МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

**Лабораторная работа №\_\_3\_**

по дисциплине«Методы машинного обучения в автоматизированных системах обработки информации и управления»

Тема: « Обработка признаков»

ИСПОЛНИТЕЛЬ: \_\_Се Цзявэнь\_\_\_\_

ФИО

группа ИУ5И-22М\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_\_"\_\_\_\_\_\_\_\_\_202\_ г.

ПРЕПОДАВАТЕЛЬ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ФИО

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_15\_"\_\_\_\_\_05\_\_\_\_2024\_ г.

Москва - 2024

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Задания

* 1. Выбрать один или несколько наборов данных (датасетов) для решения следующих задач. Каждая задача может быть решена на отдельном датасете, или несколько задач могут быть решены на одном датасете. Просьба не использовать датасет, на котором данная задача решалась в лекции.
  2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
  3. масштабирование признаков (не менее чем тремя способами);
  4. обработку выбросов для числовых признаков (по одному способу для удаления выбросов и для замены выбросов);
  5. обработку по крайней мере одного нестандартного признака (который не является числовым или категориальным);
  6. отбор признаков:
  7. один метод из группы методов фильтрации (filter methods);
  8. один метод из группы методов обертывания (wrapper methods);
  9. один метод из группы методов вложений (embedded methods).

# Программы

* Текстовое описание набора данных:

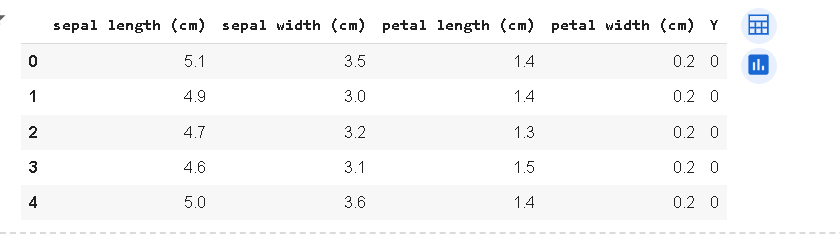
Этот набор данных содержит данные измерений трех различных сортов ириса (Сетоза, Версиколор и Вирджиния), по 50 образцов, собранных для каждого сорта, всего 150 образцов. Каждый образец имеет четыре характеристики:

Длина чашелистика: длина в сантиметрах от основания чашелистика до его кончика.

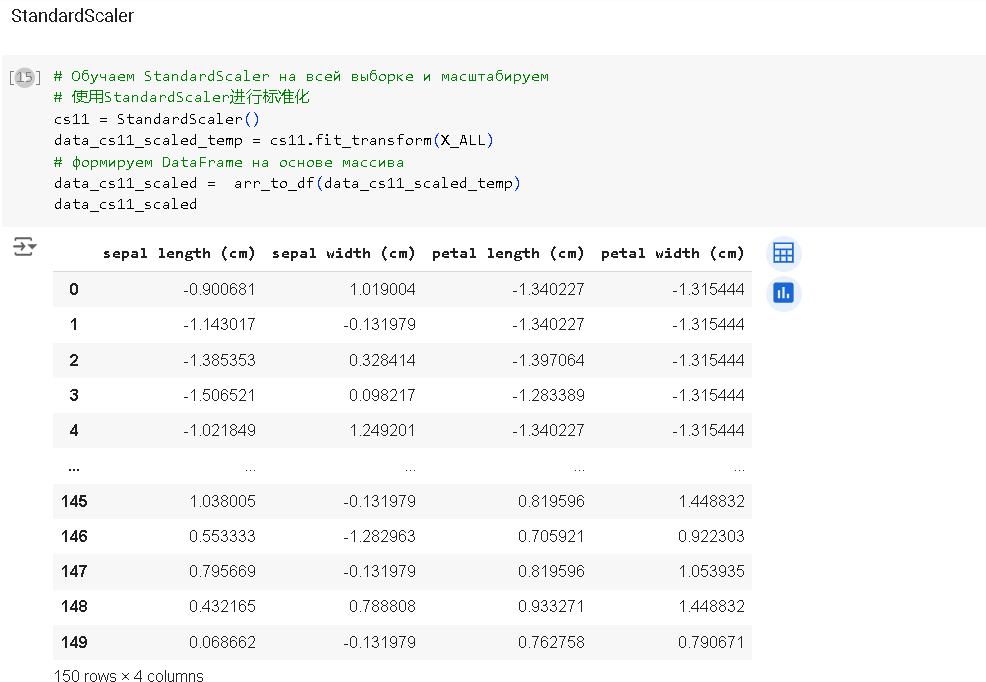
Ширина чашелистика: ширина в сантиметрах в самом широком месте.

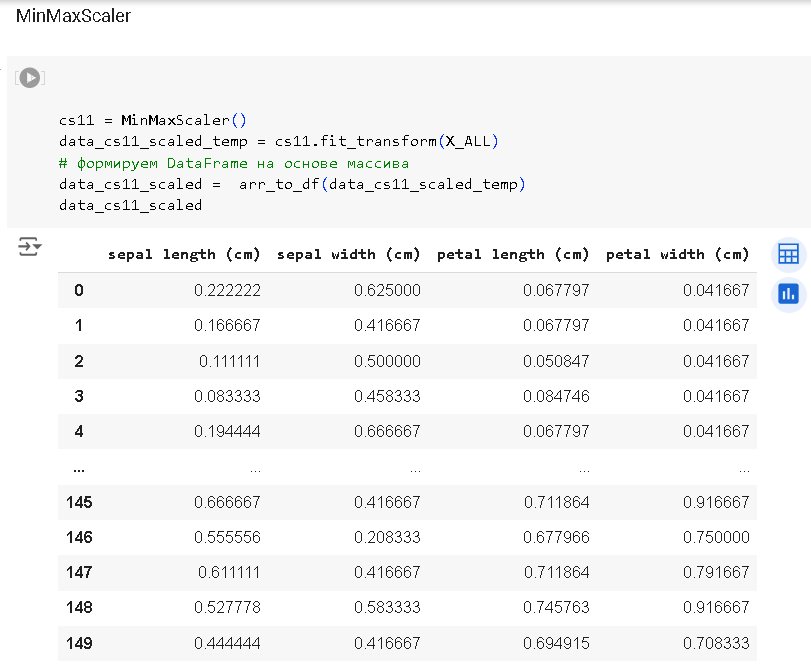
Длина лепестка: длина в сантиметрах от основания лепестка до его кончика.

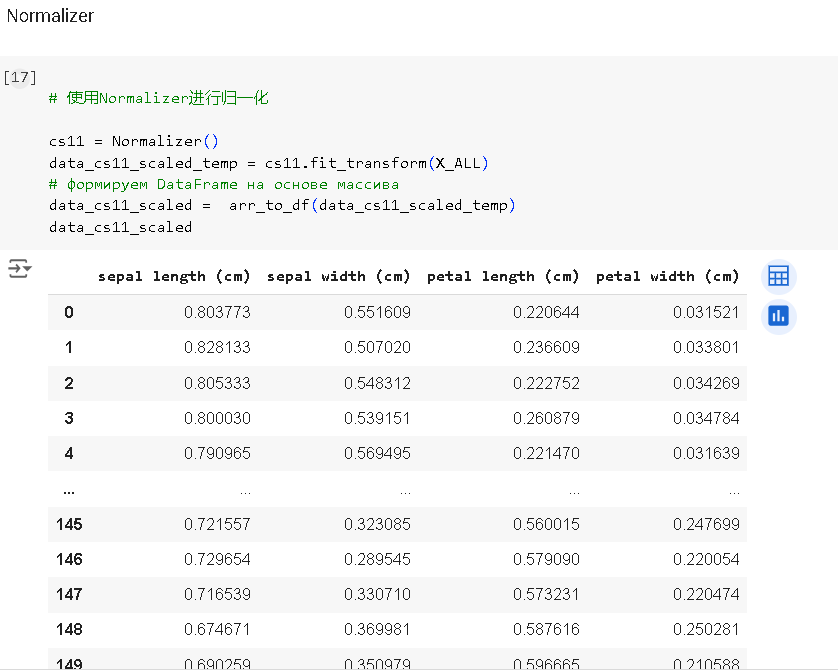
Ширина лепестка: Ширина в сантиметрах в самом широком месте.



# Масштабирование функций

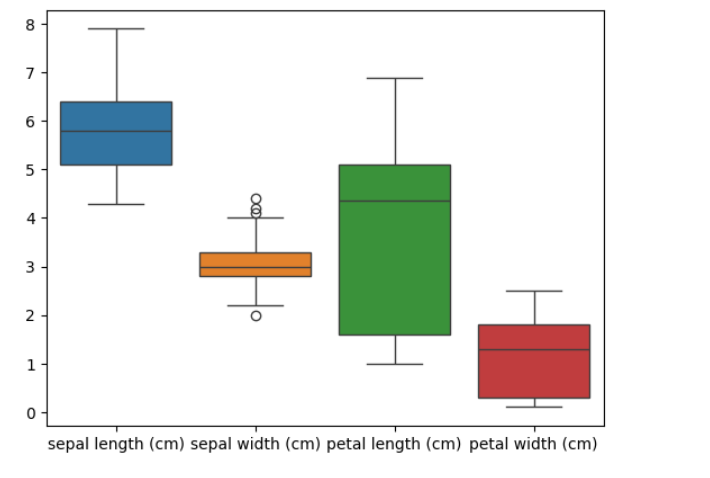




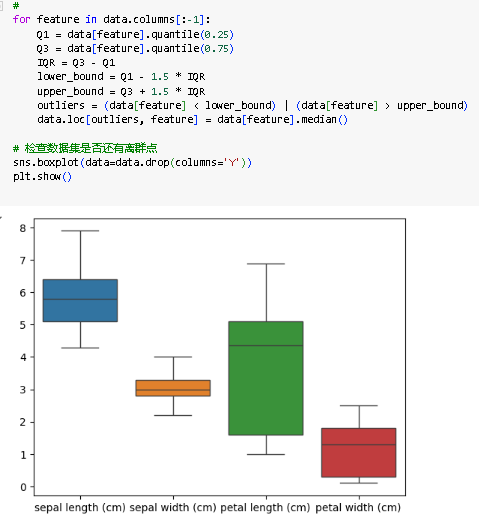


## Обработка выбросов

Используйте коробчатые диаграммы для отображения диапазона и распределения данных для каждого объекта.

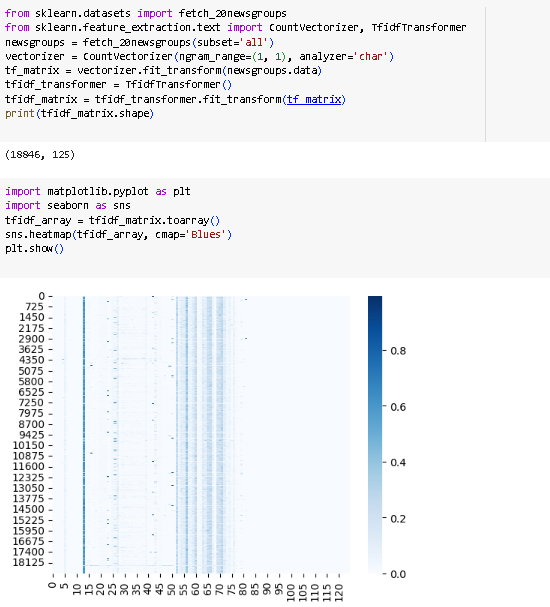


Найдите все выбросы и замените их медианой

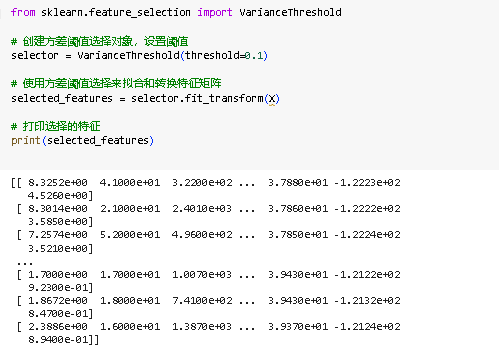


# обработку по крайней мере одного нестандартного признака

Один из способов построения матрицы TF-IDF по буквам с использованием набора данных 20 групп новостей — это использование N-грамм букв, где N представляет длину последовательности последовательных букв, которые необходимо учитывать. Буквенные N-граммы — это распространенный метод преобразования текста в числовое представление, которое представляет каждый документ в виде вектора.



# Отбор признаков (Feature selection)



# Вывод

Для набора данных радужной оболочки можно выполнить ряд задач предварительной обработки данных и выбора признаков. Среди них масштабирование признаков является одним из распространенных методов предварительной обработки данных. Числовые признаки можно масштабировать с использованием таких методов, как стандартизация, мин-максное масштабирование и регуляризация, чтобы обеспечить согласованность масштабов между различными признаками. Обработка выбросов также является важным этапом предварительной обработки. Метод Z-оценки можно использовать для обнаружения и замены выбросов в числовых характеристиках, чтобы избежать негативного влияния этих выбросов на обучение модели. Для нестандартных функций, таких как категориальные функции, такие как «Виды», можно выполнить горячее кодирование для преобразования их в числовые функции, чтобы модель могла обрабатывать эти функции. Кроме того, при выполнении выбора функций можно использовать несколько методов для выбора лучшего подмножества функций, включая методы фильтрации, методы упаковки и методы внедрения. Благодаря этим методам предварительной обработки данных и выбора признаков можно максимально полно извлечь информацию из набора данных и повысить производительность модели.