Домашнее задание
по теме «Кластеризация k-means»

Формулировка задания

Выполните подготовку данных для решения задачи кластеризации. Проведите кластеризацию алгоритмов KMeans. Обратите внимание на подбор гиперпараметров n_clusters, init, max_iter, algorithm.

Для кластеризации использовать библиотеку scikit-learn.

Результирующий код должен быть читаемым, с единой системой отступов и адекватными названиями переменных.

Описание плана работы

1) Загрузите данные из дополнительных материалов или по ссылке: https://www.kaggle.com/datasets/youssefaboelwafa/clustering-penguins-species

2) EDA(Exploratory Data Analysis) можно использовать из задачи Снижения размерности. Обратите внимание на выбросы и требования к нормализации данных.

3) Примените метод кластеризации KMeans с подбором гиперпараметров в цикле.

4) Оцените качество кластеризации методом локтя. Сделайте вывод.

5) Оцените качество кластеризации методом силуэта. Сделайте вывод.

6) Визуализируйте кластеры и центры кластеров. Обратите внимание, что для отображения на графике может понадобиться метод снижения размерности.

7) Сохраните модель в файл joblib.

Перечень необходимых инструментов

- Python
- scikit-learn
- pandas
- venv
- Jupiter Notebook

- IDE VS Code
- GigaIDE

Форма предоставления результата

1. В поле ссылки загрузить ссылку на удаленный репозиторий с доступом для наставника.
2. В поле файла загрузить архив с папкой, в которой разместить отчет со скриншотами по заданию и решение задачи. Решение должно быть представлено в формате .ipynb или .py.

Шкала оценивания

- 1.0 – отлично
- 0.7–0.9 – хорошо
- 0.5–0.6 – удовлетворительно
- Менее 0.5 – задание не выполнено