Лабораторные работы 15-17. Кластеризация

# k-Means

## Задача 1: Сегментация

Используйте встроенный в библиотеку seaborn датасет "penguins" для задачи кластеризации. Ваша задача - сегментировать пингвинов с использованием K-means.

Попробуйте иерархическую классификацию и DBSCAN на тех же данных и сравните результаты.

import seaborn as sns

# Load the data

penguins = sns.load\_dataset("penguins")

# Preprocess the data (for instance, filling missing data)

penguins.head()

## Задача 2: Кластеризация текстовых данных (новостных заголовков)

Вам нужно скачать [News Aggregator Dataset](https://www.kaggle.com/uciml/news-aggregator-dataset) с Kaggle, предварительно обработать текст новостных заголовков и провести кластеризацию с использованием K-Means. Целью этой задачи является группировка похожих новостей в одном кластере.

## Задача 3:

Примените алгоритм K-Means к датасету [Wine Recognition Data Set](https://archive.ics.uci.edu/ml/datasets/wine), чтобы выделить кластеры вин на основе их химического анализа.

## Задача 4:

Разделите датасет [Wholesale Customers Data](https://archive.ics.uci.edu/ml/datasets/wholesale+customers) на категории. В этом датасете содержатся данные о расходах клиентов оптового дистрибьютора на разные категории продуктов. Цель - сегментировать клиентов по их годовым расходам на разные категории продуктов.

# Иерархические алгоритмы

## Задача 1:

Примените иерархическую кластеризацию к набору данных [Iris](https://archive.ics.uci.edu/ml/datasets/Iris). Используйте дендрограмму для определения оптимального числа кластеров и сравните свои результаты с фактическими метками.

## Задача 2:

Текстовый анализ. Используйте датасет [20 News Groups](http://qwone.com/~jason/20Newsgroups/), который включает в себя около 20 000 сообщений, сгруппированных примерно на 20 новостных групп. Ваша задача заключается в иерархической кластеризации данных с использованием любой техники векторизации текста.

## Задача 3:

Получите данные о фильмах и их рейтинге (ссылка на данные: [Movies Dataset](https://www.kaggle.com/rounakbanik/movie-recommender-systems/data)). Ваша задача - с помощью иерархической кластеризации группировать фильмы по жанрам и средним рейтингам.

## Задача 4:

Используйте набор данных [Seeds](https://archive.ics.uci.edu/ml/datasets/seeds) для кластеризации различных сортов пшеницы. Данный набор данных содержит измерения геометрических свойств семян семи различных видов пшеницы. Ваша задача провести иерархическую кластеризацию и визуализировать результаты с помощью дендрограммы.

# DBSCAN

## Задача 1:

Используйте DBSCAN для кластеризации набора данных [Iris](https://archive.ics.uci.edu/ml/datasets/Iris). Сравните полученные результаты с изначальными метками. DBSCAN хорош для этой задачи, поскольку он может находить кластеры произвольной формы.

## Задача 2:

Примените DBSCAN для кластеризации датасета [Moons](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make\_moons.html) из библиотеки sklearn. Этот набор данных идеально подходит для демонстрации преимуществ DBSCAN, потому что он содержит два визуально различимых "месяца", и каждый из них должен стать отдельным кластером.

## Задача 3:

Используйте пакет DBSCAN в Python для кластеризации геолокационных данных (например, [Geolife GPS Trajectories](https://archive.ics.uci.edu/ml/datasets/Geolife+Trajectories+1.3)). Целью будет выявление основных мест сбора/перехода на основе GPS-координат.

## Задача 4:

Используйте DBSCAN для кластеризации датасета с уникальной структурой - [Mouse (3y)](https://cs.joensuu.fi/sipu/datasets/). Этот набор данных содержит координаты, которые образуют форму мыши, и DBSCAN может быть использован для обнаружения этой структуры.

# Сравнение алгоритмов