Определение качества вина

Домашнее задание

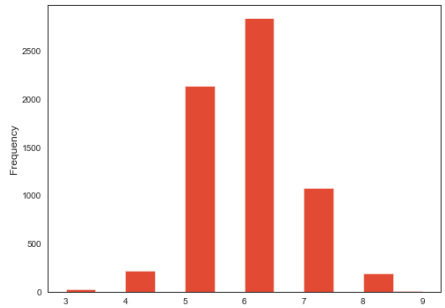
Возьмите свое решение задачи с винами (ДЗ #2) и напишите отчет по этой задаче.

1. Введение

Качество вина – это его соответствие нормативным показателям по химическому составу, окраске, прозрачности, аромату и вкусу. Так же, как и все мы имеем разные вкусы, например, в еде или одежде, так же наши предпочтения в винах тоже могут различаться. Но при определении качества вина нельзя назвать вино «плохим» только потому, что оно кому-то не нравится. Необходимо найти возможность оценки качества вина по заданным показателям, исключая субъективный фактор.

1.1. Описание проблемы и набор данных

Для анализа представлен датасет "Wines", содержащий информацию о красном и белом вариантах португальского вина "Винью Верде". Датасет состоит из 11 характеристик 6497 вин и оценок их качества по шкале от 0 до 10.



Классы вина не сбалансированы, нормальных вин гораздо больше, чем отличных или плохих.

1.2. Решение

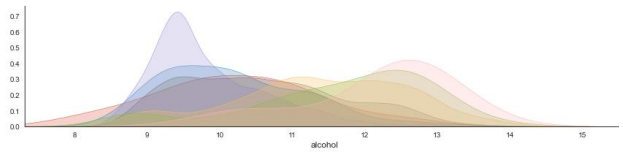
Анализируем характеристики вин, строим модель классификации и оцениваем ее качество.

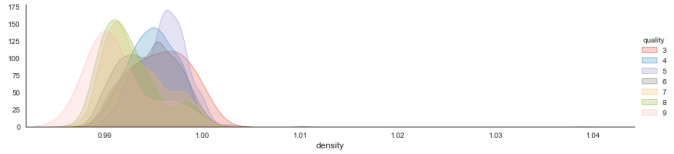
2. Реализация

2.1 Анализ данных и выделение нужных признаков

Тепловая карта корреляции признаков представлена на изображении ниже

Наиболее скоррелированны с качеством вина два показателя: процент алкоголя и плотность.





Но все распределения перекрывают друг друга, значит однозначно определить качество вина по уровню алкоголя и плотности нельзя.

2.2 Подготовка данных

Пропусков в датасете немного, заменяем имеющиеся средним арифметическим значением.

Переменная type (тип вина) - категориальная, бинаризуем ее, 1 - белое вино, 0 - красное.

Перед построением модели производим нормализацию данных.

2.3 Построение модели и оценка качества

В качестве классификатора выбран случайный лес RandomForestClassifier.

Score, полученный на обучающей выборке - ***0.85***, на тестовой выборке - ***0.6***

	3	4	5	6	7	8	9
3	0	0	3	5	0	0	0
4	0	1	19	19	0	0	0
5	0	0	311	152	5	0	0
6	0	0	103	438	50	0	0
7	0	0	8	126	79	1	0
8	0	0	1	28	9	6	0
9	0	0	0	0	1	0	0

Матрица ошибок

3. Заключение

В результате проделанной работы была построена модель, способная на наборе характеристик вина предсказывать его качество. В качестве классификатора выбран случайный лес RandomForestClassifier.

Качество модели получилось не очень хорошим, скорее всего из-за изначальной несбалансированности классов.

Возможно улучшить качество модели, если разбивать датасет на обучающую и тестовую выборки не случайным образом, а так, чтобы в каждой равномерно присутствовали все классы. А так же улучшить показатели модели можно используя для обучения датасет с гораздо большим количеством вин и их характеристик.