Липецкий государственный технический университет

Факультет автоматизации и информатики Кафедра автоматизированных систем управления

ЛАБОРАТОРНАЯ РАБОТА №4

по дисциплине «Прикладные интеллектуальные системы и экспертные системы»

Кластеризация данных

Студент Фетисов В. Д.

Группа М-ИАП-23

Руководитель Кургасов В. В.

Доцент

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 №17 n\_samples = 100

Вид классов: blobs

Random\_state = 18

cluster\_std = 2

Centers = 8

Для всех вариантов: n\_features = 2

n\_redundant = 0

n\_informative = 2

n\_cluster\_per\_class = 1

n\_classes = 4

Ход работы

На рисунках 1-10 представлен код и процесс работы программы.

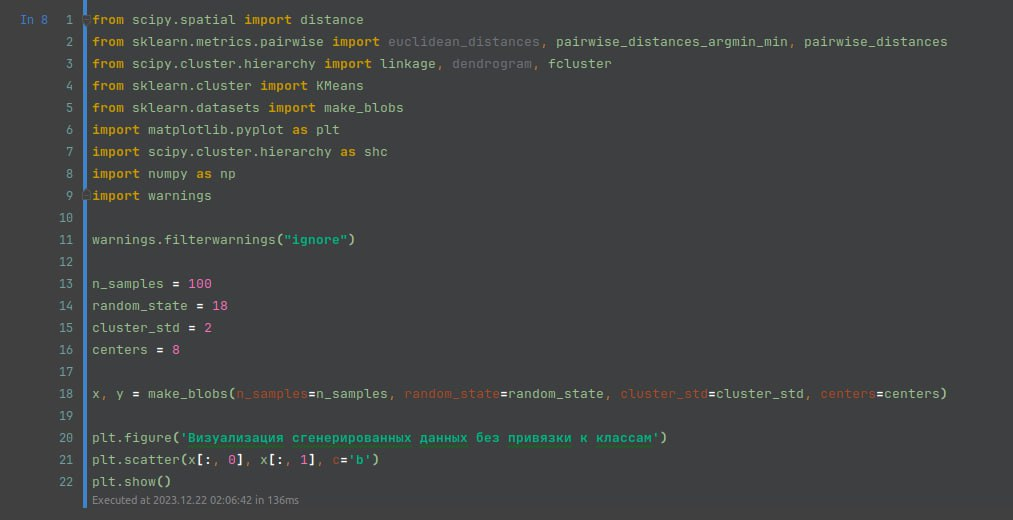


Рисунок 1 – Подключение библиотек и генерация точек

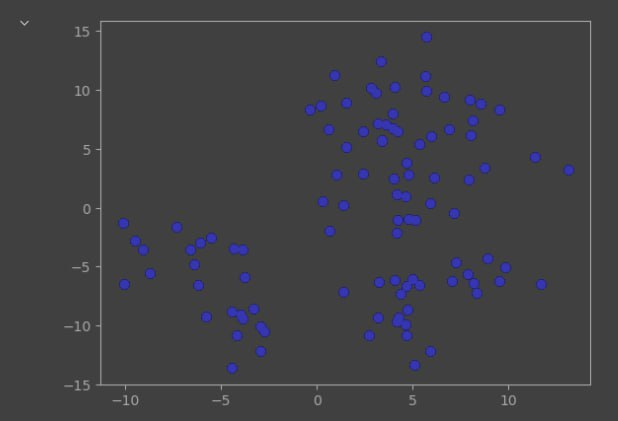


Рисунок 2 – Визуализация данных



Рисунок 3 – Иерархическая кластеризация

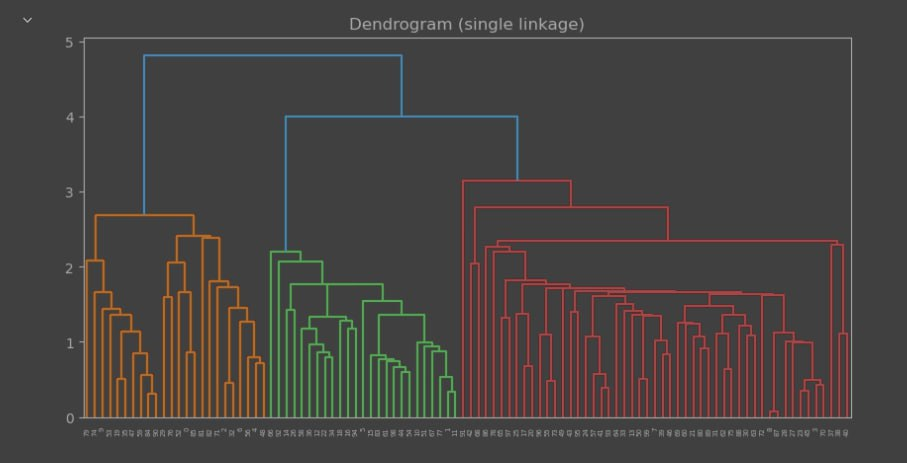


Рисунок 4 – Дендрограмма (single)

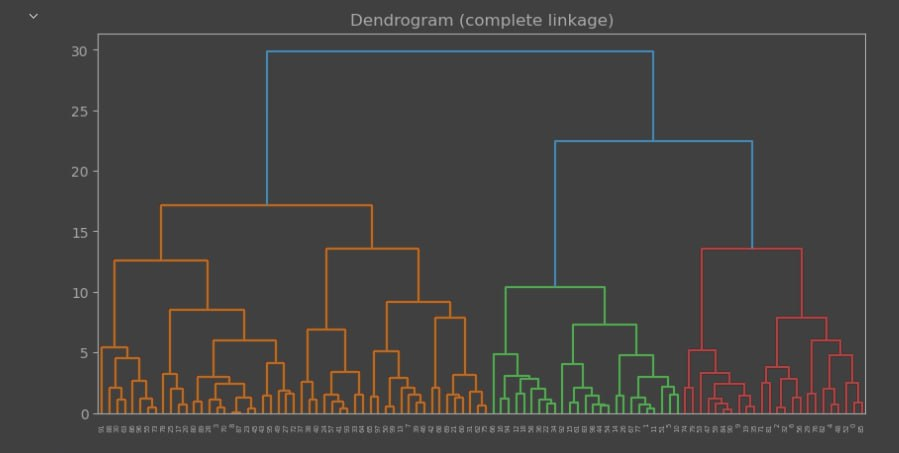


Рисунок 5 - Дендрограмма (complete)

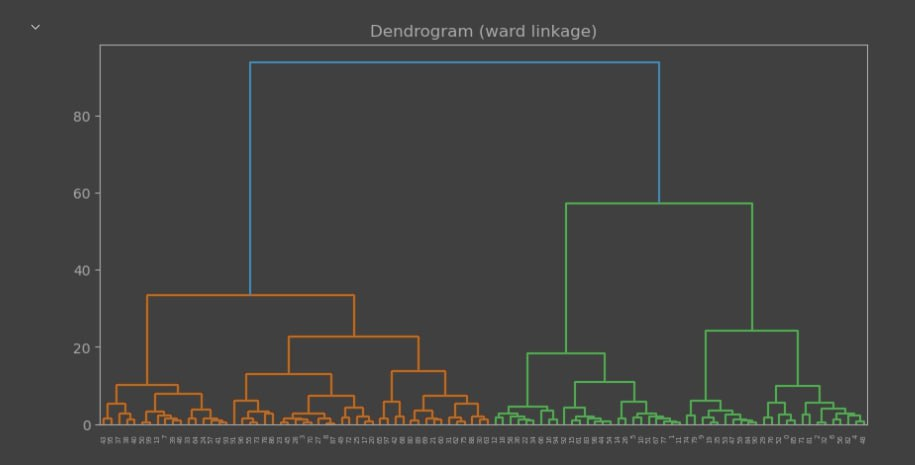


Рисунок 6 - Дендрограмма (ward)



Рисунок 7 – Визуализация разбиения на кластеры

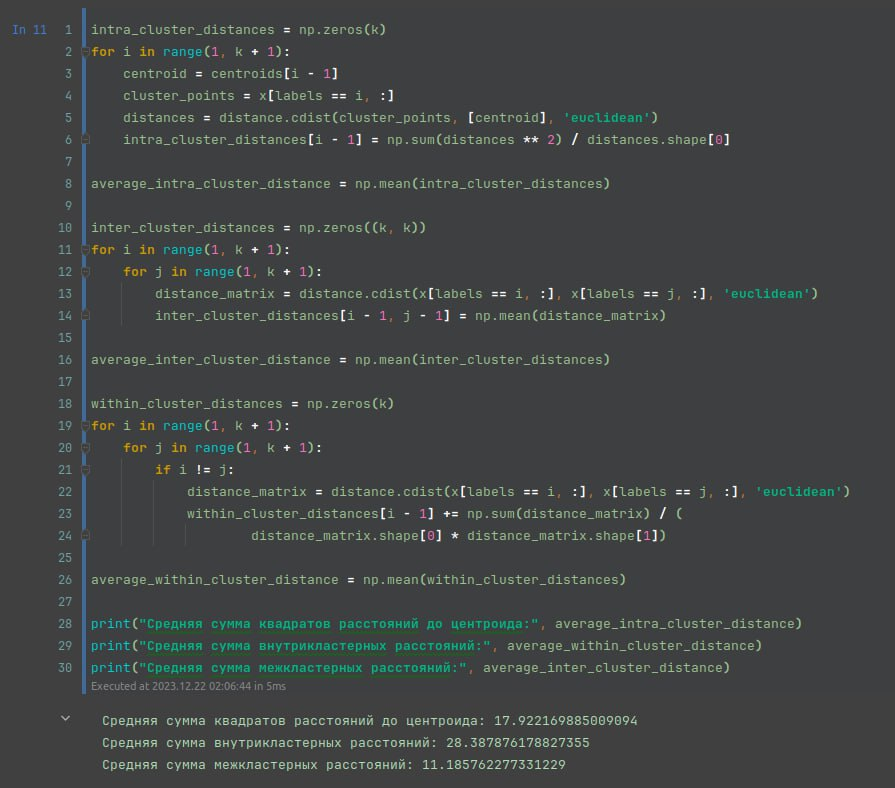


Рисунок 8 – Расчет параметров

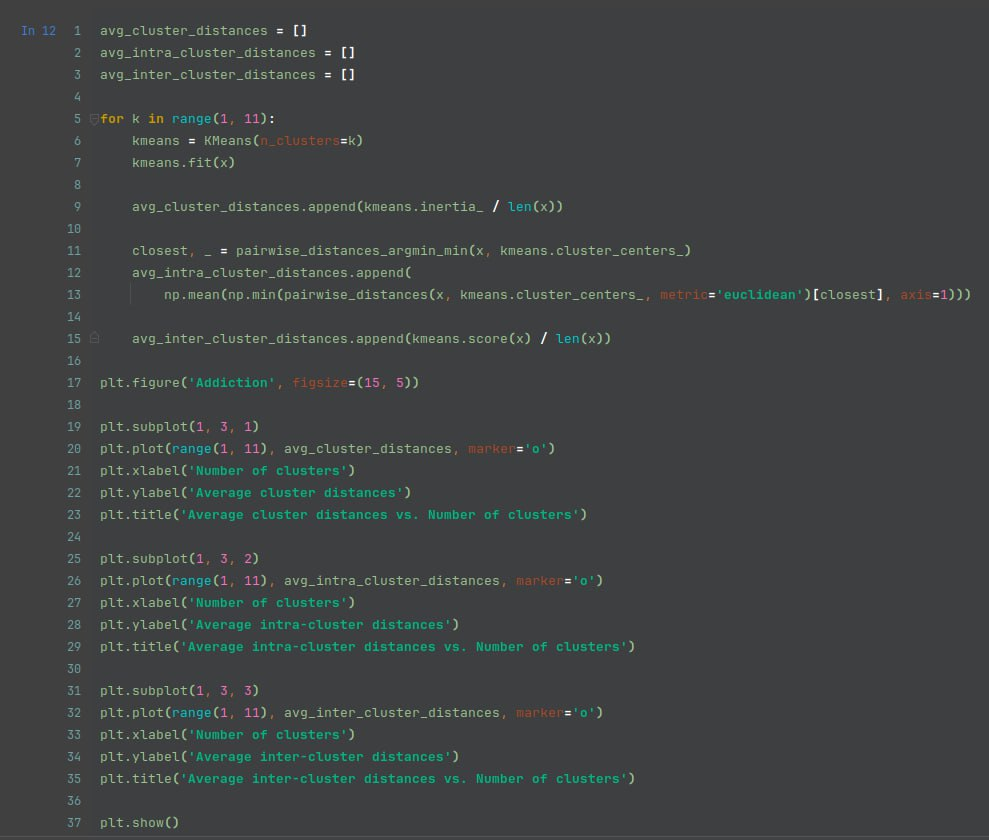


Рисунок 9 - Кластеризация

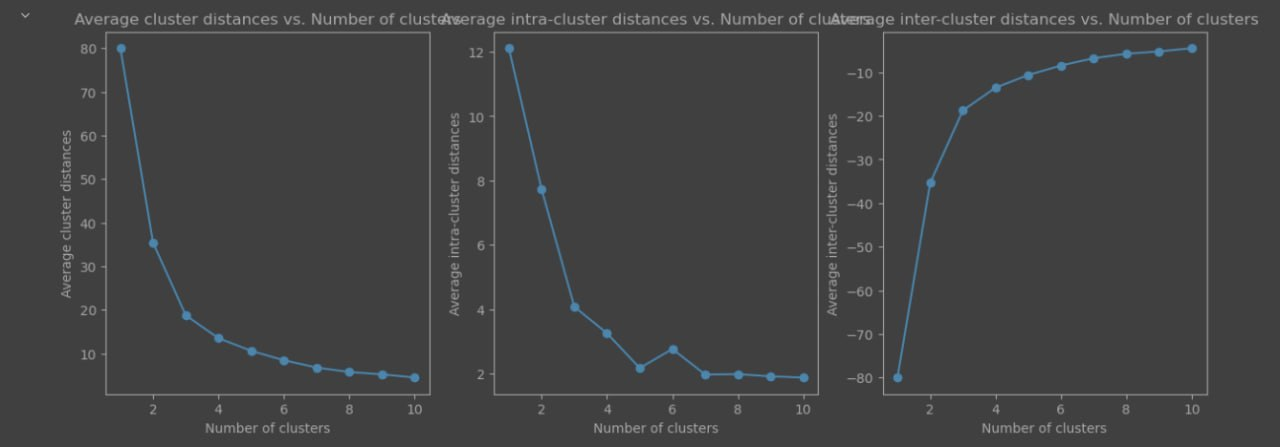


Рисунок 10 – Графики зависимостей

Вывод

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