Explainable AI Exercise with Shapley Value

Ali Torabi

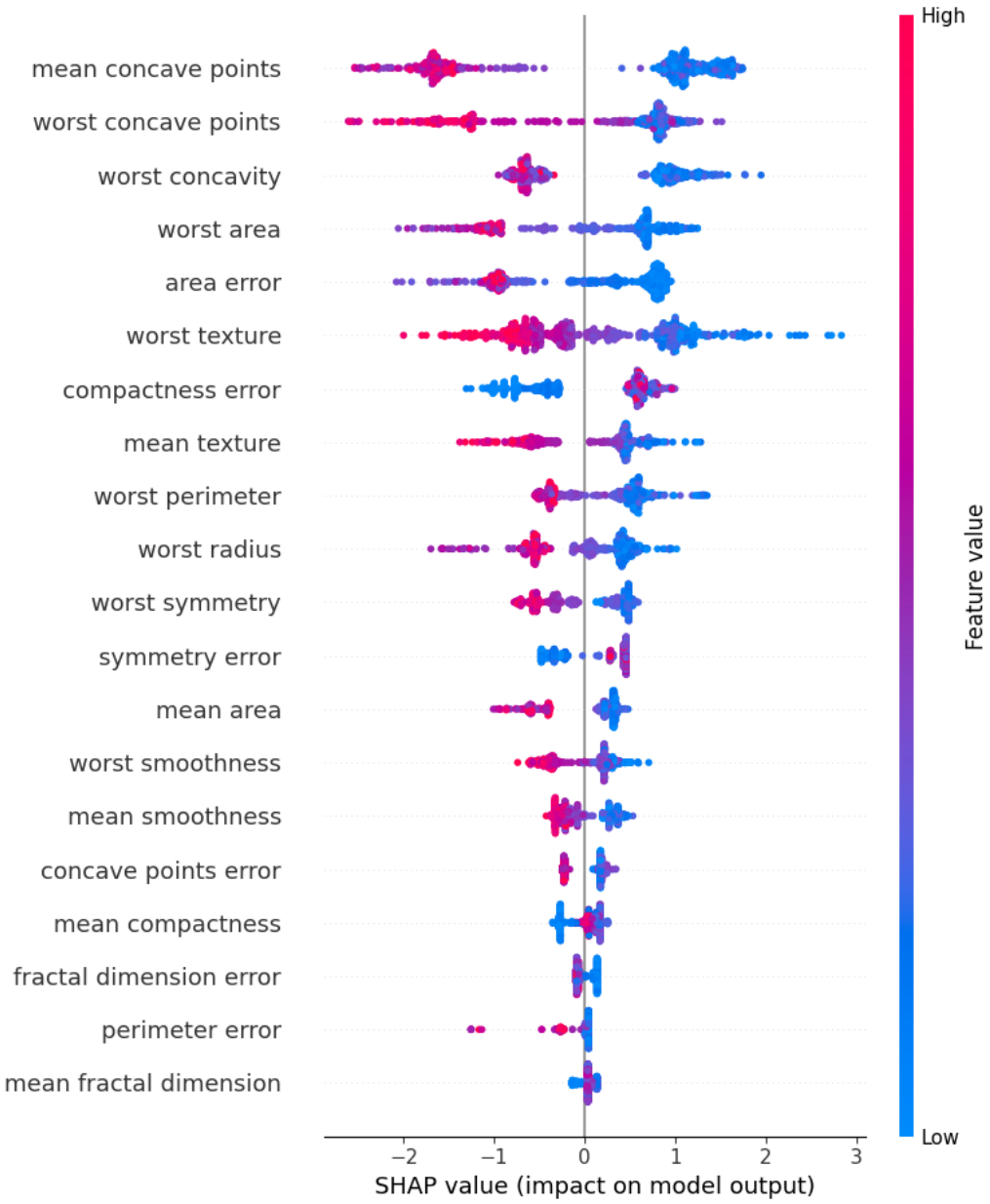
In this volunteer exercise, I'm going to use the Breast Cancer Dataset. In this dataset features get from digitized images of breast mass. This is a supervised learning case with a target distinguished by values 1 and 0, showing the sign of cancer or not, respectively. The aim is to use a method in Explainable AI, called Shapley Value, to make the model prediction more robust and interpretable to other practitioners like doctors.

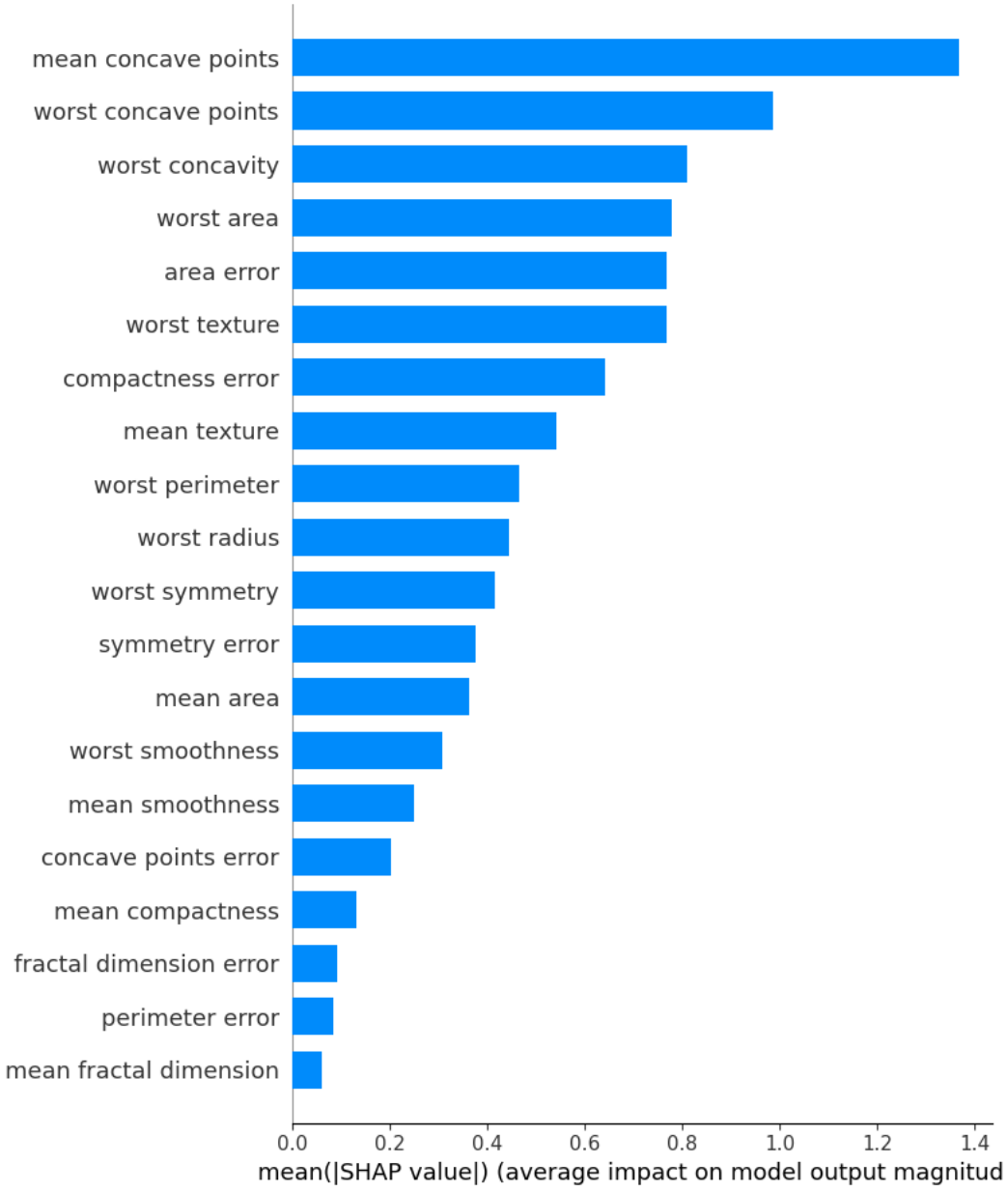
The accuracy by XgBoost learner is 95.61%. But the main idea of doing this exercise is that I'm trying to work with Shapley value as a method to find contribution of each features on final predictions. This is a method uses in Explainable AI in which it helps to understand the impact of each part of the model on predictions.

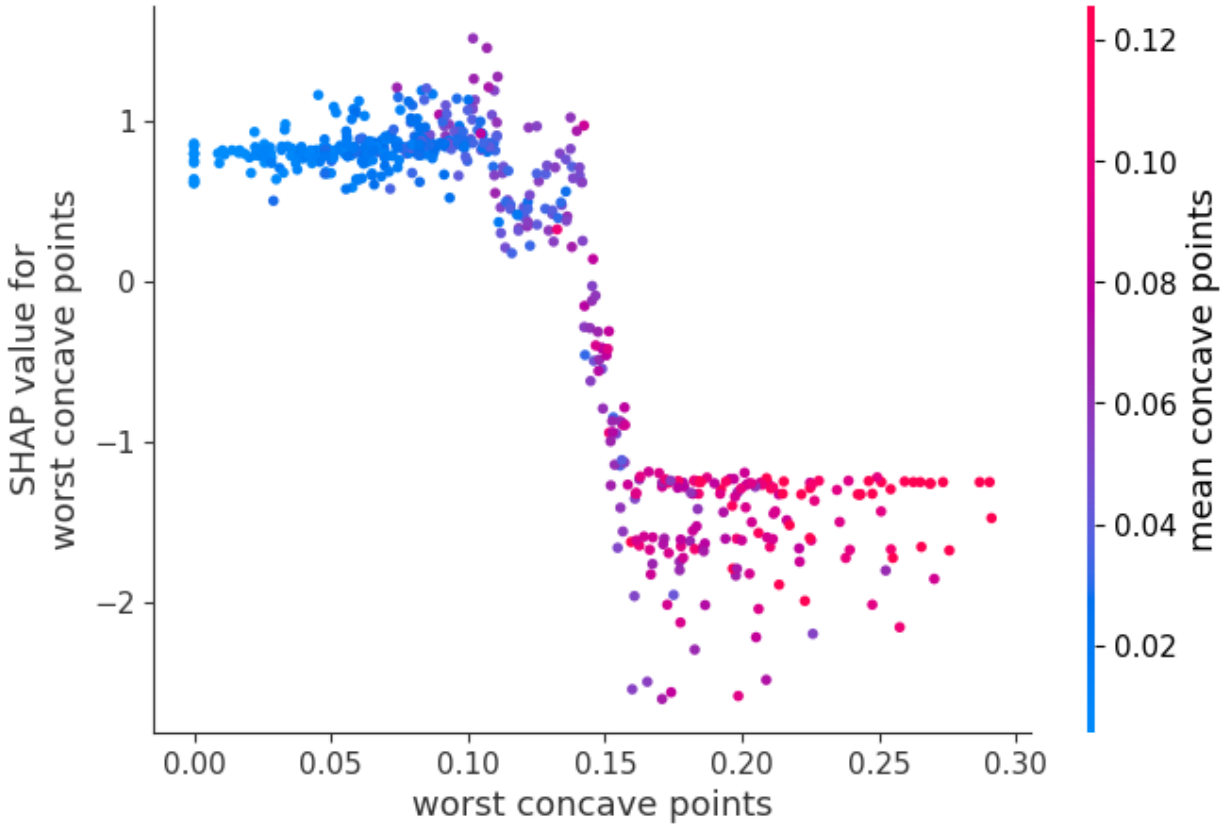
The most important plot, named Summary plot is as follows:

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. The colors show the value of the feature from low to hight (blue to red).

If you want to see the effect of one specific feature on the prediction made by the model, the dependence plot will be used. For example, in the last diagram, as the worst concave point gets higher, the impact on the possibility of having cancer will be getting higher, too.


mean concave points
worst concave points
worst concavity
worst area
area error
worst texture
compactness error
mean texture
worst perimeter
worst radius
worst symmetry
symmetry error
mean area
worst smoothness
mean smoothness
concave points error
mean compactness
fractal dimension error
perimeter error
mean fractal dimension
−2
−1
0
1
2
3
SHAP value (impact on model output)
High
Feature value
Low


mean concave points
worst concave points
worst concavity
worst area
area error
worst texture
compactness error
mean texture
worst perimeter
worst radius
worst symmetry
symmetry error
mean area
worst smoothness
mean smoothness
concave points error
mean compactness
fractal dimension error
perimeter error
mean fractal dimension
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
mean(|SHAP value|) (average impact on model output magnitud



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