## Препроцессинг датасета

### Задание

Необходимо провести чистку/восстановление признаков и отдельных экземпляров. Использовать тепловую матрицу для удаления сильно зависимых признаков, найти и выбрать наиболее важные для настройки модели признаки посредством алгоритма Gain Ratio (коэффициент усиления) из алгоритма классификации с помощью деревьев решений C4.5. Рассказать о проделанной работе и обосновать каждое преобразование датасета.

Датасет: Качество вина

### Пояснение параметров

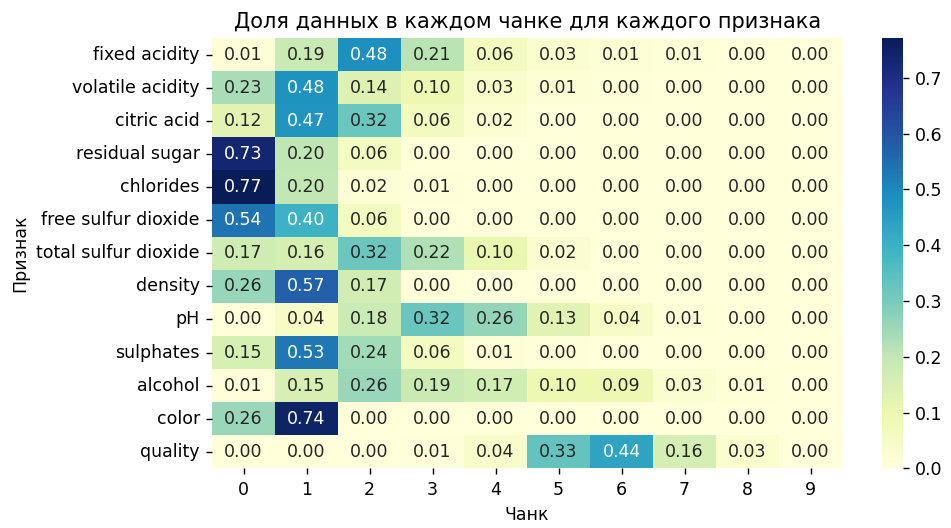
|  |  |  |
| --- | --- | --- |
| Название | Смысл | Плотность распределения |
| 1. fixed acidity Фиксированная кислотность | Количество нелетучих кислот в вине (в основном винная и яблочная кислоты). Эти кислоты важны для вкуса и стабильности вина |  |
| 2. volatile acidity  Летучая кислотность | Количество летучих кислот (в основном уксусной кислоты). Высокое содержание летучих кислот может указывать на порчу вина |  |
| 3. citric acid  Лимонная кислота | Количество лимонной кислоты, которая добавляется в вино для повышения кислотности и свежести вкуса |  |
| 4. residual sugar  Остаточный сахар | Количество сахара, оставшегося после завершения брожения. Влияет на сладость вина |  |
| 5. chlorides  Хлориды | Количество соли в вине. Высокое содержание хлоридов может указывать на низкое качество вина |  |
| 6. free sulfur dioxide  Свободный диоксид серы | Количество свободного SO₂, который используется как консервант для предотвращения окисления и роста бактерий |  |
| 7. total sulfur dioxide  Общий диоксид серы | Общее количество SO₂ в вине, включая свободный и связанный диоксид серы |  |
| 8. density  Плотность | Плотность вина, которая зависит от содержания алкоголя, сахара и других веществ. Обычно измеряется относительно плотности воды |  |
| 9. pH  Уровень pH | Мера кислотности вина. Низкий pH указывает на высокую кислотность, а высокий pH — на низкую кислотность |  |
| 10. sulphates  Сульфаты | Количество сульфатов (солей серной кислоты), которые используются как консерванты и антиоксиданты |  |
| 11. alcohol  Алкоголь | Содержание алкоголя в вине, выраженное в процентах |  |
| 12. color  Цвет вина | Категориальный параметр, указывающий на тип вина (красное или белое) |  |
| 13. quality  Качество вина | Оценка качества вина, обычно на основе экспертной оценки. Часто используется шкала от 0 до 10 |  |

### 1. Чистка и восстановление данных

* Заполнить отсутствующие данные
* Удалить дублирующиеся данные
* Удалить неверные данные
* Усреднить выбросы
* Привести данные к единому формату

При чистке данных решил ничего не делать с выбросами, так как они могут указывать на исключительно хорошие вина. Неверных или отсутствующих данных не было. Дублирующиеся строчки удалил.

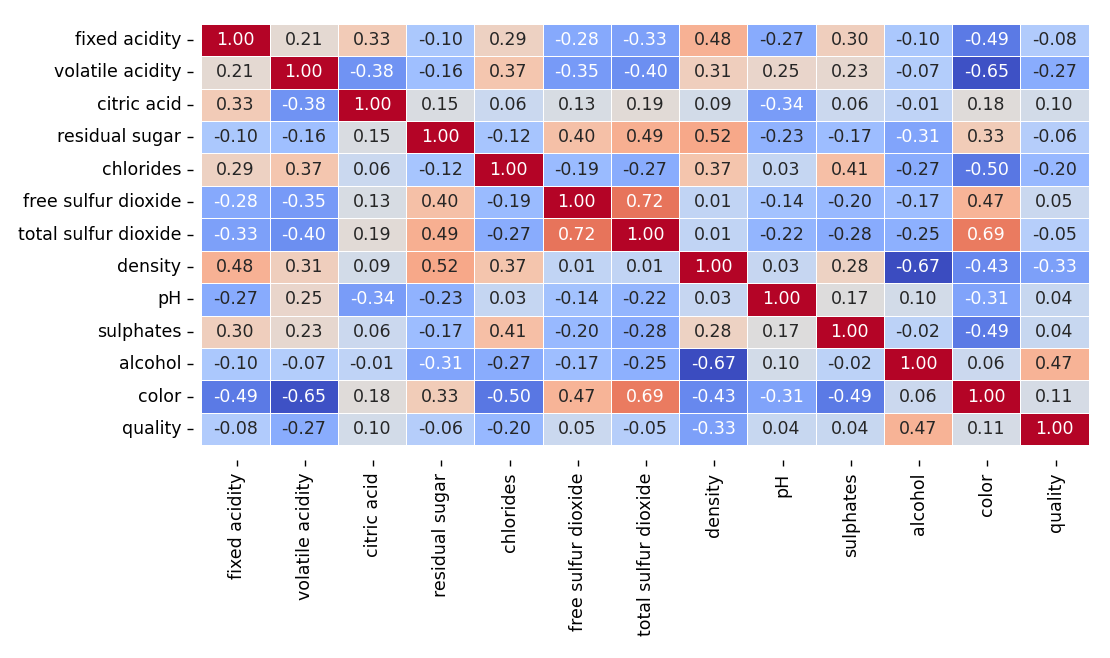
Также нормализовал и дискретизовал некатегориальные признаки, чтобы в дальнейшем применить gain ratio. Количество чанков для дискретизации взял 10.

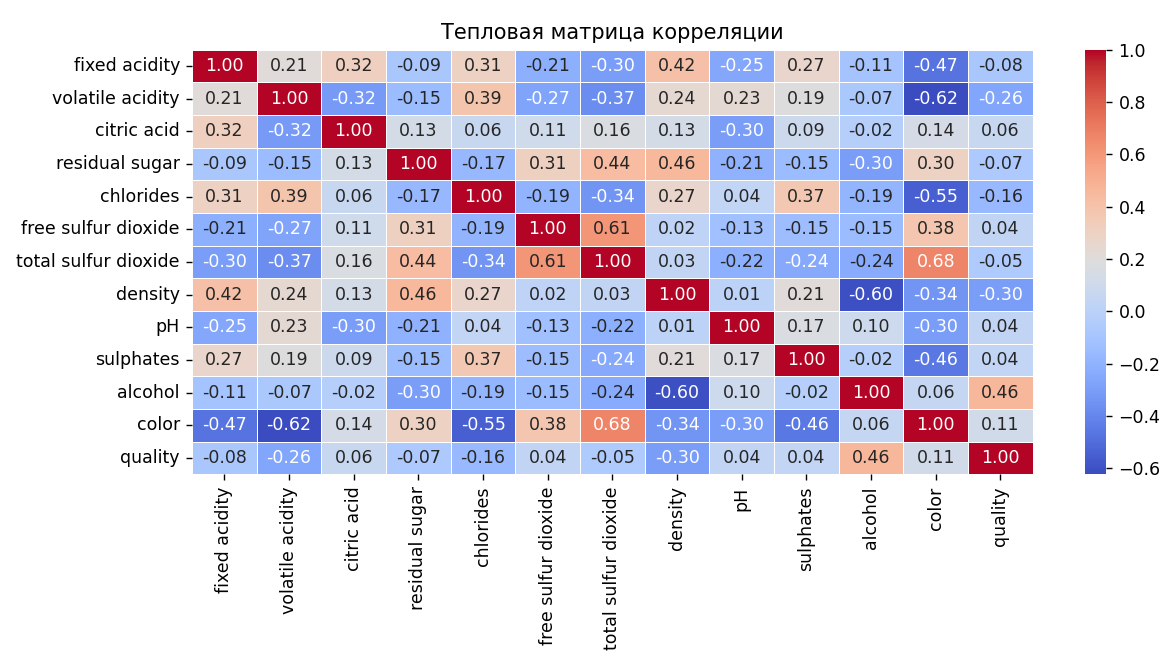


### 2. Анализ корреляции признаков

* Построить тепловую матрицу корреляции (heatmap) для визуализации зависимости между признаками
* Определить сильно коррелирующие признаки
* Удалить сильно коррелирующие признаки, чтобы избежать мультиколлинеарности (это улучшит качество модели и ускорит обучение)

Тепловая матрица, построенная с помощью библиотеки:

Тепловая матрица, построенная вручную:



Как видно по тепловой матрице, “общий диоксид серы” и “цвет вина” больше всего коррелируют с остальными признаками. Их можно удалить.

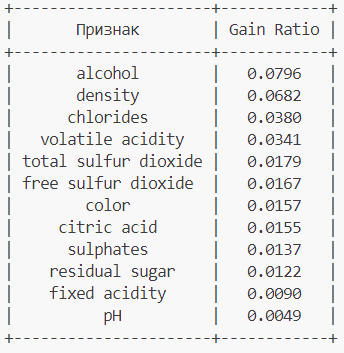
А на качество вина больше всего влияет “алкоголь”, “плотность”, “летучая кислотность” и “хлориды”

### 3. Выбор наиболее важных признаков

* Использовать алгоритм Gain Ratio для оценки важности признаков
* Выбрать признаки с наибольшим значением Gain Ratio, так как они наиболее информативны для модели

Как и ожидалось, на качество вино вина больше всего влияет “алкоголь”, “плотность”, “летучая кислотность” и “хлориды”.

Ручное вычисление Gain Ratio:



Вычисление Gain Ratio с помощью библиотечных функций (результат аналогичный):

