|  |
| --- |
|  |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА - Московский технологический университет»**  **РТУ МИРЭА** |

Институт Информационных Технологий

Кафедра Прикладной Математики (ПМ)

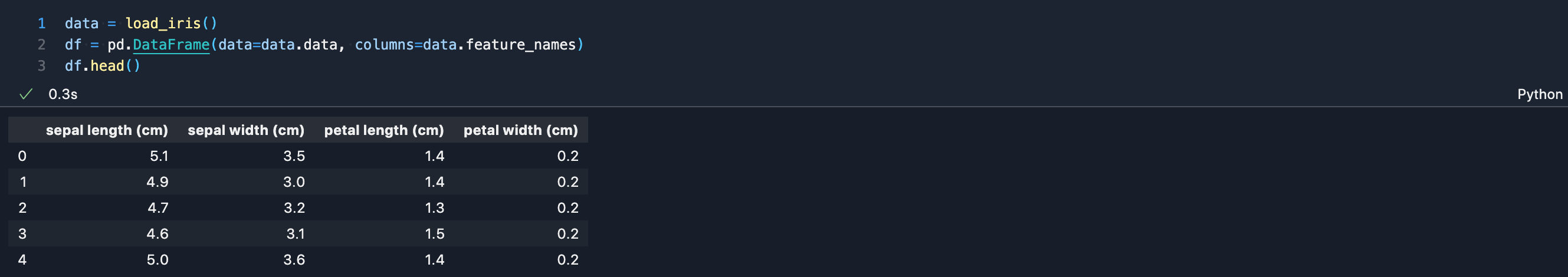
**ПРАКТИЧЕСКАЯ РАБОТА № 11**

|  |  |
| --- | --- |
| Выполнил студент группы ИКБО-08-19  Борисов А.В. | (\_\_\_\_\_\_\_\_\_\_)  *подпись* |
| Принял Ассистент кафедры ПМ  Высоцкая А.А. | (\_\_\_\_\_\_\_\_\_\_)  *подпись* |
| Практическая работа выполнена | « » 2022 г. |
| «Зачтено» | « » 2022 г. |

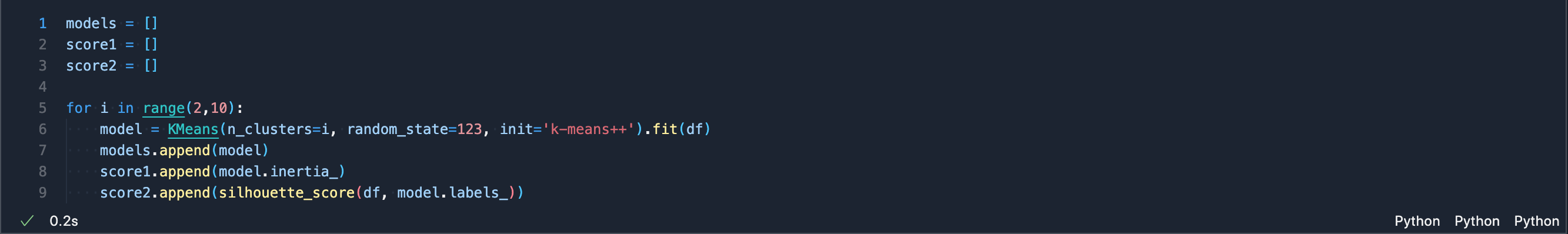
Москва 2022

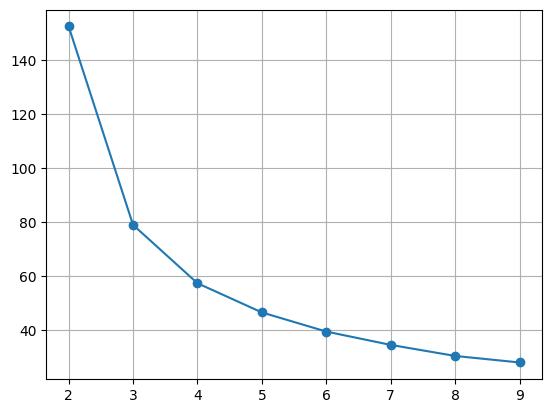
1. ***Найти данные для кластеризации.***

Для кластеризации были выбраны данные о размерах ирисов:

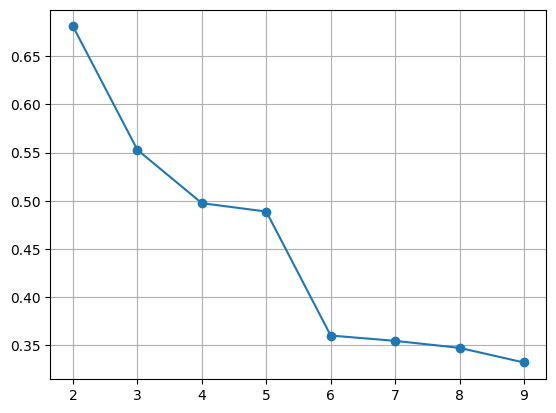
******

1. ***Провести кластеризацию данных с помощью алгоритма k-means. Использовать «правило локтя» и коэффициент силуэта для поиска оптимального количества кластеров.***

******

******

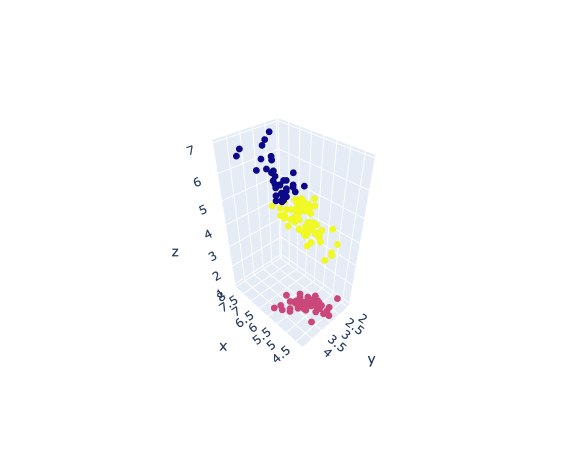
Исходя из правила локтя можно сказать, что в данном наборе данных есть 3 отдельных кластера (так оно и есть, класса в нем 3).

******

А вот коэффициент силуэта максимален при разделении исходного набора данных на два кластера. Это можно объяснить тем, что два класса в данном наборе довольно схожи (это и подтверждает график далее, где синий и желтый классы довольно близко друг к другу), однако возьмем число известное классов – 3.

******

После применения алгоритма k-means был построен трехмерный точечный график, который показывает распределение признаков классов по трем осям координат, и на основе данных полученных в результате работы данного алгоритма на графике были отмечены цветами 3 разных класса. Отчетливо видно, что два из трех классов очень схожи по своим признакам, поэтому коэффициент силуэта наибольший при разделении на два кластера.

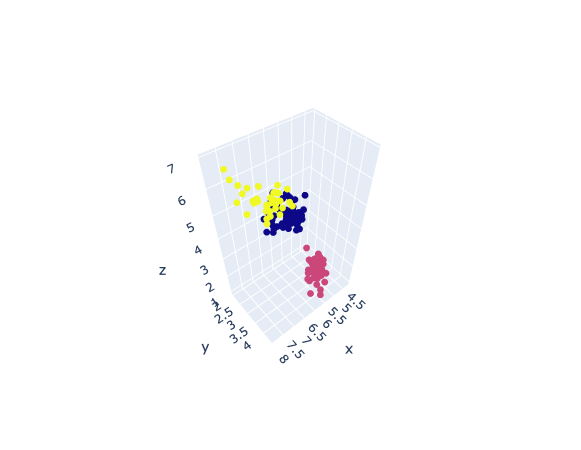
******

1. ***Провести кластеризацию данных с помощью алгоритма иерархической кластеризации.***

Для реализации алгоритма иерархической кластеризации тоже было выбрано распределение на три класса, после был построен трехмерный график и на нем были отмечены различными цветами классы, распределение которых мы получили после работы алгоритма.

В целом, результат работы очень схож на таковой у предыдущего алгоритма: тоже два очень близких по параметрам класса трудно отличимы друг от друга.

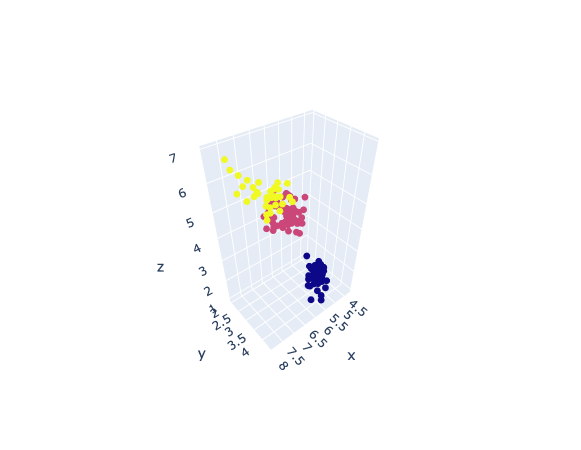
******

******

1. ***Провести кластеризацию данных с помощью алгоритма DBSCAN.***

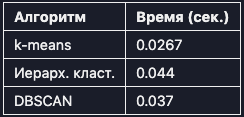
Алгоритм DBSCAN отличается от предыдущих тем, что сам находит число предполагаемых кластеров в данных. Эмпирическим путем были подобраны параметры eps и min\_samples, при которых распределение на кластеры наиболее схоже с предыдущими результатами (будем считать что они заведомо верны, ведь число классов в исходных данных мы знаем, соответственно количество кластеров, на которые делятся эти данные, тоже).

******

******

1. ***Сравнить скорость работы алгоритмов. Результаты изобразить в виде таблицы.***

Для расчета времени работы алгоритмов была применена стандартная библиотека time. Для сравнения бралось только время работы каждого алгоритма, не учитывая остальные операции, связанные с обработкой данных, полученных при их помощи. Результаты приведены в таблице далее:

******

Отчетливо видно, что существенно различается время работы у k-means и остальных алгоритмов. Иерархическая кластеризация и DBSCAN уже не имеют таких больших различий во времени обработки данных (в данном конкретном примере).