Лабораторная работа № 6

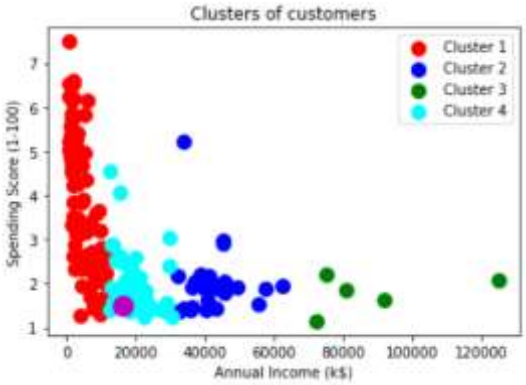
Тема: Алгоритмы кластеризации.

Для работы используйте данные **подходящие для кластеризации** на свой выбор.

Еще для кластеризации **подходят датасеты для многоклассовой классификации**, только нужно удалить столбец с метками классов, что бы имитировать обучение без учителя. Если ничего не найдете, то можно использовать *Country-data.csv* в папке с примерами на дисктейшен.

Примеры есть в папке примеров на дискстейшене и еще ! рекомендую смотреть примеры на сайте www.kaggle.com .

1. Из датасета выберите наиболее важные параметры, характеризующие цель исследования и сформируйте из них матрицу X.
2. Проверьте X на пропуски и закодируйте категориальные данные, если это необходимо.
3. Нормализуйте значения в матрице X функцией MinMaxScaler().
4. С помощью метода локтя определите оптимальное количество кластеров и разделите данные на кластеры методом **K-means**.
5. Визуализируйте результаты кластеризации, выбрав для визуализации два параметра из матрицы X.
6. Разделите данные на кластеры **методом иерархической кластеризации**, выберите с помощью дендрограммы оптимальное количество кластеров.
7. Визуализируйте результаты кластеризации методом иерархической кластеризации.
8. Оцените качество кластеризации методами K-means и иерархической кластеризации, рассчитав пару метрик качества кластеризации (модуль **sklearn.metrics**). Например, силуэт для выборки **silhouette_score()** и др.
9. Из датасета выберите любой конкретный объект (если вы делаете модель на датасете *Country-data.csv*, то выберите любую страну) и визуализируйте этот объект в виде точки отличного цвета и размера на графике кластеров (*пример на рисунке, точка пурпурного цвета*).



Вопросы:

1. Что решают задачи кластеризации в машинном обучении?
2. Расскажите принцип работы метода K-means.
3. Как можно выбрать оптимальное количество кластеров в K-means?
4. Расскажите принцип работы метода иерархической кластеризации.
5. Для чего можно использовать дендрограмму в методе иерархической кластеризации?
6. Какие метрики используют для оценки качества кластеризации?