|  |  |
| --- | --- |
|  | |
| МИНистерство Науки и высшего образования российской федерации | |
| Федеральное государственное бюджетное образовательное учреждениевысшего образования"МИРЭА - Российский технологический университет"РТУ МИРЭА | |
| Институт информационных технологий (ИТ) | |
| Кафедра прикладной математики (ПМ) | |

|  |  |
| --- | --- |
| **ОТЧЕТ**  **ПО ПРАКТИЧЕСКИМ РАБОТАМ** | |
| **по дисциплине** | |
| «Технологии и инструментарий анализа больших данных» | |
|  | |
| Выполнили студенты группы ИВБО-07-19 | Гридасов Е.А. |
| Принял ассистент кафедры ПМ | Горячев А.А. |

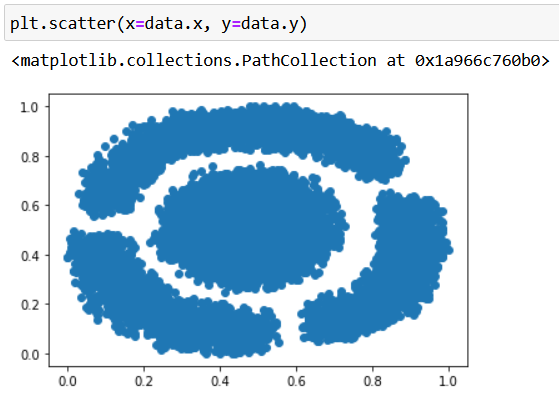
|  |  |  |
| --- | --- | --- |
| Практическая работа выполнена | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |
|  |  |  |
| «Зачтено» | «\_\_»\_\_\_\_\_\_\_\_\_2022 г. |  |

Москва 2022

# ВЫПОЛНЕНИЕ ЗАДАЧ

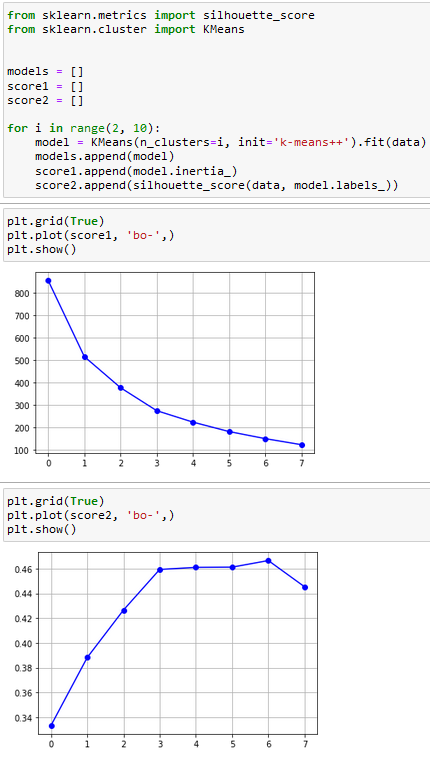
Целью работы было углубление в тему кластеризации с использованием различных моделей и метрик.

Для выполнения задания был скачан датасет в форме логотипа гугл хром. Была проведена нормализация. Датасет состоит из колонок по x, y, а также классу для проверки.



**Рисунок 1 – Датасет**

Первым делом была проведена класторизация с помощью k-means, использовано правило локтя, построен график по коэффициенту силуета.



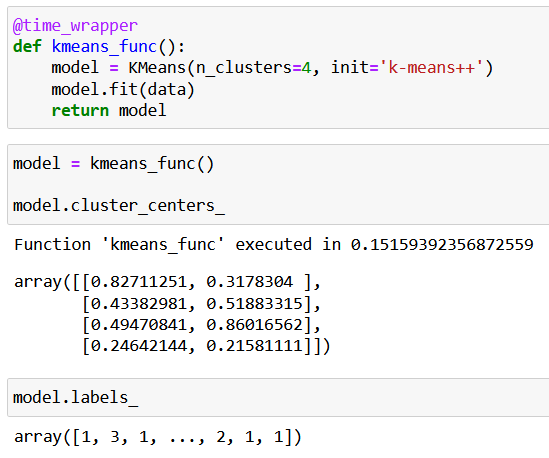
**Рисунок 2 — Правило локтя и силует**

Взяв погрешность метода к-средних, по графикам было решено выбрать 4 кластера.

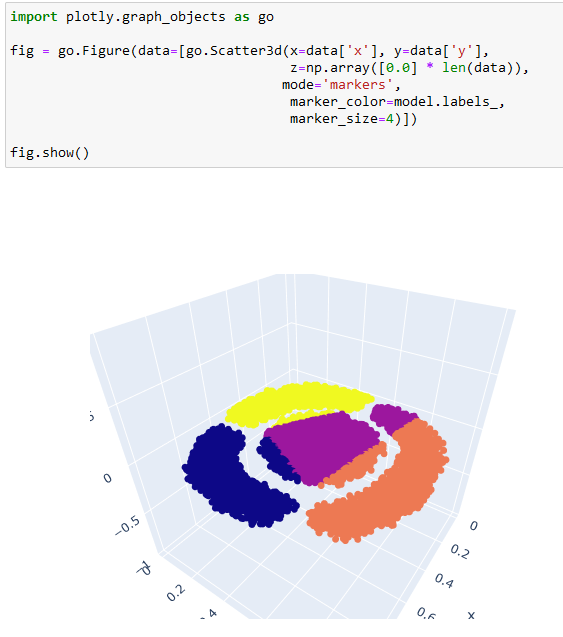
Дальше были построены все остальные модели и графики по ним с обрамлением в виде декоратора по времени.



**Рисунок 3 — Декоратор времени**

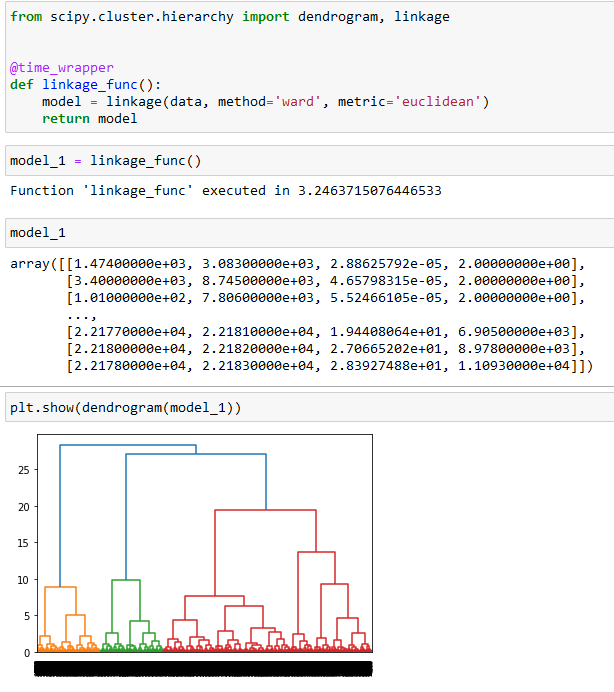


**Рисунок 4 — Модель к-средних**

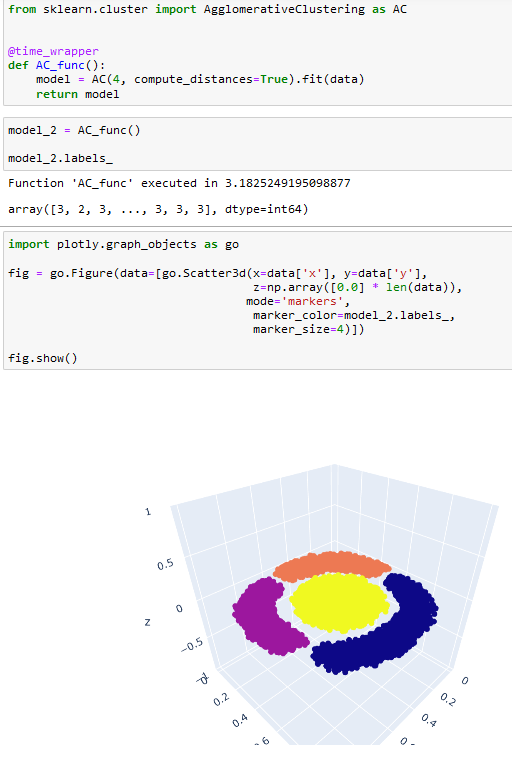


**Рисунок 5 — Результат визуализации**

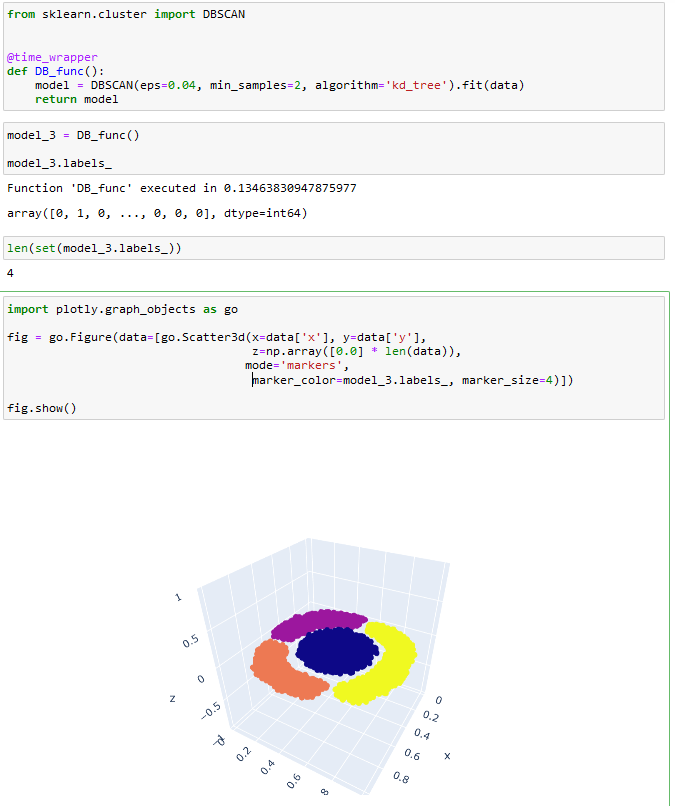
Для интереса, была построена дендрограмма с помощью метода linkage из библиотеки scipy. Получилось 4 кластера.



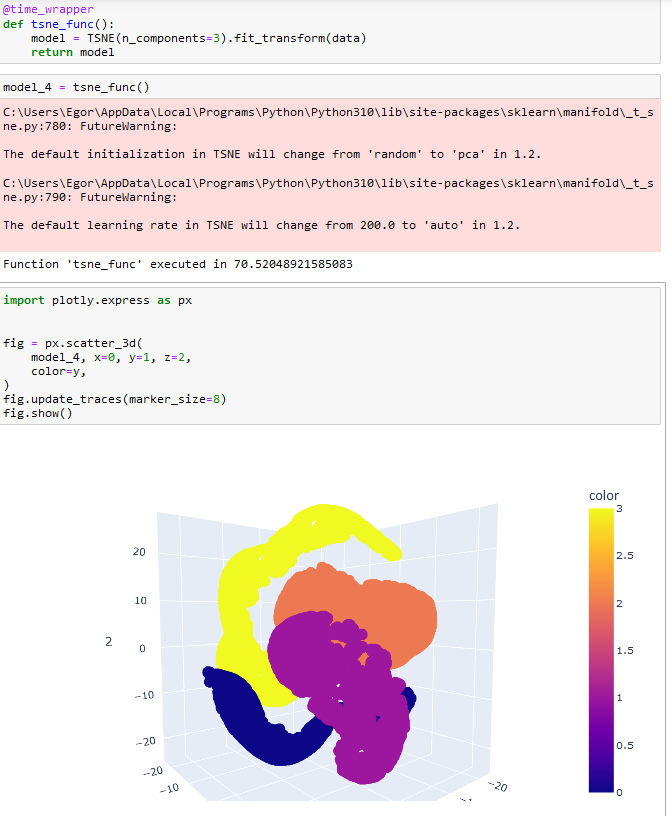
**Рисунок 6 — Результат использования методов denrogram и linkage из scipy**



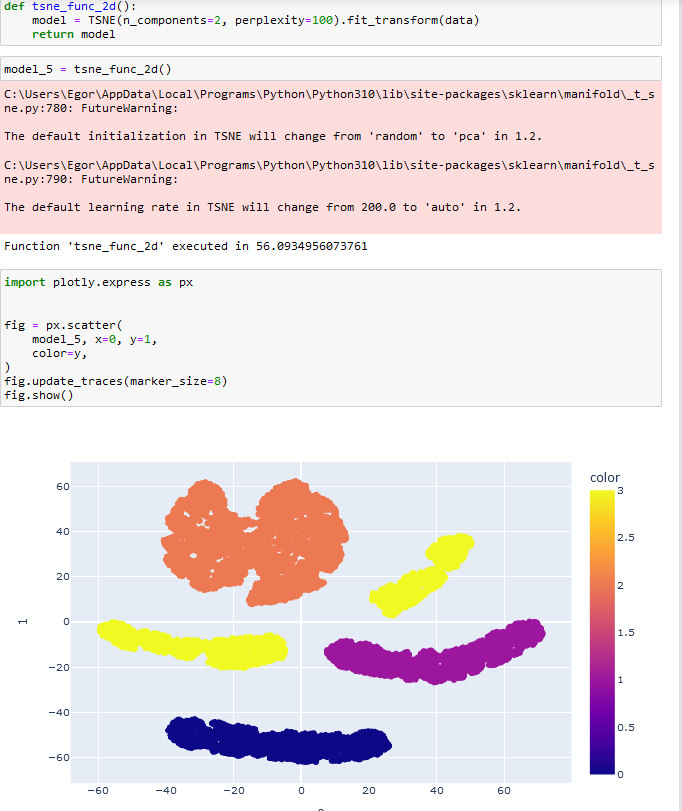
**Рисунок 7 — Модель AgglomerativeClustering**



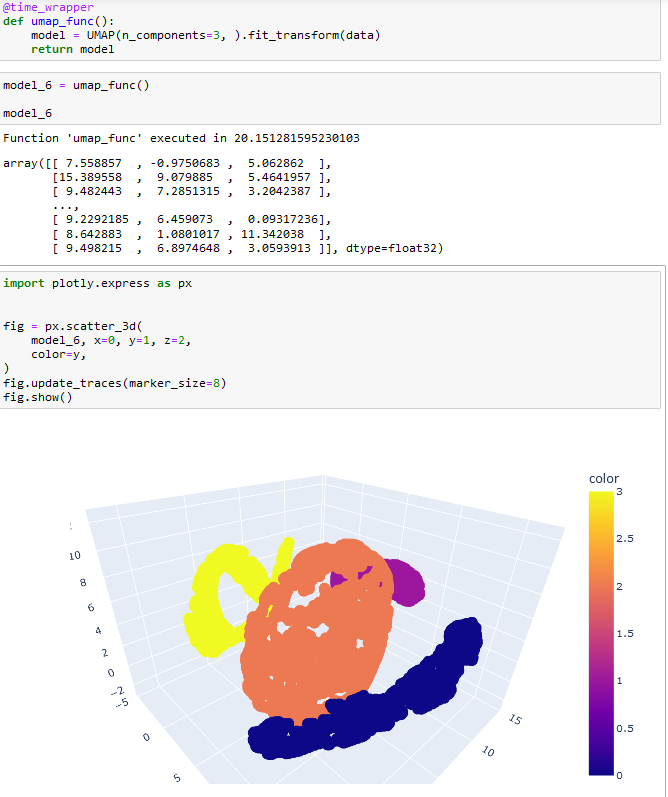
**Рисунок 8 — Модель DBSCAN**



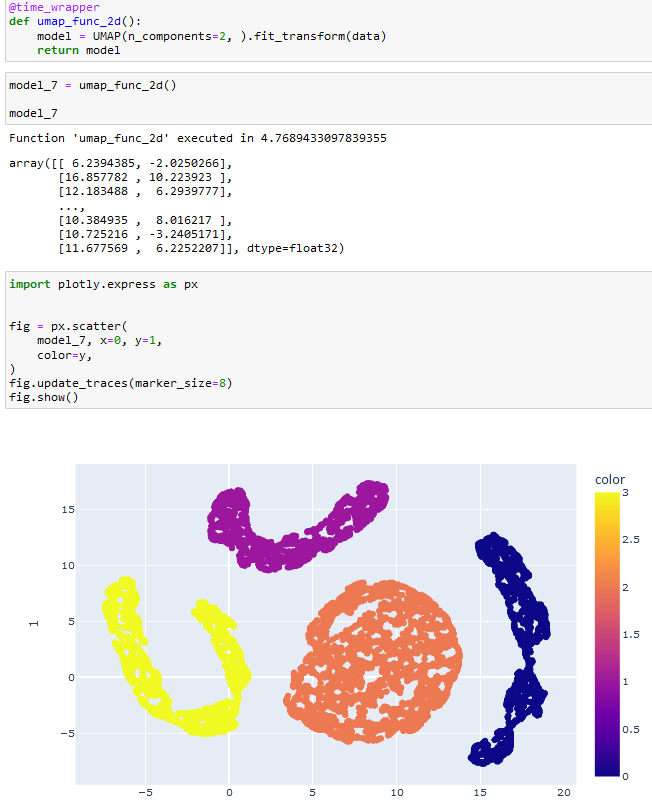
**Рисунок 9 — Модель TSNE 3D**



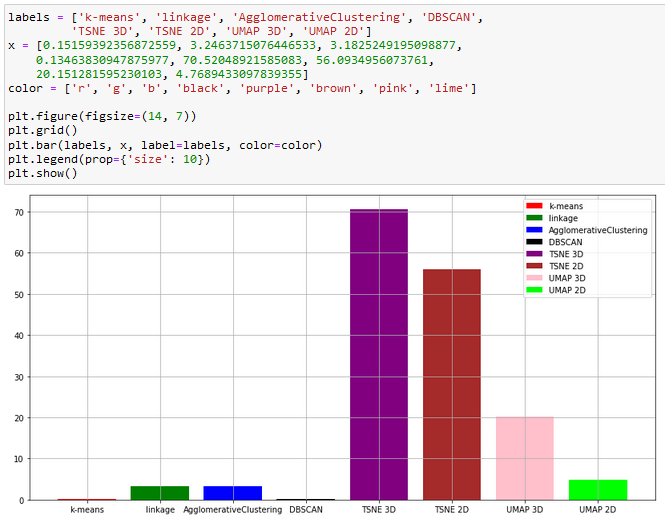
**Рисунок 10 — Модель TSNE 2D**



**Рисунок 11 — Модель UMAP 3D**



**Рисунок 12 — Модель UMAP 2D**



**Рисунок 13 — График сравнения времени обучения моделей**

Самый времязатратной оказалась TSNE 3D за счет количества гиперпараметров. Самой быстрой — k-means. При этом TSNE и UMAP преобразовали кластеры в свои плоскости, отличные от изначальных, но поймав нужное количество кластеров.

# ВЫВОДЫ

В ходе выполнения работы были изучены библиотеки matplotlib, plotly, sklearn, scipy. Освоена тема кластеризации данных.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Python Documentation [Электронный ресурс] - <https://www.python.org/>
2. Jupyter Documentation [Электронный ресурс] - <https://docs.jupyter.org/en/latest/index.html>
3. Pandas Documentation [Электронный ресурс] - <https://pandas.pydata.org/docs/getting_started/index.html>
4. Matplotlib Documentation [Электронный ресурс] - <https://matplotlib.org/>
5. Plotly Documentation [Электронный ресурс] - <https://plotly.com/python/>
6. Numpy Documentation [Электронный ресурс] - <https://numpy.org/doc/>
7. Stats Documentation [Электронный ресурс] - <https://docs.scipy.org/doc/scipy/reference/stats.html>