they had identified in the previous steps as potentially useful to improve the model. Participants adopted different strategies in their attempt to improve the model: (i) removal of less important features (P1, P2, P4), (ii) removal of most "common" features (P3), (iii) removal of features that appear to be noise in the dataset (P5, P6) and (iv) change of hyperparameters (P5).

P1, P2 and P4 removed some features they considered less important. Based on the explanation provided by the visualizations resulting from their strategy, participants were able to easily evaluate the impact of the changes they made. Although P1, P2 and P4 adopted the same strategy, they chose different features to remove, and thus the resulting impact also differed, as can be seen in there comments:

"Cool, it compares with the previous one [execution]. I worsened the results ... the most features for each instance have completely changed. (P1)"

"Ah, that's cool... So this is the impact I had deleting that feature [...] Cool, so I can see that deleting that feature was good for my model. (P2)"

"Here [Evaluation metrics visualization] it gives me the information about what has improved and what has gotten worse. [...] It is noticeable that the rest of the features also differed in importance. [...] Improved in micro but got worse in macro. (P4)"

P5 and P6 decided to remove features that seemed to them to be noise in the dataset, and P5 also made changes in the hyperparameters. P6 took out 3 features he identified as noise and inspected the result, but noticed that it did not significantly impact the model. When inspecting the evaluation metrics P6 commented: *"0.2% worse is statistically the same"*.

In the case of P5, the performed changes had had a higher (negative) impact on the model, as he expressed: *"Wow, I made the whole thing worse. See... I removed a very important word for it [the model], see what happened ?!"*. As he explored the different visualizations, he commented on the various impacts he noticed of the changes:

"Gosh, accuracy is the same value. Got a little worse too."

"It seems like the number of [features with] importance zero has decreased."

"It makes more sense now, the words you use to predict."

Based on P5 comments it seems like, even though the changes decreased the model's performance metrics, it made sense as a step to improve the model, as he mentioned: *"It improved the confidence in the model, how to explain it to others... I can explain it much better."*. He concluded, commenting on Explain-ML's support: *"It showed me why it got worse."*

P3 adopted a strategy that may not seem obvious at first, as he removed an important feature from the model. He explained as the feature was a generic word in the domain ('computer') and that he removed it to see if other features would become more discriminative. Analyzing the changes he thought it improved the model - P3: *"It has improved... the error has decreased..."*. Nonetheless, he commented he would have liked a little more information to analyze if the strategy had been successful - P3: *"I wonder if it increased the difference? When I took it [feature 'computer'] out, did this one here [pointed at feature 'science'] become more important?"*. Specifically he mentioned it would be interesting to be able to compare feature importance from one execution with the previous one.

It is interesting to notice that although all the participants were examining the same model, based on the visualization, they chose different actions they believed could improve the model. Even though some of them were not successful, participants were able to identify the low or negative impact of their changes and make considerations of what would be the next steps they would take in improving the model.

The different strategies adopted depend not only on the tool being used, but also on participants' previous experiences. Although all of them were knowledgeable ML users, they had different levels of expertise. Furthermore, we could argue that because Explain-ML provides a broad account of the model's execution, it does not guide users to any specific path to improve the model, but rather offers users an ample exploration space. This provides users with autonomy to make different decisions and conduct distinct investigations and analysis according to their contexts (i.e. dataset, classification goal or ML classifier).

5.3.4 Analysis of history results

Finally, participants were asked to explore the execution results history. Even though participants compared only 2 executions, they all considered the execution history useful in helping them having an overall view of their tuning of the model, as expressed by P2: *"So here I could have several executions and check how much is changing according to these changes that I am making in my model [...] Very cool, because then you can see this variation in terms of execution [...] I think it is cool to see the impact of my changes on the model."*

P4 also mentioned the benefits of the overall view: *"Making a chronological order of them [executions] was pretty cool. [...] It's possible to see what I did wrong and where that error occurred."*

The participants made it clear how the execution history results helped them better understand the impact of their changes in the model, as many of them noted:

"Overall, I realized that some classes have not changed at all. (P5)"

"It made the department classification worse. Removing that feature didn't improve the department classification. If the goal was to improve the department classification, it wouldn't be interesting to do that. (P3)"

"I could see the impact here, 0.1%, because I eliminated only a few features. (P1)"

"Here, for instance, even though it told me that there was a 0.5 variation, looking at the graph you can see that the impact is small, so the feature that I excluded does not impact my model much, since I can see that my results continue to be linear and parallel in the same way... But I think that in a scenario in which there is really a big impact, this will be very interesting to be able to see this variation, in terms of how much it has changed. (P2)"

Finally, based on the execution history some of them commented how it could provide them with insights regarding their next steps in tuning the model, as described by P6: *"It is possible to have an intuition of what can be improved, but that I haven't improved yet [...] It was possible to have more confidence, but I still don't feel confident with this result."*

5.3.5 Overall comments

In the last part of the interview, participants were asked about their overall view of the system and whether they thought it would be useful to them considering their activities and strategies to use ML modeling. All participants stated that they would use the tool, as illustrated in the following comments:

"I found it amazing [...] It really makes experimentation a lot easier. (P6)"

"Not only for this task [the one explored in the evaluation], but I can already see using something similar in my future applications. (P5)"

"I would super use this. I think the usefulness of this is huge. I think all the proposed views add a lot and they are quite complementary.(P2)"

As users explained why they would use Explain-ML, many of them highlighted some of the aspects they believed were the most interesting in the tool:

"I missed [in the past] a lot of information that I'm seeing here now. It's exactly what I wish I had seen at that time. If I had this framework that I have here, I think it would have been much easier to explain the model I was generating. (P5)"

"I would use it [Explain-ML] to improve the model. [...] The impact of a feature on classification, I think this is the most important [visualization].... And the metrics also... Exporting the confusion matrix is also important. (P3)"

"I would use it [Explain-ML] to know where the error came from and where it went [...] This possibility to keep the execution histories is very good, because we usually do it through a table, but in a table you forget... you don't put parameters, don't put a lot of things, and here in the system there is all that. It's a way to keep it that way, an organized way, and it's not a hassle keeping everything you need. I think everything I do is already there. (P4)"

They also stressed that it covered most of the analysis they did, as P4 mentioned *"I believe everything I do, is already there."* Some of the participants had experience with other explainability tools and compared them to Explain-ML:

"I've already used SHAP. Actually I don't use all it offers, because I know it does a lot of things, but the things I have used I think you're doing [in Explain-ML].(P2)"

"I used another tool with less features than this, which is Lemonade. Everything it has yours already has. Yours has more features [...] Parameter tuning is one thing that is not available there [in Lemonade], there we select the parameters we want and execute it with these parameters. [...] Being able to see the results for instance, and everything integrated, the word cloud, the graphs, the metrics, is not something I [usually] see. (P1)"

Nonetheless, some of them pointed out that other explainability tools cover a varied set of ML classifiers, and many said that it would be something that should be incorporated do Explain-ML:

" It [Lemonade] has already incorporated other models (cluster, frequent patterns, other machine learning algorithms), but it doesn't have as many views as it has here. So I would incorporate that, include other model types. (P1)"

"I've already used the orange tool, weka... Here you are just analyzing the random forest, there I have several things. (P3)"

"I wish this [tool] was for every other model, that I could have such a tool for any other project I was going to do, it was going to be a sensational contribution to this era of deep learning, for example. (P5)"

"I found it amazing, one of the things I would really like to see, which I commented with you, is being able to choose generic classifiers (outside of ScikitLearn). (P6)"

Some of the participants mentioned as one of Explain-ML differentials and advantages, the fact that it offers an easy-to-use interactive user interface. Participant 2 emphasized that as it did not require any programming knowledge from the user, he considered it could be interesting and useful to a wider audience than most tools:

"P2: I would use it for sure... Me, from computing, would use it... But I think it has an even greater potential when you deal with what I have already commented... When we start thinking about outsiders [from the computing field]... SHAP and these other tools have their limitations: the guy will have to generate the model, will need to implement it in python, he will need to know how to manipulate [the tool]... To use this here [Explain-ML] I do not need to know python ... So it has a very interesting potential, a wider audience."

Participant 5, on his turn, thinks that having such a usable and pleasant interface is useful, independently of the users' computing expertise:

"P5: I liked that... besides being useful, a system has to have a friendly face, like yours does. A system can be very useful sometimes, but if it does not have a cool look, doesn't have such a simplified usability... [then] the person stops using it. Seeing these graphs, seeing this facilitated interaction, it pleases the eyes to see. Even if you get stuck at some point, you want to keep moving, because it gives you good information in a way that your brain easily assimilates, so it's very cool."

Finally, it is interesting that P5 highlighted not only Explain-ML's usefulness in ML modeling, but also how he felt it could be useful for the research in the field:

"I wish it were for all the other models [...] Most 'state of art' works today don't have that. The guy [author] just says: my metric beat the state of the art, but he doesn't say why, he doesn't show why... If you had such a tool to explain the model, you would know where it is weak, what are the constraints of that model, where it is strong, where it is weak [sic]. And that is what is missing in most articles today. (P5)"

Next, we present a summary of benefits perceived by the participants (Table 5.1) and a summary of improvement suggestions made by them (Table 5.2) for Explain-ML, based on participants comments during the evaluation.

Table 5.1: Summary of Benefits

Task: Initial model training (T1)	
Benefit	View
Hyperparameter tuning: Participants thought it was useful to have the hyperparameter tuning integrated into the Explain-ML tool.	Hyperparameter Tuning
Task: Analysis of model execution visualizations (T2)	
Benefit	View
Analyze confidence in the model: Participants were able to explain how they could use the metrics to analyze their confidence in the model.	Evaluation Metrics
Better understanding of which metrics to look at closely: Learning (or confirming) that the dataset was imbalanced helped users to further analyze and better comprehend the previous evaluation metrics values, knowing about which metrics to look at closely	Class Distribution

Verify issues regarding the skewness: Participants used the class distribution to verify issues regarding the skewness (i.e., imbalance) of the dataset used.	Class Distribution
Better understanding of the model: Participants highlighted how Feature Importance have assisted them to better understand how their model made decisions; which were the most discriminative features of the model, as well as the noisy ones in the classification task.	Feature Importance
Help in selecting features for the model: Participants pointed out feature selection as an important aspect for dimensionality reduction and for potential improvements.	Feature Importance
Help in understanding the instance: The participants highlighted that visualizations 'By Instance' have assisted them to understand how the model made decisions for each instance, which were the most discriminative features for particular instances, as well as which features cause confusion to the classifier.	Instance Analysis
Help in identifying biases and understanding the certainty level of model: Participants highlighted that visualizing the probabilities helped them to identify biases and understand the certainty level of model foreach classification.	Instance Analysis
Help in identifying which classes were the easiest and hardest to predict: The analysis of the results by class allowed participants to identify which classes were the easiest and hardest to predict.	Analysis by Class
Help in identifying biases: Users were able to identify biases in the data	Analysis by Class
Help in identifying the need for class balancing: The participants reported that the "By Class" visualizations helped them identify the need for class balancing.	Analysis by Class
Support in making decisions: The visualization by class supported participants in making decisions about the classifier, regarding which the next step would be and also about their overall strategies towards improving the model.	Analysis by Class
Task: Model Improvement (T3)	
Benefit	View

Help in identifying the impact of changes on the model: The Participantes were able to identify the impact of their changes and make considerations of what would be the next steps they would take in improving the model.	All Previous Visualizations
Ample exploration space: Explain-ML provided users with autonomy to make different decisions and conduct distinct investigations and analysis according to their contexts (i.e. dataset, classification goal or ML classifier).	All Previous Visualizations
Task: Analysis of history results (T4)	
Benefit	View
Overall view the model tuning: Participants considered the execution history useful in helping them having an overall view of their tuning of the model.	Execution History
Help understanding the impact of changes: The execution history results helped the participants to better understand the impact of their changes in the model.	Execution History
Insights regarding next steps: The visualizations provide participants with insights regarding their next steps in tuning the model	Execution History

Table 5.2: Summary of Suggestions

Task: Initial model training (T1)	
Suggestion	View
Range for number of samples: Improve users' decisions during tuning, such as including an option to allow users to define the number of samples to be considered within a range set for a hyperparameter.	Hyperparameter Tuning
Task: Analysis of model execution visualizations (T2)	
Suggestion	View
Option to configure which metrics users would like to see (or not) in the interface: Some participants thought that not all metrics depicted were equally important(P3) or even that some would not be needed (P4, P5).	Evaluation Metrics
Add information to the metrics: Participants added that we should add information to the metrics that would allow them to analyze their variability, such a standard deviation and confidence interval.	Evaluation Metrics

Add metrics specific to the ML classifier	Evaluation Metrics
Add metrics for each one of the classes	Evaluation Metrics
Add more automatic strategies to feature selection: Participants made suggestions about add more options to feature selection like: • Eliminate the top X words • Eliminate a particular percentage of the less important features and • Eliminate features with importance equals zero	Instance Analysis
Automatic class balance strategies: Participants made suggestions about add support to class balance	Analysis by Class
Overall Suggestions	
Incorporate a varied set of ML classifiers	
Tree visualization when using the Random Forest	
Add visualizations of features per class	

Chapter 6

Final Remarks

In this final chapter we present a discussion on the results, the main contributions and limitations of this work. It also contains our final considerations and possible future works to complement it.

6.1 Discussion

Our evaluation of Explain-ML allowed us to explore in depth the perspectives the participants had of the tool, as well as how the proposed solution contributed to ML modeling and their activities in general. Based on participants perspectives and experience (albeit short) with Explain-ML, we collected positive indicators about our intended solution, as well as how it was presented to users.

Participants explicitly expressed the advantages they perceived in Explain-ML taking a **holistic approach** and integrating steps ranging from hyperparameter tuning, to model execution, the exploration of its visualizations/explanation, all the way to, and model improvement (including comparing different executions). In exploring the explanations offered, they considered them complementary and informative. Based on the insights they gained through their exploration and that were expressed in their interviews, it is possible to see that the tool allowed them to understand the outcomes of the model, as well as how it was working. Furthermore, they were able to identify relevant strategies they could take to improve the model, and make changes in that direction. In other words, Explain-ML's **black-box explanation** approach, based on **outcome explanation** and **model inspection** was well received by users and generated positive indicators of how it could be successfully used in ML modeling.

The current implemented version of Explain-ML offers only Random Forest modeling. The reason for this decision was that, in a human-centered machine learning approach it made sense to evaluate one model, and its potential use and adoption by users, before investing in including all the other models. Nonetheless, the visualizations included in

the tool to explain it are not specific to RF model and could be equally useful with many other ML models. Participants expressed great interest and rated Explain-ML as very relevant and useful, which indicate that it would be valuable to continue this research and the development of the tool.

The goal to continue developing Explain-ML as a **model agnostic** tool was also reinforced by the (spontaneous) suggestions made by many of the participants to include other classification models in the tool. Furthermore, throughout the evaluation participants made other relevant suggestions on how to improve not only the explanations – suggesting more information in some of the visualizations, or even new visualizations, but also on the interactive process, commenting on aspects that would be interesting to allow users to tailor to their own needs. Moving forward with our research and the development of Explain-ML, all of them will be carefully considered.

Finally, participants praised the focus on the tool’s interactivity – not requiring any coding in any step of the process, and usability. They pointed out that these decisions would benefit not only the intended knowledgeable user, but could also allow for a broader use of Explain-ML.

6.2 Conclusions and Future Work

We have presented the design, implementation and evaluation of Explain-ML, an Interactive Visual Tool for Explainable Machine Learning that follows the principles of Human-Centered Machine Learning. Explain-ML strong focus on the final user’s perceptions and needs has resulted in a interactive and integrative holistic design aimed at providing the user with an ample exploration space for interpretability questions. This space can be flexibly exploited considering the users’ own goals, background and expectations. Our qualitative evaluation also demonstrated that we have achieved the vast majority of the goals we have set for our tool, as ultimately demonstrated by the users’ enthusiasm in using Explain-ML in their own daily activities.

We believe that we have advanced the state-of-the-art in Explainable Learning by putting the user into the center of the stage, from the inception, to the design to the evaluation of the tool. This is perhaps better demonstrated when we allow the knowledge acquired by the users using our complementary set of visual explanations (including the historical views) to feed back the models in a virtuous cycle of “understand-by-doing”.

Future tasks involve taking into the account several suggestions often mentioned in the experiments and implement new features in the Explain-ML, such as:

- Indicate the metric used for calculating the feature importance;
- Inclusion of visualizations for statistical significance tests;
- Inclusion of visualization of features importance per class, with word cloud;
- Add more automatic strategies to feature selection, like:
 - Eliminate the top X words
 - Eliminate a particular percentage of the less important features and
 - Eliminate features with importance equals zero
- Inclusion of other ML models, taking into consideration their specificities (hyperparameters, specific visualizations);
- Determine a suitable way to display the generated trees or a part of them (when using Random Forests) - that was already planned, but were left for a second moment;
- Add an appropriate way of viewing the combinations/interactions between the features and their importance - that was already planned, but were left for a second moment;
- Show metrics variation, by fold;
- Display the number of words in each instance;
- Inclusion of automatic class balance strategies;
- Display the time taken to perform the model training;
- Display a comparative of the confusion matrix (improve or worsen) related with the previous execution;
- Display indicator of the amount of values to be selected within the chosen range for tuning;
- Display the estimated time to be taken for training the model;
- Add the standard error for each metric, to indicate the variations between folds;
- Enable selection of desired evaluation metrics to be displayed in the views;
- Display metrics by class;
- Add a tab with a terminal, so that the user can make more actions that the tool does not provide.

Explain-ML's ease-of-use, with no need for any coding, along with its appreciated aesthetics, were highlighted several times as major positive assets. This means that a natural venue for investigation is experimenting the tool with less 'knowledgeable' users in real-world ML projects and different tasks. Another venue for investigation is a direct comparative (and quantitative) study between Explain-ML and other tools with similar goals, such as LIME, SHAP and Lemonade.

The evaluation allowed us to obtain relevant indicators on the explainability proposal, through the various views; on the holistic view; integration and interdependence of the views. The evaluation also generated interesting suggestions about the interface and interaction. Some of them were already foreseen, but they were left for a second moment, because we are doing a user-centered process, so we wanted to evaluate the main ideas in this version.

References

- [1] Julius Adebayo and Lalana Kagal. Iterative orthogonal feature projection for diagnosing bias in black-box models. *arXiv preprint arXiv:1611.04967*, 2016.
- [2] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.
- [3] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.
- [4] Leo Breiman et al. Statistical modeling: The two cultures. *Statistical science*, 16(3):199–231, 2001.
- [5] J.M Carroll. Introduction to this Special Issue on “Scenario-Based System Development”. *Interacting with Computers*, 13(1):41–42, 09 2000.
- [6] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 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*, pages 1721–1730. ACM, 2015.
- [7] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.
- [8] Paulo Cortez and Mark J Embrechts. Opening black box data mining models using sensitivity analysis. In *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 341–348. IEEE, 2011.
- [9] Çağatay Demiralp. Clustrophile: A tool for visual clustering analysis. In *KDD 2016 Workshop on Interactive Data Exploration and Analytics*, pages 37–45, 2016.
- [10] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- [11] John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):8, 2018.

- [12] Strumbelj Erik and Igor Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11(Jan):1–18, 2010.
- [13] Jerry Alan Fails and Dan R Olsen Jr. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces*, pages 39–45. ACM, 2003.
- [14] Rebecca Fiebrink and Marco Gillies. Introduction to the special issue on human-centered machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):7, 2018.
- [15] Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, et al. Human-centred machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 3558–3565. ACM, 2016.
- [16] Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. of Comput. and Graphical Statistics*, 24(1):44–65, 2015.
- [17] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. *AI Magazine*, 38(3):50–57, 2017.
- [18] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *arXiv preprint arXiv:1805.10820*, 2018.
- [19] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2018.
- [20] Patrick Hall. *An introduction to machine learning interpretability*. O’Reilly Media, Incorporated, 2019.
- [21] Patrick Hall and Navdeep Gill. *Introduction to Machine Learning Interpretability*. O’Reilly Media, Incorporated, 2018.
- [22] Qi Han, Weimeng Zhu, Florian Heimerl, Steffen Koch, and Thomas Ertl. A visual approach for interactive co-training. In *KDD 2016 Workshop on Interactive Data Exploration and Analytics*, pages 46–52, 2016.
- [23] Giles Hooker. Discovering additive structure in black box functions. In *Proc. of ACM SIGKDD*, pages 575–580. ACM, 2004.

- [24] Josua Krause, Adam Perer, and Enrico Bertini. Using visual analytics to interpret predictive machine learning models. In *Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, 2016.
- [25] Josua Krause, Adam Perer, and Enrico Bertini. A user study on the effect of aggregating explanations for interpreting machine learning models. In *ACM KDD Workshop on Interactive Data Exploration and Analytics*, 2018.
- [26] Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5686–5697. ACM, 2016.
- [27] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*, 2017.
- [28] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.
- [29] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.
- [30] Sabine Madsen and Lene Nielsen. Exploring persona-scenarios - using storytelling to create design ideas. In Dinesh Katre, Rikke Orngreen, Pradeep Yammiyavar, and Torkil Clemmesen, editors, *Human Work Interaction Design*, pages 57–66, Germany, 2010. Springer Science+Business Media.
- [31] Christoph Molnar. A guide for making black box models explainable. URL: <https://christophm.github.io/interpretable-ml-book>, 2018.
- [32] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.
- [33] Jennifer Preece, Helen Sharp, and Yvonne Rogers. *Interaction design: beyond human-computer interaction*, page 408. John Wiley & Sons, 2019.
- [34] Rogers Yvonne Preece, Jenny and Helen Sharp. *Interaction Design: Beyond Human - Computer Interaction*. Wiley Publishing, 5th edition, 2019.
- [35] Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. Emerging perspectives in human-centered machine learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, page W11. ACM, 2019.

- [36] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*, pages 1135–1144, 2016.
- [37] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [38] Mary Beth Rosson and John M Carroll. Scenario-based usability engineering. In *Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 413–413, 2002.
- [39] Sameer Singh, Marco Tulio Ribeiro, and Carlos Guestrin. Programs as black-box explanations. *arXiv preprint arXiv:1611.07579*, 2016.
- [40] Daniel Smilkov, Shan Carter, D Sculley, Fernanda B Viégas, and Martin Wattenberg. Direct-manipulation visualization of deep networks. In *KDD 2016 Workshop on Interactive Data Exploration and Analytics*, pages 115–119, 2016.
- [41] Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. Interpreting black-box classifiers using instance-level visual explanations. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, page 6. ACM, 2017.
- [42] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD*, pages 465–474, 2017.
- [43] Ryan Turner. A model explanation system. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016.
- [44] Marina M-C Vidovic, Nico Görnitz, Klaus-Robert Müller, and Marius Kloft. Feature importance measure for non-linear learning algorithms. *arXiv preprint arXiv:1611.07567*, 2016.
- [45] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J Smith, Kalyan Veeramachaneni, and Huamin Qu. Atmseer: Increasing transparency and controllability in automated machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 681. ACM, 2019.

Appendix A

Interview Consent Form

A.1 English

Consent Form

This interview is being conducted online through an instant messaging system. All communication will be recorded for research purposes.

Introduction: This Consent Form contains information about your participation in the 'Random Forest Results Explanability' survey as a voluntary participant in the interview stage. To ensure that you are informed about your participation in this survey, we ask that you read this Consent Form. If you have any questions, do not hesitate to ask us or contact us through the email addresses indicated. You must also sign this document and will receive a copy of it.

Objective: The objective of this interview is to raise requirements to explain the results of Random Forest algorithms, in order to assess the reliability of the machine learning model.

General information about the research: You will be asked about your experience with Random Forest algorithms, interpretation of results and parameterization. The interview audio will be recorded for further analysis by the researchers.

Use of collected data: The data collected here will be used in Bárbara Gabrielle C.O. Lopes's dissertation, student on the Computer Science Department, in UFMG, under the guidance of professors Raquel O. Prates and Marcos Gonçalves, and for possible scientific publications. Any data used for publication will be presented in order to guarantee the anonymity of the interview participants. Your name or any information that may identify you will be anonymized. We guarantee the confidentiality of these records, committing ourselves to keep the files in our custody for possible future work, or the minimum period of 5 (five) years.

Privacy: Information that can identify research participants will not be released. Your name will not appear in any report resulting from this study. If you wish, you can request a copy of the data generated (audio and transcript).

If you decide not to participate in the survey: You are free to decide, at any time, whether you wish to participate in this survey or not. Your decision will not affect your academic life (if you are a UFMG student) or any relationship with the interviewers, teachers or the institution behind it. Your refusal will not incur any penalty.

Payment: Participants will not receive payment to participate in this research, since it must be volunteer.

If you have a problem or have other questions: If you have a problem that you think may be related to your participation in this research, or if you have any questions about it, you can contact the researchers at any time, during your participation or later, by the emails indicated in the header of this term.

New conditions: If you wish, you can specify new conditions in order to enable your participation in this evaluation.

A.2 Portuguese

Termo de Consentimento Livre e Esclarecido

Esta entrevista está sendo realizada on-line por meio de um sistema de mensagem instantânea. Toda a comunicação será gravada para fins de pesquisa.

Introdução: Este Termo de Consentimento contém informações sobre a sua participação na pesquisa 'Explicabilidade de Resultados de *Random Forest*' como participante voluntário da etapa de entrevistas. Para assegurar que você esteja informado sobre a sua participação nesta pesquisa, pedimos que leia este Termo de Consentimento. Caso tenha alguma dúvida, não hesite em perguntar aos pesquisadores responsáveis ou contatá-los através dos endereços de email indicados acima. Você também deverá assinar o termo do qual receberá uma cópia.

Objetivo: O objetivo dessa entrevista é levantar requisitos para explicação dos resultados de algoritmos Random Forest, de forma a aferir a confiabilidade do modelo de aprendizagem.

Informação geral sobre a pesquisa: Você será questionado sobre sua experiência com algoritmos Random Forest, interpretação dos resultados e parametrização dos mesmos. Essa entrevista terá o áudio gravado para posterior análise dos pesquisadores.

Utilização dos dados coletados: Os dados aqui coletados serão utilizados para a dissertação da aluna Bárbara Gabrielle C.O. Lopes, do Departamento de Ciência da Computação, UFMG, sob orientação dos professores Raquel O. Prates e Marcos Gonçalves, e para eventuais publicações científicas. Quaisquer dados utilizados para publicação serão apresentados de forma a garantir o anonimato dos participantes da entrevista. Seu nome

ou quaisquer informações que possam lhe identificar serão anonimizados. Garantimos a confidencialidade desses registros, comprometendo-nos a manter os arquivos sob nossa guarda para eventuais trabalhos futuros, em um prazo de no mínimo 5 (cinco) anos.

Privacidade: Informações que possam identificar os participantes da pesquisa não serão divulgadas. O seu nome não constará em nenhum relatório resultante deste estudo. Caso deseje, poderá solicitar uma cópia dos dados gerados por você (áudio e transcrição).

Se você decidir não participar na pesquisa: Você é livre para decidir, a qualquer momento, se deseja participar ou não desta pesquisa. Sua decisão não afetará sua vida estudantil (caso seja estudante da UFMG) e nem qualquer relacionamento com os entrevistadores, professores ou a instituição por trás desta. Sua recusa não acarretará qualquer penalidade.

Compensação: A participação nesta pesquisa é voluntária, e não será oferecida nenhuma remuneração aos seus participantes.

Se tiver algum problema ou se tiver outras perguntas: Se você tiver algum problema que pensa que pode estar relacionado com sua participação nesta pesquisa, ou se tiver qualquer pergunta sobre a pesquisa, poderá entrar em contato com os pesquisadores a qualquer momento, durante a sua participação ou depois, pelos e-mails indicados no cabeçalho deste termo.

Novas condições: Caso deseje, você pode especificar novas condições que devem ser atendidas de forma a viabilizar sua participação nesta avaliação.

Appendix B

Interview Script

B.1 English

Table B.1: Interview Script (English)

Block	Main Items
Beginning of the interview	• Research presentation. • Consent form (leave a copy with the participant). • Remind the participant that there are no wrong answers, that we will not make any value judgment on this data. • Record audio (check if it is recording).
I. Profile	Name, Age, Profession, Education

II. Contextualization	Let's start by contextualizing the participant regarding the purpose of the interview and the scope of the research. • Explanability in ML systems – Definition. – Importance (Reliability of results). – Examples. • Random Forest. • Text Classification.
III. Previous experience with Random Forest algorithms	Here, we will ask about the participant's experience with Machine Learning algorithms, and especially with Random Forest algorithms... • Have you used Random Forests? If the participant has already used RF: – In which context? To solve which problem? – Did you use RF as a black box? – What programming language/tools did you use (for RF)? • Have you used other Machine Learning algorithms? Which ones?

IV. Random Forest specifics	If the participant has already used RF: Here we are going to ask about more specific RF characteristics, in order to raise requirements for its explainability. • How to identify a reliable result? What metrics do you use? • How do you measure whether the RF performed well in relation to the metrics used? What values do you use as a reference and where do you get them from? • How is the RF parameterization set? Do you use acquired knowledge from colleagues, literature or other? • How do you look for help to answer questions about parameterization? • Do you imagine any resource that could be used to help a less experienced user to define this parameterization? • Quais você considera os parâmetros mais sensíveis de RF (Parâmetros que se alterados, modificam muito o resultado)? • Importance of visualizing the tree's structure.
------------------------------------	--

V. About LIME	• Introduce Lime. – Explain tool. – Show examples of Explanations generated by it. – Scope (generality, Machine Learning algorithms for which it is applicable). – Use (programming language, how it can be used, scikit learn). • Do you believe that an explanation of ML results could help you make a decision about the learning model? • Do you consider the explanation provided by LIME sufficient to assess confidence in the RF results? • Would you add any information to this explanation (specific to RF or general to ML)?
VI. Considerations	Do you have any other comments regarding what was discussed? • About LIME. • About other tools / techniques known for explainability. • About explicabilidade. • About Random Forest.
End of the interview	• Thank the participant. • Check audio recording. • Leave a copy of the consent form with the participant.

B.2 Portuguese

Table B.2: Roteiro da Entrevista (Português)

Bloco	Principais Itens
Início da entrevista	• Apresentação da pesquisa. • Assinatura do termo de consentimento (deixar uma cópia também com o participante). • Lembrar que não existem respostas erradas, que não iremos fazer nenhum julgamento de valor com esses dados. • Gravar áudio (verificar se está gravando).
I. Perfil	Nome, Idade, Profissão, Formação acadêmica (área).
II. Contextualização	Vamos começar contextualizando o entrevistado a respeito do objetivo da entrevista e escopo do trabalho. • Explicabilidade em sistemas de ML – Definição. – Importância (Confiabilidade dos resultados). – Exemplos. • Random Forest. • Classificação de textos.

III. Experiência com algoritmos de Random Forest	Aqui, vamos perguntar sobre a experiência do entrevistado com algoritmos de Machine Learning, e, principalmente, com algoritmos de Random Forest. . . • Já utilizou o Random Forest? Se o participante já trabalhou com RF: – Em qual contexto, qual problema? – Utilizou RF como caixa preta? – Quais linguagens de programação/ferramentas utilizadas (para RF)? • Já utilizou outros algoritmos de Machine Learning? quais?
--	--

IV. Especificidades da Random Forest	Se o participante já trabalhou com RF: Aqui vamos perguntar sobre características mais específicas de RF, em busca de levantar requisitos para sua explicabilidade. • Como identificar um resultado confiável? Quais as métricas utilizadas? • Como você analisa se a RF teve um bom desempenho em relação as métricas usadas? Que valores usa como referência e de onde você os extrai? • Como é setada a parametrização da RF? Utiliza conhecimento de colegas, literatura ou outro? • Como busca ajuda para sanar as dúvidas sobre parametrização? • Você imagina algum artifício que poderia ser usado para auxiliar um usuário menos experiente a definir essa parametrização? • Quais você considera os parâmetros mais sensíveis de RF (Parâmetros que se alterados, modificam muito o resultado)? • Relevância da visualização da estrutura das árvores.
--------------------------------------	--

V. Sobre o LIME	• Apresentação do Lime. – Explicação da ferramenta. – Exibir exemplos de Explicações geradas por ele. – Escopo (generalidade, algoritmos de Machine Learning para os quais é aplicável). – Utilização (linguagem de programação, como pode ser usado, scikit learn). • Você acredita que uma explicação dos resultados de ML poderia te auxiliar a tomar decisão a respeito do modelo de aprendizagem? • Você considera a explicação fornecida pelo LIME suficiente para aferir confiança nos resultados de RF? • Você acrescentaria alguma informação nesta explicação (específica para RF ou geral para ML)?
VI. Demais considerações	Você tem algum outro comentário em relação ao que foi conversado? • Sobre o LIME. • Sobre outras ferramentas/técnicas conhecidas para explicabilidade. • Sobre explicabilidade. • Sobre Random Forest.
Finalização da entrevista	• Agradecer. • Conferir gravação do áudio. • Deixar uma cópia do termo de consentimento com o participante.

Appendix C

Prototyping

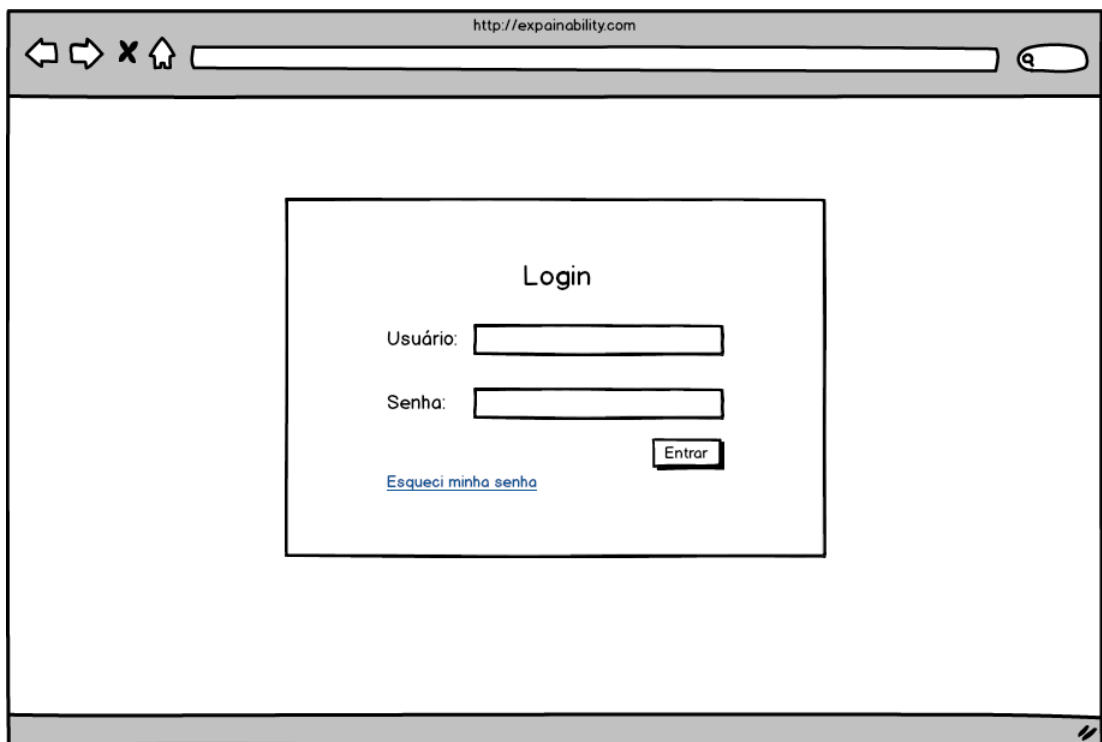


Figure C.1: Prototype: Login.

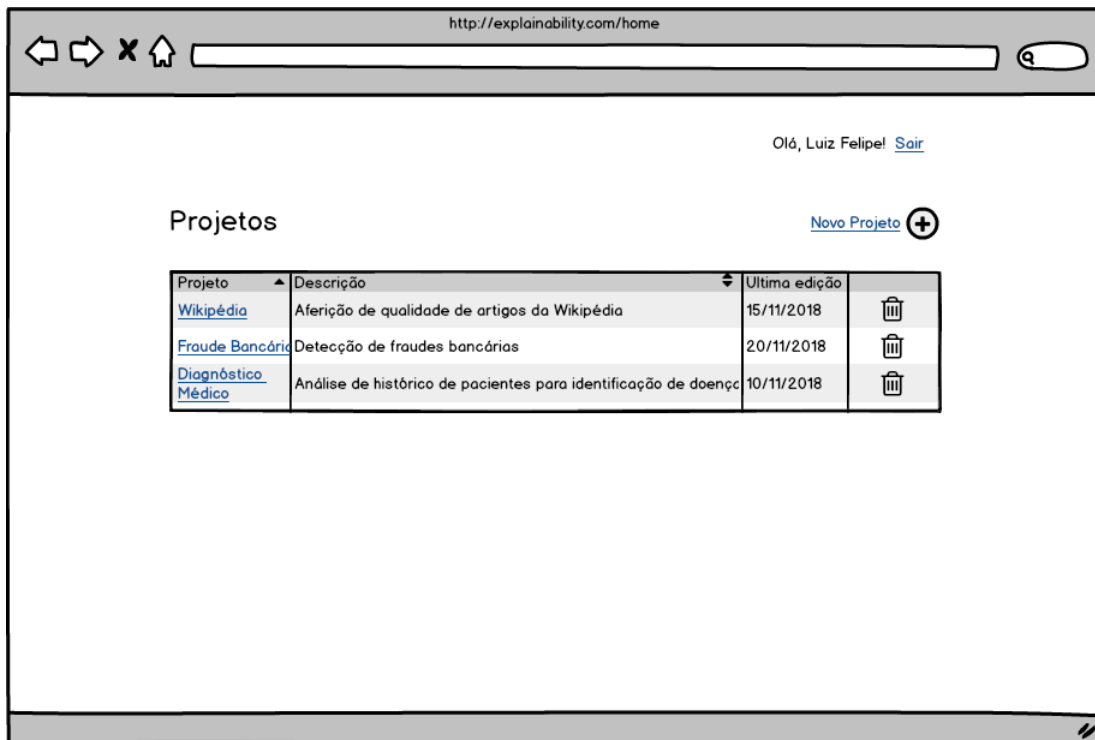


Figure C.2: Prototype: Projects.

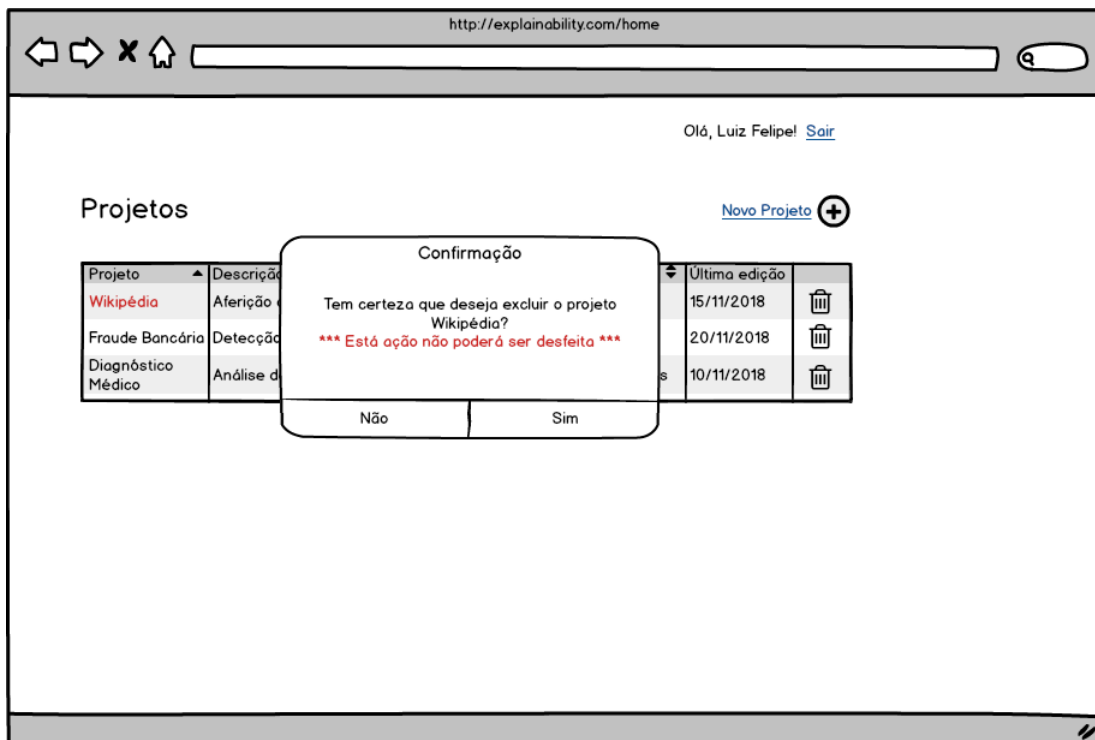


Figure C.3: Prototype: Delete Project.

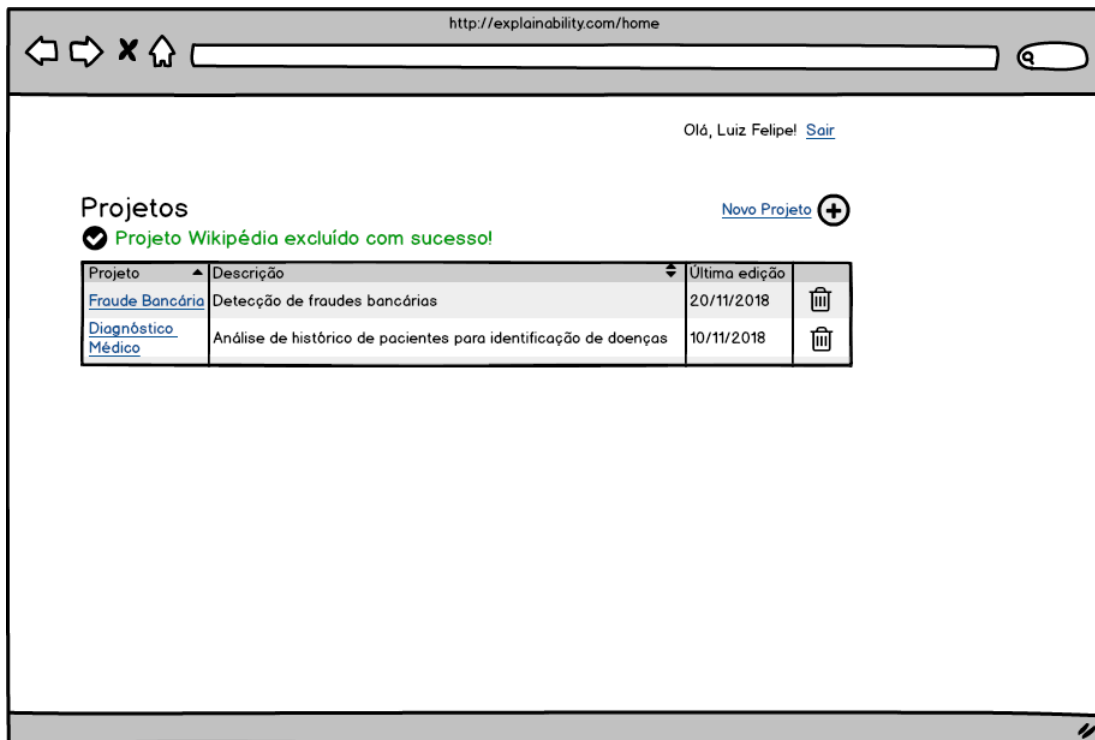
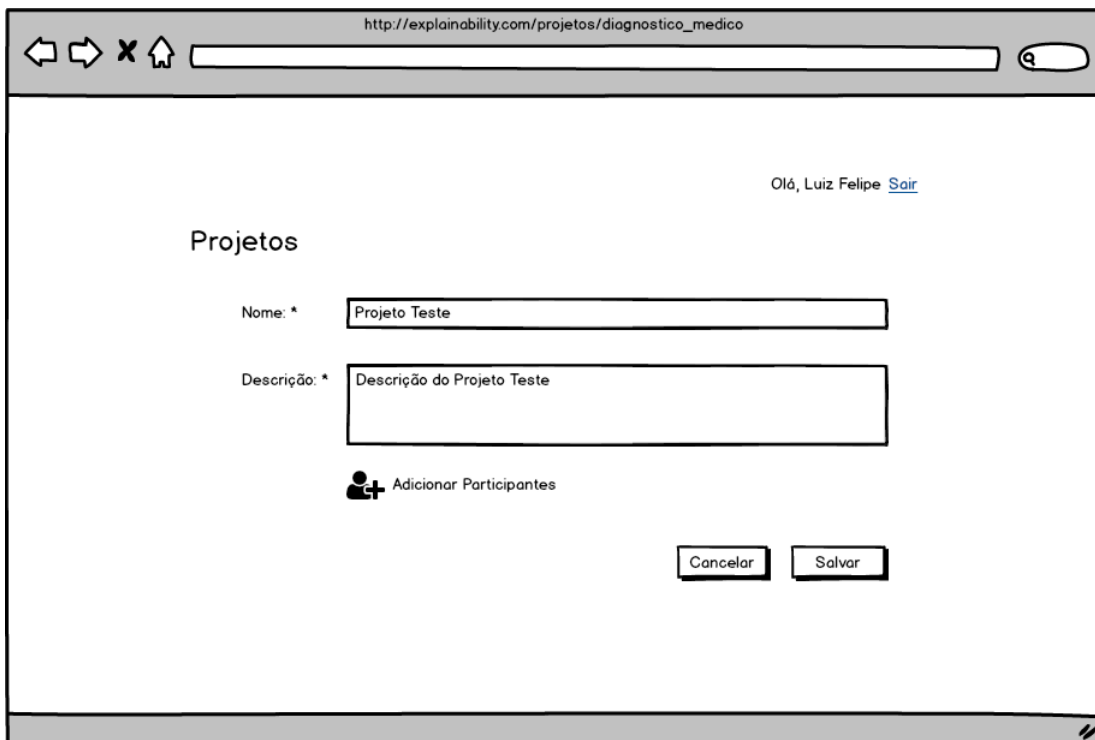


Figure C.4: Prototype: Deleted Project.

Figure C.5: Prototype: Save Project. *New Project*

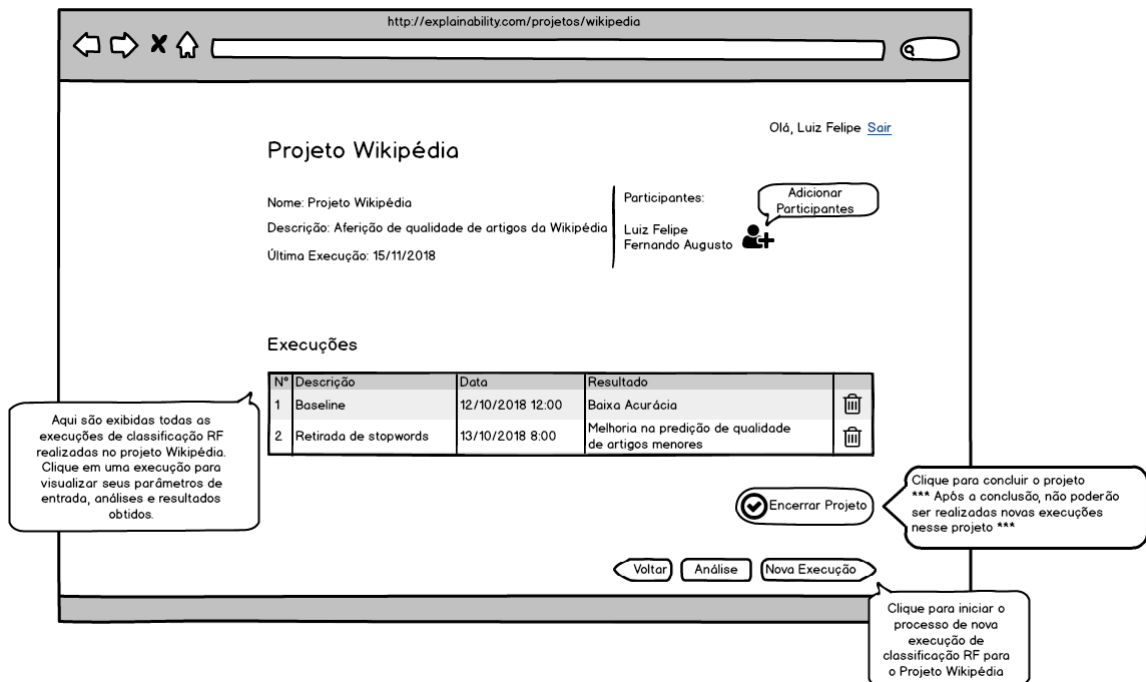
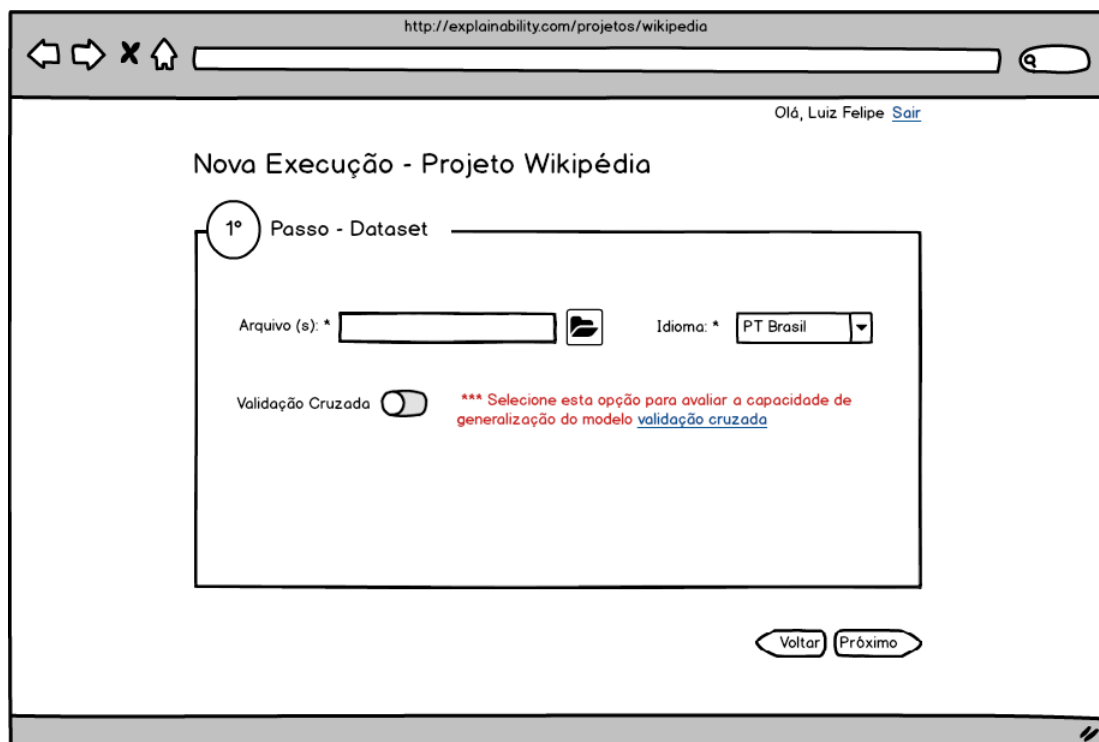


Figure C.6: Prototype: Project.

Figure C.7: Prototype: Step 1.1. *New Execution*

http://explainability.com/projetos/wikipedia

Olá, Luiz Felipe [Sair](#)

Nova Execução - Projeto Wikipédia

1º Passo - Dataset

Arquivo (s): *

Idioma: *

Validação Cruzada ☒ *** Selecione esta opção para avaliar a capacidade de generalização do modelo

Quantidade de folds para realização da validação cruzada

Nº Folds:

Divisão Automática ☒ Desmarque esta opção caso o dataset inserido já esteja dividido em folds.

[Voltar](#) [Próximo](#)

Figure C.8: Prototype: Step 1.2. *New Execution*

http://explainability.com/projetos/wikipedia

Olá, Luiz Felipe [Sair](#)

Nova Execução - Projeto Wikipédia

2º Passo - Processamento

Retirar stop words ☒ Insira sua própria lista de stopwords
*** Selecione esta opção para retirar stopwords dos textos [stopwords](#)

Extrair radicais ☒ Insira sua própria lista de radicais
*** Selecione esta opção para extrair os radicais das palavras dos textos [stemmer](#)

Usar TFIDF ☒ *** Selecione esta opção para selecionar as palavras por ordem de relevância [TFIDF](#)

[Voltar](#) [Próximo](#)

Figure C.9: Prototype: Step 2. *New Execution*

The screenshot shows a web browser window with the address bar displaying `http://explainability.com/projetos/wikipedia`. The page title is "Nova Execução - Projeto Wikipédia". In the top right corner, there is a greeting "Olá, Luiz Felipe" and a link "Sair". The main content area is titled "3º Passo - Parametrização" and contains the following elements:

- Tunnar Parâmetros:** A toggle switch is currently turned on. To its right, red text reads: "*** Selecione esta opção para que sejam buscados os melhores valores de parâmetros para o modelo".
- Método de Tuning:** A dropdown menu with "Grid Search" selected. Below it, "Random Search" is listed as an option.
- Nível de Aleatoriedade:** A horizontal slider with a black dot in the middle. To its right, red text reads: "*** Valores mais altos exigirão maior número de árvores para poder modelar todas as relações e são computacionalmente mais caras, resultando em melhoria dos resultados".

At the bottom right of the form, there are two buttons: "Voltar" and "Próximo".

Figure C.10: Prototype: Step 3.1. *New Execution*

http://explainability.com/projetos/wikipedia

Olá, Luiz Felipe [Sair](#)

Nova Execução - Projeto Wikipédia

3º Passo - Parametrização

Nº de árvores: * *** Quanto maior, melhor o resultado e maior a perda de processamento. Note que às vezes a melhoria no resultado não compensa a perda de processamento

Critério: *
 *** Selecione a função que medirá a qualidade do split

Profundidade máxima da árvore: * *** Quanto maior, maior a complexidade do modelo e maior overfit de cada árvore nos dados

Mínimo de features para split dos nós: *

Mínimo de features para folha: *

Máximo de features: *

Quantidade de jobs: * *** Número de jobs que se deseja rodar em paralelo

Figure C.11: Prototype: Step 3.2. *New Execution*

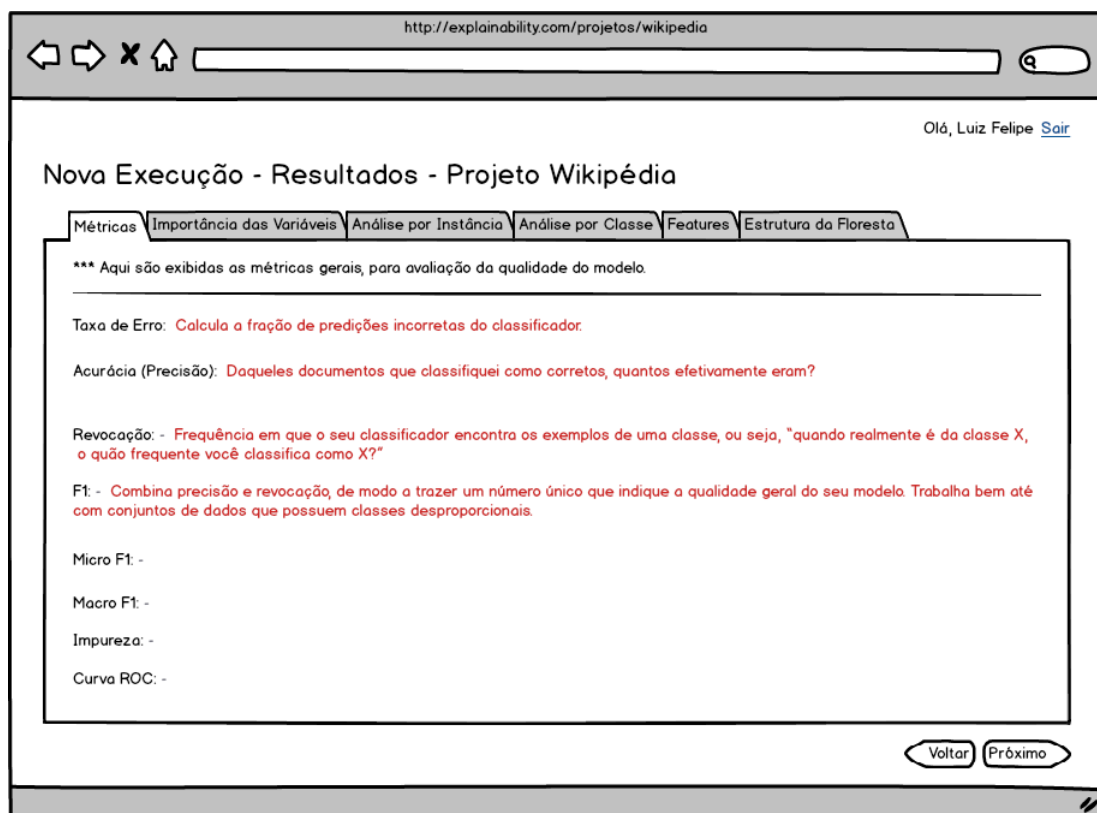


Figure C.12: Prototype: Metrics. *Results for the New Execution*

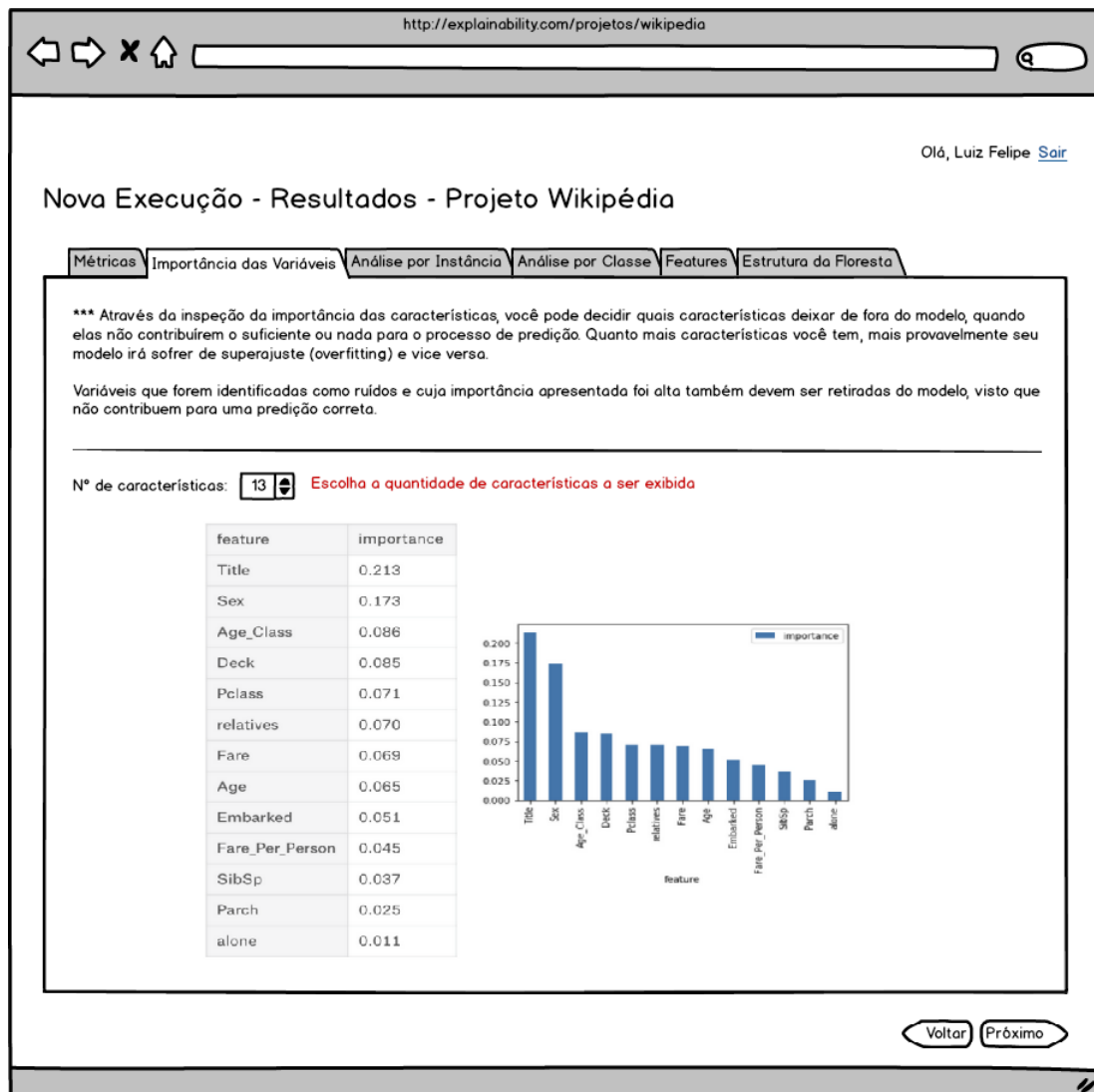


Figure C.13: Prototype: Feature Importance. *Results for the New Execution*

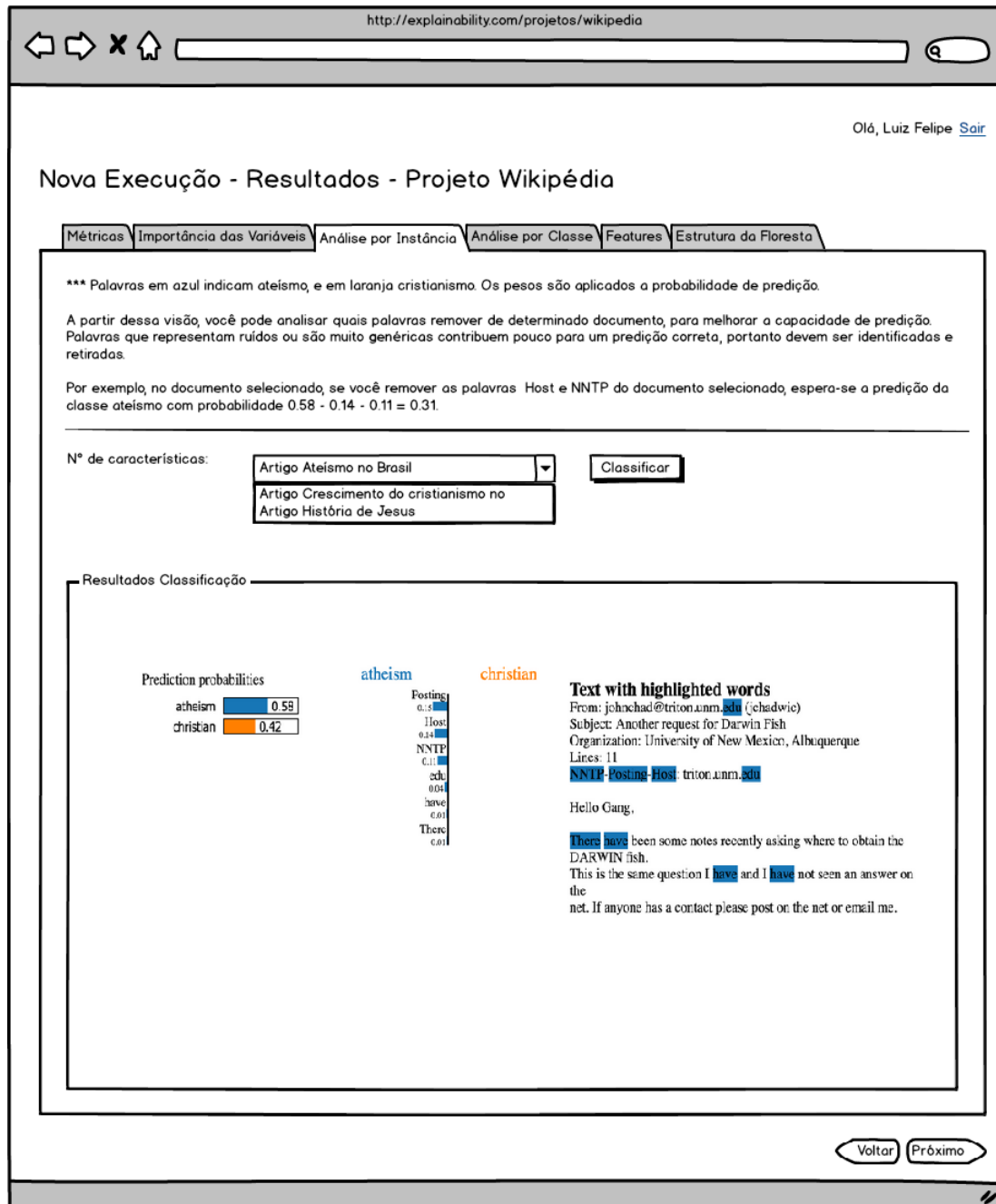


Figure C.14: Prototype: Analysis by Instance. *Results for the New Execution*

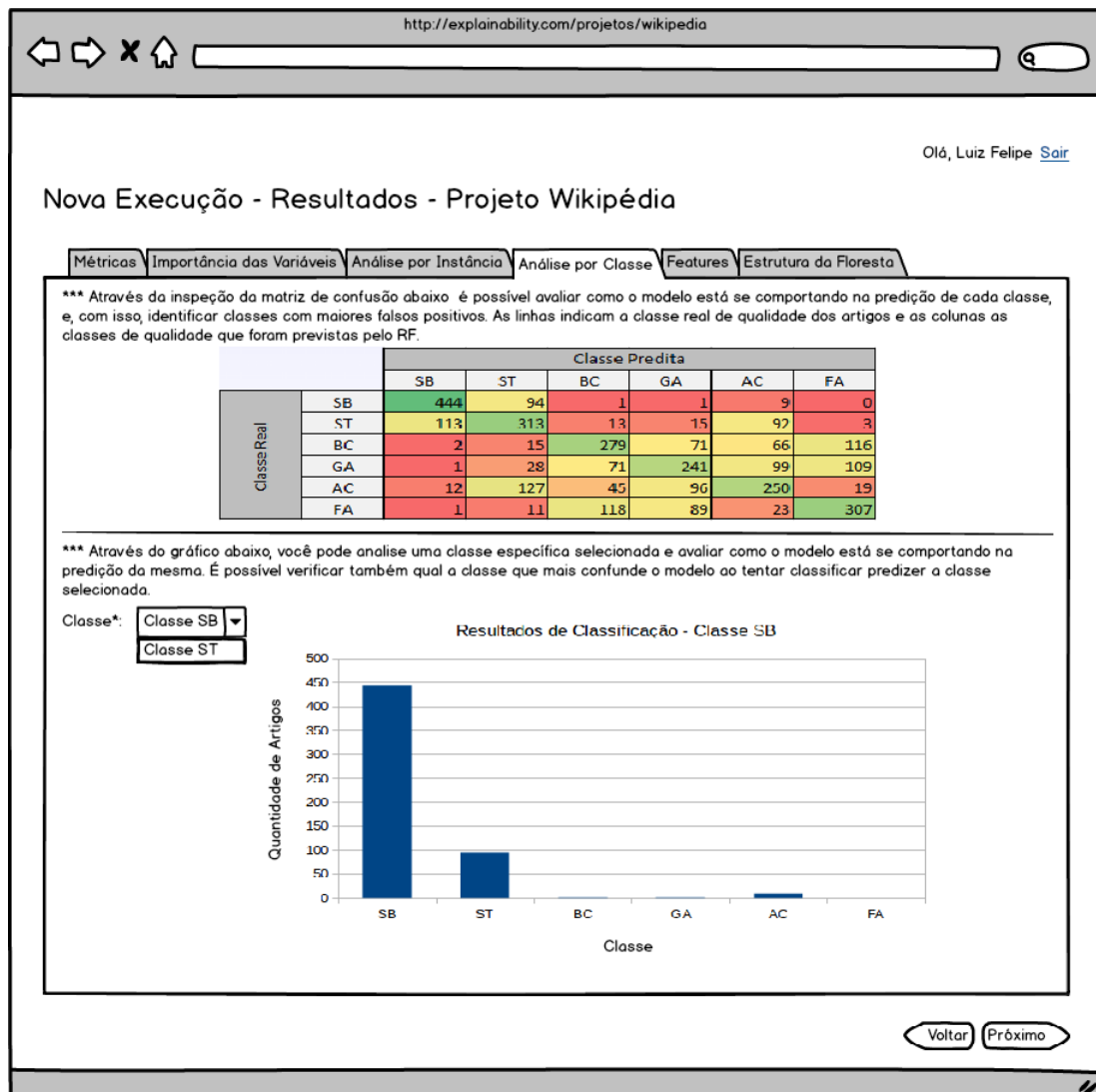


Figure C.15: Prototype: Analysis by Class. *Results for the New Execution*

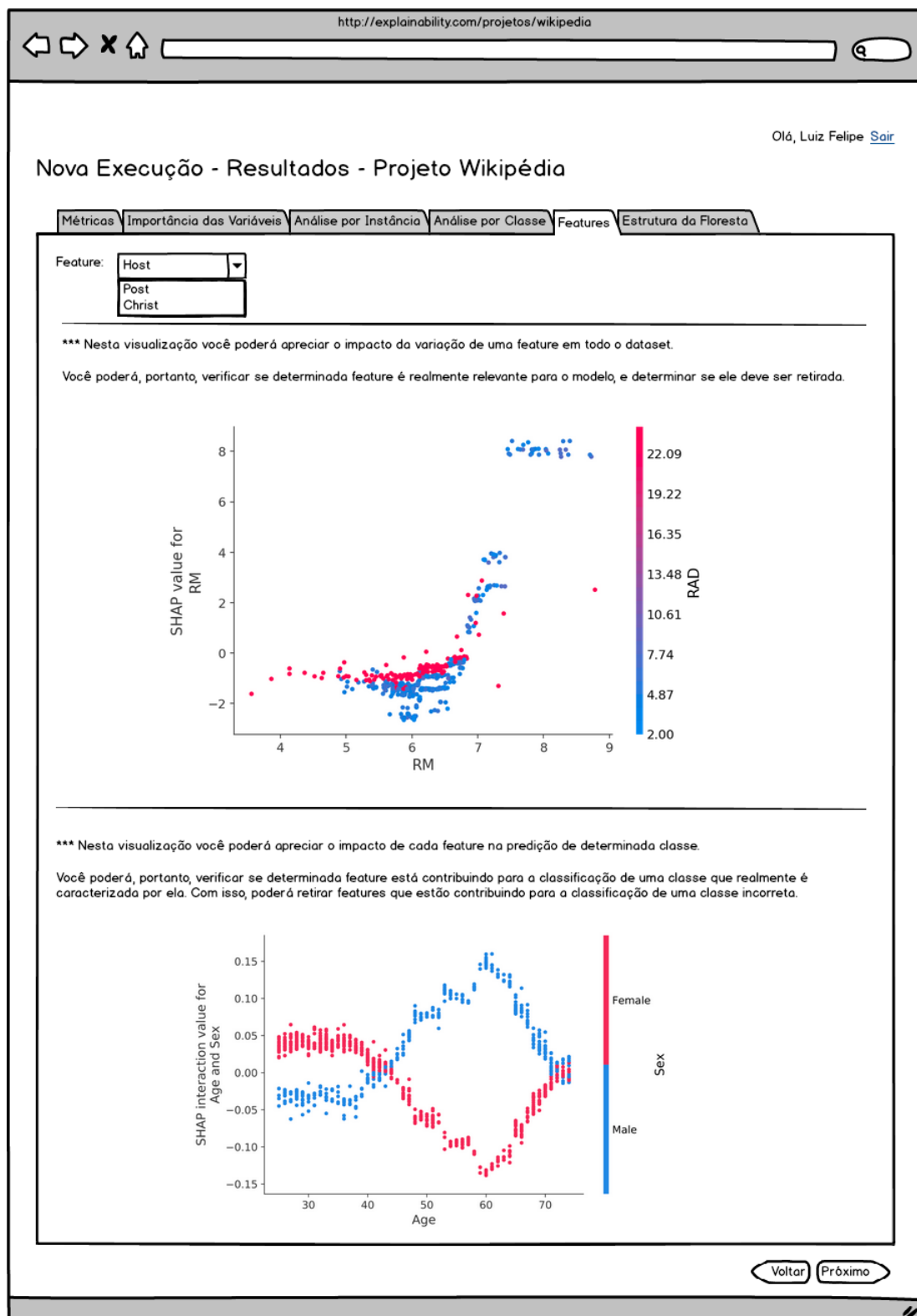


Figure C.16: Prototype: Features. *Results for the New Execution*

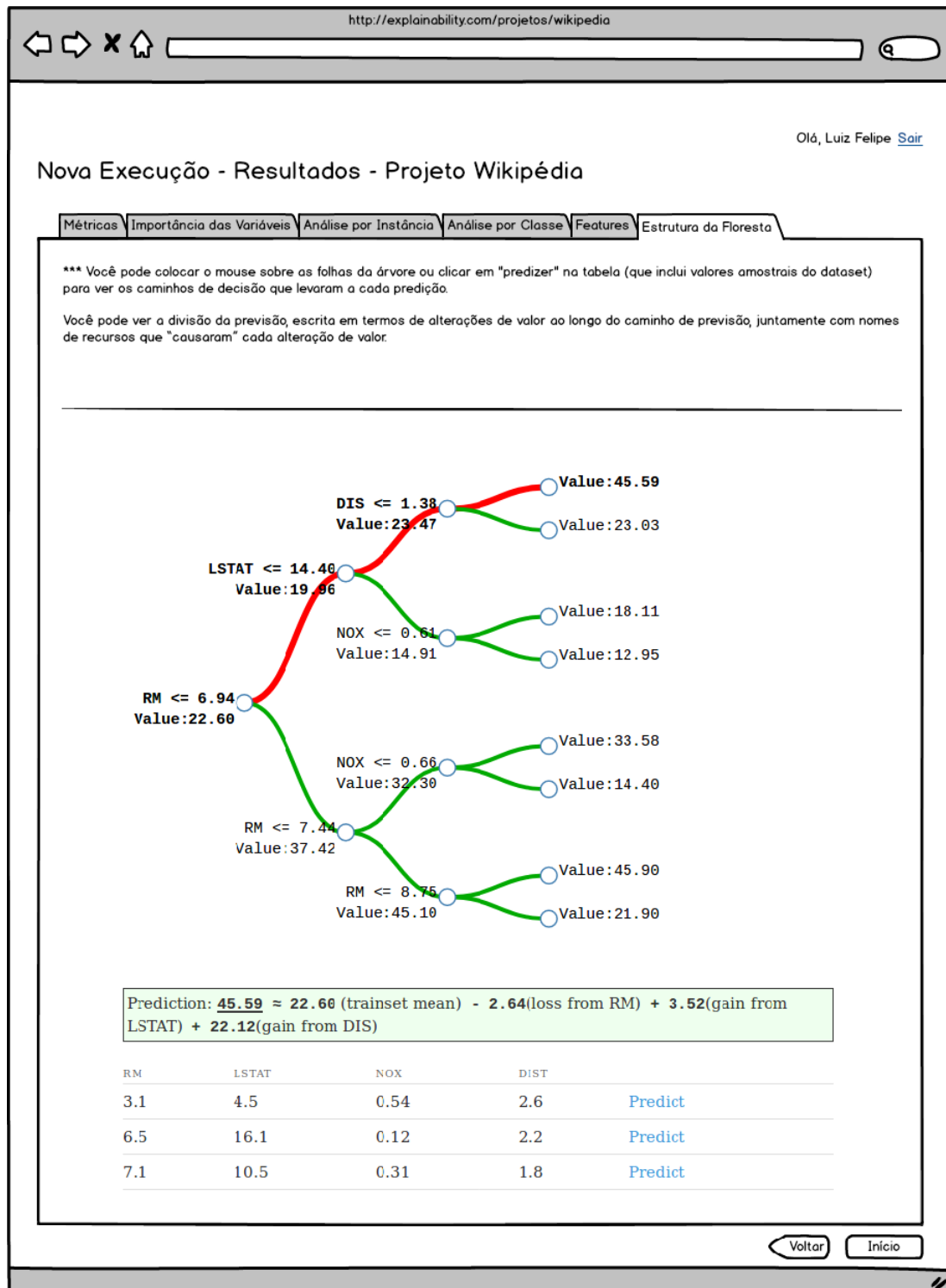


Figure C.17: Prototype: Forest Structure. *Results for the New Execution*

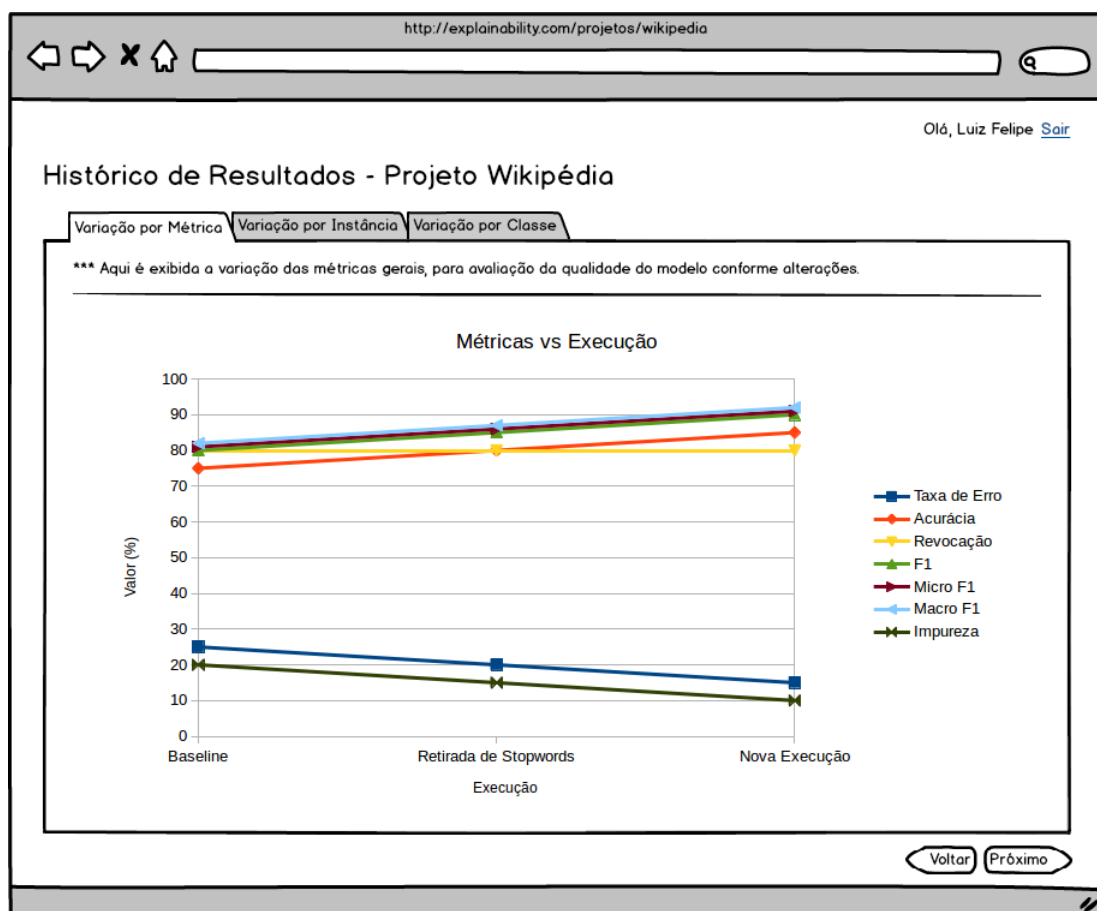


Figure C.18: Prototype: Metric Variation. *Execution History*

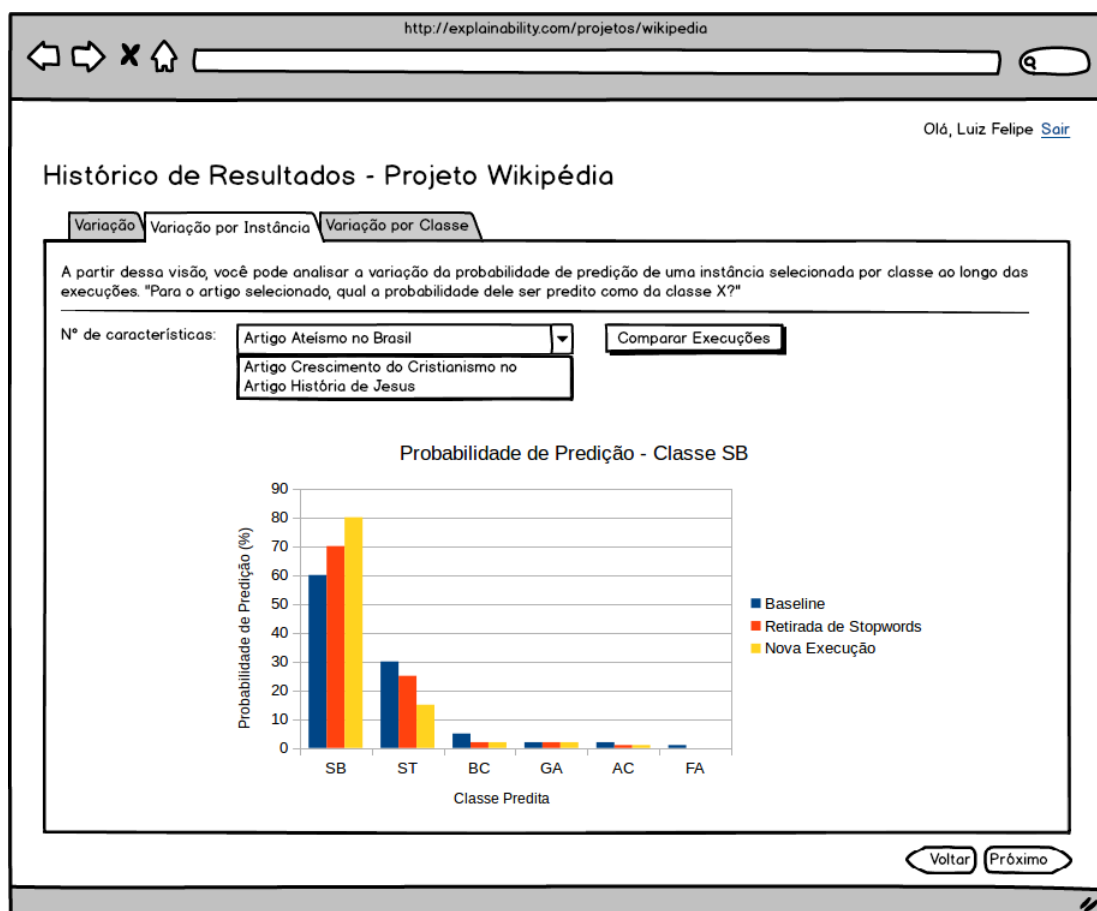


Figure C.19: Prototype: Instance Variation. *Execution History*

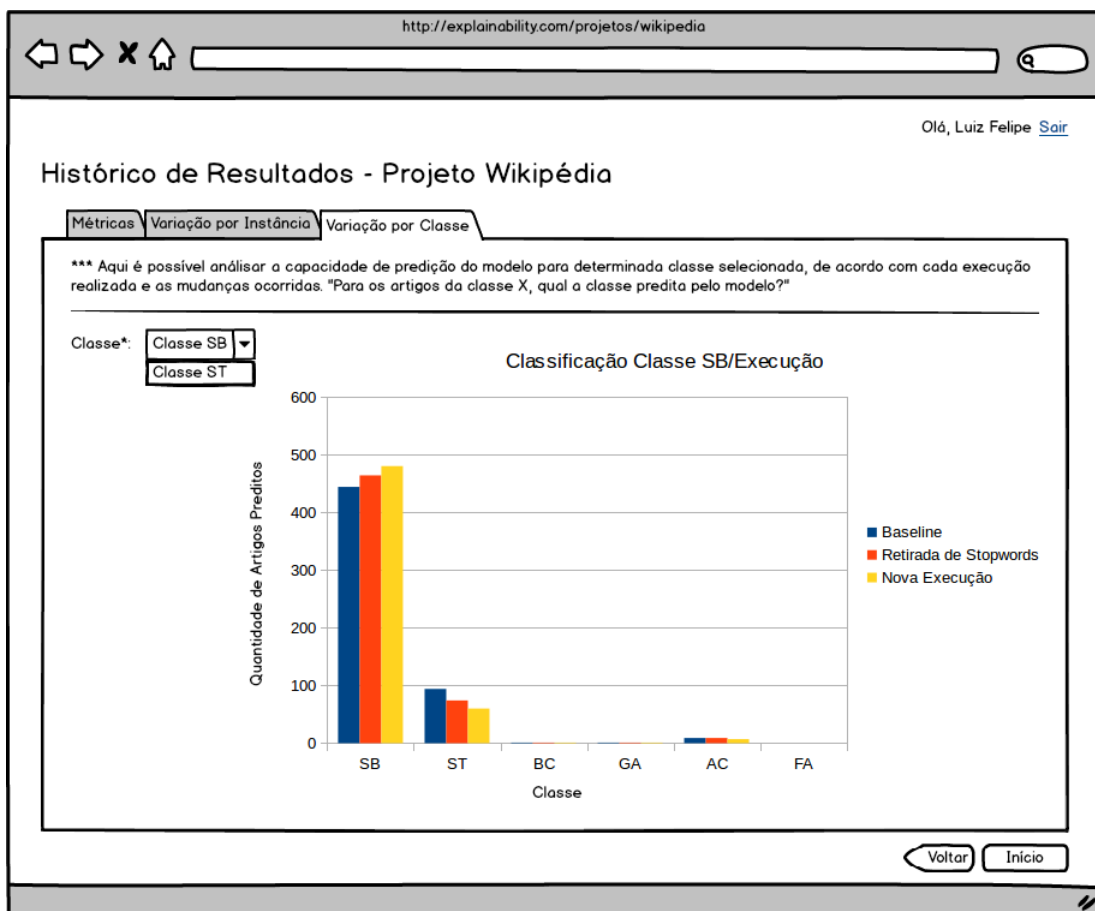


Figure C.20: Prototype: Class Variation. *Execution History*

Appendix D

Prototype Evaluation Script

D.1 English

Table D.1: Prototype Evaluation Script (English)

Task identifier	Task Description	Comments
1 Login	Log in to the system	No need to fill in user data.
2 New Project	Create a new project	Just browse the screens, there is no need to fill in textual data.
3 Wikipedia Project	Access the Wikipedia Project	
4 New Execution	Create a new execution for the Wikipedia project	Just browse the screens, there is no need to fill in textual data.
5 Results	Analyze the results of the new execution of the Wikipedia project	
6 History	Analyze the results history for the Wikipedia project	

D.2 Portuguese

Table D.2: Roteiro da Avaliação de Protótipo (Português)

Identificador da tarefa	Descrição da tarefa	Comentários
1 Login	Logar no sistema	Não há necessidade de preencher os dados do usuário
2 Novo Projeto	Criar um novo projeto no sistema	Apenas navegue nas telas, não é necessário preencher dados textuais
3 Projeto Wikipédia	Acessar o projeto Wikipédia	
4 Nova Execução	Criar uma nova execução para o projeto Wikipédia	Apenas navegue nas telas, não é necessário preencher dados textuais
5 Resultados	Analisar os resultados da nova execução do projeto Wikipédia	
6 Histórico	Analisar o histórico de resultados do projeto Wikipédia	

Appendix E

Evaluation Scenario

E.1 English

Scenario

You are participating in a research project that involves the automatic classification of UFMG web pages. For that, you will have to use Random Forest (RF) models for text classification. Since the RF model can be considered a black box, the task of interpreting its results is somewhat challenging. You heard about a new system, ExplainRF, which supports the task of explaining machine learning models using RF. So, you decide to use it to facilitate the assessment of the reliability of your machine learning model. Before creating the classification model for UFMG pages, you decide to understand how ExplainRF works. To check if ExplainRF will be useful for you, you decided to test it with the WEBKB database, since it contains university pages, it is easy to access and well known by researchers in the Machine Learning industry.

You then decide to develop a learning model for WEBKB in order to classify, as assertively as possible, the documents on the web pages according to their category. When using ExplainRF for training and analyzing your learning model, you want to determine whether it is possible to:

- Create a machine learning model for the chosen problem;
- Determine your level of confidence in your machine learning model and in the results it provides;
- Identify possible improvements to the learning model and
- Make assertive improvements in your learning model.

By reviewing the WEBKB database description, you will find information that the WEBKB database contains WWW-pages collected from computer science departments of various universities in January 1997 by the World Wide Knowledge Base (Web->Kb)

project of the CMU text learning group. The 8,282 pages were manually classified into the following categories:

- student (1641)
- faculty (1124)
- staff (137)
- department (182)
- course (930)
- project (504)
- other (3764)

The class other is a collection of pages that were not deemed the “main page” representing an instance of the previous six classes. (For example, a particular faculty member may be represented by home page, a publications list, a vitae and several research interests pages. Only the faculty member’s home page was placed in the faculty class. The publications list, vitae and research interests pages were all placed in the other category.)

E.2 Portuguese

Cenário

Você está participando de um projeto de pesquisa que envolve a classificação automática das páginas web da UFMG. Para isso, você terá que utilizar modelos Random Forest (RF) para classificação de textos. Visto que o modelo RF pode ser considerado uma caixa preta, a tarefa de interpretação dos seus resultados é um tanto quanto desafiadora. Você ficou sabendo de um novo sistema, o ExplainRF, que apoia na tarefa de explicabilidade dos modelos de aprendizagem de máquina utilizando RF. Assim, você resolve utilizá-lo para facilitar a aferição da confiabilidade do seu modelo. Antes de criar o modelo de classificação para as páginas da UFMG, você resolve entender como funciona o ExplainRF. Para verificar se o ExplainRF será útil para você, você resolveu testá-lo com a base de dados WEBKB, visto que ela contém páginas de universidade, é de fácil acesso e bastante conhecida por pesquisadores do ramo de Machine Learning.

Você resolve então desenvolver um modelo de aprendizagem para a WEBKB em vista de classificar, com a maior assertividade possível, os documentos das páginas Web

fornecidos de acordo com sua categoria. Ao utilizar o ExplainRF para treinamento e análise do seu modelo de aprendizagem, você pretende determinar se é possível:

- Criar um modelo de aprendizagem para o problema escolhido;
- Determinar seu nível de confiança no seu modelo de aprendizagem e nos resultados fornecidos por ele;
- Identificar possíveis melhorias para o modelo de aprendizagem e
- Realizar melhorias assertivas no seu modelo de aprendizagem.

Revendo a descrição da base WEBKB, você encontra a informação de que a base WEBKB contém páginas da WWW coletadas de departamentos de ciência da computação de várias universidades em janeiro de 1997 pelo projeto World Wide Knowledge Base (Web-> Kb) do grupo de aprendizado de texto da Carnegie Mellon University (CMU). As 8.282 páginas foram classificadas manualmente nas seguintes categorias:

- student (1641)
- faculty (1124)
- staff (137)
- department (182)
- course (930)
- project (504)
- other (3764)

A classe ‘other’ é uma coleção de páginas que não foram consideradas a “página principal”, representando uma instância das seis classes anteriores. (Por exemplo, um determinado membro do corpo docente pode ser representado por uma página inicial, uma lista de publicações, um currículo vitae e várias páginas de interesses de pesquisa. Somente a página inicial do membro do corpo docente foi colocada na classe ‘faculty’. A lista de publicações, as páginas de interesses do currículo vitae e de pesquisa foram todos colocados na categoria ‘other’.)

Appendix F

Evaluation Script

F.1 English

Table F.1: Evaluation Script (English)

Block	Main Items
Beginning of the interview	• Ask about the participant's experience with text classification models and Random Forest (how much he used these models, when was the last time, how he usually perform model tuning). • Explain the system (focus: model explanation) • Objective: Evaluate the system and how much the explanations/views presented help in the interpretation and fine tuning of the model. • Consent form (leave a copy with the participant). • Remind the participant that there are no wrong answers, that we will not make any value judgment on this data. • Record audio (check if it is recording). • Introduce the evaluation scenario.

I. For each System Tab	Explain the visualization to the user, if necessary, and write down the problems and difficulties of interpretation observed. Interview (During): • What did you understand about the system/visualization? • Is the visualization easy to understand? • What does this tell you about the model? • Is it necessary to understand the prediction model? What information does this visualization add to the model? • Would you change anything? Would you add any information? After viewing all the explanations: • Conclusion on the prediction model: – What did you understand about the model? – What problems did you identify? – Would you make any changes to the prediction model? Which ones? • Opinion on the views: – Were the views helpful? – Which ones? – How do visualizations help you?
-------------------------------	---

II. Prediction model improvements	• Collect indicators to support improvement • Choose 2 improvements with the greatest impact. Justify the improvements (How did you identify them?) • Run a new execution making the chosen improvements: – Was it possible to visualize the changes made? – Did the changes have any impact on the prediction model? – Was it possible to improve the prediction model? – Was it possible to increase confidence in the prediction model?
III. View History (2 executions for the chosen problem)	• Has the history helped in understanding the process and making improvements in the learning model? • What are the effects observed through changes in executions? – Was it possible to visualize the changes made? item Did the changes have any impact on the prediction model? – Was it possible to improve the prediction model? – Was it possible to increase confidence in the prediction model? • Would you make any further improvements?

Final Interview	• What did you think of the system? • Did the system/interpretation help you? • What did you find more/less interesting? • Sugestions? • Do you know other explainability systems (e.g. Lime,...)? Have you used it? • Any other comments about the system or your experience with it?
------------------------	--

F.2 Portuguese

Table F.2: Roteiro da Avaliação (Português)

Bloco	Principais Itens
Início da entrevista	• Questionar sobre experiência do participante com modelos de classificação de texto e Random Forest (quanto já usou, qual a última vez, como faz o tuning de modelos normalmente) • Explicar o sistema (foco: explicação do modelo) • Objetivo do sistema: avaliar o sistema e quanto as explicações/visualizações apresentadas ajudam na interpretação e fine tuning do modelo • Termo de consentimento (deixar uma cópia também com o participante). • Lembrar que não existem respostas erradas, que não iremos fazer nenhum julgamento de valor com esses dados. • Gravar áudio (verificar se está gravando). • Apresentação do cenário de avaliação.

I. Para cada Aba do Sistema	Explicar ao usuário sobre a visualização, caso haja necessidade, e anotar problemas e dificuldades de interpretação observados. Entrevista (Durante): • O que entendeu sobre o sistema/visualização? • A visualização é de fácil compreensão? • O que isso te diz sobre o modelo? • É algo necessário para entender o modelo de predição? Que informação esta visualização te acrescenta sobre o modelo? • Você alteraria algo? Adicionaria alguma informação? Após a visualização de todas as explicações: • Conclusão sobre o modelo de predição: – O que você entendeu sobre o modelo? – Que problemas identificou? – Realizaria alguma alteração no modelo de predição? Quais? • Opinião sobre as visualizações das explicações: – As visualizações foram úteis? – Quais? – No que te auxiliaram?
------------------------------------	---

II. Melhoria do modelo de predição	• Coletar indicadores de apoio à melhoria • Escolher 2 melhorias de maior impacto. Justificar as melhorias (Como as identificou?) • Rodar uma nova execução realizando as melhorias escolhidas: – Foi possível visualizar as alterações realizadas? – As alterações tiveram impacto no modelo de predição? – Foi possível melhorar o modelo? – Foi possível aumentar a confiança no modelo de predição?
III. Histórico de Visualizações (2 execuções para o problema escolhido)	• O histórico ajudou no entendimento do processo e realização de melhorias no modelo de aprendizagem? • Quais os efeitos observados através das alterações nas execuções? – Foi possível visualizar as alterações realizadas? – As alterações tiveram impacto no modelo de predição? – Foi possível melhorar o modelo? – Foi possível aumentar a confiança no modelo de predição? • Faria alguma melhoria adicional?

Entrevista Final	• O que achou do sistema? • O sistema/interpretação te ajudou? • O que achou mais/menos interessante? • Sugestões? • Conhece outros sistemas de explicabilidade (e.g. Lime, ...)? Já usou? • Outros comentários sobre o sistema ou sua experiência com ele?
-------------------------	---

Appendix G

Evaluation Results

Table G.1: Evaluation Results (Translation)

Task: Initial model training				
Original	Comment	(Portuguese)	Translation (English)	Participant
	"Sim, eu faço com scripts mesmo... As vezes eu uso grid search de alguma ferramenta."		"Yes [I perform hyperparameter tuning], I just use scripts... Sometimes I use grid search available in some tool."	P1
	"Seria legal colocar uma tag com o número de valores... Aqui você falou o valor mínimo e o valor máximo, mas você não especificou a quantidade de números que você quer amostrar dentro..."		"It would be cool to include a tag with the number of values... Here [referring to the parameters in the hyperparameter tuning screen] you included the minimum and maximum values, but you did not specify how many numbers you want to sample..."	P6
Task: Analysis of model execution visualizations				
Original	Comment	(Portuguese)	Translation (English)	Participant
	"O micro é uma média geral e o macro é uma média por classes... Já dá para ter uma noção, por exemplo, a gente sabe que essa base aqui é um pouco desbalanceada, porque ele não conseguiu aprender de algumas das classes."		"Micro is a general average and the macro is an average per class... One can already get a sense..., for example, we know that this dataset here is imbalanced because it [the classifier model] couldn't learn from some of the classes."	P4
	"Pelo que eu tô vendo tem as métricas mais usadas."		"From what I see, the most used metrics are available."	P5

“Tem umas que são específicas para classificador que talvez seja o caso de colocar ali. Sei lá, por exemplo, como que foi a profundidade média de árvore, como que foi o índice de gin das árvores. O problema é que quando criar muita árvore vai ficar meio difícil de visualizar, né?! Mas uma forma de pelo menos apresentar a distribuição dos índices, importância de feature”	"There are some [metrics] specific to the classifier, that might be the case to include..., for example, what was the average depth of the tree, or the gini index of the trees. The problem is that when you generate too many trees, it could be difficult to visualize, right? But at least one way to present the distribution of indices, feature relevance [...]"	P6
“Deu para ver que a base é desbalanceada, como eu tinha falado antes.”	"I can now clearly see that the base is very imbalanced as I had mentioned before."	P4
“Se as classes forem desbalanceadas, precisa olhar aquelas métricas, micro e macro F1.”	"Indeed, if classes are imbalanced, you need to look at those metrics, micro and macro F1."	P1
“Eu olho esse balanceamento em termos de volume, quantos documentos estão em cada classe, para saber qual métrica vai me dar a informação mais interessante sobre se meu modelo é efetivo ou não. Porque, por exemplo, falar de acurácia em base desbalanceada não resolve nada...”	"I look at this distribution, how many documents are in each class, to know which metric will be the most interesting one to inform me whether my model is effective or not. Because, for instance, looking at accuracy on a very imbalanced dataset doesn't mean much..."	P2
“College me surpreendeu (não apareceu entre as mais importantes), talvez seja porque tenha essa palavra em todas as classes, espalhada em todas as classes, então não vai ajudar a discriminar.”	"The word “college” surprised me (it didn't appear in the top of the rank among the most discriminative), maybe it's because it is spread across all classes, so it won't help discriminate."	P1

“Eu uso nessas explicações para entender como o modelo toma as decisões dele para definir se uma instância é de uma classe ou de outra. E quando eu olho para um gráfico desse jeito... Toda vez que aparece uma palavra muito significativa, ele consegue fazer o mapeamento de forma muito assertiva. Aí você pega palavras aqui no final que não necessariamente tem esse mapeamento assertivo.”	"I can use these explanations to understand how the model makes its decisions and to define whether an instance is from one class or another. And when I look at a graph like this... Every time a very meaningful word comes up, I can map it very assertively to the decision model. Then you can get which words, here at the end of the list, do not necessarily have this assertive capability."	P2
“É legal também porque você pode ver se existe alguma palavra que pode estar colocando algum ruído na classificação. Você pode descer naquilo e ver em quais instâncias aquela palavra tem um peso alto de importância. Se eu conseguir baixar esse modelo, eu posso entender a palavra que avacalha esse negócio e ver como resolver.”	"It's also cool because you can see whether there are any words that may be introducing some noise in the classification model. You can go “down on that” [scroll down] and see in which instances that word has a high importance weight. If I can download this model, I can understand the word that confounds the model and how to solve it."	P2
“Eu achei sensacional isso aqui... Porque agora, ligando com o que você me falou (na ferramenta), eu posso falar assim... Ah, vamos supor que ele está errando por causa de uma feature student, aí você fala: Vou tirar student... E ela está entre as mais importantes.”	"I think this is amazing, because now, with what you [the tool] told me, I can say... Hmm, let's suppose the model is misclassifying because of the feature [word] “student”, then you say: I'm taking “student” off... And it's among the most important features."	P5

“Parece que o modelo está legal porque todas essas primeiras features que estão aparecendo aqui tem tudo a ver com a ideia do dataset que você passou no começo”	"It looks like the model is OK because all these top-ranked features shown here are related to the main subject of the dataset as you described it at the beginning."	P5
"Números com 5 dígitos é o sexto item mais importante do modelo... Por que seria legal responder, né?! O quê que são esses dígitos para essa base?"	"Numbers with five digits (“parse-fivedigits”) correspond to the sixth most important feature of the model...! It would be nice to understand why, right?! What are the meaning of these numbers in this dataset?"	P6
“Eu geralmente olho essa tabelinha de importância de features... Para saber quais as features mais discriminativas, que mais ajudam a classificar... E também é legal ver as menos discriminativas, para se eu quiser fazer um feature selection, essas são as menos importantes, então posso eliminar elas do modelo e elas não vão fazer muita falta.”	"I usually look at this feature importance table to find out which features are the most discriminative, the ones that really help the classifier ... but it's also very nice to see the least discriminative features – so, if I want to make a feature selection, I know these are the ones I can eliminate from the model without much of an impact."	P1
“Eu eliminaria essas palavras do vocabulário, para gerar um modelo menor, com menos features, mais fácil de interpretar e mais rápido de rodar também.”	"I would eliminate these words from the vocabulary to generate a smaller model with fewer features, easier to interpret and faster to run as well."	P1
“Se eu tenho 7154 features que não são importantes, eu poderia querer ver que features são essas para eu não usá-las.”	"If I have 7154 features that are not important, I could want to see which features are these so I don't use them in the model."	P2

“Achei legal porque você pode fazer um feature selection usando isso. Por exemplo, você estabiliza que com uma importância menor que 0.002 você não quer manter no seu modelo, então você consegue tirar muitas features daqui... É uma redução de dimensionalidade boa.”	"It's cool! You can perform feature selection based on this. For instance, you can set a threshold on 0.002 and you do not keep values smaller than this. Then you can remove a lot of features ... It produces a “good” reduction of dimensionality."	P4
“Aí eu consigo ver, para cada objeto, quais foram as features, os atributos, mais importantes para classificar ele. Consigo ver o valor relativo dessa importância.”	"Then I can see, for each object, what were the features, attributes, most important to classify it. I can see the relative value of this importance."	P1
“Eu achei essa aqui super legal. Eu consigo ver o que foi decisivo para o modelo prever aquela instância como sendo daquela classe.”	"I found this one super cool. I can see what was decisive for the model to predict that instance as being in that class."	P2
“Essa visualização me permite ver quais as features dessa instância específica que ele usou para classificar...”	"This view allows me to see what features of this particular instance it [the model] used to classify [the instance]."	P5
"Eu consigo ver as probabilidades de cada classe. Consigo ver quais ele tem mais certeza do resultado."	"I can see the probabilities of each class. I can see for which ones it [the model] is most certain of the result."	P1
"Em termos do modelo, essa probabilidade de predição é uma coisa muito bacana, porque você consegue ver para onde que esse negócio tá sendo mais tendencioso."	"In terms of the model, this prediction probability is very cool, because you can see where this thing [the trained model] is being biased the most."	P2
"Isso é muito legal, seria o tanto que o modelo... Que ele ficou em dúvida."	"This is very cool, it would be how much the model ... How much it was in doubt."	P5

"Eu procuro olhar para ver o que tá gerando ruído. Mas a gente costuma fazer mais automatizado. Eu pegaria um script e já eliminaria as top X palavras."	"I try to look and see what is generating noise. But we usually do it more automatically. I would take a script and already eliminate the top X words."	P1
"Ou eu não elimino, ou eu testo alguns valores... As 10% piores, 1% piores, 50% piores e ver se o resultado altera muito. Se não piorar o resultado, mantenho eliminado. Eu faço isso para todas as instâncias de teste."	"Either I don't eliminate it, or I test some values... The 10% worst, 1% worst, 50% worst and see if the result changes a lot. If it doesn't get worse, I'll keep it out. I do this for all test instances."	P1
"Eu só gostaria de poder rancar features que são 0 de importância, por exemplo"	"I just wish I could eliminate features that have importance zero, for example"	P2
"Você gera a importância das features e o que eu faço normalmente é reduzir... 30% das features que menos contribuíram eu jogo fora."	"[...] what I usually do is reduce... 30% of the least important features I throw away."	P4
"É um conceito que eu já estou habituado, da matriz de confusão. Gostei das cores, eu acho que elas ressaltam ranges do intervalo que direcionam um pouco melhor a sua análise."	"It's a concept I'm already used to, the confusion matrix. I liked the colors, I think they highlight ranges that better direct your analysis."	P2
"A classe other é realmente difícil de classificar, ela está espalhada aqui. Parece que a classe faculty é a mais fácil de classificar, acerta mais."	"The class 'other' is really hard to classify, it is spread here [pointed to the line for 'other' in the confusion matrix]. It seems that the 'faculty' class is the easiest to classify, it gets it right more often."	P1

“Ele é extremamente enviesado pela quantidade de exemplos, eu vi pela classe faculdade e departamento. Departamento foi a que ele mais errou e faculdade foi o que ele mais acertou.”	"It [the model] is extremely biased by the amount of examples... I noticed based on the classes 'college' and 'department'. 'Department' was the one misclassified the most and 'college' was the one it [the model] got right the most."	P6
"Eu sempre quero saber o seguinte: Onde o meu modelo está errando mais e porque. Então aqui eu consigo ver quais classes ele está errando mais, e talvez na etapa anterior eu veja o porque ele está errando, por causa de algumas características que ele está escolhendo. Então isso vai ajudar a eu talvez fazer um balanceamento de classe..."	"I always want to know the following: Where is my model missing the most and why? So here [seeing the class visualization] I can see which classes it's getting wrong the most, and maybe in the previous step [By Instance] I'll see why it's missing... because of some of the features it's choosing. So this might help me balance the classes."	P5
“Tinha a opção de balancear? De cada classe, pegar a mesma quantidade de documentos... Eu diminuiria esse impacto... Eu balancearia, mas tem que balancear na mesma proporção, senão vai diferenciar do mundo real.”	"Did the tool have the option to balance? From each class, take the same amount of documents... I would lessen this impact... I would balance, but it would have to balance at the same rate, otherwise it would be different from the real world."	P3
“A partir da análise, eu tenho que melhorar bastante meu classificador para departamento e curso, né? Os que estão mais errando.”	"From the analysis, I have to greatly improve my classifier for 'department' and 'course', right? The ones that are getting misclassified the most."	P3

“Te permite tomar decisões de, ao invés de fazer um classificador multiclass, fazer um essemble, e tivesse classificadores separados onde você colocasse a coisa mais discriminada analisada separadamente. Isso permite pro pesquisador tomar esse tipo de decisão.”	"It allows you to make decisions like, instead of making a multiclass classifier, making an essemble, and having separate classifiers where you could put the most discriminating thing analyzed separately. This [the visualizations] allows the researcher to make this kind of decision."	P2
“O programa foi bom para visualizar os dados e propor melhorias.”	"The tool was good for visualizing the data and proposing improvements."	P3
“Deu para entender o resultado. O modelo eu consigo visualizar bem com a nuvem de palavras, para texto é muito bom.”	"I could understand the result. I can see the model well with the word cloud, for text [classification] it is very good."	P1
As visualizações foram úteis? “Com certeza, para compreensão foram [...] A análise por classe, matriz de confusão, com certeza. Na importância das features, eu acho que [n]esse gráfico se dá uma boa noção de como que é a distribuição da importância das features”	"Sure [the visualizations were helpful], for understanding, of course [...] The analysis by class, matrix of confusion, sure. Regarding the importance of features, I think this graph gives a good idea of how the importance of features is distributed."	P6
“Elas são bastante úteis, [...] Principalmente quando você tem esse volume muito grande de features, é difícil você entender na mão como esses resultados... como o classificador está usando essas informações para gerar os resultados.”	"They [the visualizations] are very useful, [...] Specially when you have this very large volume of features, it's hard for you to understand these results... how the classifier is using this information to generate the results."	P2

“Eu super usaria isso... Eu acho que a utilidade disso é enorme... Eu acho que todas as visualizações propostas agregam muito e elas são bastante complementares.”	"I would super use this. I think the usefulness of this is huge. I think all the proposed visualizations add a lot and they are quite complementary."	P2
“Eu acho que as informações que eu obtive aqui me dariam condições de refinar o meu modelo com muito mais facilidade do que eu faria se eu tivesse que ficar fazendo teste, treino... Aqui eu tenho uma ferramenta que ajuda a me guiar... Além de fazer o... de setar os hiperparâmetros, eu ainda teria que fazer uma avaliação assim grande para chegar aonde eu chegaria simplesmente rodando a sua ferramenta. Ela ajuda de mais não só fazendo o fine tuning dos parâmetros, como fazendo o fine tuning do modelo.”	"I think the information I got here would enable me to refine my model much more easily than I would if I had to do testing, training... Here I have a tool that helps me. [...] Besides doing ... tuning the hyperparameters I would have to perform a large evaluation just to get to the same point that I would by just executing your tool. It really helps not only fine tuning the parameters, but also fine tuning the model.	P5
“Fazer a importância de features por classe, quais features são mais importantes para cada classe. Eu acho que a ferramenta tem tudo que é relevante, mas é só uma sugestão caso queira adicionar.”	"Show the importance of features by class... which features are more important for each class. I think the tool has everything that is relevant, but it's only a suggestion in case you want to add."	P1
“Queria ver as features de maior impacto por classe.”	"I wanted to see the features with higher impact per class."	P3
“Não sei se existe o conceito de importância das features por classe, talvez seja o caso de colocar”	"I don't know if there is a concept of the importance of features per class, so maybe it would be the case to include it."	P6

“Eu acho que o trabalho tem um potencial enorme para expandir, se gente considerar entrar nisso de como as features interagem, principalmente quando tiver outros modelos.”	"I think this work has a huge potential to expand if we consider getting into how features interact, especially when you have other [machine learning] models."	P2
"Uma coisa que eu gostaria de ver nos modelos que eu sei que é muito difícil é, ou ver a árvore, ou um pedaço dela. Eu até já olhei no próprio arquivo do modelo, no XML, mas é muito difícil de visualizar. Eu procuraria mais perto da raíz primeiro, onde estão as features mais importantes, só para tentar ver como a combinação de features está ajudando. O problema de olhar a importância é que você só vê como as features isoladamente contribuem pro modelo. Você pode podar a árvore... Não seria o modelo final, mas seria uma simplificação do modelo... O mais importante vai estar próximo da raíz... Enfim, é algo para se pensar."	"One thing I'd like to see in the models, which I know is very difficult, is either the whole tree or a piece of it. I've even looked in the model file itself, in XML, but it's very hard to visualize. I would look closer to the root first, where the most important features are, just to try to see how the combination of features is helping. The problem with looking at importance is that you only see how features in isolation contribute to the model. You can prune the tree... It wouldn't be the final model, but it would be a simplification. The most important thing will be near the root. Anyway, it's something to think about."	P1

Task: Model Improvement

Original Comment (Portuguese)	Translation (English)	Participant
“A, legal, ele compara com o anterior... Eu piorei o resultado... Nas features por instância: Já mudou completamente”	"Cool, it compares with the previous one [execution]. I worsened the results ... the most features for each instance have completely changed."	P1

"Ah, que massa... Então esse é o impacto de eu ter excluído aquela feature..." "Legal, então eu consigo ver que excluir aquela feature foi bom pro meu modelo."	"Ah, that's cool... So this is the impact I had deleting that feature [...] Cool, so I can see that deleting that feature was good for my model."	P2
"Aqui me dá a informação do que melhorou e o que piorou" "Deu para observar que o resto das features também sofreu uma diferença na importância" "Melhorou em micro mas piorou em macro"	"Here [Evaluation metrics visualization] it gives me the information about what has improved and what has gotten worse. [...] It is noticeable that the rest of the features also differed in importance. [...] Improved in micro but got worse in macro."	P4
"Nossa, eu piorei o negócio todo... Tá vendo, eu fui tirar uma palavra super importante para ele, tá vendo o quê que deu?"	"Wow, I made the whole thing worse. See... I removed a very important word for it [the model], see what happened ?!"	P5
"Caraca, e a acurácia ficou o mesmo valor. Piorou pouco também."	"Gosh, accuracy is the same value. Got a little worse too."	P5
"Parece que diminuiu o número de importâncias que tem zero"	"It seems like the number of [features with] importance zero has decreased."	P5
"Muito mais sentido agora as palavras que você usa para a predição"	"It makes more sense now, the words you use to predict."	P5
"Melhorou a confiança no modelo, como explicar para outras pessoas... Eu consigo explicar muito melhor"	"It improved the confidence in the model, how to explain it to others... I can explain it much better."	P5
	"It showed me why it got worse."	P5
"Então melhorou... Diminuiu o erro..."	"It has improved... the error has decreased..."	P3
"Será que aumentou a diferença? Quando eu tirei (feature computer), será que essa daqui (feature science) ficou mais importante?"	"I wonder if it increased the difference? When I took it [feature 'computer'] out, did this one here [pointed at feature 'science'] become more important?"	P3

Task: Analysis of history results		
Original Comment (Portuguese)	Translation (English)	Participant
“Então aqui eu poderia ter várias execuções e verificar o quanto está variando de acordo com essas mudanças que eu vou fazendo no meu modelo... Eu quero que meu erro sempre diminua e a acurácia ou outra métrica aumente.” “Muito massa isso, porque aí você consegue ver essa variação em termos de execução.” “Acho que isso é legal para ver o impacto da minha alteração no modelo...”	"So here I could have several executions and check how much is changing according to these changes that I am making in my model [...] Very cool, because then you can see this variation in terms of execution [...] I think it is cool to see the impact of my changes on the model."	P2
Fazer uma ordem cronológica delas ficou bem legal (variação das métricas)” “Dá para ver o que eu errei, e para onde foi esse erro.”	"Making a chronological order of them [executions] was pretty cool. [...] It's possible to see what I did wrong and where that error occurred."	P4
“No geral, eu percebi que algumas classes não mudaram nada.”	"Overall, I realized that some classes have not changed at all."	P5
“Ele piorou a classificação para departamento... Tirando aquela feature não melhorou a classificação para departamento... Se o objetivo fosse melhorar a classificação para departamento, não seria interessante fazer isso”	"It made the department classification worse. Removing that feature didn't improve the department classification. If the goal was to improve the department classification, it wouldn't be interesting to do that."	P3
"Deu para ver o impacto aqui, 0.1%, porque eu eliminei poucas features mesmo."	"I could see the impact here, 0.1%, because I eliminated only a few features."	P1

“Aqui, por exemplo, embora você me falou lá que teve uma variação de 0.5, olhando o gráfico assim você vê que o impacto é pouco, então a feature que eu excluí não impacta muito meu modelo, já que eu posso ver que meus resultados continuam lineares e paralelos da mesma forma... Mas eu acho que em um cenário em que isso realmente tiver um impacto grande, que isso vai ser muito interessante para conseguir ver essa variação, em termos de quanto isso mudou.”	"Here, for instance, even though it told me that there was a 0.5 variation, looking at the graph you can see that the impact is small, so the feature that I excluded does not impact my model much, since I can see that my results continue to be linear and parallel in the same way... But I think that in a scenario in which there is really a big impact, this will be very interesting to be able to see this variation, in terms of how much it has changed."	P2
“É possível ter intuições do que pode melhorar, mas ainda não melhorei” “Foi possível criar mais confiança, mas ainda não sinto confiante com esse resultado.”	"It is possible to have an intuition of what can be improved, but that I haven't improved yet [...] It was possible to have more confidence, but I still don't feel confident with this result."	P6

Task: Overall comments

Original Comment (Portuguese)	Translation (English)	Participant
“Eu achei incrível [...] Realmente facilita bastante para experimentação”	"I found it amazing [...] It really makes experimentation a lot easier."	P6
“Não só nessa tarefa, mas eu já estou vendo nas minhas aplicações futuras usar algo parecido."	"Not only for this task [the one explored in the evaluation], but I can already see using something similar in my future applications."	P5
“Eu super usaria isso... Eu acho que a utilidade disso é enorme... Eu acho que todas as visualizações propostas agregam muito e elas são bastante complementares."	"I would super use this. I think the usefulness of this is huge. I think all the proposed views add a lot and they are quite complementary."	P2

“Eu perdia muita informação que eu tô vendo aqui, que eu podia ter tido lá. E é exatamente o que eu queria ter visto lá, naquela época. Se eu tivesse esse arcabouço que eu tenho aqui, eu acho que seria muito mais fácil explicar o modelo que eu estava gerando.”	"I missed [in the past] a lot of information that I'm seeing here now. It's exactly what I wish I had seen at that time. If I had this framework that I have here, I think it would have been much easier to explain the model I was generating."	P5
“Eu utilizaria ela [a ferramenta] para melhorar o modelo” “O impacto de uma feature na classificação, acho que essa é a mais importante... Acho que também as métricas... Exportar a matriz de confusão também é importante”	"I would use it [Explain-ML] to improve the model. [...] The impact of a feature on classification, I think this is the most important [visualization].... And the metrics also... Exporting the confusion matrix is also important."	P3
“Eu usaria, para saber questões de erro daonde saiu para onde foi” “Essa questão de você poder manter os históricos das execuções eu acho muito bom, porque normalmente a gente faz isso por meio de tabela, só que tabela você esquece... você não coloca parâmetros, não coloca um monte de coisas, e aqui no sistema tem tudo isso. É uma forma de manter assim, uma forma organizada, e não é trabalhoso manter tudo que você precisa.” “Acho que tudo que eu faço assim, já está aí. Algumas análises, mas aí seria muito fora de escopo, entre classificadores, mas foge muito do escopo aqui.” “O que a gente faz muito é gerar a matriz de confusão, ver a importância das features... Tem tudo aqui, sabe?!”	"I would use it [Explain-ML] to know where the error came from and where it went [...] This possibility to keep the execution histories is very good, because we usually do it through a table, but in a table you forget... you don't put parameters, don't put a lot of things, and here in the system there is all that. It's a way to keep it that way, an organized way, and it's not a hassle keeping everything you need. I think everything I do is already there."	P4

“Acho que tudo que eu faço assim, já está aí.”	"I believe everything I do, is already there."	P4
“Eu já usei o SHAP... Na verdade eu não uso tudo que ele tem, porque eu sei que ele faz muita coisa, mas as coisas que eu usei eu acho que vocês estão fazendo.”	"I've already used SHAP. Actually I don't use all it offers, because I know it does a lot of things, but the things I have used I think you're doing [in Explain-ML]."	P2
“Usei outra ferramenta com menos recursos que essa que é o Lemonade. Tudo que ele tem o seu já tem, o seu tem mais coisas. Ajuste de parâmetros é uma coisa que não tem lá, lá a gente seleciona os parâmetros que a gente quer e roda uma vez só com esses parâmetros.” “A possibilidade de ver por instância os resultados, e tudo integrado, a nuvem de palavras, os gráficos as métricas, não é uma coisa que eu vejo.”	"I used another tool with less features than this, which is Lemonade. Everything it has yours already has. Yours has more features [...] Parameter tuning is one thing that is not available there [in Lemonade], there we select the parameters we want and execute it with these parameters. [...] Being able to see the results for instance, and everything integrated, the word cloud, the graphs, the metrics, is not something I [usually] see."	P1
"Usei outra ferramenta com menos recursos que essa que é o Lemonade. Tudo que ele tem o seu já tem, o seu tem mais coisas. Ele já incorporou outros modelos (cluster, padrões frequentes, outros algoritmos de machine learning), mas porém não tem tantas visualizações quanto tem aqui. Então eu incorporaria isso mesmo, incluir outros tipos de modelo."	"It [Lemonade] has already incorporated other models (cluster, frequent patterns, other machine learning algorithms), but it doesn't have as many views as it has here. So I would incorporate that, include other model types."	P1
“Já utilizei a ferramenta orange, o weka... Aqui você só está analisando o random forest, lá eu tenho várias coisas”	"I've already used the orange tool, weka... Here you are just analyzing the random forest, there I have several things."	P3

"Eu queria que isso fosse para todos os outros modelos, que eu pudesse ter uma ferramenta dessa para qualquer outro projeto que eu fosse fazer, ía ser uma contribuição sensacional para essa época de deep learning, por exemplo..."	"I wish this [tool] was for every other model, that I could have such a tool for any other project I was going to do, it was going to be a sensational contribution to this era of deep learning, for example."	P5
"Eu achei incrível, uma das coisas que eu realmente gostaria de ver, que eu comentei com você, é ter a possibilidade de escolher classificadores genéricos (fora do sklearn)."	"I found it amazing, one of the things I would really like to see, which I commented with you, is being able to choose generic classifiers (outside of ScikitLearn)."	P6
"Eu usaria, com certeza usaria... Eu, da computação, usaria... Mas eu acho que ela tem um potencial ainda maior quando você lida com aquilo que eu já comentei... Quando a gente começar a pensar em pessoas de fora... O SHAP e essas outras ferramentas tem essa limitação: O cara vai ter que fazer um modelo, vai precisar implementar em python, ele vai precisar saber manipular... Para usar isso aqui eu não preciso saber python... Então ela tem um potencial muito interessante, eu acho que esse público é mais amplo."	"I would use it for sure... Me, from computing, would use it... But I think it has an even greater potential when you deal with what I have already commented... When we start thinking about outsiders [from the computing field]... SHAP and these other tools have their limitations: the guy will have to generate the model, will need to implement it in python, he will need to know how to manipulate [the tool]... To use this here [Explain-ML] I do not need to know python ... So it has a very interesting potential, a wider audience."	P2

"Gostei que assim, além de ser útil, o sistema tem que ter essa cara amigável, essa interface legal que o seu sistema tem... O sistema pode ser muito útil as vezes, e a ferramenta não ter um visual legal, não ter essa usabilidade tão simplificada... A pessoa deixa de usar... Vendo esses gráfico, vendo essa interação facilitada, agrada aos olhos ver... Mesmo que você fique travado em algum momento, você quer continuar mexendo, porque ela te dá uma informação boa de forma que o seu cérebro assimila facilmente, então é bem legal."	"I liked that... besides being useful, a system has to have a friendly face, like yours does. A system can be very useful sometimes, but if it does not have a cool look, doesn't have such a simplified usability... [then] the person stops using it. Seeing these graphs, seeing this facilitated interaction, it pleases the eyes to see. Even if you get stuck at some point, you want to keep moving, because it gives you good information in a way that your brain easily assimilates, so it's very cool."	P5
"Eu queria que isso fosse para todos os outros modelos, que eu pudesse ter uma ferramenta dessa para qualquer outro projeto que eu fosse fazer, ia ser uma contribuição sensacional para essa época de deep learning, por exemplo... A maioria dos estados da arte hoje não tem isso, o cara só fala: minha métrica bateu o estado da arte, mas não fala o porquê, não mostra o porquê... Se você tivesse uma ferramenta dessa para explicar o modelo, você saberia onde ele é fraco, quais são os constraints daquele modelo, onde ele é forte, onde ele é fraco. E isso é o que tá faltando na maioria dos artigos atualmente."	"I wish it were for all the other models [...] Most 'state of art' works today don't have that. The guy [author] just says: my metric beat the state of the art, but he doesn't say why, he doesn't show why... If you had such a tool to explain the model, you would know where it is weak, what are the constraints of that model, where it is strong, where it is weak [sic]. And that is what is missing in most articles today."	P5