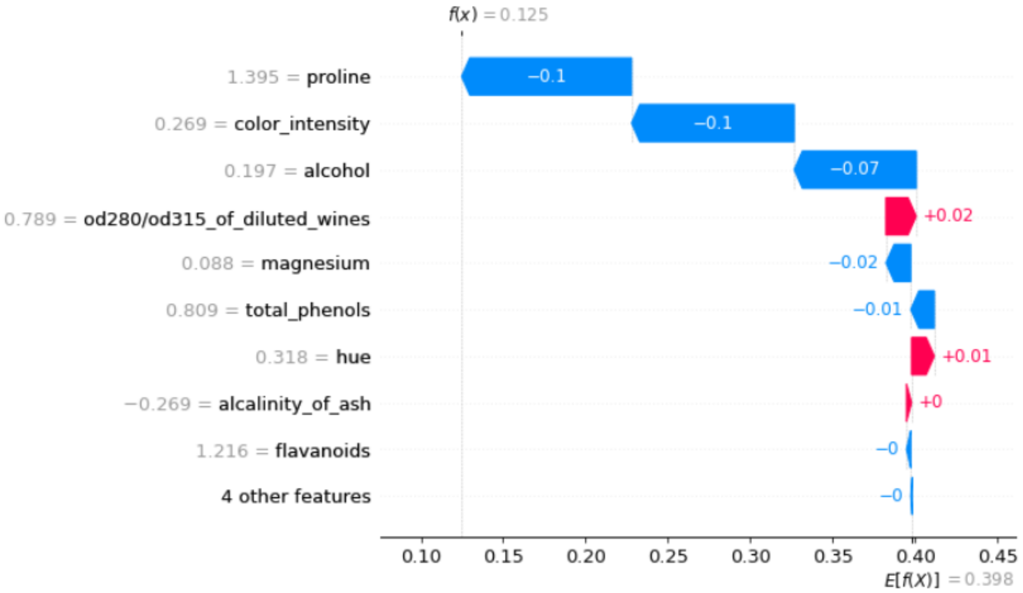
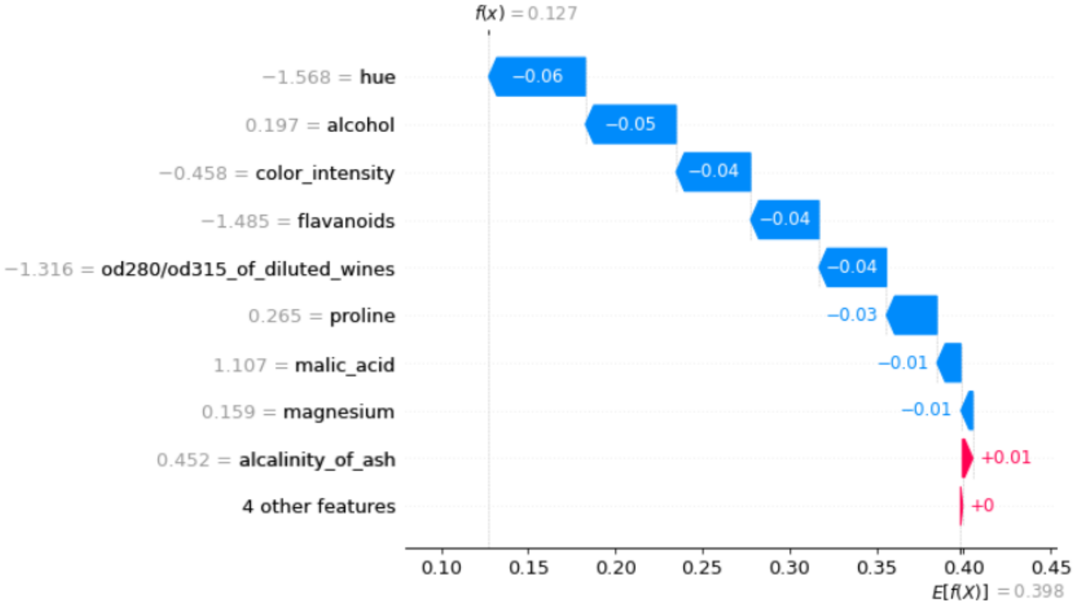
Homework 2

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. First model that was used in experiments was a random forest classifier with 100 trees. Later, in section Task 6, it was compared to k nearest neighbour classifier with parameter k equal to 5.

Task 4

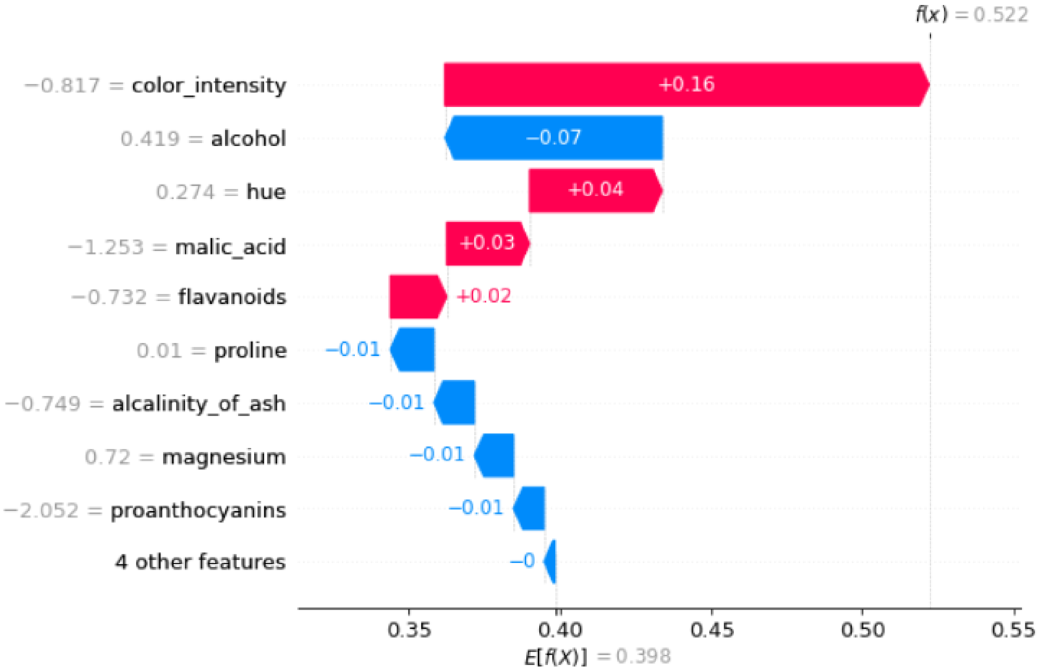




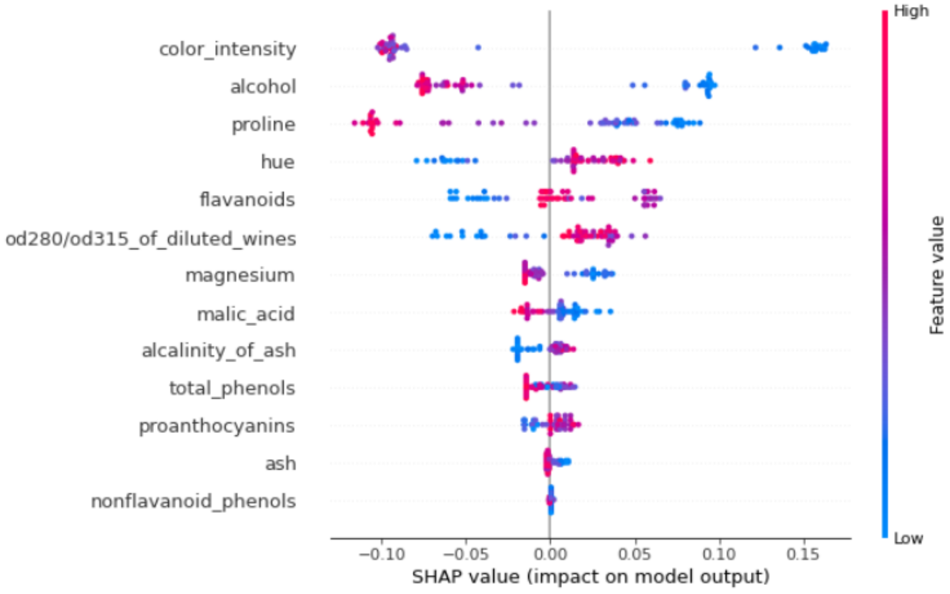
Above examples were generated using the first model. They present how features impact probability of wine being from cultivator encoded with "1". Based on training data this probability without any additional information (without knowledge about any features) is equal to 0.398. As we can see the biggest impact on this result in the first case have “proline” and “color_intensity”, while in the second case “hue” and “alcohol”. Although the model's prediction is similar (around 0.13), reasons for this prediction are different.

Task 5

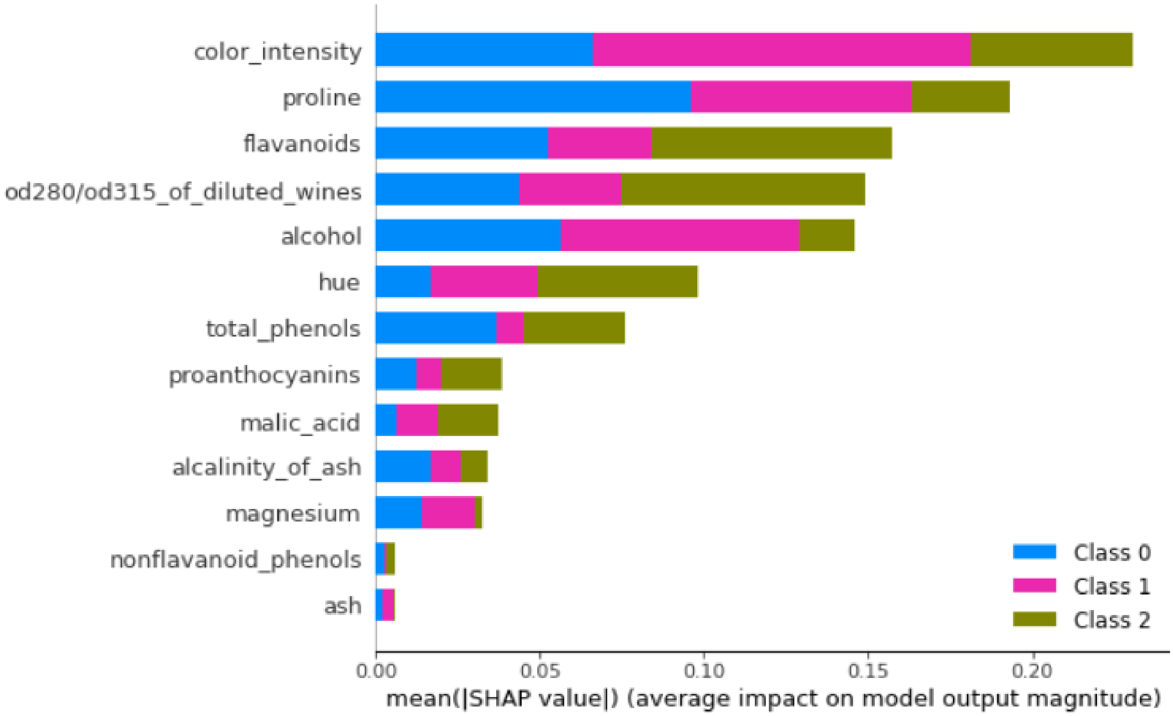
What is more interesting is that in the great majority of cases “color_intensity” has a strong **negative** impact on probability of predicting wine class “1” (like in above examples), but there are some cases (like the one below) where the same feature has the strongest **positive** impact.



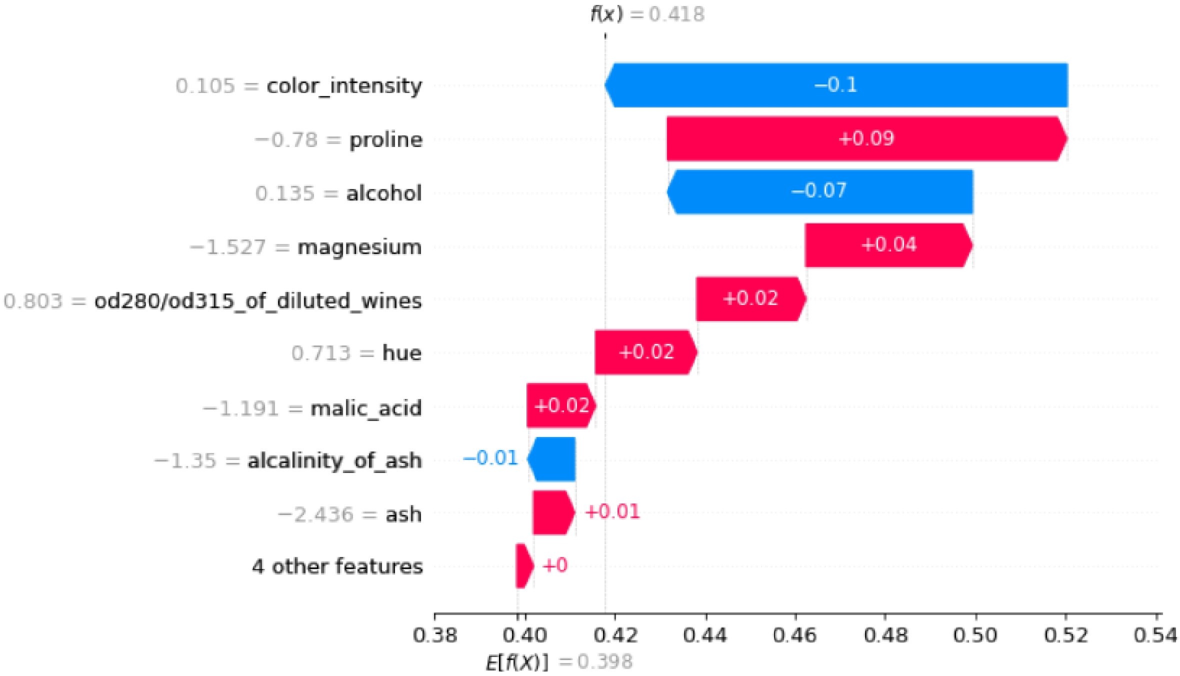
And that’s not a useless feature, as we can see below it has great predictive power for class “1”.

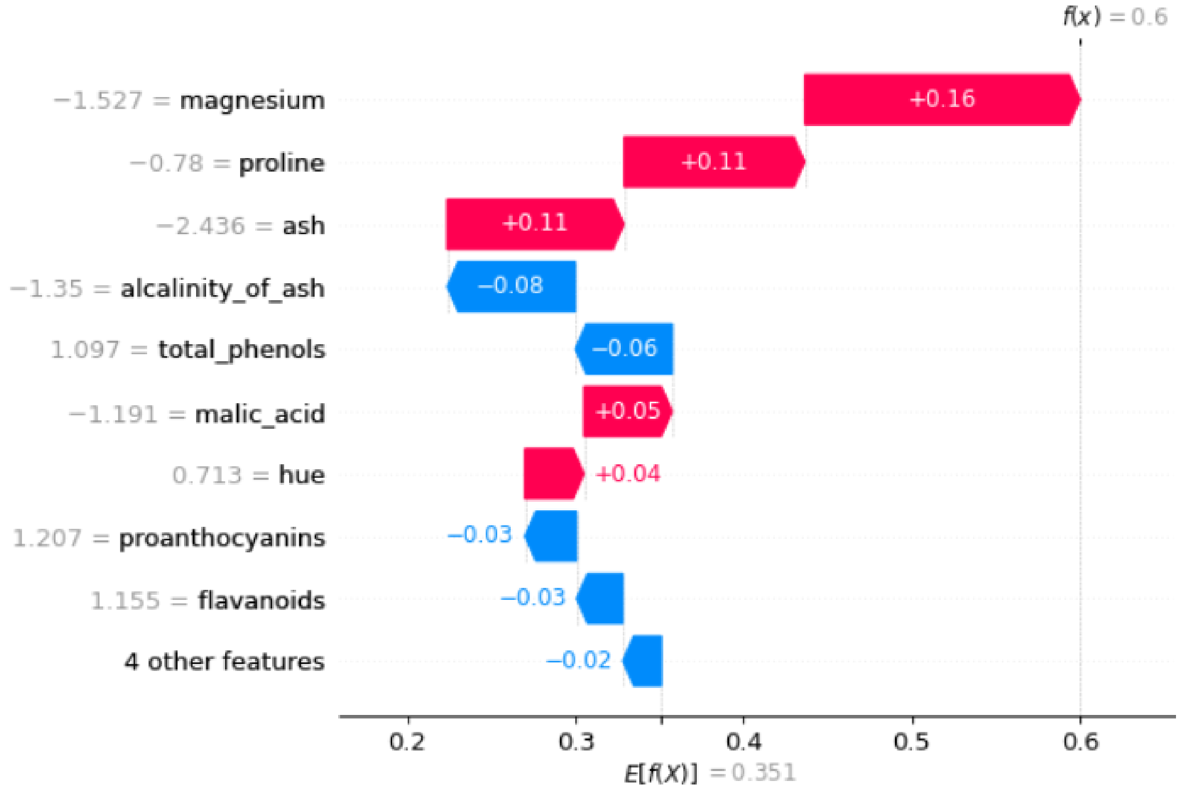


Actually it was an understatement. When we look at all classes we realise that this feature has also the second greatest SHAP value for class “0”, and one of the greatest for class “2” as well.



Task 6





Here we can compare SHAP values for different models. The only similarity between the first plot for the first model and second for the second model is their base probability, which is a little less than 0.4. Unfortunately everything else (final prediction, order of features, their impact) differ.