Министерство науки и высшего образования Российской Федерации

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ

ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Институт прикладной математики и компьютерных наук

Кафедра прикладной информатики

**Отчет по лабораторной работе №2 по дисциплине «Интеллектуальные системы»**

Выполнил

студент группы № 1462

\_\_\_\_\_\_\_\_\_\_\_\_А.В. Шабля

Проверил

доцент, канд. техн. наук

\_\_\_\_\_\_\_\_\_\_\_\_С.В. Аксёнов

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Томск – 2020

**Описание**

Для выполнения работы был выбран набор данных Wine Quality Data Set (white)

<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>

Размер выборки: 4898 элемента.

Attribute Information:

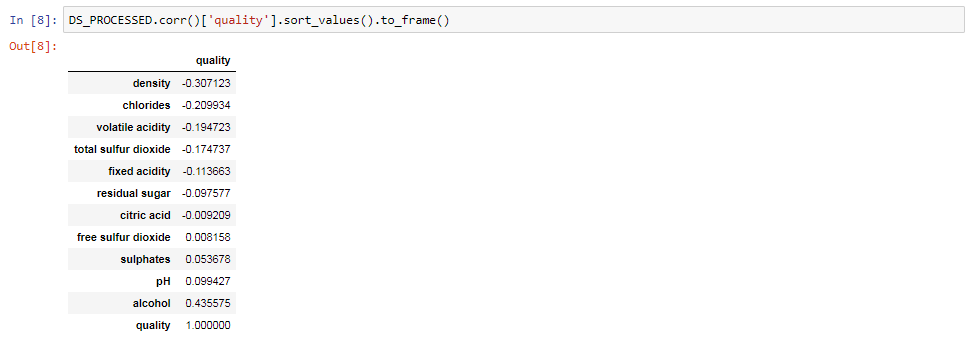
* fixed acidity
* volatile acidity
* citric acid
* residual sugar
* chlorides
* free sulfur dioxide
* total sulfur dioxide
* density
* pH
* sulphates
* alcohol

Output variable (based on sensory data):

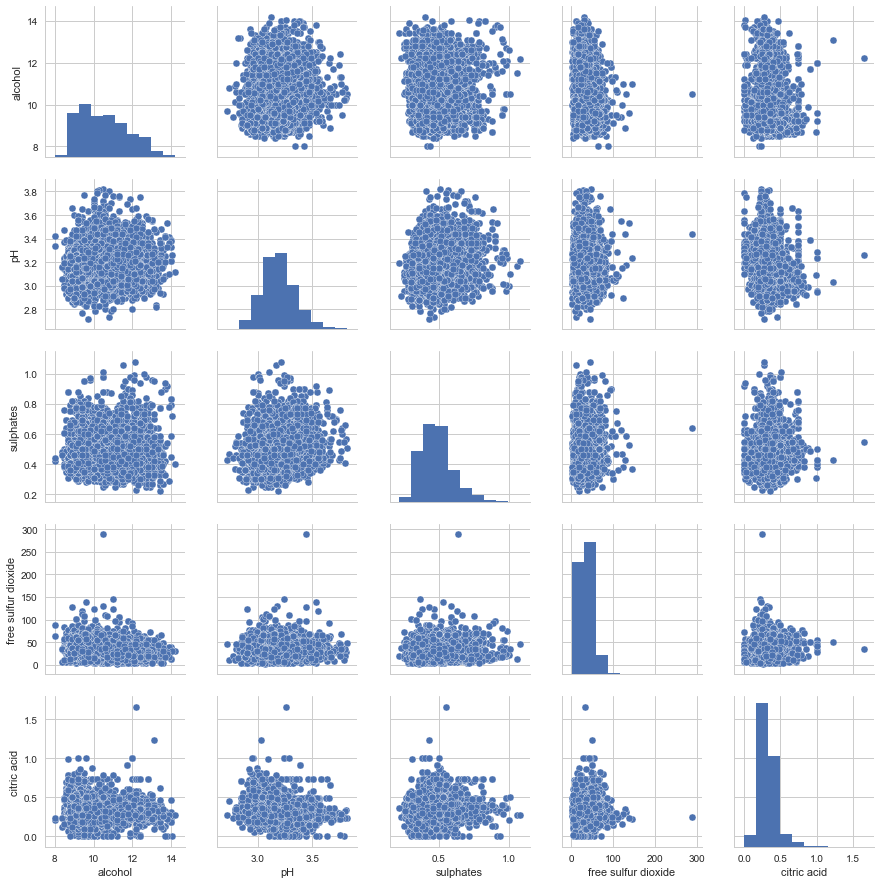
* quality (score between 0 and 10)

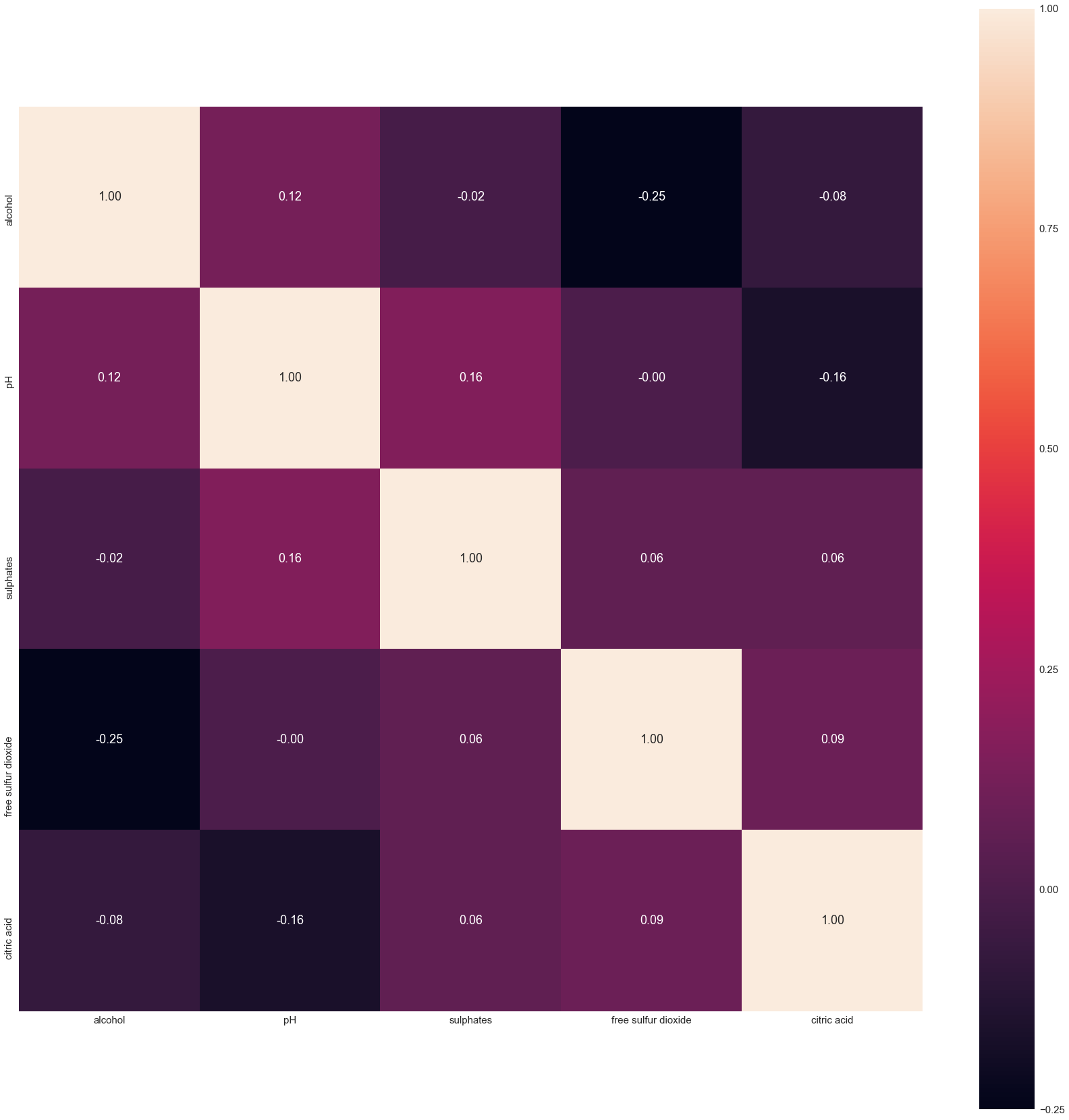
Для удобства, файл датасета был пересохранен, так как, хоть он и был в формате .csv (разделители - запятые), данные были в виде обычной таблицы. Затем были убраны все кавычки, так как данные в кавычках не считываются. В остальном, данные оказались целостными и без пропусков, поэтому дополнительных обработок не потребовалось.

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

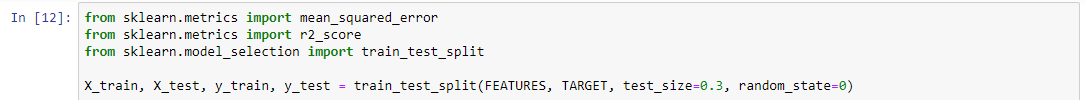


Тут показана корреляция от меньшей к большей, соответственно будем использовать нижнюю часть списка.





После того как отобрали признаки которые будем использовать, разделим выборку на тестовую и обучающую (30/70)

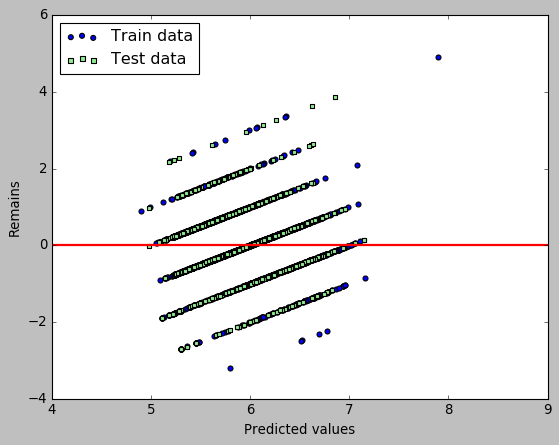


**Линейная регрессия**

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.



MSE train: 0.603

MSE test: 0.664

R2 train: 0.220

R2 test: 0.179

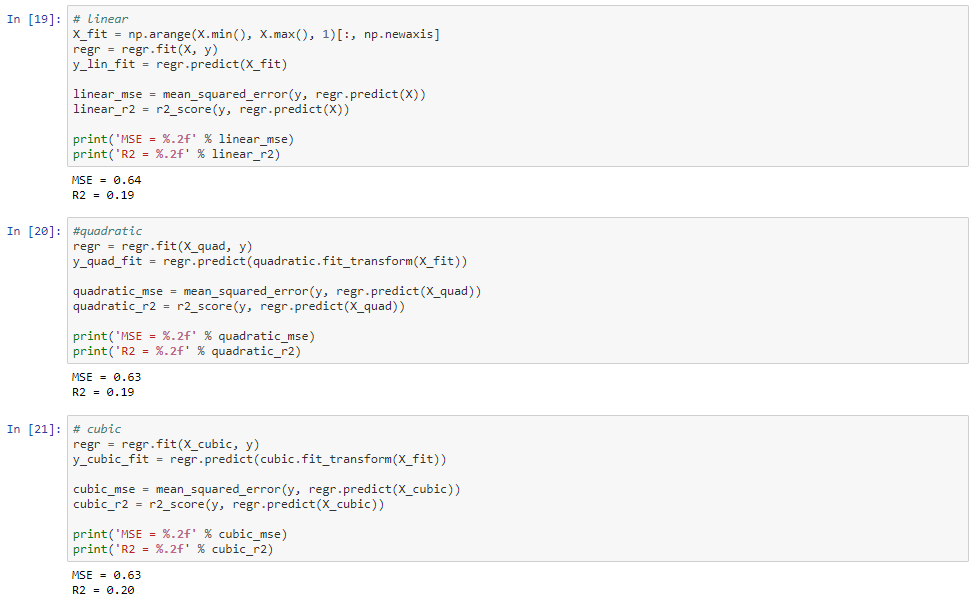
Модель показала не очень плохие результаты на выбранных атрибутах. При этом в датасете нет атрибутов, имеющих более высокий коэффициент корелляции.

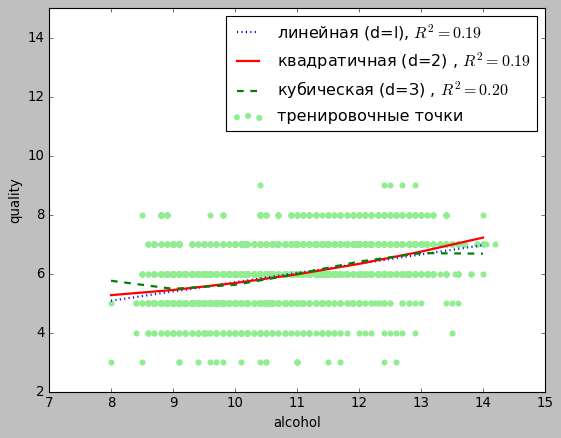
**Полиномиальная регрессия**

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.

Теперь мы смоделируем связь между параметрами ‘alcohol’ и ‘quality’ с использованием полинома второй степени (квадратичного) и третьей степени (кубического ) и сравним его с линейной подгонкой.







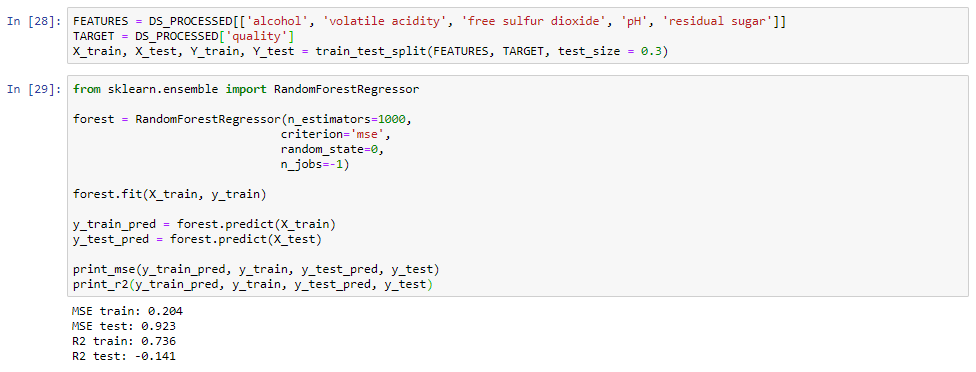
**Случайный лес**

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

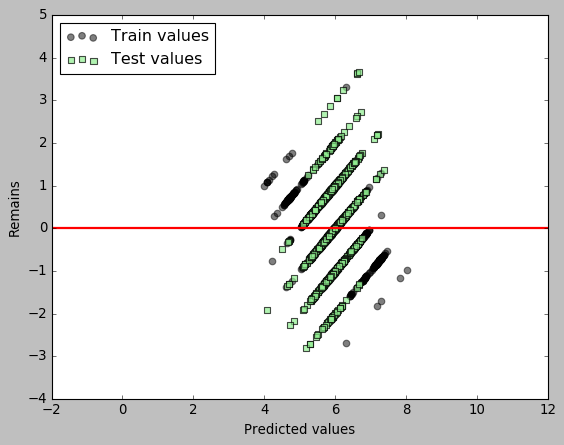
Для случайного леса выберем атрибуты основываясь на их важности. Для этого сначала обучим модель на всех атрибутах и посмотрим, насколько хорошо каждый атрибут разделял датасет.



Теперь смоделируем связь между параметрами, показавшими лучшую степень значимости (‘alcohol’, ‘volatile acidity’, ‘free sulfur dioxide’, ‘pH’, ‘residual sugar’) и ‘quality’.



К сожалению, мы видим, что случайный лес демонстрирует тенденцию к переподгонке под тренировочные данные.



**Выводы:**

На практике были изучены три алгоритма регрессии.

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 все модели показали плохие показатели, лучше всего проявил себя случайный лес (на тренировочных данных). Это объясняется тем, что даже у атрибутов с максимальным уровнем корреляции были слабо заметны зависимости для целевого атрибута.