POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

AXOM: Combination of Weak Learners' eXplanations to Improve Robustness of Ensemble's eXplanations

Tesi di Laurea Magistrale in
Computer Engineering - Ingegneria Informatica

Author: **Riccardo Pala**

Student ID: 969598
Advisor: Prof. Daniele Loiacono
Co-advisors: Prof. Esteban García-Cuesta
Academic Year: 2022-23

Abstract

Machine learning is a field of artificial intelligence that, day by day, is becoming part of every aspect of human life. This is due to its excellent characteristics, including flexibility and relative ease of deployment, all accompanied by an extraordinary capacity for prediction.

However, the underlying algorithms construct complex models, often incomprehensible for humans, causing a difficulty to provide an interpretation of the reasons that contributed to a given output. This becomes a problem of critical importance when the decisions derived from such systems strongly affect humans' lives and that is the main reason why, in many applications, machine learning techniques still struggle to find a place in use.

The development of XAI (eXplainable Artificial Intelligence) techniques have in recent years greatly improved the interpretability of models, which have thus progressed from acting as black-box models to ensuring a behaviour that is comprehensible to humans. However, the novelty of such approaches is reflected in the presence of several issues related to notions such as fairness, confidence, accessibility and many others. Among these, by placing a particular focus on the *trustworthiness* of an XAI algorithm, recent studies have shown that under this aspect, even the most widely adopted algorithms present various problematics since, for some complex models, the explanations lack *robustness*, which is tightly related to concept just mentioned.

In this work, research and analysis is conducted on the application of these techniques to *ensemble models*, i.e. models derived from the combination of many individual ones. The promise is to use aggregation to make explanations more robust and consequently more reliable, alongside the predictive abilities of the model. In particular, we argue that a combination through discriminative averaging of ensembles weak learners explanations can improve the robustness of explanations in ensemble models. This approach has been implemented and tested with post-hoc SHAP method and Random Forest ensemble with successful results. The improvements obtained have been measured quantitatively and some insights about explicability robustness on ensemble methods are presented.

Keywords: Computer Science, Engineering, Artificial Intelligence, Machine Learning, Explainability, XAI, Robustness, Trustworthiness, Decision Tree, Random Forest, SHAP

Abstract in lingua italiana

L'apprendimento automatico è un campo dell'intelligenza artificiale che, giorno dopo giorno, sta entrando a far parte di ogni aspetto della vita umana. Ciò è dovuto alle sue eccellenti caratteristiche, tra cui la flessibilità e la relativa facilità di implementazione, il tutto accompagnato da una straordinaria capacità di previsione.

Tuttavia, gli algoritmi sottostanti costruiscono modelli complessi, spesso incomprensibili per gli esseri umani, causando la difficoltà di fornire un'interpretazione delle ragioni che hanno contribuito a un determinato risultato. Questo diventa un problema di importanza critica quando le decisioni derivate da tali sistemi influenzano fortemente la vita degli esseri umani e questo è il motivo principale per cui, in molte applicazioni, le tecniche di apprendimento automatico faticano ancora a trovare impiego.

Negli ultimi anni, lo sviluppo di tecniche XAI (eXplainable Artificial Intelligence) ha migliorato notevolmente l'interpretabilità dei modelli, che sono così passati dall'agire come modelli black-box a garantire un comportamento comprensibile all'uomo. Tuttavia, la modernità di questi approcci si riflette nella presenza di diverse problematiche legate a nozioni come equità, fiducia, accessibilità e molte altre. Tra questi, ponendo attenzione sulla *affidabilità*, studi recenti hanno dimostrato che anche gli algoritmi più adottati presentano diverse problematiche poiché, per alcuni modelli complessi, le spiegazioni mancano di *robustezza*, che è strettamente legata al concetto appena citato.

In questo lavoro, vengono condotte ricerche e analisi sull'applicazione di queste tecniche ai "modelli ensemble", cioè ai modelli derivati dalla combinazione di molti predittori individuali. La promessa è quella di utilizzare l'aggregazione per rendere le spiegazioni più robuste e di conseguenza più affidabili, insieme alle capacità predittive del modello. In particolare, sosteniamo che una combinazione attraverso una media discriminativa di insiemi di spiegazioni di weak learners può migliorare la robustezza delle spiegazioni negli ensemble. Questo approccio è stato implementato e testato con il metodo post-hoc SHAP e l'ensemble Random Forest con risultati positivi. I miglioramenti ottenuti sono stati misurati quantitativamente e sono state presentate alcune intuizioni sulla robustezza delle spiegazioni nei metodi ensemble.

Parole chiave: Informatica, Ingegneria, Intelligenza Artificiale, Machine Learning, Spiegabilità, XAI, Robustezza, Affidabilità, Albero Decisionale, Foresta Casuale, SHAP

Contents

Abstract **i**

Abstract in lingua italiana **iii**

Contents **v**

1 Introduction **1**
1.1 Rationale 1
1.2 Objectives 2
1.3 Contributes 2

2 State of the Art **5**
2.1 Decision Tree 5
2.2 Ensemble Methods and Bagging 6
2.3 Random Forest 7
2.4 eXplainable Artificial Intelligence (XAI) 8

3 Development **15**
3.1 Robustness Metric Definition 17
3.2 Combination of Weak eXplanations 20
3.2.1 Why ensembles? 20
3.2.2 Why combine Weak eXplanations? 22
3.3 Averaging on the eXplanations Of the Majority (AXOM) 23
3.4 Datasets 25
3.5 Experimental Design 26

4 Results **29**

5 Conclusions **37**

6 Future Work **39**

Bibliography **41**

A SHAP - Equivalence between RF global explanation and average of weak explanations **45**

List of Figures **47**

List of Tables **49**

Acknowledgements **51**

1 | Introduction

1.1. Rationale

Machine learning (ML) is increasingly becoming an important aspect of various fields, with applications ranging from image recognition and natural language processing to medical diagnosis and fraud detection, in which decisions can heavily affect humans' lives. However, with the increasing use of AI in real-world applications, concerns have arisen about the transparency, trustworthiness, and reliability of these systems. Indeed, ML algorithms used in AI can be categorized as white-box or black-box. White-box models provide results that are understandable for experts in the domain. Black-box models, on the other hand, are extremely hard to explain and can hardly be understood even by domain experts. This implies that, while often characterized by significantly greater predictive capabilities along the ability to recognize more complex learnable patterns, the latter type of models hardly finds a place in use when dealing with critical real-world domains. As a result, the concept of explainable artificial intelligence (XAI) has become a crucial area of research. The importance of XAI lies in its ability to provide understandable and interpretable reasoning for the decisions made by AI systems. With XAI, users can comprehend the decision-making process of AI systems and identify any potential biases or errors.

There are several works in the literature that highlight the importance of model transparency, interpretability and explainability. However, recent studies have shown that, alongside the broad variety of XAI methods, there are still a number of concerns regarding the *robustness* of the explanation values that they are able to produce, particularly in situations in which models are feed with inputs that lay outside the data distribution they were trained on. Furthermore, explanations provided by models may be sensitive to small changes in the input data, leading to unreliable explanations. All these concerns justify the need to carry out investigations in order to develop explanation methods capable of improve this quality.

1.2. Objectives

This work is devoted to the conduction of a study aimed at the developing of efficient and effective methods for the application of model-agnostic XAI techniques to model ensembles. The objective is to find a way to exploit the excellent prediction capabilities and improved robustness of this category of models in order to enable the production of explanation values that are consequently more robust to small perturbations in the input data and therefore more trustable.

In this work, we decided to carry out the achievement of this goal by pursuing responses to the following queries:

- How can we define and measure the robustness of model explanations?
- How robust the existing XAI techniques can be considered?
- How can we incorporate the concept of robustness into the design and development of procedures for producing better explanations?

From a practical point of view, a case study is conducted by first reviewing the current state of ML model and XAI methods used to calculate explanation values, realizing then a comparison between the robustness of explanations obtained from the single models, the explanations obtained applying the XAI techniques straight to the ensemble model and finally the explanations obtained by the application of our proposed solution.

What we are aiming to is the production of explanation values that result more robust to small deviations in the input. This means that, given a certain data point x and a slightly perturbed version of it x', we expect the explanations y and y', respectively produced from the two inputs just mentioned, to differ marginally. To achieve this objective, the idea is to take advantage of the decomposability of the ensemble models in order to exploit the explanation values provided by the single weak estimators for the purpose of the building of a more robust global explanation by means of some form of combination.

1.3. Contributes

Concerning researches in the field of explainable artificial intelligence, it can be argued that the content of this paper highlights a point of view that has so far rarely (if ever) been explored. Typically, when trying to explain the reasons for the output of a model ensemble, we consider the model as a whole while overlooking the individual contributions of the weak learners that compose it. For this purpose, we propose to produce a global

explanation of the ensemble resulting from combining the explanations of a **subset** of weak learners. The underlying concept is based on the idea that the global explanations obtained through the straightforward application of the XAI methods to the ensemble suffer from some undesirable influence from the weak estimators who, at the prediction stage, provided a label different from the majority. Considering these influences as *noise*, what makes this work promising is that applying a discriminative selection of explanations to be taken into account can yield, through their combination, a de-noised global explanation deprived of the bad influence of the weak learners who provided an incorrect prediction.

2 | State of the Art

Technological advances in the field of artificial intelligence have enabled us to achieve extraordinary objectives. Nowadays, the use of AI techniques allows us to perform tasks in a better way, in terms of speed, reproducibility and scalability [32]. For this and other reasons, nowadays a number of decision-making problems in the real-world domain are addressed with the help of Machine Learning models. Their reliability, accuracy and ease of deployment are the pivotal properties that justify their widespread use. For the purpose of this work, the focus is on *supervised learning* models, i.e. models that, through the use of a labelled data set, are subjected to a training phase that consists of finding the best values to assign to its parameters in order to minimise the prediction error. The resulting model will thus be a tool capable of providing predictions on unlabelled data based on the decision patterns learned from the dataset it was trained on.

2.1. Decision Tree

In this work, experiments are conducted focusing on *Decision Tree* (DT), which is a family of supervised predictive models that can be used for both classification and regression cases. They are tree-structured models composed by nodes and branches, where the inner nodes represent the features of a dataset, the branches represent the decision rules and each leaf node represents an outcome. There are several algorithms that allows the construction of a Decision Tree model, such as CART [10], CHAID [25], MARS [18] and many others. In this work we will focus on the former one, which is the algorithm deployed in our experiments. The CART (Classification And Regression Trees) algorithm is a type of classification algorithm that is required to build a decision tree on the basis of **Gini's impurity index** (other possibilities are entropy and information gain). To construct a decision tree, the CART algorithm starts with the specification of a feature that will become the root node. Gini's formula is used to measure this impurity and identify how well a feature classifies the given data. The feature with the least impurity is selected as the node at any level. This process is iterated at every node at each depth level until all the data is classified.

Once the tree is constructed, to predict the class of a certain data point the algorithm starts with the root node of the tree. The algorithm compares the values of the root attribute with the attribute of the record and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and moves further. It continues the process until it reaches the leaf node of the tree, within which the input prediction will be contained. Decision Tree is one of the most commonly used models due to its speed and ease of interpretation, mixed with a reduced need for input data cleaning, compared to most models. However, it suffers from some disadvantages, such as the increasing complexity of the algorithm in the case of multi-labelled classification problems, the inability to learn excessively complex patterns and the tendency to overfit the data [10].

2.2. Ensemble Methods and Bagging

When dealing with problems such as the overfitting of training data, one outstanding solution is surely represented by the so-called *ensembles* of models. There are in fact several techniques that allow the contributions of individual models to be put together to provide more accurate and robust predictions. Ensembles are designed to increase the accuracy of the single models [19]. To better understand, human being tends to apply the same kind of reasoning when dealing with real-life decision processes, by seeking several opinions before making any important decision. We weigh the individual opinions, and combine them to reach our final decision [34, 36]. Indeed, the main idea behind the ensemble methodology is to weigh several individual estimators and combine them in order to obtain an estimator that outperforms every one of them. The individual models that we combine are known as *weak learners*. We call them weak learners because they either have a high bias or high variance, which is the reason why they cannot learn efficiently and perform poorly. A high-bias model results from not learning data well enough. It is not related to the distribution of the data, hence future predictions will be unrelated to the data and thus incorrect. On the other hand, a high variance model results from learning the data too well. It varies with each data point, hence it is impossible to predict the next point accurately. As we know from the bias-variance trade-off, an underfit model has high bias and low variance, whereas an overfit model has high variance and low bias. In either case, there is no balance between this two quantities. To overcome this problem, ensemble learning tries to balance this bias-variance trade-off by reducing one of this two quantities.

Bagging. Among all the ensemble methods, *Bagging* [7], also known as *bootstrap aggregating*, is a powerful model which helps to prevent overfitting by means of the aggregation of multiple versions of a predicted model. We use it for combining weak learners of high variance, to produce a more balanced model. Each model is trained individually, and combined using an averaging process. The primary focus of bagging is to achieve less variance than any model has individually. As the name says, the method consists in two main steps: Bootsrapping and Aggregation. The first one involves resampling subsets of data with replacement from an initial dataset. In other words, subsets of data are taken from the initial dataset by means of a drawing process in which an individual data point can be sampled multiple times. These subsets of data are called bootstrapped datasets or, simply, bootstraps. Each one of them is used to train a separate weak learner. In the aggregation phase, the individual weak learners are trained independently from each other in a way that each estimator makes independent predictions. The results of those predictions are aggregated at the end to get the overall one, using either max voting (used for classification problems, consists in taking the most occurring prediction) or averaging (used for regression problems, consists in taking the average of the predictions).

2.3. Random Forest

One relevant case of ensemble method is that of *tree ensembles*, that improve the generalization capability of single decision trees, which are usually prone to overfitting. To circumvent this problem, in fact, tree ensembles combine several trees to obtain an aggregated prediction by means of a majority voting [5]. In this work, particular attention is placed on *Random Forest* (RF), a very successful machine learning tool that exploits the combination of independent decision trees to build up a more powerful learner. Indeed, a Random Forest is a classifier consisting of a collection of tree-structured classifiers in which each tree casts a unit vote for the most popular class to assign to the input sample.

There are several works in the literature that introduce different approaches for constructing the ensemble. To cite a few examples, in [7] each tree is trained on a randomly selected version (without replacement) of the training set (bagging), in [15] the split at each node is selected from the best K splits, in [8] several training sets are generated as a consequence of randomising the outputs of the original one, in [24] each tree is grown by randomly selecting a subset of features, and in [21] the splitting points in each node of the tree are selected randomly instead of using the optimal split based on the training data. In the case of this work, experiments have been carried out using the Random Forest as a reference model, as specified in [9], where *bagging* (Bootstrapping + AGGregation)

is used in conjunction with a random selection of features. This particular method has been shown to come with a significant increment in the classification accuracy as well as in the ability to generalize of the model with respect to Decision Tree [9, 24]. One of the key ideas behind Random Forest is the use of random feature selection. Instead of using all the available features to construct each decision tree, the algorithm selects a random subset of features for each tree. This helps to reduce the correlation between the individual trees and increase the diversity of the ensemble. In addition, the use of bootstrapping also helps to increase the diversity of the trees. As already mentioned, the bootstrapping technique consists of constructing a new training set for each tree in the ensemble by dragging and replacing the data points that make up the ensemble. In each bootstrap training set, about one-third of the instances are left out, which are going to be used to conduct the out-of-bag estimates to assess the performance of the corresponding weak tree. To make a prediction for a new input instance, the model applies each of the individual decision trees to the input and combines their predictions. In a classification problem, the final prediction is the class that receives the most votes from the individual trees. In a regression problem, the final prediction is the average of the predictions made by the individual trees. One of the main advantages of this ensemble is its ability to handle high-dimensional data with complex interactions between the features. The algorithm can also handle missing data and noisy data, making it a robust and reliable method for a wide range of applications. In addition, the use of multiple decision trees reduces the risk of overfitting, which is a common problem in many other machine learning algorithms.

2.4. eXplainable Artificial Intelligence (XAI)

However, there are certain areas of competence in which addressing the problem of understanding how a model produced a certain output becomes a matter of great importance [22]. Many algorithms construct models that are opaque for humans [11], while explanations that support the output provided by a model become crucial in fields such as, for example, medicine, where an expert needs a lot of information about the features that contributed to the making of a certain decision [39]. Furthermore, XAI techniques not only improve the trustworthiness and transparency of ML models by providing explanations and simplifications of so-called “black-box models” but can also act as tools for extracting new knowledge from them.

Having said that, if some models enjoy intrinsic explainability (e.g. Decision Trees), others, more complex, act as real *black boxes* during their decision making process [5]. A striking case is provided by Deep Neural Networks, which are highly praised for their

ability to learn complicated relationships between inputs and outputs, at the cost, nevertheless, of an increasing difficulty in interpreting the reasons for certain choices. Speaking of ensembles of models, specifically ensembles of trees, this technique, while effective against overfitting, makes the interpretation of the resulting model more complex than that of each of its component trees [5].

In general, it is well-known that there is a clear trade-off between the performance of a machine learning model and its ability to produce explainable and interpretable predictions. In fact, although they enjoy greater transparency than black-box models, intrinsically explainable models are not as powerful and fail to reach the state of the art when compared to the former [27].

Post-hoc Explainability. When a model does not meet the requirements to be considered intrinsically explainable, it is necessary to apply methods to explain the reasons for its decisions. In this case, we are referring to post-hoc explainability techniques, which can help to interpret the output of a model by giving a measurement of how much a certain feature of the input data contributed to the final decision. These can be divided into *model-agnostic*, which are applicable to any type of model, or *model-specific*, which are tailored or specifically created to explain certain ML models. In this work, we will focus our studies mainly on the former category.

Model-agnostic post-hoc explainability techniques are designed to adapt their use to any type of model. The main types of approach consist of improving the interpretability of a model by simplifying it, or by extracting knowledge directly from it, with consequent visualisation. Following the taxonomy stated in [5], we can divide these techniques into the following main categories:

- **Explanation by simplification.** The opaque model is approximated by a simplified version of it, which is easier to interpret. The difficulty in applying this technique lies in the need to consider models that are flexible enough to be approximated effectively, regardless of their original complexity [6]. Within this category we also find some *local explanations* techniques, among which we mention *Local Interpretable Model-Agnostic Explanations* (LIME). The functioning of LIME consists in identifying an interpretable model that is able to faithfully approximate the predictions of the original model, within a neighbourhood of the input data point whose output we intend to explain [35].
- **Feature relevance explanation.** This type of techniques aims to rank features according to their relevance in the determination of the output of a given model.

Among the various techniques belonging to this category, one that is certainly worth mentioning is *SHapley Additive exPlanations* (SHAP). SHAP is a technique derived from Game Theory, which uses the formula for calculating the players' Shapley Values to attribute to each feature a value indicating its importance in the prediction process, following the principles of local accuracy, missingness and consistency [29].

SHapley Additive exPlanations (SHAP). We include a separate paragraph to talk in details about SHAP, which is the XAI technique used as a reference method for the experiments conducted in this work. One of the main challenges in interpreting the predictions of complex machine learning models is understanding the contribution of each input feature to the final prediction. SHAP addresses this challenge by providing a way to assign an importance value to each feature based on its contribution to the prediction. The importance values are calculated using game theory concepts, specifically the Shapley values which are a way to calculate the contributions of each player in a cooperative game and so to find the most fair way to distribute the final prize basing on them. In a ML setting, the feature values of a data instance act as players in a coalition. In such a case, the Shapley value is the average marginal contribution of a feature value across all possible partitions of the feature space. Let $\mathcal{A}$ be the set of features of the input space, the formula to calculate SHAP explanation value for a feature $a \in \mathcal{A}$ of the input x, coming from [29], is the following:

$$\phi_a(x) = \sum_{\mathcal{S} \subseteq \mathcal{A} \setminus \{a\}} \frac{|\mathcal{S}|!(|\mathcal{A}| - |\mathcal{S}| - 1)!}{|\mathcal{A}|!} \left[f_{\mathcal{S} \cup \{a\}}(x_{\mathcal{S} \cup \{a\}}) - f_{\mathcal{S}}(x_{\mathcal{S}}) \right] \tag{2.1}$$

where $\mathcal{S}$ is a partition of set $\mathcal{A}$ (the sum is built upon every possible partition of the feature set) and $f(\cdot)$ is the prediction function of the model we want to explain. The formula works by computing the SHAP values for each input feature of a given instance, which indicates the amount by which the feature affects the predicted output. To calculate such an influence, each marginal contribution of the feature is calculated as the difference between the prediction for feature set $\mathcal{S} \cup \{a\}$ and the prediction for the set $\mathcal{S}$. These marginal contributions are then averaged through a weighted mean over all possible partitions $\mathcal{S} \subseteq \mathcal{A} \setminus \{a\}$ to obtain the feature-specific SHAP explanation, that can be a positive or negative value basing on the type of influence of the feature on the prediction. The final SHAP explanation for a given input x is the set of all feature-specific SHAP explanations, thus written as:

$$\phi(x) = \left\{ \phi_a(x) \;\middle|\; \forall a \in \mathcal{A} \right\} \tag{2.2}$$

One of the key advantages of this technique is its ability to provide local and global explanations for individual predictions. Local explanations can be generated by calculating the SHAP values for each feature for a specific input instance, while global ones can be generated by aggregating the SHAP values across multiple samples or across the entire dataset. The global explanations can be visualized using various techniques, such as a feature importance plot or a dependence plot. Since its introduction, there have been many developments and variations of the SHAP technique. For example, one popular extension is the use of TreeSHAP [30], which provides a more efficient way to calculate the SHAP values for decision tree-based models. Another extension is the use of KernelSHAP [29], which provides a more accurate way to estimate the SHAP values for models with non-linear interactions between the features.

Robustness of Explanations. Although the aforementioned XAI techniques enjoy many desirable properties, including the near absence of the need to make modifications to existing models in order to be applied, they also have several limitations, mainly concerning their *robustness*. In [33], robustness is defined as the ability of a model to provide consistently correct or incorrect output, like the original output, when given slightly perturbed versions of the starting data point as input. When transposing this quality to XAI algorithms, many of them fail to produce explanations that respect this principle. To cite an example, in [26] is highlighted the inability of most of the saliency methods to be invariant to simple transformations. In [31], a taxonomy is defined concerning the analysis of the robustness of an XAI algorithm. In particular, three sources of instability in explanations are identified as causes of greatest interest, namely input perturbations, model changes and hyperparameters selection. Focusing on the former element, intuitively, robustness means that similar inputs must produce similar explanation values, that is, regardless of its method of representation, an explanation, in order to be considered valid, must remain (almost) constant in its vicinity [2].

It is important to specify that there is a substantial difference between the robustness of an ML model and that of an XAI algorithm. When we refer to the former, we usually want to measure the ability of a model, which has performed well in the training phase, to behave equally well in the deployment environment which often differs from the environment in which training data was gathered [38], that is, from a practical point of view, to enjoy more or less the same accuracy even during the testing phase. Concerning the second one, instead, we expect an XAI method to produce explanations that, on average, do not vary excessively (with respect to distance from the original input) within the neighborhood of a data point of interest, thus remaining consistent in the face of imperceptible changes in the input. It is important to specify that "imperceptible" means that the

perturbation does not lead to change in the prediction. Having said that, when we talk about robustness we typically refer to either *mean* or *adversarial* robustness. In evaluating the mean robustness of an explanation algorithm, Speaking of adversarial robustness, instead, we refer to the production of inputs that are specifically designed by an adversary to force a machine learning system to produce erroneous outputs [12], that is an input that *maximises* the error of the system. However, this type of robustness measurement appears to be more suitable for applications in the field of cybersecurity with regard to decision-making models, in order to assess the robustness of a model to adversarial attacks [37] (a scenario which hardly happens when we deal with model explanations). For these reasons, in this work to characterize the trustworthiness of XAI methods the choice fell on the mean robustness.

Random Forest Intrinsic Explainability. As already mentioned, Random Forest is one of the most widely used category of model, which boasts many successful applications in the field of ML. Concerning its intrinsic explainability, however, this decreases as the number of weak learners in the ensemble increases. Random Forest is able to provide a measure of the importance of each feature regarding the prediction, however this is not sufficient to consider the model transparent. It is, in fact, necessary to have a better understanding of how much each feature contributes to a certain label assignment and especially to understand it on a sample-by-sample basis. One way to overcome the inherent lack of explainability of Random Forest models is to apply XAI model-agnostic algorithms, such as SHAP, to obtain explanations that match the characteristics we are looking for. At [2] the authors observed that such an application yields values that only partially meet the expectations. Although the method provides the contribution of every feature for each data point in most cases the corresponding values have low robustness to small perturbations of the input. This can be interpreted as a symptom of a lack of trustworthiness of these explanations. On top of that, Decision Tree is a model with very good intrinsic explainability but by design it creates hard decision boundaries meaning that small changes in the input can lead to abrupt changes in the explanations. Random Forest relies on the combination of several weak learners to create a smoother decision boundary that better adapts to the real one. Hence, it is expected that the softer boundaries that provide a more robust model will also provide more robust explicabilities.

RF-specific Explainaibility Methods. We include a specific subsection to discuss the inherent explainability for Random Forest in order to provide a global overview of efforts done on XAI specifically for this technique. There have been several works devoted to the development of explanation techniques. The following are some examples of the

current state of the art in the domain of Random Forest explainability and, more generally, tree ensembles, mostly taken from [35]. Starting from methods of *simplification*, in [17] counterfactual sets are extracted from the model to create a more transparent version of the same, [16] poses the idea of training a less complex, and thus more intrinsically explicable, model on samples randomly extracted from the test set labelled by the ensemble, in [14] it is explained how to construct a Simplified Tree Ensemble Learner (STEL) on the basis of rules extracted from the ensemble and selected through feature selection and complexity-driven criteria, finally [23] presents as a solution that of training two models, one more complex (i.e. opaque) dedicated to prediction and a simpler one (i.e. transparent) that will be used to extract explanations, whose simultaneous use is governed by statistical divergence measures. Speaking, instead, of methods related to feature importance, among the earliest works on the subject we can find [10] and [3] in which the influence of a feature is analysed by means of random permutations of Out-Of-Bag (OOB) samples and measured through the use of metrics such as MDA (Mean Decrease Accuracy) and MIE (Mean Increase Error), which is followed by the work in [4] that makes use of feature importance measurement and partial dependency plots as tools to provide humans with information regarding the underlying learning processes in order to successfully extract knowledge from them.

Nevertheless, although these can be considered successful applications in the field of tree ensembles' explainability, in all cases it can be seen that robustness is not a property that is contemplated as a measuring instrument for their effectiveness. An XAI algorithm, in order to be considered robust, must be able to produce explanations that do not vary excessively as a result of small perturbations of the original data point.

3 | Development

In the following chapter, the work done on the design and conduction of the experiments and the data used are presented. By examining the literature, one can see that robustness is a property for which a measurement method has not yet been unanimously defined. Furthermore, among the various works that can be found on this subject, very few concern the specific application of the concept of robustness to model explanations. However, as shown in the coming sections, it was still possible to identify the most suitable metric for the conducted experiments by making use of the notion of *local Lipschitz continuity*, which was initially used as a starting point to formally assess the robustness of the classic model explanations, and subsequently as a key tool to conduct a comparative analysis between the latter and the developed procedures. The comparisons immediately showed that combining the explanations of the weak learners in an ensemble leads to the production of values that significantly improve this characteristic, so it was decided to explore such a path until the development of the procedure called AXOM (Averaging on the eXplanations Of the Majority), with which remarkable results were obtained. Before going into details with the various steps that led to the development of the solution, we include some specific paragraphs that can help the reader to better understand some critical parts of the work.

Benchmark models and method. We decided to conduct the experiments using Decision Tree and Random Forest as the reference models and SHAP as the reference XAI algorithm. The main reason behind this choice is that the simplicity and the authors' good confidence with these tools were key features in understanding the mechanisms behind the process of evaluating the robustness of explanations, given the novelty of the study. It was possible in this way to relate the results back to some well-known behaviors of the models and method, so as to provide better insights into the chosen approach. In addition, given the low computational capacity available at the time the experiments were performed, the speed of DT, RF, and TreeSHAP was crucial to be able to conduct the work in reasonable time.

Zones of explanation constancy. SHAP is a method designed in such a way that, when applied to Decision Tree (and, consequently, tree ensembles), as long as one feed the

algorithm with input data that activate the same decision branch, it will always provide constant explanation values. We will refer to the portions of space within which the XAI method outputs a constant value as *zones of explanation constancy*. When dealing with tree ensembles, two data points x_i and x_j do **not** fall in the same zone of explanation constancy if the two inputs activate two different decision branches in at least one of the weak trees of the ensemble. Note that this does not necessarily mean that the two input points produce a different prediction. Indeed, in Fig. 3.1 is shown a very simple example in which this kind of behavior is achieved.

2-Dimensional plots. In this work, heatmaps were produced basing the values obtained from 10000 perturbed x_j points, 100 for each row and 100 for each column. In order to make the representation in two dimensions possible (and **only for the purpose of the plots**), all graphical analyses were conducted by constructing the surroundings of the points of interest with perturbed points only along **two axes** of the feature space. In this regards, it is important to mention is that, for every heatmap produced in this work, the choice of the features in question was partially arbitrary, using as the only criterion a compromise between the need to analyse significant features and that of choosing axes along which the perturbations did not cause changes in predictions too frequently (which produce areas that do not fall within the cases of interest for the purposes of the robustness calculation, see Section 3.1 for a more detailed discussion in this regard), so as to provide clean and significant plots in their entirety.

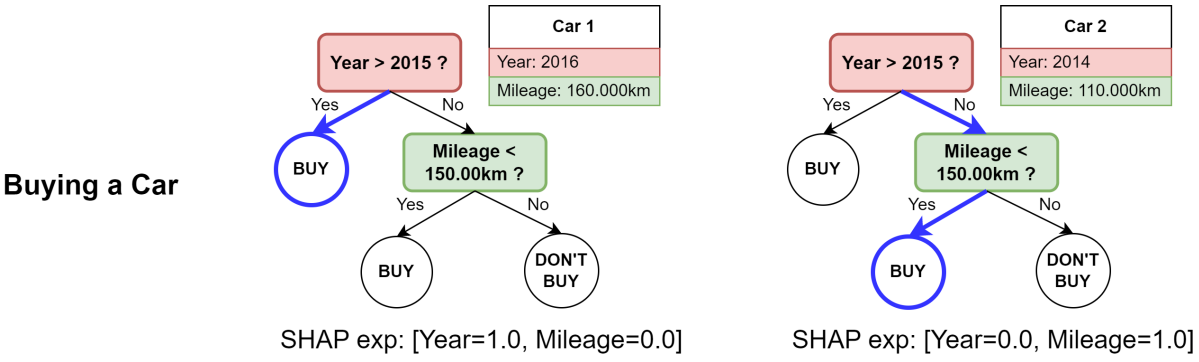

Figure 3.1: **Zones of explanation constancy** - As explained, as far as two data points activate the same prediction branches, and thus the same predicted label, they will produce the same SHAP explanation values, while two data points that activate different decision branches lead to the production of different SHAP explanation values. However, this latter case does not necessarily imply the fact that the predicted labels are different. This simple example illustrates how two inputs that share the same prediction label can lead to the production of two different explanations.

3.1. Robustness Metric Definition

As a first step, we need to rigorously define a method to quantify the robustness property of an explanation. The choice, basing on the work presented in [2], fell on the notion of *Lipschitz continuity*, in this case to be locally applied. It is, in fact, defined as follows:

Definition 3.1.1. *$f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is **locally Lipschitz** if for every x_0 there exist $\delta > 0$ and $M \in \mathbb{R}$ such that $||x - x_0|| < \delta$ implies $||f(x) - f(x_0)|| \leq M||x - x_0||$.*

Making use of this notion, the paper analyses, for each sample x_i of the test set, a circle $B_\epsilon(x_i)$ of radius ϵ of the data point in search of the maximum variation of the explanation value $\hat{L}(x_i)$, basing on the formula:

$$\hat{L}(x_i) = \max_{x_j \in B_\epsilon(x_i)} \frac{||g(x_i) - g(x_j)||_2}{||x_i - x_j||_2} \tag{3.1}$$

where the function $g(\cdot)$ of interest is the one implemented by the XAI algorithm, in our case the SHAP function expressed in 2.2.

The choice of searching for the maximum value, however, results in measurements that are unreliable for the purpose of a balanced and fair calculation of robustness around a given data point. The reason lies in the fact that, especially in the specific case where SHAP explanations are applied to a Random Forest model (composed by several weak Decision Trees), the values of interest are eclipsed by the peak measured in the area closest to x_i. This is because within each of the *zones of explanation constancy* (see discussion at the beginning of Chapter 3) **different** from the one to which the x_i point belongs, the value of fraction in (3.1) increases as the distance of x_j with x_i decreases, because of the denominator. In simple words, when $||x_i - x_j||_2$ has a very low value, but large enough to change decision branch even in only one of the weak trees in the ensemble, the value of $\hat{L}(x_i)$ diverges. Arguably, this is a very frequently encountered case, also basing on the number of estimators of which our ensemble is composed, just consider the case when the data point x_i around which we want to compute the robustness of our algorithm, is on the boundary of one of the above-mentioned zones. Fig. 3.2 shows, for each of the examined datasets, an example comparing the values obtained by first considering the difference (top) of the SHAP explanation values and then their incremental ratio (bottom). Specifically, in both cases each x_m point of the map represents a 2-dimensional version of the p-dimensional point x_j. Being a_x and a_y the two features that vary along

the two axes, we define the correspondence between x_j and x_m as:

$$x_j(x_m) = \{x_{j,1}, ..., x_{j,a_x-1}, x_{m,a_x}, x_{j,a_x+1}, ..., x_{j,a_y-1}, x_{m,a_y}, x_{j,a_y+1}, ..., x_{j,p}\} \tag{3.2}$$

In this way we can formally define the computation of the *heat* value of each point of the two maps. The explanations' difference map follows:

$$H_d(x_m) = ||g(x_i) - g(x_j(x_m))||_2 \tag{3.3}$$

while the explanations' incremental ratio map follows:

$$H_r(x_m) = \frac{||g(x_i) - g(x_j(x_m))||_2}{||x_i - x_j(x_m)||_2} \tag{3.4}$$

where, again, $g(\cdot)$ is the SHAP explanation function defined in (2.2). It can be observed that the tendency of the values in the incremental ratio is to tend towards infinity as one approaches the centre of the space, although the differences in SHAP explanation are sometimes negligible. This behavior is the consequence of the presence of the normalization factor $||x_i - x_j||_2$ at the denominator of (3.1), which value tends to 0 as x_j approaches x_i. This is reflected in the fact that using such a metric concentrate in penalize the models that creates a lot of explanation boundaries (especially if close to the x_i point) instead of assessing the real robustness within the neighborhood.

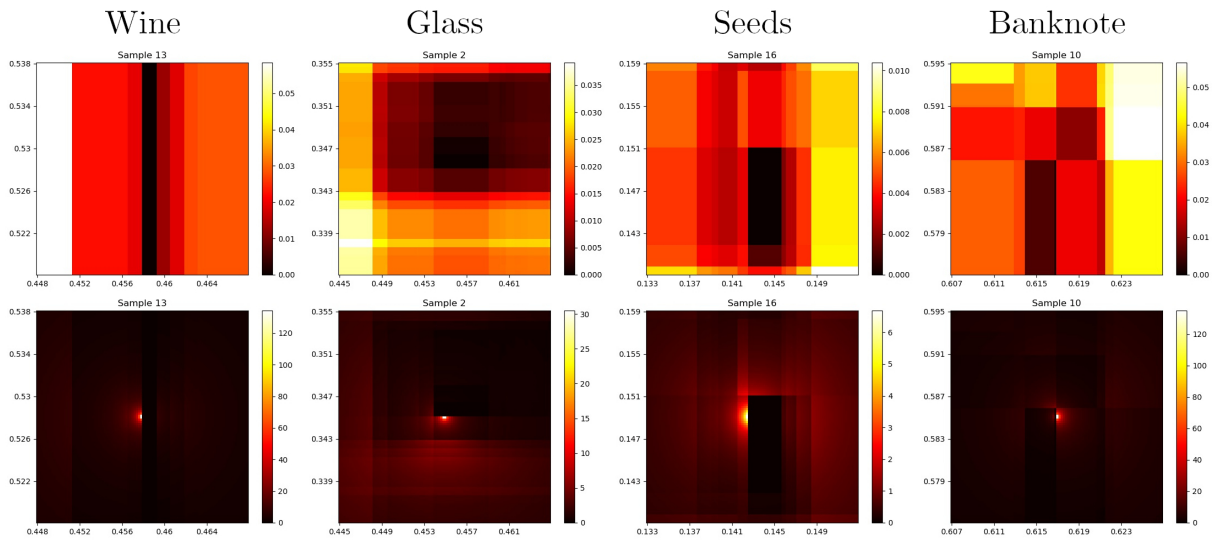

Figure 3.2: **SHAP explanations difference and incremental ratio heatmaps comparison** - The picture shows a comparison between the difference (top, see eq (3.3)) and the incremental ratio (bottom, see eq (3.4)) of the SHAP explanations produced by Random Forest in a neighbourhood of some x_i points of interest, on each of the four datasets. It can be seen that proximity to the centre of the search space is a factor that leads to the production of such high values that the differences between SHAP explanations in the surrounding areas are obscured. This shows that assessing the robustness of SHAP explanations through the maximum value of $L(x_i)$ is not a sufficiently fair method.

Nevertheless, when it comes to considering the difference in explanation values between one data point and another, it remains necessary to take into account the distance between these two. Intuitively, we expect that as the two considered points get further away, the explanations provided will also diverge more and more, and not balancing this growth would be "punitive" toward the more distant x_j points. Therefore, to avoid removing this contribution while trying to ensure that all points within the space surrounding x_i are covered in the robustness calculation, we modify the robustness criteria to calculate the **average** value of the incremental ratio between the explanation of x_i and the explanation of the x_j points around it. For this purpose, the definition of a discriminated finite-sample neighborhood is additionally provided. Let $\mathcal{X}$ denote the input space to which all x_i points belong, let $\mathcal{A}$ be the set of features of $\mathcal{X}$ and let $f(\cdot)$ be the prediction function of the model. Define, for every x_i sample of the considered test set, a discriminative discretization of its surrounding, in which points are evenly distributed, as:

$$\mathcal{N}_{f,\epsilon}(x_i) = \left\{ x_j \in \mathcal{X} \;\middle|\; |x_{i,a} - x_{j,a}| \leq \epsilon \;\forall a \in \mathcal{A},\; f(x_i) = f(x_j) \right\} \tag{3.5}$$

where, finally, $x_{i,a}$ indicates the value of the feature a of data point x_i. Given that, we now want to calculate the robustness of the SHAP explanations on data point x_i by means of the following formula:

$$\bar{L}(x_i) = \frac{1}{|\mathcal{N}_{f,\epsilon}(x_i)|} \sum_{x_j \in \mathcal{N}_{f,\epsilon}(x_i)} \frac{||g(x_i) - g(x_j)||_2}{||x_i - x_j||_2} \tag{3.6}$$

It can be seen that the computation of this value is significantly lighter than the one defined in (3.1) while preserving a good exhaustiveness of the analysis of the surrounding of the point considered. It is important to note that this function provides robustness values that are not particularly meaningful when taken individually, i.e., it is difficult to assess the quality of a result based on a set of values derived from a single model, as the values obtained may differ significantly when changing dataset or XAI method. In fact, the values obtained are only meaningful when used to conduct a comparative analysis between the robustness of explanations of different models applied to the same dataset with the same XAI method.

On top of that, as one can notice from the definition of the neighborhood in (3.5), we decided to include in the robustness calculation only the perturbed samples whose label predicted by the model is the same as the original sample. The reason behind this choice is that, intuitively, we only expect robust explanation values as long as these only account for a single output value. Indeed, it is reasonable to expect that a perturbed data point whose

label differs from that of the original data point will produce an explanation that differs substantially from that of the original data point, since different outputs are understood with different explanations. To fix the ideas, given $f(\cdot)$ and $g(\cdot)$, respectively the model's prediction function and explanation function, taken a point x_i and a perturbed version of it x_j, if $f(x_i) \neq f(x_j)$ then we expect $g(x_i)$ and $g(x_j)$ to also differ substantially. Thus, considering perturbed samples with a label different from the original one in the robustness calculation would lead to an unfair robustness calculation, which would reward algorithms that produce robust values when this property is undesirable.

3.2. Combination of Weak eXplanations

3.2.1. Why ensembles?

As explained earlier, the robustness of the explanations turns out to be a crucial property for them to be considered reliable and trustworthy. It is intuitively reasonable to think that if an ensemble of models performs well on a given dataset, it is then likely that most of the weak learners of which it is composed will contain extractable knowledge that can be considered useful in terms of explaining a certain decision. Indeed, it can be argued that producing explanations coming from the output of an ensemble of several weak learners is desirable for humans to understand a prediction, just think, in a practical case, of the tendency of many people to feel more confident about a medical opinion when it is the result of the union of the opinions of multiple experts. Taking into consideration that this type of aggregation has been widely shown to increase the robustness of the predictions, it is logical to expect that the same improvements will be observed on the explanations as well. Again, *zones of explanation constancy* in SHAP play a key role in justifying the cause of this improvement. Indeed, they can be identified as the cause of the abrupt changes in the SHAP explanations provided by Decision Tree models, since the low complexity of the decision branches structure brings to less frequent, although more sudden, change in explanation values, while, on the other hand, we may expect that explain ensemble outputs will lead to the production of explanations in which, due to the *overlapping* of weak SHAP explanations around the point of interest, changes in values are smoother, which is reflected in the absence of large differences in explanation values between two relatively close points. To prove that Random Forest produces explanations that are the result of such overlapping, we present here this notationally-adapted version

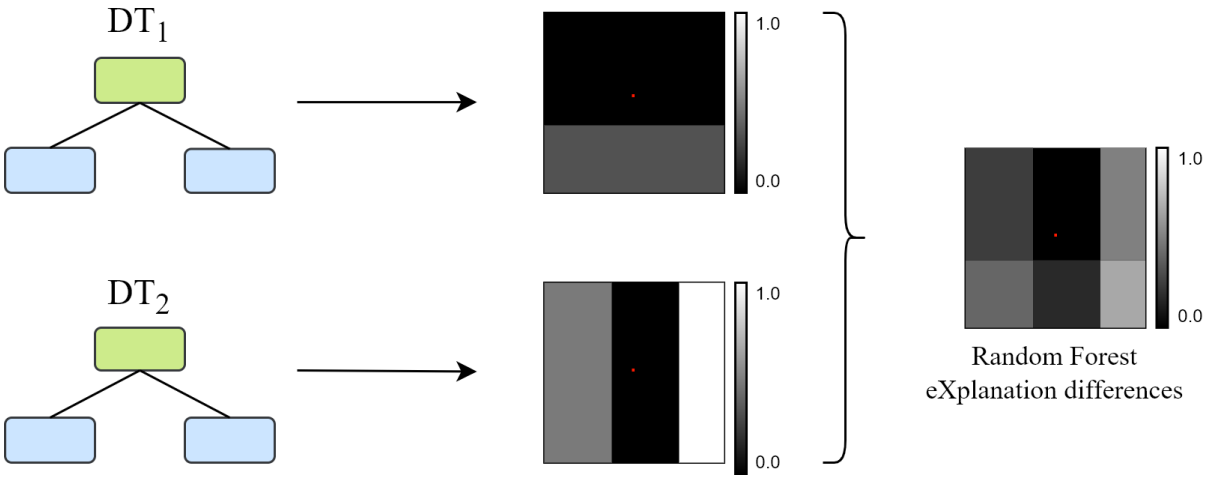

Figure 3.3: **From DT greymaps to RF greymaps** - A simple illustration of how the production of a RF greymap is influenced by the SHAP explanations of the individual weak learners that compose the ensemble. In this toy case, RF is an ensemble of only two DT models, which produce two greymaps that, combined through averaging, lead to the production of a greymap in which the progression of the SHAP explanation difference function as defined in (3.3) (the point in red in the center is x_i, while the neighborhood is composed by the x_j perturbed points) is significantly smoother, thus reflecting a more desirable behavior in terms of algorithm robustness.

of SHAP formula, for a multi-labelled classification setting:

$$\phi_{k,a}(x) = \sum_{\mathcal{S} \subseteq \mathcal{A} \setminus \{a\}} \frac{|\mathcal{S}|!(|\mathcal{A}| - |\mathcal{S}| - 1)!}{|\mathcal{A}|!} \left[f_{k,\mathcal{S} \cup \{a\}}(x_{\mathcal{S} \cup \{a\}}) - f_{k,\mathcal{S}}(x_{\mathcal{S}}) \right] \tag{3.7}$$

where $\mathcal{A}$ is the set of all the features of the dataset, $a \in \mathcal{A}$ is the feature for which we want to compute the explanation value and $f_{k,\mathcal{S}}(x_{\mathcal{S}})$ is a function, built upon a partition $\mathcal{S}$ of the set of feature, that returns 1 if, for the data point $x_{\mathcal{S}}$ (which is the data point x considering only the features in $\mathcal{S}$), the label k is provided by the model (note that this particular notation makes explicit the fact that we are provided with a set of explanations, one for each feature, that are *label specific*). By looking at the formula, one can see that the XAI function is linear with respect to the values predicted by the weak learners. For this reason, averaging the explanations produced by the weak learners provides the same value as directly applying the explanation algorithm to the entire ensemble which, for its part, produces explanations based on predictions that are the result of averaging the predictions of the decision trees that compose it. See Appendix A for further details.

Furthermore, to illustrate the implications of such behavior, we present two instances: a very simple hand-crafted example in Fig. 3.3 and a practical example over the case study datasets in Fig. 3.4. These two pictures show a comparison built basing on the variation of the value of the explanations in the neighbourhood of certain points of interest x_i (the central point of the square maps), computed through the formula defined in (3.3) for each x_j belonging to the neighbourhood of x_i. Indeed, what can be seen is that

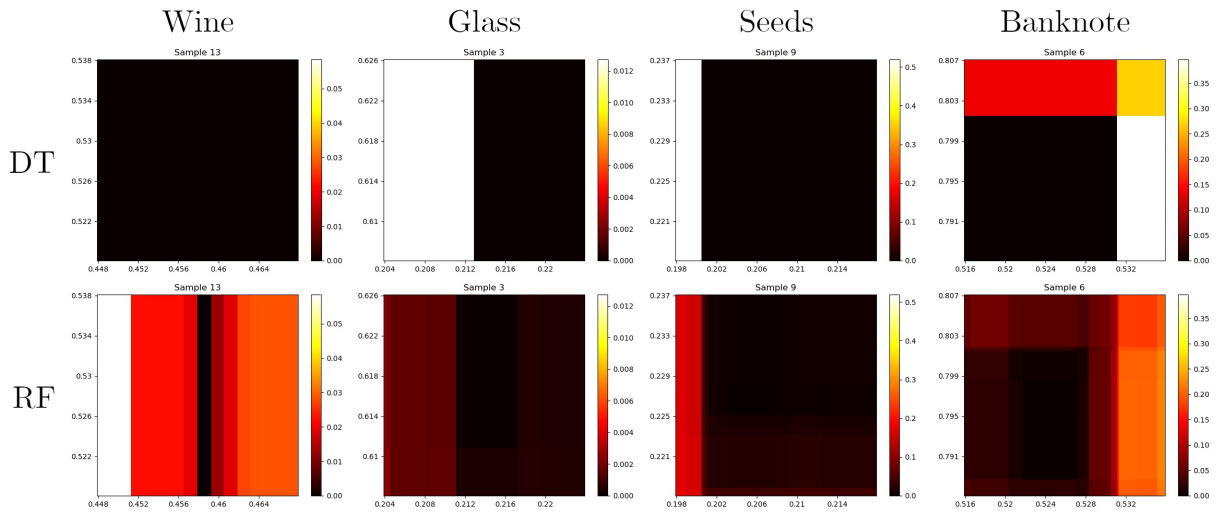

Figure 3.4: **DT-RF comparison between heatmaps of SHAP explanations difference** - Comparison between DT and RF regarding differences in the SHAP explanation values (see eq (3.3)) around some points of interest of the different datasets. For all four datasets analyzed, it can be seen that RF, w.r.t. DT, provides SHAP explanations that, while suffering from a smaller presence of areas where they are constant, enjoy a smoother progression in the values which is reflected in SHAP explanations that are overall more stable to perturbations with no abrupt numerical differences for small perturbations.

the greymaps and heatmaps related to the explanations produced by Random Forest depict more gradual color changes, which correspond to differences in values that follow a smoother (thus more desiderable) progression.

3.2.2. Why combine Weak eXplanations?

Nevertheless, obtaining this behavior by means of the direct application of XAI algorithms to model ensembles leads to the production of explanations that are smoother but do not provide significant improvements on the robustness. On top of that, it comes natural to wonder the reason behind this behavior. One way to try to overcome this problem could be to exploit some form of combination of the individual weak models explanations. In this regard, the trivial solution is to combine the contributions of the individual models through a simple average. However, as we already shown, when it comes to linear explanation algorithms like SHAP the simple average combination does not produce any variation in the original explanation values with respect to the straight application of the XAI method to the ensemble.

By intuition, a probable reason behind the lack of significant improvements in Random Forest explanations is that progressively more complex models, though often better performing, still tend to be more sensitive to the noise generated by the explanations of the weak learners that provided a wrong label. Consequently, it could be argued that in a complex ensemble in which outputs are generally produced by contributions from

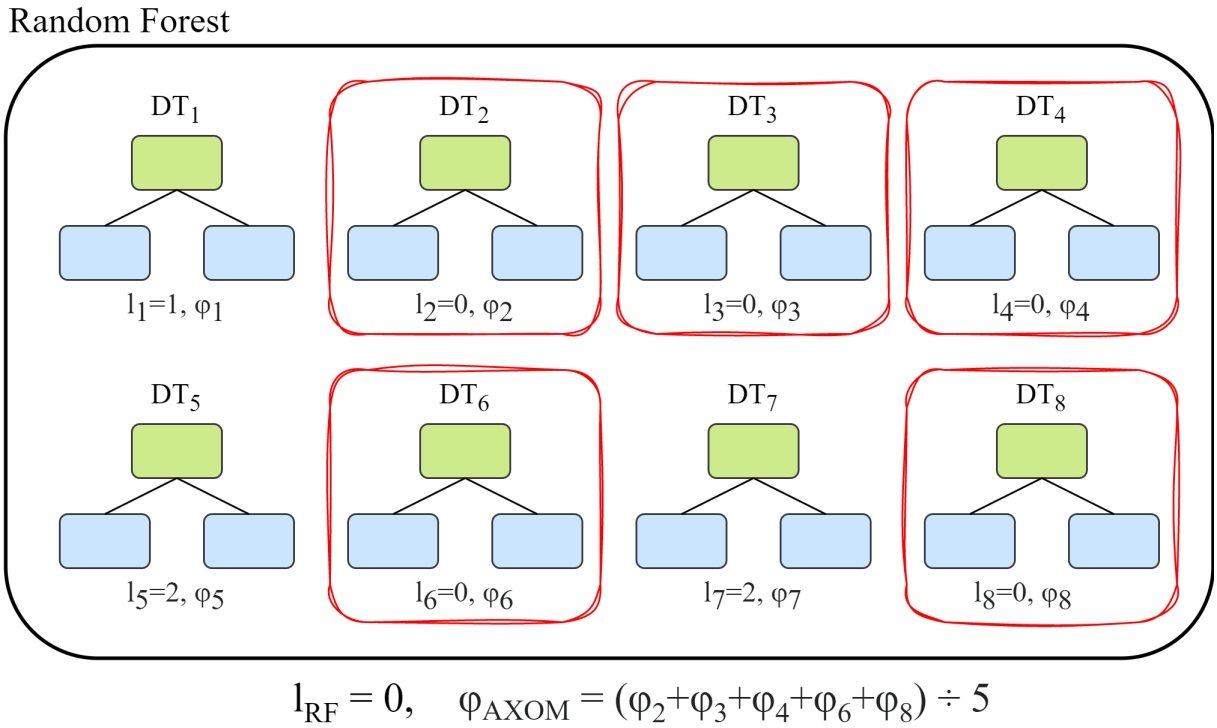

Figure 3.5: **AXOM functioning illustration** - A toy example to illustrate the functioning of AXOM: the ensemble (RF) consists of eight weak learners (DT) each of whom casts a vote l_i on the prediction of the input, with associated explanation ϕ_i. Ensemble output is chosen according to a majority vote, in which $l_{RF} = 0$ wins. AXOM, to generate the output explanation, considers averages only the explanations of the weak learners who were part of the majority, hence expressing a prediction consistent with the final ensemble prediction.

the individual models that are widely diversified, it may be reasonable to "reward" those explanations from the models that contributed positively to the final decision. In simple terms, in the case of classification problems, a non-trivial variation of the above-proposed solution is to consider in the averaging only those weak explanations that come from the weak learners who provided as output the same label as the ensemble output. Intuitively, a weak learner who contributed positively to the final ensemble output will be able to provide more relevant explanations regarding the decision made. The next section will be devoted to an in-depth analysis of this solution, called AXOM, including the implications of such a form of combination.

3.3. Averaging on the eXplanations Of the Majority (AXOM)

The high explicability of the weak learners that make up the ensemble is certainly the most appealing property when it comes to robustness of explanations. However, it must be considered that XAI algorithms such as SHAP, that are based on model output values,

are highly dependent on the prediction accuracy for a given dataset. Although Decision Tree's transparency positively affects the trustworthiness of explanations, the (relatively) low accuracy in the predictions of weak learners is reflected in the production of explanations that are less robust. As mentioned above, combining the explanations of individual learners of an ensemble can help improve this quality. Specifically, we consider only to combine the explanations of weak learners that *positively* contributed to the ensemble output, while by "positively" we mean that the classification of the weak learner matches with the obtained by the random forest ensemble (see Fig. 3.5 for a more clear graphical explanation).

In algorithm 3.1 the AXOM evaluation algorithm for each data point x is presented. The method receives as parameters the ensemble e (a Random Forest trained model), the data point x and the SHAP explainer σ. $\phi_w \in \mathbb{R}^{1\times p}$ contains the SHAP explanations of the weak learner w, being p the number of features, and an explanation is added to $\Phi \in \mathbb{R}^{n\times p}$, being n the number of selected weak learners, if the label provided by the weak learner l_w is equal to that predicted by the ensemble l_e. The final explanation $axom\_shap \in \mathbb{R}^{1\times p}$ is the mean of all selected weak explanations Φ for sample x.

Algorithm 3.1 AXOM procedure to calculate single-sample explanations for an ensemble

procedure AXOM_SHAP_EXPLANATION(e, x, σ)
  $l_e \leftarrow e$.PREDICT(x) ▷ Ensemble label prediction for x
  $W \leftarrow e$.estimators ▷ Store the ensemble's weak learners set
  $\Phi \leftarrow$ **new** LIST( )
  for w **in** W **do**
    $l_w \leftarrow w$.PREDICT(x) ▷ Weak learners label prediction for w, x
    if $l_w = l_e$ **then**
      $\phi \leftarrow$SHAP_EXPLANATION(w, x, σ) ▷ SHAP weak explanation for w, x
      Φ.APPEND(ϕ)
    end if
  end for
  $axom\_shap \leftarrow \frac{1}{|\Phi|}\sum_{\phi\in\Phi}\phi$ ▷ Mean of SHAP weak explanations
  return $axom\_shap$
end procedure

This method ensures that the obtained explanation is free from the noise resulting from the explanations of the weak learners that provided a different label from the ensemble. Arguably, this improves the quality of the explanation only in the case where the ensemble has provided correct output. In this regard, when a sample-specific explanation is produced, we expect to obtain data that support the decision that was actually made. Such information is most clearly extractable from the majority weak learners, which makes the method useful for the purposes of understanding what led to that decision.

3.4. Datasets

We decided to test the methods on four commonly used datasets from **UCI Machine Learning Repository**. Specifically, the data used as a benchmark came from the following:

- **Wine** [1]. The data come from the results of a chemical analysis of wines derived from 3 different cultivars.
- **Glass Identification** [20]. The data regard the study of classification of 7 types of glass.
- **Seeds** [13]. The examined group comprised kernels belonging to 3 different varieties of wheat.
- **Banknote Authentication** [28]. Data were extracted from images that were taken from genuine and forged banknote-like specimens (2 classes).

In Table 3.1 some other specific information as well as the accuracy of the tested models are specified. All of these datasets address a classification task with multivariate data points. Features of all datasets were entirely numerical, mostly real-valued, with some integer exceptions. To ensure consistency and comparability across all variables, datasets were standardized with values between 0 and 1 using a min-max scaling method, in order to reduce the potential bias in analysis and facilitate the interpretation of results across multiple variables. In this way, influence of each variable in the analysis was equalized, allowing for the interpretation of robustness results on a common scale. It is important to mention that among these four datasets, each one enjoys a good balancing of the classes, except for Glass Identification. Indeed, classes 0 and 1 alone represent almost the 70% of the samples, while there are no samples at all belonging to class 4.

Datasets	N. of features	Training set size	Test set size	Accuracy	
				DT	RF
Wine	13	160	18	88.9%	100%
Glass	10	192	22	81.8%	95.5%
Seeds	7	189	21	85.7%	95.2%
Banknote	5	1234	138	98.6%	99.3%

Table 3.1: Descriptions of the datasets with accompanying information on DT and RF performance on them.

On top of this, as one can notice, the number of features of the tested datasets was limited to 13. The reason behind this choice is related to computational power needs. In particular, given the choice of using 10000 points to evaluate the robustness around the neighbourhood area of interest, datasets with at most $\lfloor \log_2(10000) \rfloor = 13$ features were

chosen in order to guarantee an adequate search in the entire feature space, that is, with at least two perturbations along each feature axis.

3.5. Experimental Design

With regard to the conducted tests on robustness, the choice of the radius of the neighbourhood of the data points of the test set to be analysed was $\epsilon = 0.01$. This value defines the perturbation area to be analyzed and it is constant for all the experiments. It is important to mention that all data samples are normalised to 0-1 range and thereof the perturbation is of 1%. The same experiments were done for Decision Trees and Random Forest models. Both were trained on each of the four above-mentioned datasets in order to, by means of a brief grid-search validation, obtain the best model based on accuracy metric. Two functions were then defined for calculating the value of $\bar{L}$, one that performs this calculation through the explanations obtained directly from the Decision Tree and Random Forest models and one that performs it on Random Forest through the previously defined AXOM algorithm. For each data point x_i of the different test sets 10000 perturbed x_j samples are randomly generated, on which the variation of the explanation value is calculated through the formula in (3.6). In algorithm 3.2, the method used for calculating the robustness of the explanations of a generic sample x_i is reported.

Algorithm 3.2 Procedure to calculate mean robustness of explanation for a sample x_i

procedure COMPUTE_MEAN_ROBUSTNESS(e, x_i, σ)
  $\epsilon \leftarrow 0.01$
  $n_{points} \leftarrow 10000$
  $g_{x_i} \leftarrow$ EXPLAIN(e, x_i, σ) ▷ Explanation of model prediction for sample x_i
  $M \leftarrow$ **new** LIST( )
  $\mathcal{N}_{f,\epsilon} \leftarrow$ GRID$(x_i, \epsilon, n_{points}, e)$ ▷ Neighborhood of 10k points with same label as x_i
  for x_j **in** $\mathcal{N}_{f,\epsilon}$ **do**
    $g_{x_j} \leftarrow$ EXPLAIN(e, x_j, σ) ▷ Explanation of model prediction for sample x_j
    $\mu \leftarrow \frac{||g_{x_i} - g_{x_j}||_2}{||x_i - x_j||_2}$ ▷ Incremental ratio between x_i and x_j exps
    M.APPEND(μ)
  end for
  $\bar{L} \leftarrow \frac{1}{|M|} \sum_{\mu \in M} \mu$ ▷ Overall robustness: average of all x_j contributions
  return $\bar{L}$
end procedure

The method receives as parameters the model e (Decision Tree or Random Forest), the data point of interest x_i and the SHAP explainer σ. First of all, the search parameters *epsilon* and n_{points} are fixed (in the algorithm we reported the values used in our experiments) and the SHAP explanation g_{x_i} for point x_i is calculated. After that, we build a

grid $\mathcal{N}_{f,\epsilon}$ around the data point x_i which is going to be used to carry out the evaluation of the robustness in the neighborhood (we recall that the neighborhood is composed by the x_j point with same label as x_i, see (3.5) for its definition). For each $x_j \in \mathcal{N}_{f,\epsilon}$ the SHAP explanation g_{x_j} is produced and used to evaluate μ, that is the variation of explanation value, as defined in (3.6). Finally, the mean robustness is calculated as the average of all the μ values. Note that this procedure is valid for both classic and AXOM SHAP explanations. Indeed, the method EXPLAIN refers in a general way to the method for which one needs to asses the robustness.

4 | Results

Robustness comparison. Table 4.1 shows the $\bar{L}$ results in the form of mean and standard deviation for each model and dataset, while Fig. 4.1 present them in a more detailed way by means of box plots (**Note**: $\bar{L}$ indicates the variation of explanation values in the neighborhood, thus a lower mean value is associated with a higher robustness). It is possible to observe from the mean and standard deviation values that the AXOM procedure provides explanations that on average are more robust in each of the four analyzed datasets compared with RF. However, when comparing the robustness values of AXOM with those of Decision Tree, the former provides significant improvements only in the datasets Glass and Banknote, while in Seeds the same average robustness can be observed (although the standard deviation values are indicative of a better reliability of AXOM explanations) and in Wine DT overperforms AXOM, enjoying a seemingly perfect robustness. In the later parts of this chapter we will analyze in detail the reasons for these anomalous behaviors.

Model	Wine		Glass		Seeds		Banknote	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Decision Tree	0.00	0.00	1.89	1.78	0.65	2.02	1.75	3.32
Random Forest	0.55	0.51	1.75	1.87	0.77	0.72	1.58	1.57
AXOM	0.47	0.44	1.27	0.72	0.65	0.67	1.28	1.34

Table 4.1: Mean and standard deviation of the $\bar{L}(x_i)$ values calculated for each sample x_i of the various test sets. Lower values of $\bar{L}$ denote better robustness to perturbations.

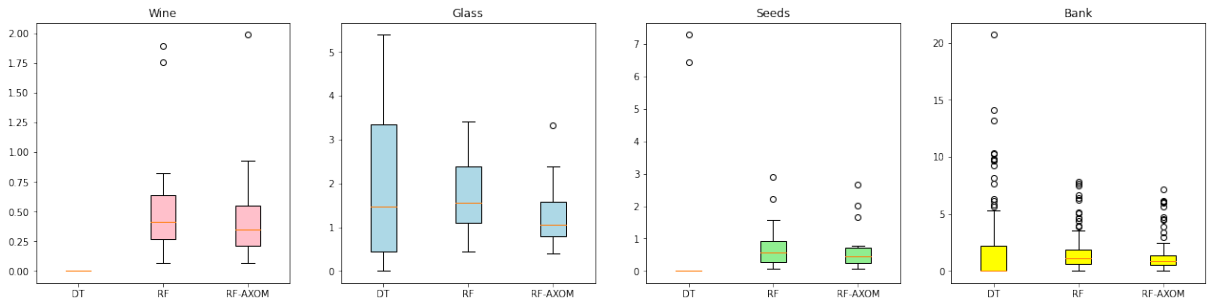

Figure 4.1: **Robustness comparison through box plots** - Box plots constructed from the $\bar{L}(x_i)$ values of the x_i samples of the test sets of each of the four analysed datasets. By isolating the outliers, it can be seen that in all cases the AXOM box plots represent a significant improvement in the robustness values with respect to the global RF while regarding DT, except for the Wine dataset on which it presents an apparently perfect behaviour due to the decision boundaries too far from the analysed test points.

Statistical analysis. To verify the reliability of these results, two-sample t-test was used (One-tailed Student's T-tests were carried out for dataset BANKNOTE, in which population is over 30 samples and therefore considerable as normally distributed, while Wilcoxon's T-tests was carried out for the others, since the distributions of the series were not normal). This made it possible to assess the probability that the improvement in mean robustness was due to chance. As we expected, Table 4.2 shows that AXOM significantly improves robustness over RF for all datasets (see row RF vs. AXOM), with p-values all below the 0.05 threshold. Regarding the DT vs. AXOM comparison, neglecting for the moment the case of the WINE dataset which we will discuss later, it can be observed that the equality in mean robustness in SEEDS is nevertheless reflected in a statistical improvement in favor of AXOM, which possesses values deviating from the mean with less magnitude. Indeed, Fig.4.1 shows that for this dataset the box is squeezed on zero, but there are outliers with a very larger value with respect to the other two models. What can be observed, still, is that the DT vs. AXOM comparison always gives better results than the DT vs. RF comparison, again delineating the better effectiveness of the procedure.

Comparison	**p values**			
	WINE	**GLASS**	**SEEDS**	**BANKNOTE**
DT vs. RF	<0.001	0.774	0.008	0.3023
DT vs. AXOM	<0.001	0.113	0.008	0.0656
RF vs. AXOM	0.042	0.007	0.030	0.0444

Table 4.2: Two samples mean T-test values when comparing the robustness of RF, DT and AXOM, for each dataset.

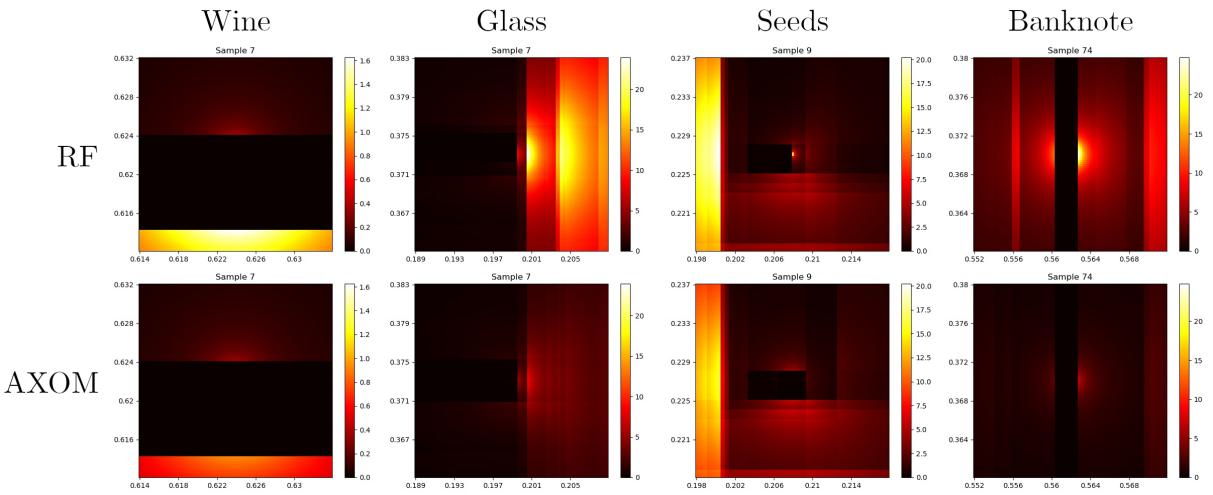

Figure 4.2: **RF-AXOM Robustness comparison through incremental ratio heatmaps** - Sample-specific comparison between the robustness heatmaps of explanations (see eq (3.4)) of RF (top) and AXOM (bottom). RF and AXOM produce explanations that vary proportionally equally, except for the absence of some boundaries in AXOM due to the absence of some weak learners' explanations in the average. The ranges of values of heatmaps produced by AXOM, are smaller than those of RF.

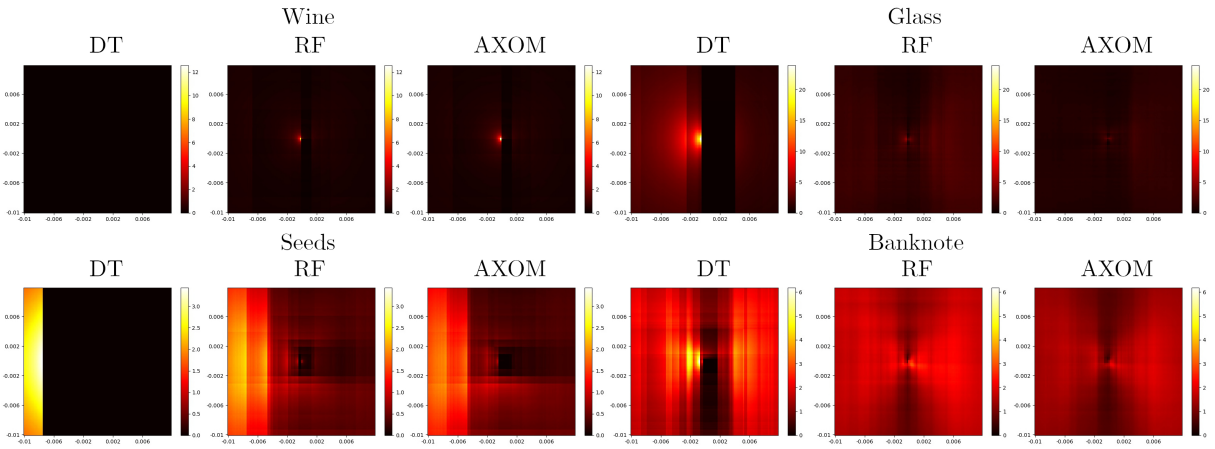

Figure 4.3: **DT-RF-AXOM full test set Robustness comparison** - Comparison of the robustness heatmaps of the explanations (see eq (4.2)) for DT, RF and AXOM for the entire dataset. The plot was produced by centering in (0, 0) all the x_i samples in the dataset so that all the samples could be fit into the same box of size $2\epsilon \times 2\epsilon$. It is clear from the plots that, in general, RF and AXOM enjoy better smoothness and robustness in value changes than DT, with AXOM in particular possessing darker plots in color than RF, indicative of better average value robustness.

Heatmaps plots. To get some insights about the results, we present two graphical illustrations by means of heatmaps constructed according to the incremental ratio of explanation values within the neighborhood of the x_i points of interest.

Fig. 4.2 shows the heatmaps of the robustness of the neighborhood of all four test sets, focusing on RF (top) and AXOM (bottom), taking a representative sample from each of the four datasets as an example. Each $H(x_m)$ value of the map is computed as defined in (3.4). While being not completely general, with this plot the improvement in robustness brought by AXOM can be clearly appreciated. One can see that the two sets of SHAP explanations vary proportionally very similarly with each other around the x_i points, with the only difference represented by the fact that the heatmaps of AXOM exhibit darker colors, indicating lower value variations as well as more desirable behavior.

Fig. 4.3 shows the same type of comparison (including also DT), but this time by overlapping (through averaging) the heatmaps of all test samples. To be more specific, being $\mathcal{T}$ the set of all test points of a given dataset, all $x_i \in \mathcal{T}$ test samples were centered in (0, 0) so that all samples could be fit in the same box with axes bounded inside $(-\epsilon, \epsilon)$, in which the x_m points vary. As explained in (3.2), there is a correspondence between x_j and x_m, this time slightly differently defined as:

$$x_j(x_m) = \left\{x_{j,1}, ..., x_{j,a_x-1}, x_{j,a_x} + x_{m,a_x}, x_{j,a_x+1}, ..., x_{j,a_y-1}, x_{j,a_y} + x_{m,a_y}, x_{j,a_y+1}, ..., x_{j,p}\right\} \tag{4.1}$$

Making use of this definition, the *heat* value in correspondence of each generic x_m point

of the map is given by the following formula:

$$H(x_m) = \frac{1}{|\mathcal{T}|} \sum_{x_i \in \mathcal{T}} \frac{||g(x_i) - g(x_j(x_m))||_2}{||x_i - x_j(x_m)||_2} \tag{4.2}$$

It is possible to see from the colors of the plots that RF and AXOM (except for the usual anomalous case represented by the WINE dataset) always exhibit more desirable behavior than DT, with significantly smaller explanation values that vary considerably more smoothly. Comparing RF vs. AXOM, also in this case one can evidently delineate a similarity in explanation variance but with much smaller value ranges for the latter.

DT explanations' robustness. Before understanding the reasons of the apparently good robustness of DT, it is worth to recall that the accuracy of the RF model is better than DT (see Table 3.1), and so intuitively also the expected quality of the explanations. Having said that, AXOM improves robustness for all datasets except for WINE dataset, when compared with DT. This is due a fortuitous behavior of the Decision Tree model for the experimental design parameters. Indeed, it can be observed from Table 4.1 that the obtained robustness is 0. That would mean that the robustness is "perfect", i.e. all the SHAP explanations for the perturbed data have exactly the same value as the original data point. However, the production of explanations that are constant over a large portion of the feature space is only desirable behaviour if the problem to be explained is simple, which represents a contradiction, as the need for explanations grows with the complexity of the problem. This explains the need to construct explanations in such a way that they are capable of modelling more complex behaviour and thus, analogous to matters concerning the accuracy of a model, justifies the possibility of constructing "ensembles of explanations". To recall the constancy of the explanations, at Fig. 3.4 can be observed that for WINE dataset the differences in explanations of DT are zero-valued (and so explanation values are constant). This happens when all the perturbed points belonging to the neighborhood fall into the same branch of the sample under analysis, meaning that the decision surface of the DT branches have a large margin between them, or at least larger than the perturbation parameter that is in our case $\epsilon = 0.01$. This can be also appreciated for the other datasets (GLASS, SEEDS, and BANKNOTE) by comparing the robustness values obtained by RF and DT on the same datasets and samples. Indeed, what can be deduced from the box plots is that DT tends to provide explanations that enjoy perfect robustness only in samples that are sufficiently distant (i.e. far enough away not to be affected by the perturbations) from the decision surfaces where a branch change occurs, whereas for data points that are more "unlucky" in this respect, the value of $\bar{L}$ is significantly larger, meaning that DT is not able to provide SHAP explanations that

smoothly change inside the neighborhood of those samples. An example of this behaviour is shown in Fig. 4.4 through two representative samples of WINE test set. Specifically, by setting $\epsilon = 0.2$ (in the experiments $\epsilon = 0.01$ was set), and thus enlarging the range of points plotted in the heatmaps concerning the difference in explanations as defined in (3.3), it can be seen that DT provides explanations that, although enjoying (apparent) perfect robustness in a relatively large neighbourhood, suffer abrupt changes in value as soon as a change in the decision branch is reached, with values significantly higher than those of AXOM. We recall that in DT a change of decision branch does not necessarily imply a change of predicted label. In fact, according to the algorithm used to produce the heatmaps, if such areas appeared in this plot they would be characterized by the color grey.

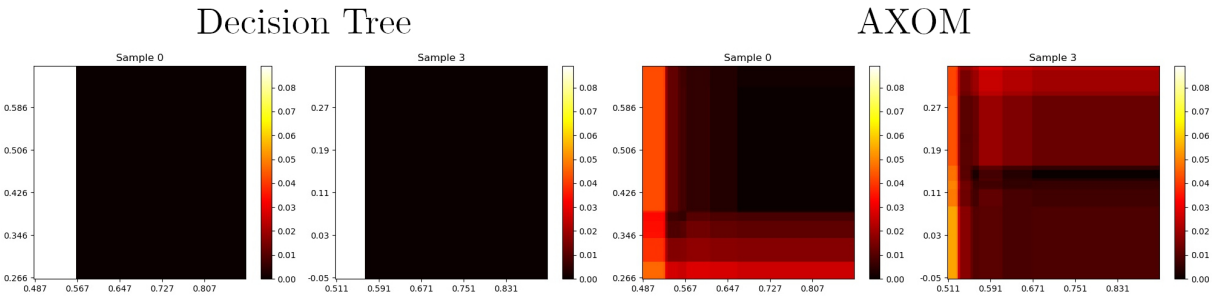

Figure 4.4: **DT-AXOM explanations smoothness comparison in Wine dataset** - Comparison of DT and AXOM explanations difference heatmaps for two representative samples of WINE dataset. The plots were produced using a maximum perturbation $\epsilon = 0.2$ to show that tweaking a data point enough to change decision branch of DT, results in abrupt value changes in the SHAP explanations produced (even if there is no change in the prediction label, in which case we would observe a grey area), thus proving that DT only enjoys *apparent* good robustness. In contrast, AXOM, while suffering value changes even in the face of smaller perturbations, produces explanations that vary more smoothly, following a more desirable behaviour.

SHAP values comparison. Finally, we show in detail the values resulting from the production of explanations through the classical method and AXOM. Fig. 4.5 shows the explanations produced from the predicted labels of four representative samples of the different test sets, while Fig. 4.6 shows the multi-label explanation values of the entire test sets. As one can observe, RF and AXOM tend to distribute the "merit" of the produced output across all features, confirming the fact that such explanations are capable of modeling and explaining more complex behaviours, compared to DT which tends instead to load as much responsibility as possible into fewer features. At first glance, the values of the full test-set explanations of RF and AXOM appear to have high similarity, especially when compared with those produced by DT. To identify the reason for this behavior, it is necessary to keep in mind that AXOM's explanations are the result of de-noising RF's explanations, where the noise is represented by the explanations of the weak estimators that provided an incorrect label (considering the ensemble prediction as ground truth). Clearly, the fewer components that provide a prediction different from

the ensemble prediction, the more similar the RF and AXOM explanations will be to each other. In this regard, Table 4.3 shows the average percentage of weak learners who provided a label different from the ensemble on the test data, i.e., the percentage of weak explanations discarded by AXOM. It is possible to see that Glass is the dataset in which there is more indecision within voting, while in Bank we observe behavior tending toward voting unanimity, which is reflected respectively in a lower and higher similarity of explanations, as observable in Fig. 4.6. In fact, although at first sight the differences are not easy to detect, we can see, for example, that in Glass the explanations of classes 2 and 4 differ significantly between RF and AXOM for almost all features, in addition to the fact that the ranges of values are higher in the case of AXOM. Seeds and Wine on the other hand, as expected from the values, exhibit intermediate behavior. On top of that, once the reasons for these behaviors are identified, the fact that the AXOM explanations are the result of combining explanations of weak learners who "guessed" the prediction makes the hypothesis that these values are more credible in explaining an output reasonable.

	Wine	**Glass**	**Seeds**	**Banknote**
Weak Mislabeling Percentage	12.1%	24.9%	14.0%	2.5%

Table 4.3: Weak learner's mislabeling percentage, for each datasets. The values represent the average number of weak learners that cast a vote different with respect to the final ensemble prediction

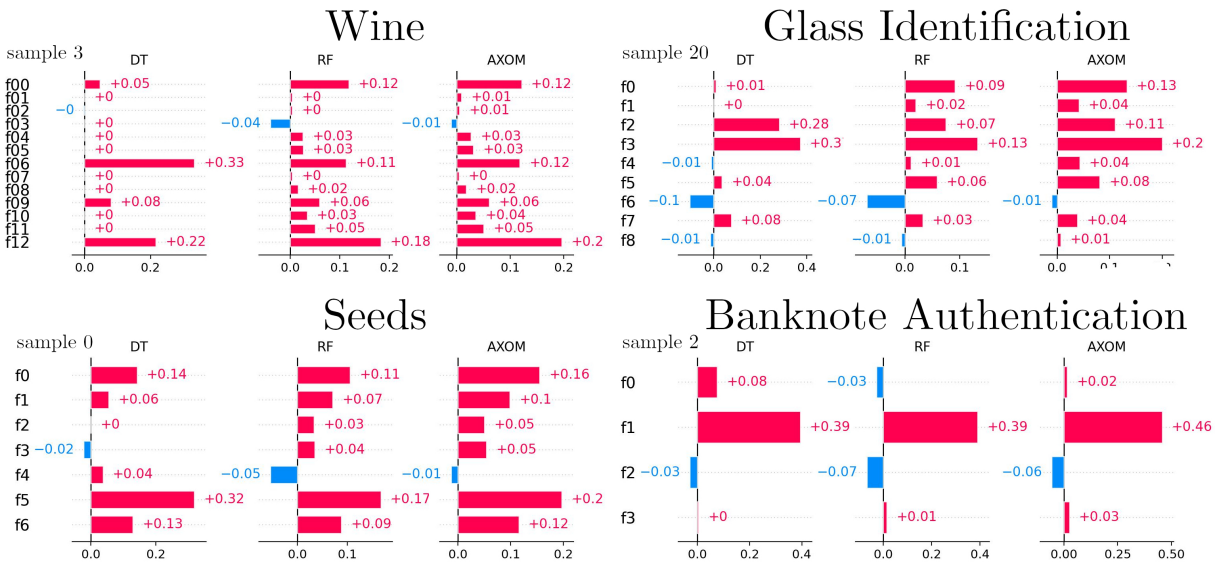

Figure 4.5: **Sample-based SHAP values comparison** - Comparison of the SHAP values produced by DT, RF and AXOM to explain a representative sample of each of the datasets. It can be seen that DT substantially differentiates the values between the various features, while RF and AXOM tend to distribute the responsibilities of the output more widely, while retaining significant differences between their values.

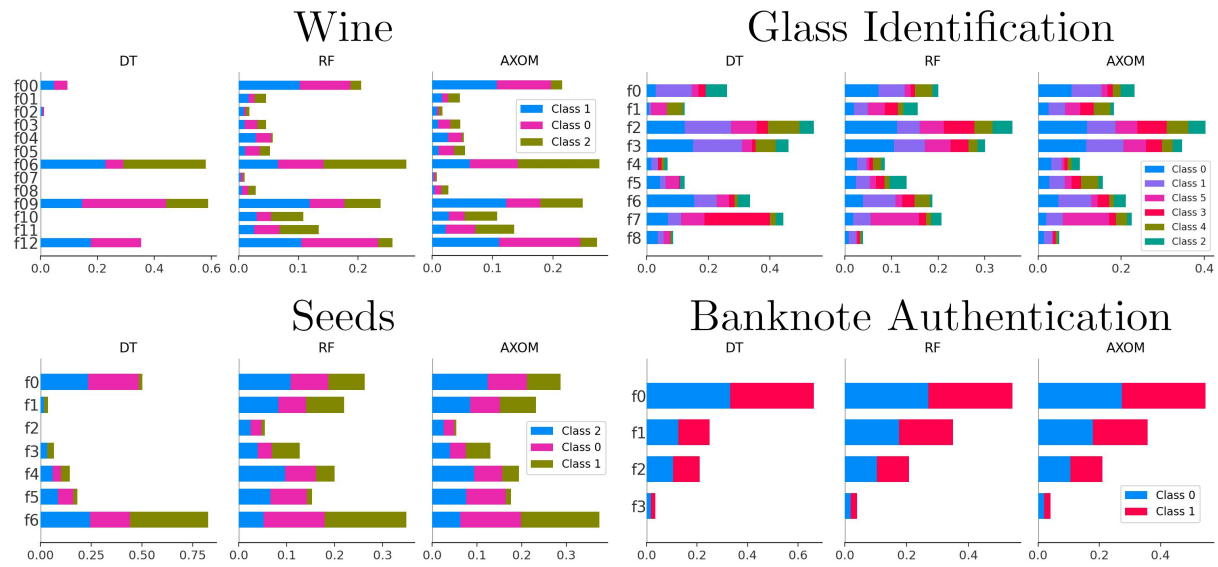

Figure 4.6: **Full test set SHAP values comparison** - Comparison of the multi-output SHAP explanations produced by DT, RF and AXOM, for each of the datasets. As for the single-output explanations of the individual samples, it is possible to see a substantial difference between DT explanations and those of the ensembles, which are more similar. This is a symptom of a radical difference in the interpretation of the test set by DT and RF, while showing that AXOM, although producing more robust values, allows similar knowledge to be extracted as RF.

5 | Conclusions

There is a growing concern about the reliability of the explanations offered by some XAI methods. This concern is also linked to the need to build trust in artificial intelligent systems that can be integrated into our way of life, thus directing studies towards improving the trustworthiness of models. By analysing the elements on which such characteristic is based, robustness of explanation values was identified as a pivotal property. In this work, in order to steer progress in this direction, we first presented as a solution the establishment of an unambiguous and justifiably fair criterion for measuring the robustness of model explanations, based on the assessment of the variation of explanation values around a point, and then proposed a procedure for calculating the SHAP explanations of model ensembles as the result of averaging the explanation values of weak learners who contributed positively to the final prediction. This approach has proven to be a method that significantly improves the robustness of model explanations compared to explanations obtained through the direct application of XAI methods to the ensemble under consideration. We can confirm that weak learners, who enjoy greater explainability than the complex model, taken individually can play a key role in explaining the decisions of the ensemble to which they belong. In particular, the application of a combination of individual weak explanations may lead to the production of more robust global explanations through the reduction of variance in the explanations, achieved by a selection of weak learners who provided a prediction consistent with the global output (producing explanations that were consequently consistent with each other), thus eliminating noise deriving from the explanations of weak learners who provided a different prediction, thereof considered by the ensemble as incorrect. We envisage that this approach is not limited to Random Forest and SHAP and that it is natural to extend it to other types of ensembles, such as Bagging or Gradient Boosting, as well as to other post-hoc XAI techniques.

6 | Future Work

The development of explanation methods that produce more robust, and therefore reliable, values could gradually encourage the deployment of Machine Learning tools in fields where they currently struggle to find possibilities for use. The fact that the reliability of results is still a major shortcoming of the models despite their high accuracy, should prompt scientists to spend increasing efforts on researches in the field of eXplainable Artificial Intelligence.

On the other hand, the development of this work and the achieved results pave the ground for future investigations concerning the reliability of XAI algorithms applied to model ensembles. It has been shown that aggregation can be used to produce "ensembles of explanations" which are characterised by more robust values than those produced by applying the methods directly to the global ensemble. However, although the work presents a successful application, there are a multitude of research paths to follow in order to improve the work, as well as make it more general.

Extension of the Analysis. It was possible to observe the performance of AXOM, in terms of robustness, compared to standard procedures for producing explanations, with perturbations of the given inputs of 1%, that is, with $\epsilon = 0.01$. An interesting extension of the analysis could be to evaluate the results of the same experiments by applying different values of the perturbation range (however within reasonable limits). It would then be possible to capture further insights into the behaviour of this property.

Procedure Generalization In this work, experiments were conducted considering Decision Tree and Random Forest as the reference models and SHAP as the explanation algorithm. Nevertheless, it is easy to realize that the existing work can be easily generalized to other types of model ensembles (e.g. bagging of models, gradient boosting etc.) as well as to additional model-agnostic XAI algorithms (e.g LIME). It would be useful to understand how the robustness of the procedure changes depending on the Machine Learning tools to which it is applied.

Performance Improvements. In ensembles of models composed of a large number of weak learners, calculating a distinct explanation value for each of the weak learners may be a computationally intensive task to perform, depending on the XAI algorithm used and the model being explained. It would be interesting to develop heuristics that would allow such computation to be performed in a more lightweight manner.

Application to Regression Problems. Given the impossibility of treating the output as categorical classes, an interesting idea to apply the procedure to regression problems is that of weighting the average of the explanations with weight values that decay as the output provided by the individual model diverges from that of the ensemble. Certainly one problem arising from the latter solution is that this would introduce the need to carefully tune the values relative to the intensity of the decay, as well as the magnitude of the function to be used. This could be solved by trying to evaluate parameter by parameter the mean robustness of the explanations of the samples present within a validation set.

Making Use of Another Combination Method. The type of aggregation of weak learners' explanations that is shown in this work is a discriminative average, which considers in the calculation only those values produced by weak learners who "agree" with the ensemble output. It is possible that the application of different tools for aggregating numerical values will lead to even better robustness results. Depending on the ensemble used, it might be useful to use prediction probabilities to weight the mean, as well as to discriminate estimators by criteria different from the one used in this work.

Bibliography

[1] S. Aeberhard. Wine. UCI Machine Learning Repository, 1991.

[2] D. Alvarez-Melis and T. S. Jaakkola. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018. URL http://arxiv.org/abs/1806.08049.

[3] K. J. Archer and R. V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52(4):2249–2260, 2008. ISSN 0167-9473. doi: https://doi.org/10.1016/j.csda.2007.08.015. URL https://www.sciencedirect.com/science/article/pii/S0167947307003076.

[4] L. Auret and C. Aldrich. Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, 35:27–42, 2012. ISSN 0892-6875. doi: https://doi.org/10.1016/j.mineng.2012.05.008. URL https://www.sciencedirect.com/science/article/pii/S0892687512001987.

[5] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2019.12.012. URL https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[6] V. Belle and I. Papantonis. Principles and practice of explainable machine learning. *Frontiers in Big Data*, 4, 2021. ISSN 2624-909X. doi: 10.3389/fdata.2021.688969. URL https://www.frontiersin.org/articles/10.3389/fdata.2021.688969.

[7] L. Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

[8] L. Breiman. Randomizing outputs to increase prediction accuracy. *Machine Learning*, 40:229–242, 2000.

[9] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL http://dx.doi.org/10.1023/A%3A1010933404324.

[10] L. Breiman, J. Friedman, C. J. Stone, and R. Olshen. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984. doi: https://doi.org/10.1201/9781315139470.

[11] N. Burkart and M. F. Huber. A survey on the explainability of supervised machine learning. *CoRR*, abs/2011.07876, 2020. URL `https://arxiv.org/abs/2011.07876`.

[12] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. J. Goodfellow, A. Madry, and A. Kurakin. On evaluating adversarial robustness. *CoRR*, abs/1902.06705, 2019. URL `http://arxiv.org/abs/1902.06705`.

[13] M. Charytanowicz, J. Niewczas, P. Kulczycki, P. Kowalski, and S. Lukasik. seeds. UCI Machine Learning Repository, 2012.

[14] H. Deng. Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics*, 2019. URL `https://doi.org/10.1007/s41060-018-0144-8`.

[15] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40:139–157, 2000. ISSN 1573-0565. doi: 10.1023/A:1007607513941. URL `https://doi.org/10.1023/A:1007607513941`.

[16] P. Domingos. Knowledge discovery via multiple models. *Intelligent Data Analysis*, 2(1):187–202, 1998. ISSN 1088-467X. doi: https://doi.org/10.1016/S1088-467X(98)00023-7. URL `https://www.sciencedirect.com/science/article/pii/S1088467X98000237`.

[17] R. R. Fernández, I. Martín de Diego, V. Aceña, A. Fernández-Isabel, and J. M. Moguerza. Random forest explainability using counterfactual sets. *Information Fusion*, 63:196–207, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.07.001. URL `https://www.sciencedirect.com/science/article/pii/S1566253520303134`.

[18] J. H. Friedman. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1):1 – 67, 1991. doi: 10.1214/aos/1176347963. URL `https://doi.org/10.1214/aos/1176347963`.

[19] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2012. doi: 10.1109/TSMCC.2011.2161285.

[20] B. German. Glass Identification. UCI Machine Learning Repository, 1987.

[21] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63:3–42, 2006.

[22] B. Goodman and S. Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 10 2017. doi: 10.1609/aimag.v38i3.2741. URL https://ojs.aaai.org/index.php/aimagazine/article/view/2741.

[23] S. Hara and K. Hayashi. Making tree ensembles interpretable: A bayesian model selection approach. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 77–85. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/hara18a.html.

[24] T. K. Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995. doi: 10.1109/ICDAR.1995.598994.

[25] G. V. Kass. An exploratory technique for investigating large quantities of categorical data. *Journal of The Royal Statistical Society Series C-applied Statistics*, 29:119–127, 1980.

[26] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un)reliability of saliency methods, 2017. URL https://arxiv.org/abs/1711.00867.

[27] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 2021. ISSN 1099-4300. doi: 10.3390/e23010018. URL https://www.mdpi.com/1099-4300/23/1/18.

[28] V. Lohweg. banknote authentication. UCI Machine Learning Repository, 2013.

[29] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. *CoRR*, abs/1705.07874, 2017. URL http://arxiv.org/abs/1705.07874.

[30] S. M. Lundberg, G. G. Erion, and S. Lee. Consistent individualized feature attribution for tree ensembles. *CoRR*, abs/1802.03888, 2018. URL http://arxiv.org/abs/1802.03888.

[31] S. Mishra, S. Dutta, J. Long, and D. Magazzeni. A survey on the robustness of feature importance and counterfactual explanations. *CoRR*, abs/2111.00358, 2021. URL https://arxiv.org/abs/2111.00358.

[32] C. Molnar. *Interpretable Machine Learning*. Bookdown, 2 edition, 2022. URL https://christophm.github.io/interpretable-ml-book.

[33] J. X. Morris, E. Lifland, J. Y. Yoo, and Y. Qi. Textattack: A framework for adversarial attacks in natural language processing. *CoRR*, abs/2005.05909, 2020. URL https://arxiv.org/abs/2005.05909.

[34] R. Polikar. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3):21–45, Third 2006. ISSN 1558-0830. doi: 10.1109/MCAS.2006.1688199.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL http://arxiv.org/abs/1602.04938.

[36] L. Rokach. Ensemble-based classifiers. *Artificial Intelligence Review*, 2010. ISSN 1573-7462. doi: 10.1007/s10462-009-9124-7. URL https://doi.org/10.1007/s10462-009-9124-7.

[37] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Comput. Surv.*, 54 (5), may 2021. ISSN 0360-0300. doi: 10.1145/3453158. URL https://doi.org/10.1145/3453158.

[38] A. Subbaswamy, R. Adams, and S. Saria. Evaluating model robustness and stability to dataset shift. In A. Banerjee and K. Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2611–2619. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/subbaswamy21a.html.

[39] E. Tjoa and C. Guan. A survey on explainable artificial intelligence (XAI): towards medical XAI. *CoRR*, abs/1907.07374, 2019. URL http://arxiv.org/abs/1907.07374.

A | SHAP - Equivalence between RF global explanation and average of weak explanations

The SHAP explanation of label k and feature i for a Random Forest model is:

$$\phi_{k,a}(x) = \sum_{\mathcal{S} \subseteq \mathcal{A} \setminus \{a\}} \frac{|\mathcal{S}|!(|\mathcal{A}| - |\mathcal{S}| - 1)!}{|\mathcal{A}|!} \left[f_{k,\mathcal{S} \cup \{a\}}(x_{\mathcal{S} \cup \{a\}}) - f_{k,\mathcal{S}}(x_{\mathcal{S}}) \right] \tag{A.1}$$

where the function $f_{k,\mathcal{S}}(x_{\mathcal{S}})$ is the prediction function constructed on the feature set $\mathcal{S}$ and returns 1 if x is classified with the label k and 0 otherwise. By calling $f_{w,k,\mathcal{S}}(x_{\mathcal{S}})$ the exact same prediction function, but related to the weak learner w of the ensemble, we can define the equation that links the predictions of the weak Decision Trees to that of the Random Forest model as:

$$f_{k,\mathcal{S}}(x_{\mathcal{S}}) = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} f_{w,k,\mathcal{S}}(x_S) \tag{A.2}$$

where $\mathcal{W}$ is the set of weak learners that compose the ensemble. By substituting this value in place of $f_{k,\mathcal{S}}(x_{\mathcal{S}})$ in the equation (A.1) we obtain:

$$\phi_{k,a}(x) = \sum_{\mathcal{S} \subseteq \mathcal{A} \setminus \{a\}} \frac{|\mathcal{S}|!(|\mathcal{A}| - |\mathcal{S}| - 1)!}{|\mathcal{A}|!} \cdot \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \left[f_{w,k,\mathcal{S} \cup \{a\}}(x_{\mathcal{S} \cup \{a\}}) - f_{w,k,\mathcal{S}}(x_{\mathcal{S}}) \right] \tag{A.3}$$

Given the linearity of the sum operator, we can take out the sum constructed on W, thus obtaining:

$$\phi_{k,a}(x) = \frac{1}{|\mathcal{W}|} \sum_{w \in \mathcal{W}} \sum_{\mathcal{S} \subseteq \mathcal{A} \setminus \{a\}} \frac{|\mathcal{S}|!(|\mathcal{A}| - |\mathcal{S}| - 1)!}{|\mathcal{A}|!} \left[f_{w,k,\mathcal{S} \cup \{a\}}(x_{\mathcal{S} \cup \{a\}}) - f_{w,k,\mathcal{S}}(x_{\mathcal{S}}) \right] \tag{A.4}$$

which corresponds to the formula for calculating the explanations of an ensemble by averaging the explanations of its weak learners.

List of Figures

3.1 Zones of explanation constancy 16
3.2 SHAP explanations difference and incremental ratio heatmaps comparison 18
3.3 From DT greymaps to RF greymaps 21
3.4 DT-RF comparison between heatmaps of SHAP explanations difference . . 22
3.5 AXOM functioning illustration 23

4.1 Robustness comparison through box plots 29
4.2 RF-AXOM Robustness comparison through incremental ratio heatmaps . . 30
4.3 DT-RF-AXOM full test set Robustness comparison 31
4.4 DT-AXOM explanations smoothness comparison in Wine dataset 33
4.5 Sample-based SHAP values comparison 35
4.6 Full test set SHAP values comparison 35

List of Tables

3.1 Datasets description 25

4.1 $\bar{L}$ robustness comparison 29
4.2 Two samples mean T-tests 30
4.3 Weak learner's mislabeling percentage 34

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Esteban García-Cuesta and Professor Daniele Loiacono, who believed in me and accompanied me through every step of the realisation of this thesis with willingness, patience and kindness, providing me with valuable insights and excellent advices throughout my research journey.

I would also like to express my deepest gratitude and sincere appreciation to Politecnico di Milano for giving me with the means and opportunities to successfully carry out and complete my studies. Your support and encouragement have been instrumental in enabling me to achieve my academic and personal goals. My time at Politecnico di Milano has been a transformative experience, and I am proud to be a student of such a prestigious institution. Thank you for the knowledge, skills, and experiences that will stay with me for a lifetime.