Исследование качества португальских вин

1. Введение

Португальские вина обладают не плохим вкусом, большим разнообразием. Потребитель ценит данный продукт. Цена за единицу товара ниже, чем у французских или испанских вин такого же качества. Для российского торгового импортера португальских вин требуется дополнительный анализ, чтобы закрепить свои позиции на рынке России. В данном аналитическом отчете выявлены ключевые характеристики, предъявляемые к производителям португальских винам для повышения качества выпускаемой продукции на потребительский рынок.

2. Описание проблемы и набор данных

2.1. Физико-химические показатели португальских вин.

В распоряжении есть объединенный набор данных по белым и красным португальским винам различного качества. Главными критериями отбора качественной продукции являлись физико-химические характеристики вин. В исходных данных было представлено 6497 различных проб марок вин. Среди них большинство составили белые вина - 4898 проб. Общее количество входных переменных (на основе физико-химических тестов) - 11.

Качество (выходная переменная) оценивалось по десятибалльной шкале. Характерные физико-химические показатели проб марок вин , по которым проводилась аналитика, представлены в таблице №1

Основные физико-химические характеристики португальских вин датасета.

Таблица №1

type	volatile acidity	chlorides	density	alcohol	quality
white	0.270	0.045	1.00100	8.8	6
white	0.260	0.030	0.99026	12.6	8
white	0.150	0.044	0.99166	10.2	5
white	0.380	0.061	0.99760	9.4	5
white	0.380	0.051	0.99097	12.7	6
red	0.630	0.077	0.99740	9.4	5
red	0.420	0.084	0.99880	8.7	6
red	0.380	0.097	0.99620	11.4	6
red	0.365	0.088	0.99660	11.0	5
red	0.310	0.067	0.99549	11.0	6

2.2 Качество продукции - целевой признак.

Целевым признаком является качество вин.

Зависимость распределения алкоголя для различного качества белых и красных вин представлена на рисунке 1.

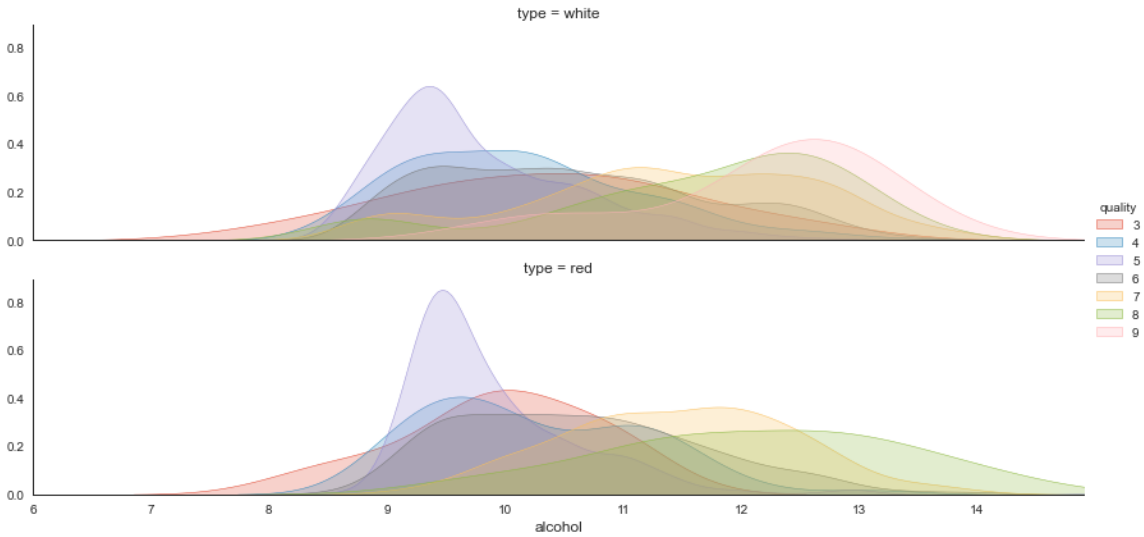

Рис.1

Зависимость распределения алкоголя для различного качества белых и красных вин представлена на рисунке 2.

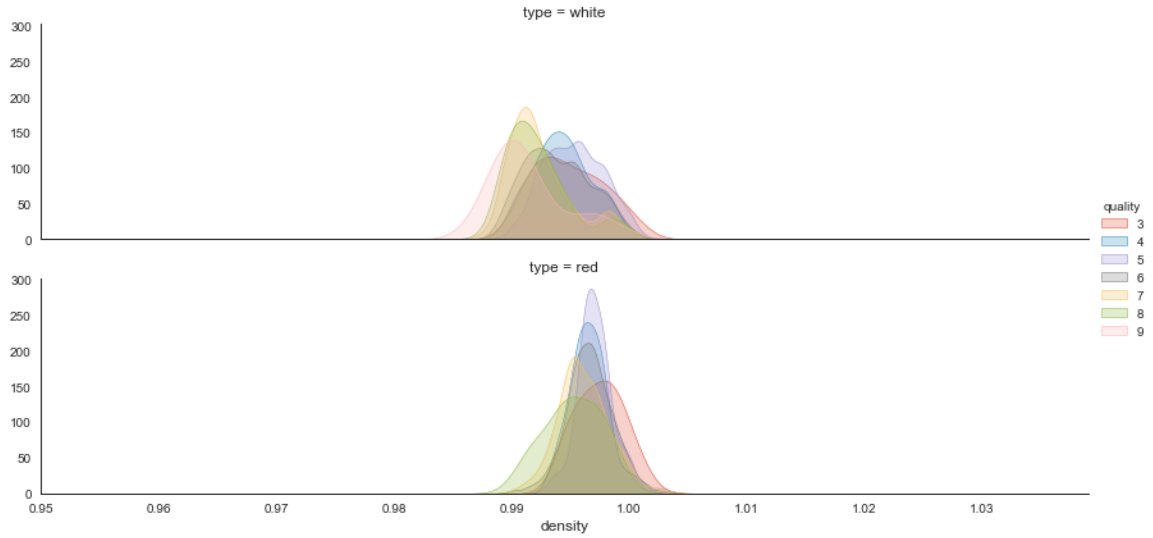

Рис.2

3. Предобработка данных

Перед моделированием был предварительно обработан датасет. Для первичной итерации при знакомстве с датасетом стандартным образом были удалены дубликаты, заполнены пропущенные значения в переменных.

Также перед тем, как выбрать модель, было выявлено, какие признаки являются важными для целевого признака. Для этого с помощью метода OneHotEncoding были дополнительно переведены категориальные переменные в новые переменные числового формата.

Зависимость важности атрибутов представлена на рисунке 3.

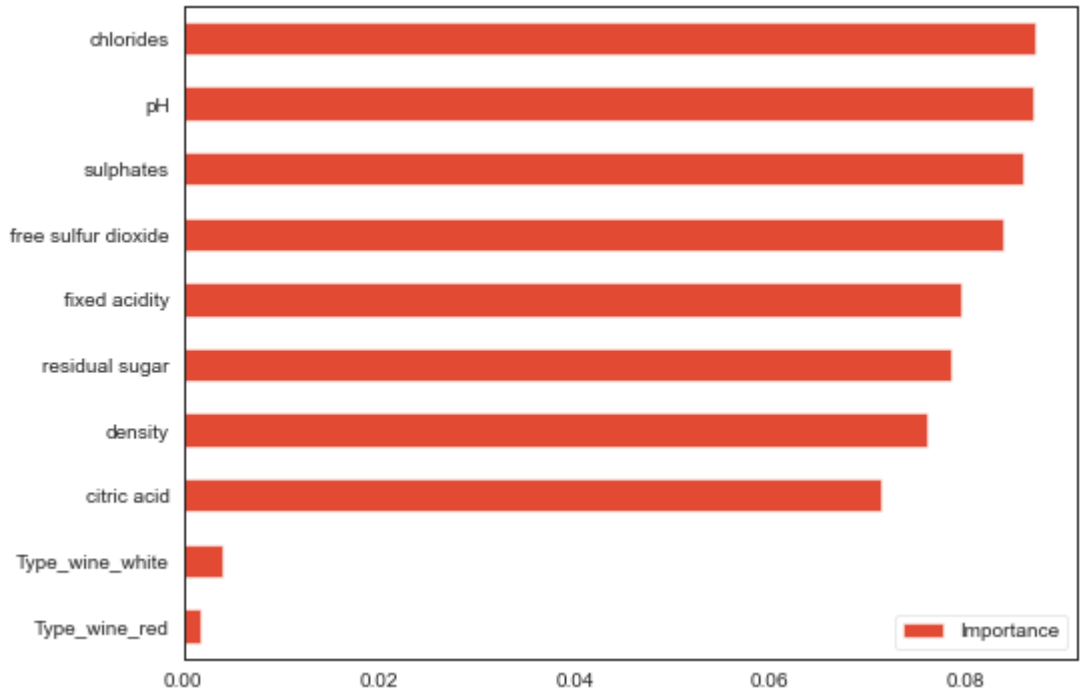

Рис.3

4. Обучение модели и анализ результата.

Целью моделирования является нахождение показателей, которые характеризует наше решение по нахождению важных признаков, как оптимальное для прогноза качества вин.

4.1 Выбор модели и обучение

В качестве первой модели в нашем первичном исследовании был выбран Случайный Лес с показателем n_estimators = 100.

Данные были нормализованы и разбиты на тестовые (30%) и тренировочные данные, проведено обучение.

4.2 Общий показатель моделирования

В результате обучения получено значение общего показателя 51,28%.

5. Обсуждение результатов моделирования

Общий показатель оказался не большим, для выбранной модели Случайного Леса.

Требуется дальнейшая работа с данными, чтобы оптимизировать (повысить значение) общий показатель.

6. Дальнейшие действия (предложения)

Следует повторно провести исследования, для этого по возможности обогатить датасет свежими данными, а также подробней провести предобработку данных.

Предварительно, основным влияющим фактором вне зависимости от типа вина оказался алкоголь. Также на качество влияет и плотность.