Условия успешной реализации потругальских вин.

1. Введение

Португальские вина обладают высоким качеством, большим разнообразием - от легких белых слегка игристых до серьезных красных, которые побеждают на мировых дегустациях. Потребитель высоко ценит данный продукт и выберет неоднократно марку вина хорошего качества. И, главное - цена за единицу товара гораздо выгоднее, чем у французских или испанских вин такого же качества.

Следовательно рынку португальских вин требуется дополнительный анализ, чтобы закрепить свои позиции.

В данном аналитическом отчете выявлены требования, предъявляемые к португальским винам для повышения качества выпускаемой продукции на потребительский рынок.

2. Описание датасета

2.1 Физико-химические показатели португальских вин.

В распоряжении оказался 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 представлены в таблице 1.

Таблица 1. Типовые Физико-химические характеристики португальских вин объединенного датасета.

type	fixed acidity	citric acid	residual sugar	density	alcohol	quality
white	6.20	0.28	2.20	$99.17 \cdot 10^{-2}$	10.50	6
white	5.70	0.25	1.10	$99.10 \cdot 10^{-2}$	11.10	6
white	7.30	0.25	6.65	$99.24 \cdot 10^{-2}$	11.10	5

red	7.50	0.11	1.50	$99.68 \cdot 10^{-2}$	9.6	5
red	8.40	0.19	2.20	$99.74 \cdot 10^{-2}$	9.2	4
red	10.3	0.42	2.00	$99.82 \cdot 10^{-2}$	11.5	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.

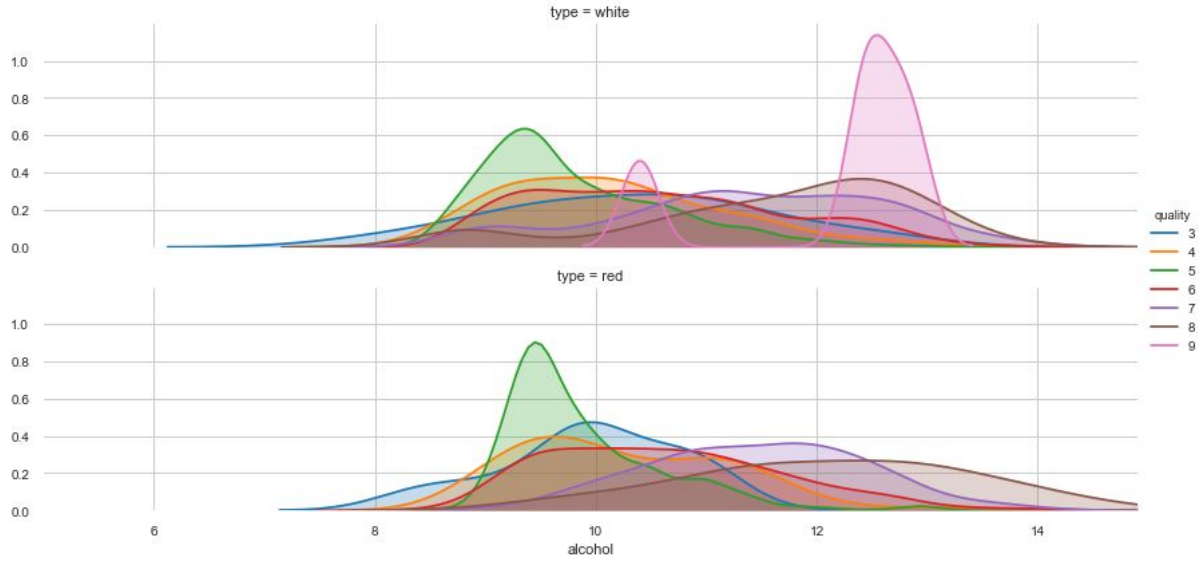

Рисунок 1. Качество продукции и содержание алкоголя по типам португальских вин.

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

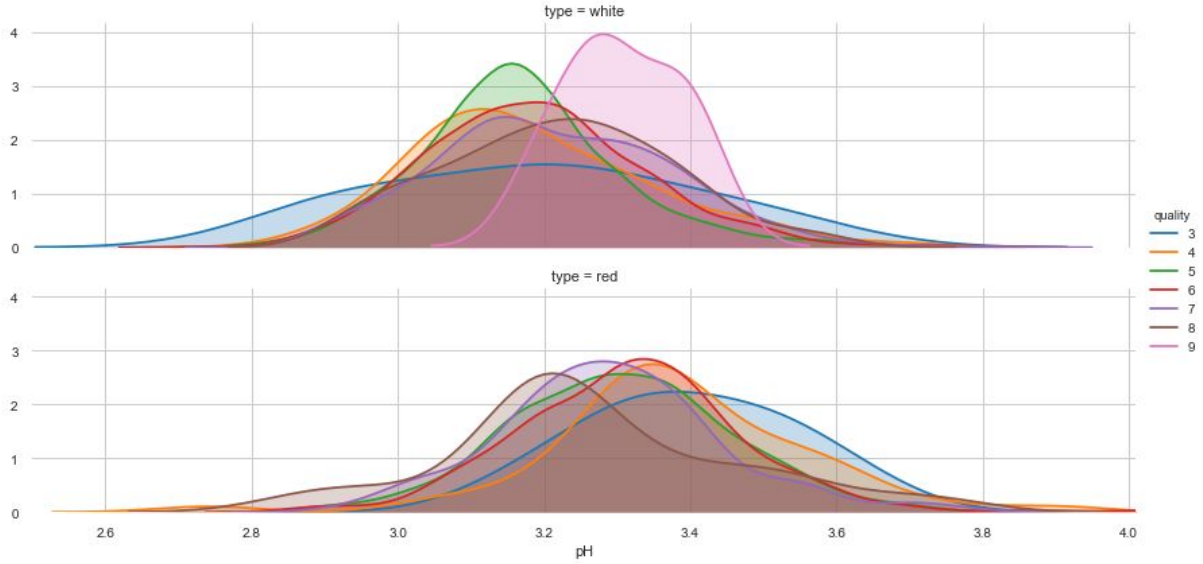

Рисунок 2. Качество продукции и водородный показатель по типам португальских вин.

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 были удалены явные выбросы по 25% и 75% - квартили.

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.

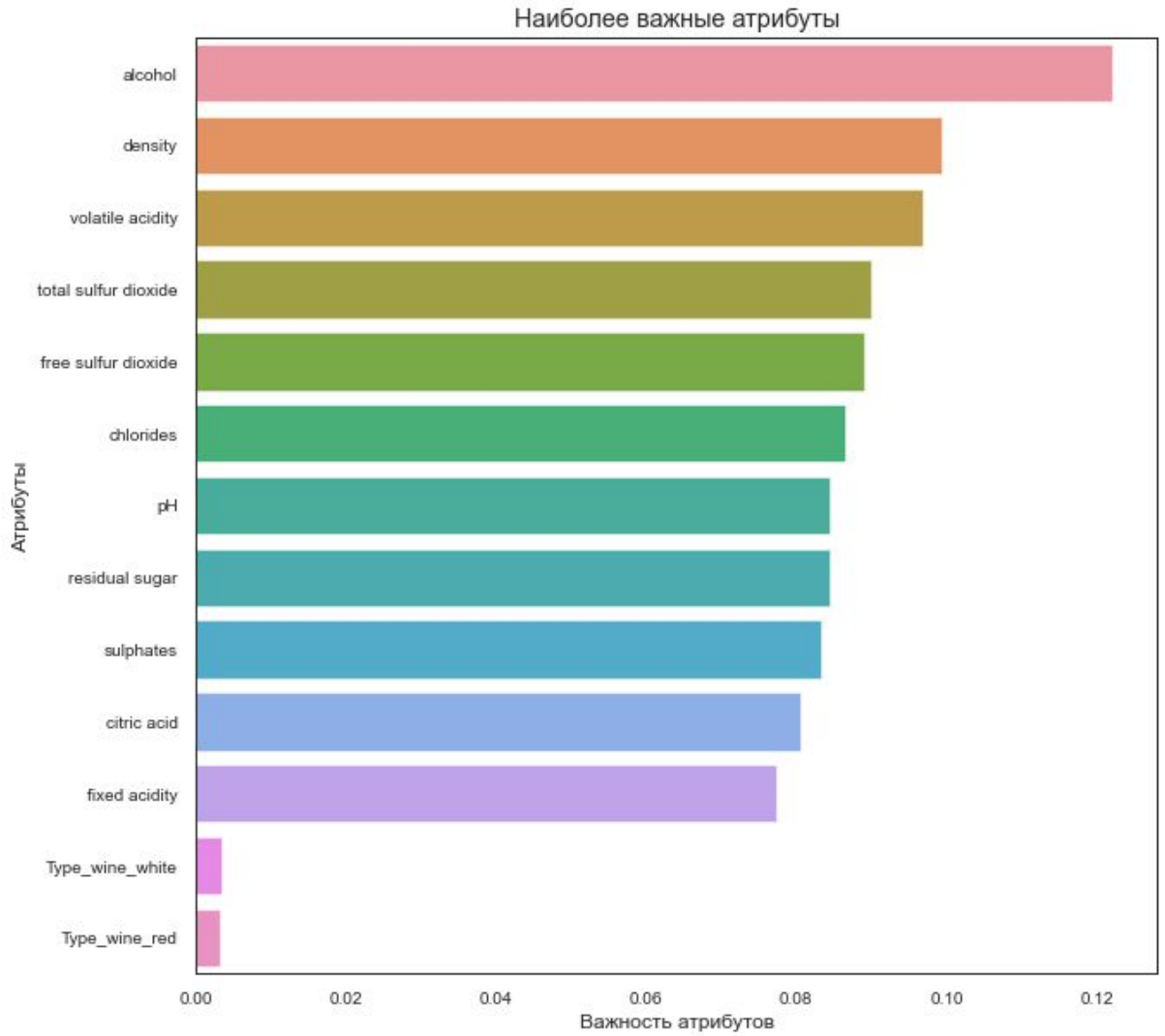

Рисунок 3. Важность атрибутов и алкоголь.

4. Обучение модели и анализ результата.

Целью моделирования является нахождение показателя, который характеризует наше решение по нахождению важных признаков, как оптимальное.

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 39.27%.

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.