**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 3](#_Toc107143349)

[2. Opis zawartości 3](#_Toc107143350)

[2.1. Przygotowane dane: 3](#_Toc107143352)

[2.2. Zmienna kategoryczna 4](#_Toc107143353)

[2.3. Opis za pomocą charakterystyk statystycznych 4](#_Toc107143354)

[2.4. Pozostałe informacje, użyte w programie 5](#_Toc107143355)

[3. Cel badań 5](#_Toc107143356)

[4. Metoda k najbliższych sąsiadów. 7](#_Toc107143357)

[4.1. Metoda K Najbliższych sąsiadów (K-NN) 7](#_Toc107143358)

[4.2. Unormowanie zmiennych numerycznych 7](#_Toc107143359)

[4.3. Wybór najefektywniejszego parametru k 8](#_Toc107143360)

[5. Badanie docelowe 9](#_Toc107143361)

[5.1. Utworzenie sztucznych danych badanych 9](#_Toc107143362)

[5.2. Przeprowadzenie badania 9](#_Toc107143363)

[6. Opis wyników i wnioski 9](#_Toc107143364)

[7. Kod źródłowy 10](#_Toc107143365)

# 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.

# Opis zawartości

## 

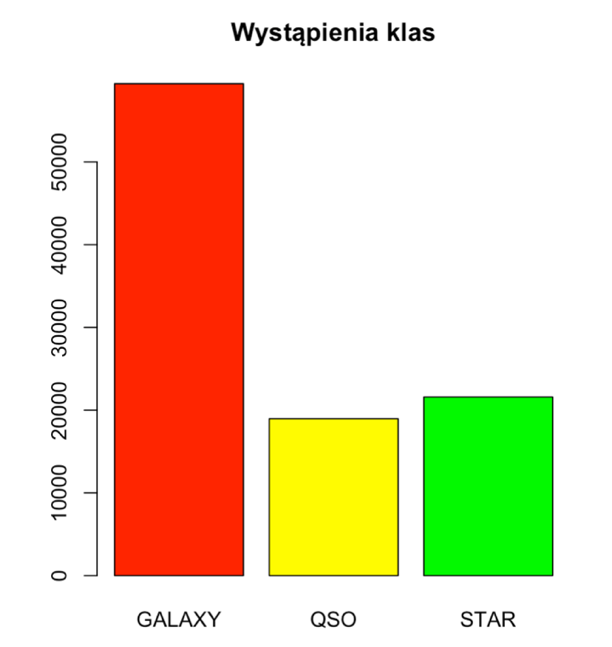
## Przygotowane dane:

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)

## 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



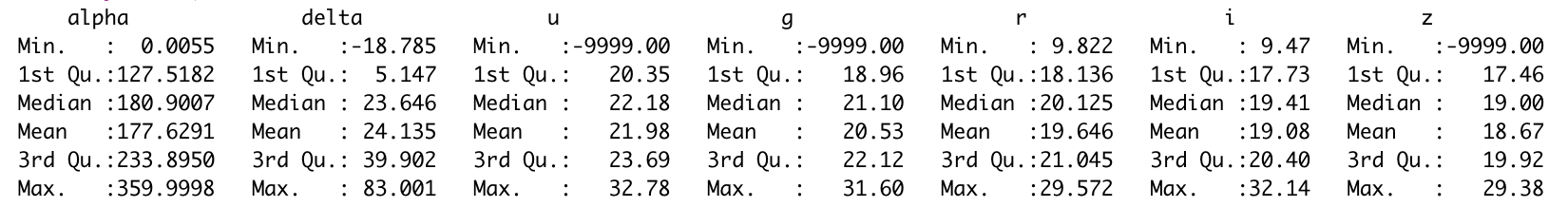
* **GALAXY** - Galaktyka
* **STAR** - Gwiazda
* **QSO** – Kwazar

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.

## 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,



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

Obraz zawierający stół

Opis wygenerowany automatycznie

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

## 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.

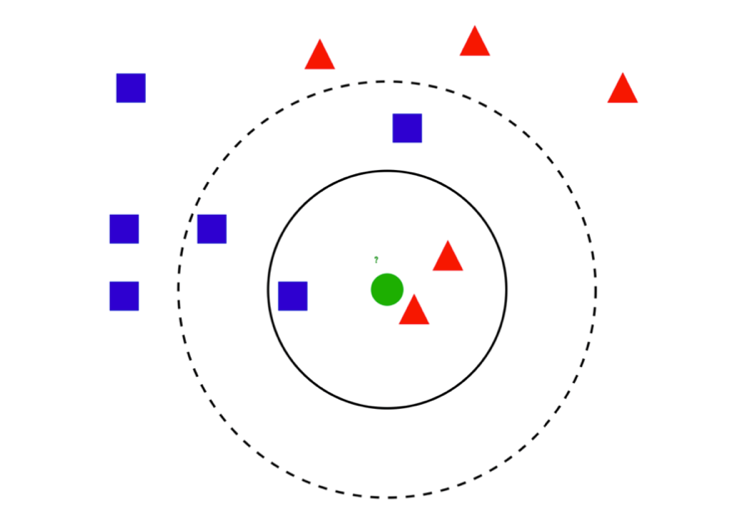
# Cel badań

* 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.

# Metoda k najbliższych sąsiadów.

## 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

Obraz zawierający stół

Opis wygenerowany automatycznie

**K** – zmienna kategoryczna (objaśniana = zależna)

**Xj** – zmienne numeryczne (objaśniająca = niezależna)

## 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).

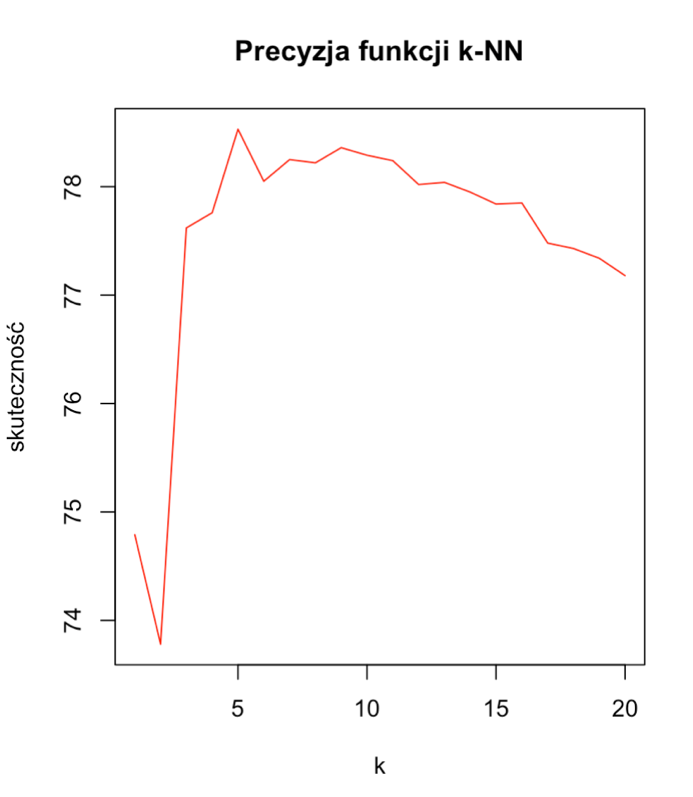
Następnie dla danego obiektu bez cechy K z cechami (y1,y2, . . . , yd) normujemy cechy numeryczne (dodając ten obiekt do wszystkich obiektów) otrzymując (y ′ 1 ,y′ 2 , . . . , y′ d ) i obliczamy odległość euklidesową do każdego obiektu z danych:

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

## 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:



Obraz zawierający stół

Opis wygenerowany automatycznie

# Badanie docelowe

## Utworzenie sztucznych danych badanych

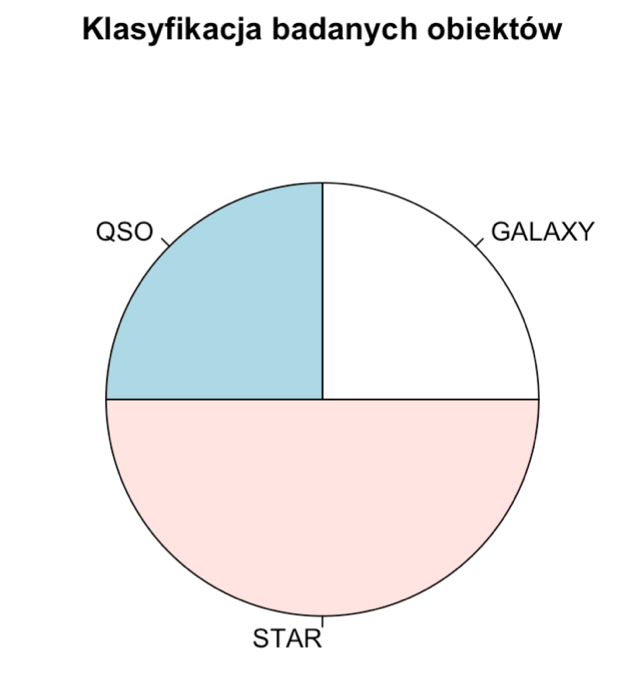
Poniżej prezentujemy stworzone przez nas dane badane, które stanowią część testową i składają się ze zmiennych numerycznych.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

## 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.

# 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:

Obraz zawierający tekst

Opis wygenerowany automatycznie

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.

# 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")