Wine quality

Постановка задачи

Имеется данные о красном и белом португальском вине «Виню Верде). Данные включают информацию о физико-химических свойствах:

- type: тип (красное или белое)
- fixed acidity: фиксированная кислотность
- volatile acidity: летучая кислотность
- citric acid: лимонная кислота
- residual sugar: остаточный сахар
- chlorides: хлориды
- free sulfur dioxide: свободный диоксид серы
- total sulfur dioxide: общий диоксид серы
- density: плотность
- pH: кислотность
- sulphates: сульфаты
- alcohol: крепость вина

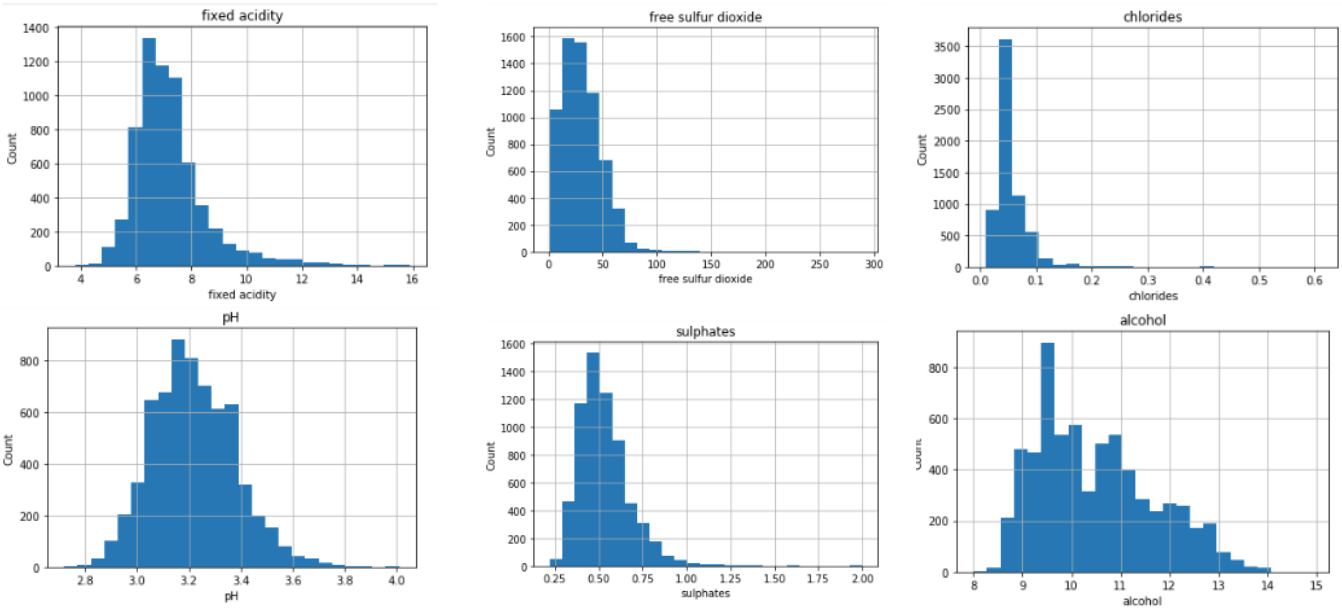
Для каждого вида вина выставляется оценка от 0 до 10.

	type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

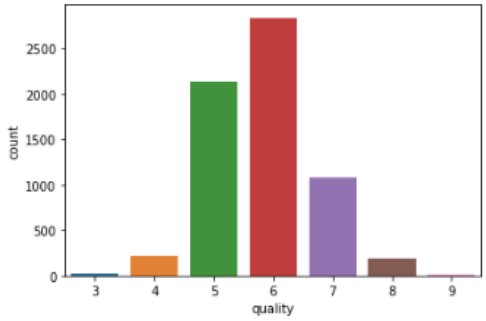
Задача: предсказать оценку вина, зная его физико-химические свойства и цвет (белое или красное)

Анализ данных

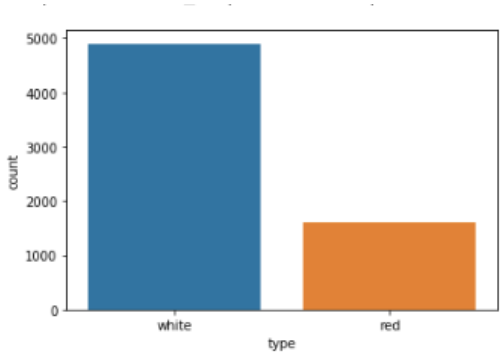
Для решения задачи важно понять, какие именно физико-химические свойства играют самую значимую роль при выставлении оценки. Для этого рассмотрено распределение признаков - распределение для всех свойств имеют вид нормального, гауссовского, можно говорить о некоторых наиболее часто встречающихся значений.



Распределение оценки вина говорит о том, что чаще всего встречается оценка 6, оценки 0, 1, 2 и 10 не ставились



С помощью тепловой карты, установлено, что с оценкой вина очень связаны признаки alcohol, density,chlorides, volatile acidity, type. Поскольку люди предпочитают белое вино, и переменная сильно не влияет на на категорию, то type не учитываем.



Модель обучения и результат

Для построения моделей были выбраны:

- методы регрессии:
  - линейная и регрессия,
  - случайный лес,
- модели классификации:
  - классификация опорных векторов,
  - дерево решений,
  - случайный лес,
  - метод ближайших соседей.

Модели регрессии сработали плохо, на тестовой вероятность корректной работы составляет не более 20%. Модели классификации дали более хорошие результаты. Лучше всего справился алгоритм случайный лес, оценка - 60% тестовой выборке

Для улучшения качества решение рекомендуется увеличить объем данных, возможно, увеличив рассматриваемые параметры.