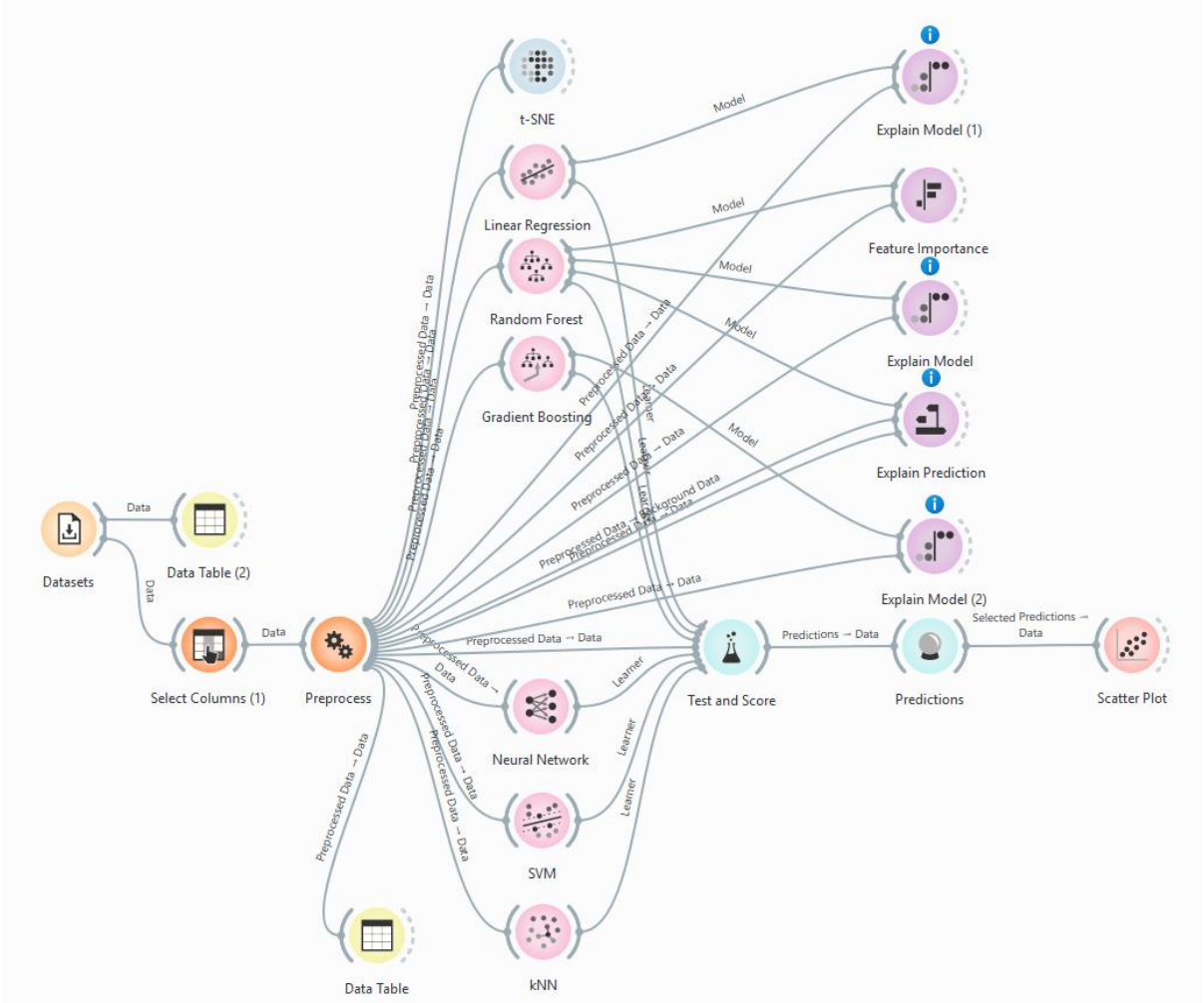
Uczenie maszynowe – Interpretowalność modeli (SHAP)

Zbiór danych: Wine Quality (Red)
Autor: Mateusz Łopaciński
Data: 18.10.2025

1. Wprowadzenie

W analizie wykorzystano zbiór **Wine Quality – Red**, zawierający **1599 próbek wina** opisanych przez **11 cech fizykochemicznych**, w tym m.in. zawartość alkoholu, kwasowość i poziom siarczanów. Zmienną docelową jest **ocena jakości w skali od 3 do 8 punktów**.

Celem analizy było porównanie dokładności i interpretowalności kilku modeli regresyjnych z wykorzystaniem metody **SHAP (SHapley Additive Explanations)**, która pozwala określić, w jakim stopniu poszczególne cechy wpływają na wynik predykcji. Takie podejście umożliwia wgląd w sposób, w jaki model podejmuje decyzje.



Rysunek 1. Schemat analizy przeprowadzonej w środowisku Orange.

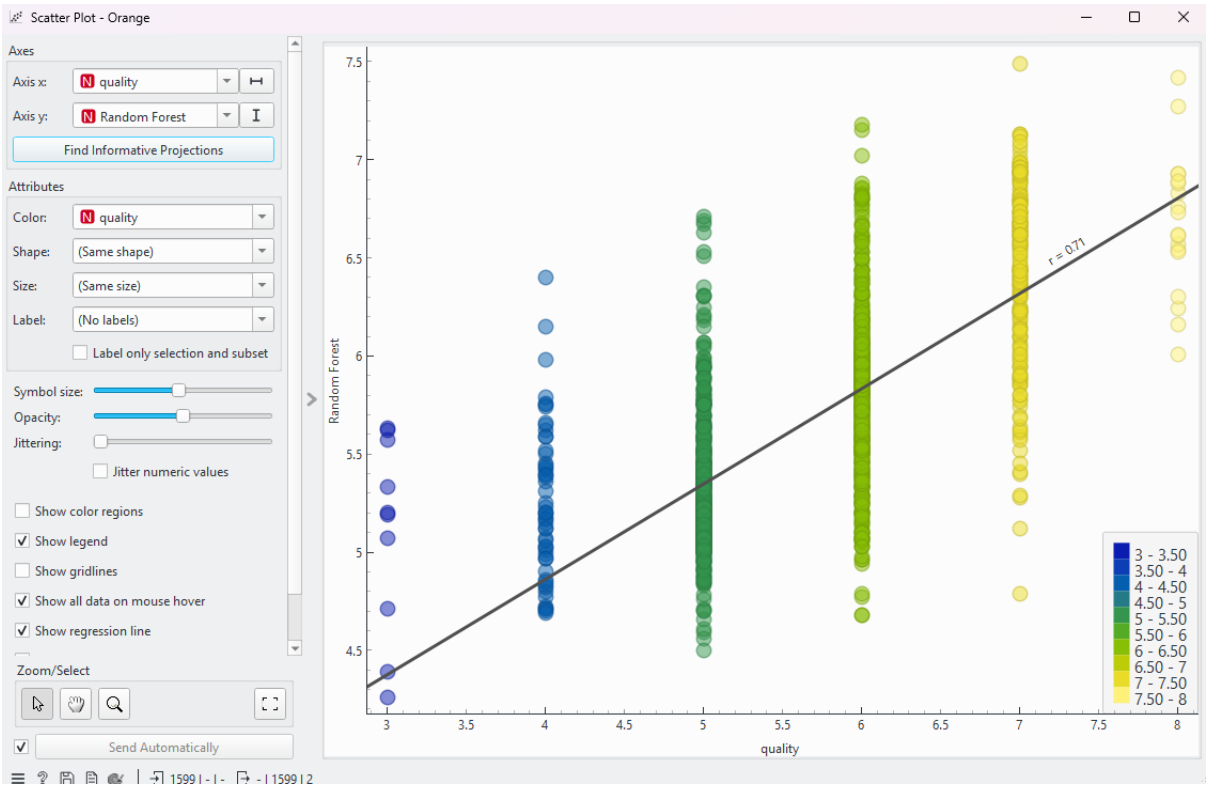
2. Modelowanie i ocena wyników

W analizie przetestowano sześć modeli regresyjnych: **regresję liniową**, **Random Forest**, **Gradient Boosting**, **sieć neuronową**, **SVM** oraz **k-NN**. Dokładność modeli oceniono metodą 10-krotnej walidacji krzyżowej, wykorzystując miary **MSE**, **MAE** oraz **R^2.**

Najlepsze wyniki uzyskał model **Random Forest**, osiągając **R^2 = 0,50 i MSE = 0,32**.

Regresja liniowa oraz **SVM** uzyskały umiarkowaną skuteczność (**R^2 ≈ 0,35–0,39**), natomiast **Gradient Boosting** w domyślnej konfiguracji osiągnął bardzo niskie **R^2** (**≈0,05**), co świadczy o niedostatecznym dopasowaniu modelu do danych.

Poniższy wyjkres pokazuje, że model Random Forest dobrze oddaje ogólny trend między rzeczywistą a przewidywaną jakością (**r = 0,71**), choć widoczne jest pewne rozproszenie punktów, świadczące o umiarkowanej dokładności predykcji.

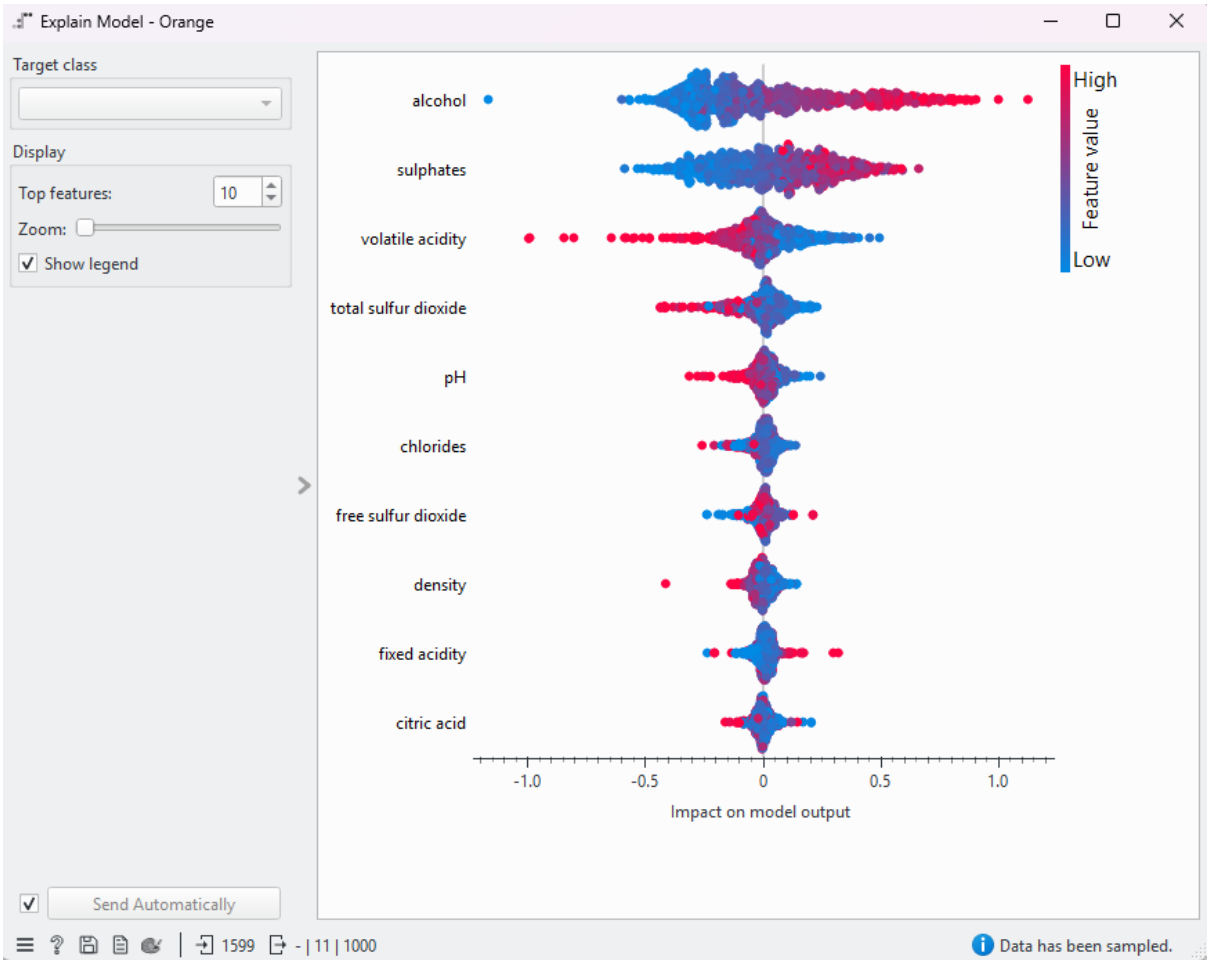


Rysunek 2. Porównanie rzeczywistych i przewidywanych wartości jakości wina dla modelu Random Forest

3. Modelowanie i ocena wyników

Analiza wartości **SHAP** dla modelu **Random Forest** wskazuje, że największy wpływ na przewidywaną jakość mają trzy cechy: **alcohol**, **sulphates** oraz **volatile acidity**. Wyższa zawartość alkoholu i siarczanów sprzyja wyższej ocenie jakości, natomiast większa kwasowość lotna obniża przewidywaną ocenę. Pozostałe zmienne, takie jak *total sulfur dioxide*, *chlorides* czy *pH*, mają mniejsze, ale zauważalne znaczenie.

Wyniki są spójne z wiedzą chemiczną — lepsze wina charakteryzują się wyższym poziomem alkoholu, umiarkowaną kwasowością i odpowiednim poziomem siarczanów. Model Random Forest dobrze uchwycił te zależności, co potwierdza, że nauczył się istotnych, nieliniowych relacji pomiędzy cechami a oceną jakości.

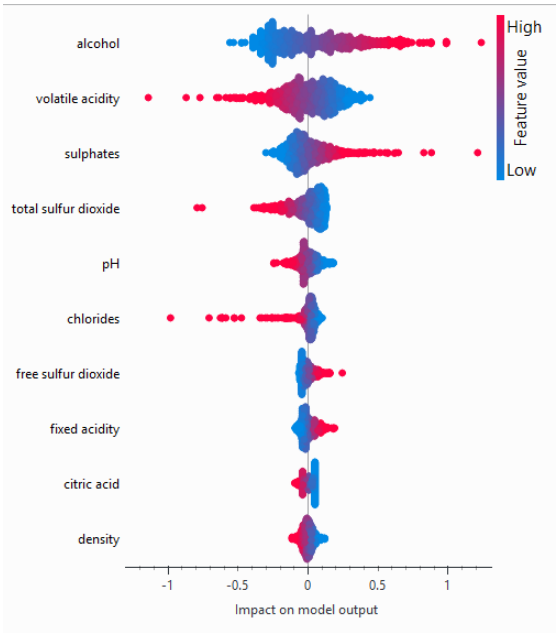


Rysunek 3. Wykres wpływu cech na wynik predykcji modelu Random Forest (Explain Model)

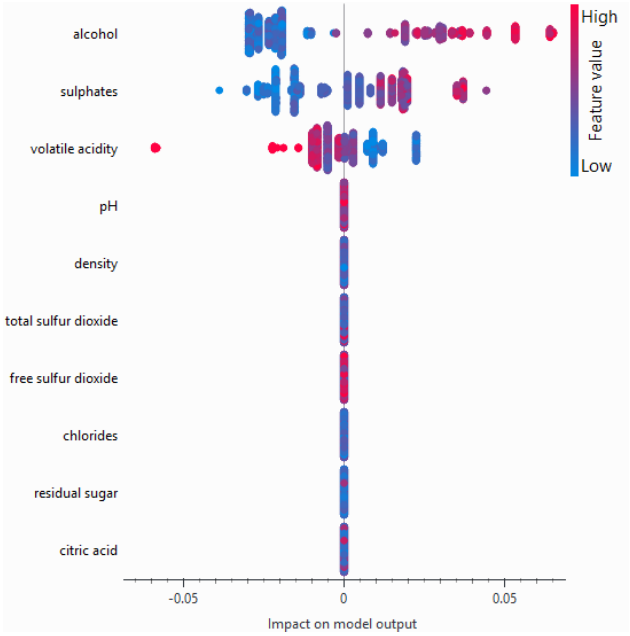
Porównanie z innymi modelami pokazuje różnice w sposobie odwzorowywania wpływów:

- **Regresja liniowa** wykazała proste, monotoniczne zależności między cechami a wynikiem, bez efektów nieliniowych ani interakcji.
- **Gradient Boosting** uzyskał bardzo niskie wartości SHAP, co wskazuje, że nie zdołał uchwycić istotnych zależności — zgodnie z niskim R^2 obserwowanym w ewaluacji. Poziome przerwy na wykresie (*Rysunek 5*) wynikają z dyskretnego charakteru decyzji modelu Gradient Boosting, który przypisuje zbliżone wartości predykcji wielu obserwacjom należącym do tych samych liści drzew. Zjawisko to odzwierciedla prostą strukturę modelu i ograniczoną liczbę rozpoznanych zależności w danych.

Wraz ze wzrostem jakości modelu wartości SHAP stają się bardziej wyraziste i spójne, co świadczy o trafnym odwzorowaniu zależności w danych.



Rysunek 4. Explain Model - Regresja Liniowa



Rysunek 5. Explain Model – Gradient Boosting

4. Interpretacja lokalna

Analiza lokalna pozwala wyjaśnić, w jaki sposób model **Random Forest** uzasadnia przewidywania dla pojedynczych przypadków. Wykorzystano narzędzie *Explain Prediction*, które przedstawia wpływ poszczególnych cech na wynik predykcji dla wybranego wina (*Rysunek 6*).

Model rozpoczął obliczenia od wartości bazowej **5,59**, odpowiadającej średniej jakości w całym zbiorze danych. Po uwzględnieniu wartości cech chemicznych przewidywana ocena jakości wyniosła **5,01**, co oznacza, że analizowana próbka została oceniona jako nieco słabsza od przeciętnej.

Na **obniżenie** wyniku największy wpływ miały:

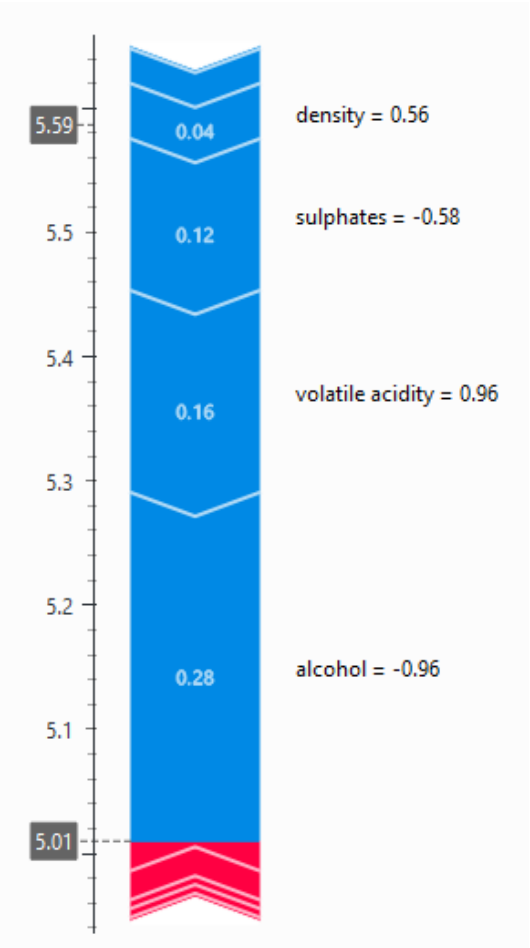
- **niska zawartość alkoholu (−0,96)**,
- **wysoka kwasowość lotna (+0,96)**,
- **wyższa gęstość (−0,56)**,
- **umiarkowany poziom siarczanów (−0,58)**.

Natomiast niewielki, korzystny wpływ miały cechy widoczne na przybliżonym wykresie (*Rysunek 7*):

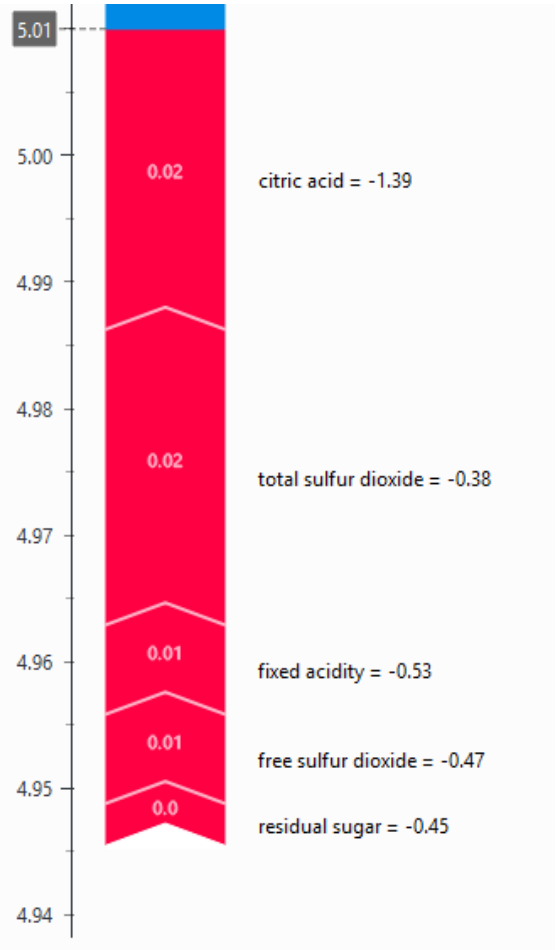
- **cytryniany (−1,39)**,
- **całkowity dwutlenek siarki (−0,38)**,
- **kwasowość stała (−0,53)**,

- **wolny dwutlenek siarki (−0,47)**,
- **zawartość cukru (−0,45)**.

Sumarycznie negatywne oddziaływanie głównych cech (niski alkohol, wysoka kwasowość, większa gęstość) spowodowało spadek przewidywanej jakości w stosunku do średniej wartości modelu. Wynik jest spójny z wiedzą ekspercką — niska zawartość alkoholu i większa kwasowość są typowe dla win o niższej jakości.



Rysunek 6. Explain Prediction – Znaczące cechy



Rysunek 7. Explain Prediction – Mniej znaczące cechy

5. Wnioski końcowe

Przeprowadzona analiza potwierdziła, że model **Random Forest** osiągnął najwyższą skuteczność w przewidywaniu jakości czerwonego wina spośród wszystkich porównywanych algorytmów.

Zastosowanie metody **SHAP** pozwoliło wskazać kluczowe czynniki wpływające na wynik predykcji – **zawartość alkoholu**, **siarczany** oraz **kwasowość lotną**. Wpływy te są zgodne z wiedzą chemiczną, co potwierdza wiarygodność modelu i poprawność odwzorowania zależności między cechami a oceną jakości.

Porównanie z **regresją liniową** i **Gradient Boosting** wykazało, że modele o wyższej skuteczności predykcyjnej generują bardziej wyraźne i spójne rozkłady wartości SHAP, co przekłada się na większą interpretowalność ich wyników. Analiza SHAP pozwoliła nie tylko ocenić jakość modeli, ale również zrozumieć, które cechy w największym stopniu kształtują decyzje modelu przy przewidywaniu jakości wina.