Comparative Analysis of Explainability in Machine Learning Algorithms

Gaurav Pradeep
Roll Number: 47
220962308

Navya Kulhari
Roll Number: 50
220962318

Aditya Aggarwal
Roll Number: 59
220962374

***Abstract*—This study aims to evaluate and compare the explainability and performance of multiple machine learning (ML) algorithms, specifically Logistic Regression, Decision Tree, SVM, Neural Network, and XGBoost, across three datasets (Pima, Crime, and Breast Cancer). Using interpretability methods such as LIME and SHAP, this analysis explores the trade-offs between model accuracy and interpretability. The insights gained provide guidance on the best algorithms for applications that prioritize transparency and help establish standardized methods for assessing model explainability in ML.**

***Index Terms*—Explainable AI, Machine Learning, Interpretability, LIME, SHAP, Pima Dataset, Crime Dataset, Breast Cancer Dataset**

I. Introduction

In recent years, machine learning (ML) has witnessed widespread adoption across various industries due to its ability to uncover complex patterns and make accurate predictions from large datasets. However, the deployment of ML models in critical sectors such as healthcare, finance, and criminal justice raises significant concerns regarding their transparency and accountability. In such high-stakes environments, the decisions made by ML models can have profound impacts on individuals and society, necessitating a clear understanding of how these models arrive at their predictions.

Traditional ML algorithms like Logistic Regression offer inherent interpretability, allowing stakeholders to easily understand feature contributions to the outcome. In contrast, more complex models such as Neural Networks, Support Vector Machines (SVMs), and ensemble methods like XGBoost often function as "black boxes," obscuring their internal decision-making processes. This opacity can hinder trust in the model, complicate compliance with regulatory requirements, and make it challenging to identify and correct biases.

Explainable Artificial Intelligence (XAI) has emerged as a crucial area of research aiming to bridge the gap between model complexity and interpretability. By providing tools and techniques to explain model predictions, XAI facilitates transparency and enables stakeholders to gain insights into the inner workings of complex models. Notably, methods such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations) have gained prominence for their ability to provide post-hoc explanations of model predictions.

This study explores the balance between model accuracy and explainability by evaluating various ML algorithms using LIME and SHAP across three diverse datasets: Pima, Crime, and Breast Cancer. Through comprehensive analysis, we aim to identify optimal models that deliver both high performance and adequate interpretability, thereby guiding practitioners in selecting suitable models for applications where transparency is as critical as accuracy.

II. Objectives

The primary objectives of this study are as follows:

- **Objective 1:** To evaluate the explainability of various machine learning algorithms using established interpretability techniques. This involves applying methods such as LIME and SHAP to models like Logistic Regression, Decision Trees, SVM, Neural Networks, and XGBoost to assess how effectively they can elucidate the reasoning behind individual predictions and overall model behavior.
- **Objective 2:** To analyze and compare the performance of these ML models using comprehensive evaluation metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. By doing so, we aim to understand the trade-offs between model accuracy and interpretability, identifying whether high-performing models necessarily compromise transparency.
- **Objective 3:** To apply LIME and SHAP to assess the quality of the explanations generated by the models, focusing on specific metrics such as fidelity (the degree to which the explanations approximate the original model) and sparsity (the simplicity of the explanation in terms of the number of features used). This will help in quantifying the robustness and practicality of the explanations provided.
- **Objective 4:** To provide evidence-based recommendations on the most suitable algorithms for high-stakes applications that require a balance of both accuracy and transparency, including considerations for regulatory compliance, ethical standards, and user trust.

III. Literature Survey

The quest for interpretable machine learning models has gained significant momentum in recent years, driven by the increasing complexity of models and the critical need for transparency in decision-making processes. According to Doshi-Velez and Kim (2017) [4], interpretability is essential for

ensuring that ML models align with human values, ethical standards, and regulatory requirements.

LIME, proposed by Ribeiro et al. (2016) [2], is a model-agnostic interpretability technique that provides local explanations by approximating the complex model with an interpretable surrogate model around the vicinity of the instance being predicted. This method has been widely adopted due to its flexibility and ease of use across different model types.

SHAP, introduced by Lundberg and Lee (2017) [3], offers a unified approach to interpreting model predictions by leveraging concepts from cooperative game theory. SHAP values quantify the contribution of each feature to the prediction, ensuring consistency and local accuracy. It provides both global interpretability by aggregating feature contributions across the dataset and local interpretability for individual predictions.

Several studies have highlighted the challenges associated with interpreting complex models. For instance, Gunning and Aha (2019) [5] emphasized the importance of XAI in making AI systems more transparent and trustworthy. However, as noted by Lipton (2018) [6], there is often a trade-off between model complexity and interpretability, and achieving both remains a significant challenge.

Metrics such as fidelity, which measures how well the explanations approximate the original model, and sparsity, which assesses the simplicity of explanations, have been proposed to evaluate interpretability methods. Nevertheless, as Molnar (2019) [7] points out, the lack of standardized metrics and evaluation frameworks complicates the comparison of interpretability methods across different models and datasets.

This study builds upon existing research by systematically evaluating both performance and explainability of various ML models using LIME and SHAP across multiple datasets. By doing so, we aim to contribute to the development of standardized methods for assessing model explainability, facilitating better comparisons and informed decision-making in the selection of ML models for critical applications.

IV. Methodology

A. Data Sources

To conduct a comprehensive analysis, we selected three datasets with varying characteristics to represent different types of classification and regression problems. These datasets were chosen to evaluate the models across diverse domains and data complexities.

- **Pima Indians Diabetes Dataset** [8]: Sourced from the UCI Machine Learning Repository, this dataset comprises 768 instances and 8 numerical attributes related to medical diagnostic measurements. The features include attributes such as the number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, serum insulin, body mass index (BMI), diabetes pedigree function, and age. The binary target variable indicates whether a patient shows signs of diabetes according to World Health Organization criteria. This dataset was chosen due to its moderate size and relevance to healthcare, where interpretability is critical.
- **Communities and Crime Dataset** [9]: Also obtained from the UCI Machine Learning Repository, this dataset contains socio-economic, law enforcement, and demographic data for different communities in the United States. It includes 1,994 instances and 128 features after excluding non-predictive and identifier attributes. The goal is to predict the per capita violent crime rate, making it a regression problem. This dataset was selected for its complexity and high-dimensional feature space, which poses challenges for both modeling and interpretability.
- **Breast Cancer Wisconsin (Diagnostic) Dataset** [10]: This dataset contains 569 instances with 30 numerical features computed from digitized images of fine needle aspirate (FNA) of breast masses. The features describe characteristics of the cell nuclei present in the images, such as radius, texture, perimeter, area, smoothness, and compactness. The binary target variable indicates whether the tumor is malignant or benign. The dataset is widely used in medical research, emphasizing the necessity for accurate and interpretable models in healthcare diagnostics.

B. Data Preprocessing

Each dataset underwent a series of preprocessing steps to ensure data quality and suitability for model training. Missing values were handled appropriately; for the Pima dataset, zeros in certain features (e.g., blood pressure, BMI) were replaced with the mean of the non-zero values. For the Crime dataset, features with excessive missing values were removed, and the remaining missing values were imputed using mean or median as appropriate. Features were standardized or normalized when necessary to facilitate model convergence and improve performance.

C. Models and Evaluation Techniques

We evaluated the following machine learning algorithms, selected for their varying levels of complexity and interpretability:

- **Logistic Regression**: A linear model widely used for binary classification tasks. It provides coefficients that directly indicate the influence of each feature on the prediction, making it inherently interpretable.
- **Decision Tree**: A tree-based model that makes decisions based on feature thresholds, resulting in a set of if-then rules. Decision Trees are easily interpretable but can overfit the data if not properly pruned.
- **Support Vector Machine (SVM)**: An algorithm that finds the hyperplane that best separates classes in the feature space. While effective in high-dimensional spaces, SVMs with non-linear kernels are less interpretable due to the complexity of the decision boundaries.
- **Neural Network**: A multi-layer perceptron model capable of capturing complex non-linear relationships. Neural Networks are considered black-box models due to their intricate internal structures, making interpretability a challenge.

- **XGBoost**: An optimized gradient boosting algorithm that builds an ensemble of weak learners (decision trees). It is known for its superior performance on structured data but lacks inherent interpretability.

For the regression task (Crime dataset), we utilized:

- **Linear Regression**: Provides a straightforward interpretation of feature coefficients.
- **Support Vector Regression (SVR)**: Extends SVM to regression problems.
- **Random Forest Regressor**: An ensemble of decision trees that improves predictive accuracy at the expense of interpretability.
- **XGBoost Regressor**: Applies gradient boosting to regression problems.

D. Explainability Techniques

To assess the interpretability of the models, we employed the following explainability techniques:

- **LIME (Local Interpretable Model-Agnostic Explanations)**: LIME explains individual predictions by perturbing the input and observing changes in the output, fitting a simple interpretable model (e.g., linear regression) locally around the instance of interest. This provides insights into which features most influenced a specific prediction.
- **SHAP (SHapley Additive exPlanations)**: SHAP computes Shapley values from cooperative game theory to quantify the contribution of each feature to the prediction. It provides consistent and locally accurate feature attributions, offering both global and local interpretability.

E. Evaluation Metrics

To comprehensively evaluate model performance and interpretability, we employed the following metrics:

- **Performance Metrics for Classification**: Accuracy, Precision, Recall, F1-Score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics provide a balanced view of model performance, especially in datasets with class imbalance.
- **Performance Metrics for Regression**: Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R^2) score.
- **Explainability Metrics**:
  - **Fidelity**: Measures how well the interpretability method approximates the original model's predictions. High fidelity indicates that the explanations are representative of the model's behavior.
  - **Sparsity**: Refers to the number of features used in the explanation. A sparser explanation is often more interpretable, as it involves fewer features.

F. Experimental Setup

All experiments were conducted using Python 3.8 with scikit-learn for model implementation, and the LIME and SHAP libraries for interpretability analysis. The datasets were split into training and testing sets using an 80/20 split, ensuring that the distribution of classes was maintained (stratified sampling for classification tasks). Hyperparameter tuning was performed using grid search with cross-validation where necessary to optimize model performance.

V. Results and Discussion

The performance and explainability results for each dataset are presented in Tables I-A and I-B (Pima Dataset), Table II (Crime Dataset), and Tables III-A and III-B (Breast Cancer Dataset). These tables provide a comprehensive overview of each model's predictive performance and the quality of the explanations generated by LIME and SHAP.

A. Pima Indians Diabetes Dataset

1) Model Performance: As shown in Table I-A, the SVM model achieved the highest accuracy (75.32%) and F1-Score (63.46%), indicating superior predictive performance compared to other models. Logistic Regression and Decision Tree models demonstrated lower accuracy but offer higher interpretability.

TABLE I
Model Performance for Pima Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	71.43%	60.87%	51.85%	56.00%	82.30%
Decision Tree	72.08%	63.41%	48.15%	54.74%	66.57%
SVM	**75.32%**	66.00%	**61.11%**	**63.46%**	79.24%
Neural Network	74.03%	**65.91%**	53.70%	59.18%	**81.70%**
XGBoost	74.68%	65.31%	59.26%	62.14%	80.56%

2) Explainability Analysis: Table I-B presents the explainability metrics for the Pima dataset. The LIME Fidelity scores indicate that Logistic Regression had a higher fidelity (0.6354) compared to the Decision Tree (0.2658), suggesting that LIME explanations better approximated the Logistic Regression model. LIME and SHAP sparsity values were consistent across models, with the Decision Tree exhibiting slightly higher SHAP sparsity, indicating more complex explanations.

TABLE II
Explainability Metrics for Pima Dataset

Model	LIME Fidelity	LIME Sparsity	SHAP Sparsity
Logistic Regression	0.6354	8	8
Decision Tree	0.2658	8	16
SVM	0.5985	8	8
Neural Network	0.5977	8	8
XGBoost	0.5132	8	8

3) Discussion: The SVM model, despite its higher accuracy, poses challenges for interpretability. The Logistic Regression model offers a balance between performance and transparency, making it suitable for medical applications where understanding feature contributions is essential. The Decision Tree model, while interpretable, showed lower fidelity in LIME explanations, potentially due to overfitting or complexity in the tree structure.

B. Communities and Crime Dataset

1) Model Performance: Table II illustrates that the Random Forest Regressor achieved the lowest MAE (0.0877) and highest R^2 (0.7350), indicating strong predictive capabilities. Linear Regression performed poorly with a negative R^2 (-2.2493), reflecting its inadequacy in capturing the complex relationships in the data.

TABLE III
MODEL PERFORMANCE AND EXPLAINABILITY FOR CRIME DATASET

Model	MAE	MSE	RMSE	R^2	SHAP Sparsity
Linear Regression	0.3067	0.1635	0.4044	-2.2493	122
Decision Tree	0.1652	0.0443	0.2105	0.1198	53
SVR	0.1034	0.0152	0.1234	0.6975	122
Random Forest	**0.0877**	**0.0133**	**0.1155**	**0.7350**	122
XGBoost	0.0949	0.0177	0.1330	0.6485	102

2) Explainability Analysis: The SHAP sparsity values suggest that the Random Forest and SVR models required a larger number of features (122) to explain the predictions, highlighting the complexity of their decision-making processes. The Decision Tree model, with a SHAP sparsity of 53, provided more concise explanations at the expense of predictive performance.

3) Discussion: The Random Forest model's superior performance comes with increased complexity in explanations, which may hinder interpretability. In contrast, the Decision Tree offers simpler explanations but with lower predictive accuracy. This underscores the trade-off between performance and interpretability in regression tasks involving high-dimensional data.

C. Breast Cancer Wisconsin Dataset

1) Model Performance: As shown in Table III-A, the SVM and Neural Network models achieved the highest accuracy (97.37%) and F1-Score (96.30%), closely followed by Logistic Regression. All models demonstrated high performance due to the well-structured nature of the dataset.

TABLE IV
MODEL PERFORMANCE FOR BREAST CANCER DATASET

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	96.49%	97.50%	92.86%	95.12%	**99.60%**
Decision Tree	92.98%	90.48%	90.48%	90.48%	92.46%
SVM	**97.37%**	100.00%	92.86%	96.30%	99.47%
Neural Network	**97.37%**	100.00%	92.86%	96.30%	99.57%
XGBoost	95.61%	100.00%	88.10%	93.67%	99.34%

2) Explainability Analysis: Table III-B indicates that the LIME Fidelity scores were relatively low across all models, with the Decision Tree having the lowest fidelity (0.0820). The sparsity values suggest that explanations involved a moderate number of features, which is acceptable given the importance of comprehensive feature analysis in medical diagnostics.

TABLE V
EXPLAINABILITY METRICS FOR BREAST CANCER DATASET

Model	LIME Fidelity	LIME Sparsity	SHAP Sparsity
Logistic Regression	0.1689	10	30
Decision Tree	0.0820	10	26
SVM	0.1899	10	30
Neural Network	0.1697	10	30
XGBoost	0.1832	10	27

3) Discussion: In this dataset, the high performance of complex models like SVM and Neural Networks does not significantly compromise interpretability when using explainability techniques. The Logistic Regression model remains a strong candidate for applications requiring transparency, offering both high accuracy and straightforward explanations.

D. Overall Analysis

The results across all datasets highlight the inherent trade-offs between model complexity, predictive performance, and interpretability. Complex models tend to offer better predictive performance but require advanced interpretability methods to make their decision processes transparent. Simpler models like Logistic Regression and Decision Trees provide inherent interpretability but may not always achieve the highest accuracy.

The use of LIME and SHAP proved effective in providing post-hoc explanations for complex models, although the fidelity and sparsity metrics indicate varying degrees of approximation quality and explanation simplicity. SHAP generally provided more consistent explanations across models, while LIME's fidelity varied significantly, especially with models like Decision Trees.

These findings emphasize the importance of selecting models based on the specific requirements of the application domain. In high-stakes fields where transparency is critical, a slight compromise in accuracy may be acceptable in favor of interpretability. Conversely, in applications where predictive performance is paramount and explanations are less critical, complex models may be preferred.

VI. CONCLUSION

This study comprehensively evaluated the interplay between model accuracy and interpretability across different machine learning algorithms and datasets. Our findings demonstrate that while complex models such as Neural Networks, SVMs, and ensemble methods like XGBoost achieve superior predictive performance, they inherently lack transparency. Explainability techniques like LIME and SHAP are indispensable tools for interpreting these models, although they introduce additional layers of complexity and potential approximation errors.

Conversely, simpler models like Logistic Regression and Decision Trees offer straightforward interpretability, making them suitable for applications where transparency and ease of explanation are paramount. However, this often comes at the

cost of reduced predictive accuracy, particularly in datasets with complex, non-linear relationships.

Our analysis underscores the necessity for a balanced approach in model selection, taking into account the specific needs of the application domain. In critical sectors such as healthcare and finance, where decisions must be interpretable and justifiable, the use of simpler models or the application of robust explainability techniques is essential. Moreover, the development of standardized evaluation metrics for interpretability, as demonstrated in this study, facilitates better comparisons and informed decision-making.

Future work should focus on integrating interpretability directly into the model training process, potentially through the development of inherently interpretable models that do not compromise on accuracy. Additionally, expanding the scope of explainability metrics and exploring their correlations with human-understandable explanations can further enhance the practical utility of explainable AI.

REFERENCES

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

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, "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, 2016.

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[4] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[5] D. Gunning and D. Aha, "DARPA's explainable artificial intelligence (XAI) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[6] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[7] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, Lulu.com, 2019.

[8] UCI Machine Learning Repository: Pima Indians Diabetes Dataset. Available at: https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes

[9] UCI Machine Learning Repository: Communities and Crime Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

[10] UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)