Лабораторная работа 5.
ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ. ЗАДАЧА КЛАСТЕРИЗАЦИИ

1. Изучение примеров.

● Изучите примеры: Lab5_ML Ex1 Base Clustering.ipynb**,** Lab5_ML Ex2 Метод_кластеризации_k_means.ipynb, Lab5_ML_Ex3 K_means_Hierarch.ipynb

2. Загрузка и подготовка данных.

2.1. Сгенерировать 3 датасета с использованием функции[1] make_classification[2] и 2 датасета с использованием функции make_blobs[3] (см. Lab5_ML Ex1 Base Clustering.ipynb). Данные необходимо сгенерировать так, чтобы на них можно было получить хорошее качество кластеризации. Количество кластеров должно быть различным и не менее трёх.

2.2. В соответствии с индивидуальным вариантом загрузите предобработанный датасет в формате CSV для решения задачи классификации. Удалите метку класса.

3. Решение задачи кластеризации.

3.1. Реализовать следующие алгоритмы кластеризации на синтетических данных: k-means, иерархическая кластеризация, DBSCAN, EM-алгоритм, Affinity Propagation.

3.2. Реализовать следующие алгоритмы кластеризации на данных для задачи классификации: k-means, DBSCAN, EM-алгоритм, Affinity Propagation.

3.3. Для соответствующего алгоритма кластеризации (используемых в п. 3.1 и п. 3.2) подобрать оптимальные гиперпараметры. В случае алгоритма кластеризации k-means используйте «метод локтя» и график силуэтов для.

3.4. Провести визуализацию работы всех алгоритмов кластеризации. Если это возможно, то вывести номер кластера. Опишите качество кластеров по их внешнему виду.

3.5. Если это возможно, то вывести номер кластера, создав дополнительный столбец в датасете для задачи классификации. Найдите характеристики каждого из кластеров с помощью библиотеки Pandas (см. Lab5_ML Ex2 Метод_кластеризации_k_means.ipynb).

4. Оценка качества моделей. Каждый реализованный алгоритм кластеризации оцените 2 внешними и 2 внутренними метриками оценки качества.

---

[1] **make_classification** и **make_blobs** — это функции из библиотеки Sklearn.datasets, которые используются для генерации синтетических наборов данных для задач классификации и кластеризации соответственно.

[2] **Функция make_classification** генерирует случайный набор данных для классификации, который можно использовать для тестирования алгоритмов ML. Можно настроить количество классов, признаков и образцов.
https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html

[3] **Функция make_blobs** создает набор данных с определенным количеством кластеров, где каждый кластер имеет свои центры. Это подходит для задач кластеризации. https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_blobs.html

5. Реализация алгоритма кластеризации k-means.

5.1. Самостоятельно разработайте и реализуйте алгоритм кластеризации k-means с возможностью подсчета суммы квадратов расстояний между точками и соответствующими центроидами.

5.2. Поместите разработанный алгоритм кластеризации k-means в существующий файл библиотеки алгоритмов ML и подключите его к основной программе.

5.3. Проведите кластеризацию двух сформированных датасетов (см. п. 2.1, для k-means и любой другой) и датасета для классификации с использованием собственной реализации алгоритма k-means и k-means из библиотеки Scikit-learn.

5.4. Произведите визуализацию построенных моделей и покажите распределение кластеров.

5.5. Выполните оценку качества полученных моделей кластеризации. Сравните полученные результаты (Образец 1).

Образец 1

	Внутренние метрики		**Внешние метрики**	
Алгоритм ML	**Y1**	**Y2**	**Y3**	**Y4**
k-means из Sklearn	0.XX	0.XX	0.XX	0.XX
программный k-means				

6. Создание таблицы результатов.

● Создайте таблицу, выведите в ней наименования используемых алгоритмов кластеризации, наименования и значения вычисленных метрик оценки качества (Y1 и т.д.,) для синтетических данных и датасета для задачи классификации (Образец 2).

Образец 2

	Внутренние метрики		**Внешние метрики**	
Алгоритм ML	**Y1**	**Y2**	**Y3**	**Y4**
…	0.XX	0.XX	0.XX	0.XX
…				

7. Вывод. Напишите вывод о выполненной Лабораторной работе №5, в котором выберите лучшую модель кластеризации для синтетических данных и данных задачи классификации. Обоснуйте свое решение.