Analityka i eksploracja danych
Laboratorium - Część 2

Karol Jaskółka 241306

12.01.2022

Spis treści

1 Wstęp **3**

2 kNN **4**
2.1 Parametry klasyfikatora 4
2.2 Redukcja wymiarowości danych 5

3 Drzewo decyzyjne **6**
3.1 Parametry klasyfikatora 6
3.2 Redukcja wymiarowości danych 7
3.3 Zastosowanie niesymetrycznych kosztów błędów 8
3.4 Składanie modeli 8

4 Las losowy **9**
4.1 Parametry klasyfikatora 9
4.2 Redukcja wymiarowości danych 10
4.3 Zastosowanie niesymetrycznych kosztów błędów 11

5 MLP **12**
5.1 Parametry klasyfikatora 12
5.2 Redukcja wymiarowości danych 13

6 SVM **14**
6.1 Parametry klasyfikatora 14
6.2 Redukcja wymiarowości danych 15
6.3 Zastosowanie niesymetrycznych kosztów błędów 16

7 Podsumowanie **17**
7.1 Porównanie algorytmów 17
7.2 Wnioski 17

1 Wstęp

Do realizacji laboratorium został wybrany zbiór danych **spam.dat**, przeznaczony do klasyfikacji w oparciu o dane wysoko wymiarowe.

Celem zadania jest zbudowanie klasyfikatora i jego dostrojenie w celu uzyskania najlepszych wyników. Jako kryterium jakości przyjmujemy:

- minimalizację stopy błędów 'yes' → 'no' (FNR)
- przy zapewnieniu stopy błędów 'no → 'yes' (FPR) poniżej 0.5%

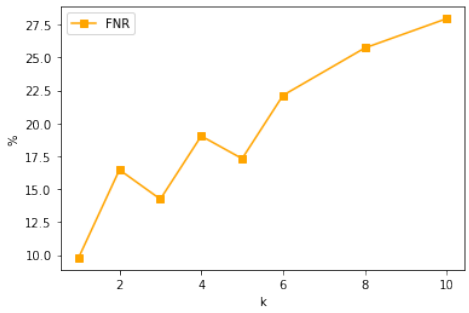
Do przeprowadzenia eksperymentów wykorzystano następujące algorytmy:

- kNN
- Drzewo decyzyjne
- Las losowy
- SVM
- MLP

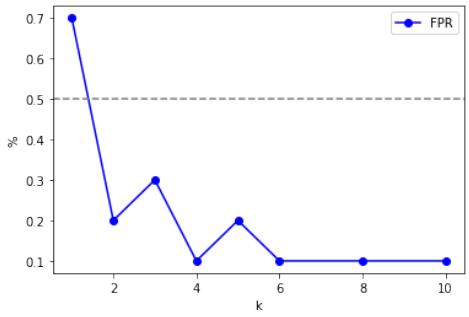
2 kNN

2.1 Parametry klasyfikatora

W przypadku algorytmu kNN badanym parametrem była liczba najbliższych sąsiadów k ∈ [1, 2, 3, 4, 5, 6, 8, 10].



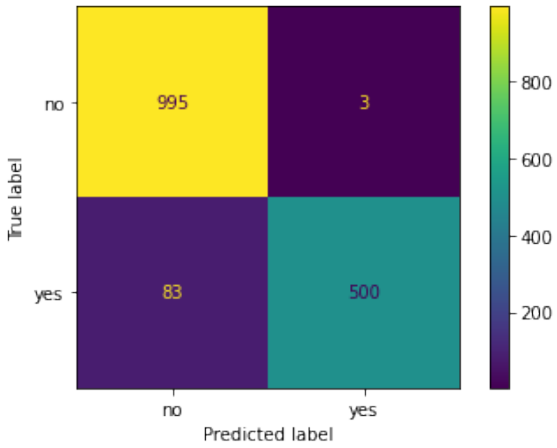
(a) FNR



(b) FPR

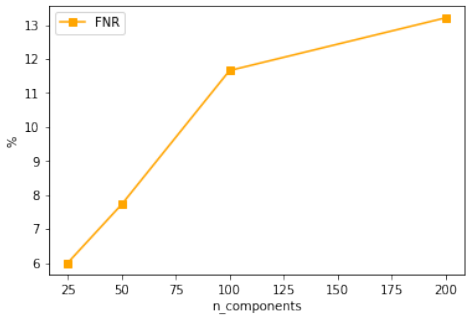
Klasyfikator uzyskał najlepsze wyniki dla $\mathbf{k = 3}$.

- FNR - 14.24%
- FPR - 0.3%

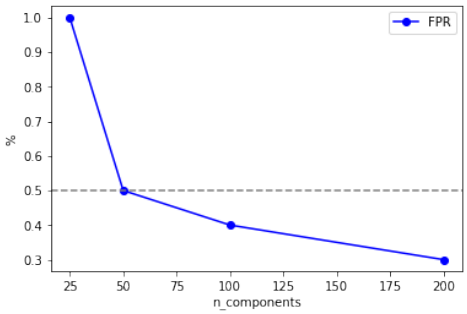


2.2 Redukcja wymiarowości danych

Do kolejnego badania został użyty klasyfikator z liczbą sąsiadów równą 3. W celu redukcji wymiarowości wykorzystano algorytm PCA, zbadany pod kątem liczby komponentów n ∈ [25, 50, 100, 200].



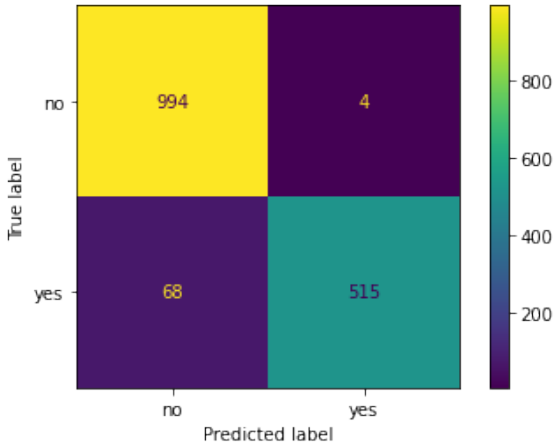
(a) FNR



(b) FPR

Klasyfikator uzyskał najlepsze wyniki dla **100** komponentów PCA.

- FNR - 11.66%
- FPR - 0.4%

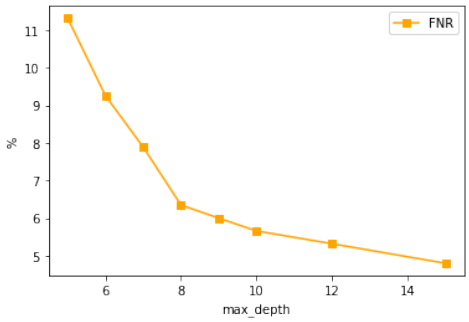


Dzięki redukcji wymiarowości działanie klasyfikatora polepszyło się o ponad 2.5% w kontekście wskaźnika FNR, utrzymując FPR poniżej 0.5%.

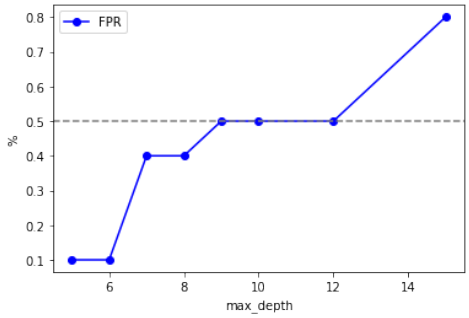
3 Drzewo decyzyjne

3.1 Parametry klasyfikatora

W przypadku algorytmu drzewa decyzyjnego badanym parametrem była maksymalna wysokość drzewa $\in [5, 6, 7, 8, 9, 10, 12, 15]$.



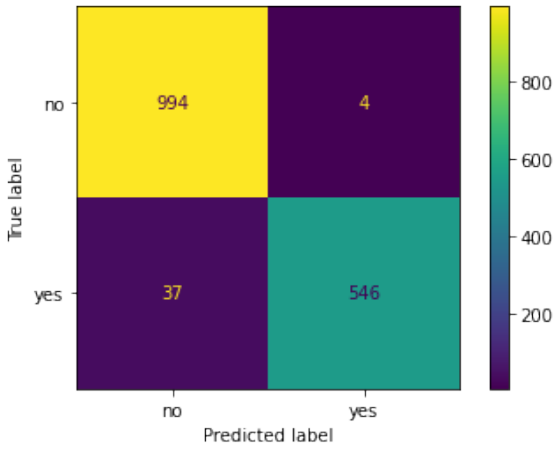
(a) FNR



(b) FPR

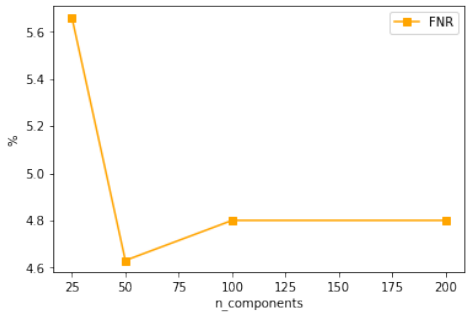
Klasyfikator uzyskał najlepsze wyniki dla wysokości równej **8**.

- FNR - 6.35%
- FPR - 0.4%

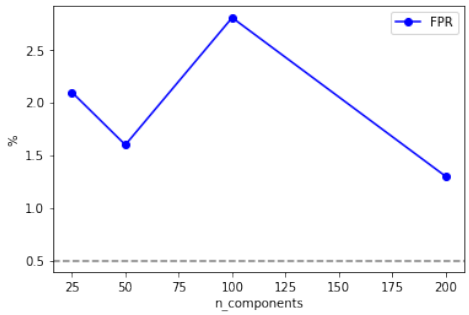


3.2 Redukcja wymiarowości danych

Do badania został użyty klasyfikator z maksymalną wysokościom równą 8. W celu redukcji wymiarowości wykorzystano algorytm PCA, zbadany pod kątem liczby komponentów n ∈ [25, 50, 100, 200].

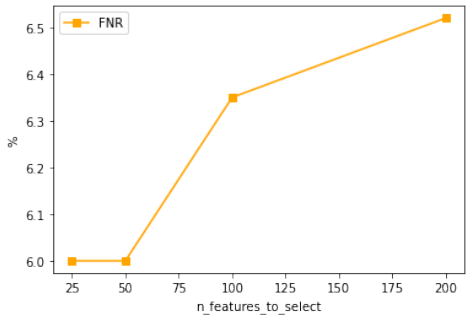


(a) FNR

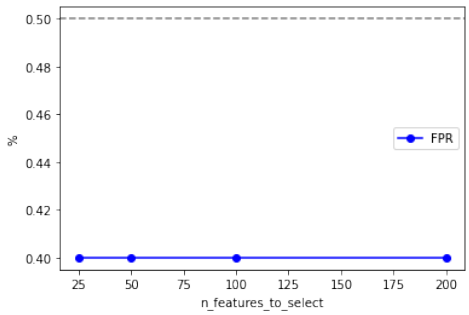


(b) FPR

Drugim algorytmem był RFE w którym za argument liczby cech zostały podane analogiczne wartości do PCA.



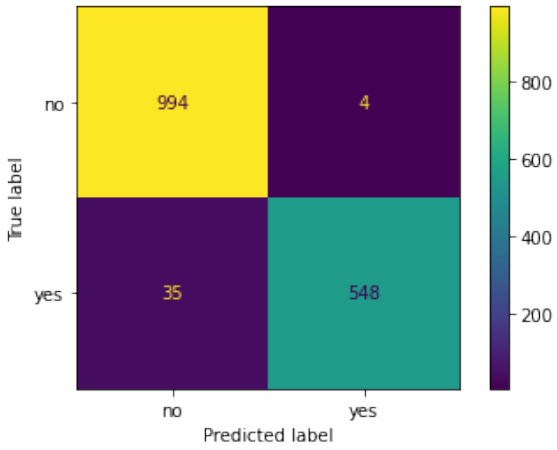
(a) FNR



(b) FPR

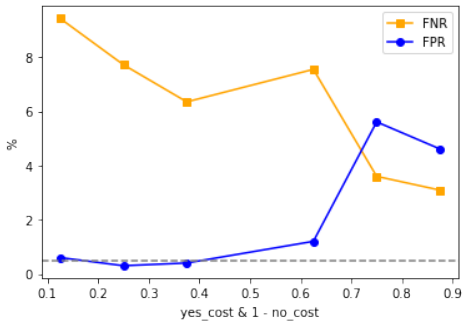
W przypadku PCA żaden wynik nie zmieścił się w docelowym przedziale FPR poniżej 0.5 %. Natomiast dla **RFE** uzyskano poprawę wskaźnika FNR przy zachowaniu FPR na poziomie 0.4 % dla liczby cech równej **25** oraz **50**.

- FNR - 6.0%
- FPR - 0.4%



3.3 Zastosowanie niesymetrycznych kosztów błędów

Kolejnym aspektem było wykorzystanie niesymetrycznych kosztów błędów. W tym celu wykorzystano parametr class_weight. Badanie zostało przeprowadzone na sześciu lustrzanych parach klas 'yes' oraz 'no'. Wagi prezentują się następująco (0.125 - 0.875, 0.25 - 0.75, 0.375 - 0.635, 0.635 - 0.375 itd.)



Najlepsze rezultaty osiągnięto dla wagi 'yes' równej 0.375 oraz 'no' równej 0.625, niemniej jednak wynik ten był równy temu z części 3.1.

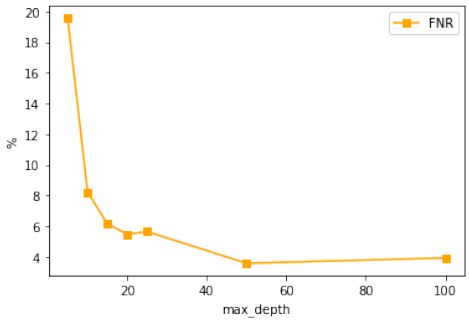
3.4 Składanie modeli

W tej części został wykorzystany algorytm AdaBoost. Pomimo, że uzyskana wartość FNR wyniosła jedyne 4.8 %, to wartość FPR przekroczyła dopuszczalną granicę i wyniosła 0.8 %.

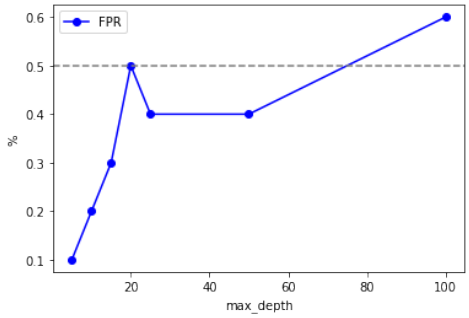
4 Las losowy

4.1 Parametry klasyfikatora

W przypadku lasu losowego badanym parametrem była maksymalna wysokość drzew klasyfikatorów $\in [5, 10, 15, 20, 25, 50, 100]$.



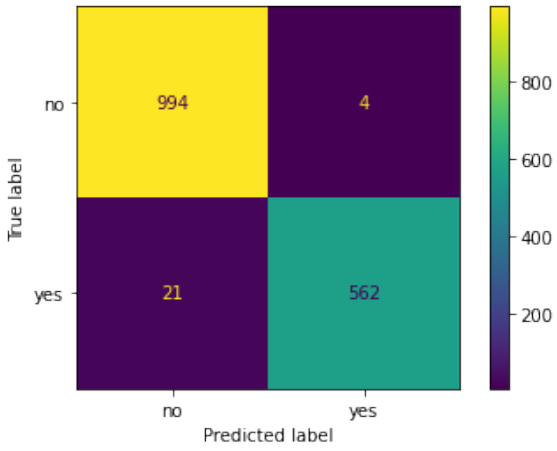
(a) FNR



(b) FPR

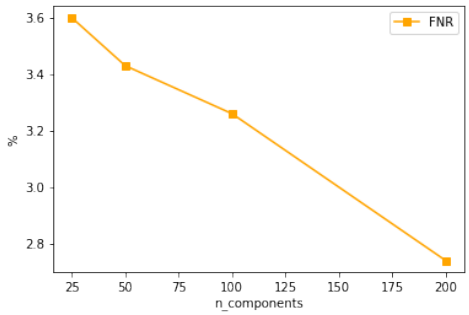
Klasyfikator uzyskał najlepsze wyniki dla wysokości równej **25**.

- FNR - 3.6%
- FPR - 0.4%

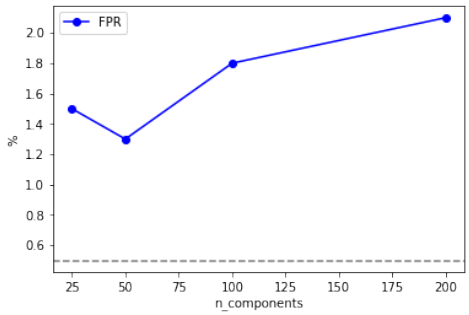


4.2 Redukcja wymiarowości danych

Do badania zostały użyte klasyfikatory z maksymalną wysokościom równą 25. W celu redukcji wymiarowości wykorzystano algorytm PCA, zbadany pod kątem liczby komponentów n ∈ [25, 50, 100, 200].

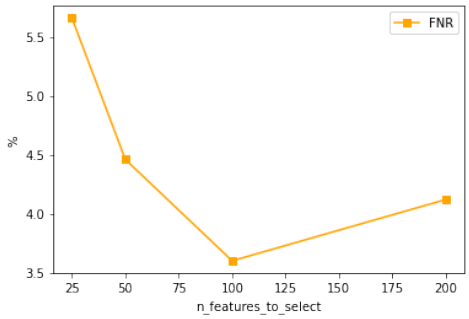


(a) FNR

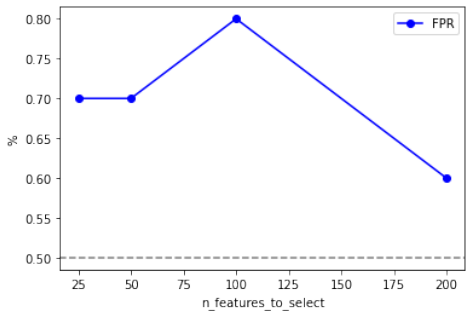


(b) FPR

Drugim algorytmem był RFE w którym za argument liczby cech zostały podane analogiczne wartości do PCA.



(a) FNR

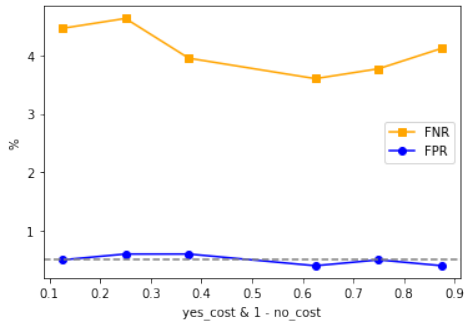


(b) FPR

W obu przypadku żaden wynik nie zmieścił się w docelowym przedziale FPR poniżej 0.5 %.

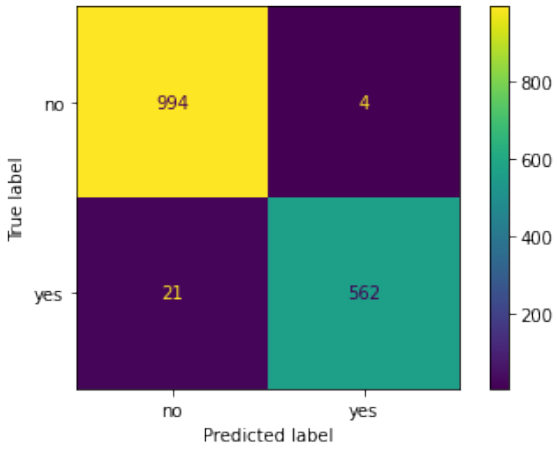
4.3 Zastosowanie niesymetrycznych kosztów błędów

Kolejnym aspektem było wykorzystanie niesymetrycznych kosztów błędów. W tym celu wykorzystano parametr class_weight. Badanie zostało przeprowadzone na sześciu lustrzanych parach klas 'yes' oraz 'no'. Wagi prezentują się następująco (0.125 - 0.875, 0.25 - 0.75, 0.375 - 0.635, 0.635 - 0.375 itd.)



Najlepsze rezultaty osiągnięto dla wagi 'yes' równej 0.625 oraz 'no' równej 0.375, wynik ten był równy temu z części 4.1.

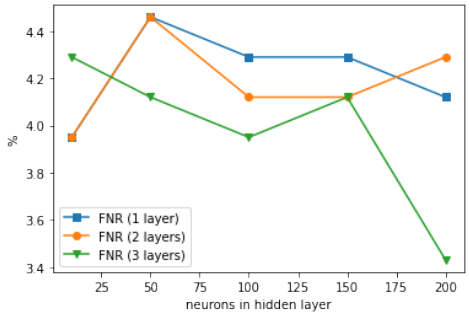
- FNR - 3.6%
- FPR - 0.4%



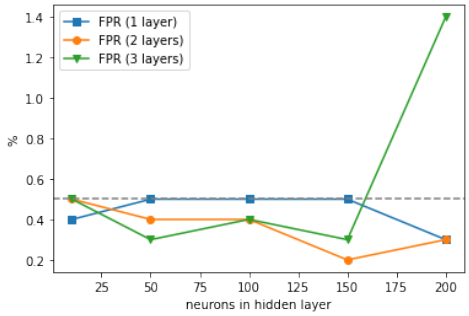
5 MLP

5.1 Parametry klasyfikatora

W przypadku algorytmu MLP badanymi parametrami były liczba neuronów w warstwie ukrytej ∈ [10, 50, 100, 150, 200] oraz liczba warstw ukrytych ∈ [1, 2, 3]



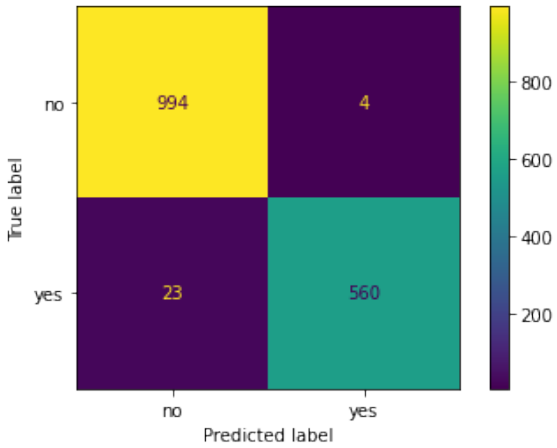
(a) FNR



(b) FPR

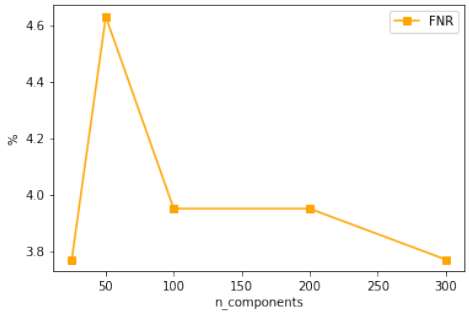
Klasyfikator uzyskał najlepsze wyniki dla **trzech warstw ukrytch** z liczbą neuronów równą **100**.

- FNR - 3.95%
- FPR - 0.4%

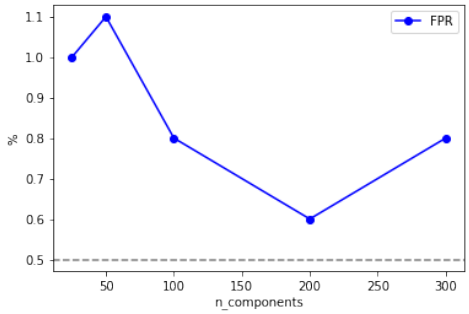


5.2 Redukcja wymiarowości danych

Do kolejnego badania został użyty klasyfikator z liczbą warstw ukrytych równą 3 i liczbą neuronów w każdej warstwie równą 100. W celu redukcji wymiarowości wykorzystano algorytm PCA, zbadany pod kątem liczby komponentów n $\in$ [25, 50, 100, 200, 300].



(a) FNR



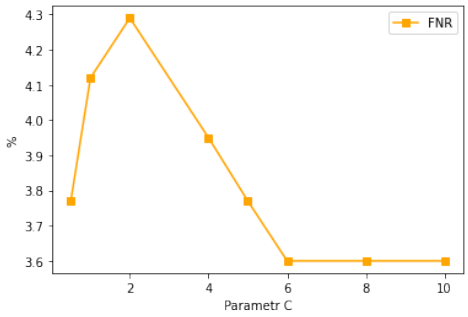
(b) FPR

Dzięki redukcji wymiarowości działanie klasyfikatora w kontekście wskaźnika FNR polepszyło się nieznacznie w dwóch przypadkach i wyniosło 3.77%, aczkolwiek we wszystkich wskaźnik FPR wyniósł powyżej 0.5% co dyskwalifikuje otrzymane wyniki.

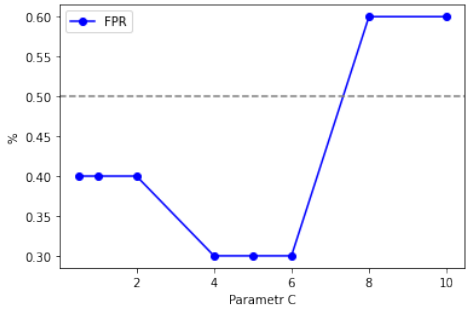
6 SVM

6.1 Parametry klasyfikatora

W przypadku algorytmu SVM zbadany został parametr C $\in$ [0.5, 1, 2, 4, 5, 6, 8, 10]



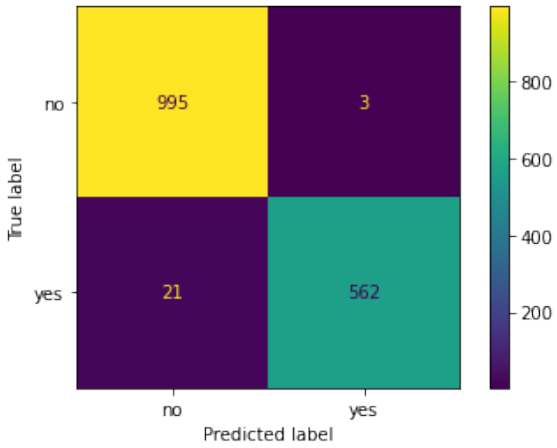
(a) FNR



(b) FPR

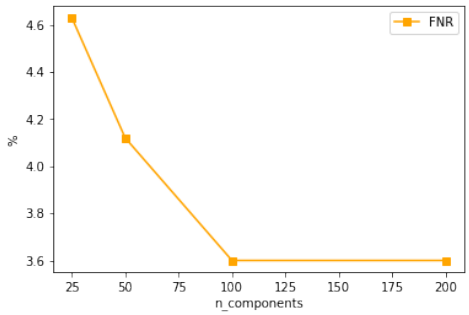
Klasyfikator uzyskał najlepsze wyniki dla $\mathbf{C = 6}$.

- FNR - 3.6%
- FPR - 0.3%

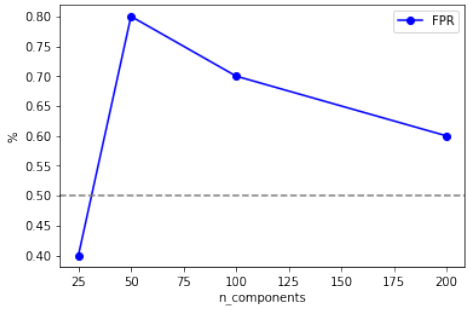


6.2 Redukcja wymiarowości danych

Do badania został użyty klasyfikator z parametrem C równym 6. W celu redukcji wymiarowości wykorzystano algorytm PCA, zbadany pod kątem liczby komponentów n ∈ [25, 50, 100, 200].

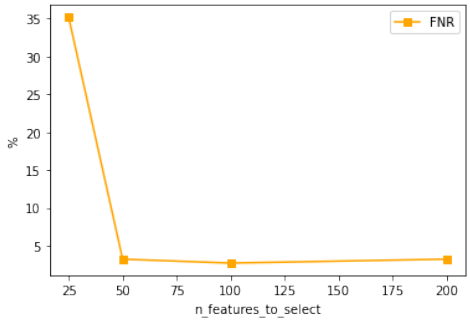


(a) FNR

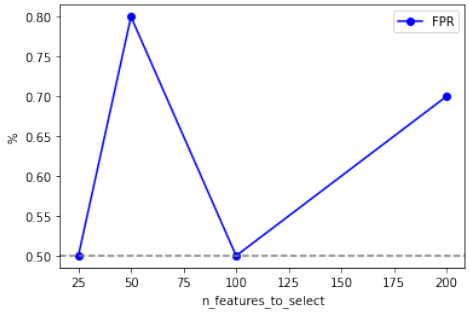


(b) FPR

Drugim algorytmem był RFE w którym za argument liczby cech zostały podane analogiczne wartości do PCA.

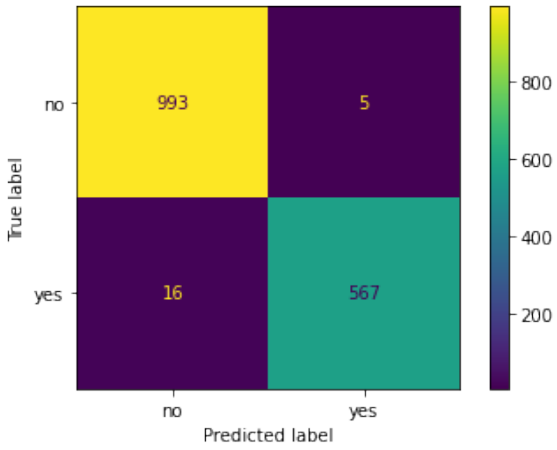


(a) FNR



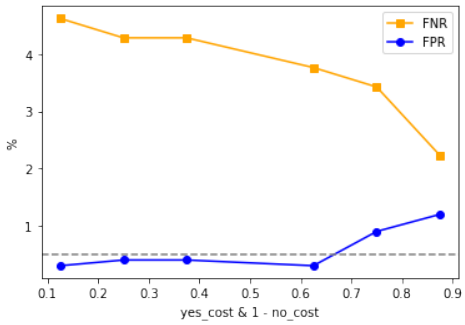
(b) FPR

W obu przypadkach redukcja cech spowodała wzrost wskaźnika FPR powyżej dopuszczalnej granicy. Aczkolwiek wartym odnotowania jest fakt, że algorytm RFE dla liczby cech 100 uzyskał wynik FNR równy jedyne 2.74 % przy FPR wynoszącym 0.5 %. W związku z tym jego macierz została umieszczoną poniżej, mimo przekroczenia progu.



6.3 Zastosowanie niesymetrycznych kosztów błędów

Kolejnym aspektem było wykorzystanie niesymetrycznych kosztów błędów. W tym celu wykorzystano parametr class_weight. Badanie zostało przeprowadzone na sześciu lustrzanych parach klas 'yes' oraz 'no'. Wagi prezentują się następująco (0.125 - 0.875, 0.25 - 0.75, 0.375 - 0.635, 0.635 - 0.375 itd.)



Najlepsze rezultaty osiągnięto dla wagi 'yes' równej 0.625 oraz 'no' równej 0.375, niemniej jednak wynik FPR równy 3.77 % jest nieznacznie gorszy od tego z części 6.1. Zwiększając dalej wagę klasy 'yes' wyniki wskaźnika FNR spadły poniżej 3.0 %, aczkolwiek spowodowało to przekroczenie ustalonej granicy FPR.

7 Podsumowanie

7.1 Porównanie algorytmów

W poniższej tabeli zostały wzięte pod uwagę najlepsze uzyskane wyniki przez każdy z dostrojonych klasyfikatorów.

Algorytm	FNR	FPR	parametry
kNN	11.66 %	0.4 %	liczba sąsiadów - 3, komponenty PCA - 100
Drzewo decyzyjne	6.0 %	0.4 %	max wys. drzewa - 8, cechy RFE - 25/50
Las losowy	3.6 %	0.4 %	max wys. drzew - 25, 'yes' - 0.625, 'no' - 0.375
MLP	3.95 %	0.4 %	warstwy ukryte - 3, po 100 neuronów
SVM	3.6 %	0.3 %	parametr C - 6

7.2 Wnioski

Uzyskanie samej wysokiej czułość (niska wartość FPR) jest zadaniem osiągalnym przez wszystkie algorytmy, niemniej jednak zwiększanie specyficzności (minimalizacja FNR) powoduje utratę czułości. Dlatego też, zachowanie odpowiedniego balansu okazało się dla niektórych metod trudniejsze niż dla pozostałych.

Algorytm k najbliższych sąsiadów okazał się zdecydowanie słabszy od pozostałych, gdyż wskaźnik FNR w jego przypadku był wyższy niż 10 %. Algorytm drzewa decyzyjnego był z kolei jedynym z pozostałych, który nie był w stanie pokonać granicy 5 %.

Pozostałe trzy - las losowy, MLP i SVM uzyskały zbliżone do siebie wyniki, które można uznać za satysfakcjonujące. W ich przypadku możliwa byłaby minimalizacji wskaźnika FNR do wartości bardzo niskich, aczkolwiek wiązałoby się to z przekroczeniem granicy FPR ustalonej jako 0.5 %.