Szymon Cimochowski
Informatyka II st.
238069

Inteligencja obliczeniowa
Zgłębianie danych

Przedmiotem badania jest analiza danych dotyczących spożywania marihuany przez wyselekcjonowaną grupę ludzi poddanych kontroli. Naszym celem jest sprawdzenie, jakie aspekty kryją się za tym, czy palono ww. rodzaj narkotyku w ostatnim roku, czy też nie.

1. Baza danych

Bazą, która posłuży do zgłębiania danych jest zestawienie spożycia różnych rodzajów narkotyków przez ludzi (Drug Consumption – https://archive.ics.uci.edu/ml/datasets/Drug+consumption+(quantified)). Została ona odpowiednio spreparowana, aby spełniała założenia naszego badania. Składa się ona z kilkunastu kolumn.

Nazwa kolumny	Opis
Age	Wiek osoby badanej wyrażony w przedziale • 18-24 • 25-34 • 35-44 • 45+
Gender	Płeć (Female, Male)
Education	Rodzaj wykształcenia • Left school • Student • Pro certificate/diploma • University degree • Masters degree • Doctorate degree
Country	Kraj, w którym obecnie zamieszkuje • Australia • Canada • UK • USA • Other
Nscore	Wynik testu cechy charakteru - Neurotyk (jest melancholikiem i cholerykiem) (12 – min, 60 – max)
Escore	Ekstrawertyk (choleryk i sangwinik) (16 – min, 59 – max)
Oscore	Otwarty na doświadczenia (24 – min, 60 – max)
Ascore	Ugodowy (12 – min, 60 – max)
Cscore	Sumienny (12 – min, 60 – max)
Impulsive	Impulsywny (w skali: 1 – 10)
SS	Czy ma doznania wzrokowe? (w skali: 1 – 11)
SmokeYearAgo	Czy badany palił marihuanę rok temu, czy nie? (yes/no)

Postanowiliśmy zrezygnować z kolumny Ethnicity, gdyż zróżnicowanie badanych pod względem koloru skóry było nieznaczne (92 % badanych zostało ustalonych jako ludzi o kolorze skóry białym, podczas gdy pozostałe mniejszości etniczne nie stanowiły więcej, niż 1 %).

Usunięto również pozostałe kolumny klasowe, aby skupić się wyłącznie na problemie palenia marihuany przez ludzi.

2. Przetwarzanie bazy danych

Pierwszym zadaniem było przerobienie klasy SmokeYearAgo, aby spełniała założenia zadania dotyczące klasy. Należało zmienić wartości factorów, które zawierała dana kolumna na dwie, które odpowiadałyby na pytanie, czy osoba paliła marihuanę rok temu, korzystając z dokumentacji zawartej przy bazie danych.

```
cannabis$SmokeYearAgo = ifelse(cannabis$SmokeYearAgo == "CL0" |
                               cannabis$SmokeYearAgo == "CL1" |
                               cannabis$SmokeYearAgo == "CL2",
                               "no", "yes")
```

Kolejnym krokiem było spreparowanie danych, aby spełniały założenia przy klasyfikatorze kNN. W związku z tym utworzyłem dwie tabele. Jedna będzie wykorzystana przy klasyfikatorach, druga natomiast posłuży do metod asocjacyjnych (którą przytoczę w późniejszej części sprawozdania).

```
> str(cannabis)
'data.frame':   1885 obs. of  12 variables:
 $ Age         : num  2 1 2 0 2 3 3 2 2 3 ...
 $ Gender      : num  1 0 0 1 1 1 0 0 1 0 ...
 $ Education   : num  2 5 2 4 5 0 4 0 2 4 ...
 $ Country     : num  2 2 2 2 2 1 3 2 1 2 ...
 $ Nscore      : num  39 29 31 34 43 29 31 24 42 33 ...
 $ Escore      : num  36 52 45 34 28 38 32 52 55 40 ...
 $ Oscore      : num  42 55 40 46 43 35 43 40 39 36 ...
 $ Ascore      : num  37 48 32 47 41 55 41 41 48 47 ...
 $ Cscore      : num  42 41 34 46 50 52 48 52 49 43 ...
 $ Impulsive   : num  4 3 2 2 4 2 4 5 2 2 ...
 $ SS          : num  3 6 8 3 6 2 7 5 2 4 ...
 $ SmokeYearAgo: Factor w/ 2 levels "no","yes": 1 2 2 1 2 1 1 1 1 1 ...
```

Oczywistym jest, że przy pierwszej tabeli każda wartość musi być numeryczna. Jedynym factorem jest kolumna „SmokeYearAgo”.
Baza danych nie zawierała pustych wartości, więc pod tym względem można było zachować spokój.
Ze względu na to, iż wszystkie wartości zawarte w „surowej” bazie danych zawierały liczby rzeczywiste, oczywistym zabiegiem była zamiana wartości rzeczywistych na wartości naturalne w celu analizy danych. Poniżej przykład jednej ze spreparowanych kolumn.

```
nscoreResults <- c(-3.46436,-3.15735,-2.75696,-2.52197,-2.42317,
                   -2.34360,-2.21844,-2.05048,-1.86962,-1.69163,
                   -1.55078,-1.43907,-1.32828,-1.19430,-1.05308,
                   -0.92104,-0.79151,-0.67825,-0.58016,-0.46725,
                   -0.34799,-0.24649,-0.14882,-0.05188,0.04257,
                   0.13606,0.22393,0.31287,0.41667,0.52135,0.62967,
                   0.73545,0.82562,0.91093,1.02119,1.13281,1.23461,
                   1.37297,1.49158,1.60383,1.72012,1.83990,1.98437,
                   2.12700,2.28554,2.46262,2.61139,2.82196,3.27393)

for(x in 1:1885) {
  for(y in 1:49) {
    if (cannabis$Nscore[x] == nscoreResults[y])
      cannabis$Nscore[x] <- (11 + y)
  }
}
```

3. Klasyfikacja danych

Podział bazy na zbiór treningowy i testowy.

Poniżej załączony jest fragment kodu dotyczący podziału bazy danych na zbiór treningowy i testowy w stosunku 70 % do 30 %.

```
set.seed(2137)
ind <- sample(2, nrow(cannabis), replace=TRUE, prob=c(0.7, 0.3))
cannabis.training <- cannabis[ind==1,]
cannabis.test <- cannabis[ind==2,]
realAnswers <- cannabis.test[,12]
```

C4.5 / ID3 Drzewo

Kod odpowiedzialny za stworzenie struktury drzewa

```
ctree <- ctree(SmokeYearAgo ~ ., data = cannabis.training)

ctree.predicted <- predict(ctree, cannabis.test[,1:11])
ctree.confMat <- table(ctree.predicted, realAnswers)[2:1, 2:1]
ctree.accuracy <- mean(realAnswers == ctree.predicted)
ctree.tpr <- calc_tpr(ctree.confMat)
ctree.fpr <- calc_fpr(ctree.confMat)

plot(ctree, type = "simple")
```

Ze względu na rozległą wielkość drzewa, ilustracja zostanie ujęta na załączonym do sprawozdania obrazku o nazwie „drzewo.png”.

Wyniki

```
               realAnswers
ctree.predicted yes  no
            yes 237  69
            no   54 201
```

Dokładność wyniosła: 78%.

Naive Bayes

Kod odpowiedzialny za uruchomienie klasyfikatora NaiveBayes:

```
naive.model <- naiveBayes(SmokeYearAgo ~ ., data = cannabis.training)

naive.predicted <- predict(naive.model, cannabis.test[,1:11])
naive.confMat <- table(naive.predicted, realAnswers)[2:1, 2:1]
naive.accuracy <- mean(realAnswers == naive.predicted)
naive.tpr <- calc_tpr(naive.confMat)
naive.fpr <- calc_fpr(naive.confMat)
```

Wyniki

```
                realAnswers
naive.predicted yes  no
            yes 225  47
            no   66 223
```

Dokładność wyniosła: 79,9%.

kNN

Zanim przejdziemy do klasyfikatora kNN, należy uprzednio znormalizować bazę danych, której będziemy używać. Do pomocy posłuży nam funkcja do normalizacji danych.

```
normalize <- function(vec) {
  (vec - min(vec)) / ((max(vec)) - min(vec))
}
```

Za to sam proces normalizacji bazy wygląda następująco:

```
cannabis.norm <- normalize(cannabis[1:11])
cannabis.norm <- cbind(cannabis.norm, cannabis[12])
cannabis.norm.training <- cannabis.norm[ind==1,]
cannabis.norm.test <- cannabis.norm[ind==2,]
```

Poniżej sam kod, który obsługuje klasyfikator kNN:

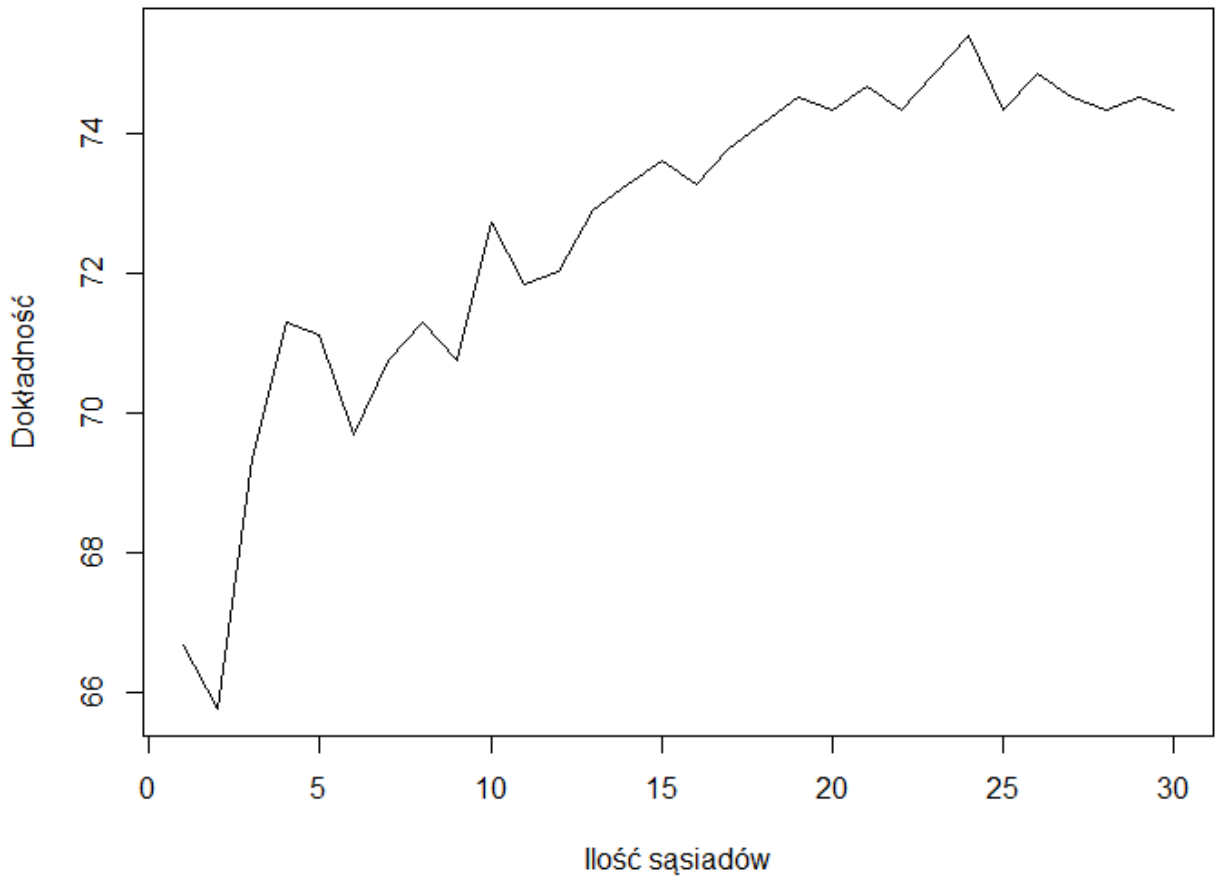
```
knn.model <- knn(cannabis.norm.training[,1:11],
                 cannabis.norm.test[,1:11],
                 cl = cannabis.norm.training[,12], k = 24,
                 prob = FALSE)
knn.predicted <- knn.model
knn.confMat <- table(knn.predicted, realAnswers)[2:1, 2:1]
knn.accuracy <- mean(realAnswers == knn.predicted)
knn.tpr <- calc_tpr(knn.confMat)
knn.fpr <- calc_fpr(knn.confMat)
```

Wyniki:

```
            realAnswers
knn.predicted yes  no
          yes 221  70
          no   70 200
```

Dokładność wyniosła: 75,04%.

Dlaczego wybrano jako k=24? Przeprowadziliśmy badania w związku z najlepszym wynikiem kNN poprzez sprawdzanie, przy jakiej wartości będzie najbardziej satysfakcjonujący wynik. Ze względu na dużą ilośc danych, z których trudno gołym okiem jest cokolwiek wywnioskować, sam klasyfikator przy jeszcze większej liczbie k – sąsiadów najprawdopodobniej osiągnąłby jeszcze lepsze wyniki. My jednak na rzecz badań ograniczyliśmy zakres do 30.



Random Forest

Kod odpowiedzialny za uruchomienie klasyfikatora Random Forest:

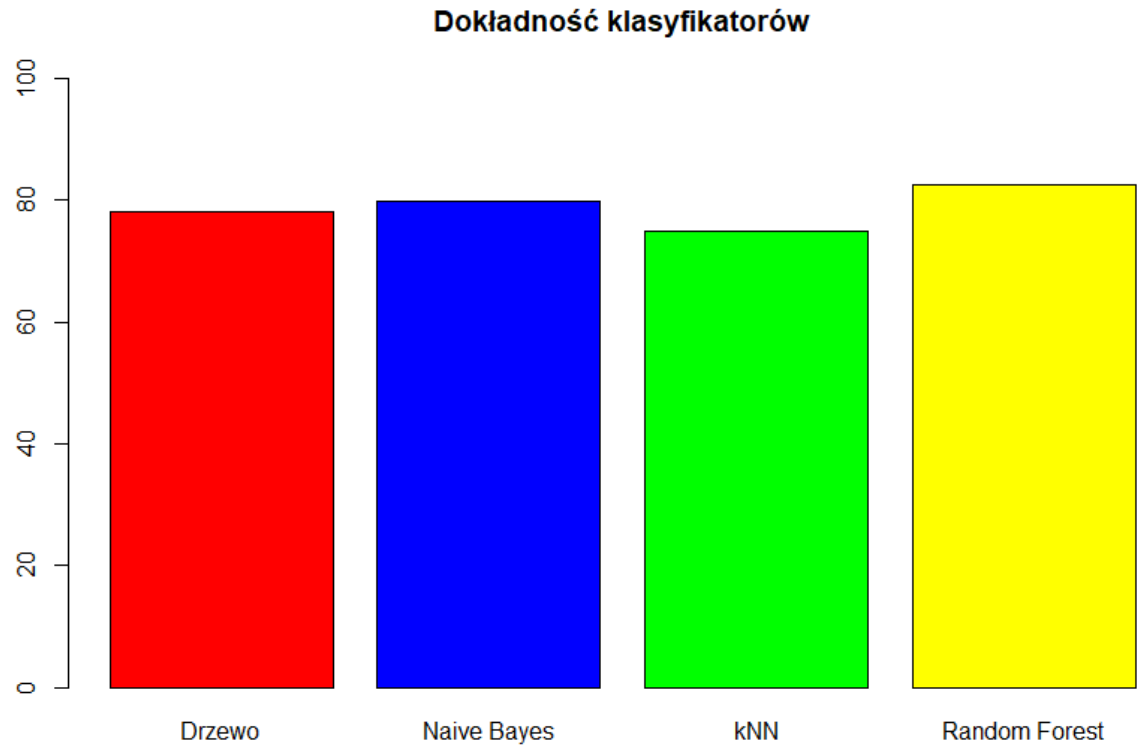
```
rf.model <- randomForest(SmokeYearAgo ~ ., data = cannabis.training,
                         ntree=100, proximity=T)
rf.predicted <- predict(rf.model, cannabis.test[,1:11])
rf.confMat <- table(rf.predicted, realAnswers)[2:1, 2:1]
rf.accuracy <- mean(realAnswers == rf.predicted)
rf.tpr <- calc_tpr(rf.confMat)
rf.fpr <- calc_fpr(rf.confMat)
```

Wyniki:

```
             realAnswers
rf.predicted yes  no
         yes 239  46
         no   52 224
```

Dokładność wyniosła: 82,53%.

Porównanie klasyfikatorów przedstawionych w sprawozdaniu



Największą dokładnością wykazał się Random Forest, dla którego standardową ilosc drzew ustawiono wartość 100. Najgorszym klasyfikatorem jest kNN, lecz jak wymieniliśmy wyżej, w przypadku uwzględnienia większej ilości k-sąsiadów, osiągnąłby prawdopodobnie lepsze wyniki.

Głównym problemem przy klasyfikatorach były wartości odpowiadające za cechy charakteru, które nie do końca mogą stwierdzić, czy osoba paliła marihuanę w przeciągu ostatniego roku, czy też nie.

Ewaluacja klasyfikatorów

Oznaczenie wartości TP, FP, TN, FN.

TP jest człowiekiem, który rzeczywiście palił czystą marihuanę w przeciągu roku, **FP** jest człowiekiem u którego stwierdzono zażywanie marihuany w przeciągu roku, lecz w rzeczywistości nie palił marihuany (był biernym palaczem), **TN** jest człowiekiem, który rzeczywiście nie palił marihuany w ciągu ostatniego roku, a **FN** jest człowiekiem, u którego stwierdzono, że nie palił marihuany zeszłego roku, lecz w rzeczywistości palił (możliwa jest substancja, która zamaskowała obecność marihuany we krwi badanego).

Obliczanie TPR (recall, sensitivity) i FPR (fall-out, false alarm).

W celu obliczenia TPR jak i FPR, napisałem funkcję obliczającą ww. wartości, biorąc pod uwagę tabelki z wartościami TP, FP, TN i FN. Wraz z funkcjami, dołączona jest legenda.

```r
calc_tpr <- function(t){
  tpr = t[1]/(t[1]+t[2])
  return(tpr)
}

calc_fpr <- function(t) {
  fpr = t[3]/(t[3]+t[4])
  return(fpr)
}

# O.B. LEGENDA
# [1] - TP
# [2] - TN
# [3] - FP
# [4] - FN
```

```
Klasyfikator       FPR       TPR
      Drzewo 0.2555556 0.8144330
 Naive Bayes 0.1740741 0.7731959
         kNN 0.2592593 0.7594502
Random Forest 0.1703704 0.8213058
```

Zależności FNR i TNR od TPR i FPR

FNR = 1 – TPR
TNR = 1 – FPR

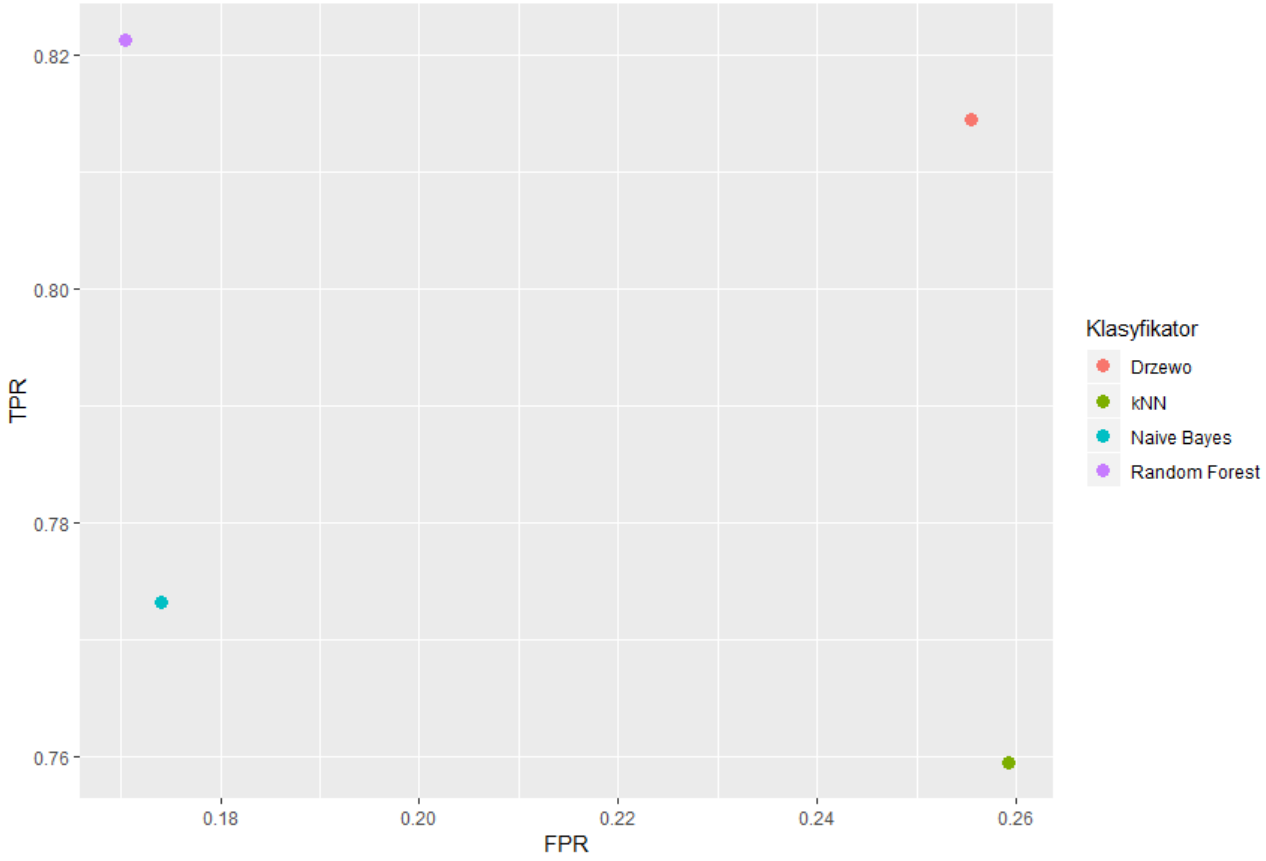
Określenie błędów pierwszego i drugiego rodzaju i powiązanie z TPR, FPR, TNR i FNR.

Błąd piewszego rodzaju to ludzie palący marihuanę w przeciągu ostatniego roku, u których błędnie zdiagnozowano, że nie palili danego narkotyku. Błąd drugiego rodzaju, to ludzie niepalący, którzy zostali zdiagnozowani jako palacze w ciągu ostatniego roku. Im więcej wyników pierwszego rodzaju, tym mniejsza będzie wartość TNR, a większa FPR. Im więcej błędów drugiego rodzaju, tym mniejsza będzie wartość TPR, a większa FNR.

Który z błędów w bazie jest gorszy do popełnienia? Pierwszego, czy drugiego rodzaju?

Wg naszych badań, gorszym do popełnienia jest błąd pierwszego rodzaju. W przypadku, gdy nie można wykryć substancji toksycznych, które są zawarte wraz z marihuaną, nie będzie można przeprowadzić dodatkowych badań. Tutaj akurat naszym celem było porównanie cech charakteru do spożywania marihuany, czyli skutki uboczne zażywania i sprawdzenie skali problemu, więc przy niemożliwości wykrycia marihuany, nie jesteśmy w stanie stwierdzić, czy narkotyk rzeczywiście brał czynny udział w zmianie charakteru osoby badanej.

Obliczenie pary (FPR, TPR) i zaznaczenie punktów na wykresie.



Im wyżej znajduje się klasyfikator, tym lepiej. Bierzemy również pod uwagę najmniejszą wartość FPR (związaną z błędami pierwszego rzędu). W związku z tym, najlepszym kandydatem okazuje się być Random Forest, który wcześniej wykazał się największą dokładnością co do rezultatów. Moglibyśmy również brać pod uwagę klasyfikator drzewa, lecz zawiera duży współczynnik FPR, który szybko eliminuje go z walki o najlepszy klasyfikator.

4. Grupowanie metodą k – średnich.

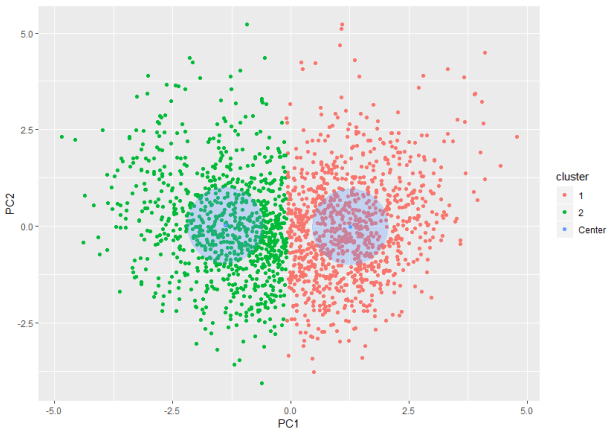
Poniżej załączam kod odpowiedzialny za wcześniejsze spreparowanie danych do uruchomienia algorytmu i utworzenie wykresów.

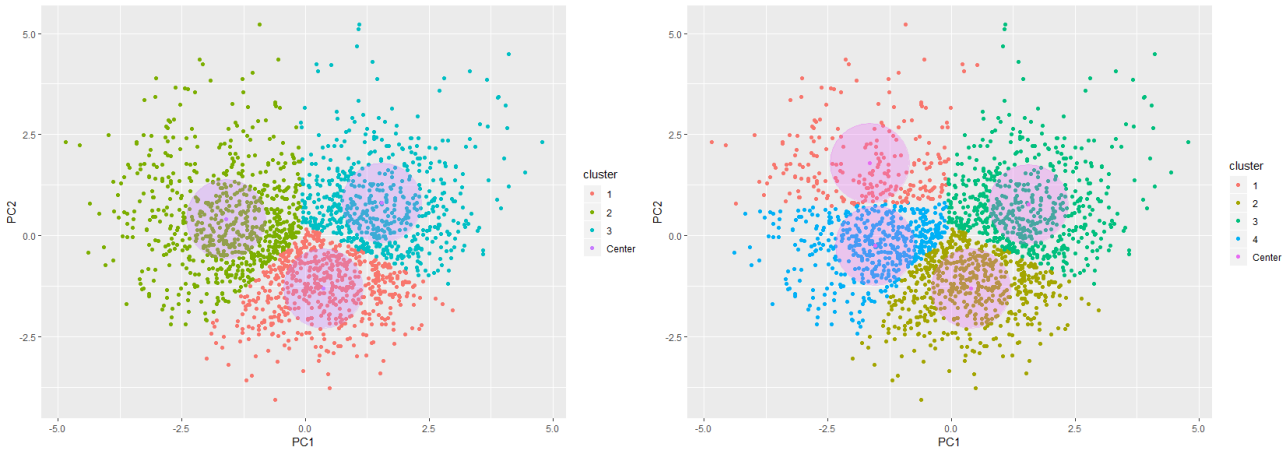
```
cannabis.log <- log(cannabis[,c(1:11)])

replace_faults <- function(xd) {
  xd[is.infinite(xd)] <- NA
  replace_value <- mean(xd, na.rm = TRUE)
  xd[is.na(xd)] <- replace_value
}

replace_means <- sapply(cannabis.log, replace_faults)
replace_means <- as.data.frame(replace_means)

cannabis.log$Age[is.infinite(cannabis.log$Age)] <- replace_means["Age", ]
cannabis.log$Education[is.infinite(cannabis.log$Education)] <- replace_means["Education", ]
cannabis.log$Country[is.infinite(cannabis.log$Country)] <- replace_means["Country", ]
```

```
cannabis.log$Gender = cannabis$Gender

cannabis.stand <- scale(cannabis.log, center = TRUE)
cannabis.pca <- prcomp(cannabis.stand)
cannabis.final <- predict(cannabis.pca)[,1:2]
cannabis.test.final <- as.data.frame(cannabis.final)

k <- kmeans(cannabis.test.final, 4)
cannabis.test.final$cluster = factor(k$cluster)
centers = as.data.frame(k$centers)
```

```
k.predicted_plot <- ggplot(cannabis.test.final, aes(x=PC1, y=PC2, color=cluster)) +
  geom_point() +
  geom_point(data=centers, aes(x=PC1,y=PC2, color='Center')) +
  geom_point(data=centers, aes(x=PC1,y=PC2, color='Center'), size=36, alpha=.3, show.legend=FALSE)

k.real_plot <- ggplot(cannabis.test.final, aes(x=PC1, y=PC2, color=cannabis$SmokeYearAgo)) +
  geom_point()
```

Poniżej przedstawiamy również podziały na 2, 3 i 4 klastry w porównaniu do prawdziwych wyników.





W porównaniu z rzeczywistymi wynikami można zaobserwować, że wyniki pokrywają się w mniej/więcej 75 %. Oprócz tego, w miarę prawidłowy sposób algorytm oddzielił klastry od siebie. Pozwoliliśmy sobie również zaznaczyć centra klastrów. Algorytm mimo skupienia danych dał radę podzielić na 3 i 4 klastry. Prawdopodobnie odpowiada to czasom, w jakim ostatnio palono marihuanę z początkowej bazy danych.

5. Reguły asocjacyjne

Na początek fragmenty kodu odpowiadające za spreparowanie danych do uruchomienia algorytmu apriori (tutaj posłużymy się drugą bazą danych, którą przygotowaliśmy specjalnie na tą okazję).

```
cannabis.rules <- cannabis.raw

cannabis.rules$Age = as.factor(cannabis.rules$Age)
cannabis.rules$Gender = as.factor(cannabis.rules$Gender)
cannabis.rules$Education = as.factor(cannabis.rules$Education)
cannabis.rules$Country = as.factor(cannabis.rules$Country)
cannabis.rules$Nscore = as.factor(cannabis.rules$Nscore)
cannabis.rules$Escore = as.factor(cannabis.rules$Escore)
cannabis.rules$Oscore = as.factor(cannabis.rules$Oscore)
cannabis.rules$Ascore = as.factor(cannabis.rules$Ascore)
cannabis.rules$Cscore = as.factor(cannabis.rules$Cscore)
cannabis.rules$Impulsive = as.factor(cannabis.rules$Impulsive)
cannabis.rules$SS = as.factor(cannabis.rules$SS)
cannabis.rules$SmokeYearAgo = as.factor(cannabis.rules$SmokeYearAgo)

cannabis.rules.disc <- discretizeDF(cannabis.rules)

cannabis.rules.disc <- as(cannabis.rules.disc, 'transactions')

rules <- apriori(cannabis.rules.disc,
                 parameter = list(minlen=2, supp=0.1, conf=0.8),
                 appearance = list(rhs=c("SmokeYearAgo=no", "SmokeYearAgo=yes"), default="lhs"),
                 control = list(verbose=F))
```

```
rules.sorted <- sort(rules, by="lift")
subset.matrix <- is.subset(rules.sorted, rules.sorted)
subset.matrix[lower.tri(subset.matrix, diag=T)] <- FALSE
redundant <- colSums(subset.matrix, na.rm=T) >= 1
rules.pruned <- rules.sorted[!redundant]

print(inspect(head(rules.pruned, by = "lift")))
```

```
    lhs                                rhs                    support confidence     lift count
[1] {Age=45+,
     Country=UK}                    => {SmokeYearAgo=no}  0.1389920  0.8646865 1.839655   262
[2] {Age=18-24,
     Gender=Male,
     Country=USA}                   => {SmokeYearAgo=yes} 0.1151194  0.9730942 1.836119   217
[3] {Age=18-24,
     Education=Student,
     Country=USA}                   => {SmokeYearAgo=yes} 0.1129973  0.9638009 1.818583   213
[4] {Age=18-24,
     Country=USA}                   => {SmokeYearAgo=yes} 0.1660477  0.9630769 1.817217   313
[5] {Age=18-24,
     Gender=Male,
     Education=Student}             => {SmokeYearAgo=yes} 0.1183024  0.9529915 1.798187   223
[6] {Age=18-24,
     Gender=Male}                   => {SmokeYearAgo=yes} 0.2037135  0.9458128 1.784642   384
```

Z przedstawionej wyżej tabelki jesteśmy w stanie zaobserwować, że w głównej mierze palaczami marihuany są osoby młode w przedziale wiekowym od 18 do 24 lat, mężczyźni, studenci, mieszkający w Stanach Zjednoczonych. W przypadku osób starszych, wyraźnie zaobserować można, iż osoby powyżej 45 roku życia nie palili marihuany w ciągu ostatniego roku.

6. Dodatek

Jako dodatkową obserwację chcielibyśmy zobaczyć, jak wygląda na diagramach kołowych sytuacja z paleniem marihuany w ciągu roku w UK i USA.