# Отчёт по проекту Wine quality

## 1. Бизнес-анализ

### 1.1 Цель

Предсказать категорию качества вина

### 1.2 Описание

В исходном датасете из Kaggle приведены объективные данные по винам (численные и категориальные характеристики: кислотность, сахар, спиртуозность, сульфаты и пр.) и соответствующая субъективная категория качества.

Данные были получены здесь:

<https://www.kaggle.com/datasets/rajyellow46/wine-quality>

Исходные данные содержат характеристики вина:

* type
* fixed acidity
* volatile acidity
* citric acid
* residual sugar
* chlorides
* free sulfur dioxide
* total sulfur dioxide
* density
* pH
* sulphates
* alcohol

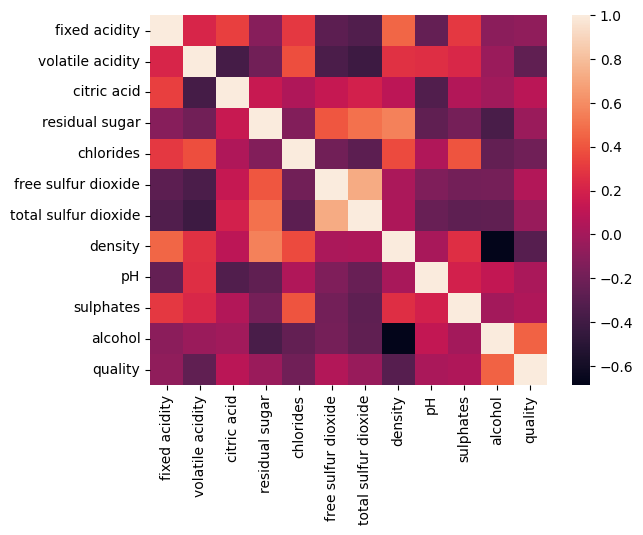
А также соответствующую этим данным субъективную оценку -- целочисленное значение от 3 до 9.

## 2. Анализ данных

Исследования данных показали, что

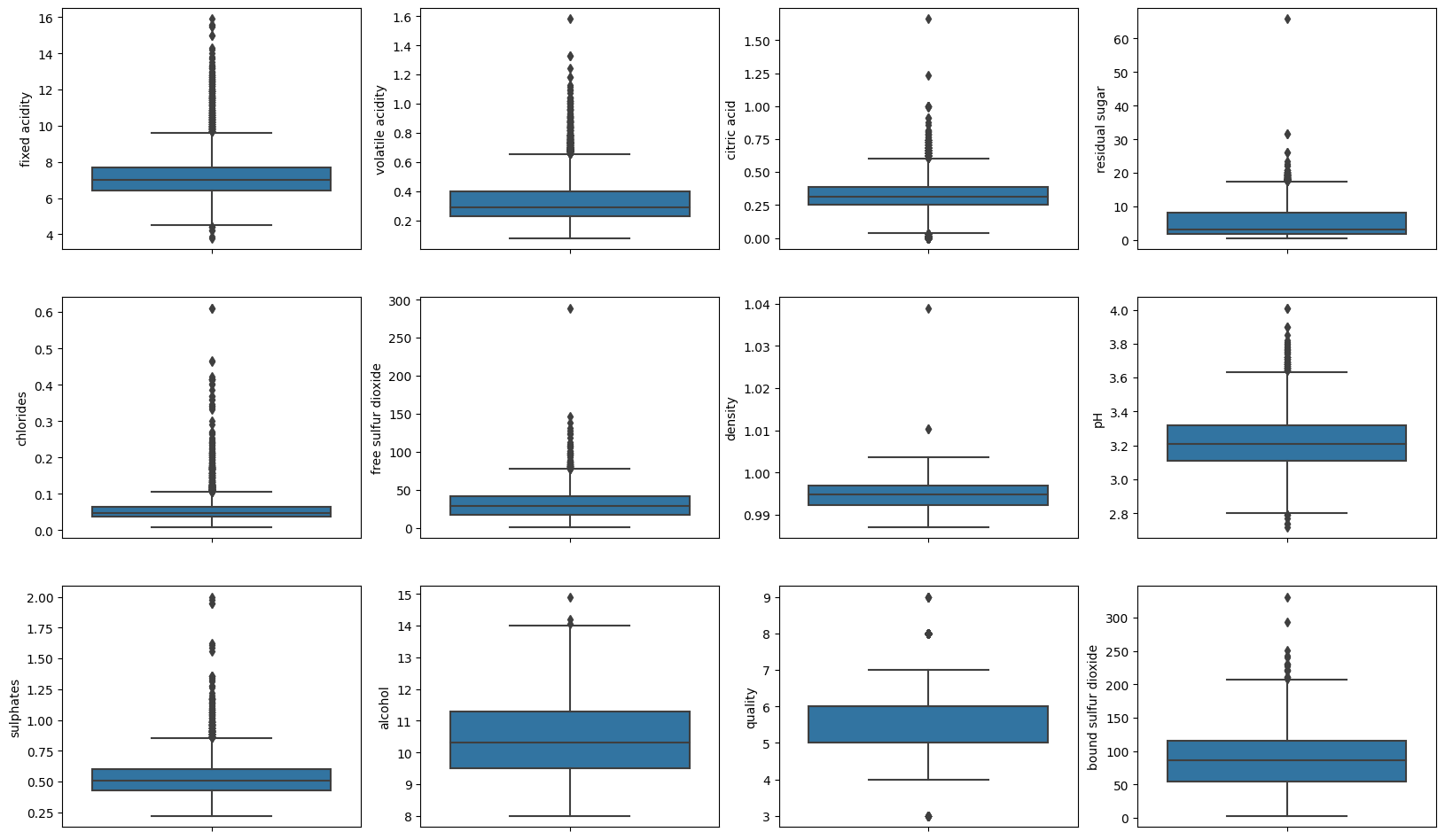
данные содержат небольшое количество пропущенных значений.

есть достаточно большие корреляции между свободным и общим содержанием диоксида серы, а также плотностью и спиртуозностью:



Было принято решение сделать искусственный параметр “связанный диоксид серы” (bound sulfur dioxide) как разность общего и свободного диоксида серы. Также было решено удалить параметр “плотность” (density).

Анализ выбросов показал небольшое количество выбросов по некоторым параметрам



## 3. Подготовка данных

Помимо удаления density и замены общего содержания диоксида серы на связанное, к категориальному параметру “тип” (type) применено One-hot encoding. Были устранены выбросы и данные с пропусками.

Обнаружен дисбаланс классов вин:

|  |  |
| --- | --- |
| **Класс** | **Количество вхождений** |
| 6 | 2808 |
| 5 | 2122 |
| 7 | 1074 |
| 4 | 213 |
| 8 | 192 |
| 3 | 27 |
| 9 | 5 |

Поэтому необходимо применить SMOTE-оверсэмплинг для балансировки классов

## 4. Моделирование

Использовано два типа моделей: перцептрон

model = Sequential([

L.Dense(13,activation='linear'),

L.Dense(20,activation='selu'),

L.BatchNormalization(),

L.Dropout(0.3),

L.Dense(20,activation='linear'),

L.Dense(output\_length,activation='softmax')

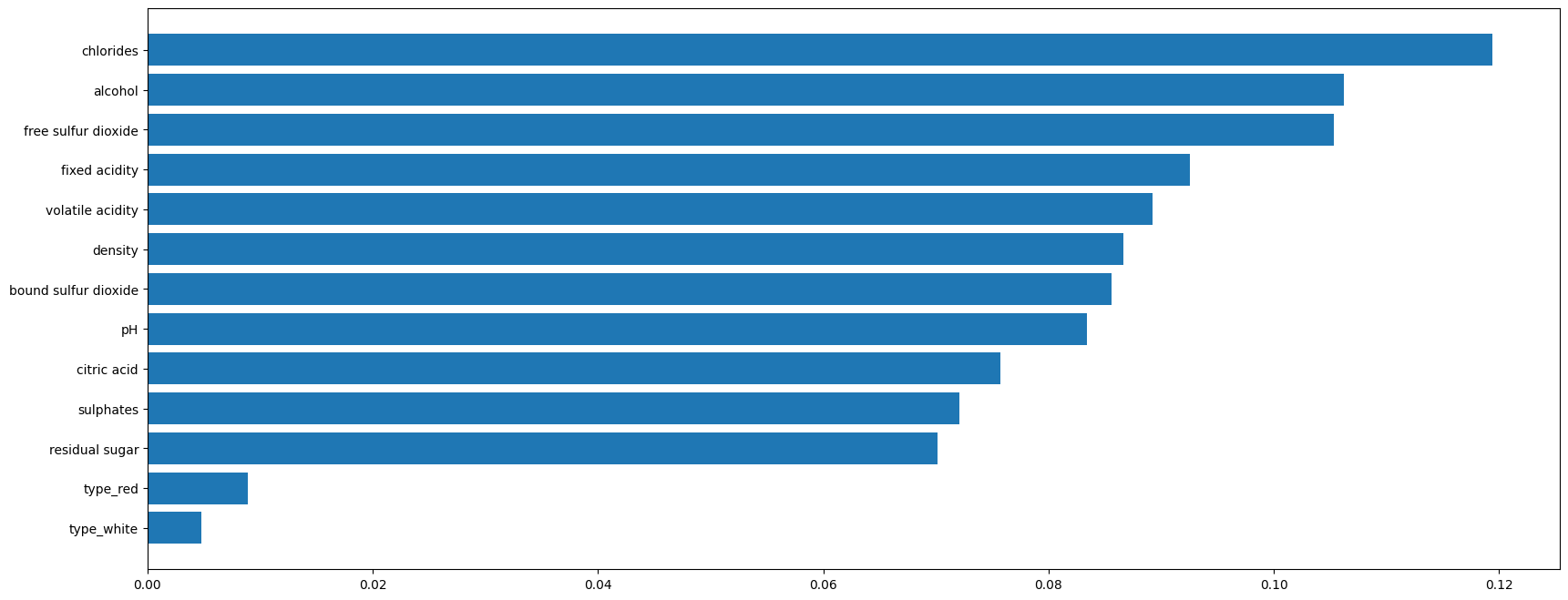
])

и случайный лес.

## 5. Оценка

Перцептрон показал точность не более 0,47, в то время как случайный лес около 0,86. Решено использовать случайный лес.

Наибольшее значение имеет параметр “содержание хлоридов”, затем “алкоголь” и “свободный диоксид серы”



## 6. Внедрение

В результате получена модель случайного леса, сохранённая в формате Pickle

Подготовка данных для прогнозирования будет включать следующие шаги:

1. Создание синтетического признака bound sulfur dioxide как разность total sulfur dioxide и free sulfur dioxide

2. Применить OHE к типу вина (type)

3. Удаление колонки density

4. Удаление пустых данных