Feature Selection and Explainable AI

Tomer Porat (060961448), Ziv ben-David (308142488)

March 13, 2023

Abstract

Explainable Artificial Intelligence (XAI) is a field of Artificial Intelligence (AI) that promotes a set of tools, techniques, and algorithms that can generate high-quality interpretable, intuitive, human-understandable explanations of AI decisions. Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. In this project we try to integrate between Feature Selection and XAI, and investigate if we can use XAI tools for choosing main features. A popular tool of XAI is the SHAP library. SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. For our research we will use SHAP in order to know if it can be used for feature selection too.

1 Problem description

1.1 Explainable AI

As black-box Machine Learning (ML) models are increasingly being employed to make important predictions in critical contexts, the demand for transparency is increasing from the various stakeholders in AI [7]. The danger is on creating and using decisions that are not justifiable, legitimate, or that simply do not allow obtaining detailed explanations of their behaviour [4]. Explanations supporting the output of a model are crucial, e.g., in precision medicine, where experts require far more information from the model than a simple binary prediction for supporting their diagnosis [9]. Other examples include autonomous vehicles in transportation, security, and finance, among others.

In order to avoid limiting the effectiveness of the current generation of AI systems, eXplainable AI (XAI) [4] proposes creating a suite of ML techniques that 1) produce more explainable models while maintaining a high level of learning performance (e.g., prediction accuracy), and 2) enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners [3]. In other words, Explainable artificial intelligence is the process of understanding how and why a machine learning model makes its predictions. It can also help machine learning developers and data scientists to better understand and interpret a models' behavior.

The SHAP framework [5] based on shapley values is one of the XAI tools. SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions.

1.2 Feature Selection

Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model. Statistical-based feature selection methods involve evaluating the relationship between each input variable and the target variable using statistics and selecting those input variables that have the strongest relationship with the target variable. These methods can be fast and effective, although the choice of statistical measures depends on the data type of both the input and output variables. As such, it can be challenging for a machine learning practitioner to select an appropriate statistical measure for a dataset when performing filter-based feature selection.

1.3 Goal and Motivation

We are interested in checking whether it is possible to use SHAP as a tool for Feature Selection. Today the data scientist has several tools to perform feature selection and we want to investigate whether SHAP can also be used for this. The advantage of using SHAP is the ability to choose any learning model that the data scientist wants to, thanks to the ability of SHAP to work with any learning model. If SHAP will turn out as a suitable tool for feature selection, the data scientist can earn twice because SHAP will give explainability for the chosen models and feature selection too.

2 Solution overview

As said before, we want to use SHAP library in order to know if it can be used as a tool for feature selection. We have selected four popular datasets. Each dataset was preprocessed, and we executed four machine-learning models on it. For each model, we used SHAP and presented a list of all dataset's features ordered by importance. By the listed features, it is possible to know which of them is more important in the learning process and which is less. In this way, it is possible to evaluate which of the features is unnecessary and learning can be done more efficiently and quickly without it. In addition, we calculated the accuracy for each model while training it with all the dataset's features. We did this to test whether the accuracy improved or at least did not change after the feature selection.

The four popular datasets we used are[1]:

- Dataset 1 - Iris Dataset : Data of different iris plants classified into three species.

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

- Dataset 2 - Titanic Passengers Dataset : This data contains the detailes of the passengers classified to survived or to not survived from the disaster.

- Dataset 3 - Cancer Detection Dataset: Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

- Dataset 4 - Stress Detection in Sleep Dataset: Sleeping parameters captured for sleep study be an IoT sensor.

For each dataset, we ensured that there isn't any missing value, and if so, we completed them. We have removed irrelevant features such as Names or IDs. We divided each dataset into 'train' and 'test', and ensured that the ratio between all categories in 'train' was the same as in 'test'.

After the pre-processing for each dataset, we executed four different machine-learning models on it using the SHAP library[1]: KNN, SVM, Decision Tree and Logistic Regression. For each model, SHAP shows a list of the features sorted by importance. In addition, we also executed another model for each dataset - the XGBOOST model. XGBOOST is a tree-based machine-learning model, and therefore it also presents feature importance as part of its learning[2].

After SHAP plotted the lists of features sorted by importance, we compared the lists of the different models and checked if there are features with low importance values in most of the lists. We removed those features from the list of features and executed the learning models again to check if the accuracy changed for the better or not.

3 Experimental evaluation

3.1 Iris Dataset

In figure 1 we can see the plot of SHAP for all four models.

Table 1 shows the accuracy for each model.

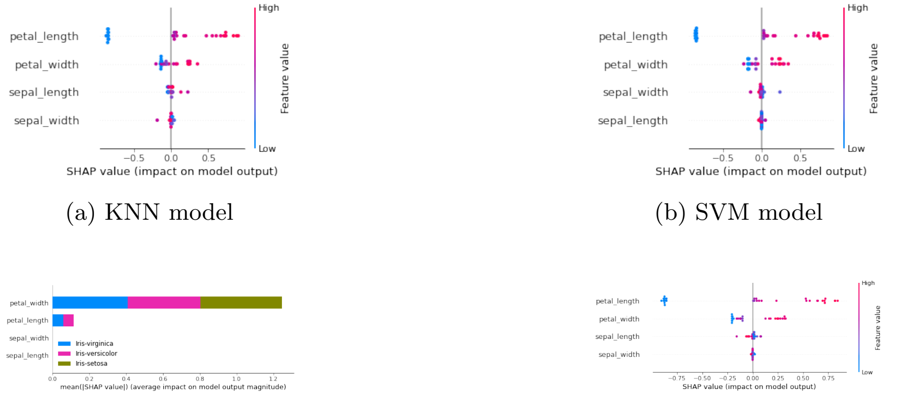
	KNN	SVM	Decision Tree	Logistic Regression
accuracy	0.97	0.97	0.90	0.93

Table 1: Accuracy for models before feature selection

From those graphs we can see that half of the models ranked 'sepal length' as the least important, and half of the models ranked 'sepal width' as the least important. We can try drop one of them and test the dataset again. However, the XGBOOST model ranked 'sepal length' as the least important too. Therefore, we choose to select all features except 'sepal length'.

Table 2 shows the accuracy for each model after the feature selection.

[2] At first thought, we wanted to compare SHAP's and XGBOOST's feature importance. However, we realized that a more accurate comparison would be based on the use of machine-learning known metrics such as accuracy



(a) KNN model

(b) SVM model

(c) Decision Tree model

(d) Logistic Regression model

Figure 1: SHAP feature selection for the iris dataset

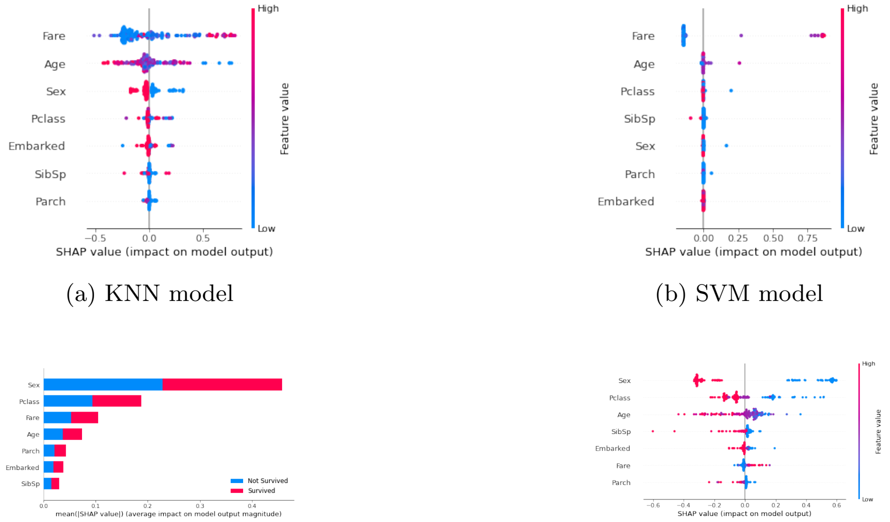
	KNN	SVM	Decision Tree	Logistic Regression
accuracy	0.97	0.97	0.90	0.93

Table 2: Accuracy for models after feature selection

As we can see, the feature selection we used by SHAP succeeded and yielded the same accuracy with less features.

3.2 Titanic Dataset

In figure 2 we can see the plot of SHAP for all four models.



(a) KNN model

(b) SVM model

(c) Decision Tree model

(d) Logistic Regression model

Figure 2: SHAP feature selection for the titanic dataset

Table 3 shows the accuracy for each model.

From the graphs above its seems that the feature 'parch' is less important. Therefore, we choose to select all features without 'parch' and train the models again.

Table 4 shows the accuracy for each model after the feature selection.

	KNN	SVM	Decision Tree	Logistic Regression
accuracy	0.74	0.68	0.82	0.80

Table 3: Accuracy for models before feature selection

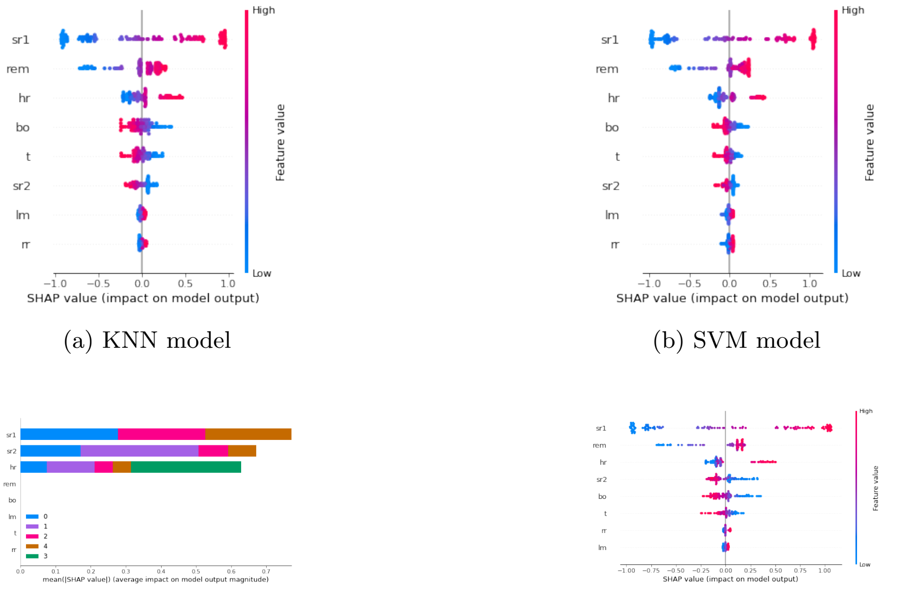
	KNN	SVM	Decision Tree	Logistic Regression
accuracy	0.75	0.68	0.81	0.81

Table 4: Accuracy for models after feature selection

As we can see, the feature selection we used by SHAP succeeded and yielded the same or more accuracy with less features (except the decision tree that decresed from 0.82 into 0.81).

3.3 Stress Detection In Sleep Dataset

In figure 3 we can see the plot of SHAP for all four models.



(a) KNN model

(b) SVM model

(c) Decision Tree model

(d) Logistic Regression model

Figure 3: SHAP feature selection for the stress dataset

Table 5 shows the accuracy for each model.

	KNN	SVM	Decision Tree	Logistic Regression
accuracy	1.00	1.00	0.98	1.00

Table 5: Accuracy for models before feature selection

From the graphs above we can see that the models ranked 'rr' as the least important. Therefore, we choose to select all features except 'rr'.

Table 6 shows the accuracy for each model after the feature selection.

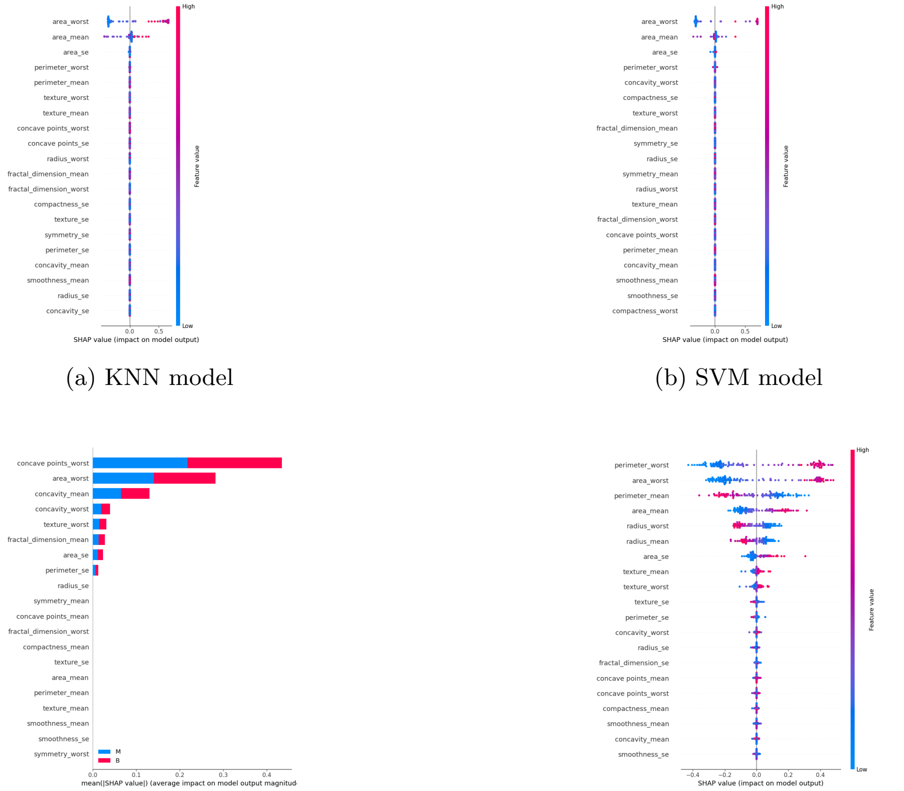
As we can see, the feature selection we used by SHAP succeeded and yielded the same accuracy with less features.

	KNN	SVM	Decision Tree	Logistic Regression
accuracy	1.00	1.00	0.98	1.00

Table 6: Accuracy for models after feature selection

3.4 Cancer Detection Dataset

In figure 4 we can see the plot of SHAP for all four models.



(a) KNN model

(b) SVM model

(c) Decision Tree model

(d) Logistic Regression model

Figure 4: SHAP feature selection for the cancer dataset

Table 7 shows the accuracy for each model.

	KNN	SVM	Decision Tree	Logistic Regression
accuracy	0.94	0.94	0.93	0.94

Table 7: Accuracy for models before feature selection

From the graphs above we can see that there are some features that even not appeared in the top 20 important features. Therefore, we chose to select all features without the following: compactness_worst, concave points_se, concavity_se, fractal_dimension_se, radius_mean, smoothness_worst, symmetry_worst.

Table 8 shows the accuracy for each model after the feature selection.

As we can see, the feature selection we used by SHAP succeeded and yielded the same or more accuracy with less features.

	KNN	SVM	Decision Tree	Logistic Regression
accuracy	0.94	0.94	0.95	0.94

Table 8: Accuracy for models after feature selection

4 Related work

The SHAP library is a popular tool for the explainability of models [8]. However, SHAP used for other things too as we can see in [2], [1]. In addition, there are some uses of SHAP and XGBOOST together, like interpretation framework [6]. Nevertheless, we did not find a paper dealing with our topic. We took inspiration from the uses of SHAP and XGBOOST, and from there we came up with the idea of trying to use them as a tool.

5 Conclusions and Future Work

In this project we tried to integrate between Feature Selection and XAI, and investigate if we can use SHAP (as a XAI tool) for feature selection. The experiments above shows that using feature selection by SHAP's feature importance yields the same or more accuracy value (except for a single exception). Therefore, we can conclude that there is a high probability that SHAP can be a tool for selecting features. However, in order to prove this claim completely, we will have to research SHAP with more models and with more types of datasets. The next step will be to test the same approach on regression datasets and with appropriate machine-learning models.

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.