# Лабораторная работа №1 - Изучение и предобработка данных

Наборы данных: lab1_var2.csv, iris.csv

1. **Изучение набора данных iris.csv с использованием Pandas и Seaborn:**
   - 1.1. Загрузить данные из файла как Pandas DataFrame
   - 1.2. Вызвав у датафрейма метод head, проверить корректность загруженных данных
   - 1.3. Вызвав у датафрейма метод describe, получить характеристики. Опишите полученный результат.
   - 1.4. Видоизмените полученный датафрейм таким образом, чтобы метка классов были следующими: 0 - Iris-setosa, 1 - Iris-versicolor, 2 - Iris-virginica. Сохраните полученный датафрейм в отдельный файл формата csv.
   - 1.5. Визуально оцените набор данных, построив изображение, содержащее графики ядерной оценки плотности каждого признака (кроме признака класса), диаграмму рассеяния и двумерную ядерную оценку плотности для каждых признаков. Наблюдения разных классов должны быть выделены отдельным цветом (рекомендуемая палитра ‘tab10’ или ‘Set1’). Пример построения: https://seaborn.pydata.org/examples/pair_grid_with_kde.html . Опишите полученный график, что на нем изображено, какие выводы о данных можно сделать.
   - 1.6. На одном изображении постройте гистограммы распределения для каждого признака (для построения нескольких диаграмм на одном изображении, необходимо создать **subplot** из **matplotlib**, и для каждой диаграммы задать параметр **ax**, указав нужную ячейку. subplot возвращает два параметра: саму фигуру с изображением и список ячеек. Например, изображение с 4 ячейками записанных в ряд: **fig, axs = plt.subplots(1,4)**. Указание ячейки в параметре диаграммы делается следующим образом: **ax=axs[0]**). Затем последовательно модифицируйте изображение:
     - 1.6.1. Постройте гистограммы для разного количества столбцов: 5,10,15,20,30. Выберите на ваш взгляд такое количество столбцов, который лучшие образом описывает форму распределения признаков.
     - 1.6.2. Сделайте на каждой гистограмме разделение по цвету согласно классу. Проведите это в двух режимах, когда

гистограммы накладываются/суммируются и когда пересекаются. Далее используйте режим с пересечением.

1.6.3. Постройте гистограммы, чтобы вместо столбцов изображались ступеньки.

1.6.4. Добавьте на гистограммы график ядерной оценки плотности.

**2. Изучение набора данных iris.csv с использованием NumPy:**

2.1. Загрузите данные из файла как массив NumPy

2.2. Выведите первые 10 наблюдений набора данных.

2.3. Рассчитайте характеристики полученные методом describe в п. 1.3 с использованием методов NumPy

**3. Изучение набора данных вашего варианта:**

3.1. Оцените и опишите набор данных вашего варианта с использованием методов в п. 1

**4. Преобразование данных:**

4.1. Получите из датафрейма из п. 1.4 столбец с названием классов. Используя LabelEncoder и OneHotEncoder получите различные способы кодирования меток класса. В чем различия полученных кодировок?

4.2. Для датафрейма из п. 1.4, получите все столбцы признаков (столбцы не содержащие метки классов). Преобразуйте полученные столбцы в массив NumPy.

4.3. Для массива NumPy из п. 4.2 примените StandardScaler, MinMaxScaler, MaxAbsScaler и RobustScaler. Для каждого из результатов постройте гистограммы по каждому признаку без разделения по классам. В чем различия между такими преобразованиями данных?

4.4. Согласно варианту, самостоятельно реализуйте StandardScaler или MinMaxScaler с использованием NumPy. Проверьте корректность работы на вашем наборе данных, сравните результаты между вашей реализацией и реализацией из Sklearn, а также рассчитав минимальное, максимальное, среднее значение и дисперсию, после преобразования.