Министерство образования и науки Российской Федерации

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО

ПОЛИТЕХ
Санкт-Петербургский
политехнический университет
Петра Великого

Отчет по Лабораторной работе № 3.
по дисциплине “Машинное обучение”

Выполнила
студентка гр. 3530202/00201 Козлова Е. А.

Руководитель Селин И. А.

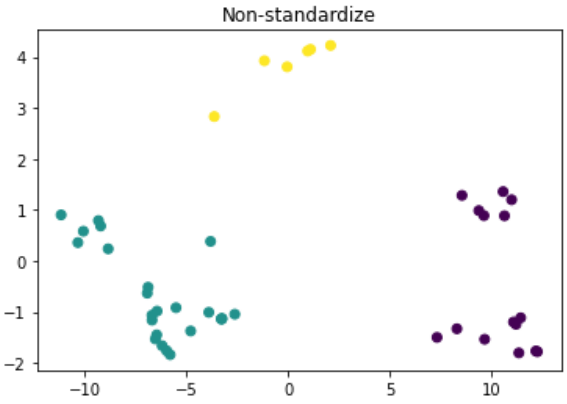
Санкт-Петербург
2023

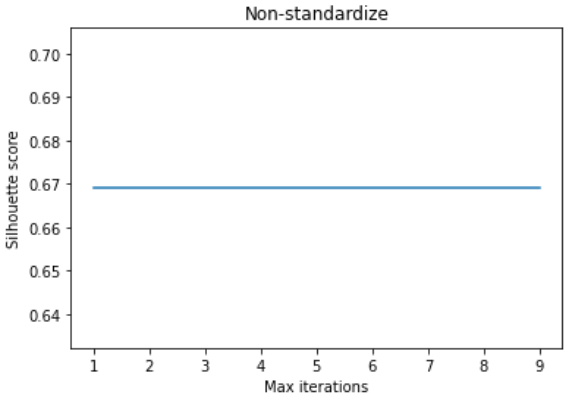
Оглавление

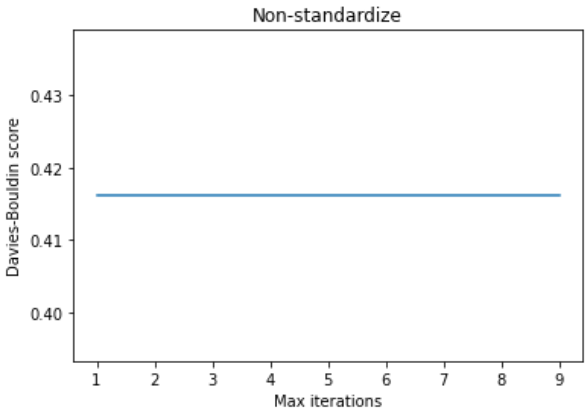
Задание 1

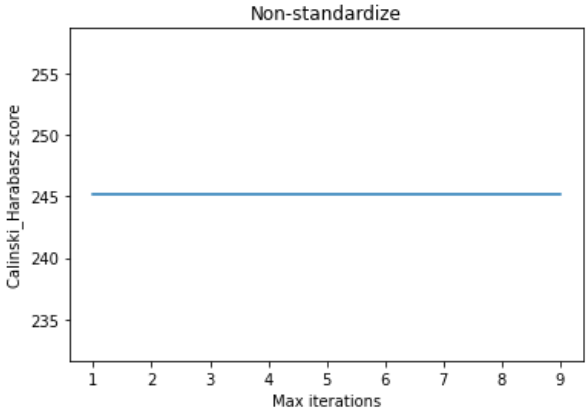
Разбейте множество объектов из набора данных pluton.csv на 3 кластера с помощью k-means. Сравните качество разбиения в зависимости от максимального числа итераций алгоритма и использования стандартизации.

Без стандартизации:

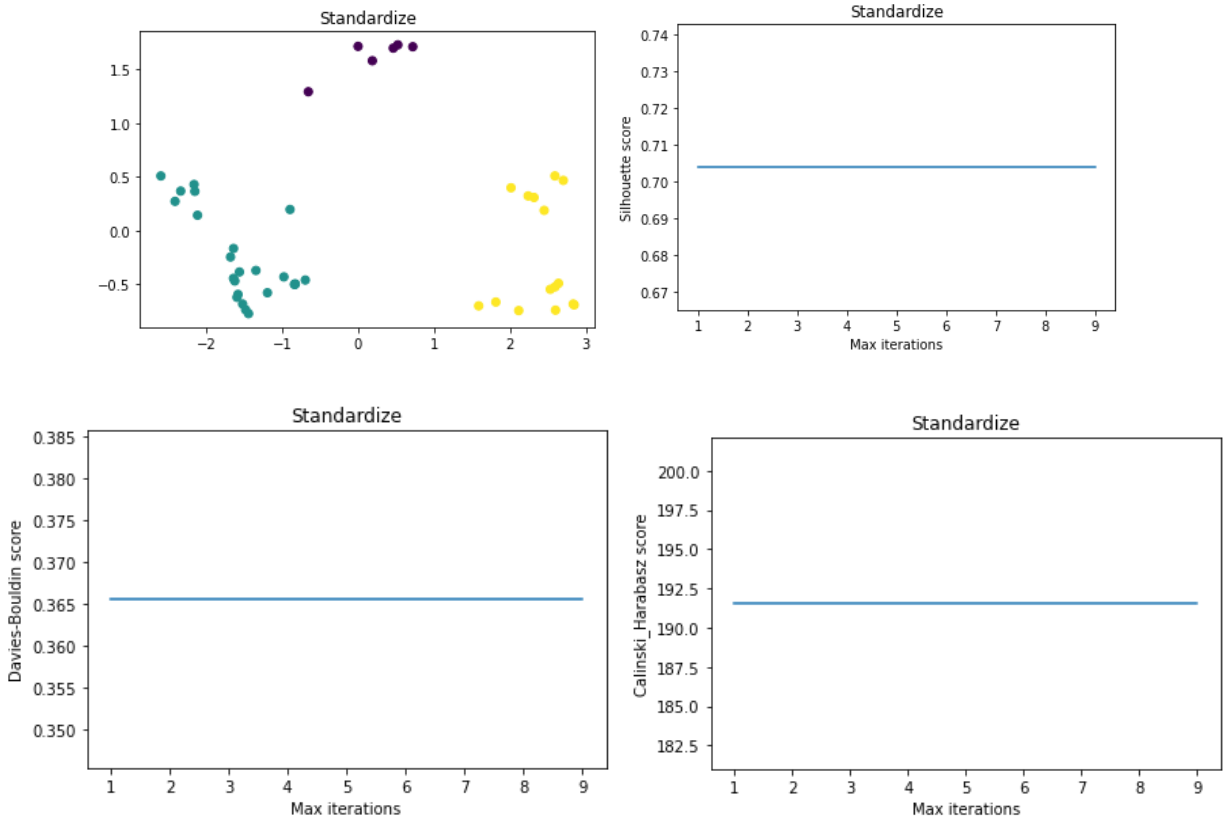








Со стандартизацией:

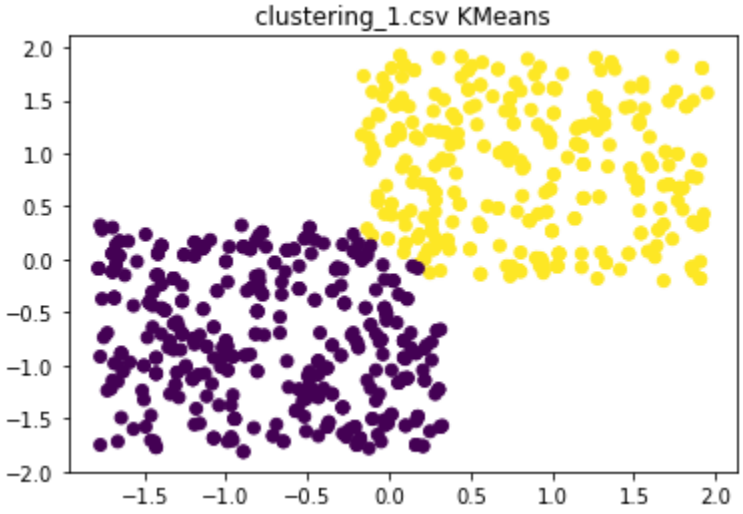


Вывод: из визуализации данных видно, что текущее число кластеров не оптимально, и увеличение их числа может значительно улучшить качество кластеризации. Однако увеличение максимального количества итераций не оказывает влияния на качество кластеризации. При сравнении метрик мы замечаем, что при использовании нестандартизированных данных лучший результат показывает индекс Calinski-Harabasz, в то время как при использовании стандартизированных данных наилучшие результаты достигаются с помощью индексов Дэвиса-Болдуина и Силуэтов.

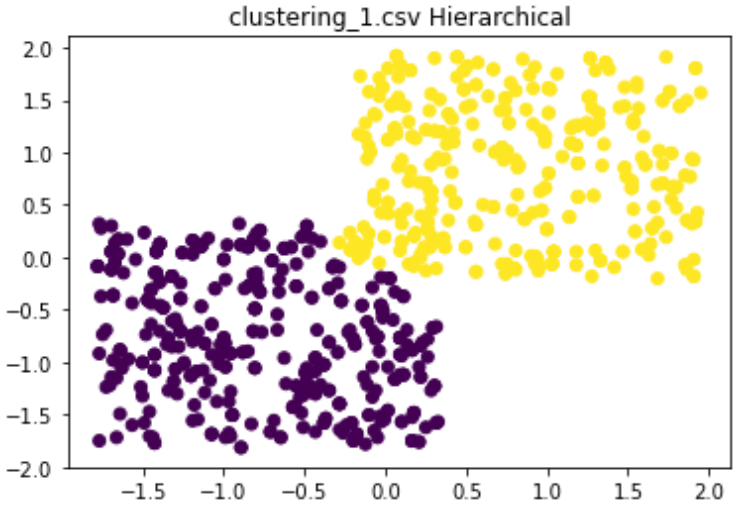
Задание 2

Разбейте на кластеры множество объектов из наборов данных clustering_1.csv, clustering_2.csv и clustering_3.csv с помощью k-means, DBSCAN и иерархической кластеризации. Определите оптимальное количество кластеров (где это применимо). Какой из методов сработал лучше и почему?

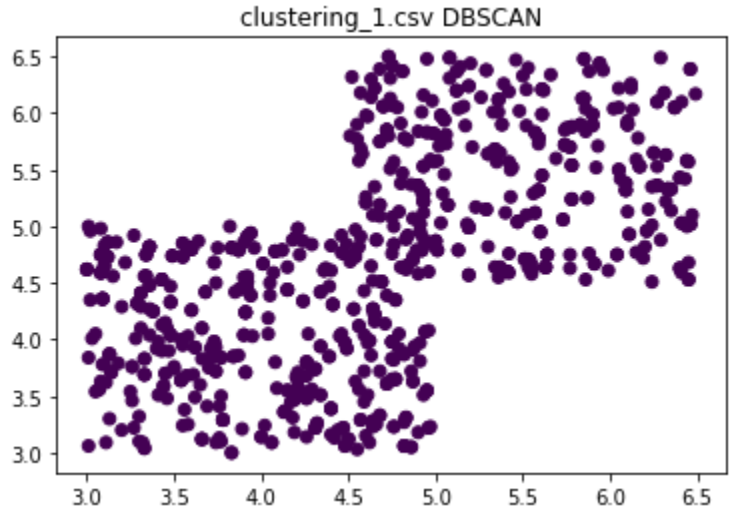
Clustering_1.csv
KMeans:



Hierarchical:



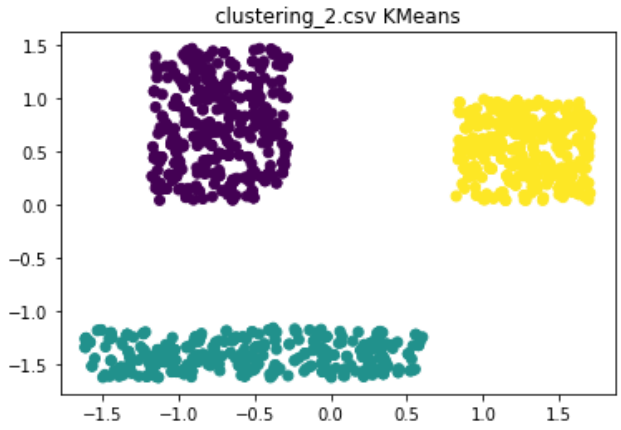
DBSCAN:



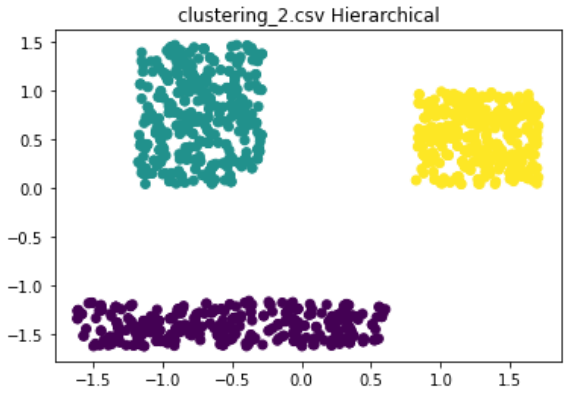
Оптимальное количество кластеров:
Исследование проводилось с помощью Silhouette score и результаты таковы:

- KMeans - 2
- Hierarchical - 2
- DBSCAN - 1

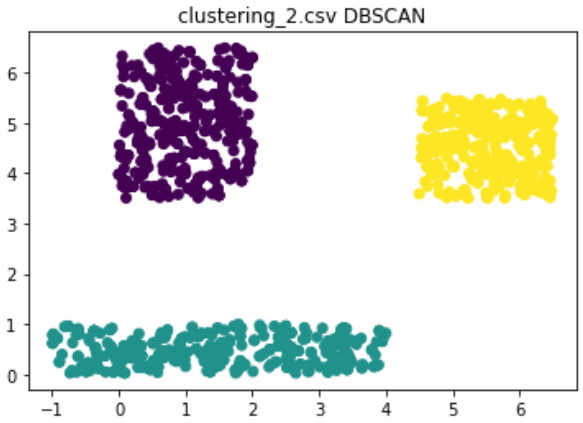
Clustering_2.csv
KMeans:



Hierarchical:



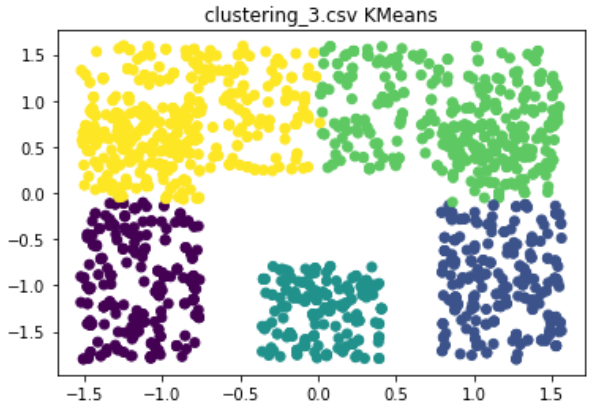
DBSCAN:



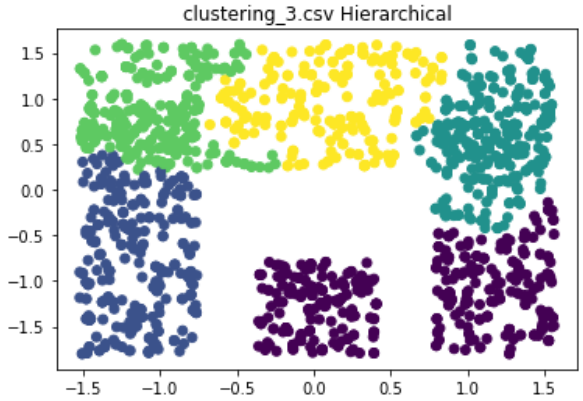
Оптимальное количество кластеров:
Исследование проводилось с помощью Silhouette score и результаты таковы:

- KMeans - 3
- Hierarchical - 3
- DBSCAN - 3

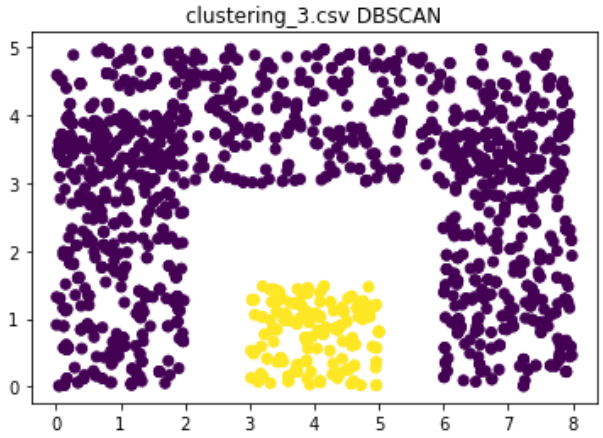
Clustering_3.csv
KMeans:



Hierarchical:



DBSCAN:



Оптимальное количество кластеров:
Исследование проводилось с помощью Silhouette score и результаты таковы:

- KMeans - 5
- Hierarchical - 5
- DBSCAN - 2

Вывод: из графиков становится очевидным, что DBSCAN может столкнуться с трудностями в выявлении кластеров, которые плотно соседствуют друг с другом. Эта проблема связана с тем, что DBSCAN определяет граничные точки, которые имеют меньше соседей, но при этом находятся близко к какой-либо внутренней точке. В то время как алгоритмы K-Means и иерархическая кластеризация справляются с этой задачей наилучшим образом.

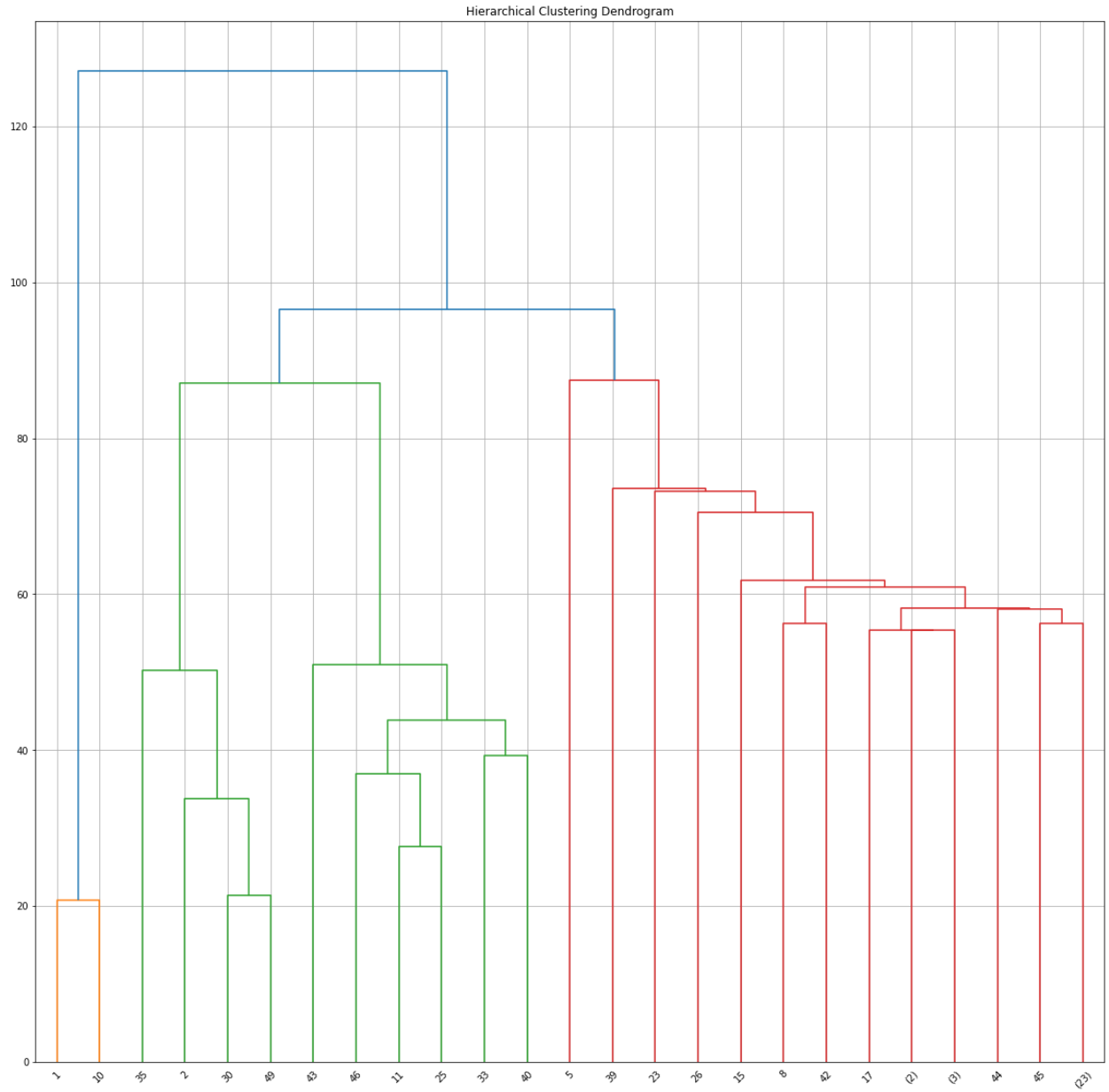
Задание 3

Осуществите сжатие цветовой палитры изображения (любого, на ваш выбор). Для этого выделите n кластеров из цветов всех пикселей изображения и зафиксируйте центра этих кластеров. Создайте изображение с цветами из сокращенной палитры (цвета пикселей только из центров выделенных кластеров). Покажите исходное и сжатое изображения.

Вывод: с помощью алгоритма K-Means было получено изображение с сокращенной палитрой цветов. Как мы можем заметить, из-за того, что в изначальном изображении присутствовало большее количество цветов, в полученном изображении была потеряна плавность перехода. Если посмотреть на изображение, то можно заметить полное изменение цветов со светлых на темные.

Задание 4

Постройте дендрограмму для набора данных votes.csv (число голосов, поданных за республиканцев на выборах с 1856 по 1976 год). Строки представляют 50 штатов, а столбцы - годы выборов (31). Проинтерпретируйте полученный результат.



Вывод: на данной дендрограмме прекрасно видно, что у нас образовалось 4 кластера. Так же можем увидеть, как эти кластеры объединяются или разделяются между собой.