Министерство науки и высшего образования РФ

ФГБОУ ВО «Удмуртский государственный университет»

Институт математики, информационных технологий и физики

Кафедра теоретических основ информатики

Отчет по дисциплине «Машинное обучение»

Лабораторная работа №3 “Решение задачи кластеризации”

Выполнил:

студент группы ОМ-02.04.02.04-11

Данилов Владимир Александрович

Принял:

К.ф-м.н.

Федоров Дмитрий Леонидович

Ижевск, 2023 г.

**Содержание**

[Постановка задачи 3](#_Toc151253360)

[Краткая теория 3](#_Toc151253361)

[Результаты тестирования модели 5](#_Toc151253362)

[Анализ результатов и выводы 8](#_Toc151253363)

# Постановка задачи

Требуется написать программу, решающую задачу бинарной кластеризации. Для этого требуется выполнить следующее:

1. Загрузить набор данных по анализу и прогнозированию инфаркту как датафрейм библиотеки pandas. Источник данных: https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis.
2. Провести кластеризацию несколькими методами, используя библиотеку sklearn.
   1. Методом k средних (выбрать значение 3 <= k <= 10);
   2. Методом сдвига к среднему
   3. Методом DBSCAN;
   4. Методом иерархического слияния
3. Для каждого метода провести визуализацию данных. Провести анализ методом силуэтных коэффициентов.
4. Построить гистограммы с результатами вычислений.

# Краткая теория

Кластеризация — это процесс разбиения данных на группы (кластеры), так чтобы объекты в одной группе были похожи друг на друга, но отличались от объектов в других группах. Цель кластеризации — найти естественные группы в данных, то есть кластеры, которые соответствуют реальным структурам в данных. Кластеризация используется во многих областях, включая data mining, машинное обучение, анализ данных и визуализацию данных.

Метод алгоритма “k средних” пытается сгруппировать наблюдения в k групп, причем каждая группа имеет примерно равную дисперсию. Количество групп к задается пользователем в качестве гиперпараметра. В частности, в алгоритме к средних:

1. В случайных позициях создается к "центральных" точек, по одной на кластер.

2. Для каждого наблюдения:

• вычисляется расстояние между каждым наблюдением и к центральными точками;

• наблюдение назначается кластеру ближайшей центральной точки.

3. Центральные точки передвигаются в средние значения (т. е. в центры) своих соответствующих кластеров.

4. Шаги 2 и 3 повторяются до тех пор, пока ни одно из наблюдений не изменит свою принадлежность кластеру.

На этом этапе алгоритм считается достигшим схождения и останавливается

В алгоритме сдвига к среднему заключена простая идея, которую, правда, несколько трудно объяснить. Поэтому наилучшим подходом может быть аналогия. Представьте себе очень туманное футбольное поле (т. е. двухмерное пространство), на котором находятся 100 человек (т. е. наши наблюдения). Из-за тумана человек может видеть только на небольшое расстояние. Каждую минуту каждый человек оглядывается и делает шаг в сторону большинства людей, которых он видит. С течением времени люди начинают собираться, поскольку они неоднократно предпринимают шаги в сторону все более крупных скоплений людей. В конечном счете получаются кластеры людей в пределах поля. Люди назначаются кластерам, в которых они оказываются.

Алгоритм DBSCAN руководствуется идеей о том, что кластеры будут областями, где много наблюдений плотно упакованы вместе, и не делает никаких допущений о форме кластера. В частности, в алгоритме DBSCAN:

1. Выбирается случайное наблюдение хi.

2. Если хi, имеет минимальное число близких соседей, то мы рассматриваем его как часть кластера.

3. Шаг 2 повторяется рекурсивно для всех соседей хi затем для соседа и т. д. Это основные наблюдения кластера.

4. После завершения шага 3 выбирается новая случайная точка (т. е. перезапуск шага 1).

Когда все это будет завершено, мы получим набор ключевых наблюдений для ряда кластеров. Наконец, любое наблюдение, близкое к кластеру, но не являющееся ключевым образцом, считается частью кластера, в то время как любое наблюдение, не близкое к кластеру, помечается как выброс.

Агломеративная кластеризация — это мощный и гибкий алгоритм, который выполняет кластеризацию иерархически. В агломеративной кластеризации все наблюдения начинаются как одноэлементные кластеры. Далее кластеры, удовлетворяющие некоторым критериям, объединяются. Этот процесс повторяется, выращивая кластеры до тех пор, пока не будет достигнута некоторая конечная точка. В библиотеке scikit-leam В классе AgglomerativeClustering используется параметр СВЯЗИ linkage для определения стратегии слияния, которая сводит к минимуму следующее:

* дисперсию объединенных кластеров по методу Уорда (ward);
* среднее расстояние между наблюдениями от пар кластеров по методу средней связи (average);
* максимальное расстояние между наблюдениями от пар кластеров по методу полной связи (complete).

# Результаты тестирования модели

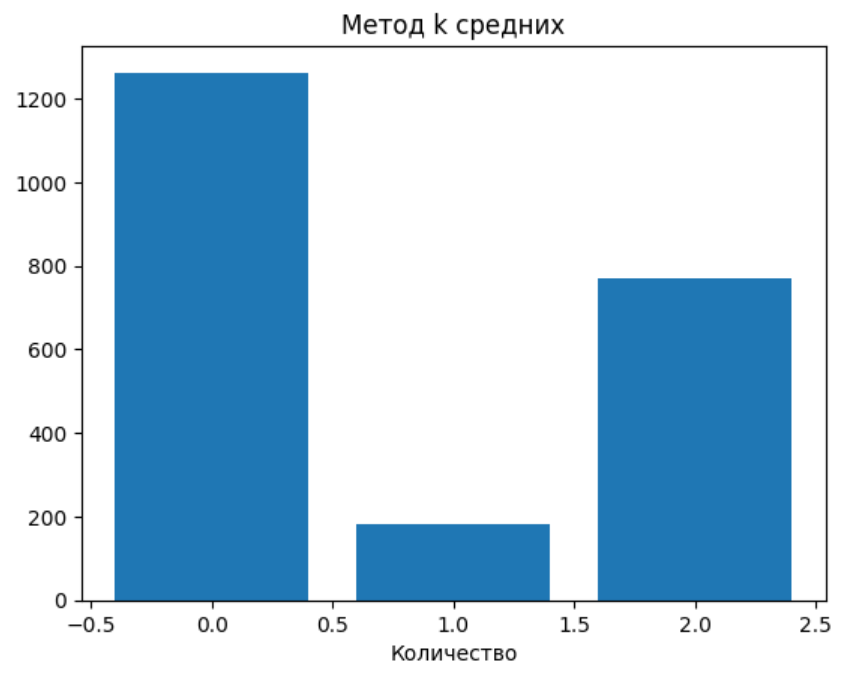
**Метод k средних:**

0.22709455175594823

0: 1262

1: 183

2: 771

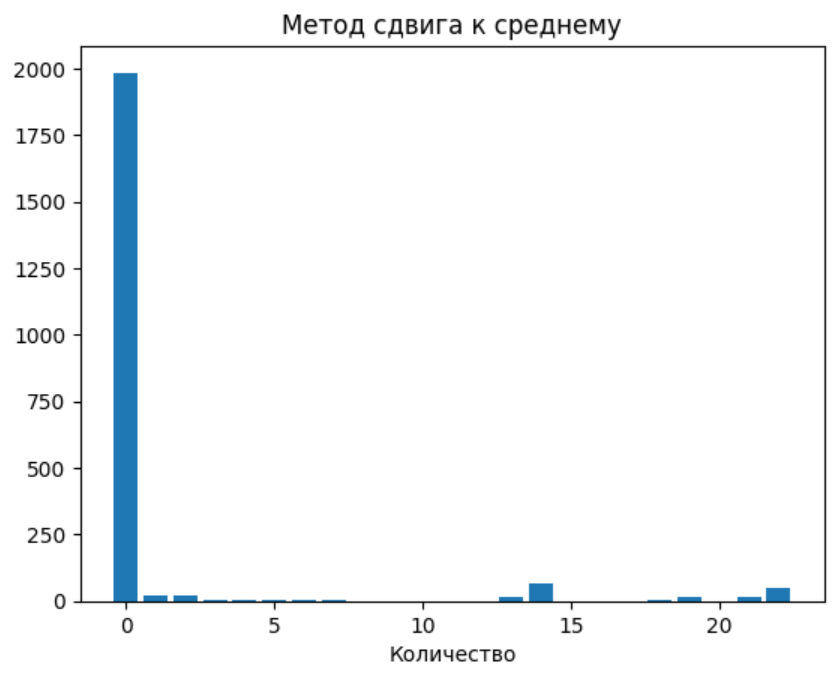


**Метод сдвига к среднему:**

0.5817500491982808

0: 100

1: 50



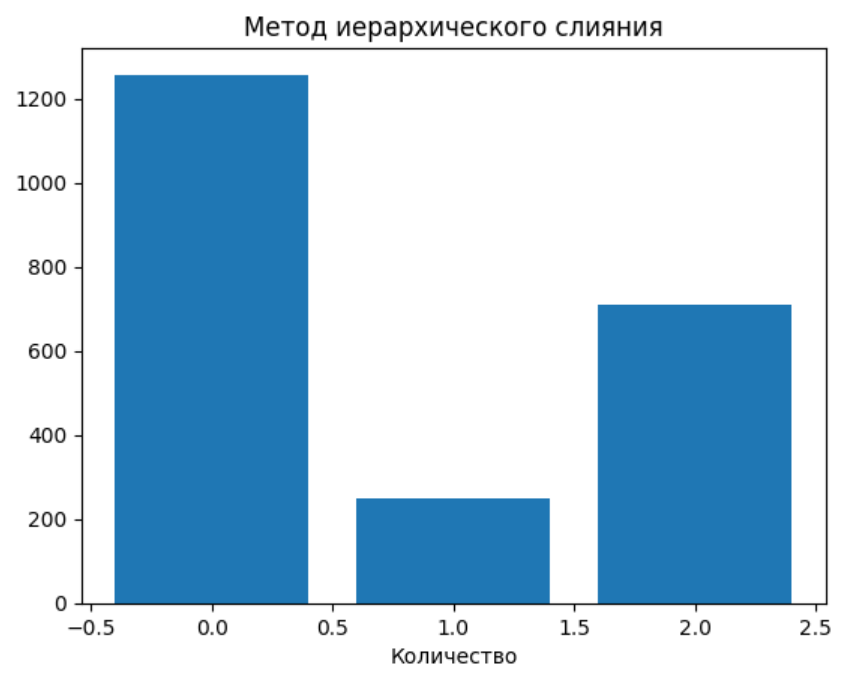
**Метод иерархического слияния:**

0.19473470003716356

0: 1257

1: 250

2: 709



# Анализ результатов и выводы