ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

***«*САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ПЕТРА ВЕЛИКОГО»**

Институт компьютерных наук и технологий

**Высшая школа программной инженерии**

Изображение выглядит как текст

Автоматически созданное описание

**ЛАБОРАТОРНАЯ РАБОТА №3**

по дисциплине «Машинное обучение»

Студент А. М. Потапова

гр. 3530202/90202

Руководитель И. А. Селин

Санкт-Петербург

2022 г

**Содержание**

[Задание 1 2](#_Toc117672037)

[Задание 2 4](#_Toc117672038)

[Задание 3 7](#_Toc117672039)

[Задание 4 9](#_Toc117672040)

# **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.

**Ход работы**

* Используем модуль StandardScaler из библиотеки sklearn, чтобы создать стандартизированный датасет. Сравним, как выглядят данные в обычном и стандартизированном датасетах.

Обычный:

Изображение выглядит как текст

Автоматически созданное описание

Стандартизированный:

Изображение выглядит как текст

Автоматически созданное описание

* Сравним две модели k-means. Первая будет использовать нестандартизированные данные, а вторая – стандартизированные. Будем сравнивать результаты разбиения двух моделей в зависимости от максимального числа итераций. Для оценки качества разбиения будем использовать атрибут класса k-means под названием inertia. Выведем таблицу опытов (первый столбец – максимальное число операций, второй – inertia для нестандартизированных данных, третий - inertia для стандартизированных)

Изображение выглядит как текст

Автоматически созданное описание

**Вывод**

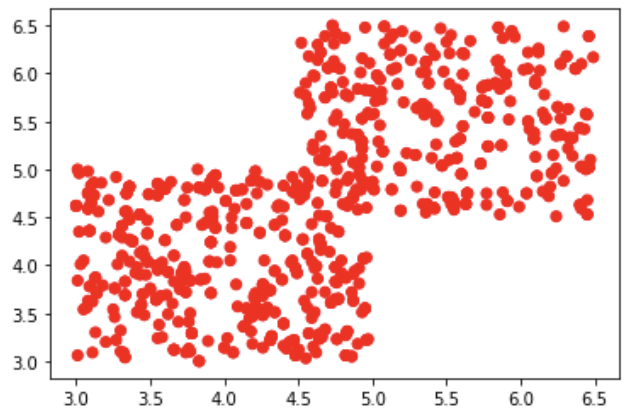
Исходя из полученных результатов можно заметить, что для наших моделей не изменяется метрика inetria, что говорит о быстрой сходимости. Также, заметим, что значение метрики inetria для стандартизированного датасета сильно меньше, чем для не стандартизированного, что также говорит о том, что предпочтительнее использовать стандартизированные данные.

# **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?

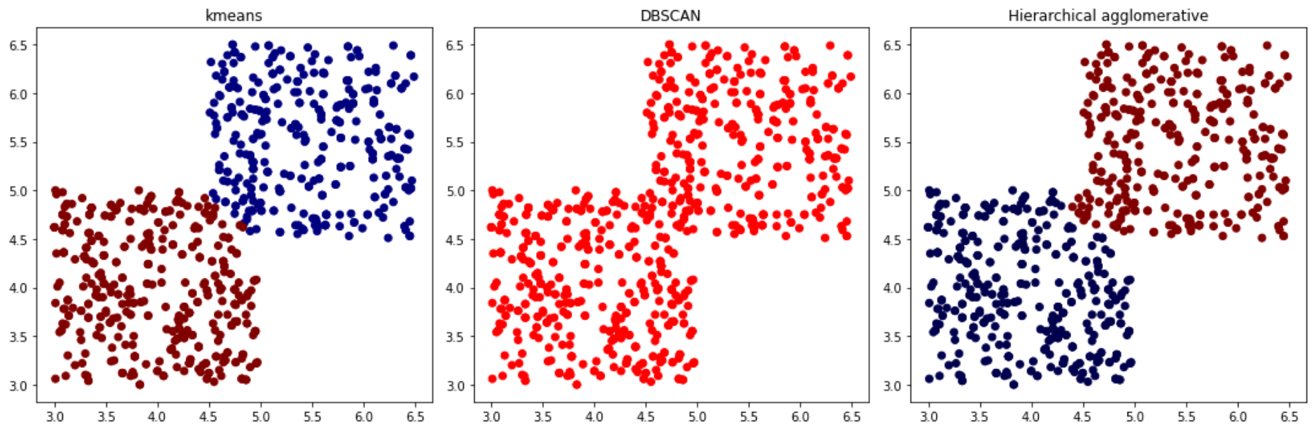
* Разбиение на кластеры множества объектов clustering\_1.csv.

*Изначальные данные:*



Предположим, оптимальным решением будет разбиение полученного датасета на два кластера.

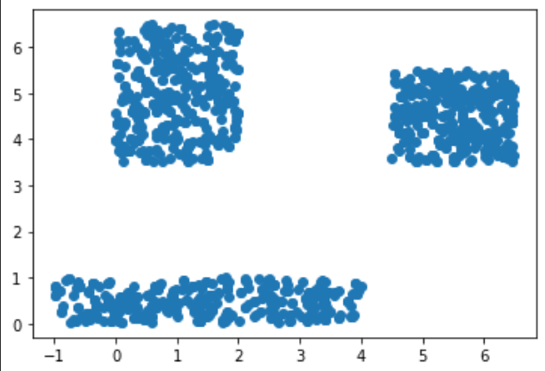
*Результаты разбиения:*



Как мы видим, k-means и иерархическая кластеризация справились со своей задачей хорошо, в то время как DBSCAN не смог выделить 2 кластера из-за того, что границы кластеров пересекались.

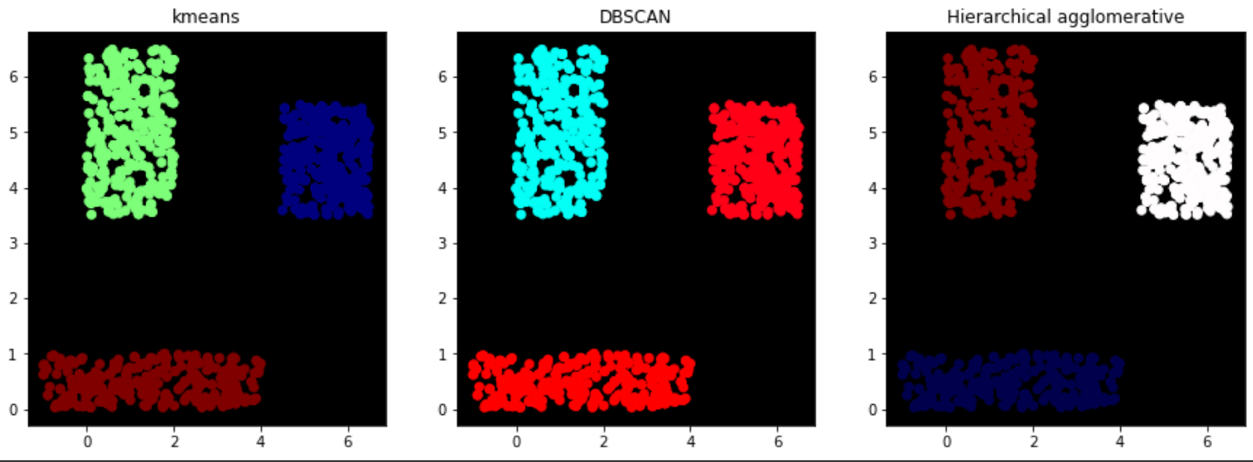
* Разбиение на кластеры множества объектов clustering\_2.csv.

*Изначальные данные:*



Предположим, оптимальным решением будет разбиение полученного датасета на три кластера.

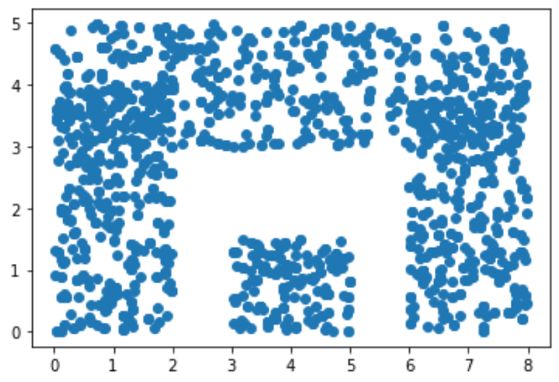
*Результаты разбиения:*



В этом случае все методы отлично справились с разбиением.

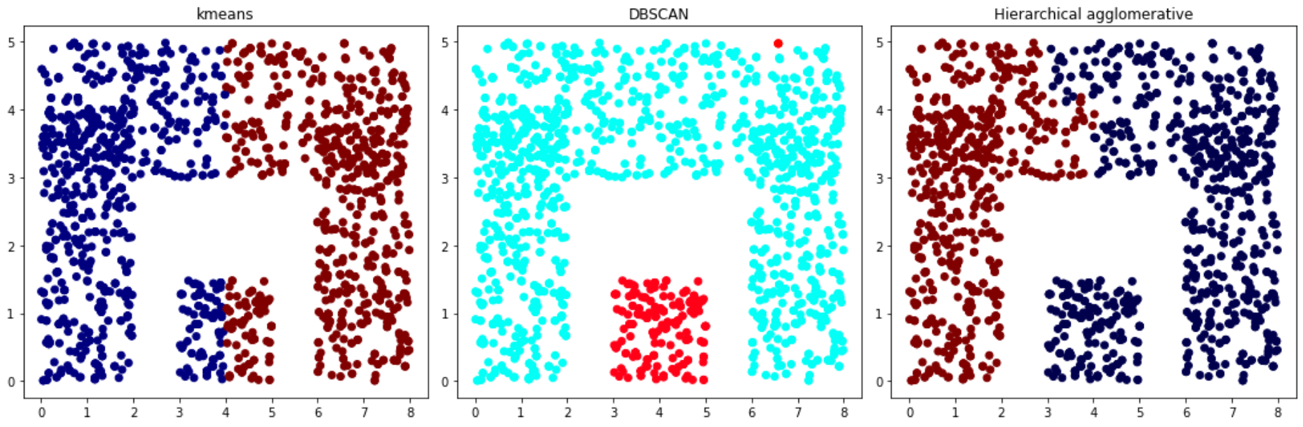
* Разбиение на кластеры множества объектов clustering\_2.csv.

*Изначальные данные:*



Предположим, оптимальным решением будет разбиение полученного датасета на два кластера (один окружен другим).

*Результаты разбиения:*



Заметим, ни один из методов не справился с данной задачей идеально. K-means и иерархическая кластеризация смогли выделить 2 кластера, но кластеры оказались не такими, какими мы ожидали. Метод DBSCAN выделил кластеры почти так, как мы хотели, но поместил лишний элемент в другой кластер.

**Вывод**

На датасетах clustering\_1 и clustering\_2 методы k-means и иерархическая кластеризации показали себя одинаково хорошо. В датасете clustering\_3 оба эти метода выделили 2 кластера, но не такие, какие мы предполагали. Метод DBSCAN показал себя хорошо на датасетах clustering\_2 и clustering\_3, но на датасете clustering\_1 не смог выделить 2 кластера из-за пересечения. У нас не получилось выделить лучший метод. Целесообразность использования того или иного метода зависит от начального датасета, который мы хотим разбить на кластеры.

# **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.

*Исходное изображение:*



*Изображение с цветами из сокращенной палитры при выделении 10 кластеров:*

Изображение выглядит как текст

Автоматически созданное описание

*Изображение с цветами из сокращенной палитры при выделении 30 кластеров:*

Изображение выглядит как текст, укладывает, внутренний, млекопитающее

Автоматически созданное описание

*Изображение с цветами из сокращенной палитры при выделении 100 кластеров:*

Изображение выглядит как укладывает, внутренний, млекопитающее, спит

Автоматически созданное описание

**Вывод**

Исходя из полученных результатов можно сделать вывод, что использование палитры из 100 цветов вполне достаточно, чтобы приблизиться к исходному изображению.

# **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.

Данные полученные из votes.csv:

Изображение выглядит как текст, монитор, черный, закрыть

Автоматически созданное описание

Поскольку при построении дендрограммы возникла ошибка, в связи с наличием пустых значений, заполним их средними для текущего года значениями по штатам. Получили следующие данные:

Изображение выглядит как текст, монитор, закрыть, снимок экрана

Автоматически созданное описание

Для построения дендрограммы воспользуемся функцией dendrogram из библиотеки sklearn.

Изображение выглядит как текст

Автоматически созданное описание

*Полученная дендрограмма*



**Вывод**

Исходя из полученных результатов можно сделать вывод, что чем меньше расстояние между двумя штатами, тем меньше разница в голосах за все года. На основе этого происходит разделение штатов на группы. В части групп голоса за республиканцев доминируют, в другой части - нет.