Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика, искусственный интеллект и системы управления»

Кафедра «Системы обработки информации и управления»



**Лабораторная работа № 1**

**по дисциплине «Методы машинного обучения»**

Создание "истории о данных"

ИСПОЛНИТЕЛЬ:

студентка ИУ5-23М

Морозевич М.А.

ПРЕПОДАВАТЕЛЬ:

Гапанюк Ю.Е.

\_\_\_ "\_\_\_\_\_\_\_\_" 2024 г.

Москва, 2024

# **Задание**

Необходимо изучить различные методы визуализации данных и создать историю на основе данных для набора данных:

* История должна содержать не менее 5 шагов (где 5 - рекомендуемое количество шагов). Каждый шаг содержит график и его текстовую интерпретацию.
* На каждом шаге наряду с удачным итоговым графиком рекомендуется в юпитер-ноутбуке оставлять результаты предварительных "неудачных" графиков.
* Не рекомендуется повторять виды графиков, желательно создать 5 графиков различных видов.
* Выбор графиков должен быть обоснован использованием методологии data-to-viz. Рекомендуется учитывать типичные ошибки построения выбранного вида графика по методологии data-to-viz. Если методология Вами отвергается, то просьба обосновать Ваше решение по выбору графика.
* История должна содержать итоговые выводы. В реальных "историях о данных" именно эти выводы представляют собой основную ценность для предприятия.

# **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.

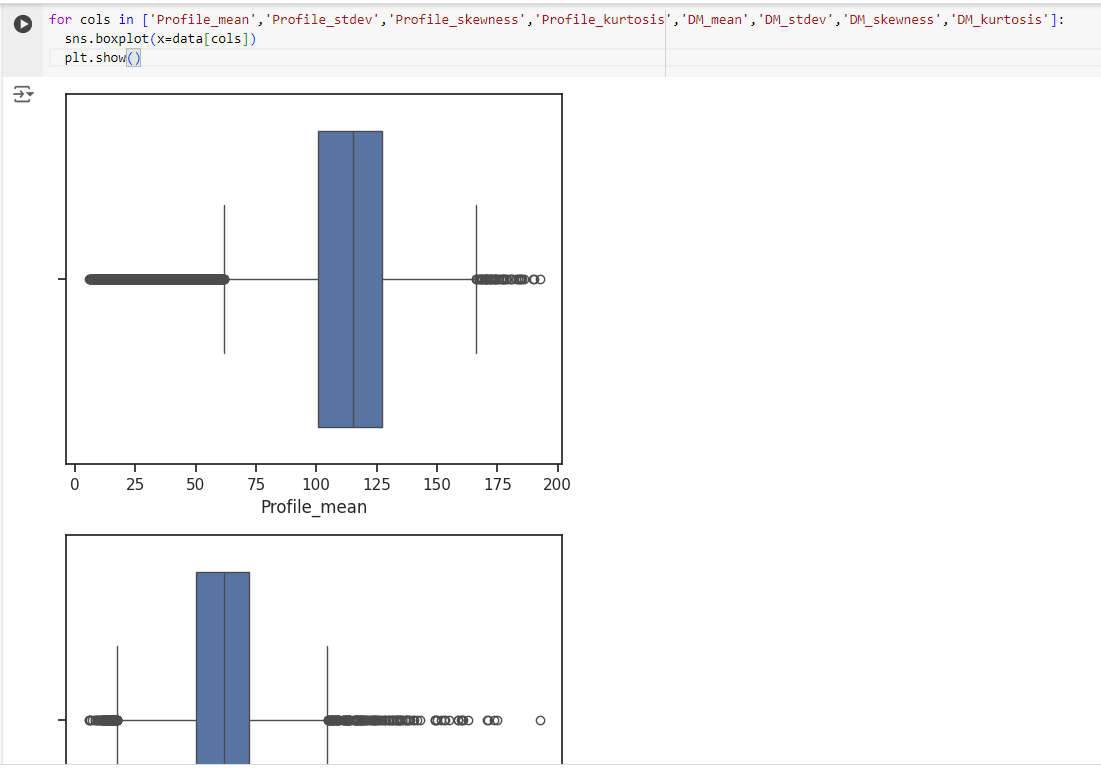
Названия и описания колонок набора данных:

* 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.

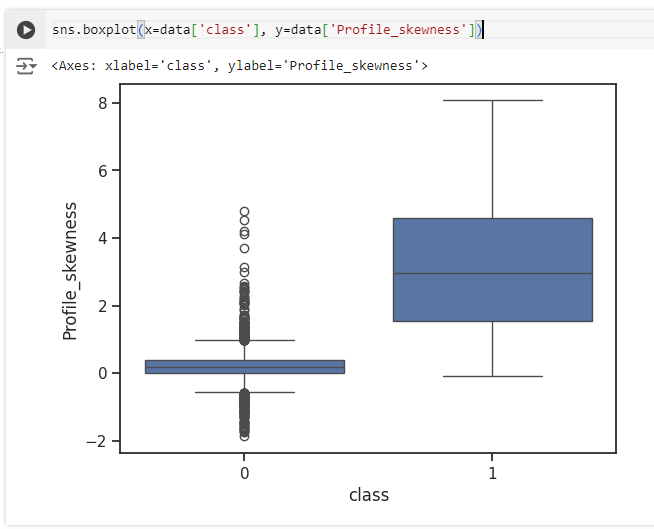
# **Исследование распределений признаков**

Построим графики «ящик с усами» для всех признаков.

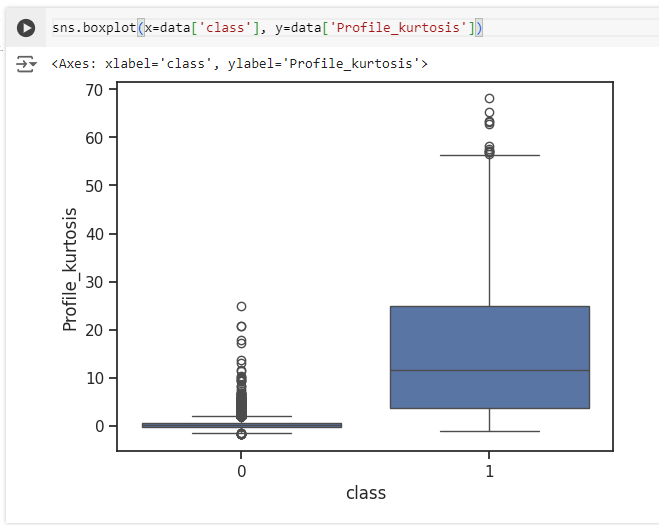


Из графиков можно отметить, что все признаки имеют ассиметричные распределения, далекие от нормального. Более того, многие значения могут быть выбросами.

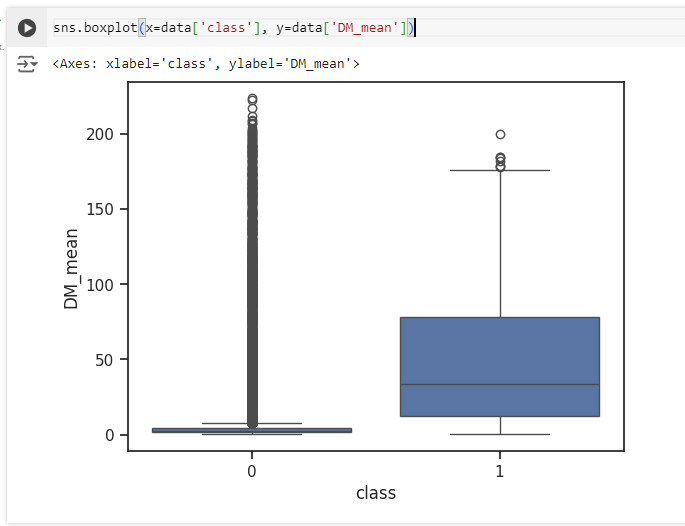
Проанализируем несколько признаков относительно целевого: «Profile\_skewness», «Profile\_kurtosis» и «DM\_mean».



Из графика видно, что значения для положительной оценки расположены около значений [2,5) признака «Profile\_skewness».

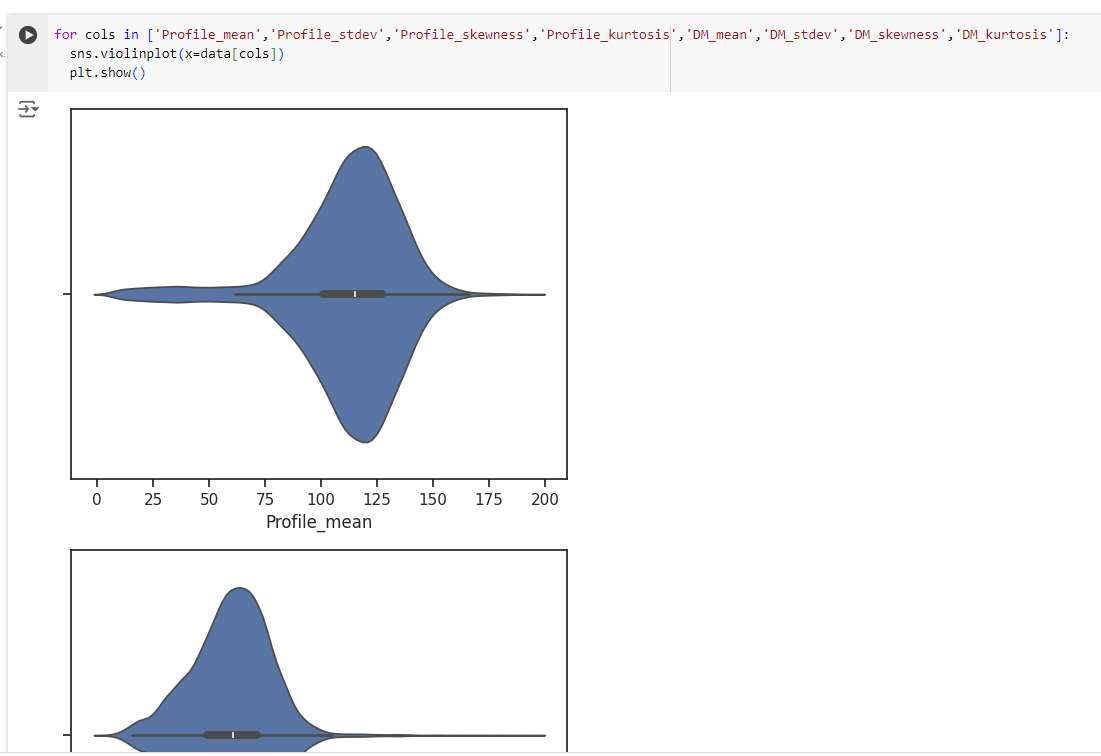


Из графика видно, что значения для положительной оценки расположены около значений [5,25] признака «Profile\_kurtosis».



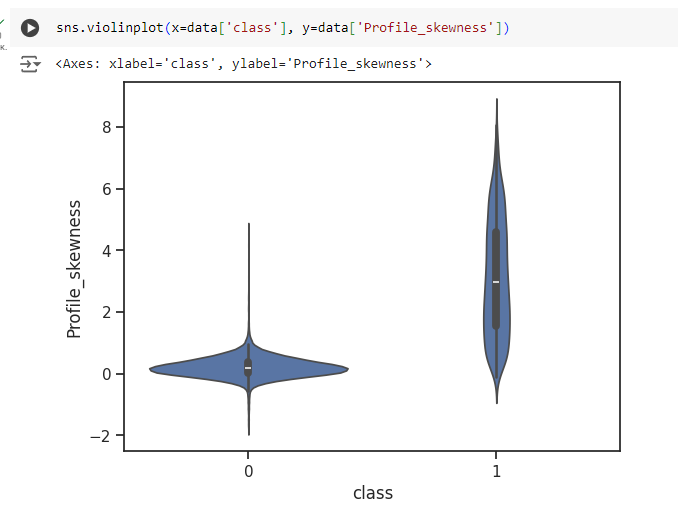
Из графика видно, что значения для отрицательной оценки расположены вдоль всего диапазона признака «DM\_mean».

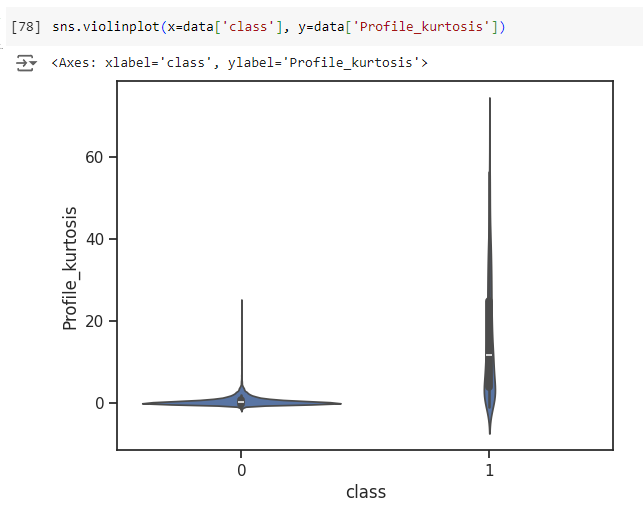
Построим графики-скрипки для всех признаков.



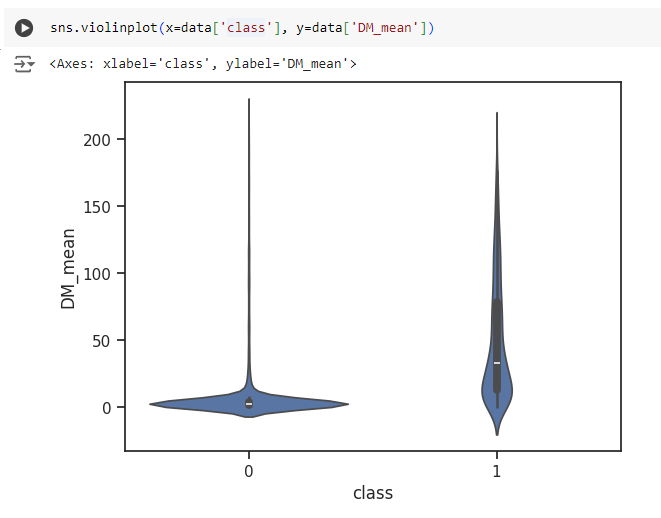
Из графиков можно отметить, что все признаки имеют ассиметричные плотности.

Проанализируем несколько признаков относительно целевого: «Profile\_skewness», «Profile\_kurtosis» и «DM\_mean».





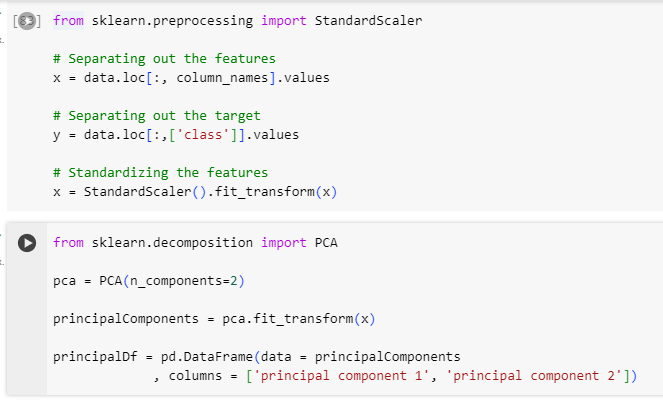
По графикам видно, что для отрицательного оценки большинство значений признака находится около 0.

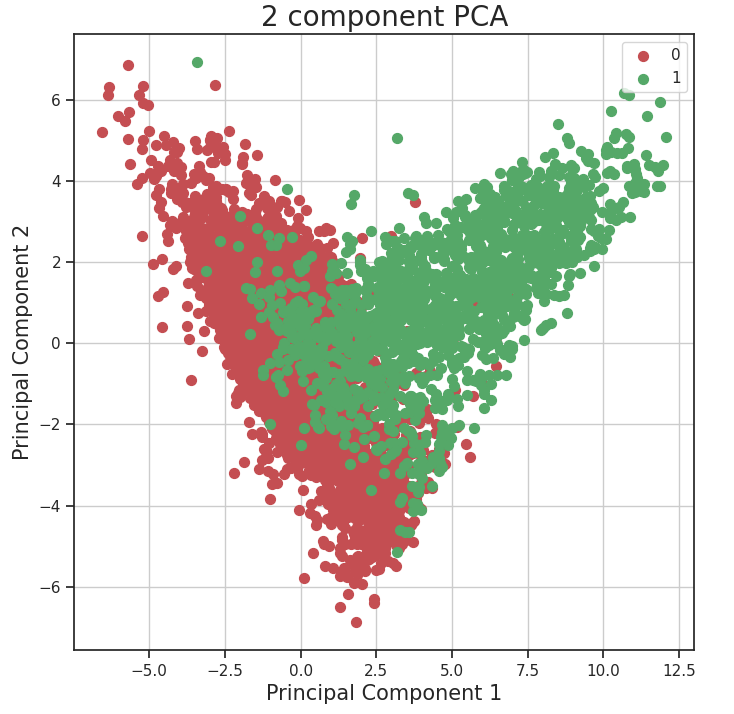


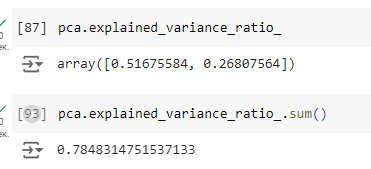
По графику видно, что большая часть значений признака распределена около 0 для обоих значений целевого признака.

1. **Исследование взаимосвязи между признаками**

Выполним PCA для создания меньшего числа новых признаков, которые будут содержать большую часть информации.

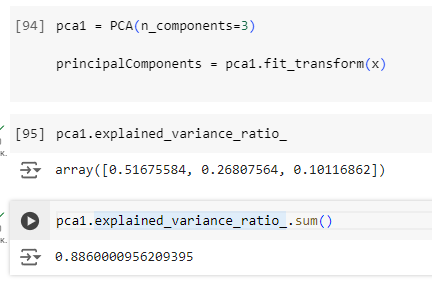






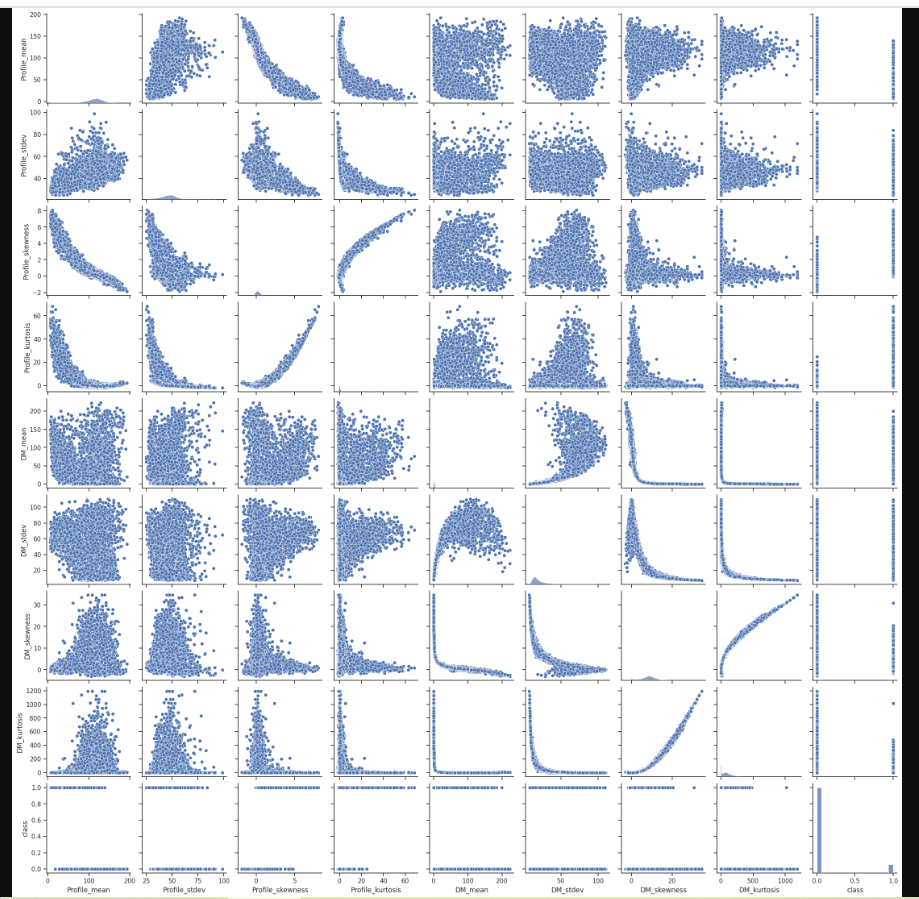
По результату анализа можно выделить 2 новых признака, которые будут содержать 78,48% информации из 8 исходных признаков.

Для увеличения объема информации попробуем PCA для 3 признаков.



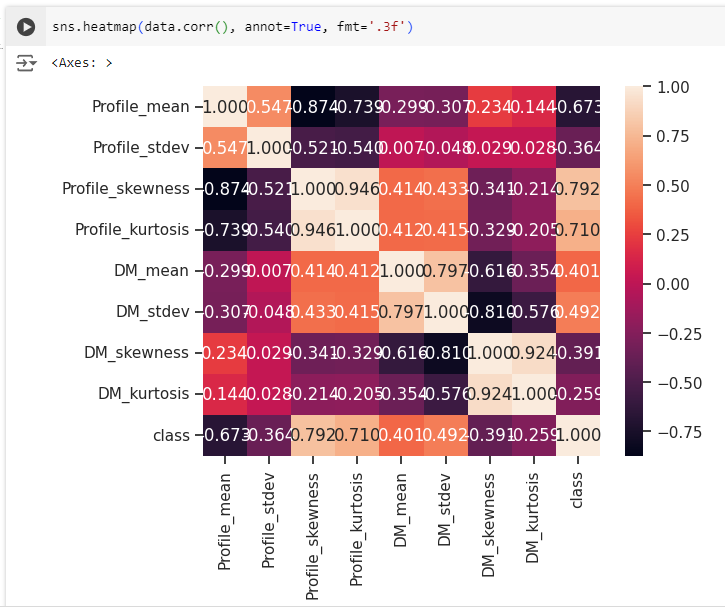
Добавление третьего признака позволило добавить еще 10,11% информации.

Следующим шагом построим Correlogram.



По графикам можно отметить, что между многими признаками наблюдается корреляция. Например, признаками «DM\_kurtosis» и «DM\_skewness». При этом можно выделить две группы коррелирующих признаков «Profile\_» и «DM\_», так как они описывают кривые и связаны между собой, но между признаками групп связь меньше.

Представим корреляции между признаками в виде тепловой карты.



Из нее можно отметить, что наибольшие по модулю значения корреляции наблюдаются в группах, а наименьшие – из разных групп.

# **Вывод**

Таким образом, можно отметить, что распределения признаков являются ассиметричными, и у большинства признаков есть выбросы. Более того, можно выделить две группы «Profile\_» и «DM\_» с сильными корреляциями между признаками, что соответствует логике создания набора данных с помощью двух кривых. При этом между признаками из групп слабая корреляция. Из этого можно сделать вывод, что в будущем можно будет снизить размерность набора данных.

С помощью PCA было продемонстрировано, что при выделении новых признаков можно получить признаки с количеством информации: первый – 51,66%, второй – 26,81%, третий – 10,11%. Соответственно, 8 признаков можно снизить до 3 признаков с сохранением до 88% информации.