Politechnika Wrocławska

Wydział Zarządzania

METODA K-NN

Stellar Classification Dataset

Projekt – Techniki eksploracji danych

Grupa: Środa TN 9:15-11:00

Autor:

Damian Kędzierski 260493

Prowadzący: dr hab. inż. Zbigniew Michna

WROCŁAW lato 2022

Spis treści

1. Opis zestawu danych

Do naszego projektu wykorzystaliśmy zestaw danych klasyfikacji obiektów w kosmosie, który zaczerpnęliśmy ze strony https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17.

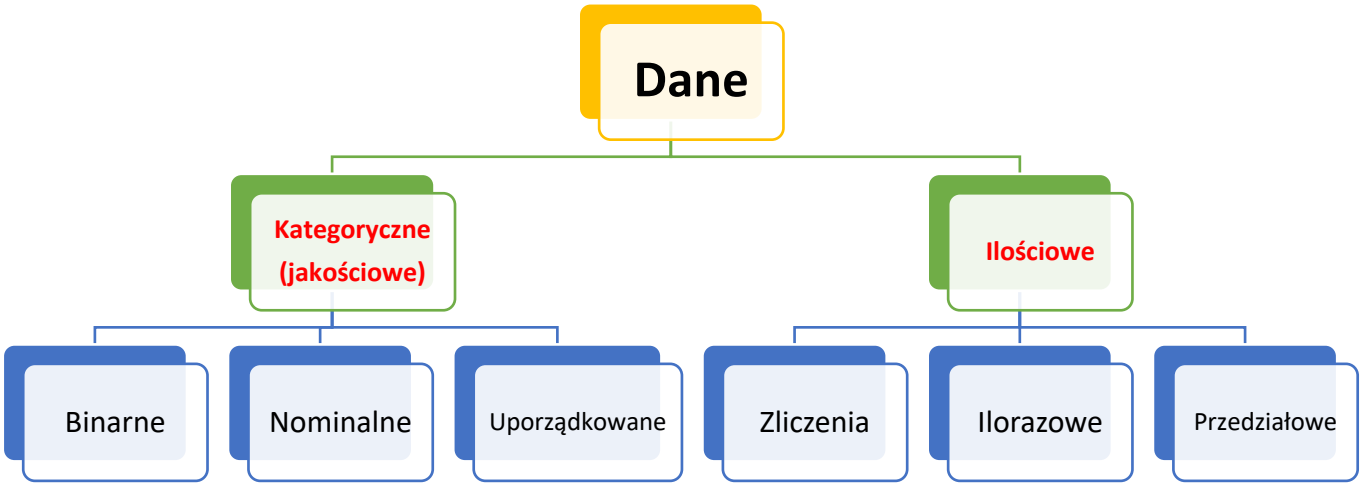
Krótki opis klasyfikacji gwiazd, czyli przybliżenie tematyki:

W astronomii klasyfikacja gwiazd to klasyfikacja gwiazd na podstawie ich cech spektralnych. Schemat klasyfikacji galaktyk, kwazarów i gwiazd jest jednym z najbardziej fundamentalnych w astronomii. Wczesne katalogowanie gwiazd i ich rozmieszczenie na niebie doprowadziło do zrozumienia, że tworzą one naszą własną galaktykę. Ten satelita danych ma na celu klasyfikację gwiazd, galaktyk i kwazarów na podstawie ich charakterystyki spektralnej.

Kolumny cech i klasy:

Dane składają się ze 100 000 obserwacji przestrzeni kosmicznej wykonanych przez SDSS (Sloan Digital Sky Survey). Każda obserwacja jest opisana przez 17 kolumn cech i 1 kolumnę klasy, która identyfikuje ją jako gwiazdę, galaktykę lub kwazar.

2. Opis zawartości



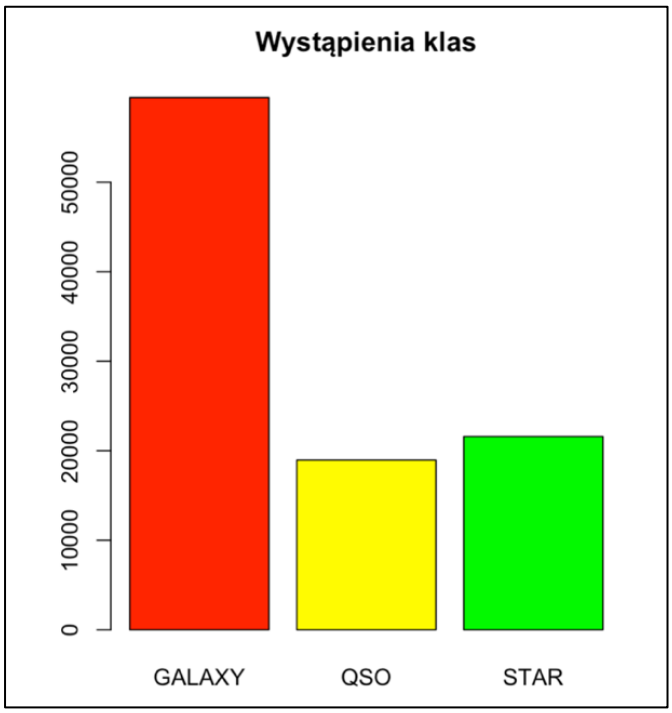
2.1. Przygotowane dane:

Zdecydowaliśmy się zredukować liczbę cech z 17 do 7, które są naszym zdaniem najważniejszymi zmiennymi ilościowymi. Ponadto przekształciliśmy zmienną kategoryczną do formatu ramki danych (uporządkowanej), dzięki czemu zdiagnozowaliśmy ją jako zmienna kategoryczna uporządkowana.

1. alfa = kąt wzniesienia w prawo (w epoce J2000)
2. delta = kąt deklinacji (w epoce J2000)
3. u = filtr ultrafioletowy w układzie fotometrycznym
4. g = Zielony filtr w układzie fotometrycznym
5. r = Czerwony filtr w układzie fotometrycznym
6. i = Filtr bliskiej podczerwieni w układzie fotometrycznym
7. z = Filtr podczerwieni w układzie fotometrycznym
8. class = klasa obiektu (galaktyka, gwiazda lub obiekt kwazara)

2.2. Zmienna kategoryczna

Po wczytaniu i przygotowaniu danych jako zmienną kategoryczną (uporządkowaną) użyliśmy zmienną *class*, która klasyfikuje dane do jednej z poniższych kategorii, za pomocą zmiennych ilościowych



- **GALAXY** - Galaktyka
- **STAR** - Gwiazda
- **QSO** – Kwazar

Na podstawie podanego wykresu należy wysnuć wnioski, że w wyraźnej większości cechami zmiennej kategorycznej są galaktyki. Następnie w kolejności są prezentują się gwiazdy, a w najmniejszej obiekty kwazarów.

2.3. Opis za pomocą charakterystyk statystycznych

Za pomocą charakterystyk statystycznych sprawdziliśmy poszczególne zmienne numeryczne, które uzyskaliśmy z funkcji „summary”. Stanowią one najważniejsze dane statyczne i są to kolejno:

- Wartość minimalna i maksymalna
- pierwszy kwartyl,
- mediana,
- średnia arytmetyczna,
- trzeci kwartyl,

```
     alpha             delta                u                 g                r                i                z          
 Min.   :  0.0055   Min.   :-18.785   Min.   :-9999.00   Min.   :-9999.00   Min.   : 9.822   Min.   : 9.47   Min.   :-9999.00  
 1st Qu.:127.5182   1st Qu.:  5.147   1st Qu.:   20.35   1st Qu.:   18.96   1st Qu.:18.136   1st Qu.:17.73   1st Qu.:   17.46  
 Median :180.9007   Median : 23.646   Median :   22.18   Median :   21.10   Median :20.125   Median :19.41   Median :   19.00  
 Mean   :177.6291   Mean   : 24.135   Mean   :   21.98   Mean   :   20.53   Mean   :19.646   Mean   :19.08   Mean   :   18.67  
 3rd Qu.:233.8950   3rd Qu.: 39.902   3rd Qu.:   23.69   3rd Qu.:   22.12   3rd Qu.:21.045   3rd Qu.:20.40   3rd Qu.:   19.92  
 Max.   :359.9998   Max.   : 83.001   Max.   :   32.78   Max.   :   31.60   Max.   :29.572   Max.   :32.14   Max.   :   29.38  
```

Poniższe dane uzyskaliśmy natomiast z biblioteki „psych”, dzięki funkcji „describe” a najistotniejsze z nich poza wyszczególnionymi dzięki funkcji „summary” to:

- odchylenie standardowe,
- zasięg,
- współczynnik krzywej,
- współczynnik kurtozy

```
      vars     n   mean    sd median trimmed   mad      min    max    range    skew kurtosis   se
alpha    1 1e+05 177.63 96.50 180.90  177.13 78.88     0.01 360.00   359.99   -0.03    -0.54 0.31
delta    2 1e+05  24.14 19.64  23.65   23.45 25.60   -18.79  83.00   101.79    0.18    -1.04 0.06
u        3 1e+05  21.98 31.77  22.18   22.09  2.47 -9999.00  32.78 10031.78 -313.84 98991.44 0.10
g        4 1e+05  20.53 31.75  21.10   20.73  1.98 -9999.00  31.60 10030.60 -314.27 99171.20 0.10
r        5 1e+05  19.65  1.85  20.13   19.75  1.85     9.82  29.57    19.75   -0.51    -0.38 0.01
i        6 1e+05  19.08  1.76  19.41   19.16  1.81     9.47  32.14    22.67   -0.40    -0.23 0.01
z        7 1e+05  18.67 31.73  19.00   18.81  1.74 -9999.00  29.38 10028.38 -314.75 99374.39 0.10
```

Dodatkowo warto zwrócić uwagę na daną alfa, czyli na kąt wzniesienia w prawo, dla której zmienne numeryczne osiągają ekstremum.

2.4. Pozostałe informacje, użyte w programie

- library(tidyverse)
- library(ggplot2)
- library(class)

Narzędzia „tidyverse” tworzą pakiety, w których zaimplementowano filozofię podejścia do analizy danych. W wypadku wykresów i pakietu „ggplot2” została wykorzystana gramatyka grafiki Wilkinsona. Narzędzia te są odpowiedzią na zapotrzebowanie tych osób, które chcą relatywnie szybko i łatwo przygotować dane, poddać je analizie i ostatecznie wyniki wizualizować. Nie oznacza to, ze bez tych pakietów takich operacji nie jesteśmy w stanie zrobić. Jednak redukują one złożoność całej procedury, przez co składnia jest bardziej przejrzysta i intuicyjna.

Funkcje obsługujące cechy obiektów „class” to klasa (rodzaj) obiektu. „Class” jako zmienna to klasa, jaką reprezentuje obiekt; ważna w programowaniu obiektowym.

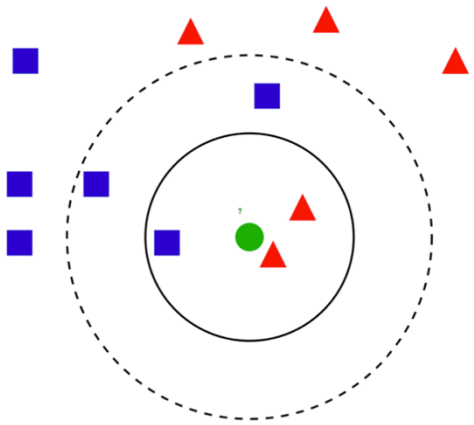
3. Cel badań

- Głównym celem badań jest klasyfikacja danych do jednej z cech zmiennej kategorycznej, czyli przedstawienie, które obiekty stanowią galaktykę, gwiazdę lub kwazar.
    - Pierwszy punkt w naszym badaniu to stworzenie funkcji normującej, żeby uzyskać jak najbardziej precyzyjne wyniki wartości naszych zmiennych do pozostałych obliczeń.
    - Następnie priorytetem jest wyszczególnienie ze zbioru danych bazowych oraz danych badanych i tym samym utworzyć zbiór treningowy z danych bazowych oraz zbiór testowy z danych badanych.
    - Ostatni cel stanowi zestawienie klasyfikacji przy użyciu metody k najbliższych sąsiadów i wybór parametru k o najwyższej skuteczności.

4. Metoda k najbliższych sąsiadów.

4.1. Metoda K Najbliższych sąsiadów (K-NN)

Metoda k najbliższych sąsiadów (k-NN) jest zaliczana do metod klasyfikacji (algorytm nadzorowany uczenia maszynowego). Obiekty mają przypisane jakieś cechy (atrybuty) - zmienne kategoryczne i zmienne numeryczne. Dla danego obiektu bez cechy kategorycznej i dla ustalonego k wybieramy k najbliższych obiektów (np. na podstawie odległości euklidesowej dla zmiennych unormowanych numerycznych) i przypisujemy temu obiektowi cechę kategoryczną, która występuje najczęściej spośród k najbliższych obiektów.



- Zmienną **kategoryczną** (zmienną celu, zależna) jest → {trójkąt, kwadrat}
- Zmienna **numeryczna** (zmienna niezależna) → współrzędne punktu na płaszczyźnie (x,y)

Obliczamy **odległość pomiędzy punktem zielonym** bez przypisanej cechy w celu określenia (predykcji)

Dokładny algorytm k-NN

obiekt	K	X_1	X_2	...	X_d
1	k_1	x_{11}	x_{12}	...	x_{1d}
2	k_2	x_{21}	x_{22}	...	x_{2d}
⋮	⋮	⋮	⋮	⋮	⋮
n	k_n	x_{n1}	x_{n2}	...	x_{nd}

K – zmienna kategoryczna (objaśniana = zależna)

Xj – zmienne numeryczne (objaśniająca = niezależna)

4.2. Unormowanie zmiennych numerycznych

Zmienne Xj należy unormować, aby zapewnić większą elastyczność danych pozbyć się ewentualnych niespójności ich złożoności. Przykładowo zmienne mogą być w różnych jednostkach (np. długość, waga), a po unormowaniu za pomocą skalowania pierwotnych danych do pewnych przedziałów, najbardziej przydatnych, które mają wartość między 0-1 i ponadto są bez jednostek.

Normujemy metodą min-max (w każdej kolumnie).

$$x'_{ij} = \frac{x_{ij} - min_{1 \le i \le n} x_{ij}}{max_{1 \le i \le n} x_{ij} - min_{1 \le i \le n} x_{ij}}$$

Następnie dla danego obiektu bez cechy K z cechami (y1,y2, . . . , yd) normujemy cechy numeryczne (dodając ten obiekt do wszystkich obiektów) otrzymując (y ′ 1 ,y′ 2 , . . . , y′ d) i obliczamy odległość euklidesową do każdego obiektu z danych:

$$d_i = \sqrt{\sum_{j=1}^{d} (x'_{ij} - y'_{ij})^2}$$

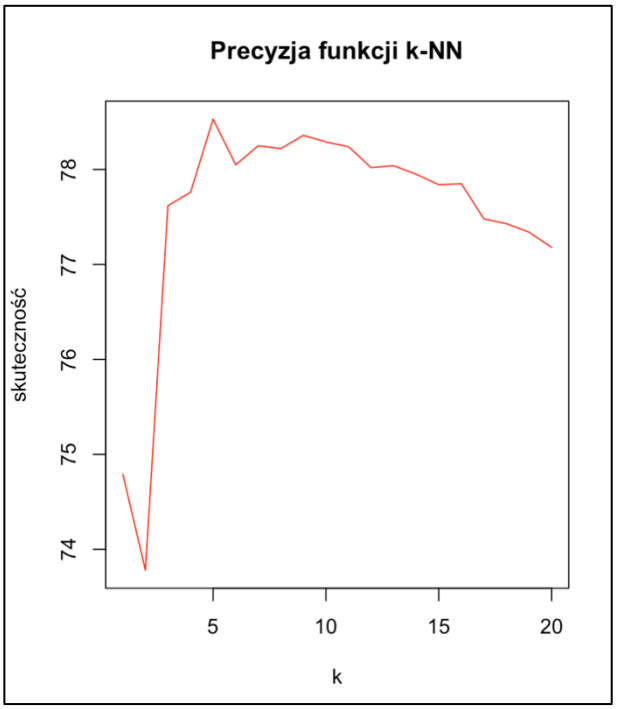
Następnie dla ustalonego k znajdujemy k najbliższych obiektów i przypisujemy cechę zmiennej objaśniane (zależnej, kategorycznej) tych obiektów, których jest najwięcej pośród k najbliższych obiektów.

4.3. Wybór najefektywniejszego parametru k

W celu uzyskania najbardziej precyzyjnego k iterowaliśmy za pomocą pętli „for” po i elementach. Efektem było otrzymanie k = 5, które ma najwyższą skuteczność i wynosi ona 78,53%.

Precyzja naszego k w funkcji k najbliższych sąsiadów prezentuje się na poniższym wykresie następująco:

```
"krok numer:  1 , skuteczność k =  74.79"
"krok numer:  2 , skuteczność k =  73.78"
"krok numer:  3 , skuteczność k =  77.62"
"krok numer:  4 , skuteczność k =  77.76"
"krok numer:  5 , skuteczność k =  78.53"
"krok numer:  6 , skuteczność k =  78.05"
"krok numer:  7 , skuteczność k =  78.25"
"krok numer:  8 , skuteczność k =  78.22"
"krok numer:  9 , skuteczność k =  78.36"
"krok numer:  10 , skuteczność k =  78.29"
"krok numer:  11 , skuteczność k =  78.24"
"krok numer:  12 , skuteczność k =  78.02"
"krok numer:  13 , skuteczność k =  78.04"
"krok numer:  14 , skuteczność k =  77.95"
"krok numer:  15 , skuteczność k =  77.84"
"krok numer:  16 , skuteczność k =  77.85"
"krok numer:  17 , skuteczność k =  77.48"
"krok numer:  18 , skuteczność k =  77.43"
"krok numer:  19 , skuteczność k =  77.34"
"krok numer:  20 , skuteczność k =  77.18"
```



5. Badanie docelowe

5.1. Utworzenie sztucznych danych badanych

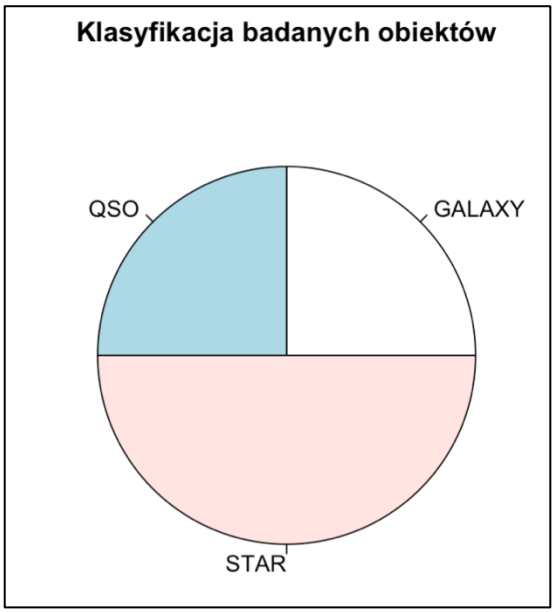
Poniżej prezentujemy stworzone przez nas dane badane, które stanowią część testową i składają się ze zmiennych numerycznych.

Cechy	alfa	delta	u	g	r	i	z
1.	133.68911	30.494632	21.37902	25.21130	19.99501	20.96573	19.99371
2.	22.05256	9.497881	23.89214	21.35644	20.18345	21.84956	15.76756
3.	119.17502	38.757654	22.01756	21.47564	18.85662	18.18365	17.77564
4.	42.07561	32.746019	21.86580	18.68362	17.18674	16.57467	16.13549

5.2. Przeprowadzenie badania

Badanie zaczęliśmy od funkcji normującej, a następnie stworzyliśmy tabelę zmiennych numerycznych. Zmienne te stanowią nasze dane bazowe i tworzą część uczącą. Tych zmiennych jest 100 000. Natomiast w danych badanych, które stanowią część testową możemy wyróżnić 4 zmienne.

6. Opis wyników i wnioski



Wykres klasyfikacji badanych obiektów przedstawia, zakwalifikowane dane badane do odpowiednich klas. Widać, że gwiazd jest dwa razy więcej niż galaktyk i kwazarów. Prezentuje to tabela poniżej:

odp		
GALAXY	QSO	STAR
1	1	2

Dzięki unormowaniu zmiennych numerycznych mogliśmy wyznaczyć optymalną wartość parametru k. Przy pomocy metody k-najbliższych sąsiadów uzyskaliśmy precyzję parametru k. Tutaj należy dodać, że nasze wnioski mogłyby być inne, gdyby padł wybór na przybliżoną wartość aktualnego parametru. Skuteczność naszego badania wynosi 78,53%, zatem jest wysoka i należy całą analizę interpretować w opisany przez nas sposób. Istnieje odsetek niepoprawnie dostosowanych kategorii o czym należy pamiętać.

7. Kod źródłowy

```
library(tidyverse)

library(ggplot2)

library(class)


#wczytanie danych

star<- read.csv("star_classification.csv", header=TRUE, sep=",", dec=".",
na.strings="NA")


#Przygotowane dane

dane = star[,c(2:8, 14)]

dane$class = ordered(as.factor(dane$class)) #Zmienna kategoryczna uporządkowana


#Podstawowe informacje o danych

names(dane)

str(dane)

dim(dane)

class(dane)


#Częstość występowania każdej klasy

freq <- table(dane$class)

barplot(freq, col = c("red", "yellow", "green"),  main = "Wystąpienia klas")



#statystyki opisowe zmiennych numerycznych

summary(dane[,-8])


library(psych)

describe(dane[,-8])
```

```
### K-NN
# Losowe 90% wierszy
random <-sample(1:nrow(dane),size=0.9*nrow(dane))
  #Funkcja normująca
f_nor <-function(x) { (x-min(x))/(max(x)-min(x)) }
  # Normuje kolumny zmiennych numerycznych i zapisuje jako ramka danych bez kolumny 8
dane.norm<-as.data.frame(lapply(dane[,c(-8)],f_nor))
  # Zbiór treningowy spośród 90% losowo wybranych wierszy
dane.train<-dane.norm[random,]
  # Zbiór testowy z pozostałych 10% wierszy
dane.test<-dane.norm[-random,]
  # Kolumnę prawdziwej klasyfikacji dla zbioru treningowego
dane.train.category <- dane[random,8]
  #Kolumna prawdziwej klasyfikacji dla zbioru testowego
dane.test.category <- dane[-random,8]

#Poszukiwanie najlepszego k = i {i = 1,...,20}
skutecznosc <- 0
for(i in 1:20) {
  # Klasyfikuje obiekty ze zbioru testowego dla k = i
  d = knn(dane.train,dane.test,cl=dane.train.category,k=i)
  # Tworzy macierz błędów (1-szy wiersz prawidłowa klasyfikacja)
  error_matrix <- table(d,dane.test.category)
  # Funkcja licząca częstość prawidłowych prognoz k (efektywność k)
  f_correct <- function(x){ sum(diag(x)/(sum(rowSums(x)))) * 100 }

      print(paste("krok numer: ", i, ", skuteczność k = ",f_correct(error_matrix)))
      skutecznosc[i] = f_correct(error_matrix)
}
skutecznosc
plot(skutecznosc, type = "l", xlab = "k", ylab = "skuteczność", main = "Precyzja
funkcji k-NN", col = "red")
max = max(skutecznosc)
which(skutecznosc == max) #najlepsze k
```

```
###DOCELOWE BADANIE
#funkcja normująca
nor <- function(x){(x - min(x))/(max(x)-min(x))}

#stworzenie tabeli zmiennych numerycznych (dane bazowe)
dane1 <- dane[, c(-8)]
#wprowadzenie badanych wartości (dane badane)
star1=c(133.689107, 30.4946318, 21.37902, 25.21130 ,19.99501, 20.96573, 19.99371)
star2=c(22.052556, 9.4978808, 23.89214, 21.35644, 20.18345, 21.84956, 15.76756)
star3=c(119.175021, 38.7576541, 22.01756, 21.47564, 18.85662, 18.18365, 17.77564)
star4=c(42.075615, 32.7460194, 21.86580, 18.68362, 17.18674, 16.57467, 16.13549 )
#połączenie danych bazowych z danymi badanymi
dane2=rbind.data.frame(dane1, star1, star2, star3, star4)

#normalizacja oraz rzutowanie typu (na ramkę danych)
dane_norm=as.data.frame(lapply(dane2[,c(-8)], nor))

#podział danych na część UCZĄCĄ (dane bazowe) i TESTOWĄ (dane badane)
dane.train=dane_norm[c(1:100000),]
dane.test=dane_norm[c(100001:100004),]
#Cechy, które będziemy nadawać badanym danym (klasa - zmienna kategoryczna)
dane.target.type <- dane[,8]

#klasyfikacja obiektów dla najlepszego k = 5
result <- knn(dane.train, dane.test, cl=dane.target.type,k=5)
table(result)

pie(table(result), main = "Klasyfikacja badanych obiektów")
```