Politechnika Warszawska

WSI
Laboratorium 4

Bartosz Czerwiński - 331165

6 maja 2025

Spis treści

1. Wstęp

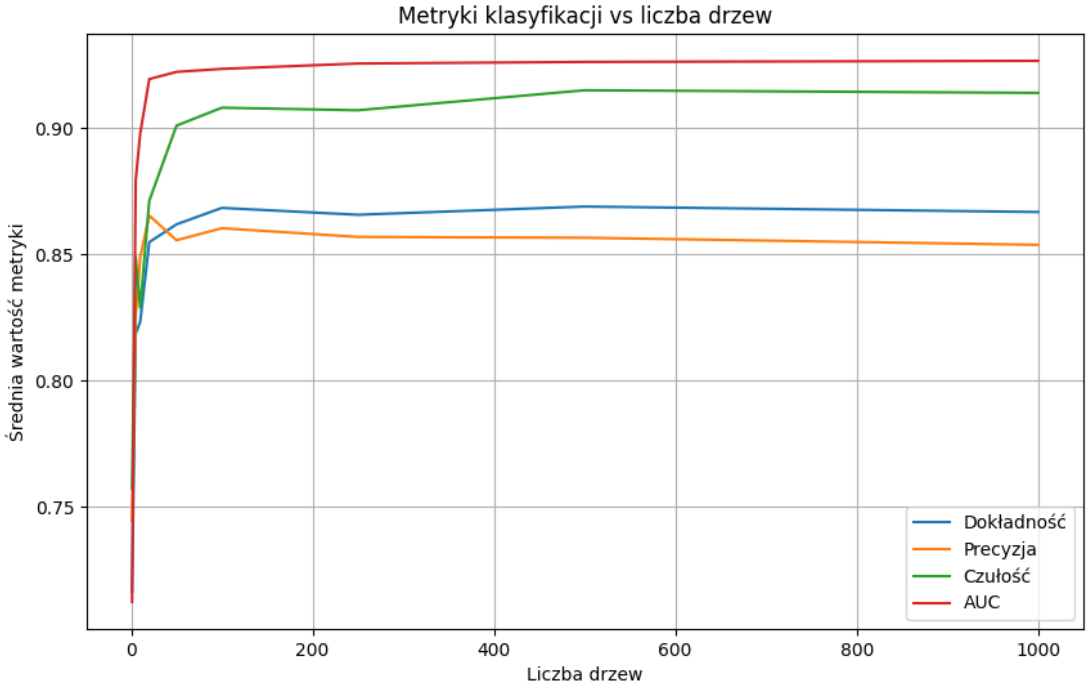
Celem laboratorium było zaimplementowanie lasu losowego oraz jego zastosowanie do rozwiązania problemu klasyfikacji – przewidywania chorób serca na podstawie wybranych cech pacjentów.

2. Przygotowanie danych

Zbiór danych zawierał cechy kategoryczne, dlatego należało zamienić je na wartości liczbowe. W tym celu skorzystano z biblioteki `pandas`. Następnie dane zostały podzielone na zbiór treningowy i testowy w stosunku 80:20.

3. Testy

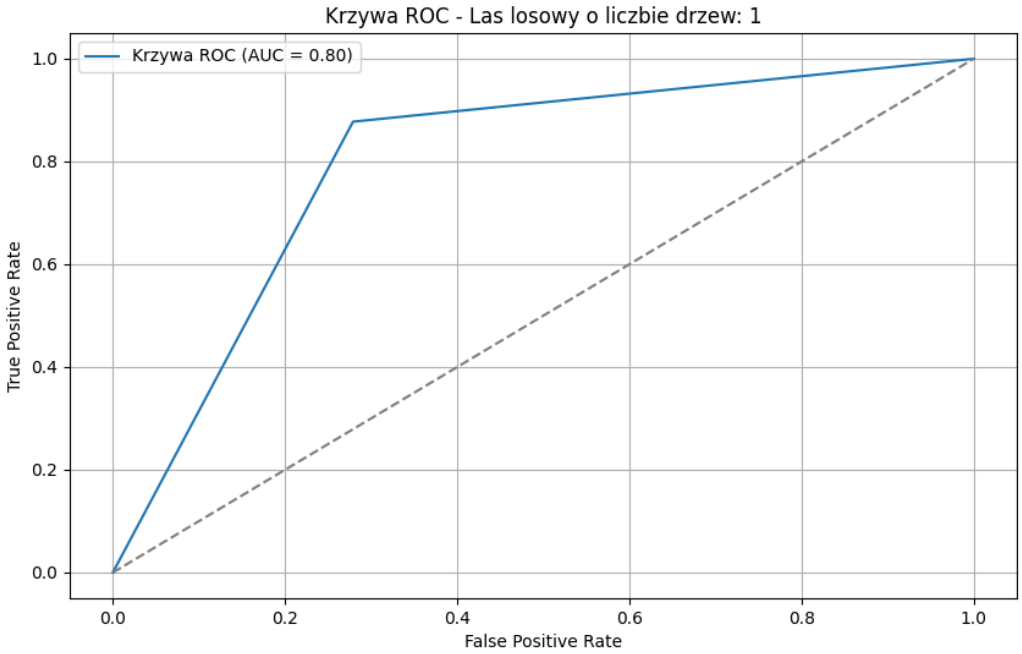
Testy zostały przeprowadzone dla różnych liczb drzew w lesie losowym: $1, 5, 10, 20, 50, 100, 250, 500, 1000$. Dla każdej liczby drzew przeprowadzono trening na 10 różnych zbiorach treningowych, a następnie obliczono na zbiorach testowych wartości metryk: dokładność, precyzja, czułość, pole pod krzywą ROC. Dla jednego z testów dla każdej liczby drzew narysowano także krzywą ROC.



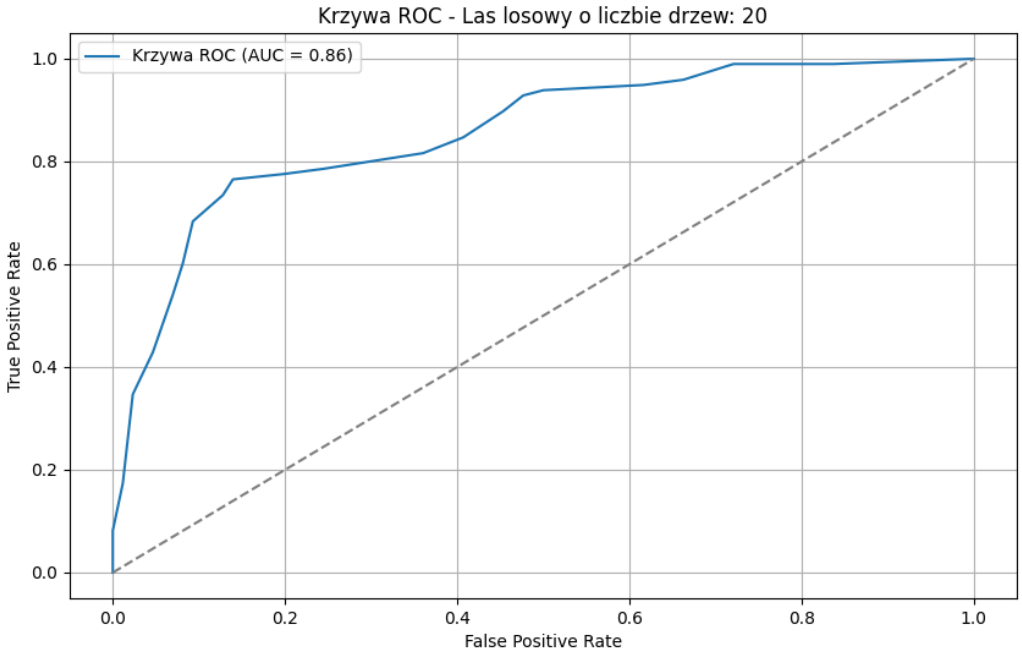
Rys. 1: Wpływ liczby drzew na metryki

Można zauważyć, że dla małej liczby drzew metryki mocno się zmieniają, a wraz ze wzrostem liczby drzew metryki rosną i się stabilizują.

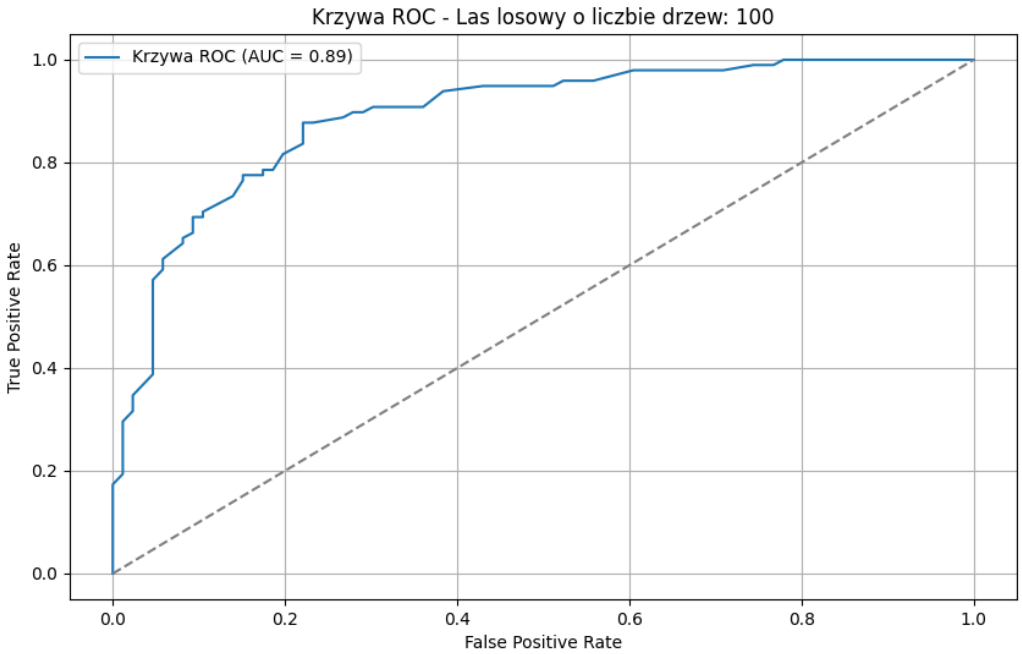
4. Krzywa ROC



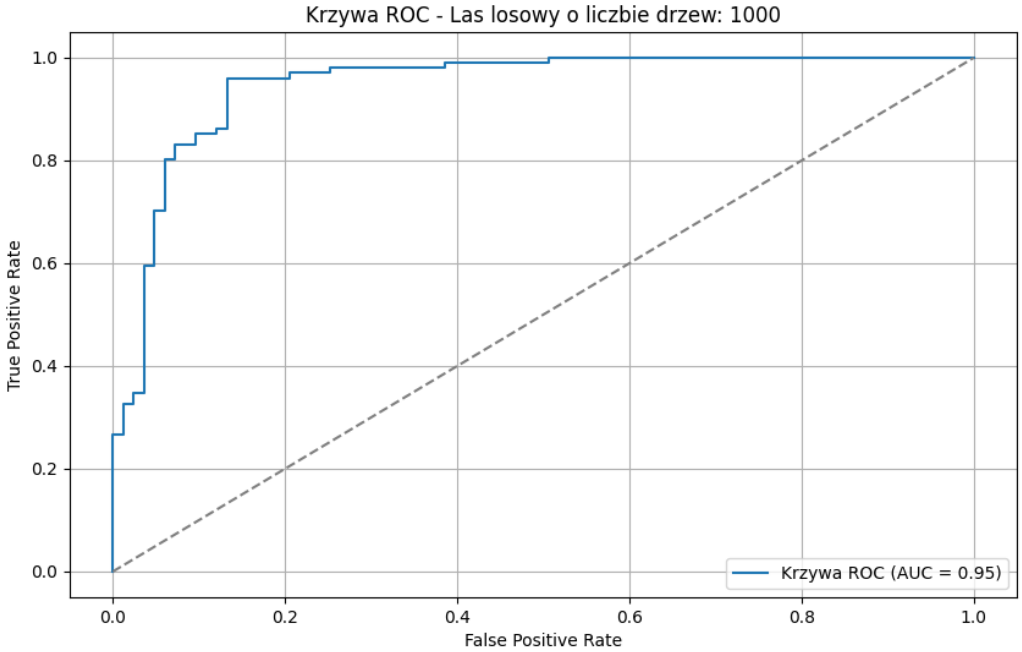
Rys. 2: Krzywa ROC dla 1 drzewa



Rys. 3: Krzywa ROC dla 20 drzew



Rys. 4: Krzywa ROC dla 100 drzew



Rys. 5: Krzywa ROC dla 1000 drzew

Krzywa ROC przedstawia zależność między czułością a odsetkiem błędnych alarmów (zdrowych uznanych za chorych). Zbliżanie się krzywej do lewego górnego rogu oznacza, że model jest idealny - wykrywa wszystkich chorych i nie ma żadnych fałszywych alarmów. Wraz ze zbliżaniem się wykresu do lewego górnego rogu zwiększa się AUC - pole pod wykresem. Można zauważyć, że wraz ze zwiększaniem liczby drzew, krzywa ROC zbliża się coraz bardziej do idealnej krzywej, co prowadzi do zwiększenia pola pod wykresem.

5. Znaczenie wyników

Obliczone następujące parametry lasu losowego:

- Dokładność (accuracy) - stosunek prawidłowych przewidywań do wszystkich przewidywań;
- Precyzja (precision) - stosunek poprawnie przewidzianych przypadków pozytywnych (wartość 1) do wszystkich przypadków przewidzianych pozytywnie;
- Czułość (recall) - stosunek poprawnie przewidzianych przypadków pozytywnych do wszystkich przypadków, które w rzeczywistości były pozytywne.

Gdyby podany model miał być wykorzystywany w rzeczywistej sytuacji wykrywania chorób serca, maksymalizowana powinna być czułość, ponieważ jest ona wrażliwa na brak przewidywania przypadku pozytywnego (przewidywanie braku choroby, gdy ta faktycznie istnieje). Celem jest zmniejszenie takich przypadków tak bardzo, jak to tylko możliwe, dlatego należy maksymalizować właśnie tę metrykę.