Wine Type and Quality Classification

1. Вступление

Вино - это алкогольный напиток, изготовленный из сброженного винограда. Дрожжи потребляют сахар в винограде и превращают его в этанол, углекислый газ и тепло. Это приятный дегустационный алкогольный напиток, любимый целлебрат. Безусловно, будет интересно проанализировать физико-химические свойства вина и понять их связь и значение с качеством вина и классификацией типов.

Набор данных связан с красным и белым вариантами вина "Виньо Верде". Винью Верде является уникальным продуктом из региона Минью (северо-запад) Португалии. Среднее в алкоголе, оно особенно ценится благодаря своей свежести (особенно летом). Этот набор данных общедоступен только для исследовательских целей, для более подробной информации читайте Cortez и др., 2009. . В связи с проблемами конфиденциальности и логистики, доступны только физико-химические (входные) и сенсорные (выходные) переменные (например, нет данных о типах винограда, марке вина, продажной цене вина и т.д.).

2. Постановка задачи

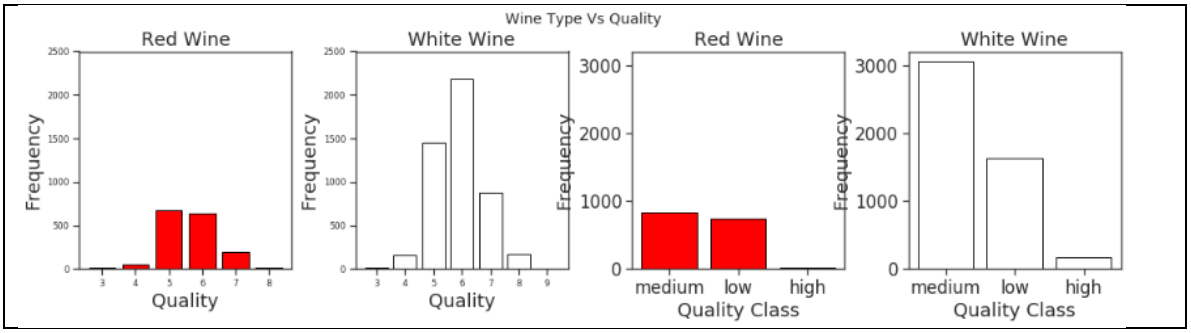
2.1 Предсказать, является ли каждый образец вина красным или белым.

2.2 Прогнозируйте качество каждого образца вина, которое может быть низким, средним или высоким.

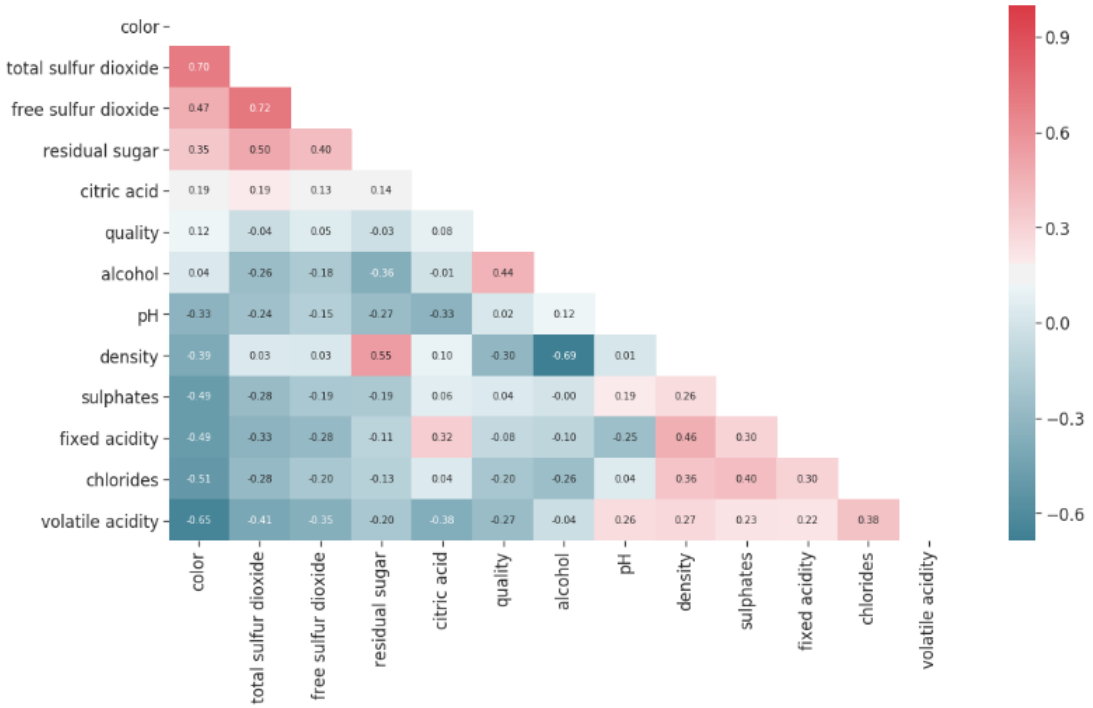
3. Исследование (анализ) датасета

3.1 Размер датасета: (6497, 14)

3.2 После обработки датасета мы получили распределение данных по качества типам и визуализировали дисбаланс классами, особенно в случаях с высоким качеством.



3.3 Оценка корреляции данных показала, что, несмотря на слабую корреляцию большинства данных являются некоторые имеют явную сложную зависимость, например остаточный сахар имеет половинное отношение к общему диоксиду серы и 0,40 со свободным, это хороший признак того, что чем больше остаточного сахара, тем больше диоксида серы добавляется виноделом. 0,5 указывает на то, что белое вино, как правило, имеет больше остаточного сахара, чем красное вино.



Очевидно, что для повышения точности классификации целесообразно использовать специально подготовленные данные. Цель состоит в том, чтобы найти подпространство признаков, оптимизирующее классовую разделимость. Мы остановились на методе главных компонент (Principal component analysis), поскольку он показал набольшую точность на тестовой выборке: 99.33% против 96.29% у Linear Discriminant Analysis (KNeighbors Classifier).

4. Классификационные модели и результаты

Мы рассмотрели различные подходы к реализации классификаторов моделей, а также использовать гиперпараметризацию, перекрестную валидацию и сравнение результатов между различными показателями ошибок.

Поскольку первая задача бинарна и как показано выше, существую корреляция между данными, мы разработали собственный классификатор использующий метод главных компонент, как поставщика данных и классификатор, который мы подбирали в поиске наилучшего результата. Кроме того, мы разработали нейронную сеть, состоящую из 6 слоёв, для классификации мы использовали DNN classifier.

4.1 Мы начали с Logistic Regression с использованием подбора оптимальных параметров GridSearchCV(). Наилучшая точно была достигнута в 99.38% на всех метриках качества и 99.28% на ROC AUC score.

4.1.1 Эти показатели мы с параметрами классификатора:

4.1.2 Нейронная сеть (DNN Regressor) показала среднюю точность: **99,60**% на всех метриках качества **99,99**% на на ROC AUC score.

Сравнение методов классификации показывает, явное преимущество в точности классификации:

	CV Accuracy	Accuracy	ROC AUC Score	ROC Area
LogisticRegression	0.995164	0.9938	0.992825	0.998977
DNN Regressor	0.995962	0.9954	0.999886	0.999886

4.2 Для задач определения качества вина мы использовали классификаторы, для которых так же проводили поиск оптимальных параметров.

Мы отказались от использования метода главных компонент, как источника данных для классификатора, поскольку это приводит к существенному снижения точности.

№	Classifier	CV Accuracy	Accuracy	ROC AUC Score	ROC Area
1.	Decision Tree	0.713540	0.7262	0.723468	0.806923
2.	Random Forest Classifier	0.805416	0.8213	0.807632	0.908012
3.	XGBoost (eXtreme Gradient Boosting)	0.782592	0.8190	0.809842	0.892034
4.	KNeighborsClassifier	0.785493	0.8090	0.796836	0.887450
5.	GradientBoostingClassifier	0.789168	0.8198	0.810580	0.874790
6.	AdaBoostClassifier	0.704642	0.7193	0.696634	0.693622
7.	LogisticRegression	0.712959	0.7270	0.724063	0.809575
8.	LinearSVC	0.712766	0.7278	0.726263	0.799150

Как видно, наилучшая точность достигнута на 2, 3 и 5 классификаторе, проведя стекинг моделей мы получили наилучшую точность на паре Random Forest Classifier и Gradient Boosting Classifier:

	precision	recall	f1-score	support
low	0.79	0.78	0.78	476
medium	0.84	0.87	0.85	778
high	0.93	0.33	0.49	39
avg / total	0.82	0.82	0.82	1293

Метрика ROC AUC Score by Classes:

{'low': 0.8258, 'medium': 0.8085, 'high': 0.6663}

ROC AUC Score: 81.06%

Как мы и ожидали стекинг моделей позволил несколько улучшить точность классификации.

5. Заключение

Мы достигли своей цели, получив хорошие модели, смогли создать классификационные модели как для типа, так и для качества вина, проходя все этапы в стандартных моделях машинного обучения и интеллектуального анализа данных, таких как модель CRISP-DM,

Однако наша модель имеет разрыв около 13%. Частично это можно объяснить тем, что качественные заметки на самом деле являются заметками, выделенными экспертом, который учитывает другие признаки, которые не являются физико-химическими, и поэтому не присутствуют в этом наборе данных. Изложение записки является сенсорной оценкой, основанной на процессе дегустации, если назвать лишь некоторые из них:

- Другие релевантные физические свойства, не присутствующие в данных, такие как прозрачность или вязкость.
- Некоторые из характеристик, которые мы имеем, могут влиять на запах и вкус вина, но, у нас нет всех необходимых данных, например, сорт или тип винограда.
- У нас нет танина, одной из основных характеристик вина. В вине именно присутствие фенольных соединений добавляет вину горечь.
- Традиционно производимые в португальском регионе Верде вина - это молодые вина, обладающие высокой кислотностью и заметной свежестью. Поэтому время охраны не столь важно для классификации этого вина, но урожай, как и в других винах, тоже.

Мы пришли к выводу, что по физико-химическим признакам можно предсказать качество вина и его тип. Несмотря на то, что предсказание типа имеет только образовательную основу, предсказание качества дает некоторые практические возможности, например:

- Винный магазин или крупный дистрибьютор может квалифицировать новые вина еще до их приобретения и, таким образом, лучше оценить их стоимость покупки и возможность продажи.
- Результат модели, ее интерпретация и все оценки EDA дают методы и правила принятия решений, которые могут помочь виноделам искать вина лучшего качества.

6. Ссылки

- Wines Type and Quality Classification Exercises
  https://www.kaggle.com/mgmarques/wines-type-and-quality-classification-exercises