Лабораторная работа №5 «Кластеризация»

Работу выполнила студентка группы 5140201/30301 Фазылова Алика

Задание 1.

Разбейте множество объектов из набора данных pluton в пакете «cluster» на 3 кластера методом центров тяжести (kmeans). Сравните качество разбиения в зависимости от максимального числа итераций алгоритма.

Решение

Датасет pluton содержит 45 объектов с четырьмя признаками: Pu238, Pu239, Pu240 и Pu241. Проведем кластеризацию, разбив данные на 3 кластера методом центров тяжести (kmeans) (рисунки 1-3). Проведем процедуру для разного количества итераций. Квадраты на графике являются результатами кластеризации по двум признакам.

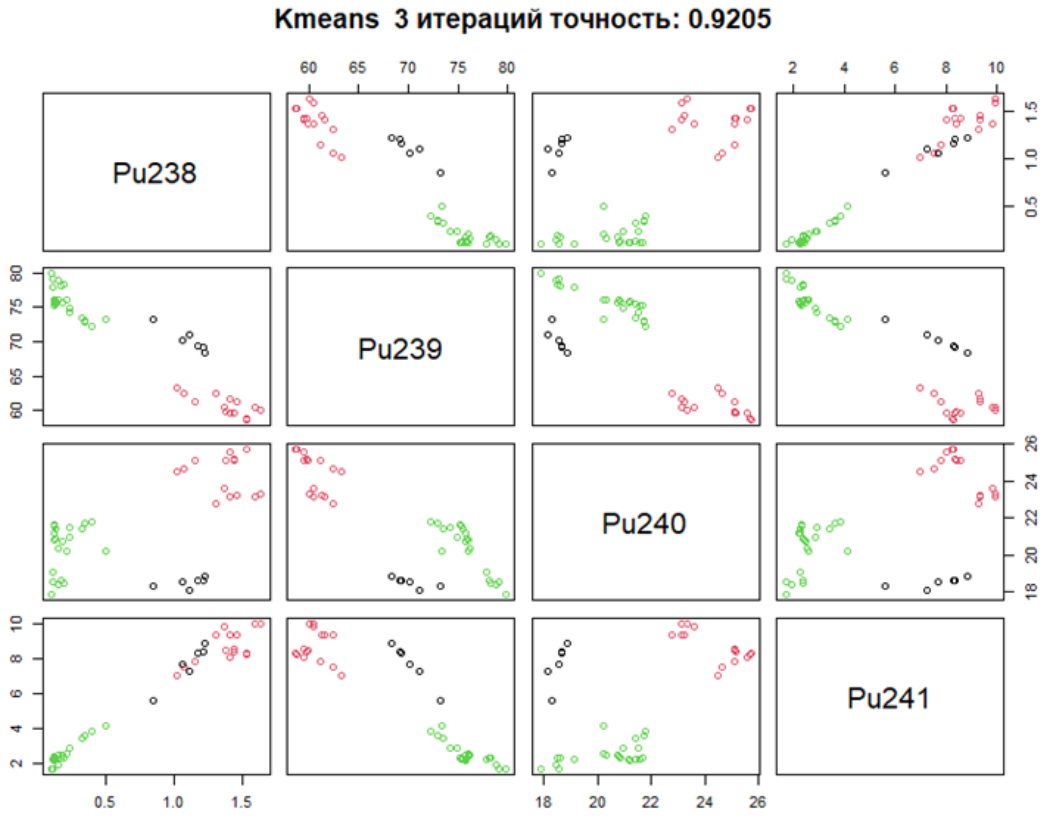

Рисунок 1 – Кластеризация 3 итерации

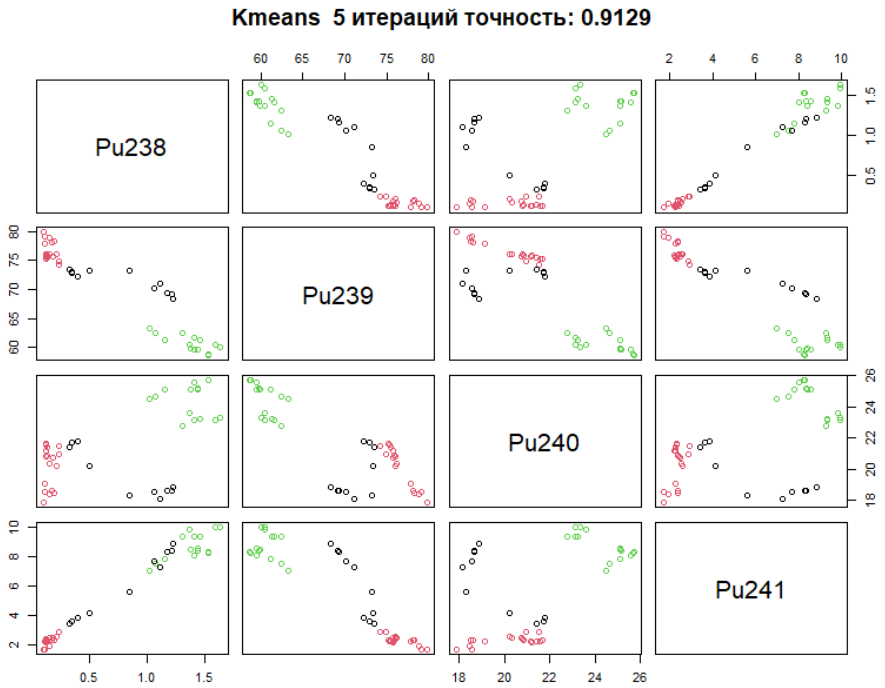

Рисунок 2 – Кластеризация 5 итераций

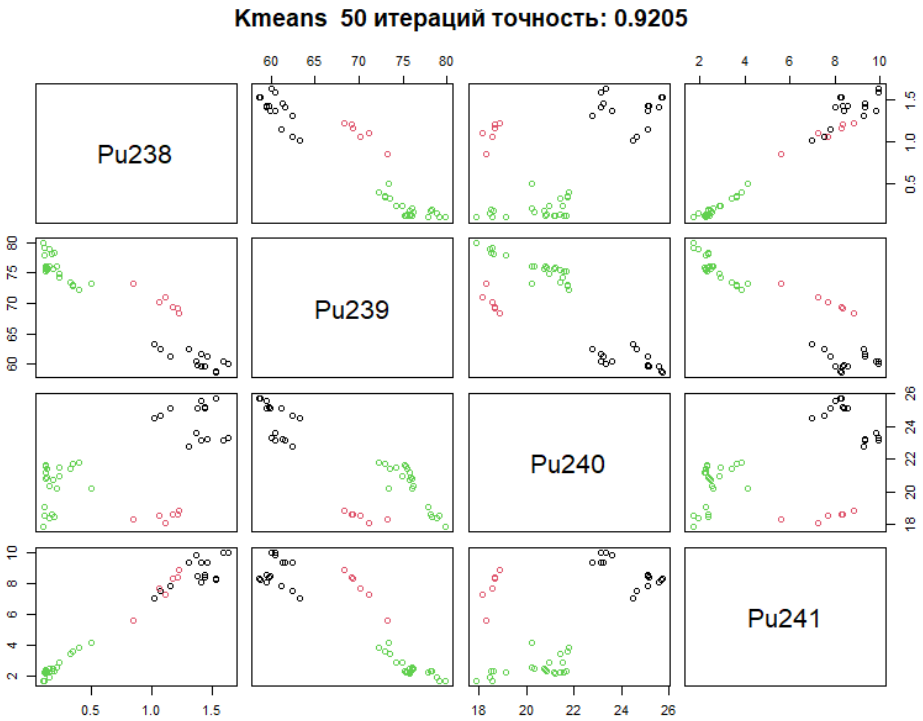

Рисунок 3 – Кластеризация 50 итераций

От изменения количества итераций точность модели сильно не изменяется, варьируясь в значениях 0.9129-0.9205. Так что для данного датасета достаточно будет трех итераций (точность 0.9205)

Листинг кода 1 задачи:

```
library(cluster)
data(pluton)
iterations <- c(3, 5, 10,50, 100)
for (i in iterations) {
  cl <- kmeans(pluton, iter.max = i, centers = 3)
  totss <- cl$totss
  betweenss <- cl$betweenss
  accuracy <- round(betweenss / totss,4)
  plot(pluton,
       col = cl$cluster,
       main = paste("Kmeans ", i, "итераций", 'точность:',accuracy))
}
```

Задание 2.

Сгенерируйте набор данных в двумерном пространстве, состоящий из 3 кластеров, каждый из которых сильно “вытянут” вдоль одной из осей. Исследуйте качество кластеризации методом clara в зависимости от 1) использования стандартизации; 2) типа метрики. Объясните полученные результаты.

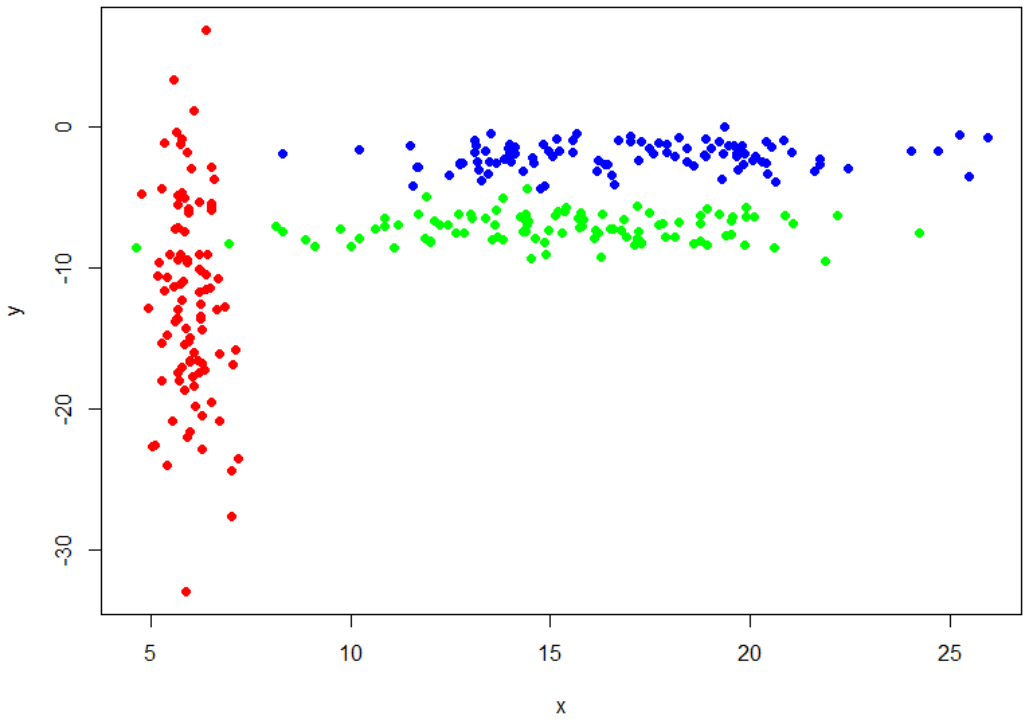

Рисунок 4 – Исходные данные

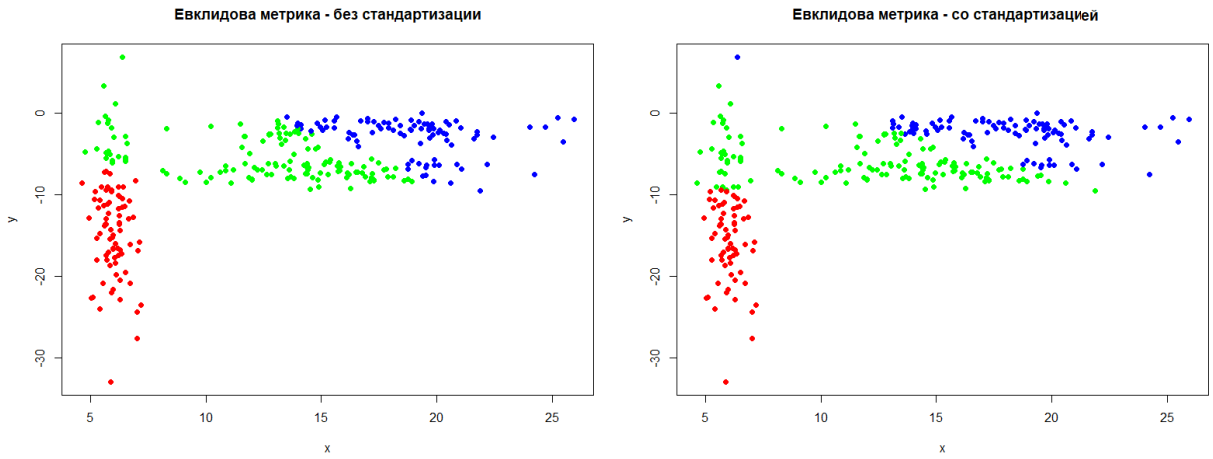

Рисунок 5 – Евклидова метрика

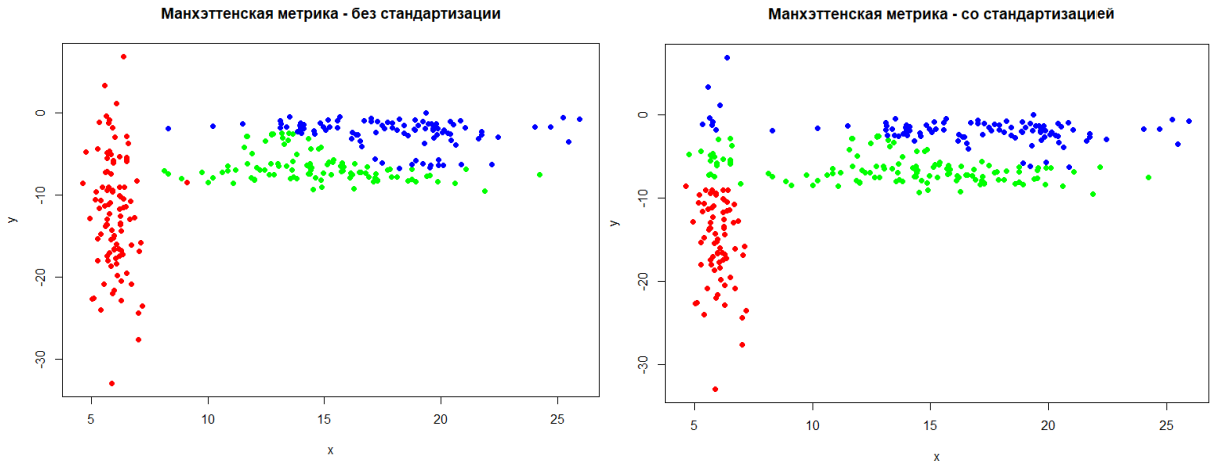

Рисунок 6 – Манхэттенская метрика

Из графиков можно сделать вывод, что самая оптимальная модель с Манхэттенской метрикой и без стандартизации.

Листинг кода 2 задачи:

```
library(cluster)

col = c(rep("#FF0000", 100), rep("#00FF00", 100), rep("#0000FF", 100))
data2 <- data.frame(x = c(rnorm(100, 6, 0.5), rnorm(100, 15, 3.5), rnorm(100, 17, 3.5)), y =
c(rnorm(100, -13, 7), rnorm(100, -7, 1), rnorm(100, -2, 1)))
data2 <- cbind(data2, col)

plot(data2$x, data2$y, xlab="x", ylab="y", col = data2$col, pch = 19)

model = clara(data2[, 1:2], k = 3, metric = "euclidean", stand = FALSE)
colors <- c("#FF0000", "#00FF00", "#0000FF")
plot(data2[, 1:2], col = colors[model$clustering], xlab = "x", ylab = "y", main = "Евклидова
метрика - без стандартизации", pch = 19)
```

```
model = clara(data2[, 1:2], k = 3, metric = "manhattan", stand = FALSE)
plot(data2[, 1:2], col = colors[model$clustering], xlab = "x", ylab = "y", main = "Манхэттенская метрика - без стандартизации", pch = 19)

model = clara(data2[, 1:2], k = 3, metric = "euclidean", stand = TRUE)
plot(data2[, 1:2], col = colors[model$clustering], xlab = "x", ylab = "y", main = "Евклидова метрика - со стандартизацией", pch = 19)

model = clara(data2[, 1:2], k = 3, metric = "manhattan", stand = TRUE)
plot(data2[, 1:2], col = colors[model$clustering], xlab = "x", ylab = "y", main = "Манхэттенская метрика - со стандартизацией", pch = 19)
```

Задание 3.

Постройте дендрограмму для набора данных votes.repub в пакете «cluster» (число голосов, поданных за республиканцев на выборах с 1856 по 1976 год). Строки представляют 50 штатов, а столбцы - годы выборов (31). Проинтерпретируйте полученный результат.

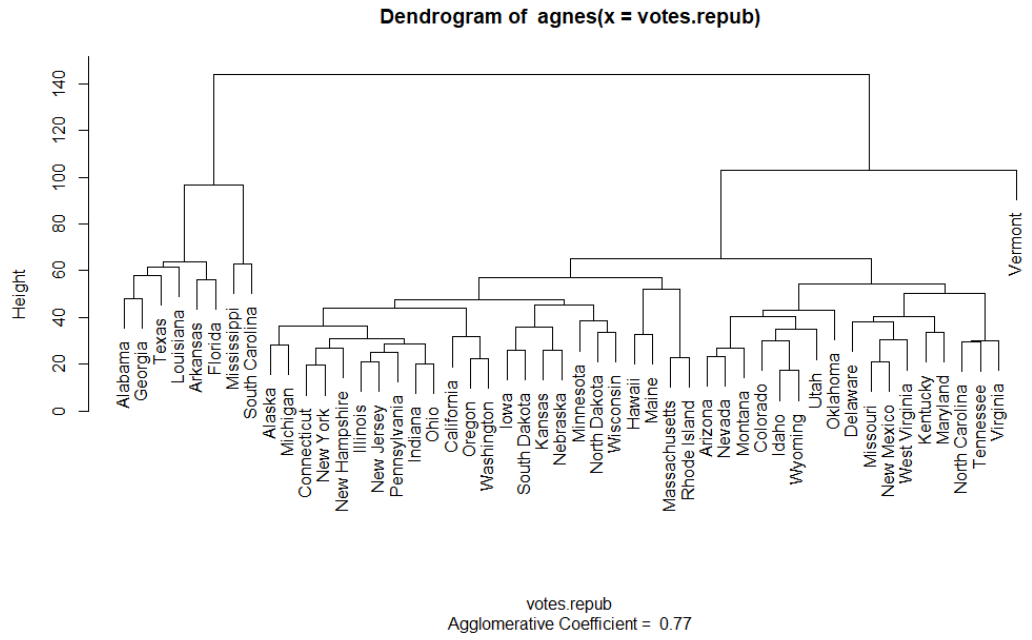

Рисунок 7 – Дендрограмма для набора votes.repub

На дендограмме выделяются три основных кластера по среднему, за всё время, количеству голосов за республиканцев на выборах. Штат Vermont выделяется среди всех большинством голосов. Штаты Alabama, Georgia, Texas, Louisiana, Arkansas, Florida, Mississipi, South Carolina объединяются в кластер с средним количеством голосов. Остальные штаты с меньшим

Листинг кода 3 задачи:

```
library(cluster)
data3<-votes.repub
plot(agnes(votes.repub))
```

Задание 4.

Постройте дендрограмму для набора данных animals в пакете «cluster». Данные содержат 6 двоичных признаков для 20 животных. Переменные - [, 1] war теплокровные; [, 2] fly летающие; [, 3] ver позвоночные; [, 4] end вымирающие; [, 5] gro живущие в группе; [, 6] hai имеющие волосяной покров. Проинтерпретируйте полученный результат.

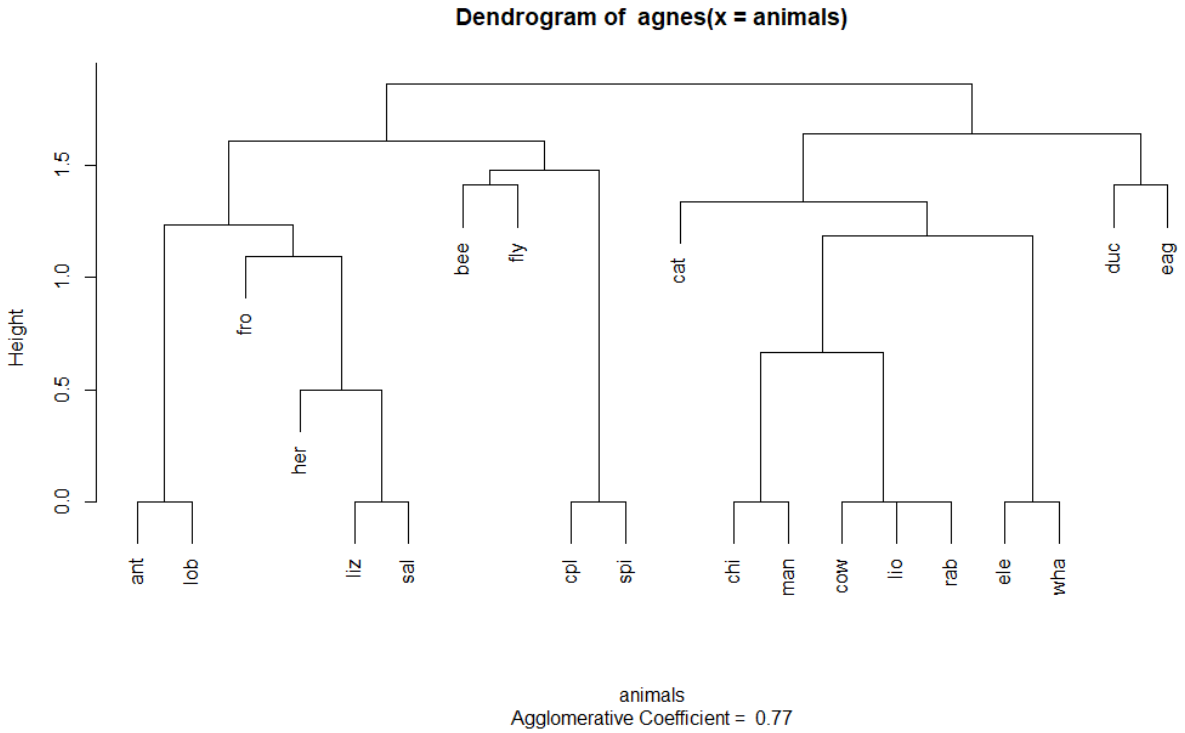

Рисунок 8 – Дендрограмма для набора animals

Из данной диаграммы видно, что кластеры формируются в зависимости от схожести признаков видов животных. Чем больше схожести, тем вероятнее, что животные окажутся в одном кластере.

Листинг кода 4 задачи:

```
library(cluster)
data4<-animals
plot(agnes(animals))
```

Задание 5.

Рассмотрите данные из файла seeds_dataset.txt, который содержит описание зерен трех сортов пшеницы: Kama, Rosa and Canadian. Признаки: 1. область A, 2. периметр P, 3. компактность C = 4*pi*A/P^2, 4. длина зерна, 5. ширина зерна, 6. коэффициент ассиметрии, 7. длина колоска.

Проведем кластеризацию (рисунок 9) методом Clara используя Манхэттенскую метрику без стандартизации. Точность такой модели равна 91,45%.

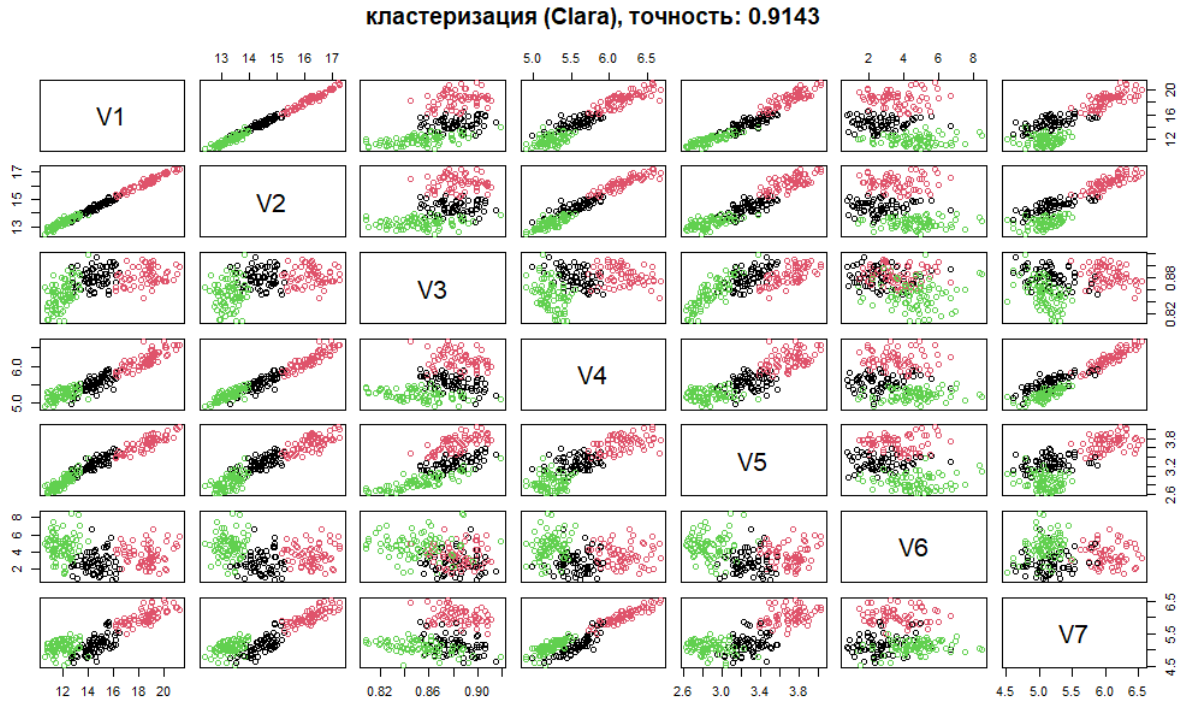

Рисунок 9 – Кластеризация для набора seeds

Построим дендрограмму (рисунок 10).

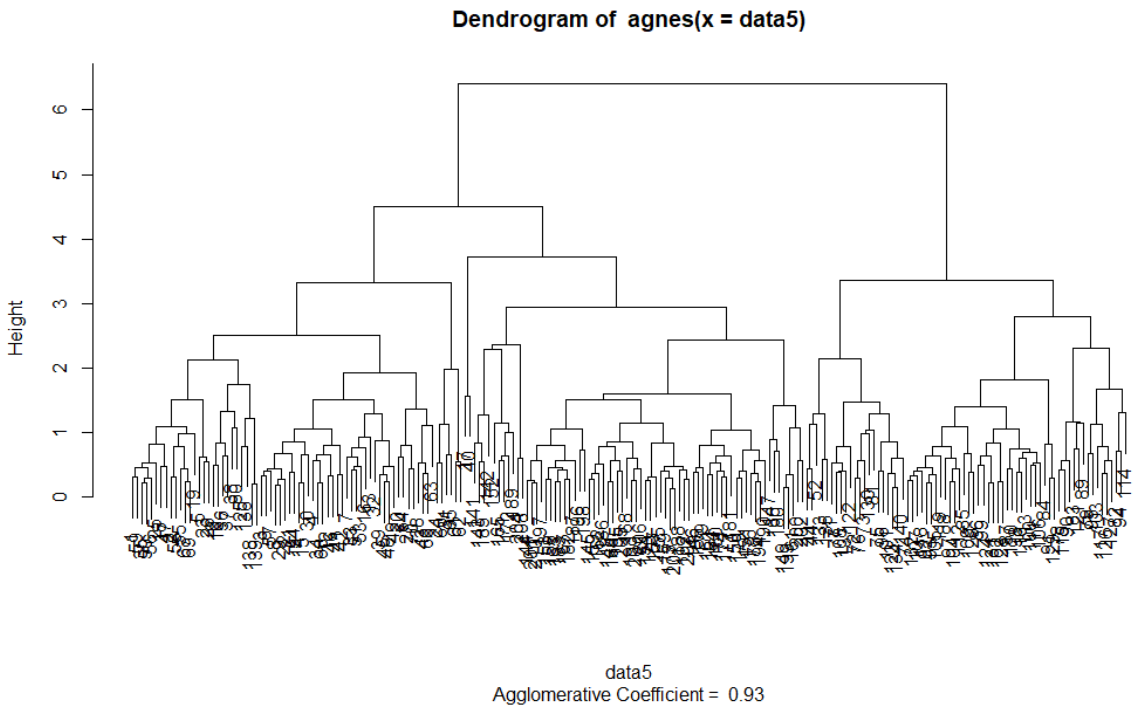

Рисунок 10 – Дендрограмма для набора seeds

Из дендрограммы видно, что данные делятся на три основных кластера.

Листинг кода 5 задачи:

```
library(cluster)
data5<-read.table("seeds_dataset.txt", sep = "", header=FALSE)
data5 <- na.omit(data5)
label <- data5[, 8]
features <- data5[,-8]

model <- clara(features, 3,metric = "manhattan", stand = FALSE)
accuracy <- round(mean(model$cluster == label),4)
plot(features, col = model$cluster, main = paste("кластеризация (Clara), точность:",accuracy))
plot(agnes(data5))
```