



Prognozowanie cukrzycy u kobiet: wybrane metody uczenia maszynowego

Projekt
Statystyczna analiza danych

Klaudia Sołek

1. Cel

Celem pracy jest zbadanie jakości prognozowania wybranymi metodami uczenia maszynowego:

- Metoda k najbliższych sąsiadów (KNN)
- Ważona metoda k najbliższych sąsiadów (KKNN)
- Regresja logistyczna
- Liniowa analiza dyskryminacyjna (LDA)

2. Opis i wstępna analiza danych

Dane „diabetes” dotyczą 768 kobiet opisanych za pomocą 9 zmiennych:

- pregnant – ilość ciąż
- glucose - stężenie glukozy w osoczu (test tolerancji glukozy)
- pressure - rozkurczowe ciśnienie krwi (mm Hg)
- triceps - grubość fałdu skórniego tricepsa (mm)
- insulin – stężenie insuliny po 2 godzinach (mu U/ml)
- mass - wskaźnik masy ciała (masa w kg/ (wzrost w m)²)
- pedigree – funkcja rodowodu cukrzycy
- age – wiek (lata)
- diabetes - zmienna kategoriowa (test na cukrzycę): pos lub neg

Cukrzyca (diabetes mellitus) to przewlekła choroba metaboliczna wynikająca z zaburzonego wydzielania lub działania insuliny - hormonu produkowanego przez trzustkę.

Dane Światowej Organizacji Zdrowia (WHO) wskazują, że według szacunków w 2014 r. na świecie żyło 422 mln dorosłych z cukrzycą (dla porównania - w 1980 r. było ich 108 mln). Według International Diabetes Federation w 2040 roku ma być już 642 milionów osób z cukrzycą.

2.1 Uzupełnienie braków danych

pregnant	glucose	pressure	triceps	insulin	mass	pedigree	age	diabetes
Zmienna może przyjmować wartość 0	Braki danych zastąpione średnią					Nie występują braki danych		

Zmienna	glucose	pressure	triceps	insulin	mass
Średnia	121,69	72,39	29,11	155,71	32,46
Wartość zastępująca brak danych	122	72	29	156	32,5

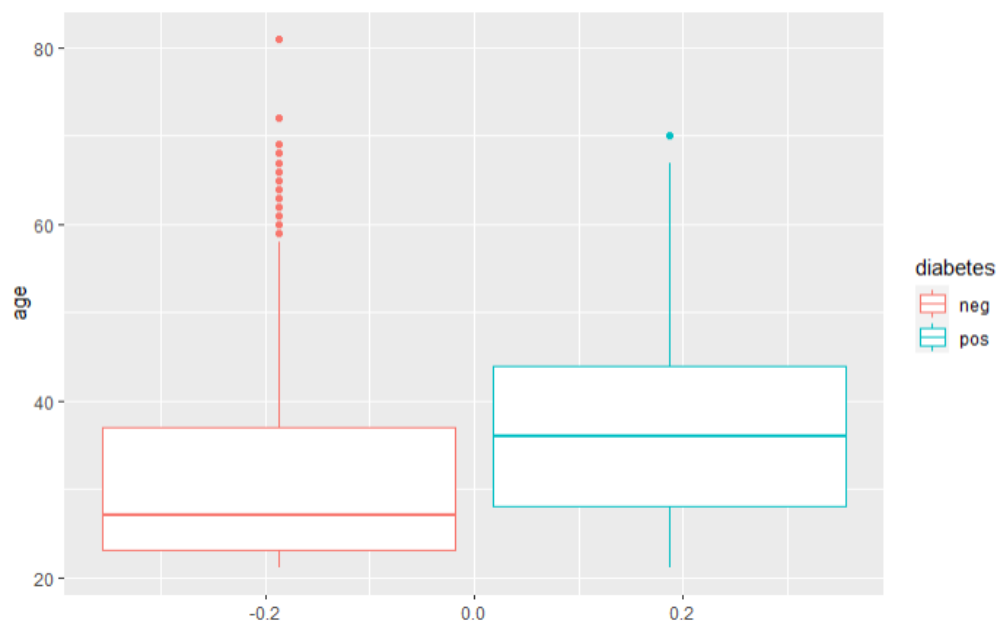
2.2 Podstawowe statystyki

pregnant	glucose	pressure	triceps	insulin	mass
Min. : 0.000	Min. : 44.00	Min. : 24.00	Min. : 7.00	Min. : 14.0	Min. : 18.20
1st Qu.: 1.000	1st Qu.: 99.75	1st Qu.: 64.00	1st Qu.: 25.00	1st Qu.: 121.5	1st Qu.: 27.50
Median : 3.000	Median : 117.00	Median : 72.00	Median : 29.00	Median : 156.0	Median : 32.40
Mean : 3.845	Mean : 121.69	Mean : 72.39	Mean : 29.11	Mean : 155.8	Mean : 32.46
3rd Qu.: 6.000	3rd Qu.: 140.25	3rd Qu.: 80.00	3rd Qu.: 32.00	3rd Qu.: 156.0	3rd Qu.: 36.60
Max. : 17.000	Max. : 199.00	Max. : 122.00	Max. : 99.00	Max. : 846.0	Max. : 67.10

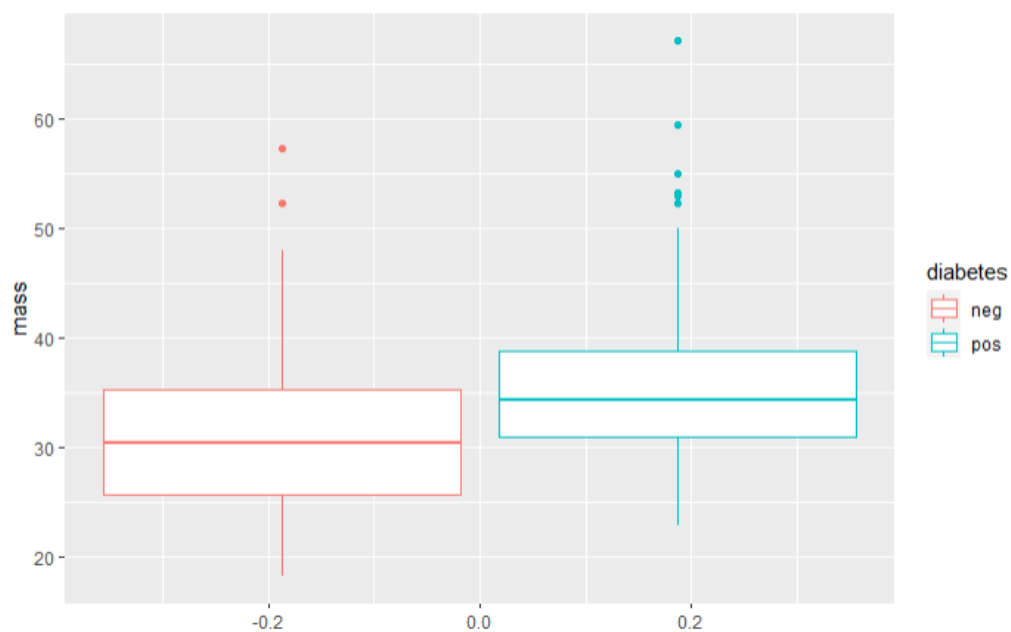
pedigree	age	diabetes
Min. : 0.0780	Min. : 21.00	neg: 500
1st Qu.: 0.2437	1st Qu.: 24.00	pos: 268
Median : 0.3725	Median : 29.00	
Mean : 0.4719	Mean : 33.24	
3rd Qu.: 0.6262	3rd Qu.: 41.00	
Max. : 2.4200	Max. : 81.00	

Zmienna kategoryczna „diabetes” została zmieniona na typ factor. Negatywny wynik testu na cukrzycę otrzymało 500 kobiet, pozostałe 268 są chore na cukrzycę. Kobiety są w przedziale wiekowym 21-81. Średnia wieku to 33 lata. Zmienna „pregnant” osiąga zaskakująco wysoką wartość maksymalną 17 ciąż. Szeroki zakres zmiennych (*pregnant* 0-17, *glucose* 44-199, *pressure* 24-122, *triceps* 7-99, *insulin* 14-846, *mass* 18.20-67.1, *age* 21-81) oznacza, że mamy do czynienia z bardzo zróżnicowanymi kobietami.

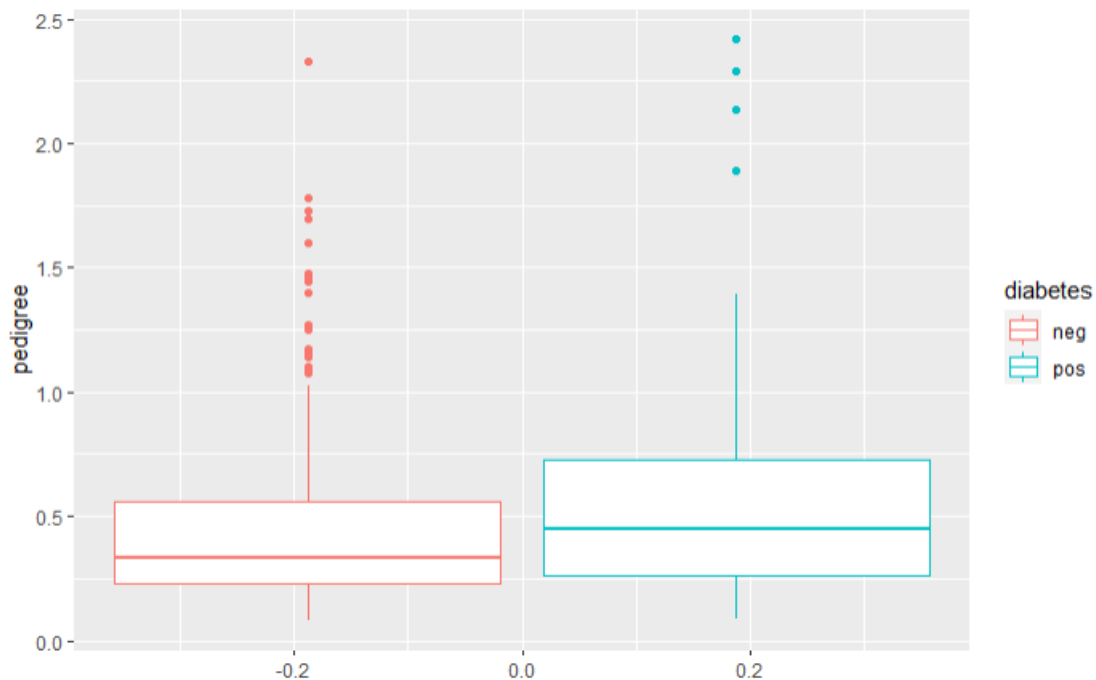
2.3 Analiza wpływu poszczególnych zmiennych na zmienną wynikową „diabetes”



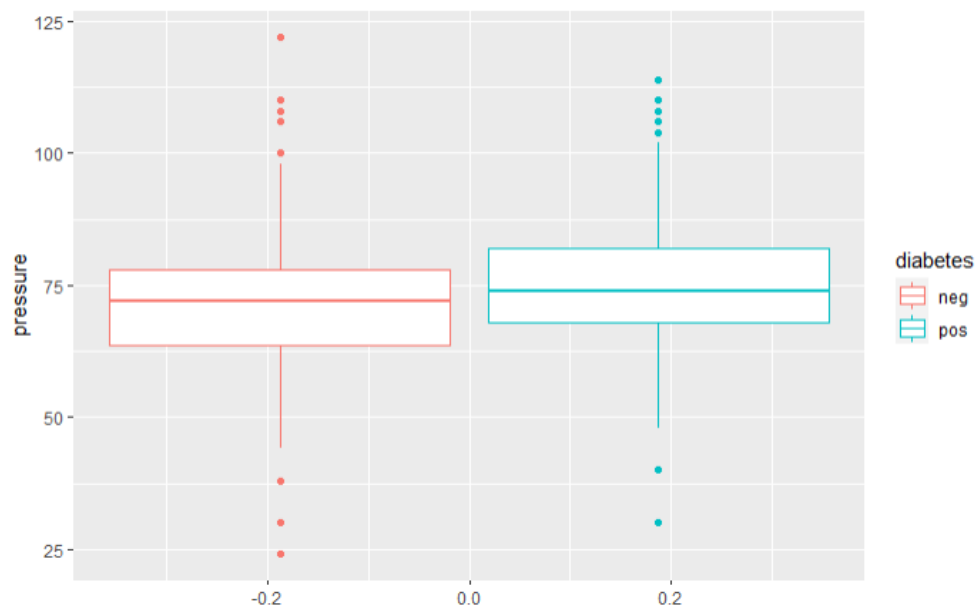
Mediana wieku kobiet chorych na cukrzycę jest na podobnym poziomie co trzeci kwartył wieku kobiet z negatywnym wynikiem jest także o około 10 lat większa od mediany zdrowych kobiet, co oznacza, że chore kobiety były przeważnie starsze od zdrowych. Długi „wąs” w przypadku negatywnego testu świadczy o tym, że występują wartości odstające.



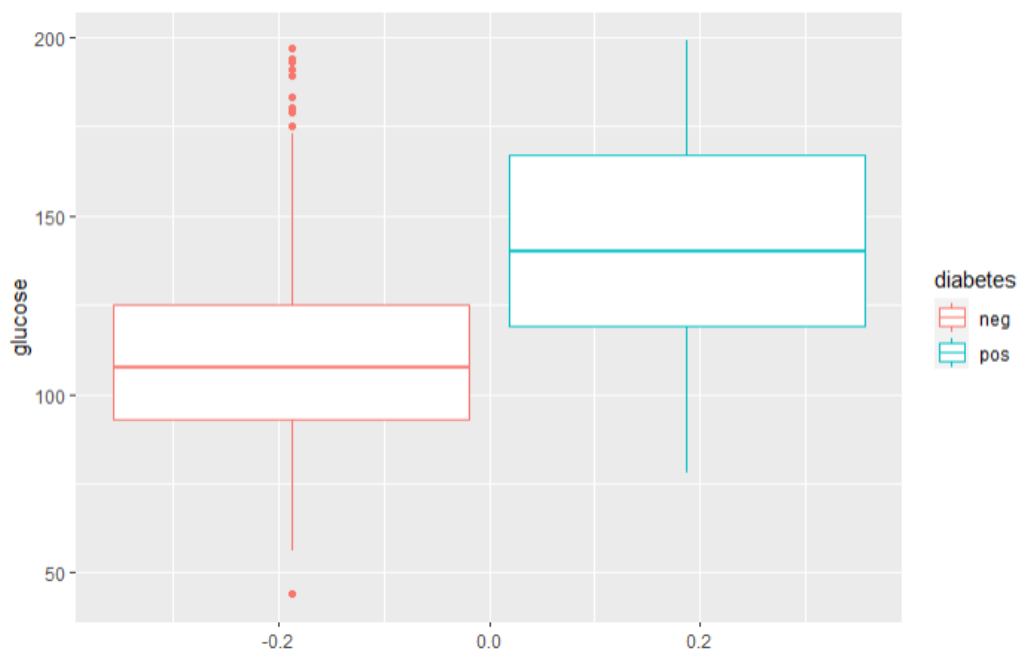
Mediana masy zdrowych kobiet jest na podobnym poziomie co pierwszy kwartył masy kobiet z pozytywnym wynikiem testu, na tej podstawie można wnioskować, że masa ma wpływ na wynik testu, kobiety o większej masie mają większe szanse na zachorowanie na cukrzycę.



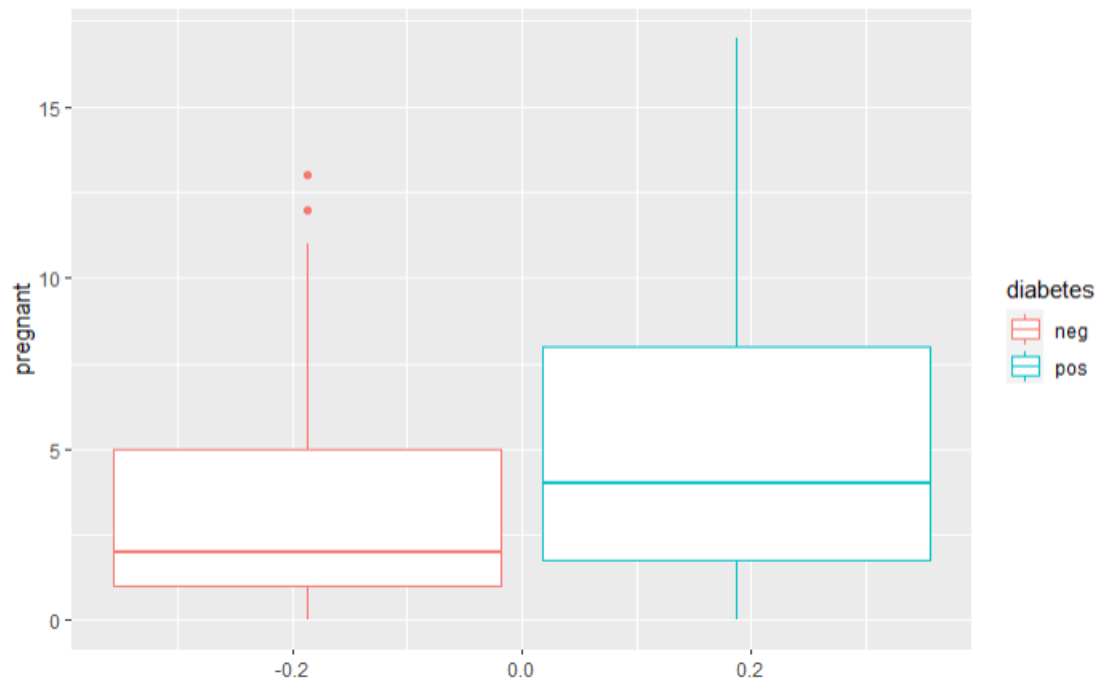
Zarówno w przypadku pozytywnego jak i negatywnego wyniku dane są rozproszone. Mediana dla pozytywnego wyniku testu jest nieznacznie większa. Funkcja rodowodu cukrzycy ma stosunkowo mały wpływ na wynik testu.



Rozkurczowe ciśnienie krwi ma najmniejszy wpływ na wynik testu w porównaniu z resztą zmiennych. Mediany są do siebie zbliżone, w obu przypadkach „wąsy” są długie co świadczy o rozproszenie danych.



Stężenie glukozy w osoczu ma największy wpływ na wynik testu w porównaniu z innymi zmiennymi. Mediany różnią się od siebie o około 30. Pierwszy kwartył osób z pozytywnym wynikiem jest na podobnym poziomie co trzeci kwartył dla negatywnego testu. Im większe stężenie glukozy w osoczu tym większa szansa na pozytywny wynik testu.



Mediana liczby ciąż dla kobiet z pozytywnym wynikiem testu jest większa od tych z negatywnym.
Wraz ze wzrostem liczby ciąż rośnie szansa zachorowania na cukrzycę.

2.4 Analiza tabel częstości

```
{r}
table(dane$pregnant, dane$diabetes)
```

	neg	pos
0	73	38
1	106	29
2	84	19
3	48	27
4	45	23
5	36	21
6	34	16
7	20	25
8	16	22
9	10	18
10	14	10
11	4	7
12	5	4
13	5	5
14	0	2
15	0	1
17	0	1

Kobiety o liczbie ciąż od 0 do 6 w większości uzyskiwały negatywny wynik testu. Natomiast w przypadku liczby ciąż: 7, 8, 9, 11 i 14 w większości przypadków wynik testu był pozytywny.

```

####{r}
table(dane$diabetes, dane$glucose)
####

      44 56 57 61 62 65 67 68 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97
neg   1  1  2  1  1  1  1  3  4  1  3  4  2  2  2  3  3  5  6  3  6  9  6  3  7  8  6  9  9  8  6  7  10  8  7
pos   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  1  0  0  0  1  1  0  0  1  0  2  0  1  1  0  3  0  2

      98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123
neg   3 17 13  8  9  8  3 10 13  8 10  7  6 11  9  3  9  4  6  9  4  7  8  4 12  7
pos   0  0  4  1  4  1  3  3  1  3  3  5  0  3  4  2  2  6  1  2  2  4  3  2  5  2

      124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149
neg    6  7  7  5  5  8  4  2  3  3  2  2  4  6  2  6  2  3  3  4  3  1  4  4  1  0
pos    5  7  2  0  6  6  3  3  2  2  4  2  4  2  3  2  3  2  2  2  4  4  5  3  3  1

      150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175
neg    2  3  1  2  5  1  0  2  2  1  0  1  1  0  1  3  1  0  0  0  0  0  0  1  0  1
pos    1  3  3  0  1  4  3  0  6  1  1  2  5  3  2  1  2  3  4  1  2  3  1  5  2  1

      176 177 178 179 180 181 182 183 184 186 187 188 189 190 191 193 194 195 196 197 198 199
neg    0  0  0  2  1  0  0  1  0  0  0  0  0  1  0  1  1  1  0  0  1  0  0
pos    2  1  1  3  4  5  1  2  3  1  4  2  3  1  0  1  2  2  3  3  1  1

```

Interesującym faktem jest to, że kobiety o poziomie glukozy 197, 194, 193 i 191 uzyskały negatywny wynik testu a dla stężenia glukozy w osoczu 78 wynik jest pozytywny.

2.3 Analiza współczynników zmienności i korelacji

Do obliczeń przyjmujemy zmienną „diabetes” jako wartość liczbową (numeric value).

zmienna <chr>	współczynnik_zmienności <dbl>
pregnant	0.8763413
glucose	0.2501131
pressure	0.1671113
triceps	0.3020200
insulin	0.5458200
mass	0.2118164
pedigree	0.7021514
age	0.3537882
diabetes	0.3535701

Współczynnik zmienności wszystkich zmiennych przekracza 10%, więc nie wykluczamy żadnej zmiennej.

```

##{r}
dane$diabetes <- as.numeric(dane$diabetes)
round(cor(dane),3)

```

```

      pregnant glucose pressure triceps insulin mass pedigree age diabetes
pregnant    1.000   0.128   0.209   0.082   0.056 0.022  -0.034 0.544   0.222
glucose      0.128   1.000   0.219   0.193   0.420 0.231   0.137 0.267   0.493
pressure     0.209   0.219   1.000   0.192   0.073 0.281  -0.002 0.325   0.166
triceps      0.082   0.193   0.192   1.000   0.158 0.543   0.102 0.126   0.215
insulin      0.056   0.420   0.073   0.158   1.000 0.166   0.098 0.137   0.215
mass         0.022   0.231   0.281   0.543   0.166 1.000   0.153 0.025   0.312
pedigree     -0.034   0.137  -0.002   0.102   0.098 0.153   1.000 0.034   0.174
age          0.544   0.267   0.325   0.126   0.137 0.025   0.034 1.000   0.238
diabetes     0.222   0.493   0.166   0.215   0.215 0.312   0.174 0.238   1.000

```

Współczynnik korelacji nie przekracza poziomu 0.9, co świadczy o braku współliniowości.

3. Macierz błędu

		Test na cukrzyce	
		prawdziwy	fałszywy
Wynik testu (prognoza)	pozytywny	TP (prawdziwie pozytywny)	FP (fałszywie pozytywny)
	negatywny	FN (fałszywie negatywny)	TN (prawdziwie negatywny)

Do oceny macierzy błędu służą parametry takie jak:

- Dokładność (ang. Accuracy)
- Czułość (ang. sensitivity)
- Specyficzność (ang. specificity)

Wzór na dokładność:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Dokładność pozwala ocenić jakość klasyfikacji testu. Dowiadujemy się, jaka część testów na cukrzyce, ze wszystkich zaklasyfikowanych, została zaklasyfikowana poprawnie. Stosunek sumy prawdziwie pozytywnych i prawdziwie negatywnych do wszystkich klasyfikowanych przypadków.

Wzór na czułość:

$$TPR = \frac{TP}{TP + FN}$$

Czułość to miara wskazująca w jakim procencie klasa faktycznie pozytywna została pokryta przewidywaniem pozytywnym (procent osób chorych na cukrzycę, dla których test diagnostyczny wskazuje wynik pozytywny).

Wzór na specyficzność:

$$TPR = \frac{TN}{TN + FP}$$

Specyficzność to miara wskazująca w jakim procencie klasa faktycznie negatywna została pokryta przewidywaniem negatywnym (procent osób zdrowych, dla których test diagnostyczny wskazuje wynik negatywny).

Wszystkie parametry (dokładność, czułość, specyficzność) powinny osiągać jak największą wartość (dążyć do 1).

4. Metoda k najbliższych sąsiadów (KNN)

Metoda k najbliższych sąsiadów (KNN, k-Nearest Neighbors) dla obiektu, wyznacza k jego najbliższych sąsiadów (tj. punktów o najmniejszej odległości według zadanej metryki), a następnie wyznacza wynik w oparciu o głos większości tych obiektów. Najważniejszym problemem praktycznym jest wybór właściwej wartości k. W miarę wzrostu k metoda staje się mniej elastyczna i tworzy granicę decyzyjną zbliżoną do liniowej.

Na początku zmieniamy zmienną kategoryczną „diabetes” na zmienną zero-jedynkową.

```
```{r}
str(dane)
```

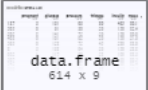
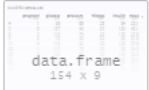
'data.frame':  768 obs. of  9 variables:
 $ pregnant: int  6 1 8 1 0 5 3 10 2 8 ...
 $ glucose  : int  148 85 183 89 137 116 78 115 197 125 ...
 $ pressure: int  72 66 64 66 40 74 50 72 70 96 ...
 $ triceps  : int  35 29 29 23 35 29 32 29 45 29 ...
 $ insulin  : int  156 156 156 94 168 156 88 156 543 156 ...
 $ mass     : num  33.6 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 32.5 ...
 $ pedigree: num  0.627 0.351 0.672 0.167 2.288 ...
 $ age      : int  50 31 32 21 33 30 26 29 53 54 ...
 $ diabetes: Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 1 2 2 ...
```

Dane zostały podzielone na zbiór uczący (80%) i zbiór testowy (20%). Zbiór uczący zawiera 614 obserwacji a testowy 154.

```

set.seed(9)
index <- sample(nrow(dane), 614, replace = F)
uczacy <- dane[index,]
testowy <- dane[-index,]
uczacy
testowy

```

Description: df [614 x 9]

| | pregnant
<int> | glucose
<int> | pressure
<int> | triceps
<int> | insulin
<int> | mass
<dbl> | pedigree
<dbl> | age
<int> | diabetes
<fctr> |
|-----|-------------------|------------------|-------------------|------------------|------------------|---------------|-------------------|--------------|--------------------|
| 187 | 8 | 181 | 68 | 36 | 495 | 30.1 | 0.615 | 60 | 1 |
| 565 | 0 | 91 | 80 | 29 | 156 | 32.4 | 0.601 | 27 | 0 |
| 262 | 3 | 141 | 72 | 29 | 156 | 30.0 | 0.761 | 27 | 1 |
| 408 | 0 | 101 | 62 | 29 | 156 | 21.9 | 0.336 | 25 | 0 |
| 595 | 6 | 123 | 72 | 45 | 230 | 33.6 | 0.733 | 34 | 0 |
| 3 | 8 | 183 | 64 | 29 | 156 | 23.3 | 0.672 | 32 | 1 |
| 652 | 1 | 117 | 60 | 23 | 106 | 33.8 | 0.466 | 27 | 0 |
| 507 | 0 | 180 | 90 | 26 | 90 | 36.5 | 0.314 | 35 | 1 |
| 542 | 3 | 128 | 72 | 25 | 190 | 32.4 | 0.549 | 27 | 1 |
| 556 | 7 | 124 | 70 | 33 | 215 | 25.5 | 0.161 | 37 | 0 |

1-10 of 614 rows

Previous 1 2 3 4 5 6 ... 62 1

```

summary(uczacy)

```

| pregnant | glucose | pressure | triceps | insulin | mass |
|---------------|---------------|----------------|---------------|---------------|---------------|
| Min. : 0.00 | Min. : 56.0 | Min. : 24.00 | Min. : 7.00 | Min. : 14.0 | Min. :18.20 |
| 1st Qu.: 1.00 | 1st Qu.: 99.0 | 1st Qu.: 64.00 | 1st Qu.:25.00 | 1st Qu.:120.0 | 1st Qu.:27.60 |
| Median : 3.00 | Median :117.0 | Median : 72.00 | Median :29.00 | Median :156.0 | Median :32.40 |
| Mean : 3.95 | Mean :120.9 | Mean : 72.22 | Mean :28.91 | Mean :156.5 | Mean :32.35 |
| 3rd Qu.: 6.00 | 3rd Qu.:139.0 | 3rd Qu.: 80.00 | 3rd Qu.:32.00 | 3rd Qu.:156.0 | 3rd Qu.:36.50 |
| Max. :17.00 | Max. :199.0 | Max. :122.00 | Max. :99.00 | Max. :846.0 | Max. :59.40 |

| pedigree | age | diabetes |
|----------------|---------------|----------|
| Min. :0.0840 | Min. :21.00 | 0:406 |
| 1st Qu.:0.2470 | 1st Qu.:24.00 | 1:208 |
| Median :0.3840 | Median :29.00 | |
| Mean :0.4766 | Mean :33.64 | |
| 3rd Qu.:0.6378 | 3rd Qu.:41.00 | |
| Max. :2.4200 | Max. :81.00 | |

W zbiorze uczącym prawie dwukrotnie więcej kobiet uzyskało negatywny wynik testu. W porównaniu do statystyk wszystkich danych zwiększyła się minimalna wartość stężenia glukozy oraz minimalna wartość funkcji rodowodu ciąży a zmniejszyła się maksymalna masa. Pozostałe min max pozostały bez zmian.

```
summary(testowy)
```

```

pregnant      glucose      pressure      triceps      insulin      mass
Min.   : 0.000   Min.   : 44.0   Min.   : 40.00   Min.   : 8.00   Min.   : 40.0   Min.   :18.20
1st Qu.: 1.000   1st Qu.:102.0   1st Qu.: 64.00   1st Qu.:27.00   1st Qu.:125.2   1st Qu.:26.70
Median : 3.000   Median :119.5   Median : 72.00   Median :29.00   Median :156.0   Median :32.15
Mean   : 3.429   Mean   :124.9   Mean   : 73.03   Mean   :29.92   Mean   :153.0   Mean   :32.90
3rd Qu.: 5.000   3rd Qu.:145.8   3rd Qu.: 80.00   3rd Qu.:33.00   3rd Qu.:156.0   3rd Qu.:37.48
Max.   :14.000   Max.   :198.0   Max.   :110.00   Max.   :60.00   Max.   :579.0   Max.   :67.10

pedigree      age      diabetes
Min.   :0.0780   Min.   :21.00   0:94
1st Qu.:0.2370   1st Qu.:24.00   1:60
Median :0.3375   Median :29.00
Mean   :0.4530   Mean   :31.64
3rd Qu.:0.5995   3rd Qu.:37.00
Max.   :2.2880   Max.   :67.00

```

W przypadku zbioru testowego również większość kobiet nie uzyskała pozytywnego wyniku testu na cukrzyce.

Następnie dane w zbiorze uczącym oraz testowym zostały zestandaryzowane.

Description: df [614 x 9]

| | pregnant
<dbl> | glucose
<dbl> | pressure
<dbl> | triceps
<dbl> | insulin
<dbl> | mass
<dbl> | pedigree
<dbl> |
|-----|-------------------|------------------|-------------------|------------------|------------------|---------------|-------------------|
| 187 | 1.16868488 | 1.999448678 | -0.35391011 | 0.81481866 | 3.86759321 | -0.337060558 | 0.41768886 |
| 565 | -1.13955000 | -0.994036392 | 0.65133649 | 0.01084928 | -0.00517247 | 0.008089844 | 0.37543382 |
| 262 | -0.27396192 | 0.669010869 | -0.01882791 | 0.01084928 | -0.00517247 | -0.352067097 | 0.85834863 |
| 408 | -1.13955000 | -0.661426940 | -0.85653341 | 0.01084928 | -0.00517247 | -1.567596774 | -0.42439383 |
| 595 | 0.59162616 | 0.070313855 | -0.01882791 | 1.84849357 | 0.84021001 | 0.188168315 | 0.77383854 |
| 3 | 1.16868488 | 2.065970568 | -0.68899231 | 0.01084928 | -0.00517247 | -1.357505225 | 0.58972726 |
| 652 | -0.85102064 | -0.129251816 | -1.02407451 | -0.67826732 | -0.57637685 | 0.218181394 | -0.03202555 |
| 507 | -1.13955000 | 1.966187732 | 1.48904199 | -0.33370902 | -0.75916225 | 0.623357953 | -0.49079462 |
| 542 | -0.27396192 | 0.236618581 | -0.01882791 | -0.44856179 | 0.38324651 | 0.008089844 | 0.21848651 |
| 556 | 0.88015552 | 0.103574800 | -0.18636901 | 0.47026036 | 0.66884870 | -1.027361362 | -0.95258190 |

1-10 of 614 rows | 1-8 of 9 columns

Previous 1 2 3 4 5 6 ... 62 Next

Description: df [154 x 9]

| | pregnant
<dbl> | glucose
<dbl> | pressure
<dbl> | triceps
<dbl> | insulin
<dbl> | mass
<dbl> | pedigree
<dbl> |
|----|-------------------|------------------|-------------------|------------------|------------------|---------------|-------------------|
| 4 | -0.8294247 | -1.129658160 | -0.55235753 | -0.759595522 | -0.79397719 | -0.6266015715 | -0.862061101 |
| 5 | -1.1709525 | 0.381185538 | -2.59449930 | 0.558463186 | 0.20120144 | 1.3296758040 | 5.531056364 |
| 6 | 0.5366866 | -0.279808580 | 0.07599379 | -0.100566168 | 0.03982112 | -0.9526478007 | -0.759578313 |
| 7 | -0.1463691 | -1.475893174 | -1.80906015 | 0.228948509 | -0.87466735 | -0.2483879456 | -0.617910929 |
| 11 | 0.1951588 | -0.468664042 | 1.48978424 | -0.100566168 | 0.03982112 | 0.6123740996 | -0.789720309 |
| 20 | -0.8294247 | -0.311284490 | -0.23818187 | 0.009272058 | -0.76708047 | 0.2211186246 | 0.229079174 |
| 23 | 1.2197422 | 2.238264250 | 1.33269642 | -0.100566168 | 0.03982112 | 0.8992947814 | -0.006028399 |
| 25 | 2.5858535 | 0.570041000 | 1.64687207 | 0.338786735 | -0.09466248 | 0.4819556080 | -0.599825731 |
| 40 | 0.1951588 | -0.437188132 | -0.08109404 | 1.876521895 | 0.72568747 | 0.5471648538 | 2.824305075 |
| 47 | -0.8294247 | 0.664468731 | -1.33779667 | -0.100566168 | 0.03982112 | -0.4179319848 | 0.334576162 |

1-10 of 154 rows | 1-8 of 9 columns

Previous 1 2 3 4 5 6 ... 16 Next

Klasyfikacja KNN dla k = 3 - zbiór uczący

```
library(class)
knn.uczacy <- knn(train = uczacy[, -9], # zmienne objaśniające - zbiór uczący
                  test = uczacy[, -9], # zmienne objaśniające - zbiór, na którym weryfikujemy model
                  cl = uczacy[, 9],    # zmienna wynikowa ze zbioru uczącego
                  k = 3)              # liczba sąsiadów
knn.uczacy
```

```
[1] 1 0 1 0 1 1 0 1 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 1 1 1 0
[54] 0 1 0 0 0 0 1 0 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 0 0 1
[107] 1 0 0 0 0 1 1 0 0 1 0 0 0 0 1 1 1 0 0 1 1 0 0 0 1 0 1 0 0 0 1 0 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0
[160] 1 1 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 0 0 1 1 0 0 1 0 0 0 0
[213] 0 0 1 0 0 1 0 1 1 0 0 0 1 1 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 1 1 1 0 0 1 0 0 1 0 0 0
[266] 0 0 1 0 0 0 0 1 1 1 0 0 0 1 0 0 0 0 0 1 0 1 1 0 0 1 1 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0
[319] 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 1 0 1 0 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0
[372] 0 0 1 0 0 0 0 0 1 0 1 0 1 1 0 0 1 0 0 1 1 1 0 0 1 1 1 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 1 1 1 0 0 0
[425] 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 0 1 0 1 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 1 1
[478] 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 1 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 1 0 1
[531] 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 1 0 0 0 1 1 0 0
[584] 1 0 0 0 0 0 1 0 1 0 1 0 0 1 1 0 1 0 1 1 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 1 0 0
```

Levels: 0 1

Macierz błędów dla zbioru uczącego:

| knn.uczacy | 0 | 1 |
|------------|-----|-----|
| 0 | 363 | 52 |
| 1 | 43 | 156 |

Prawdziwie negatywny: 363

Prawdziwie pozytywny: 156

Fałszywie pozytywny: 52

Fałszywie negatywny: 43

Więcej razy uzyskaliśmy prawdziwie negatywny niż pozytywny, mniej razy fałszywie negatywny niż pozytywny.

```
[1] 0.8452769
[1] 0.8940887
[1] 0.75
```

Dokładność: 84,53%

Czułość: 89,41%

Specyficzność: 75%

Wysoka czułość świadczy o małym odsetku chorych osób, które nie zostały rozpoznane.

Klasyfikacja KNN dla $k = 3$ - zbiór testowy

```
##{r}
knn.testowy <- knn(train = uczacy[, -9],
                  test = testowy[, -9],
                  cl = uczacy[, 9],
                  k = 3)
knn.testowy
##
[1] 0 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 1 0 1 0 1 1 0 0 1 0 0 0 1 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 0 1 0 0 0 1
[54] 0 0 0 1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
[107] 1 1 1 0 0 0 0 0 0 0 1 1 1 1 0 1 0 0 1 0 0 0 0 0 1 1 1 1 1 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
Levels: 0 1
```

```
knn.testowy  0  1
             0 78 31
             1 16 29
```

Prawdziwie negatywny: 78

Prawdziwie pozytywny: 29

Fałszywie pozytywny: 31

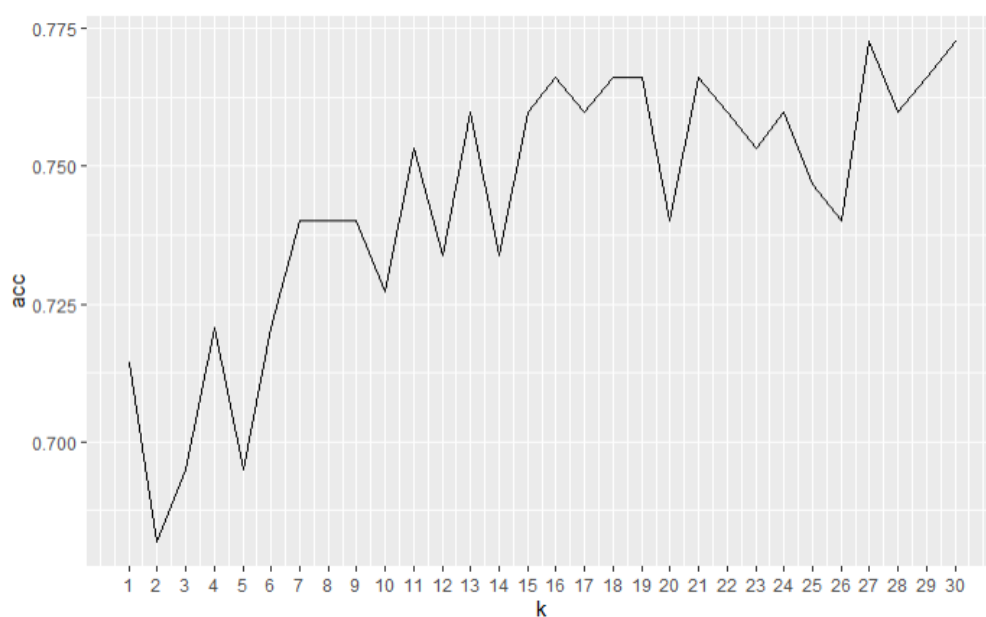
Fałszywie negatywny: 16

Tak samo jak w przypadku zbioru uczącego więcej razy uzyskaliśmy prawdziwie negatywny niż pozytywny, mniej razy fałszywie negatywny niż pozytywny.

```
[1] 0.6948052
[1] 0.8297872
[1] 0.4833333
```

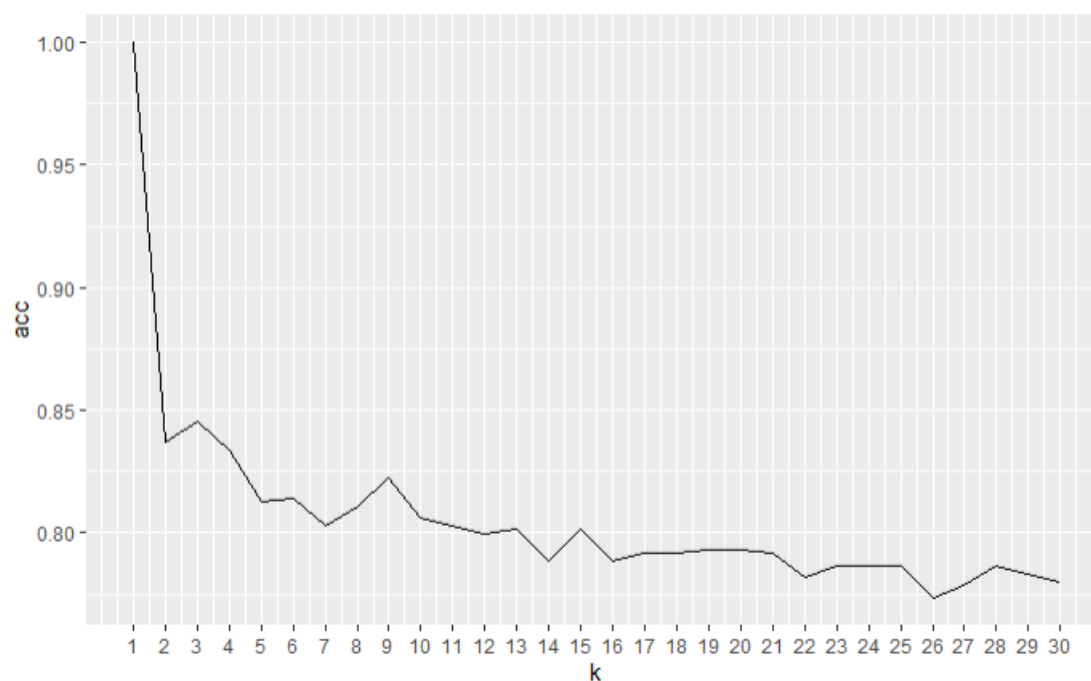
Uzyskano niższe parametry niż dla zbioru uczącego. Specyficzność poniżej 50% oznacza, że ilość fałszywie pozytywnych jest większa od prawdziwie pozytywnych

Wybór liczby sąsiadów – zbiór testowy



Najmniejsza dokładność dla dwóch sąsiadów wynosi 66,88%, natomiast najkorzystniejszą opcją jest przyjęcie $k = 27$, dokładność osiąga wtedy 77,27%.

Porównanie dokładności w zbiorze uczącym:



W zbiorze uczącym mała liczba sąsiadów daje największą dokładność. Dla k równego 1 osiągamy dokładność 100%, powyżej 15 sąsiadów dokładność jest poniżej 80%.

5. Ważona metoda k najbliższych sąsiadów (KKNN)

Ważona metoda k najbliższych sąsiadów jest pewnym udoskonaleniem metody k najbliższych sąsiadów. „Głosowanie” dotyczące klasyfikacji wciąż uwzględnia k sąsiadów. Waga „głosu” nie jest taka sama, a zależy od odległości sąsiada od klasyfikowanego obiektu.

Dane zestandaryzowane, zmienna „diabetes” jako typ factor.

```
##{r}
str(uczacy)

R Console
data.frame
614 x 9

'data.frame': 614 obs. of 9 variables:
 $ pregnant: num [1:614, 1] 1.169 -1.14 -0.274 -1.14 0.592 ...
 .. attr(*, "scaled:center")= num 3.95
 .. attr(*, "scaled:scale")= num 3.47
 $ glucose : num [1:614, 1] 1.9994 -0.994 0.669 -0.6614 0.0703 ...
 .. attr(*, "scaled:center")= num 121
 .. attr(*, "scaled:scale")= num 30.1
 $ pressure: num [1:614, 1] -0.3539 0.6513 -0.0188 -0.8565 -0.0188 ...
 .. attr(*, "scaled:center")= num 72.2
 .. attr(*, "scaled:scale")= num 11.9
 $ triceps : num [1:614, 1] 0.8148 0.0108 0.0108 0.0108 1.8485 ...
 .. attr(*, "scaled:center")= num 28.9
 .. attr(*, "scaled:scale")= num 8.71
 $ insulin : num [1:614, 1] 3.86759 -0.00517 -0.00517 -0.00517 0.84021 ...
 .. attr(*, "scaled:center")= num 156
 .. attr(*, "scaled:scale")= num 87.5
 $ mass : num [1:614, 1] -0.33706 0.00809 -0.35207 -1.5676 0.18817 ...
 .. attr(*, "scaled:center")= num 32.3
 .. attr(*, "scaled:scale")= num 6.66
 $ pedigree: num [1:614, 1] 0.418 0.375 0.858 -0.424 0.774 ...
 .. attr(*, "scaled:center")= num 0.477
 .. attr(*, "scaled:scale")= num 0.331
 $ age : num [1:614, 1] 2.1712 -0.5471 -0.5471 -0.7118 0.0295 ...
 .. attr(*, "scaled:center")= num 33.6
 .. attr(*, "scaled:scale")= num 12.1
 $ diabetes: Factor w/ 2 levels "0","1": 2 1 2 1 1 2 1 2 2 1 ...
```

Klasyfikacja KKNN dla k = 3 zbiór uczący

```
##{r}
library(kknn)
kknn.uczacy <- kknn(formula = uczacy[, 9]~., # formuła ze zmienną wynikową Y ze zbioru uczącego
                    train = uczacy[, -9], # zmienne objaśniające - zbiór uczący
                    test = uczacy[, -9], # zmienne objaśniające - zbiór, na którym weryfikujemy model
                    k = 3) # liczba sąsiadów

kknn.uczacy

call:
kknn(formula = uczacy[, 9] ~ ., train = uczacy[, -9], test = uczacy[, -9], k = 3)

Response: "nominal"
```

Prognoza wartości Y(diabetes):

```
##{r}
kknn.uczacy.wyniki <- fitted(kknn.uczacy)
kknn.uczacy.wyniki

[1] 1 0 1 0 0 1 0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 1 1 1 0 0 1 1 1 1 1 1 0 0 0 0 0 1 0 0 0 0 0 1 1 1 1 1 0
[54] 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 0 1 1 0 1 1 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 1
[107] 1 0 0 1 0 0 1 1 0 1 0 0 0 0 1 1 1 0 0 1 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0 1 1 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0
[160] 1 1 0 1 0 0 0 0 1 1 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 1 1 0 0 0 0 0 1 0 1 0 1 0 0 1 0 0 0 1 0 0 1 0 0 0 0
[213] 0 0 1 0 0 1 0 1 1 0 1 0 1 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 1 1 0 0 0 0 0 1 1 1 1 0 0 0 0 0 1 0 0 1
[266] 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 1 0 0 0 0 0
[319] 0 0 0 1 1 0 1 1 0 0 1 1 1 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 1
[372] 1 0 1 0 0 0 1 0 1 0 1 0 1 1 0 0 1 0 0 1 1 1 0 0 1 1 1 0 0 1 0 0 0 0 1 1 0 0 1 0 1 0 1 0 0 0 0 1 1 0 0 0
[425] 0 0 0 0 1 0 0 0 1 0 0 0 0 1 1 0 1 1 0 0 0 0 1 0 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 1 1 0 1 0 1 1
[478] 0 0 0 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 1 1 0 1 0 1 0
[531] 1 1 0 1 0 0 1 0 1 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 1 1 0 0 0 0 0 1 0 1 1 0 0 0 0 1 1 0 0
[584] 0 0 0 0 0 0 0 0 1 0 1 0 0 1 1 0 1 0 0 1 1 0 1 1 1 0 1 1 1 0 1 1 1 0

Levels: 0 1
```

Macierz błędów zbioru uczącego

```
kknn.uczacy.wyniki    0    1
                     0 406    0
                     1    0 208
```

Prawdziwie negatywny: 406

Prawdziwie pozytywny: 208

Fałszywie pozytywny: 0

Fałszywie negatywny: 0

Nie osiągnięto żadnych fałszywie pozytywnych oraz fałszywie negatywnych testów na cukrzyce.

```
[1] 1
[1] 1
[1] 1
```

Dokładność, czułość oraz specyficzność dla trzech sąsiadów w metodzie KKNN osiąga 100% co oznacza, że u wszystkich chorych kobiet została rozpoznana cukrzyca, a żadnej zdrowej kobiecie nie przypisano pozytywnego testu na cukrzyce.

Prognoza na zbiorze testowym:

```
##{r}
kknn.testowy <- kknn(formula = uczacy[, 9]~.,
  train = uczacy[, -9],
  test = testowy[, -9],
  k = 3)

kknn.testowy.wyniki <- fitted(kknn.testowy)
kknn.testowy.wyniki
t4 <- table(kknn.testowy.wyniki, testowy$diabetes)
t4

[1] 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 1 0 1 1 0 1 0 0 0 0 0 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 0 1 1 0 0 1
[54] 0 0 0 1 0 0 0 1 0 1 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0
[107] 1 1 1 0 1 0 0 0 0 0 1 1 1 1 0 1 0 0 1 0 1 0 0 1 1 1 1 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Levels: 0 1
```


Macierz błędu dla zbioru testowego

```
kkn.n.testowy.wyniki  0  1
                        0 76 26
                        1 18 34
```

Prawdziwie negatywny: 76

Prawdziwie pozytywny: 34

Fałszywie pozytywny: 26

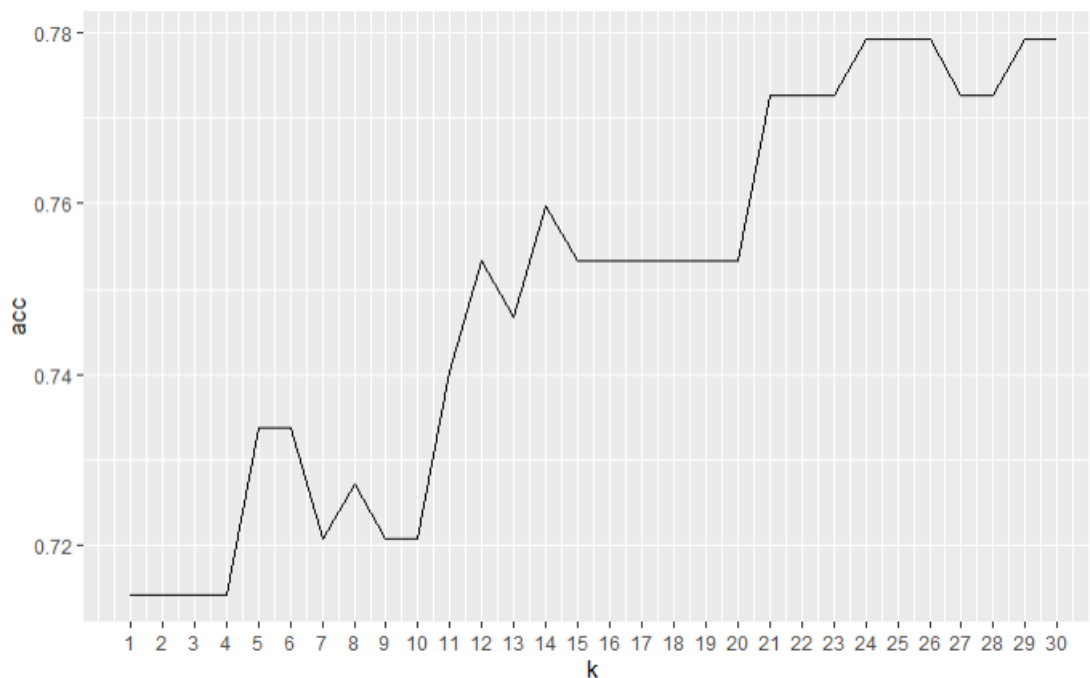
Fałszywie negatywny: 18

Fałszywie negatywne oraz pozytywne nie przekraczają ilości prawdziwie negatywnych i pozytywnych.

```
[1] 0.7142857
[1] 0.8085106
[1] 0.5666667
```

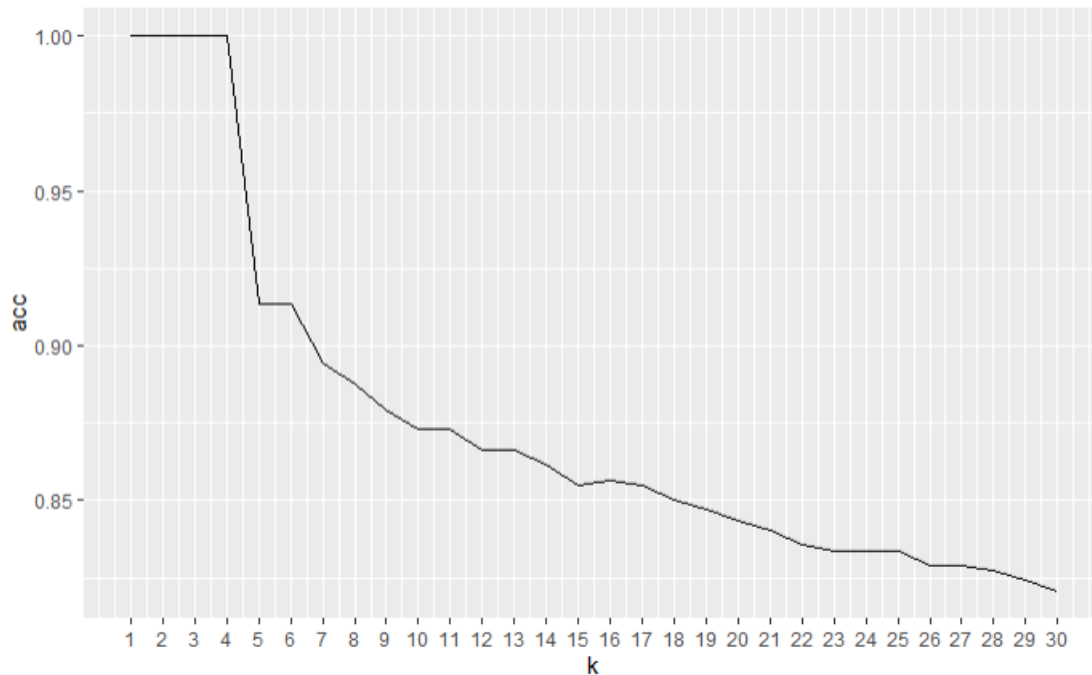
Podobnie jak w metodzie k najbliższych sąsiadów osiągnęliśmy niską specyficzność dla zbioru testowego na poziomie 56,67% co oznacza, że dużo zdrowych kobiet uzyskało fałszywie pozytywny wynik testu na cukrzyce.

Maksymalizacja dokładności na zbiorze testowym – wybór liczby sąsiadów:



Osiągnięto najniższy poziom dokładności równy 71,43% dla $k \in \{1,4\}$. Największą dokładność 77,92% uzyskamy przy przyjęciu $k \in \{24, 25, 26, 29, 30\}$.

Analiza dokładności zbioru uczącego:



Odwrótnie jak w przypadku zbioru testowego dla $k \in \langle 1,4 \rangle$ osiągnięto najwyższy poziom dokładności równy 100%. Gdy $k > 4$ dokładność zaczyna spadać i osiąga najniższy poziom 82,08% dla 30 sąsiadów.

6. Regresja logistyczna

Regresja logistyczna jest techniką regresyjną co oznacza, że jest ona zestawem narzędzi statystycznych służących do oszacowania zależności między zmiennymi. W regresji logistycznej, na podstawie zestawu cech ilościowych i jakościowych chcemy przewidzieć wartość zmiennej jakościowej.

Model regresji logistycznej:

```
##{r}
library(stats)
glm_model <- glm(diabetes ~ .,
  data = uczacy,
  family = "binomial")
summary(glm_model)
```

Call:
glm(formula = diabetes ~ ., family = "binomial", data = uczacy)

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|---------|--------|--------|
| -2.4807 | -0.7099 | -0.3886 | 0.6856 | 2.2153 |

Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) |
|-------------|----------|------------|---------|--------------|
| (Intercept) | -0.92848 | 0.11071 | -8.386 | < 2e-16 *** |
| pregnant | 0.46834 | 0.12258 | 3.821 | 0.000133 *** |
| glucose | 1.21123 | 0.13622 | 8.891 | < 2e-16 *** |
| pressure | -0.07176 | 0.11863 | -0.605 | 0.545267 |
| triceps | 0.07879 | 0.13554 | 0.581 | 0.561035 |
| insulin | -0.12982 | 0.11144 | -1.165 | 0.244054 |
| mass | 0.56597 | 0.13716 | 4.126 | 3.68e-05 *** |
| pedigree | 0.23337 | 0.11175 | 2.088 | 0.036775 * |
| age | 0.10546 | 0.12686 | 0.831 | 0.405790 |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 786.18 on 613 degrees of freedom
Residual deviance: 560.56 on 605 degrees of freedom
AIC: 578.56

Number of Fisher Scoring iterations: 5

Eliminujemy nieistotne zmienne: pressure, triceps, insulin oraz age.

Tworzymy nowy model:

```
##{r}
library(stats)
glm_model2 <- glm(diabetes ~ .,
  data = uczacy2,
  family = "binomial")
summary(glm_model2)
```

Call:
glm(formula = diabetes ~ ., family = "binomial", data = uczacy2)

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|---------|---------|--------|--------|
| -2.7327 | -0.7055 | -0.3926 | 0.6940 | 2.1931 |

Coefficients:

| | Estimate | Std. Error | z value | Pr(> z) |
|-------------|-----------|------------|---------|--------------|
| (Intercept) | -9.276532 | 0.808235 | -11.478 | < 2e-16 *** |
| pregnant | 0.147025 | 0.030069 | 4.890 | 1.01e-06 *** |
| glucose | 0.038941 | 0.004043 | 9.631 | < 2e-16 *** |
| mass | 0.084665 | 0.017153 | 4.936 | 7.97e-07 *** |
| pedigree | 0.695424 | 0.336901 | 2.064 | 0.039 * |

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 786.18 on 613 degrees of freedom
Residual deviance: 563.10 on 609 degrees of freedom
AIC: 573.1

Number of Fisher Scoring iterations: 5

W nowym modelu wszystkie zmienne są istotne. Najmniej istotna jest zmienna pedigree, pozostałe zmienne mają istotność na podobnym poziomie, co oznacza, że na wynik testu na cukrzycę największy wpływ ma liczba ciąż, stężenie glukozy w osoczu oraz masa kobiety.

Prognoza na zbiorze uczącym i testowym:

```

{r}
p <- predict(glm_model2, newdata = uczacy2, type = "response")
prog_train <- ifelse(p > 0.5, 1, 0)
head(prog_train, 20)

```

| | | | | | | | | | | | | | | | | | | | |
|-----|-----|-----|-----|-----|---|-----|-----|-----|-----|-----|-----|-----|----|-----|-----|-----|-----|-----|-----|
| 187 | 565 | 262 | 408 | 595 | 3 | 652 | 507 | 542 | 556 | 549 | 274 | 758 | 42 | 748 | 547 | 478 | 569 | 342 | 464 |
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |

```

{r}
p2 <- predict(glm_model2, newdata = testowy2, type = "response")
prog_test <- ifelse(p2 > 0.5, 1, 0)
head(prog_test, 20)

```

| | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|-----|-----|
| 4 | 5 | 6 | 7 | 11 | 20 | 23 | 25 | 40 | 47 | 48 | 55 | 58 | 63 | 84 | 92 | 95 | 99 | 100 | 101 |
| 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

Macierz błędów dla zbioru uczącego, dokładność, czułość oraz specyficzność:

```

{r}
tab_train <- table(prog_train, uczacy2$diabetes)
tab_train

```

| prog_train | 0 | 1 |
|------------|-----|-----|
| 0 | 357 | 96 |
| 1 | 49 | 112 |

```

{r}
(tab_train[1,1] + tab_train[2,2])/sum(tab_train)
tab_train[1,1]/(tab_train[1,1] + tab_train[2,1])
tab_train[2,2]/(tab_train[2,2] + tab_train[1,2])

```

```

[1] 0.7638436
[1] 0.8793103
[1] 0.5384615

```

Prawdziwie negatywny: 357

Prawdziwie pozytywny: 112

Fałszywie pozytywny: 96

Fałszywie negatywny: 49

Wystąpiła niska specyficzność na poziomie 53,85% dla zbioru uczącego dla regresji logistycznej, w przypadku KNN (75%) oraz KNNN (100%) była o wiele większa, co oznacza, że w tej metodzie więcej kobiet dostało fałszywie pozytywny wynik testu na cukrzycę.

Macierz błęd dla zbioru testowego, dokładność, czułość, specyficzność:

```
##{r}
tab_test <- table(prog_test, testowy2$diabetes)
tab_test

prog_test 0 1
0 85 24
1 9 36

##{r}
(tab_test[1,1] + tab_test[2,2])/sum(tab_test)
tab_test[1,1]/(tab_test[1,1] + tab_test[2,1])
tab_test[2,2]/(tab_test[2,2] + tab_test[1,2])
##
```

```
[1] 0.7857143
[1] 0.9042553
[1] 0.6
```

Prawdziwie negatywny: 85

Prawdziwie pozytywny: 36

Fałszywie pozytywny: 24

Fałszywie negatywny: 9

Co ciekawe wszystkie parametry (dokładność, czułość, specyficzność) osiągają większe wartości dla zbioru testowego, inaczej niż w poprzednich metodach, gdzie prognozy na zbiorach uczących były lepsze od tych przeprowadzonych na zbiorze testowym.

7. Liniowa analiza dyskryminacyjna (LDA)

Podobnie, jak wcześniej poznane metody, analiza dyskryminacyjna jest (statystyczną) metodą uczenia maszynowego z nauczycielem. To oznacza, że służy ona do możliwie najlepszego podziału obiektów wielowymiarowych na kilka rozłącznych grup.

Jest to metoda geometryczna, która koncentruje się na znalezieniu takiego kierunku rzutowania punktów na hiperpłaszczyznę, by jednocześnie:

- maksymalizować odległość między średnimi w grupach,
- minimalizować wariancję wewnątrzgrupową.

Im dalej od siebie będą położone punkty centralne i im mniejszy będzie rozrzut, tym mniej pokrywać się będą ich rozkłady.

Założenia dla zmiennych ilościowych:

- Równość wariancji w grupach,
- Rozkład normalny w grupach.

Sprawdzenie pierwszego założenia o równości wariancji w grupach:

Test Levene'a - test, którego zadaniem jest ocena czy wariancja w zbiorze danych jest równa w grupach. Jeżeli $p < 0,05$ to oznacza, że wariancje są niejednorodne (heterogeniczne), czyli występują różnice pomiędzy wariancjami w porównywanych grupach. Natomiast w przypadku $p > 0,05$, przyjmujemy założenie o homogeniczności wariancji.

H0: Wariancje w grupach są równe.

H1: Wariancje w grupach nie są równe.

```
##{r}
leveneTest(age ~ diabetes, dane)
leveneTest(pedigree ~ diabetes, dane)
leveneTest(mass ~ diabetes, dane)
leveneTest(insulin ~ diabetes, dane)
leveneTest(triceps ~ diabetes, dane)
leveneTest(pressure ~ diabetes, dane)
leveneTest(glucose ~ diabetes, dane)
leveneTest(pregnant ~ diabetes, dane)
##

Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1 2.2252 0.1362
766
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1 11.798 0.000625 ***
766
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1 1.407 0.2359
766
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1 0.0028 0.9581
766
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1 1.3469 0.2462
766
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1 0.1139 0.7359
766
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1 23.499 1.513e-06 ***
766
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Levene's Test for Homogeneity of Variance (center = median)
  Df F value Pr(>F)
group 1 22.747 2.212e-06 ***
766
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Założenie o równości wariancji w grupach nie jest spełnione dla zmiennych: *pedigree*, *pregnant* oraz *glucose*, dlatego nie sprawdzamy drugiego założenia o rozkładach normalnych w grupach i nie wykorzystujemy LDA.

8. Podsumowanie

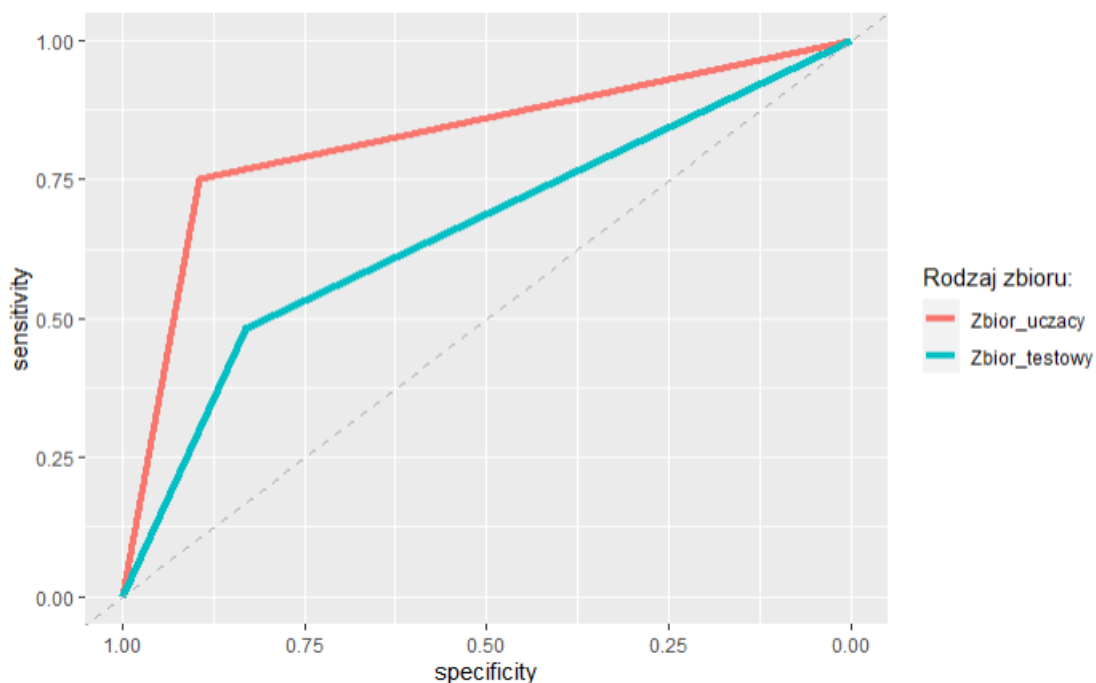
| | | Dokładność w zbiorze uczącym | Dokładność w zbiorze testowym |
|--------|----------------------|------------------------------|-------------------------------|
| METODA | KNN | 0,8452769 | 0,6948052 |
| | KKNN | 1 | 0,7142857 |
| | Regresja logistyczna | 0,7638436 | 0,7857143 |

Największą dokładność uzyskano dla KKNN w zbiorze uczącym a najmniejszą dla KNN dla zbioru testowego.

Krzywa ROC - opisuje zachowanie modelu dla różnych punktów odcięcia. Na osi x są poziomy specyficzności, a na osi y poziomy czułości. Model jest tym lepszy im bardziej krzywa ROC unosi się nad krzywą $y = x$.

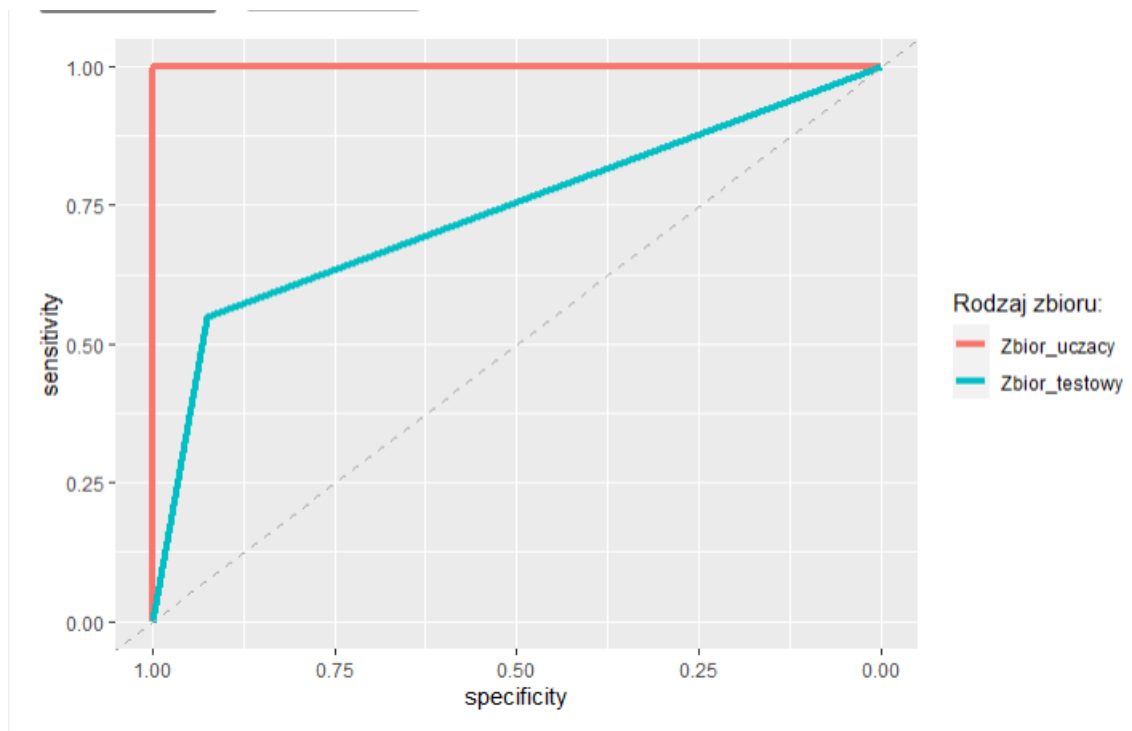
Położenie krzywej ROC zależy od jakości klasyfikatora. AUC to pole pod krzywą ROC. Dla AUC = 0.5 klasyfikator jest losowy. Dla AUC = 1 klasyfikator jest idealny.

Krzywa ROC dla KNN:



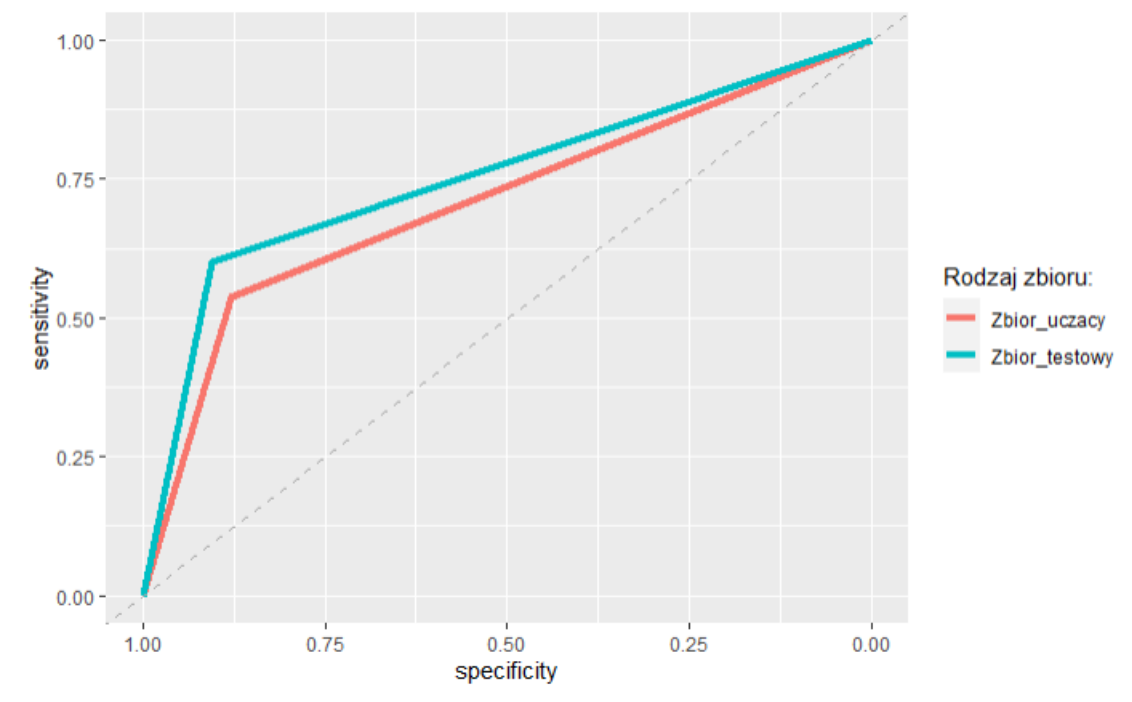
Krzywa ROC dla zbioru uczącego jest wyżej nad krzywą $y = x$ niż krzywa dla zbioru testowego.

Krzywa ROC dla KKNN:



Krzywa ROC dla zbioru uczącego osiąga najlepszy możliwy poziom. Krzywa dla zbioru uczącego najbardziej odbiega od krzywej ROC dla zbioru testowego niż w innych metodach.

Krzywa ROC regresja logistyczna:



Krzywa ROC dla zbioru testowego jest wyżej nad krzywą $y = x$ niż krzywa dla zbioru uczącego.


```

{r}
auc(ROC.knn.u)
auc(ROC.knn.t)

Area under the curve: 0.822
Area under the curve: 0.6566

{r}
auc(ROC.kknn.u)
auc(ROC.kknn.t)

Area under the curve: 1
Area under the curve: 0.7378

{r}
auc(ROC.r1.u)
auc(ROC.r1.t)

Area under the curve: 0.7089
Area under the curve: 0.7521

```

AUC KNN zbiór uczący = 0,822

AUC KNN zbiór testowy = 0,6566

AUC KNN zbiór uczący = 1

AUC KNN zbiór testowy = 0,7378

AUC regresja liniowa zbiór uczący = 0,7089

AUC regresja liniowa zbiór testowy = 0,7521

Uzyskano idealny klasyfikator KNN w zbiorze uczącym, ponieważ AUC wynosi 1. Najmniejszy wynik AUC osiągnięto dla zbioru testowego KNN.

9. Źródła

<https://www.poradnikzdrowie.pl/zdrowie/cukrzyca/cukrzyca-rodzaje-cukrzycy-przyczyny-objawy-leczenie-powiklania-aa-WXha-YTn5-FNbo.html>

<https://pogotowiestatystyczne.pl/slowniki/test-levene/>

<https://www.statystyczny.pl/macierz-bledow-raport-dokladnosc-czulosc-precyzja/>

<https://mathspace.pl/matematyka/ocena-jakosci-klasyfikacji-czesc-2/>

https://pl.wikipedia.org/wiki/Czu%C5%82o%C5%9B%C4%87_i_swoisto%C5%9B%C4%87

<https://datascience.eu/pl/uczenie-maszynowe/zrozumienie-auc-krzywa-roc/>

<https://algolytics.pl/tutorial-jak-ocenic-jakosc-i-poprawnosc-modeli-klasyfikacyjnych-czesc-4-krzywa-roc/>

https://upel.agh.edu.pl/wz/pluginfile.php/166677/mod_resource/content/1/sad_w4.pdf

https://upel.agh.edu.pl/wz/pluginfile.php/165887/mod_resource/content/1/sad_wyk%C5%82ad3.pdf