Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**ЛАБОРАТОРНАЯ РАБОТА №6**

**Дисциплина: Обработка больших данных**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ К. А. Корнилов

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_В. И. Шиян

**Тема:** Кластеризация и классификация.

**Цель**: Закрепить знания, об алгоритмах классификации и кластеризации данных, ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научиться визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты.

**Задание:**

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. 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.
9. Для наивного Байесовского классификатора: решите задачу с помощью наивного Байесовского классификатора; проанализируйте точность полученных решений для тестовых данных
10. Для деревьев решений: примените метод деревьев решений для задачи классификации (для того же набора данных); исследовать дерево решений; если позволяет размерность, постройте его график. Проанализировать точность полученных решений для тестовых данных (с известным значением переменной отклика), сравните результаты с ранее полученными.
11. Выполнить классификацию с помощью случайного леса, сопоставить результат с результатом дерева решения, прокомментировать результат сравнения. Сопоставьте результаты с результатами Байесовского классификатора. Проанализируйте полученные результаты, сделайте вывод относительно результатов кластерного анализа (ЛР 6.1), оформите общий отчёт по ЛР 6.1 и 6.2.

**Ход работы**:

1. Был выполнен дискриптивный анализ набора данных о сердечных заболеваниях людей.

Были вычислены среднее значение, медиана и мода для каждого из признаков в наборе данных.

Код вычисления:

means = apply(data,2,mean)

means

medians = apply(data,2,median)

medians

getmode <- function(v) {

uniqv <- unique(v)

uniqv[which.max(tabulate(match(v, uniqv)))]

}

modas = apply(data,2,getmode)

modas

Результат вычисления представлен на рисунке 1.

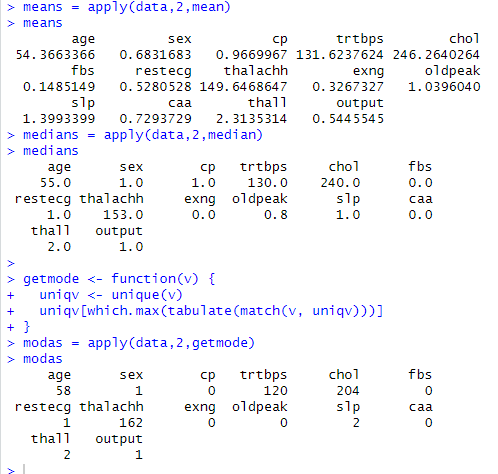


Рисунок 1 – Результат вычисления центральных тенденций для набора данных

Были вычислены значения дисперсии, среднего квадратичного отклонения и квартилей. Результат вычисления представлен на рисунке 2.

Код вычисления:

#Считаем дисперсия и стандартное отклонения

variances = apply(data,2,var)

variances

sds = apply(data,2,sd)

sds

# Считаем квартили,мин и макс

mins = apply(data,2,min)

mins

maxs = apply(data,2,max)

maxs

quartilies = apply(data,2,quantile)

quartilies

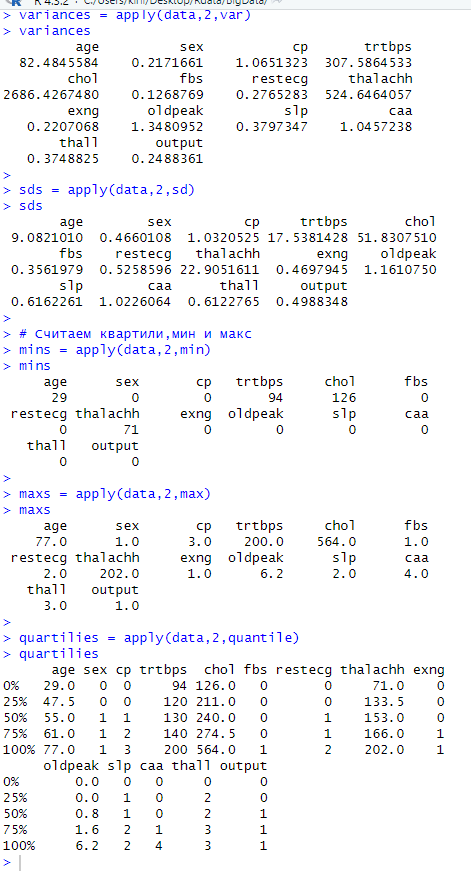


Рисунок 2 – Результат вычисления тенденций изменения и квартилей.

Были построены графики распределения каждой из величин и их боксплоты. Результат представлен на рисунке 3-4.

Код:

par(mfrow=c(3,3))

for(i in 1:(length(colnames(data))-5)){

boxplot(data[,i],main=colnames(data)[i])

}

par(mfrow=c(3,2))

for(i in (length(colnames(data))-4):(length(colnames(data)))){

boxplot(data[,i],main=colnames(data)[i])

}

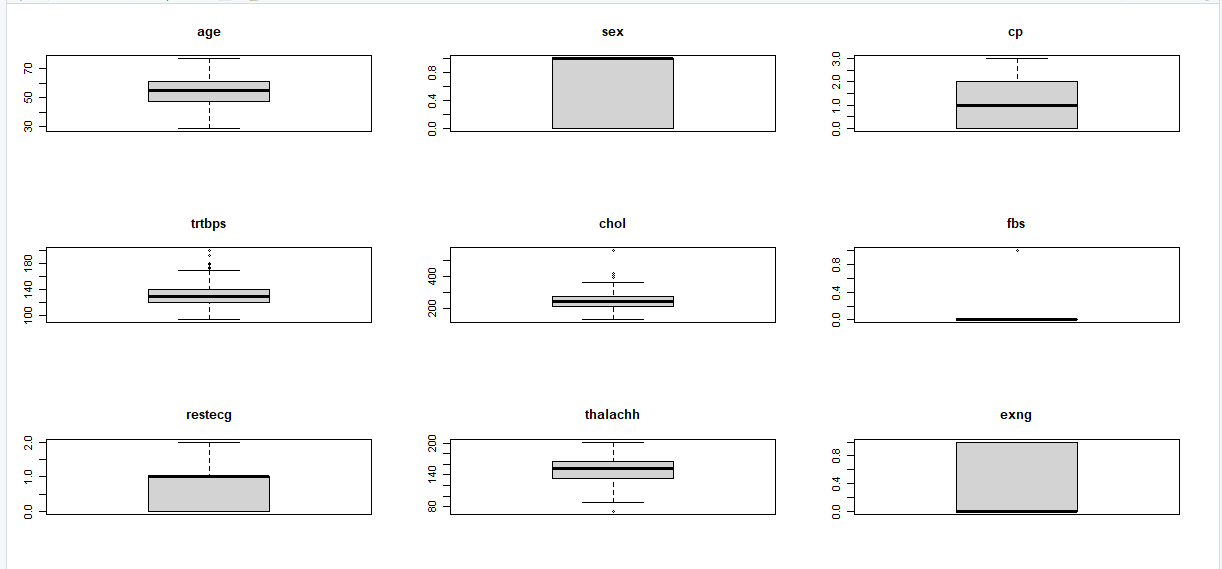
#Смотрим распределение

par(mfrow=c(4,4))

for(i in 1:(length(colnames(data)))){

hist(data[,i],xlab='values',main=colnames(data)[i],col=rainbow(length(unique(data[,i]))))

}



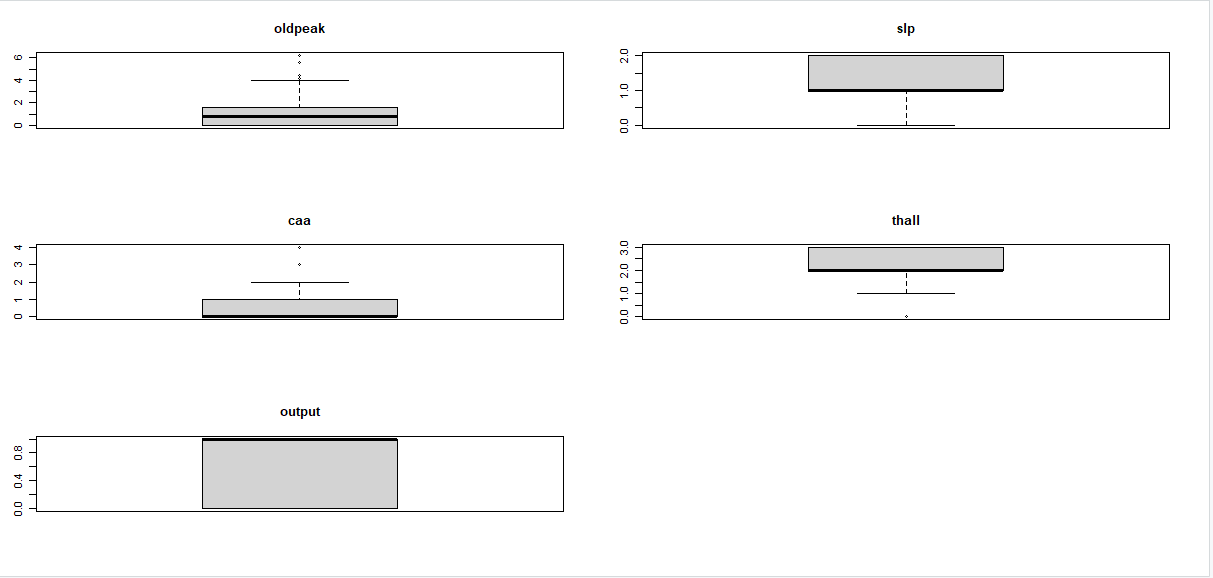


Рисунок 3 – Результат построения боксплота для набора данных

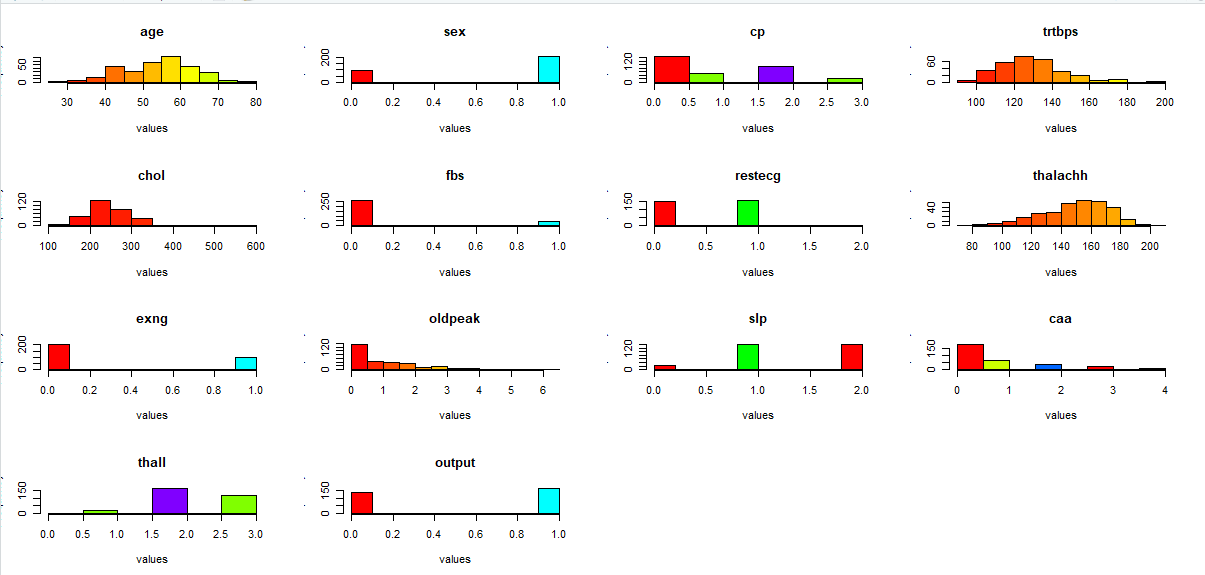


Рисунок 4 – Результаты построения гистограммы для набора данных

1. Было оценено оптимальное число кластеров. С данной целью были построены диаграммы: “Метод локтя”,”Метод силуэта”,”Статистика разрыва” и алгоритм консенсуса.

Перед этими действиями все данные были стандартизированы и нормированы. Код стандартизирования и нормирования:

source('lab6/desc\_analysis.R')

rm(list=ls())

data\_scaled = scale(data,center=mins,scale=maxs-mins)

data\_scaled

data

library(BBmisc)

data\_sc\_norm = normalize(data\_scaled,method="range",range=c(0,1))

data\_sc\_norm

data\_k=data\_sc\_norm[,-(length(colnames(data\_sc\_norm)))]

data\_k

Код построения диаграмм:

#Построения локтя

library (factoextra)

library (cluster)

fviz\_nbclust(data\_k, kmeans, method = "wss")

#Построили метод силуэта

set.seed(123)

fviz\_nbclust(data\_k, kmeans, method = "silhouette") +

labs(subtitle = "Silhouette method")

#

set.seed(123)

fviz\_nbclust(data, kmeans, method = "wss")

gap\_stat <- clusGap(data\_k, FUN = kmeans, K.max =12)

#plot number of clusters vs. gap statistic

fviz\_gap\_stat(gap\_stat)

#Алгоритм на основе консенсуса

library(parameters)

n\_clust <- n\_clusters(data.frame(data\_k),

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

standardize = FALSE)

n\_clust

plot(n\_clust)

Результаты построения диаграмм представлены на рисунках 5,6,7,8.

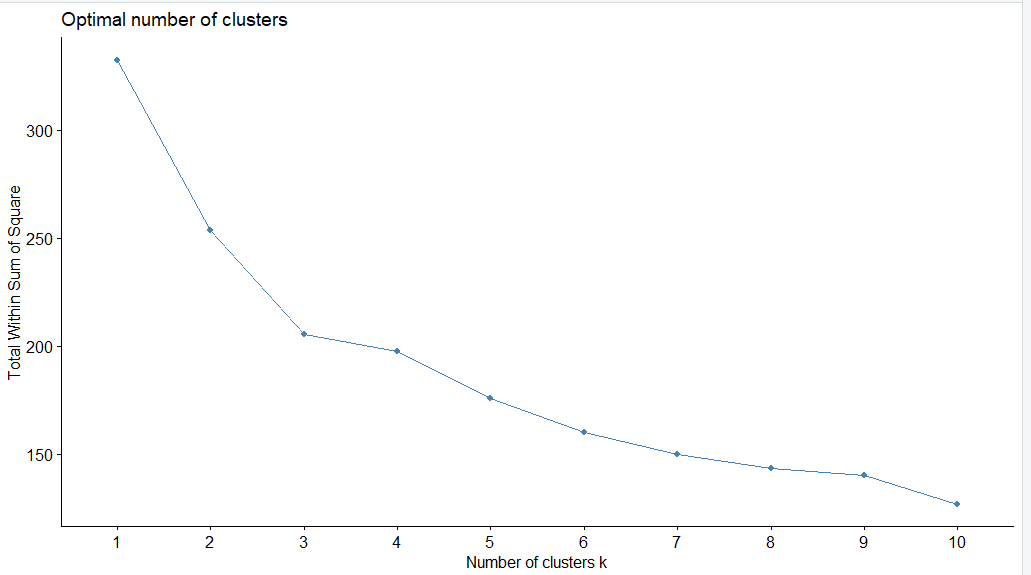


Рисунок 5 – Результат построения метода локтя

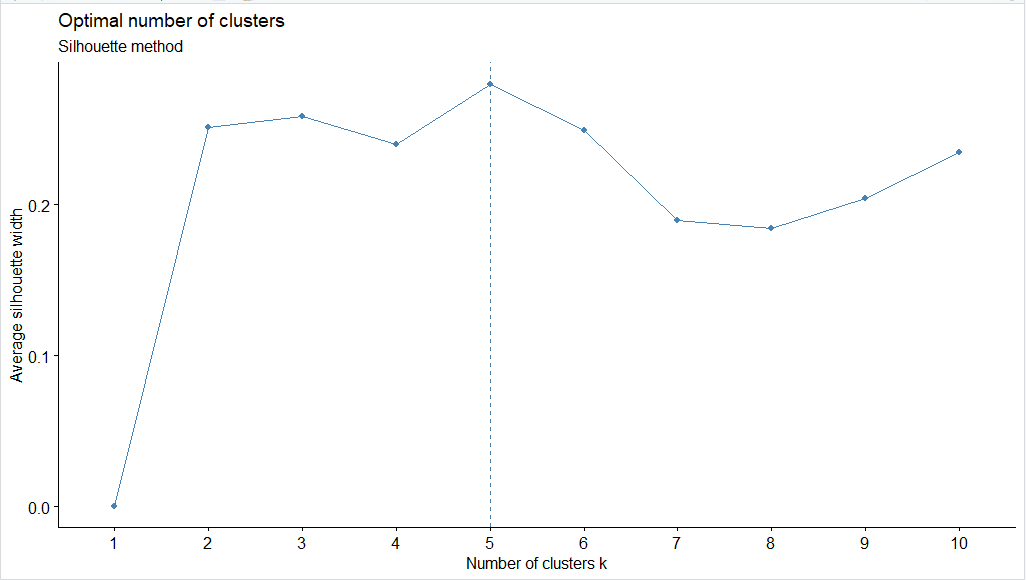


Рисунок 6 – Результат построения метода силуэта

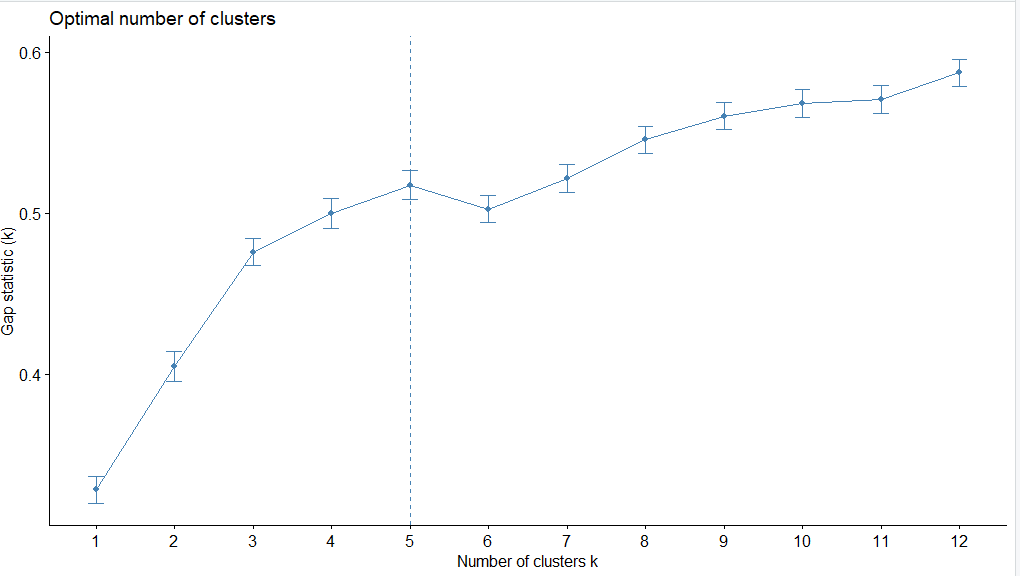


Рисунок 7 – Результат построения статистики разрыва

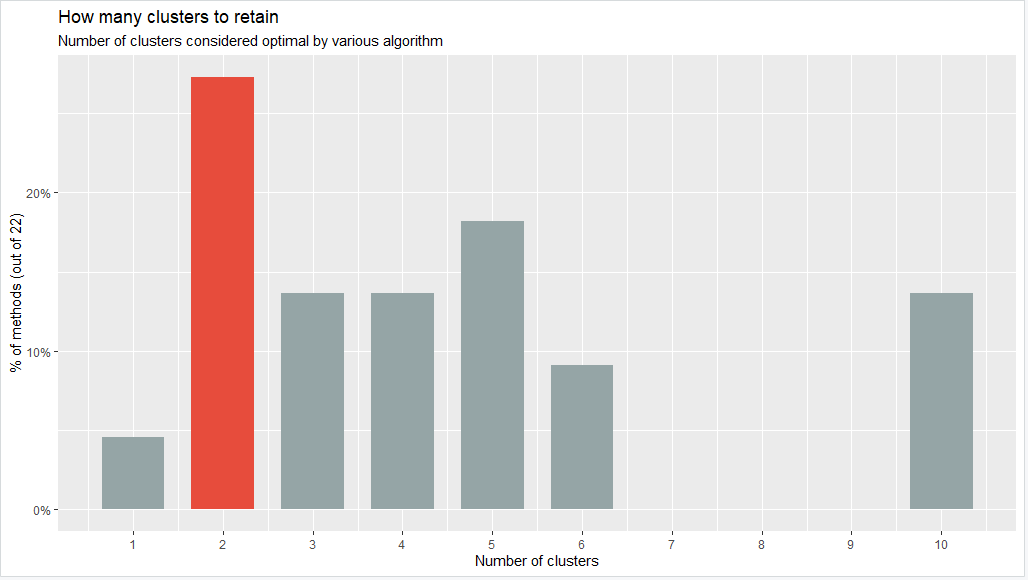


Рисунок 8 – Результат построения консенсуса

Согласно методу локтя оптимальное число кластеров находится в районе 4-5 штук. Аналогично согласно методу силуэта и статистики разрыва оптимальное число кластеров находится в райное 5. Однако согласно алгоритму консенсуса оптимальное число кластеров равно 2. Основываясь на данных статистиках построим кластеризацию по 5 и 2 кластерам и сравним их.

1. Построим иерархические кластеризации по 2 и 5 кластерам.

Код построения:

dist.datas=dist(data\_k)

labels\_datas=data\_sc\_norm[,(length(colnames(data\_sc\_norm)))]

clust.datas=hclust(dist.datas,'ward.D')

plot(clust.datas,labels\_datas,cex=0.5)

rect.hclust(clust.datas,k=5,border="red")

dist.datas=dist(data\_k)

labels\_datas=data\_sc\_norm[,(length(colnames(data\_sc\_norm)))]

clust.datas=hclust(dist.datas,'ward.D')

plot(clust.datas,labels\_datas,cex=0.5)

rect.hclust(clust.datas,k=2,border="red")

Результаты построения иерархической кластеризации представлены в виде деревьев на рисунках 9 и 10.

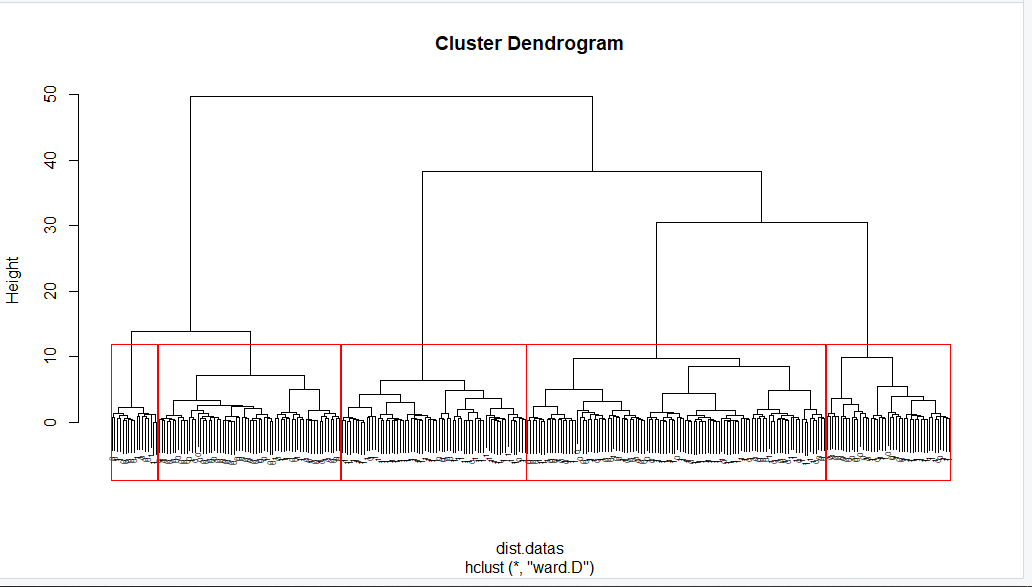


Рисунок 9 – Иерархическая кластеризация по 5 кластерам

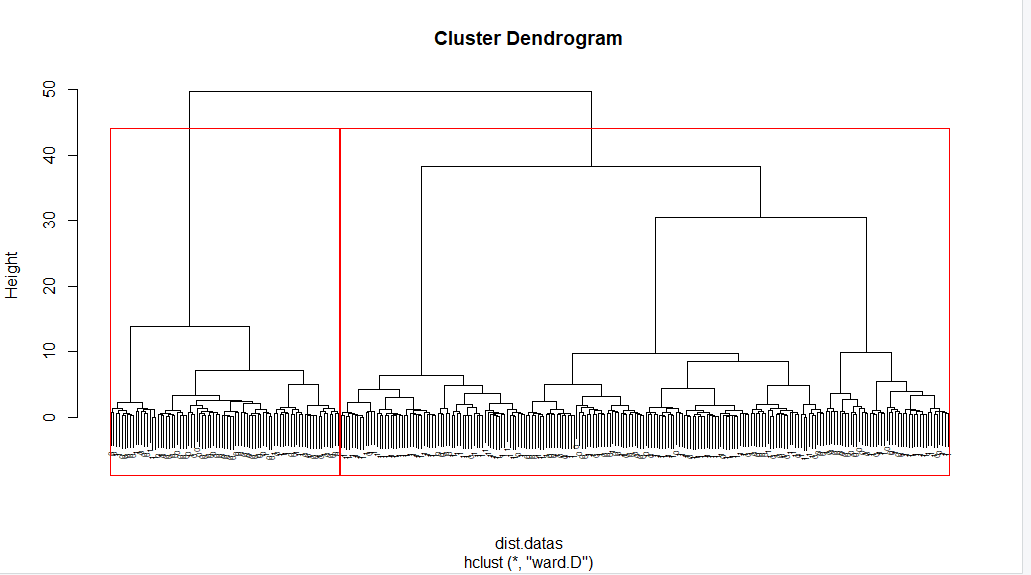


Рисунок 10 ­– Иерархическая кластеризация по 2 кластерам

Если мы выведем представителей данных группы, то можно увидеть, что иерархическая кластеризация в обоих случаях разделяет данные не совсем верно относительно целевой переменной. В кластеры попадают люди, у которых и есть, и отсутствует болезни сердца. Однако заметно, что в обоих разделениях в кластерах каких-то людей все-таки больше. Подобные проблемы с разделениями могут быть связаны в наличии достаточно большего количества данных, имеющих 2 значения: 1 и 0. Пример полученного кластера представлен на рисунке 11:

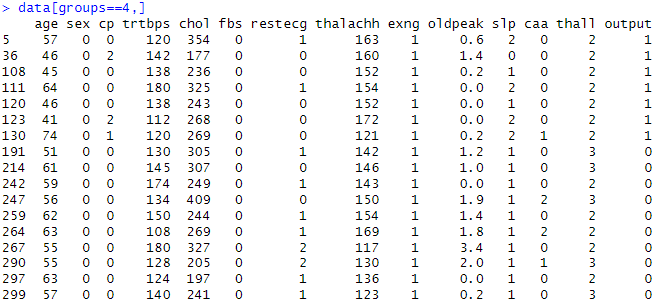


Рисунок 11 – Пример кластера

1. Были построены диаграмма со столбчатыми диаграммами и боксплотами групп. Также был проведен сравнительный анализ.

Результат построения диаграмм для кластеризации из 5 кластеров представлен на рисунке 12.

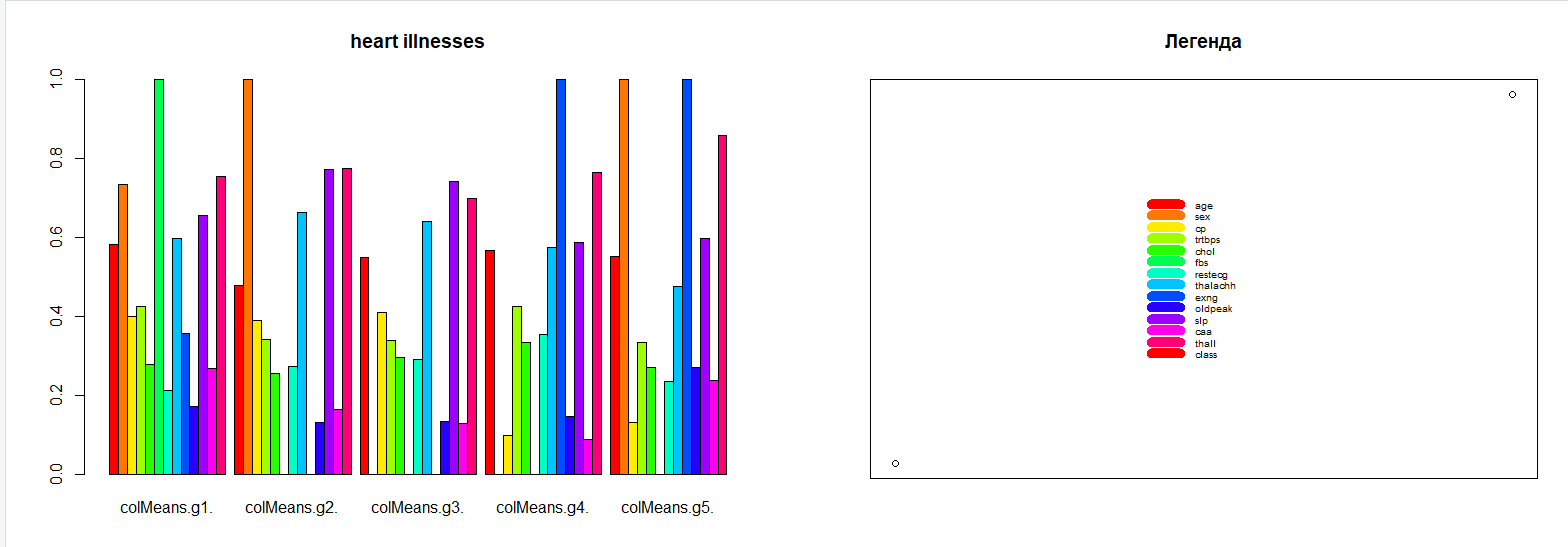


Рисунок 12 – Результат построения столбчатой диаграммы для 5 кластеров

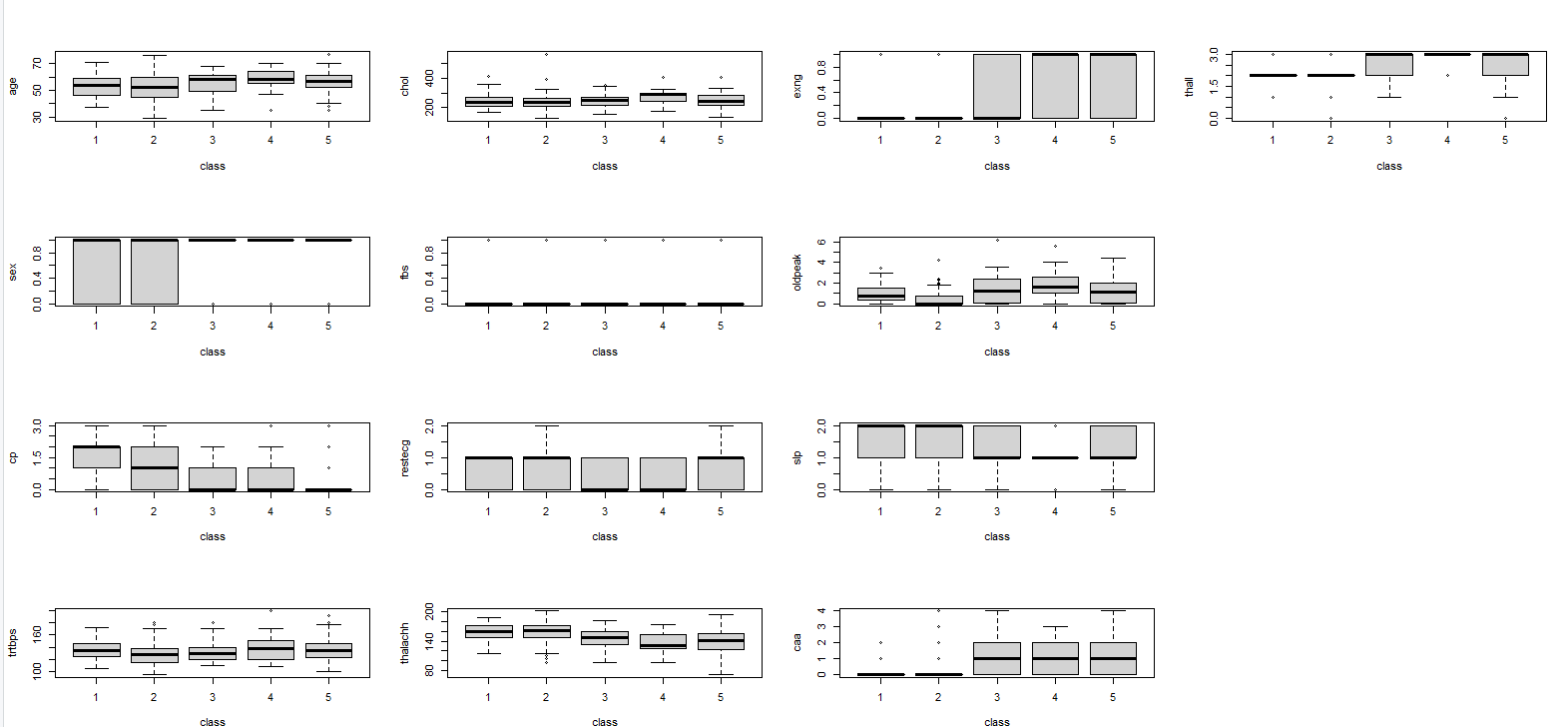


Рисунок 13 – Результат построения боксплотов для кластеризации для 5 кластеров

Согласно столбчатой диаграмме видно, что у каждого из кластеров значение одного или двух параметров сильно превышает значение этих же параметров у других кластеров. Аналогичную ситуацию можно видеть и на боксплоте, где у разных кластеров разбросы значений могут сильно отличаться.

Результат построения данных диаграмм для 2-ух кластеров представлен на рисунках 14-15.

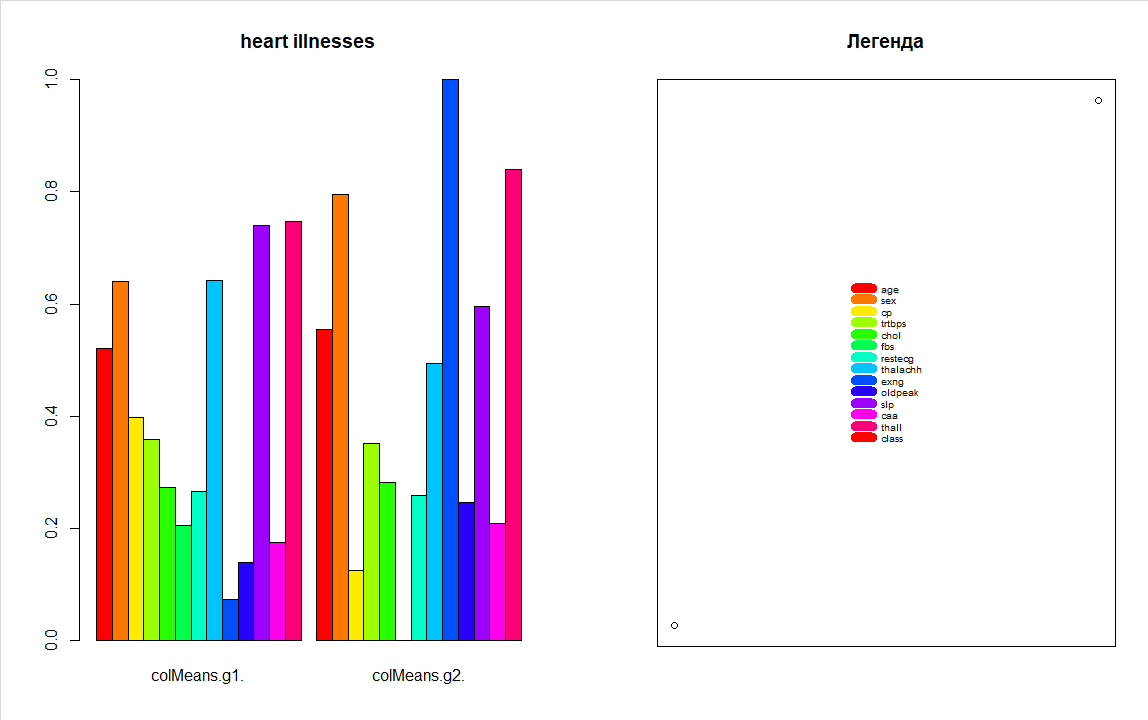


Рисунок 14 – Результат построения столбчатой диаграммы для 2 кластеров

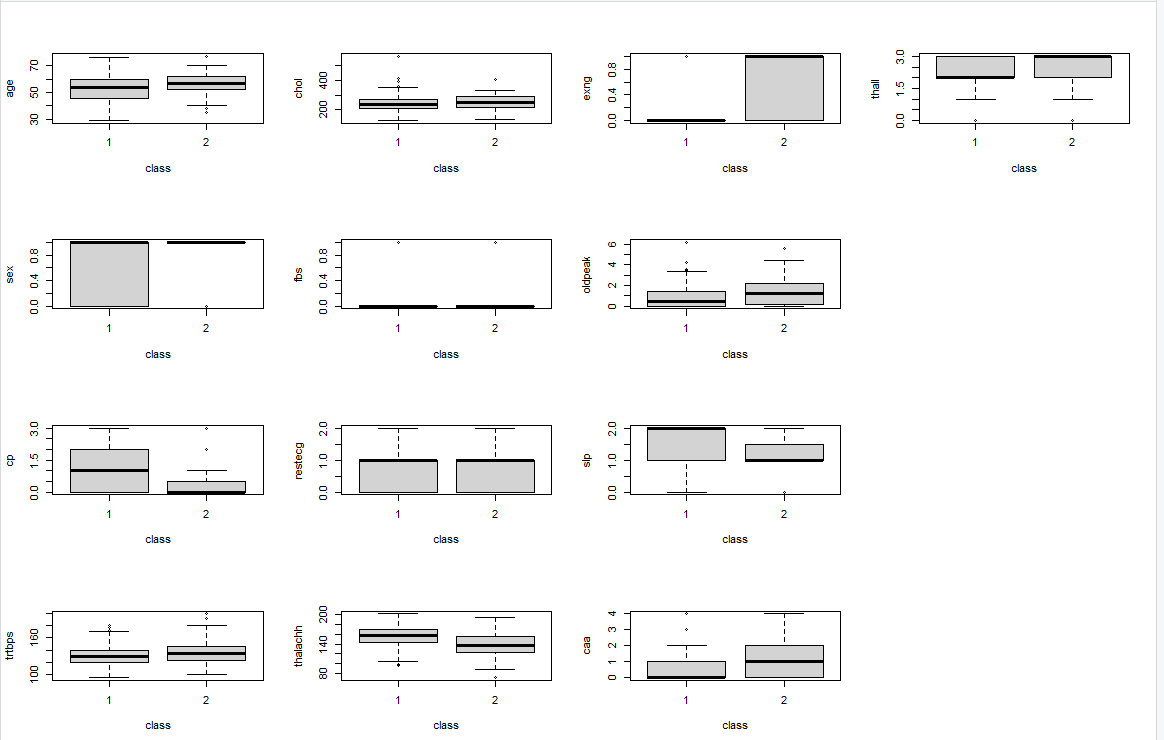


Рисунок 15 – Результат построения боксплота для 2-ух кластеров

На данных диаграммах мы можем увидеть ситуацию аналогичную ситуации для 5 кластеров – имеются параметры, значения которых сильно различаются от кластера к кластеру.

1. Была выполнена кластеризация датасета на 2 и 5 кластеров с использованием алгоритма k-means. Был проведен анализ качества выполнения кластеризации.

Код:

km = kmeans(data\_sc\_norm,5)

fviz\_cluster(km, data\_k,palette="Set2", ggtheme = theme\_minimal())

print(km)

data[km$cluster == 1,]

data[km$cluster == 2,]

data[km$cluster == 3,]

data[km$cluster == 4,]

data[km$cluster == 5,]

km = kmeans(data\_sc\_norm,2)

fviz\_cluster(km, data\_k,palette="Set2", ggtheme = theme\_minimal())

print(km)

data[km$cluster == 1,]

data[km$cluster == 2,]

Результат выполнения кластеризации представлен на рисунках 16 и 17.

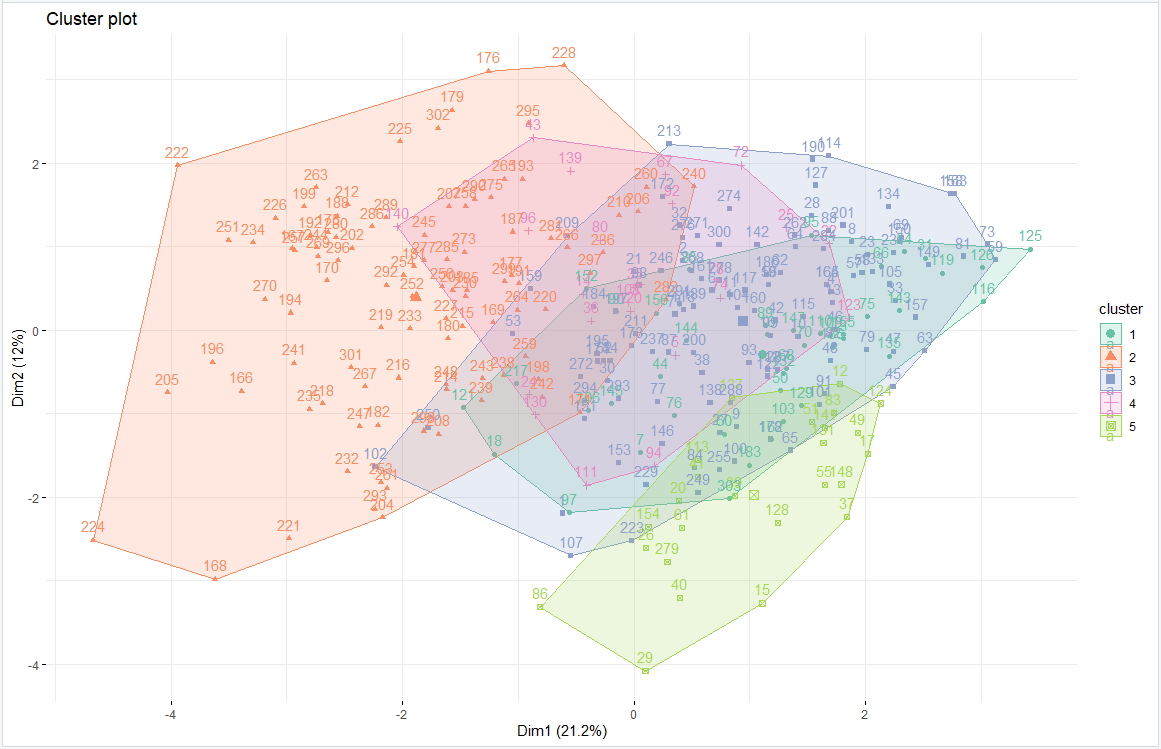


Рисунок 16 – Результат выполнения кластеризации на 5 кластеров

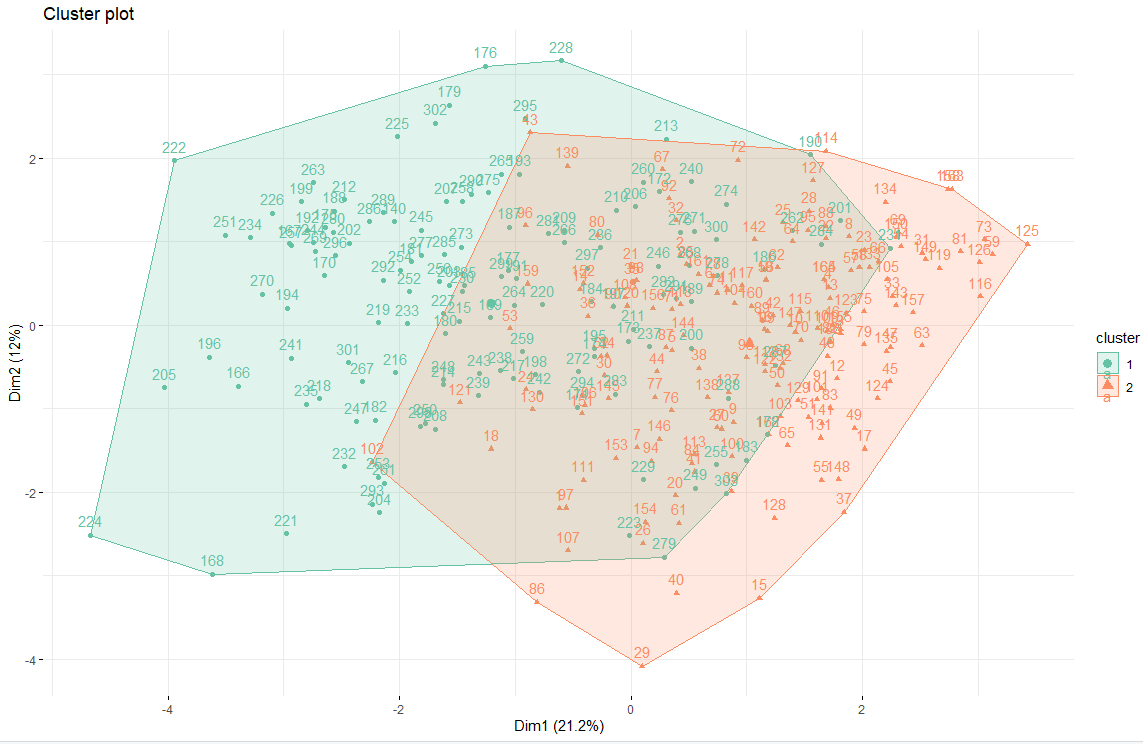
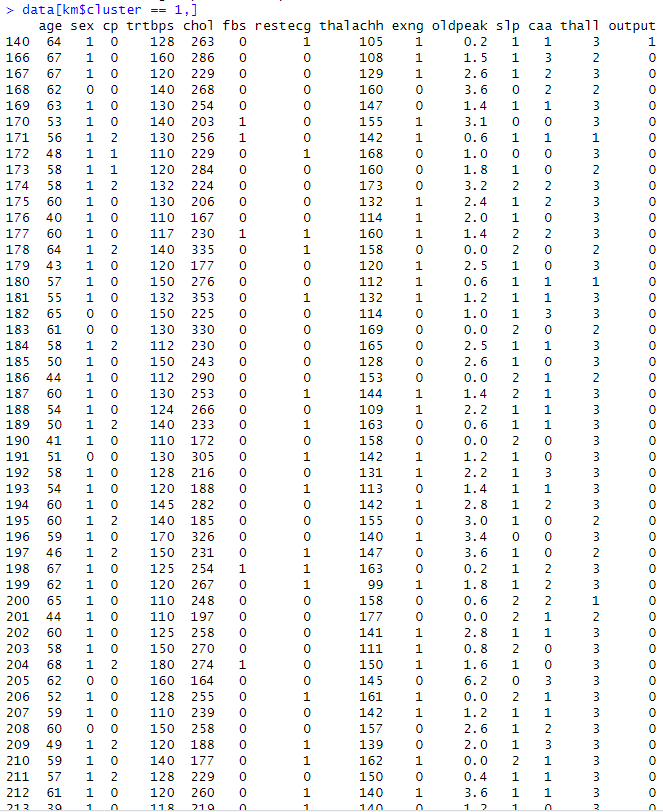
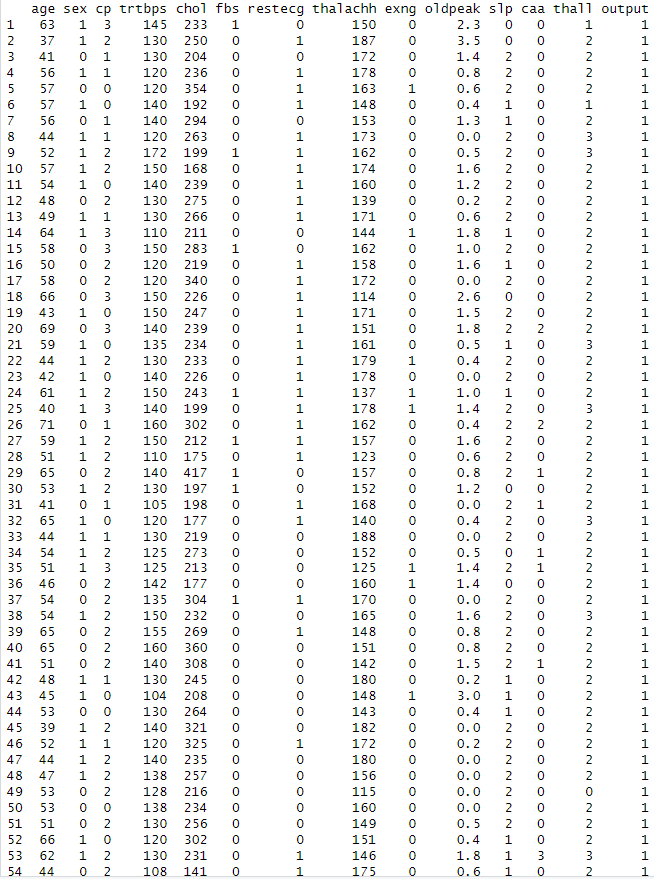


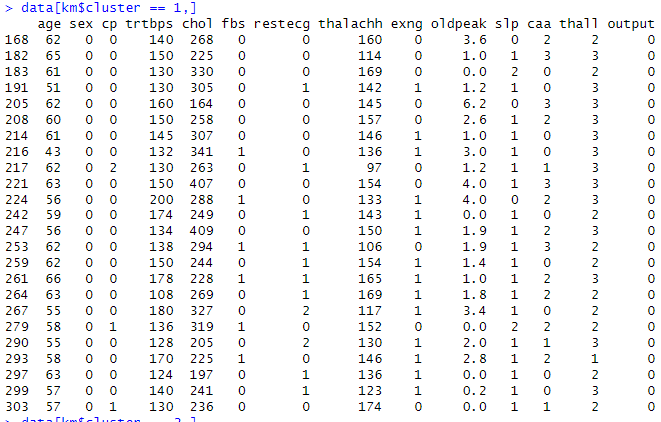
Рисунок 17 – Результат выполнения кластеризации на 2 кластеров

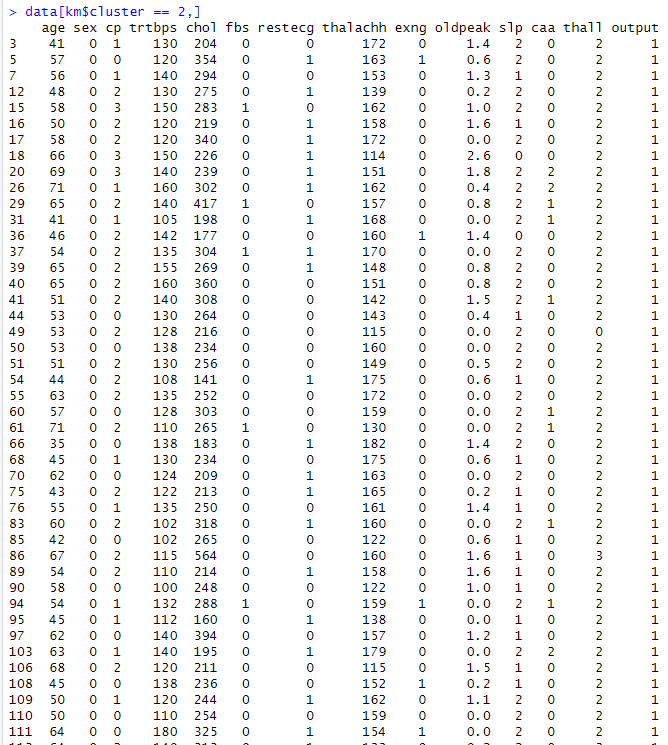
Как видно по данному изображения при разделении набора данных на 5 кластеров наблюдается множество пересечений и вхождений кластеров друг в друга. При разделении же на 2 кластера видно более четкое разделение с меньшими пересечениями. Данное утверждение подтверждается выводом сформировавшихся групп. При разделении на 2 кластера мы можем видеть, что в первый кластер попала лишь 1 ошибочная запись, при разделении на 5 кластеров в последний кластер попало 4 ошибочных записей. Хотя ошибка относительного общего количества не велика, можно предположить, что разделение на 2 кластера может оказаться немного более точным.

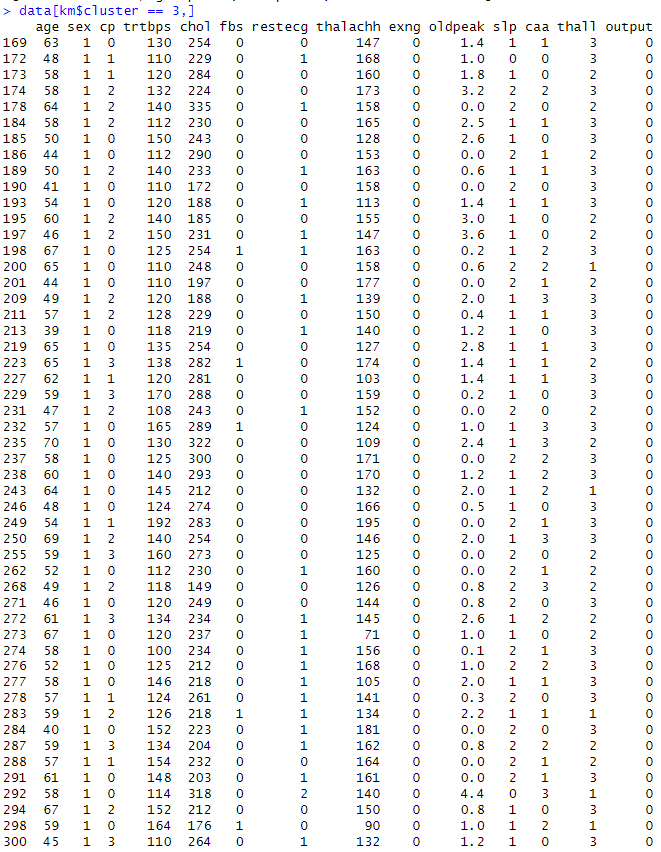


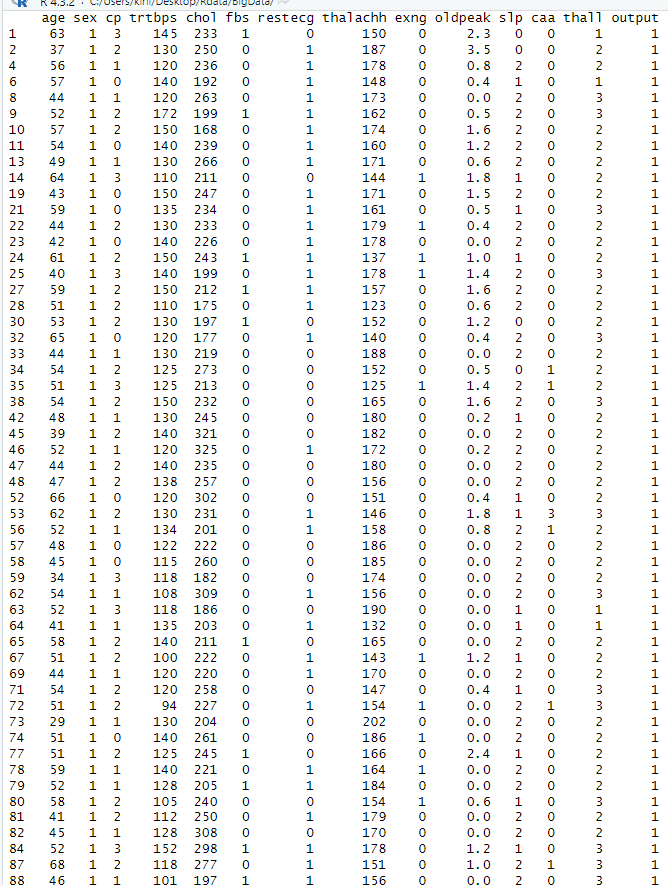


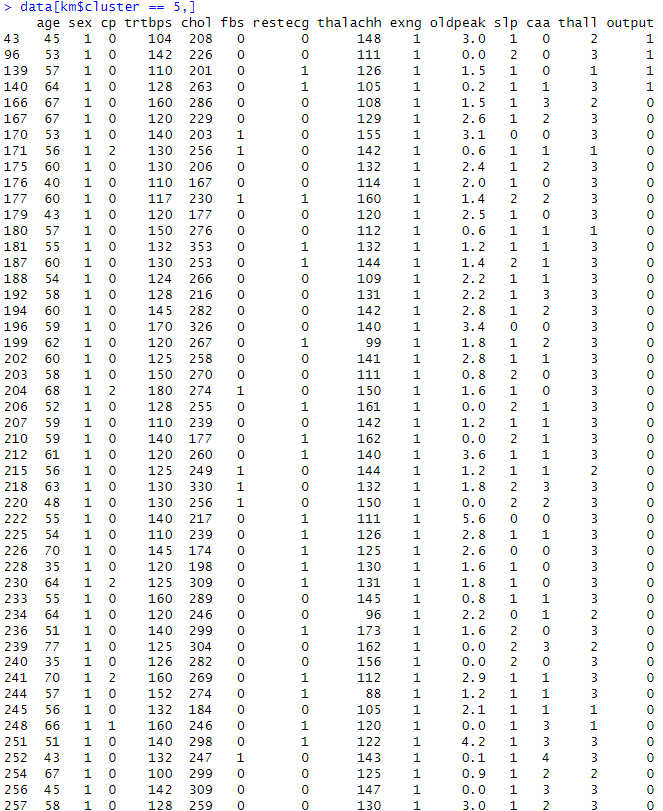
Рисунки 17-18 – Результаты для 2-ух кластеров











Рисунки 19-23 – Результаты разделения на 5 кластеров

1. Был построен scatterplot для набора данных, включая построение зависимостей для уже разделенных данных. Как видно на рисунках 24-26 данные при кластеризации на 2 кластера кучкуются более однородно, чем при 5 кластерах.

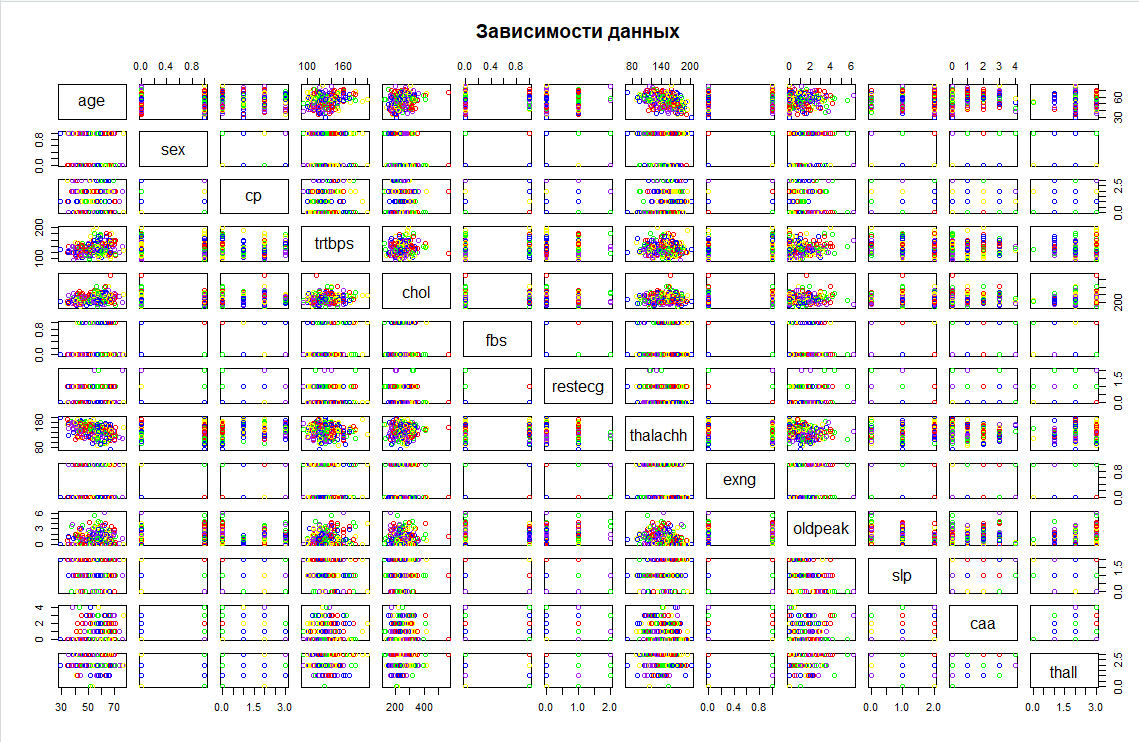


Рисунок 24 – Расположение данных при их попарных зависимостях

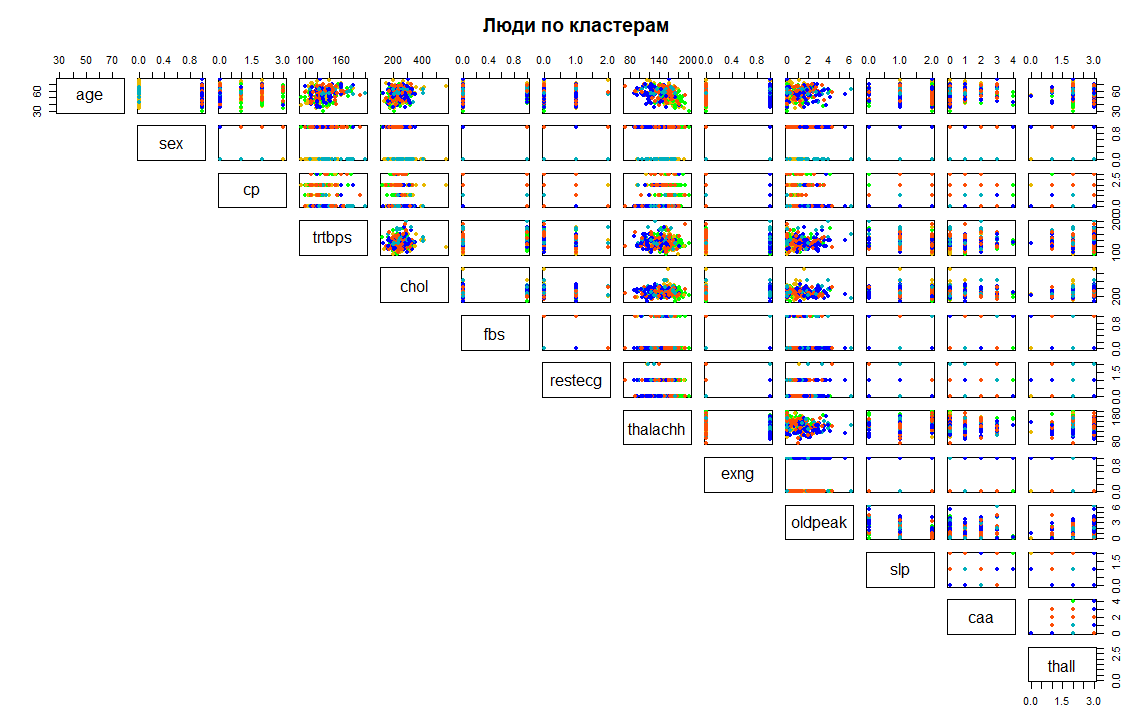


Рисунок 25 – Расположение данных при их попарных зависимостях для 5 кластеров

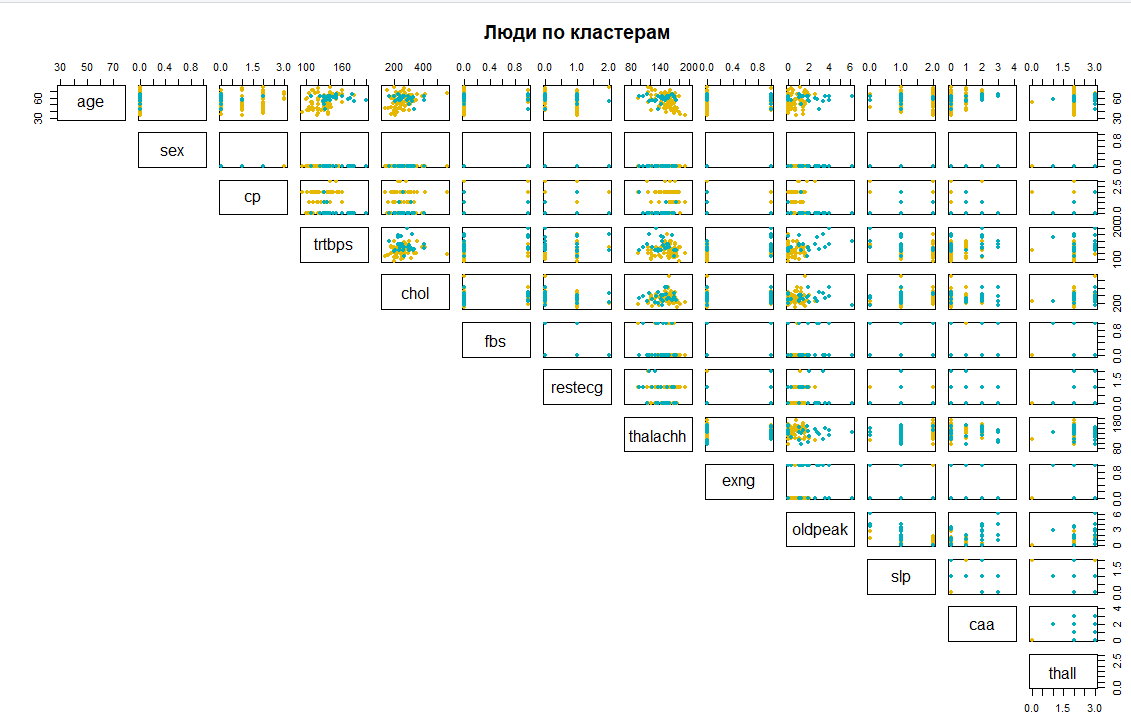


Рисунок 26 – Расположение данных при их попарных зависимостях для 2 кластеров

1. Была построена трехмерную кластеризацию по scatterplot3d для кластеризации по 2 и 5 кластерам. Результаты построения представлены на рисунках 27-28. Как видно данные располагаются на передней и задней гранях фигуры.

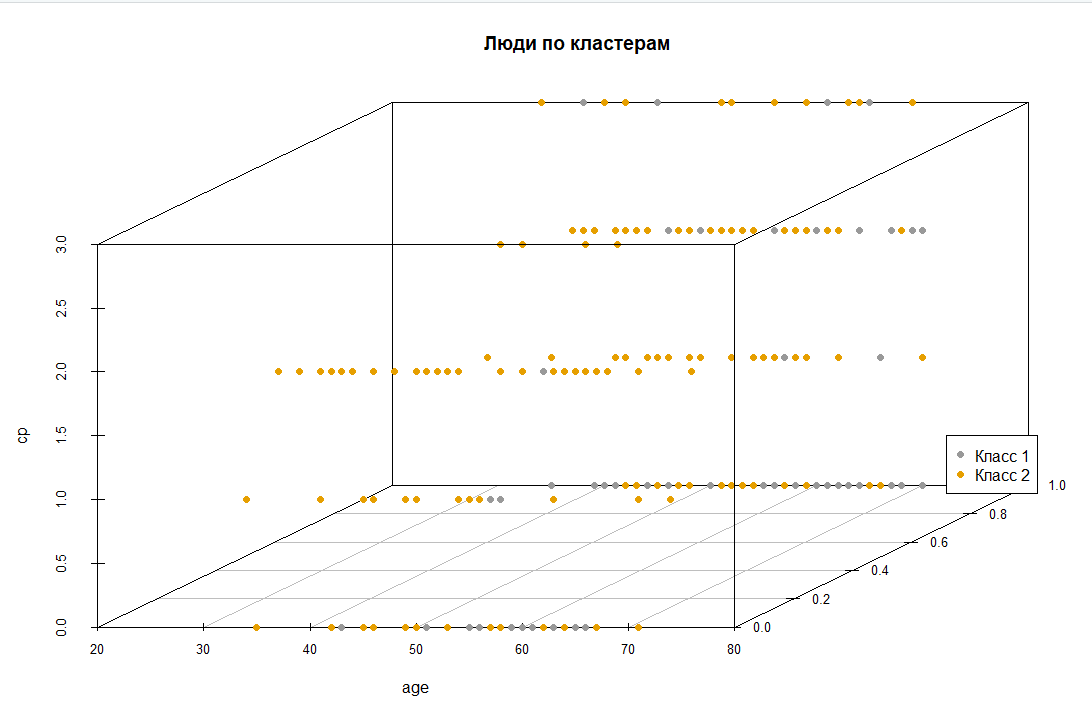


Рисунок 27 – 3d отображения кластеризации по 2-ум кластерам

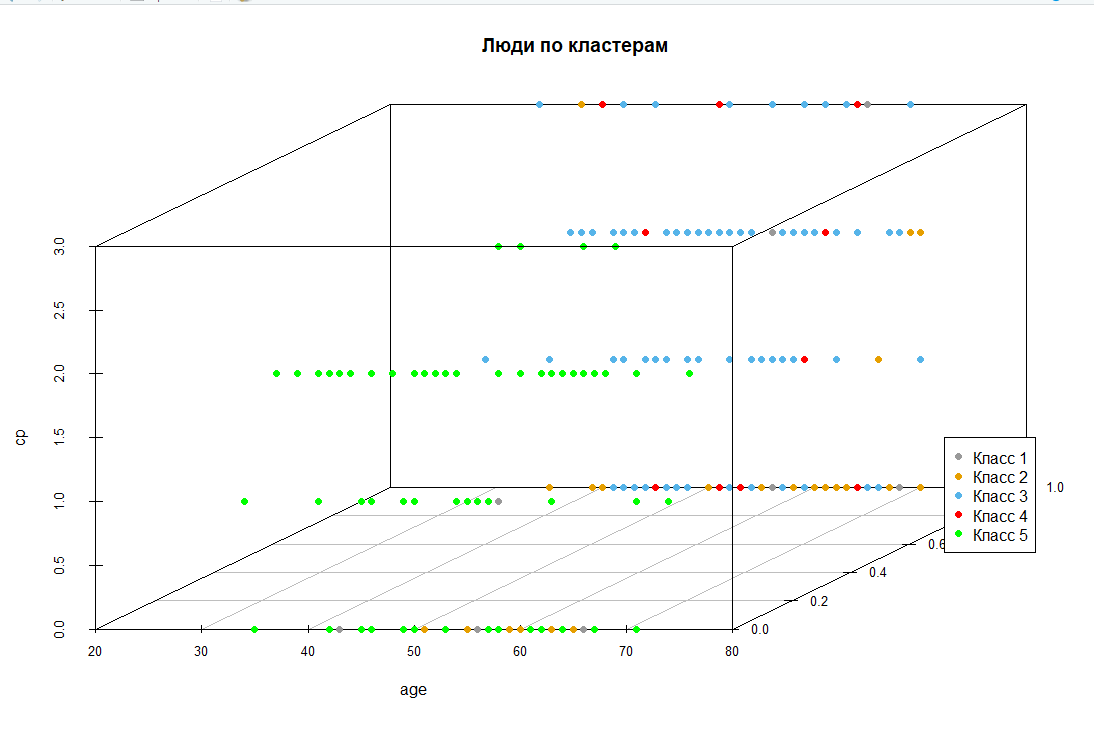


Рисунок 28 – 3d отображения кластеризации по 5 кластерам

1. Были добавлены найденные при в результате кластерного анализа классы (groups), как вектор-столбец в предыдущий DATASET. Набор данных был разделен на две части (обучающую и тестовую).

Код:

#Добавим к набору данных найденные кластеры

#Для 2 кластеров

rm(list=ls())

data\_2\_new = data.frame(cbind(data\_sc\_norm[,-14],factor(km2$cluster),data\_sc\_norm[,14]))

colnames(data\_2\_new)[c(14,15)]=c("groups","output")

data\_2\_new$groups

data\_2\_new$groups = scale(data\_2\_new$groups,center=min(data\_2\_new$groups),scale=max(data\_2\_new$groups)-min(data\_2\_new$groups))

data\_2\_new[,14]=factor(data\_2\_new[,14]/(max(data\_2\_new[,14])-min(data\_2\_new[,14])))

data\_2\_new[,15]=factor(data\_2\_new[,15])

data\_2\_new[,14]

#Для 5 кластеров

data\_5\_new = data.frame(cbind(data\_sc\_norm[,-14],factor(km5$cluster),data\_sc\_norm[,14]))

colnames(data\_5\_new)[c(14,15)]=c("groups","output")

data\_5\_new$groups

data\_5\_new$groups = scale(data\_5\_new$groups,center=min(data\_5\_new$groups),scale=max(data\_5\_new$groups)-min(data\_5\_new$groups))

data\_5\_new[,14]=factor(data\_5\_new[,14]/(max(data\_5\_new[,14])-min(data\_5\_new[,14])))

data\_5\_new[,15]=factor(data\_5\_new[,15])

data\_5\_new[,14]

data\_5\_new

#Формируем тренировочные и тестовые данные

set.seed(1234)

ind = sample(2,nrow(data\_2\_new),replace=TRUE,prob=c(0.7,0.3))

#Для 2 кластеров

trainData2=data.frame(data\_2\_new[ind==1,])

testData2=data\_2\_new[ind==2,]

nrow(testData2)

nrow(trainData2)

nrow(data\_2\_new)

#Для 5 кластеров

trainData5=data\_5\_new[ind==1,]

testData5=data\_5\_new[ind==2,]

nrow(testData5)

nrow(trainData5)

nrow(data\_5\_new)

1. Для наивного Байесовского классификатора: была решена задача с помощью наивного Байесовского классификатора с и без учета групп, полученных при кластеризации. Была проанализирована точность полученных решений для тестовых данных.

Код:

#Наивный байес

#для 2

library(klaR)

naive\_heart\_2 <- NaiveBayes(trainData2$output ~ ., data = trainData2)

naive\_heart\_2$tables

summary(trainData2)

opar=par()

layout(matrix(1:16, 4, 4, byrow = TRUE))

plot(naive\_heart\_2,lwd = 2)

legend("topleft",legend=c("yes","no"),lty=1:2,cex=0.5)

data

summary(data)

pred <- predict(naive\_heart\_2, testData2[,-15])$class

(table(Факт = testData2$output, Прогноз = pred))

naive\_heart\_2 <- NaiveBayes(trainData2$output ~ ., data = trainData2[,-14])

naive\_heart\_2$tables

opar=par()

layout(matrix(1:16, 4, 4, byrow = TRUE))

plot(naive\_heart\_2,lwd = 2)

legend("topleft",legend=c("yes","no"),lty=1:2,cex=0.5)

pred <- predict(naive\_heart\_2, testData2[,c(-14,-15)])$class

(table(Факт = testData2$output, Прогноз = pred))

#для 5

naive\_heart\_5 <- NaiveBayes(trainData5$output ~ ., data = trainData5)

naive\_heart\_5$tables

summary(trainData5)

opar=par()

layout(matrix(1:16, 4, 4, byrow = TRUE))

plot(naive\_heart\_5,lwd = 2)

legend("topleft",legend=c("yes","no"),lty=1:2,cex=0.5)

data

summary(data)

pred <- predict(naive\_heart\_5, testData5[,-15])$class

(table(Факт = testData5$output, Прогноз = pred))

naive\_heart\_5 <- NaiveBayes(trainData5$output ~ ., data = trainData5[,-14])

naive\_heart\_5$tables

opar=par()

layout(matrix(1:16, 4, 4, byrow = TRUE))

plot(naive\_heart\_5,lwd = 2)

legend("topleft",legend=c("yes","no"),lty=1:2,cex=0.5)

pred <- predict(naive\_heart\_5, testData5[,c(-14,-15)])$class

(table(Факт = testData5$output, Прогноз = pred))

Результаты построения классификатора для 2 кластеров представлены на рисунках 29,30,31,32. Как видно по данным рисункам наличие групп, полученных с помощью кластеризации, увеличивает точность классификатора, поскольку появляется признак еще больше различающий объекты разных классов. Также видно, что в пределах одного класса имеется как минимум 1 признак сильно выбивающийся по значению по сравнению с этим же значением у другого класса.

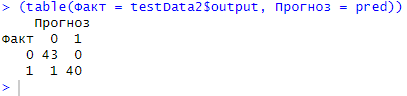


Рисунок 29 – Результаты тестирования для классификатора с группами

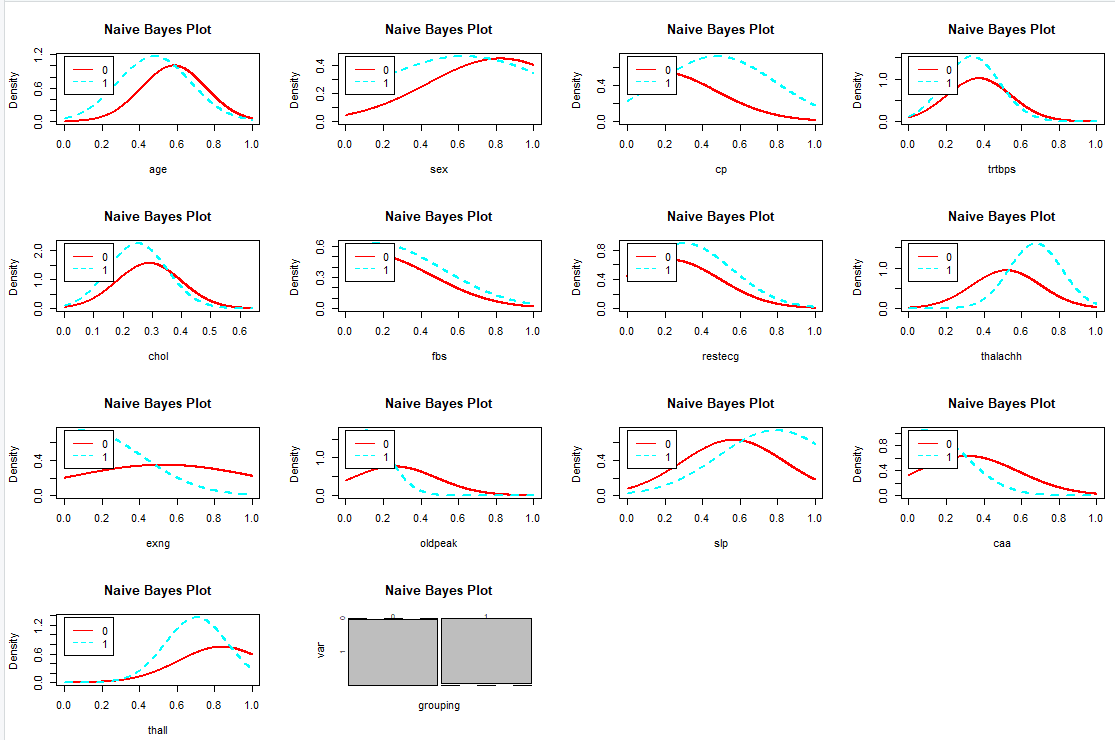


Рисунок 30 – Результат распределения признаков по классам

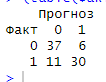


Рисунок 31 – Результаты тестирования классификатора без групп

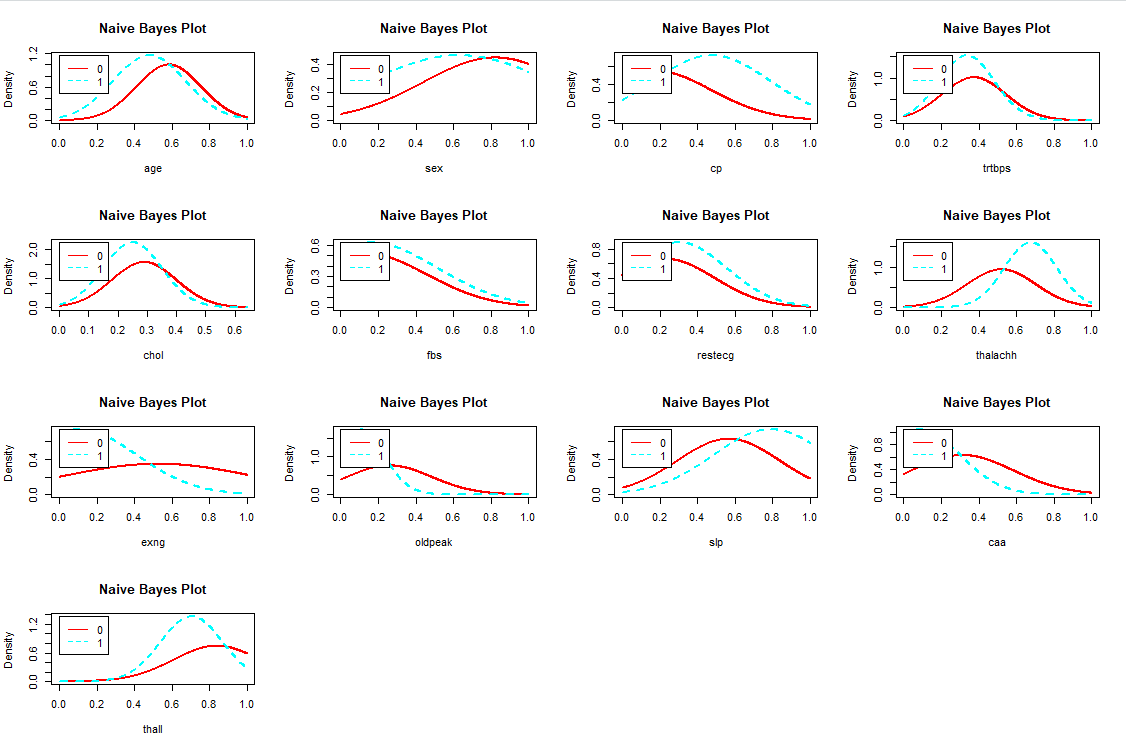


Рисунок 32 – Результаты распределения признаков по классам без групп

Результаты построения для 5 кластеров представлены на рисунках 33,34. Как видно кластеризация на 5 групп оказалась немного хуже по точности, чем кластеризация на 2 группы. Распределение же признаков осталось примерно таким же.



Рисунок 33 – Результаты тестирования для 5 кластеров с группами

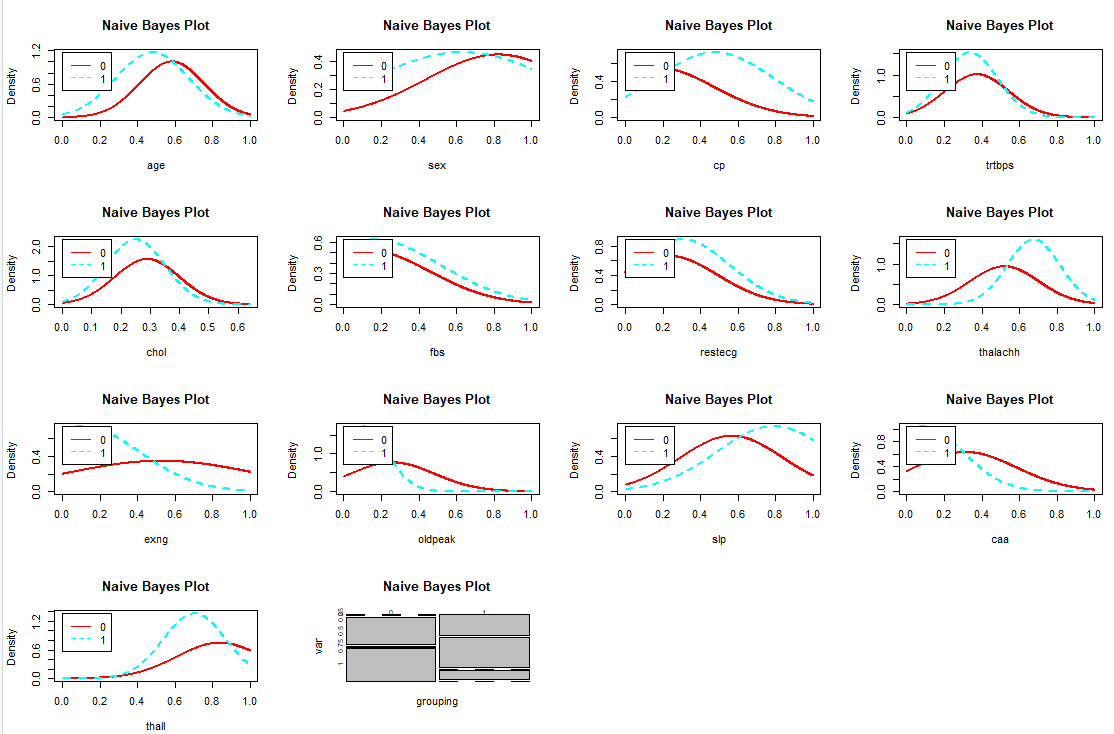


Рисунок 34 – Результат распределения признаков по классам

1. Для деревьев решений: был применен метод деревьев решений для задачи классификации (для того же набора данных); было исследовано дерево решений и построен его график. Проанализирована точность полученных решений для тестовых данных (с известным значением переменной отклика), сравнен результаты с ранее полученными.

Код построения:

library(party)

heart\_2\_tree=ctree(trainData2$output~.,data=trainData2)

heart\_2\_tree

pred <- predict(heart\_2\_tree, testData2[,-15])

table(pred, testData2[,15])

plot(heart\_2\_tree,text = "vertical", ymax = 1)

#Аналогично но без групп

heart\_2\_tree=ctree(trainData2$output~.,data=trainData2[,-14])

heart\_2\_tree

pred <- predict(heart\_2\_tree, testData2[,c(-14,-15)])

table(pred, testData2[,15])

plot(heart\_2\_tree,text = "vertical", ymax = 1)

#Деревья для 5

library(party)

heart\_5\_tree=ctree(trainData5$output~.,data=trainData5)

heart\_5\_tree

pred <- predict(heart\_5\_tree, testData5[,-15])

table(pred, testData5[,15])

plot(heart\_5\_tree,text = "vertical", ymax = 1)

#Аналогично но без групп

heart\_5\_tree=ctree(trainData5$output~.,data=trainData5[,-14])

heart\_5\_tree

pred <- predict(heart\_5\_tree, testData5[,c(-14,-15)])

table(pred, testData5[,15])

plot(heart\_5\_tree,text = "vertical", ymax = 1)

Результаты построения дерева представлены на рисунках 35,36,37,38. Как видно дерево без учета кластеров работает намного хуже, поскольку номер кластера предоставляет признак несущий много информации об итоговом классе объекта. Также видно, что без учета кластеров дерево работает хуже, чем байес.



Рисунок 35 – Результат тестирования для дерева по 2 группам

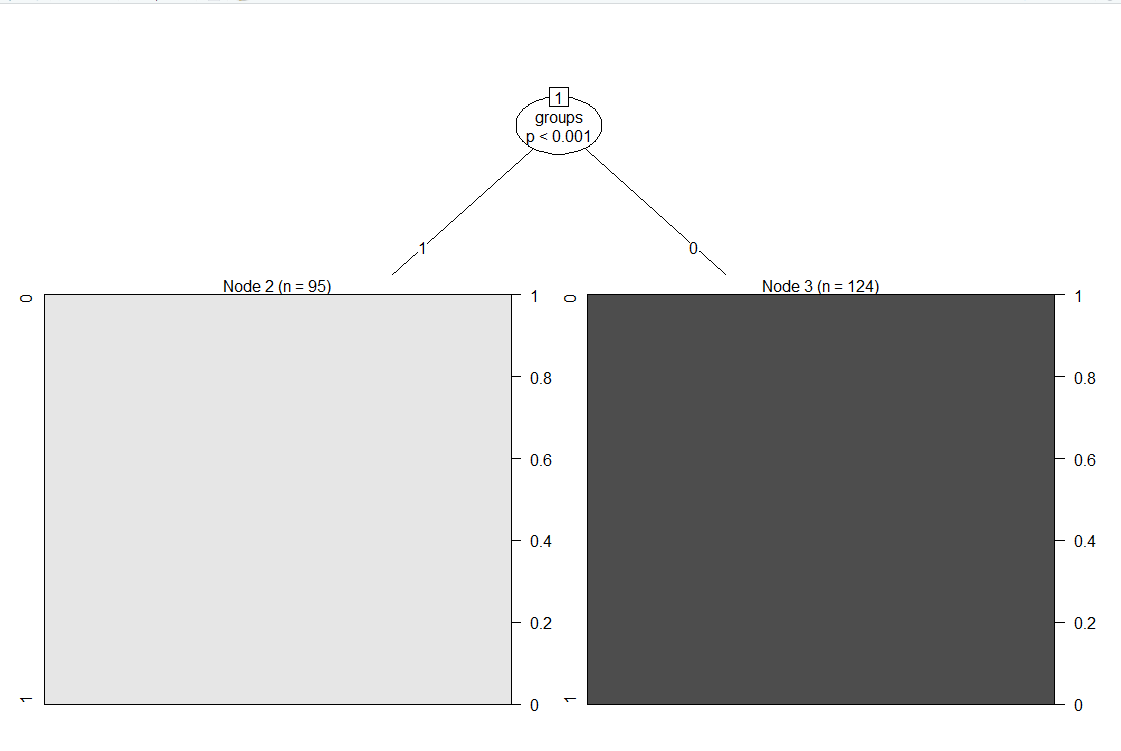


Рисунок 36 – Дерево для 2 групп



Рисунок 37 – Результат тестирования без учета кластеров

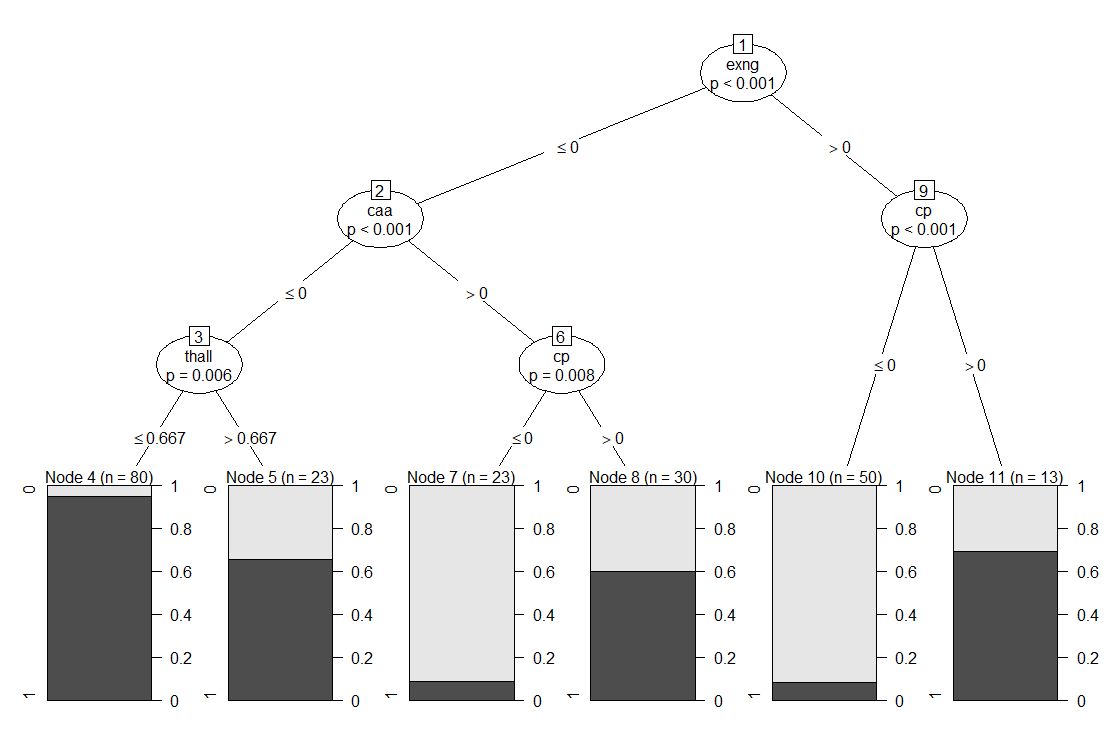


Рисунок 38 – Дерево без учета кластеров

Результаты построения для 5 кластеров представлены на рисунках 39,40. Согласно рисунку видно, что дерево для 5 групп работает хуже, чем для 2 групп. Однако также видно, что оно использовало больше признаков для построения классификации, чем дерево на 2 группах.



Рисунок 39 – Результат тестирования для 5 групп

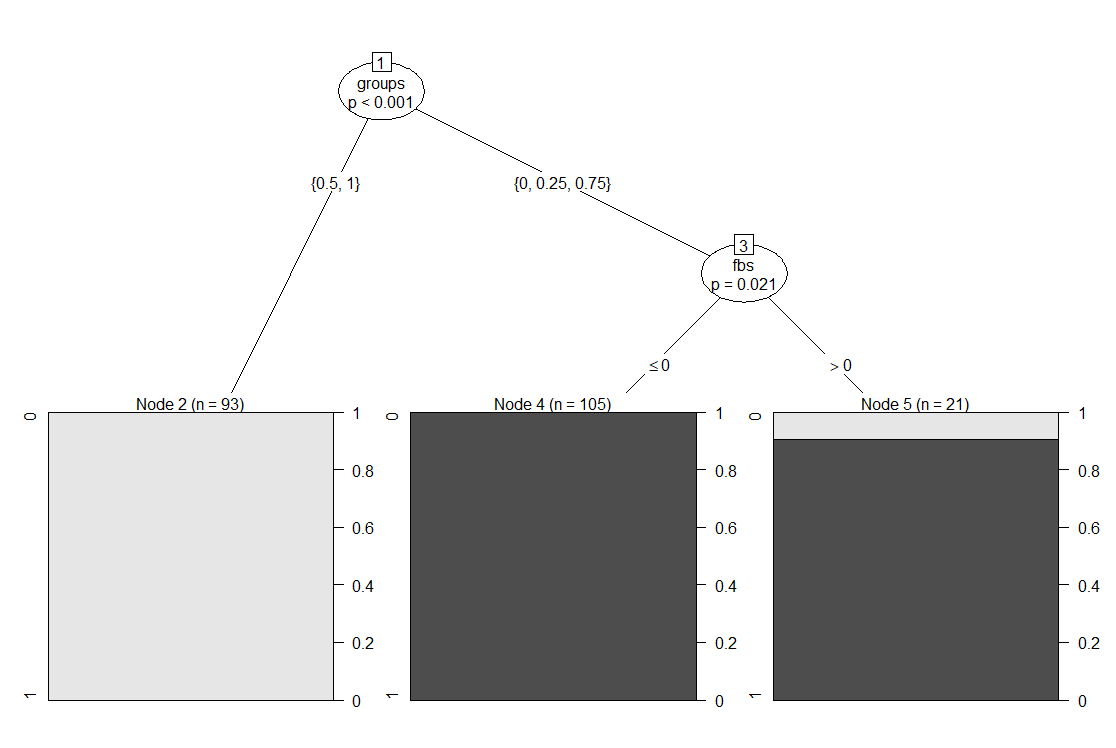


Рисунок 40 – Дерево для 5 групп

1. Была выполнена классификация с помощью случайного леса, сопоставлен результат с результатом дерева решения, прокомментирован результат сравнения. Были сопоставлены результаты с результатами Байесовского классификатора.

Код:

#Случайный лес

library(randomForest)

#Для 2

rf=randomForest(trainData2$output ~ .,data=trainData2, ntree=100, proximity=TRUE)

table(predict(rf,testData2[,-15]), testData2$output)

rf=randomForest(trainData2$output ~ .,data=trainData2[,-14], ntree=100, proximity=TRUE)

table(predict(rf,testData2[,c(-14,-15)]), testData2$output)

#Для 5

rf=randomForest(trainData5$output ~ .,data=trainData5, ntree=100, proximity=TRUE)

table(predict(rf,testData5[,-15]), testData5$output)

rf=randomForest(trainData5$output ~ .,data=trainData5[,-14], ntree=100, proximity=TRUE)

table(predict(rf,testData5[,c(-14,-15)]), testData5$output)

Результаты построения классифкатора представлены на рисунках 41,42,43. Согласно данным рисункам можно сделать вывод, что классификатор на 2 группах работает лучше всего. Также видно, что без учета групп классификатор работает лучше, чем Байес без групп.



Рисунок 41 – Результат построения для 2 групп



Рисунок 42 – Результат построения без учета групп



Рисунок 43 – Результат построения для 5 групп

**Вывод:** Были закреплены знания, об алгоритмах классификации и кластеризации данных. Были получены навыки визуализирования результатов работы функций кластерного анализа и классификаторов, интерпретации полученные результаты