**Домашнее задание «Разработка отчетов по исследованию»**

**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**

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

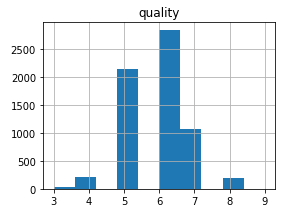
переменная) оценивалось по десятибалльной шкале.

Характерные физико-химические показатели проб марок вин представлены в таблице:

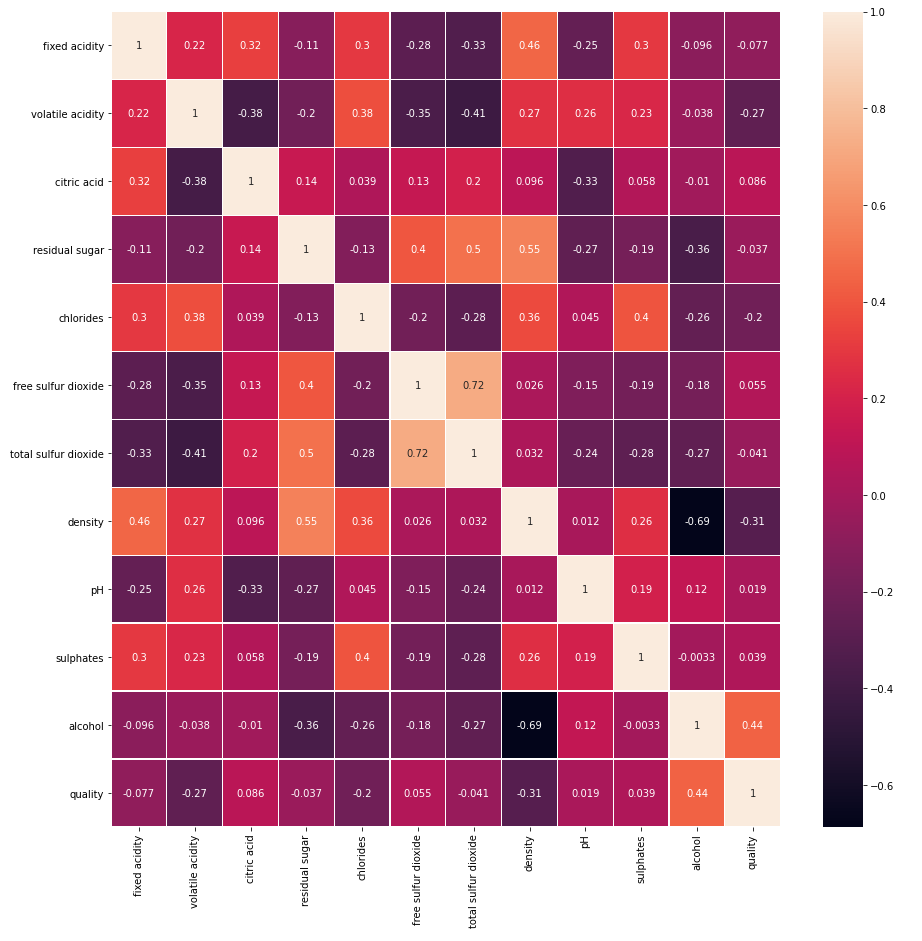
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **type** | **fixed acidity** | **volatile acidity** | **citric acid** | **chlorides** | **free sulfur dioxide** | **total sulfur dioxide** | **density** | **sulphates** | **alcohol** | **quality** |
| white | 7.0 | 0.27 | 0.36 | 0.045 | 45.0 | 170.0 | 1.0010 | 0.45 | 8.8 | 6 |
| white | 6.3 | 0.30 | 0.34 | 0.049 | 14.0 | 132.0 | 0.9940 | 0.49 | 9.5 | 6 |
| white | 8.1 | 0.28 | 0.40 | 0.050 | 30.0 | 97.0 | 0.9951 | 0.44 | 10.1 | 6 |
| white | 7.2 | 0.23 | 0.32 | 0.058 | 47.0 | 186.0 | 0.9956 | 0.40 | 9.9 | 6 |
| white | 7.2 | 0.23 | 0.32 | 0.058 | 47.0 | 186.0 | 0.9956 | 0.40 | 9.9 | 6 |

Качество продукции - целевой признак.

Гистограмма распределения качества:



Тепловая карта корреляции признаков:



**Описание проведенной работы**

Перед моделированием был предварительно обработан датасет. Целью моделирования является нахождение показателя, который характеризует наше

решение по нахождению важных признаков, как оптимальное.

В качестве моделей в нашем первичном исследовании был выбран случайный лес с показателем n\_estimators = 50 и логистическая регрессия.

Данные были нормализованы и разбиты на тестовые (20%) и тренировочные данные,

проведено обучение.

Показатель качества моделей оказался небольшим, для выбранной модели Случайного Леса и логистической регрессии 0.53

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

общий показатель.

**Предложения по дальнейшим действиям**

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.