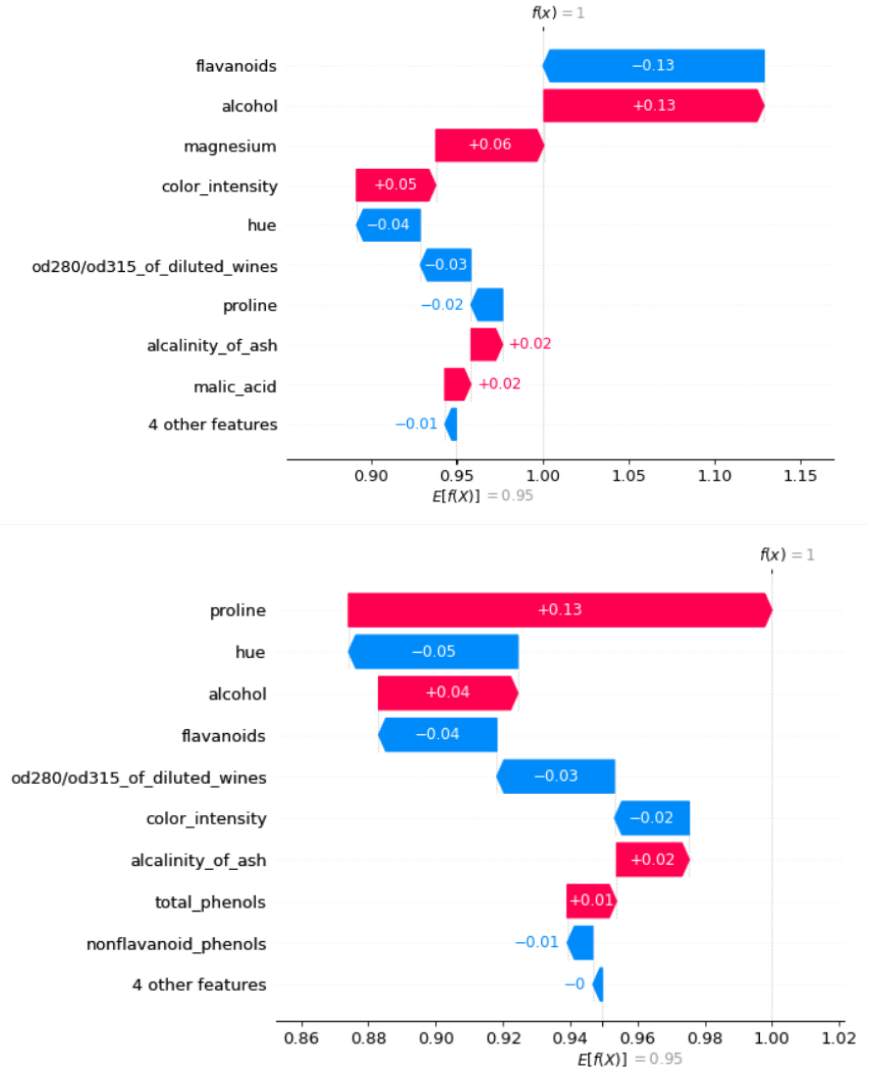
Homework 2

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 dataset, by trying to predict wine class (class is in fact encoding of one of three different wine cultivars from the same region of Italy) using data from chemical analysis of wine.

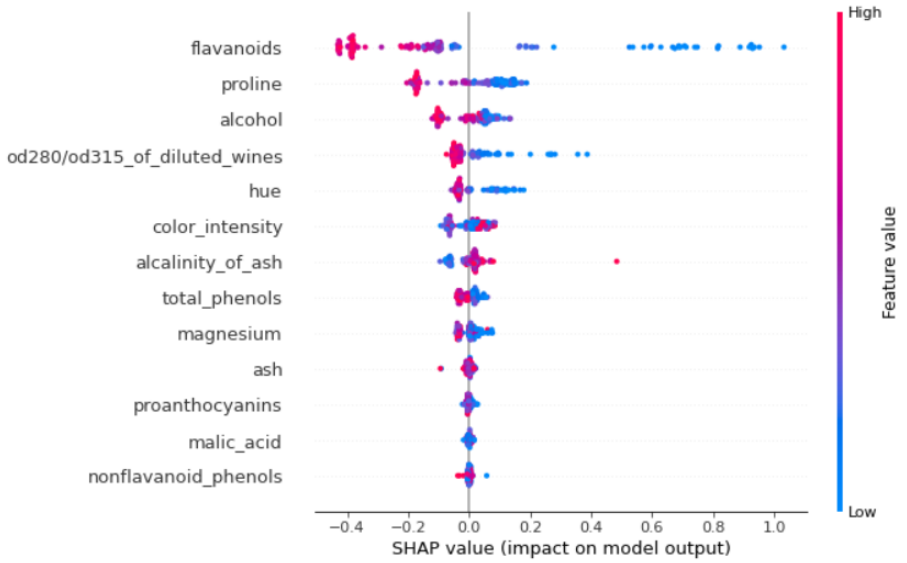
Task 4



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).

Task 5

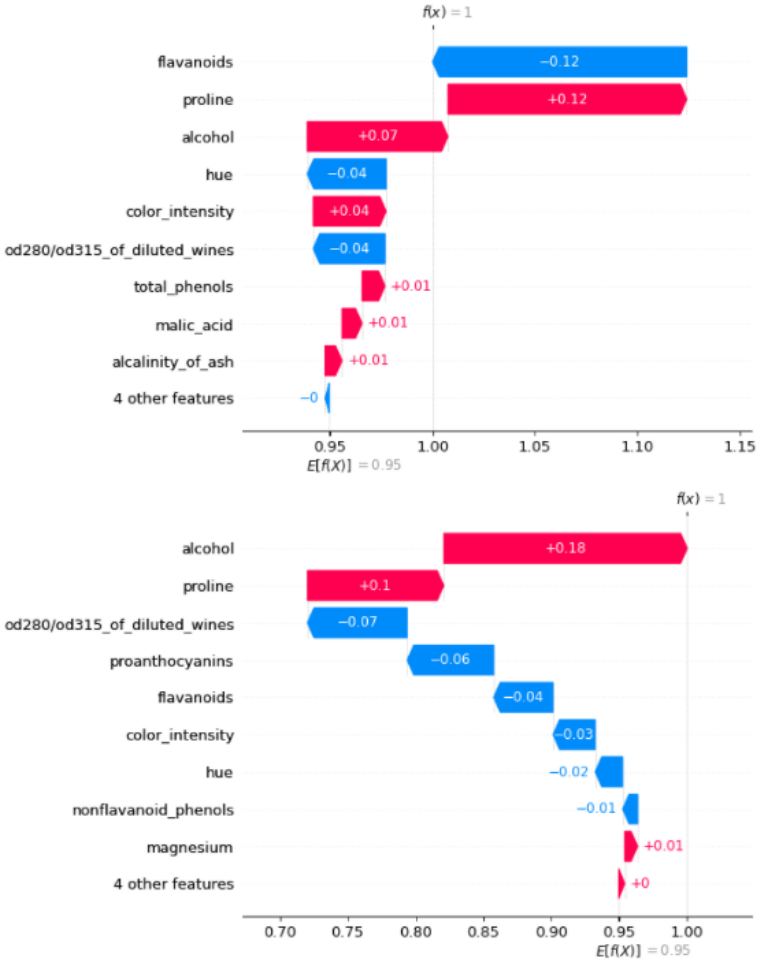
What is more interesting in the above plots is that proline (a kind of organic acid), which is the most descriptive feature with positive impact in the second example, has negative impact in the first example. Therefore the same feature may have positive or negative impact on predicting the same class, depending on the analysed example. Moreover that's not a useless feature, here we have plot that utilises all the data:



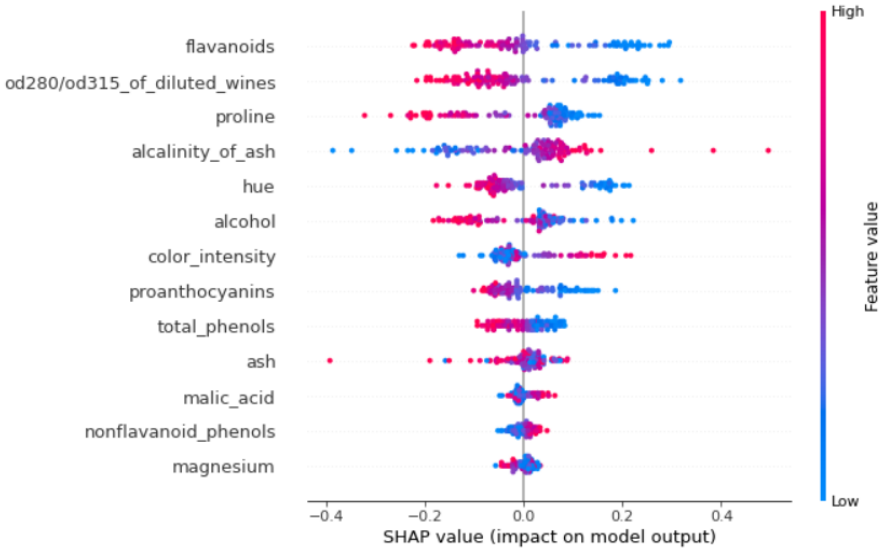
Proline clearly is very descriptive and in general has a negative impact on wine class. Since class of wine is just an encoding, we can interpret that the following way. The more proline wine has the more likely it comes from a wine cultivator encoded with number 0. Wines with very little proline are more likely to come from cultivator encoded with number 2, and wines with average amount of proline are more likely to come from cultivator encoded with number 1.

Task 6

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.