T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL
ARTIFICIAL INTELLIGENCE HEAD OF THE DEPARTMENT

QUANTITATIVE EVALUATION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES ACROSS VARIED DATASETS AND MACHINE LEARNING METHODS

MASTER'S THESIS

FEYZA AYDOĞAN

ISTANBUL 2024

T.C.
BAHCESEHIR UNIVERSITY
GRADUATE SCHOOL
ARTIFICIAL INTELLIGENCE HEAD OF THE DEPARTMENT

QUANTITATIVE EVALUATION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES ACROSS VARIED DATASETS AND MACHINE LEARNING METHODS

MASTER'S THESIS

THESIS ADVISOR

Assoc. Prof. TEVFİK AYTEKİN

ISTANBUL 2024

T.C.

BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL

MASTER THESIS APPROVAL FORM

Program Name:	Artificial Intelligence
Student's Name and Surname:	Feyza Aydoğan
Name Of The Thesis:	Quantitative Evaluation of Explainable Artificial Intelligence Techniques Across Varied Datasets and Machine Learning Methods
Thesis Defense Date:	04.06.2024

This thesis has been approved by the Graduate School which has fulfilled the necessary conditions as Master thesis.

Assoc. Prof. Yücel Batu Salman

Institute Director

This thesis was read by us, quality and content as a Master's thesis has been seen and accepted as sufficient.

	Title/Name	**Institution**	**Signature**
Thesis Advisor's	Assoc. Prof. Tevfik Aytekin	Bahçeşehir University	
Member's	Asst. Prof. Tarkan Aydın	Bahçeşehir University	
Member's	Prof. Dr. Nizamettin Aydın	İstanbul Technical University	

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Feyza Aydoğan

Signature:

ABSTRACT

QUANTITATIVE EVALUATION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE TECHNIQUES ACROSS VARIED DATASETS AND MACHINE LEARNING METHODS

Feyza Aydoğan
Master's Program in Artificial Intelligence
Supervisor: Assoc. Prof. Tevfik Aytekin

May 2024, 47 pages

The quantitative assessment of explainable artificial intelligence (XAI), a field of paramount importance in contemporary research, is critically significant. This study establishes a methodology to evaluate explainable artificial intelligence methods, engaging with 10 datasets from distinct domains for binary classification and regression tasks. Evaluation includes a comparison of machine learning models like XGBoost, LightGBM, and CatBoost, using four XAI techniques: TreeExplainer, LIME, SamplingExplainer and PermutationExplainer. These techniques are assessed against seven metrics: faithfulness, monotonicity, model parameter randomization, identity, separability, stability, and duration. The findings are indicative, showing that each XAI method displays unique strengths and weaknesses, emphasizing that there is no one-size-fits-all solution in XAI techniques. TreeExplainer, for instance, excels in certain metrics where LIME may not, highlighting the situational superiority of some methods over others. This nuanced performance further illustrates the essential need for standardized evaluation metrics within the field of XAI, drawing parallels to the established metrics in conventional machine learning to ensure a level of transparency, comprehensibility, and trust that is paramount for the wider acceptance of XAI applications. The contribution of this analysis to XAI research is twofold: it proposes metrics that could streamline the evaluation of XAI methods, and it offers insights into the applicability of XAI across various domains. By facilitating informed decisions about method selection, this study promotes the advancement and effective utilization of XAI.

Key Words: Explainable Artificial Intelligence, Quantitative Evaluation Metric, Machine Learning

ÖZ

AÇIKLANABİLİR YAPAY ZEKA TEKNİKLERİNİN ÇEŞİTLİ VERİ SETİ VE MAKİNE ÖĞRENMESİ YÖNTEMLERİ ÜZERİNDE NİCEL DEĞERLENDİRİLMESİ

Feyza, Aydoğan
Yapay Zeka Yüksek Lisans Programı
Tez Danışmanı: Doç. Dr. Tevfik Aytekin

Mayıs 2024, 47 sayfa

Günümüzde büyük bir öneme sahip olan açıklanabilir yapay zekanın nicel olarak değerlendirilmesi büyük bir öneme sahiptir. Bu çalışma, ikili sınıflandırma ve regresyon görevleri için çeşitli alanlardan gelen on veri seti ile açıklanabilir yapay zeka metodlarını değerlendirecek bir yöntem oluşturur. Değerlendirme, XGBoost, LightGBM ve CatBoost gibi makine öğrenimi modellerinin karşılaştırılmasını ve TreeExplainer, LIME, SamplingExplainer ve PermutationExplainer olmak üzere dört XAI tekniğinin kullanımını içerir. Bu teknikler sadakat, monotonluk, model parametre rastgeleştirme, özdeşlik, ayrılabilirlik, istikrar ve süre olmak üzere yedi ölçüt üzerinden değerlendirilir. Bulunan sonuçlar göstermektedir ki her XAI metodu eşsiz güçlü ve zayıf yönler sergiler ve XAI tekniklerinde her duruma uyan tek bir çözümün olmadığını vurgular. Örneğin, TreeExplainer belirli ölçütlerde, LIME'ın başarısız olduğu alanlarda başarılı olabilir, bu da bazı metodların diğerlerine göre duruma bağlı üstünlüğünü ortaya koyar. Bu ayrıntılı performans, XAI alanında standartlaştırılmış değerlendirme ölçütlerinin zorunluluğunu daha da belirtir ve geleneksel makine öğrenimindeki kurulmuş metriklerle paralellik çizer ki bu da XAI uygulamalarının daha geniş kabul görmesi için şeffaflık, anlaşılabilirlik ve güven seviyesinin en üst düzeyde olmasını sağlar. Bu analizin XAI araştırmalarına katkısı iki yönlüdür: XAI metodlarının değerlendirilmesini kolaylaştırabilecek ölçütler önerir ve XAI'nin çeşitli alanlarda uygulanabilirliği hakkında içgörüler sunar. Metod seçimi hakkında bilgilendirilmiş kararlar almayı kolaylaştırarak bu çalışma, XAI'nin ilerlemesini ve etkili kullanımını destekler.

Anahtar Kelimeler: Açıklanabilir Yapay Zeka, Nitel Değerlendirme Ölçütü, Makina Öğrenmesi

To my family

ACKNOWLEDGEMENTS

I am profoundly grateful to my supervisor, Assoc. Prof. Tevfik Aytekin, for his invaluable guidance, advice, and insights during my research journey. His mentorship was crucial in the successful completion of this thesis.

I must extend my heartfelt thanks to my family and friends. Their unwavering support and belief in me were my pillars of strength throughout this challenging and rewarding process.

This thesis would not have been achievable without the support, patience, understanding, and guidance of my cherished mentors, family, and friends. I am deeply thankful to each one of them for their indispensable contributions to both my personal and academic endeavors.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

XAI	Explainable Artificial Intelligence
LIME	Local Interpretable Model-Agnostic Explanations
SHAP	Shapley Additive Explanations
TCAV	Testing with Concept Activation Vectors
PDP	PArtial Dependence Plots
ICE	Individual Explanations
DT	Decision Tree
TDT	depth of the tree
ACD	Average Weighted Class Depth
MEMC	MEan Evaluation of Metrics Change
EBM	Explainable Boosting Machine
ANN	Artificial Neural Network
DNN	Deep Neural Network
ML	Machine Learning
VI	Variable Importance
PERM	Permutation-based Variable Importance
FIRM	Feature Importance Ranking Measure
SMOTE	Synthetic Minority Over-sampling Technique

Chapter 1

Introduction

Explainable Artificial Intelligence has become a hot topic in recent years, mostly because AI systems are becoming more and more complicated and commonplace across many industries. As outlined by Barredo Arrieta et al. (2020), there has been a growth in the sophistication of AI-powered systems, and as decisions made by these systems eventually impact human lives, there is a growing need to comprehend how AI approaches are used to make those judgments. However, using these decisions directly may cause certain risks. The risk, as Barredo Arrieta et al. (2020) highlight, is in making and implementing judgments that are illogical, illegitimate, or that merely make it impossible to provide in-depth explanations for their actions. While these obstacles are not insurmountable, they do emphasize how crucial it is to create AI systems that are not only intelligent but also transparent and ethical. The doubt of bias in AI further complicates this landscape. As noted by Guidotti et al., (2018), biases may be inherited by decision models trained on data, potentially resulting in unfair and incorrect conclusions. Addressing these biases necessitates a deeper understanding of the underlying mechanisms of AI decision-making. It becomes essential to identify the specific factors that necessitate the push for explainability within these complex systems.

Explainability may not be a complete requirement in all AI systems. The following factors determine the requirement for explainability: (a) the level of functional opacity brought on by the intricacy of AI algorithms and (b) how resilient the application domain is to mistakes (Adadi & Berrada, 2018). These elements highlight the complexity of AI systems and the vital requirement for transparency across a range of fields. This need is not only technical but also deeply rooted in ethical and legal considerations. Legal academics and social scientists seem to have developed new worries about "algorithms" that are centered around opacity (Burrell, 2016), further emphasizing the societal implications of opaque decision-making processes. Also, new AI approaches that can explain and interpret judgments are required due to demands from the social, ethical, and legal spheres (Adadi & Berrada, 2018). This demand signifies a fundamental shift towards the development of AI systems that are

not only efficient but also accountable and understandable. The implications of entrusting AI with critical decisions further highlight the stakes involved. Entrusting crucial judgments to a system that is incapable of self-explanation or human explanation adds an obvious layer of risk to the already problematic situation when it comes to high-stakes choices (Carvalho, Pereira, & Cardoso, 2019). The demand for AI systems that can explain their reasoning has increased in light of these dangers, particularly as AI continues to be used in high-stakes industries. As outlined by Arya et al. (2019), the need from society for artificial intelligence systems to explain their predictions has coincided with the growing use of AI systems in high-stakes fields. This social demand for explainability is not only about reducing risks, but also about improving the decision-making process itself, making it more transparent and ultimately more effective. When there is a discrepancy between what a model can explain and what a decision-maker needs to know, explainability becomes necessary (Burkart & Huber, 2021). This discrepancy highlights the fundamental issue with explainability, which is closing the knowledge gap between AI and human comprehension. One could interpret the renewed interest in XAI as a reaction to this difficulty. The area of explainable artificial intelligence has seen a resurgence due to the need for reliable, equitable, strong, and high-performing models for practical applications (Linardatos, Papastefanopoulos, & Kotsiantis, 2020) demonstrating the widespread agreement on the significance of creating AI systems that are not only strong but also just and explainable.

XAI has multiple purposes. These were also mentioned in certain articles. XAI suggests developing a set of machine learning methods that will create more interpretable models while preserving a high degree of learning performance and making it possible for people to comprehend, properly trust, and effectively manage the new breed of artificially intelligent companions (Barredo Arrieta et al., 2020). Additionally, Adadi & Berrada (2018) stated that enabling explainability aims to make algorithmic judgments, along with any data that informs them, understandable to non-technical stakeholders and end users (Adadi & Berrada, 2018).

The XAI domain can be used in many different domains. The need increases especially in domains where decision-making is critical and explainability is needed. According to Adadi & Berrada (2018), XAI is used in different domains such as transportation, healthcare, legal, finance, and military. This broad application, which

spans industries and affects many aspects of daily life, highlights the need for and appeal of explainable AI. Nevertheless, there are many obstacles in the way of obtaining broad interpretability. While it is true that interpretability in AI/ML is a difficult problem, this does not imply that all AI/ML approaches are opaque. It is true that certain algorithms are easier to understand than others, and that interpretability and accuracy sometimes have to be traded off (Adadi & Berrada, 2018).

In the field of explainable artificial intelligence, numerous methodologies have been developed to enhance the interpretability and transparency of AI models. These include Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive Explanations (SHAP), Integrated Gradients, Testing with Concept Activation Vectors (TCAV), and Feature Importance. Additionally, Partial Dependence Plots (PDP), Individual Conditional Expectation (ICE) plots, Global Surrogate Models, Counterfactual Explanations, and Anchors (Holzinger, Saranti, Molnar, Biecek, & Samek, 2022) are pivotal in providing insights into both local and global model behaviours, facilitating a deeper understanding of machine learning decisions.

The field of explainable AI represents a unique area of research. Unlike the evaluation of traditional machine learning models, such as classification or regression, which relies on specific metrics, there is no universally accepted metric for assessing XAI methods. This situation highlights the complexity of creating and evaluating XAI techniques, as they aim to enhance the interpretability of AI systems in various ways, making it challenging to establish a standard measure of their effectiveness. According to Barredo Arrieta et al. (2020), to assess the explainability of machine learning models and suggest future research avenues for improving their comprehensibility, there is a need for study on the concepts and metrics involved. This requirement highlights the significance of ongoing research and development in this field and represents a crucial step toward improving the instruments and processes for assessing AI explainability. More quantitative, universal XAI measures are required to complement the current measurement protocols and community-proposed tools (Barredo Arrieta et al., 2020). The demand for more standardized and universal metrics reflects the growing need for an integrity framework for assessing and comparing the explainability of AI systems. What is needed is a standardized process to evaluate, assess, and measure how explainable various boosting strategies are so that scientists can compare them (Burkart & Huber, 2021). Interpretable machine learning research remains a relatively

tiny part of machine learning research overall, as compared to the concentration on improving performance metrics, building machine learning algorithms and models themselves, and so on (Carvalho, Pereira, & Cardoso, 2019). This reveals the need to create metrics to compare XAI methods. The area of machine learning interpretability research urgently needs to shift its emphasis from developing new explanation techniques to comparing and evaluating the current approaches (Poursabzi-Sangdeh, Goldstein, Hofman, Wortman Vaughan, & Wallach, 2021). According to Nauta et al. (2023), appropriate XAI evaluation criteria become more and more necessary as the variety of XAI approaches increases. As the field of Explainable Artificial Intelligence (XAI) continues to expand, the necessity for the formulation and implementation of precise evaluation metrics becomes increasingly critical. This development is essential to ensure the effectiveness and applicability of XAI methods in a variety of domains, thereby enhancing their reliability and transparency.

In this research, four distinct Explainable Artificial Intelligence (XAI) methodologies, namely LIME (Local Interpretable Model-agnostic Explanations), TreeExplainer, SamplingExplainer in SHAP library, and PermutationExplainer in SHAP library, will be systematically analyzed and compared across a spectrum of evaluation metrics. The study delineates seven evaluation metrics that are rigorously defined for a comprehensive assessment: Faithfulness, Monotonicity, Model Parameter Randomization, Identity, Separation, Stability, and Duration. These metrics serve as a robust framework for evaluating the interpretability and reliability of the XAI methods under scrutiny.

Furthermore, the investigation will employ three advanced machine learning models—XGBoost, LightGBM, and CatBoost—across both classification and regression tasks to establish a broad testing ground. The application of these models will facilitate a multifaceted analysis of the XAI techniques, ensuring the findings are robust and applicable across various contexts.

The empirical evaluation will be conducted on ten diverse datasets, encompassing a wide range of domains to ensure the generalizability of the results. Through this rigorous analysis, the study aims to uncover nuanced insights into the comparative effectiveness and limitations of the XAI methods, guided by the defined evaluation metrics. Such a comprehensive approach will contribute significantly to the field of Explainable AI by providing a detailed understanding of how different XAI

methods perform across varied models and datasets, thereby guiding future research and applications in creating more transparent, interpretable, and trustworthy AI systems.

Chapter 2

Literature Review

The explainability of models stands out as a critical issue. The fact that the decisions made by AI systems can be explained provides a better understanding of the model results. It is of great importance that the models are explainable when any important decision needs to be made in environments where critical decisions are made, such as in the field of health and finance. Because when making a decision, knowing why that decision was made will make the result of the model more reliable. More than one explainable AI method has been developed such as LIME, Kernel Shap, Tree Shap, Anchors etc. These explainable AI methods are divided into model-specific techniques versus model-agnostic techniques. Tools for model-specific interpretation are made specifically to interpret models with attributes and functionalities. They are limited to use with a single class of algorithms. Model-agnostic interpretation techniques are applicable to any machine learning model (Dwivedi et al., 2023). On the other hand, explainability methods are divided into local and global interpretation. A model's general behaviour and overall analysis are examined in global interpretation approaches. Analysing each of the model's individual predictions and choices is known as "local interpretation" (Dwivedi et al., 2023).

As the field of XAI has become so widespread, the reliability and evaluation of the XAI methods developed have become an important research area. Various studies have been conducted in this field, major definitions have been put forward and evaluation metrics appropriate to these definitions have been developed. For the purpose of evaluating interpretability, three primary experimental levels are suggested. Application-grounded assessment necessitates carrying out end-user studies inside of a functioning application. Human-grounded assessment is the process of carrying out less complex human-subject tests while preserving the primary features of the intended application. Evaluation with a functional foundation doesn't need human experimentation. A formal definition of interpretability, such as the depth of a decision tree, is used as a stand-in for explanation quality evaluation in this kind of assessment (Carvalho, Pereira, & Cardoso, 2019).

Specific definitions have been derived for XAI evaluation metrics. These evaluation metrics have been defined by different studies. According to Carvalho, Pereira, & Cardoso (2019), the three quantitative interpretability indicators are represented as identity, separability, stability. Identity requires that the explanations for identical items be the same. The term separability refers to It is impossible for non-similar items to have identical explanations. Stability dictates that comparable items need comparable justifications. In addition to this, three quantitative interpretability indicators are explained as completeness, correctness, and compactness. For an explanation to be considered complete, the audience must confirm its validity. If an explanation is accurate, confidence should be built in it. Compactness dictates that the justification be brief.

According to Belle & Papantonis (2021), comprehensibility, fidelity, accuracy, scalability, and generality are listed as the evaluation criteria. The degree to which extracted representations are understandable to humans is known as comprehensibility, and it touches on the previously discussed characteristics of transparency. The degree to which extracted representations faithfully depict the opaque models from which they were taken is known as fidelity. The capacity of extracted representations to forecast unknown cases with accuracy is known as accuracy. The method's scalability refers to its capacity to handle opaque models with sizable input spaces and weighted connection counts. Generality is the degree to which specific training regimens or limitations on opaque models are necessary for the procedure.

As outlined by Rosenfeld (2021), in order to evaluate XAI more effectively, research proposes a paradigm change that centers on metrics that measure the explanation itself and its suitability in light of the XAI purpose. Four metrics are proposed as R, F, S and D. According to Singh (2021), many XAI properties are explained which are continuity, sensitivity, selectivity, soundness, faithfulness, fidelity, robustness, tractability, unambiguity, interpretability, consistency, stability and explicitness etc.

According to Coroamă & Groza (2022), two difficulties are highlighted in their paper's summary of the literature's current assessment techniques and metrics. A significant portion of the study stays inside theoretical concepts that haven't been used or embraced in practical applications. Nonetheless, certain criteria may have broad

applicability. Applicable XAI evaluation metrics are presented which are D, R, F, S, simplicity, sensitivity, completeness, soundness, stability, robustness, computational cost, monotonicity, perturbation-based metrics, diversity, correctness, confidence, fidelity, representativity, consistency, faithfulness etc. On the other hand, subjective XAI metrics are transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, persuasiveness, efficiency, satisfaction, comprehensibility and justifiability.

One of the study provides a thorough assessment of the evaluateability and interpretability approaches, which helps meet the growing need for XAI evaluation methods. Assessing with or without users is a topic that is connected to the evaluation of plausibility. XAI quality properties are stated as correctness, completeness, consistency, continuity, contrastively, covariate complexity, compactness, composition, confidence, context, coherence, contrability. It is stated how to evaluate these methods quantitatively. Correctness refers to how loyal and true the explanation is in relation to the predicted model. Correctness can be measured with applying model parameter randomization check, explanation randomization check, white box check, controlled synthetic data check, single deletion and incremental deletion. A measure of completeness is how well the explanation explains the prediction model f. It is calculated with preservation check, deletion check, fidelity, and predictive performance. Consistency ensures that similar inputs have similar justifications. This can be quantified with implementation invariance. Continuity takes into account the smoothness or continuity of the explanation function. It can be calculated with stability for slight variations, fidelity for slight variations and connectedness. Contrastivity seeks to make comparisons with different targets or occurrences easier by addressing how discriminative an explanation is and it is measured with target sensitivity, target disciriminativeness and data randomization check (Nauta et al., 2023).

Quantitative XAI evaluation methods have been described in more than one study. It is very important that these are applicable. In addition, in many studies, mathematical results have been derived using these methods or similar methods and comparisons have been made between XAI methods. According to Arya et al. (2019), AI Explainability 360 tool which is open-source software is introduced. It includes two evaluation metrics which are faithfulness and monotonicity. In another study, synthetic data was generated and then three different evaluation metrics which are faithfulness,

monotonicity and incompleteness were developed mathematically. Then, SHAP and LIME were compared (Oblizanov, Shevskaya, Kazak, Rudenko, & Dorofeeva, 2023). According to El Shawi, Sherif, Al-Mallah, & Sakr (2019), six distinct assessment measures which are identity, stability, separability, similarity, time and bias detection were created theoretically and used to the health sector. The LIME, SHAP, and Anchor techniques were compared. The findings demonstrate that no one interpretability method can offer the optimal outcomes across all measurements and data kinds.

One of the studies addresses some of SHAP's limitations and takes into consideration the interdependence of features, investigating LIME, the comprehensive approach, SHAP, and the coalition technique. Metrics have been developed: time per instance, the measurement of the average deviation in influence provided by a technique from the Complete method, and the distribution of feature significance assigned by each explanation (Doumard et al., 2022). Subsequent to this research, another study was conducted that expanded the metrics by including robustness, readability, and pairwise feature interaction. This additional study has crafted a comprehensive guide detailing the appropriate models to employ in various scenarios (Doumard et al., 2023).

In a comprehensive study referenced as (Duell, Fan, Burnett, Aarts, & Zhou, 2021), an in-depth metric analysis was conducted on top features using Explainable AI (XAI) methods such as SHAP, LIME, and Anchors. The research meticulously evaluated how these different XAI techniques influence the identification of significant features within a given dataset. The empirical results of the tests highlighted that the top features identified by each XAI method were markedly different.

In another study, the issue of determining a precise and unambiguous set of measures for the assessment of local linear explanations is one that we tackle in this study. Every measure has been included into the LEAF open-source Python framework. Metrics are conciseness, local fidelity, local concordance, reiteration similarity and prescriptivity. SHAP and LIME are used in this work and comparisons made between these two methods (Amparore, Perotti, & Bajardi, 2021).

One of the papers provides objective measures for evaluating explainable AI (XAI) tools, such as LIME and SHAP. By measuring the complexity of the XAI approach using decision tree (DT) depth, it eliminates subjective evaluation. The research presents a strategy wherein DT creation is informed by feature significance

ratings obtained from SHAP and LIME. The evaluation of DT complexity, and consequently XAI complexity, is done using two metrics: total depth of the tree (TDT) and average weighted class depth (ACD). The results demonstrate the greater complexity and scalability of SHAP, providing guidance on applicability for different document sizes and pointing out aspects to enhance black-box models such as feedforward neural networks (AHMED & ALPKOÇAK, 2022).

According to M El-gezawy, Abdel-Kader, & Ali, (2023), a new metric called Mean Evaluation of Metrics Change (MEMC) was developed to assess Explainable AI approaches' performance globally. MEMC is used to evaluate XAI approaches, including intrinsic (LIME, SHAP, and ANCHORS) and post-hoc (Random Forest, XGBoost, Logistic Regression, Decision tree, EBM, ANN, and DNN) techniques for classification in order to determine which is the best methodology.

Another paper includes a comparative assessment of three distinct model-agnostic XAI techniques: AraucanaXAI, LIME, and SHAP. Four quantitative assessment metrics which are identity, fidelity, separability, and time to calculate an explanation are used to compare the various XAI techniques (Buonocore, Nicora, Dagliati, & Parimbelli, 2022).

According to Yalcin, Fan, & Liu (2021), a binary classification-focused approach for quantitatively evaluating the accuracy of XAI algorithms is presented, utilizing datasets with a known explanatory ground truth. Formal language grammars are used to construct datasets, and certain attributes are used to classify strings as positive. In a string, symbols denote explanatory ground truth only since they augment the attribute. Three conclusions are drawn from the study, which uses SHAP and LIME for feature attribution: explanation accuracy declines with dataset complexity, explanation accuracy is more closely correlated with higher classification accuracy, and SHAP performs better in explanation accuracy than LIME.

In a study, various explainable AI techniques including Variable Importance (VI), Permutation-based Variable Importance (perm), SHAP, and Feature Importance Ranking Measure (FIRM) have been employed. The machine learning models utilized encompass linear models, decision trees, neural networks, and ensemble methods. For evaluation metrics, the study considers feature importance, consistency, stability, robustness, computation time, fairness, and regulatory compliance (Lozano-Murcia, Romero, Serrano-Guerrero, & Olivas, 2023).

A comprehensive literature review has been conducted on evaluation methods for Explainable Artificial Intelligence. It has been observed that there are multiple evaluation metrics available, emphasizing the critical role of quantitative methods in this context. Through the development of specific mathematical methods, these metrics have been made applicable, allowing for the rigorous quantitative assessment of XAI approaches. Different XAI methods have been compared across these metrics in various studies, showcasing the importance of quantitative methods in providing objective, measurable insights into the effectiveness of explainability approaches. This review not only highlights the diversity and applicability of evaluation metrics in the field of XAI but also underscores the essential contribution of quantitative methods in advancing our understanding and assessment of XAI techniques.

Chapter 3

Theoretical Background

This section delves into the theoretical background of Explainable Artificial Intelligence methods, aiming to clarify their operational principles, significance, and impact on AI interpretability. By examining how XAI facilitates a better understanding between complex AI decisions and human comprehension, this introduction sets the stage for further analysis. It provides a concise overview intended to enhance the reader's grasp of XAI's role in advancing artificial intelligence research and application.

3.1 LIME (Local Interpretable Model-Agnostic Explanations)

LIME is an interpretable model-based local approximation approach that can faithfully explain the predictions of any classifier or regressor (Ribeiro, Singh, & Guestrin, 2016).

$$explanation(x) = argmin_{g \in G} L(f, g, \pi_x) + \Omega(g) \quad (1)$$

The original input point is represented by x. In G, which represents the family of interpretable models, f is a complicated model while g is a basic model. A proximity metric called π_x indicates how close the area is to x. In the vicinity of x, additional data points are generated, and the data points are weighted based on their distance from x, denoted by π_x. New data points are predicted with complicated model f. The goal of the loss function, denoted by L, is to reduce the total square difference between the simple model g's prediction and the label derived from the complex model f. $\Omega(g)$ guarantees that a straightforward explanation with few variables is obtained by regularizing the complexity of g (Molnar, 2023).

3.2 Shapley Values

A coalitional game theory technique called Shapley values indicates how to divide the "payout" among the characteristics in an equitable manner. A value function val of the players in S is used to define the Shapley value. A feature value's contribution to the payment, weighted and totaled over all potential feature value combinations, is the Shapley value (Molnar, 2023).

$$\phi_j = \sum_{S \subseteq 1,\dots M \setminus \{j\}} \frac{|S|!(M - |S| - 1)!)}{M!} [val(S \cup \{j\}) - val(S)] \tag{2}$$

where val is the prediction model, S is the subset of the features, M is the number of features, val(S) gives the prediction result for the subset.

3.3 SHAP (Shapley Additive Explanations)

Calculating the Shapley value takes a long time. Because there are two thousand different coalitions, it is computationally costly to compute the Shapley value exactly. People want to have partial explanations (Molnar, 2023). One approach is SHAP, which was presented by Lundberg & Lee (2017). By calculating the contribution of each characteristic to the prediction, SHAP seeks to explain the prediction of an instance x. One new approach that SHAP offers is the representation of the Shapley value explanation as a linear model and additive feature attribution technique. Shapley values and LIME are connected in that perspective (Molnar, 2023).

$$g(z^{'}) = \phi_0 + \sum_{j=1}^{M} \phi_j z_j^{'} \tag{3}$$

where g is the explanation model, $z' \in \{0,1\}^M$ is the coalition vector that shows whether feature is present or absent, M is number of features, Φ_0 is average output of the model, Φ_j is Sharply value for j feature.

All four qualities (Efficiency, Symmetry, Dummy, and Additivity) can only be satisfied by Shapley values. Considering that SHAP computes Shapley values, it also meets these. Local accuracy mandates that the sum of the feature attributions plus the base value (which represents the model's output in the absence of any features) should equal the actual output of the model for the given input. Another characteristic is missingness. If the reduced inputs indicate the presence of features, then missingness necessitates that features absent from the initial input have no effect. Consistency is the third attribute. According to consistency, an input's attribution shouldn't go down if a model is altered so that a simplified input's contribution rises or remains constant independent of the other inputs (Lundberg & Lee, 2017). The alternative kernel-based estimate method for Shapley values, KernelSHAP, was first put out by the SHAP developers and was motivated by local surrogate models. Furthermore, an effective estimate method for tree-based models called TreeSHAP was suggested by them (Molnar, 2023).

3.3.1 KernelExplainer. When computing the kernel SHAP, a random coalition is formed, signifying that certain features are designated as 1 (present) or 0 (absent). Subsequently, absent features are replaced with values extracted from the training dataset. Then each coalition, weights are calculated. The biggest weights are assigned to small coalitions (i.e., few 1s) and large coalitions (i.e., many 1s). It can discovered the isolated major influence of a single characteristic on the prediction if the coalition is made up of only that one feature. It can discovered the overall effect of a feature (primary effect plus feature interactions) if a coalition includes all but one of the features. Since several coalitions may be formed with just half of the features, it cannot infered anything about the contribution of any one feature if the coalition only has half of the traits (Molnar, 2023).

$$\pi_x(z') = \frac{M-1}{\binom{M}{|z'|}|z'|(M-|z'|)} \tag{4}$$

where M is the maximum coalition size and |z'| is number of present features.

Then weighted regression model is created. By optimizing the loss function L, linear model g is trained. The estimated coefficients of the model, represented as ϕ_j, are the Shapley values, which are obtained by optimizing the familiar sum of squared errors typically used in linear models.

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j \tag{5}$$

where g is the sum of shapley values for present features.

$$L(\hat{f}, g, \pi_x) = \sum_{z' \in Z} [\hat{f}(h_x(z')) - g(z')]^2 \pi_x(z') \tag{6}$$

where f is a complicated model while g is a basic model.

3.3.2 TreeExplainer. Despite SHAP values' theoretical benefits, their practical application is limited by difficulties in efficiently estimating conditional expectations and the exponential complexity of their calculation, prompting the development of faster, tree-specific SHAP value estimation methods. Complexity of the SHAP values directly using trees is $O(TL2^M)$. An innovative algorithm is introduced that computes the aforementioned values in polynomial rather than exponential time. Specifically, this algorithm operates in $O(TLD^2)$ time and requires $O(D^2 + M)$ memory, with D equating to logL for balanced trees. Here, T represents the count of trees, L denotes the maximum leaf count in any tree, and M stands for the feature count.

Recursively monitoring the percentage of all feasible subsets that flow into each of the tree's leaves is the intuition behind the polynomial time approach. In Equation 2, it is applied to all 2^M subsets S. Merely monitoring the number of subgroups that flow through every branch of the tree would seem like an acceptable approach.

Every possible subset size is monitored in the improved version throughout the recursion procedure. The UNWIND approach reverses these expansions and is made to function in perfect harmony with the EXTEND technique, which extends subsets according to a predetermined ratio of ones to zeros. When moving down the tree, EXTEND is used, and UNWIND recalculates the extensions to precisely compute

weights for features along a leaf's path and corrects for multiple splits on the same feature. This optimizes memory usage.

$$\Phi_{i,j} = \sum_{S \subseteq N \setminus \{i,j\}} \frac{|S|!(M - |S| - 2)!}{2(M-1)!} \nabla_{ij}(S)$$

$$\nabla_{ij}(S) = f_x(S \cup \{i,j\}) - f_x(S \cup \{i\}) - f_x(S \cup \{j\}) + f_x(S)$$

$$= f_x(S \cup \{i,j\}) - f_x(S \cup \{j\}) - [f_x(S \cup \{i\}) - f_x(S)].$$

$$\phi_{i,i} = \phi_i - \sum_{j \neq i} \phi_{i,j} \tag{7}$$

SHAP interaction values can be understood as the disparity between the SHAP values of feature i when feature j is included and when it's excluded (Lundberg, Erion, & Lee, 2019). Python library is created using this methodology (Lrjball, n.d.).

3.3.3 SamplingExplainer. The sampling explainer is a model-agnostic approach used to compute Shap values. This method does not directly calculate Shap values using all features. Instead, it selects a specific number of samples, which is divided into a two-stage process. In the first stage, samples are evenly distributed among the features, and Shap values are calculated for each feature. Additionally, the variance of values within each sample is recorded. In the second stage, the sample numbers are adjusted for each feature based on their variances. For all features involved in the model's prediction, combinations are determined for the distributed sample numbers. These combinations determine which features are included. Missing features are filled with values obtained from the background dataset. For each feature combination, the effect of adding or removing the features of interest on the model is calculated. This process is used to understand the contribution of the desired features to the output. Python library is created using this methodology (Lrjball, n.d.).

3.3.4 PermutationExplainer. The permutation explainer can compute Shapley values for any model; it is not dependent on any particular model. Iterating over all possible combinations of the characteristics, both forwards and backwards, is how it operates. This technique reduces the number of model assessments required by allowing changes to be made to one feature at a time. As such, it guarantees efficiency

irrespective of the quantity of original model executions required to approximate the feature attribution values. The difference between the base value and the output of the model for each explained case is therefore properly represented by the SHAP values that were produced, notwithstanding their approximation. Python library is created using this methodology (Lrjball, n.d.).

Chapter 4

Experimental Setup

In this section, an in-depth exploration of seven distinct quantitative evaluation metrics specifically designed to assess the efficacy of explainable artificial intelligence methods is presented. These metrics are crucial for providing a rigorous, objective framework to measure how effectively these XAI methods can be understood, interpreted, and trusted by users. Evaluating explainability in a quantitative manner aims to establish a standard benchmark that facilitates the comparison of different XAI approaches, highlighting their strengths and identifying potential areas for improvement. On the other hand, the dataset and machine learning techniques used will be defined in this section. Lastly, the stages of the experiment will be explained in detail.

4.1 Evaluation Metrics

4.1.1 Faithfulness. Faithfulness is a metric used to evaluate correctness. This metric involves removing, obscuring, or modifying a single characteristic in the input and then assessing how the predictive model's output has changed. Calculate the correlation using the significance score of the explanation (Nauta et al., 2023). The faithfulness metric is based on the Pearson sample correlation coefficient between the feature weights and their approximate contributions to the change in the model prediction when they are fixed or removed. It assesses the degree of correspondence between explaining the significance of each feature. Values around 1 show that the weight distribution is right, whereas values near 0 show that the characteristics' effect on the prediction does not match the weight set (Oblizanov, Shevskaya, Kazak, Rudenko, & Dorofeeva, 2023). The mathematical formula explaining faithfulness is shown in Formula 5.

$$faithfullness = \frac{1}{n}\sum_{i=1}^{n} correlation(|feature\_weights[i]|, |y_{predict[i]} - y_{[feature\_removed\_predictions[i][j]}|) \quad (8)$$

where n is the number of elements in the dataset, $y_{predict[i]}$ is the model prediction for the i-th element in the dataset, $y_{feature\_removed\_predictions[i][j]}$ is the model prediction for the i-th element when the j-th feature is disabled, and feature_weights[i] is the absolute value of the feature weights for the i-th element.

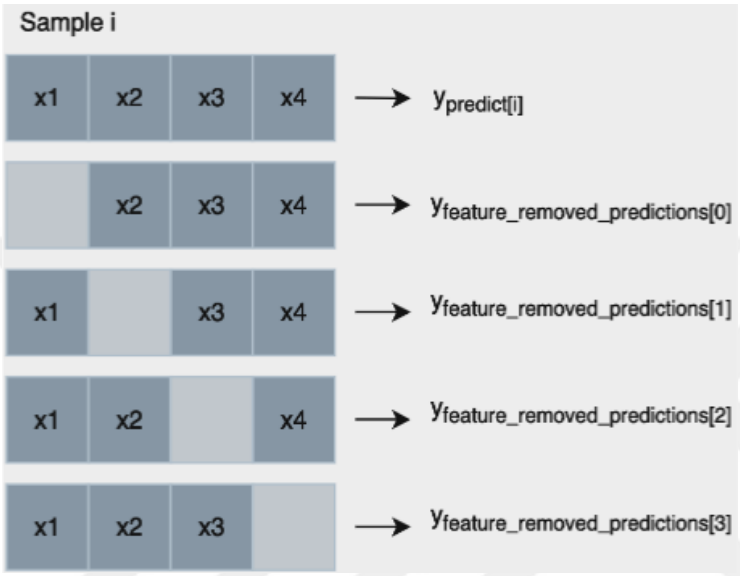

Figure 1. Faithfulness

The calculation of faithfulness on an individual sample is depicted in Figure 1. Firstly, the model generates a prediction for the given sample. Subsequently, model predictions are computed by iteratively extracting each feature. The removal of a feature does not imply leaving it unattended; various methods can be employed for this purpose. Specifically, the extracted feature may be replaced with 0, or alternatively, the average of that feature in the training data can substitute the extracted feature. Notably, a distinct methodology was adopted in the current study. In the selected method, all instance values in train dataset are substituted for the extracted features and prediction is made. This is then averaged. In this way, $y_{feature\_removed\_predictions[i][j]}$ is reached. Subsequently, as delineated in Formula 8, a list is generated by computing the difference between results obtained when no features are extracted, and results obtained when features are extracted one by one. The correlation

between the feature weights and the created list is computed. This process is iteratively applied to all data points, and the average is taken to calculate overall faithfulness.

4.1.2 Monotonicity. Monotonicity is another metric used to evaluate correctness. It is added features to the input one at a time under the sequence of explanation and see how the predictive model's output changes with each additional input (Nauta et al., 2023). Based on this idea, the monotonicity measure assesses how accurate a XAI-obtained series of characteristics is by ranking them according to increasing weight. The XAI approach permitted a distortion of feature priority if monotonicity was not observed: a feature with less effect was given more weight, or vice versa (Oblizanov, Shevskaya, Kazak, Rudenko, & Dorofeeva, 2023).

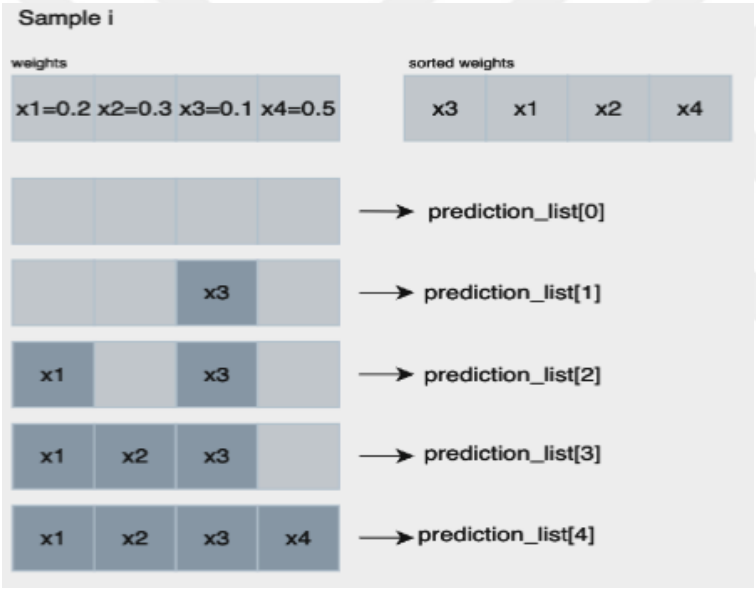

Figure 2. Monotonicity

First, feature weights are sorted in increasing order. Afterwards, first a prediction is made without adding any features. Then, the features are added one by one, and the prediction is made. Idle features are made using samples in the train data, as in faithfulness. Expectation is that as each feature is added, the difference with the prediction of the previous feature will gradually increase.

$$monotonicity = \frac{1}{n}\sum_{i=1}^{n} is\_monotonic(diff(abs(diff(y\_prediction\_list)) \tag{9}$$

where n is the number of elements in the dataset, y_prediction_list is the prediction results of the features added one by one.

4.1.3 Model parameter randomization. Model parameter randomization is another metric to evaluate correctness. This method checks to see whether the explanation changes when you arbitrarily alter the prediction model's parameters (Nauta et al., 2023). In the proposed methodology, the initial step involves the computation of the weight for the foundational model through the application of Explainable Artificial Intelligence (XAI) methods. This resultant weight is considered as the baseline model weight. Subsequently, an iterative process is employed wherein the internal parameters of the model undergo adjustments at a specified rate, such as 1 percent, 2 percent, and so forth. Following each adjustment, the weights are recalculated to reflect the modifications. To facilitate a meaningful analysis, the Euclidean distance between the base model weight and the weights corresponding to the altered parameters is computed.

$$D = \sqrt{\sum_{i=1}^{n}(W_{base,i} - W_{changed,i})^2} \tag{10}$$

where n is the number of features in the dataset, $W_{base,i}$ is the weight without changing parameters, $W_{changed,i}$ is the weight with changing parameters.

Upon completion of the aforementioned steps, an interpretative insight is sought through the creation of a graphical representation. This graphical depiction is based on the Euclidean distance and the rate of parameter change. The resulting graph serves as a visual aid to elucidate the relationship between alterations in model parameters and the associated impact on model weights, thereby contributing to a comprehensive understanding of the model's behavior under varied conditions. In order to understand the relationship, a correlation is taken between rate and Euclidean distance.

The most important point is that to make a comment, the performances of the models should not be too far apart when their parameters are changed. For this purpose,

the deviation rates in the base model are found by calculating the mean absolute percentage by using F1 score for the classification models and root mean square error for regression models. If the ratio is not high, it can be stated that the percentage change and Euclidian distance relationship will be interpreted.

$$Dev_{base} = \frac{1}{n}\sum_{i=1}^{n}\frac{|Metric_{base,i} - Metric_{changed,i}|}{Metric_{base,i}} \tag{11}$$

where n is the number of parameter changes, $Metric_{base,i}$ is the model performance (F1 score or root mean square error) without changing parameters, $Metric_{changed,i}$ is the model performance with changing parameters.

4.1.4 Identity. According to this measure, two similar cases have to have the same explanations if they exist (Carvalho, Pereira, \& Cardoso, 2019). In the proposed methodology, the XAI model undergoes a dual execution, wherein weights are determined independently for each instantiation. Subsequently, the Euclidean distance is computed between the weight vectors corresponding to each instance. If the resulting distances fall below a predetermined threshold, it is deemed that the weights are identical for the given instance. The threshold is determined as 1e-8. This process is iteratively conducted for all instances within the dataset, facilitating the derivation of a comprehensive identity score. The identity score serves as a global metric reflecting the degree of similarity between the weight vectors across all instances, thereby providing a quantitative assessment of the consistency or uniformity in the model's behavior. Score range is from 0 to 100.

$$true = \sum_{i=1}^{n}\begin{cases}1 & \text{if } |\text{dis}_i| < threshold \\ 0 & \text{otherwise}\end{cases}$$

$$\text{score} = \frac{\text{true}}{\text{n}} \tag{12}$$

where n is number of instance, dis_i is the euclidian distance for instance i, threshold is identity limit.

4.1.5 Separability. According to this measure, two distinct cases must have distinct explanations if they exist (Carvalho, Pereira, & Cardoso, 2019). If two instances are not equal to each other, the weights obtained from the XAI models are expected to be different from each other. According to El Shawi, Sherif, Al-Mallah, & Sakr (2019), a score is obtained by checking this between instances that are different from each other.

$$\text{wrong} = \sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} 1(\exp[i] = \exp[j])$$

$$\text{score} = 1 - \frac{\text{wrong}}{\text{n}} \tag{13}$$

where n is number of instance and exp is explanations.

Another approach has been created to measure this metric. In the presented methodology, the initial step involves the computation of Euclidean distances among the instance data points. Subsequently, a further computation is undertaken to determine the Euclidean distance between the respective weights associated with these instances. Following this, a correlation analysis is performed on these two distances. The rationale underlying this approach lies in the presumption that instances positioned at greater Euclidean distances from one another are anticipated to exhibit greater disparities in their associated weights. This correlation analysis seeks to quantify the degree of association between the spatial separation of instances and the dissimilarity in their weights, thereby elucidating potential patterns or relationships within the dataset.

$$score = corr(\|\mathbf{X}_1 - \mathbf{X}_2\|, \|\mathbf{W}_1 - \mathbf{W}_2\|) \tag{14}$$

where $\|X_1 – X2\|$ is Euclidean distance between instances, $\|W_1 – W2\|$ is Euclidean distance between weights.

4.1.6 Stability. According to this measure, instances that are part of the same class have to have similar explanations (Carvalho, Pereira, & Cardoso, 2019). According to El Shawi, Sherif, Al-Mallah, & Sakr (2019), after determining the weights for each instance, a subsequent step involves the application of k-means

clustering, where the number of clusters corresponds to the predetermined number of labels. Following this clustering process, an examination is conducted to assess the congruence between labels and clusters. The anticipated outcome is the alignment of instances sharing the same label within the confines of a common cluster. This method is used for classification tasks.

$$\text{error} = \sum_{i=1}^{n} ||\text{labels}_i - \text{cluster\_labels}_i|| \qquad \text{score} = 1 - \frac{\text{error}}{\text{n}} \tag{15}$$

where n is number of instance, $labels_i$ is label of the instance and $cluster\_labels_i$ is the clustering labels of the weights.

4.1.7 Duration. Duration is a metric that indicates how long each XAI model takes for each instance.

4.2 Dataset Description

Explainable AI (XAI) finds applications across diverse domains, with pivotal implications in critical sectors like finance and health. This study focuses on datasets within analogous domains to explore ten distinct datasets, encompassing five for binary classification and five for regression analysis. The inclusion of both classification and regression datasets aims to evaluate XAI metrics across varied data types, facilitating comprehensive insights. The datasets span domains such as finance, health, and education, each characterized by varying feature dimensions. By incorporating datasets with differing feature counts, this study aims to discern the impact of feature variability on XAI evaluation metrics. The selection of open-source datasets ensures transparency and accessibility in the research process. Through this endeavor, we seek to enrich understanding regarding the efficacy of XAI techniques in different contexts, contributing to the development of robust evaluation methodologies for interpretable AI systems.

Table 1
Datasets

Dataset Name	Feature Count	Type
Telco Customer Churn Dataset (BlastChar., 2018)	21	Binary classification
Employee Attrition Dataset (Patel, 2018)	35	Binary classification
Diabetes Dataset (Akturk, 2020)	9	Binary classification
Heart Disease Dataset (Lapp, 2019)	14	Binary classification
Stroke Prediction Dataset (Fedesoriano, 2021)	12	Binary classification
House Price Prediction Dataset (Kaggle)	81	Regression
Student Performance Dataset (Ansodariya, 2022)	34	Regression
US Cars Dataset (Alsenani, 2020)	13	Regression
Medical Cost Personal Datasets (Choi, 2018)	7	Regression
Crab Age Prediction (Sidhu, 2021)	9	Regression

4.3 Machine Learning Models

This study employs tree-based classification and regression models, including XGBoost, LightGBM, and CatBoost, for both classification and regression tasks. These machine learning algorithms are chosen for their effectiveness in handling complex datasets and their widespread adoption in various domains. By utilizing these models across diverse tasks, the research aims to investigate their performance and interpretability, thus contributing to advancing the understanding and applicability of tree-based methods in explainable AI research.

4.4 Experimental Design

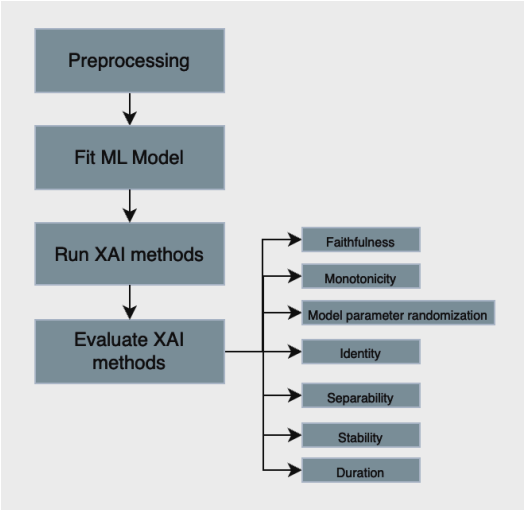

Figure 3. Experimental Design

The experimental design comprises four distinct stages: preprocessing, model fitting, execution of XAI methods, and evaluation metric analysis. These stages are meticulously structured to ensure methodological rigor and comprehensiveness in investigating the effectiveness and interpretability of the employed techniques.

4.4.1 Pre-processing. The pre-processing step involves preparing the data for training. In this stage, the dataset is loaded, and object-type data is transformed into numerical representations. The data is then split into training and testing sets. Subsequently, missing values in the dataset are filled. For numerical columns, the mean value is often applied, while for categorical columns, the most frequently occurring value is used. Standard scaling is applied to numerical columns to normalize their values, while one-hot encoding is applied to categorical columns to convert them into a numerical format suitable for machine learning algorithms. In classification tasks, datasets may often exhibit class imbalance, where some classes have significantly fewer samples than others. To address this issue, Synthetic Minority Over-sampling Technique (SMOTE) can be applied to balance the class distribution by generating synthetic samples for the minority class.

By following these steps, the data is prepared into a format ready for training, ensuring that it is appropriately transformed, balanced, and suitable for use with machine learning algorithms.

4.4.2 Model fitting. Model parameter optimization is conducted using Optuna to select the optimal model parameters. Following parameter selection, the model is trained, and predictions are generated. Subsequently, model performance is evaluated for both classification and regression tasks. For classification, metrics such as accuracy, precision, recall, F1-score, and ROC AUC score are assessed, while for regression, mean absolute error, mean squared error, root mean square error, and R-squared score are examined. This rigorous evaluation framework ensures a comprehensive understanding of the model's efficacy across various performance indicators, facilitating informed decision-making in practical applications.

4.4.3 XAI methods. In this study, feature weights for each instance are computed using the implementations of TreeExplainer, LIME, SamplingExplianer, and PermutationExplainer techniques. Ready-made Python libraries are utilized for this purpose. Through the implementation of these techniques using Python libraries, feature weights are computed for each instance, providing insights into the relative importance of features in the model's predictions.

4.4.4 Evaluation metric. In this study, evaluations are conducted on models using weights generated by XAI techniques, based on several metrics. These metrics include faithfulness, monotonicity, model parameter randomization, identity, separability, stability, and duration. Scores are derived according to these metrics for each machine learning model, dataset, and XAI method under consideration.

Based on the computed scores for each XAI method across these metrics, comparisons will be made to determine which XAI method performs better according to specific evaluation criteria. This analysis will provide insights into the strengths and weaknesses of different XAI techniques, guiding their selection and usage in practical applications.

Chapter 5

Result and Discussions

Evaluation metrics were tested on three different machine learning techniques (XGBoost, LightGBM, and CatBoost). Results were obtained using 4 different explainable AI methods (TreeExplainer, Lime, SamplingExplainer, PermutationExplainer) on 10 different datasets.

5.1 Faithfulness

In the investigation, the faithfulness score was computed across the XGBoost, LightGBM, and CatBoost machine learning methodologies, each analyzed with four distinct explainable AI techniques. The results revealed that the SamplingExplainer method consistently yielded the highest faithfulness score across all machine learning models.

Specifically, when employing the XGBoost machine learning approach, the SamplingExplainer method demonstrated the highest faithfulness score among the four explainable AI techniques.

Table 2
Faithfulness result for XGBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.92	0.71	0.97	0.96
Employee Attrition Dataset	0.91	0.65	0.90	0.89
Diabetes Dataset	0.92	0.78	0.96	0.93
Heart Disease Dataset	0.77	0.05	0.82	0.80
Stroke Prediction Dataset	0.88	0.66	0.96	0.96
House Price Prediction Dataset	0.99	0.49	0.99	0.98
Student Performance Dataset	0.94	0.53	0.95	0.95
US Cars Dataset	0.94	0.16	0.95	0.94
Medical Cost Personal Datasets	0.97	0.86	0.97	0.96
Crab Age Prediction	0.69	0.70	0.75	0.71
Average	**0.89**	**0.56**	**0.92**	**0.91**

Similarly, when employing LightGBM and CatBoost, the SamplingExplainer method also emerged as the optimal choice in terms of faithfulness score.

Table 3
Faithfulness result for LigthtGBM

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.92	0.71	0.96	0.95
Employee Attrition Dataset	0.90	0.66	0.91	0.91
Diabetes Dataset	0.93	0.75	0.97	0.94
Heart Disease Dataset	0.68	-0.01	0.80	0.78
Stroke Prediction Dataset	0.93	0.87	0.99	0.99
House Price Prediction Dataset	0.95	0.47	0.96	0.95
Student Performance Dataset	0.95	0.57	0.96	0.97
US Cars Dataset	0.92	0.26	0.94	0.93
Medical Cost Personal Datasets	0.97	0.85	0.97	0.96
Crab Age Prediction	0.80	0.78	0.83	0.80
Average	**0.90**	**0.59**	**0.93**	**0.92**

Table 4
Faithfulness result for CatBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.74	0.59	0.84	0.80
Employee Attrition Dataset	0.76	0.44	0.78	0.76
Diabetes Dataset	0.84	0.69	0.91	0.88
Heart Disease Dataset	0.69	0.13	0.82	0.79
Stroke Prediction Dataset	0.92	0.79	0.97	0.97
House Price Prediction Dataset	0.95	0.52	0.97	0.96
Student Performance Dataset	0.95	0.59	0.96	0.96
US Cars Dataset	0.93	0.15	0.94	0.94
Medical Cost Personal Datasets	0.97	0.86	0.97	0.97
Crab Age Prediction	0.73	0.59	0.77	0.74
Average	**0.85**	**0.54**	**0.89**	**0.88**

Table 5 shows the average faithfulness score of three machine learning methods in 4 different explainable ai methods. As shown in Table 5, an overarching analysis of the aggregated results underscored the SamplingExplainer method's supremacy in terms of faithfulness across all machine learning models and datasets.

Table 5
Faithfulness result

TreeExp	Lime	SamplingExp	PermutationExp
0.88	0.56	0.91	0.90

5.2 Monotonicity

The investigation focused on the evaluation of monotonicity scores, achieved through the application of the XGBoost, LightGBM, and CatBoost machine learning algorithms, supplemented by four distinct explainable artificial intelligence methodologies. The analysis revealed that the TreeExplainer method outperforms other explainable AI techniques in generating the highest monotonicity score when used with XGBoost and CatBoost algorithms. Conversely, within the LightGBM framework, the PermutationExplainer method emerged as the most effective for achieving superior monotonicity scores.

Table 6
Monotonicity result for XGBoost

Dataset Name	Feature Count	Tree Exp	Lime	Sampling Exp	Permutation Exp
Telco Customer Churn Dataset	21	0	0	0	0
Employee Attrition Dataset	35	0	0	0	0
Diabetes Dataset	9	0.026	0	0.047	0.021
Heart Disease Dataset	14	0	0	0	0
Stroke Prediction Dataset	12	0	0	0	0
House Price Prediction Dataset	81	0	0	0	0
Student Performance Dataset	34	0	0	0	0
US Cars Dataset	13	0	0	0	0
Medical Cost Personal Datasets	7	0.290	0.036	0.143	0.188
Crab Age Prediction	9	0.001	0	0	0
Average		**0.032**	**0.004**	**0.019**	**0.021**

Table 7
Monotonicity result for LightGBM

Dataset Name	Feature Count	Tree Exp	Lime	Sampling Exp	Permutation Exp
Telco Customer Churn Dataset	21	0	0	0	0
Employee Attrition Dataset	35	0	0	0	0
Diabetes Dataset	9	0.047	0.005	0.057	0.026
Heart Disease Dataset	14	0	0	0	0
Stroke Prediction Dataset	12	0	0	0.002	0
House Price Prediction Dataset	81	0	0	0	0
Student Performance Dataset	34	0	0	0	0
US Cars Dataset	13	0	0	0	0
Medical Cost Personal Datasets	7	0.278	0.015	0.209	0.310
Crab Age Prediction	9	0.013	0.003	0.034	0.037
Average		**0.034**	**0.002**	**0.030**	**0.037**

Table 8
Monotonicity result for CatBoost

Dataset Name	Feature Count	Tree Exp	Lime	Sampling Exp	Permutation Exp
Telco Customer Churn Dataset	21	0	0	0	0
Employee Attrition Dataset	35	0	0	0	0
Diabetes Dataset	9	0.010	0.000	0.026	0.016
Heart Disease Dataset	14	0	0	0	0
Stroke Prediction Dataset	12	0	0	0	0
House Price Prediction Dataset	81	0	0	0	0
Student Performance Dataset	34	0	0	0	0
US Cars Dataset	13	0	0	0	0
Medical Cost Personal Datasets	7	0.227	0.048	0.122	0.113
Crab Age Prediction	9	0.001	0.000	0.002	0.001
Average		**0.024**	**0.005**	**0.015**	**0.013**

An essential consideration highlighted by this study is the inverse relationship between the number of features and the likelihood of attaining a high monotonicity score. This correlation is substantiated by the data presented in Tables 6, 7, and 8, where monotonicity scores were assessed for feature counts of 7, 9, and 12, respectively.

As shown in Table 9, upon a comprehensive evaluation of the aggregated results across all employed machine learning models and datasets, it is discernible that the TreeExplainer method consistently furnishes the most favorable monotonicity scores. This outcome underscores the efficacy of the TreeExplainer technique in enhancing the interpretability of machine learning models through superior monotonicity, thereby affirming its integral role in the domain of explainable AI.

Table 9
Monotonicity result

TreeExp	Lime	SamplingExp	PermutationExp
0.030	0.004	0.021	0.024

5.3 Model Parameter Randomization

When performing model parameter randomization, the steps are taken as follows. By changing the parameters of the machine learning models, euclidian distance is taken between the explainable AI weights that are produced without changing the parameters and the explainable AI weights whose parameters have been changed at a certain scale. The changed parameters are as follows: max depth, learning rate, min child weight, gamma, alpha, lambda, etc. The applied scaling is 0.01, 0.02, 0.1, 0.2, 0.3, 0.4, 0.5. When applying this scale, the parameters are randomly increased or decreased at this rate.

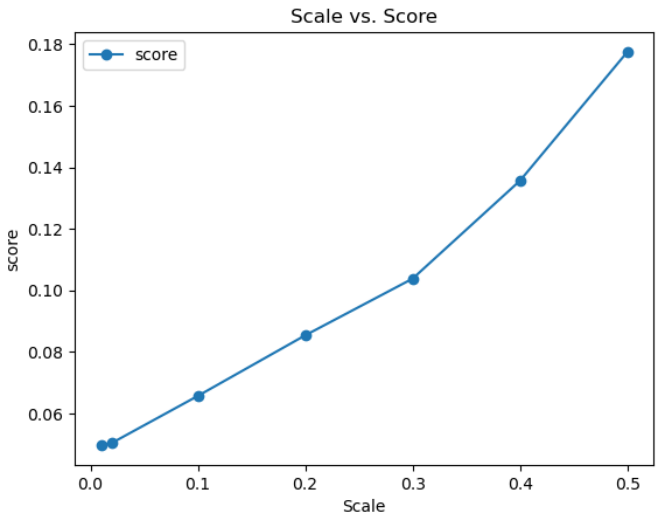

Figure 4. Scaling and Euclidian distance relation of Telco Customer Churn Dataset with XGBoost and TreeExplainer

In Figure 4, model parameter randomization was applied to the Telco Customer Churn Dataset using XGBoost and TreeExplainer. The scale shows the amount of scale applied to the model parameters (increasing or decreasing the learning rate by 0.1). The score section shows the Euclidian distance between the weights resulting from the model whose parameters have not been changed at all and the model whose parameters have been changed up to scale. Looking at the figure, it shows that as the scale ratio increases, the resulting weights differ from each other. The results indicate a positive correlation between the degree of parameter modification in the machine learning model and the resultant variation in the weights assigned by explainable AI methods.

This correlation suggests that the weights are sensitive to and affected by the extent of parameter alteration. To recognize this relationship, the relationship between scale and score was characterized by calculating the correlation. The correlation of the relationship shown in Figure 4 is 0.99.

The important point here is that as the scale changes, the model results must be similar to each other to take Euclidian distance and make comparisons. Because if the model results are not similar to each other, the change between the weights will not be due only to parameter changes. Since the model itself changes, the weights will be different.

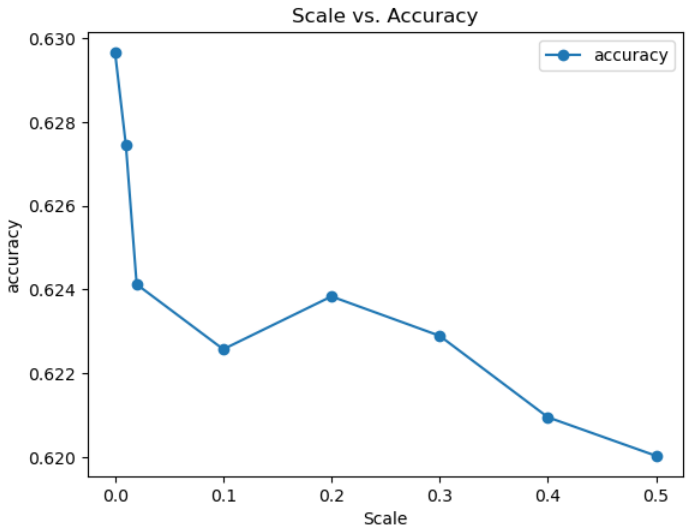

Figure 5. Scaling and F1-score relation of Telco Customer Churn Dataset with XGBoost and TreeExplainer

Figure 5 shows this relationship. It is shown how much the F1 score changes as you change the scale. A limitation was placed by calculating the mean absolute percentage between the F1-score between the base model and the models with modified parameters. If this rate is more than 6 percent, it was decided that no comparison should be made. In the example shown in Figure 5, the mean absolute percentage is 1 percent. For this reason, model parameter randomization calculations can be made. Table 11, Table 13, and Table 15 shows mean absolute percentage values according to datasets and models. Model parameter randomization was not calculated for datasets with values greater than 6 percent mean absolute percentage, and the model parameter randomization results are shown in Table 10, Table 12, and Table 14.

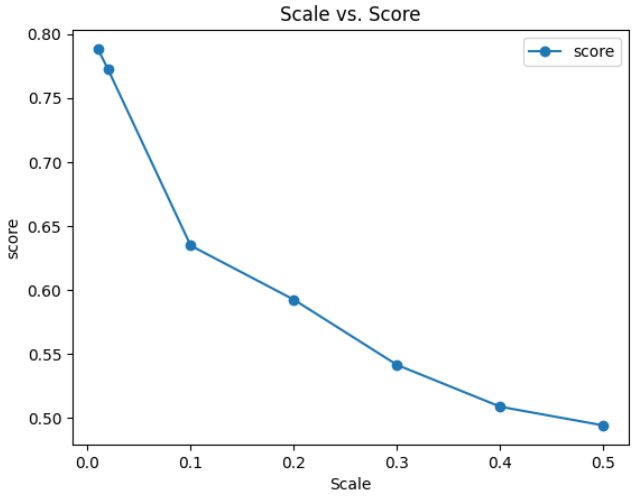

Figure 6. Scaling and Euclidian distance relation of Student Performance Dataset with CatBoost and SamplingExplainer

To give another example, in Figure 6, model parameter randomization was applied to the Student Performance Dataset using Catboost and SamplingExplainer. As the scale ratio increases, the resulting weights show that they are similar to each other. In other words, it seems that a slight change in this model increases the difference between the weights greatly. However, it seems that this difference is not much for extremely large parameter changes. The correlation of the relationship shown in Figure 6 is -0.94.

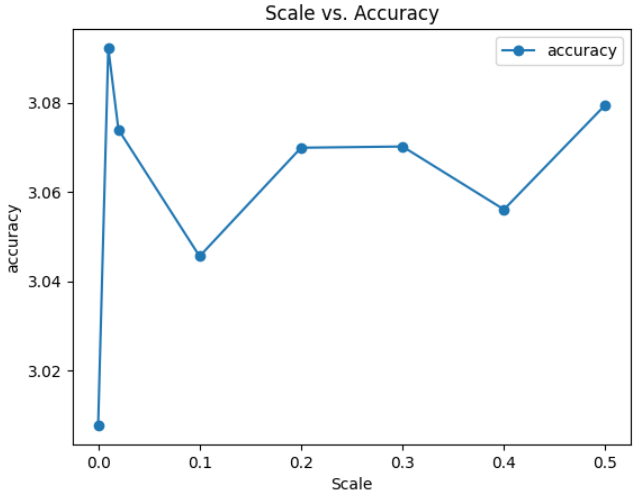

Figure 7. Scaling and root mean square error relation of Student Performance Dataset with CatBoost and SamplingExplainer

In the example shown in Figure 7, the mean absolute percentage is 2 percent. For this reason, model parameter randomization calculation can be made.

Looking at the model parameter randomization results in Table 10, the results are close to each other. The correlation between scaling and Euclidian distance is positive.

Table 10
Model Parameter Randomization Correlation for XGBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.99	0.99	0.97	1.00
Employee Attrition Dataset	0.99	0.98	0.98	1.00
Diabetes Dataset	0.99	0.99	0.99	0.99
Heart Disease Dataset	1.00	0.99	1.00	1.00
Student Performance Dataset	0.67	0.67	0.55	0.56
US Cars Dataset	0.99	0.95	0.97	0.97
Crab Age Prediction	0.98	0.91	0.98	0.97
Average	**0.94**	**0.93**	**0.92**	**0.93**

Table 11
Model Parameter Randomization Mean Absolute Percentage for XGBoost

Dataset Name	Tree Exp (%)	Lime (%)	Sampling Exp (%)	Permutation Exp (%)
Telco Customer Churn Dataset	1	1	1	1
Employee Attrition Dataset	4	5	5	4
Diabetes Dataset	5	5	5	5
Heart Disease Dataset	0	0	0	0
Stroke Prediction Dataset	8	8	5	7
House Price Prediction Dataset	7	8	9	8
Student Performance Dataset	4	3	3	4
US Cars Dataset	3	3	3	3
Medical Cost Personal Datasets	15	11	20	10
Crab Age Prediction	3	3	3	2

Looking at the model parameter randomization results in Table 12, the correlation between scaling and Euclidian distance is positive.

Table 12
Model Parameter Randomization Correlation for LightGBM

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.97	0.95	0.86	0.96
Employee Attrition Dataset	0.99	0.98	1.00	0.99
Diabetes Dataset	0.83	0.94	0.86	0.96
Heart Disease Dataset	0.96	0.99	0.98	0.98
Stroke Prediction Dataset	0.99	1.00	1.00	0.99
Student Performance Dataset	1.00	1.00	0.99	0.99
US Cars Dataset	1.00	1.00	1.00	1.00
Crab Age Prediction	0.25	0.98	-0.23	0.83
Average	**0.87**	**0.98**	**0.81**	**0.96**

Table 13
Model Parameter Randomization Mean Absolute Percentage for LightGBM

Dataset Name	Tree Exp (%)	Lime (%)	Sampling Exp (%)	Permutation Exp(%)
Telco Customer Churn Dataset	1	1	1	1
Employee Attrition Dataset	3	3	3	3
Diabetes Dataset	5	4	4	5
Heart Disease Dataset	0	0	0	0
Stroke Prediction Dataset	3	3	3	4
House Price Prediction Dataset	9	8	7	8
Student Performance Dataset	2	2	2	3
US Cars Dataset	2	1	2	2
Medical Cost Personal Datasets	11	10	16	15
Crab Age Prediction	1	3	1	2

Looking at the model parameter randomization results in Table 14, the correlation between scaling and Euclidian distance is generally negative. There is only a positive correlation in Lime, but it is not high either.

Table 14
Model Parameter Randomization Correlation for CatBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	-0.66	-0.03	-0.34	0.00
Heart Disease Dataset	-0.83	-0.81	-0.11	-0.91
Student Performance Dataset	-0.94	-0.78	-0.94	-0.93
US Cars Dataset	0.24	0.82	0.38	-0.74
Average	**-0.55**	**0.23**	**-0.25**	**-0.65**

Table 15
Model Parameter Randomization Mean Absolute Percentage for CatBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	2	1	1	1
Employee Attrition Dataset	20	18	20	20
Diabetes Dataset	9	9	7	9
Heart Disease Dataset	3	5	3	4
Stroke Prediction Dataset	9	8	8	8
House Price Prediction Dataset	12	12	10	13
Student Performance Dataset	2	2	2	2
US Cars Dataset	6	5	6	6
Medical Cost Personal Datasets	15	16	11	13
Crab Age Prediction	11	10	8	11

A review of the aggregate data across various explainable AI methods reveals a consistent correlation between the scaling factor and the Euclidean distance among models. This relationship suggests that the assignment of weights by the explainable AI methods is sensitive to the degree of change in the model parameters. Notably, the LIME model exhibits the most pronounced correlation in this context.

Table 16
Model parameter randomization result

TreeExp	Lime	SamplingExp	PermutationExp
0.6	0.8	0.63	0.61

5.4 Identity

This study embarked on a comprehensive analysis of identity scores derived from the application of three advanced machine learning algorithms: XGBoost, LightGBM, and CatBoost. Each algorithm was evaluated in conjunction with four distinct explainable AI methodologies to ascertain their effectiveness in yielding optimal identity scores. Findings unequivocally indicate that the TreeExplainer method consistently outperformed its counterparts in facilitating the highest identity scores across all tested machine learning frameworks.

Table 17
Identity result for XGBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	100	0	0	0
Employee Attrition Dataset	100	0	0	0
Diabetes Dataset	100	0	0	0
Heart Disease Dataset	100	0	0	0
Stroke Prediction Dataset	100	0	0	0
House Price Prediction Dataset	100	0	0	0
Student Performance Dataset	100	0	0	0
US Cars Dataset	100	0	0	0
Medical Cost Personal Datasets	100	0	0	0
Crab Age Prediction	100	0	0	0
Average	**100**	**0**	**0**	**0**

Table 18
Identity result for LightGBM

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	100	0	0	0
Employee Attrition Dataset	100	0	0	0
Diabetes Dataset	100	0	0	0
Heart Disease Dataset	100	0	0	0
Stroke Prediction Dataset	100	0	0	0
House Price Prediction Dataset	100	0	0	0
Student Performance Dataset	100	0	0	0
US Cars Dataset	100	0	0	0
Medical Cost Personal Datasets	100	0	0	0
Crab Age Prediction	100	0	0	0
Average	**100**	**0**	**0**	**0**

Table 19
Identity result for CatBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	100	0	0	0
Employee Attrition Dataset	100	0	0	0
Diabetes Dataset	100	0	0	0
Heart Disease Dataset	100	0	0	0
Stroke Prediction Dataset	100	0	0	0
House Price Prediction Dataset	100	0	0	0
Student Performance Dataset	100	0	0	0
US Cars Dataset	100	0	0	0
Medical Cost Personal Datasets	100	0	0	0
Crab Age Prediction	100	0	0	0
Average	**100**	**0**	**0**	**0**

As shown in Table 20, a generalized analysis encompassing various machine learning methodologies and datasets, it has been empirically ascertained that the TreeExplainer method consistently secures the highest average identity scores.

Table 20
Identity result

TreeExp	Lime	SamplingExp	PermutationExp
100	0	0	0

5.5 Separability

The analysis focused on evaluating separability scores obtained by employing the XGBoost, LightGBM, and CatBoost machine learning algorithms, each analyzed in conjunction with four different explainable AI methods. Results indicate that the TreeExplainer method consistently achieved the highest Separability score across all three machine learning platforms. When considering the average performance across various datasets and machine learning methodologies in Table 24, the TreeExplainer method was identified as the most effective in enhancing separability scores. It is noteworthy that the SamplingExplainer and PermutationExplainer methods also demonstrated competitive performance, closely approximating the scores achieved by the TreeExplainer method.

Table 21
Separability result for XGBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.50	0.51	0.47	0.46
Employee Attrition Dataset	0.12	0.16	0.13	0.14
Diabetes Dataset	0.42	0.36	0.46	0.46
Heart Disease Dataset	0.46	0.40	0.37	0.38
Stroke Prediction Dataset	0.46	0.48	0.47	0.47
House Price Prediction Dataset	0.65	0.36	0.66	0.66
Student Performance Dataset	0.39	0.30	0.40	0.40
US Cars Dataset	0.30	0.13	0.29	0.29
Medical Cost Personal Datasets	0.27	0.22	0.27	0.27
Crab Age Prediction	0.89	0.82	0.91	0.90
Average	**0.45**	**0.37**	**0.442**	**0.443**

Table 22
Separability result for LightGBM

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.57	0.55	0.49	0.48
Employee Attrition Dataset	0.08	0.14	0.12	0.12
Diabetes Dataset	0.42	0.38	0.48	0.47
Heart Disease Dataset	0.45	0.43	0.37	0.39
Stroke Prediction Dataset	0.43	0.43	0.44	0.44
House Price Prediction Dataset	0.53	0.34	0.56	0.56
Student Performance Dataset	0.35	0.31	0.35	0.35
US Cars Dataset	0.35	0.21	0.34	0.35
Medical Cost Personal Datasets	0.27	0.22	0.27	0.27
Crab Age Prediction	0.91	0.79	0.90	0.90
Average	**0.44**	**0.38**	**0.432**	**0.432**

Table 23
Separability result for CatBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.54	0.58	0.46	0.44
Employee Attrition Dataset	0.26	0.17	0.13	0.12
Diabetes Dataset	0.42	0.36	0.48	0.48
Heart Disease Dataset	0.60	0.49	0.35	0.36
Stroke Prediction Dataset	0.40	0.42	0.42	0.42
House Price Prediction Dataset	0.57	0.38	0.55	0.55
Student Performance Dataset	0.39	0.34	0.40	0.40
US Cars Dataset	0.32	0.26	0.33	0.33
Medical Cost Personal Datasets	0.27	0.22	0.28	0.28
Crab Age Prediction	0.76	0.79	0.81	0.80
Average	**0.45**	**0.40**	**0.419**	**0.418**

Table 24
Separability result

TreeExp	Lime	SamplingExp	PermutationExp
0.44	0.38	0.431	0.432

5.6 Stability

Upon evaluating to ascertain the stability scores using the XGBoost, LightGBM, and CatBoost machine learning algorithms, each complemented by four distinct explainable AI methods, it was observed that the PermutationExplainer method consistently yielded the highest stability scores across all machine learning frameworks.

Table 25
Stability result for XGBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.50	0.50	0.72	0.72
Employee Attrition Dataset	0.73	0.73	0.73	0.73
Diabetes Dataset	0.76	0.77	0.77	0.78
Heart Disease Dataset	0.98	0.71	1	1
Stroke Prediction Dataset	0.66	0.77	0.72	0.73
Average	**0.73**	**0.70**	**0.788**	**0.793**

Table 26
Stability result for LightGBM

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.50	0.50	0.71	0.71
Employee Attrition Dataset	0.73	0.73	0.73	0.73
Diabetes Dataset	0.75	0.77	0.77	0.77
Heart Disease Dataset	0.94	0.71	0.98	0.98
Stroke Prediction Dataset	0.69	0.77	0.69	0.69
Average	**0.72**	**0.70**	**0.7766**	**0.7769**

Table 27
Stability result for CatBoost

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.51	0.64	0.71	0.72
Employee Attrition Dataset	0.73	0.73	0.73	0.73
Diabetes Dataset	0.70	0.77	0.77	0.77
Heart Disease Dataset	0.92	0.80	1	1
Stroke Prediction Dataset	0.69	0.77	0.70	0.70
Average	**0.71**	**0.74**	**0.7827**	**0.7833**

As shown in Table 28, a comprehensive analysis across various machine learning algorithms and datasets reveals that the PermutationExplainer method consistently achieves the highest average stability score. Furthermore, the SamplingExplainer method also demonstrates a performance closely aligned with that of the PermutationExplainer method, indicating its effectiveness in providing stable results.

Table 28
Stability result

TreeExp	Lime	SamplingExp	PermutationExp
0.72	0.71	0.783	0.784

5.7 Duration

For each instance, the duration required to achieve explainability was meticulously documented. Across three distinct machine learning models, the TreeExplainer method emerged as the most efficient, delivering explainability insights in the shortest timeframe. Conversely, the SamplingExplainer method was identified as the most time-consuming approach for generating explainability results. As shown in Table 31, this pattern held true on average, highlighting consistent performance trends across the evaluated methodologies.

Table 29
Duration result for XGBoost (seconds)

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.0001	0.0500	0.6928	0.0318
Employee Attrition Dataset	0.0005	0.0964	0.7308	0.0764
Diabetes Dataset	0.0004	0.0363	0.0920	0.0783
Heart Disease Dataset	0.0004	0.0387	0.1526	0.0569
Stroke Prediction Dataset	0.0009	0.0762	0.4255	0.1138
House Price Prediction Dataset	0.0003	0.1722	3.0341	0.0423
Student Performance Dataset	0.0004	0.0538	0.3454	0.0341
US Cars Dataset	0.0004	0.3437	3.0350	0.0410
Medical Cost Personal Datasets	0.0001	0.0254	0.0784	0.0264
Crab Age Prediction	0.0002	0.0318	0.1485	0.0589
Average	**0.0004**	**0.0924**	**0.8735**	**0.0560**

Table 30
Duration result for LightGBM (seconds)

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.00003	0.0476	0.7047	0.0364
Employee Attrition Dataset	0.0002	0.0833	0.6715	0.0747
Diabetes Dataset	0.0001	0.0378	0.0943	0.1072
Heart Disease Dataset	0.0003	0.0619	0.2031	0.1148
Stroke Prediction Dataset	0.00004	0.0370	0.3031	0.0431
House Price Prediction Dataset	0.0002	0.1822	3.7640	0.0673
Student Performance Dataset	0.0001	0.0602	0.3604	0.0279
US Cars Dataset	0.0003	0.3305	4.1336	0.0580
Medical Cost Personal Datasets	0.0001	0.0361	0.0956	0.0622
Crab Age Prediction	0.0001	0.0490	0.1913	0.1048
Average	**0.0001**	**0.0926**	**1.0522**	**0.0696**

Table 31
Duration result for CatBoost (seconds)

Dataset Name	TreeExp	Lime	SamplingExp	PermutationExp
Telco Customer Churn Dataset	0.00004	0.0410	1.0641	0.0192
Employee Attrition Dataset	0.0002	0.0682	1.0144	0.0160
Diabetes Dataset	0.0001	0.0176	0.0885	0.0195
Heart Disease Dataset	0.0002	0.0221	0.1847	0.0196
Stroke Prediction Dataset	0.0005	0.0239	0.3991	0.0157
House Price Prediction Dataset	0.0013	0.2927	15.6924	0.0470
Student Performance Dataset	0.0001	0.0637	0.7463	0.0220
US Cars Dataset	0.0014	0.2491	12.0166	0.0406
Medical Cost Personal Datasets	0.00002	0.0143	0.1063	0.0200
Crab Age Prediction	0.0027	0.0182	0.1882	0.0250
Average	**0.0006**	**0.0811**	**3.1500**	**0.0245**

Table 32
Duration result

TreeExp	Lime	SamplingExp	PermutationExp
0.004	0.0887	1.6919	0.050

5.8 Discussion

Table 33 presents a systematic ranking of the performance of various explainable AI models across a suite of evaluation metrics.

Table 33
Overall result

Evaluation Metric	TreeExp	Lime	SamplingExp	PermutationExp
Faithfulness	3	4	1	2
Monotonicity	1	4	3	2
Identity	1	2	2	2
Separability	1	4	3	2
Stability	3	4	2	1
Duration	1	3	4	2
Average	**1.6667**	**3.5000**	**2.5000**	**1.8333**

In the dimension of faithfulness, the ranking from highest to lowest performance is as follows: SamplingExplainer, PermutationExplainer, TreeExplainer, and LIME. For the monotonicity metric, TreeExplainer leads, followed by PermutationExplainer, SamplingExplainer, and LIME. TreeExplainer also emerges as the superior model in terms of identity, with the other models demonstrating comparable performance. In

assessing separability, the order of performance is TreeExplainer, PermutationExplainer, SamplingExplainer, and LIME. For stability, the PermutationExplainer method is the most proficient, succeeded by SamplingExplainer, TreeExplainer, and LIME. Regarding the duration for explainability, TreeExplainer is the most expedient, followed by PermutationExplainer, LIME, and SamplingExplainer.

In the faithfulness dimension, the results favor the SamplingExplainer method, suggesting that it may be more aligned with human intuition and logic, and could potentially be more trustworthy in scenarios where the explanation's fidelity to the model's behavior is critical. Conversely, LIME's lower ranking might indicate that it struggles in certain contexts, possibly due to the local approximation approach it employs.

The superior performance of TreeExplainer in the monotonicity and identity metrics might be attributed to its model-specific nature, being tailored for tree-based algorithms, which can leverage the inherent structural information in such models. The high ranking of TreeExplainer in these metrics suggests that it could be highly reliable for interpreting tree-based models, such as those used in financial or medical decision-making processes where understanding feature interactions and contributions is crucial.

For separability, TreeExplainer's prominent position may highlight its proficiency in differentiating the explanations for distinct instances. This ability is crucial in complex datasets where discerning the unique contribution of individual features to the model's predictions is necessary for understanding and trust.

The PermutationExplainer method's top ranking in stability is noteworthy. It implies that it is less likely to produce widely varying explanations for similar instances, which is an important trait for user trust and the iterative model development process. The stability of an explanation method is especially important in regulated industries, where consistency of explanation can be as important as accuracy.

The duration metric, where TreeExplainer leads, is particularly relevant in real-time or operational environments where the speed of explanation generation can be a critical factor. A model's ability to quickly provide explanations is valuable for dynamic systems that require immediate human interpretability, such as autonomous vehicle decision systems or real-time trading algorithms.

When these rankings are synthesized into an overall performance score and averaged, the models rank as follows: TreeExplainer, PermutationExplainer, SamplingExplainer, and LIME. It is noteworthy that the TreeExplainer model is model-agnostic but is exclusively applicable to tree-based machine learning algorithms. In cases where alternative algorithms are to be utilized, the PermutationExplainer model is recommended as the most suitable.

It is critical to acknowledge that these rankings are not absolute and may vary based on the specific dataset and the context of the application. Different domains and problem sets may place varying levels of importance on each of these metrics, and as such, the selection of an XAI method should be context-dependent. Furthermore, as the field of XAI continues to evolve, the development of new methods or the improvement of existing ones could alter these rankings. Therefore, ongoing research and validation across diverse datasets and use cases remain essential to ascertain the robustness and generalizability of these findings.

Chapter 6

Conclusion

This research aims to develop a methodology for the evaluation of the rapidly evolving and diversifying field of explainable artificial intelligence (XAI) methods. The evaluation of XAI models remains a significant challenge in the literature, underlining the necessity for reliable and reproducible metrics to measure the effectiveness of XAI methods. The study seeks to address this challenge by offering a comprehensive analysis across 10 datasets from various domains, encompassing both binary classification and regression problems, and utilizing a variety of machine learning models including XGBoost, LightGBM, and CatBoost. Subsequently, four different XAI methods—TreeExplainer, LIME, SamplingExplainer, and PermutationExplainer—were evaluated against seven distinct evaluation metrics: faithfulness, monotonicity, model parameter randomization, identity, separability, stability, and duration.

Findings reveal that each XAI method excels in specific metrics, thereby highlighting the strengths and weaknesses of each method. Overall, the TreeExplainer method demonstrated the best performance, while the LIME method showed lower performance compared to the other examined methods. The PermutationExplainer method emerged as the second-best performer, followed by the SamplingExplainer method. These results, generalized from the datasets analyzed, have the potential to significantly contribute to other studies in the XAI field.

The research establishes a solid foundation for the evaluation of XAI models, emphasizing the importance of developing standardized evaluation metrics for XAI methods, similar to those in machine learning models. Such standardization would enhance the transparency, comprehensibility, and reliability of XAI applications, thus increasing their general acceptance and applicability. Furthermore, by providing a comparative analysis of the advantages and limitations of different XAI methods, our work facilitates more informed choices and applications of these methods.

In conclusion, this thesis contributes significantly to the current research in XAI, offering a roadmap for future studies. The proposed evaluation metrics could accelerate scientific progress in the field by enabling more effective evaluation of XAI

methods. Additionally, our study provides valuable insights into how XAI can be more effectively utilized across different domains, aiming to maximize the potential of XAI.

REFERENCES

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160

AHMED, N. A., & Alpkoçak, A. (2022). A quantitative evaluation of explainable AI methods using the depth of decision tree. Turkish Journal of Electrical Engineering and Computer Sciences, 30(6), 2054-2072.

Akturk, M. (2020). Diabetes dataset. Retrieved from https://www.kaggle.com/datasets/mathchi/diabetes-data-set

Alsenani, D. (2020). US cars dataset. Retrieved from https://www.kaggle.com/datasets/doaaalsenani/usa-cers-dataset

Amparore, E., Perotti, A., & Bajardi, P. (2021). To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods. PeerJ Computer Science, 7, e479.

Ansodariya, D. (2022). Student performance dataset. Retrieved from https://www.kaggle.com/datasets/devansodariya/student-performance-data

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, 82-115.

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... & Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012.

Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. Frontiers in big Data, 39.

BlastChar. (2018). Telco customer churn. Retrieved from https://www.kaggle.com/datasets/blastchar/telco-customer-churn

Buonocore, T. M., Nicora, G., Dagliati, A., & Parimbelli, E. (2022). Evaluation of XAI on ALS 6-months mortality prediction. In CLEF (Working Notes) (pp. 1228-1235).

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research, 70, 245–317. doi:10.1613/jair.1.12228

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big data & society, 3(1), 2053951715622512

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. Electronics, 8(8), 832. doi:10.3390/electronics8080832

Choi, M. (2018). Medical Cost Personal Datasets. Retrieved from https://www.kaggle.com/datasets/mirichoi0218/insurance

Coroama, L., & Groza, A. (2022, September). Evaluation Metrics in Explainable Artificial Intelligence (XAI). In International Conference on Advanced Research in Technologies, Information, Innovation and Sustainability (pp. 401-413). Cham: Springer Nature Switzerland.

Doumard, E., Aligon, J., Escriva, E., Excoffier, J. B., Monsarrat, P., & Soulé-Dupuy, C. (2022, March). A comparative study of additive local explanation methods based on feature influences. In 24th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data ((DOLAP 2022) (Vol. 3130, No. paper 4, pp. 31-40). CEUR-WS. org.

Doumard, E., Aligon, J., Escriva, E., Excoffier, J. B., Monsarrat, P., & Soulé-Dupuy, C. (2023). A quantitative approach for the comparison of additive local explanation methods. Information Systems, 114, 102162.

Duell, J., Fan, X., Burnett, B., Aarts, G., & Zhou, S.-M. (2021). A comparison of explanations given by explainable artificial intelligence methods on analysing electronic health records. 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). doi:10.1109/bhi50953.2021.9508618

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... & Ranjan, R. (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. ACM Computing Surveys, 55(9), 1-33.

El Shawi, R., Sherif, Y., Al-Mallah, M., & Sakr, S. (2019). Interpretability in healthcare a comparative study of local machine learning interpretability techniques. 2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS). doi:10.1109/cbms.2019.00065

Fedesoriano. (2021). Stroke prediction dataset. Retrieved from https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022). Explainable AI methods-a brief overview. In International workshop on extending explainable AI beyond deep models and classifiers (pp. 13-38). Springer, Cham

Kaggle. (n.d.). House prices - advanced regression techniques. Retrieved from https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Lapp, D. (2019). Heart disease dataset. Retrieved from https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable ai: A review of machine learning interpretability methods. Entropy, 23(1), 18

Lozano-Murcia, C., Romero, F. P., Serrano-Guerrero, J., & Olivas, J. A. (2023). A Comparison between Explainable Machine Learning Methods for Classification and Regression Problems in the Actuarial Context. Mathematics, 11(14), 3088.

Lrjball. (n.d.). Lrjball/SHAP: A game theoretic approach to explain the output of any machine learning model. Retrieved from https://github.com/lrjball/shap

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

M El-gezawy, A. M., Abdel-Kader, H., & Ali, A. H. (2023). A New XAI Evaluation Metric for Classification. IJCI. International Journal of Computers and Information, 10(3), 58-62.

Molnar, C. (2023a). Interpretable machine learning. Retrieved from https://christophm.github.io/interpretable-ml-book/lime.html

Molnar, C. (2023). Interpretable machine learning. Retrieved from https://christophm.github.io/interpretable-ml-book/shapley.html#shapley

Molnar, C. (2023). Interpretable machine learning. Retrieved from https://christophm.github.io/interpretable-ml-book/shap.html

Molnar, C. (2023a). Interpretable machine learning. Retrieved from https://christophm.github.io/interpretable-ml-book/shap.html#kernelshap

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., ... & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. ACM Computing Surveys, 55(13s), 1-42

Oblizanov, A., Shevskaya, N., Kazak, A., Rudenko, M., & Dorofeeva, A. (2023). Evaluation Metrics Research for Explainable Artificial Intelligence Global Methods Using Synthetic Data. Applied System Innovation, 6(1), 26.

Patel, P. (2018). Employee attrition. Retrieved from https://www.kaggle.com/datasets/patelprashant/employee-attrition

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021, May). Manipulating and measuring model interpretability. In Proceedings of the 2021 CHI conference on human factors in computing systems (pp. 1-52)

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

Rosenfeld, A. (2021, May). Better metrics for evaluating explainable artificial intelligence. In Proceedings of the 20th international conference on autonomous agents and multiagent systems (pp. 45-50).

Sidhu, G. S. (2021). Crab age prediction. Retrieved from https://www.kaggle.com/datasets/sidhus/crab-age-prediction

Singh, V. (2021). Explainable ai metrics and properties for evaluation and analysis of counterfactual explanations: Explainable ai metrics and properties for evaluation and analysis of counterfactual explanations.

Yalcin, O., Fan, X., & Liu, S. (2021). Evaluating the correctness of explainable AI algorithms for classification. arXiv preprint arXiv:2105.09740.