# Предсказание качества вина на основе его физико-химических признаков

## 1. Введение

Оценка качества готового продукта - всегда непростая задача. Она требует проведения дорогостоящих тестов, привлечения высокооплачиваемых специалистов и, конечно, времени. А что если продукт получился плохим? И как понять, что нужно изменить, чтобы сделать его лучше? В этом исследовании мы используем модели машинного обучения, чтобы на примере вина прогнозировать качество продукта ещё до его производства.

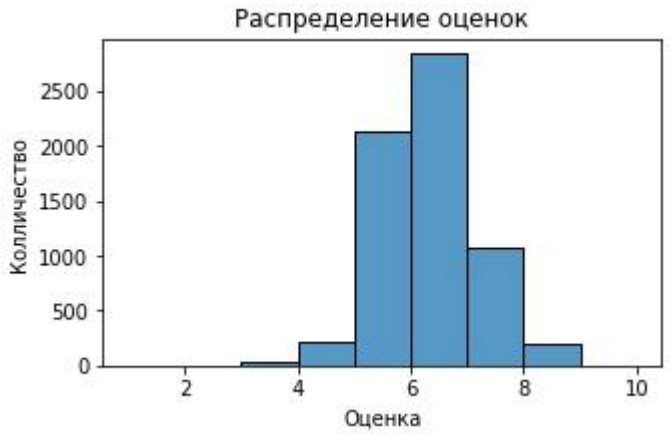
## 2. Данные

Мы используем набор данных с сайта Kaggle, который содержит физические и химические характеристики для 6500 образцов красного и белого португальского вина "Vinho Verde",  а также их качественную оценку по десятибалльной шкале, полученную от различных дегустаторов.

В наборе представлены следующие характеристики:

| **Характеристика** | **Ед. изм.** | **Среднее** | **Минимальное** | **Максимальное** |
|---|---|---|---|---|
| Фиксированная кислотность | г/дм3 | 7.22 | 3.80 | 15.90 |
| Летучая кислотность | г/дм3 | 0.34 | 0.08 | 1.58 |
| Лимонная кислота | г/дм3 | 0.32 | 0.00 | 1.66 |
| Остаточный сахар | г/дм3 | 5.44 | 0.60 | 65.80 |
| Хлориды | г/дм3 | 0.06 | 0.01 | 0.61 |
| Свободный диоксид серы | г/дм3 | 30.53 | 1.00 | 289.00 |
| Общий диоксид серы | г/дм3 | 115.74 | 6.00 | 440.00 |
| Плотность | г/см3 | 0.99 | 0.99 | 01.04 |
| Уровень pH | г/дм3 | 3.22 | 2.72 | 04.01 |
| Сульфаты | г/дм3 | 0.53 | 0.22 | 2.00 |
| Алкоголь | об. % | 10.49 | 8.00 | 14.90 |

Посмотрим на график распределения оценок, поставленных дегустаторами:




На этом графике видно, что дегустаторы чаще оценивали качество вина как среднее (оценки от 5 до 7), и реже как плохое (от 0 до 4) или хорошее (от 8 до 10)

## 3. Исследование

В начале исследования мы подготовили данные для работы моделей машинного обучения:

- Нашли пропуски и заменили их на средние значения - некоторые модели могут работать хуже, если встретят неизвестное значение признака
- Сделали  нормализацию данных - многие модели лучше работают, когда данные имеют один масштаб, например от 0 до 1

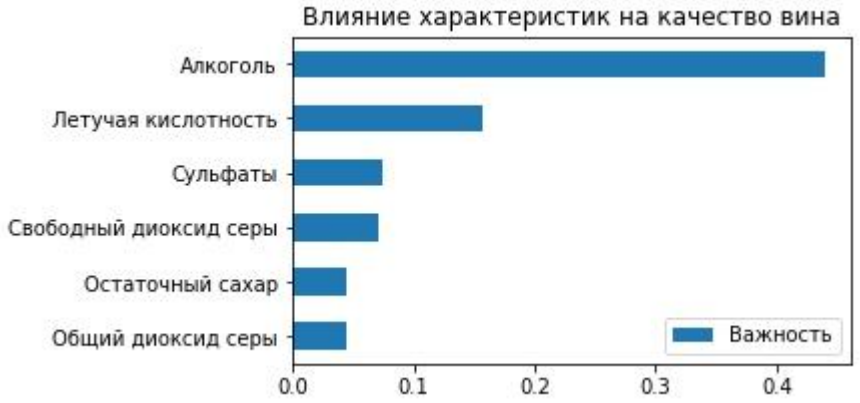
Потом мы разделили готовые данные на набор для обучения (80%) и набор для тестирования (20%). Это необходимо, чтобы избежать переобучения (ситуация, когда модель показывает хорошие результаты только на тех данных, на которых обучалась).

После этого мы обучили несколько моделей (линейная регрессия, случайный лес, градиентный бустинг) и сравнили их качество с помощью метрики RMSE (среднеквадратичная ошибка). Чем ниже значение этой ошибки, тем выше точность модели.

Лучший результат показала модель “случайный лес”. С её помощью можно наиболее точно предсказать качество вина.

## 4. Результаты

Посмотрим на то, какие характеристики вина оказались самыми важными для нашей модели:




Как видно из графика, больше всего на качество вина влияют алкоголь и летучая кислотность.

## 5. Выводы

Используя модели машинного обучения, производители смогут настраивать физические и химические свойства вина, чтобы получить продукт лучшего качества. Также этот способ поможет более точно оценивать уже готовое вино и использовать эти результаты для улучшения системы ценообразования и продвижения продукта.