МГТУ им. Н. Э. Баумана

кафедра ИУ5

курс «Технологии машинного обучения»

Рубежный контроль №2

**«Технологии использования и оценки моделей**

**машинного обучения»**

Вариант 22

ВЫПОЛНИЛ:

Сукиасян В. М.

Группа ИУ5-61Б

ПРОВЕРИЛ:

Гапанюк Ю. Е.

Москва, 2020 г.

|  |  |  |
| --- | --- | --- |
| **Номер варианта** | **Номер задачи** | **Номер набора данных, указанного в задаче** |
| 22 | 1 | 6 |

**Задача №2.** Кластеризация данных(по вариантам).



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

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

**Наборы данных:**

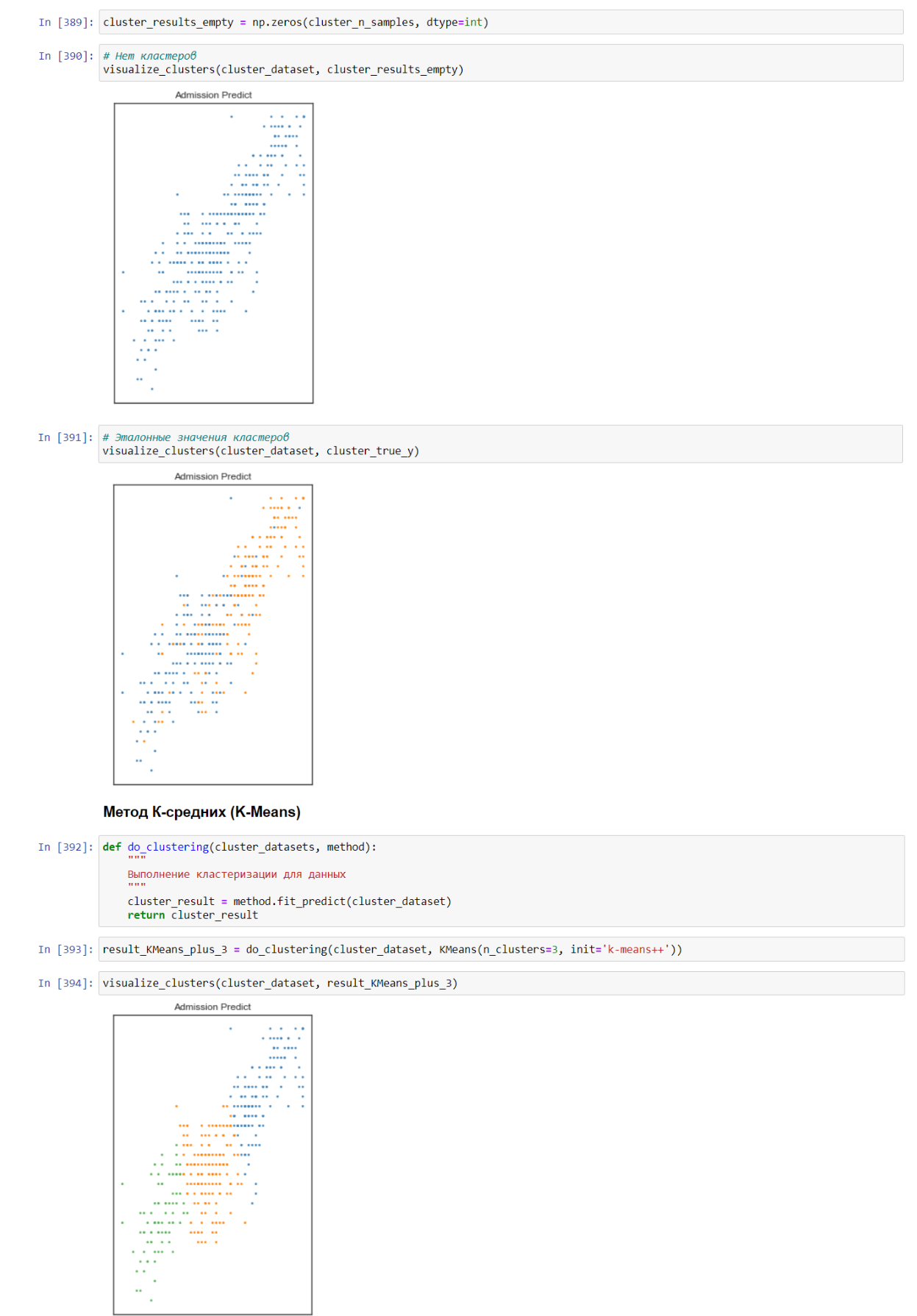
1. <https://www.kaggle.com/mohansacharya/graduate-admissions>

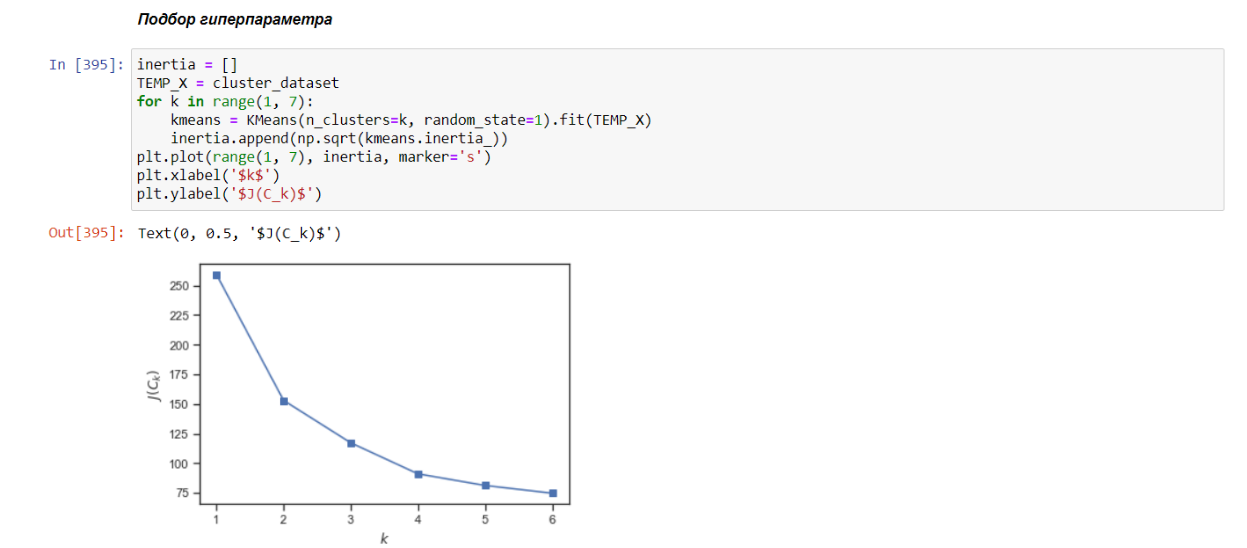
файл *Admission\_Predict.csv*

**Выполнение работы**



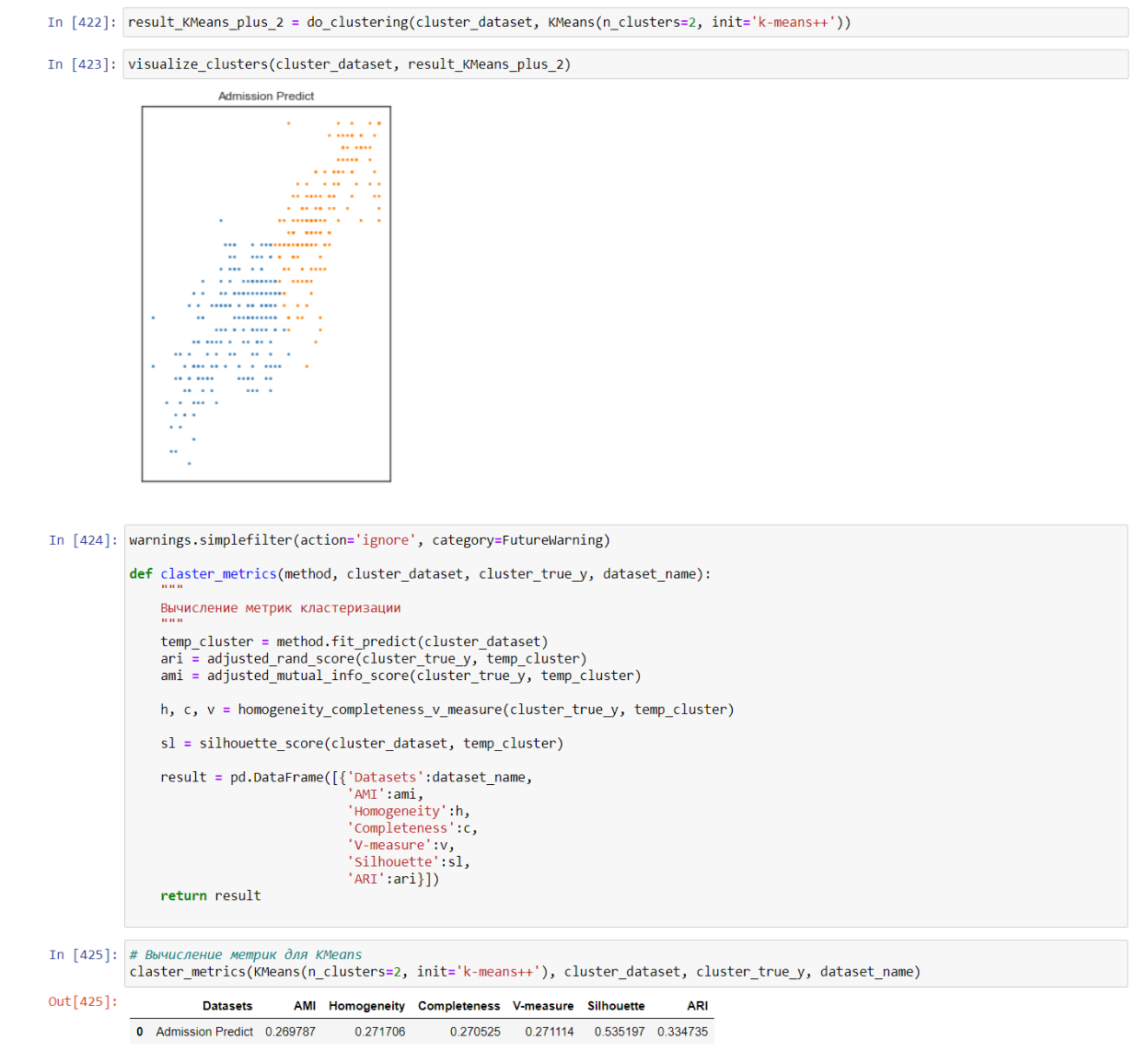


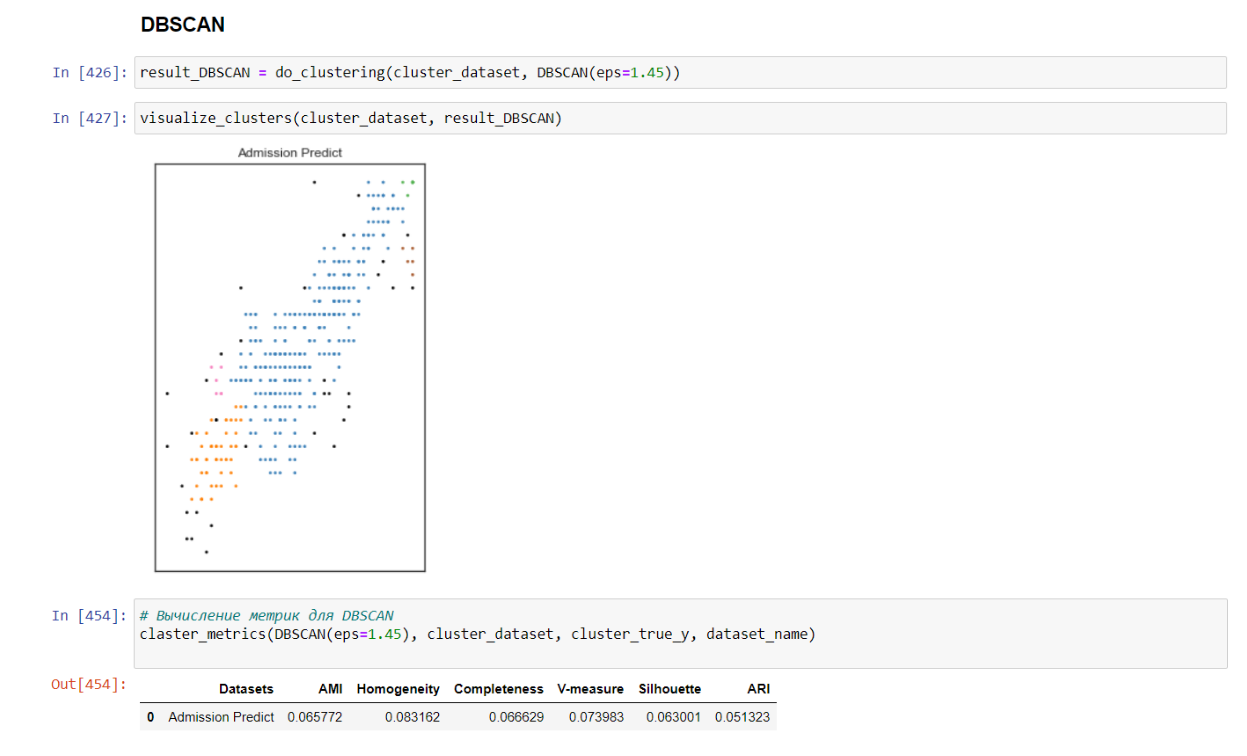




* качестве гиперпараметра K (количество кластеров) было выбрано число 3. Как видно по графику, после K = 2 уменьшение инерции замедляется.

Результат с количеством кластеров К = 2:





**Метрики для оценки качества кластеризации:**

*Adjusted Rand Index* –применяется,когда известны истинные метки классов.Отчасти метрика напоминает accuracy, так как сравнивает полученные метки классов с известными истинными классами.

*Adjusted Mutual Information*

*Homogeneity* -каждый кластер содержит только представителей

единственного класса (под классом понимается истинное значение метки

кластера).

*Completeness* -все элементы одного класса помещены в один и тот жекластер.

*V-measure* –среднее гармоническое отHomogeneityиCompleteness. *Коэффициент силуэта* –не требует знания истинных значений метоккластеров. Показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров.

**Вывод:**

Сравнивая полученные метрики двух алгоритмов кластеризации – K-Means и DBSCAN, можно сказать, что оба этих алгоритма далеки от идеальных для текущего датасета, но K-Means лучше справился с задачей разбиения исходного набора данных на кластеры.