Explainable AI Exercise with Shapley Value

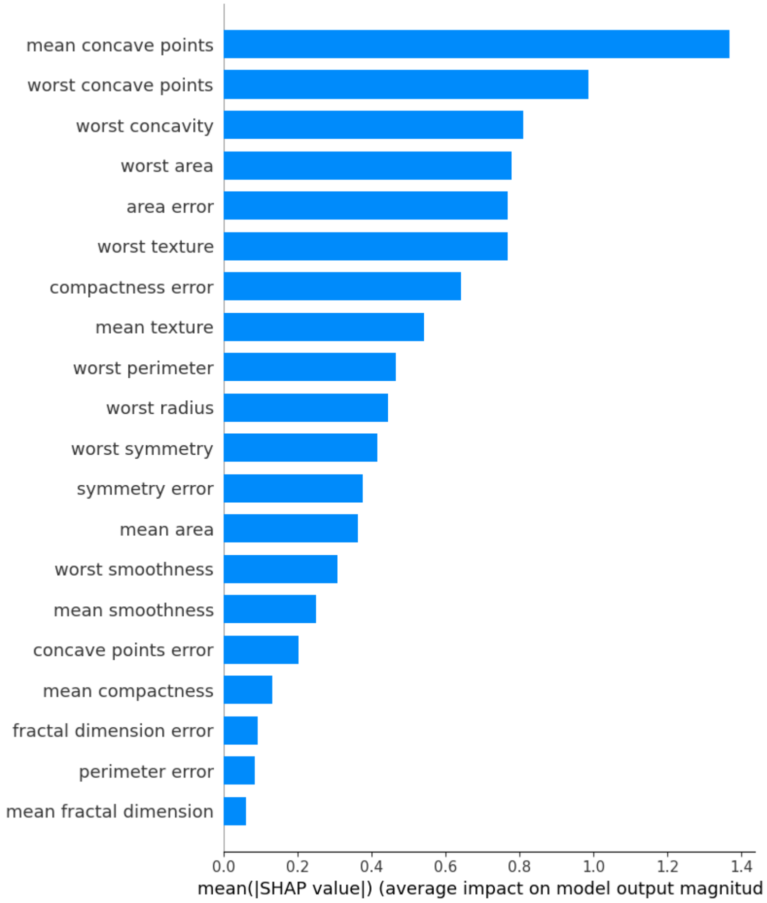
Ali Torabi
Soudabeh Bolouri

In this volunteer exercise, we are going to use the Breast Cancer Dataset. In this dataset features get from digitized images of breast mass. This is a supervised learning case with a target distinguished by values 1 and 0, showing the sign of cancer or not, respectively. The aim is to use a method in Explainable AI, called Shapley Value, to make the model prediction more robust and interpretable to other practitioners like doctors. In this code, it creates and trains an XGBoost model, then creates an Explainer object to compute feature attributions using SHAP. With this approach, it computes feature importance to the prediction without using or knowing the model architecture (model agnostic). It means it can be used to interpret any machine learning model.

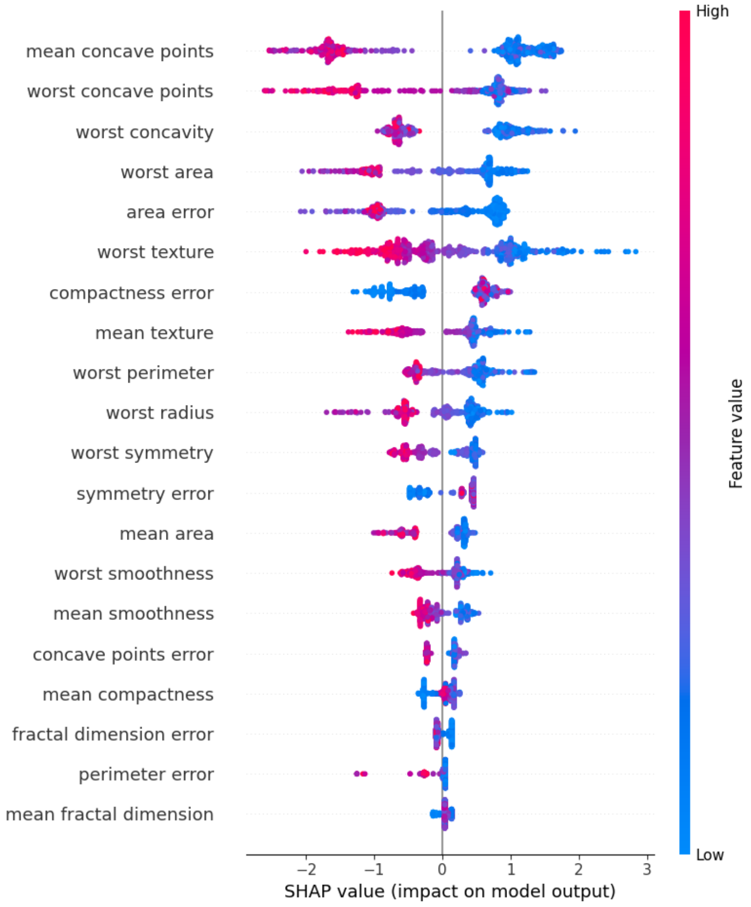
How SHAP value computed?

SHAP values are based on game theory and assign an importance value to each feature in a model. Features with positive SHAP values positively impact the prediction, while those with negative values have a negative impact. The magnitude is a measure of how strong the effect is [2]. Shapley values rely on examining how each feature influences the predicted value of a model by generating many predictions based on a partial set of the features used by the model and comparing the results of the predicted values [1]. There are two ways to compute SHAP values: coalition and paths. In this experience it uses paths. It’s a dynamic programming approach that generate a series of paths, each step in a path adding another feature to generate prediction.
The accuracy by XgBoost learner is 95.61%. But the main idea of doing this exercise is that I'm trying to work with Shapley value as a method to find contribution of each features on final predictions. This is a method uses in Explainable AI in which it helps to understand the impact of each part of the model on predictions.
The most important plot, named Summary plot is as follows:
In summary plot, we can see the feature importance in terms of Shapley value. The features are ordered accordingly to their importance. So, the first one is the most important one as has a high Shapley value range. In summary plot, we can see that

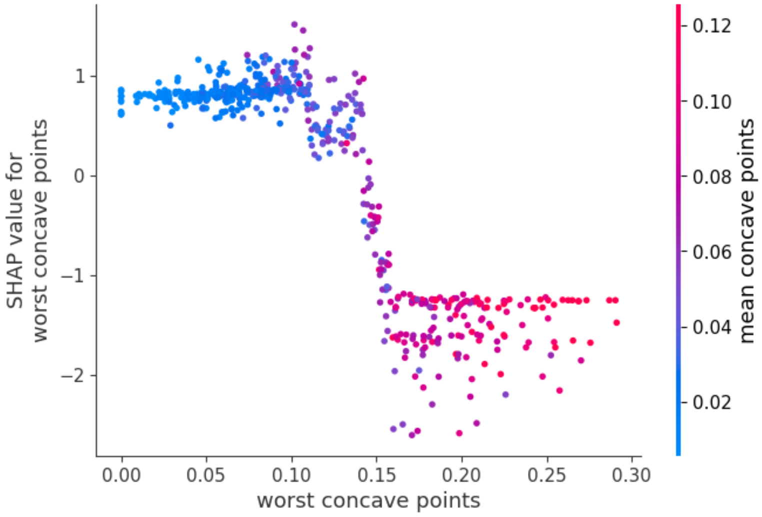
mean concave points, worst concave points, and area error are the three most influential features that can be impact on prediction.



Another type of summary plot is also shown in this report. The colors show the value of the feature from low to hight (blue to red). The diagram is also called Beeswarm plot which is used for visualizing the global attributions of features. These features are split into individual rows, with each row plotting all of the individual Shapley values along the x-axis, for all values of that feature. Each value of a feature is rendered as a point, so larger clusters of points (the Beeswarm) show where many feature values had a similar Shapley value. SHAP also colors each point with a normalized heat mapping from the feature's lowest to highest values [1]. If you look at the features in this plot, obviously it shows that mean concave points is mostly high with a negative SHAP value.Even this true for words concave points and worst texture. This means higher mean concave points, for example, tend to negatively affect the result.



If you want to see the effect of one specific feature on the prediction made by the model, the dependence plot will be used. A dependence plot is a scatter plot that shows the effect a single feature has on the predictions made by the model [3].



Each dot is a single prediction (row) from the dataset.The x-axis is the value of the feature (worst concave points).The y-axis is the SHAP value for that feature, which represents how much knowing that feature's value changes the output of the model for that sample's prediction. Each dot is a single prediction (row) from the dataset. The color corresponds to a second feature that may have an interaction effect with the feature we are plotting (by default this second feature is chosen automatically). If an interaction effect is present between this other feature and the feature we are plotting it will show up as a distinct vertical pattern of coloring. For example, for worst concave points values 0.15 with mean concave points between 0.06 and 0.1 are less likely to have cancer than worst concave points greater than 0.16 but mean concave points less than 0.06. This suggests an interaction effect between worst concave points and mean concave points.

This exercise shows only one aspect of explainable AI, which aims to make models more robust, accountable, and trustworthy.

References:

[1] Michael Munn David Pitman - Explainable AI for Practitioners_ Designing and Implementing Explainable ML Solutions-OReilly Media 2022

[2] https://www.datacamp.com/tutorial/introduction-to-shap-values-machine-learning-interpretability#

[3] https://shap-lrjball.readthedocs.io/en/latest/example_notebooks/plots/dependence_plot.html#:~:text=Simple dependence plot,-A dependence plot&text=Each dot is a single,model for that sample's prediction.