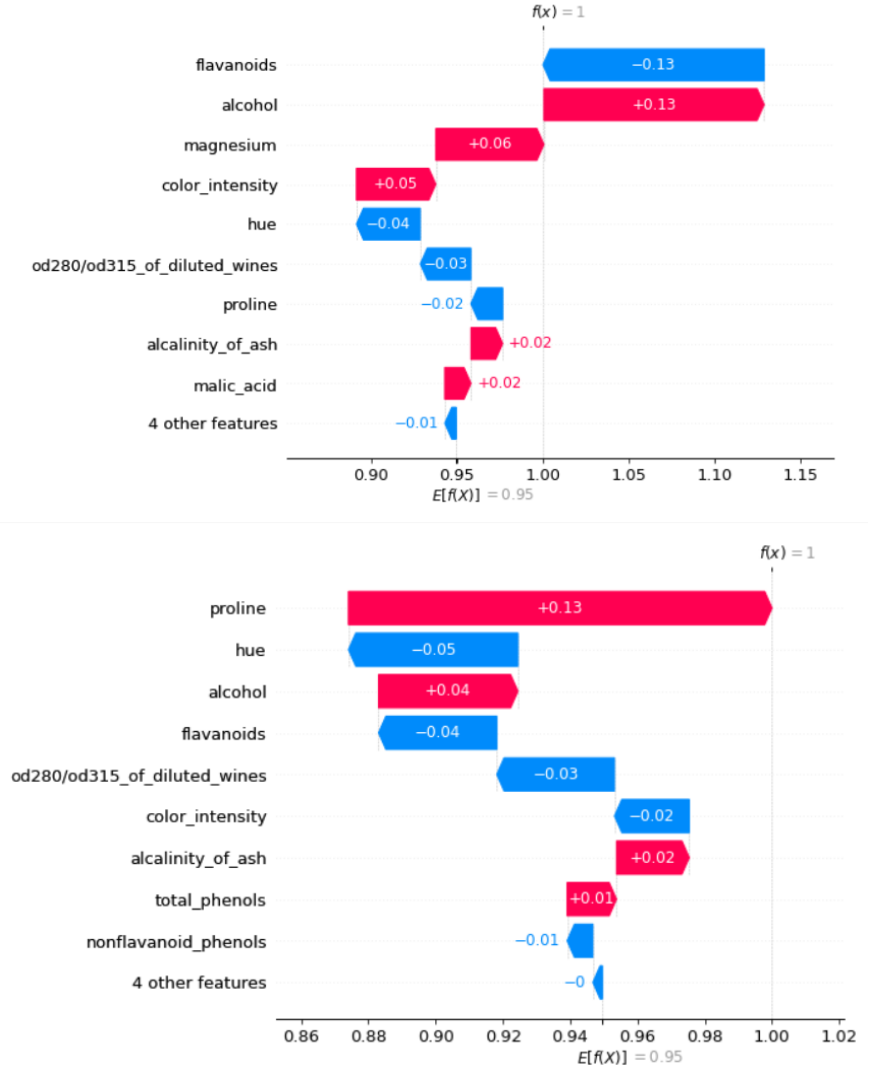
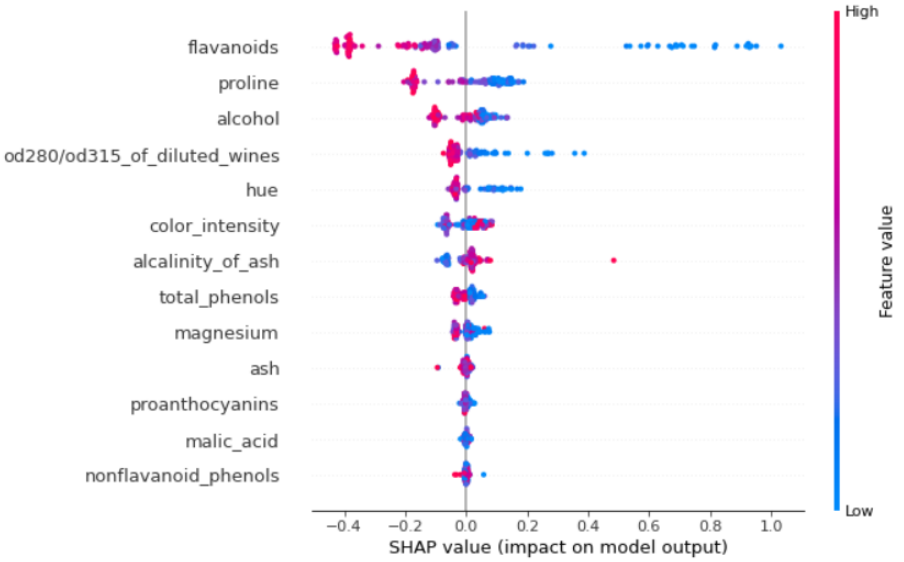
Homework 2

Since SHAP measures local feature's interpretations, different observations may result in different feature importance. Following document describes those differences. Experiments were conducted on UCI ML wine recognition datasets.

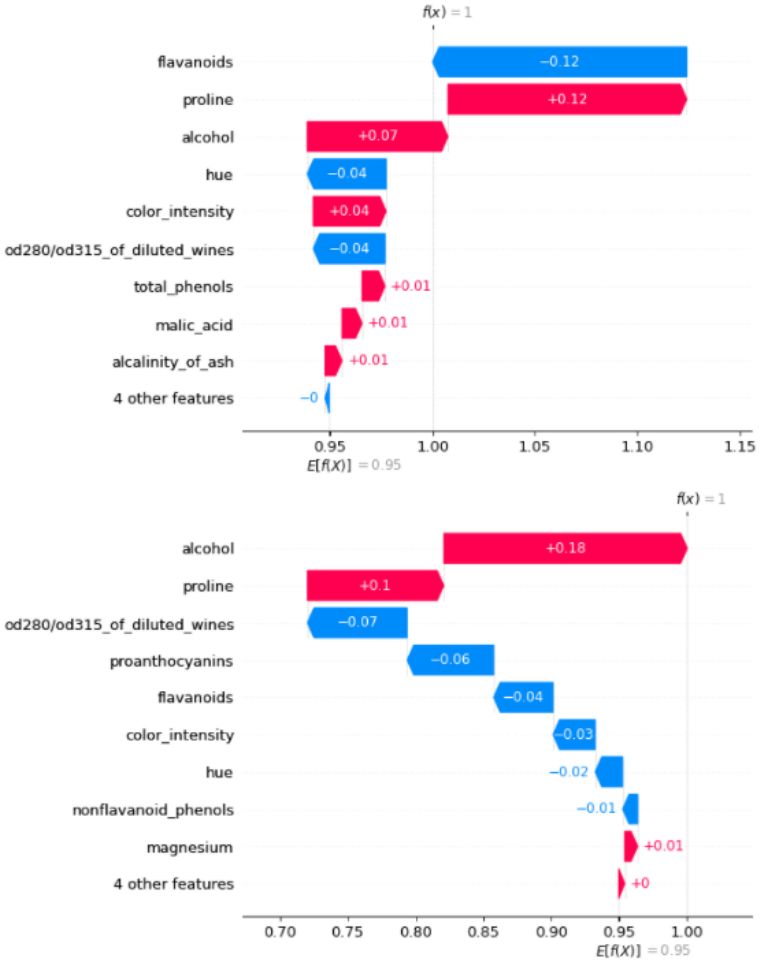


Above examples were generated by a gradient boosting classifier that uses 200 regression trees. If we compare three most important features for those two samples we get (flavanoids, alcohol, magnesium) and (proline, hue, alcohol). Intersection has only one element (alcohol), which has quite a different impact (+0.13 vs + 0.04). What is more interesting is that proline, which is the most descriptive feature with positive impact in the second example, have negative impact in the first example. Therefore the same feature may have positive or negative impact on predicting the same class, depending on the analysed example. And that's not useless feature, here we have plot that utilises all the data:



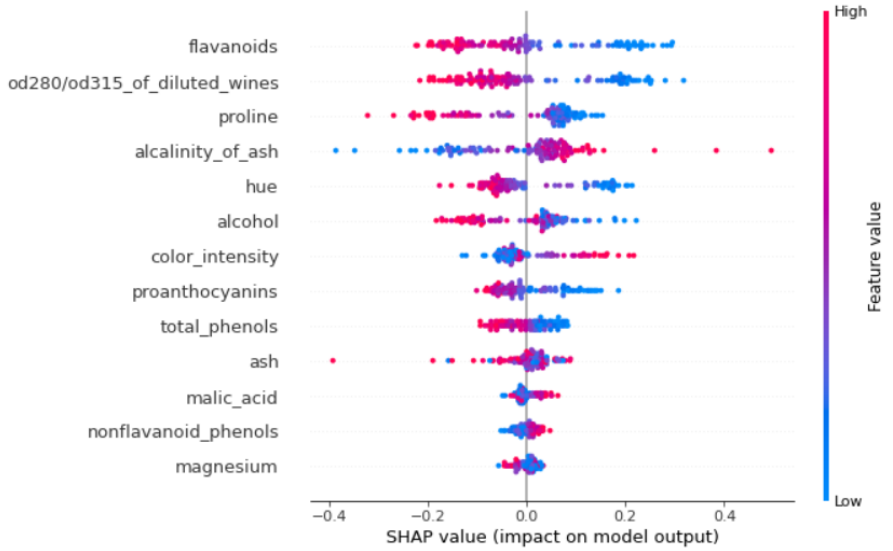
Proline clearly is very descriptive and in general has a negative impact on wine class (the higher proline value, the lower wine class).

We can compare SHAP values for different models. Here is a comparison of previous model (higher) with multi-layer perceptron classifier (lower):



We can see that the order and impact of features is different. For example alcohol which is the third most important feature for the first model has over 2 times the influence on the

second model, where it is most important. Surprisingly while comparing those two models I found very little differences, no more than within predictions of the same model. But when we look on data computed on whole dataset for second model differences are more visible:



Apart from changed order of features we can see that SHAP values are more stretched out and colors are more separated. That means that the second model uses features better, and extracts more information from them.