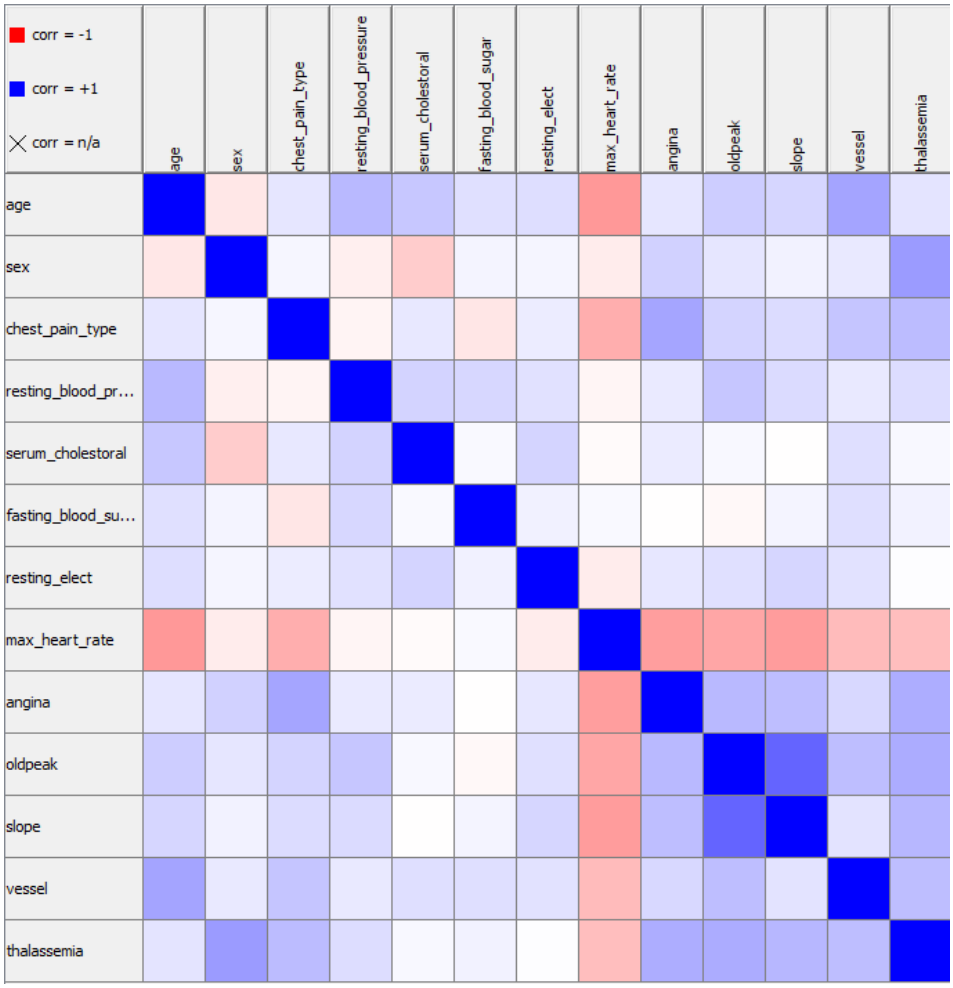
Eksploracja Danych – Raport 2 Klaudia Pełczyńska

Analizowany zbiór danych znajduje się w pliku *heart_disease.csv*, który zawiera dane pacjentów kardiologicznych. Predyktorami są zmienne: **age, sex, chest_pain_type, serum_cholestoral, fasting_blood_sugar, resting_elect, max_heart_rate, angina, oldpeak, slope, vessel, thalassemia**, a zmienna celu jest zmienna **heart_disease.** Ziarno generatora wynosi 308285.

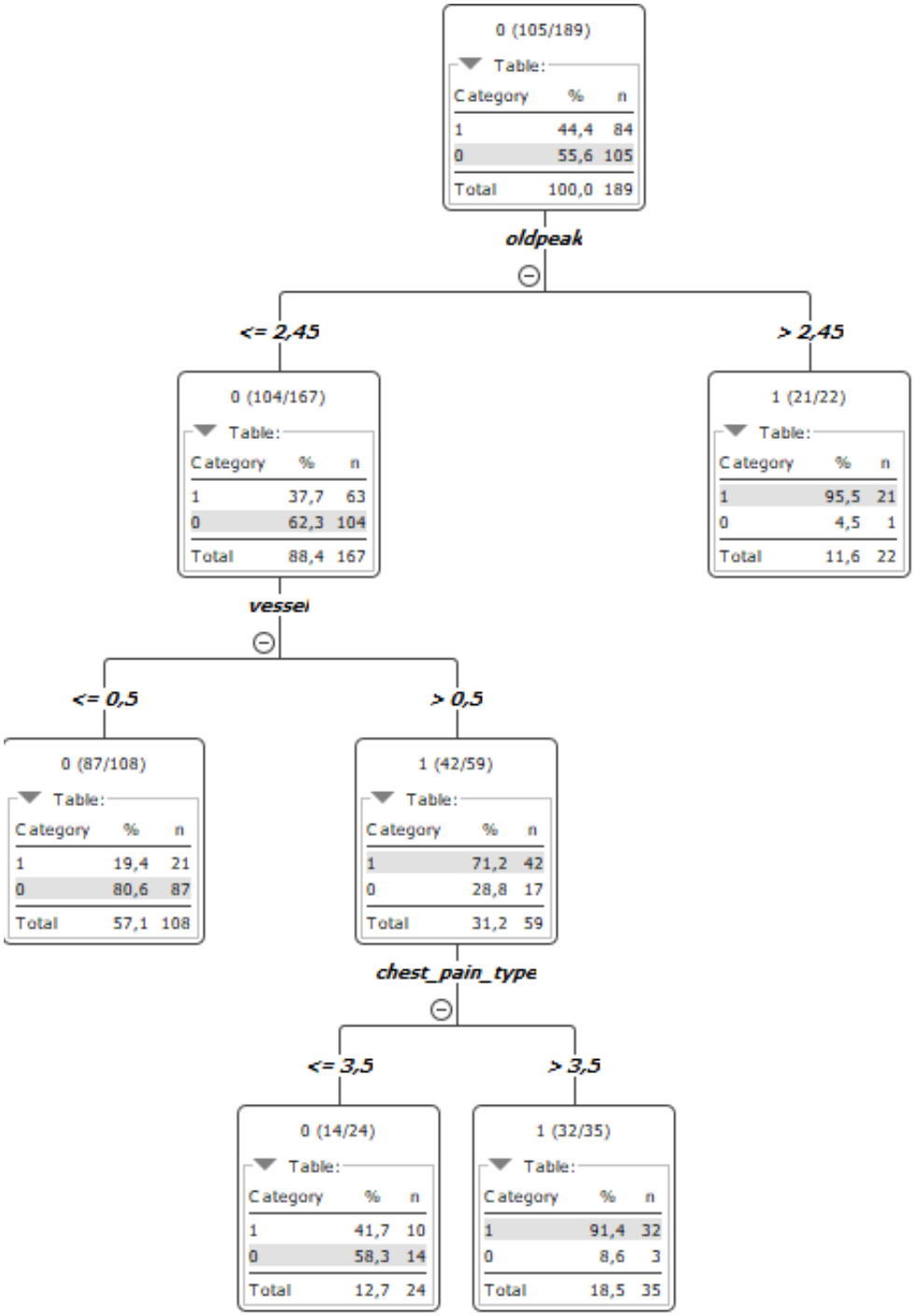


Rysunek 1. Diagram korelacji

Na podstawie powyższego diagramu korelacji zostały wybrane następujące predyktory: **age, sex, chest_pain_type, resting_blood_pressure, serum_cholestoral, fasting_blood_sugar, resting_elect, angina, oldpeak, vessel**. Odrzuciliśmy następujące zmienne:

- slope – ze względu na korelację ze zmienną oldpeak (0.61)
- max_heart_rate – ze względu na korelację z zmiennymi age (-0.40), chest_pain_type (-0.32), angina (-0.38), oldpeak (-0.35)
- thalassemia – ze względu na korelację z zmiennymi sex (0.39), angina (0.32), oldpeak (0.32)

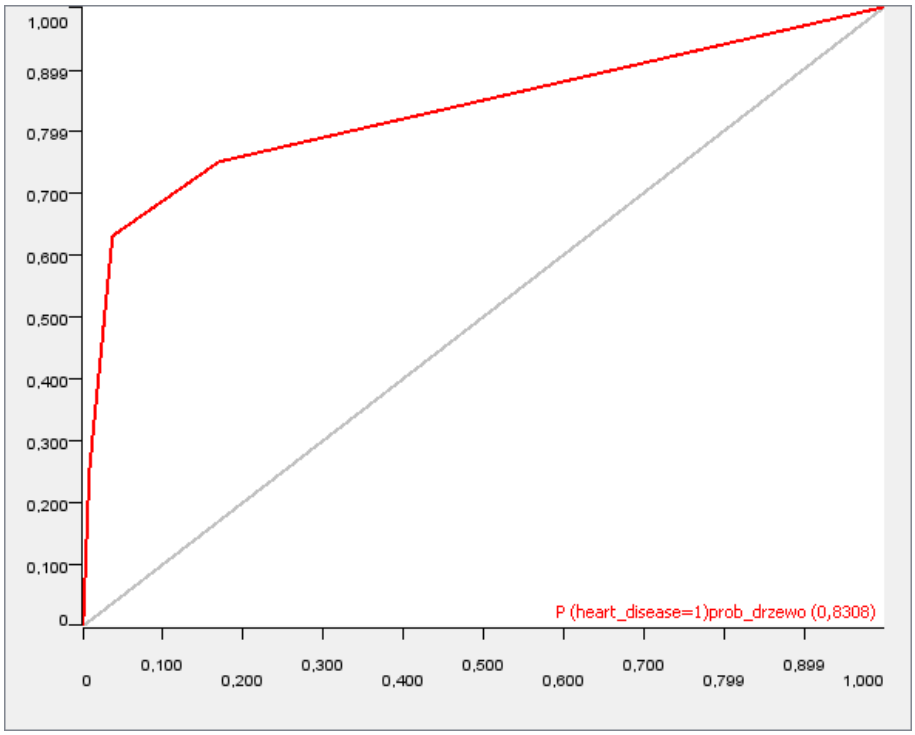
Drzewo C5.0



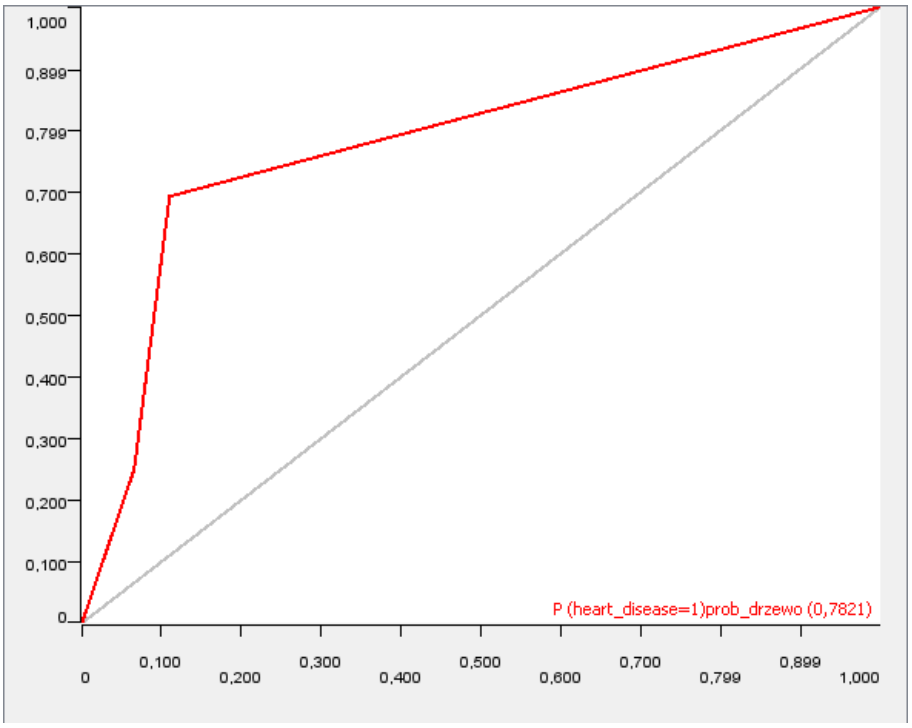
Rysunek 2. Drzewo C5.0

Pierwsza wersja drzewa C5.0 została zbudowana wykorzystując wszystkie wymienione wyżej predyktory. W powyższym algorytmie istnieje możliwość dostosowania minimalnej liczby rekordów w liściu, która w tym przypadku została ustalona na 15 ze względu na małą liczbę obserwacji. Przy takiej wartości model zwracał najlepsze wyniki.

Jak możemy zauważyć po wprowadzeniu danych model sprawdza czy zmienna **oldpeak** jest nie większa niż 2.45 jeśli tak, to drzewo sprawdza czy zmienna **vessel** jest nie większa niż 0.5, jeśli nie to sprawdza zmienną **chest_pain_type**. Na samym końcu zwraca on klasę do której należy większość próbek w danym liściu.



Rysunek 3. Krzywa ROC dla zbioru uczącego



Rysunek 4. Krzywa ROC dla zbioru testowego

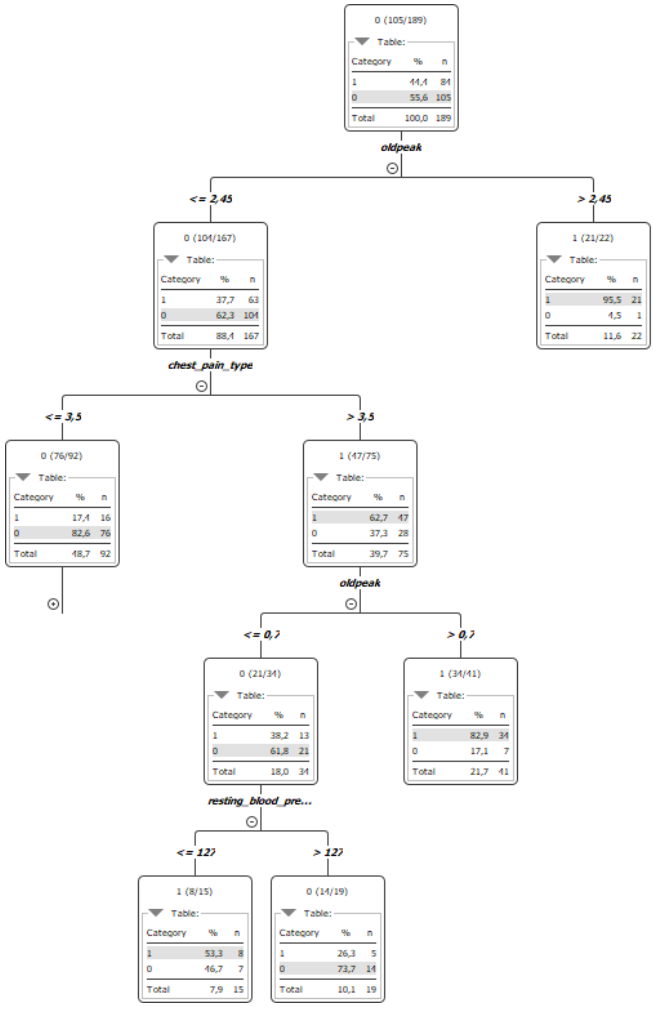
Trafność dla próby uczącej	81%
Czułość dla próby uczącej	63%
Swoistość dla próby uczącej	96%
Trafność dla próby testowej	80%
Czułość dla próby testowej	69%
Swoistość dla próby testowej	89%

Tabela 1. Wartości statystyk testowych w podziale na dane uczące i dane testowe

W tabelce powyżej widzimy, że drzewo w pierwszej wersji osiągnęło trafność równą 81% na danych uczących i 80% na danych testowych, co wskazuje na nieznaczne przetrenowanie. Czułość dla danych uczących wyniosła 63%, a na próbie testowej 69%, natomiast swoistość wyniosła odpowiednio 96% i 89%. Wartości statystyk testowych różnią się od siebie o mniej niż 10%, co wskazuje, że model nie jest aż tak mocno przetrenowany.

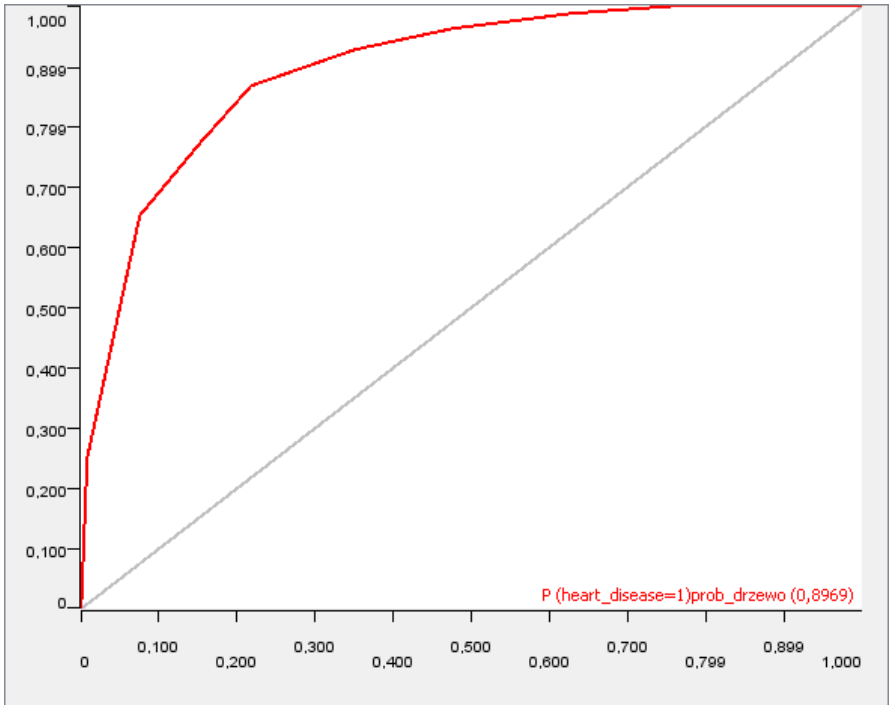
Zależy nam na wykrywaniu czy pacjent jest chory, dlatego moim celem było poprawienie czułości, czyli odsetkiem chorych, którzy zostali poprawnie wykryci. W tym celu zbudowałam drugi model tym razem jednak odrzuciłam również zmienne **angina** i **vessel**.

Ze względu na wielkość drzewa został pokazany tylko jego fragment.

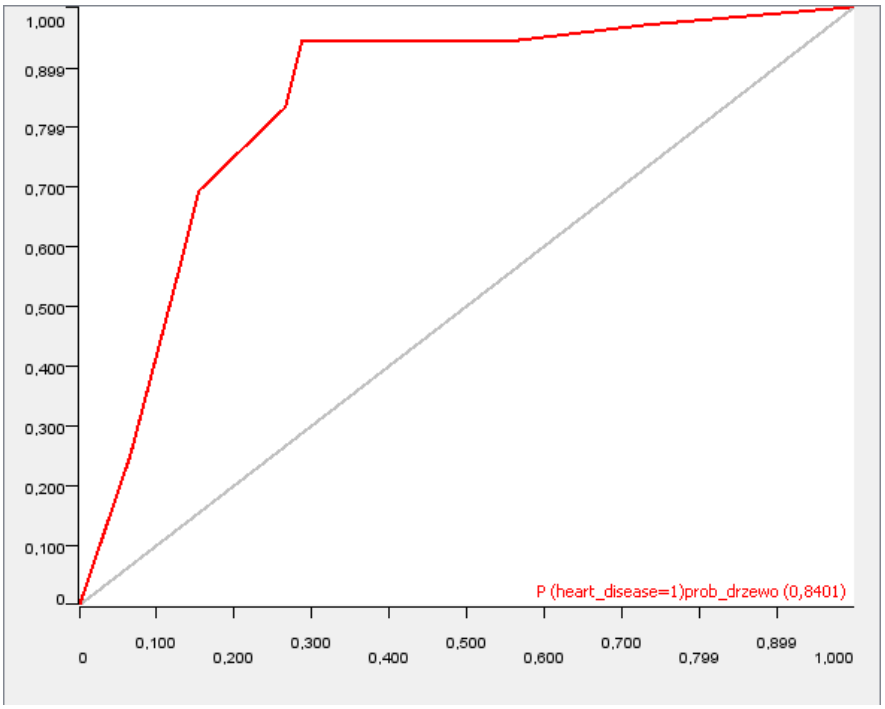


Rysunek 5. Drzewo C5.0

Model, którego fragment widać na rysunku 5 jest bardziej skomplikowany niż poprzednie drzewo C5.0. Jest również trochę bardziej przetrenowany (jednak różnica jest mniejsza niż 10%), co widać w tabeli 2.



Rysunek 6. Krzywa ROC dla zbioru uczącego



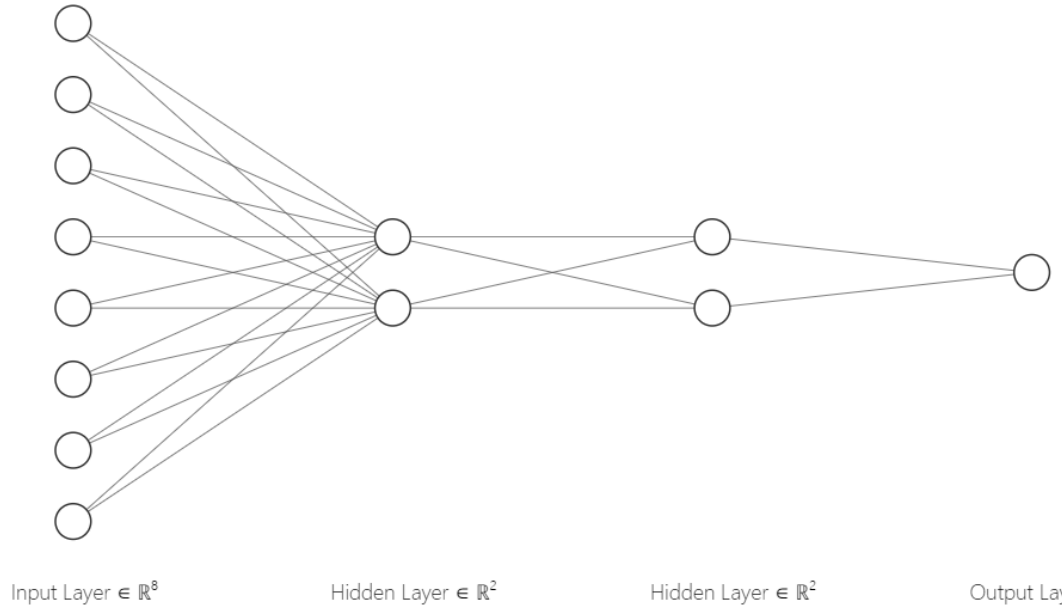
Rysunek 7. Krzywa ROC dla zbioru testowego

Trafność dla próby uczącej	82%
Czułość dla próby uczącej	87%
Swoistość dla próby uczącej	78%
Trafność dla próby testowej	78%
Czułość dla próby testowej	83%
Swoistość dla próby testowej	73%

Tabela 2. Wartości statystyk testowych w podziale na dane uczące i dane testowe

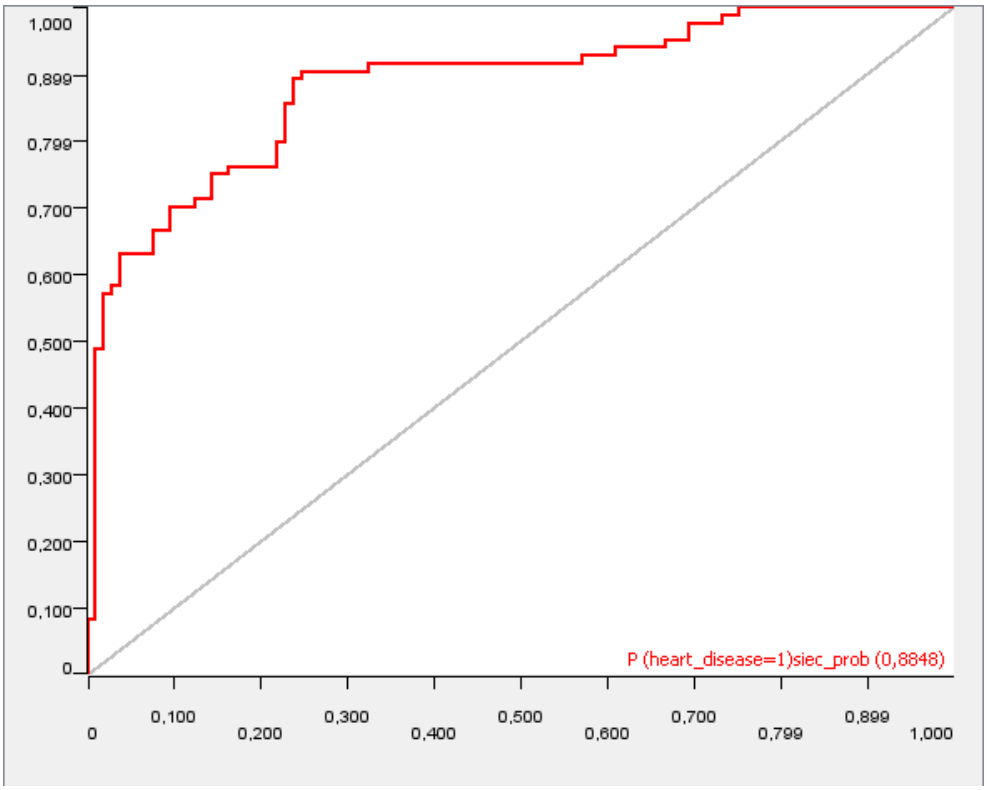
Druga wersja drzewa C5.0 jest trochę bardziej przetrenowana, jednak ten model dużo lepiej wykrywa klasę pozytywną (osoby chore). W przeciwieństwie do pierwszej wersji, która o wiele lepiej wykrywała klasę negatywną (osoby zdrowe). Również pole pod krzywą w zbiorach uczących jak i w zbiorach testowych jest wyższe dla drugiego modelu, co znaczy, że jest on lepszym modelem.

Sieć MLP

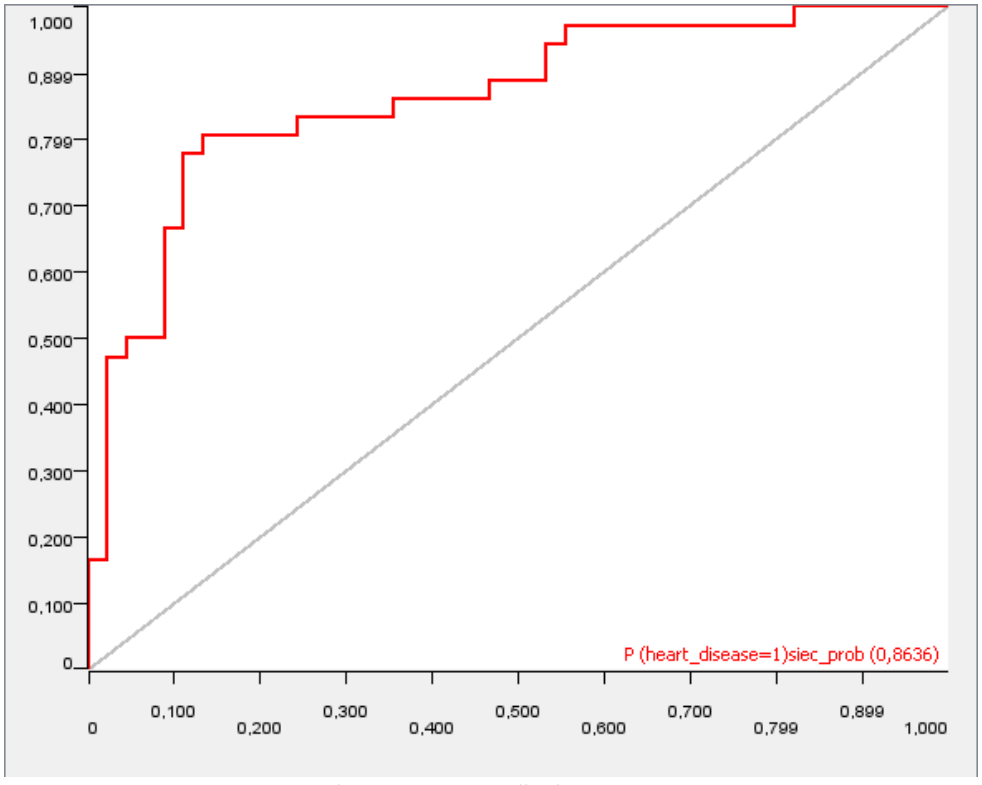


Rysunek 8. Sieć neuronowa

Zastosowany model sieci neuronowej składa się z warstwy wejściowej posiadającej 8 wejść, dwóch warstw ukrytych składających się z 2 neuronów i jednego wyjścia. Dla lepszych wyników nie zastosowałam predyktorów angina i vessel.



Rysunek 9. Krzywa ROC dla zbioru uczącego



Rysunek 10. Krzywa ROC dla zbioru testowego

Trafność dla próby uczącej	79%
Czułość dla próby uczącej	82%
Swoistość dla próby uczącej	77%
Trafność dla próby testowej	83%
Czułość dla próby testowej	78%
Swoistość dla próby testowej	87%

Tabela 3. Wartości statystyk testowych w podziale na dane uczące i dane testowe

Trafność sieci MLP na danych uczących wyniosła 79%, natomiast na danych testowych 83%. Czułość na próbie uczącej wyniosła 82%, a na próbie testowej 78%, swoistość wyniosła odpowiednio 77% i 87%. Spadek wartości czułości świadczy o nieznacznym przetrenowaniu.

Wnioski

Porównując otrzymane wyniki uważam, że najlepszym uzyskanym modelem jest druga wersja drzewa C5.0. Pomimo, że jest najbardziej przetrenowanym modelem jego czułość na danych testowych była najwyższa, na czym najbardziej nam zależy.