Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования «Кубанский государственный университет»

Кафедра вычислительных технологий

**ОТЧЕТ**

о выполнении лабораторной работы №6

по дисциплине “Обработка больших данных”

Выполнил: ст. гр. 36/1

Придава А.А.

Проверил: преподаватель

Шиян В. И.

Краснодар

2024

**Постановка задачи:**

Часть первая - (ЛР 6.1) - предназначена для разведочного анализа данных и определения количества результирующих выходных групп.

1. Выполнить дескриптивный анализ данных (здесь приветствуются дополнительные исследования).
2. Оценить оптимальное число кластеров, для этого построить диаграмму "Метод силуэта", “Метод локтя”,  "Статистику разрыва" и Алгоритм консенсуса.
3. Выполнить иерархическую кластеризацию вашего набора данных, построив **дендрограмму**. Подробно обосновать Ваш выбор числа групп.
4. Построить диаграмму со столбчатыми диаграммами (рис. 6.8) и  боксплотами групп (рис. 6.12). Провести сравнительный анализ полученных групп.
5. Выполнить кластеризацию своего датасета по  k-means (рис.6.9, 6.10).
6. Выполнить построение scatterplot  (рис. 6.13) с помощью функций plot или pairs.
7. Построить трехмерную кластеризацию по scatterplot3d (6.16)
8. В целом: выполнить шаги 1-3,5 анализа для своего набора данных (если какие-то из шагов нерелевантны вашему набору данных, объяснить почему).

Часть вторая - (ЛР 6.2) - использует предположение о количестве выходных классов Вашего набора данных, классификация строится из этого предположения – см следующую ЛР.

Данные берутся из предыдущего набора (ЛР 6 часть 1). Вам необходимо добавить найденные при в результате кластерного анализа классы (groups), как вектор-столбец в предыдущий DATASET, разделить его на две части (обучающую и тестовую) обучить на обучающей выборке классификатор, а затем применить классификатор к тестовым данным. Важно: понадобится преобразовать вектор groups в фактор.

Для наивного Байесовского классификатора:

1. Решите задачу с помощью наивного Байесовского классификатора;

2. Проанализируйте точность полученных решений для тестовых данных

Для деревьев решений:

1. Примените метод деревьев решений для задачи классификации (для того же набора данных).

2. Исследуйте дерево решений; если позволяет размерность, постройте его график.

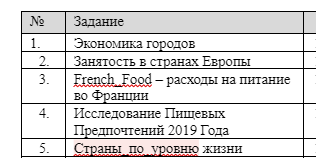
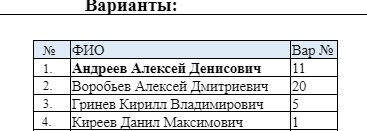
3. Проанализируйте точность полученных решений для тестовых данных (с известным значением переменной отклика), сравните результаты с ранее полученными.

4. Выполнить классификацию с помощью случайного леса, сопоставить результат с результатом дерева решения, прокомментировать результат сравнения.

5. Сопоставьте результаты с результатами Байесовского классификатора.

6. Проанализируйте полученные результаты, сделайте вывод относительно результатов кластерного анализа (ЛР 6.1), оформите общий отчёт по ЛР 6.1 и 6.2.

**Выполняемый вариант:**



Вариант – 5, набор данных – Страны по уровню жизни.

**Часть датасета:**

"страна","рождаем","смертн","деск\_см","длит\_муж","длит\_жен","доход","регион"

"Albania",24.7,5.7,30.8,69.6,75.5,600,1

"Bulgaria",12.5,11.9,14.4,68.3,74.7,2250,1

"Czechoslovakia",13.4,11.7,11.3,71.8,77.7,2980,1

"Former\_E.\_Germany",12,12.4,7.6,69.8,75.9,-9999,1

"Hungary",11.6,13.4,14.8,65.4,73.8,2780,1

"Poland",14.3,10.2,16,67.2,75.7,1690,1

"Romania",13.6,10.7,26.9,66.5,72.4,1640,1

"Yugoslavia",14,9,20.2,68.6,74.5,-9999,1

"USSR",17.7,10,23,64.6,74,2242,1

Значения 9999 и -9999 обозначают, что значение неизвестно или не определено.

**Ход работы:**

1. **Проведение дескриптивного анализа.**

Оценим описательную статистику данного датасета. По суммаризации данных были получены следующие значения:

рождаем смертн

Min. : 9.70 Min. : 2.20

1st Qu.:14.70 1st Qu.: 7.70

Median :29.00 Median : 9.50

Mean :29.46 Mean :10.73

3rd Qu.:42.55 3rd Qu.:12.30

Max. :52.20 Max. :25.00

деск\_см длит\_муж длит\_жен доход

Min. : 4.50 Min. :38.10 Min. :41.20 Min. : 80

1st Qu.: 13.05 1st Qu.:55.40 1st Qu.:56.75 1st Qu.: 475

Median : 43.00 Median :63.40 Median :67.60 Median : 1690

Mean : 55.28 Mean :61.38 Mean :66.03 Mean : 5741

3rd Qu.: 86.50 3rd Qu.:68.50 3rd Qu.:75.45 3rd Qu.: 7325

Max. :181.60 Max. :75.90 Max. :81.80 Max. :34064

По полученным значениям можно сделать вывод, что данные будет необходимо стандартизировать, так как они имеют многократное различие.

Дополнительно, я вывел график 10 лучших стран по показателю дохода.

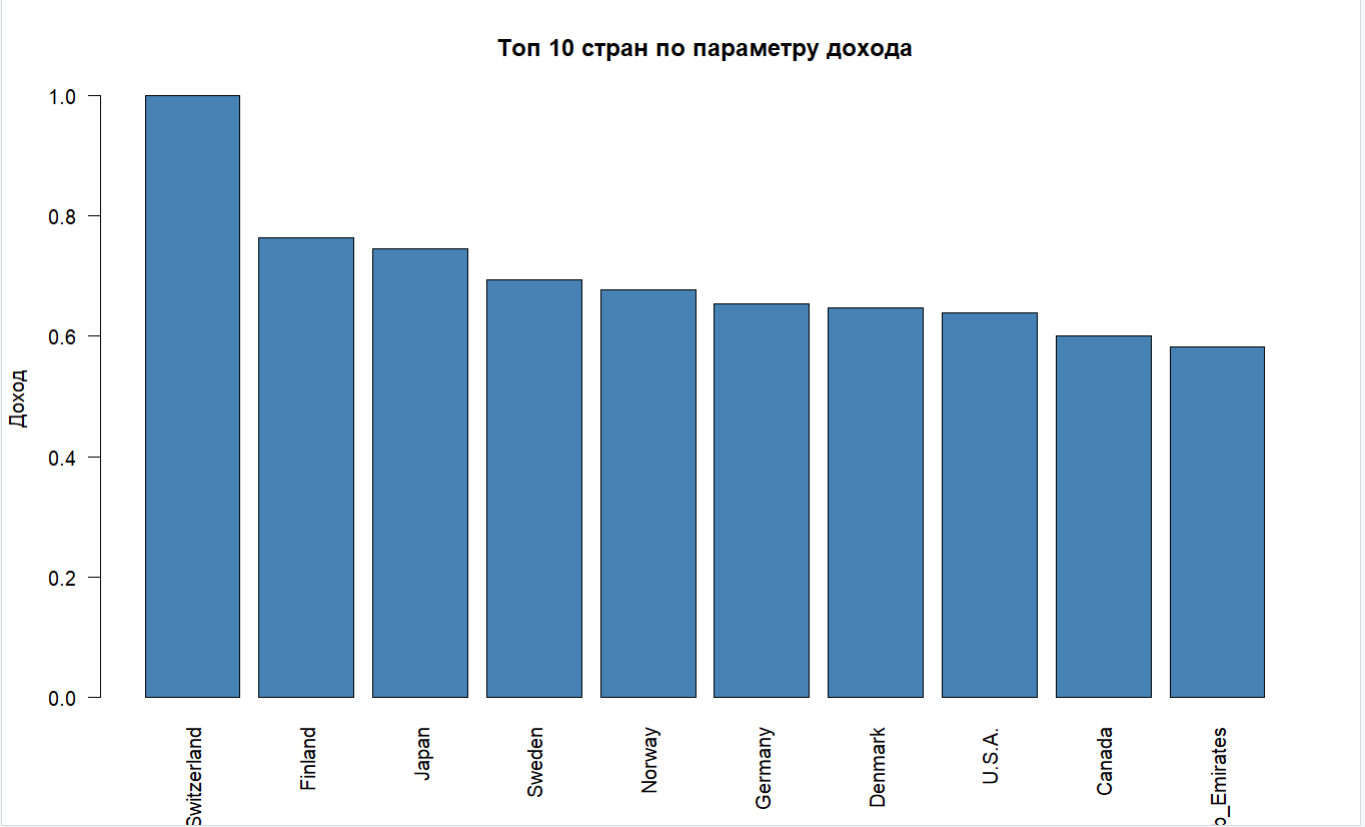


Рисунок 1 – 10 стран по параметру дохода (значения стандартизованы)

Также я дополнительно вывел график всех стран с длительностью мужской и женской жизни.

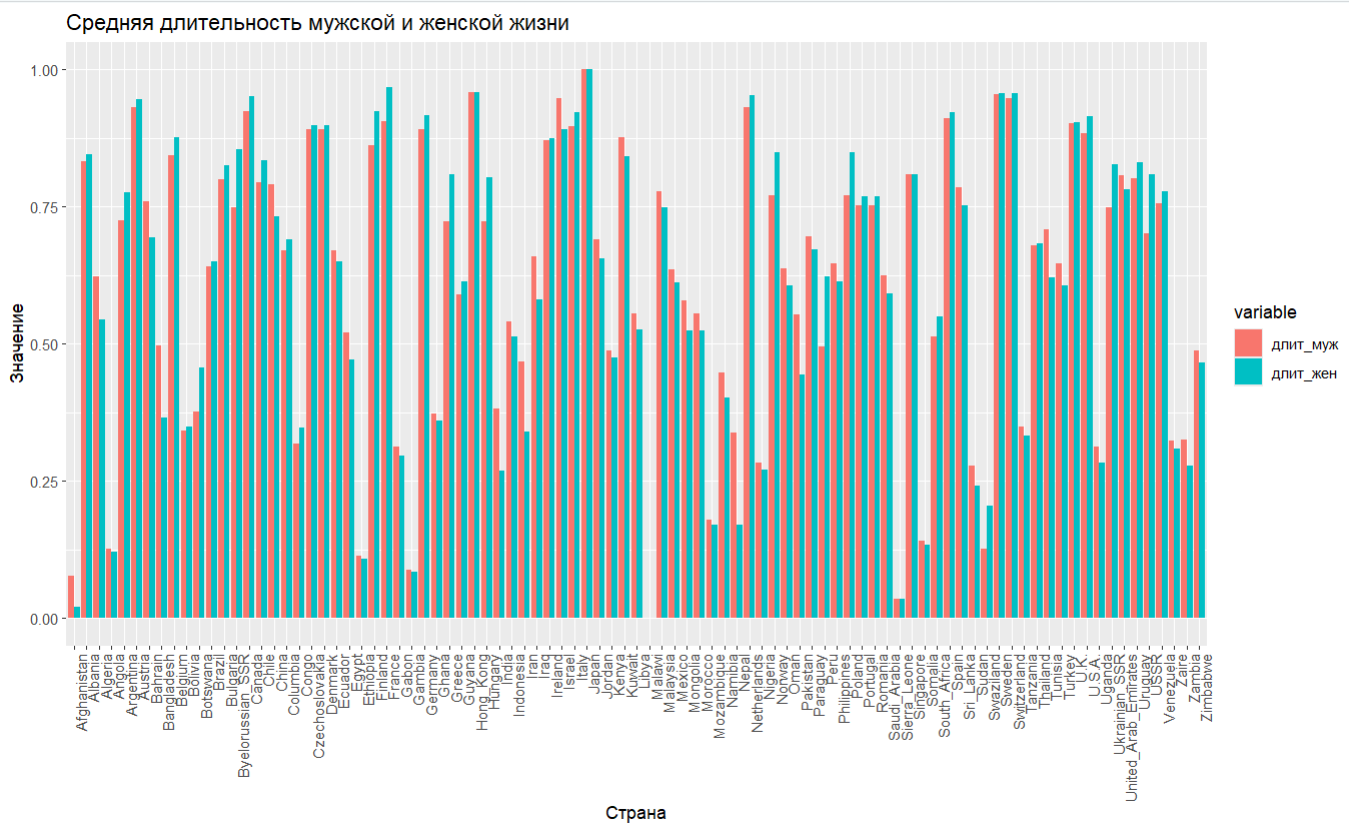


Рисунок 2 – График средней длительности мужчины и женщины в странах(значения нормализованы)

На данном графике можно явно увидеть сильные различия между странами, выделяются группы у которых высокие значения индексов, а также те у которых эти значения очень маленькие, это говорит нам о том, что по данным критериям данные можно разбить по нескольким кластерам.

1. **Определение оптимального числа кластеров.**

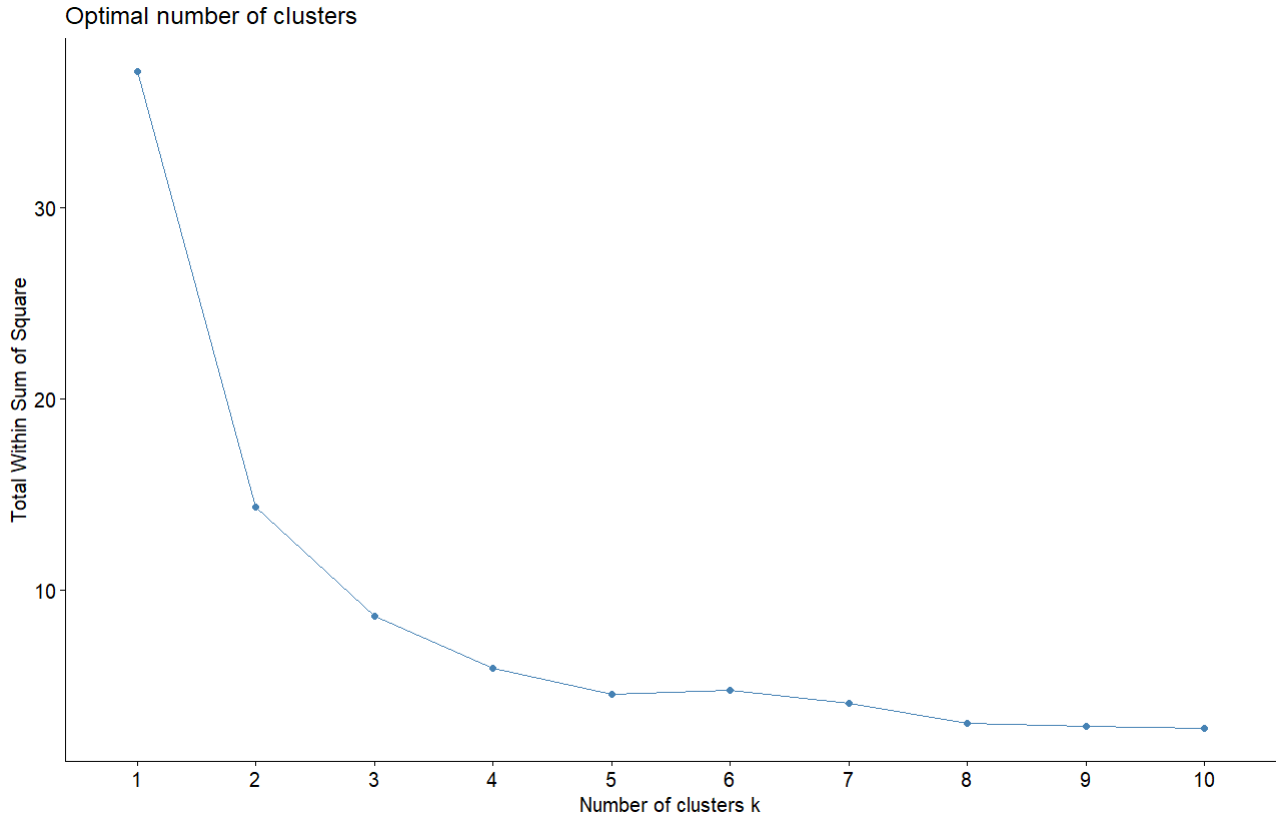


Рисунок 3 – Метод локтя

По методу локтя возьмём точку 4 – как точку, когда линия начала выравниваться, тогда 4 и будет оптимальным числом кластеров по этому методу.

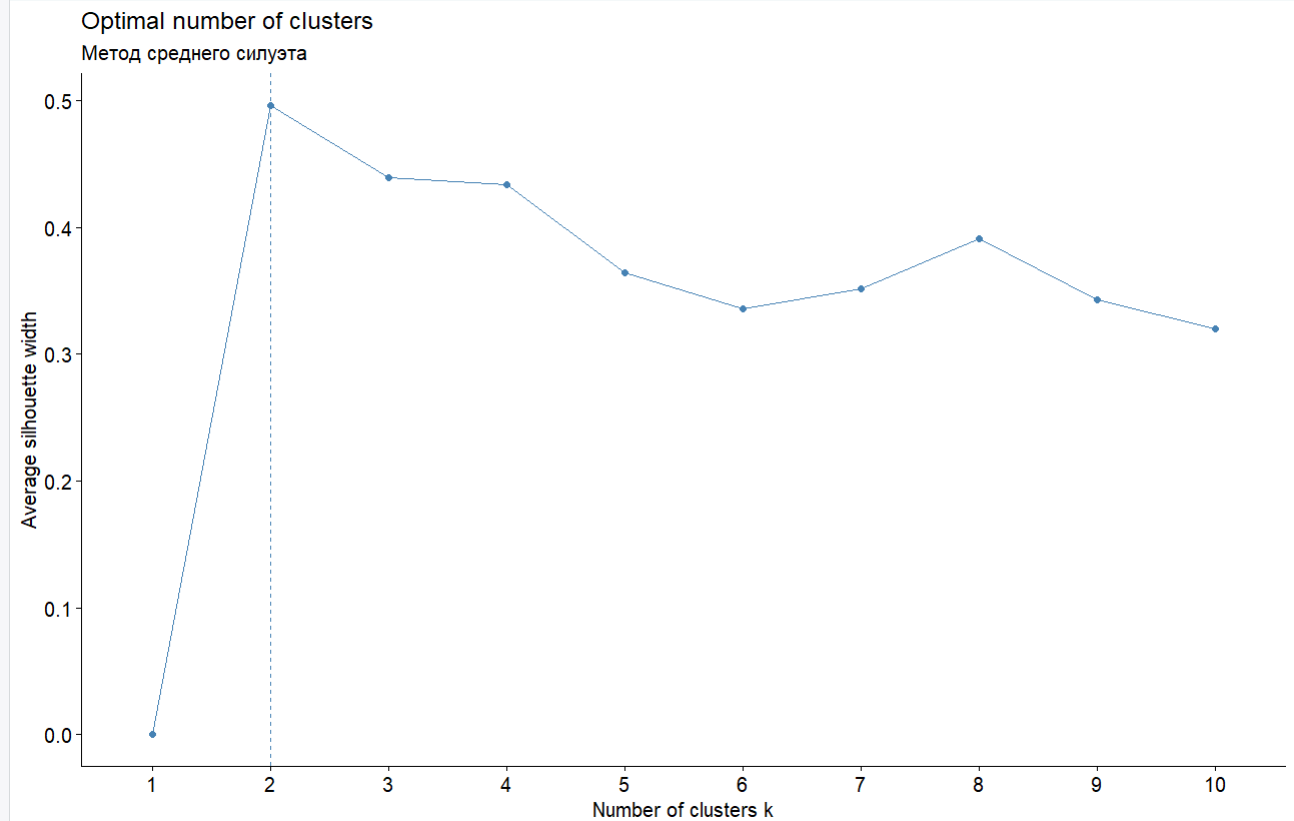


Рисунок 4 – Метод среднего силуэта

По данному методу оптимальным количеством кластеров будет 2.

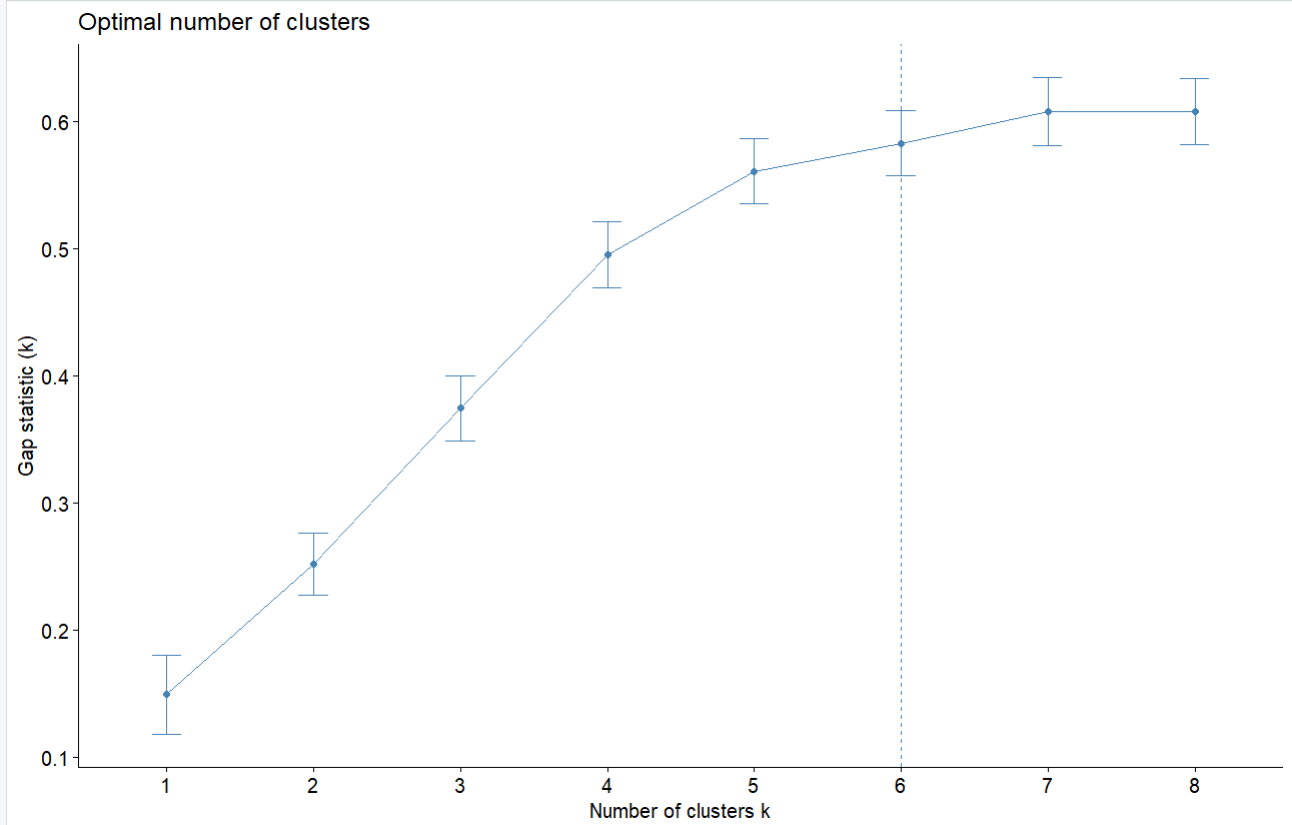


Рисунок 5 – По статистике разрыва

С помощью метода статистики разрыва оптимальное число кластеров равняется 6.

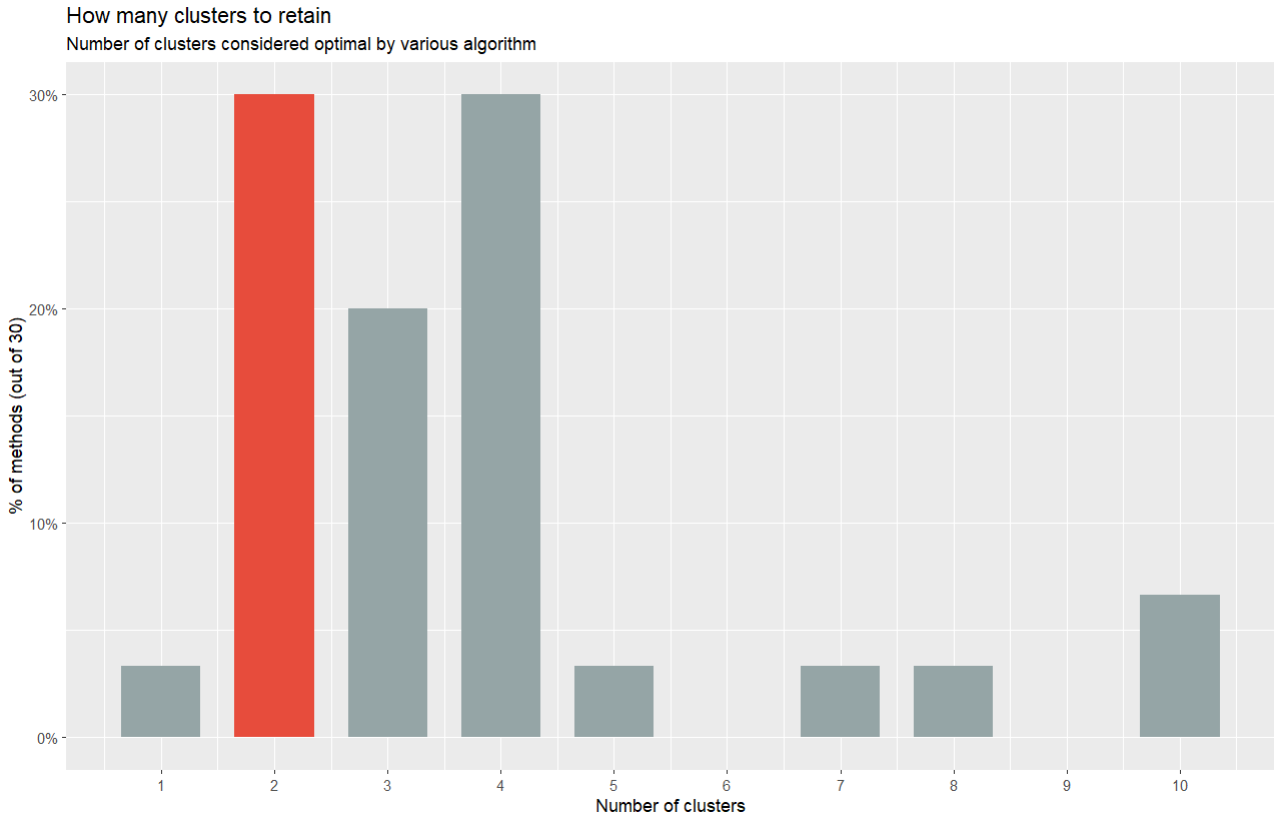


Рисунок 6 – Метод консенсуса

Будем ориентироваться на число 4 и 6 (6 – число, равное количеству регионов, и его мы получили с помощью статистики разрыва) для числа кластеров.

1. **Построение дендраграммы**

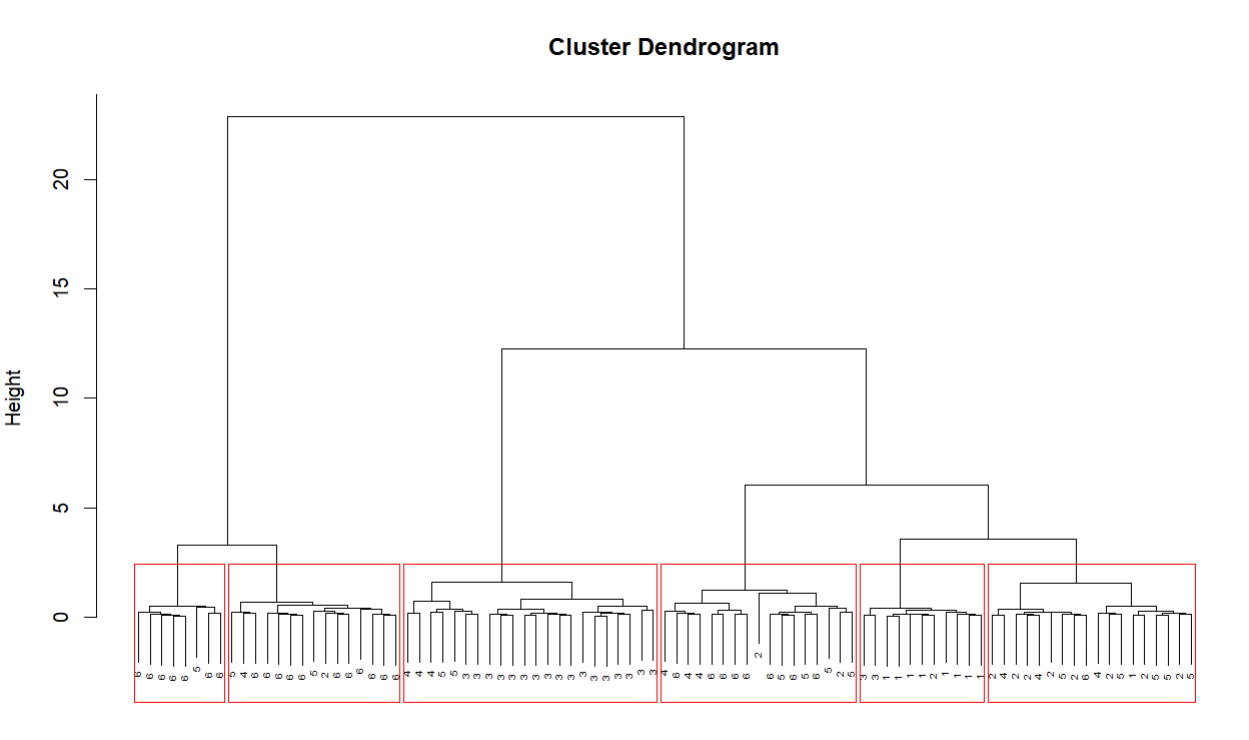
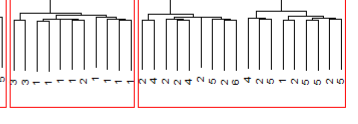
****

Рисунок 6 – Построенная дендрограмма, разделённая на 6 кластеров

Попробуем разделить дендрограмму на 6 кластеров, видим, что в кластерах присутствует множество смешанных значений:



Тогда уменьшим количество, объединив этот кластер.

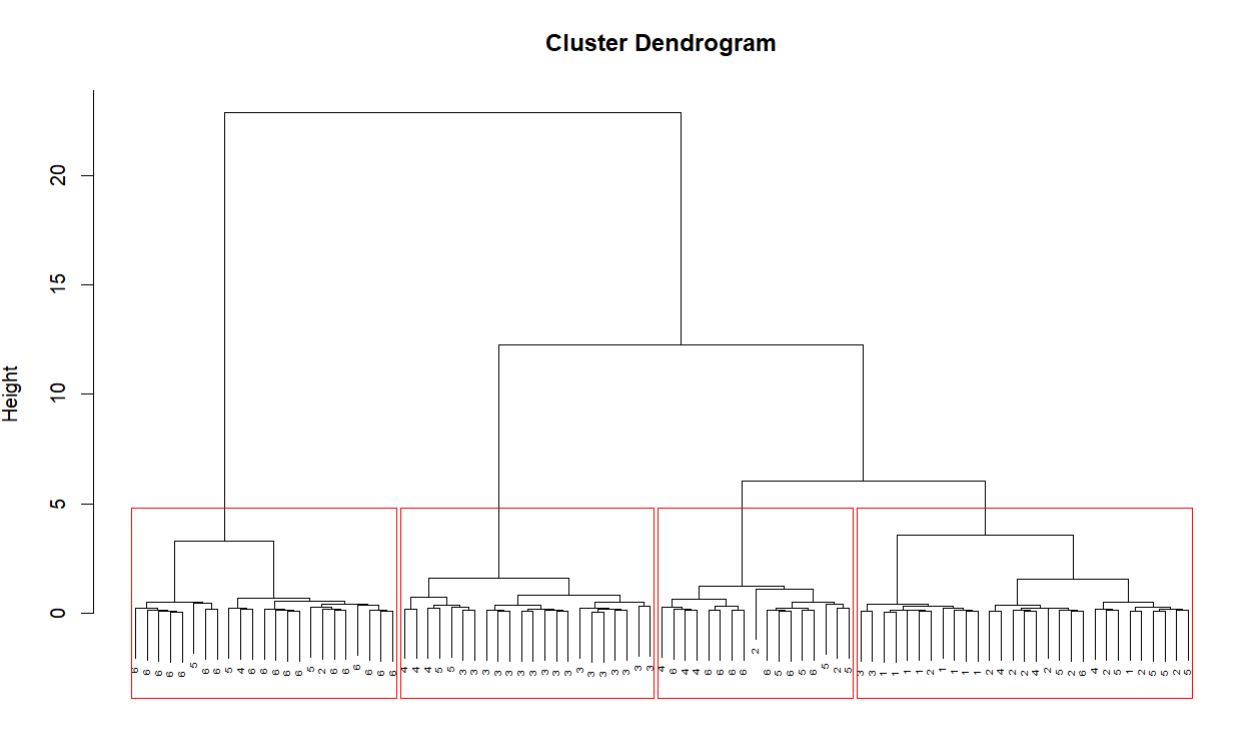


Рисунок 7 – Построенная дендрограмма, разделённая на 4 кластера

При разделении на 4 кластера, мы получили более лучшие группы, разделённые по регионам, группы получились более равномерные.

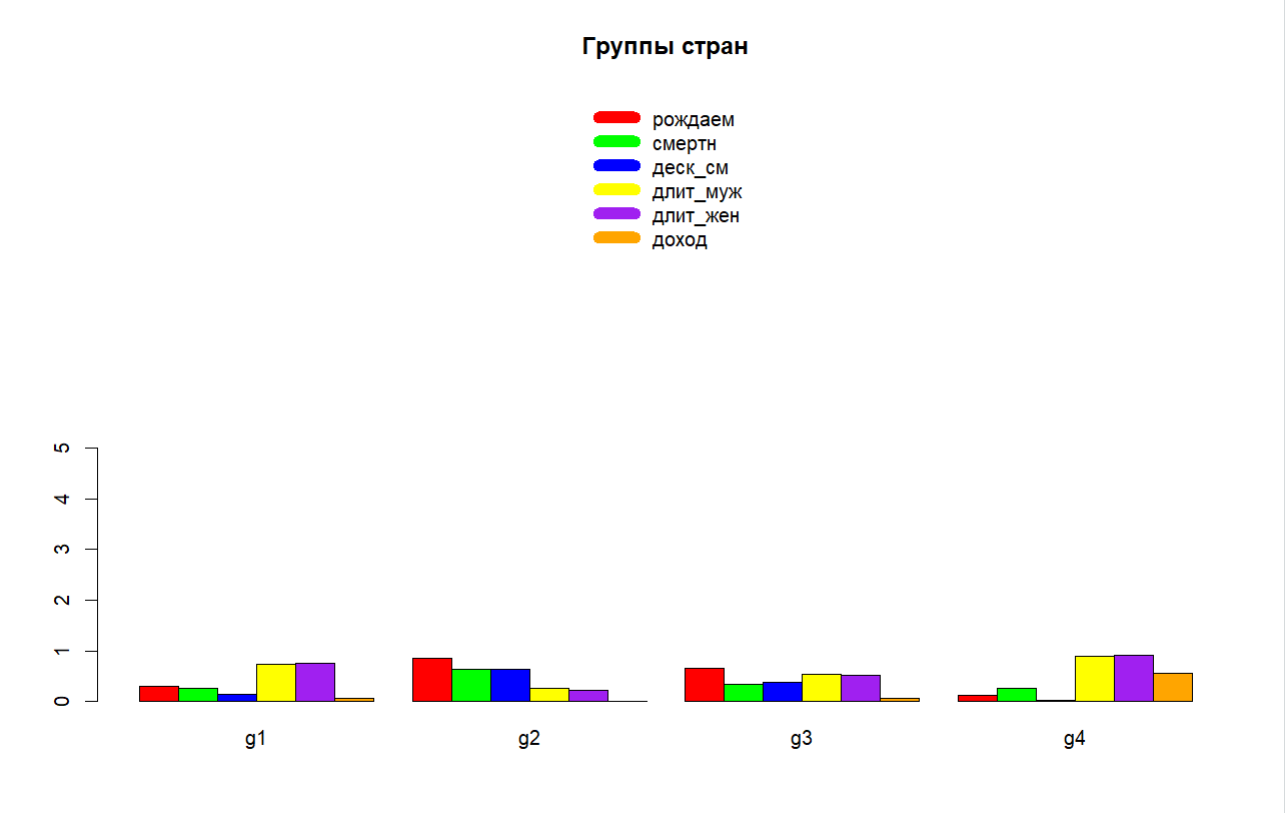


Рисунок 8 – группы стран

Выведя параметры групп стран в виде столбчатой диаграммы, видим явное различие групп по доходам, длительности жизней. Также на данной диаграмме уже можно сразу заметить некоторые зависимости, такие как, что при высоком доходе, низкая рождаемость и низкая детская смертность, а при низком доходе, рождаемость высокая и также высокая детская смертность.

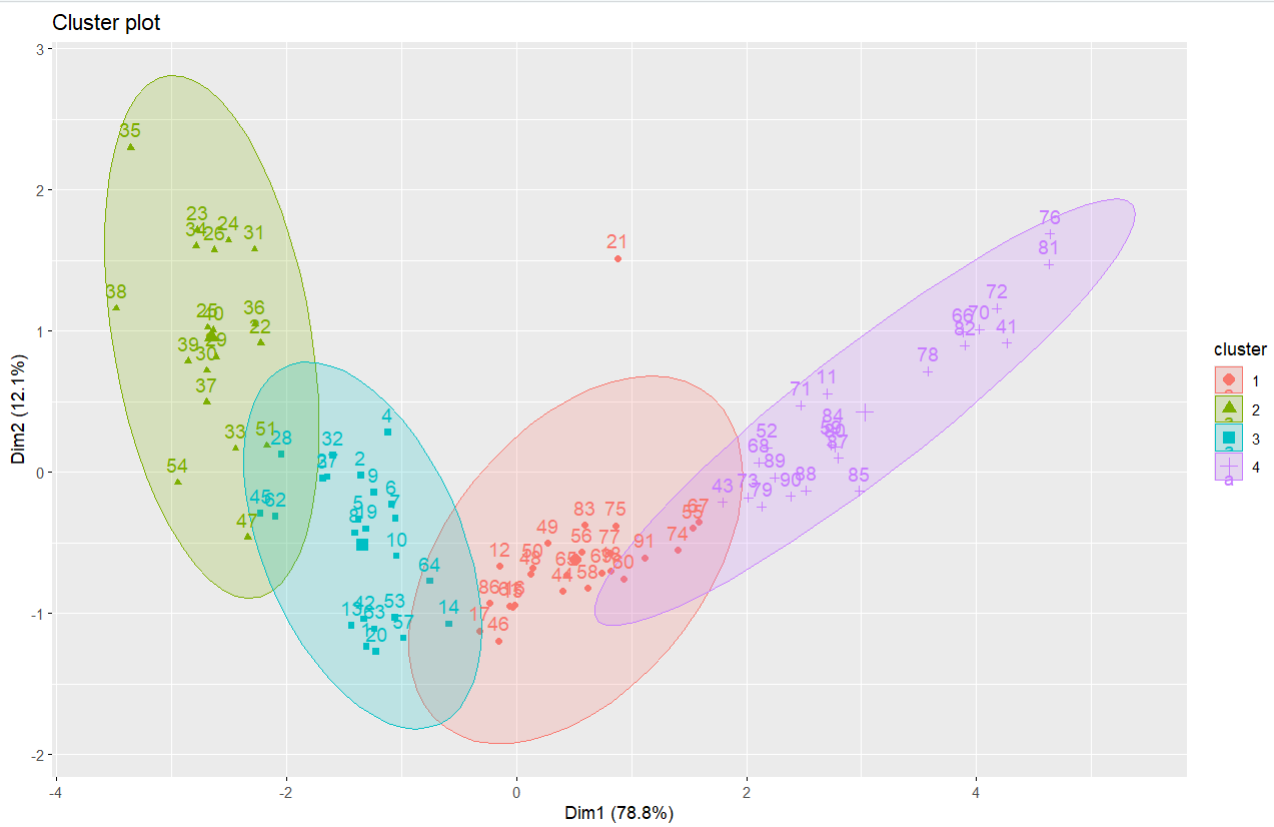


Рисунок 9 – K-means кластеризация

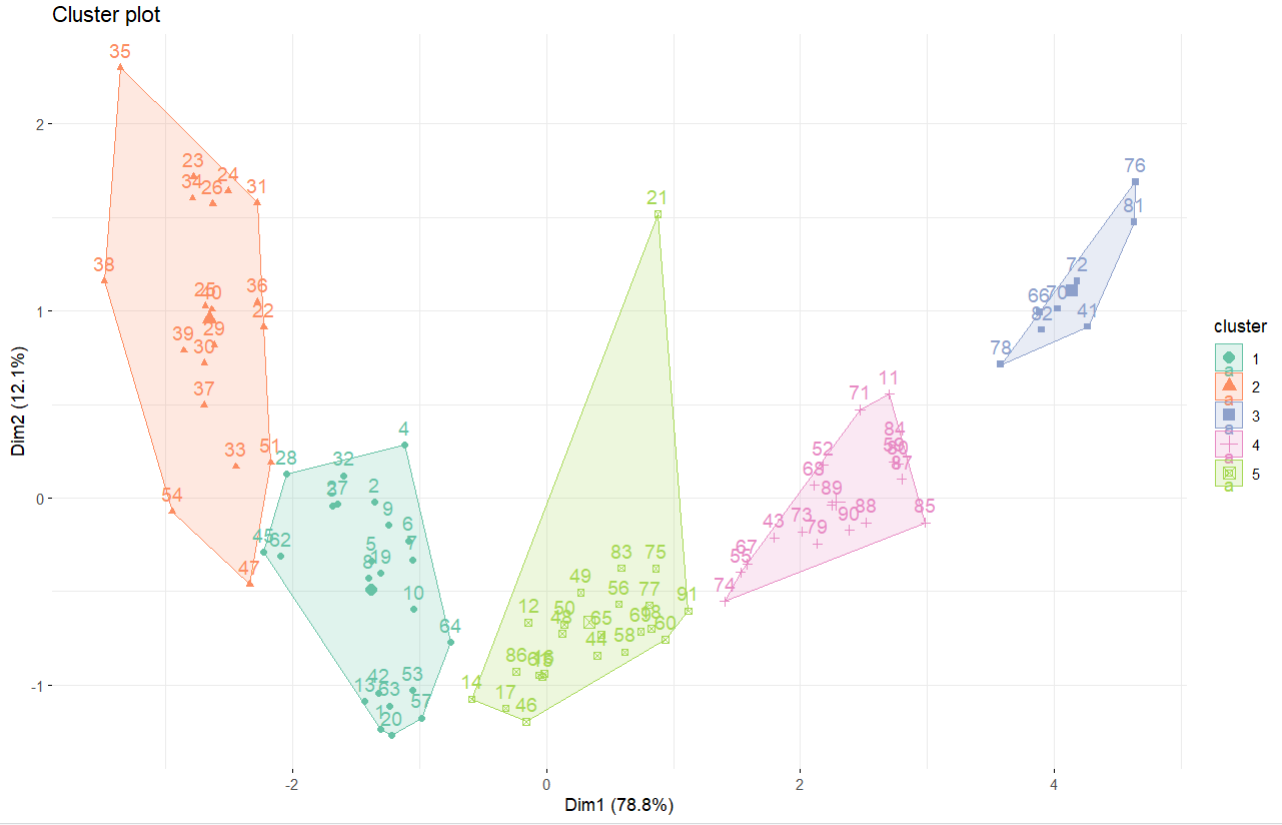


Рисунок 10 – Изменен вид k-mean

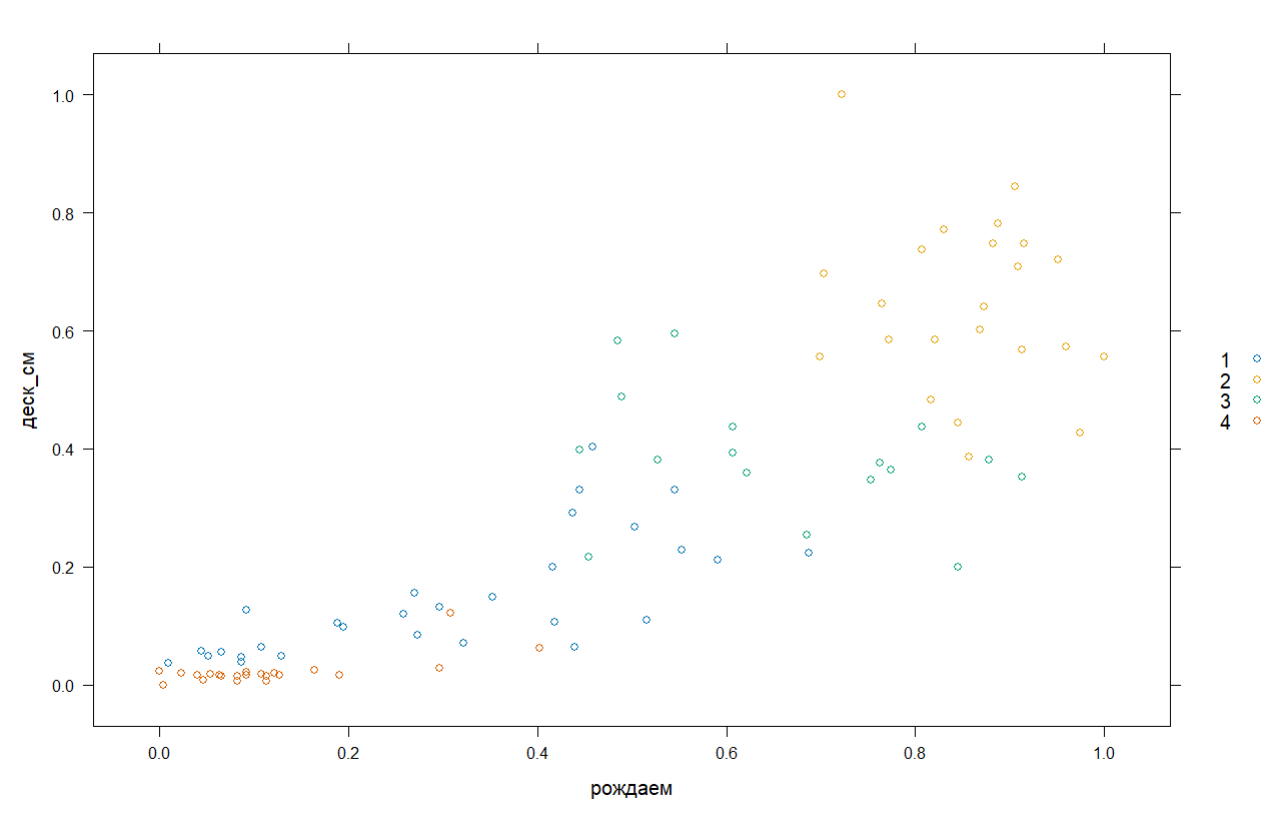


Рисунок 11 – Зависимость детской смертности от рождаемости по группам

Видим чёткую зависимость детской смертности от рождаемости, обычно, чем больше рождаемость в странах, тем и выше детская смертность, это во многом относит нас к странам Африки, где очень плохая ситуация с бедностью, но высокая рождаемость.

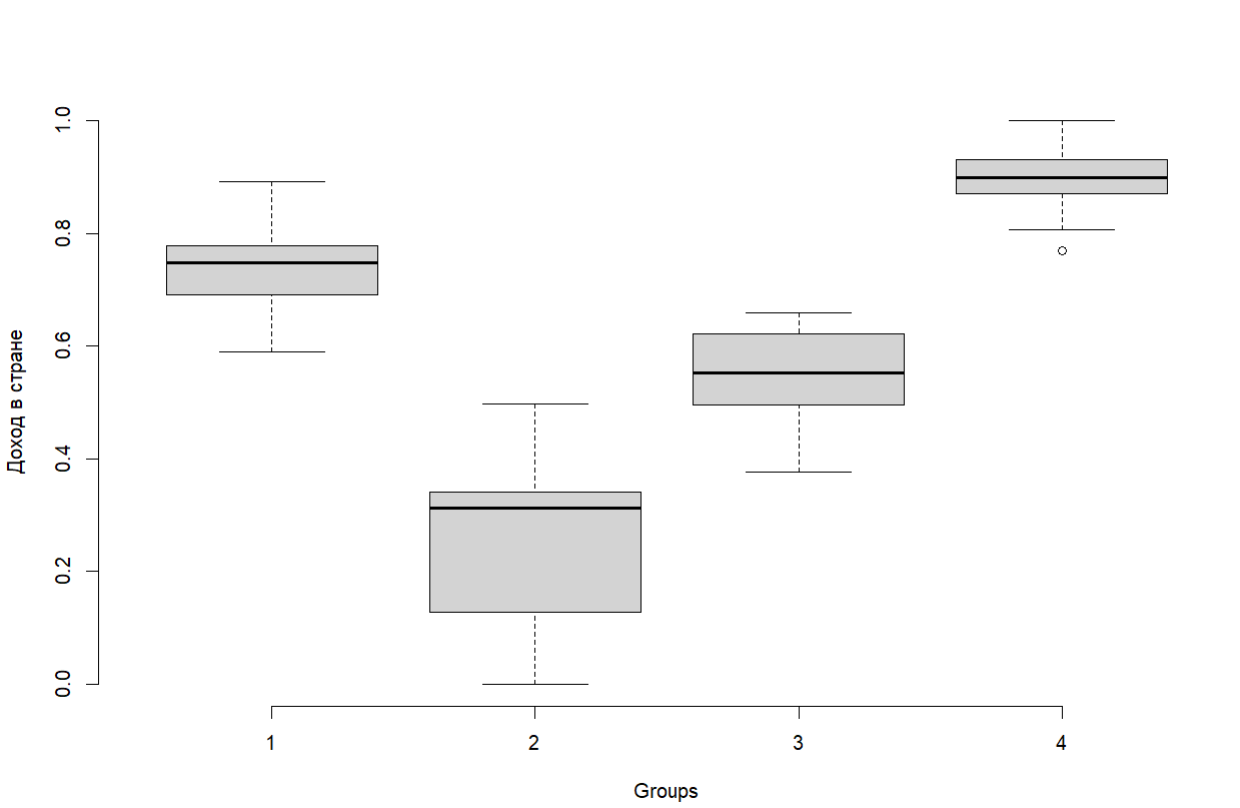


Рисунок 12 – Доход в стране по группам

По выводу боксплотов чётко видно различия между различными группами в доходах по стране.

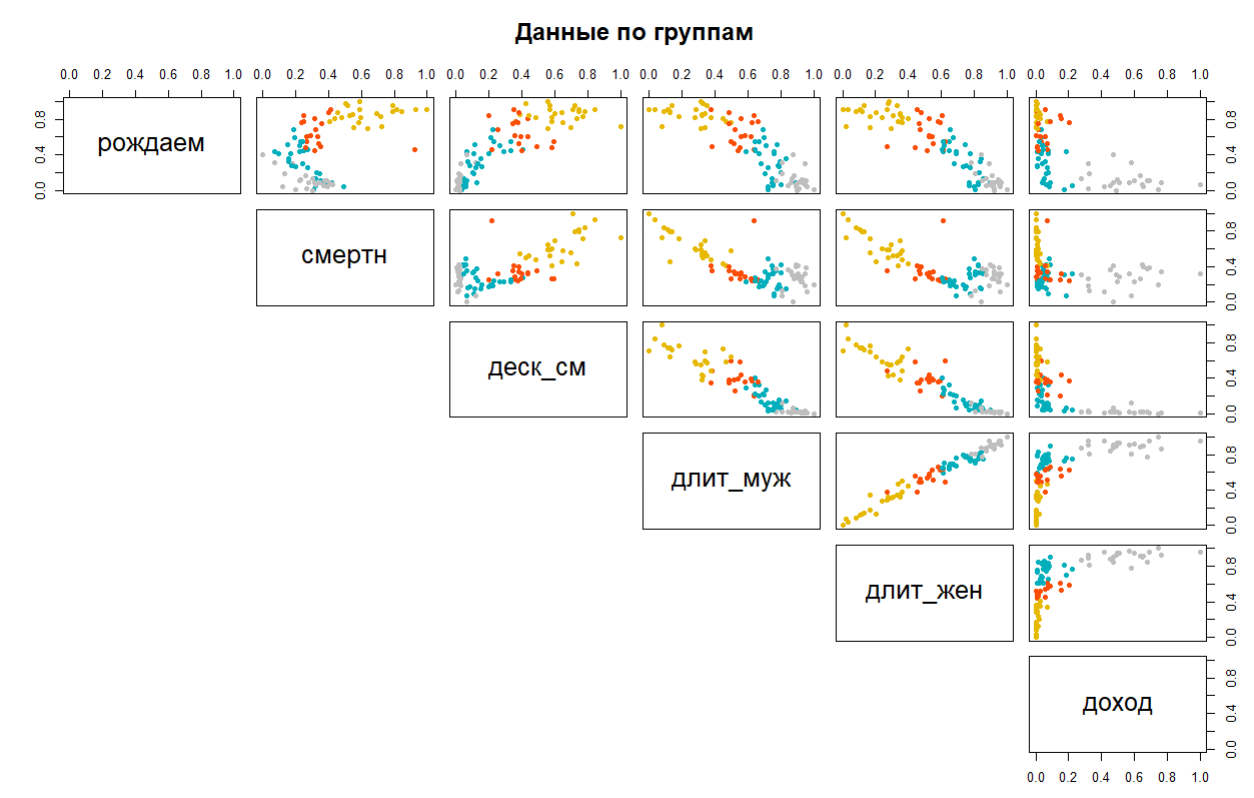


Рисунок 13 – Скаттерплот, отражающий особенности зависимостей по группам

По скаттерплоту далее мы продолжаем видеть чёткую корреляцию переменных рождаемости, смертности и длительности жизни мужчин и женщин.

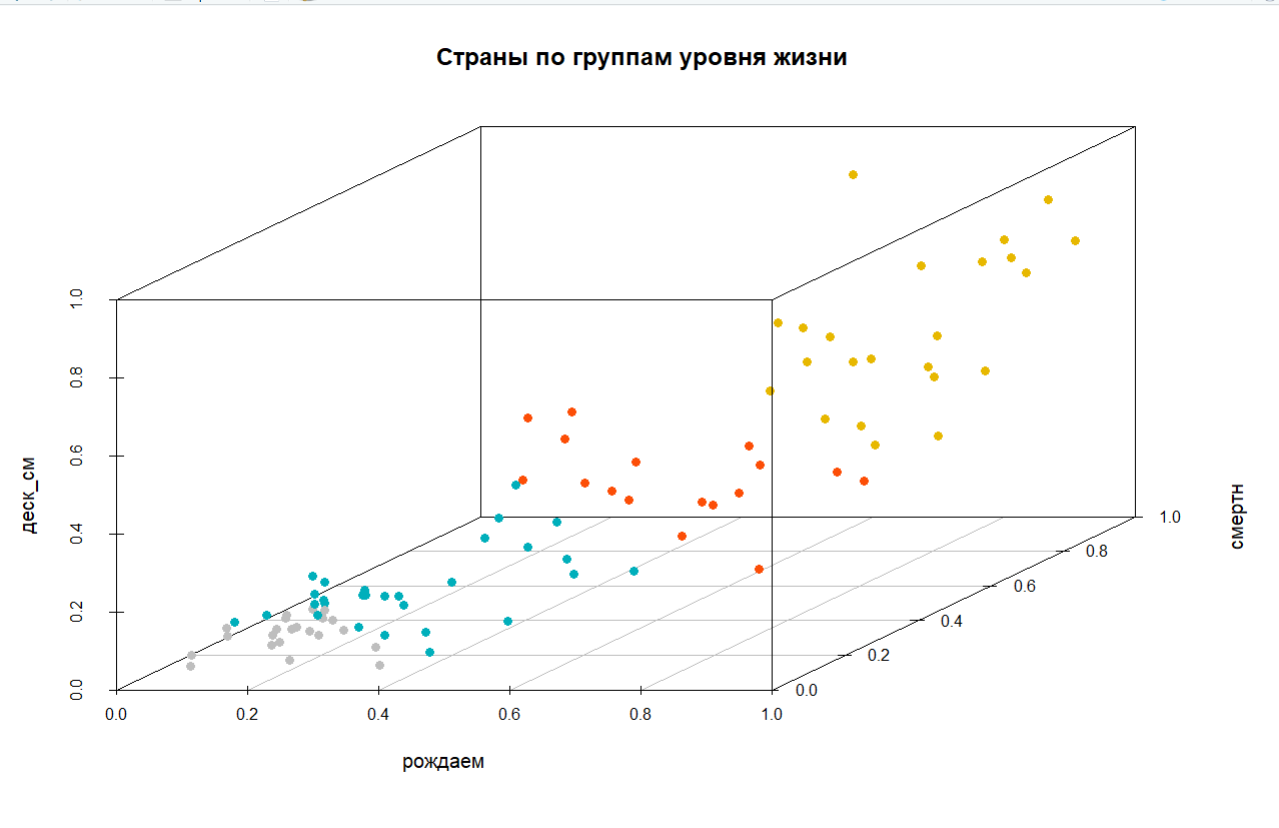


Рисунок 14 – 3D график

Сравнение графиков одних переменных от других даёт нам множество возможностей для анализа, например, видно, что обычно, чем больше рождаемость и меньше доход, тем больше детская смертность.

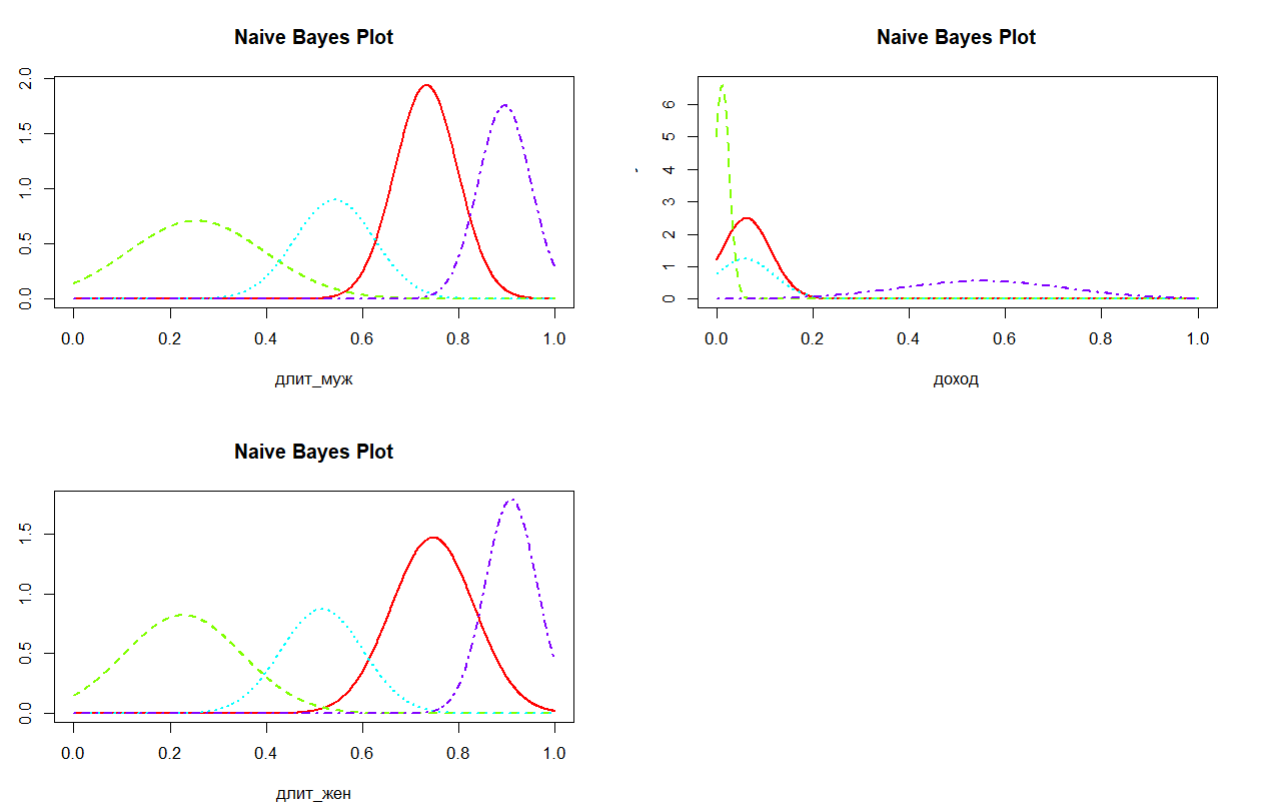


Рисунок 15 – Графики наивного Байеса

Классификатор метода наивного Байеса составил нам следующие таблицы вероятностей по всем признакам:

$рождаем

[,1] [,2]

1 0.3049899 0.19403402

2 0.8554476 0.08412704

3 0.6586851 0.15535977

4 0.1178610 0.10167141

$смертн

[,1] [,2]

1 0.2614943 0.1036214

2 0.6393974 0.1617462

3 0.3454592 0.1573037

4 0.2681419 0.1151725

$деск\_см

[,1] [,2]

1 0.14511575 0.10160477

2 0.64375322 0.14460672

3 0.38612283 0.10689721

4 0.02338176 0.02488481

$длит\_муж

[,1] [,2]

1 0.7332604 0.06553417

2 0.2544283 0.14218674

3 0.5423280 0.08276978

4 0.8959836 0.05490317

$длит\_жен

[,1] [,2]

1 0.7467301 0.08663578

2 0.2268152 0.12300240

3 0.5160243 0.08555083

4 0.9095387 0.05408677

$доход

[,1] [,2]

1 0.06129862 0.05100915

2 0.01138387 0.01529524

3 0.05798576 0.06046932

4 0.55154832 0.17348943

Средние значения параметров (первый столбец) и их стандартные отклонения (второй столбец) для каждого выделенного класса.

Так, можно установить, что у 4-ой группы длительность жизни мужчин и женщин наибольшие, доход наибольший, но малая рождаемость при низкой смертности и детской смертности. А у второй группы наибольшая детская смертность и наименьший доход при наибольшей рождаемости.

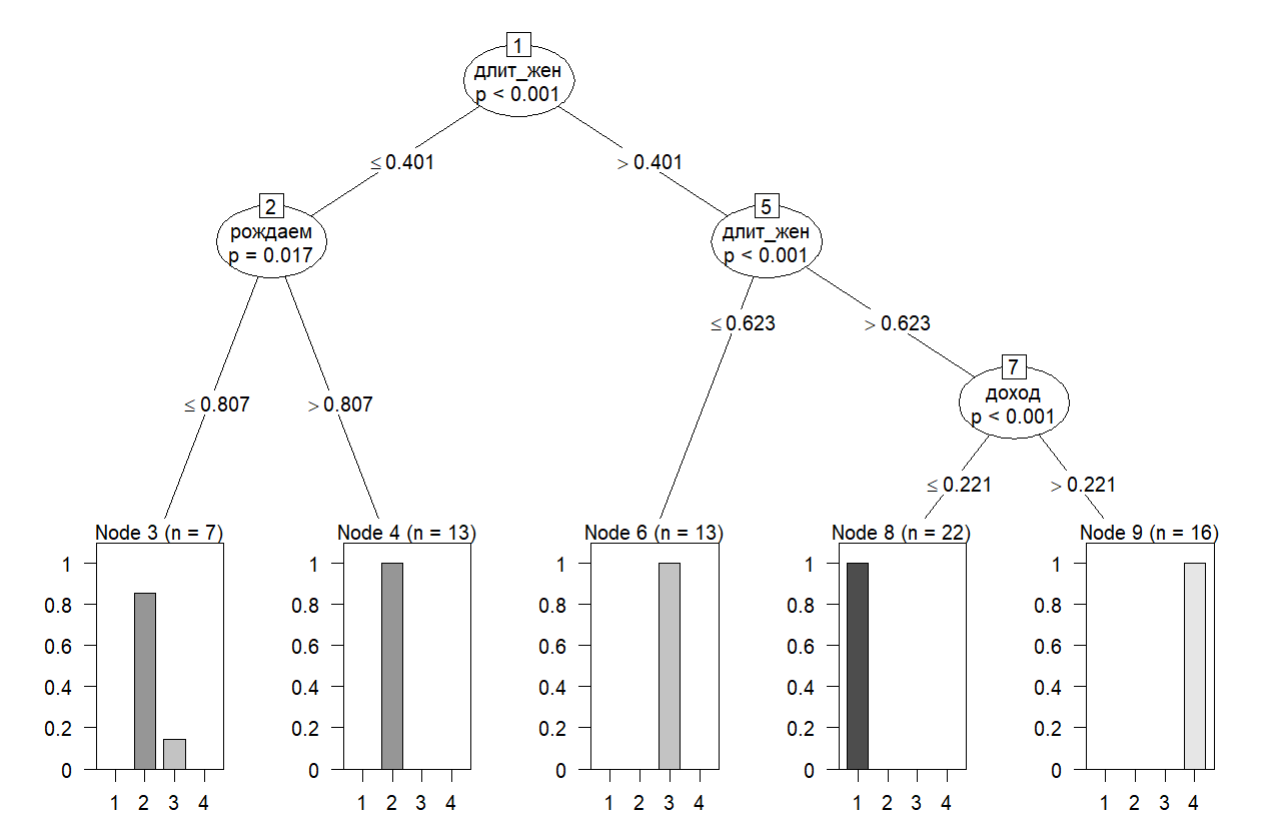


Рисунок 16 – График дерева решений

Для дерева решений было получена ошибка в менее чем 20% случаев при классификации для элемента из третьей группы.

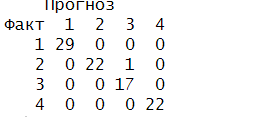


Рисунок 17 – Результат по Байесу

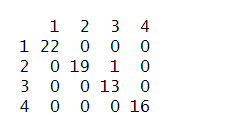


Рисунок 18 – Результат дерева решений

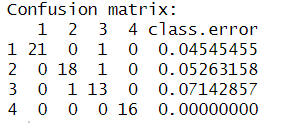


Рисунок 19 –Результаты предсказания случайного леса

Сравнивая полученные результаты, увидим, что все алгоритмы допустили по одной ошибке, за исключением случайного леса, в данном алгоритме было допущено 2 ошибки.

**Вывод:** Проведя анализ этого набора данных, были сделаны выводы о том, что данные 90 стран можно разделить на 4 группы. В каждой группе прослеживаются явные зависимости некоторых переменных друг от друга, в частности, детская смертность сильно зависит от дохода и рождаемости. Чем выше доходы, то тем ниже детская смертность, при этом чем выше рождаемость, тем и выше детская смертность. Кроме того есть зависимость длительности жизни мужчин и женщин от дохода, чем выше доход, тем выше длительность жизни.

Был проведён анализ с помощью различных алгоритмов: наивный алгоритм Байеса, дерево решений, случайный лес.

**Код скриптов на R:**

**Нормализация и преобразование файла:**

# Загрузка библиотек

library(readxl)

library(dplyr)

library(tidyr)

library(scales)

df <- read\_excel("zzz.xls")

# Удаление строк с пропущенными значениями

df <- df[!(df$страна == 9999 | df$рождаем == 9999 | df$рождаем == -9999 |

df$смертн == 9999 | df$смертн == -9999 | df$деск\_см == 9999 |

df$деск\_см == -9999 | df$длит\_муж == 9999 | df$длит\_муж == -9999 |

df$длит\_жен == 9999 | df$длит\_жен == -9999 | df$доход == 9999 |

df$доход == -9999 | df$регион == 9999 | df$регион == -9999), ]

summary(df)

# Нормализация числовых переменных

df$рождаем <- rescale(df$рождаем, to = c(0, 1))

df$смертн <- rescale(df$смертн, to = c(0, 1))

df$деск\_см <- rescale(df$деск\_см, to = c(0, 1))

df$длит\_муж <- rescale(df$длит\_муж, to = c(0, 1))

df$длит\_жен <- rescale(df$длит\_жен, to = c(0, 1))

df$доход <- rescale(df$доход, to = c(0, 1))

# Запись преобразованного DataFrame в файл "zzz.csv"

write.csv(df, "zzz.csv", row.names = FALSE)

**Дескриптивный анализ:**

# Импортируем данные из CSV файла

data <- read.csv("zzz.csv", header = TRUE)

# Смотрим на первые несколько строк

head(data)

# Получаем общее описание данных

summar <- summary(data)

print(summar)

# Выводим размер датафрейма

dim(data)

# Находим уникальные значения для каждого столбца

sapply(data, function(x) length(unique(x)))

# Проверяем наличие пропущенных значений

sum(is.na(data))

# Получаем основные статистики для числовых столбцов

summary(data[, sapply(data, is.numeric)])

# Сортируем данные по убыванию значения

data <- data[order(-data$доход), ]

# Выбираем топ 10 стран

top\_10 <- head(data[, c("страна", "доход")], 10)

# Создаем столбчатую диаграмму

barplot(top\_10$доход,

names.arg = top\_10$страна,

xlab = "",

ylab = "Доход",

main = "Топ 10 стран по параметру дохода",

col = "steelblue",

las=2

)

# Создаем новый датафрейм с нормализованными значениями

data\_plot <- data[, c("страна", "длит\_муж", "длит\_жен")]

# Преобразуем данные из широкого формата в длинный

library(reshape2)

data\_long <- melt(data\_plot, id.vars = "страна")

# Создаем столбчатую диаграмму

library(ggplot2)

ggplot(data\_long, aes(x = страна, y = value, fill = variable)) +

geom\_bar(stat = "identity", position = "dodge") +

labs(x = "Страна", y = "Значение", title = "Средняя длительность мужской и женской жизни") +

theme(axis.text.x = element\_text(angle = 90, hjust = 1))

**Нахождение оптимального числа кластеров:**

# Импортируем данные из CSV файла

data <- read.csv("zzz.csv", header = TRUE)

head(data)

data <- data[, c( -8)]

head(data)

# Определение числовых и нечисловых столбцов

num\_cols <- sapply(data, is.numeric)

num\_cols

# Создание нового датафрейма с только числовыми столбцами

num\_data <- subset(data, select = -num\_cols)

dist.countries <- dist(num\_data)

clust.countries <- hclust(dist.countries, "ward.D")

# install.packages("factoextra")

library (factoextra)

library (cluster)

# Метод локтя

fviz\_nbclust(num\_data, kmeans, method = "wss")

# Silhouette method

fviz\_nbclust(num\_data, kmeans, method = "silhouette") +

labs(subtitle = "Метод среднего силуэта")

# Метод статистики разрыва

#посчитать статистику разрыва, базирующуюся на числе кластеров K.max = 6:

gap\_stat <- clusGap(num\_data, FUN = kmeans, nstart =4,K.max =8, B = 100)

#plot number of clusters vs. gap statistic

fviz\_gap\_stat(gap\_stat)

#Алгоритм на основе консенсуса

# install.packages('parameters')

# install.packages('NbClust')

# install.packages('mclust')

library(parameters)

library(NbClust)

library(mclust)

# Применение кластеризации методом консенсуса

# install.packages('parameters')

library(parameters)

n\_clust <- n\_clusters(data,

package = c("easystats", "NbClust", "mclust"),

standardize = FALSE)

n\_clust

plot(n\_clust)

**Построение дендрограмм и других графиков:**

# Импортируем данные из CSV файла

data <- read.csv("zzz.csv", header = TRUE)

head(data)

data\_wthout\_region <- data[, c(-8)]

# Определение числовых и нечисловых столбцов

num\_cols <- sapply(data\_wthout\_region, is.numeric)

# Создание нового датафрейма с только числовыми столбцами

num\_data <- data\_wthout\_region[, num\_cols]

head(num\_data)

dist.countries <- dist(num\_data)

clust.countries <- hclust(dist.countries, "ward.D")

plot(clust.countries, data$регион, cex=0.5)

rect.hclust(clust.countries, k=4, border="red")

# Вектор groups содержит номер кластера, в который попал классифицируемый объект

groups <- cutree(clust.countries, k=4)

# Для каждой группы определяем средние значения характеристик и строим датафрейм.

# в 1-ом кластере

g1<-colMeans(num\_data[groups==1, 1:6])

# во 2-ом кластере

g2<-colMeans(num\_data[groups==2, 1:6])

# в 3-ем кластере

g3<-colMeans(num\_data[groups==3, 1:6])

# в 4-ем кластере

g4<-colMeans(num\_data[groups==4, 1:6])

df<-data.frame(g1,g2,g3, g4)

df1<-t(df)

df<-t(df1)

barplot(df, ylim=c(0,12),

main = "Группы стран", axes = FALSE,

col=c("red","green","blue","yellow", "purple", "orange"), beside=TRUE)

axis(2, at = 0:5, labels = 0:5)

legend("top", legend = rownames(df), col=c("red","green","blue","yellow", "purple", "orange"), lwd=10, bty = "n")

# КЛАСТЕРИЗАЦИЯ K-MEANS

km.res <- kmeans(num\_data, 5, nstart = 10)

fviz\_cluster(km.res, num\_data, ellipse.type = "norm")

# Изменить палитру и темуы

fviz\_cluster(km.res, num\_data,

palette = "Set2", ggtheme = theme\_minimal())

library (lattice)

xyplot(деск\_см ~ рождаем,

data = data,

group = groups,

auto.key = TRUE

)

data["Groups"] <- groups

boxplot(

длит\_муж ~ Groups,

data = data,

ylab = "Доход в стране",

frame = FALSE,

col = "lightgray"

)

pairs(num\_data)

pairs(

num\_data,

main= "Уровень жизни в стране",

col = c("red","green","blue", "grey")

)

my\_cols <- c("#00AFBB", "#E7B800", "#FC4E07", "grey")

pairs(

num\_data,

main= "Данные по группам",

pch = 19, cex = 0.8,

col = my\_cols[data$Groups],

lower.panel=NULL

)

# Трёхмерный скаттерплот

# install.packages("scatterplot3d")

library("scatterplot3d")

# Трехмерная классификация ирисов:

colors <- c("#00AFBB", "#E7B800", "#FC4E07", "grey")

colors <- colors[as.numeric(data$Groups)]

head(num\_data)

s3d <- scatterplot3d(

num\_data[1:3],

main= "Страны по группам уровня жизни",

pch = 16,

color=colors)

legend(

s3d$xyz.convert(7.5, 3, 4.5),

legend = data$Groups,

col = colors,

pch = 16)

**Наивный Байес, деревья решений, случайный лес:**

#install.packages("klaR")

library(klaR)

#ЧАСТЬ 2

data <- read.csv("zzz.csv", header = TRUE)

data <- data[,-8]

groups

my\_data <- data

my\_data$Group<- c(as.factor(groups))

naive\_df <- NaiveBayes(my\_data$Group ~ ., data = my\_data)

naive\_df$tables

naive\_df$tables$доход

naive\_df

#делаем графики по байсу

opar=par()

opar

layout(matrix(c(1,2,3,4), 2, 2))

plot(naive\_df, lwd = 2, legendplot = FALSE)

legend("topleft",lty=1:3, cex=0.5)

#восстановление графика

par=opar

# Предсказание - результаты по Байесу

pred <- predict(naive\_df, my\_data)$class

(table(Факт = my\_data$Group, Прогноз = pred))

# Классификация Decision Tree

#install.packages("party")

library(party)

# Разделяем данные на два датасета

set.seed(1234)

ind <- sample(2, nrow(my\_data), replace=TRUE, prob=c(0.7, 0.3))

trainData <- my\_data[ind==1,]

testData <- my\_data[ind==2,]

nrow(trainData)

nrow(testData)

nrow(my\_data)

# Строим модель

my\_data

myFormula <- Group ~ рождаем + смертн + деск\_см + длит\_муж + длит\_жен + доход

df\_ctree <- ctree(myFormula, data=trainData)

df\_ctree

predict(df\_ctree)

table(predict(df\_ctree, newdata=testData), testData$Group)

predict(df\_ctree)

plot(df\_ctree)

# Алгоритм Random Forest

# install.packages("randomForest")

library(randomForest)

rf <- randomForest(Group ~ .,data=trainData, ntree=100, proximity=TRUE)

table(predict(rf, newdata=testData), testData$Group)

print(rf)