Przedmiot: Uczenie Maszynowe

Laboratorium: 1

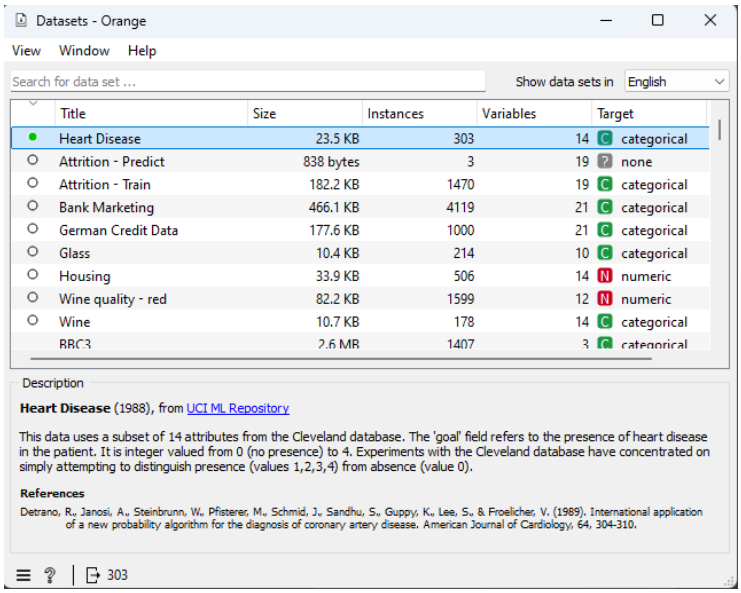
Kierunek: Informatyka – Data Science

Autor: Bartłomiej Jamiołkowski

Do realizacji zadania ‘ZADANIE LAB FINALNE’ został wykorzystany zbiór danych ‘Heart Disease’ pobrany z repozytorium online programu Orange. Jego wybór umotywowany jest występowaniem w nim ciekawych i zróżnicowanych przykładów, co ukazują zamieszczone wykresy widgetów ‘ICE’.

Omawiany zbiór danych skupia się on na problemie klasyfikacji pacjentów pod względem ryzyka występowania u nich chorób serca. W klasyfikacji wykorzystywane są cechy medyczne a także wyniki badań pacjentów. Domyślną zmienną kategoryczną jest ‘diameter narrowing’ oznaczająca zwężenie naczyń krwionośnych. Przyjmuje ona wartości 0 – brak zwężenia naczyń lub 1 – zwężenie naczyń co jest przesłanką do stwierdzenia występowania choroby serca.

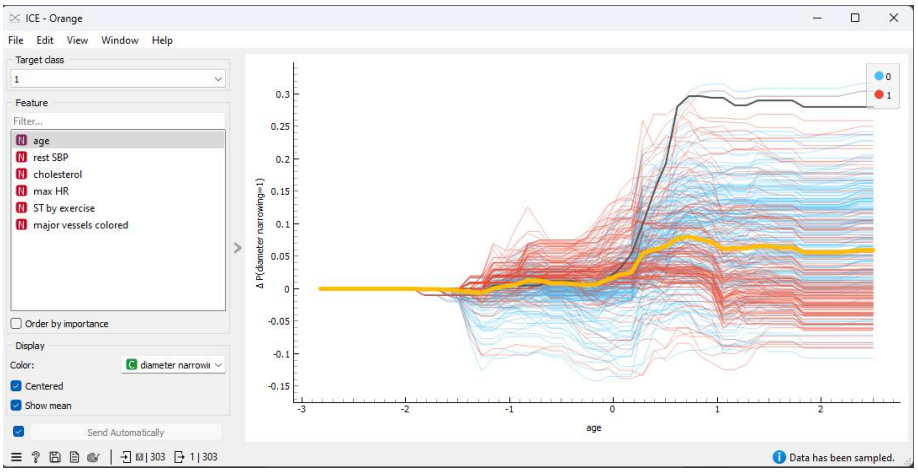
Wspomniany zestaw danych składa się z 303 obserwacji i 14 zmiennych. Został on podzielony na zbiór uczący (70% danych) i testowy (30% danych) za pomocą widgetu ‘Test and Score’.



Rysunek 1 Pobranie zbioru danych Heart Disease z repozytorium online

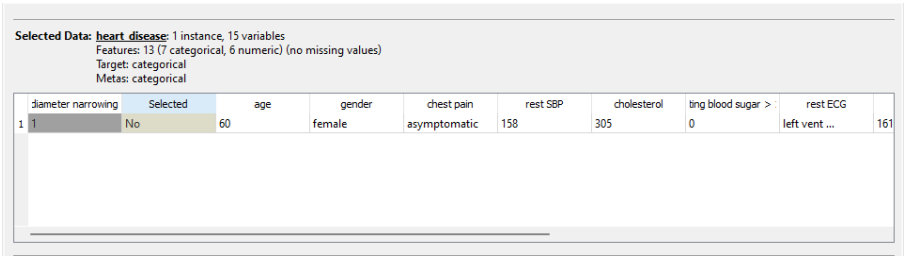
Wybór interesujących dwóch przykładów w pierwszej kolejności opiera się na analizie wykresów z widgetów ‘ICE’, gdzie analizowaną zmienną był wiek (z ang. age). Domyślnie

był do nich podpięty model ‘Random Forest’ z 100 drzew jako parametrem. Według zaleceń z laboratorium pożądane są zróżnicowane przykłady znacznie odstające od reszty. Poniższy rysunek ukazuje pierwszy zaznaczony przykład.

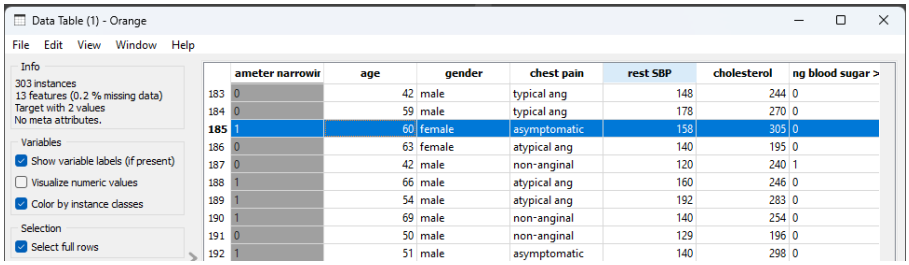


Rysunek 2 Wybór przykładu nr 1 za pomocą widgetu ICE dla modelu Random Forest

Po oznaczeniu przykładu zostały wyświetlone informacje o nim w celu identyfikacji jego położenia w tabeli danych.

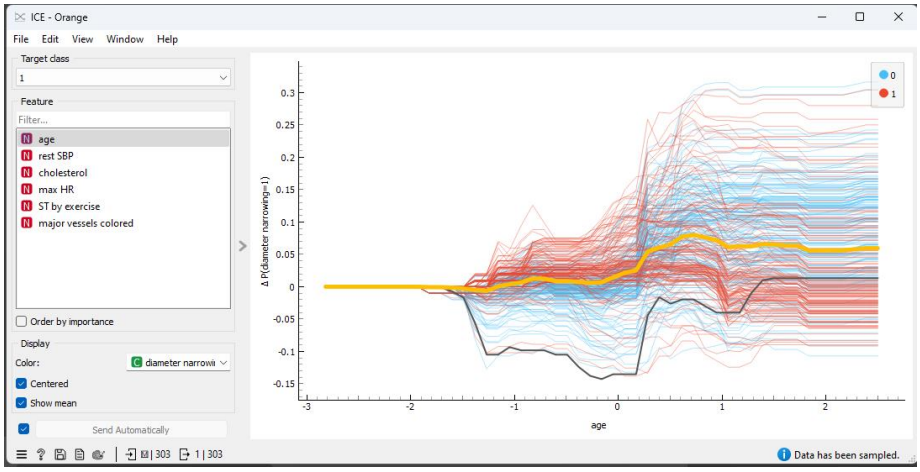


Rysunek 3 Szczegóły wybranego przykładu z wykresu ICE

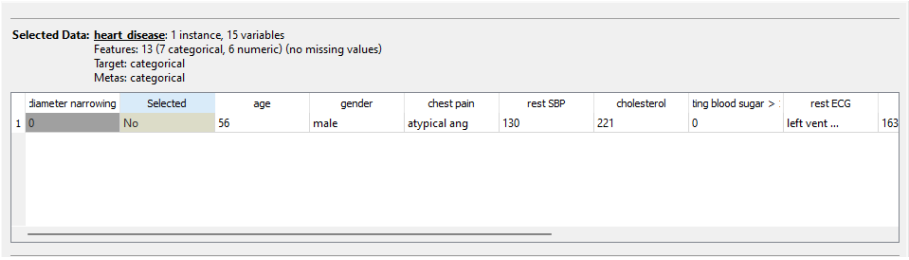


Rysunek 4 Wybór pierwszego przykładu ze zbioru

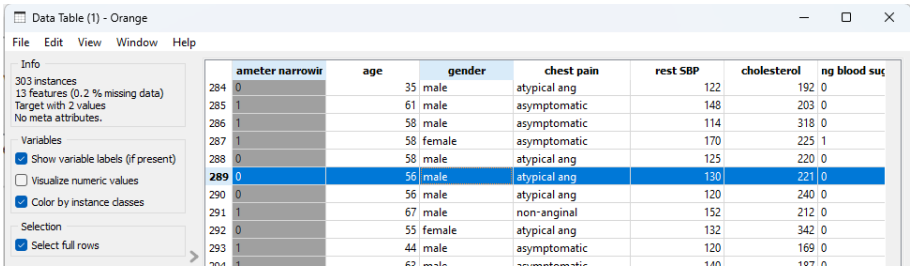
Podobnie postąpiono podczas wyboru drugiego przykładu. W tym wypadku zamiast obserwacji, gdzie ‘diameter narrowing’ = 1, wybrano obserwację należącą do klasy 0.



Rysunek 5 Wybór przykładu nr 2 za pomocą widgetu ICE dla modelu Random Forest

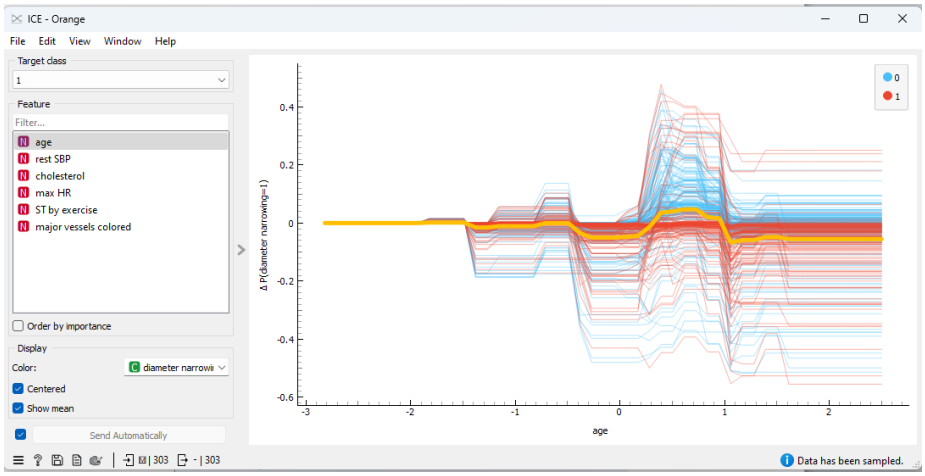


Rysunek 6 Szczegóły wybranego przykładu z wykresu ICE

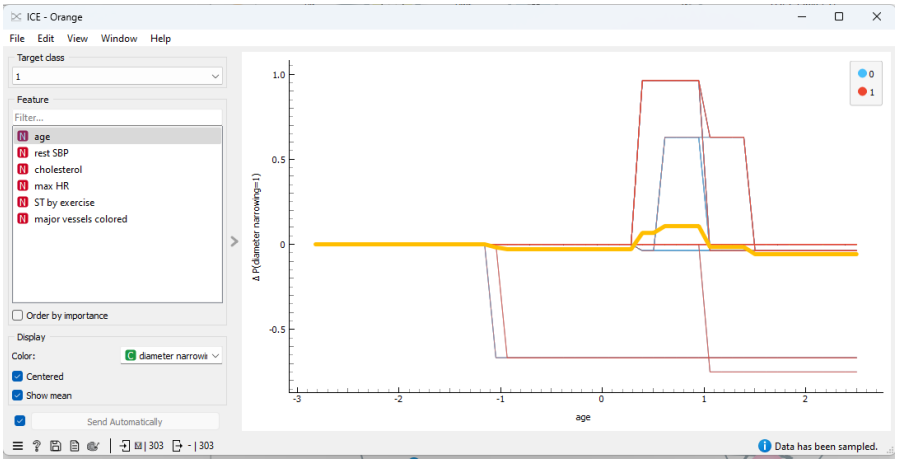


Rysunek 7 Wybór drugiego przykładu ze zbioru

Dodatkowo zamieszone są dwa wykresy z widgetu ‘ICE’ pokazujące jak predykcje przynależności obserwacji do poszczególnych klas zmieniały się wraz ze zmianą wartości zmiennej ‘age’ odpowiednio dla modeli ‘Gradient Boosting’ z 100 drzew jako parametrem i ‘Decision Tree’. Ponieważ wykresy bazujące na modelu Ranodm Forest wydawały się najbardziej rozbudowane, to z na nich został oparty wybór przykładów.

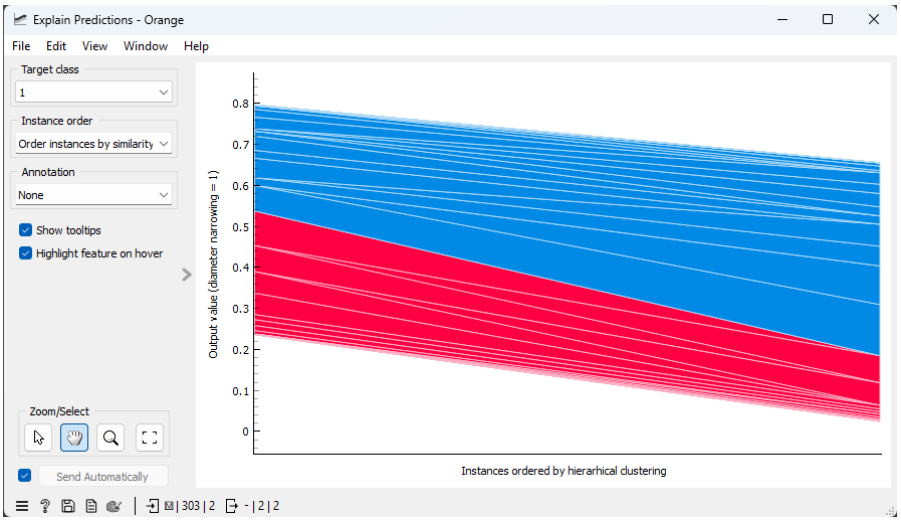


Rysunek 8 Wykres z widgetu ICE dla modelu Gradient Boosting



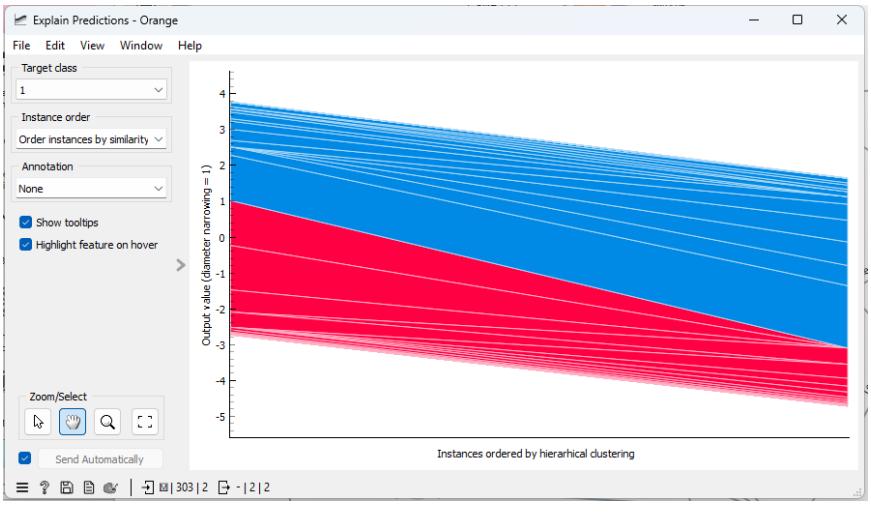
Rysunek 9 Wykres z widgetu ICE dla modelu Decision Tree

Kolejną czynnością w zadaniu było zwizualizowanie wartości SHAP dla każdej cechy dwóch analizowanych przykładów za pomocą widgetu ‘Explain Predictions’.

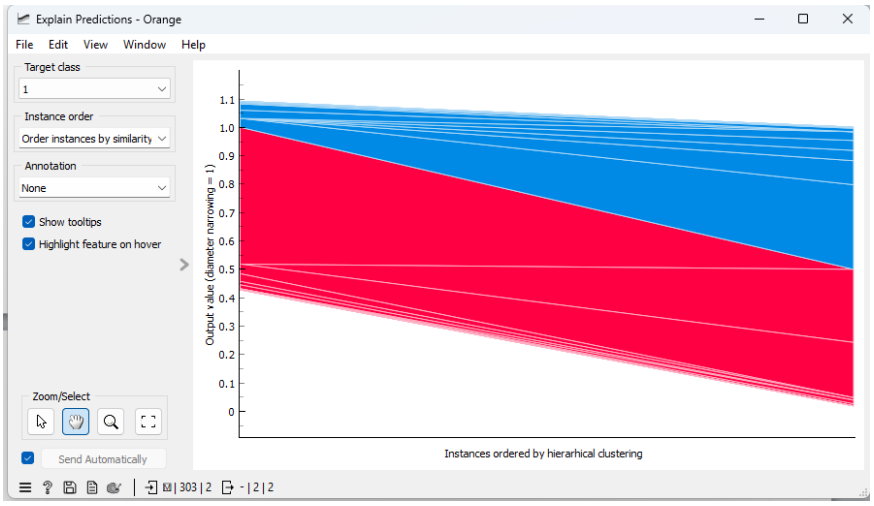


Rysunek 10 Wartości SHAP dla każdej cechy w widgecie Explain Predictions dla modelu Random Forest

Na zamieszczonym powyżej rysunku widać wartości SHAP dla każdej cechy związanej z dwoma przykładami. Po lewej stronie jest przykład osoby, u której stwierdzono zwężenie naczyń krwionośnych (dominuje kolor czerwony). Odpowiednio po prawej stronie znajduje się przykład osoby bez tego objawu choroby. Im bardziej wartości SHAP są oddalone od 0 w przypadku konkretnej zmiennej, tym większy wpływ ma ta zmienna na predykcję danej klasy. Analizując interaktywną wizualizację można dojść do wniosku, że największy wpływ na to, że dana osoba została zaklasyfikowana do grupy osób ze zwężonymi naczyniami miały wpływ czynniki takie jak: zaawansowany wiek, występujące bóle w klatce piersiowej i podwyższony cholesterol. W kontraście na fakt, że osoba została uznana za zdrową miały wpływ zmienne takie jak: częstsze zabarwienie odczynnika, brak bólu w klatce piersiowej, czyniski poziom odcinka ST w teście wysiłkowym.



Rysunek 11 Wartości SHAP dla każdej cechy w widgecie Explain Predictions dla modelu Gradient Boosting

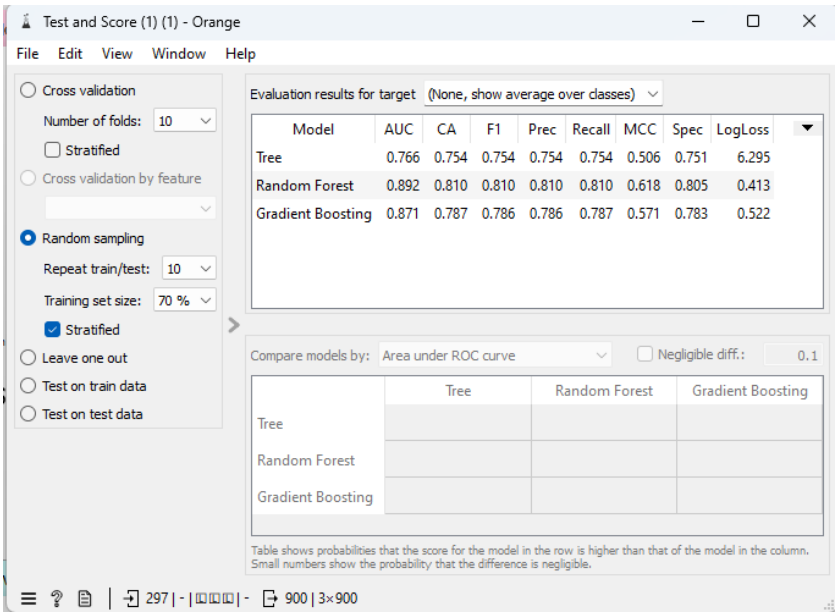


Rysunek 12 Wartości SHAP dla każdej cechy w widgecie Explain Predictions dla modelu Decision Tree

Powyżej znajdują się wykresy wartości SHAP dla modeli Gradient Boosting oraz Decision Tree. O ile wartości SHAP poszczególnych cech się zmieniają o tyle można zauważyć, że Decision Tree przypisuje duże wartości SHAP dla pojedynczych zmiennych, podczas gdy w modelach Random Forest i Gradient Boosting wartości SHAP dla poszczególnych zmiennych są bardziej podobne.

Można wnioskować, że model Decision Tree podejmuje decyzje na podstawie kluczowych cech, co może być przydaten w kontekście prostych problemów. W przypadku bardziej złożonych problemów ten model może dać gorsze rezultaty niż modele Random Forest i Gradient Boosting. Oba wymienione modele uwzględniają wpływ większej liczby zmiennych, co może poprawić dokładność predykcji klas.

Kolejnym etapem zadania jest sprawdzenie skuteczności predykcji poszczególnych modeli. Do przetestowania tego aspekyu wykorzystywany jest widget ‘test and Score’ zamieszony poniżej.

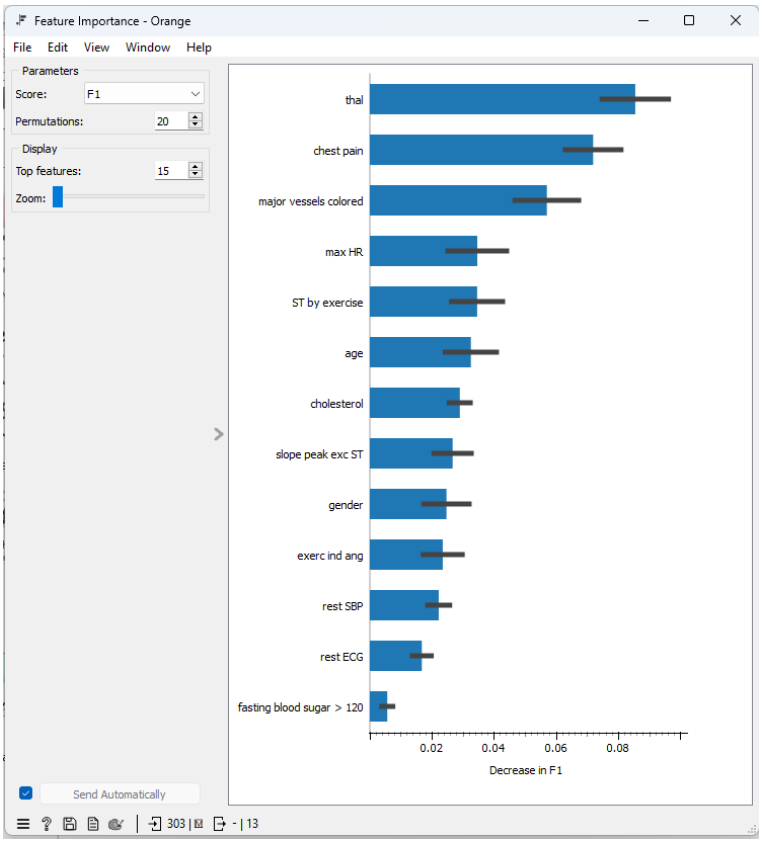


Rysunek 13 Wyniki testów jakości predykcji dla poszczególnych modeli

Z przeprowadzonych testów wynika, że najsilniejszym modelem jest Random Forest (RF), drugim modelem jest Gradient Boosting (GB) a najsłabszym modelem w tym wypadku jest Decision Tree (DT). Jest to wytłumaczalne biorąc pod uwagę prostszą naturę tego modelu.

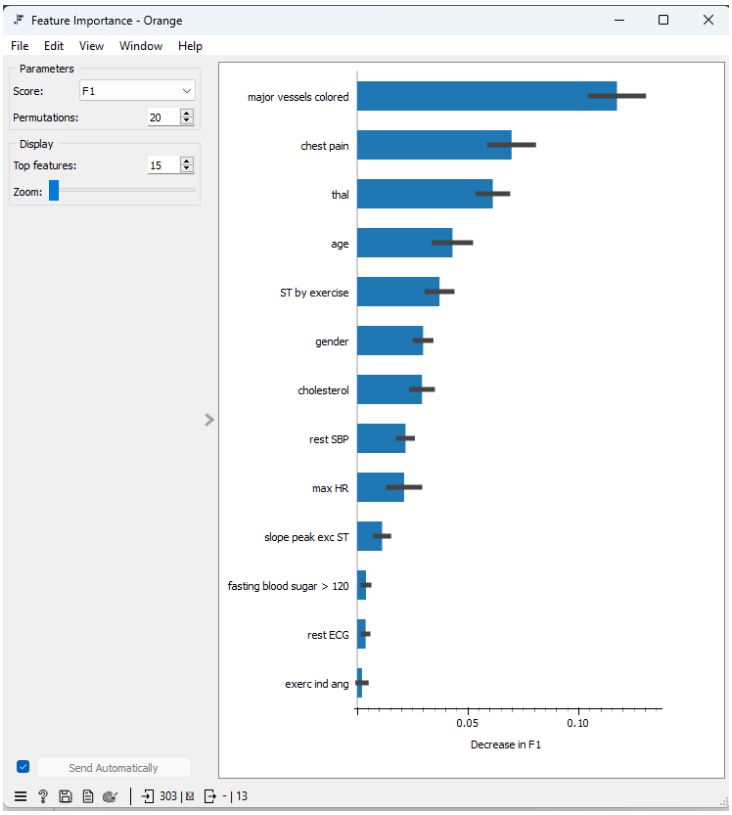
Następnym etapem zadania jest zbadanie wybór lasów losowych (boosting i random subsets) na zmianę rankingu cech Shapleya. Analiza jest przeprowadzona przy użyciu Feature importance i Explain model.

1. Feature importance - jako Score podczas analizy wybrałem F1



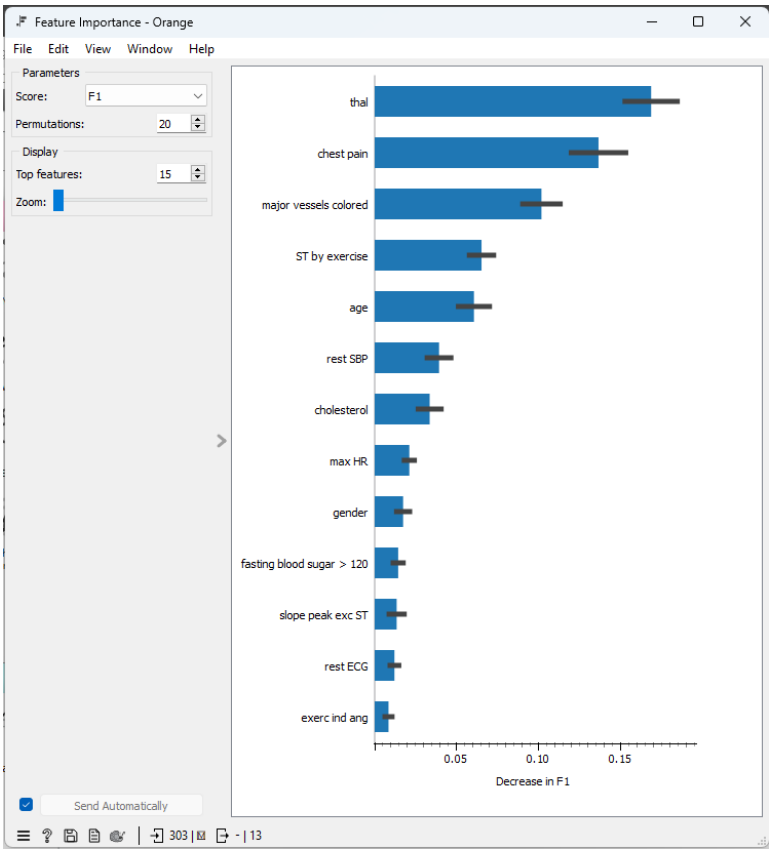
Rysunek 14 Wyniki Feature Importance dla modelu Random Forest

Widać, że w przypadku Random Forest usunięcie cechy ‘tahl’ spowodowałoby spadek F1 o około 8%. trochę podobna sytuacja ma miejsce w przypadku zmiennej ‘chest pain’, której usunięcie będzie skutkowało spadkiem F1 o około 6%.



Rysunek 15 Wyniki Feature Importance dla modelu Gradient Boosting

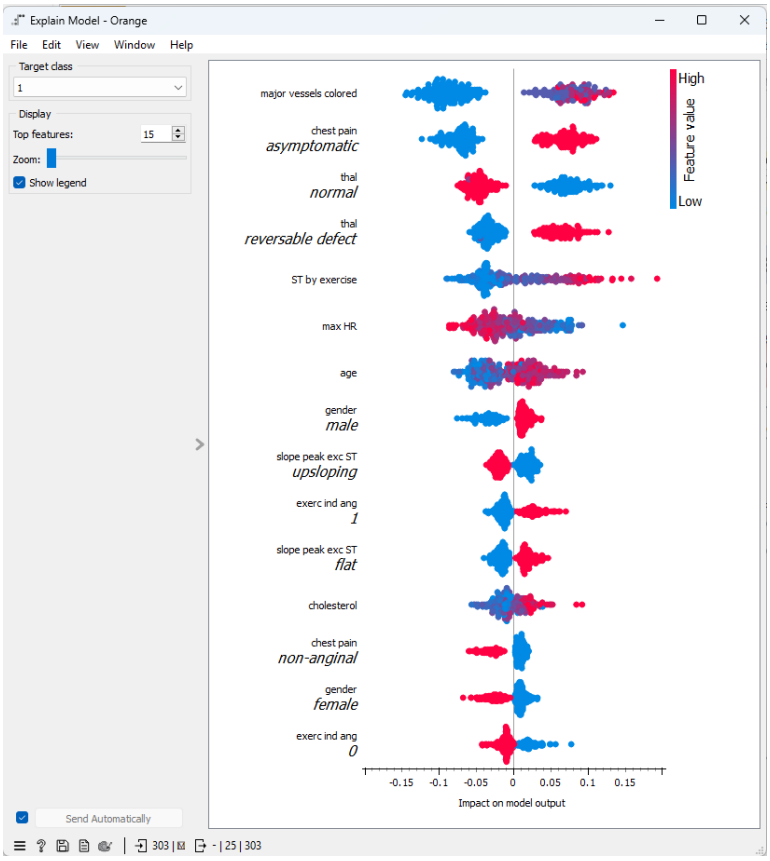
W przypadku Gradient Boosting widać, że usunięcie cechy ‘major vessel colored’ spowoduje spadek F1 o około 12%, trochę podobna sytuacja ma miejsce w przypadku zmiennej ‘chest pain’, której usunięcie będzie skutkowało spadkiem F1 o około 7 %.



Rysunek 16 Wyniki Feature Importance dla modelu Decision Tree

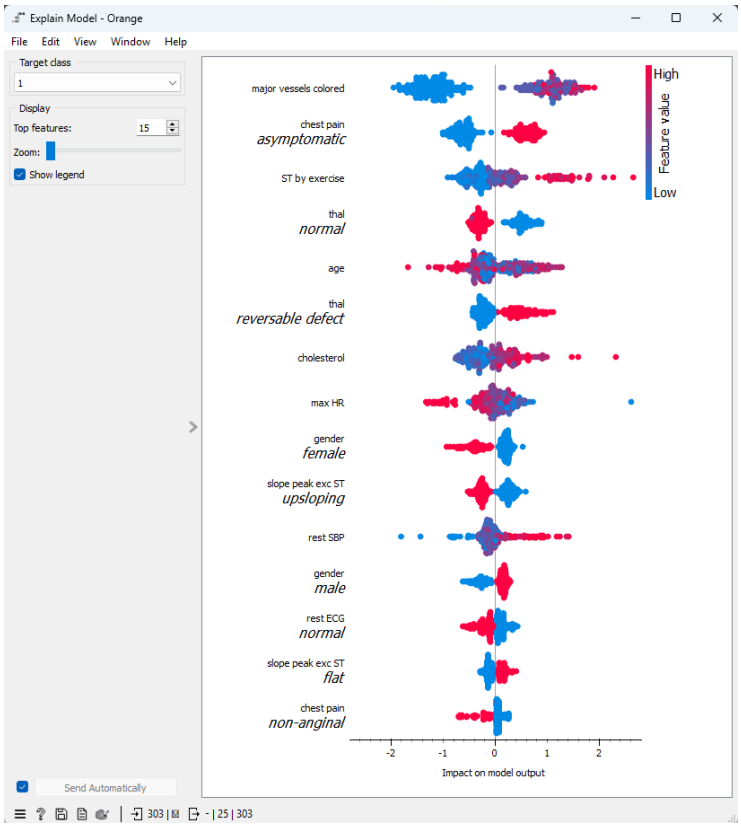
W przypadku Decision Tree widać, że usunięcie cechy ‘thal’ spowoduje spadek F1 o około 15%, trochę podobna sytuacja ma miejsce w przypadku zmiennej ‘chest pain’, której usunięcie będzie skutkowało spadkiem F1 o około 13 %.

2) Explain Model



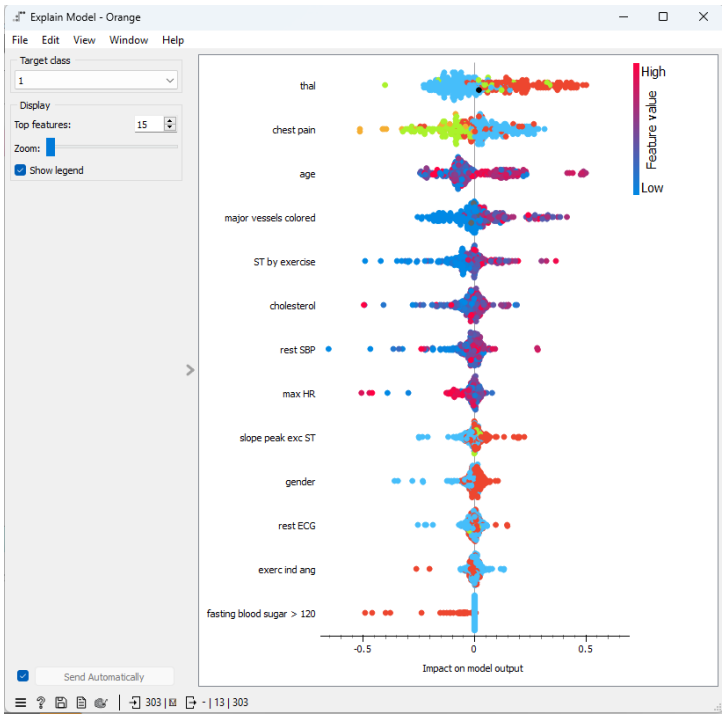
Rysunek 17 Wyniki Explain Model dla modelu Random Forest

Analizując wykres można dojść do wniosku, że największy wpływ na zakwalifikowanie osoby do grupy osób ze zwiększonym ryzykiem choroby w modelu Random Forest mają zmienne: ‘major vessel colored’, ‘chest pain’ oraz ‘thal’. Niska wartość pierwszej zmiennej ma negatywny wpływ na przyporządkowanie danej osoby do grupy osób chorych. Podobnie sytuacja wygląda ze zmienną ‘chest pain’. Większa wartość zmiennej ‘thal’ wskazuje na podwyższone ryzko choroby, ponieważ występują zaburzenia krwi. Można wnioskować, że wpływ różnych cech jest bardziej rozłożony. Jest to spowodowane faktem, że Random Forest jest algorytmem opartym na wielu drzewach decyzyjnych.



Rysunek 18 Wyniki Explain Model dla modelu Gradient Boosting

Analizując wykres można dojść do wniosku, że największy wpływ na zakwalifikowanie osoby do grupy osób ze zwiększonym ryzykiem choroby w modelu Gradient Boosting mają zmienne: ‘major vessel colored’, ‘chest pain’ oraz ‘thal’. Niska wartość zmiennej ma negatywny wpływ na przyporządkowanie danej osoby do grupy osób chorych. Różnicą w stosunku do modelu Random Forest jest trochę gorsza jakość klasyfikacji chorych w przypadku zmiennej ‘thal’.

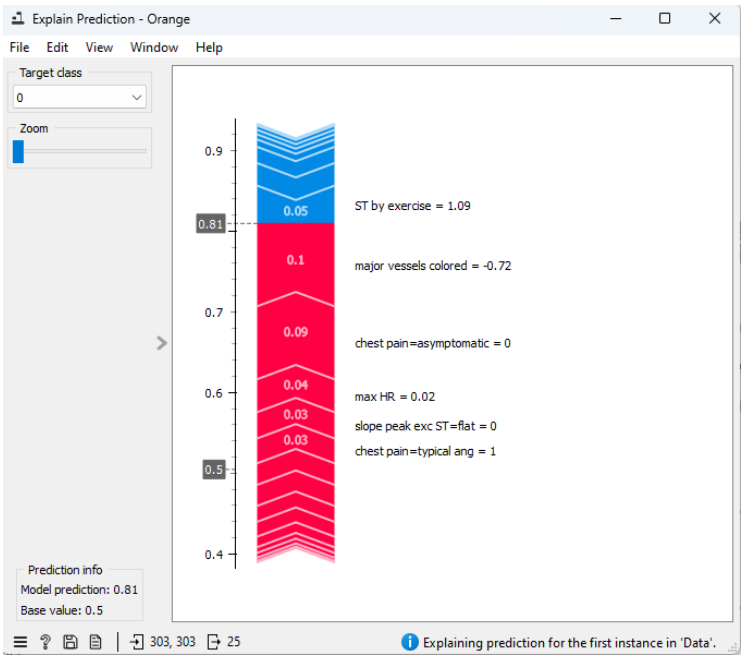


Rysunek 19 Wyniki Explain Model dla modelu Decision Tree

Analizując wykres można dojść do wniosku, że największy wpływ na zakwalifikowanie osoby do grupy osób ze zwiększonym ryzykiem choroby w modelu Gradient Boosting mają inne zmienne niż w pozostałych dwóch modelach. Są to zmienne: ‘thal’, ‘chest pain’ oraz ‘age’. Widać, że liczba błędów klasyfiakcji jest większa dla Decision Tree w stosunku do pozostałych dwóch modeli.

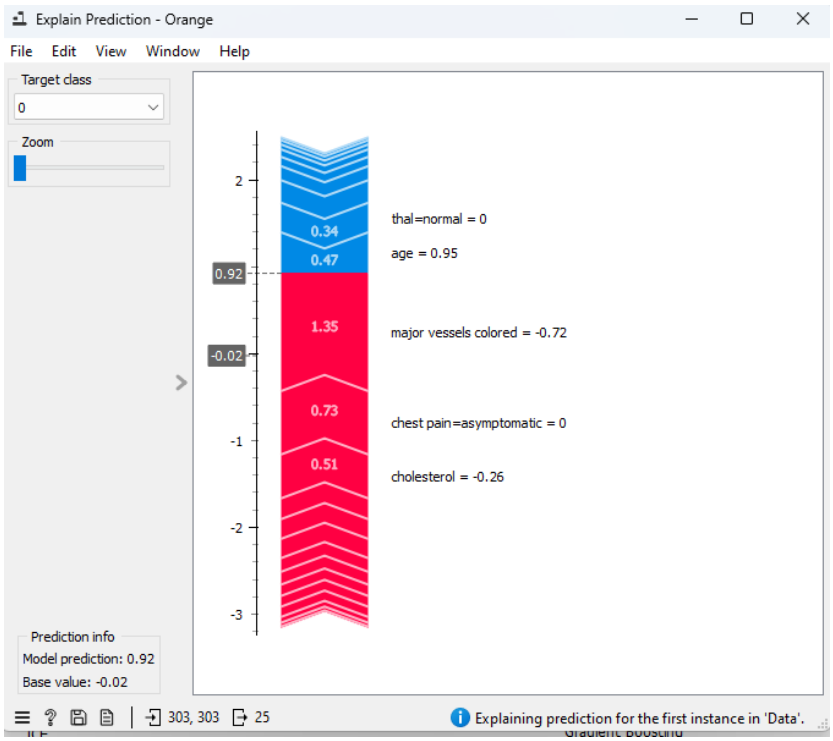
Obserwując wszystkie trzy wizualizacja można zauważyć, że Decision Tree najgorzej sobie radzi z klasyfikacją, ponieważ w przypadku większości zmiennych te obserwacje nie sa tak odseparowane od siebie jak to miało w przypadku Random Forest czy Gradient Boosting. Random Forest daje bardziej równomierny rozkład wpływu cech. Zmiany rankingu wynikają z różnic występujących między tymi algorytmami: Random Forest tworzy wiele niezależnych drzew na losowych podzbiorach danych i cech, co skutkuje bardziej równomiernym rozkładem wpływów. Gradient Boosting jest algorytmem który tworzy kolejne drzewa w celu poprawy wcześniejszych błędów, co powoduje większą koncentrację na najważniejszych zmiennych..

Ostatnim etapem projektu jest wytłumaczenie (Explain prediction) dla obu modeli dlaczego wybrane (2) przykłady testowe zostały zaklasyfikowane do danych klas oraz porównanie wyników działania dla różnych modeli.



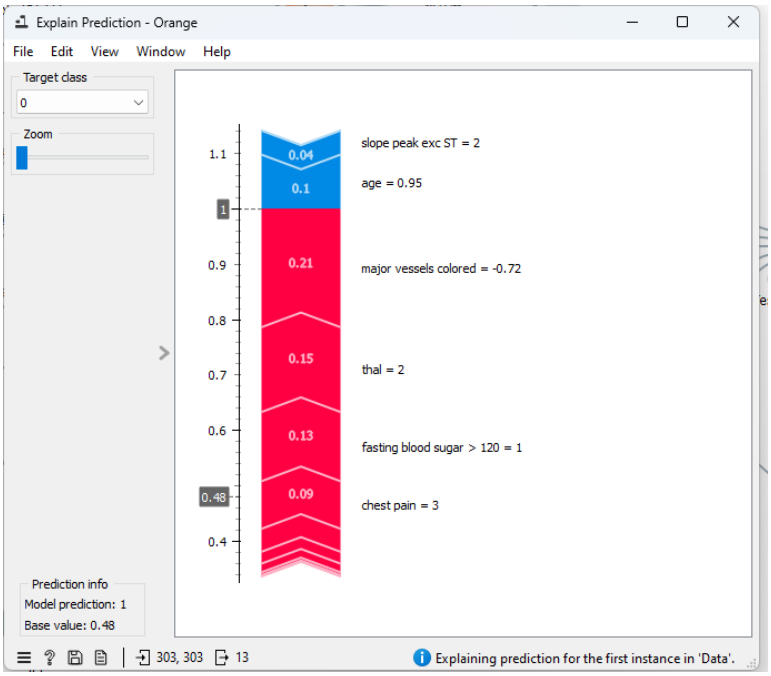
Rysunek 20 Wyniki Explain Prediction dla modelu Random Forest

Powyższy rysunek wskazuje, że model Random Forest w 81% przewidział, że dana osoba zostanie zaklasyfikowana do grupy osób, które nie maja zwężonych naczyń krwionośnych. Największy dodatni wpływ maja zmienne ‘chest pain’ oraz ‘major vesel colored’.



Rysunek 21 Wyniki Explain Prediction dla modelu Gradient Boosting

Powyższy rysunek wskazuje, że model Gradient Boosting w 92% przewidział, że dana osoba zostanie zaklasyfikowana do grupy osób, które nie maja zwężonych naczyń krwionośnych. Największy dodatni wpływ maja zmienne ‘major vessel colored ‘oraz ‘chest pain’.



Rysunek 22 Wyniki Explain Prediction dla modelu Decision Tree

Powyższy rysunek wskazuje, że model Gradient Boosting w 100% przewidział, że dana osoba zostanie zaklasyfikowana do grupy osób, które nie maja zwężonych naczyń krwionośnych. Taki wynik wskazuje, że model Decision Tree się przetrenował. Największy dodatni wpływ maja zmienne ‘major vessel colored ‘oraz ‘thal’.

Porównując wyniki predykcji można wnioskować, ze najlepiej z problemme klasyfikacji poradził sobie model Gradient Boosting, kolejny był Random Forest. Model Decion Tree się przetrenował.