Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра «Системное программирование»

**ОТЧЁТ**

по лабораторной работе №9

на тему «Плотностная кластеризация»

Выполнил
Студент группы КЭ-120
_____ Д.А. Снегирева
«____»_____2020 г.
Email: dashasneg@mail.ru

Челябинск 2021

## ЗАДАНИЕ

1. Выполните кластеризацию набора 2-х или 3-мерных данных с помощью алгоритма DBSCAN (предполагается, что полученные кластеры не будут выпуклыми), используя различные значения параметров $MinPts$ (из интервала 3..9) и $Eps$.

Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

2. Выполните кластеризацию зашумленного набора данных из задания 8 с помощью алгоритма DBSCAN, используя различные значения параметров $MinPts$ (из интервала 3..9) и $Eps$.

Выполните визуализацию полученных результатов в виде точечных графиков, на которых цвет точки отражает принадлежность кластеру.

СОДЕРЖАНИЕ

# 1 КРАТКИЕ СВЕДЕНИЯ О НАБОРАХ ДАННЫХ И СРЕДСТВАХ РЕАЛИЗАЦИИ

В данной работе был использован набор данных по рукописным цифрам UCI ML hand-written digits datasets (https://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits), состоящий из десяти классов. Изображения рукописных цифр в наборе представлены матрицей 8 x 8 (интенсивности белого цвета для каждого пикселя). Далее эта матрица "разворачивается" в вектор длины 64, получается признаковое описание объекта. С помощью PCA размерность была снижена до 2 признаков.

Также, был использован набор данных Ирисы Фишера (https://archive.ics.uci.edu/ml/datasets/iris). После понижения размерности с помощью PCA до 2 признаков, данные образуют 3 группы, которые имеют вытянутую форму.

В качестве средств реализации были использованы библиотеки scikit-learn и scikit-learn-extra.

Репозиторий задания: https://github.com/DasHaSneg/BigDataMiningCourse

Каталог задания: 9 density clustering

## 2 ВИЗУАЛИЗАЦИЯ РЕЗУЛЬТАТОВ

На рисунке 1 приведен набор данных ирисы Фишера, сниженный до 2 признаков с помощью PCA. Можно заметить, что кластеры имеют вытянутую форму.

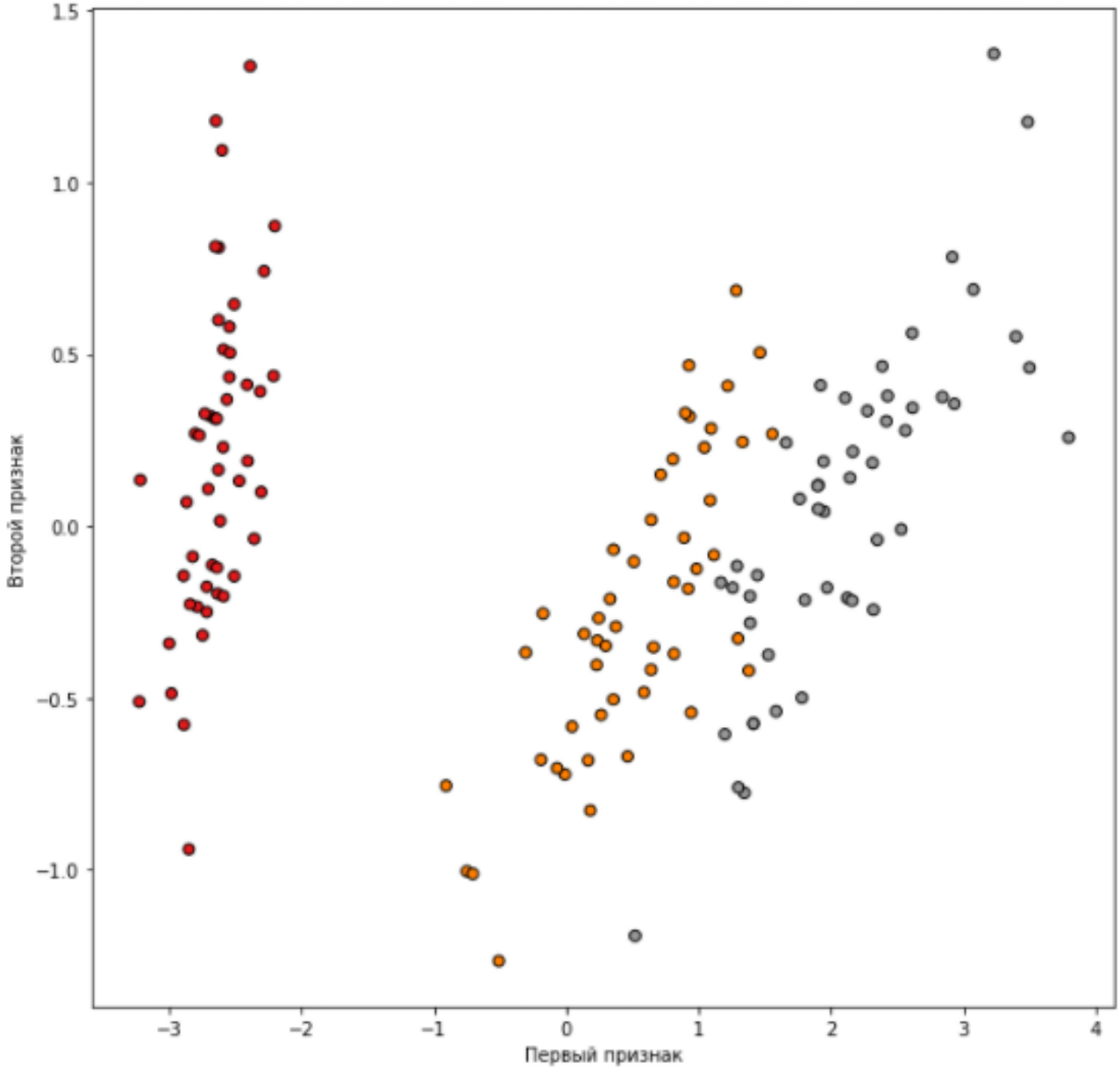


Рис. 1. Набор данных с ирисами Фишера

Далее была выполнена кластеризация данного набора с помощью алгоритма DBSCAN со значениями параметра min_samples от 3 до 9 с шагом 1 и параметра eps от 0.3 до 0.4 с шагом 0.05. Результаты приведены на рисунке 2.

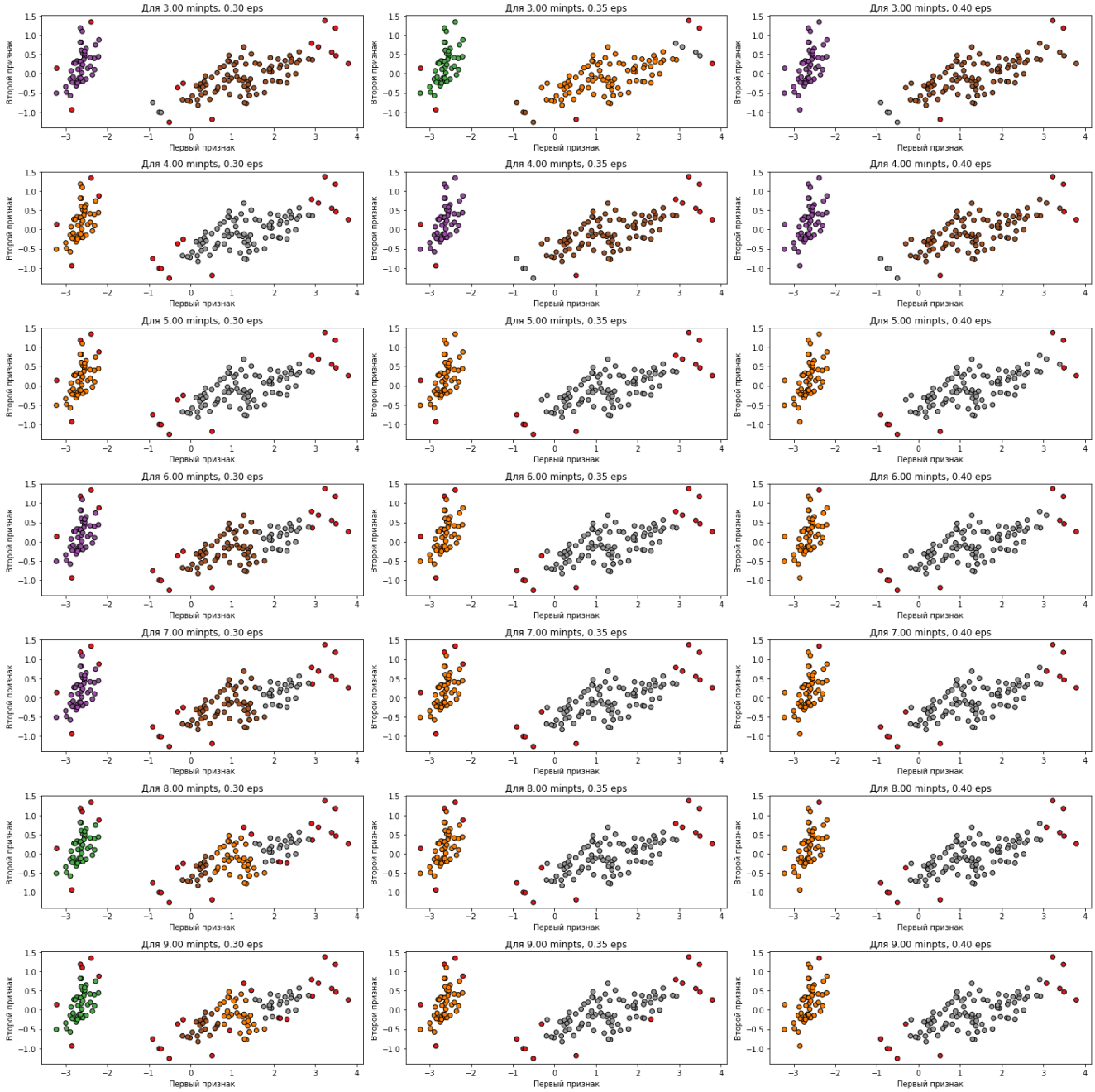


Рис. 2. Результаты кластеризации второго набора данных с помощью DBSCAN

Как видно из рисунка, наиболее удачными параметрами оказались min_samples = 6 и eps = 0.30.

На рисунке 3 приведен набор данных с рукописными цифрами, сниженный до 2 признаков с помощью PCA. Можно заметить, что даже на глаз рукописные цифры неплохо разделяются на кластеры (разные цвета точек означают разные кластеры).

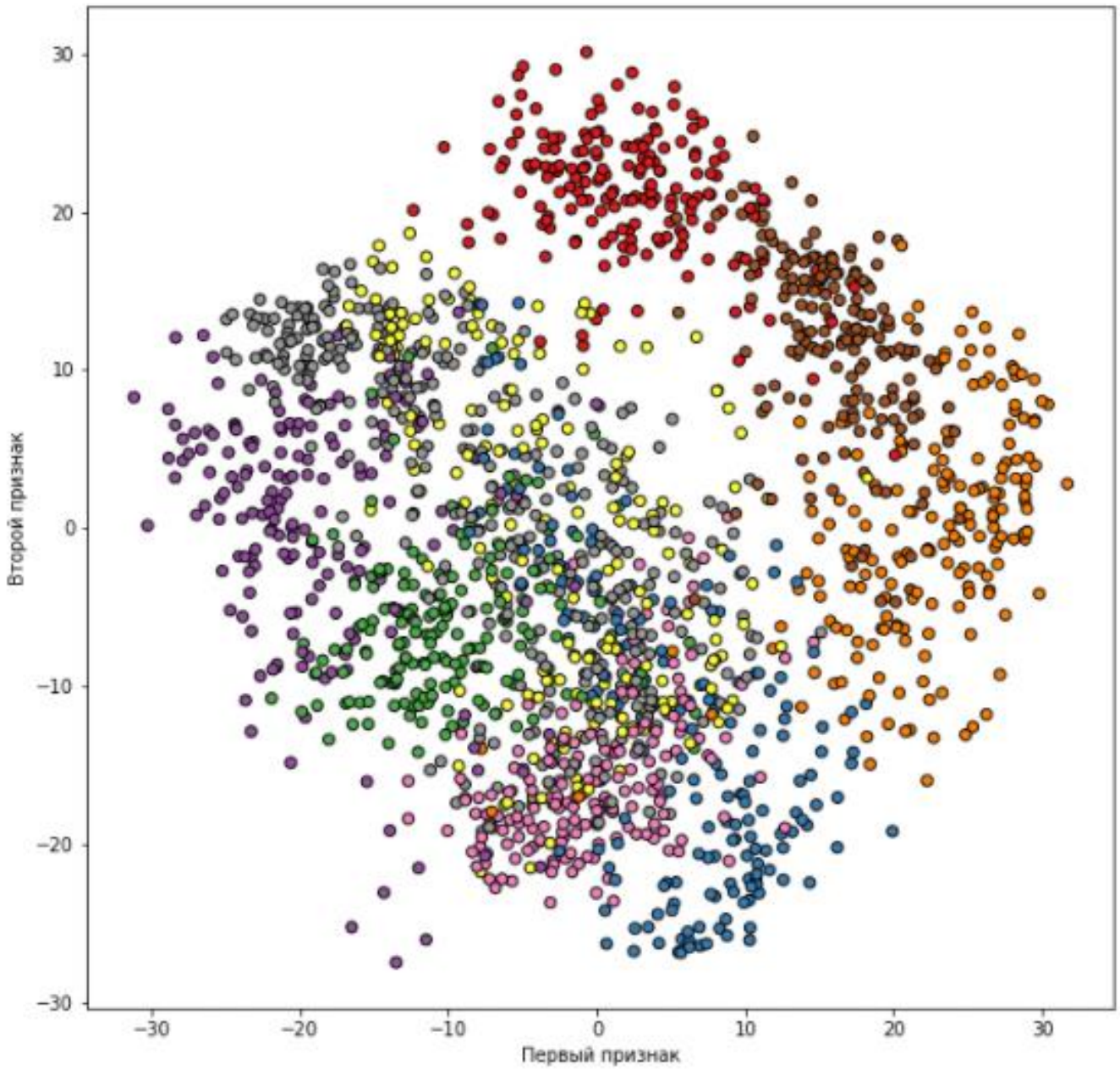


Рис. 3. Набор данных с рукописными цифрами

Затем был использован метод DBSCAN для кластеризации данного набора данных с параметрами min_samples от 3 до 9 с шагом 1 и eps от 1.7 до 1.8 с шагом 0.05. Результаты приведены на рисунке 4.

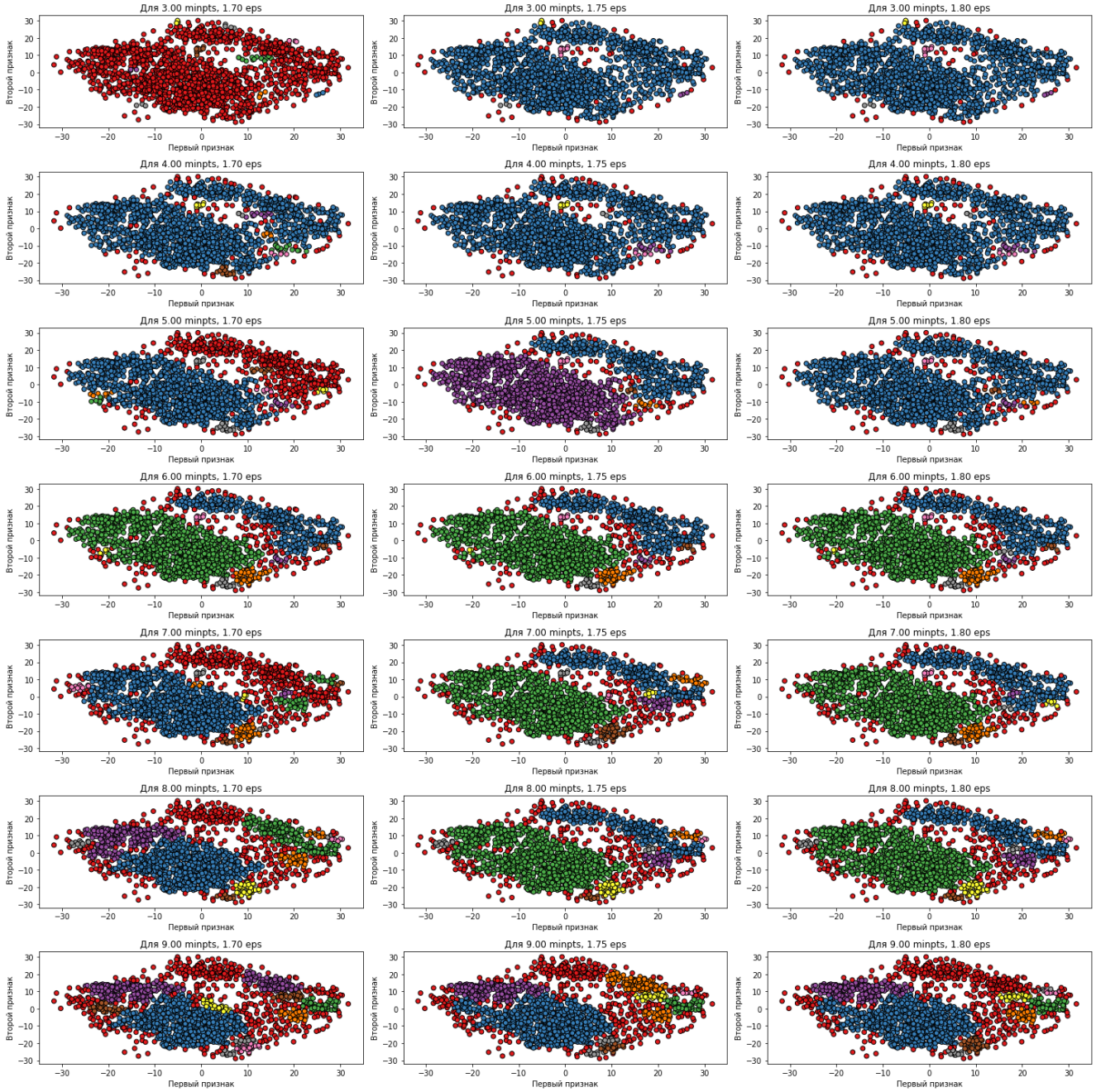


Рис. 4. Результаты кластеризации первого набора данных с помощью DBSCAN

Из рисунка видно, что ни одна из комбинаций параметров не привела к кластерам похожим ну нужные. Поэтом можно сделать вывод, что данный метод не подходит для выпуклых кластеров.