Оценка качества вина по физико-химическому составу

1. Введение

Вино является популярным алкогольным напитком, востребованным по всему миру. Вкус и качество вина зависят от множества факторов: даже из одного и того же сорта винограда, собранного с одной территории, могут получится вина, отличающиеся друг от друга. В этой задаче необходимо изучить взаимосвязь состава вина и его оценки.

2. Описание данных

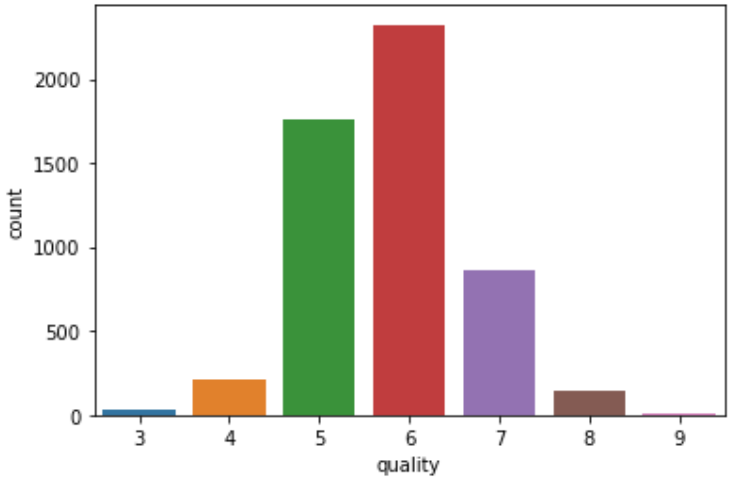
В датасете собраны данные о физико-химическом составе различных вин. Набор данных включает в себя 6497 строк и 12 признаков. Целевая переменная Quality (качество) оценивается по шкале от 0 до 10 баллов.

Таблица 1. Физико-химические параметры вин

type	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
white	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
white	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
white	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
white	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Значения целевой переменной распределены неравномерно: вин среднего качества много, вин с хорошими оценками (3, 4) или плохими оценками (8, 9) мало, вина с очень плохими или очень хорошим оценками отсутствуют

Рисунок 1. Распределение оценок качества вина



3. Предобработка данных

Перед обучением моделей данные были обработаны следующим образом:

- пропущенные значения заполнены средними значениями соответствующих признаков,

- признаки с наиболее ассиметричным распределением были логарифмированы,

- добавлен новый признак “sweetness”: все вина распределены по категориям «сухое», «полусухое», «полусладкое», «сладкое»,

- категориальные признаки переведены в числовые с помощью One-hot encoding,

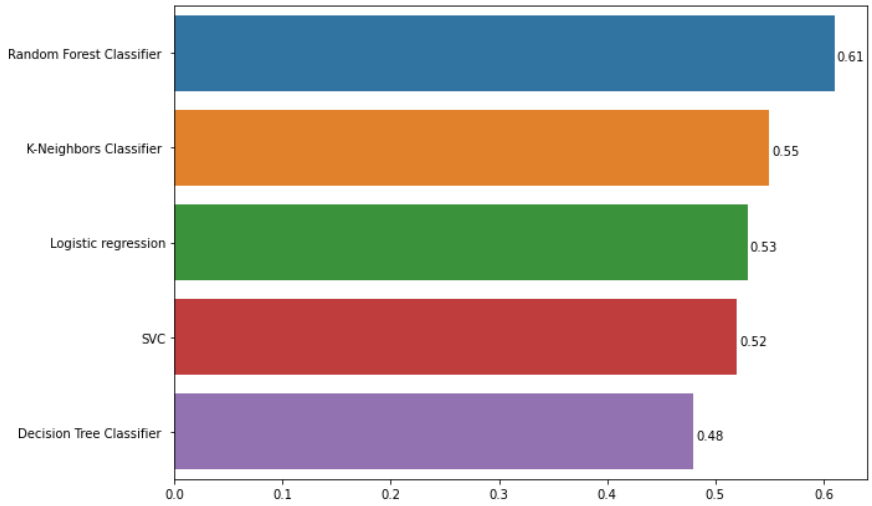
- удалены дубликаты строк,

- применен oversampling: по алгоритму SMOTE добавлены объекты наиболее редких классов.

4. Обучение моделей

Для обучения датасет был разделен на тренировочный и валидационный в соотношении 3:1. К данным применялись различные классификаторы из пакета sklearn. В качестве ключевой метрики рассматривался accuracy score. Наилучший результат показал алгоритм Random Forest Classifier с показателем accuracy равным 0,61.

Рисунок 2. Accuracy score для классификаторов sklearn



5. Дальнейшие действия

Показатель accuracy не очень высок, требуется оптимизация модели либо данных. Необходимы повторные исследования данных, более продвинутый feature engineering, подбор гиперпараметров основной модели.