Projekt - Reprezentacja i Analiza Danych

1. Uwagi wstępne:
Poniższe sprawozdanie nie zawiera wszystkich danych uzyskanych podczas analizy.
Tworząc to sprawozdanie starałem się zachować klarowność informacji uzyskanych w matlabie, jednak wydaje mi się że wygodniejszym będzie obejrzenie ich w wyniku skryptu.

2. Podstawowe parametry zbioru danych oraz dobór atrybutów
W zbiorze znajduje się 408 obiektów, należących do 4 klas
Parametry dla poszczególnych klas:

Klasa 1:

	dat1	dat2	dat3	dat4	dat5	dat6	dat7	dat8	dat9	dat10	dat11	dat12
mean	-168.89	143.06	8.997	24.743	8.7351	125.19	33.498	-0.025	-0.10461	30.995	71.817	16.741
median	-146.15	143.34	8.8317	24.275	8.6867	124.68	34.328	-0.021675	-0.1344	33.032	71.753	15.846
std	914.84	9.2101	2.2287	13.848	1.2625	10.311	11.053	0.096233	1.0248	13.211	10.896	11.702
var	8.3692e+05	84.826	4.9673	191.77	1.5939	106.31	122.18	0.0092607	1.0502	174.53	118.73	136.94

Klasa 2:

	dat1	dat2	dat3	dat4	dat5	dat6	dat7	dat8	dat9	dat10	dat11	dat12
mean	-67.585	144.5	9.07	61.614	8.8821	123.6	59.73	0.0065165	-0.092993	6.8222	84.25	17.136
median	-220.93	144.4	8.8433	61.208	8.79	124.47	58.864	-0.005183	-0.14258	6.838	86.061	14.916
std	1281.3	9.1603	1.9541	8.7035	1.5428	10.15	12.785	0.11879	1.2372	4.0593	12.2	13.133
var	1.6417e+06	83.911	3.8185	75.751	2.3803	103.02	163.46	0.01411	1.5307	16.478	148.84	172.47

Klasa 3:

	dat1	dat2	dat3	dat4	dat5	dat6	dat7	dat8	dat9	dat10	dat11	dat12
mean	22.468	142.23	9.4626	53.583	9.0772	123.79	13.654	0.010299	-0.01642	40.8	51.705	16.425
median	15.55	142.13	9.569	54.492	9.0425	123.43	10.945	0.0087511	-0.006695	40.862	51.047	13.221
std	1140.4	9.8921	2.1175	11.966	1.6987	9.3412	10.057	0.11501	1.1474	10.26	5.6518	11.894
var	1.3005e+06	97.854	4.4837	143.18	2.8856	87.258	101.14	0.013227	1.3166	105.27	31.942	141.47

Klasa 4:

	dat1	dat2	dat3	dat4	dat5	dat6	dat7	dat8	dat9	dat10	dat11	dat12
mean	117.48	142.49	9.608	21.596	8.823	123.51	58.717	0.014836	0.033571	91.789	96.357	16.945
median	85.291	142.79	9.5678	21.646	8.9036	123.65	59.986	0.016109	0.0045263	91.777	96.417	16.554
std	1091.7	8.6636	2.2203	11.113	1.7016	9.8548	12.646	0.096551	0.9778	6.3863	8.9413	11.156
var	1.1919e+06	75.059	4.9299	123.51	2.8956	97.117	159.92	0.0093221	0.95609	40.785	79.947	124.45

Średnie wartości atrybutów dla danych klas:

	dat1	dat2	dat3	dat4	dat5	dat6	dat7	dat8	dat9	dat10	dat11	dat12
klasa1	-168.89	143.06	8.997	24.743	8.7351	125.19	33.498	-0.025	-0.10461	30.995	71.817	16.741
klasa2	-67.585	144.5	9.07	61.614	8.8821	123.6	59.73	0.0065165	-0.092993	6.8222	84.25	17.136
klasa3	22.468	142.23	9.4626	53.583	9.0772	123.79	13.654	0.010299	-0.01642	40.8	51.705	16.425
klasa4	117.48	142.49	9.608	21.596	8.823	123.51	58.717	0.014836	0.033571	91.789	96.357	16.945

Odchylenia standardowe średnich wartości atrybutów między klasami, średnie wartości atrybutów dla całego zbioru, oraz stosunek odchyleń do ogólnych średnich.

```
means_std =

  122.5682    1.0121    0.2972   20.1868    0.1451    0.7858   22.1218    0.0181    0.0654   35.7653   19.0634    0.3040

means_total =

  -28.5611  143.1399    9.2700   42.5463    8.8877  124.0018   41.1016    0.0019   -0.0489   39.2295   75.1827   16.8143

rel_std =

   -4.2914    0.0071    0.0321    0.4745    0.0163    0.0063    0.5382    9.4169   -1.3390    0.9117    0.2536    0.0181
```

Atrybuty o rel_std większym niż 0.5:

```
attr_with_big_std =

     1     7     8     9    10
```

Pary atrybutow o wartosci bezwzglednej korelacji wiekszej niz 0.5

8 1

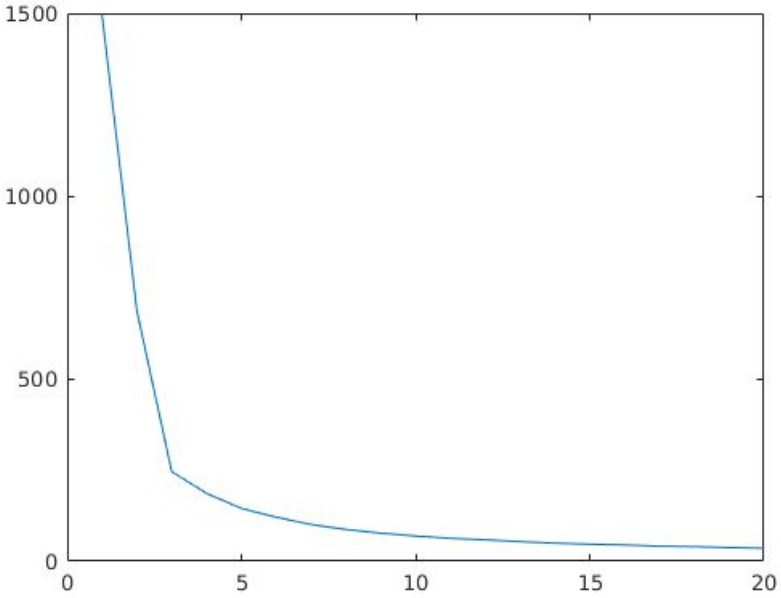
9 1

10 4

11 7

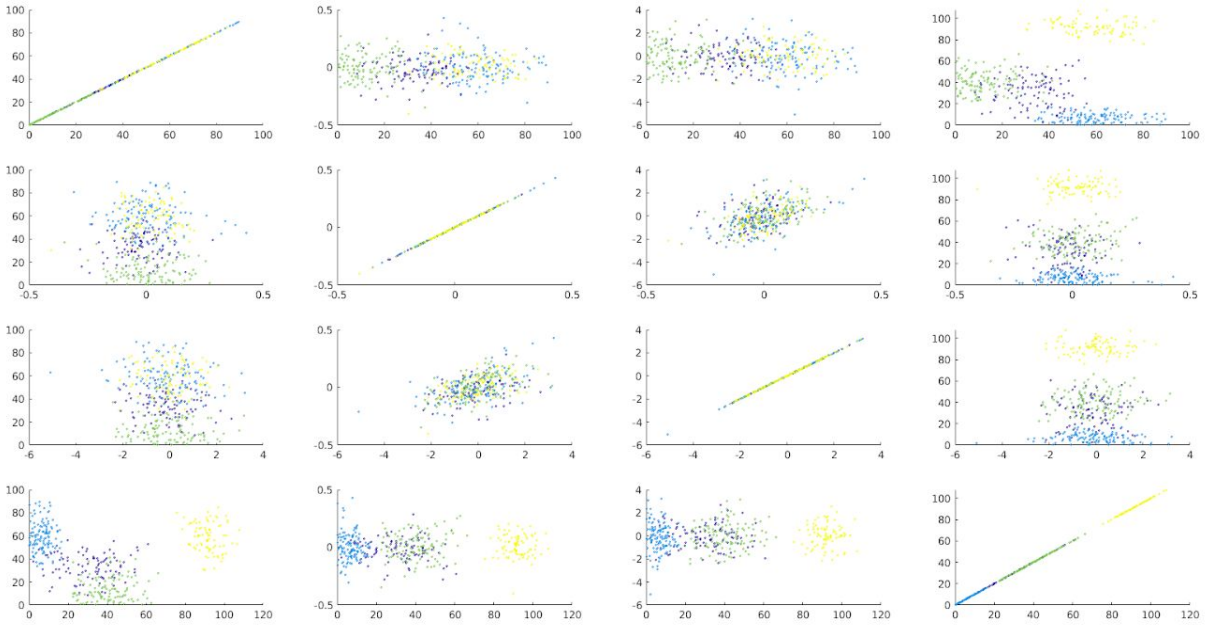
Wnioski:
Po przeprowadzeniu podstawowej analizy atrybutów dla klas wybrałem następujące klasy do dalszej analizy: 7, 8, 9, 10, ponieważ są to atrybuty najbardziej różniące się między klasami oraz słabo skorelowane, w związku z czym powinny dać dobre rezultaty w klasyfikacji.

3. Dobór ilości klas
Doboru ilości klas dokonałem metodą kmeans, używając atrybutów dobranych w poprzednim punkcie. Rezultaty są następujące:



Odległość spada bardzo szybko aż do klasyfikacji dla 3 klas, przy której prędkość spadku odległości przy dodawaniu klas znacząco spada.

Potwierdza to macierz wykresów punktowych dla 4 klas określonych w wejściowym zbiorze danych - widać że 2 z tych klas są dobrze określone, a pozostałe 2 słabo.



4. Przydatność różnych klasyfikatorów do klasyfikacji zbioru danych
    1. Klasyfikator jednego sąsiada:

        Wynik klasyfikacji dla zbioru testowego - dobrane atrybuty

c_tab =

22	1	5	0
4	32	0	0
5	0	28	0
0	0	0	25

ilosc bledow: 15
Wynik klasyfikacji dla zbioru testowego - wszystkie atrybuty
c_tab =

21	4	4	0
6	27	2	0
14	2	17	0
2	0	1	22

ilosc bledow: 35

2. Klasyfikator wielu sąsiadów - zbiór po standaryzacji (konieczna aby uzyskać sensowne rezultaty) :
    a. Z selekcją atrybutów
        '1-NN'

c_tab =

18	2	9	0
6	29	0	0
9	0	24	0
0	0	0	25

ilosc bledow: 26
'3-NN'
c_tab =

21	1	6	1
3	32	0	0
7	0	26	0
0	0	0	25

ilosc bledow: 18
'5-NN'
c_tab =

22	2	4	1
3	32	0	0

6 0 27 0
0 0 0 25

ilosc bledow: 16
'7-NN'
c_tab =
21 2 5 1
3 32 0 0
3 0 30 0
0 0 0 25

ilosc bledow: 14

b. Bez selekcji atrybutów
'1-NN'
c_tab =

25 3 1 0
4 31 0 0
2 0 31 0
1 0 0 24

ilosc bledow: 11
'3-NN'
c_tab =

25 3 0 1
2 33 0 0
4 0 29 0
0 0 0 25

ilosc bledow: 10
'5-NN'
c_tab =

25 2 1 1
1 34 0 0
1 0 32 0
0 0 0 25

ilosc bledow: 6
'7-NN'
c_tab =
25 2 1 1

1	34	0	0
1	0	32	0
0	0	0	25

ilosc bledow: 6

3. Metoda prototypów - niestandaryzowany zbiór danych
   a. Z doborem atrybutów
      metoda prototypów - treningowy
      c_tab =

50	7	9	0
5	79	0	0
9	0	67	0
0	0	0	60

błędy: 30

metoda prototypów - testowy

c_tab =

19	2	7	1
2	33	0	0
4	0	29	0
0	0	0	25

błędy: 16

   b. Bez doboru atrybutów
      metoda prototypów - treningowy

c_tab =

8	30	8	20
0	49	6	29
2	34	6	34
1	25	0	34

Błędy: 189
metoda prototypów - testowy

c_tab =

0	15	3	11
2	13	1	19
1	15	3	14
1	10	0	14

Błędy: 92

4. Metoda prototypów - standaryzowana
    a. Wszystkie atrybuty:
    metoda prototypów - treningowy
    c_tab =

61	3	2	0
0	82	2	0
2	1	73	0
0	0	0	60

Błędy: 10

metoda prototypów - testowy

c_tab =

27	1	1	0
0	34	1	0
2	0	31	0
0	0	0	25

Błędy: 5

5. Klasyfikator Bayesa
    a. Dobrane atrybuty
    Zbiór uczący:

49	6	11	0
2	82	0	0
8	0	68	0
0	0	0	60

Zbiór testowy:

c_tab =

20	2	7	0
1	34	0	0
4	0	29	0
0	0	0	25

ilosc bledow: 14

b. Wszystkie atrybuty

65	1	0	0
1	83	0	0
1	0	75	0
0	0	0	60

wynik klasyfikacji dla zbioru testowego

c_tab =

28	0	1	0
0	35	0	0
2	0	31	0
0	0	0	25

ilosc bledow: 3

6. Drzewo decyzyjne:
   a. Dobrane atrybuty

   wynik klasyfikacji dla zbioru uczącego

59	2	5	0
1	83	0	0
2	0	74	0
0	0	0	60

   wynik klasyfikacji dla zbioru testowego

22	2	5	0
1	34	0	0
9	0	24	0
0	0	0	25

   ilosc bledow: 17

   b. Wszystkie atrybuty

   wynik klasyfikacji dla zbioru uczącego

66	0	0	0
0	84	0	0
2	0	74	0

0 0 0 60

wynik klasyfikacji dla zbioru testowego

26 0 3 0
1 34 0 0
1 0 32 0
0 0 0 25

ilosc bledow: 5

Wnioski:

Zdecydowanie najlepszym klasyfikatorem dla tego zbioru danych jest klasyfikator Bayesa - daje on jednocześnie bardzo dobre wyniki klasyfikacji, i nie jest tak kosztowny obliczeniowo jak np. klasyfikator 5 lub 7 najbliższych sąsiadów. Jest to o tyle zaskakujące, że wiele z atrybutów w zbiorze jest ze sobą silnie skorelowanych, a klasyfikator Bayesa traktuje wszystkie zdarzenia jak niezależne. Porównywalne z klasyfikatorem Bayesa wydaje się jedynie drzewo decyzyjne, chociaż klasyfikator N najbliższych sąsiadów również daje dobre rezultaty pod warunkiem uprzedniej standaryzacji zbioru(czego można było się spodziewać). Klasyfikator jednego sąsiada daje akceptowalne rezultaty jedynie pod warunkiem wcześniejszego doboru atrybutów. Co ciekawe standaryzacja wpływa w tym wypadku negatywnie na wyniki.