**Лабораторная работа 2**

Описание датасета

Эти данные являются характеристиками красного вина. Набор данных загружается из репозитория машинного обучения UCI. В нём 1599 записей и 12 атрибутов.

Attribute Information:

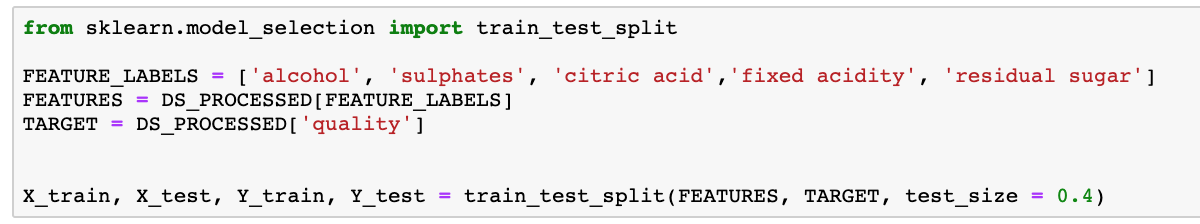


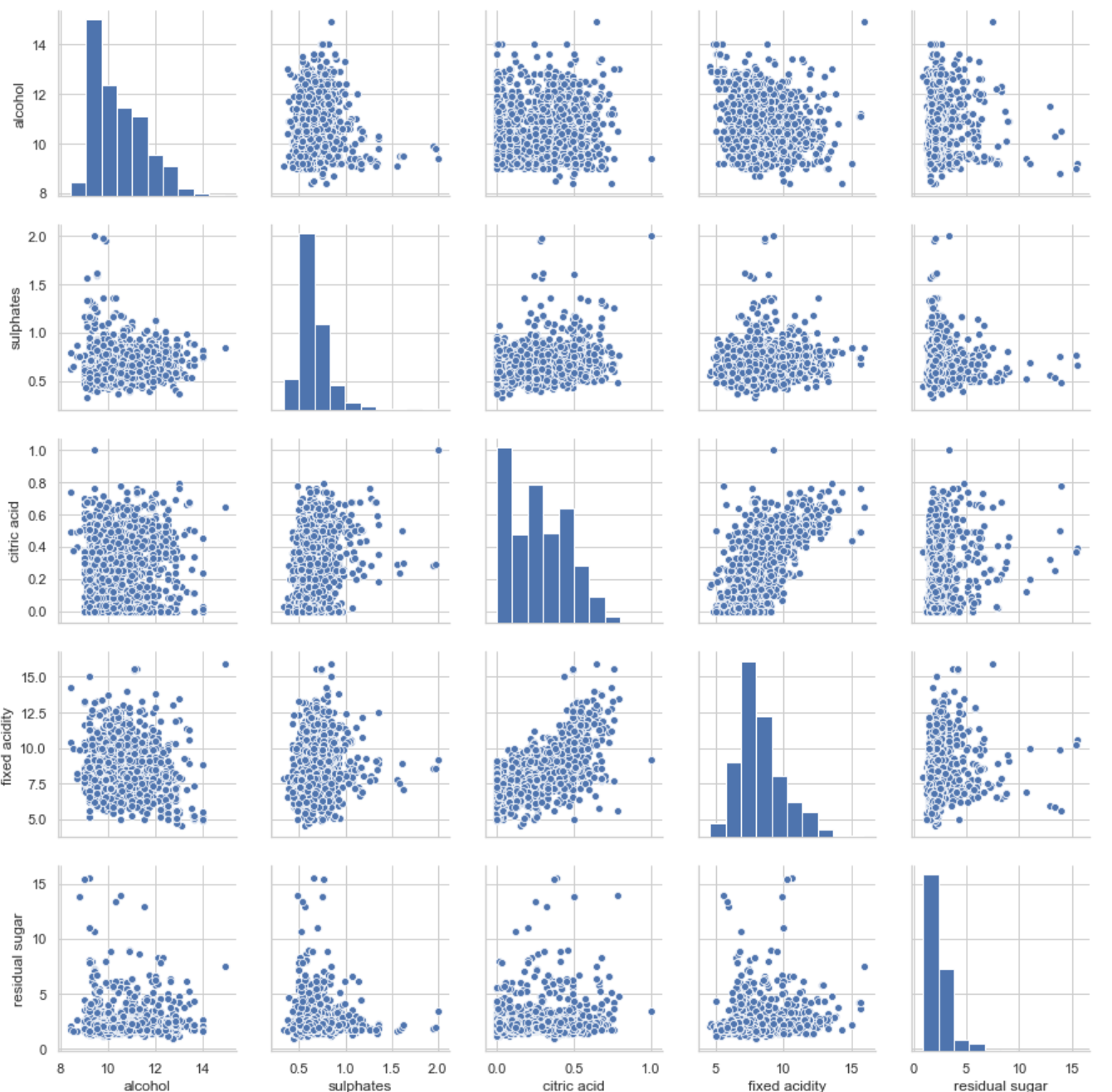
Для отбора признаков которые будем использовать в модели посмотрим что наиболее коррелирует с атрибутом «качество»

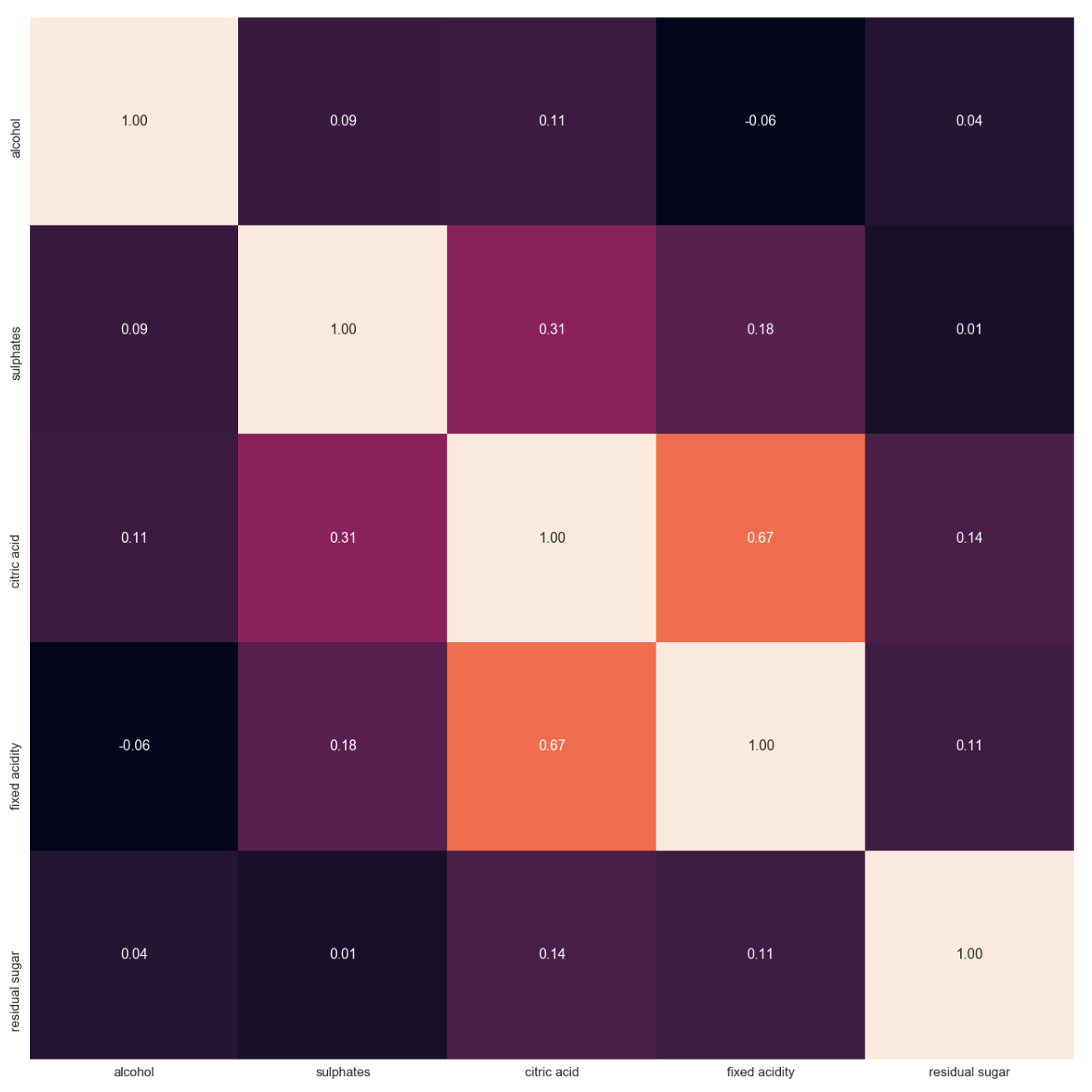


Тут показана корреляция от меньшей к большей, соответственно будем использовать нижнюю часть списка.

После того как отобрали признаки которые будем использовать, разделим выборку на тестовую и обучающую (40/60)





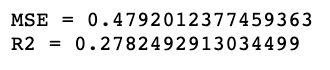


Линейная регрессия

Линейная регрессия может пониматься как нахождение оптимально подогнанной прямой линии, проходящей через точки образцов тренировочных данных.

средневзвешенная квадратичная ошибка (mean squaгed епог, MSE) - просто усредненное значение функции стоимости SSE, которую мы минимизируем для подгонки линейной регрессионной модели.

R2 - коэффициент детерминации, это доля дисперсии отклика, которая охвачена моделью.

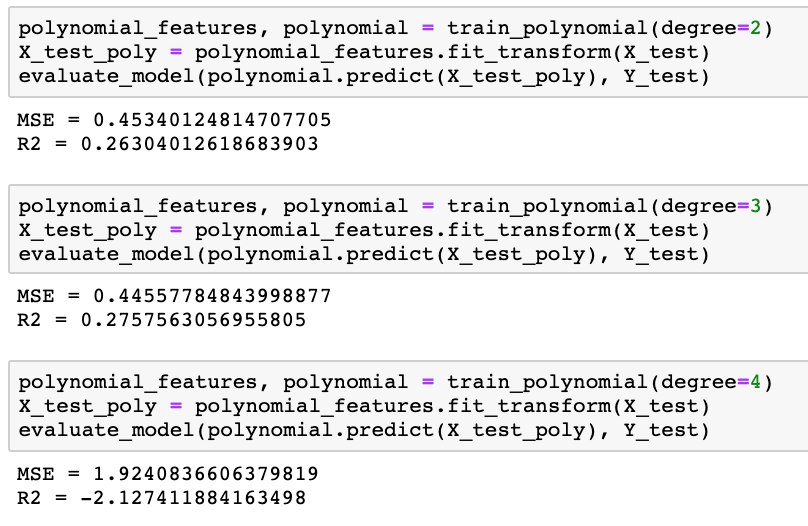


Модель показала не очень хорошие результаты на выбранных атрибутах. При этом в датасете нет атрибутов, имеющих более высокий коэффициент корелляции, поэтому вряд ли линейная регрессия сможет дать значительно лучшие результаты на этих данных.

Полиномиальная регрессия

Здесь degree обозначает степень полинома. Несмотря на то что мы можем использовать полиномиальную регрессию для моделирования нелинейных связей, она по­ прежнему рассматривается как модель множественной линейной регрессии, ввиду линейных коэффициентов регрессии w.

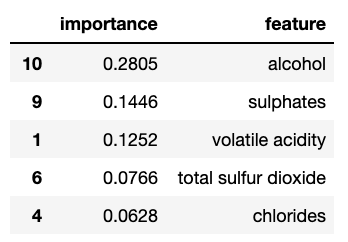
Теперь обсудим вопрос использования класса-преобразователя для полиномиальных признаков PolynomialFeatures библиотеки scikit-learn с целью добавления в задачу простой регрессии.



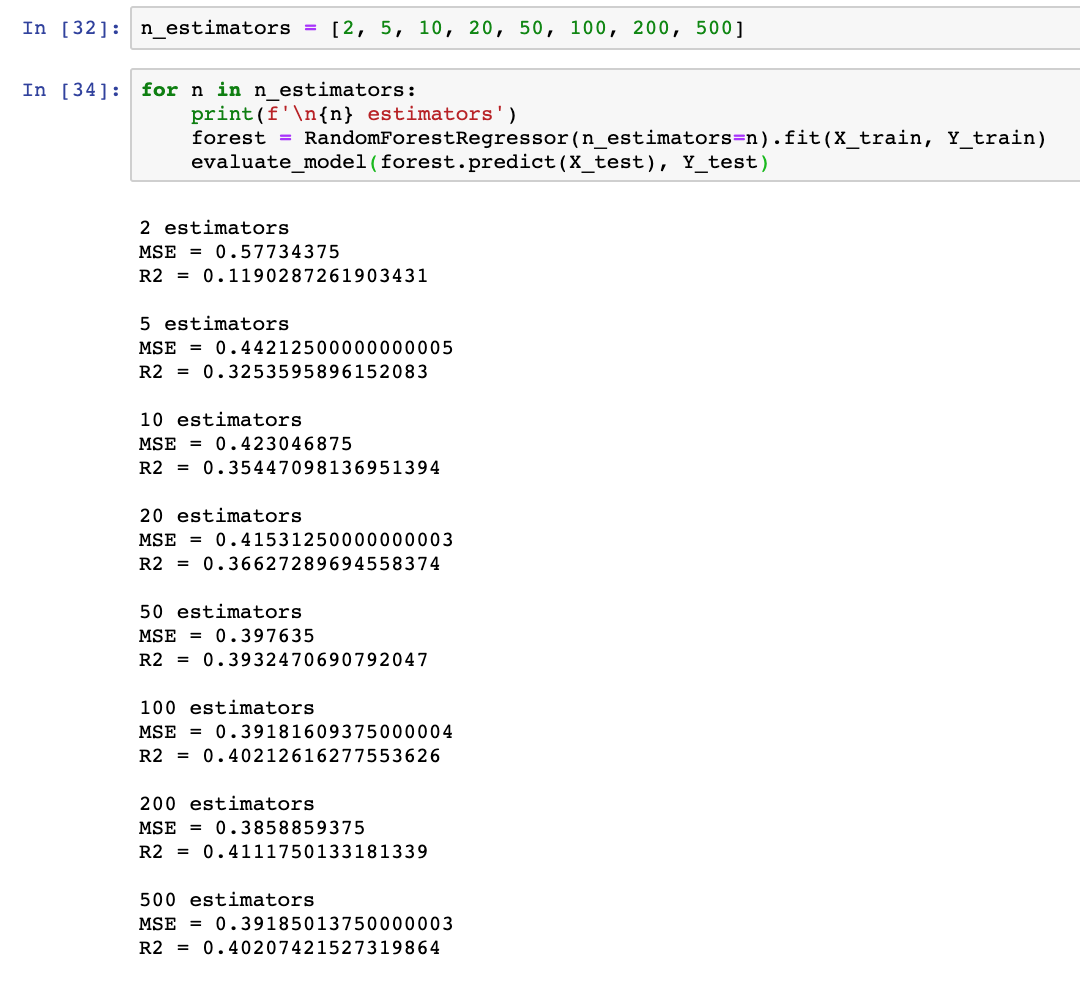
Случайный лес

Преимущества случайных лесов состоят в том, что они менее чувствительны к выбросам в наборе данных и не требуют большой настройки параметров.

Для случайного леса выберем атрибуты основываясь на их важности. Для этого сначала обучим модель на всех атрибутах и посмотрим, насколько хорошо каждый атрибут разделял датасет.



Единственный параметр в случайных лесах, с которым нам нужно экспериментировать, - это, как правило, число деревьев в ансамбле.



Вывод

Мы построили нашу первую модель, реализовав линейную регрессию. Затем мы увидели, каким образом применять на практике линейные модели библиотеки scikit-learn для задачи регрессии. Чтобы определить предсказательную способность регрессионных моделей, мы вычислили средневзвешенную сумму квадратичных ошибок и связанную с ней метрику R2. Мы также представили несколько подходов для моделирования нелинейных связей, включая полиномиальное преобразование признаков и регрессоры на основе случайных лесов.

На выбранном датасете лучше всего проявил себя случайный лес, в то время как линейная и полиномиальная регрессия показали не очень хорошие показатели. Это объясняется тем, что даже у атрибутов с макимальным уровнем корреляции были слабо заметны линейные и полиномиальные зависимости для целевого атрибута.