Ćw. 4 - klasyfikator ID3 Aleksander Drwal

Wartości wszystkich wyników są średnią z 20 uruchomień algorytmu. Dane były podzielone w sposób losowy na zbiór trenujący i testujący w stosunku $3:2$. Zbiory na których algorytm był testowany to Breast cancer i mushroom.

Wyniki dla zbioru Breast cancer

$$Macierz\ pomyłek \approx \begin{bmatrix} 10.7 & 17.25 \\ 21.35 & 63.2 \end{bmatrix}$$

$$Dokładność \approx 66\%$$

$$Czułość \approx 79\%$$

$$Swoistość \approx 33\%$$

$$Precyzja \approx 72\%$$

Wyniki dla zbioru mushroom

$$Macierz\ pomyłek \approx \begin{bmatrix} 1672.4 & 0 \\ 0 & 1555.6 \end{bmatrix}$$

$$Dokładność \approx 100\%$$

$$Czułość \approx 100\%$$

$$Swoistość \approx 100\%$$

$$Precyzja \approx 100\%$$

Hipotezy różnicy w wynikach

Przetestowałem, czy różnicą w wynikach algorytmu może być fakt, że w zbiorze `Breast cancer` klasa `no-recurrence-events` jest dominująca (201 sztuk vs. 85). Po usunięciu wierszy i wyrównaniu ilości klas, model osiąga wyniki jednak podobne wyniki.

$$Macierz\ pomyłek \approx \begin{bmatrix} 26 & 12.5 \\ 10.9 & 22.3 \end{bmatrix}$$

$$Dokładność \approx 67\%$$

$$Czułość \approx 70\%$$

$$Swoistość \approx 64\%$$

$$Precyzja \approx 67\%$$

Tak samo było w przypadku zbioru `mushroom`, z tym, że różnica tam była jeszcze mniejsza.

W zbiorze Breast cancer, wiek był podany jako przedział, aby sprawdzić, czy może to wpływać na wyniki algorytmu, zamieniłem każdą wartość wieku z przedziału na konkretną

liczbę.

$$Macierz\ pomyłek \approx \begin{bmatrix} 58.2 & 24.5 \\ 7.8 & 21.7 \end{bmatrix}$$

$$Dokładność \approx 58\%$$

$$Czułość \approx 73\%$$

$$Swoistość \approx 24\%$$

$$Precyzja \approx 70\%$$

Wpłynęło to jednak negatywnie na wyniki - algorytm przeuczył się na zbiorze trenującym.

Aby przetestować, czy większą ilość danych w zbiorze mushroom, ma wpływ na tego lepsze wyniki, to usunąłem ze zbioru około 97% losowo wybranych wierszy.

$$Macierz\ pomyłek \approx \begin{bmatrix} 49 & 1.2 \\ 0 & 38.9 \end{bmatrix}$$

$$Dokładność \approx 98\%$$

$$Czułość \approx 100\%$$

$$Swoistość \approx 97\%$$

$$Precyzja \approx 97\%$$