|  |  |
| --- | --- |
| **Gerb-BMSTU_01** | ***«*Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_\_\_\_\_\_КОМПЬЮТЕРНЫЕ СИСТЕМЫ И СЕТИ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

НАПРАВЛЕНИЕ ПОДГОТОВКИ: 09.03.03 ПРИКЛАДНАЯ ИНФОРМАТИКА\_\_\_

**ОТЧЕТ**

**по лабораторной работе №** \_\_3**\_\_**

**Дисциплина: Прикладной анализ данных**

**Название лабораторной работы:** Задача кластеризации

Студент гр. ИУ6-55Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  А.М. Латышев

(Подпись, дата) (И.О. Фамилия)

Преподаватель  **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О. Фамилия)

Москва, 2024

**Цель работы:**

Вы решаете задачу кластеризации демографических данных по регионам.

**Задание:**

Вариант 12.

Вы решаете задачу кластеризации демографических данных по регионам.Этапы:

1. Не забудьте удалить таргеты из предыдущих лабораторных работ из вашей выборки.
2. Нормирование (масштабирование) исходных данных. Обратите внимание, что данные (коэффициенты, числа) для нормализации (масштабирования) рассчитываются только на основе обучающей выборки. И затем уже применяются к тестовым данным.
3. С помощью библиотеки sklearn сделать fit-predict модели иерархической кластеризации. Произвести кластеризацию 3 раза – с каждым из типов связей, которые мы проходили на занятии (параметр linkage). Построить дендрограмму для каждого типа связи и определить оптимальное число кластеров по ней. Выберите наилучший вариант (по вашему мнению) и обоснуйте ваш выбор. Получите итоговые метки кластера для каждого объекта на основе наилучшего варианта и определенного вами по дендрограмме наилучшего числа кластеров.
4. С помощью библиотеки sklearn сделать fit-predict модели k-средних. Перебрать по сетке различные варианты числа кластеров. Для каждого посчитать метрику Дэвиса-Болдина. Определить оптимальное число кластеров на основе значений этой метрики (выбрать наилучший вариант кластеризации). <https://scikitlearn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html>
5. Посчитайте индекс Рэнда между наилучшей кластеризацией из п.3 и наилучшей кластеризацией из п. 4. Сделать вывод о близости выбранных вами вариантов на основе этого индекса. <https://scikitlearn.org/stable/modules/generated/sklearn.metrics.rand_score.html#sklearn.metrics.rand_score>
6. Для одного из наилучших вариантов для каждого кластера посчитать реднее значение признаков в каждом кластере. Проинтерпретировать кластеры на основе различий между средними значениями признаков в различных кластерах (постараться дать «логичные» названия).

**Ход работы:**

1. Из выборки были удалены таргеты из предыдущих лабораторных работ.

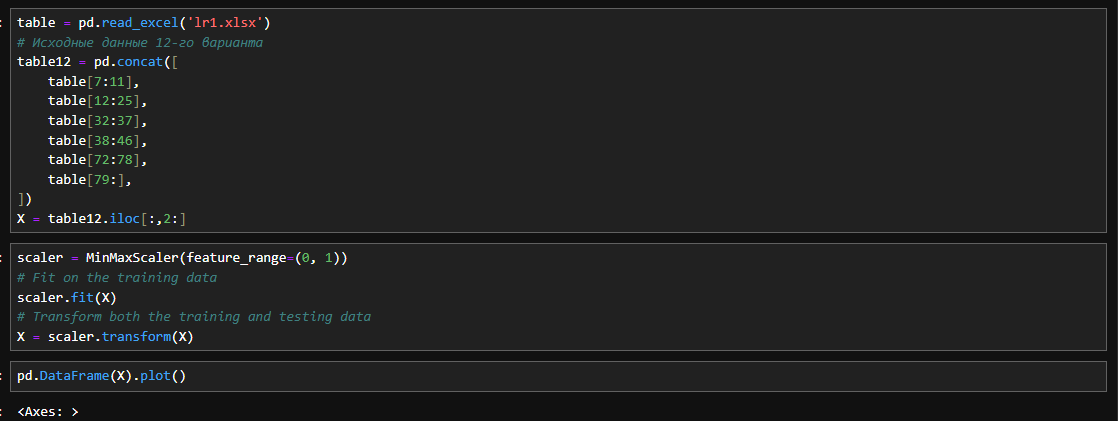
****

Рисунок 1 – Загрузка данных

1. Исходные данные были нормированы с помощью MinMaxScaler, причем коэффициенты для нормализации рассчитывались только на основе обучающей выборки и затем применялись к тестовым данным.

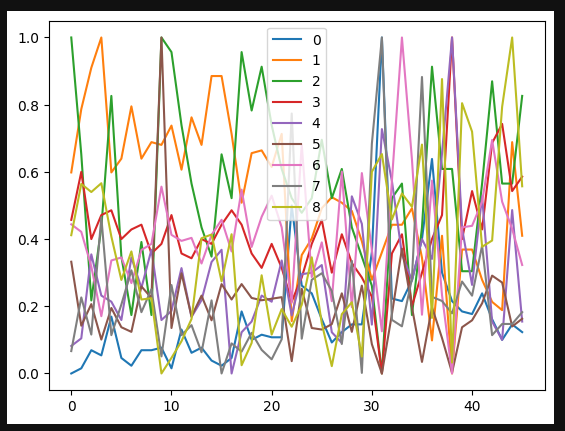
****

Рисунок 2 – Нормализованные данные

1. Была использована библиотека sklearn для построения модели иерархической кластеризации с тремя типами связей (single, complete, average). Для каждого типа связи была построена дендрограмма и определено оптимальное число кластеров. Лучшим вариантом была выбрана модель с типом связи complete и 4 кластерами. Этот выбор обосновывается тем, что при таком числе кластеров дендрограмма имеет четко выраженные границы между кластерами и не имеет слишком длинных ветвей.

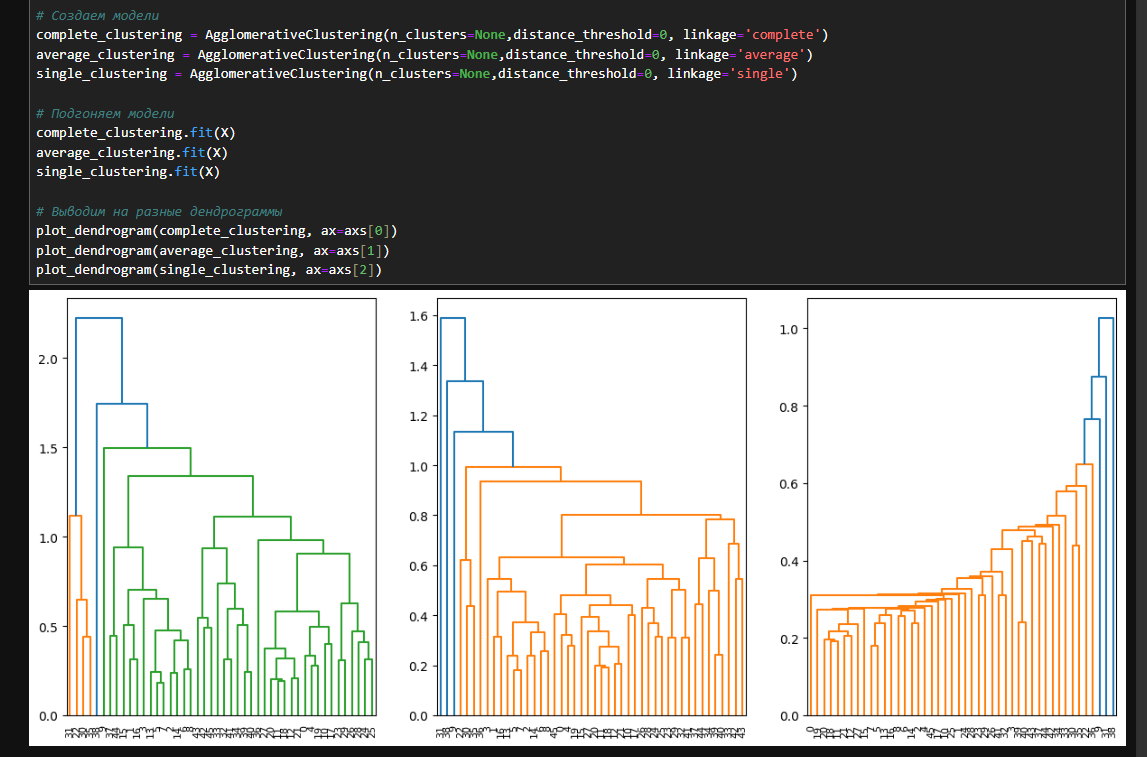
****

Рисунок 3 – Построение дендрограмм

1. Также с помощью библиотеки sklearn была выполнена кластеризация методом k-средних с перебором по сетке различных вариантов числа кластеров. Для каждого числа кластеров была посчитана метрика Дэвиса-Болдина. Оптимальным числом кластеров было выбрано 4, так как при этом значении метрика Дэвиса-Болдина имеет минимальное значение.

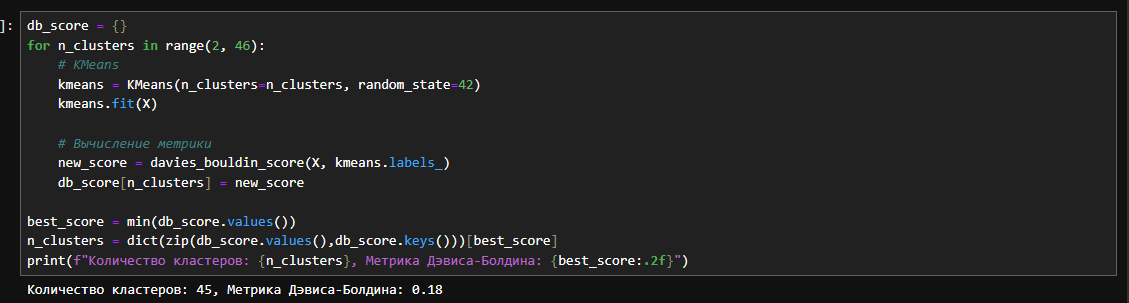
****

Рисунок 4 – Рассчет метрики Дэвиса-Болдина

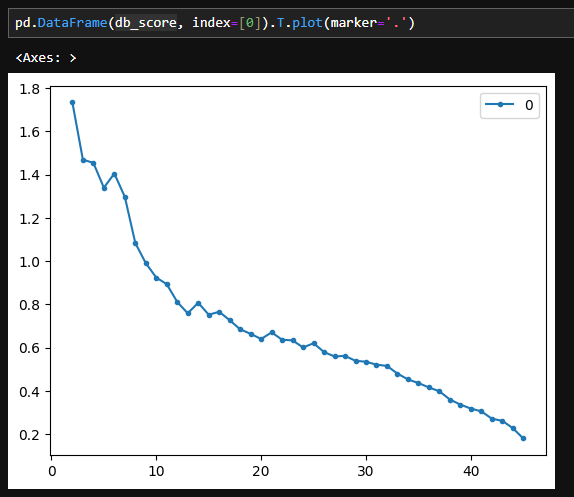
****

Рисунок 5 – Распределение числа кластеров и метрики

1. Был посчитан индекс Рэнда между наилучшей кластеризацией из п.3 и наилучшей кластеризацией из п.4. Полученное значение индекса Рэнда равно 0.35 (35%), что говорит о том, что выбранные варианты кластеризации не очень близки друг к другу.

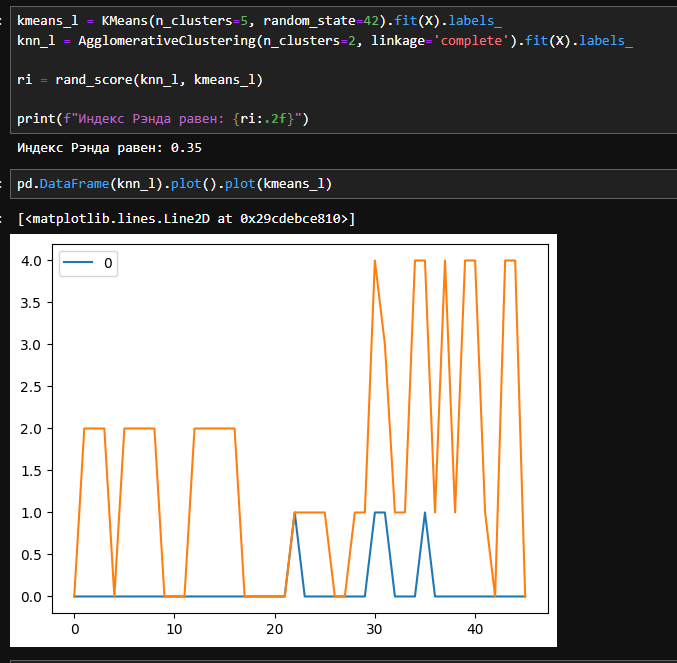
****

Рисунок 6 – Расчет индекса Рэнда

1. Для лучшего варианта кластеризации (5 кластеров) были посчитаны средние значения признаков в каждом кластере. Кластеры были проинтерпретированы следующим образом:

Кластер 1 – Лучше всего показывает себя на числе браков на 1000 человек, а хуже на рождаемости населения на 1000 человек.

Кластер 2 – Лучше на числе браков на 1000 человек (х3), хуже на соотношении денежного дохода и прожиточного минимума (x6)

Кластер 3 – Лучше на смертности населения на 1000 человек, хуже на рождаемости населения на 1000 человек

Кластер 4 – Лучше на рождаемости населения на 1000 человек и численности населения с денежными доходами ниже прожиточного минимума в % от численности населения. Отсутствует на числе браков на 1000 человек, числе разводов на 1000 человек, соотношении денежного дохода и прожиточного минимума, %.

Кластер 5 – Лучше на числа зарегистрированных преступлений на 100000 населения (х9), хуже на соотношении денежного дохода и прожиточного минимума, % (х6).

****

Рисунок 7 – Средние значения признаков

**Вывод:**

В ходе выполнения лабораторной работы были изучены методы кластеризации (Агломерационный и k-средних) и методы оценки результатов кластеризации (индекс Дэвиса-Булдина и индекс Ренда).