13.12.2018r

Zadanie domowe nr 2
Zgłębianie danych

Dawid Ćwik
Informatyka II
gr. 1
238137

1. Wstęp

Wybrana baza danych https://www.kaggle.com/sammy123/lower-back-pain-symptoms-dataset

Baza zawiera **310 obserwacji** pogrupowanych w **13 kolumnach** oraz **jedną kolumnę klasową**. Tematem jest ból pleców w dolnym odcinku. Liczby przedstawiają kąty pomiędzy odcinkami kręgosłupa oraz ustawienia kości połączonych z kręgosłupem jak np. miednica. **Baza ma na celu zidentyfikowanie złego ułożenia kręgosłupa**.

NAZWY KOLUMN	
ENG	PL
pelvic_incidence	Padanie miednicy
pelvic tilt	Pochylenie miednicy
lumbar_lordosis_angle	Lordoza lędźwiowa
scral_slope	Nachylenie sakralne
pelvic_radius	Kąt miednicy
degree_spondylolisthesis	Stopień zwyrodnienia
pelvic_slope	Miednicze zbocze
Direct_tilt	Kierunkowe pochylenie
thoracic_slope	Piersiowe nachylenie
cervical_tilt	Szyjne pochylenie
sacrum_angle	Kąt kości krzyżowej
scoliosis_slope	Nachylenie boczne
Class_att	Kolumna klasowa

Ostatnia kolumna zawiera wartosci abnormal/normal, które pełnią funkcje kolumny klasowej, na podstawie której będzie wnioskowany wynik badań

2. Przetwarzanie / obróbka / łączenie / dzielenie baz danych

Baza danych nie zawiera określonych nagłówków kolumn, jedynie ich opisy. Przy ilości kilkunastu kolumn, ciężkie jest operowanie na niej posiadając nagłówki „COL_1", „COL_2". Nadałem danym nagłówki opisów w wersji polskiej, co również było sporym wyzwaniem przy specjalistycznych określeniach.

W tabeli kolumna 14 zawierała uwagi do obserwacji i wnioski autora. Nie jest ona pomocna przy pracy dlatego należało ją usunąć.

Tabela klasowa zawierała ocenę w języku angielskim abnormal/normal, co nie dawało latwości pracy z danymi, jako iż jest to napis. Przyjąłem więc jako dane w kolumnie znaczącej **1 – wada kręgosłupa, 0 – brak wady kręgosłupa**

Col12	Class_att	X14
43.5123	Abnormal	*NA*
16.1102	Abnormal	*NA*
19.2221	Abnormal	Prediction is done by using binary classification.
18.8329	Abnormal	*NA*
24.9171	Abnormal	*NA*
9.6548	Abnormal	Attribute1 = pelvic_incidence (numeric)
25.9477	Abnormal	Attribute2 = pelvic_tilt (numeric)
26.3543	Abnormal	Attribute3 = lumbar_lordosis_angle (numeric)
40.0276	Abnormal	Attribute4 = sacral_slope (numeric)
21.4320	Abnormal	Attribute5 = pelvic_radius (numeric)
18.3437	Abnormal	Attribute6 = degree_spondylolisthesis (numeric)

Możliwość dokonywania obliczeń warunkowana jest posiadaniem kolumny zawierającej typ danych **factor**. Należało ustawić taki typ danych w kolumnie klasowej

```
for ( i in 1:nrow(data) ) {
  if (data["Ocena"][i,] == "Abnormal") data["Ocena"][i,] <- as.numeric(1)
  else if (data["Ocena"][i,] == "Normal") data["Ocena"][i,] <- as.numeric(0)
}

data$Ocena <- factor(data$Ocena)
```

3. Klasyfikatory i ich ewaluacja

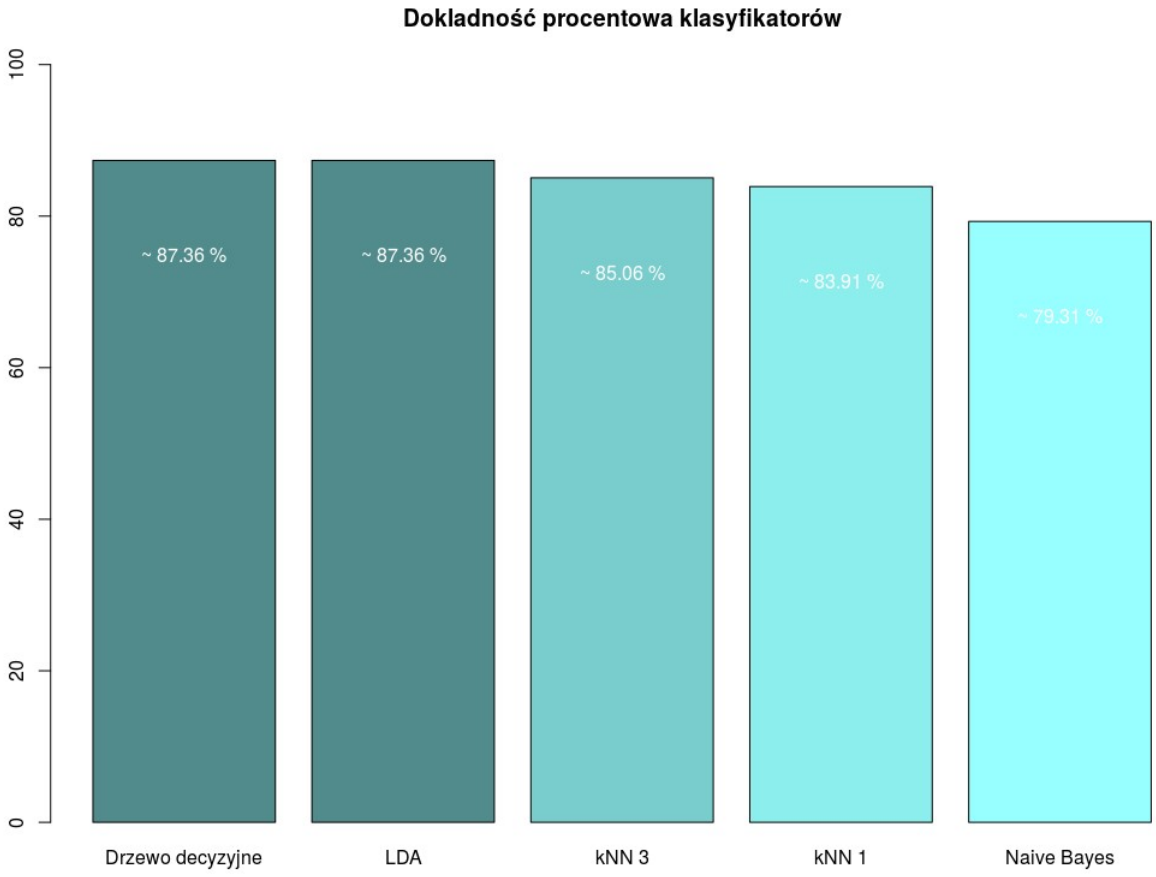
Baza danych podzielona została na zbiór testowy oraz treningowy. Dodatkowo, przed każdym badaniem klasyfikatora, kopiowane były zbiory powyższe, tak aby uniknąć ewentualnego badania zmodyfikowanej bazy. Dla klasyfikatora kNN, zgodnie z informacjami z ćwiczeń, dane zostały znormalizowane.

```
### NORMALIZACJA
normalize <- function(vector) {
  (vector - min(vector))/((max(vector))-min(vector))
}

### PRZYGOTOWANIE DANYCH TESTOWYCH ORAZ TRENINGOWYCH
set.seed(1234)
#Dane surowe
data.raw <- data
#Dane testowe
ind <- sample(2, nrow(data), replace=TRUE, prob=c(0.7, 0.3))
#zwykle
data.training <- data[ind==1,]
data.test <-data[ind==2,]
#znormalizowane
data.norm <- normalize(data[1:12])
data.norm <- cbind(data.norm, data[13])
data.norm.training <- data.norm[ind==1,]
data.norm.test <- data.norm[ind==2,]

faktyczna_ocena <- data.test[,13]
```

Badanie wykonane zostało przy użyciu 4 klasyfikatorów. Drzewem decyzyjnym, Naive Bayes, kNN oraz LDA. Ostatecznie najlepszym okazały się drzewo decyzyjne oraz LDA. Otrzymały one takie same rezultaty, około 87% dokładności.
Zastosowane zostały również 2 warianty kNN. Jeden z nich przypisuje 3 najbliższych sąsiadów, a drugi 1. Lepszy okazuje się ten który bierze pod uwagę więcej danych.
Najgorszym z nim okazał się Naive Bayes, jednak osiągnął on wynik około 79%, co moim zdaniem klasyfikuje go jako dobry algorytm.

W tabeli poniżej prezentowane są wyniki macierzy błędu, oraz graficzne zestawienie dokładności algorytmów.

KLASYFIKATOR	MACIERZ BŁĘDU
Drzewo decyzyjne	predicted_tree 0 1 0 19 1 1 10 57
LDA	predicted_lda 0 1 0 21 3 1 8 55
KNN 3	predicted_knn3 0 1 0 22 6 1 7 52
KNN 1	predicted_knn1 0 1 0 21 6 1 8 52
NaiveBayes	predicted_naive 0 1 0 25 14 1 4 44



Dzieki danym zawartym w macierzy błedu jestesmy w stanie dokładniej ewaluować klasyfikatory

TP – *prawdziwie pozytywna* – osoba z wadą poprawnie zdiagnozowana
FP – *fałszywie pozytywna* – osoba z wadą zdiagnozowana jako bez wady
TN – *prawdziwie negatywna* – osoba bez wady zdiagnozowana poprawnie
FN – *fałszywie negatywna* - osoba bez wady zdiagnozowana jako z wadą

Dzięki tym danym, możliwe było wiliczenie **TPR** i **FPR**

KLASYFIKATOR	TPR	FPR
Drzewo decyzyjne	0,66	0,02
LDA	0,72	0,06
KNN3	0,76	0,1
KNN1	0,72	0,1
Naive Bayes	0,86	0,24

Przy badaniu wyniknęła zależność pomiędzy **TPR** i **FPR** a **FNR** i **TNR**

$$TPR = TP \div (TP + FN) = 1 - FNR$$

$$FPR = FP \div (FP + TN)$$

$$TNR = TN \div (TN + FP) = 1 - FPR$$

$$FNR = FN \div (FN + TP)$$

Można wyciągnąć następujące wnioski:

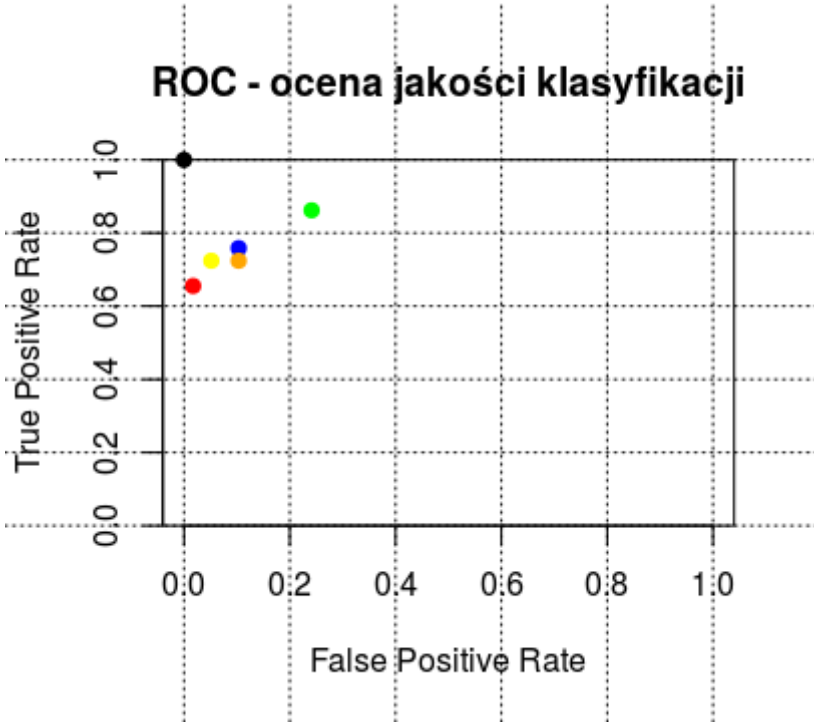
- zwiększenie **FP** zwieksza się **FPR** oraz zmniejsza **TNR**
- zwiększenie **FN** zwiększa się **FNR** oraz zmniejsza **TPR**

Błąd pierwszego rodzaju:
Błąd polegający na nieodrzuceniu hipotezy zerowej. W kontekście badanej bazy, uznanie osoby z wadą jako zdrowej (**FP**)

Błąd drugiego rodzaju:
Błąd polegający na odrzuceniu hipotezy zerowej. W kontekście badanej bazy, uznanie osoby bez wady jako chora (**FN**)

Istotność błędu:
Waga popełnienia błedu jest ciężka do określenia z powodu wpływu na zdrowie pacjenta. Wydawać się może jednak, że **niezdiagnozowana choroba jest mniej inwazyjna dla ciała człowieka niż zdiagnozowanie choroby, kóra nie istniej.** Ewentualne przyszłe lecznie może mieć negatywny wpływ na zdrowie a w konsekwencji osoba zdrowa może stać się chora.

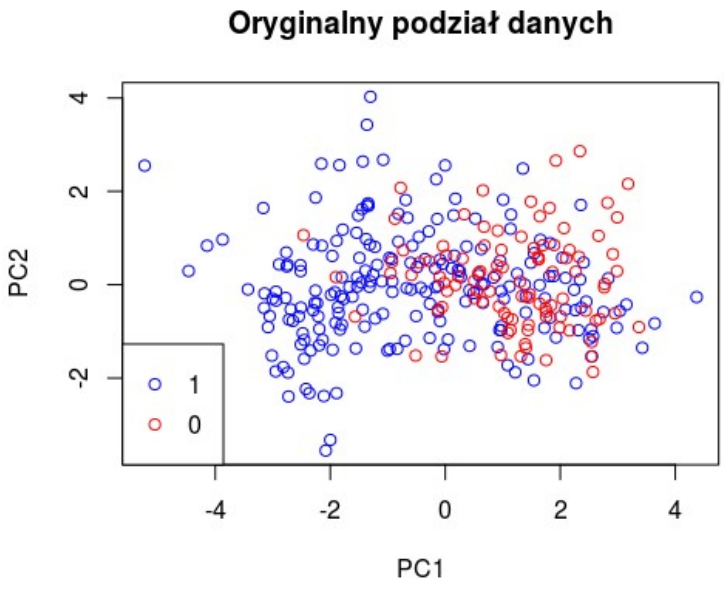
Drzewo
LDA
kNN3
kNN1
Naive
Idealny



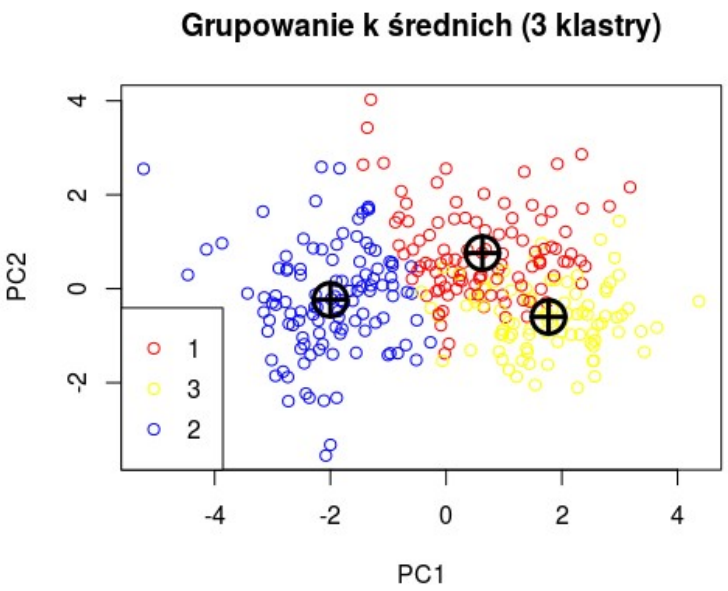
Idealny klasyfikator ma miejsce w punkcje (1,0). Wtedy TPR ma najwiekszy możliwy wymiar a FPR najmniejszy.
Wyżej powiedzieliśmy, że unikać chcemy przypadku zdiagnozowania osoby jako chorej w momencie gdy jest zdrowa, czyli błąd drugiego rodzaju. Im wyżej znajduje się znacznik, tym mniej błędów drugiego stopnia popełnił. W naszym przypadku jest to **Naive Bayes**
Jeżeli interesuje nas jak najmniejsza ilość popełnianych błędów pierwszego stopnia to im bardziej na lewo tym lepszy wynik. Tym razem okazuje się **Drzewo decyzyjne**.
Biorąc pod uwagę te 2 argumenty, wynik satysfakcjonujący powinien znajdować się jak najblizej punktu idelanego. W naszym wypadku jest to sprawa dyskusyjna jednak wyłonić możemy 2 znaczące algorytmy: **kNN3** oraz **LDA**

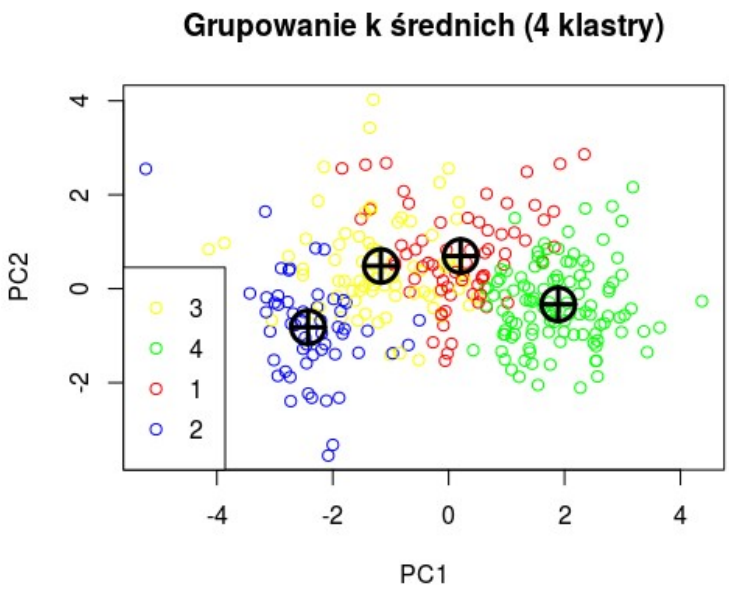
4. Grupowanie metodą k-średnich

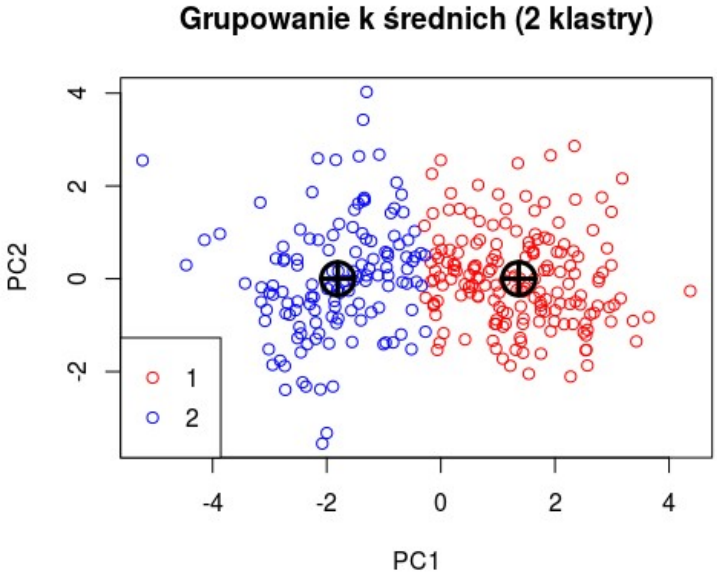
Oddzielenie klas w orginalnej bazie danych dostarczyło informacji na temat tendecyjności danych. Można zauważyć, iż te znajdujące się bliżej lewej strony, maja większą skłonność do bycia klasą oznaczoną jako chora (1). Niestety ze względu na losowość tych danych, ciężko jednoznacznie określić przestrzenie dla klas.



Korzystając z metody k-średnich dane zostały podzielone na klastry 2, 3 oraz 4 częściowe. Niestety ze wzlędu na zbyt losowe dane, ciężko określić poprawność tej metody.







Najbliżej było podzielenie na 2 klastry jednak ten sposob jest zupełnie losowy. W 4 klastrze, zauważamy jednak, że jedna z klas posiada jedynie obiekty które są chore (niebieska). Może to oznaczać, że była by szansa na bliższe określenie zbiorów. Czy ma to jednak sens, możemy również podzielić na bardzo dużą ilość klastrów i doszukiwać się pojedyńczo zależności. Korzystanie z takiej metody nie jest najlepszym pomysłem.

Poniżej przedstawiony fragement kodu odpowiedzialny za tą metodę

```
## GRUPOWANIE METODA K-SREDNICH
data.log <- log(data.norm[,1:12])
data.log <- data.log[is.finite(rowSums(data.log)),]

data.scale <- scale(data.log, center=TRUE)
data.pca <- prcomp(data.scale)
data.kmeans <- predict(data.pca)

km_2 <- kmeans(data.kmeans, 2, iter.max = 100, algorithm = c("Lloyd"), trace=FALSE)
km_3 <- kmeans(data.kmeans, 3, iter.max = 100, algorithm = c("Lloyd"), trace=FALSE)
km_4 <- kmeans(data.kmeans, 4, iter.max = 100, algorithm = c("Lloyd"), trace=FALSE)
```

5. Reguły asocjacyjne

Z powodu na brak danych słownych w bazie, a jedynie liczby, kluczowym była modyfikacja danych w kolumnach. Ze względu na brak wiedzy w zakresie poprawnego ułożenia kości względem siebie, nie podejmowałem się określenia czy dana osoba ma odchylenie od normy światowej i przypisaniu jej tej informacji. Postanowiłem policzyć średnią dla kolumny i na podstawie tego przypisać rekordom wynik: poniżej/powyżej średniej.

```
#Liczenie średnich dla kolumn
ruleData <- data[,1:13]

Padanie_miednicy.avg <- mean(ruleData$Padanie_miednicy)
Pochylenie_miednicy.avg <- mean(ruleData$Pochylenie_miednicy)
Lordoza_lędźwiowa.avg <- mean(ruleData$Lordoza_lędźwiowa)
Nachylenie_kości_krzyżowej.avg <- mean(ruleData$Nachylenie_kości_krzyżowej)
Kąt_ułożenia_miednicy.avg <- mean(ruleData$Kąt_ułożenia_miednicy)
Stopień_zwyrodnienia.avg <- mean(ruleData$Stopień_zwyrodnienia)
Zbocze_miednicy.avg <- mean(ruleData$Zbocze_miednicy)
Kierunkowe_pochylenie.avg <- mean(ruleData$Kierunkowe_pochylenie)
Nachylenie_piersiowego.avg <- mean(ruleData$Nachylenie_piersiowego)
Pochylenie_szyjnego.avg <- mean(ruleData$Pochylenie_szyjnego)
Kąt_kości_krzyżowej.avg <- mean(ruleData$Kąt_kości_krzyżowej)
Nachylenie_boczne.avg <- mean(ruleData$Nachylenie_boczne)

avg_names <-c("Padanie_miednicy", "Pochylenie_miednicy", "Lordoza_lędźwiowa", "Nachylenie_kości_krzyżowej",
              "Kąt_ułożenia_miednicy", "Stopień_zwyrodnienia", "Zbocze_miednicy", "Kierunkowe_pochylenie",
              "Nachylenie_piersiowego", "Pochylenie_szyjnego", "Kąt_kości_krzyżowej", "Nachylenie_boczne")
avg_values <- c(Padanie_miednicy.avg, Pochylenie_miednicy.avg, Lordoza_lędźwiowa.avg,
                Nachylenie_kości_krzyżowej.avg, Kąt_ułożenia_miednicy.avg, Stopień_zwyrodnienia.avg,
                Zbocze_miednicy.avg, Kierunkowe_pochylenie.avg, Nachylenie_piersiowego.avg,
                Pochylenie_szyjnego.avg, Kąt_kości_krzyżowej.avg, Nachylenie_boczne.avg)


for (column in 1:12) {
  for (row in 1:nrow(ruleData)) {
    if(ruleData[row,][column] > avg_values[column]) {
      ruleData[row,][column] <- "Powyżej średniej"
    }
    if(ruleData[row,][column] <= avg_values[column]) {
      ruleData[row,][column] <- "Poniżej średniej"
    }
  }
}
```

Następnie należało te dane przekonwertować na „factor”. W przypadku tej metody skorzystałem z metody apriori().

```
rules <- apriori(ruleData ,parameter = list(minlen=2, supp=0.005, conf=0.8),
                 appearance = list(rhs=c("Ocena=1", "Ocena=0"),default="lhs"),
                 control = list(verbose=F))
```

Kolumn z wariantami było sporo, dlatego ostatecznie funkcja stworzyła ok. 75 000 reguł. Poniżej przedstawiam jedynie wybrane z nich.

1. Gdy osoba ma pochyloną miednicę poniżej średniej (17,5), lordoza lędźwiowa poniżej średniej (51,9), pochyla się do tyłu, a kąt padania kości krzyżowej jest mniejszy niż -14.05 osoba ma nie ma wady kręgosłupa

2. Gdy osoba ma mniejszą lordoze (względem pozostałych badanych), jej odcinek szyjny jest powyżej średniej (11,9), a nachylenie odcinka piersiowego jest mniejsze niż pozostałych badanych (13.06), osoba oznaczona została jako zdrowa

Jeżeli jednak chodzi o zdiagnozowanie wad, algorym wyszukał się następujących:

1. Lordoza lędźwiowa powyżej sredniej (pogłebiona)
2. Pochylenie szyjnego odcinka poniżej średniej oraz kąt kości krzyżowej poniżej średniej
3. Nachylenie odcinka piersiowego poniżej średniej oraz kąt kości krzyżowej poniżej średniej

7. Podsumowanie

Temat projeku i jego konstrukacja pozwalały na wykorzystanie do badań bazy, która nas interesuje. Dzięki temu praca była przyjemniejsza i jej efekty przyniosły radość.
Ciężko jednak bezpośrednio odnieść się do jego wyników. Algorytmy weryfikujące uzyskiwały bardzo dobre wyniki a reguły asocjacyjne wykazały kilka ciekawych zależności dzięki którym możemy zauważyć, że nawet już przy złym kącie odcinka lędźwiowego, możemy mieć problemy z kręgosłupem.
W pracy która wykonujemy praktycznie cały czas spędzamy siedząc i w pozycji zgarbionej. Wnioski jakie nachodzą po tego typu badaniach to na ból w plecach może wpływać masa różnych czynników. Od punktu w szyji aż po ustawienie miednicy.
Ze względu na małą ilość danych, metoda klastrowania okazała się bezużyteczna.