Feature Selection and Explainable AI

Opal Peltzman (208521385), Gil Ben-David (318686177)

February 5, 2025

Abstract

Explainable Artificial Intelligence (XAI) aims to make AI models interpretable and transparent. Feature selection is crucial in machine learning for improving computational efficiency and model performance. Traditional feature selection methods, such as statistical tests and recursive elimination, often lack interpretability. In this research, we investigate whether XAI tools, particularly SHAP (SHapley Additive exPlanations), can effectively be used for feature selection.

We evaluate SHAP's feature importance across four machine learning models (KNN, SVM, Decision Tree, and Logistic Regression) and compare its feature rankings with XGBoost's built-in importance metrics. Our methodology involves applying SHAP for feature selection, removing non-significant features, and comparing model performance before and after selection. We conduct experiments on four datasets from diverse domains, assessing accuracy, R-squared, and interpretability improvements.

By analyzing SHAP's consistency across models and its impact on predictive performance, this study aims to determine whether SHAP-based feature selection enhances model explainability while maintaining or improving accuracy.

1 Problem description

1.1 Explainable AI

As Machine Learning (ML) models are increasingly deployed in critical domains, the need for transparency among various AI stakeholders is growing [7]. The primary concern with black-box models is that they can lead to decisions that are not justifiable, legitimate, or interpretable, making it difficult to understand their underlying reasoning [4]. Providing clear explanations for model outputs is essential, particularly in fields such as precision medicine, where experts require more than just binary predictions to support their diagnoses [9]. Similar concerns extend to autonomous vehicles, cybersecurity, and financial decision-making.

To address these challenges, eXplainable AI (XAI) [4] aims to develop machine learning techniques that (1) enhance model interpretability while maintaining predictive accuracy, and (2) enable users to understand, trust, and manage AI-based decision-making systems effectively [3]. In essence, XAI helps developers and data scientists interpret model behavior, fostering both transparency and trust in AI-driven applications.

A key tool in the XAI landscape is the SHAP (SHapley Additive exPlanations) framework [5], which is rooted in Shapley values from game theory. SHAP provides a structured approach for explaining machine learning predictions by attributing importance scores to individual features. This not only enhances model interpretability but also serves as a potential method for feature selection, a concept we investigate in this research.

1.2 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. This reduction can help lower computational costs and, in some cases, enhance model performance by eliminating irrelevant or redundant features.

Statistical-based feature selection methods assess the relationship between each input variable and the target variable using statistical measures. These methods are often fast and effective; however, their effectiveness depends on the choice of statistical measures, which must be carefully selected based on the data type of both input and output variables. This makes filter-based feature selection challenging, as practitioners must determine the most suitable statistical approach for a given dataset.

1.3 Goal and Motivation

Our goal is to evaluate whether SHAP can serve as an effective tool for feature selection. Data scientists have multiple tools for feature selection, and we seek to explore whether SHAP can be added to this toolkit.

A key advantage of SHAP is its model-agnostic nature, allowing it to work with any machine learning algorithm. If SHAP proves to be a suitable feature selection tool, it would provide a dual benefit: not only enhancing explainability for chosen models but also streamlining feature selection within the same framework. This could improve both model interpretability and performance, making SHAP a powerful addition to the feature selection process.

2 Solution Overview

As previously mentioned, our goal is to evaluate whether the SHAP library can serve as an effective tool for feature selection. To do this, we selected four widely used classification datasets and applied four different machine-learning models to each. We used SHAP to generate a ranked list of features based on importance for each model, allowing us to identify which features contribute the most to predictions and which are less significant. By removing low-importance features, we assessed whether models could achieve comparable or improved accuracy while benefiting from a more efficient learning process.

To establish a baseline, we first trained each model using all features and recorded its accuracy. We then repeated the training after removing features identified as less significant by SHAP and evaluated

whether accuracy improved or remained stable.

The four datasets used in our study[1] are:

- **Dataset 1 -** Iris Dataset: A dataset containing measurements of iris flowers classified into three species.
- **Dataset 2 -** Titanic Passengers Dataset: This dataset includes passenger details and their survival status from the Titanic disaster.
- **Dataset 3 -** Cancer Detection Dataset: Features extracted from digitized images of breast mass cell nuclei, used for cancer classification.
- **Dataset 4 -** Stress Detection in Sleep Dataset: Data collected via IoT sensors measuring sleep parameters to detect stress levels.

Before training, we performed data preprocessing on each dataset. This included handling missing values, removing irrelevant features such as names or IDs, and ensuring that the class distribution remained balanced between training and test sets.

For each dataset, we applied four machine-learning models using the SHAP library[1]: K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Decision Tree, and Logistic Regression. SHAP ranked the features based on their importance in model predictions.

Additionally, we trained an XGBoost model for each dataset. XGBoost, being a tree-based algorithm, inherently provides feature importance as part of its learning process[2].

After obtaining SHAP's ranked feature lists, we compared the rankings across different models to identify consistently low-importance features. We then retrained the models after removing these features and measured the impact on accuracy to determine whether feature selection based on SHAP led to performance improvements or maintained predictive performance.

3 Experimental evaluation

3.1 Iris Dataset

Figure 1 presents the SHAP feature importance plots for all four models.

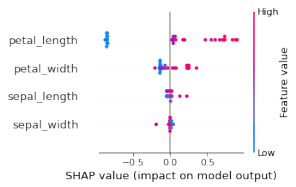
Table 1 presents the accuracy of each model before feature selection.

Model	KNN	SVM	Decision Tree	Logistic Regression
Accuracy	0.97	0.97	0.90	0.93

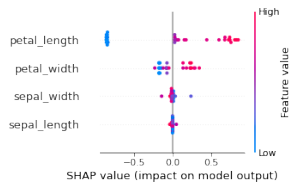
Table 1: Model accuracy before feature selection

[1]Since we wanted to focus on a specific area for our research, we chose to conduct experiments exclusively on classification datasets with dedicated machine learning models.

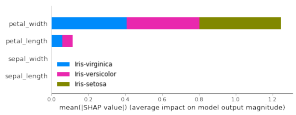
[2]Initially, we planned to directly compare SHAP's feature importance rankings with those from XGBoost. However, we later determined that a more meaningful comparison would be based on established machine-learning evaluation metrics, such as accuracy.



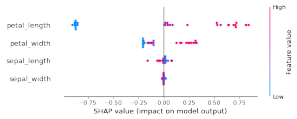
(a) K-Nearest Neighbors (KNN)



(b) Support Vector Machine (SVM)



(c) Decision Tree



(d) Logistic Regression

Figure 1: SHAP feature selection results for the Iris dataset

Analyzing the SHAP feature importance rankings, we observe that two models ranked "sepal length" as the least important feature, while the other two models ranked "sepal width" lowest. To determine which feature to remove, we cross-referenced SHAP rankings with the XGBoost model, which also identified "sepal length" as the least important feature. Based on this alignment, we chose to remove "sepal length" and retrain the models.

Table 2 presents the accuracy of each model after feature selection.

Model	KNN	SVM	Decision Tree	Logistic Regression
Accuracy	0.97	0.97	0.90	0.93

Table 2: Model accuracy after feature selection

The results demonstrate that removing "sepal length" had no impact on model accuracy. This suggests that SHAP-based feature selection successfully identified and removed a redundant feature while preserving predictive performance, leading to a more efficient model.

3.2 Titanic Dataset

Figure 2 presents the SHAP feature importance plots for all four models.

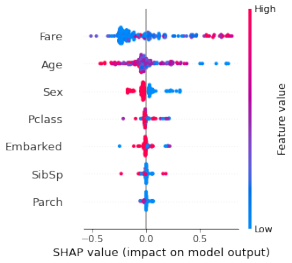
Table 3 presents the accuracy of each model before feature selection.

Model	KNN	SVM	Decision Tree	Logistic Regression
Accuracy	0.74	0.68	0.82	0.80

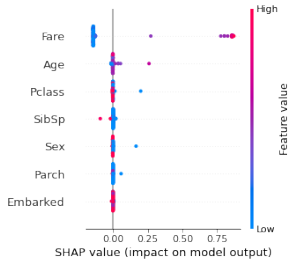
Table 3: Model accuracy before feature selection

Analyzing the SHAP feature importance rankings, we observe that the feature "parch" (number of parents/children aboard) consistently ranked as one of the least important features across all models. To test whether removing it would impact model performance, we retrained the models without "parch" and re-evaluated their accuracy.

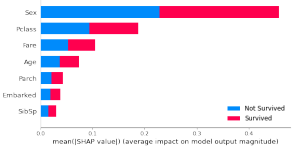
Table 4 presents the accuracy of each model after feature selection.



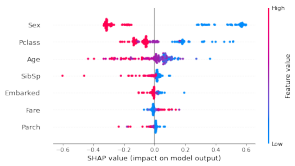
(a) K-Nearest Neighbors (KNN)



(b) Support Vector Machine (SVM)



(c) Decision Tree



(d) Logistic Regression

Figure 2: SHAP feature selection results for the Titanic dataset

Model	KNN	SVM	Decision Tree	Logistic Regression
Accuracy	0.75	0.68	0.81	0.81

Table 4: Model accuracy after feature selection

The results indicate that removing "parch" had minimal impact on model performance. KNN and logistic regression models saw slight improvements, while SVM remained unchanged. The decision tree model experienced a minor drop in accuracy from 0.82 to 0.81, suggesting that "parch" may have had a marginal role in its decision-making process. Overall, SHAP-based feature selection effectively reduced input complexity without significantly affecting model accuracy.

3.3 Stress Detection In Sleep Dataset

Figure 3 presents the SHAP feature importance plots for all four models.

Table 5 presents the accuracy of each model before feature selection.

Model	KNN	SVM	Decision Tree	Logistic Regression
Accuracy	1.00	1.00	0.98	1.00

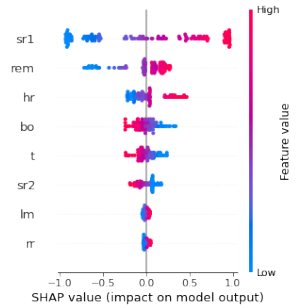
Table 5: Model accuracy before feature selection

Analyzing the SHAP feature importance rankings, we observe that all models identified the feature "rr" (respiratory rate) as the least important. Based on this finding, we removed "rr" and retrained the models to assess whether this change impacted performance.

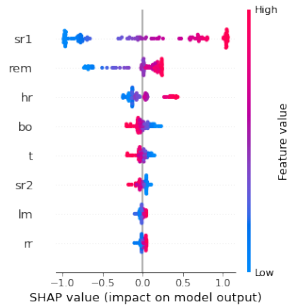
Table 6 presents the accuracy of each model after feature selection.

Model	KNN	SVM	Decision Tree	Logistic Regression
Accuracy	1.00	1.00	0.98	1.00

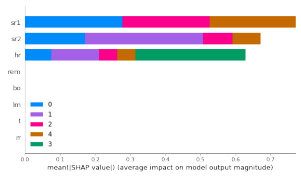
Table 6: Model accuracy after feature selection



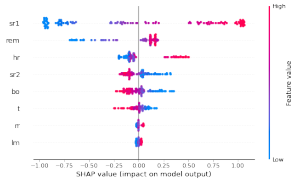
(a) K-Nearest Neighbors (KNN)



(b) Support Vector Machine (SVM)



(c) Decision Tree



(d) Logistic Regression

Figure 3: SHAP feature selection results for the Stress Detection in Sleep dataset

The results indicate that removing "rr" had no effect on model accuracy. This suggests that "rr" was not contributing meaningful predictive value to the models. Consequently, SHAP-based feature selection successfully reduced input complexity without compromising predictive performance, demonstrating its effectiveness in identifying and eliminating redundant features.

3.4 Cancer Detection Dataset

Figure 4 presents the SHAP feature importance plots for all four models.

Table 7 presents the accuracy of each model before feature selection.

Model	KNN	SVM	Decision Tree	Logistic Regression
Accuracy	0.94	0.94	0.93	0.94

Table 7: Model accuracy before feature selection

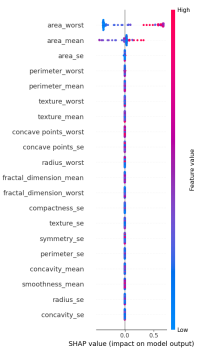
Analyzing the SHAP feature importance rankings, we observed that several features did not appear in the top 20 most important features across models. Since these features contributed minimally to predictions, we removed the following: compactness_worst, concave points_se, concavity_se, fractal_dimension_se, radius_mean, smoothness_worst, and symmetry_worst.

Table 8 presents the accuracy of each model after feature selection.

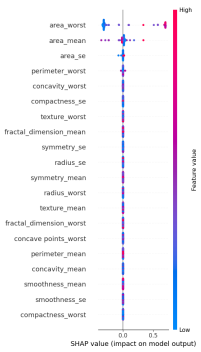
Model	KNN	SVM	Decision Tree	Logistic Regression
Accuracy	0.94	0.94	0.95	0.94

Table 8: Model accuracy after feature selection

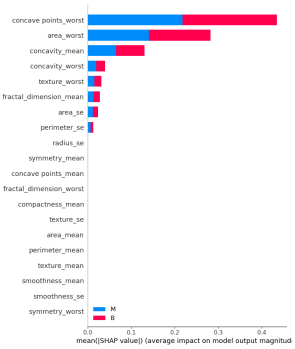
The results indicate that removing these features had no negative impact on model accuracy. In fact, the accuracy of the Decision Tree model slightly improved from 0.93 to 0.95. This suggests that



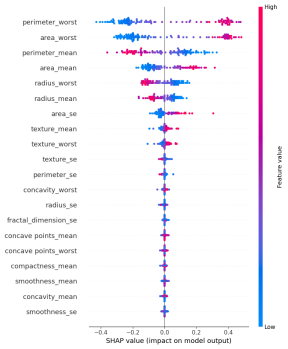
(a) K-Nearest Neighbors (KNN)



(b) Support Vector Machine (SVM)



(c) Decision Tree



(d) Logistic Regression

Figure 4: SHAP feature selection results for the Cancer Detection dataset

SHAP-based feature selection successfully identified redundant features, leading to a more efficient model without sacrificing predictive performance.

4 Related Work

The SHAP library is widely used as a tool for model explainability [8]. However, SHAP has also been applied in other domains beyond explainability, as demonstrated in [2, 1]. Additionally, SHAP has been integrated with XGBoost for interpretability purposes, such as in the interpretation framework proposed in [6].

Despite SHAP's growing adoption, we did not find any prior research specifically exploring its use as a feature selection tool across multiple machine-learning models. Our work was inspired by existing applications of SHAP and XGBoost, which led us to investigate whether SHAP could be leveraged not only for model interpretability but also as a feature selection mechanism.

5 Conclusions and Future Work

In this project, we aimed to integrate Feature Selection and XAI, specifically investigating whether SHAP, as an XAI tool, can be used for feature selection. Our experimental results indicate that

selecting features based on SHAP's feature importance rankings generally maintained or improved model accuracy, with only a single exception.

These findings suggest that SHAP demonstrates strong potential as a feature selection tool. However, to validate this claim more comprehensively, further research is required across a broader range of machine-learning models and dataset types.

As a next step, we plan to extend our experiments to regression datasets, applying SHAP-based feature selection to appropriate regression models. This will allow us to assess whether the observed benefits hold across different types of predictive tasks and further solidify SHAP's role in feature selection.

References

[1] Rafa Alenezi and Simone A. Ludwig. Explainability of cybersecurity threats data using shap. pages 01–10, 2021.

[2] Liat Antwarg, Bracha Shapira, and Lior Rokach. Explaining anomalies detected by autoencoders using SHAP. *CoRR*, abs/1903.02407, 2019.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[4] David Gunning. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, 2(2):1, 2017.

[5] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[6] Yuan Meng, Nianhua Yang, Zhilin Qian, and Gaoyu Zhang. What makes an online review more helpful: An interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(3):466–490, 2021.

[7] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. Stakeholders in explainable ai. *arXiv preprint arXiv:1810.00184*, 2018.

[8] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014.

[9] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.