Оценка качества вина по его физико-химическим свойствам

Введение

Винодельческая отрасль сейчас активно развивается, спрос на вино увеличивается год от года. Стоимость бутылки вина зависит от довольно абстрактной концепции оценки дегустаторами, мнение которых может иметь высокую степень изменчивости. Другим ключевым фактором в сертификации вина являются физико-химические тесты, которые проводятся на лабораторной основе и учитывают такие факторы, как кислотность, содержание спирта, наличие сахара и другие свойства.

В данном отчете представлены результаты исследования набора данных Kaggle, содержащего свойства сортов красного и белого вина. Были применены методы машинного обучения для предсказания качественной оценки вина на основании его физико-химических показателей.

Описание данных

Проводилось исследование набора данных, содержащего информацию о сортах красного и белого вина. Все вина были произведены в определенном районе Португалии. Имеются данные о 12 различных свойствах вин, одним из которых является качество, основанное на сенсорных данных дегустаторов, а остальными - химические свойства, такие как плотность, кислотность, содержание алкоголя и т. д. Каждый сорт вина дегустировался тремя независимыми дегустаторами, и окончательной оценкой выставлялась средняя оценка, данная дегустаторами.

Были рассмотрены следующие физико-химические свойства вин:

1. Фиксированная кислотность
2. Летучая кислотность
3. Лимонная кислота
4. Остаточный сахар
5. Хлориды
6. Свободный диоксид серы
7. Общий диоксид серы
8. Плотность
9. Водородный показатель
10. Сульфаты
11. Содержание спирта

Структура качественных оценок представлена на рисунке 1. Оценки распределены в интервале от 3 до 9.

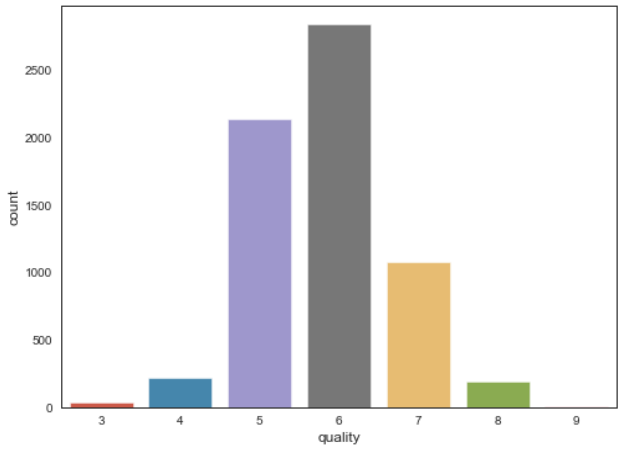

Рисунок 1 - Структура оценки качества вина

Были рассмотрены различные варианты взаимосвязей между качеством и свойствами вина. Например, была замечена корреляция между оценкой дегустаторов и количеством содержанием спирта (Рисунок 2)

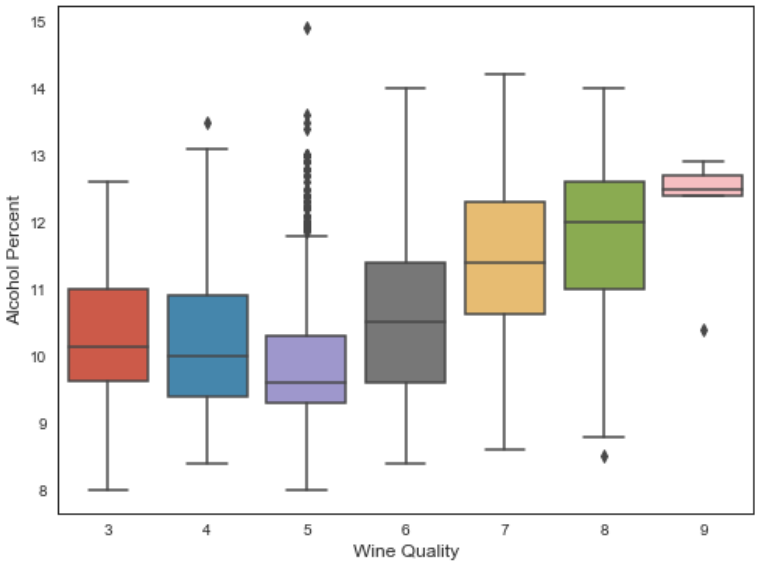

Рисунок 2 - Зависимость оценки качества вина от содержания спирта

Моделирование и оценка модели

Для предсказания оценки качества была выбрана модель логистической регрессии. Модель классифицировала вина из набора данных на «плохое» вино и «хорошее». «Хорошими» считались вина с оценкой качества выше 6.

Мы можем оценить точность модели, используя набор для валидации, где мы знаем фактический результат. Этот набор данных не использовался для обучения, поэтому он абсолютно новый для модели. Точность модели составила 80,37% для данных для обучения и 79,45% для тестовых данных, что свидетельствует о том, что модель не переобучилась.

Выводы

Было проведено аналитическое исследование взаимосвязи физико-химических свойств вина и его вкусового восприятия. Была обучена модель логистической регрессии, предсказывающая, будет ли вино востребованным, по его физико-химическим свойствам. Точность модели составили около 80%.

Использование этого анализа поможет понять, можно ли, изменяя химические и физические показатели вина, повысить качество вина на рынке.