1. Usunięcie wartości odstających

Pierwszym operacją wykonaną na zbiorze danych - zaraz po zbadaniu podstawowych jego cech - było pozbycie się wartości odstających. Obecność wartości odstających potwierdza zarówno analiza wykresu sporządzonego dla dwóch dowolnych cech czy wynik różnicy wartości średniej i mediany:

```
>> T =ABS(MEAN(TRAIN) - MEDIAN(TRAIN))

T =

  0.00000   0.00420   0.01469   0.21028   0.20882  79.65832   1.06037   0.00908
```

W celu zlokalizowania odstających próbek obliczono wartości minimalne oraz maksymalne poszczególnych cech:

```
[MV MIDX_1] = MAX(TRAIN)

77  186  186  186  186  186
186  186

[MV MIDX_2] = MIN(TRAIN)

58  642  642  642  642  393
25  539
```

Na tej podstawie zidentyfikowano próbki odstające jako nr. 642 i 168. Zbadano również ich sąsiedztwa:

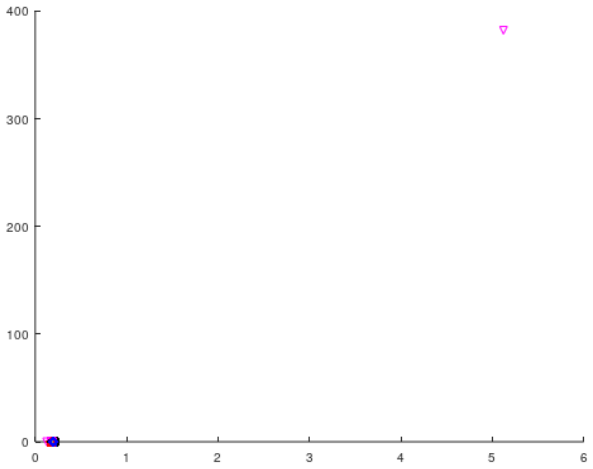
- **Próbka 642**

```
MIDX_2 = 642

>> TRAIN(MIDX-1:MIDX+1, :)

ANS =

  3.00000  0.18165  0.00007  0.00191  0.00001 -0.00000 -0.00000 -
0.00000

  3.00000  0.12500  0.00000  0.00000  0.00000  0.00000  0.00000  0.00000

  3.00000  0.18161  0.00002  0.00198  0.00000 -0.00000 -0.00000
0.00000
```

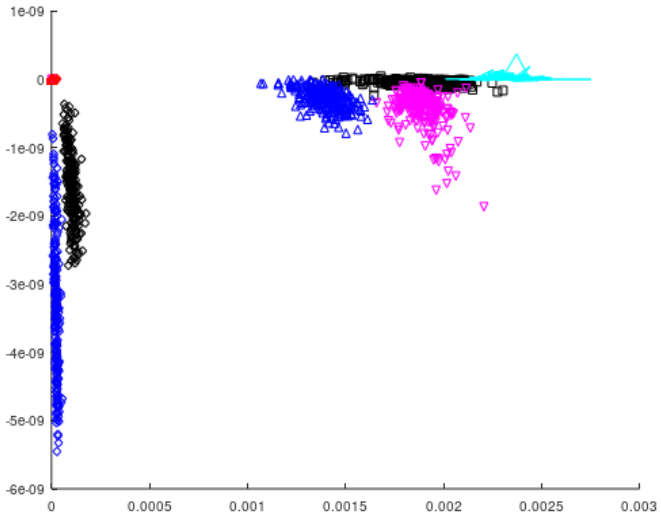
- **Próbka 186**

```
MIDX_2 = 186

>> TRAIN(MIDX-1:MIDX+1, :)

ANS =

  3.00000  0.18165  0.00007  0.00191  0.00001 -0.00000 -0.00000 -
0.00000

  3.00000  0.12500  0.00000  0.00000  0.00000  0.00000  0.00000
0.00000

  3.00000  0.18161  0.00002  0.00198  0.00000 -0.00000 -0.00000  0
```

Po usunięciu wspomnianych próbek ponowna analiza różnicy wartości średniej i mediany oraz wykresu dwóch cech wskazała na ich poprawną identyfikację i usunięcie.



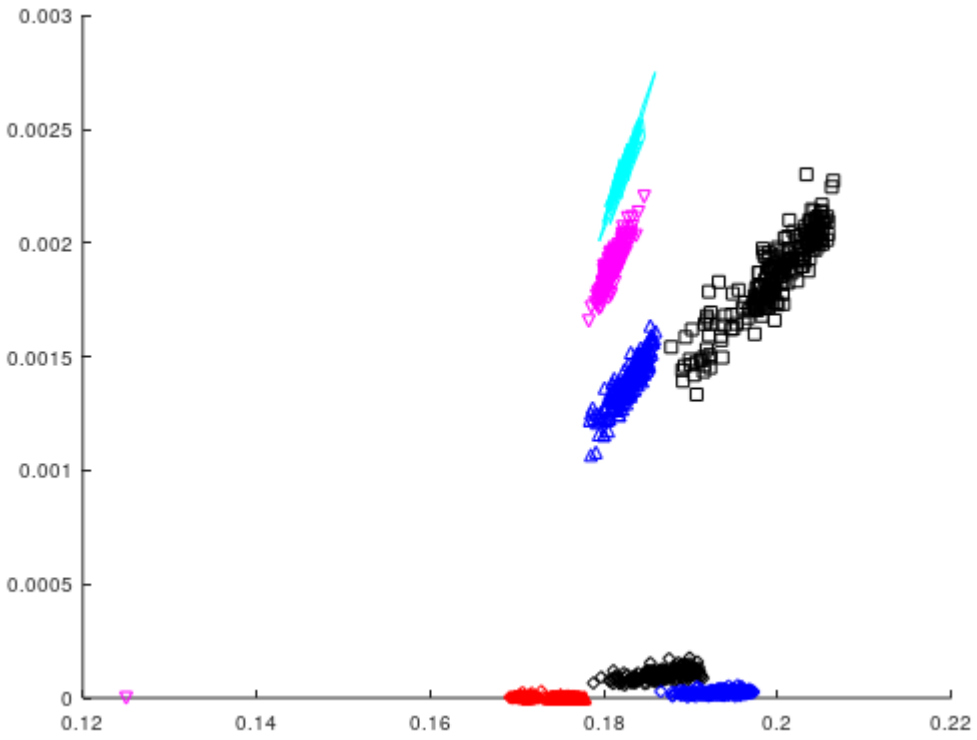
A) WYKRES KLAS PRZED USUNIĘCIEM PRÓBEK ODSTAJĄCYCH



B) WYKRES KLAS PO USUNIĘCIU PRÓBEK ODSTAJĄCYCH

2. Wybór cech do budowy klasyfikatora Bayesa

W celu wybrania najlepszego zestawu cech umożlwiającego poprawną klasyfikację próbek, badano rozkład punktów na wykresie cech - rysowanego dla różnych ich kombinacji. Wybrano zestaw gdzie zbiory punktów były najlepiej separowalne – dla cech 1 i 3.



Wykres klas dla cech 1 i 3

3. Budowa klasyfikatora

Na podstawie tak otrzymanych danych zbioru uczącego zbudujemy klasyfikator Bayesa przy różnych metodach liczenia funkcji gęstości prawdopodobieństwa PDF.

- **Z założeniem niezależności cech oraz ich normalnego rozkładu (para_indep.m)**
- **Z założeniem zależności cech oraz ich normalnego rozkładu (para_multi.m)**
- **Z wykorzystaniem okna Parzena do aproksymacji PDF (para_parzen.m)**

W celu porównania wyników klasyfikacji wyżej przedstawionymi metodami obliczono skuteczność klasyfikacji. Wykorzystana funkcja oblicza średnią wartość źle zaklasyfikowanych próbek, porównując etykiety ze zbioru testowego z tymi otrzymanymi w wyniku klasyfikacji. Wynik podano z dokładnością do 3 cyfr znaczących po przecinku.

	para_indep	para_multi	para_parzen (w =0,001)
ercf	**0.0263**	**0.00493**	**0.0285**

4. Redukcja zbioru uczącego

Rozdział ten poświęcony jest zbadaniu wpływu wielkości zbioru uczącego na jakość klasyfikatora Bayesa. W tym celu obliczono wartości średnie i odchylenia standardowe współczynnika błędu dla 5 powtórzeń oraz następujących częściach dziesiętnych zbioru początkowego: [0.1 ; 0.25 ; 0.5]. Wyniki zamieszczono w tabeli:

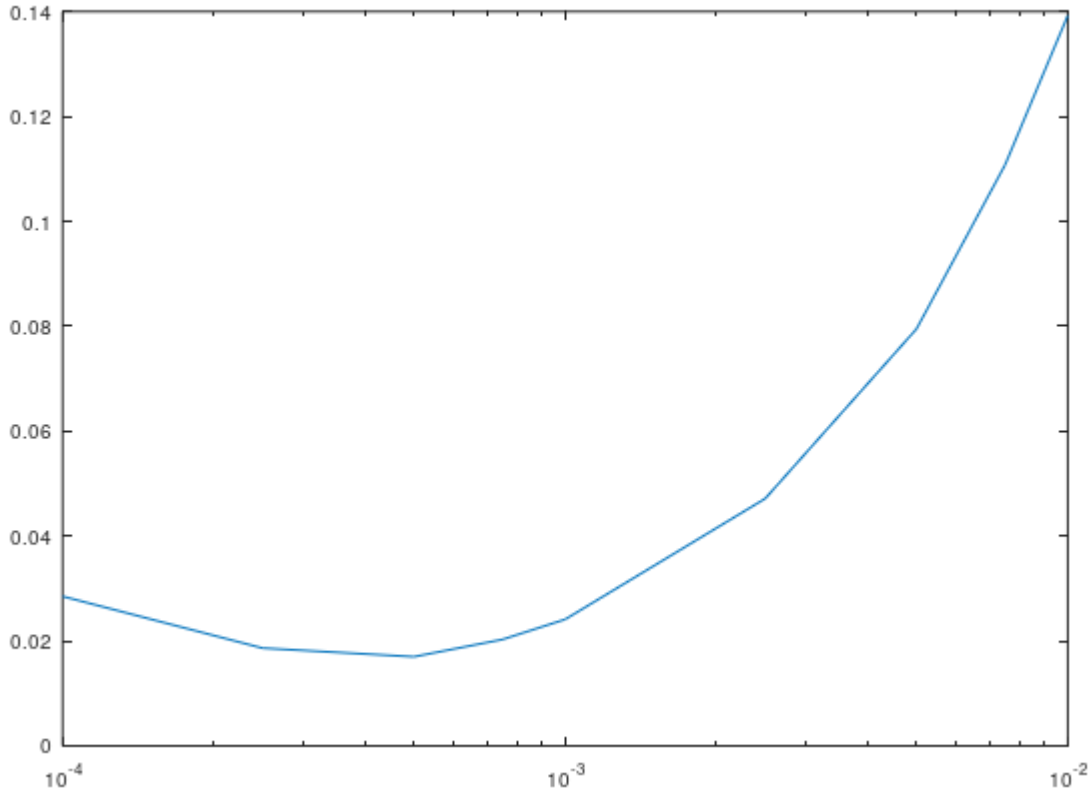
	μ_{indep}	σ_{indep}	μ_{multi}	σ_{indep}	μ_{parzen}	σ_{parzen}
10%	0.033	0.0057	0.0067	0.0016	0.100	0.0170
25%	0.028	0.0018	0.071	0.0017	0.058	0.0085
50%	0.027	0.0016	0.0056	0.0011	0.037	0.0027

Analiza powyższej tabeli pozwala łatwo zauważyć zależność jaką jest wzrost średniej wartości ercf wraz ze zmniejszaniem się zbioru uczącego. Zatem im więcej danych w zbiorze uczącym tym lepszej jakości klasyfikator otrzymamy.

5. Szerokość okna Parzena

Kolejnym etapem było zbadanie wpływu szerokości okna Parzena na jakość klasyfikatora. W tym celu ponownie obliczono wartości ercf dla różnych jego wartości. Wyniki przedstawiono w tabeli poniżej:

h	0.0001	0.00025	0.0005	0.00075	0.001	0.0025	0.005	0.0075	0.01
ercf	0.0285	0.0186	0.0170	0.0203	0.0241	0.0471	0.0795	0.111	0.1393



Na podstawie tabeli i wykresu możemy zaobserwować początkowy niewielki spadek a następnie gwałtowny wzrost błędu klasyfikatora wraz ze wzrostem szerokości okna Parzena w. Najmniejsza jego wartość zachodzi dla szerokości w = 0.0005

6. Modyfikacja wartości prawdopodobieństw apriori.

Zbadano również w jaki sposób modyfikacja prawdopodobieństw apriori wpłynie na jakość klasyfikatora. Do tej pory prawdopodobieństwo to było stałe dla każdej metody i równe 0.125. Dla tego przypadku ustalono różne, następujące wartości:
apriori = [0.165 0.085 0.085 0.165 0.165 0.085 0.085 0.165];
Obliczono wartości średnie ercf jak miało to miejsce w rozdziale 3. Poniżej przedstawiono wyniki:

	para_indep	para_multi	para_parzen (w =0,001)
Stałe wartość $p_{apriori}$	**0.0263**	**0.00493**	**0.0285**
Różne wartości $p_{apriori}$	**0.0212**	**0.00380**	**0.0290**

Analiza powyższej tabeli pozwala stwierdzić, że zmiana prawdopodobieństw apriori wpłynęła poztywnie na jakość klasyfikatora. Jednak nie w każdym przypadku jesteśmy w stanie dobrze i jednoznacznie je określić

7. Normalizacja danych

Ostatnim krokiem było zbadanie wpływu normalizacji danych na jakość klasyfikatora. W tym celu wykorzystano klasyfikator stworzony na potrzebę lab1 – cls1nn.m.
W poniższej tabeli zamieszczono wartość klasyfikacji dla zbioru znormalizowanego i nieznormalizowanego.

	Znormalizowany	Nieznormalizowany	
ercf	**0.00219**	**0.0153**	