Szymon Gruda	239661	239661@edu.p.lodz.pl
Jan Karwowski	239671	239671@edu.p.lodz.pl
Michał Kidawa	239673	239673@edu.p.lodz.pl
Kamil Kowalewski	239676	239676@edu.p.lodz.pl

Zadanie 3.: Problem Set 3

Spis treści

1. Wprowadzenie

1.1. Opis zbiorów danych

1.1.1. Zbiór Heart Disease UCI

— **age** – wiek w latach
— **sex** – płeć, gdzie 1 to mężczyzna a 0 to kobieta
— **cp (chest-pain-type)** – rodzaj bólu w klatce piersiowej, przyjmuje wartość 0, 1, 2 lub 3
— **trestbps (resting-blood-pressure)** – ciśnienie krwi w czasie spoczynku (w mm/Hg przy przyjęciu do szpitala)
— **chol (serum-cholestoral)** – cholesterol w surowicy w mg/dl
— **fbs (fasting-blood-sugar)** – poziom cukru we krwi na czco, przyjmuje wartość 1 dla poziomu większego niż 120 mg/dl, lub wartość 0 dla poziomu mniejszego
— **restecg (resting-electrocardiographic)** – wyniki eloktrokardiografu w spoczynku, przyjmuje wartość 0, 1 lub 2 stanie
— **thalach (maximum-heart-rate)** – najwyższe osiągnięte tętno
— **exang (exercise-induced-angina**) – dławica wysiłkowa, przyjmuje wartość 1, jeżeli dławica występuje, w przeciwnym razie przyjmuje wartość 0
— **oldpeak** – Obniżenie odcinka ST, wywołane przez ćwiczenie, w stosunku do odpoczynku
— **slope (the-slope-of-the-peak-exercise)** – nachylenie szczytowe odcinka ST podczas wysiłku, przyjmuje wartość 0, 1 lub 2
— **ca (number-of-major-vessels)** – liczba głównych naczyń, przyjmuje wartość 0, 1, 2, 3 lub 4
— **thal** – przyjmuje wartość 0, 1, 2 lub 3
— **target** – przyjmuje wartość 0 lub 1

1.1.2. Zbiór Gestures

Zbiór posiada 65 kolumn w tym 8*8 pomiarów z sensorów. Dana osoba miała umieszczonych 8 czujników i każdy z nich zwracał 8 wartości co daje 64 kolumny oraz ostatnia 65 kolumnę czyli *GESTURE_CLASS*

1.1.3. Zbiór Weather

— **Date** – numer dnia roku wykonania pomiaru
— **Location** – miejsce pomiaru zamienione na liczbe
— **MinTemp** – minimalna tempeteratura
— **MaxTemp** – maksymalna tempeteratura
— **Rainfall** – opad deszczu w mm
— **Evaporation** – parowanie
— **Sunshine** – liczba godzin, w których swieci słońce
— **WindGustDir** – kierunek wiatru
— **WindGustSpeed** – siła wiatru
— **WindDir9am** – kierunek wiatru o godzinie 9 rano
— **WindDir3pm** – kierunek wiatru o godzinie 15

— **WindSpeed9am** – siła wiatru o godzinie 9 rano
— **WindSpeed3pm** – siła wiatru o godzinie 15
— **Humidity9am** – wilgotność o godzinie 9 rano
— **Humidity3pm** – wilgotność o godzinie 15
— **Pressure9am** – ciśnienie atmosferyczne o godzinie 9 rano
— **Pressure3pm** – ciśnienie atmosferyczne o godzinie 15
— **Cloud9am** – zachmurzenie o godzinie 9 rano
— **Cloud3pm** – zachmurzenie o godzinie 15
— **Temp9am** – temperatura o godzinie 9 rano
— **Temp3pm** – temperatura o godzinie 15
— **RainToday** – wartość logiczna - jeśli opady do godziny 9 rano były powyżej 1mm to wartość 1 jeśli nie to 0
— **RainTomorrow** – target variable

1.2. K-najbliższych sąsiadów

Zadanie klasyfikacji, polegające na przydzielaniu obiektów do wcześniej zdefiniowanych grup można rozwiązać na różne sposoby. Powstało wiele algorytmów klasyfikujących, a spośród nich jednym z najprostszych co do zasady działania jest algorytm k-NN (*ang. k-nearest neighbors*). Klasyfikator ten nie wymaga procesu uczenia, zbiór uczący jest jedynie przechowywany w pamięci programu a przetwarzany jest dopiero podczas właściwej klasyfikacji. Działanie tego algorytmu polega na znajdowaniu najbliższych (zgodnie z pewną miarą) obiektów ze zbioru uczącego dla każdego elementu ze zbioru testowego. Następnie klasa przetwarzanego obiektu zostaje rozpoznana jako najczęstsza spośród znalezionych wcześniej "sąsiadów". Algorytm można więc przedstawić w następujących krokach:
1. Weź jeden element ze zbioru testowego
2. Znajdź k najbliższych elementów ze zbioru uczącego
3. Wybierz klasę najczęściej występującą pośród znalezionych elementów
4. Wybrana klasa jest klasą rozpoznawanego elementu ze zbioru testowego

1.3. Naiwny klasyfikator Bayesa

Naiwny klasyfikator Bayesa wykorzystuje w procesie klasyfikacji twierdzenie Bayesa:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

gdzie:
— A, B – zdarzenia
— $P(A \mid B)$ – prawdopodobieństwo zdarzenia A, o ile zajdzie B
— $P(B \mid A)$ – prawdopodobieństwo zdarzenia B, o ile zajdzie A
— $P(A)$ – prawdopodobieństwo wystąpienia zdarzenia A
— $P(B)$ – suma prawdopodobieństw wszystkich potencjalnych skutków zdarzenia: $P(B) = \sum P(B \mid A)P(B)$

Model naiwnego klasyfikatora Bayesa zakłada, że cecha danej próbki jest niepowiązana z pozostałymi cechami i wskazuje na prawdopodobieństwo przynależności do danej klasy.

Przy tym klasyfikatorze jedynym dostępnym parametrem był podział na dane treningowe i testowe. Dokładność klasyfikacji sprawdzono dla każdego ze zbiorów danych w zależności od podziału na zbiór uczący i testowy.

1.4. Maszyna wektorów nośnych

Maszyna wektorów nośnych - SVM (ang. Support Vector Machines) jako klasyfikator danych wykorzystuje przestrzeń decyzyjną, którą dzieli się budując granice separujące obiekty o różnej przynależności klasowej. Warunkiem jest to aby dzielone zbiory były liniowo separowalne, niestety jest rzadko spotykane – tylko trywialne przykłady zbiorów charakteryzują się tą własnością. Na szczęście wykorzystanie funkcji jądrowych (ang. kernel functions) umożliwia przekształcenie każdego zbioru tak, aby charakteryzował się on liniową separalnością danych. W takim wypadku można poprowadzić nieskończenie wiele płaszczyzn dzielących klasy. Aby wybrać konkretną stosuję się metodę maksymalizacji marginesu – dąży się do tego aby najbliższe dla hiperpłaszczyzny obiekty z danego obszaru były jak najbardziej od niej oddalone. Podstawowa implementacja maszyny wektorów nośnych uniemożliwia znalezienie się obiektu klasy A, po stronie hiperpłaszczyzny, na której znajdują się obiekty klasy B. Takie zachowanie powoduje pewne problemy, jak np. nadmierne dopasowanie modelu. Rozwiązaniem tego problemu jest nadanie odpowiedniej wartości parametrowi C, który jest odpowiedzialny za tolerancję zjawiska, w którym niektóre obiekty znajdują się po nieprawidłowej stronie hiperpłaszczyzny. Natomiast wartość parametru gamma (dostępnego tylko dla niektórych funkcji jądra) określa promień pojedynczego punktu, w jakiego odległości musi się znaleźć inny punkt, aby zostały one zgrupowane razem.

1.5. Drzewa decyzyjne i lasy losowe

Drzewa decyzyjne to bardzo prosty i jednocześnie skuteczny klasyfikator. Na podstawie zbioru danych uczących buduje się drzewo, gdzie każdy węzeł pośredni zawiera pewien warunek (cechę i jej wartość graniczną), według którego dzieli się zbiór uczący na dwie części (w przypadku biblioteki scikit-learn są to drzewa binarne). Następnie każdą taką część znowu dzieli się na kolejne wybierając kolejny warunek. Na każdym etapie podziały robi się w ten sposób, aby powstałe podzbiory zbioru uczącego były według pewnej miary jak najmniej zanieczyszczone, czyli aby podziały były na każdym etapie jak najlepsze. Węzły końcowe (liście) oznaczają konkretne klasy. Drzewo takie można budować aż do osiągnięcia idealnie czystych podzbiorów, co w najgorszym przypadku może oznaczać jeden liść na jeden przykład uczący, co z kolei wiąże się ze skrajnym przetrenowaniem modelu. Z tego powodu w przypadku uczenia drzew decyzyjnych zawsze wprowadza się parametry regularyzacyjne, takie jak maksymalna wysokość drzewa czy minimalna liczba przykładów uczących w liściu. Tak zbudowane drzewo decyzyjne można wykorzystać do rozpoznawania klasy nieznanego przykładu, poprzez "przejście" od korzenia drzewa aż do któregoś liścia, zgodnie z wartościami cech danego przykładu i warunkami w węzłach pośrednich.

Las losowy jest niczym innym jak zespołem drzew, gdzie każde może być wyuczone wykorzystując jedynie pewien podzbiór zbioru uczącego. W ten sposób powstaje pewna liczba klasyfikatorów, każdy trochę inny, popełniający inne błędy. W celu znalezienia klasy zadanego przykładu wybiera się tę, która jest najczęściej wskazywana przez wszystkie drzewa w lesie. W ten sposób uzyskuje się znacznie bardziej uniwersalny i model, lepiej generalizujący i lepiej radzący sobie w przypadku zbiorów z dużą wariancją.

2. Wyniki

2.1. K-najbliższych sąsiadów

2.1.1. Metryka Euklidesowa

K	Accuracy
1	0.6264
2	0.5824
3	0.6703
4	0.6703
5	0.6593
6	0.6374
7	0.6813
8	0.7143
9	0.7143
10	0.7253
11	0.7033
12	0.6703
13	0.6593
14	0.6374
15	0.6484
16	0.6484
17	0.6703
18	0.6813
19	0.6703
20	0.6593
21	0.6923
22	0.6813
23	0.7033
24	0.6593
25	0.6593
26	0.6593
27	0.6813
28	0.6813
29	0.6923

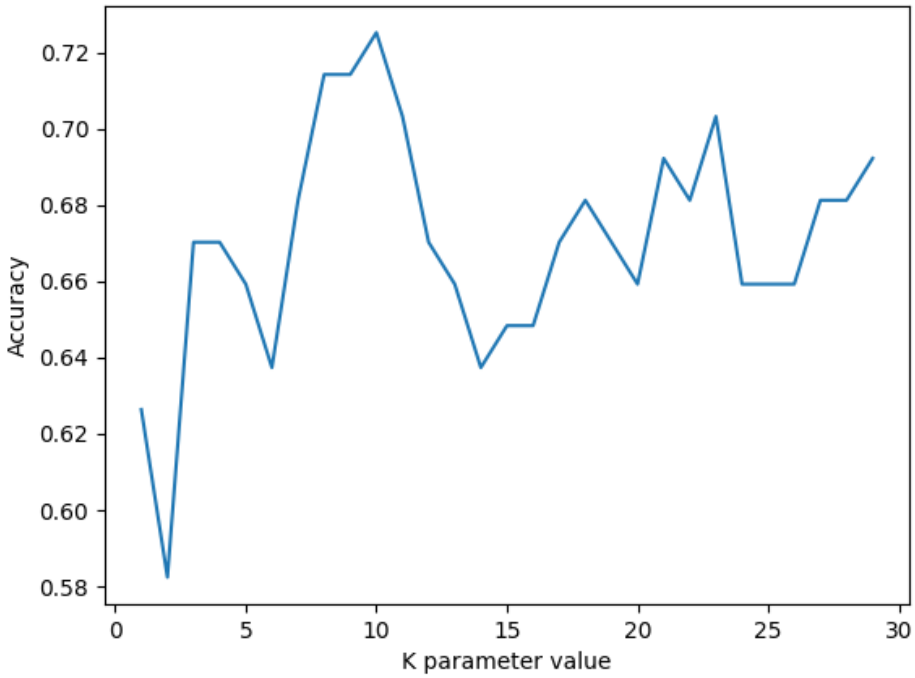
Tabela 1. Zbiór heart disease

K	Accuracy
1	0.6644
2	0.6655
3	0.6835
4	0.6855
5	0.6781
6	0.6812
7	0.6812
8	0.6838
9	0.6821
10	0.6807
11	0.6801
12	0.6738
13	0.6801
14	0.6747
15	0.6741
16	0.6712
17	0.6712
18	0.6701
19	0.6698
20	0.6701
21	0.6624
22	0.6621
23	0.6572
24	0.657
25	0.6587
26	0.6558
27	0.6592
28	0.6584
29	0.6587

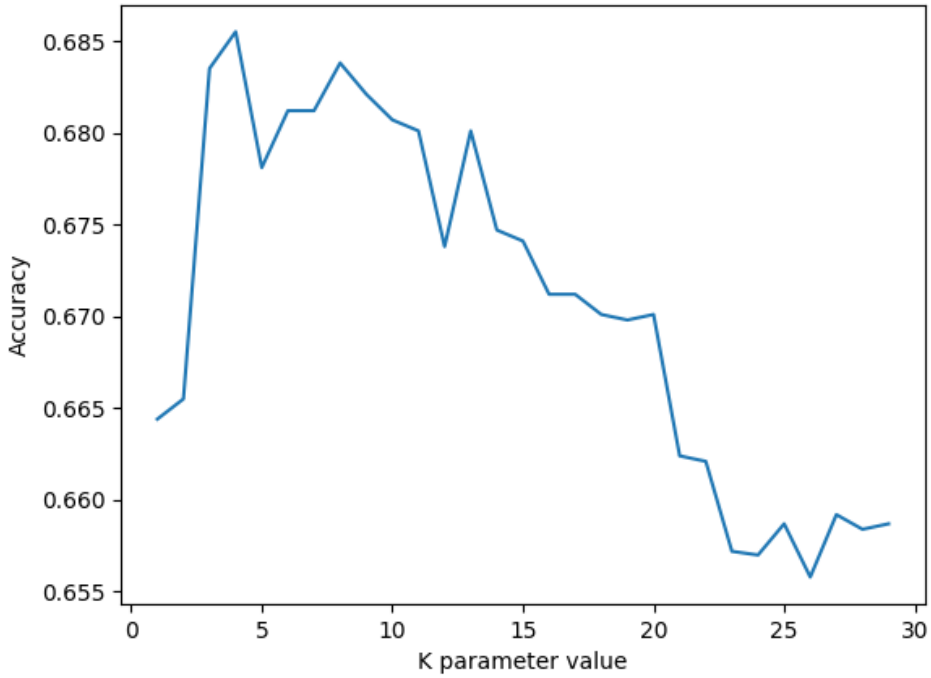
Tabela 2. Zbiór gestures

K	Accuracy
1	0.8117
2	0.8284
3	0.8348
4	0.8394
5	0.8434
6	0.8447
7	0.8472
8	0.8457
9	0.849
10	0.847
11	0.8505
12	0.8503
13	0.8516
14	0.8508
15	0.8525
16	0.85
17	0.8514
18	0.8499
19	0.8509
20	0.8496
21	0.8509
22	0.8495
23	0.8514
24	0.8502
25	0.8526
26	0.8503
27	0.8524
28	0.8511
29	0.8524

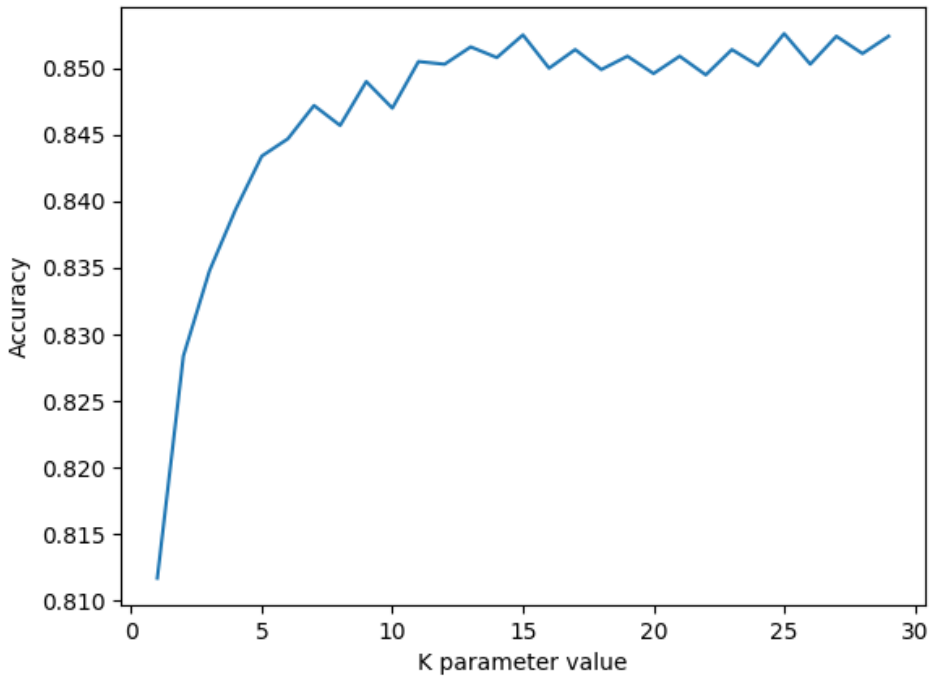
Tabela 3. Zbiór weather AUS



Rysunek 1. Zbiór heart disease



Rysunek 2. Zbiór gestures



Rysunek 3. Zbiór weather AUS

2.1.2. Metryka Manhattan

K	Accuracy
1	0.6044
2	0.6703
3	0.7253
4	0.7363
5	0.7363
6	0.7253
7	0.7363
8	0.7143
9	0.6923
10	0.7033
11	0.7033
12	0.6923
13	0.7033
14	0.7033
15	0.7253
16	0.7033
17	0.6813
18	0.6813
19	0.7143
20	0.6923
21	0.7143
22	0.7253
23	0.7033
24	0.6923
25	0.6813
26	0.6813
27	0.6813
28	0.6593
29	0.6923

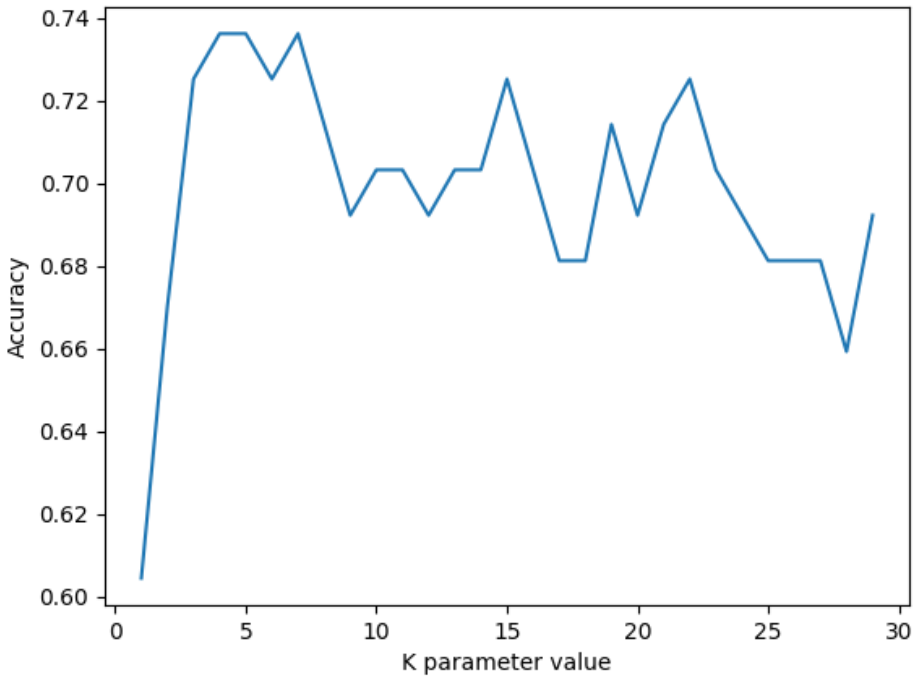
Tabela 4. Zbiór heart disease

K	Accuracy
1	0.6555
2	0.6504
3	0.6667
4	0.6564
5	0.6764
6	0.6672
7	0.6741
8	0.6675
9	0.6692
10	0.6672
11	0.6729
12	0.663
13	0.6627
14	0.6604
15	0.6604
16	0.6587
17	0.6564
18	0.6564
19	0.6558
20	0.6515
21	0.6507
22	0.6484
23	0.6478
24	0.6473
25	0.6484
26	0.6404
27	0.6424
28	0.6396
29	0.6418

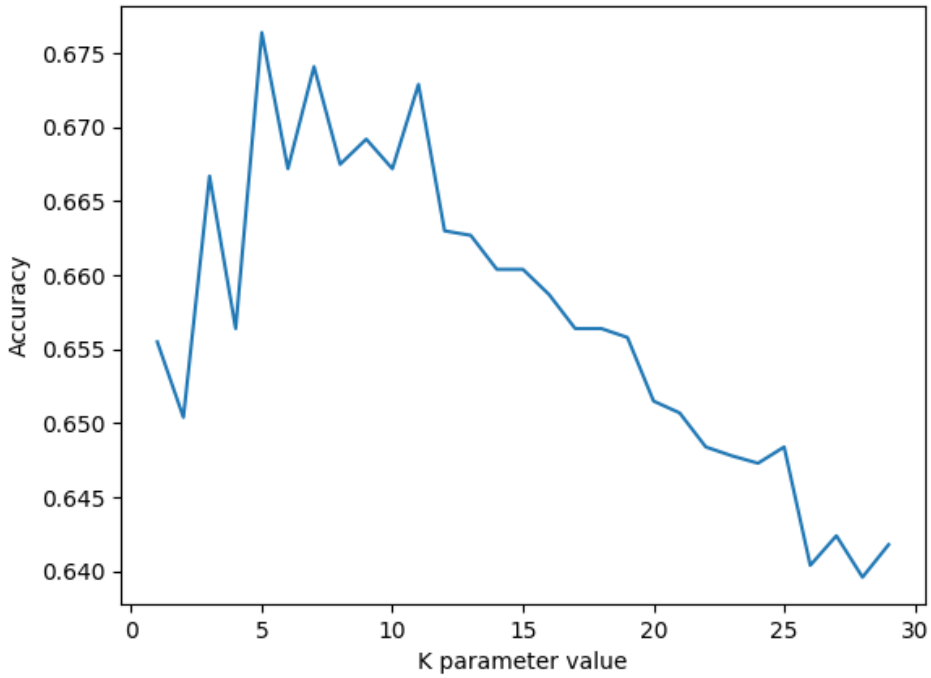
Tabela 5. Zbiór gestures

K	Accuracy
1	0.8224
2	0.8351
3	0.8424
4	0.8421
5	0.847
6	0.8446
7	0.8516
8	0.8498
9	0.8535
10	0.8505
11	0.8532
12	0.8518
13	0.854
14	0.8516
15	0.8535
16	0.8522
17	0.8546
18	0.8522
19	0.8542
20	0.8528
21	0.8545
22	0.8541
23	0.855
24	0.8537
25	0.8542
26	0.8528
27	0.854
28	0.8527
29	0.8534

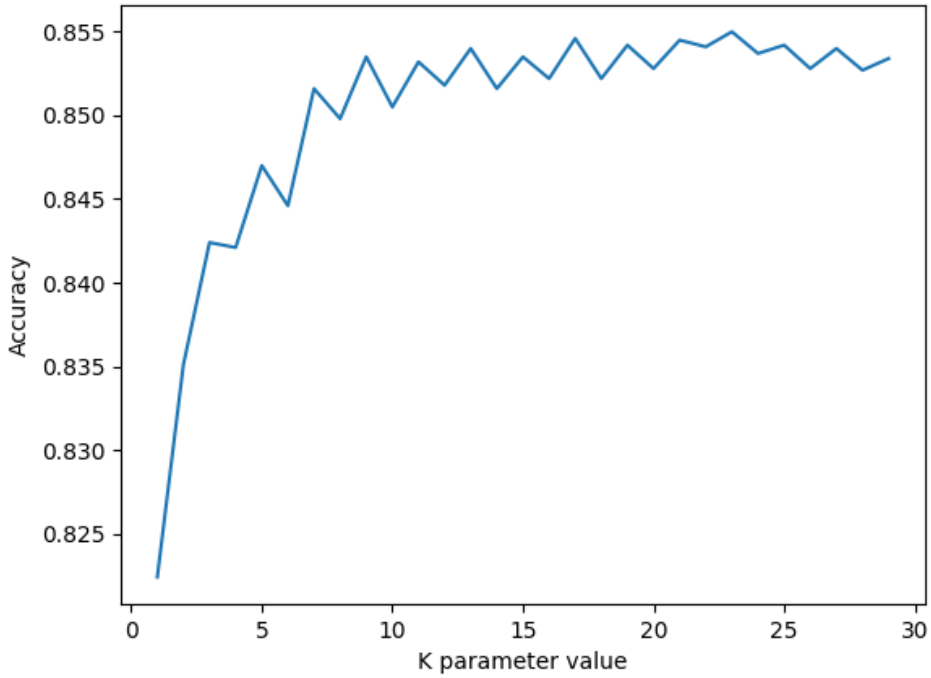
Tabela 6. Zbiór weather AUS



Rysunek 4. Zbiór heart disease



Rysunek 5. Zbiór gestures



Rysunek 6. Zbiór weather AUS

2.2. Naiwny klasyfikator Bayesa

Na początku wśród dostępnych wariantów została wybrana metoda, która najlepiej poradzi sobie z dostarczonymi zbiorami danych. Do eksperymentu został przyjęty podział 70% danych uczących i 30% testowych.

wariant klasyfikatora	**Zbiór danych**		
	heart diseases	**gestures**	**weather**
multinomial	0.7473	-	-
bernoulli	0.8022	0.4632	0.7655
gaussian	0.8132	0.8779	0.8004

Tabela 7. Porównanie dokładności dostępnych wariantów klasyfikatora Bayesa

Wybrane zbiory danych *gestures* oraz *weather* zawierają wartości ujemne i nie mogą zostać użyte w przypadku wariantu wielomianowego. Na podstawie wyników wybrany został wariant Gaussa, jako metoda, która najlepiej poradziła sobie z danymi testowymi.

W przypadku eksperymentów wykorzystujących naiwny klasyfikator Bayesa jako parametr został przyjęty podział na dane uczące i testowe.

% danych testowych	**Zbiór danych**		
	heart diseases	**gestures**	**weather**
30 %	0.8132	0.8779	0.8004
35 %	0.8037	0.8826	0.7996
40 %	0.8033	0.8859	0.8011
45 %	0.7737	0.883	0.7991
50 %	0.7961	0.8853	0.8002
55 %	0.7784	0.8804	0.802
60 %	0.7747	0.8804	0.8029
65 %	0.7543	0.8803	0.8032
70 %	0.7793	0.8802	0.8067
75 %	0.7851	0.8842	0.8063
80 %	0.786	0.884	0.8048
85 %	0.7907	0.8816	0.8064
90 %	0.7033	0.8826	0.8048

Tabela 8. Porównanie dokładności dla różnych zbiorów, dla naiwnego klasyfikatora Bayesa

2.3. Maszyna wektorów nośnych

W eksperymentach badających dokładność klasyfikatora opartego o maszynę wektorów nośnych sprawdzanie było zachowanie klasyfikatora (dokładność klasyfikacji) dla trzech różnych funkcji jądra:

1. wielomianowej - w skrócie poly
2. radialnej funkcji bazowej (ang. radial basis function) - w
3. skrócie RBF
4. sigmoidalnej - w skrócie sigmoid

Dla każdej funkcji zmieniano wartości parametru C, lub gamma.

C	Accuracy
0.1	0.5275
0.2	0.6264
0.3	0.6484
0.4	0.6813
0.5	0.6813
0.6	0.6703
0.7	0.6813
0.8	0.6813
0.9	0.6923
1.0	0.6923
1.1	0.6923
1.2	0.6923
1.3	0.7033
1.4	0.7143
1.5	0.7143
1.6	0.7253
1.7	0.7253
1.8	0.7253
1.9	0.7253
2.0	0.7253

Tabela 9. Zbiór heart funkcja - wielomian

C	Accuracy
0.1	0.5275
0.2	0.5275
0.3	0.5275
0.4	0.5714
0.5	0.6044
0.6	0.6154
0.7	0.6593
0.8	0.6593
0.9	0.6484
1.0	0.6593
1.1	0.6593
1.2	0.6813
1.3	0.6813
1.4	0.6813
1.5	0.6813
1.6	0.6813
1.7	0.6923
1.8	0.6923
1.9	0.6923
2.0	0.6813

Tabela 10. Zbiór heart funkcja - RBF

C	Accuracy
0.1	0.5275
0.2	0.5275
0.3	0.5275
0.4	0.5275
0.5	0.5275
0.6	0.5275
0.7	0.5275
0.8	0.5275
0.9	0.5275
1.0	0.5275
1.1	0.5275
1.2	0.5275
1.3	0.5275
1.4	0.5275
1.5	0.5385
1.6	0.5604
1.7	0.5495
1.8	0.5824
1.9	0.6044
2.0	0.6374

Tabela 11. Zbiór weather funkcja - Sigmoidalna

Gamma	Accuracy
1e-20	0.5275
1e-19	0.5275
1e-18	0.5275
1e-17	0.5275
1e-16	0.5275
1e-15	0.5275
1e-14	0.5275
1e-13	0.5275
1e-12	0.5275
1e-11	0.5275
1e-10	0.5275
1e-09	0.5275
1e-08	0.5275
1e-07	0.5275
1e-06	0.5275
1e-05	0.6813
0.0001	0.8242
0.001	0.7582
0.01	0.7473
0.1	0.7802

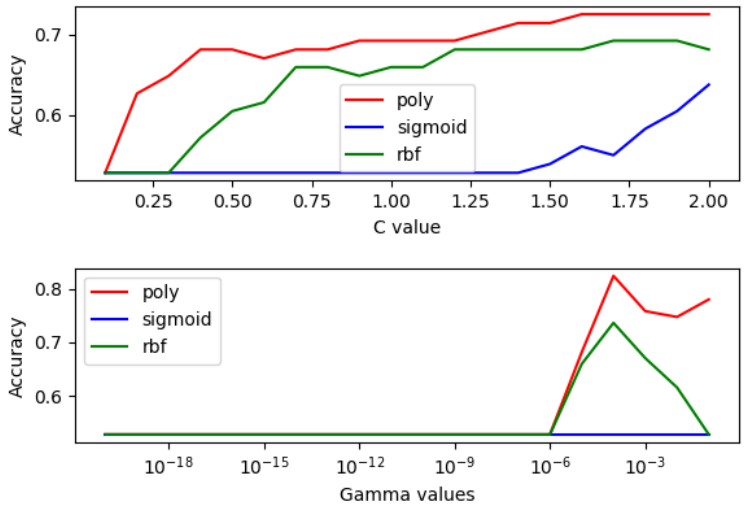
Tabela 12. Zbiór heart funkcja - wielomian

Gamma	Accuracy
1e-20	0.5275
1e-19	0.5275
1e-18	0.5275
1e-17	0.5275
1e-16	0.5275
1e-15	0.5275
1e-14	0.5275
1e-13	0.5275
1e-12	0.5275
1e-11	0.5275
1e-10	0.5275
1e-09	0.5275
1e-08	0.5275
1e-07	0.5275
1e-06	0.5275
1e-05	0.6593
0.0001	0.7363
0.001	0.6703
0.01	0.6154
0.1	0.5275

Tabela 13. Zbiór heart funkcja - RBF

Gamma	Accuracy
1e-20	0.5275
1e-19	0.5275
1e-18	0.5275
1e-17	0.5275
1e-16	0.5275
1e-15	0.5275
1e-14	0.5275
1e-13	0.5275
1e-12	0.5275
1e-11	0.5275
1e-10	0.5275
1e-09	0.5275
1e-08	0.5275
1e-07	0.5275
1e-06	0.5275
1e-05	0.5275
0.0001	0.5275
0.001	0.5275
0.01	0.5275
0.1	0.5275

Tabela 14. Zbiór weather funkcja - Sigmoidalna



Rysunek 7. Wykres dokładności klasyfikacji, w zależności od wartości parametrów C i gamma, dla zbioru heart

C	Accuracy
0.1	0.3525
0.2	0.403
0.3	0.4389
0.4	0.4655
0.5	0.4951
0.6	0.514
0.7	0.5228
0.8	0.5274
0.9	0.5283
1.0	0.528
1.1	0.5265
1.2	0.5283
1.3	0.5297
1.4	0.5317
1.5	0.5305
1.6	0.5317
1.7	0.5305
1.8	0.5303
1.9	0.5308
2.0	0.5317

Tabela 15. Zbiór gestures funkcja - wielomian

C	Accuracy
0.1	0.718
0.2	0.7811
0.3	0.8074
0.4	0.8251
0.5	0.8353
0.6	0.8479
0.7	0.8522
0.8	0.8602
0.9	0.8653
1.0	0.8687
1.1	0.871
1.2	0.8727
1.3	0.8747
1.4	0.8764
1.5	0.8796
1.6	0.8816
1.7	0.8836
1.8	0.8856
1.9	0.8884
2.0	0.8907

Tabela 16. Zbiór gestures funkcja - RBF

C	Accuracy
0.1	0.2711
0.2	0.2854
0.3	0.276
0.4	0.2603
0.5	0.2489
0.6	0.2403
0.7	0.2357
0.8	0.226
0.9	0.216
1.0	0.2135
1.1	0.2126
1.2	0.212
1.3	0.2089
1.4	0.2049
1.5	0.2006
1.6	0.1983
1.7	0.1978
1.8	0.1986
1.9	0.1955
2.0	0.1938

Tabela 17. Zbiór gestures funkcja - Sigmoidalna

Gamma	Accuracy
1e-20	0.2457
1e-19	0.2457
1e-18	0.2457
1e-17	0.2457
1e-16	0.2457
1e-15	0.2457
1e-14	0.2457
1e-13	0.2457
1e-12	0.2457
1e-11	0.2457
1e-10	0.2457
1e-09	0.2457
1e-08	0.2457
1e-07	0.2457
1e-06	0.2457
1e-05	0.2623
0.0001	0.5508
0.001	0.5636
0.01	0.5636
0.1	0.5636

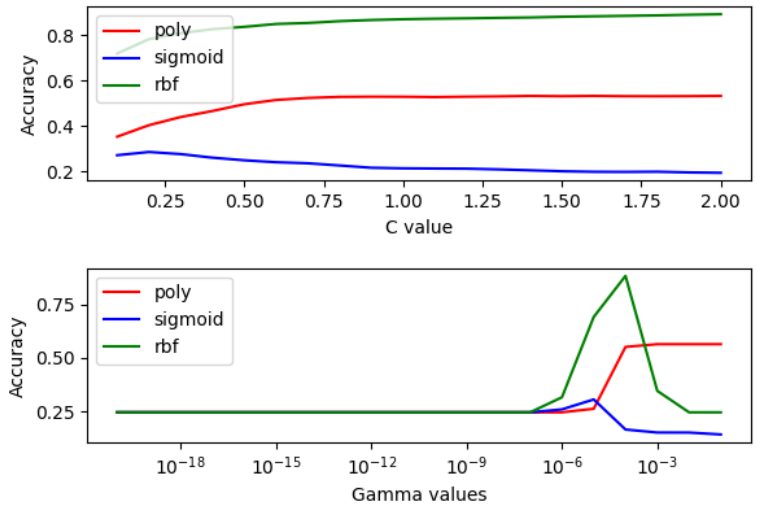
Tabela 18. Zbiór gestures funkcja - wielomian

Gamma	Accuracy
1e-20	0.2457
1e-19	0.2457
1e-18	0.2457
1e-17	0.2457
1e-16	0.2457
1e-15	0.2457
1e-14	0.2457
1e-13	0.2457
1e-12	0.2457
1e-11	0.2457
1e-10	0.2457
1e-09	0.2457
1e-08	0.2457
1e-07	0.2457
1e-06	0.3159
1e-05	0.6906
0.0001	0.8821
0.001	0.3459
0.01	0.2457
0.1	0.2457

Tabela 19. Zbiór gestures funkcja - RBF

Gamma	Accuracy
1e-20	0.2457
1e-19	0.2457
1e-18	0.2457
1e-17	0.2457
1e-16	0.2457
1e-15	0.2457
1e-14	0.2457
1e-13	0.2457
1e-12	0.2457
1e-11	0.2457
1e-10	0.2457
1e-09	0.2457
1e-08	0.2457
1e-07	0.2457
1e-06	0.2594
1e-05	0.3057
0.0001	0.1655
0.001	0.1515
0.01	0.1513
0.1	0.1424

Tabela 20. Zbiór gestures funkcja - Sigmoidalna



Rysunek 8. Wykres dokładności klasyfikacji, w zależności od wartości parametrów C i gamma, dla zbioru gestures

C	Accuracy
0.1	0.7781
0.2	0.802
0.3	0.8296
0.4	0.8384
0.5	0.842
0.6	0.8443
0.7	0.8446
0.8	0.8458
0.9	0.8461
1.0	0.8468
1.1	0.8476
1.2	0.8477
1.3	0.8484
1.4	0.8485
1.5	0.8486
1.6	0.8488
1.7	0.849
1.8	0.8491
1.9	0.8495
2.0	0.8496

Tabela 21. Zbiór weather funkcja - wielomian

C	Accuracy
0.1	0.7779
0.2	0.778
0.3	0.7807
0.4	0.7921
0.5	0.8101
0.6	0.8219
0.7	0.8306
0.8	0.8361
0.9	0.839
1.0	0.8399
1.1	0.8419
1.2	0.8433
1.3	0.8445
1.4	0.8454
1.5	0.8462
1.6	0.8463
1.7	0.8456
1.8	0.8456
1.9	0.8462
2.0	0.8463

Tabela 22. Zbiór weather funkcja - RBF

C	Accuracy
0.1	0.7779
0.2	0.7779
0.3	0.7779
0.4	0.7779
0.5	0.7779
0.6	0.7779
0.7	0.7779
0.8	0.778
0.9	0.778
1.0	0.778
1.1	0.778
1.2	0.7781
1.3	0.7785
1.4	0.7792
1.5	0.7799
1.6	0.7807
1.7	0.7821
1.8	0.7834
1.9	0.7848
2.0	0.7862

Tabela 23. Zbiór weather funkcja - Sigmoidalna

Gamma	Accuracy
1e-21	0.7779
1e-20	0.7779
1e-19	0.7779
1e-18	0.7779
1e-17	0.7779
1e-16	0.7779
1e-15	0.7779
1e-14	0.7779
1e-13	0.7779
1e-12	0.7779
1e-11	0.7779
1e-10	0.7779
1e-09	0.7779
1e-08	0.7779
1e-07	0.7779
1e-06	0.8515
1e-05	0.8548
0.0001	0.8559
0.001	0.8561
0.01	0.855

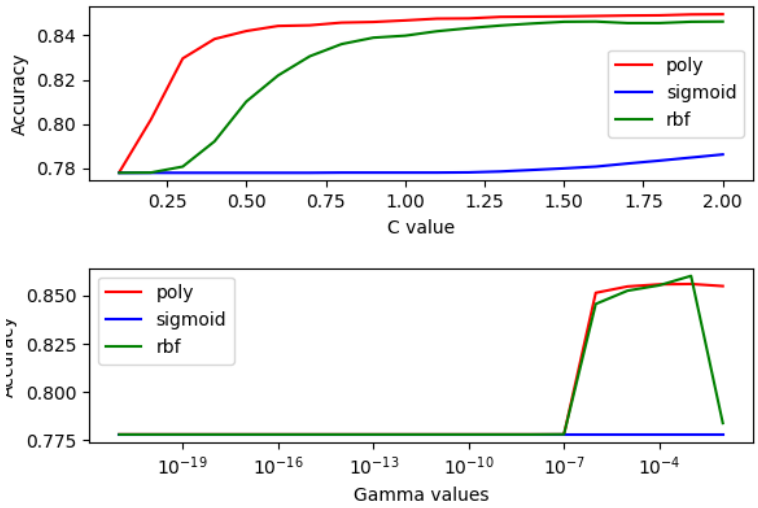
Tabela 24. Zbiór weather funkcja - wielomian

Gamma	Accuracy
1e-21	0.7779
1e-20	0.7779
1e-19	0.7779
1e-18	0.7779
1e-17	0.7779
1e-16	0.7779
1e-15	0.7779
1e-14	0.7779
1e-13	0.7779
1e-12	0.7779
1e-11	0.7779
1e-10	0.7779
1e-09	0.7779
1e-08	0.7779
1e-07	0.778
1e-06	0.8457
1e-05	0.8526
0.0001	0.8554
0.001	0.8603
0.01	0.7838

Tabela 25. Zbiór weather funkcja - RBF

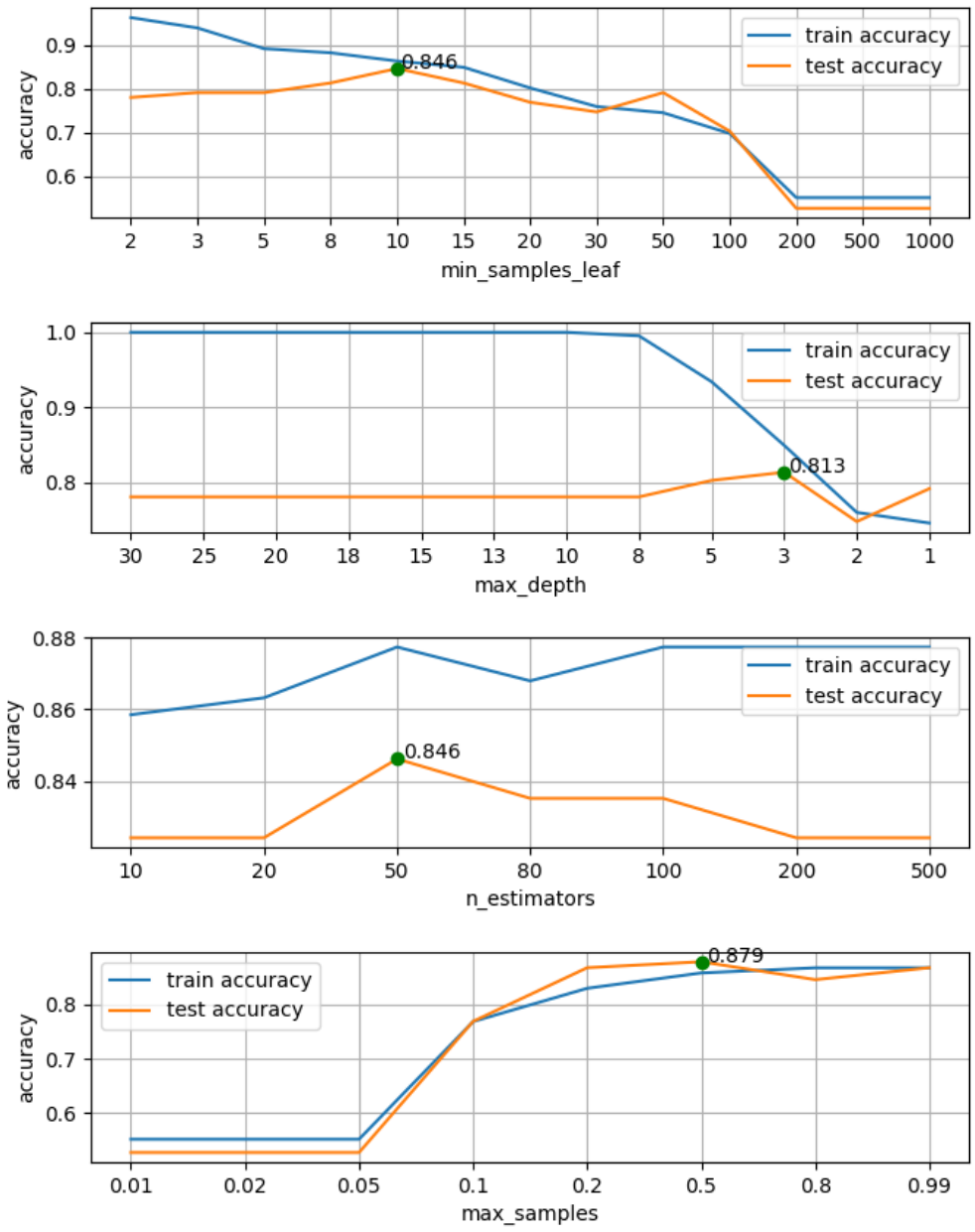
Gamma	Accuracy
1e-21	0.7779
1e-20	0.7779
1e-19	0.7779
1e-18	0.7779
1e-17	0.7779
1e-16	0.7779
1e-15	0.7779
1e-14	0.7779
1e-13	0.7779
1e-12	0.7779
1e-11	0.7779
1e-10	0.7779
1e-09	0.7779
1e-08	0.7779
1e-07	0.7779
1e-06	0.7779
1e-05	0.7779
0.0001	0.7779
0.001	0.7779
0.01	0.7779

Tabela 26. Zbiór weather funkcja - Sigmoidalna

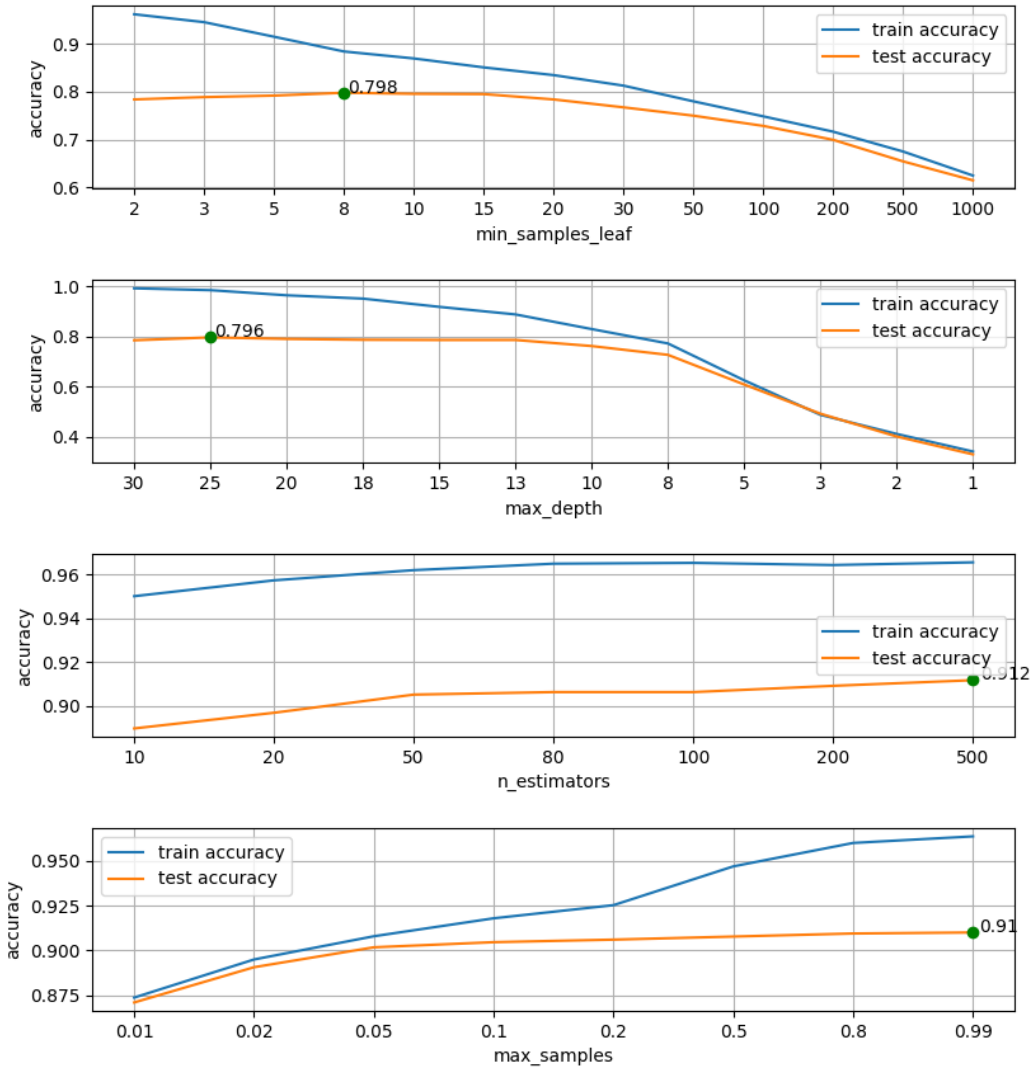


Rysunek 9. Wykres dokładności klasyfikacji, w zależności od wartości parametrów C i gamma, dla zbioru weather

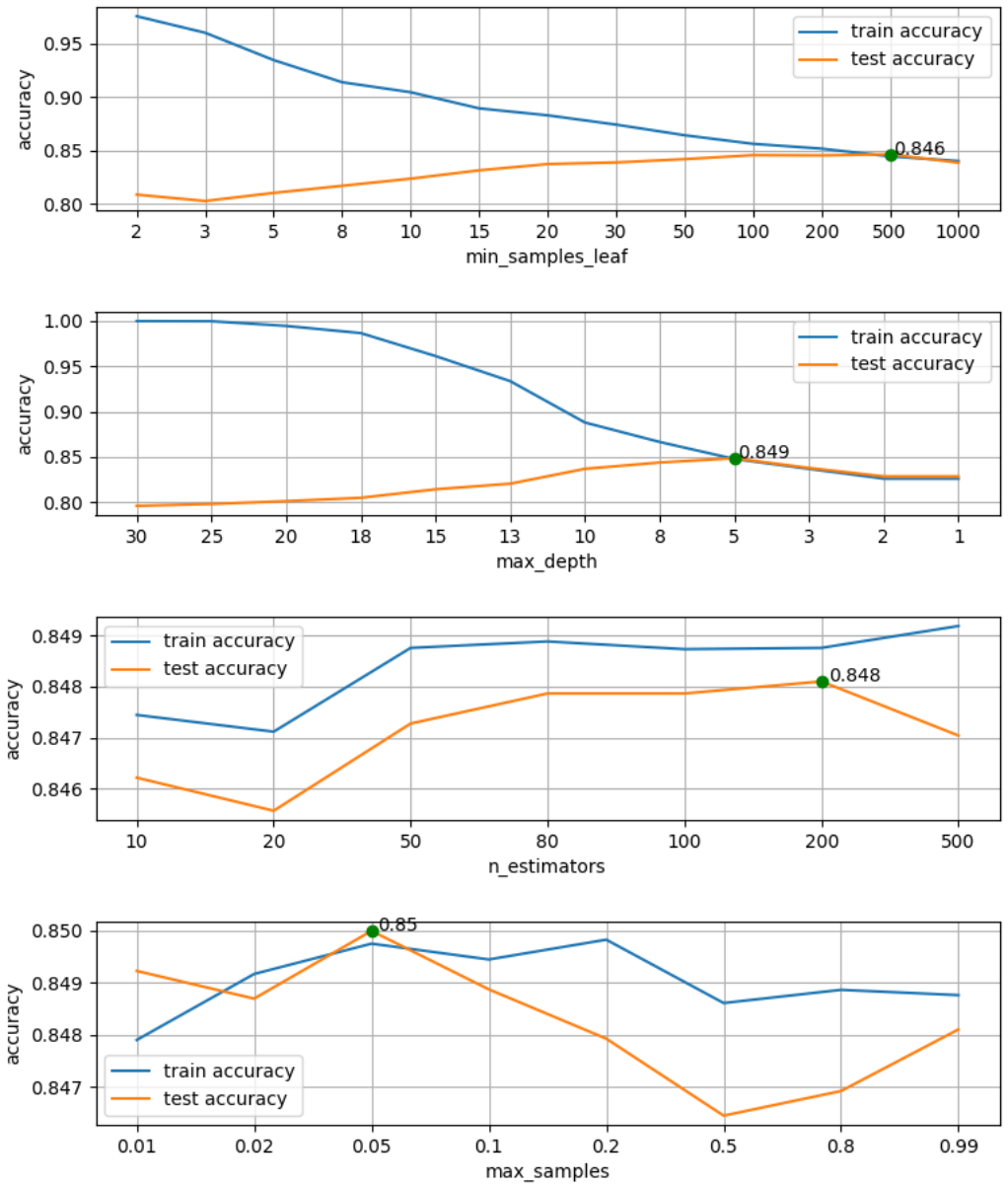
2.4. Drzewa decyzyjne i lasy losowe



Rysunek 10. Dokładność klasyfikacji drzewem decyzyjnym (dwa górne wykresy) i lasem losowym (dwa dolne wykresy) w zależności od parametrów dla zbioru *heart disease*



Rysunek 11. Dokładność klasyfikacji drzewem decyzyjnym (dwa górne wykresy) i lasem losowym (dwa dolne wykresy) w zależności od parametrów dla zbioru *gestures*



Rysunek 12. Dokładność klasyfikacji drzewem decyzyjnym (dwa górne wykresy) i lasem losowym (dwa dolne wykresy) w zależności od parametrów dla zbioru *weather AUS*

2.5. Porównanie klasyfikatorów

Klasyfikator	Acc. *hearts*	Acc. *gestures*	Acc. *weather*
K najbliższych sąsiadów	0.736	0.683	0.854
Naiwny klasyfikator Bayesa	0.813	0.878	0.800
Maszyna wektorów nośnych	0.824	0.891	0.86
Drzewa decyzyjne i lasy losowe	0.879	0.912	0.85

Tabela 27. Porównanie najlepszych wyników klasyfikacji dla różnych algorytmów

3. Dyskusja

3.1. K-najbliższych sąsiadów

3.1.1. Matryka Euklidesowa

W przypadku najmniejszego zbioru danych *Heart disease* [1] dla wartości *K=8* do *K=11* można zauważyć znaczący wzrost wartości *accuracy*. Co więcej dla *K=10* udało się uzyskać najlepszy wynik. Wraz ze wzrostem parametru K można zaobserwować delikatne unormowanie się wartości *accuracy* - nie ma tak dużych różnich jak w przypadku mniejszych wartości parametru K. Szczególną uwagę warto zwrócić na to, że dla wartości nieparzystych uzyskujemy znacząco lepsze wynik niż dla wartości parzystych. Stąd na wykresie można zobaczyć te 'skoki'. Z literatury, sposobu działania tej metody zdrowej logiki jest to dosyć uzasadnione. W przypadku tego zbioru korzystanie z wartości parametru K powyżej wartości 20 nie ma sensu gdyż wydłuża czas obliczeń a wyniki nie polepszają się. Biorąc pod uwagę zbiór [1] optymalnymi wartościami są te z przedziału od 7 do 11.

W przypadku zbioru danych *Gestures* [2], który jest zbiorem pośrednim jeżeli chodzi o jego wielkość to dla wartości *K=3* oraz *K=4* udało się uzyskać najlepszy wynik. Tak samo jak w przypadku zbioru [1] wartości parametru K powyżej wartości 20 nie polepszają wyniku oraz wydłużają czas obliczeń. Warto zwrócić uwagę na fakt, że różnice między większością wartości parametru *accuracy* jest w okolicach *0.03* wiec wyniki dla wszystkich wartości parametru K są praktycznie identyczne - są znacznie mniejsze odchylenia niż w przypadku zbioru [1]. Analizując wyniki i chcąc zoptymalizować dobór parametru K autor sugeruje użycie wartości K równej 3 do 7 z sugestią wyboru wartości nieprzystych.

W przypadku ostatniego zbioru a jednocześnie największego o nazwie *Weather AUS* [3] najlepszy wynik udało się uzyskać dla wartości *K=15*. Analogicznie jak w przypadku zbioru drugiego [2] różnice pomiędzy uzyskanymi wynikami są bardzo niewielkie. Z ciekawych rzeczy, które można zauważyć jest to, że dla większych zbiorów danych aby uzyskać lepsze wyniki można pokusić sie o delikatne zwiększenie wartość parametru K - zawsze warto to zweryfikować doświadczalnie. Sugerowane jest również korzystanie z wartości nieparzystych dla tego parametru.

3.1.2. Matryka Manhattan

W przypadku najmniejszego zbioru danych *Heart disease* [1] dla wartości *K=4* do *K=7* można zauważyć znaczący wzrost wartości *accuracy*. Co więcej dla *K=4, K=5 K=7* udało się uzyskać najlepszy wynik. Wyniki i wnioski są dosyć zbliżone do tych w których została wykorzystana metryka Euklidesowa natomiast aby uzyskać zbliżone wyniki można było użyć mniejszej wartości parametru *K*.

W przypadku zbioru danych *Gestures* [2], który jest zbiorem pośrednim jeżeli chodzi o jego wielkość to dla wartości *K=5* udało się uzyskać najlepszy wynik. Analogicznie jak w przypadku metryki Euklidesowej różnice w wynikach są bardzo nieznaczne - autor domniemuje, że jest to kwestia charakterystyki zbioru danych. Zakres sugerowaych wartości parametru *K* jest identyczny jak dla metryki Euklidesowej.

W przypadku ostatniego zbioru a jednocześnie największego o nazwie *Weather AUS* [3] najlepszy wynik udało się uzyskać dla wartości *K=17*. Analogicznie jak w dla metryki Euklidesowej różnice w wartościach *accuracy* jest niezwykle mała natomiast była wymagana wieksza wartość parametru *K* aby uzyskać maksimum wartości *accuracy*. Również zalecane jest korzystanie z wartości nieparzystych dla tego parametru - widać ewidentne 'skoki' na wykresie dla wartości nieprzystych.

3.2. Naiwny klasyfikator Bayesa

Istnieje kilka wariantów naiwnego klasyfikatora Bayesa, które mogły zostać potencjalnie wykorzystane w zadaniu - *wielomianowy*, *Bernoulliego* oraz *Gaussa*. Wariant wielomianowy (*ang. multinomial*) jest najczęściej wykorzystywany w przypadku analizy tekstu naturalnego i nie przyjmuje wartości ujemnych, a wariant *Bernoulliego* sprawdza się dla wartości binarnych, dlatego zostały odrzucone podczas wykonywania zadania. Wariantem wykorzystanym podczas klasyfikacji był *naiwny klasyfikator Gaussa*.

Analizując otrzymane wyniki *accuracy* można zauważyć, że są one zbliżone dla każdej wartości procent wykorzystanych danych jako zbiór testowy. Największe różnice można zauważyć w przypadku zbioru *Heart disease* [1], ale wynika to z jego małego rozmiaru.

Nawet dla ekstremalnych wartości procent danych testowych jak 85%, czy 90% klasyfikator osiągał wysokie wartości *accuracy*. Kluczową kwestią wpływającą na wynik klasyfikacji był zatem sam dobór zbioru, jego liczebność oraz cechy.

3.3. Maszyna wektorów nośnych

Dla zbioru *Heart disease* [1] warto rozważyć wykorzystanie jako funkcji jądra - funkcji wielomianowej, dla której zanotowano najlepsze rezultaty klasyfikacji, zarówno dla parametru *C* jak i *Gamma*. Aby móc zarekomendować wykorzystanie funkcji sigmoidalnej należałoby kontynuować eksperymenty dla innych wartości parametrów. Na wykresie 7, możemy zauważyć tendencję wzrostową dokładności klasyfikacji dla wartości parametru *C* z przedziału [1,7; 2,0], można zatem wnioskować że będzie się ona utrzymywać dla warto-

ści wyższych i może uda się uzyskać wyższą dokładność klasyfikacji niż dla funkcji wielomianowej.

Dla zbioru *Gestures* [2] rekomendowane jest wykorzystywanie funkcji RBF, manipulacja parametrem C nie wpływa w sposób szczególny na dokładności klasyfikacji danych należących do tego zbioru. Natomiast można zaobserwować dynamiczną zmianę dokładności klasyfikacji w zależności od parametru *Gamma*, którego optymalną (zbadaną) wartością dla funkcji RBF jest 0,0001, co skutkuje dokładnością działania klasyfikatora na poziomie 88%.

Dla zbioru *Weather AUS* [3] sytuacja jest najbardziej interesująca ze wszystkich badanych zbiorów. Jest to skutek tego, że przy odpowiednim dobraniu parametrów, można uzyskać zbliżoną dokładność klasyfikacji (w okolicach 79%-86%) dla dwóch różnych funkcji jądra - dla funkcji wielomianowej oraz RBF. Aby rekomendować wykorzystanie funkcji sigmoidalnej, wymagane byłoby kontynuowanie eksperymentów, tak aby określić wartość parametru C, które za skutkuje wzrostem dokładności klasyfikacji. W przedziale [1,5; 2,0] można zaobserwować delikatny wzrost dokładności dla tej funkcji.

3.4. Drzewa decyzyjne i lasy losowe

Wykresy na rysunkach 10, 11 oraz 12 prezentują wyniki klasyfikacji dla pojedynczego drzewa decyzyjnego i lasu losowego. Każdy z rysunków jest związany z jednym z trzech zbiorów danych. Każdy zawiera również 4 wykresy.

Pierwsze dwa są poświęcone pojedynczemu drzewu decyzyjnemu. Przebadana została dokładność klasyfikacji w zależności od dwóch parametrów: *min_samples_leaf* (minimalna liczba próbek w liściu), oraz *max_depth* (maksymalna wysokość drzewa). Oba te parametry służą regularyzacji drzewa decyzyjnego, które samo z siebie ma silną tendencję do przeuczania. Zarówno *min_samples_leaf* jak i *max_depth* mają bardzo podobne zadanie i wystarczy właściwie zastosować jeden z nich, aby algorytm budujący drzewo zatrzymał się wystarczająco wcześnie. Wyniki dla każdego zbioru danych pokazują tę samą tendencję – na początku (małe wartości *min_samples_leaf* i duże wartości *max_depth*) model jest silnie przetrenowany, czyli osiąga dokładność bliską 1 dla zbioru uczącego i znacznie mniejszą dla zbioru testowego. Wraz z modyfikacją wartości tych parametrów obie krzywe (dla zbioru uczącego i testowego) zbliżają się do siebie. W pewnym momencie, podczas tego zbliżania lub na samym jego końcu, dokładność klasyfikacji na zbiorze testowym osiąga swoje maksimum (zielona kropka) i później obie krzywe konsekwentnie już maleją. Zachowanie to jest zgodne z oczekiwaniami, według których jeżeli będziemy coraz bardziej ograniczać swobodę modelu to w końcu uniemożliwimy mu naukę i będzie niedouczony. Przy analizie krzywych z dwóch pierwszych wykresów na każdym rysunku, warto zwrócić uwagę, że są one tym bardziej gładkie im większy zbiór danych podlega analizie.

Jeżeli chodzi o szczegółowe wyniki eksperymentów dotyczących pojedynczego drzewa decyzyjnego, to dla pierwszego zbioru (*heart disease*) udało się osiągnąć dokładność 0.846 lub 0.813, w zależności od parametru. Ostatecznie za najlepszy, wykorzystany przy uczeniu całego lasu, uznany został

min_samples_leaf = 10. Bardzo podobnie ma się sytuacja w przypadku zbioru *gestures*, gdzie ten sam parametr przyjmujący bliską wartość równą 8, pozwolił osiągnąć dokładność klasyfikacji 0.798. W trzecim zbiorze wreszcie przewagę zdobyło ograniczeni głębokości drzewa i to właśnie ten parametr, przyjmując wartość 5, pozwolił uzyskać dokładność klasyfikacji 0.849. Najprawdopodobniej wynika to z faktu, że zbiór ten jest znacznie (kilkadziesiąt i kilkaset) razy większy od poprzednich. Tak więc można mniej ograniczyć wysokość drzew pozostawiając im wciąż dużą swobodę przy rozbudowie. Z kolei widać tutaj również, że najlepsza wartość parametru *min_samples_leaf*, również ze względu na rozmiar zbioru, jest dużo większa niż dla poprzednich zbiorów danych.

Druga para wykresów na każdym z trzech rysunków jest poświęcona wynikom dla lasu losowego. Na podstawie wyników klasyfikacji pojedynczego drzewa wybrany została najlepszy parametr wraz z najlepszą wartością (*max_depth* lub *min_samples_leaf*), który następnie był wykorzystany do regularyzacji wszystkich drzew w lesie. Tak więc jeden z parametrów lasu losowego, dotyczący charakterystyki samych drzew, wybrany został na podstawie eksperymentów dla pojedynczego drzewa. Jeżeli chodzi o parametry dotyczące samego zespołu klasyfikatorów to zbadane zostały dwa: *n_estimators* (liczba drzew w lesie), oraz *max_samples* (liczba losowych próbek ze zbioru uczącego wykorzystana do uczenia pojedynczego drzewa). Tym razem najpierw wybrana została najlepsza znaleziona wartość *n_estimators* a następnie wykorzystana przy szukaniu najlepszej wartości *max_samples*. W ten sposób ostatecznie dla każdego zbioru wybrany został najlepszy model, osiągający najwyższą dokładność klasyfikacji.

Przyglądając się wykresom dotyczącym lasu losowego warto zwrócić uwagę, że w większości przypadków osiąga on lepsze wyniki niż pojedyncze drzewo. Jest to oczywiście zachowanie jak najbardziej oczekiwane. Dla pierwszego zbioru danych najlepszy okazał się las 50 drzew, gdzie każde uczy się na podstawie połowy zbioru uczącego. Pozwoliło to osiągnąć dokładność 0.879. Jeszcze przed wybraniem dobrej wartości parametru *max_samples* las sprawował się lepiej niż pojedyncze drzewo. Ostatecznie jednak przewaga zespołu klasyfikatorów nad jednym dla tego zbioru to dokładność większa o zaledwie kilka setnych. Drugi zbiór (*gestures*) najlepiej pokazał przewagę lasu losowego nad jednym drzewem decyzyjnym. Uzyskana dokładność klasyfikacji jest większa o ponad jedną dziesiątą. Co ciekawe, ograniczenie wartości parametru *max_samples* z 1 (domyślna wartości przy testowaniu samego *n_estimators*) do 0.99, spowodowało minimalny spadek dokładności. Jednakże to właśnie wyniki dla tego zbioru pokazuje, że zespół drzew może znacząco poprawić jakość klasyfikacji, która ostatecznie wyniosła 0.91. Prawdopodobnie zbiór ten ma dużą wariancję, które jest znacząco redukowana przez wprowadzenie zespołu klasyfikatorów. Ostatni i największy zbiór danych również ostatecznie został lepiej zaklasyfikowany przez las losowy niż pojedyncze drzewo. Różnica ta jest jednak bardzo niewielka i uzyskana została dokładność 0.85. Co ciekawe, w przypadku badania wartości parametru *n_estimators* dla tego zbioru, najlepsza dokładność klasyfikacji całego zespołu drzew okazała się mniejsza nż pojedynczego drzewa. W ogóle warto zwrócić uwagę, że dla ostatniego zbioru danych wszystkie wyniki są bardzo do

siebie zbliżone. Jest to najprawdopodobniej spowodowane obciążeniem tego zbioru – większość przykładów należy do jednej klasy. Niestety wykorzystana w prezentacji wyników miara *accuracy* nie pokazuje tego w żaden sposób. Należy więc z dystansem podejść do oceny jakości klasyfikacji dla ostatniego zbioru, która jest raczej niemożliwa na podstawie tylko tej miary.

4. Wnioski

Podsumowując wykonane zadanie wnioskujemy, że:

— Wartość parametru K raczej powinna być nieparzysta ze względu na sposób działania tego klasyfikatora. Testując dobór parametru K warto zwracać uwagę na różnice bezwzględne między wynikami. Często różnice te są bardzo niewielkie a korzystając z mniejszej wartości K przyspieszamy obliczenia

— Dokładność klasyfikacji w zależności od wartości parametru C nie zmienia się tak dynamicznie jak w zależności od wartości parametru *Gamma*, na skutek czego łatwiej jest znaleźć odpowiednią wartość tego parametru. Dodatkowo zwiększenie wartości *Gamma* znacząco wydłuża czas nauki maszyny wektorów nośnych.

— Lasy losowe zazwyczaj radzą sobie lepiej niż pojedyncze drzewa decyzyjne, różnica ta jest czasami jednak bardzo niewielka w stosunku do dodatkowego czasu obliczeń wymaganego przy uczeniu całego zespołu klasyfikatorów

— Kluczowym parametrem dla klasyfikatora Bayesa jest sam dobór zbioru danych i jego charakterystyka

Literatura

[1] https://www.kaggle.com/ronitf/heart-disease-uci
[2] https://www.kaggle.com/kyr7plus/emg-4
[3] https://www.kaggle.com/jsphyg/weather-dataset-rattle-package