|  |
| --- |
| Пензенский государственный университет  Факультет вычислительной техники  Кафедра «Вычислительная техника» |
| Отчет  по лабораторной работе №  по дисциплине «»  на тему «Изучение набора данных»  Вариант № 1 |
|  |
|  |
| Проверил:  Карамышева Н.С. |
| Пенза |

**Задание.**

1. Для полученного набора данных и сделанного в лабораторной работе 1 описания провести предварительный анализ данных в соответствии с этапами 2-и и 3-й фазы методологии CRISP.

2. Описать имеющиеся данные.

3. Проверить гипотезы о наличии пропусков и выбросов в данных.

4. Выделить наиболее релевантные решаемой задаче признаки.

5. Сформировать предложения о дополнительном сборе данных (если это необходимо и возможно).

6. Оформить отчёт, содержащий информацию о выполнении пунк-тов 1-5 задания.

**Ход работы**

Так как данные были взяты с открытого источника они относятся к категории сторонних. В описании датасета говорится, что данные представляет из себя реальные частотные характеристики банкнот, вычисленные с изображений подлинных и поддельных образцов. Количество экземпляров: 1372. Частотные характеристики получены с помощью Wavelet Transform.

Таблица содержит 5 столбцов:

1 – Дисперсия вейвлет преобразования изображения.

2 – Асимметрия вейвлет преобразования изображения.

3 – Эксцесс вейвлет преобразования изображения.

4 – Энтропия изображения.

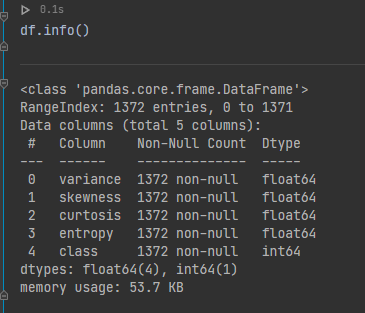
5 – Целевой класс. Подлинная купюра или поддельная.

Нашей задачей будет определить и выявить закономерности позволяющие отличить настоящие банкноты от фальшивых.

**Анализ данных**

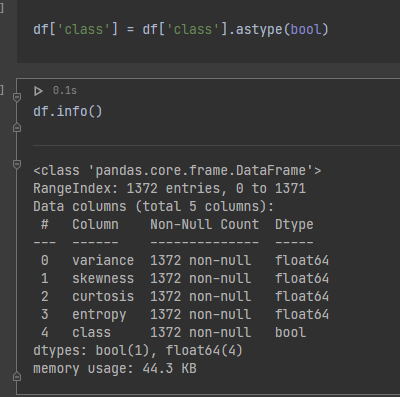
Рассмотрим данные и общие сведения о них.



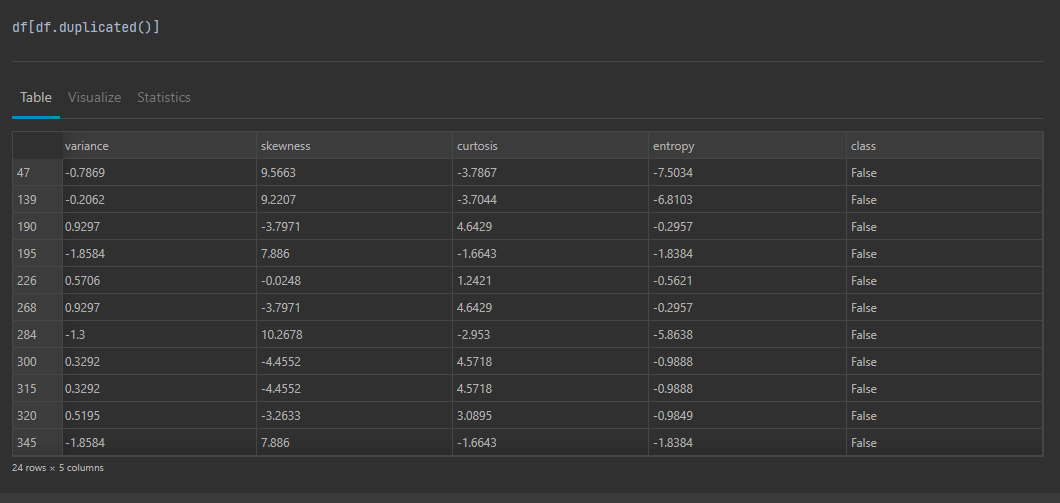


Видим, что количество записей в каждой столбце равно 1372, а значит пропусков в данных нет. Помимо этого, в таблице нет категориальных признаков.

Приведём целевой параметр к булевой переменной и уменьшим размер датасета.

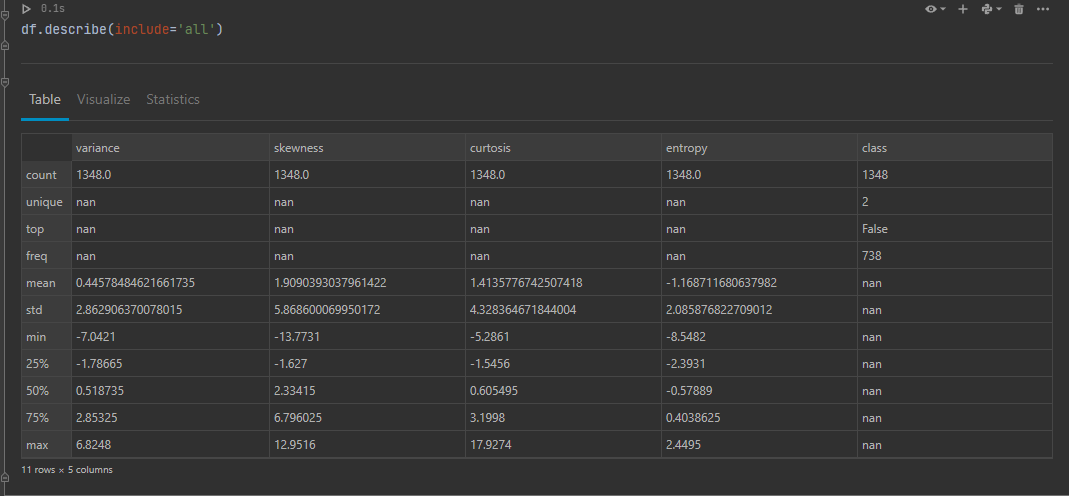


Мы уже знаем, что пропусков в данных нет, а есть ли дубликаты? Проверим это с помощью duplicated.



Действительно, 24 строки данных оказались дубликатами. Удалим их.

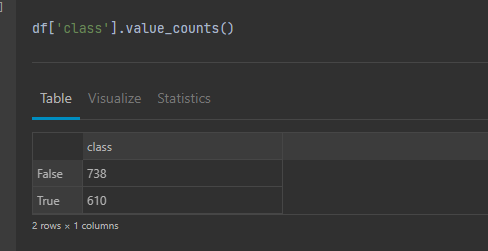
Теперь рассмотрим общие сведения о каждом параметре по отдельности.



Мы видим средние значения, стандартное отклонение, минимальные и максимальные значения для числовых признаков.

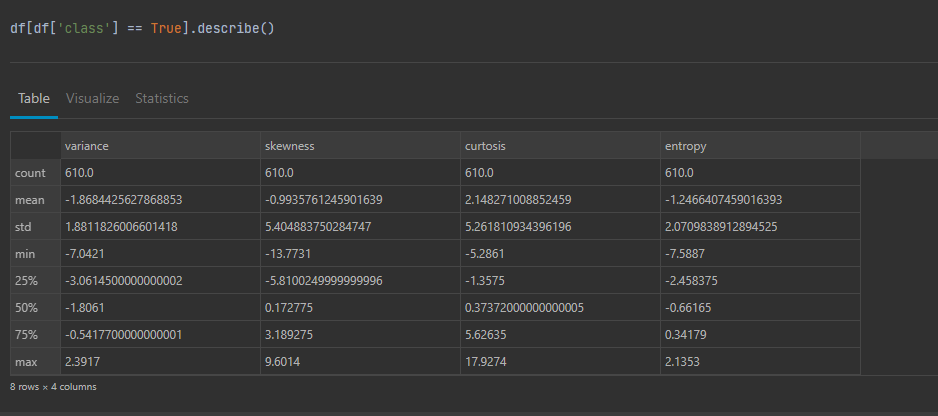
По этим данным сложно определить, какие параметры влияют на определение принадлежности класса, поэтому рассмотрим параметры для каждого класса по отдельности.

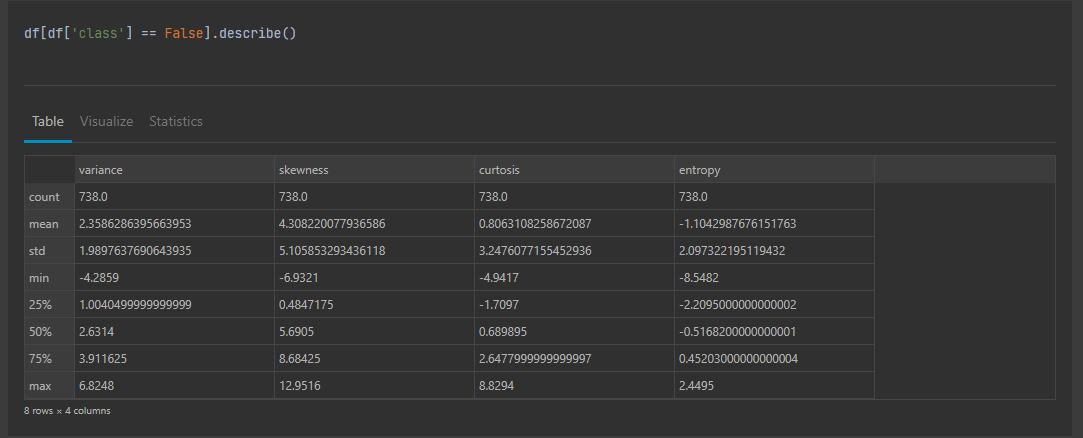
Сначала определим отношение количества фальшивых банкнот от подлинных.



Отношение близко к ½, что очень хорошо.

Детальнее исследуем каждый класс по отдельности.



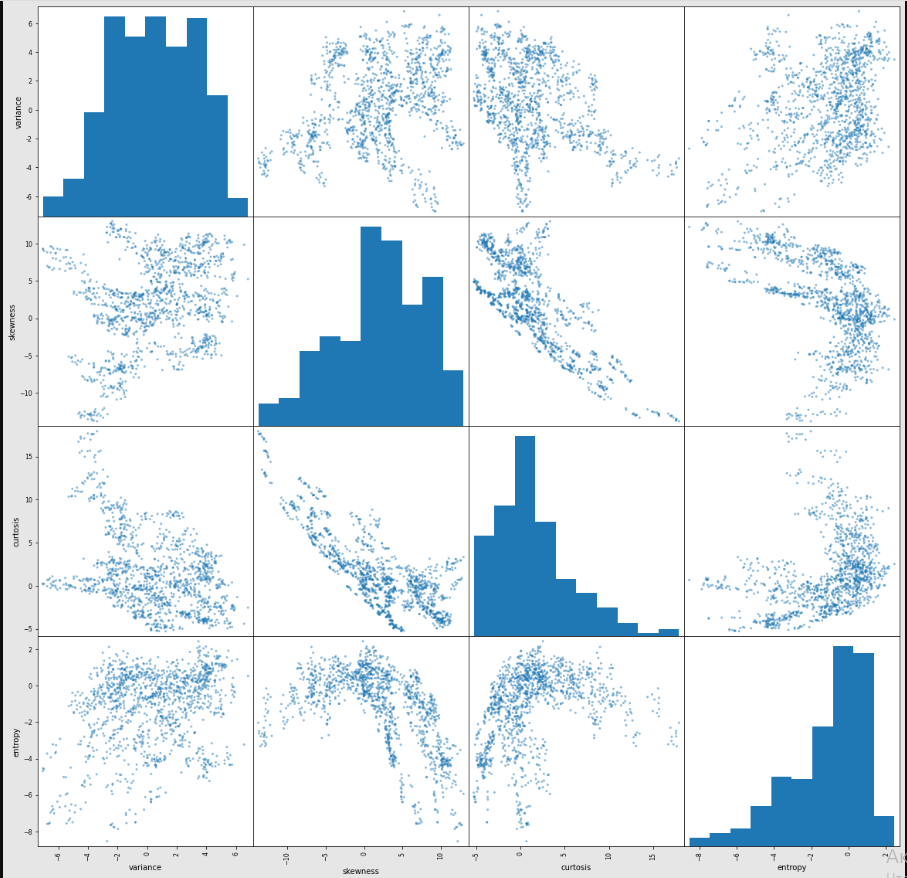


False означает, что купюра настоящая.

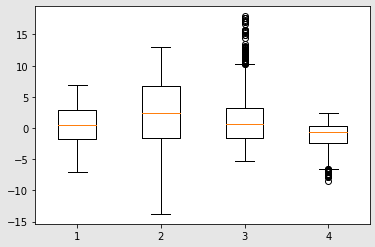
True означает, что купюра поддельная.

Теперь лучше видно, что, например, среднее значение skewness у подлинных образцов около 4, тогда как для поддельных -1. Различия есть и у параметров variance с curtosis.

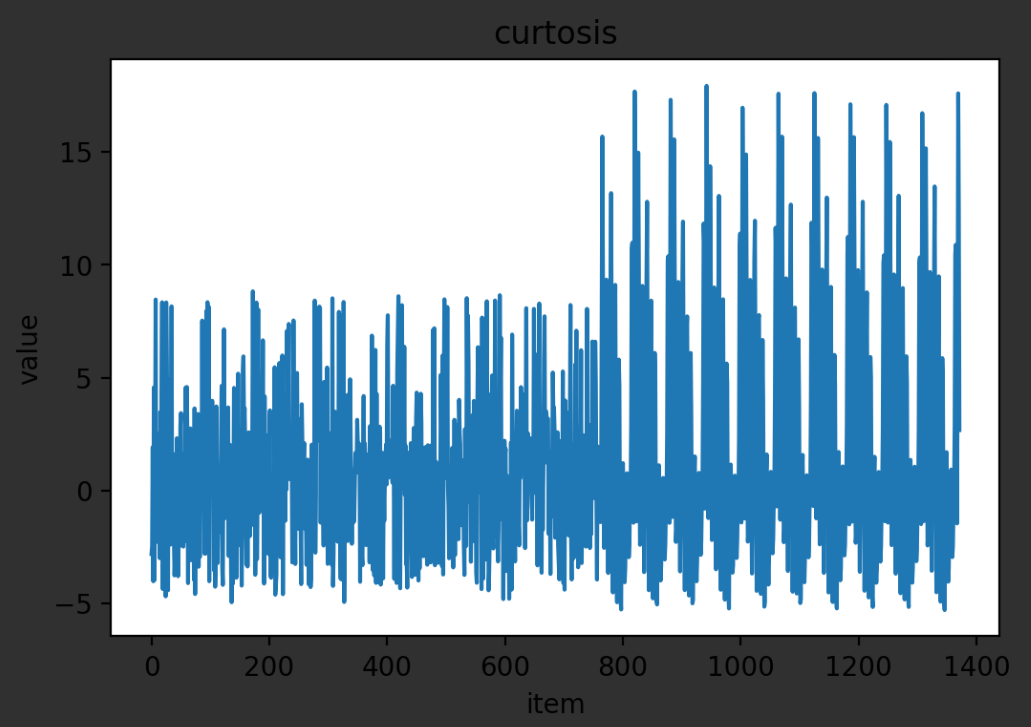
Теперь построим различных графики для лучшего понимания картины.



На диаграмме распределения видно, что признаки skewness и curtosis коррелируют между собой.

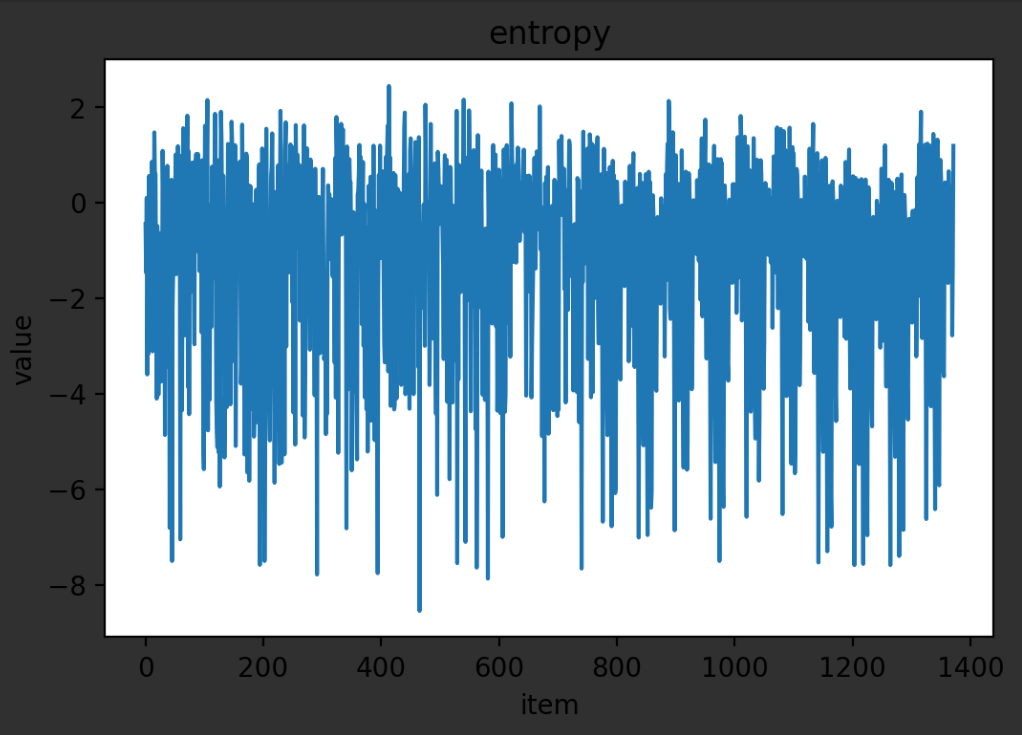


Параметр под номером 3 – это curtosis. Кружки сверху показывают, что в данном параметре есть выбросы, начиная со значения 10. Рассмотрим детальнее.

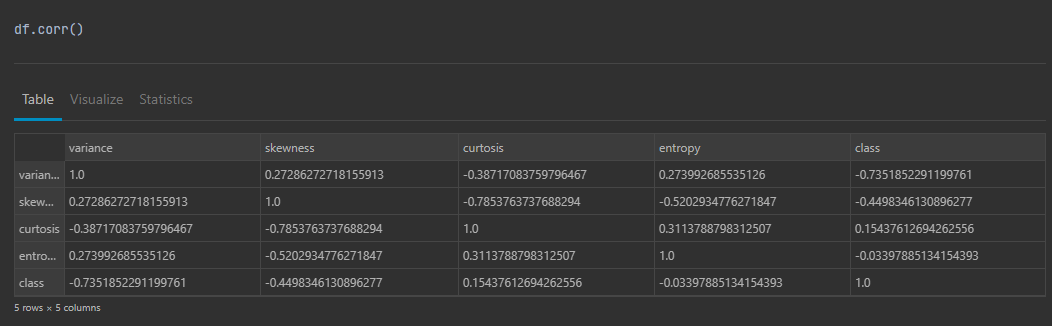


Действительно, начиная с примерно 800 образка значения имеют более высокие значения, но, стоит отметить, что такие значения могут быть присущи конкретному целевому классу, поэтому я бы не стал их изменять.

Стоит отметить параметр entropy, который действительно имеет некоторые отрицательные выбросы. Возможно, их стоит подкорректировать.



Напоследок рассмотрим корреляцию между признаками.



Видно, что коэффициент корреляции по модулю между curtosis и skewness составляет 0.78, значит признаки сильно дублируются, в дальнейшем один из параметров может быть удален.

Подводя итог, можно сделать вывод, что данные пригодны для использования и построение модели. В них отсутствуют пропуски, но присутствуют дубликаты, которые могут быть легко удалены. В параметрах curtosis и entropy есть незначительные выбросы.