Липецкий государственный технический университет

Факультет автоматизации и информатики Кафедра автоматизированных систем управления

ЛАБОРАТОРНАЯ РАБОТА №5

по дисциплине «Прикладные интеллектуальные системы и экспертные системы»

Кластеризация данных

Студент Бубырь Д.А.

Группа М-ИАП-23-1

Руководитель Кургасов В.В.

доцент, канд. пед. наук

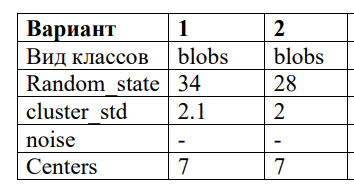
Липецк 2023 г.

Цель работы

Получить практические навыки решения задачи кластеризации фактографических данных в среде Jupiter Notebook. Научиться настраивать параметры методов и оценивать точность полученного разбиения.

Задание кафедры

1. Загрузить выборки согласно варианту задания.
2. Отобразить данные на графике в пространстве признаков. Поскольку решается задача кластеризации, то подразумевается, что априорная информация о принадлежности каждого объекта истинному классу неизвестна, соответственно, на данном этапе все объекты на графике должны отображаться одним цветом, без привязки к классу.
3. Провести иерархическую кластеризацию выборки, используя разные способы вычисления расстояния между кластерами: расстояние ближайшего соседа (single), дальнего соседа (complete), Уорда (Ward). Построить дендрограммы для каждого способа. Размер графика должен быть подобран таким образом, чтобы дендрограмма хорошо читалась.
4. Исходя из дендрограмм выбрать лучший способ вычисления расстояния между кластерами.
5. Для выбранного способа, исходя из дендрограммы, определить количество кластеров в имеющейся выборке. Отобразить разбиение на кластеры и центроиды на графике в пространстве признаков (объекты одного кластера должны отображаться одним и тем же цветом, центроиды всех кластеров – также одним цветом, отличным от цвета кластеров).
6. Рассчитать среднюю сумму квадратов расстояний до центроида, среднюю сумму межкластерных расстояний для данного разбиения. Сделать вывод о качестве разбиения.
7. Провести кластеризацию выборки методом k-средних. для k [1, 10].
8. Сформировать три графика: зависимость средней суммы квадратов расстояний до центроида, средней суммы средних внутрикластерных расстояний и средней суммы межкластерных расстояний от количества кластеров. Исходя из результатов, выбрать оптимальное количество кластеров.
9. Составить сравнительную таблицу результатов разбиения иерархическим методом и методом k-средних.

Вариант №2 

Ход работы

Загрузим выборки согласно варианту, код для генерации данных представлен на рисунке 1.



Рисунок 1 – Код для генерации данных

Отобразим на графике сгенерированные данные, для этого воспользуемся библиотекой matplotlib.pyplot. Полученный график представлен на рисунке 2.

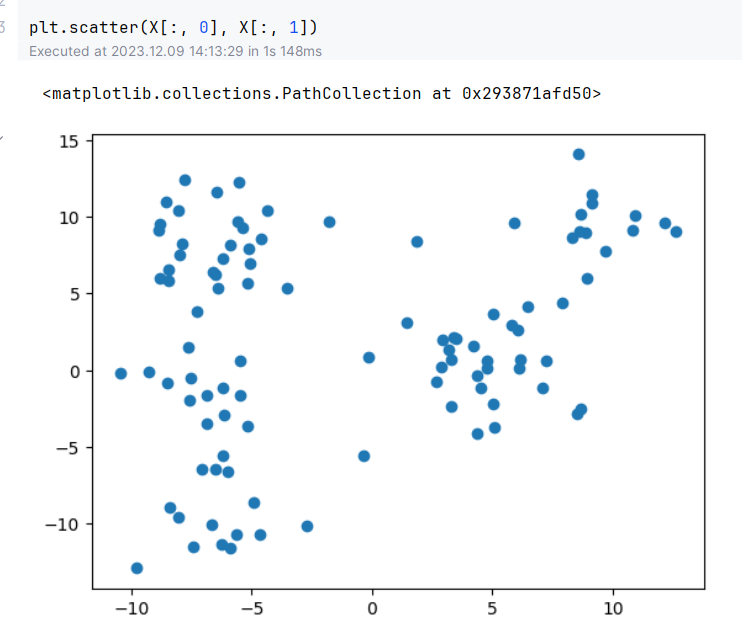


Рисунок 2 – Визуализация сгенерированных данных

Воспользуемся иерархической кластеризацией выборки с использованием различных методов вычисления расстояния.

Метод вычисления расстояния ближайшего соседа (single) и код для этого представлен на рисунке 3. Полученная дендограмма представлена на рисунке 4.

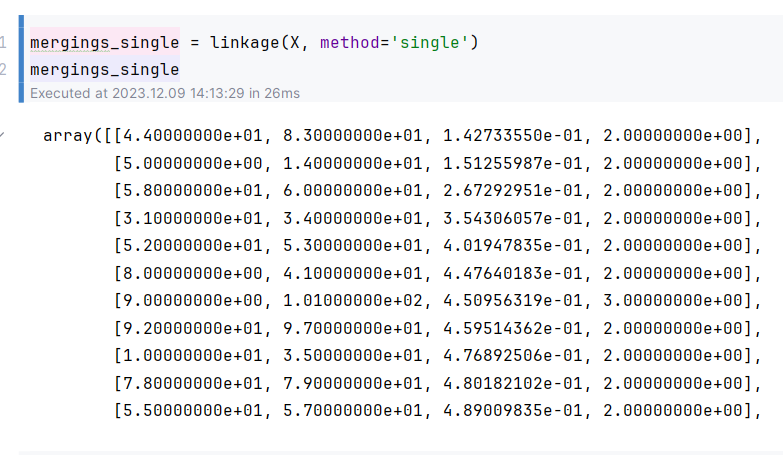


Рисунок 3 – Вычисленные расстояния ближайшего соседа (single)

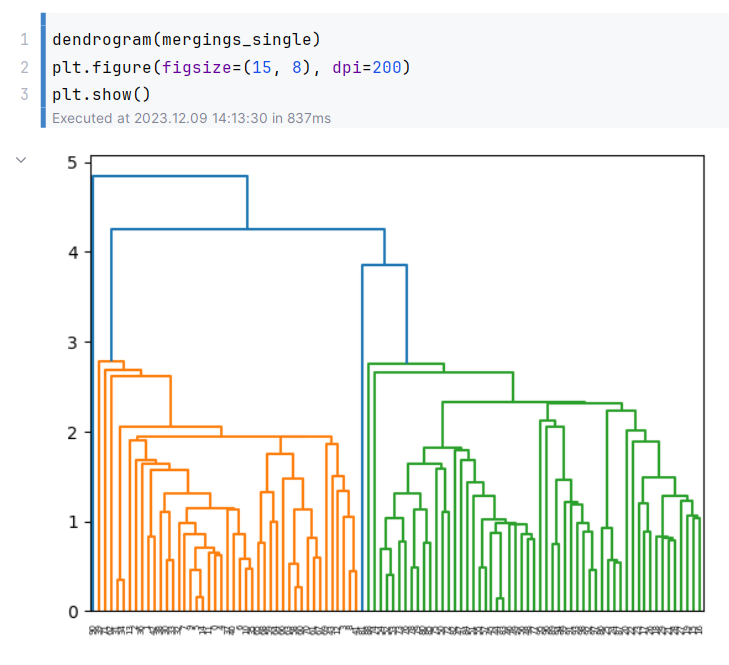


Рисунок 4 – Дендограмма для расстояния ближайшего соседа (single)

Метод вычисления расстояния дальнего соседа (complete) и код для этого представлен на рисунке 5. Полученная дендограмма представлена на рисунке 6.

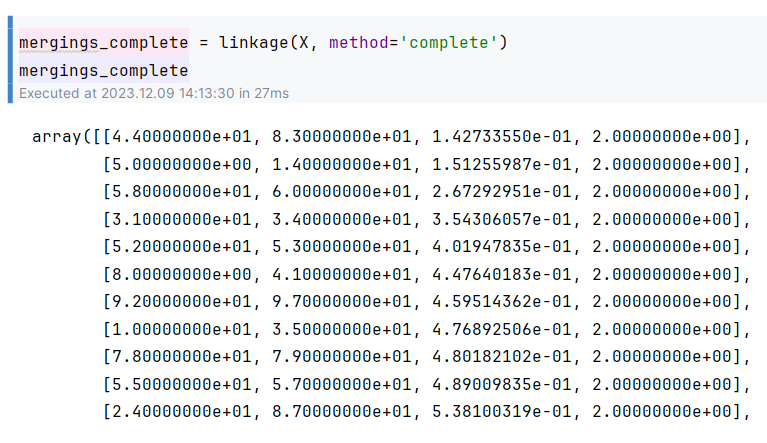


Рисунок 5 – Вычисленные расстояния дальнего соседа (complete)

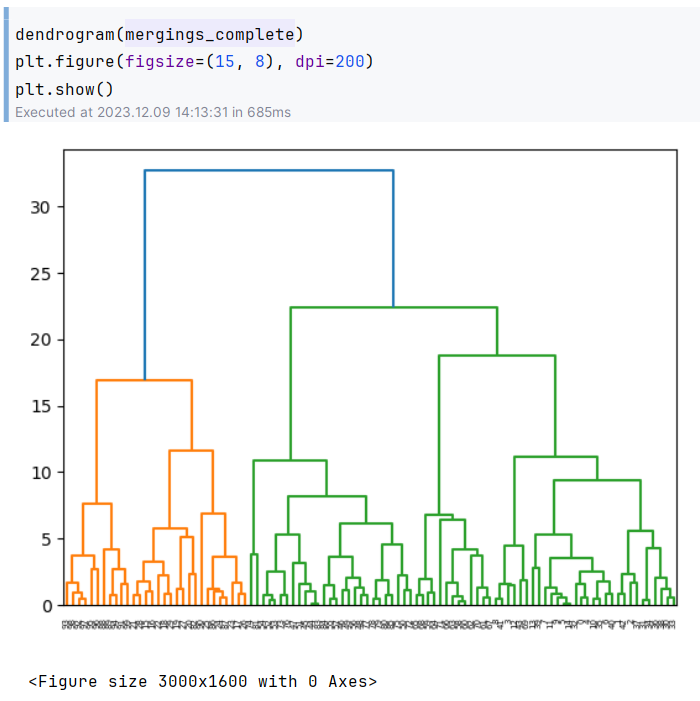


Рисунок 6 – Дендограмма для расстояния дальнего соседа (complete)

Метод вычисления расстояния Уорда (ward) и код для этого представлен на рисунке 7. Полученная дендограмма представлена на рисунке 8.

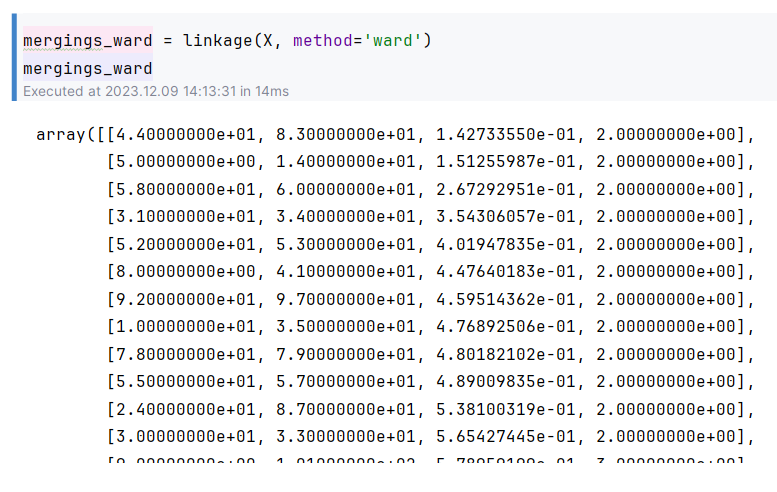


Рисунок 7 – Вычисленные расстояния с помощью метода Уорда (ward)

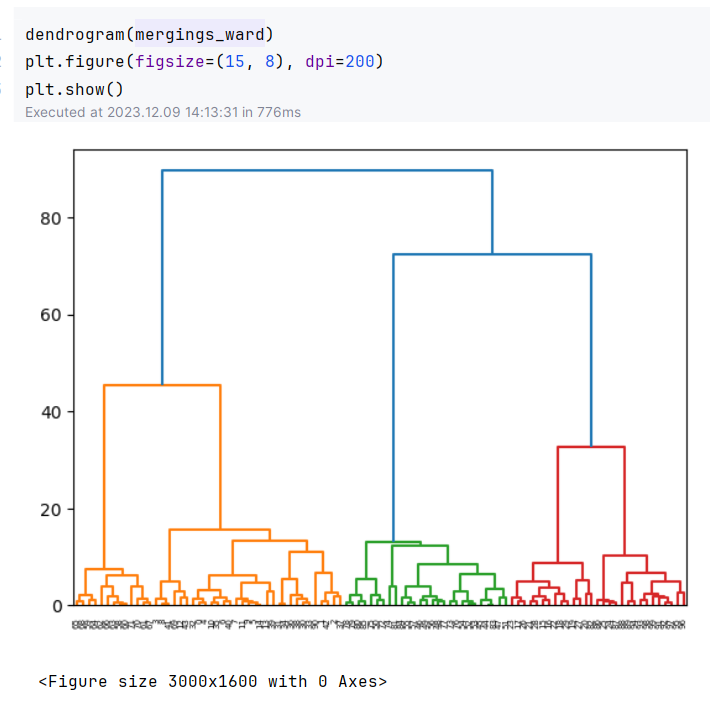


Рисунок 8 – Дендограмма для расстояния Уорда (ward)

Лучшим способом вычисления расстояния между кластерами является Уорда (ward). Определим количество кластеров в имеющейся выборке с использованием данного способа и отобразим разбиение на кластеры и центроиды на графике в пространстве признаков. Полученное разбиение представлено на рисунке 9.

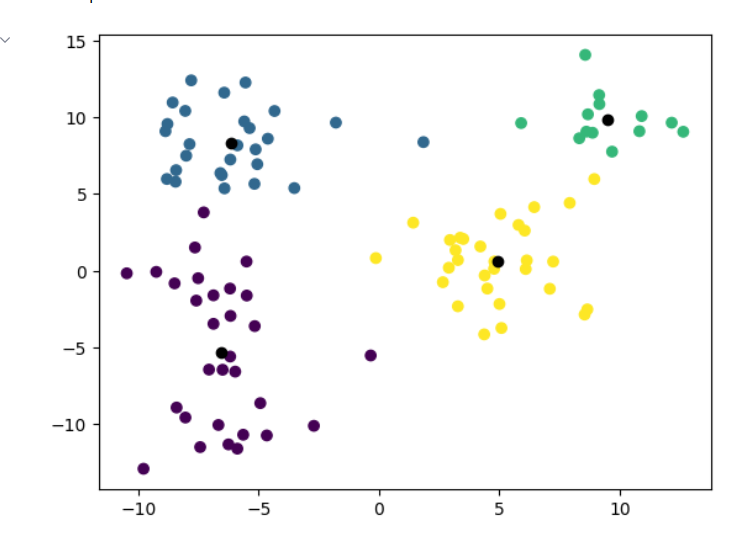


Рисунок 9 – График разбиения данных на кластеры

Рассчитаем среднюю сумму квадратов расстояний до центроида, среднюю сумму средних внутрикластерных расстояний и среднюю сумму межкластерных расстояний для данного разбиения.

Рассчитанное значение суммы квадратов расстояний до центроида и код для этого представлен на рисунке 10. Рассчитанное значение средних внутрикластерных расстояний и код для этого представлен на рисунке 11. Рассчитанное значение суммы межкластерных расстояний и код для этого представлен на рисунке 12.

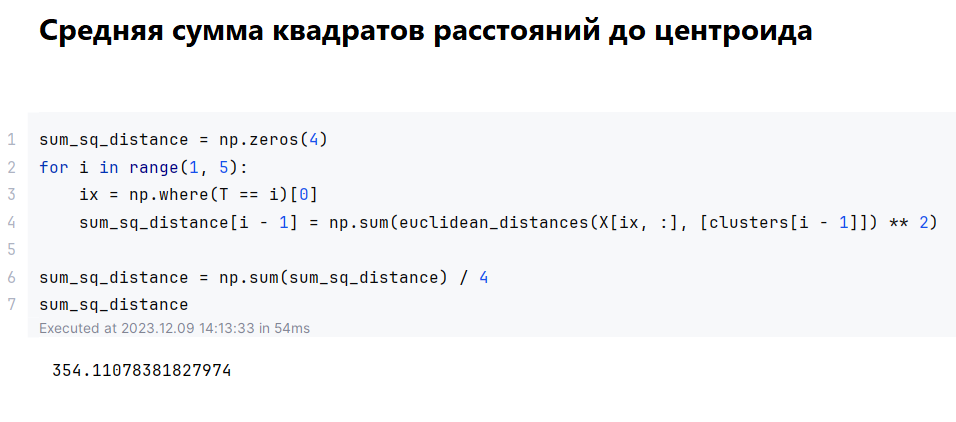


Рисунок 10 – Сумма квадратов расстояний до центроида

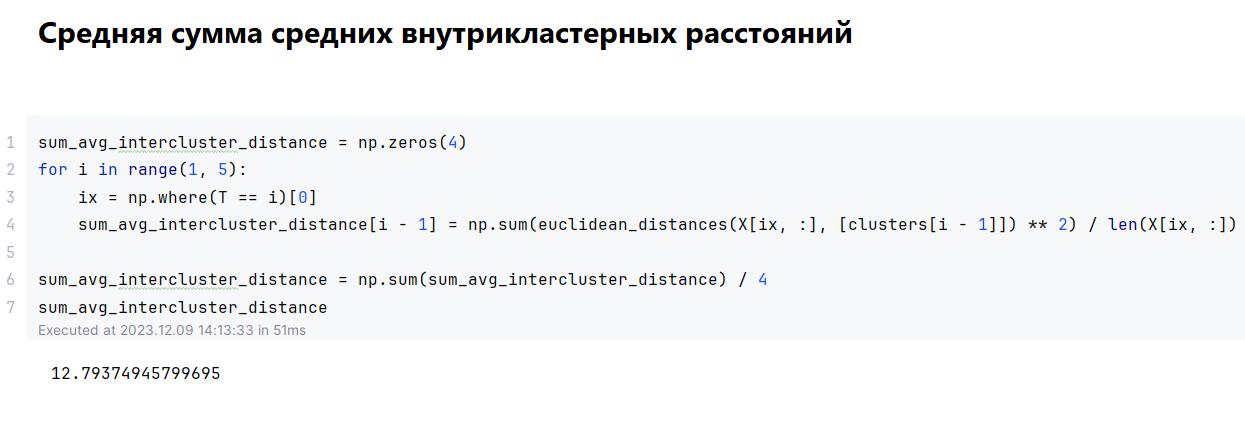


Рисунок 11 – Сумма средних внутрикластерных расстояний

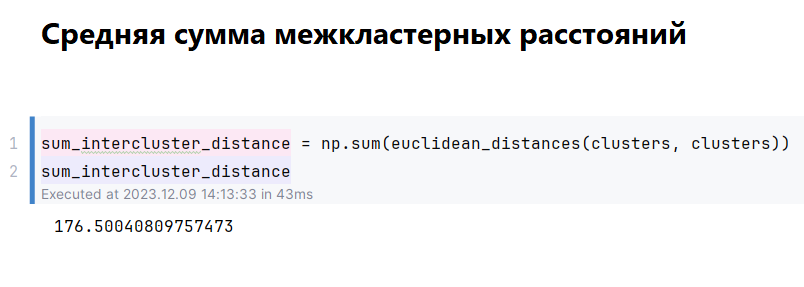


Рисунок 12 – Сумма межкластерных расстояний

Проведем кластеризацию выборки методом k-средних для k = [1, 10], а после построим три графика: зависимость средней суммы квадратов расстояний до центроида, средней суммы средних внутрикластерных расстояний и средней суммы межкластерных расстояний от количества кластеров.

Код для расчета средней суммы квадратов расстояний до центроида представлен на рисунке 13, а построенный график на рисунке 14.

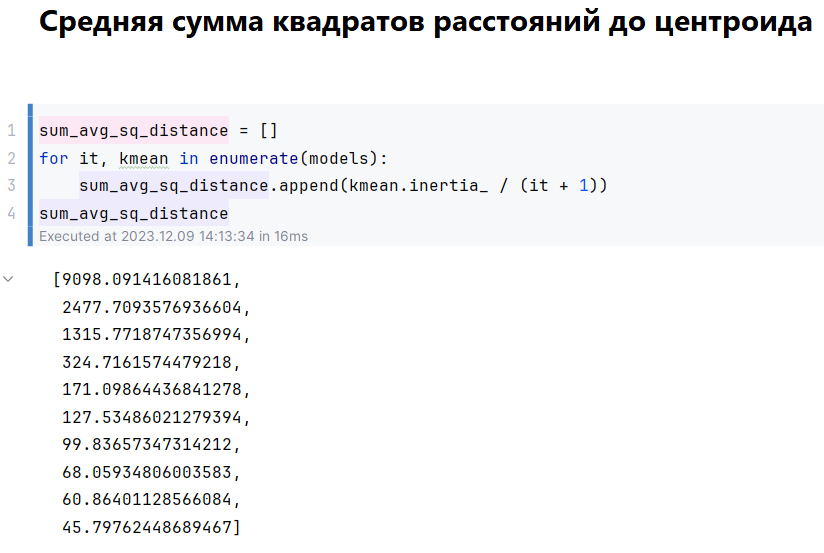


Рисунок 13 – Код для вычисления средней суммы квадратов расстояний до центроида

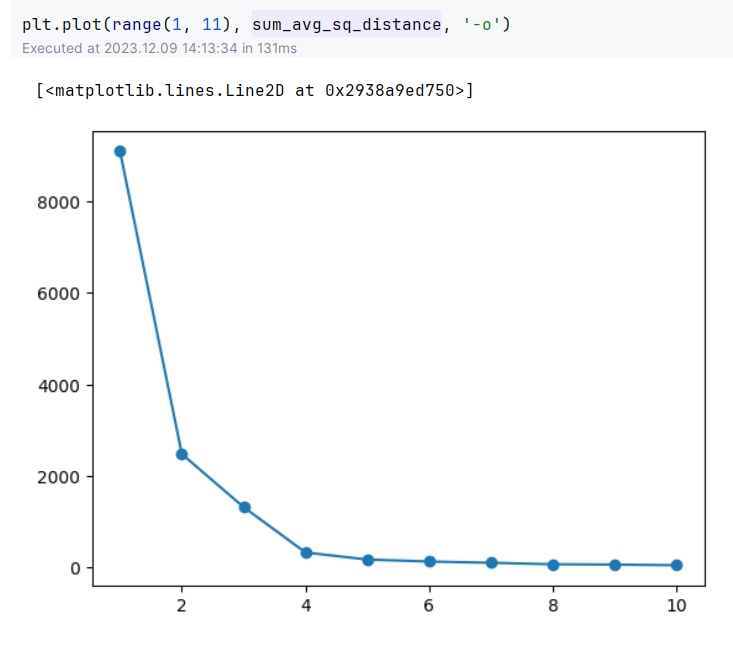


Рисунок 14 – График средней суммы квадратов расстояний до центроида Код для расчета средней суммы средних внутрикластерных расстояний

представлен на рисунке 15, а построенный график на рисунке 16.

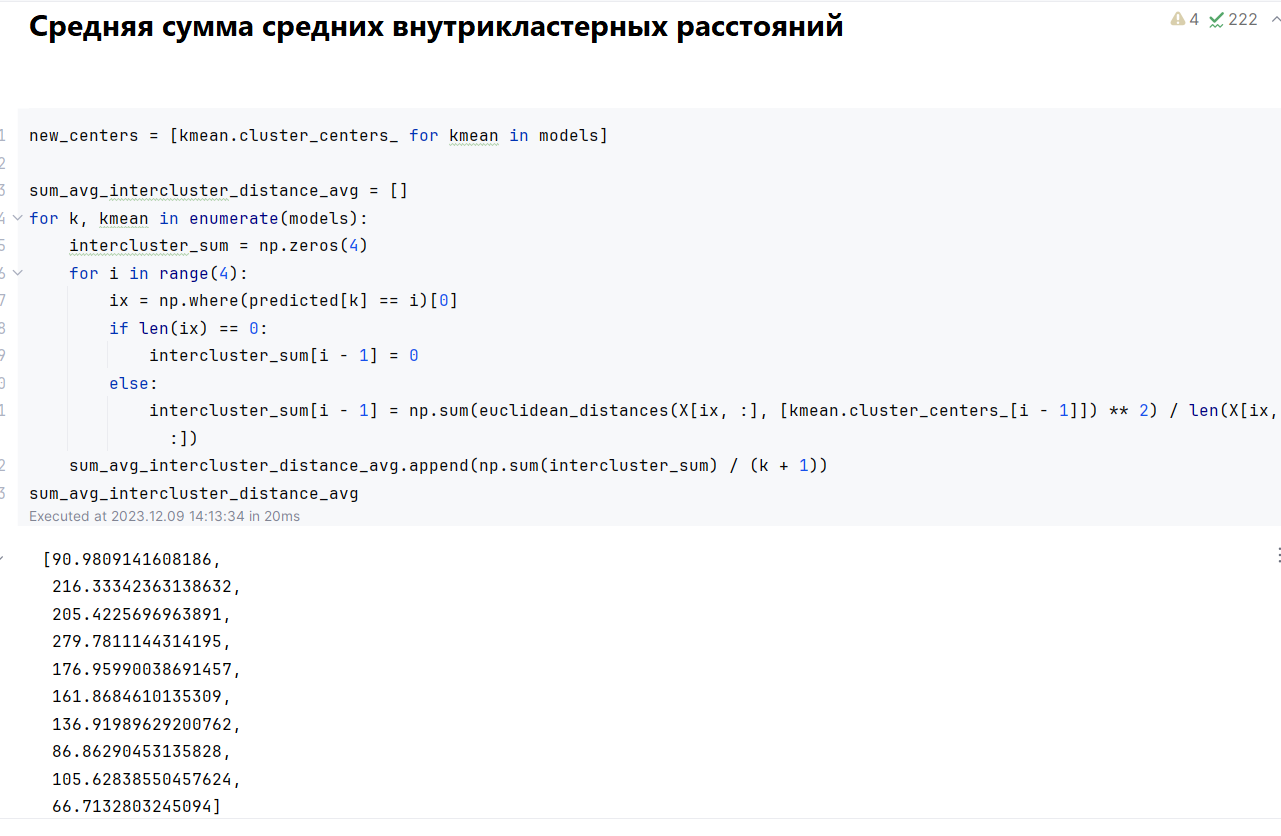


Рисунок 15 – Код для вычисления средней суммы средних внутрикластерных расстояний

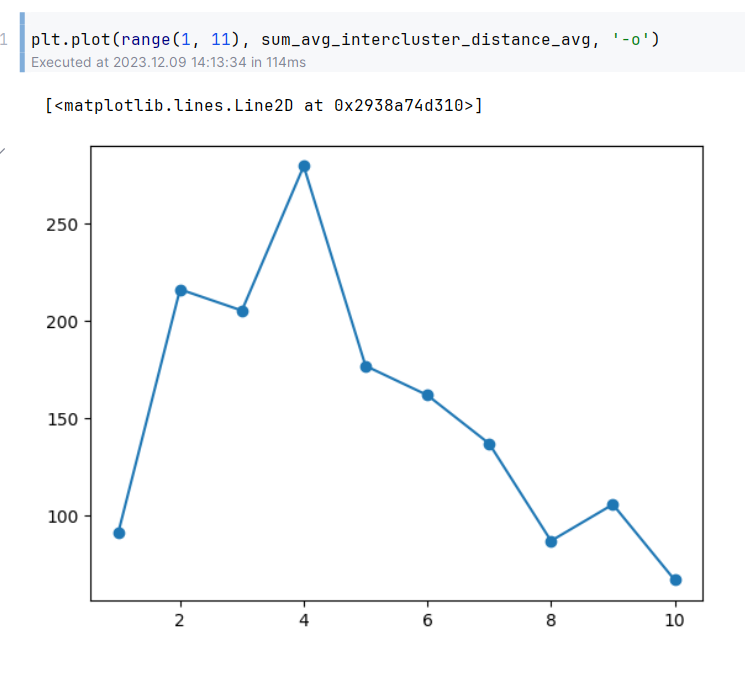


Рисунок 16 – График средней суммы средних внутрикластерных расстояний

Код для расчета средней суммы межкластерных расстояний от количества кластеров представлен на рисунке 17, а построенный график на рисунке 18.

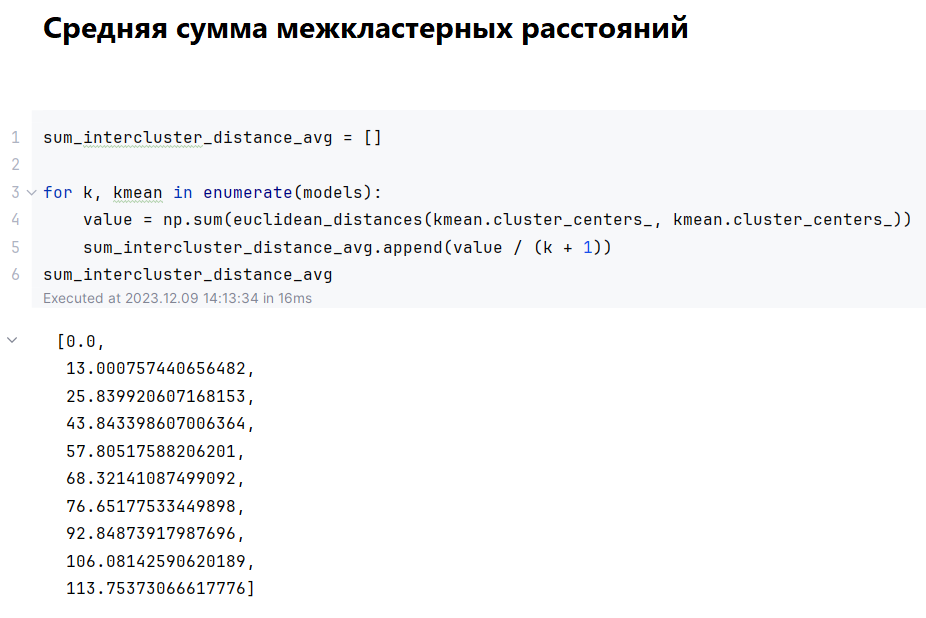


Рисунок 17 – Код для вычисления средней суммы межкластерных расстояний от количества кластеров

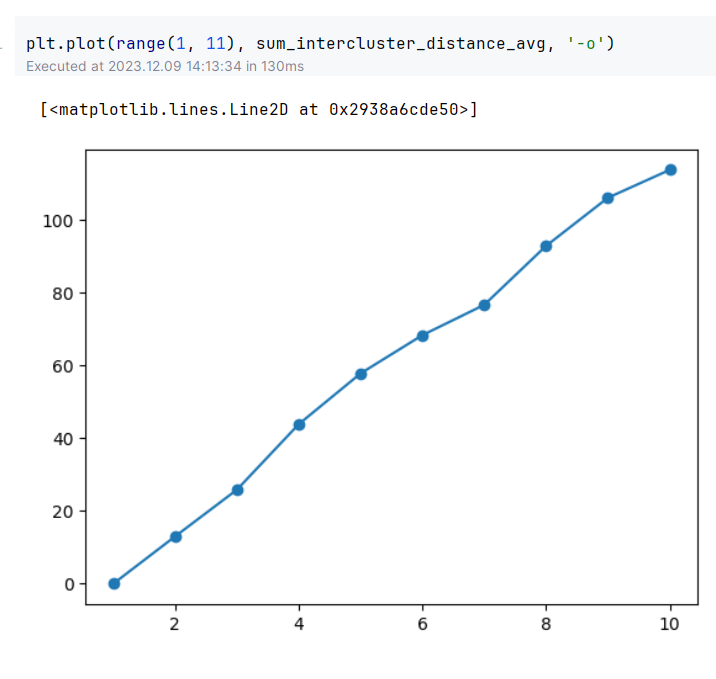


Рисунок 18 – График средней суммы межкластерных расстояний от количества кластеров

Как видно из рисунка 14 оптимальное количество кластеров составляет 4.

Составим сравнительную таблицу для ранее описанных метрик качества моделей для иерархического метода и метода k-средних. Составленная таблица представлена на рисунках 19 – 21.

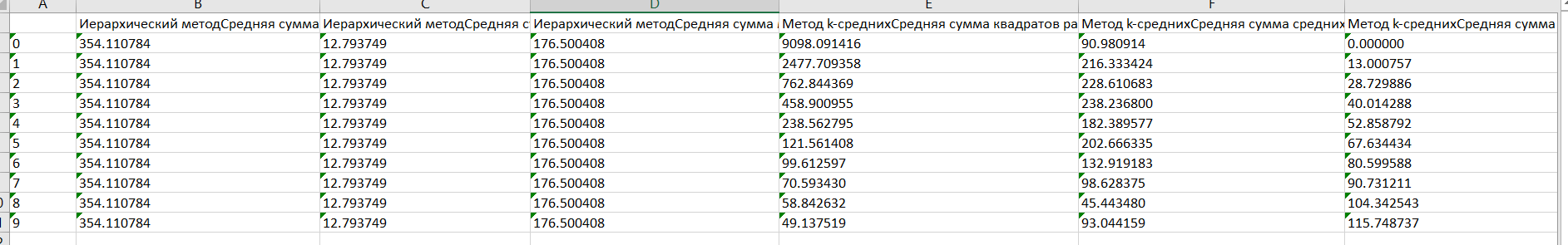


Рисунок 19 – Сводная таблица сравнения

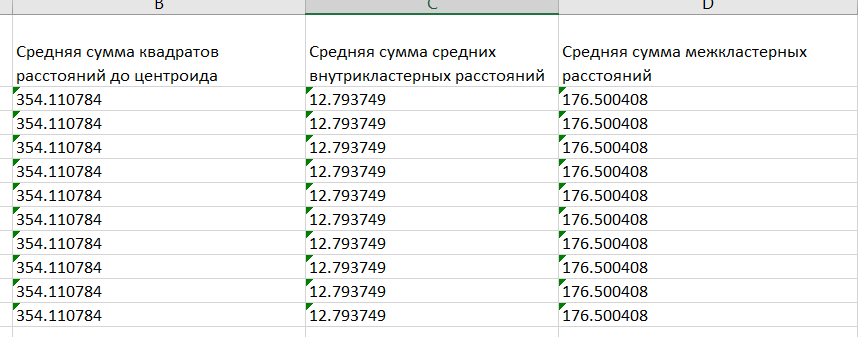


Рисунок 20 – Таблица данных для иерархического метода

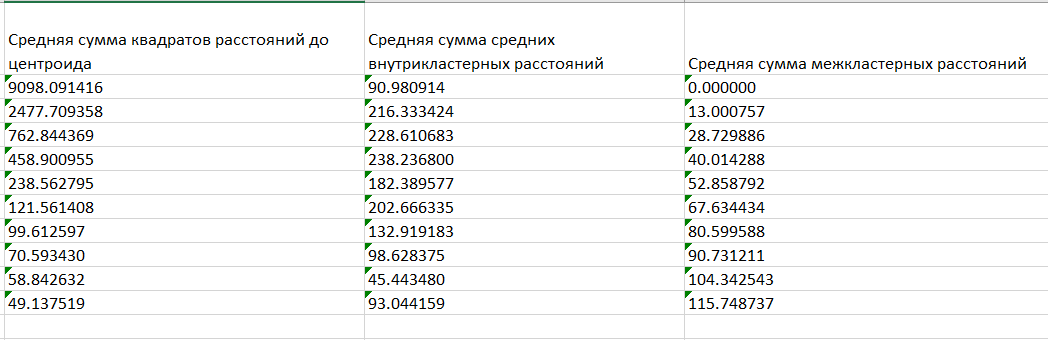


Рисунок 21 – Таблица данных для метода k-средних

Вывод

В ходе выполнения данной лабораторной работы мною были получены навыки кластеризации данных.

В рамках данной работы были применены различные методы кластеризации: иерархический метод и метод k-средних.

В ходе анализа метрик было определено, что оптимальное значение кластеров равняется четырем.

Также была составлена таблица сравнения метрик иерархическим методом кластеризации и методом k-средних.