

Feature ranking- Laboratorium nr 3

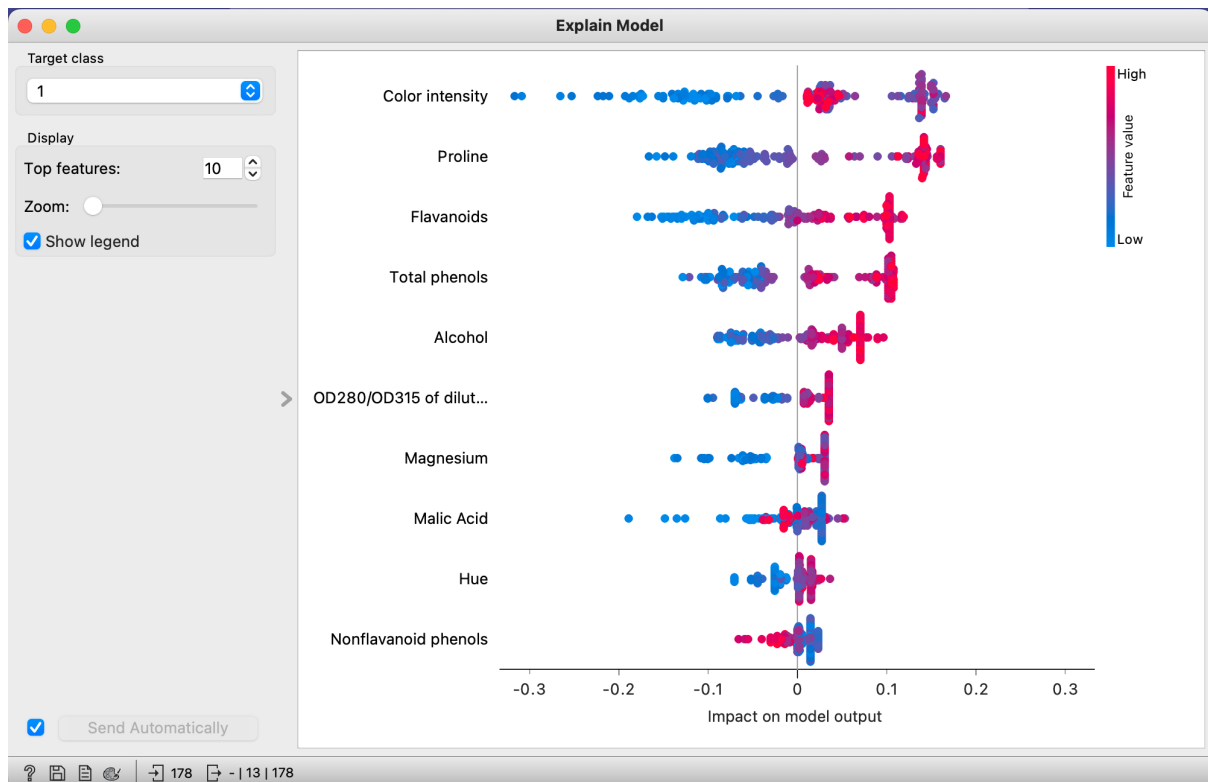
Autor: Filip Katulski

Do wykonania zadania wybrałem zbiór „Wine”, zawierający wina z 3 różnych szczepów win z południa Włoch.

Porównanie Modelu Random Forest oraz SVM:

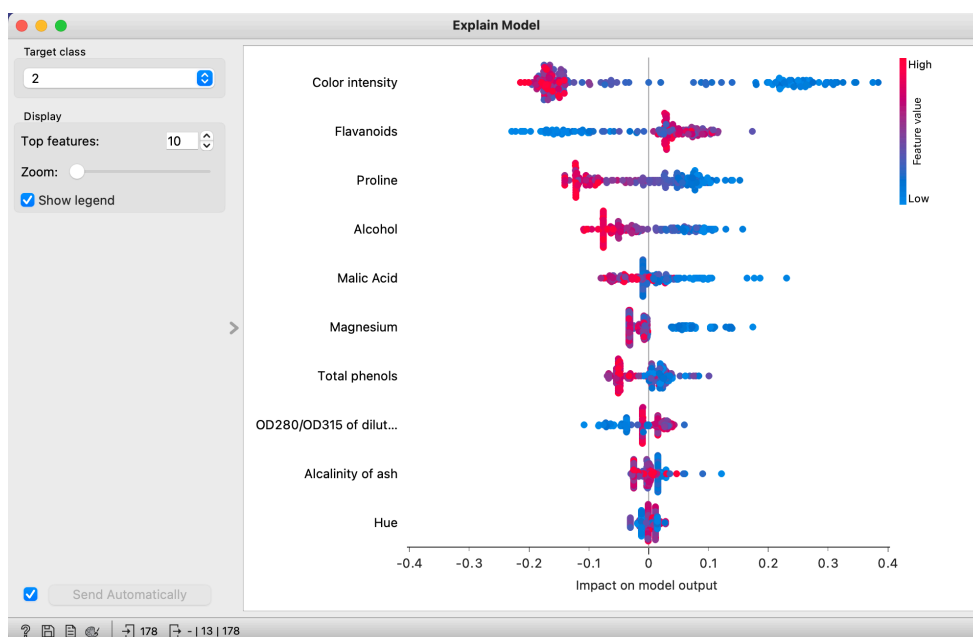
1. Algorytm Random Forest:

1. Wyjaśnienie modelu na przykładzie kategorii 1 za pomocą narzędzia Explain Model:



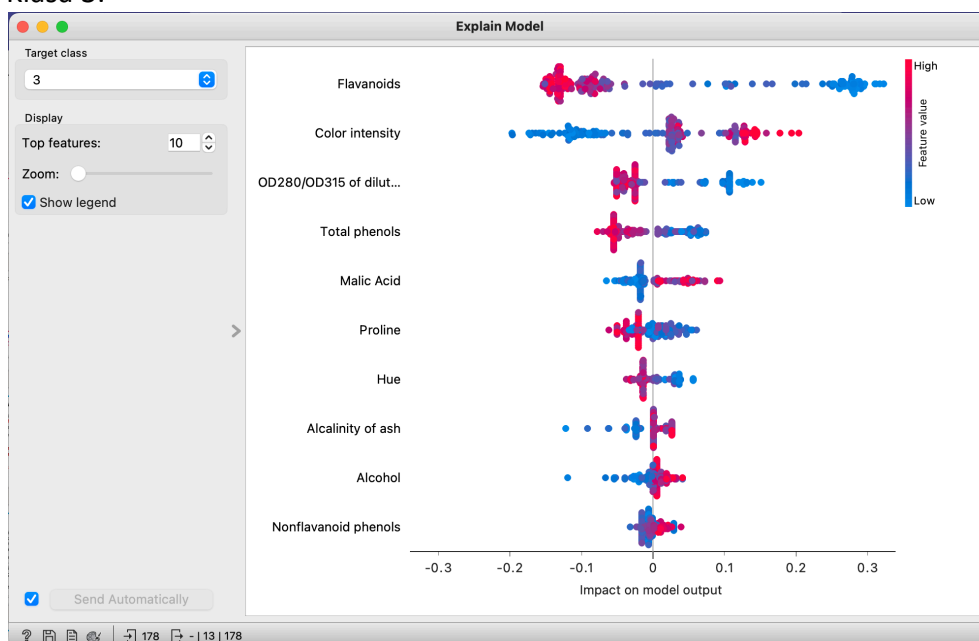
Moduł ten opisuje wartość każdej cechy (feature value) dla danej kategorii. Im bardziej czerwony punkt tym efekt jest większy. Najważniejszymi cechami dla kategorii pierwszej są: intensywność koloru, zawartość Proliny oraz Flawanoidów oraz całkowitej ilości Fenoli w winie. Wysoka intensywność koloru oraz wysoka zawartość ww. związków zwiększa szanse na zakwalifikowanie wina do klasy pierwszej.

Klasa 2:



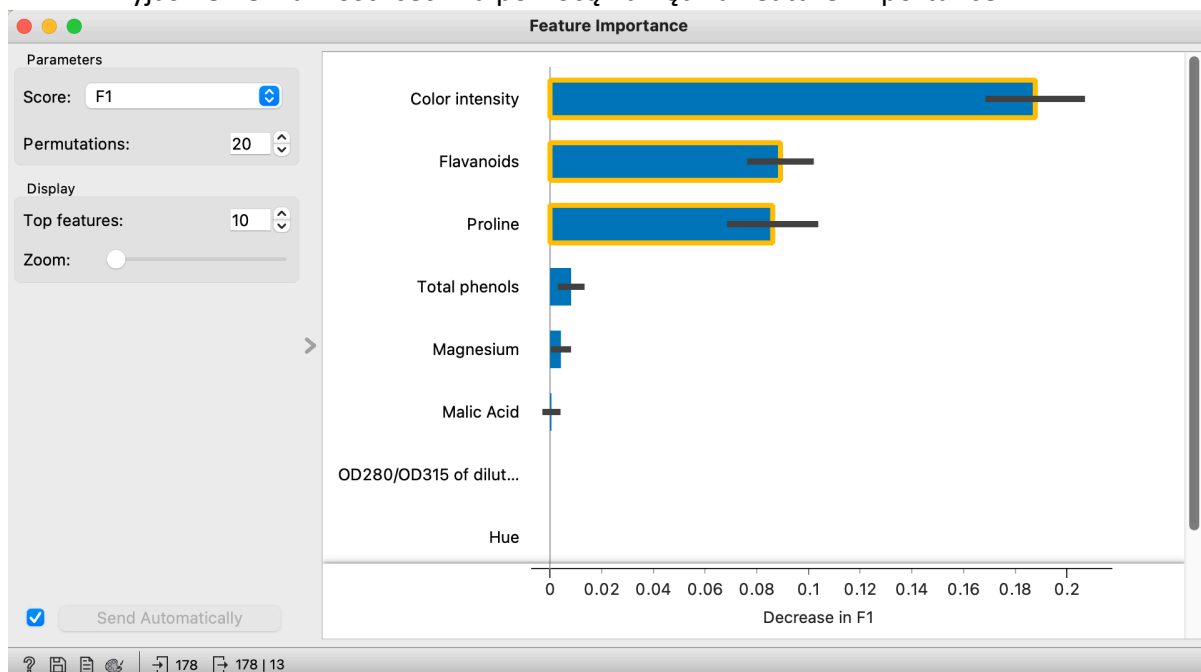
Najważniejszym wyznacznikiem tak samo jak poprzednio jest intensywność koloru i zawartość Flawanoidów oraz Proliny. Jednocześnie model sugeruje, że mniejsza ilość alkoholu wpływa na zakwalifikowanie modelu do tej klasy.

Klasa 3:



Według modelu trzeci szczep wina zawiera niewielką ilość Flawanoidów, co jest jego najważniejszą cechą. Intensywność koloru sugeruje, że im wyższa jej wartość tym większa szansa na zakwalifikowanie do klasy 3, lecz widoczne są wartości odstające.

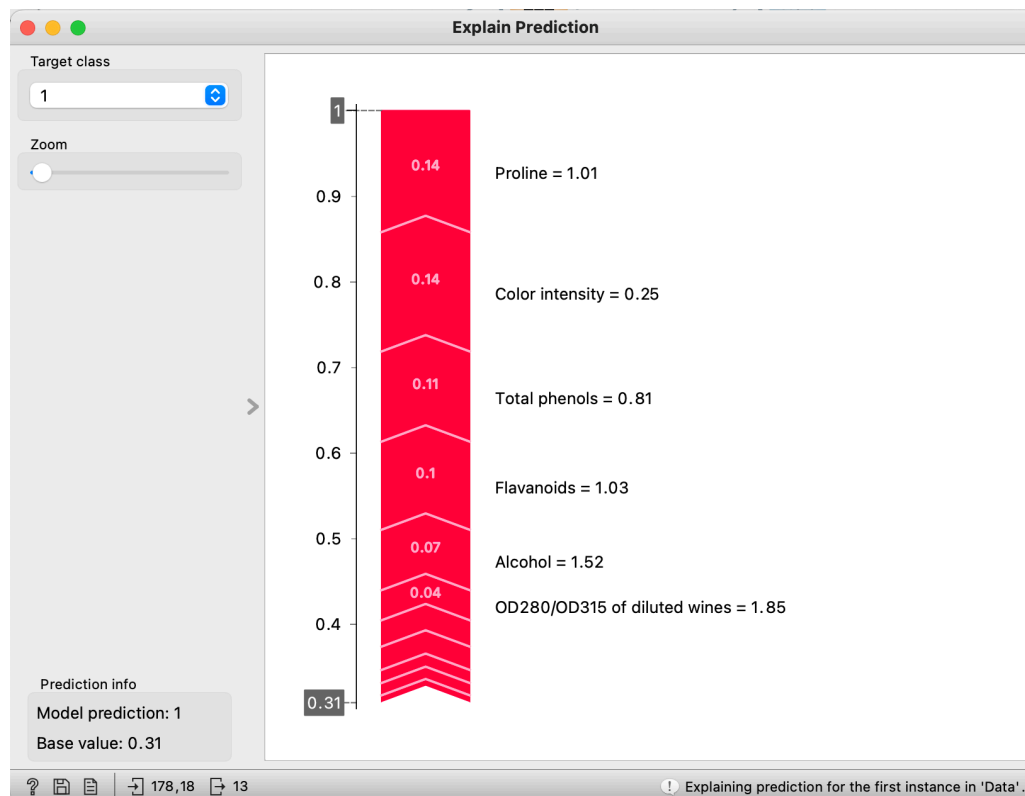
2. Wyjaśnienie ważności cech za pomocą narzędzia Feature Importance:



Narzędzie to oblicza spadek wyników (w moim przykładzie użyłem F1-score) po wyizolowaniu danej cechy przy klasyfikacji. Tutaj zgodnie z Explain Model najbardziej znaczącymi cechami są intensywność koloru oraz zawartość Flawanoidów i Proliny.

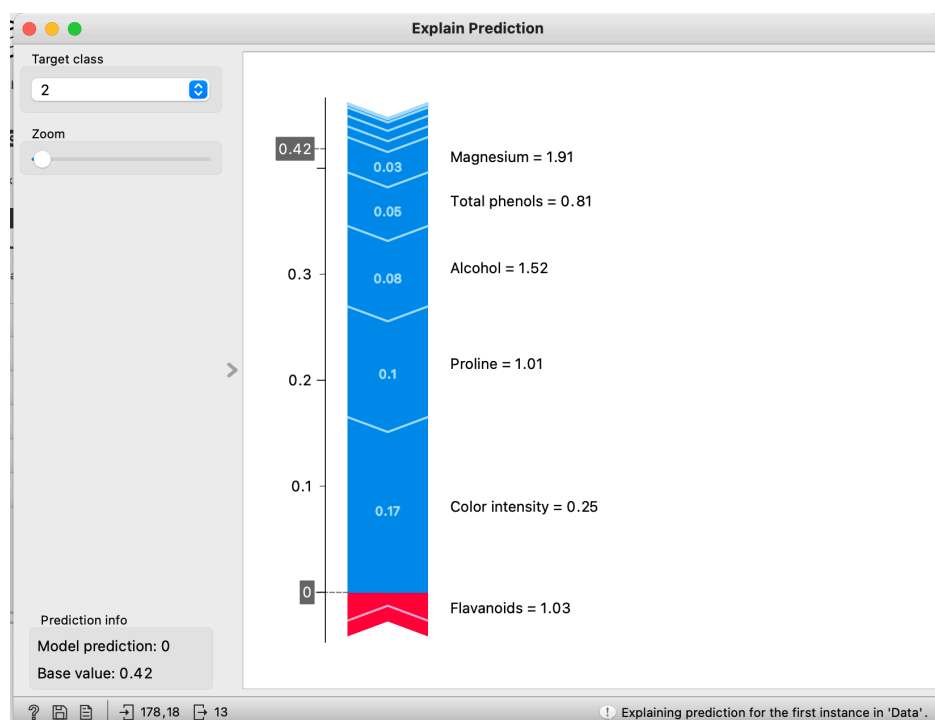
3. Wyjaśnienie predykcji za pomocą Explain Prediction

Predykcja dla pierwszej instancji w zestawie danych.

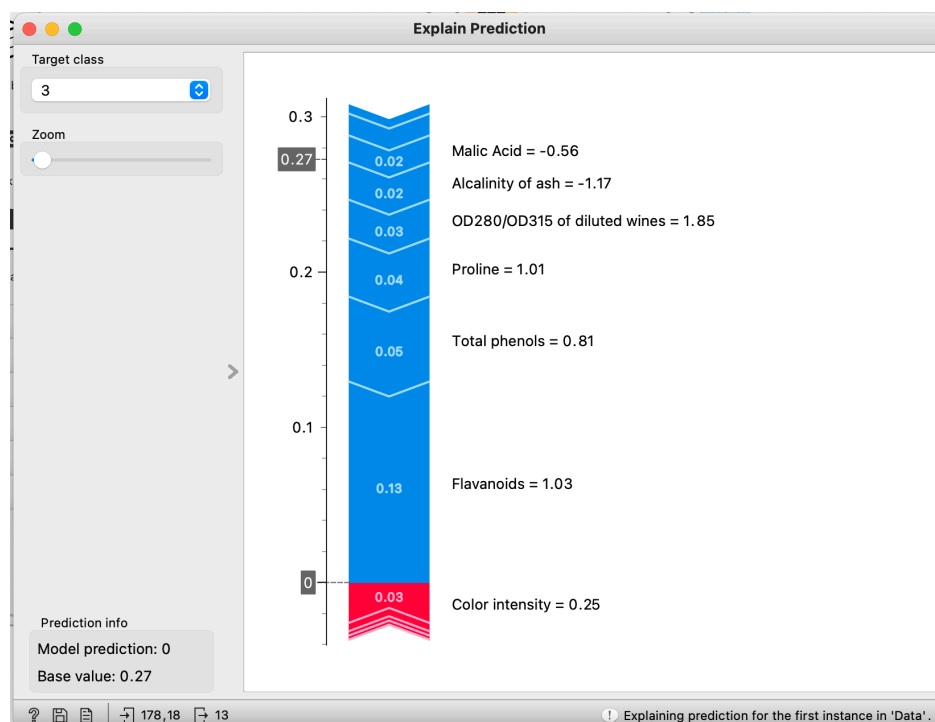


„Strzałka” odpowiada wartości Shapleya. Im wyższa wartość tym dana cecha bardziej odpowiada za zakwalifikowanie tego elementu do danej klasy.

Klasa 2:



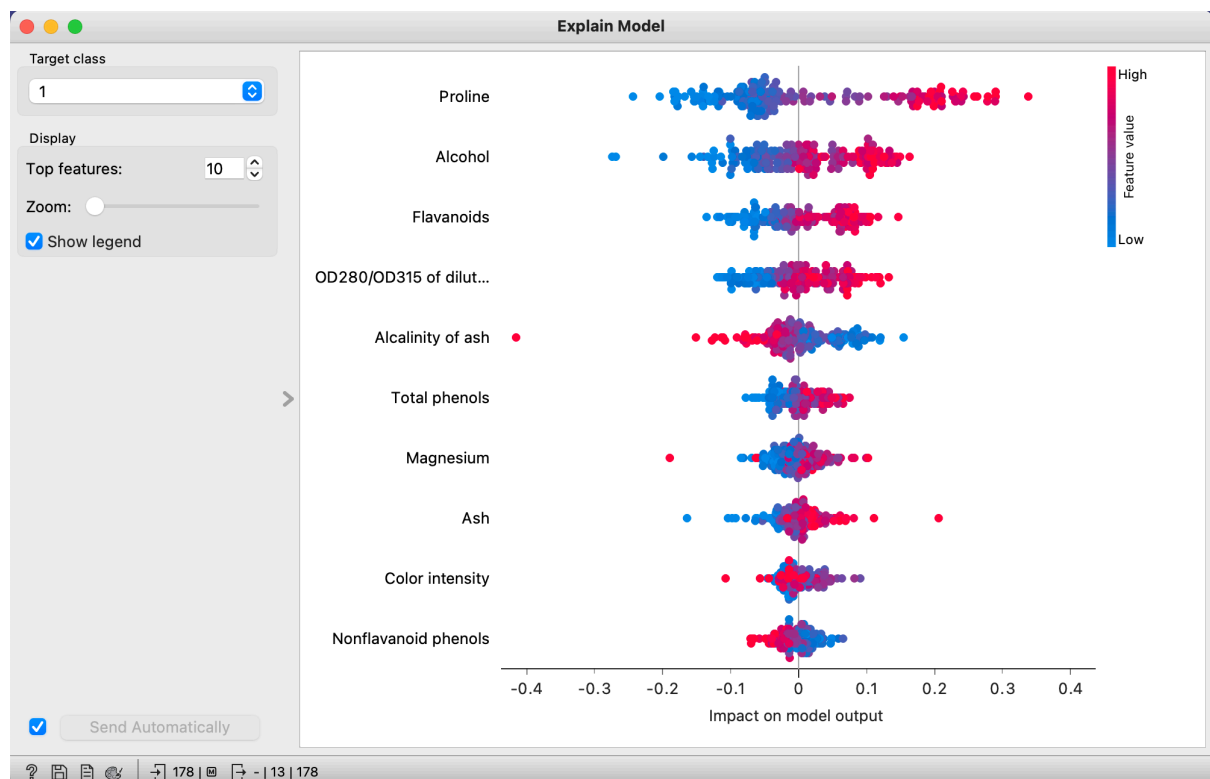
Klasa 3:



Z tego zestawu cech wynika, że pierwszy element został zakwalifikowany jako klasa 1, ze względu na wysoką zawartość Proliny, intensywność koloru oraz całk. Zawartość Fenoli.

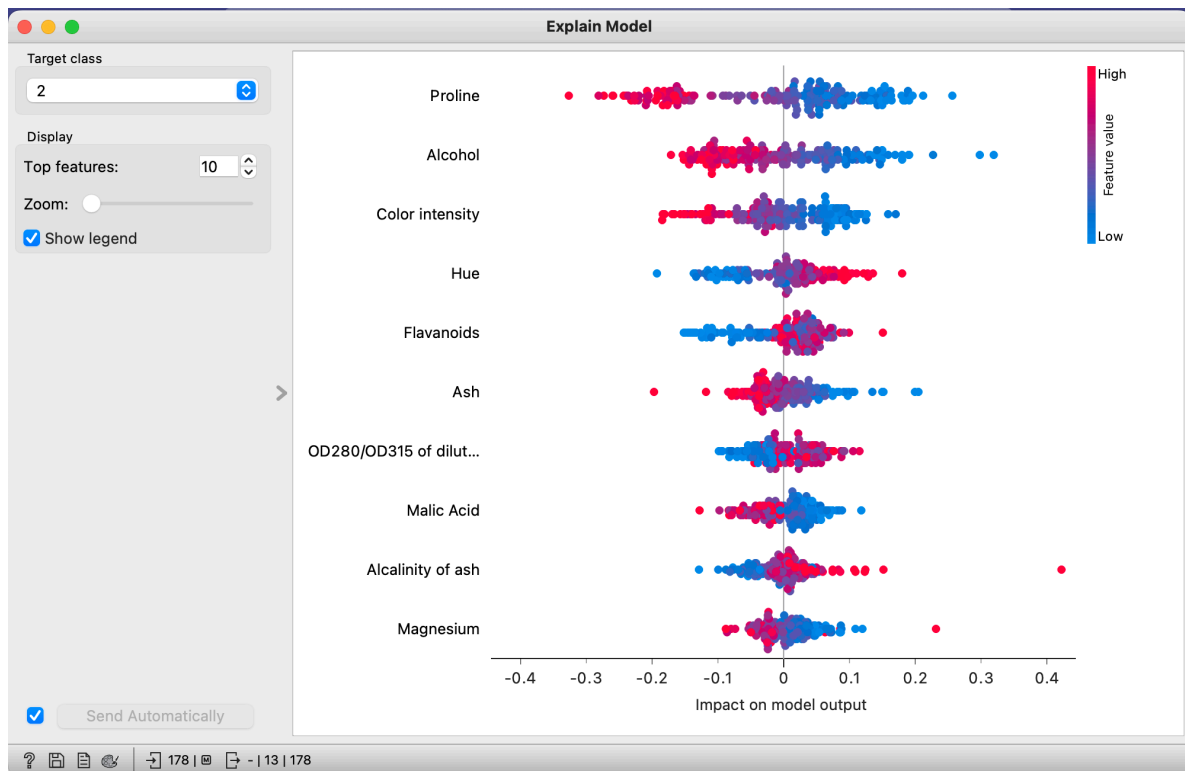
2. Model SVM:

1. Wyjaśnianie modelu za pomocą Explain Model:

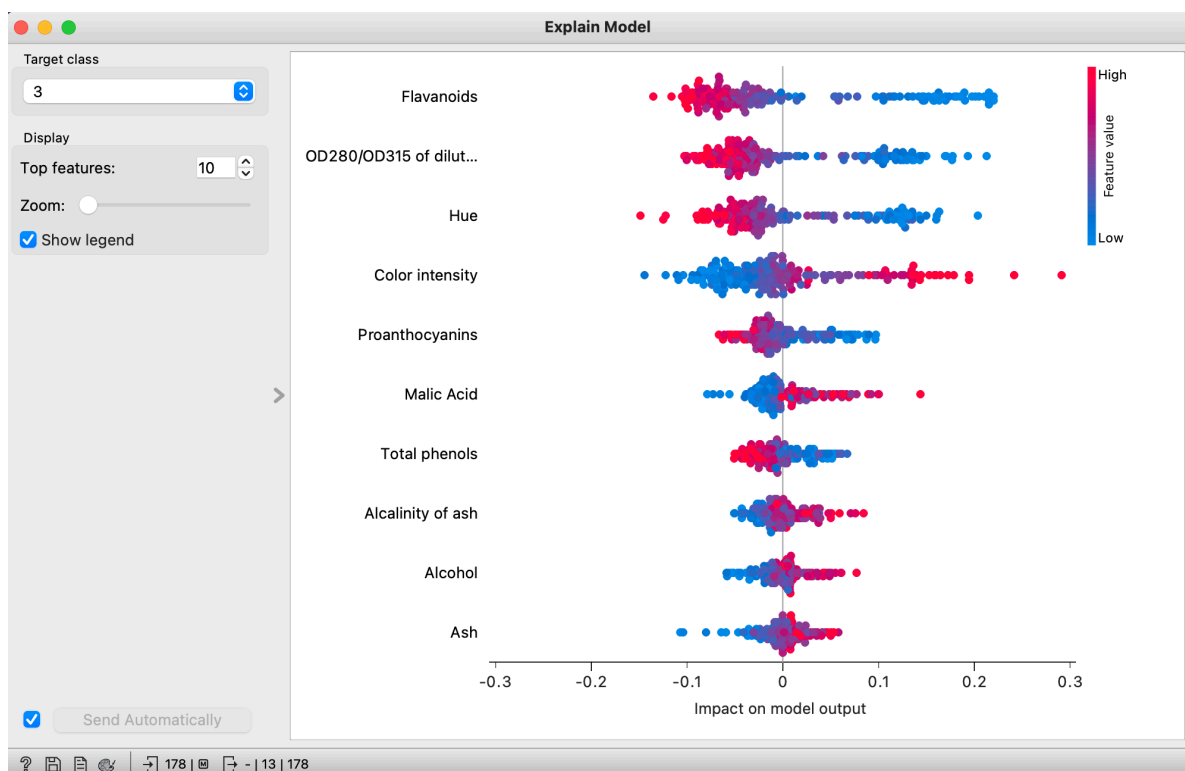


Model za bardziej znaczące cechy przyjmuje Prolinę, ilość alkoholu oraz pozostałe związki chemiczne. Intensywność koloru nie jest już tak znacząca dla klasy pierwszej, jednak nadal ważna dla klas 2 i 3.

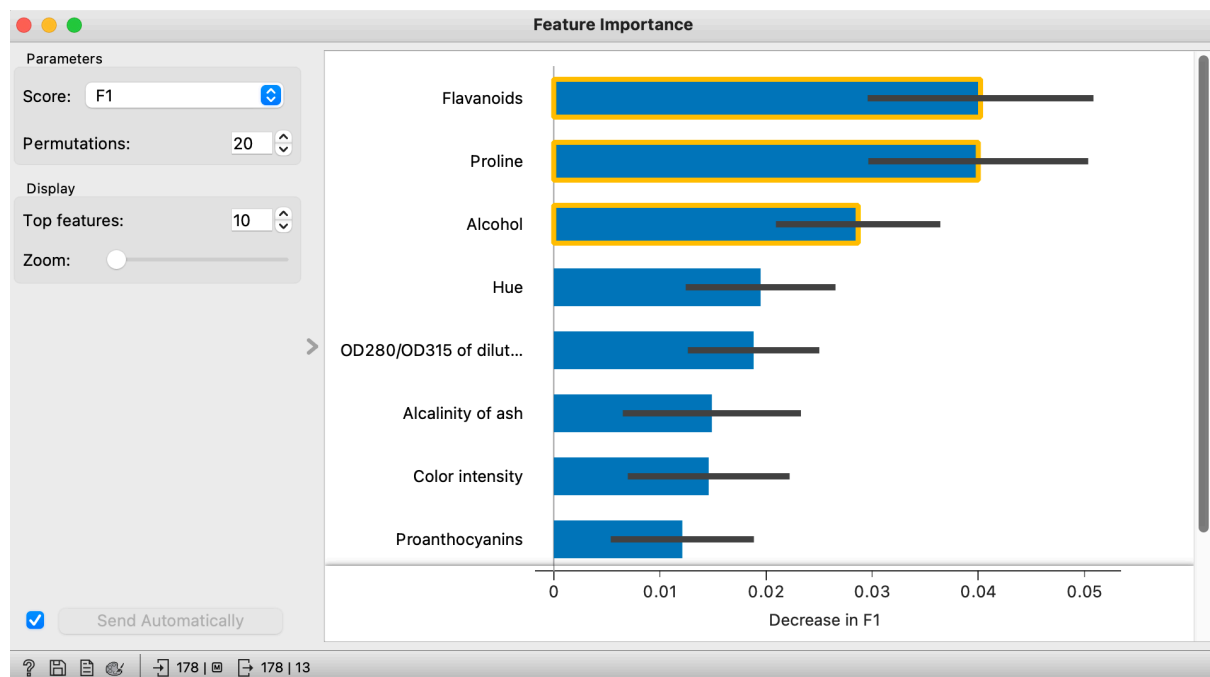
Klasa 2:



Klasa 3:

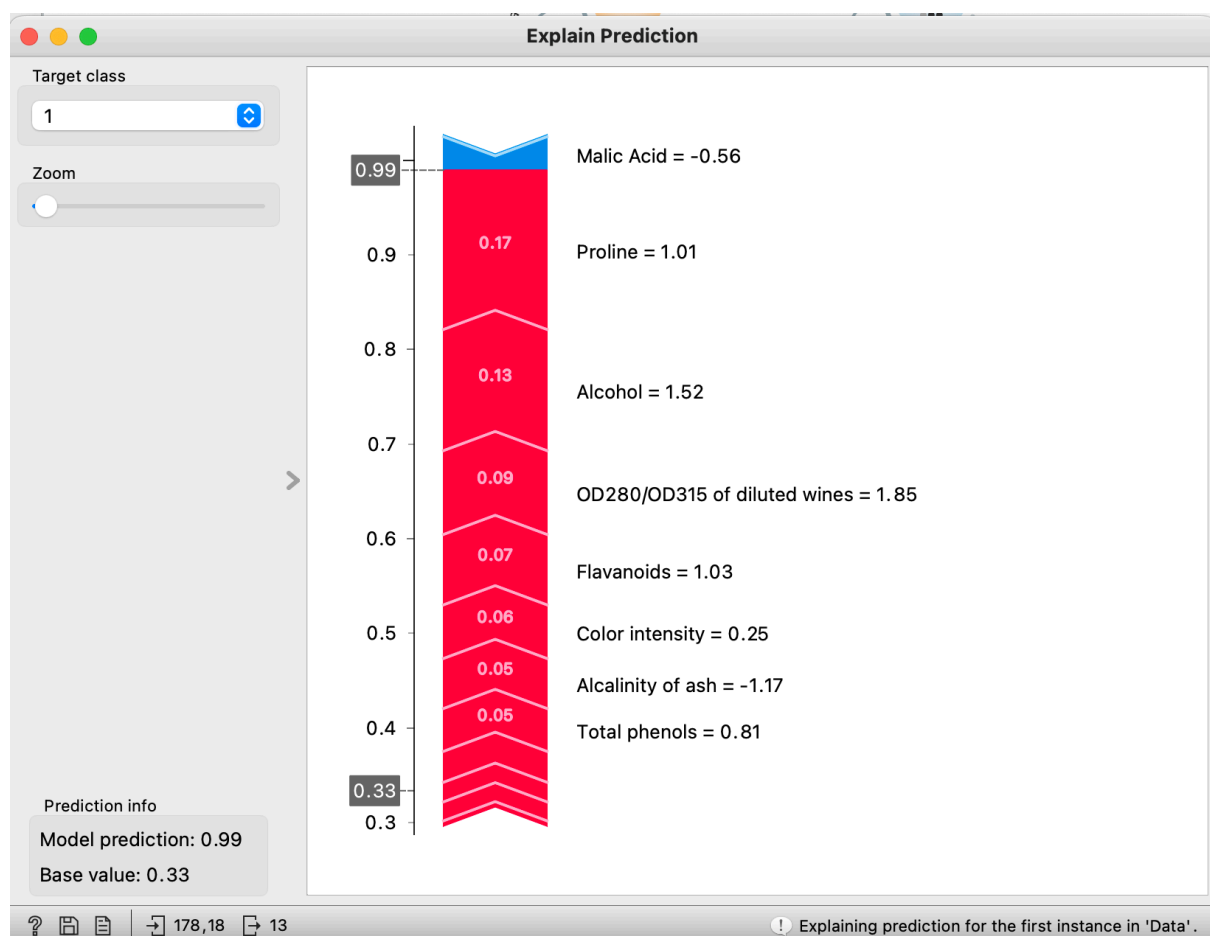


2. Feature Importance



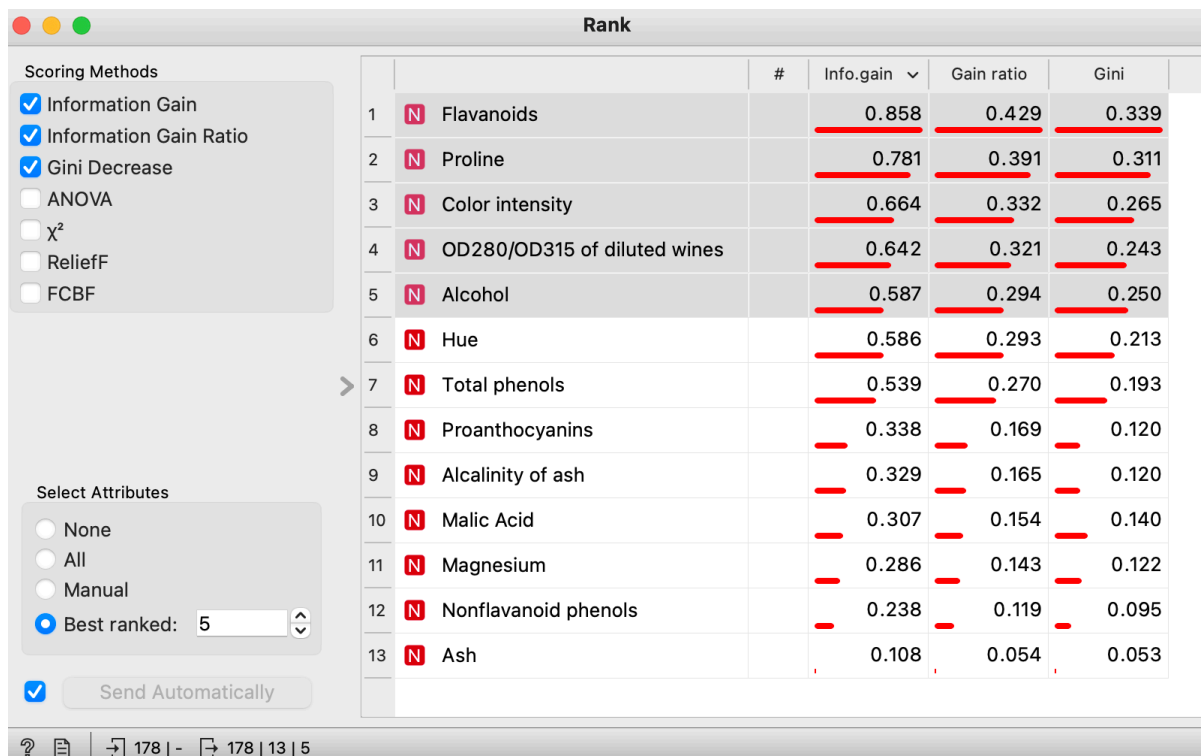
Porównując cechy za pomocą metryki F1 score (taka sama jak przy Random Forest) można zauważyć, że w przeciwieństwie do poprzedniego przykładu pierwsze trzy cechy nie mają aż takiego wpływu na spadek wyniku. Zmieniły się również najważniejsze cechy.

3. Explain Prediction:



Model SVM tak samo zakwalifikował pierwszy element zbioru jako klasę 1. Jednocześnie zawartość Proliny była najważniejszą cechą, jednak kwas jabłkowy (Malic Acid) został uznany za cechę obniżającą prawdopodobieństwo zaklasyfikowania danego elementu do szczepu 1.

Narzędzie Rank:



The screenshot shows the Rank tool interface. On the left, under 'Scoring Methods', the following are checked: Information Gain, Information Gain Ratio, and Gini Decrease. Under 'Select Attributes', 'Best ranked: 5' is selected. A 'Send Automatically' button is at the bottom left. The main table displays 13 features ranked by three metrics: Info.gain, Gain ratio, and Gini. The top three features are Flavanoids, Proline, and Color intensity.

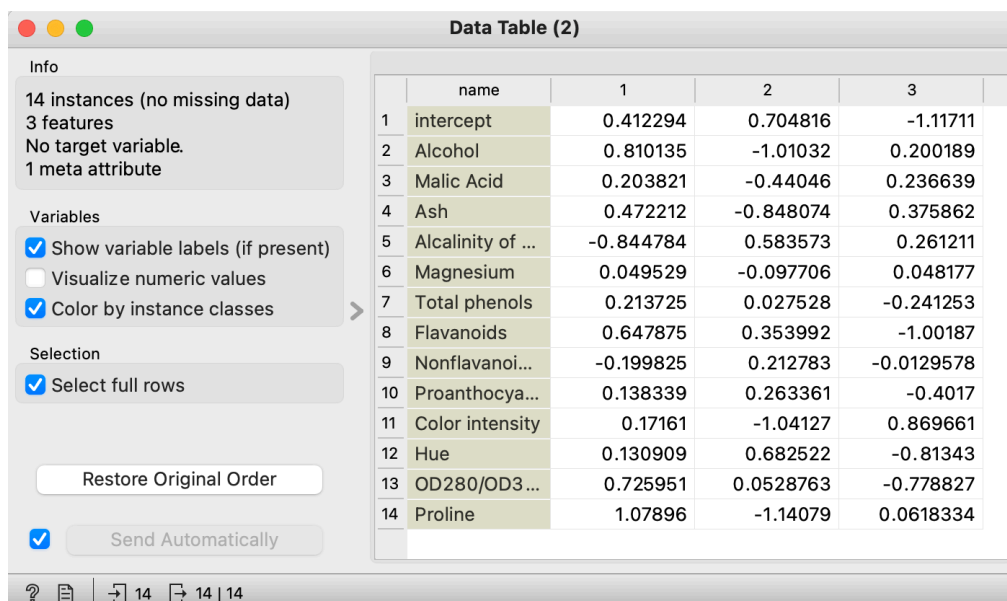
		#	Info.gain	Gain ratio	Gini
1	N Flavanoids		0.858	0.429	0.339
2	N Proline		0.781	0.391	0.311
3	N Color intensity		0.664	0.332	0.265
4	N OD280/OD315 of diluted wines		0.642	0.321	0.243
5	N Alcohol		0.587	0.294	0.250
6	N Hue		0.586	0.293	0.213
7	N Total phenols		0.539	0.270	0.193
8	N Proanthocyanins		0.338	0.169	0.120
9	N Alcalinity of ash		0.329	0.165	0.120
10	N Malic Acid		0.307	0.154	0.140
11	N Magnesium		0.286	0.143	0.122
12	N Nonflavanoid phenols		0.238	0.119	0.095
13	N Ash		0.108	0.054	0.053

Poza zastosowaniem metod opierających się o SHAP, zastosowałem również porównanie wartości cech za pomocą metryk współczynnika Giniego oraz Information Gain w narzędziu Rank.

Zawartość Flawonoidów, Proliny oraz intensywność koloru znów zostały uznane za najważniejsze cechy, jednak warto zauważyć, że zawartość alkoholu jest równie ważna.

Regresja Liniowa oraz Logistyczna (Linear and Logistic Regression)

1. Regresja logistyczna:



Info

14 instances (no missing data)
3 features
No target variable.
1 meta attribute

Variables

☒ Show variable labels (if present)
☐ Visualize numeric values
☒ Color by instance classes

Selection

☒ Select full rows

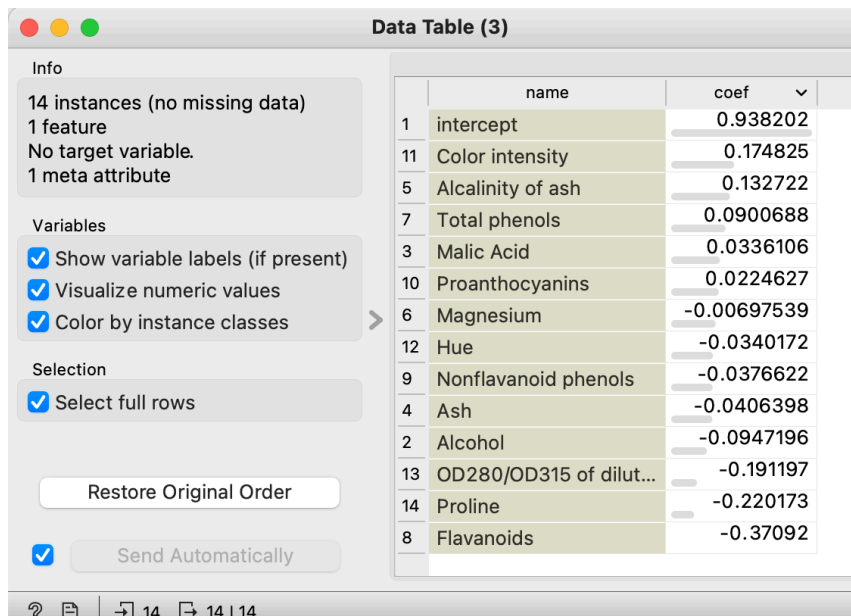
Restore Original Order

☒ Send Automatically

	name	1	2	3
1	intercept	0.412294	0.704816	-1.11711
2	Alcohol	0.810135	-1.01032	0.200189
3	Malic Acid	0.203821	-0.44046	0.236639
4	Ash	0.472212	-0.848074	0.375862
5	Alcalinity of ...	-0.844784	0.583573	0.261211
6	Magnesium	0.049529	-0.097706	0.048177
7	Total phenols	0.213725	0.027528	-0.241253
8	Flavanoids	0.647875	0.353992	-1.00187
9	Nonflavano...	-0.199825	0.212783	-0.0129578
10	Proanthocya...	0.138339	0.263361	-0.4017
11	Color intensity	0.17161	-1.04127	0.869661
12	Hue	0.130909	0.682522	-0.81343
13	OD280/OD3...	0.725951	0.0528763	-0.778827
14	Proline	1.07896	-1.14079	0.0618334

Metoda polega na przepuszczeniu danych przez Preprocessor, następnie przez narzędzie Feature Constructor. Wyznaczone współczynniki zostały zebrane dla każdego szczepu wina wraz z parametrem Intercept (przesunięciem).

2. Regresja liniowa



Info

14 instances (no missing data)
1 feature
No target variable.
1 meta attribute

Variables

☒ Show variable labels (if present)
☒ Visualize numeric values
☒ Color by instance classes

Selection

☒ Select full rows

Restore Original Order

☒ Send Automatically

	name	coef
1	intercept	0.938202
11	Color intensity	0.174825
5	Alcalinity of ash	0.132722
7	Total phenols	0.0900688
3	Malic Acid	0.0336106
10	Proanthocyanins	0.0224627
6	Magnesium	-0.00697539
12	Hue	-0.0340172
9	Nonflavanoid phenols	-0.0376622
4	Ash	-0.0406398
2	Alcohol	-0.0947196
13	OD280/OD315 of dilut...	-0.191197
14	Proline	-0.220173
8	Flavanoids	-0.37092

Model Regresji liniowej pozwala wyznaczyć wpływ każdej z cech, najważniejszą jest intensywność koloru, zasadowość popiołu w glebie oraz całkowita zawartość fenoli w glebie.