Классификация качества вина с помощью модели машинного обучения.

Исследовательский анализ данных:

Вино состоит из многих элементов, которые способствуют его качеству. Стоит разобраться, что означают эти черты в вине. Вот их небольшой обзор:

* Тип

Цвет вина зависит от цвета винограда: красное вино производится из красного и черного винограда, а белое вино — из белого винограда. При производстве белого вина семена и кожица удаляются, в то время как при производстве красного вина этого не требуется. Эта разница в процессе влияет на вкус и цвет, делая красное вино более насыщенным, а белое вино более легким и фруктовым. Этот набор данных имеет атрибут type, который идентифицирует цвет вина.

* Кислотность

Кислотность в вине очень важна, так как она влияет на вкус. Вино с высокой кислотностью можно определить как бодрящее, яркое, пикантное, угловатое, живое, хрустящее, свежее, крепкое. Идеально сочетается с едой. В то время как пониженная кислотность будет охарактеризована как жирная, несвежая, мягкая. Кислотность – важнейшая характеристика игристых вин. Помимо изменения вкуса вина, он также позволяет ему дольше храниться, поскольку выступает в качестве естественного консерванта структуры вина. Кислоты бывают разные:

Фиксированная кислотность: происходит от кислот, содержащихся в винограде. Фиксированная кислотность в наборе данных выражена в г(винная кислота)/дм3. Летучая кислотность: вызвана присутствием уксусной кислоты. Если присутствует небольшое количество летучих кислот, это усиливает запах вина. Когда летучие кислоты слишком высоки, они добавляют уксусный привкус и ухудшают качество вина. Летучая кислотность выражена в г(уксусной кислоты)/дм3 в наборе данных. Лимонная кислота: вводится в качестве добавки во время процесса брожения вина и помогает повысить общую кислотность. Он также используется для целей стабилизации. Лимонная кислота в наборе данных выражена в г/дм3. В винном мире pH (потенциал водорода) — это числовая шкала, определяющая кислотность вина. Все вина находятся на кислой стороне спектра pH, и большинство из них колеблется от 2,5 до примерно 4,5 pH (напиток с pH = 7 нейтрален, вода).

* Остаточный сахар

Сладость вина исходит от сахаров в виноградном соке: глюкозы и фруктозы. Различные уровни этих элементов способствуют образованию остаточного сахара. Другие источники сахара также могут быть добавлены в процессе ферментации, например, кукурузный сахар, дрожжи. Остаточный сахар является мерой количества сухих веществ сахара в конце процесса производства вина. Остаточный сахар в наборе данных выражается в г/дм3.

* Хлориды

Хлориды – это количество соли в вине. Это происходит из-за различных аспектов винограда и условий окружающей среды, в которых он вырос. Хлориды в наборе данных выражены в г(хлорид натрия)/дм3.

* Сульфаты

Сульфаты в вине появляются в процессе его брожения. Сульфаты в вине обратно пропорциональны кислотности и цвету. Остаточный сахар прямо пропорционален количеству сульфитов для предотвращения брожения. Сульфаты встречаются в природе, но их также можно добавлять в процессе. Они помогают защитить вино от нежелательных микробов и окисления. Существует несколько способов измерения сульфитов в вине: свободный диоксид серы, общий диоксид серы и сульфаты.

Свободный диоксид серы: чрезмерное количество свободного диоксида серы может быть заметно для потребителей, скрывая ароматы вина и препятствуя его насыщению кислородом в процессе дыхания. В высоких концентрациях придает резкий, горьковатый, металлический привкус. Диоксид серы в основном используется для уничтожения вредных бактерий, но в то же время для сохранения качества и свежести. Свободный диоксид серы в наборе данных выражен в мг/дм3.

Общий диоксид серы: добавляется для уничтожения бактерий и сохранения качества вина. Чрезмерное количество этого диоксида может повредить вино, добавив нежелательный запах. Общий диоксид серы в наборе данных выражен в мг/дм3.

Сульфаты: минеральные соли, содержащие серу. Сульфаты необходимы для производства вина. Он способствует аромату и вкусу вина. Сульфаты выражены в г (сульфат калия)/дм3 в наборе данных.

* Алкоголь

Спирт является результатом преобразования сахара дрожжами в этанол в процессе ферментации. Уровень алкоголя влияет на вкус вина, более высокий уровень алкоголя делает вкус более сильным, а более низкий уровень алкоголя делает вино более легким. Алкоголь выражен в % vol в наборе данных.

* Плотность

Плотность – это сравнение веса определенного объема вина с эквивалентным объемом воды. Он обычно используется как мера превращения сахара в спирт. В наборе данных плотность выражается в г/см3.

Другие важные аспекты атрибутов, на которые стоит обратить внимание:

1. Исследуемый набор данных не содержит информации о сортах винограда, марке вина или цене.
2. В нашем датасете присутствует 6 классов вина, по шкале от 3 до 9. Видим дисбаланс классов:

- 6 класс - 2836

- 5 класс – 2138

- 7 класс - 1079

- 4 класс - 216

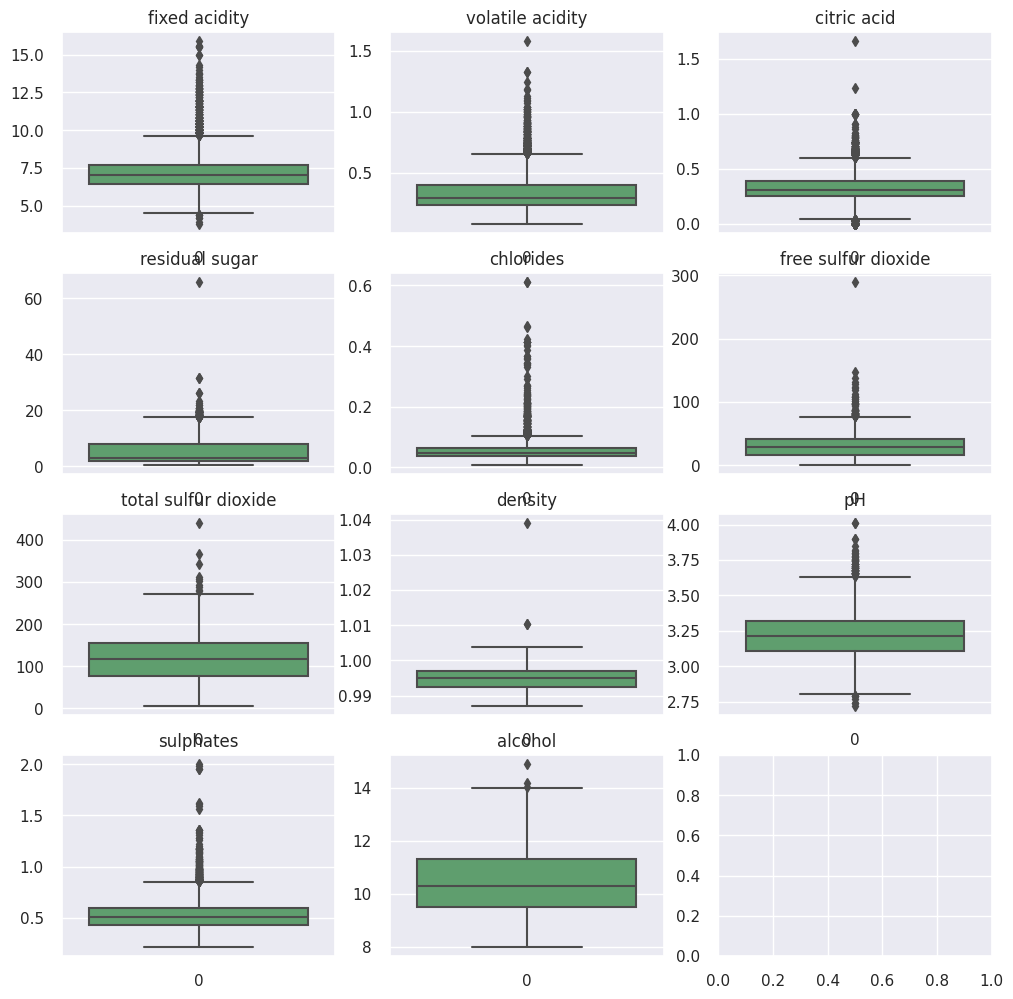
- 8 класс - 193

- 3 класс - 30

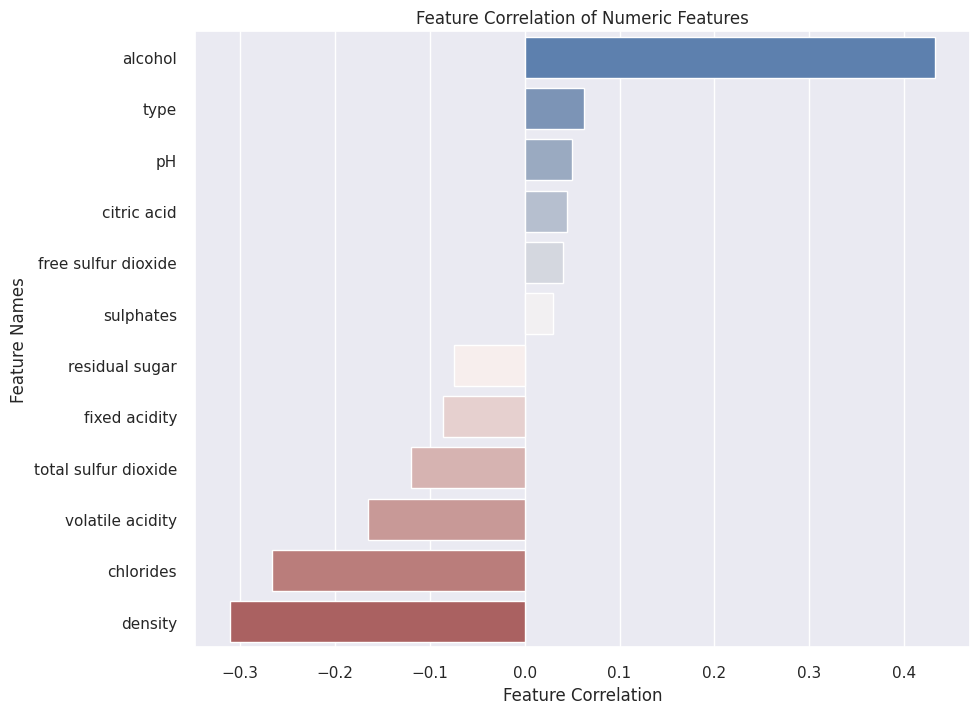
- 9 класс – 5

В последующем, при обучении модели приведем классы к балансу с помощью библиотеки imblearn.over\_sampling методом SMOTE.

1. Все пропущенные значения представляют собой очень низкий процент в числовых характеристиках. Мы заполним нулевые значения средним значением для каждого атрибута.
2. В датасете присутсвуют выбросы. Удалим выбросы для последующего обучения модели.



1. Основными предикторами для целевой переменной являются:
   * Алкоголь
   * В меньшей степени pH баланс
   * Плотность и хлориды имеет сильную отрицательную корреляцию на качество вина.



Обучение модели:

В этом руководстве мы будем использовать PyCaret для разработки конвейера машинного обучения.

В основе baseline моделей были выбраны:

- RandomForestClassifier

- ExtraTreesClassifier

- GradientBoostingClassifier

- XGBClassifier

- LGBMClassifier

Модель ExtraTreesClassifier показала результат выше по сравнению с другими моделями, качество модели составило 0.863. Из обучения был удален признак тип вина, пскольку после удаления выбросов распределене между двумя типами стало 92% одного типа и 8% другого типа. С помощью библиотеки Optuna подберем гиперпараметры для лучшего качества модели. Подобрав гиперпараметры качество модели составило - 88.83. Что показывает неплохой результат.

Сохраним модель с помощью библиотеки joblib.