

Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной
математики

Кафедра вычислительной математики и программирования

Лабораторная работа №0 по курсу «Искусственный интеллект»
Тема: Анализ и подготовка данных

Студент: В. В. Бирюков
Преподаватели: Д. В. Сошников
С. Х. Ахмед
Группа: М8О-307Б-19
Дата:
Оценка:
Подпись:

Москва, 2022

Лабораторная работа №0

Задача: Требуется определить задачу и найти под нее соответствующие данные. Необходимо вручную проанализировать данные, визуализировать зависимости, построить новые признаки, определить, хватит ли этих данных, и если не хватит найти еще.

1 Описание

В качестве задачи выбрана задача по бинарной классификации астрономических объектов согласно датасету [Predicting Pulsar Star](#).

Датасет направлен на определение пульсаров, и описывает каждый объект при помощи следующих признаков:

1. Mean of the integrated profile — среднее значение усредненного профиля импульса
2. Standard deviation of the integrated profile — отклонение усредненного профиля импульса
3. Excess kurtosis of the integrated profile — эксцесс усредненного профиля импульса
4. Skewness of the integrated profile — асимметрия усредненного профиля импульса
5. Mean of the DM-SNR curve — среднее значение кривой DM-SNR
6. Standard deviation of the DM-SNR curve — отклонение кривой DM-SNR
7. Excess kurtosis of the DM-SNR curve — эксцесс кривой DM-SNR
8. Skewness of the DM-SNR curve — асимметрия кривой DM-SNR
9. Class — целевое значение

2 Ход работы

Датасет содержит в себе данные о восьми параметрах объекта и его класс.

RangeIndex: 12528 entries, 0 to 12527

Data columns (total 9 columns):

Column Non-Null Count Dtype

```

0 Mean IP 12528 non-null float64
1 Std IP 12528 non-null float64
2 Kurtosis IP 10793 non-null float64
3 Skewness IP 12528 non-null float64
4 Mean DM-SNR 12528 non-null float64
5 Std DM-SNR 11350 non-null float64
6 Kurtosis DM-SNR 12528 non-null float64
7 Skewness DM-SNR 11903 non-null float64
8 class 12528 non-null int32

```

Все признаки являются количественными, дополнительного кодирования не требуется. Видно, что данные неполные, рассмотрим это подробнее.

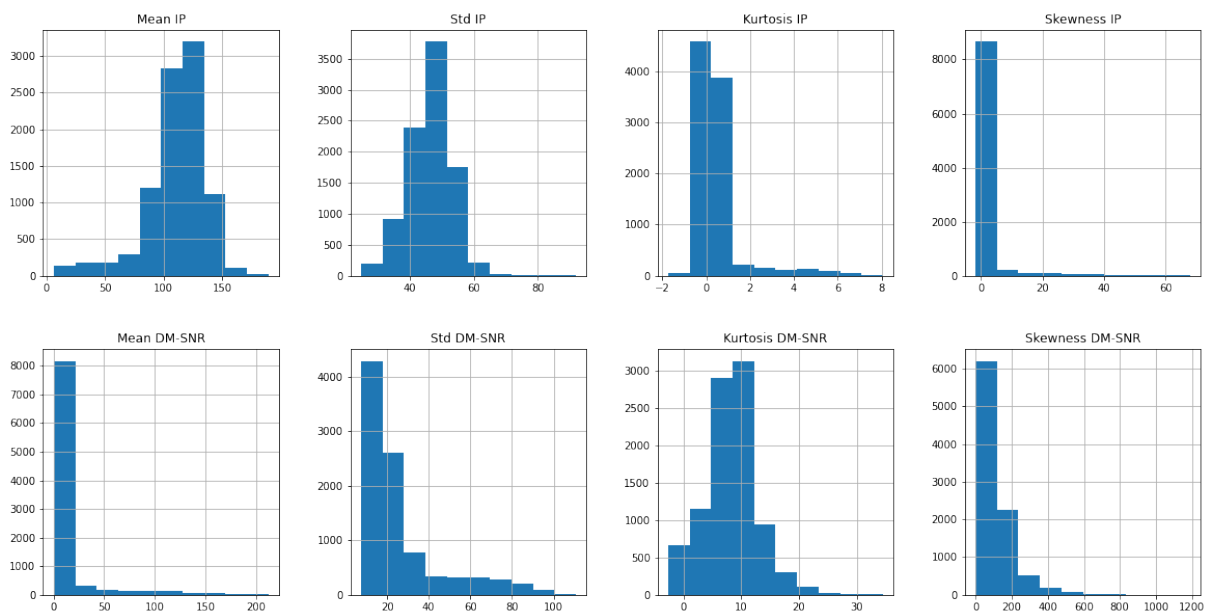
```

Mean IP 0
Std IP 0
Kurtosis IP 1735
Skewness IP 0
Mean DM-SNR 0
Std DM-SNR 1178
Kurtosis DM-SNR 0
Skewness DM-SNR 625
class 0

```

В качестве борьбы с пропусками удалим все строки, в которых они обнаружены. В результате данных станет немного меньше.

Визуализируем признаки при помощи гистограмм:

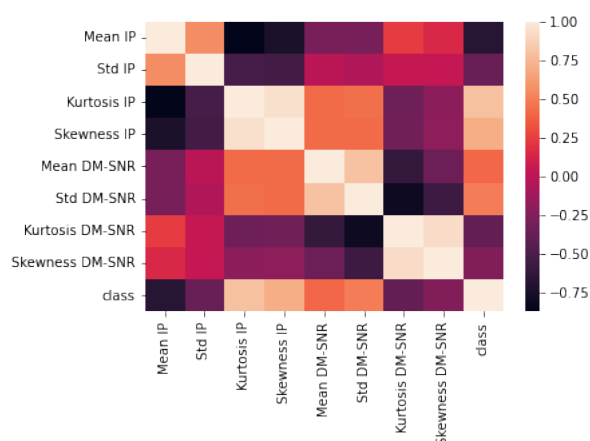


В распределениях признаков не наблюдается аномальных отклонений.

Подсчитаем коэффициенты корреляции между признаками:

	Mean IP	Std IP	Kurtosis IP	Skewness IP	Mean DM-SNR	Std DM-SNR	Kurtosis DM-SNR	Skewness DM-SNR	class
Mean IP	1.000000	0.554197	-0.872497	-0.734920	-0.299984	-0.307431	0.236010	0.146103	-0.675819
Std IP	0.554197	1.000000	-0.528370	-0.542560	-0.011061	-0.059486	0.036907	0.030959	-0.368223
Kurtosis IP	-0.872497	-0.528370	1.000000	0.944715	0.421126	0.436362	-0.344571	-0.216748	0.790866
Skewness IP	-0.734920	-0.542560	0.944715	1.000000	0.415570	0.415902	-0.328328	-0.204109	0.704743
Mean DM-SNR	-0.299984	-0.011061	0.421126	0.415570	1.000000	0.796449	-0.614526	-0.353186	0.407043
Std DM-SNR	-0.307431	-0.059486	0.436362	0.415902	0.796449	1.000000	-0.807013	-0.573260	0.493163
Kurtosis DM-SNR	0.236010	0.036907	-0.344571	-0.328328	-0.614526	-0.807013	1.000000	0.924326	-0.390352
Skewness DM-SNR	0.146103	0.030959	-0.216748	-0.204109	-0.353186	-0.573260	0.924326	1.000000	-0.258428
class	-0.675819	-0.368223	0.790866	0.704743	0.407043	0.493163	-0.390352	-0.258428	1.000000

Также в виде тепловой карты:

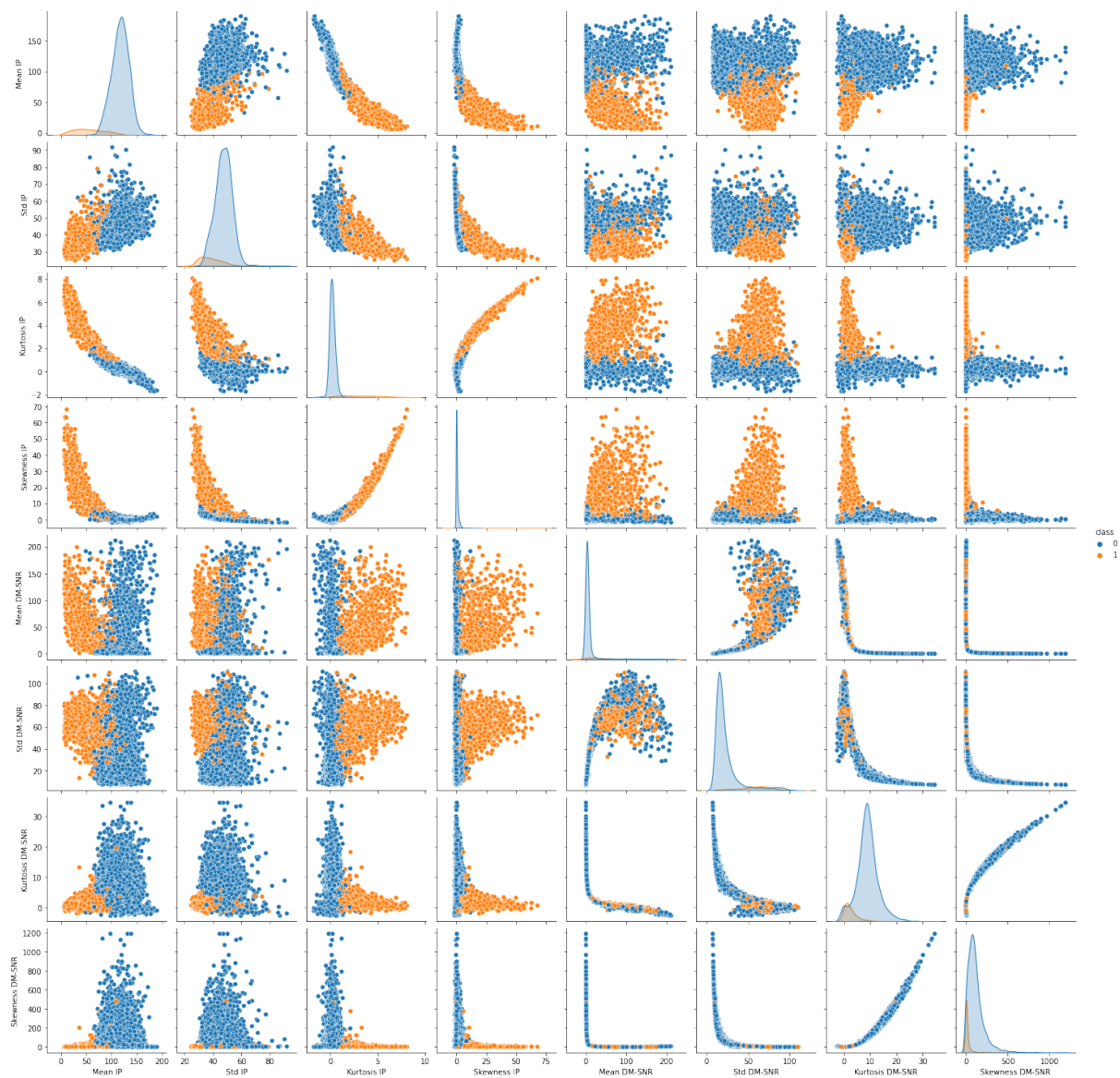


Видно, что некоторые признаки коррелируют друг с другом. Вероятно это связано с тем, что они представляют из себя статистические характеристики других данных. Отдельно отметим коэффициенты корреляции между признаками и целевым значением.

Mean IP -0.675819
Kurtosis DM-SNR -0.390352
Std IP -0.368223
Skewness DM-SNR -0.258428
Mean DM-SNR 0.407043
Std DM-SNR 0.493163
Skewness IP 0.704743
Kurtosis IP 0.790866

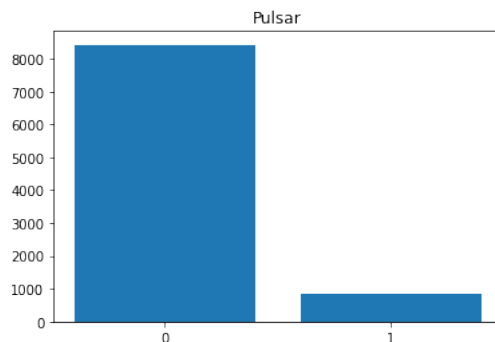
Имеются три признака с высокой корреляцией: Mean IP, Skewness IP, Kurtosis IP — их можно считать наиболее важными. Это говорит о возможности использования линейных моделей.

Также построим попарные графики зависимостей для всех признаков:



Такие графики подтверждают предыдущие выводы. На них видно и сильно коррелирующие признаки, и то, что на графиках признаков с высокой корреляцией с целевым значением, объекты разных классов можно достаточно хорошо разделить прямой. Отсюда также можно сделать вывод об отсутствии необходимости составления новых признаков.

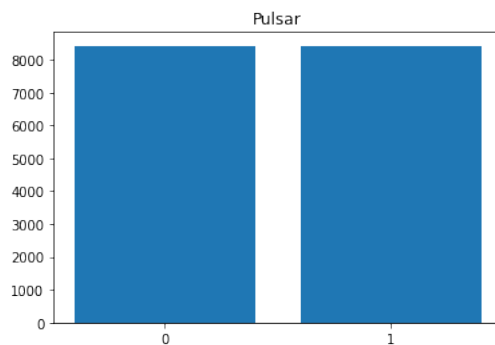
Важным параметром датасета является соотношение между классами.



Данные сильно несбалансированны, что плохо повлияет на обучение. Применим оверсэмплинг.

```
1 sample = data[data['class'] == 1]
2 while data[data['class'] == 1].shape[0] + sample.shape[0] <
3     data[data['class'] == 0].shape[0]:
4     data = pd.concat([data, sample])
5
6 data = pd.concat([data, sample.iloc[:data[data['class'] == 0].shape[0] -
7     data[data['class'] == 1].shape[0]]])
```

После оверсэмплинга:



3 Выводы

В ходе выполнения лабораторной работы был проведен анализ данных, для последующего их использования в обучении линейной модели. Данные визуализированы различными способами, что позволило сделать некоторые выводы о зависимостях между ними и применимости линейных моделей в целом. Также обнаружен дисбаланс классов, который исправлен методом оверсемплинга.