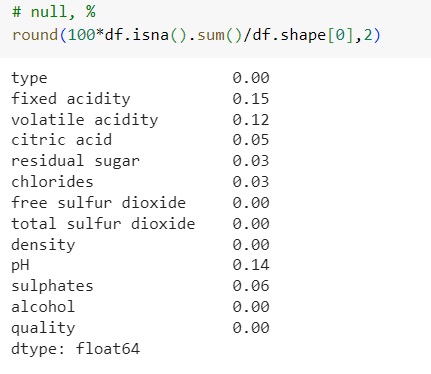
Отчет

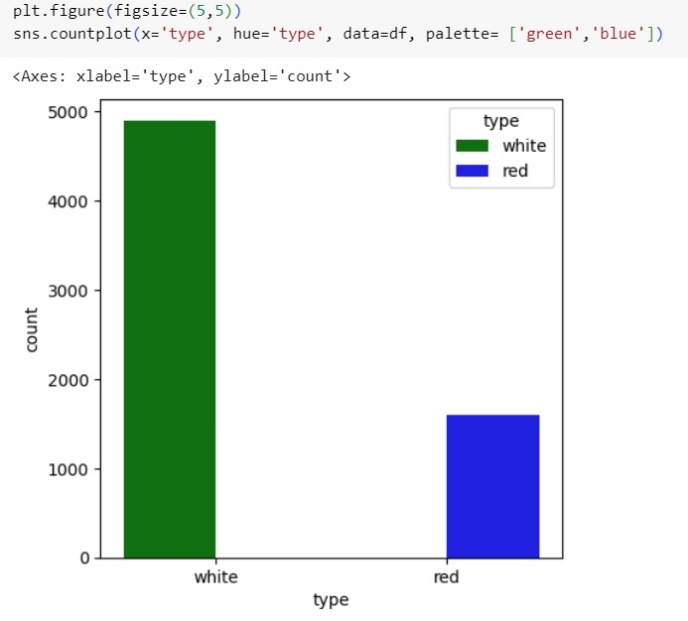
1. Анализ данных.

1.1 Первичная статистика исходных данных.

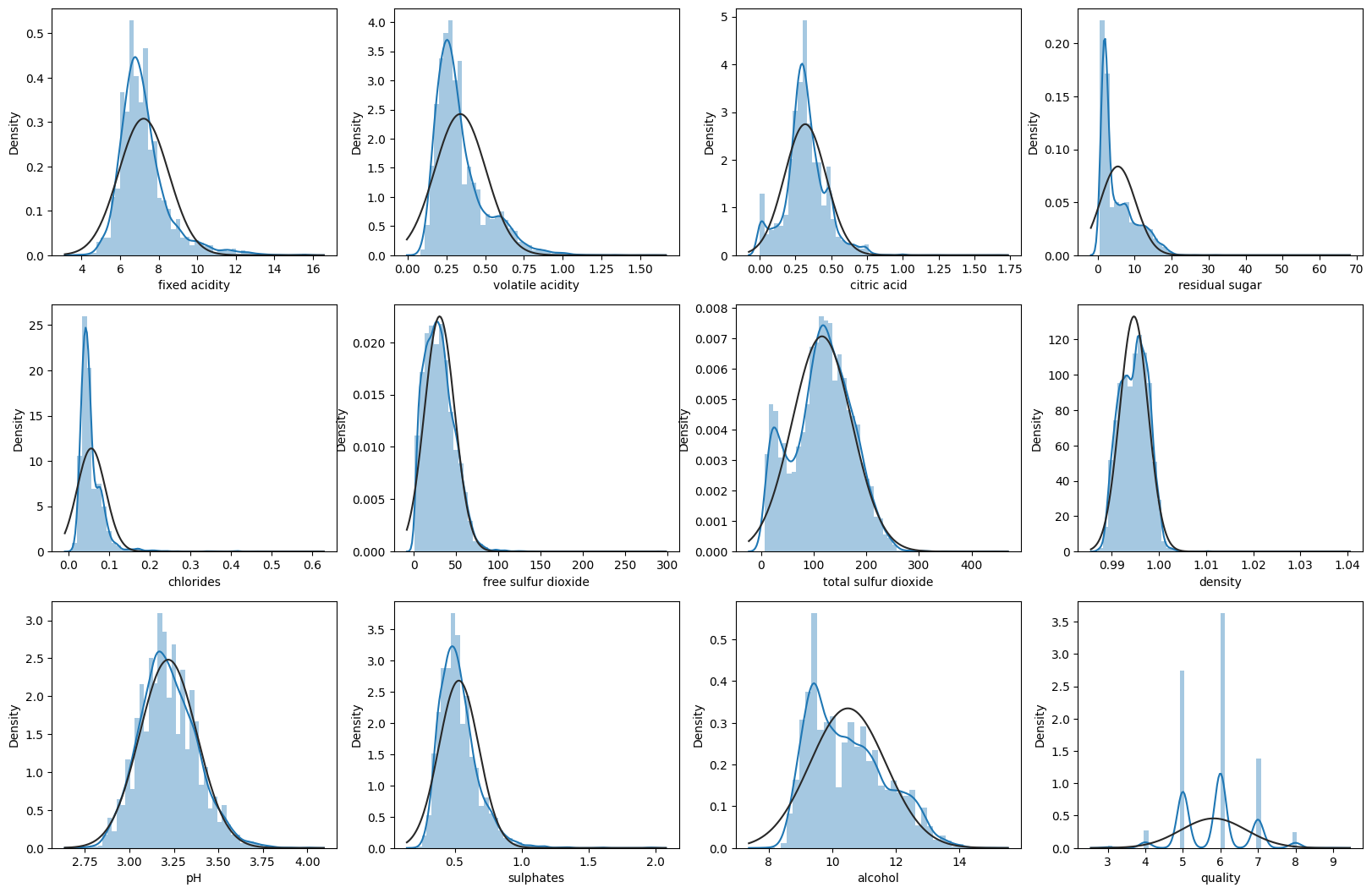
* 1. Процент пропусков

Пропущенные данные имеются в 7 колонках, однако в каждой колонке они составляют менее 1 %.

* 1. Баланс классов

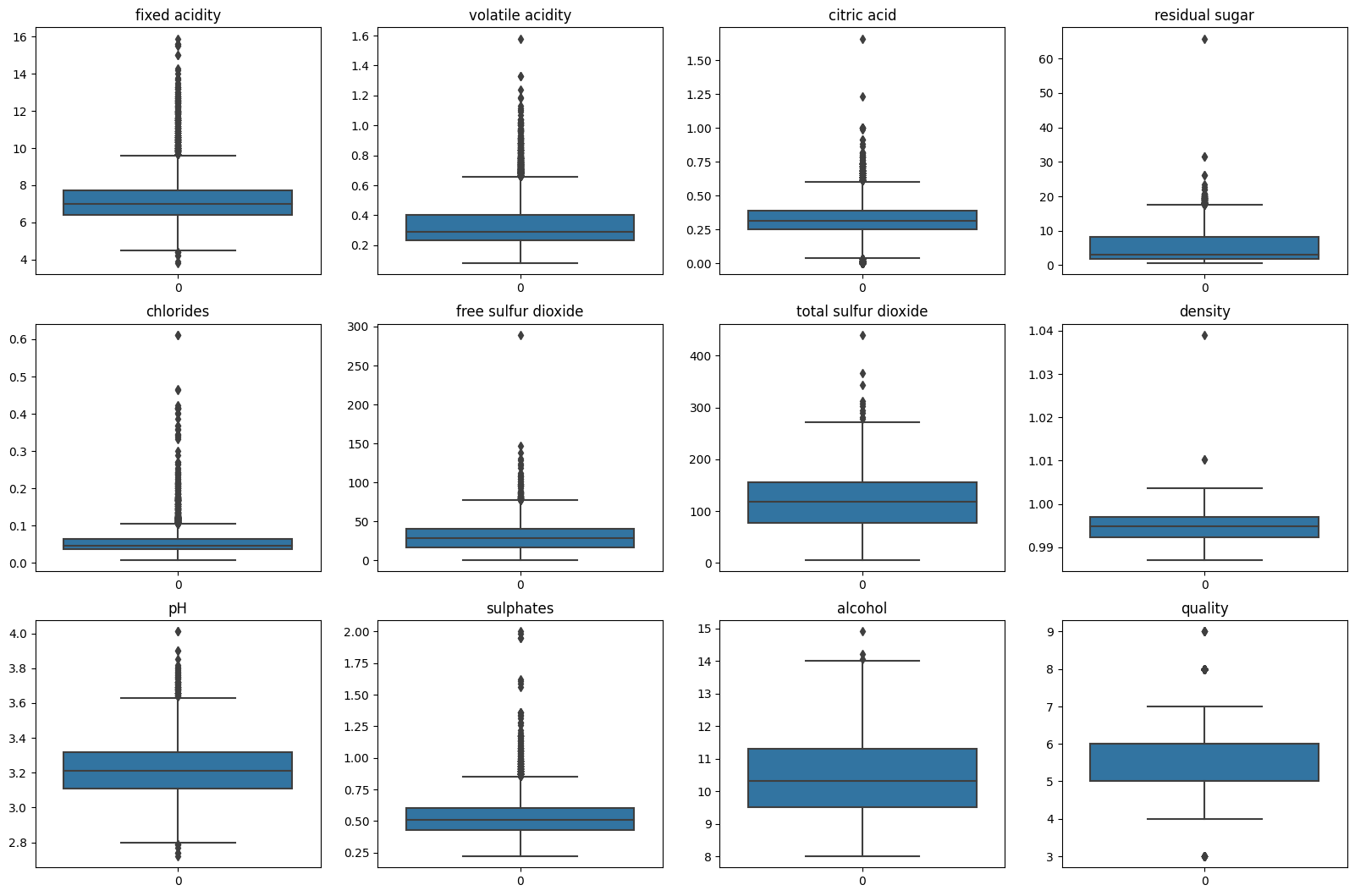


Как мы видим, наблюдается сильный дисбаланс в сторону класса ”white” .

* 1. Распределение данных

В основном данные имеют нормальное распределение со сдвигом влево, за исключением “quality”, что объясняется категориальностью этой переменной. Также можно заметить, что “alcohol” имеет множество пиков и его распределение далеко от нормального, хотя общая тенденция сохраняется.

* 1. Выбросы



Выбросы наблюдаются везде, однако они все лежат в пределах нормы показателей химического анализа. В случае «quality» оценка 9 вполне допустима.

1. Подготовка данных.

В ходе подготовки данных были выполнены следующие изменения:

* Заполнение пропусков средними значениями
* Удаление дубликатов
* Добавление синтетических данных для устранения дисбаланса классов
* Перевод текстовой категориальной переменной в числовую.

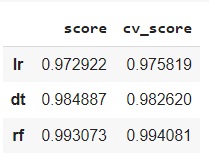
1. Модели

Были реализованы следующие модели:

* RandomForestClassifier
* LogisticRegression
* DecisionTreeClassifier

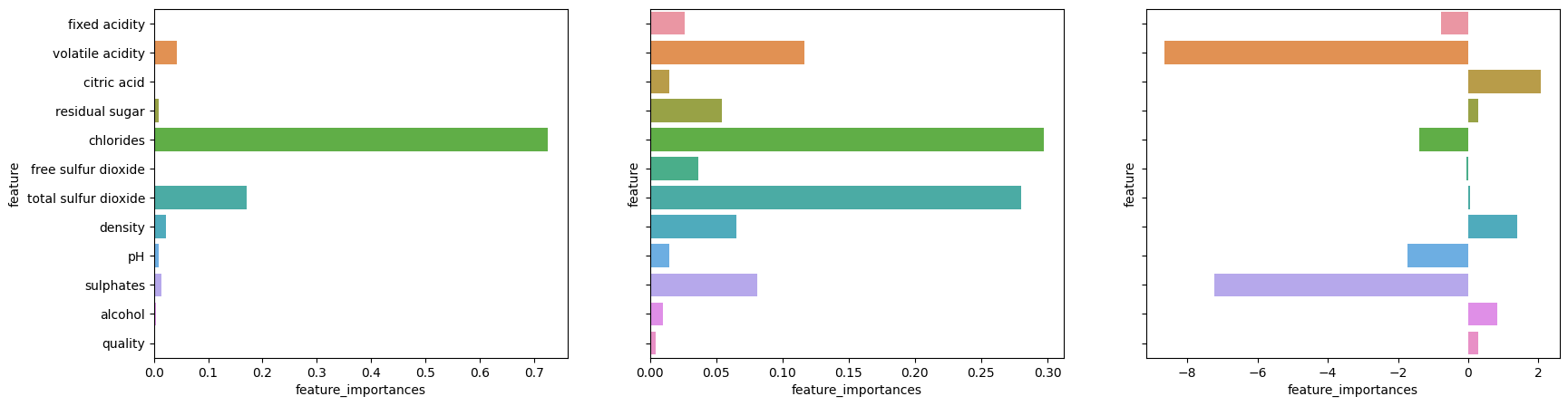
Модели использовались с параметрами по умолчанию. Подбор гиперпараметров отсутствует.

1. Результаты



Все модели показали высокую точность, что на моменте обучения, что на моменте предсказания. Незначительно лучше проявила себя модель RandomForestClassifier.

Важность признаков.

DecisionTreeClassifier RandomForestClassifier LogisticRegression

DecisionTreeClassifie и RandomForestClassifier имеют схожий результат по нахождению главных признаков – самый влияющий признак – «Chlorides». Согласно LogisticRegression чем выше значение «volatile acidity», тем вероятнее, что вино окажется красным.