Role of Classifier in Feature Importance: An Empirical Evaluation

HUANBO MENG, University of Twente, The Netherland

Feature importance methods, such as SHAP (SHapley Additive exPlanations) and permutation feature importance, are widely utilized to assess the influence of features on model outcomes. In theory, these methods should provide consistent importance rankings regardless of the classifier employed. However, the internal mechanisms of classifiers can significantly influence the resulting importance scores in practice. This research aims to investigate the impact of four feature importance measurements on five different classifiers across four datasets. Comparative analyses of feature importance ranking results will be conducted to identify inconsistencies and patterns linked to classifier choice. We can infer from the research that feature importance techniques SHAP and PFI are more closely related to linear classifiers. Feature importance ranking results are mostly influenced by the generalizability of features and the linearity of the classifier. Additionally, one can alter the ranking results to achieve more desirable results by adjusting the hyperparameter.

Keywords

Classifier, feature importance, SHAP, permutation feature importance, KNN, SVC, LIME

1 INTRODUCTION

Artificial Intelligence has become an essential part of daily human life in the form of recognition systems, recommender systems, voice recognition systems and predictive analytics. After decades of usage in scientific fields, the demand of AI has increased to a significant level, and the capability of new algorithms has reached a new peak. Despite the progress made by researchers, the demand for achieving high computing is still growing rapidly, and the complexity of the algorithms used is constantly expanding. This has led to an urgent desire to understand the conditions under which artificial intelligence systems, from research and development personnel to final products, make decisions. This push for the process and reasons of deduction has led to the emergence of a new research field: explainable artificial intelligence (XAI).

TScIT 42, January 30, 2025, Enschede, The Netherlands

Explainable Artificial Intelligence (XAI) has become a rapidly growing research area in the field of artificial intelligence. Unlike traditional AI systems, which often function as "black boxes," XAI provides insights into how models reach their decisions, the explainability is impossible in traditional AI [1, 5]. One of the core components of XAI is the concept of feature importance, which quantifies the contribution of individual features to a model's predictions [3, 16]. The importance of features plays a crucial role in parsing data and evaluating model behavior. SHapley Additive Explanation (SHAP) and permutation feature importance as BFI are the two most popular techniques for providing feature importance ranking. In an ideal state, these methods should be consistent and not affected by the type of classifier used. In practice, this is not the case. The working of these methods is influenced by the classifiers. The result should be fairness, objectivity, and reliable interpretability, no matter which classifiers the user chooses [7, 11, 13, 18].

However, in practice, classifiers have formed a unique way of influencing feature importance ranking due to their unique structural framework, poor optimization habits, and handling of feature interactions. This raises a key question: is the feature importance really an attribute of the data, or is it closely related to the selection of classifiers? This questions the reliability of widely used interpretability tools and forces us to question whether users can truly trust the rankings they produce in different classifiers.

Although the influence of classifier on feature importance score is well known, extent of this influence is not quantified. In this research we attempt to quantify it by designing eperiments with multiple datasets and classifiers. The main research question is as follows:

RQ: How to quantify the influence of classifier or ranking results of feature importance algorithm thus revealing influence of the internal working of classifier to feature importance ranking?

2 BACKGROUND

Feature importance methods like SHAP model have been a hot research topic since 2021. There have been many papers using the SHAP model for machine learning and data analysis [6, 8, 9, 19]. There are also a number of articles that propose ways to improve the SHAP model. This article proposes a new metric for calculating the importance of global features. The new metric adopts the coefficient of determination in addition to the traditional metric, in order to be used to improve the original SHAP

model [6]. In the latest article [6] on 2024, there has been research of a comparison in model performance using the most important features selected by SHAP (SHapley Additive exPlanations) values and the model's built-in feature importance list. "According to our findings, the return on investment for implementing SHAP may be relatively low, particularly when built-in feature selection methods are available, especially for large datasets. Additionally, the considerable computational expenses associated with SHAP may render it impractical for handling Big Data" [6]. This article does an excellent job of pointing out some of the limitations of SHAP as well as its instability.

For the existing solution, the research topic is rather rare in the field. There are research into the relation between feature importance and classifier build-in feature selection mechanism. But this research paper will focus more on the extent of differences in the feature importance with regards to the selection of classifiers.

3 METHODOLOGY

This section will focus on the methodology used to answer the research questions. This research is based on existing models and datasets, through pre-processing to make the datasets more applicable to the data analysis and comparisons. This study uses four datasets, five classifier models and four feature importance methods for comparative analysis.

- **Dataset**: Wine Quality Dataset [12], Mushroom Dataset [14], Diabetes Dataset [2], Breast Cancer Dataset [15].
- **Classifier**: Support Vector Classification as SVC, Logistic Regression, XGBoost, Decision Tree and K-nearest Neighbor Classification as KNN.
- **Feature importance method**: SHAP, Local Interpretable Model-Agnostic Explanations as LIME, built-in algorithms, Permutation Feature Importance as PFI.

All four datasets are binary and numerical to keep the consistency in this research. In the choice of classifiers, including both linear and non-linear classifiers to cover the diversity is the priority. To ensure generalizability as well as applicability, five out of ten well known classifiers [17] are selected.

In terms of feature importance methods, three most accepted methods in the current research environment [10] are chosen in this research topic. Finally, a comparison is made between SHAP and iterSHAP to confirm the uniformity and accuracy of SHAP values [4].

3.1 Data pre-processing

The overview tables (table 1-4) illustrate the attribute of four datasets such as data size, balance of dataset, cross validation, tree depth and train-test ratio.

In the Wine Quality Dataset, the data is pre-processed by dividing all taste levels higher than five as good quality(positive) and otherwise bad quality(negative).

Regarding the Mushroom Dataset, the balance of the dataset is 0.55 to 0.45 which is consumed as rather balanced. The train set ratio is 0.6 which differs from other datasets.

In the Diabetes Dataset, due to the size of the dataset being massive and would heavily damage the computation time, it is trimmed down to 1000 and the data was randomly selected by the ratio of 50/50 based on whether the individual had diabetes.

With respect to the Diabetes Dataset, the size is rather small compared to the other three datasets, with a ratio of 0.62 positive samples.

Table 1. Wine Quality Dataset

Size	1143 * 12
The Balance of Dataset	0.54/0.45
Cross validation	0.74
Tree depth	15
train & test	0.8/0.2

Table 2. Mushroom Dataset

Size	8124 * 23
The Balance of Dataset	0.55/0.45
Cross validation	0.79
Tree depth	6
train & test	0.6/0.4

Table 3. Diabetes Dataset

Size	1000 * 10
The Balance of Dataset	0.5/0.5
Cross validation	0.88
Tree depth	30
train & test	0.8/0.2

Table 4. Breast Cancer Dataset

Size	569 * 32
The Balance of Dataset	0.62/0.37
Cross validation	0.95
Tree depth	7
train & test	0.8/0.2

3.2 Hyper-parameter optimization

In order to keep the experiment uniform and persuasive, all five classifiers were configured with default settings as baseline configuration for this study. With regards to the SVC, (*lm: C = 1.0 lc: SVC linear*) was selected. The intention was to control the variables and avoid unnecessary bias in analysing the data.

3.3 Approach

In this study, the approach compares the result for the same feature importance measurement applied to different classifiers as shown in Figure 1. For each dataset, we will apply five different classifiers, and based on each classifier, we will acquire the results of feature importance ranking from several feature importance measurements.

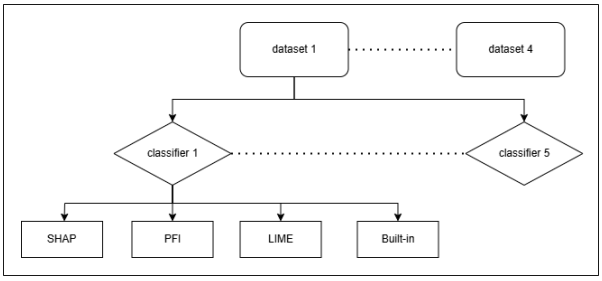

Fig. 1. Architecture of data collection

4. OBSERVATION

4.1 Visualization on wine quality dataset

This section focuses on the Wine Quality dataset, to provide a general understanding of the correlation between variables with quality, data distribution and the gradient performance of each feature in different feature importance measurements. First, by investigating the heatmap in Figure 2, a fundamental understanding of the correlation between wine attributes and the quality of wine can be established. Among all the features, the alcohol has the highest weight 0.48 that shows a strong positive correlation with the quality of the red wine. Also the citric acid and sulphates have a rather high correlation with quality. Besides the correlation with quality, the sulphates exhibit a strong relation with citric acid and chlorides. Meanwhile, alcohol has a rather weak positive correlation with the pH value. As for pH value, it shows a huge negative relation with multiple features such as fixed acidity, citric acid, and density.

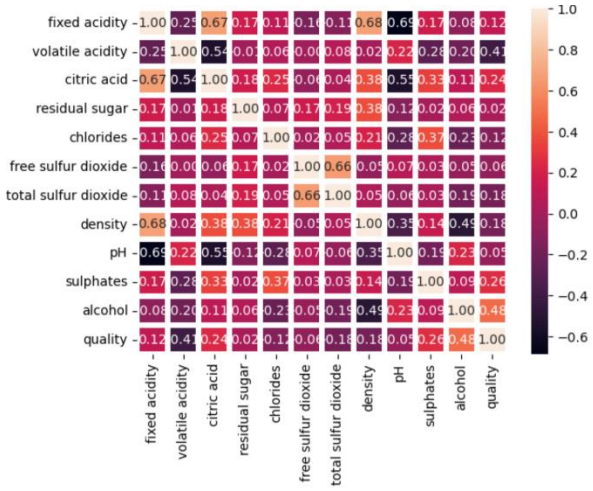

Fig. 2. Feature correlation heatmap

By taking a closer analysis of the data distribution on the features mentioned above, the feature alcohol and citric acid are significantly skewed to the right, which shows a high amount of allocation on low alcohol level and citric acid level in the wine quality dataset. For the density, it shows a strong normal distribution which also indicates the density of wine in the dataset is rather evenly distributed. With regard to volatile acidity, the majority of the values allocated around 0.5, indicate most samples exhibit a middle level of volatile acidity. The distribution appears right skew with a few instances higher than 1.00.

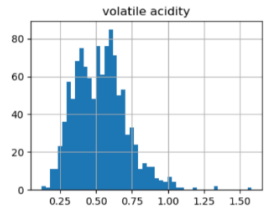

Fig. 3. Volatile acidity

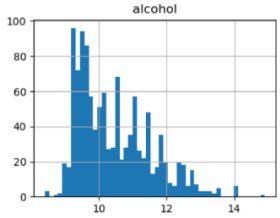

Fig. 4. Alcohol

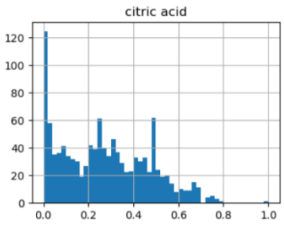

Fig. 5. Citric acid

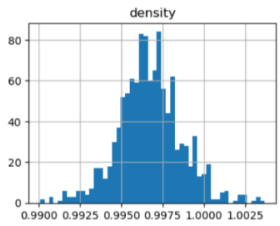

Fig. 6. density

4.1.1 *Feature importance ranking with four measurements*

The following 4 figures are the heatmaps for all four feature importance measurements applied on different classifiers.

In Figure 7, with regard to the results of SVC, LR, and XGBoost, alcohol has the highest feature importance. Meanwhile, volatile acidity and fixed acidity are secondly aligned in all three classifiers. This shows both alcohol and volatile acidity are strong predictors in this dataset. Sulphates is another important element in only XGBoost. Worth noticing, that the feature importances of KNN and decision tree under SHAP measurement are evenly distributed on a certain level.

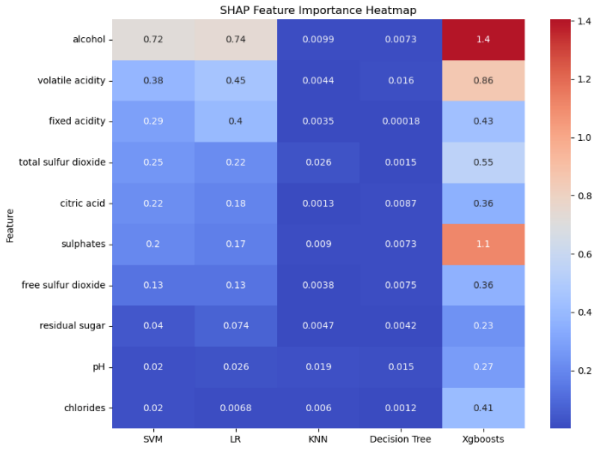

Fig. 7. SHAP on classifiers

Figure 8 shows five classifiers with PFI measurement applied. Observation shows that the alcohol feature continues to dominate all five classifiers. Which again, strongly proved the connection with wine quality. Despite the domination of alcohol, the volatile acidity is no longer outstanding compared to itself in SHAP. This could be the reason for the different inner calculations of SHAP and PFI. But the distribution of feature importance on KNN and decision tree still shows evenly except alcohol.

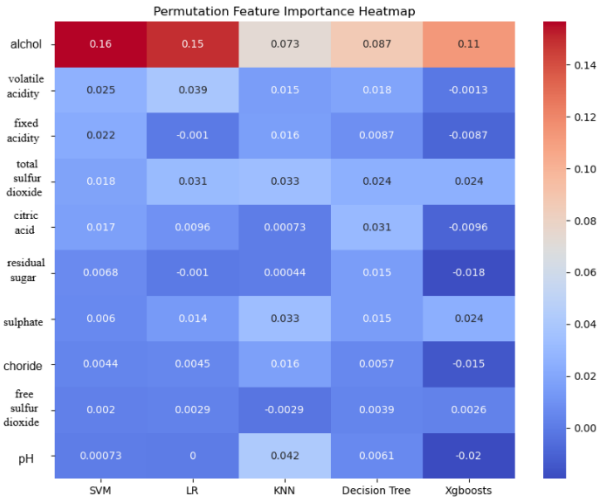

Fig. 8. Permutation Feature importance on classifiers

In figure 9, it shows the feature importance ranking with LIME. The ranking in LIME is rather chaotic compared to the other measurements. However, the general direction and concept of ranking is similar to SHAP and Permutation feature importance. Features like volatile acidity and alcohol show high weight in all five classifiers. Even though the chaos exists, the rankings in KNN and the decision tree remain flat. Due to the non-linear kernel and unique distance calculation, it is normal and expected to have such a result.

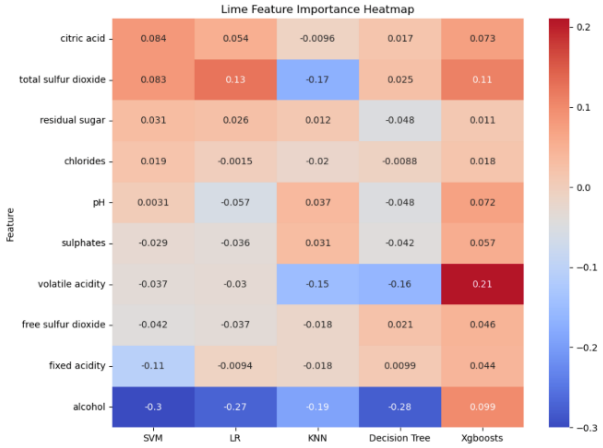

Fig. 9. LIME on classifiers

In Figure 10, it illustrates the ranking of build-in measurement on each classifier. The KNN does not include such a procedure. Hence, the comparison between SVC, LR, Decision Tree and XGBoost will be illustrated in this section. Through the observation, alcohol no longer dominates the ranking but volatile acidity. In both two linear classifiers(SVM and LR), it appears to have a huge gap compared to the second feature. Meanwhile, features in XGBoost are evenly distributed, showing a similar behavior as the decision tree.

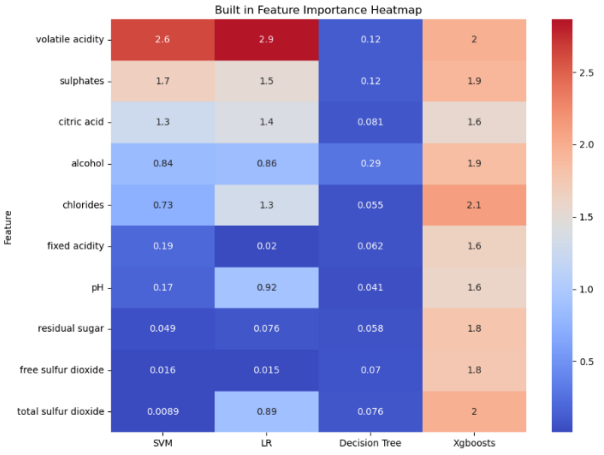

Fig. 10. Built-in measurement on classifiers

4.2 Feature importance method behavior across datasets

Due to the large size of the data collection, and the number of features in each data set varies, this section will illustrate some representative data of each dataset for comparison. (Some data will be trimmed/rounded and enlarged by 100 for better reading).

4.2.1 Wine Dataset with SHAP

In this section, it tends to present SHAP value of different classifiers from Wine Quality Dataset. Through observation from Table 5, both SVC and Logistic Regression emphasize a smaller set of dominant features, particularly alcohol and volatile acidity, which consistently rank as the most important predictors. It shows a strong linear relation between alcohol and volatile acidity. However, the Decision Tree model shows a slightly broader distribution of feature importance compared to linear models.

In general, features like alcohol and volatile acidity dominate in multiple classifiers. This shows that in SHAP ranking measurement, XGBoost and linear models (SVC, LR) heavily rely on a few dominant features, while KNN and Decision Tree distribute importance more evenly across features.

Table 5. Wine Dataset with SHAP

SVC	LR	KNN	Decision tree	XGBoost
alcohol 0.72	alcohol 0.74	total sulfur dioxide 0.255	volatile acidity 0.156	alcohol 1.40
volatile acidity 0.38	volatile acidity 0.45	pH 0.193	pH 0.151	sulphates 1.10
fixed acidity 0.29	total sulfur dioxide 0.39	alcohol 0.098	citric acid 0.08	volatile acidity 0.85
total sulfur dioxide 0.25	sulphates 0.21	sulphates 0.090	free sulfur dioxide 0.075	total sulfur dioxide 0.54
citric acid 0.22	pH 0.18	density 0.059	sulphates 0.073	fixed acidity 0.43
sulphates 0.20	citric acid 0.16	residual sugar 0.047	alcohol 0.072	chlorides 0.41
free sulfur dioxide 0.13	free sulfur dioxide 0.12	volatile acidity 0.043	density 0.042	density 0.39
residual sugar 0.04	residual sugar 0.07	free sulfur dioxide 0.037	residual sugar 0.016	free sulfur dioxide 0.35

4.2.2 Breast Cancer Dataset with LIME

In this section, feature importance method LIME was used on the Breast cancer dataset(Table 6). In this ranking result, Linear classifiers such as SVC and LR highlight a smaller subset of dominant features. Both of them have features like radius_mean and texture_se as key predictors. KNN and Decision Tree demonstrate a more dispersed distribution of feature importance. In detail, the KNN ranking results have a rather strange way of ranking all features, this could be the reason for the algorithm of KNN which calculates the distance of each feature.

In general, the ranking is comparatively similar compared to other datasets. Each feature component shows weak predictive value independent in both KNN and decision tree. Compared to other classifiers ranking results that applied with LIM in the same dataset, KNN is the most chaotic.

Table 6. Breast Cancer Dataset with LIME

SVC	LR	KNN
radius_mean 0.145	radius_worst 0.10	radius_wors 0.176
texture_se 0.0573	radius_mean 0.099	concave points_mean 0.156
symmetry_worst 0.0343	compactness_worst 0.024	perimeter_worst 0.154
concave points_worst 0.021	fractal_dimension 0.015	area_worst 0.146
smoothness_mean - 0.0242	fractal_dimension_mean 0.014787	radius_mean 0.128
perimeter_se -0.031	smoothness_worst - 0.0173	concave points_worst 0.124
texture_worst -0.050	area_se - 0.046	area_se 0.122
area_se - 0.071	texture_wors -0.053	perimeter_mean 0.118
area_worst - 0.099	perimeter_worst -0.096	radius_se 0.117

Decision tree	XGBoost
concave points_mean 0.08	texture_worst 0.210
concave points_se 0.0822	concavity_worst 0.136
perimeter_worst -0.078	area_worst 0.128
radius_worst -0.061	concave points_mean 0.119

concavity_se -0.048	texture_mean 0.111
area_mean 0.0329	concave points_worst 0.092
area_se -0.042	radius_worst 0.064
symmetry_se 0.041	smoothness_mean 0.09 0.060
concave points_worst - 0.0387	concavity_mean 0.045

4.2.3 Diabetes Dataset with PFI

In Table 7, by applying the Permutation feature importance method as measurement on Diabetes Dataset. As we can observe, the results are rather uniform and aligned compared to other feature importance methods. In particular, HbA1c_level and blood_glucose_level consistently rank as the most important features across all classifiers, indicating their strong predictive value for the target variable.

Features like age and BMI have secondary importance across most classifiers, though they are less highlighted by models like SVM and LR compared to Boost. The decision tree is the outlier compared to the other classifiers, the gender and age components have a rather high ranking which could be a reason for the non-linear module scale. But compared to the other dataset, the high degree of consistency in the top-ranked features is the highlight of this observation.

Table 7. Diabetes Dataset with PFI

SVC	LR	KNN	Decision tree	XGBoost
HbA1c_level 0.044717	HbA1c_level 0.044283	HbA1c_level 0.045600	HbA1c_level 0.05860	HbA1c_level 0.17505
blood_glucose_level 0.025883	blood_glucose_level 0.026633	blood_glucose_level 0.034167	blood_glucose_level 0.04270	blood_glucose_level 0.13055
age 0.004883	age 0.004750	age 0.007100	gender 0.00375	age 0.04545
bmi 0.001933	bmi 0.003017	bmi 0.004283	age 0.00265	bmi 0.01815
heart_disease 0.000517	heart_disease 0.000383	gender 0.003233	heart_disease 0.00155	hypertension 0.00760
hypertension 0.000367	hypertension - 0.000017	hypertension 0.002000	bmi 0.00120	heart_disease 0.00655
gender - 0.000400	gender - 0.000150	heart_disease 0.000667	hypertension 0.00120	gender 0.00255

4.2.4 Mushroom Dataset with build-in

In Table 8, data is collected by using the built-in feature importance measurement to rank all the features. In this table, Odor-related features (odor_n, odor_c) consistently rank among the most important for all classifiers, reflecting their strong forecasting relationship with the target variable.

Spore-print-color_r is highly important in SVC and LR, though its importance diminishes in the Decision Tree model. The pattern is similar to the other dataset results which enhances the correlation between linear classifier results.

Table 8. Mushroom Dataset with Build-in

SVC	LR	Decision tree
spore-print-color_r 1.730824	odor_n 3.925559	odor_n 0.616699
odor_c 1.462210	spore-print-color_r 3.848597	stalk-root_c 0.173272
gill-size 1.055080	gill-size 3.172746	stalk-root_r 0.084494
odor_n - 1.017965	odor_c 3.161368	spore-print-color_r 0.028983
odor_p 0.975459	odor_f 2.876567	gill-spacing 0.026232
gill-spacing - 0.950331		bruises 0.024130
stalk-surface-above-ring_k 0.902406		stalk-surface-below-ring_s 0.021097

4.3 Classifier behavior

The section is to analyze the classifier behavior on the global wide. Aiming to generally categorize the pattern of the same feature importance measurement applied on different classifiers.

Linear Classifiers (SVC, LR):

The results of SVC and LR end up in a strong harmony. The feature importance rankings are highly identical to the top three features. Even though there are few disorder features that have lower importance, the result is still less chaotic than other classifiers. In overall observation, both linear classifiers are highly aligned on their decision boundaries and feature importance patterns. This results in similar trends in feature importance selection and ranking.

KNN:

KNN demonstrates a discrete ranking of feature importance in four datasets. Particularly in the wine dataset, features like total sulfur dioxide and pH rank highly, meanwhile in the cancer dataset, smoothness_worst and texture_worst appear more prominently. In general, the KNN targets distance, it focuses on the neighbor instead of individual feature importance. In most cases, this can lead to feature importance measurements acting differently compared to the other classifiers.

Decision Tree:

In most of the datasets, Decision Tree tends to provide an even distribution of feature importance. Features like volatile acidity and pH in the wine dataset and concave points_worst and area_se in the cancer dataset contribute similarly.

XGBoost:

XGBoost prioritizes fewer features with high SHAP values. This is due to XGBoost focusing more on key predictors. For example, in the cancer dataset, concave points_mean and radius_wors have high SHAP values. Meanwhile its focus on alcohol and sulphates in the wine dataset.

4.4 Dataset-Specific behavior

The wine dataset primarily focuses on alcohol and volatile acidity, which have a rather clear linear relationship with quality scores.

The cancer dataset focuses on features like area_worst and concave points_mean, which are nonlinear and interact more complexly, leading to higher SHAP values for ensemble models like XGBoost.

The diabetes dataset has the most united result compared to the rest. This could be the reason that most of the features in the Diabetes dataset have a strong connection and both linear and non-linear classifiers are taking accountability by these feature relations.

4.5 The usage of iterSHAP

The iterative approach leads to a more refined and accurate model as more features are added in an optimal order based on their SHAP importance scores. In this research, IterSHAP is only used to double check the SHAP value accuracy and compare the strength of relation between each feature.

4.6 Tuning hyperparameter (Wine Dataset)

By analyzing Table 9, through tuning the hyperparameter we can manipulate the feature importance ranking result. To achieve that, first by setting C=1, it appears to have alcohol at the highest, which follows the same pattern as other classifiers' results as mentioned earlier. But when tuning the C=100, volatile acidity and sulphates are the most influential features, with significantly higher importance scores compared to other features.

Features like chlorides and citric acid gain more feature importance ratio, which is due to the model is highly sensitive to even subtle patterns associated with these features.

The importance of alcohol decreases to 0.86 at C=100 showing that the model prioritizes features differently without considering a linear relationship.

Through adjusting the hyperparameter, it is achievable to aligns on a certain degree with other classifier results. Such a manipulation can be used for tuning feature importance results in a more appropriate position with a certain C.

Table 9. Hyperparameter with c

SVM c=1	SVM c = 100
alcohol 0.72	volatile acidity 3.309644
volatile acidity 0.38	sulphates 2.987445
fixed acidity 0.29	chlorides 2.066883
total sulfur dioxide 0.25	citric acid 1.922843
citric acid 0.22	alcohol 0.863719
sulphates 0.20	pH 0.278911
free sulfur dioxide 0.13	fixed acidity 0.205237
residual sugar 0.04	residual sugar 0.041761
pH 0.02	density 0.041442
chlorides 0.02	free sulfur dioxide 0.013303

5 CONCLUSION

In this study we dive into the relationship between classifiers and feature importance scores. We attempt to discover the irregular behavior with different feature importance measurements applied on classifiers. It appears that with different classifiers, the ranking of same feature importance methods will result differently.

There are two components affecting the feature importance results. One of them is generalizability of features. In the observation section, some features are universally important across classifiers in all 4 datasets indicating strong independent predictive value. However, secondary features vary more in importance, suggesting these are influenced by classifier-specific favor. Even though it is not always like this, In the diabetes dataset we observed that most of the feature importances are highly aligned. This is due to the feature correlation strength in this particular dataset. This suggests that for a feature importance measurement, particularly as Permutation feature importance in our case, it can be classifier-agnostic with a certain suitable dataset.

The other major factor affecting the ranking result is whether the classifier model is linear or nonlinear. It has a significant impact on the ranking result. Linear models, such as Support Vector Classifier with a linear kernel and Logistic Regression are fundamentally designed to capture dominant, global patterns in data. The linear relationship between input data and the goal variable is the main emphasis of these two models. In particular, features with stronger patterns and stable correlations throughout the dataset will be given higher weight. In contrast, the ranking outcome of nonlinear models such as Decision Trees and XGBoost will rely more on deeper complex relationships. Since linearity is not assumed in these models, such a behavior will gain more ability to capture interactions between features and target for subtle patterns that may not be apparent in linear models. The distinction specifies the importance of choosing classifiers based on the dataset and target.

After demonstrating that the classifier selection has a substantial impact on feature importance ranking, it is crucial to investigate how the internal mechanisms of these classifiers can be altered to affect the ranking outcomes. A hyperparameter that is set too high causes overfitting, in which the model ignores more general patterns in favor of concentrating primarily on dominating and particular properties. Meanwhile, reducing hyperparameter C to a certain fit value balances feature importance, ensuring the model captures generalizable patterns, which is essential for future predictions on unseen data. Feature importance rankings are therefore sensitive to hyperparameters, underlining the need for careful hyperparameter optimization to avoid misinterpreting model behavior.

6 FUTURE WORK

Looking back on the goals of this research, there is plenty of room for further analysis. Paper has shown and analyzed the performance of the same feature importance method on different classifiers, although feature importance should be consistent with the database property but it is not so, they are kept in some connection with the classifier property which can be manipulated in a way. This study laid the foundation for many future studies.

In the future, using this paper as the basis for analyzing the classifiers one by one on the mathematical angle would be ideal, and this paper will cast a very good role as a cornerstone and direction in future research. Exploring the influence of each classifier on feature importance in mathematical perspective will be complex, but based on the analysis in this paper, researchers will have a rough grouping and research focus. Furthermore, researchers can focus on developing systematic approaches to hyper-parameter optimization that can consistently yield optimal feature importance rankings across different classifiers. By doing so, we can better understand the interplay between classifier parameters and feature importance, ultimately leading to more robust and interpretable models. Another point worth mentioning is that each different feature importance method performs differently on the same classifier. Although this is a little deviation from the research direction of this paper, it is still an efficient reuse of the data analyzed and collected in this paper. This research will be able to pave a clearer path for further understanding of feature importance method, and potentially bring a new classifiers-agnostic method to the field of machine learning.

ACKNOWLEDGMENTS

I would like to thank my supervisor, Faizan Ahmed, for his assistance in the research and his help through the whole module and also bringing up such an interesting topic to me. I would also like to thank Faryal Siddique for reviewing my paper.

REFERENCE:

[1] Arrieta, A.B., Rodríguez, N.D., Ser, J.D., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Inf. Fusion, 58, 82-115.
[2] Khare, A. D. (2022, October 6). *Diabetes dataset*. Kaggle. https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset
[3] Bove, C., Aigrain, J., Lesot, M., Tijus, C.A., & Detyniecki, M. (2022). Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. Proceedings of the 27th International Conference on Intelligent User Interfaces.
[4] Mourik, F.G. van (2023) *IterSHAP: an XAI feature selection method for small high-dimensional datasets*.
[5] Vilone, G., & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf. Fusion, 76, 89-106.
[6] Wang, H., Liang, Q., Hancock, J.T. et al. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. J Big Data 11, 44 (2024). https://doi.org/10.1186/s40537-024-00905-w
[7] Olivier, L., & Xuanxiang, H., & Nicholas, A., & Joao, M. (2024). From SHAP Scores to Feature Importance Scores. 10.48550/arXiv.2405.11766.
[8] Verdinelli, Isabella & Wasserman, Larry. (2023). Feature Importance: A Closer Look at Shapley Values and LOCO. 10.48550/arXiv.2303.05981.
[9] Rajyalakshmi, Kona & Gunasekaran, M.. (2024). Comparative study on improved support vector machine and random forest classifier for efficient classification of music genre based on accuracy. 020115. 10.1063/5.0197668.
[10] Lan. *Model Explainability - SHAP vs. LIME vs. Permutation Feature Importance*, *Model explainability - shap vs. Lime vs. permutation feature importance*. Available at: https://www.lanchuhuong.com/data-and-code/2022-07-22_model-explainability-shap-vs-lime-vs-permutation-feature-importance-98484efba066/
[11] Saarela, M., Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. SN Appl. Sci. 3, 272 (2021). https://doi.org/10.1007/s42452-021- 04148-9Kwon, Y., & Zou, J.Y. (2022). WeightedSHAP: analyzing and improving Shapley based feature attributions. ArXiv, abs/2209.13429.
[12] H, M. Y. (2022, January 15). *Wine quality dataset*. Kaggle. https://www.kaggle.com/datasets/yasserh/wine-quality-dataset
[13] Sundararajan, M., & Najmi, A. (2019). The many Shapley values for model explanation. International Conference on Machine Learning.
[14] Learning, U. M. (2016, December 1). *Mushroom classification*. Kaggle. https://www.kaggle.com/datasets/uciml/mushroom-classification
[15] Namdari, R. (2022, August 8). *Breast cancer*. Kaggle. https://www.kaggle.com/datasets/reihanenamdari/breast-cancer
[16] Hwang, S., Chung, H., Lee, T., Kim, J., Kim, Y.C., Kim, J., Kwak, H.W., Choi, I., & Yeo, H.Y. (2023). Feature importance measures from random forest regressor using near-infrared spectra for predicting carbonization characteristics of kraft lignin-derived hydrochar. Journal of Wood Science, 69, 1-12.
[17] Bhattacharyya, S. (2023) 10 popular ML algorithms for solving classification problems, Medium. Available at: *https://medium.com/@howtodoml/10-popular-ml-algorithms-for-solving-classification-problems-b3bc1770fbdc*
[18] Liu, Y. et al. (2022) 'Diagnosis of parkinson's disease based on Shap Value Feature Selection', Biocybernetics and Biomedical Engineering, 42(3), pp. 856–869. doi:10.1016/j.bbe.2022.06.007.
[19] Lee, YG., Oh, JY., Kim, D. et al. SHAP ValueBased Feature Importance Analysis for ShortTerm Load Forecasting. J. Electr. Eng. Technol. 18, 579–588 (2023). https://doi.org/10.1007/s42835-022-01161-9