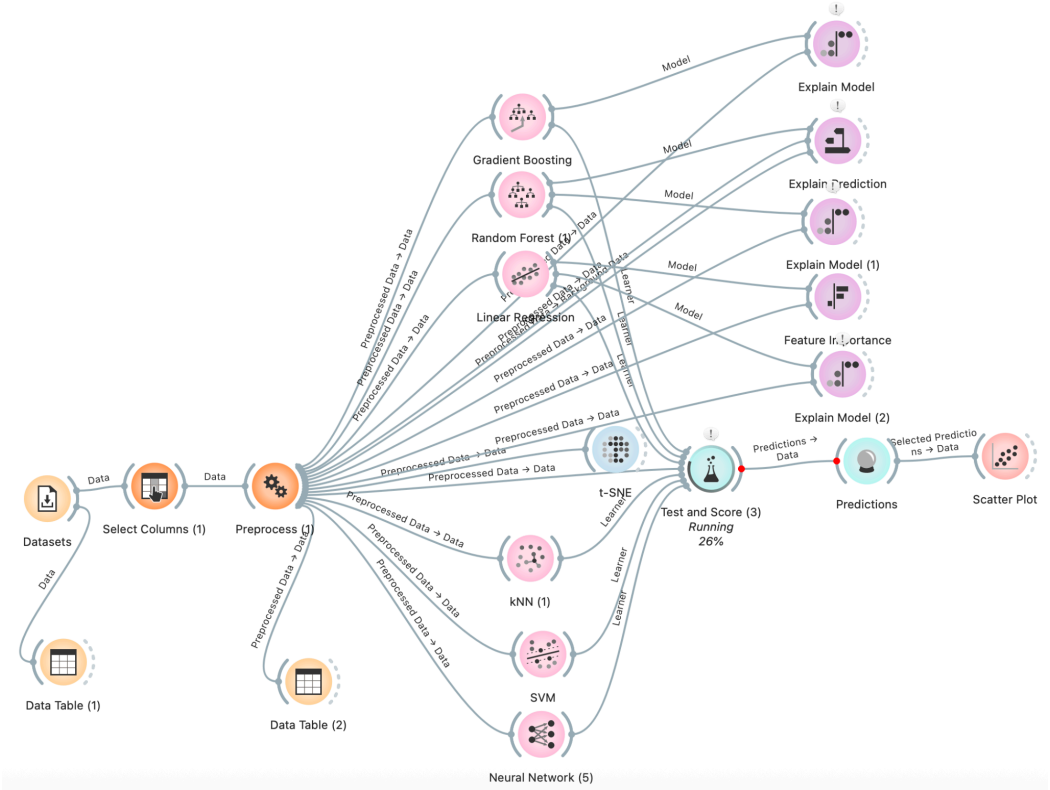
Juliusz Wasieleski, grudzień 2025

Uczenie Maszynowe lab 1 sprawozdanie

1. Wprowadzenie

W analizie wykorzystano zbiór **California Housing**, zawierający 20640 wpisów o domach w Kalifornii, opisanych przez 8 cech, takich jak liczba pokoi, wiek doomu, liczba sypialni, populacja miasta itp. Zmienną docelową jest ocena warto≤ś≤ci domu

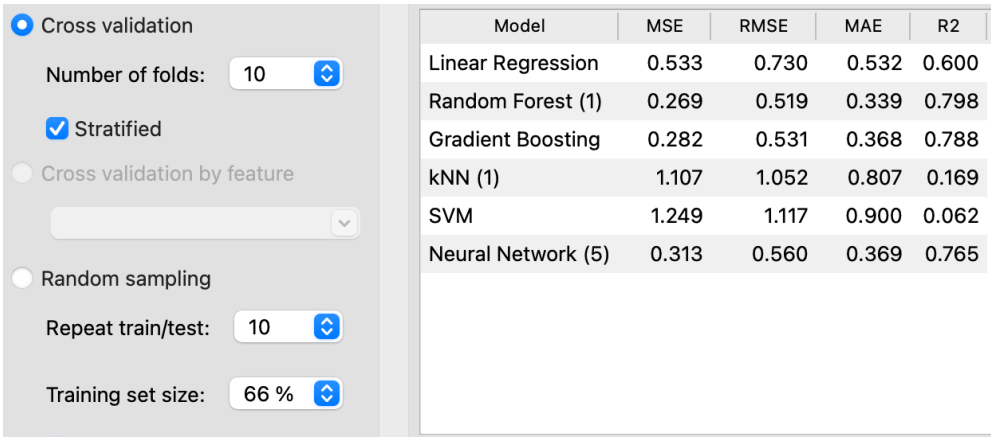
Celem analizy było porównanie dokładności i interpretowalności kilku modeli regresyjnych z wykorzystaniem metody SHAP (SHapley Additive Explanations), która pozwala określić, w jakim stopniu poszczególne cechy wpływają na wynik predykcji.Takie podejście umożliwia wgląd w sposób, w jaki model podejmuje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



Rys. 1. Schemat analizy przeprowadzonej w środowisku Orange.

2. Modelowanie i ocena wyników

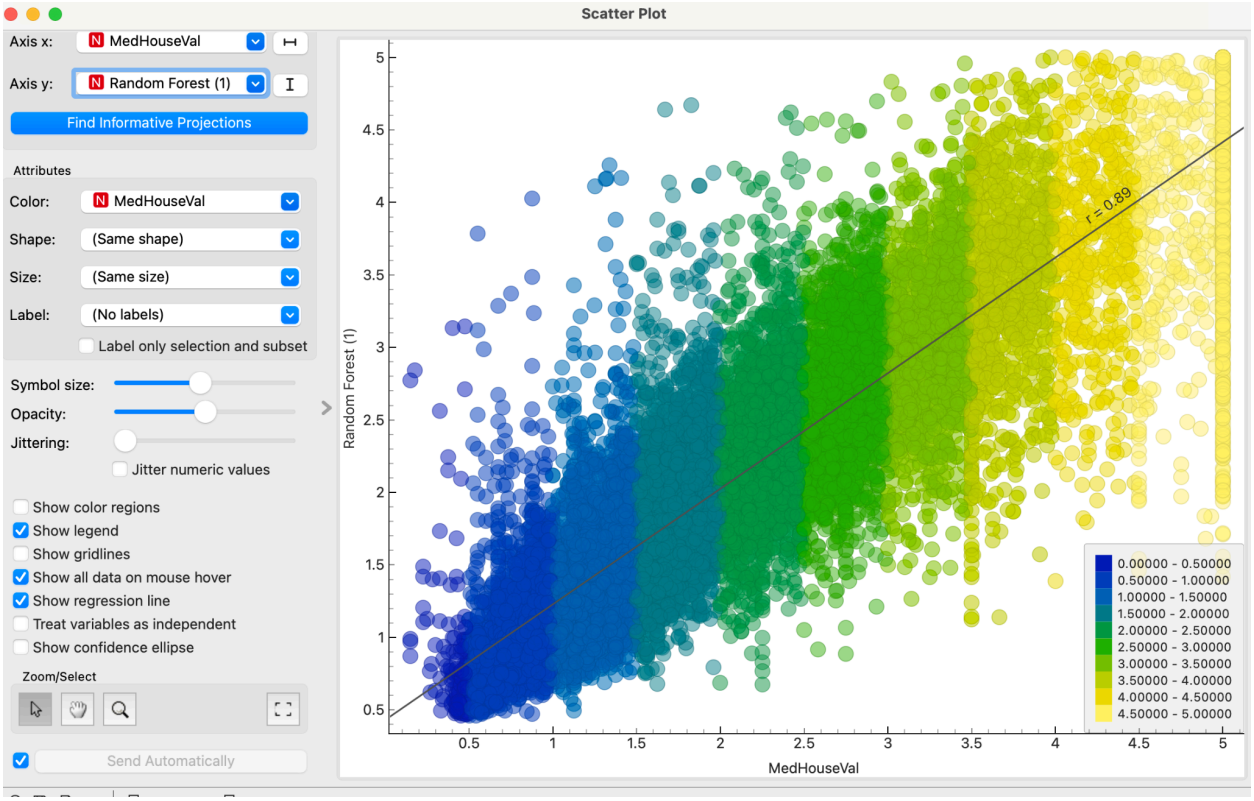
W analizie przetestowano sześć modeli regresyjnych: regresję liniową, Random Forest, Gradient Boosting, sieć neuronową, SVM oraz k-NN. Dokładność modeli oceniono metodą 10-krotnej walidacji krzyżowej, wykorzystując miary MSE, RMSE, MAE oraz R^2.



Model	MSE	RMSE	MAE	R2
Linear Regression	0.533	0.730	0.532	0.600
Random Forest (1)	0.269	0.519	0.339	0.798
Gradient Boosting	0.282	0.531	0.368	0.788
kNN (1)	1.107	1.052	0.807	0.169
SVM	1.249	1.117	0.900	0.062
Neural Network (5)	0.313	0.560	0.369	0.765

Tab. 1 Wyniki predykcji

Najlepsze wyniki uzyskał **Random Forest** R^2 = 0,798 i **MSE** = 0,269. Nie dużo gorsze wyniki dały odpowiednio: **Gradient Boosting**, **sieć neuronowa** i **regresja liniowa** (zakres 0,788 < R^2 < 0,6). Za to **kNN** i **SVM** uzyskały bardzo niskie skuteczności (R^2 < 0,2)

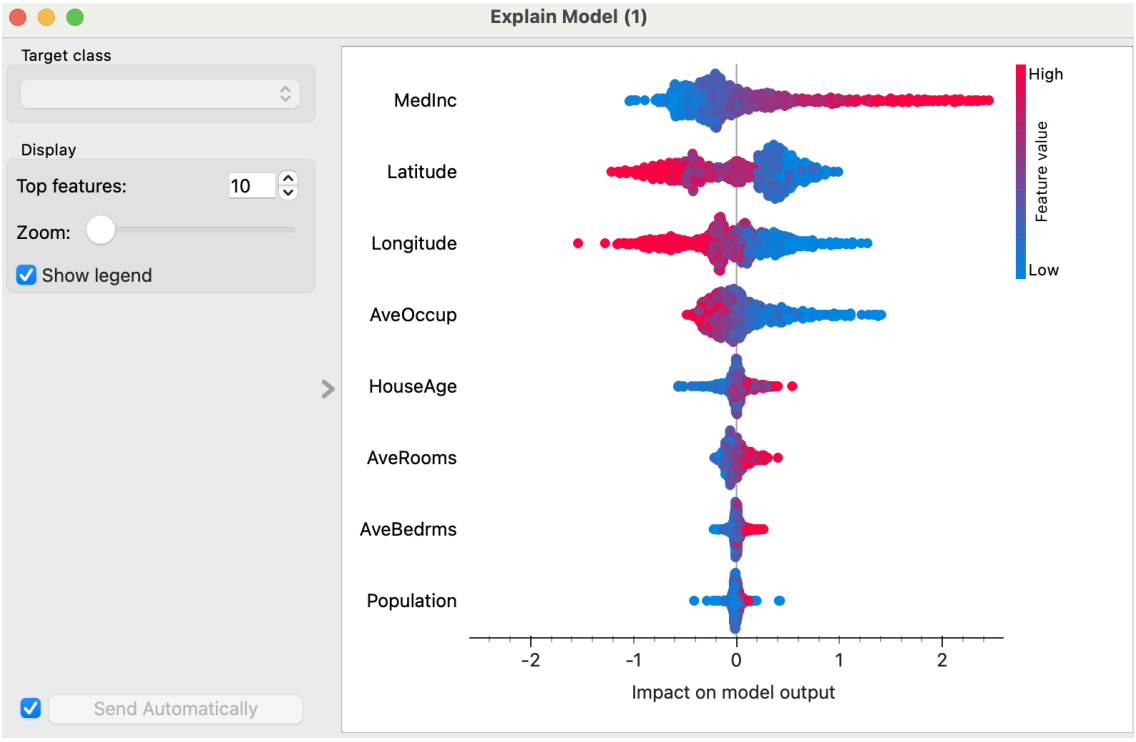


Rysunek 2. Porównanie rzeczywistych i przewidywanych wartości jakości wina dla modelu Random Forest

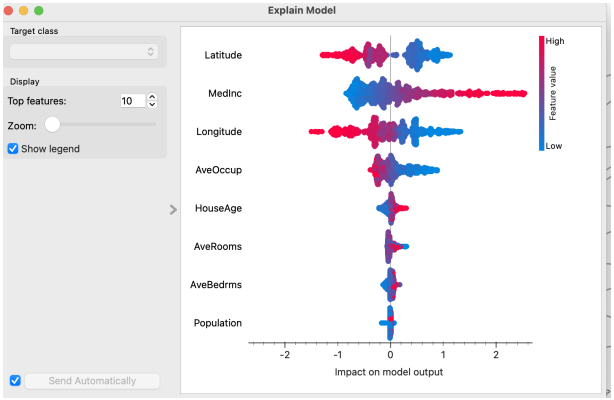
3. Analiza SHAP

Analiza wartości **SHAP** dla modelu **Random Forest** wskazuje, że największy wpływ na przewidywaną jakość mają trzy cechy: Latitude, Longitude oraz MedInc (Medium Income). Położenie geograficzne oraz średni dochód mają duży wpływ na wartość domu (położenie bardziej na północny zachód, tym niższa wartość domu, średni dochód im niższy tym niższa wartość domu) Pozostałe zmienne, a głównie liczba domowników jak i populacja miasta nie mają zbyt dużego wpływu.

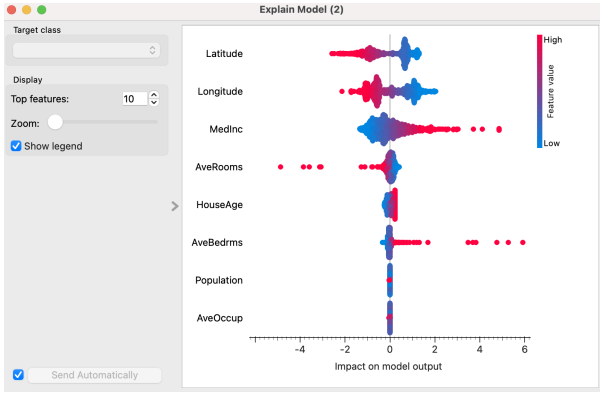
Wyniki są spójne z geografią Kaliforni, gdzie na południu i raczej we wschodniej części jest Dolina Krzemowa oraz Los Angeles) .



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



Rys. 4 SHAP dla Gradient Boosting

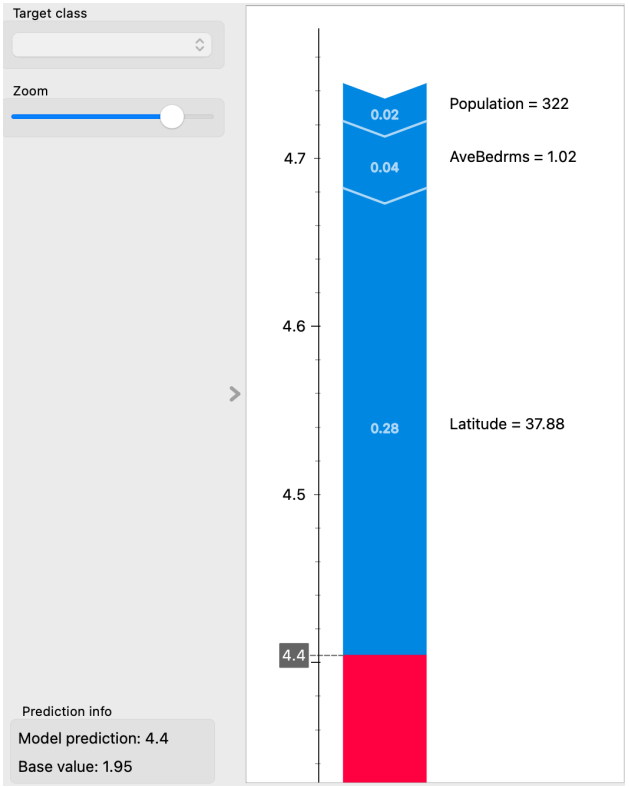


Rys. 5 SHAP dla Regresji Liniowej

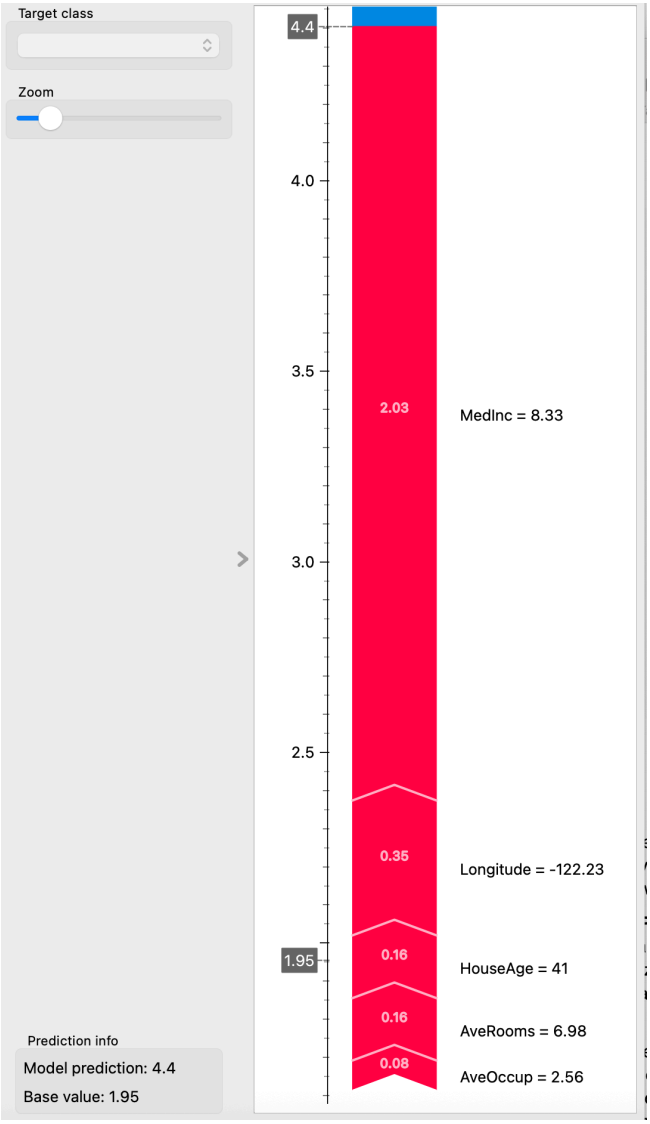
Porównanie z innymi modelami pokazuje różnice w sposobie odwzorowywania wpływów:

- Regresja liniowa wykazała prostsze, bardziej monotoniczne zależności między cechami a wynikiem, z mniejszymi efektami nieliniowymi.
- Gradient Boosting uzyskał dość dobre wartości SHAP, co wskazuje, że zdołał uchwycić istotne zależności — zgodnie z wysokim R^2 obserwowanym w ewaluacji.
- Wszystkie wartości na wykresach SHAP są ciągłe co wynika z ciągłego charakteru predykcji (nie estymujemy klas tylko wartość z przedziału, który można traktować jako ciągły)
- Wraz ze wzrostem jakości modelu wartości SHAP stają się bardziej wyraziste i spójne, co świadczy o trafnym odwzorowaniu zależności w danych.

4. Interpretacja lokalna



Rys 6. Explain Prediction – Znaczące cechy



Rys 7. Explain Prediction – Mniej znaczące cechy

Analiza lokalna pozwala wyjaśnić, w jaki sposób model Random Forest uzasadnia przewidywania dla pojedynczych przypadków. Wykorzystałem narzędzie Explain Prediction, które przedstawia wpływ poszczególnych cech na wynik predykcji dla wybranego wina (Rysunek 7).

Model rozpoczął obliczenia od wartości bazowej 1,95, odpowiadającej średniej jakości w całym zbiorze danych. Po uwzględnieniu cech domu, jego wartość wyniosła 4,4, co oznacza, że analizowana przypadek został oceniony lepiej od przeciętnego.

Na podwyższenie wyniku największy wpływ miały:

- **Średni dochód** (8,33),
- **Wysokość geograficzna** (-122 stopnie),
- **Wiek domu** (41 lat),
- **Średnia liczba pokoi** (6,98),
- **Średnia liczba lokatorów**.

Natomiast niewielki, niekorzystny wpływ miały cechy widoczne na przybliżonym wykresie (Rysunek 6):

- **Szerokość geograficzna** (37 stopni),
- **Średnia liczba pokoi**(1,22),
- **Liczba mieszkańców miejscowości** (322).

Wybrany dom leży na wschód od Oakland w północnej części stanu, gdzie domy są ogólnie tańsze. Dużo na plus dla tego domu działa średni dochód, który poprawił wynik o ponad 2 (z 1.95 do 3,98)

5. Wnioski

Przeprowadzona analiza potwierdziła, że model **Random Forest** osiągnął najwyższą skuteczność w przewidywaniu wartości domu spośród wszystkich porównywanych algorytmów.

Zastosowanie metody **SHAP** pozwoliło wskazać kluczowe czynniki wpływające na wynik predykcji – położenie geograficzne domu i średni dochód. Wpływy te są zgodne z geografią Kalifornii oraz uwarunkowaniami socjologicznymi 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.