Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика, искусственный интеллект и системы управления»

Кафедра «Системы обработки информации и управления»



**Рубежный контроль № 3**

**по дисциплине «Методы машинного обучения»**

Обработка признаков (часть 2)

ИСПОЛНИТЕЛЬ:

студентка ИУ5-23М

Морозевич М.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

\_\_\_ "\_\_\_\_\_\_\_\_" 2024 г.

Москва, 2024

# **Задание**

Для наборов данных необходимо решить задачи ниже:

* масштабирование признаков (не менее чем тремя способами);
* обработка выбросов для числовых признаков (по одному способу для удаления выбросов и для замены выбросов);
* обработка по крайней мере одного нестандартного признака (который не является числовым или категориальным);
* отбор признаков:
  + один метод из группы методов фильтрации;
  + один метод из группы методов обертывания;
  + один метод из группы методов вложений.

# **Описание данных**

HTRU2 - это набор данных, описывающий выборку потенциальных пульсаров, собранных в ходе исследования Вселенной с высоким временным разрешением (South).

Пульсары описываются двумя путями:

1. Интегрированный профиль импульса: каждый пульсар генерирует уникальную структуру излучения, известную как его профиль импульса.
2. Кривая DM-SNR: Радиоволны, излучаемые пульсарами, с более высокой частотой замедляются меньше по сравнению с волнами с более высокой частотой, что называется дисперсией.

Названия и описания колонок набора данных:

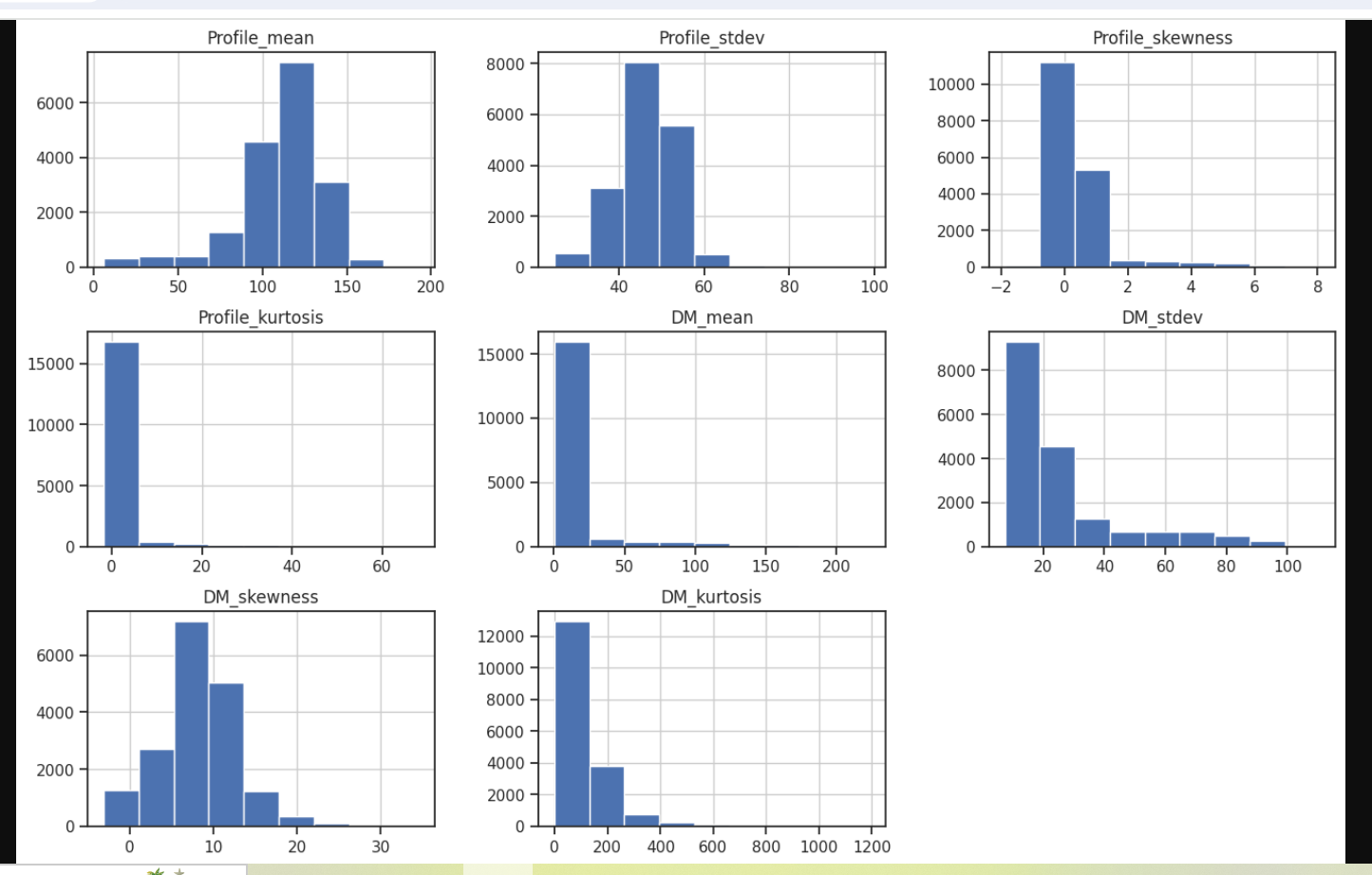
* Profile\_mean: Среднее значение интегрального профиля;
* Profile\_stdev: Стандартное отклонение интегрального профиля;
* Profile\_skewness: Чрезмерный эксцесс интегрального профиля;
* Profile\_kurtosis: Асимметрия интегрального профиля;
* DM\_mean: Среднее значение кривой DM-SNR;
* DM\_stdev: Стандартное отклонение кривой DM-SNR;
* DM\_skewness: Чрезмерный эксцесс кривой DM-SNR;
* DM\_kurtosis: Асимметрия кривой DM-SNR;
* class: класс объекта (0 или 1).

Представленный набор данных содержит 16 259 ложных примеров, вызванных радиочастотными помехами, и 1 639 реальных примеров пульсаров. Все эти примеры были проверены комментаторами-людьми.

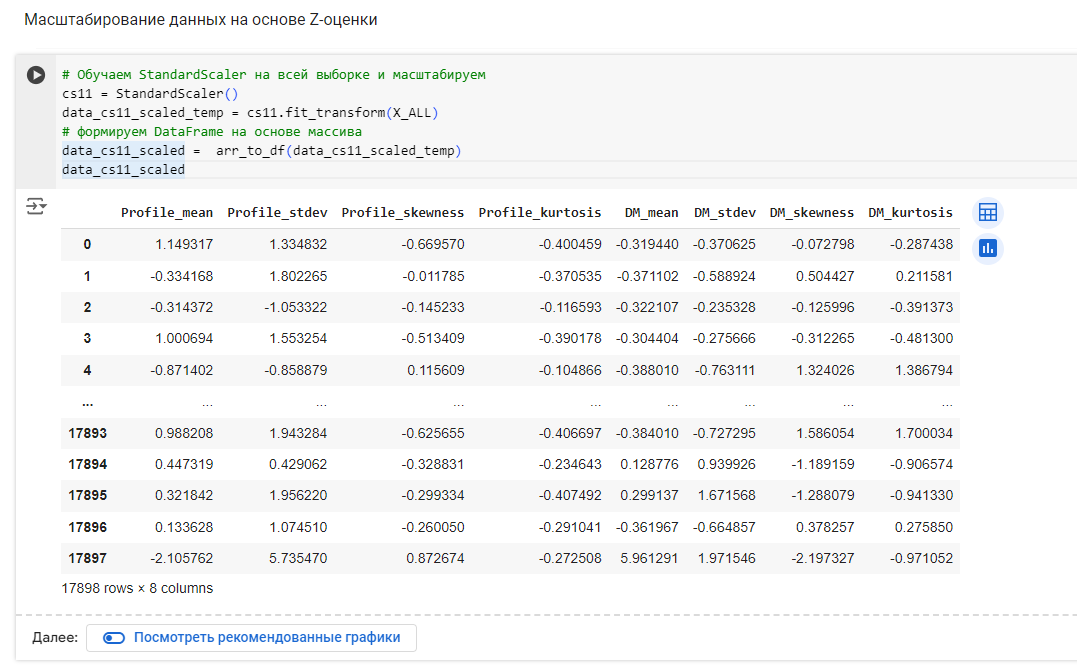
# **Масштабирование признаков**

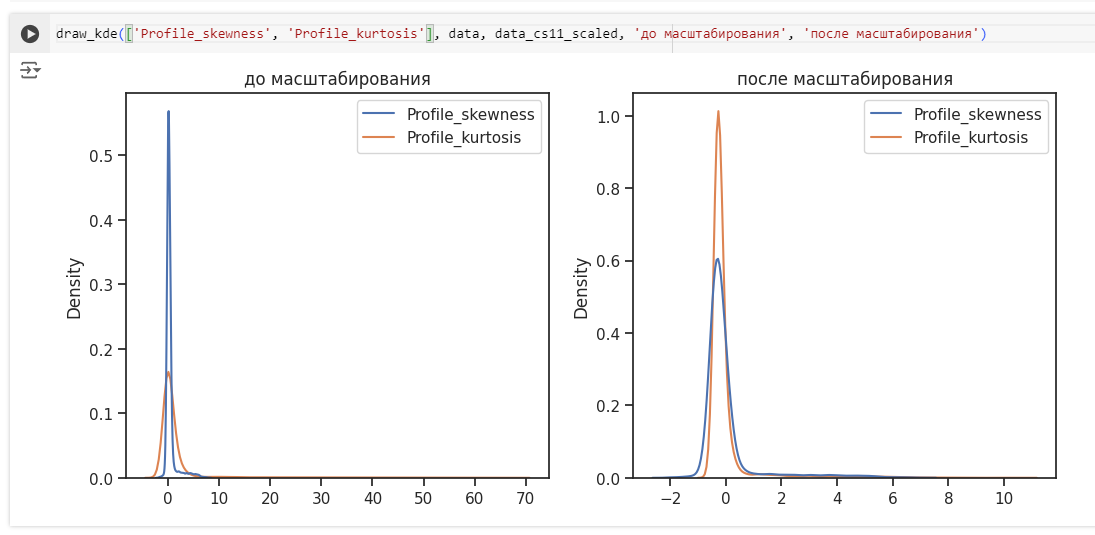
Получим данные и построим гистограммы распределения числовых признаков.



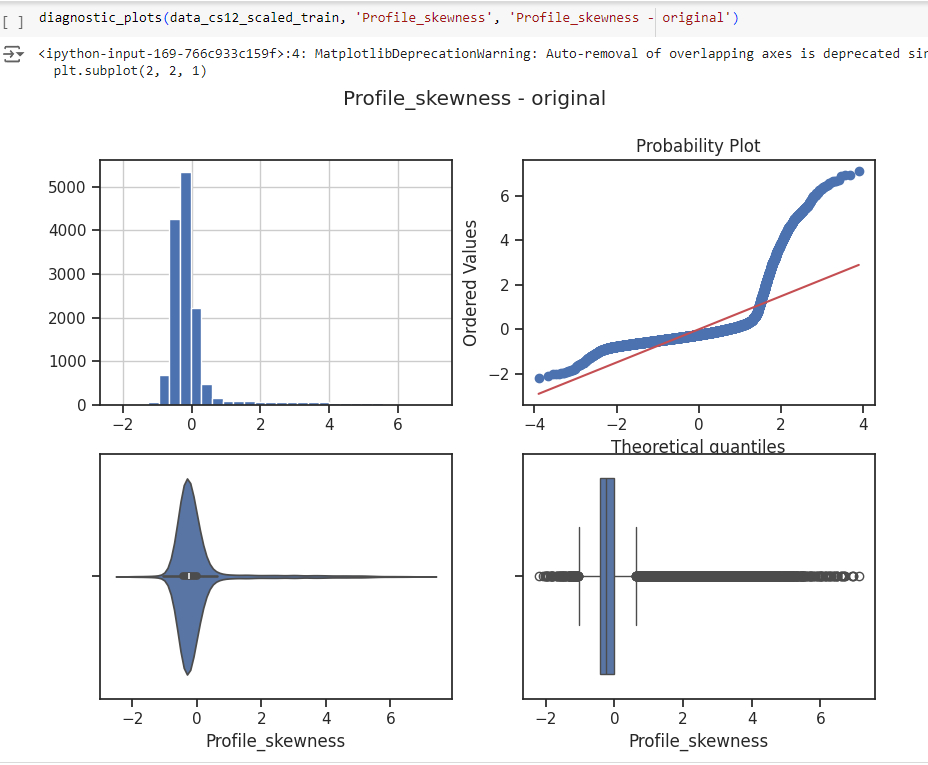
****

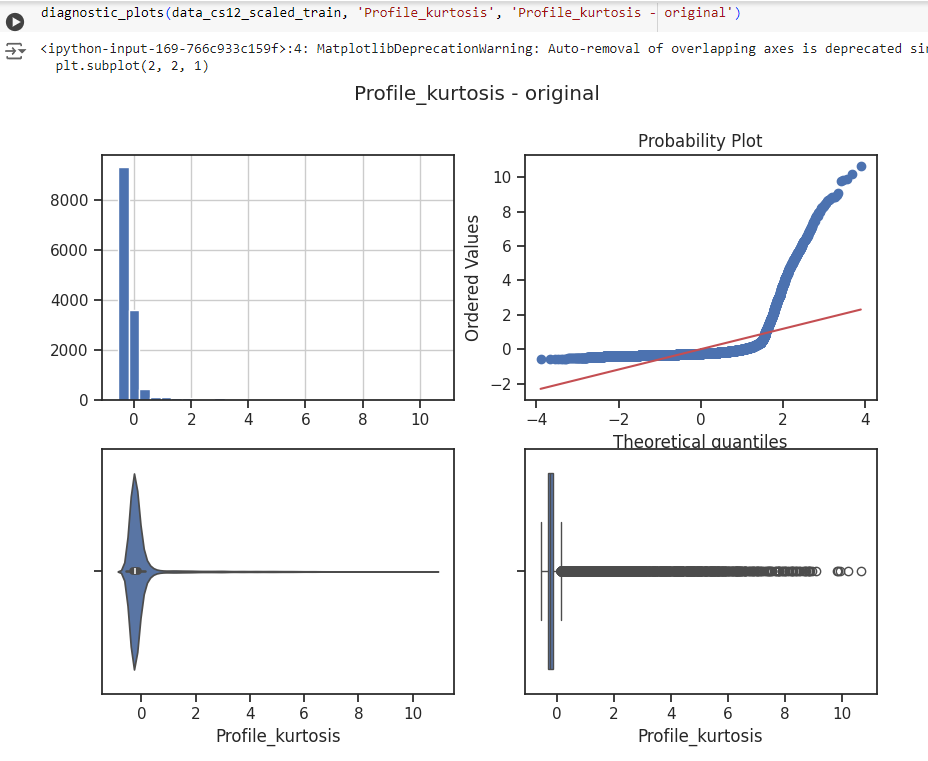
Применим масштабирование на основе Z-оценки. Построим плотности распределения для колонок «Profile\_skewness» и «Profile\_kurtosis».



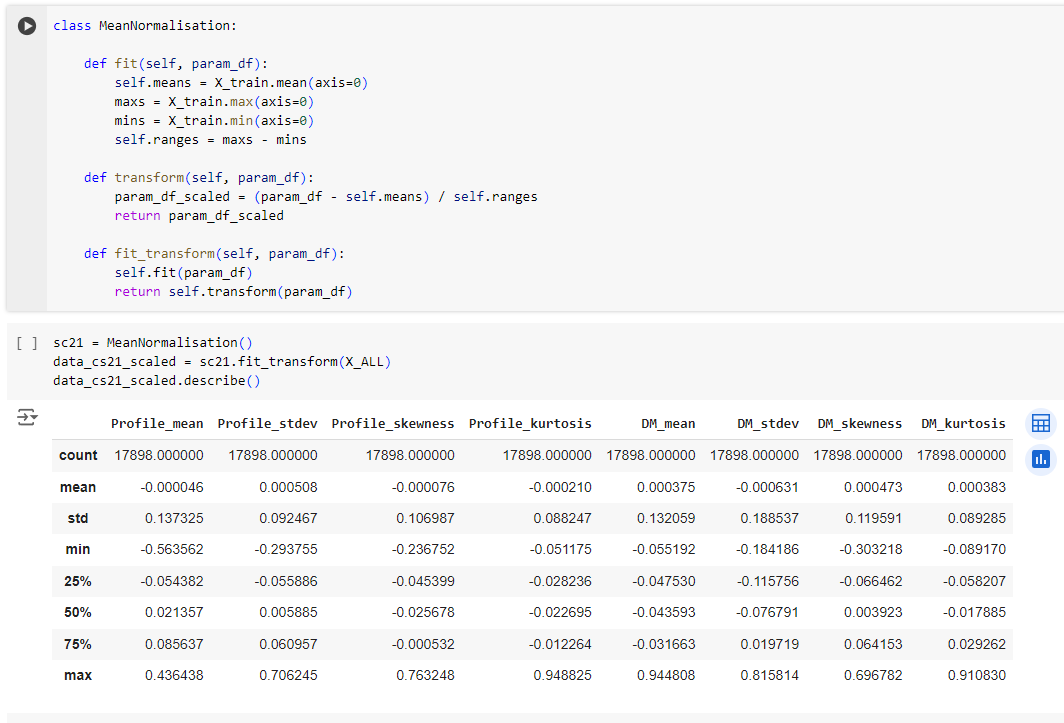


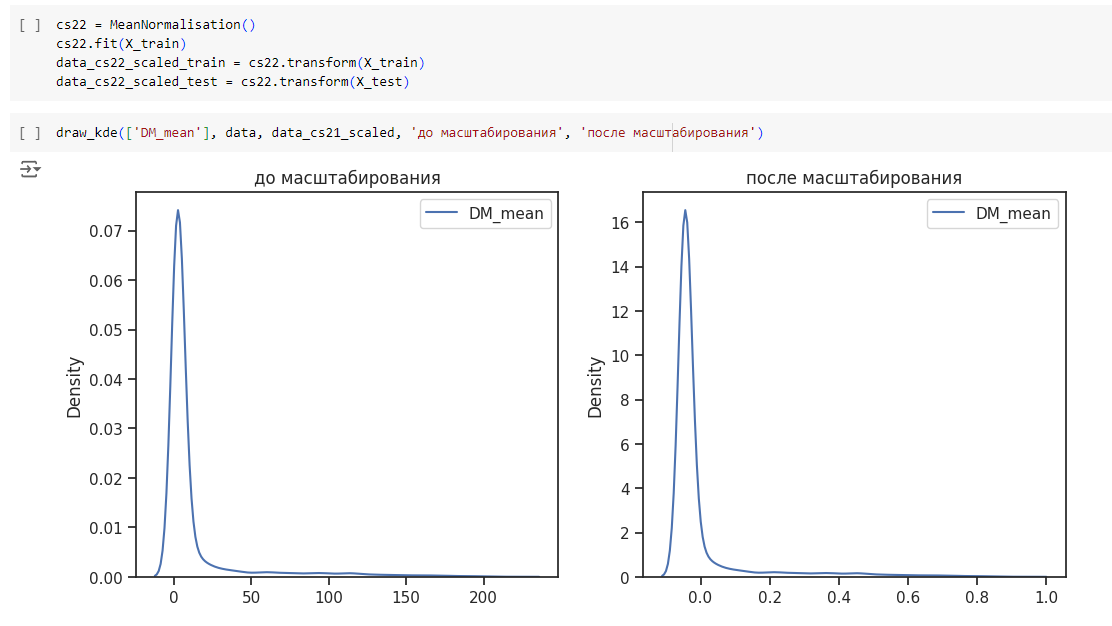
Построим диагностические графики для них.

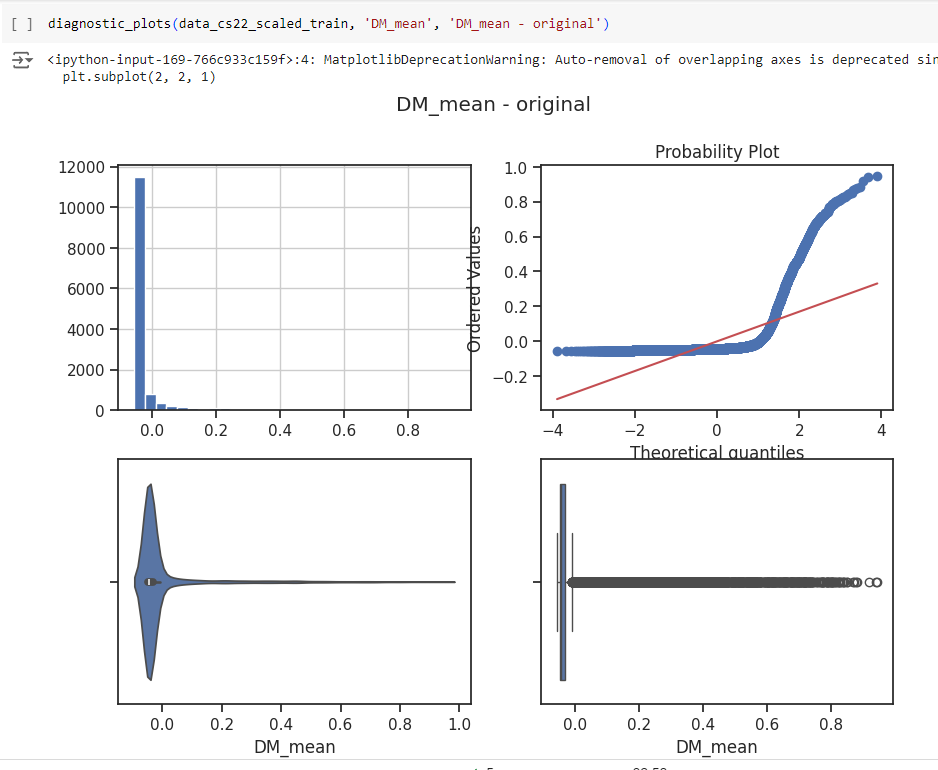




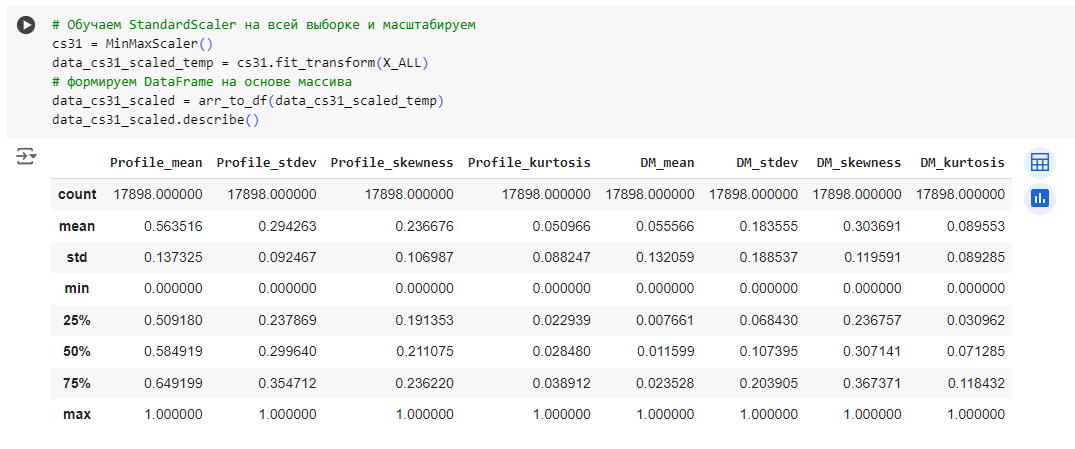
Применим масштабирование Mean Normalisation. Построим плотности распределения и диагностические графики для колонки «DM\_mean».

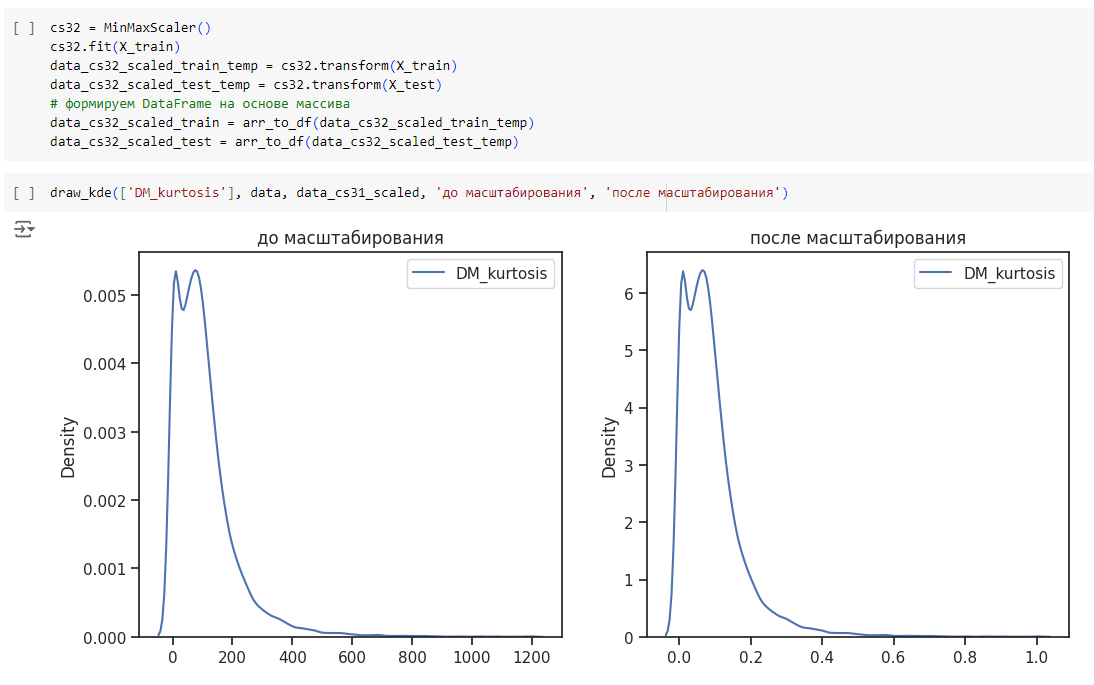


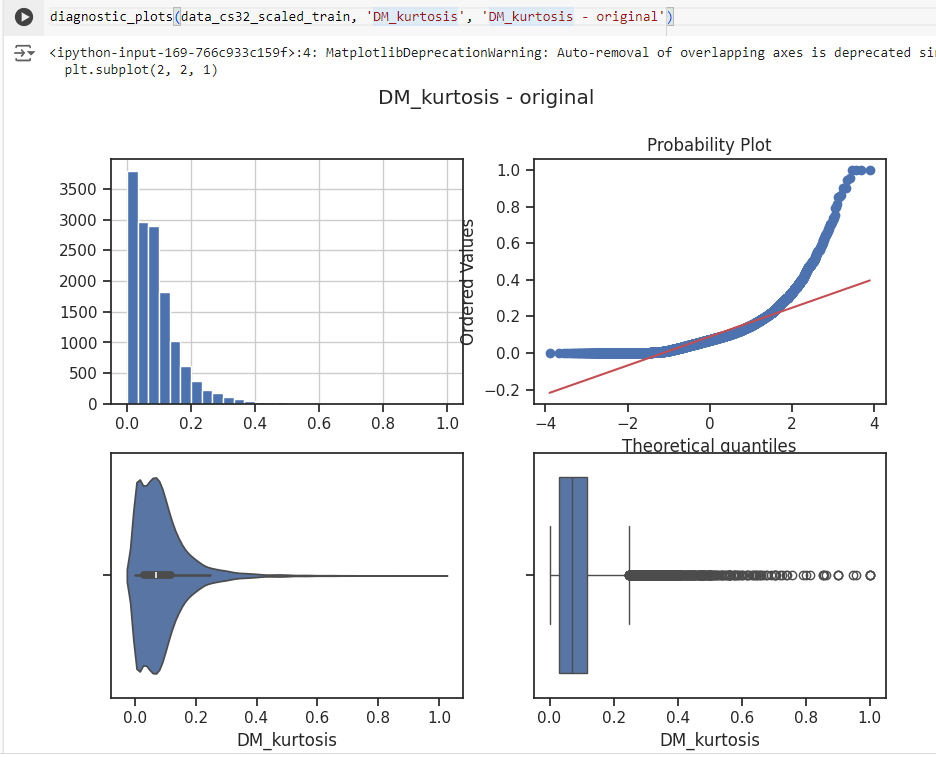




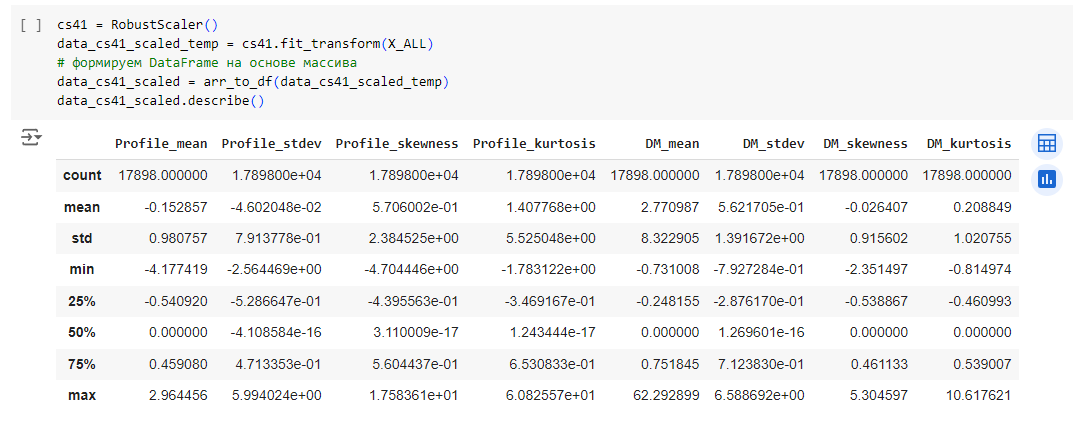
Применим MinMax-масштабирование. Построим плотности распределения и диагностические графики для колонок «DM\_kurtosis».

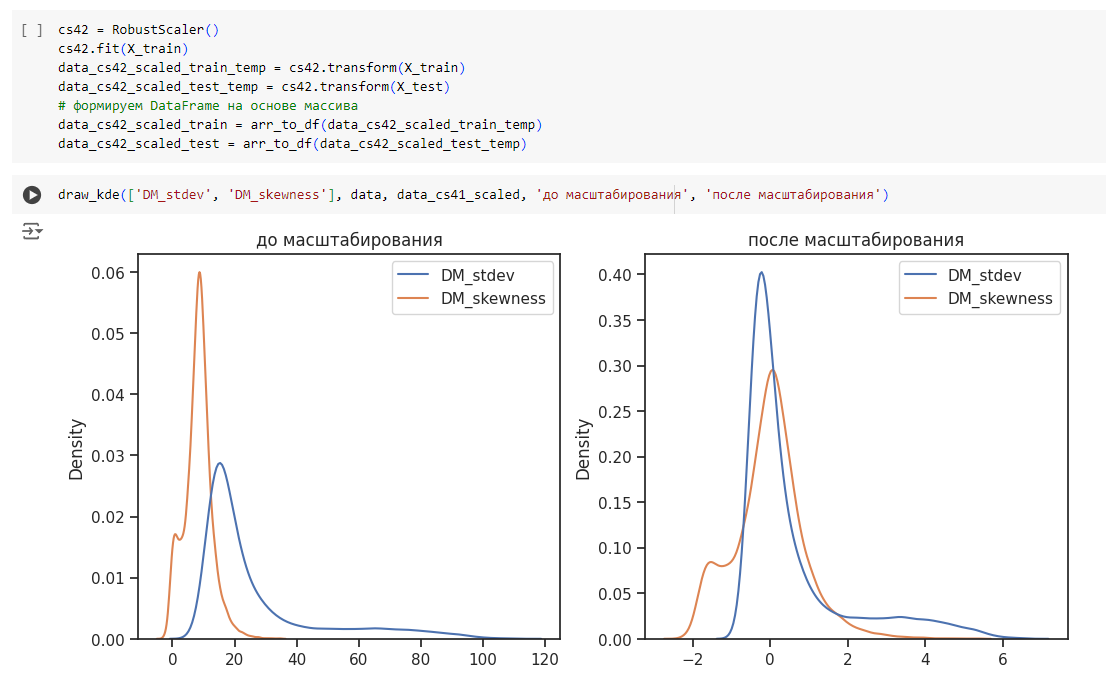


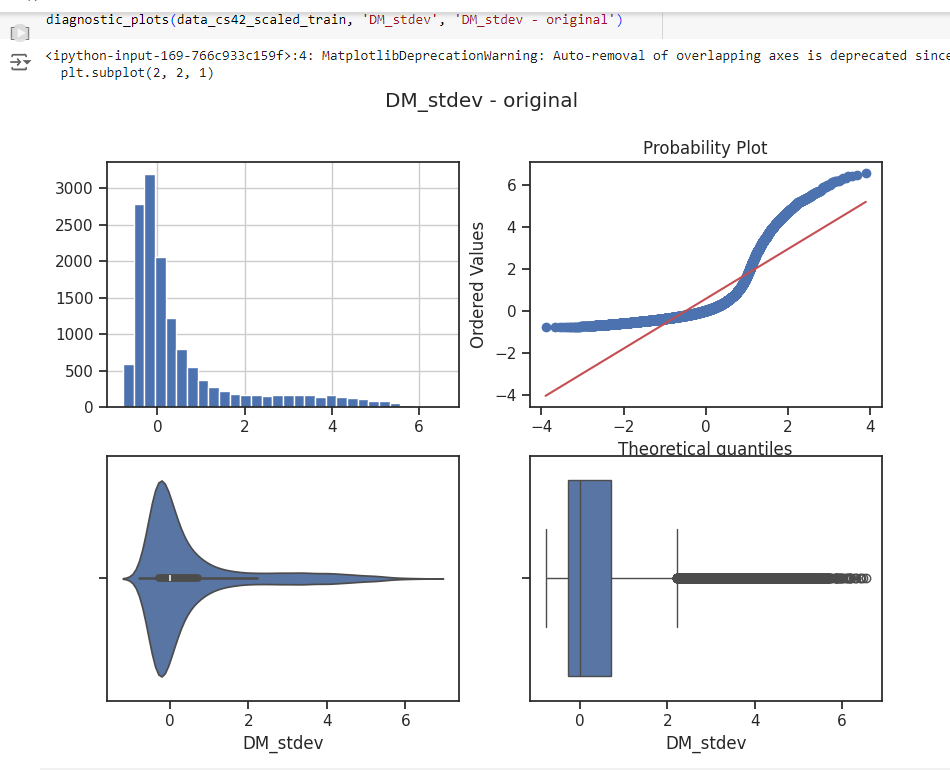


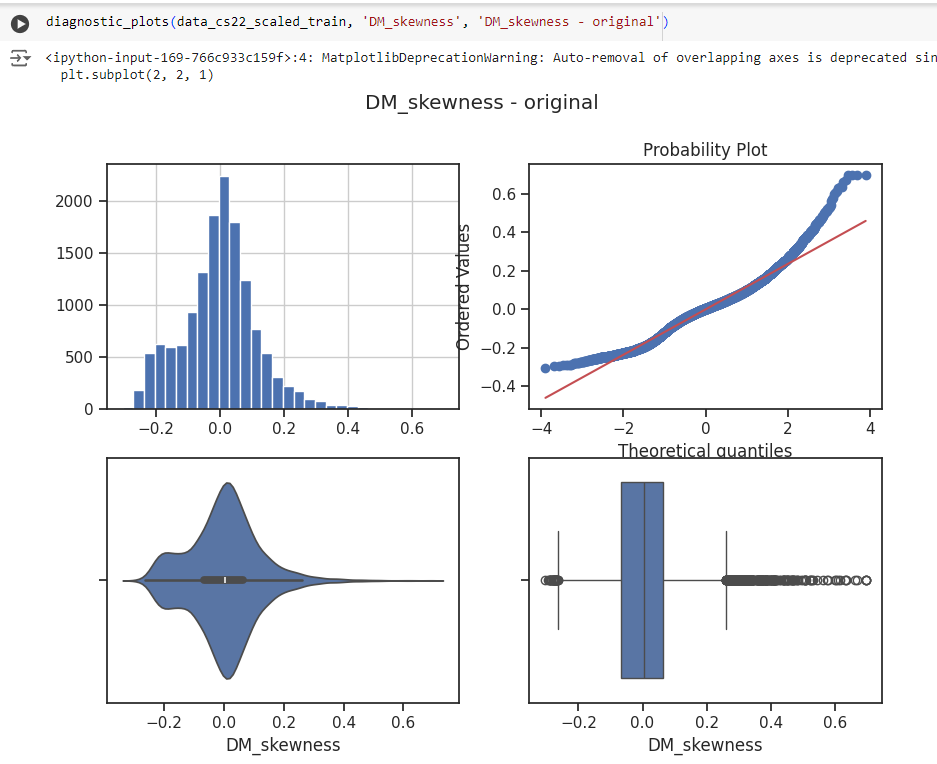


Применим масштабирование по медиане. Построим плотности распределения и диагностические графики для колонок «DM\_stdev» и «DM\_skewness».



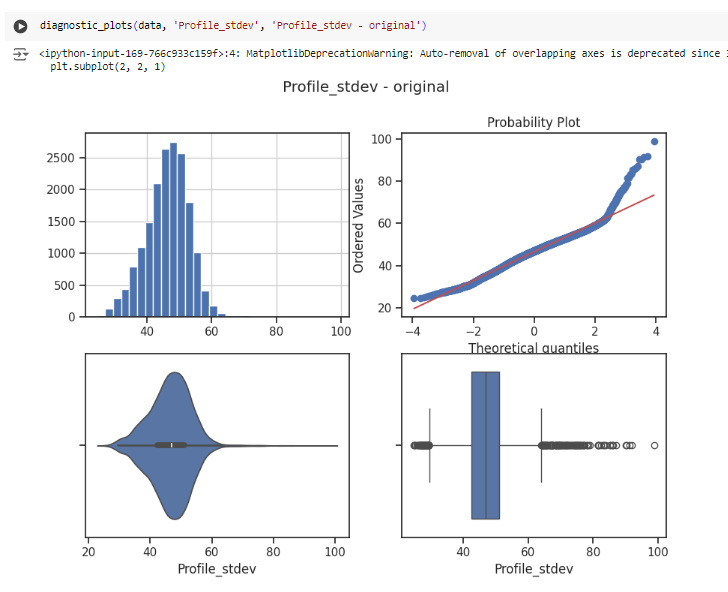


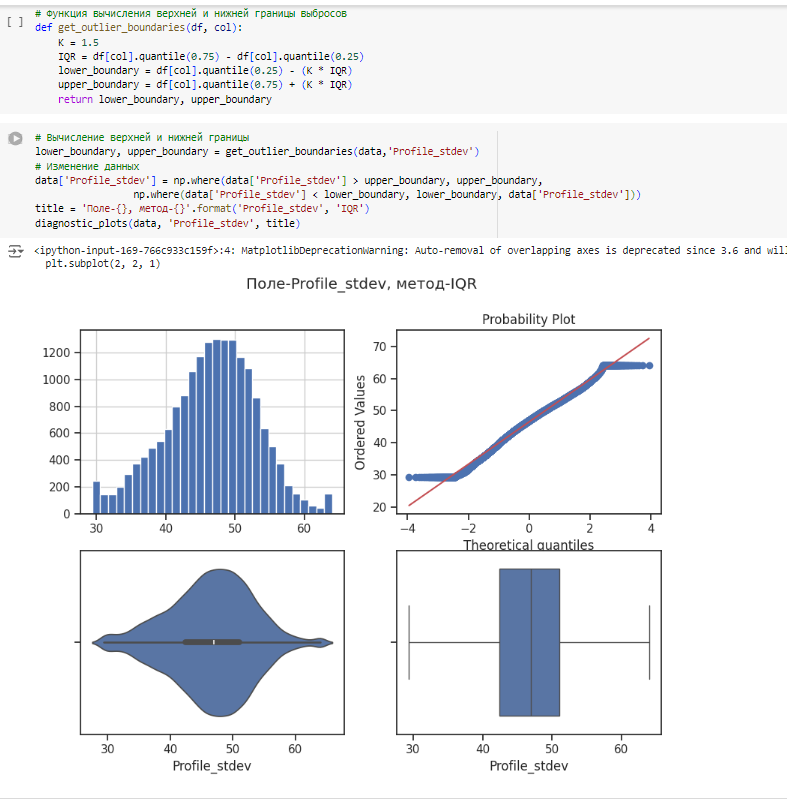


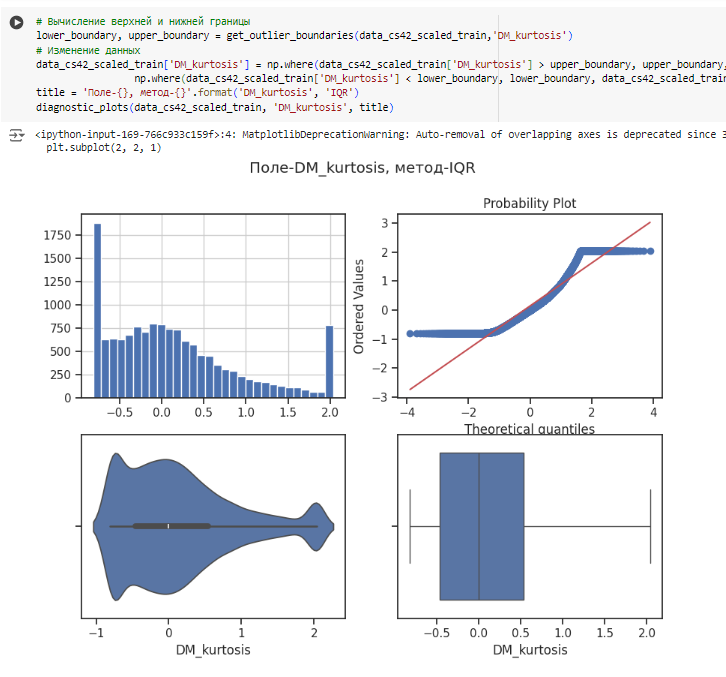


# **Обработка выбросов для числовых признаков**

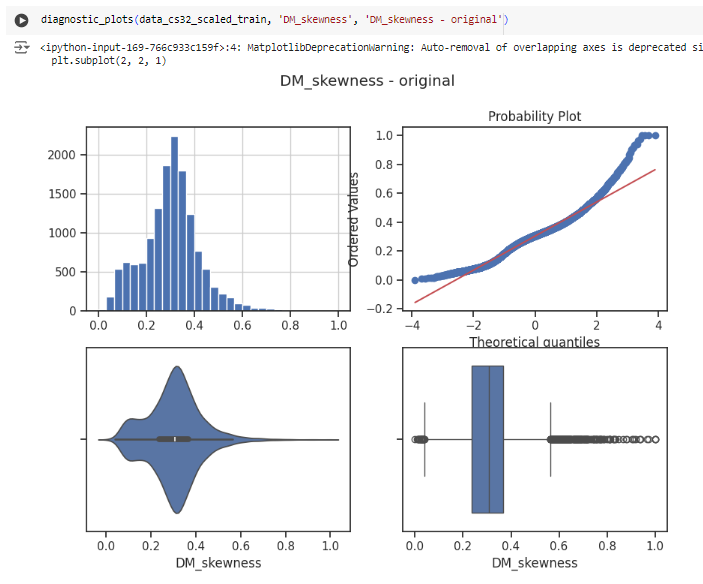
Заменим выбросы с использованием межквартильного размаха. Построим диагностические графики для колонки «Profile\_stdev» и «DM\_kurtosis».

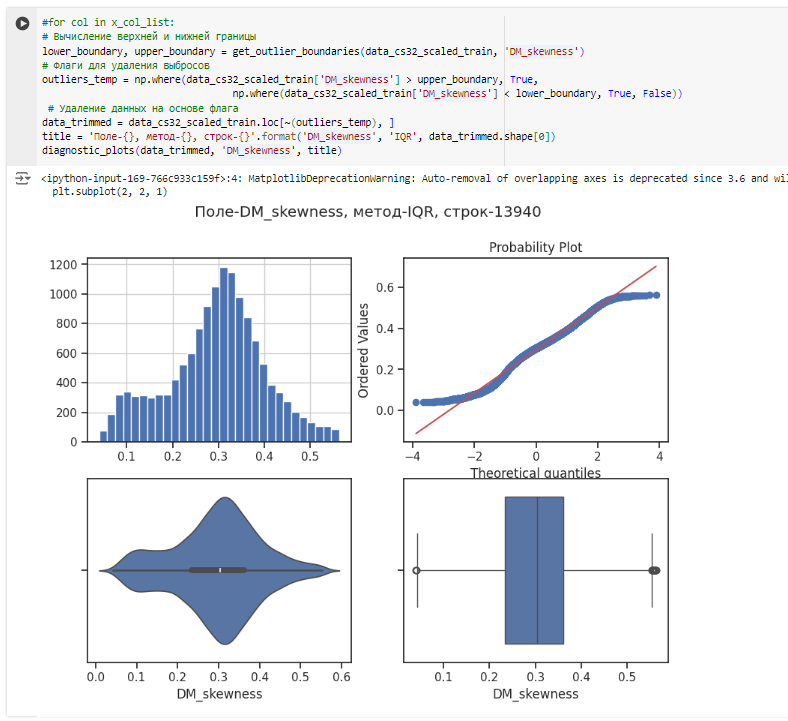






Удалим выбросы с использованием межквартильного размаха. Построим диагностические графики для колонки «DM\_skewness».



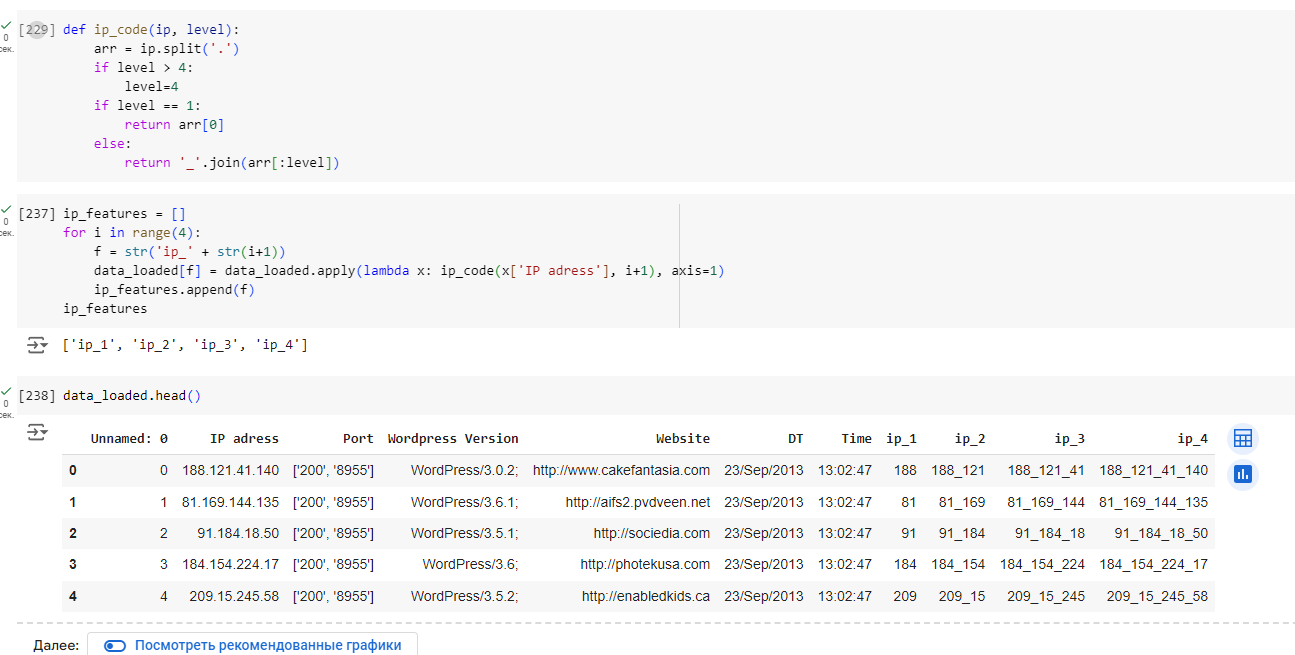


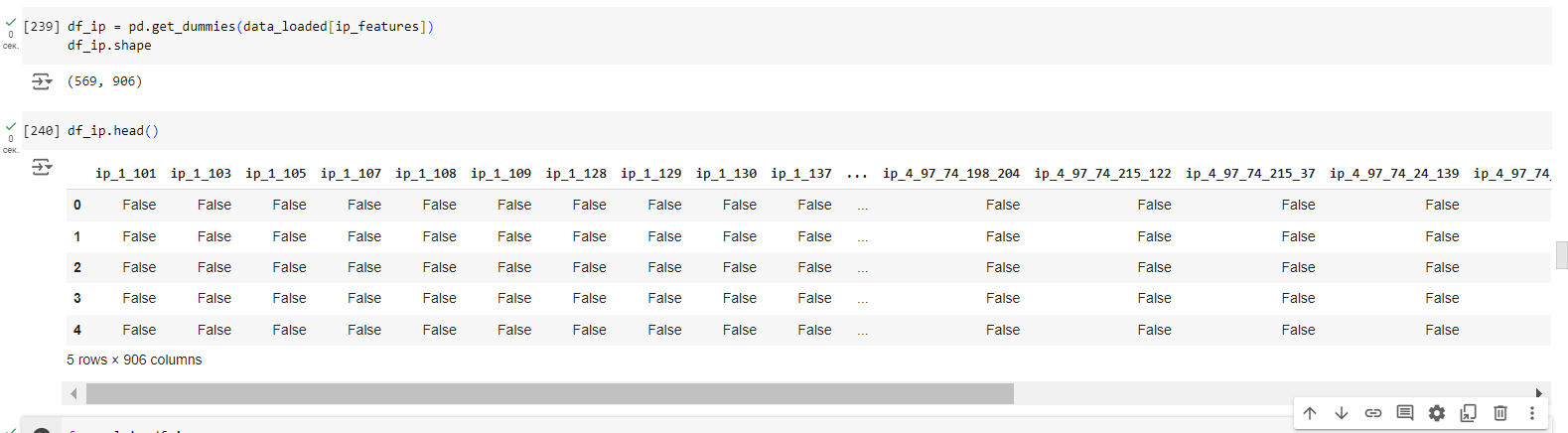
# **Обработка нестандартного признака**

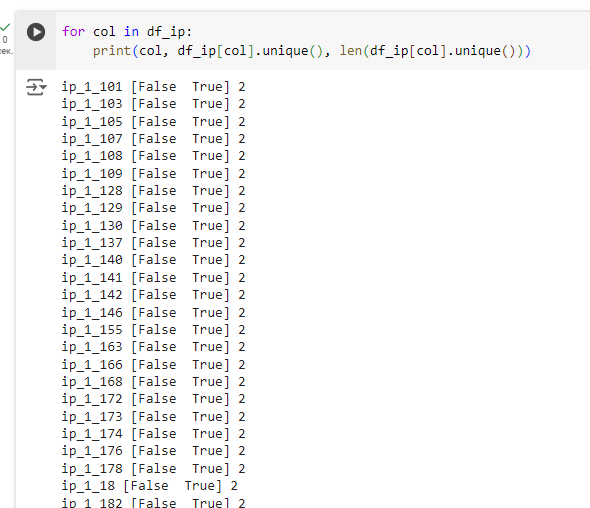
Для решения задачи обработки нестандартного признака используем другой набор данных – «WordPress DDos Log Dataset», который содержит информацию о DDoS-атаке на сервер. Сами данные взяты из журнала сервера.



Обработаем признак «IP adress», который содержит IP-адреса.





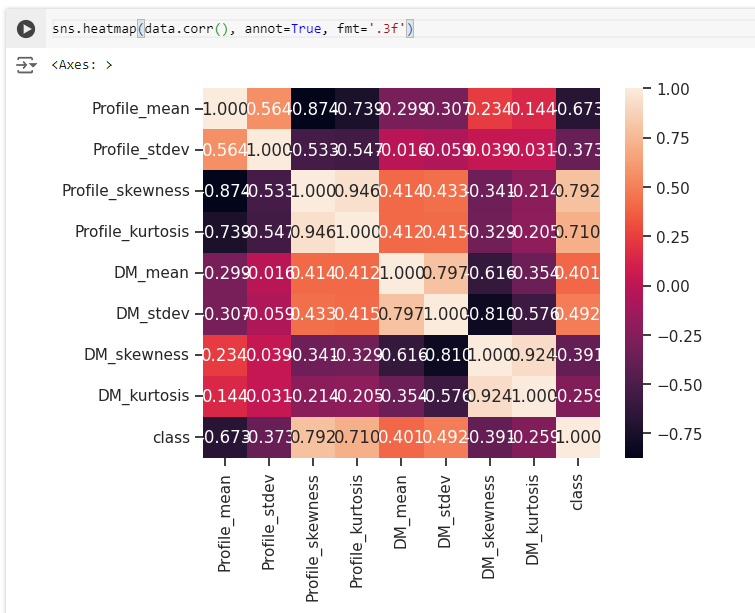


Таким образом, был обработан признак IP-адреса.

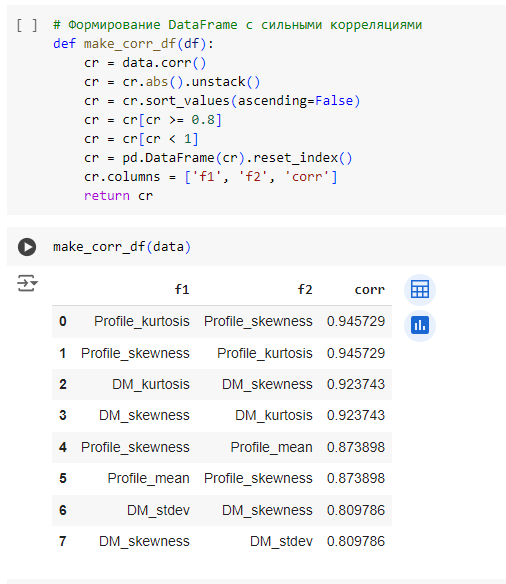
# **Отбор признаков**

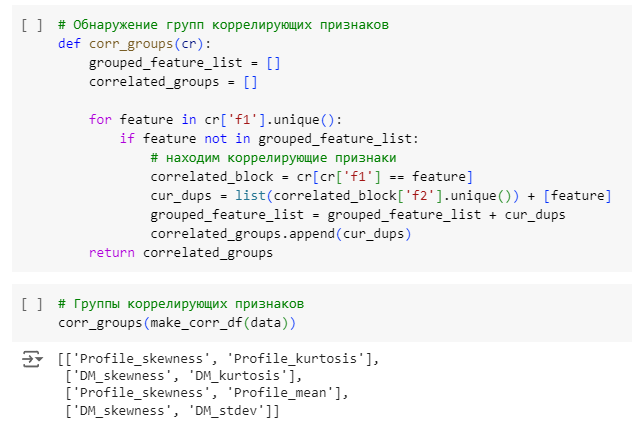
**Методы фильтрации**

Используем методы, основанные на корреляции. Построим корреляционную матрицу.

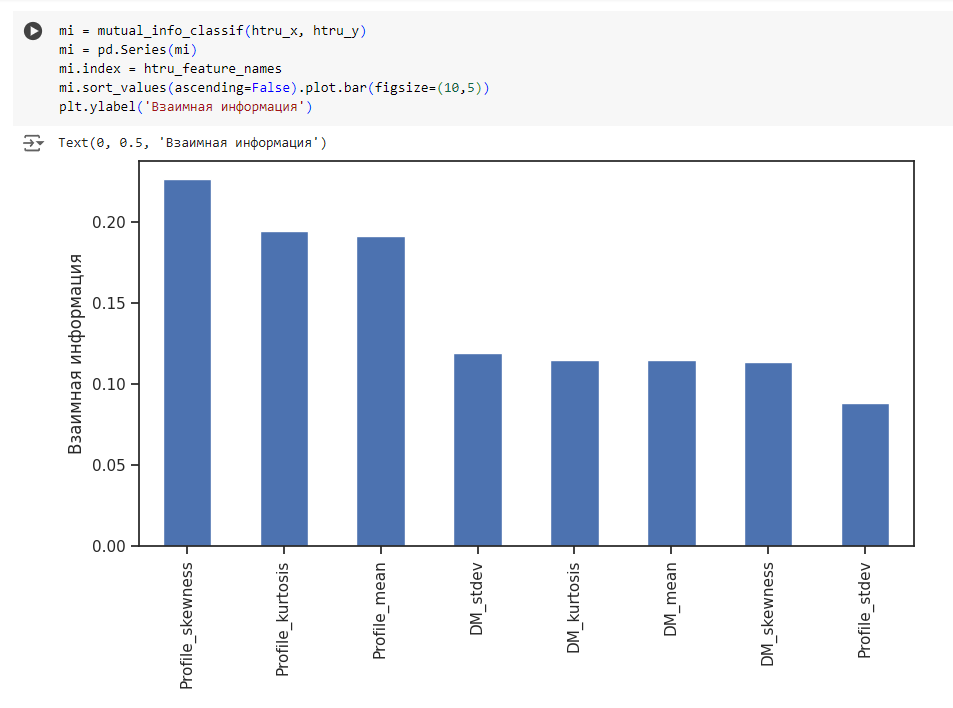


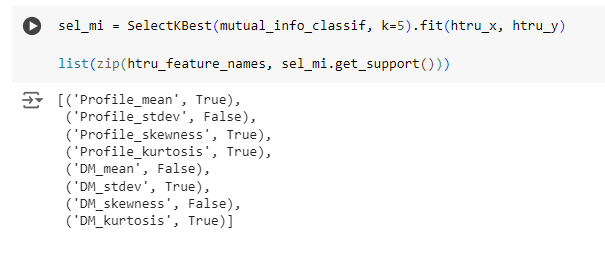
Из нее можно отметить, что признаки средне коррелируют с целевым, при этом между собой достаточно сильно. Найдем корреляции между признаками и выведем их группы.





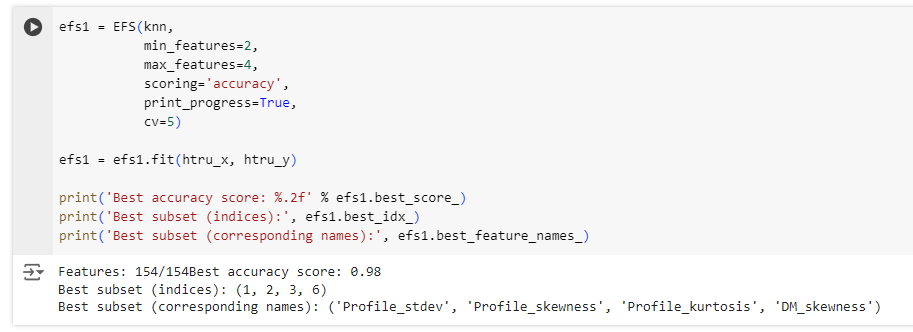
Используем методы, основанные на статистических характеристиках.

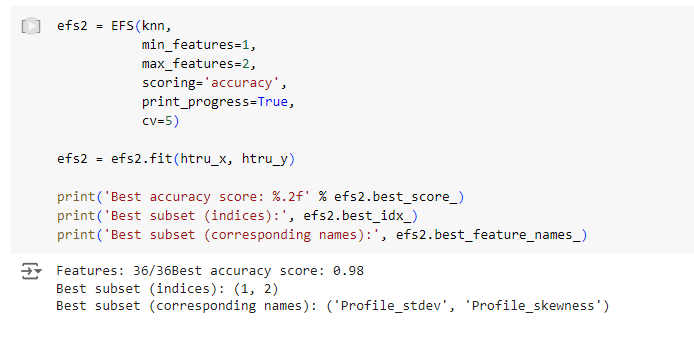




**Методы обертывания**

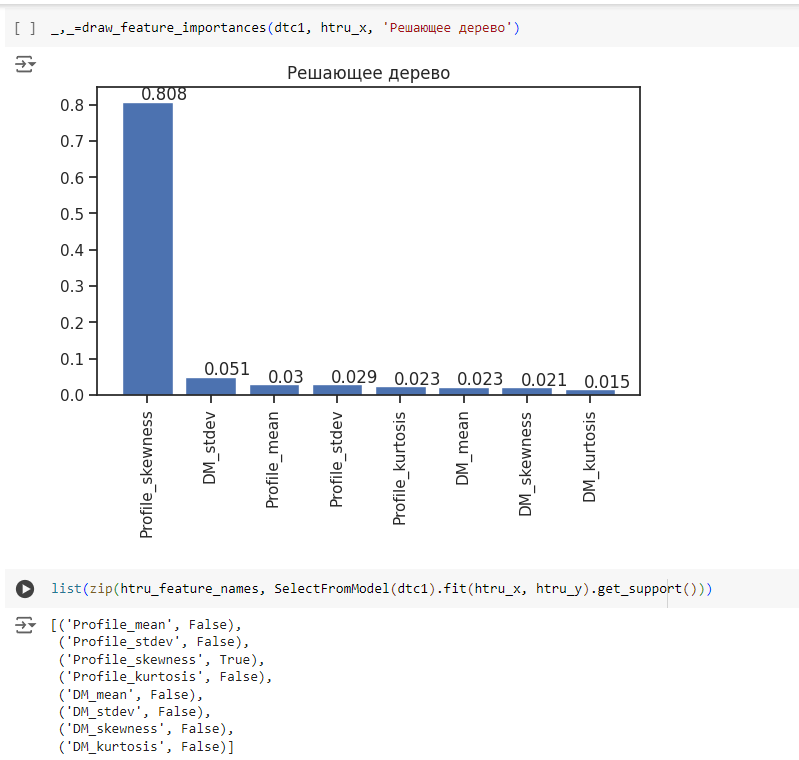
С помощью метода ExhaustiveFeatureSelector оценим хорошие 2-4 признака или 1-2 признака.

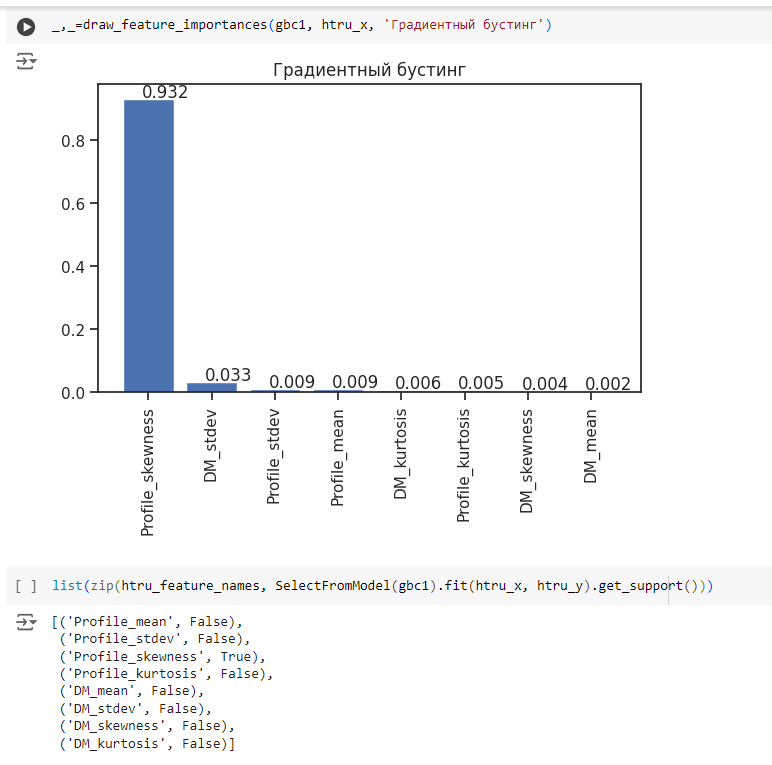
****

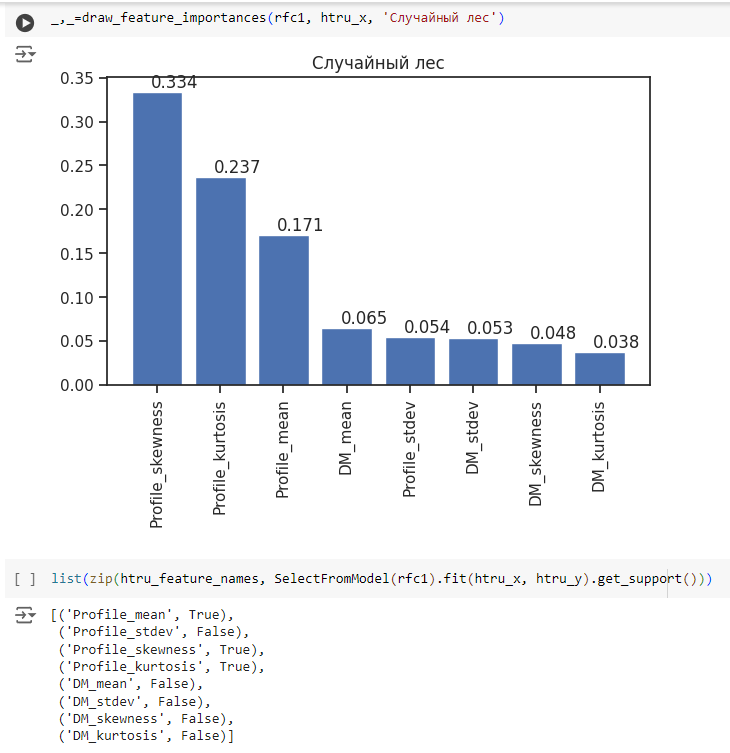
****

**Методы вложений**

Используем модель на основе решающего дерева и его ансамблевые модели.







Сведем оценки хороших признаков в таблицу 1.

Таблица 1. Хорошие признаки согласно отбору признаков

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Колонка | **SelectKBest** | **EFS1** | **EFS2** | **dtc** | **rfc** | **gbc** |
| *Profile\_mean* | Да |  |  |  |  | Да |
| *Profile\_stdev* |  | Да | Да |  |  |  |
| *Profile\_skewness* | Да | Да | Да | Да | Да | Да |
| *Profile\_kurtosis* | Да | Да |  |  |  | Да |
| *DM\_mean* |  |  |  |  |  |  |
| *DM\_stdev* | Да |  |  |  |  |  |
| *DM\_skewness* |  | Да |  |  |  |  |
| *DM\_kurtosis* | Да |  |  |  |  |  |

Из таблицы можно отметить, что все методы указали на то, что признак «Profile\_skewness» является хорошим. Более того, учитывая оценки всех методов, 8 признаков могут быть к сведены к 2-3 признакам, что значительно снизить общее количество признаков.