Московский государственный технический университет им. Н.Э. Баумана

Кафедра «Системы обработки информации и управления»

****

Лабораторная работа №3

по дисциплине

«Методы машинного обучения»

на тему

## **«Обработка признаков (часть 2)»**

Выполнил:

студент Люй Чжэ

группы ИУ5И-23М

Москва — 2024 г.

**1. Цель лабораторной работы**

изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

**2. Задание**

* 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.
* Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:

1. масштабирование признаков (не менее чем тремя способами);
2. обработку выбросов для числовых признаков (по одному способу для удаления выбросов и для замены выбросов);
3. обработку по крайней мере одного нестандартного признака (который не является числовым или категориальным);
4. отбор признаков:

4.1 один метод из группы методов фильтрации (filter methods);

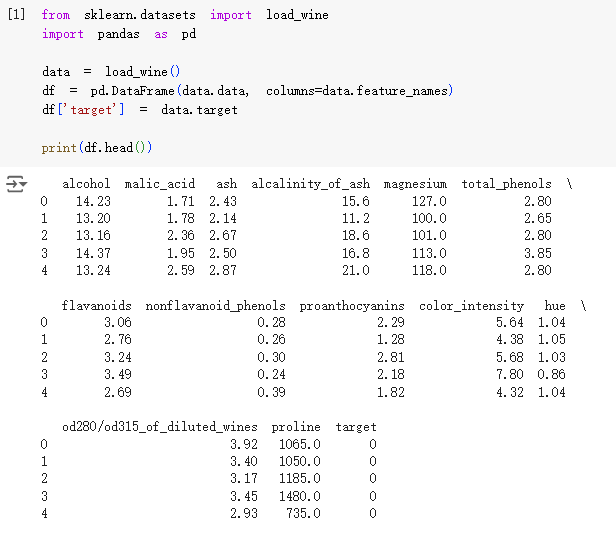
4.2 один метод из группы методов обертывания (wrapper methods);

4.3 один метод из группы методов вложений (embedded methods).

**3. Ход выполнения работы**

Я использовал набор данных «Вино» из библиотеки sklearn. Этот датасет содержит как числовые, так и категориальные признаки, что позволяет продемонстрировать различные методы предобработки данных.

**Шаг 1: Загрузка датасета**



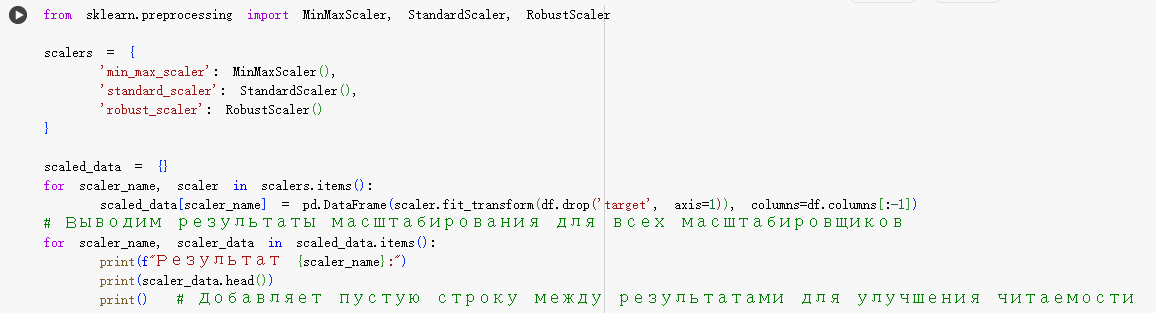
**Шаг 2: Масштабирование признаков**

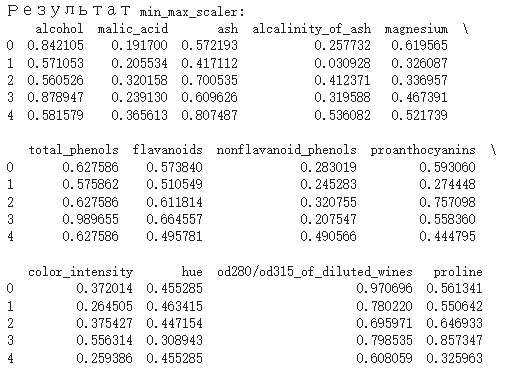
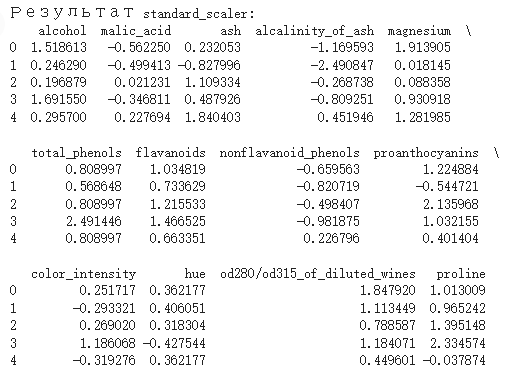
#### Методы масштабирования:

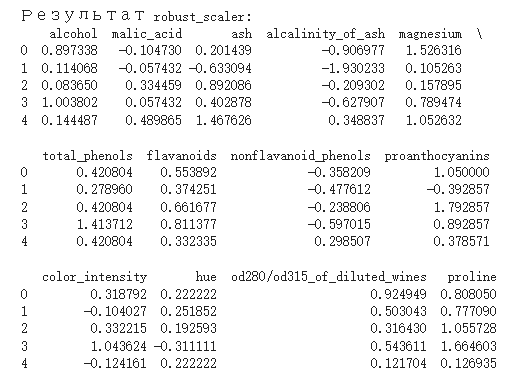
#### MinMaxScaler - приведение данных к масштабу от 0 до 1.

#### StandardScaler - масштабирование данных с нулевым средним и стандартным отклонением равным 1.

#### RobustScaler - масштабирование данных с использованием медианы и межквартильного размаха (менее чувствительно к выбросам).

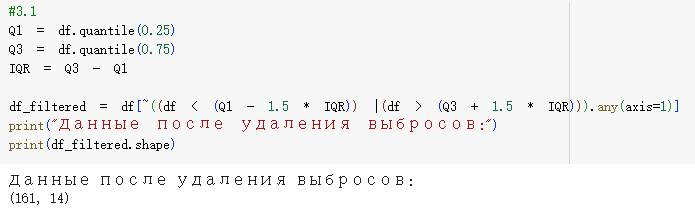




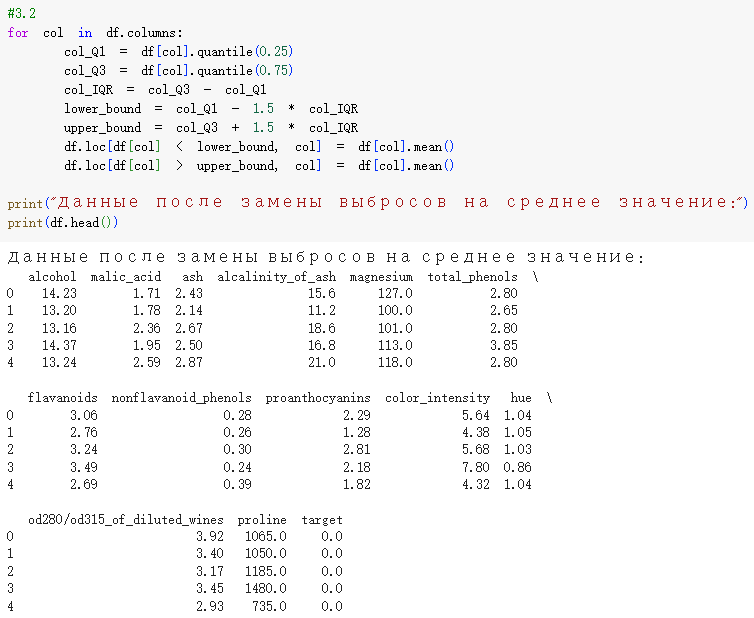


**Шаг 3: Обработка выбросов для числовых признаков**

#### Удаление выбросов с помощью IQR

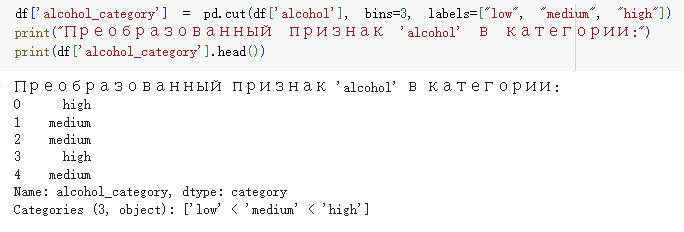


#### Замена выбросов средним значением



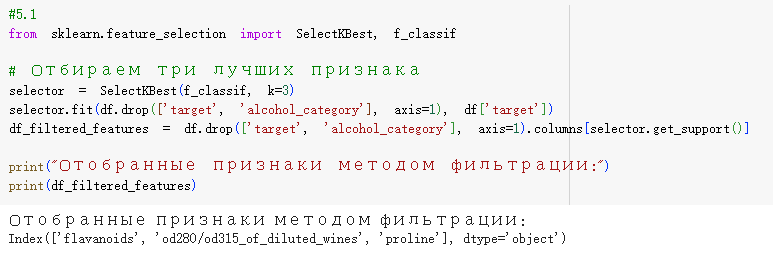
**Шаг 4: Обработка нестандартных признаков**

Для демонстрации выберем один из признаков и преобразуем его. Например, можно преобразовать числовой признак в текстовый.

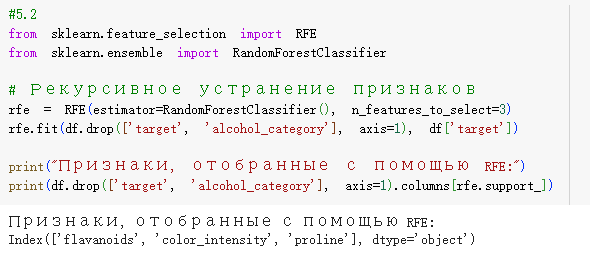


**Шаг 5: Отбор признаков**

#### Фильтрационные методы (Filter methods)



#### Методы обертывания (Wrapper methods)



#### Методы вложений (Embedded methods)

