Отчет по задаче о качестве вин с Kaggle

---

1 Введение

1.1 Описание задачи

Вино - это алкогольный напиток, приготовленный из броженного винограда. Дрожжи потребляют сахар в винограде и превращают его в этанол, углекислый газ и тепло. Это приятный на вкус алкогольный напиток, любимый многими. Безусловно, будет интересно проанализировать физико-химические свойства вина и понять их влияние на качество и тип вина.

Задачами исследования являются:

- Проанализировать влияние параметров вина на его качество и тип.
- Предсказать качество каждого образца вина.

1.2 Описание набора данных

Набор данных содержит красные и белые вина 'Vinho Verde'. Vinho verde - уникальный продукт из региона Минью в Португалии. Этот набор данных является общедоступным только для исследовательских целей, для получения дополнительной информации читайте Cortez et al., 2009. Из-за проблем конфиденциальности и логистики доступны только физико-химические (входные) и сенсорные (выходные) переменные (например, нет данных о типах винограда, марке вина, цене продажи вина и тд).

DATASET

- *Name:* **Red Wine Quality Data Set**
- *Source:* UCI Machine Learning Repository
- *Input variables:*
  1 - fixed acidity
  2 - volatile acidity
  3 - citric acid
  4 - residual sugar
  5 - chlorides
  6 - free sulfur dioxide
  7 - total sulfur dioxide
  8 - density
  9 - pH
  10 - sulphates
  11 - alcohol
- *Output variable:* quality (score between 0 and 10)
- *Data Set Characteristics:* Multivariate
- *Number of Observations:* 1599
- *Number of Attributes/Variables:* 12
- *Missing Values:* N/A

Source: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

2 Исследование

2.1 Подготовка данных для исследования.

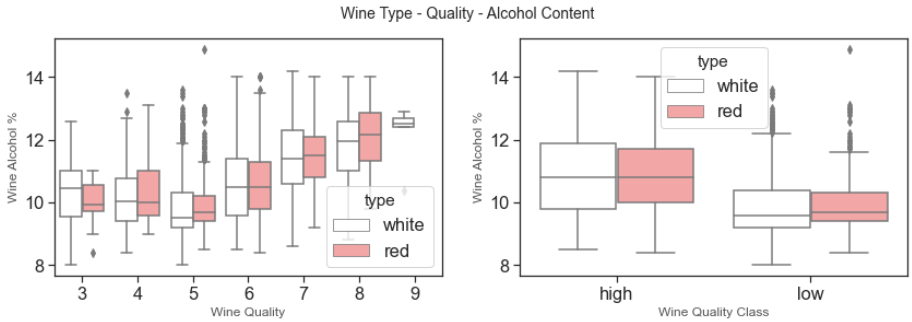
Подготовка данных включала в себя переопределение метки качества вина. Учитывая распределение значений переменной качества вина (порядка 80% значений метрики качества сосредоточены в центре шкалы (5-6 из 10)), было принято решение переопределить меру качества и сделать ее бинарной, где оценки больше 5 считаются высоким качеством, а остальные - низким.

Также, используя преобразование Бокса-Кокса, приводим все числовые переменные к нормальному распределению. И напоследок удаляем из рассмотрения переменные, которые наиболее скоррелированы с остальными (в нашем случае - это residual sugar и total_sulfur_dioxide).

2.2 Анализ влияния параметров вина на его качество и тип

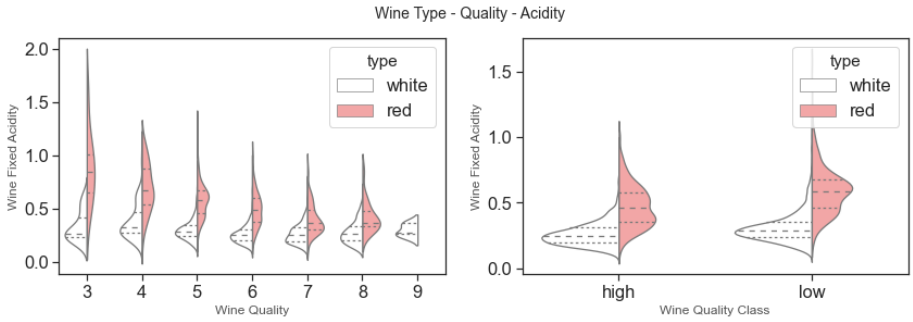
Рассмотрим как некоторые основные параметры вина связаны с его типом и качественной оценкой.

Содержание спирта

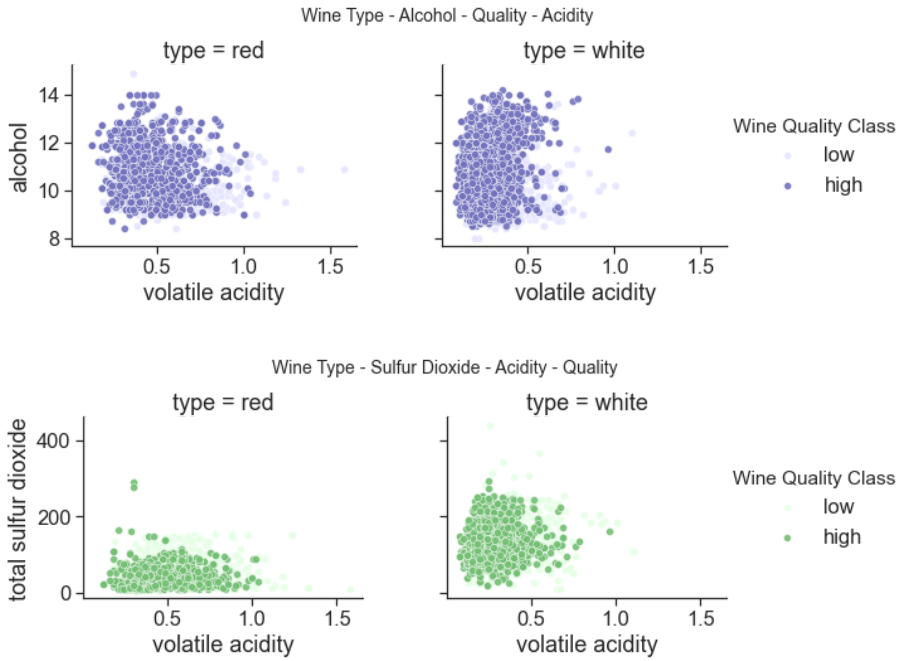


Мы можем четко наблюдать, что спирт по объемному распределению имеет отчетливую тенденцию к увеличению для более качественных образцов вина. В среднем зависимости между типом вина и количеством спирта не наблюдается.

Кислотность и диоксид серы

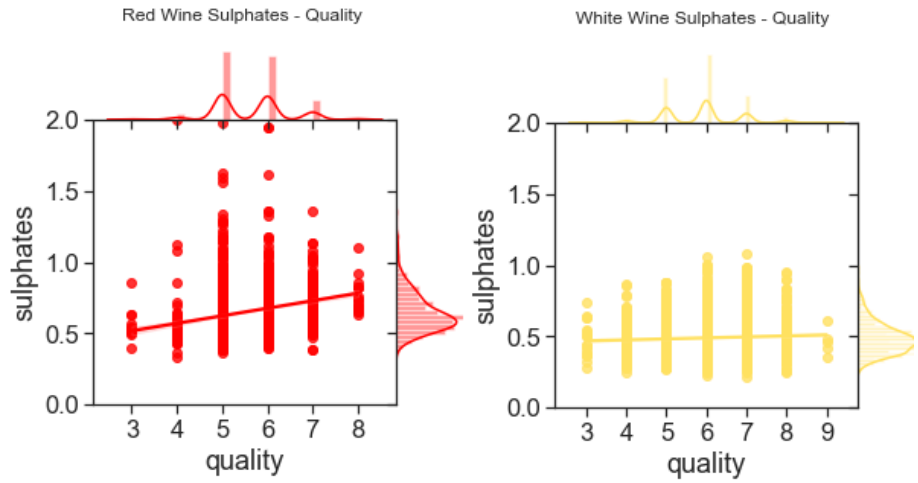


Видно, что образцы красного вина имеют более высокую кислотность по сравнению с аналогами из белого вина. Также мы можем наблюдать общее снижение кислотности с более высоким качеством для образцов красного вина.



Летучая кислотность, а также общий диоксид серы значительно ниже в высококачественных образцах вина. Помимо этого, мы также видим, что уровни летучей кислотности в образцах белого вина несколько ниже, чем в образцах красного вина. Кроме того, общий объем диоксида серы значительно больше в образцах белого вина по сравнению с образцами красного вина.

Содержание сульфатов



Несмотря на то, что, по-видимому, существует некоторая тенденция к более высокому уровню сульфатов для образцов вина более высокого качества, корреляция довольно слабая. Тем не менее, мы видим, что эта склонность вызвана более высокой концентрацией при среднем качестве, и определенно, что уровни сульфатов для красного вина намного выше, чем в белом.

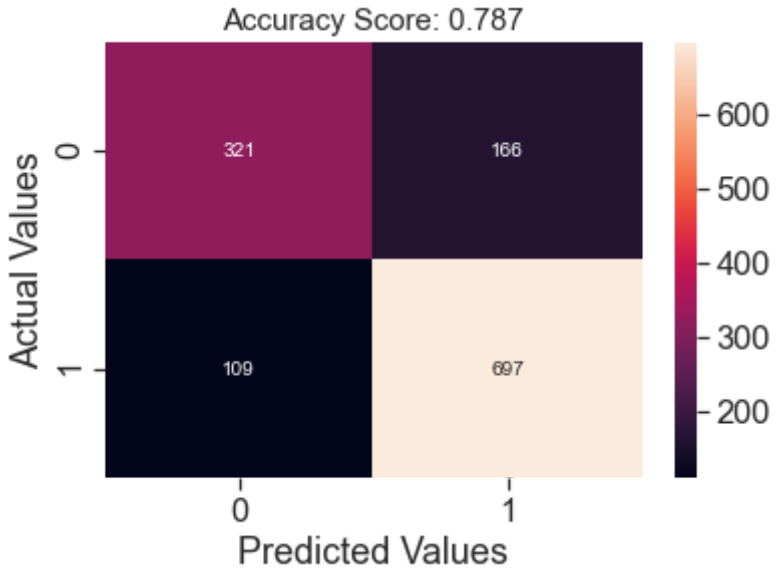
2.3 Предсказание качества вина

Мы рассмотрели три классификатора на наших данных : Логистическая регрессия, к ближайших соседей и дерево решений. Все три модели на валидации дают достаточно высокую точность.

MODEL	ACCURACY	F-SCORE
Логистическая регрессия	0.71	0.77
К-ближайших соседей	0.79	0.83
Дерево решений	0.74	0.80

На валидационной выборке метрики качества предсказания лучше у модели "К-ближайших соседей", поэтому мы делаем выбор в ее пользу для предсказания.

На следующем графике мы видим распределение правильных и неправильных предсказаний модели:



Видно, что наиболее успешно модель предсказывает вина с хорошим качеством. По видимому это обусловлено несбалансированными данными - вин с качеством >5 существенно больше по количеству в нашем датасете.

3 Выводы

В данном исследовании мы рассмотрели как влияют химические параметры вина на его качество.

Мы построили модель, которая с достаточно высоким качеством предсказывает хорошее вино будет или плохое, исходя из заданного набора химических показателей.

Вариантом продолжения исследования является деление шкалы качества вина на более чем два значения. Сложность заключается в том, как правильно выбрать границы новых классов качества, так как в текущем массиве данных значения параметра качества распределены сильно неравномерно.