# Лабораторная работа 5

# Часть 1. «Идентификация продукта»



Умная камера фотосепаратора (аппарата для сортировки сыпучих продуктов) настроена на выявление признаков подсолнечника, ячменя, пшеницы и гороха.

В файле «Train.csv» представлены признаки 4-х видов сыпучих продуктов:

* **Area** – площадь объекта;
* **MajorAxisLenght** – длина главной оси инерции;
* **MinorAxisLenght** – длина побочной оси инерции;
* **Eccentricity** – экстрецесинтет объекта;
* **Orientation** – угол ориентации объекта;
* **ConvexArea** – площадь выпуклой оболочки, построенной над объектом;
* **FilledArea** – заполненность;
* **EulerNumber** – число Эйлера;
* **EquivDiameter** – эквивалентый диметр;
* **Solidity** – твердость;

# Extent;

* **Perimeter** – периметр объекта;

**Целевая переменная Class** – класс объекта, принимает 4 значения.

Файл «Train.csv» содержит признаки объектов и их маркировку.

Ваша задача состоит в том, чтобы по имеющимся признакам классифицировать продукты по видам: ячмень, подсолнечник, пшеница и горох.

1. Загрузить данные из файла Train.csv в DataFrame.
2. Вывести статистическую информацию о наборе данных.
3. Вывести названия столбцов и строк.
4. Заменить или удалить пропущенные значения.
5. Заменить категориальные данные количественными.
6. Выбрать важные признаки с помощью корреляционного анализа (оценить их влияние на значение целевого признака)
7. Определить, как наблюдения (объекты) распределены по классам?
8. Разделите данные на признаки и ответы.
9. Построить классификаторы, используя метод ближайших k-соседей и градиентного бустинга.
10. Подобрать наилучшее значение гиперпараметра, используя поиск по сетке.
11. Построить классификатор с найденным параметром.
12. Определить важность признаков для классификации.
13. Дать описание каждого класса.
14. Загрузить данные из файла Test.csv в DataFrame.
15. По построенному в п. 11 классификатору, определить класс каждого объекта.
16. Дать описание каждого класса, полученного по тестирующей выборке.
17. Провести стандартизацию данных и для них выполнить п. 9 – 16.
18. Сравнить результаты двух классификаций.

**Часть 2**

1. Загрузить данные из файла Train.csv в DataFrame.
2. Удалить столбец Class
3. Загрузить данные из файла Test.csv в DataFrame.
4. Объединить оба DataFrame в один.
5. Решить задачу кластеризации, используя методы К-средних и DBSAN. Оптимальное количество кластеров определить методом локтя.

from sklearn.cluster import DBSCAN, KMeans

kmeans = KMeans(n\_clusters = 4, init = 'k-means++', max\_iter = 2000, n\_init = 10, random\_state = 42)

model\_dbscan\_s = DBSCAN(eps=0.01, min\_samples=100)

1. Проведите оценку полученных моделей методом силуэта.

from sklearn.metrics import silhouette\_score

print(silhouette\_score(X=x, labels=y\_pred)) # силуэт

1. Сделайте выводы