Определение качества вина

1 Общее представление

1.1 Введение

Качество вина - есть, совокупность свойств, делающих его приемлемым или желательным для потребителя, на которого производят впечатление особенности вина, вызывающие приятные ощущения. Поэтому проблема качества должна решаться с помощью технологии производства вина, которая направлена прежде всего на сохранение и развитие этих особенностей (Ларреа, 1956).

1.2 Проблема

Качество является совокупностью приятных вкусовых ощущений, непосредственно связанных с химическим составом вина. Но хорошо известно, насколько неточными и субъективными являются определение и оценка вкусовых свойств вина, с одной стороны, и трудность связать их с химическим составом вина — с другой. Необходимо иметь какой-то достаточный инструментарий для оценки качества вина по заданным характеристикам уменьшая субъективный фактор оценки.

1.3 Решение

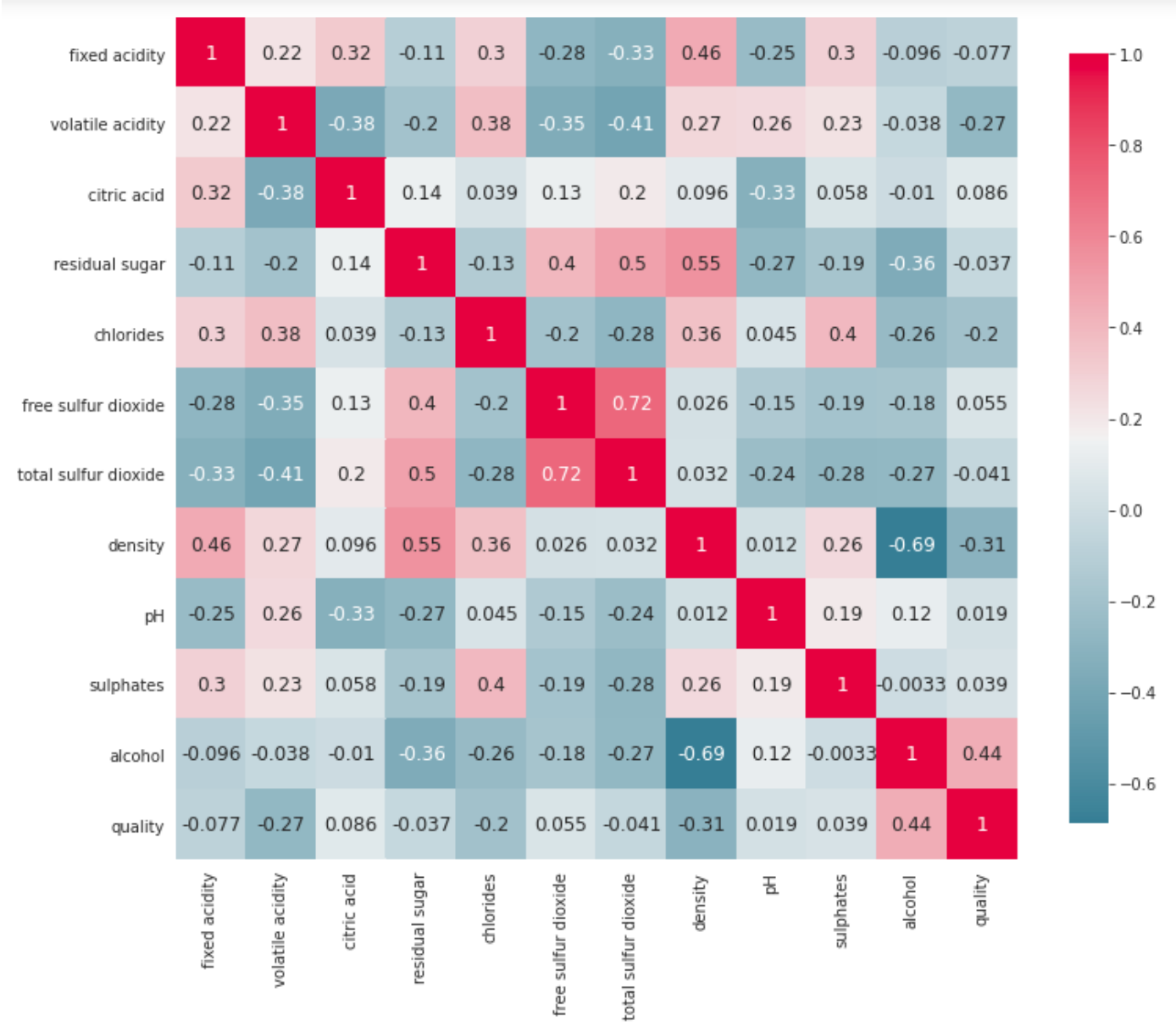
Анализируем имеющиеся в распоряжении данные характеристик вин, выделяем нужные признаки, строим модель классификации качества вин и оцениваем ее качество. На выходе получаем рабочую модель оценки качества вина без привлечения сомелье.

2 Реализация

2.1 Анализ данных и выделение нужных признаков

Для анализа представлен датасет (https://www.kaggle.com/rajyellow46/wine-quality), состоящий из 11 характеристик около 6500 вин и оценок их качества профессиональными сомелье по шкале от 0 до 10.

Характеристики и их корреляция представлены на изображении ниже.



Есть несколько признаков, скоррелированных с качеством вина, однако, корреляция не настолько велика, чтобы только лишь по ним определять качество.

Взаимная корреляция признаков не достаточно высока, чтобы исключить какие-то признаки.

Датасет не требует большого объема обработки: пропусков не много, явных выбросов по значениям нет.

Требуется нормализация характеристик датасета перед построением модели.

2.2 Построение модели и оценка качества

Строим 2 варианта модели, чтобы можно было выявить среди них наиболее удачную.

Варианты следующие:

- на простом классификаторе - KNN;
- на более сложном классификаторе - случайный лес.

Ниже представлена таблица оценки качества обоих вариантов моделей на тренировочной и валидационной выборках.

	Тестовая выборка	Валидационная выборка
KNN	0.66	0.53
Случайный лес	0.85	0.62

3 Заключение

В результате проделанной работы была построена модель, способная на наборе характеристик вина предсказывать его качество. Модель была представлена в 2-х вариантах. Лучше себя показал вариант на классификаторе "случайный лес". Однако, качество модели оказалось не слишком высокое.

Вот ряд рекомендаций по улучшению качества модели:

- сформировать сбалансированную выборку классов для обучения модели, либо реализовать функцию потерь, учитывающую несбалансированность классов;
- найти источники новых признаков и добавить их в датасет;
- собрать характеристики большего кол-ва вин.