# **Предварительный отчёт по оценке качества вина на основе физико-химические признаков**

**Цель исследования** - протестировать методы выбора признаков для предсказания качества вина. На предварительном этапе – исследовать данные и наметить дальнейшие шаги по тестированию выбора признаков.

**Исходные данные** - физико-химические характеристики португальского вина «Винью Верде», содержащиеся в файле **winequalityN.csv**.

Входные переменные (на основе физико-химических тестов):

1. type - тип вина (категориальная переменная)
2. fixed acidity - фиксированная кислотность
3. volatile acidity - летучая кислотность
4. citric acid - лимонная кислота
5. residual sugar - остаточный сахар
6. chlorides - хлориды
7. free sulfur dioxide - свободный диоксид серы
8. total sulfur dioxide - общий диоксид серы
9. density - плотность
10. рН – водородный показатель
11. sulphates - сульфаты
12. alcohol - алкоголь
13. quality - качество (оценка от 0 до 10) – выходная переменная

Данные загружены в датафрейм – **df\_raw.** В качестве препроцессинга переменная **type** была переведена в цифровые значения – **df\_type**.

Первичный анализ данных показал, что датасет имеет незначительное количество пропусков:

Data columns (total 13 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 type 6497 non-null int64

1 fixed acidity 6487 non-null float64

2 volatile acidity 6489 non-null float64

3 citric acid 6494 non-null float64

4 residual sugar 6495 non-null float64

5 chlorides 6495 non-null float64

6 free sulfur dioxide 6497 non-null float64

7 total sulfur dioxide 6497 non-null float64

8 density 6497 non-null float64

9 pH 6488 non-null float64

10 sulphates 6493 non-null float64

11 alcohol 6497 non-null float64

12 quality 6497 non-null int64

Пропуски наблюдаются в признаках: **fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, pH, sulphates**. Заполним отсутствующие значения средним – **df\_mean**.

Визуальный анализ показал, что гистограммы параметров до заполнения и после не изменились.

Пример гистограммы для параметра **fixed acidity**

A picture containing text, screenshot, diagram, plot

Description automatically generated

|  | **type** | **fixed acidity** | **volatile acidity** | **citric acid** | **residual sugar** | **chlorides** | **free sulfur dioxide** | **total sulfur dioxide** | **density** | **pH** | **sulphates** | **alcohol** | **quality** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **count** | 6497 | 6487 | 6489 | 6494 | 6495 | 6495 | 6497 | 6497 | 6497 | 6488 | 6493 | 6497 | 6497 |
| **mean** | 0.75 | 7.21 | 0.33 | 0.31 | 5.44 | 0.056 | 30.52 | 115.74 | 0.994 | 3.21 | 0.53 | 10.49 | 5.81 |
| **std** | 0.43 | 1.29 | 0.16 | 0.14 | 4.75 | 0.035 | 17.74 | 56.52 | 0.002 | 0.16 | 0.14 | 1.19 | 0.87 |
| **min** | 0 | 3.8 | 0.08 | 0 | 0.6 | 0.009 | 1 | 6 | 0.987 | 2.72 | 0.22 | 8.00 | 3 |
| **25%** | 1 | 6.4 | 0.23 | 0.25 | 1.8 | 0.038 | 17 | 77 | 0.992 | 3.11 | 0.43 | 9.5 | 5 |
| **50%** | 1 | 7 | 0.29 | 0.31 | 3 | 0.047 | 29 | 118 | 0.994 | 3.21 | 0.51 | 10.3 | 6 |
| **75%** | 1 | 7.7 | 0.4 | 0.39 | 8.1 | 0.065 | 41 | 156 | 0.996 | 3.32 | 0.6 | 11.3 | 6 |
| **max** | 1 | 15.9 | 1.58 | 1.66 | 65.8 | 0.611 | 289 | 440 | 1.038 | 4.01 | 2 | 14.9 | 9 |

Основные статистические характеристики признаков:

Гистограммы по всем признакам:

A picture containing diagram, text, line, plot

Description automatically generated

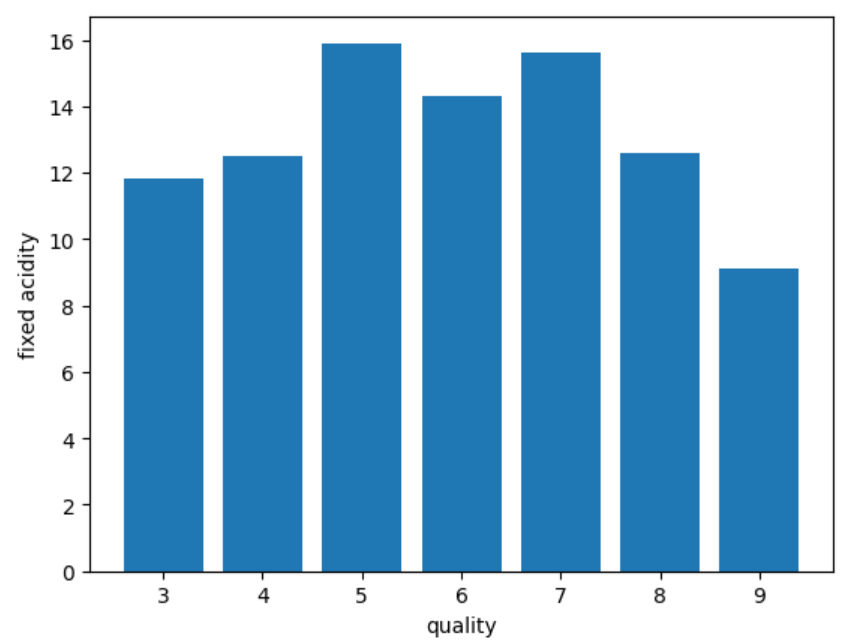
A picture containing diagram, line, plot, plan

Description automatically generated

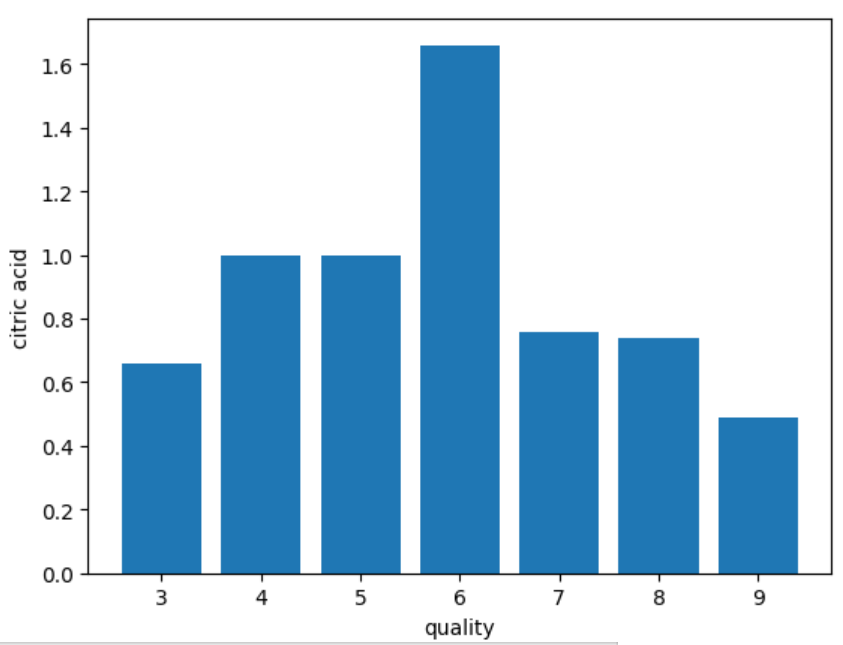
По гистограмме видим, что целевые классы **quality** не сбалансированы:

|  |  |
| --- | --- |
| Количество классов | Количество значений |
| 6 | 2836 |
| 5 | 2138 |
| 7 | 1079 |
| 4 | 216 |
| 8 | 193 |
| 3 | 30 |
| 9 | 5 |

Изучим вопрос, как различные признаки распределяются по оценкам качества вина.

A picture containing text, screenshot, plot, diagram

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, rectangle, line

Description automatically generated

A picture containing text, screenshot, rectangle, plot

Description automatically generatedA picture containing text, screenshot, plot, diagram

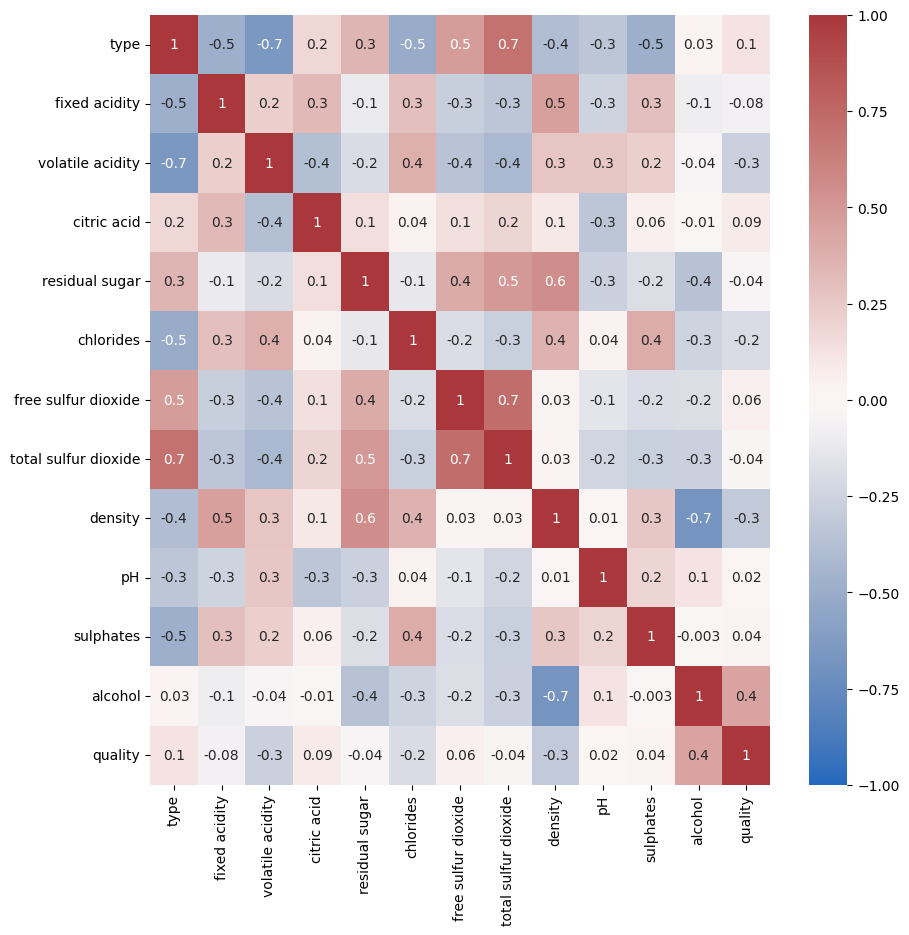
Description automatically generated

A picture containing text, screenshot, rectangle, plot

Description automatically generated

Как видим из гистограмм: признак **density** не дифференцируется по классам **quality**, слабо дифференцируются **pH** и **alcohol**. Признак **density** удалим из датасета чтобы проверить, как его наличие влияет на результат обучения – **df\_del**.

Посмотрим на взаимную корреляцию признаков: матрица не показывает нам признаков с очень сильной корреляцией.



Далее, разделим дфтафрейм на обучающую и тестовую выборки, в качестве целевого параметра Y используем признак **quality**.

Для начального этапа обучения попробуем использовать 2 варианта: обучающая выборка **df\_mean** по всем признакам с заполненными средним отсутствующими значениями и **df\_del** – датафрейм после удаления признака **density**.

Используем test\_size=0.2, random\_state=99

В качестве моделей будем использовать следующие алгоритмы:

"Logistic Regression", LogisticRegression(max\_iter=80000)

'Support Vector Classification', SVC(kernel='rbf')

'Random Forest', RandomForestClassifier()

'Decision Tree', DecisionTreeClassifier()

Аккуратность по **df\_mean:**

Logistic Regression: 0.538201 (0.020290)

Support Vector Classification: 0.433915 (0.019554)

Random Forest: 0.671548 (0.015786)

Decision Tree: 0.588406 (0.024312)

Аккуратность по **df\_del:**

Logistic Regression: 0.487017 (0.018030)

Support Vector Classification: 0.433339 (0.021755)

Random Forest: 0.651533 (0.017166)

Decision Tree: 0.567055 (0.019659)

Результаты обучения на датасете без удаления признака **density** выше. Далее будем использовать датасет **df\_mean.** Сбалансируем классы с помощью модуля SMOTE, в результате работы которого количество значений в каждом классе станет равным 2836. И нормализуем с помощью алгоритма MinMaxScaler.

Аккуратность после всех преобразований выросла:

Logistic Regression: 0.488005 (0.012395)

Support Vector Classification: 0.602733 (0.009945)

Random Forest: 0.890057 (0.006757)

Decision Tree: 0.817391 (0.009156)

В итоге, мы можем выбрать наиболее интересные алгоритмы для дальнейшего изучения вопроса по выбору признаков.

Планируем:

* изучить вопрос, как отскоки признаков могут влиять на результат,
* имеет ли смысл разделить датасет на красное и белое вино и анализировать отдельно друг от друга,
* удалить признаки в высоким значением отрицательной корреляции