# Введение

В данной работе будет описано обучение модели для определения качества вина по его физико-химическим характеристикам.

Во второй главе проблема будет описана детальнее вместе с датасетами, на которых модель будет обучена и протестирована. Третья глава описывает предварительную обработку данных и выбор признаков для модели. В четвертой главе описаны модели, которые будут обучены, и в пятой главе описаны результаты обучения. В шестой главе будут проанализированы полученные результаты и предложены пути для дальнейшего развития.

# Описание проблемы и данные

Из-за неразглашения данных и\или проблем логистики у вина известны только его физико-химический состав и вкусовые качества. Необходимо классифицировать вино по качеству от 0 до 10.

Данные представлены в виде CSV таблицы, при этом присутствует одна категориальная переменная и пропуски.

Пример данных:



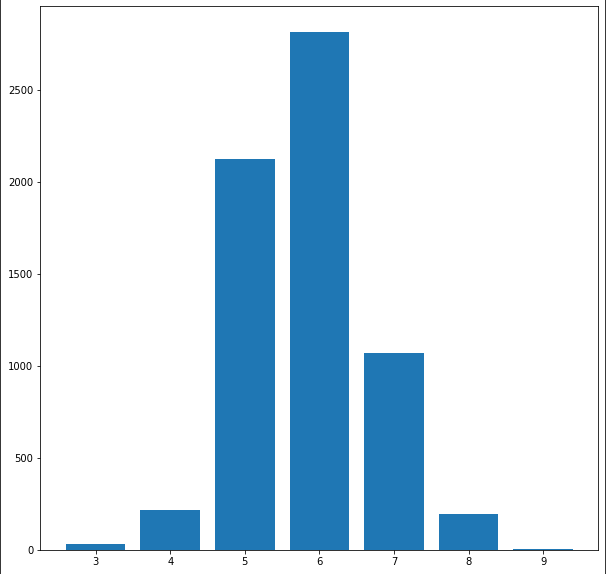
Описание переменных:

* type - тип вина (белое\красное)
* fixed acidity - фиксированная кислотность
* volatile acidity - переменная кислотность
* citric acid - цитрусовая кислотность
* residual sugar - остаточный сахар
* chlorides - хлориды
* free sulfur dioxide - свободный диоксид серы
* total sulfur dioxide - общий диоксид серы
* density - плотность
* pH
* sulphates - сульфаты
* alcohol - алкоголь

Целевая переменная:

* quality (оценка от 0 до 10)

Из представленных данных шкала оценок для обучения содержит только диапазон от 3 до 9, что сокращает маркировки классификации. При этом классы распределены неравномерно, что показывает гистограмма ниже



# Выбор признаков и предварительная обработка

В отсутствие экспертных знаний в данной области, принято решение выявлять пропуски алгоритмически. Для этого использовались алгоритмы изолирующего леса, One-Class SVM, Local Outlier Factor и Elliptic Envelope. Строки, которые каждый из алгоритмов считает выбросами менее 5% от общего количества данных, можно удалить без потери качества. Строк с пропусками также незначительное количество от общих данных, поэтому их тоже можно удалить.

Также было принято решение закодировать тип вина как One-Hot Encoding, возможно, разные характеристики по-разному влияют на качество белых и красных вин.

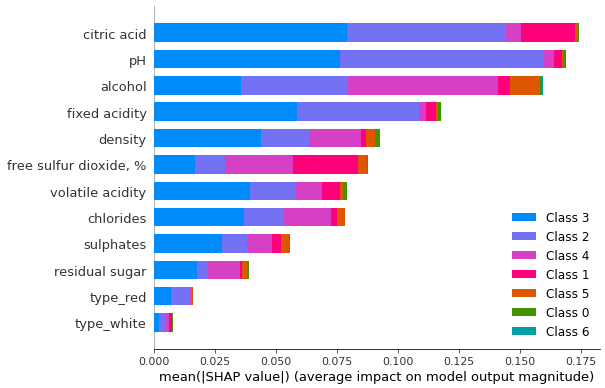
Из-за высокой корреляции между количеством свободных диоксидов серы их общим количеством был введен признак процентного их содержания.

Остальные признаки остались без изменений.

Чтобы исправить неравномерное распределение классов, был проведен оверсемплинг на тестовой выборке.

# Выбор признаков и моделей для обучения

Ниже приведена диаграмма важности признаков для выявления того или иного класса.



Как видно из диаграммы, тип вина не вносит вклада в определение качества вина, поэтому он был отброшен. Для классификации на оставшихся признаках были выбраны две модели: логистическая регрессия (линейная модель) и случайный лес (нелинейная модель).

# Результаты

Логистическая регрессия плохо справляется с данной задачей, показывая точность 48%. Случайный лес показывает на обучающей выборке точность в 99%, что говорит о переобучении. На тестовой выборке точность составляет 65%.

Ниже приведена матрица ошибок для случайного леса.

| Класс | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | 0 | 1 | 3 | 2 | 0 | 0 | 0 |
| 4 | 2 | 8 | 16 | 16 | 1 | 0 | 0 |
| 5 | 0 | 7 | 307 | 99 | 12 | 0 | 0 |
| 6 | 0 | 3 | 116 | 369 | 75 | 0 | 0 |
| 7 | 0 | 1 | 8 | 63 | 139 | 4 | 0 |
| 8 | 0 | 0 | 0 | 7 | 16 | 15 | 0 |
| 9 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

# Анализ результатов

В данном контексте модель неплохо справляется с задачей. В основном ошибки возникают в смежных классах, что допустимо для бизнеса. Увеличение точности может быть достигнуто созданием дополнительных признаков с привлечением специалистов в химическом анализе. Также стоит попробовать оптимизировать параметры случайного леса, другие виды моделей и их ансамбли.