МГТУ им. Н. Э. Баумана, кафедра ИУ5

курс “Технологии машинного обучения”

Рубежный контроль №2

# «Технологии использования и оценки моделей машинного обучения»

ВЫПОЛНИЛ:

Фонканц Р.В.

Группа: ИУ5-61Б

Вариант: 23

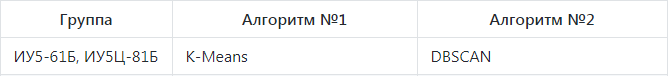
ПРОВЕРИЛ:

Гапанюк Ю.Е.

Москва 2020

**Задача 2. Кластеризация данных:**

Кластеризуйте данные с помощью двух алгоритмов кластеризации (варианты по группам приведены в таблице):



Сравните качество кластеризации с помощью следующих метрик качества кластеризации (если это возможно для Вашего набора данных):

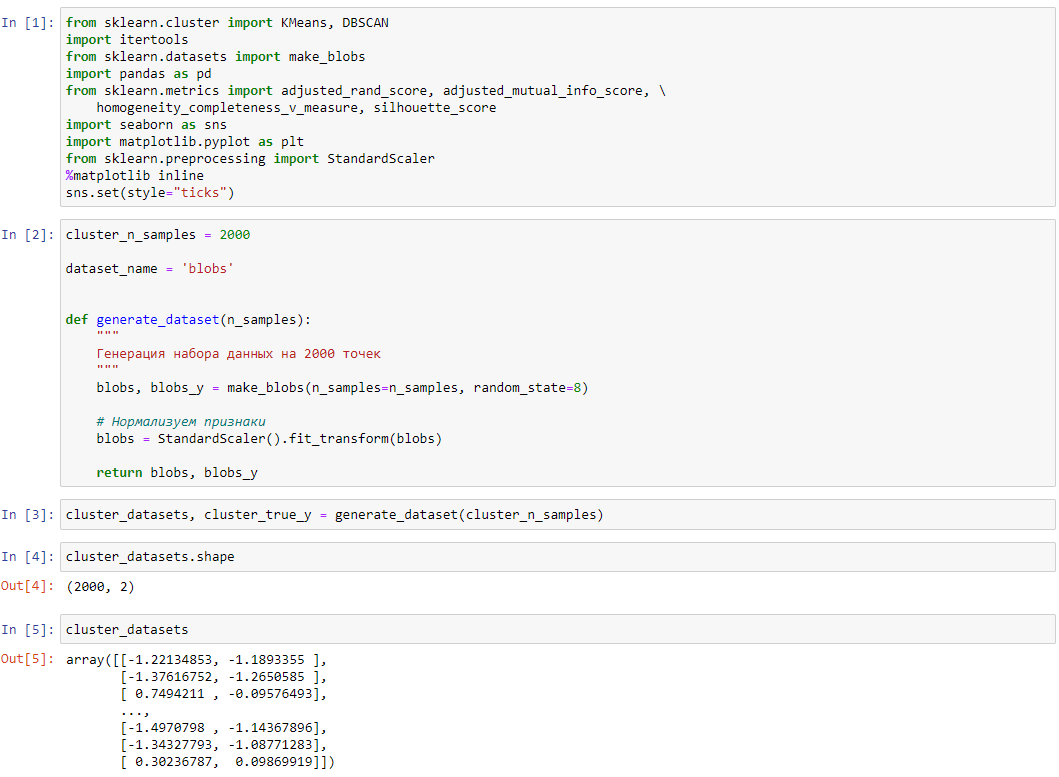
1. Adjusted Rand index Adjusted Mutual Information Homogeneity,
2. completeness,
3. V-measure,
4. Коэффициент силуэта.

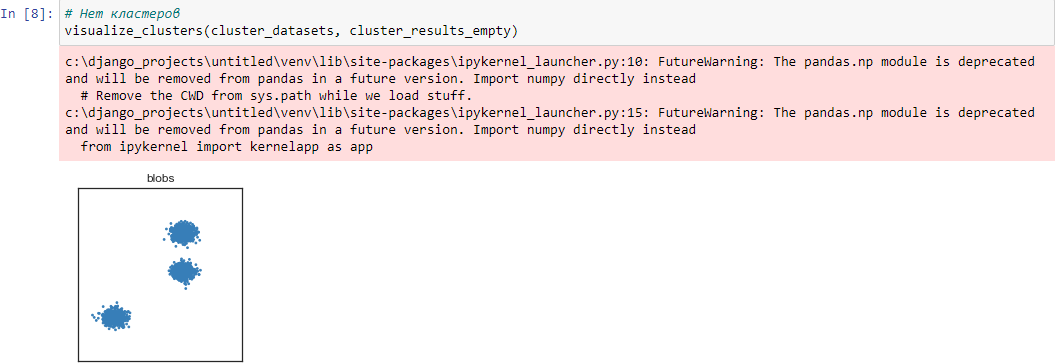
Сделайте выводы о том, какой алгоритм осуществляет более качественную кластеризацию на Вашем наборе данных.

**Набор данных (№23):** Генерация набора данных с использованием функции datasets.make\_blobs библиотеки scikit-learn на 2000 точек. Форма кластеров - 'blobs'.

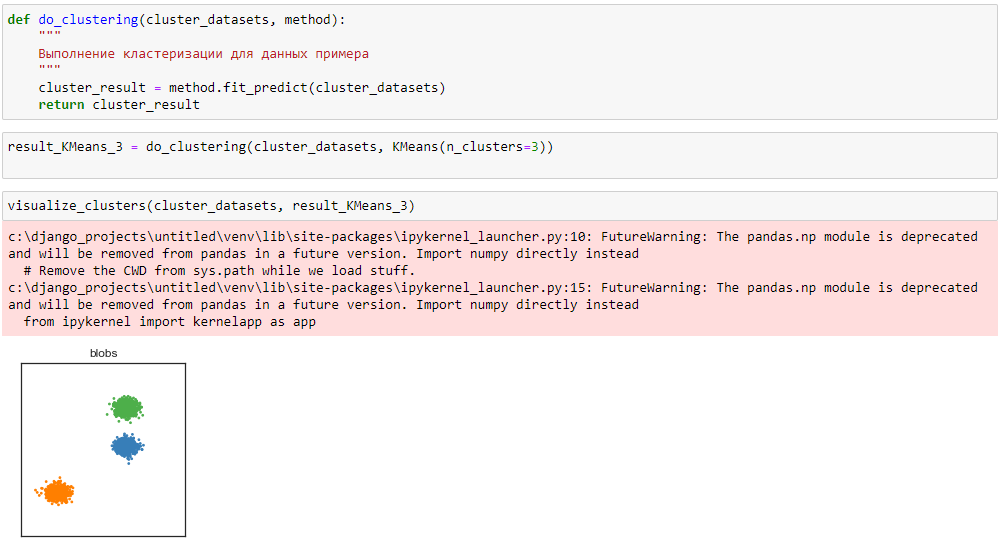
**Выполненная работа:**

## Генерация датасета, его первичный анализ и визуализация:

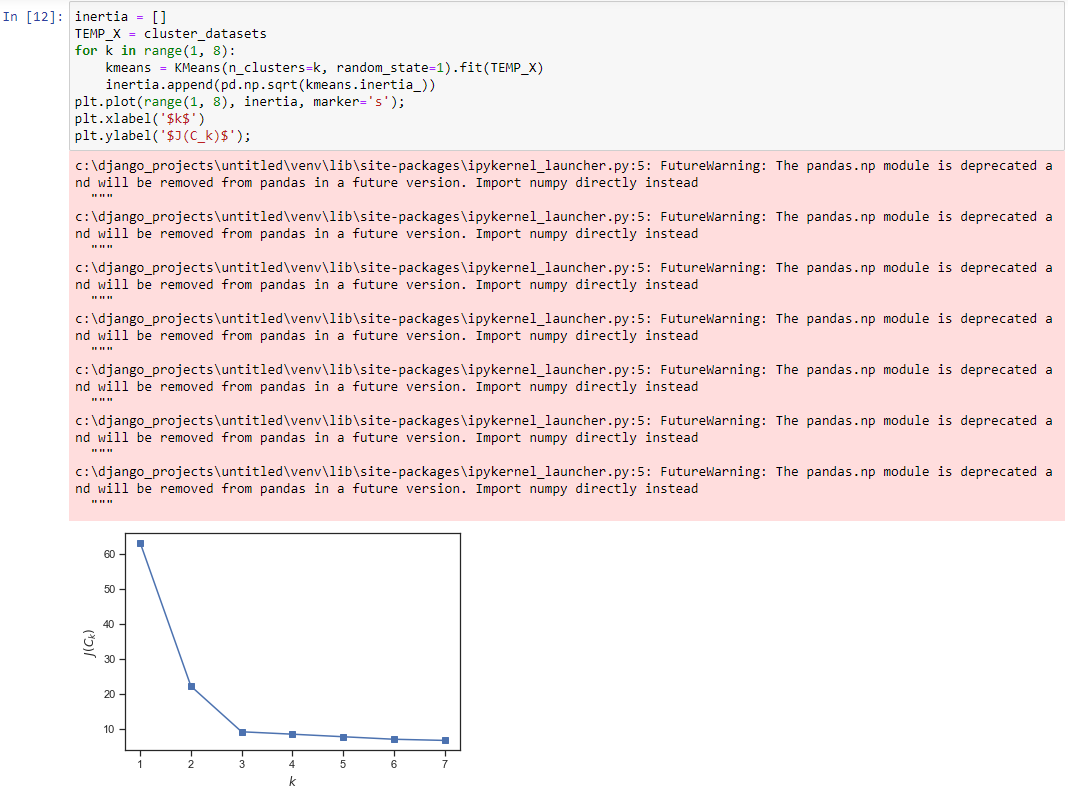


## 2. Кластеризация данных с помощью алгоритма K-Means:



В качестве гиперпараметра K (количество кластеров) было выбрано число 3. Проверим, какое значение будет эффективным на самом деле:

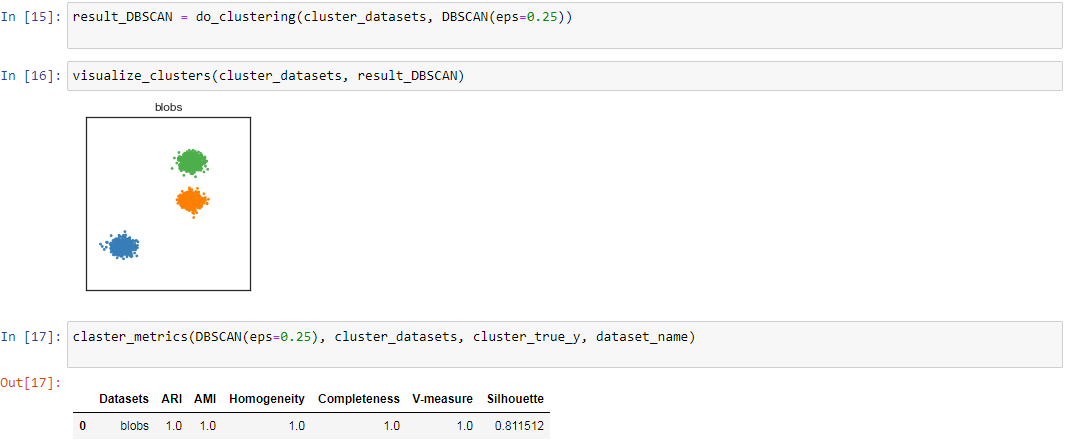


Как видно, число 3 является наиболее оптимальным значением гиперпараметра, а значит оставляем все, как есть.



## 3. Кластеризация данных с помощью алгоритма DBSCAN:

Зададим параметр eps (максимальный радиус соседства) равным 0.25. В результате получим:



В данной работе были применены два алгоритма кластеризации данных: K-Means и DBSCAN. Анализируя результаты, нельзя сказать, какой из двух алгоритмов обеспечивает более качественную кластеризацию. Полученные метрики обоих алгоритмов, как видно, имеют одинаково идеальные значения, что говорит скорее о качественном исходном наборе данных.

Данный отчет размещен в репозитории: <https://github.com/Rauger474/TMO>