Аналитический отчёт по исследованию

**«Прогнозирование качества вина по физико-химическим данным»**

Хошина Наталья, студентка курса DS университета Netology

1. **Введение**

Качество вина определяет цену вина. Качество измеряется по всем аспектам, таким как урожай, сорт, тип, место, используемый виноград и / или различные химические анализы. Каждый вид вина имеет свой состав и вкус.

Существует около пяти типов вина: красное вино, белое вино, розовое вино, игристое вино и крепленое вино. Поскольку для каждого типа существуют различные сорта и смеси, оценка качества стала сложной. В этом отчете мы хотим вывести новый метод оценки качества с помощью химического анализа.

Аналитический отчёт проведен на учебном проекте в рамках домашнего задания Netology "Методология ведения DS-проектов", в котором используется несколько методов классификации качества вина по физико-химическим данным.

Конечная цель проекта состоит в том, чтобы иметь возможность прогнозировать качество предоставленного вина на основе данных красного и белого португальского вина "Vinho Verde".

Задача решалась по данным из ресурса Kaggle. Посмотреть более полную информацию по данным, а также сами данные можно по представленной ссылке: [Ссылка на Kaggle](https://www.kaggle.com/datasets/rajyellow46/wine-quality?datasetId=35901&searchQuery=ana).

1. **Описание набора данных**

Набор данных включает в себя 1599 образцов красного вина и 4898 образцов белого вина, каждый образец содержит 12 признаков: фиксированная кислота, летучая кислота, лимонная кислота, остаточный сахар, хлорид, свободный диоксид серы, общий диоксид серы, плотность, рН, сульфаты, спирт и целевую переменную – качество (оценка от 0 до 10).

На рисунке 1 продемонстрированы первые 5 строк из описанного датасета.

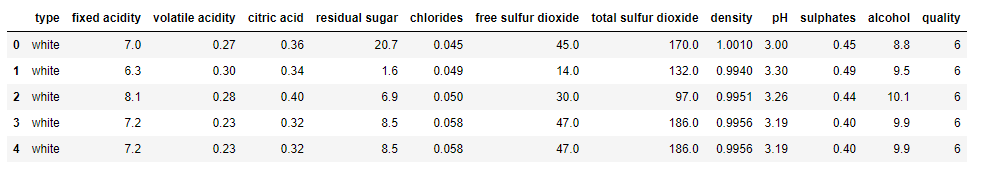


Рис.1. Пример данных из датасета

В таблице 1 представлено признаковое пространство датасета, с которым была проделана дальнейшая работа.

Таблица 1. Датасет

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№** | **Признак** | **Перевод** | **Количество ненулевых значений** | **Тип данных** |
| 1 | type | тип вина | 6497 | object |
| 2 | fixed acidity | фиксированная кислотность | 6487 | float64 |
| 3 | volatile acidity | летучая кислотность | 6489 | float64 |
| 4 | citric acid | лимонная кислота | 6494 | float64 |
| 5 | residual sugar | остаточный сахар | 6495 | float64 |
| 6 | chlorides | хлориды | 6495 | float64 |
| 7 | free sulfur dioxide | свободный диоксид серы | 6497 | float64 |
| 8 | total sulfur dioxide | общий диоксид серы | 6497 | float64 |
| 9 | density | плотность | 6497 | float64 |
| 10 | pH | pH | 6488 | float64 |
| 11 | sulphates | сульфаты | 6493 | float64 |
| 12 | alcohol | спирт | 6497 | float64 |
| 13 | quality (target) | качество (целевая переменная) | 6497 | int64 |

Взаимосвязь между типом и качеством вина, представлена на диаграмме 1.

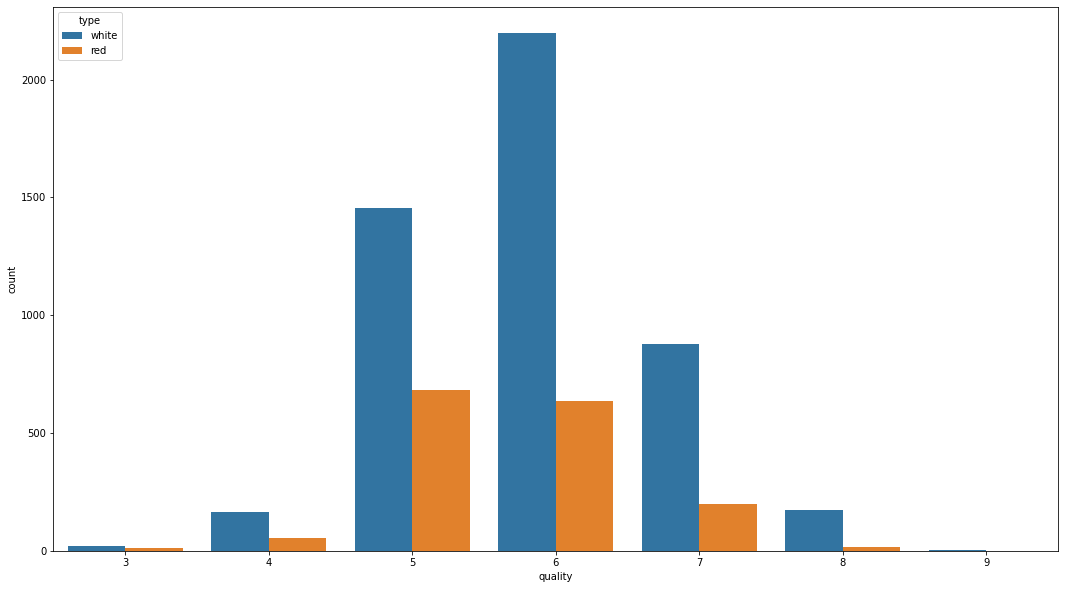


Диаграмма1. Взаимосвязь типа и качества вина.

1. **Предварительная обработка данных**

Для улучшения качества данных, были преобразован категориальный признак *“type”*, заполнены пропущенные значения. Произведен поиск и удаление выбросов.

Была произведена нормализация данных см. рисунок 2.

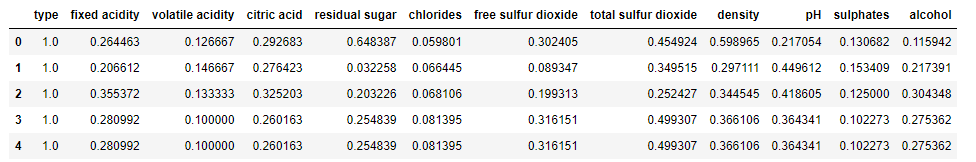


Рис.1. Пример нормализованных данных

Для улучшения качества обучения модели, была произведена балансировка классов для целевой переменной.

1. **Обучение моделей**

При обучении сравнивались 4 модели:

- Logistic Regression

- Decision Tree Classifier

- Random Forest Classifier

- Stacking Classifier

В таблице 2 приведены сравнения моделей и их качества обучения.

Таблица 2. Точность обучения моделей

|  |  |
| --- | --- |
| **Модель** | **Точность** |
| Stacking Classifier | 0.92 |
| Random Forest Classifier | 0.91 |
| Decision Tree Classifier | 0.83 |
| Logistic Regression | 0.48 |

**5. Заключение**

Наилучшую точность обучения показала модель «Stacking Classifier».

Для возможного дальнейшего улучшения качества модели, рекомендуется провести эксперименты с гиперпараметрами. Также увеличив количество данных, можно также повысить качество обучения моделей и, соответственно, их более точного предсказания.

Качественные баллы начисляются за дегустацию вина. Эти баллы не является стандартизированы и будут варьироваться от дегустатора к дегустатору. Следовательно, оценка качества вина на основе сенсорных ощущений дегустатора является субъективной.

Предложенный метод на основе модели машинного обучения, оценивает качество по химическим свойствам, может рассчитать качество вина автоматически. Цена может быть изменена в зависимости от винодельни, если они смогут обучить модель своими собственными наборами данных о качестве и цене.