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:

## Algorytm Random Forest:

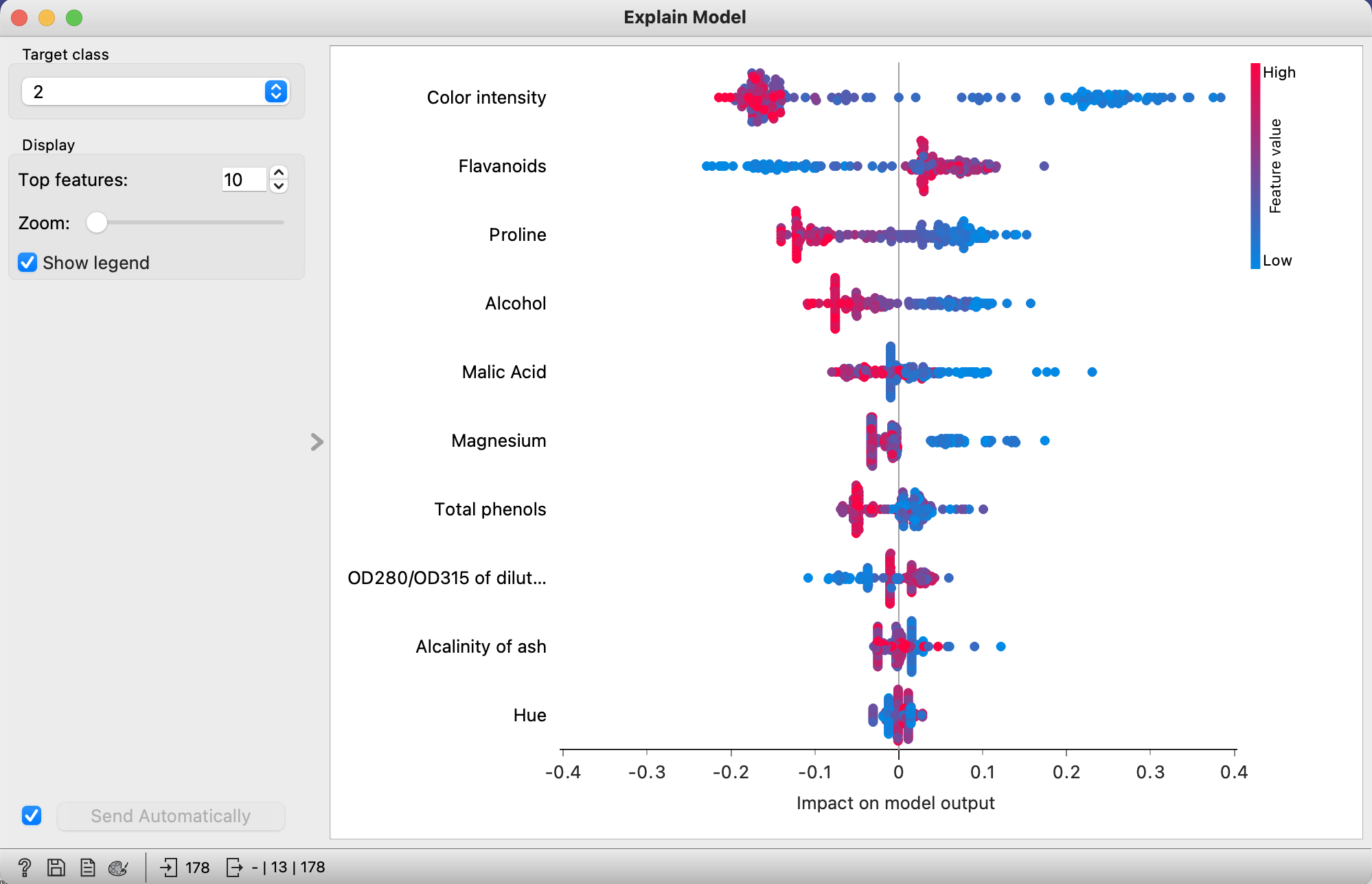
1. Wyjaśnienie modelu na przykładzie kategorii 1 za pomocą narzędzia Explain Model:

Graphical user interface

Description automatically generated

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:

Graphical user interface

Description automatically generated

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

1. Wyjaśnienie ważności cech za pomocą narzędzia Feature Importance:

Graphical user interface

Description automatically generated

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.

1. Wyjaśnienie predykcji za pomocą Explain Prediction

Predykcja dla pierwszej instancji w zestawie danych.

Graphical user interface, application

Description automatically generated

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

Klasa 2:

Graphical user interface, application

Description automatically generated

Klasa 3:

Graphical user interface, application

Description automatically generated

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.

## Model SVM:

1. Wyjaśnianie modelu za pomocą Explain Model:

Graphical user interface

Description automatically generated

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:

Graphical user interface, application

Description automatically generated

Klasa 3:

Graphical user interface

Description automatically generated

1. Feature Importance

Chart, bar chart

Description automatically generated

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 wpływy na spadek wyniku. Zmieniły się również najważniejsze cechy.

1. Explain Prediction:

Graphical user interface, application

Description automatically generated

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 prawdopobieństwo zaklasyfikowania danego elementu do szczepu 1.

# Narzędzie Rank:

Graphical user interface, application

Description automatically generated

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)

## Regresja logistyczna:

Graphical user interface, application, table

Description automatically generated

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).

## Regresja liniowa

Graphical user interface, application

Description automatically generated

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