Московский авиационный институт

(национальный исследовательский университет)

Институт №8 «Информационные технологии и прикладная математика»

Лабораторная работа №1 по искусственному интеллекту

«Данные и проблемы с ними»

6 семестр

Студент: Стифеев Евгений Михайлович

Группа: М8О-306Б-17

Руководитель: Ахмед Самир Халид

Дата: 13.04.20

Москва, 2020

Условие

Необходимо сформировать два набора данных для приложений машинного обучения. Первый датасет должен представлять из себя табличный набор данных для задачи классификации. Второй датасет должен быть отличен от первого, и может представлять из себя набор изображений, корпус документов, другой табличный датасет или датасет из соревнования Kaggle, предназначенный для решения интересующей вас задачи машинного обучения. Необходимо провести анализ обоих наборов данных, поставить решаемую вами задачу, определить признаки необходимые для решения задачи, в случае необходимости заняться генерацией новых признаков, устранением проблем в данных, визуализировать распределение и зависимость целевого признака от выбранных признаков. В отчете описать все проблемы, с которыми вы столкнулись и выбранные подходы к их решению.

Датасет mushrooms

Данный датасет представляет собой табличный датасет, состоящий из 8124 записей и 23 атрибута. Загрузив его в Pandas-объект DataFrame и вызвав для него (объекта) метод `info()` я обнаружил, что все 23 признака категориальны. Классифицировать грибы я хотел по аттрибуту class $\in \{e = edible, p = poisonous\}$. Это задача бинарной классификации при обучении с учителем.

Все категориальные аттрибуты подлежат отображению в числа. Далее, я привожу описание, как я разобрался с каждым из них:

1. class:
   Тут всё просто: два значения – отображаем в $\{0, 1\}$. Подобные отображения я делал, используя метод `map()`.
2. cap-shape:
   Множество значений этого признака такого: $bell = b, conical = c, convex = x, flat = f, knobbed = k, sunken = s$. Я решил, что такой признак, можно закодировать цифрами таким образом:
   - $sunken = 0$,
   - $flat = 1$,
   - $convex = 2$,
   - $knobbed = 3$,
   - $conical = 3.3$,
   - $bell = 3.5$,

   по смыслу этих признаков: от утопленной формы ($sunken$) до $bell$ (в виде колокола).
3. cap-surface:
   Аналогично:
   - $smooth = 0$,
   - $fibrous = 1$,
   - $scaly = 2$,
   - $grooves = 3$.
4. cap-color:
   Это аттрибут можно закодировать, пользуясь RGB-таблицей цветов. Я создал три новых признака: cap-color-r, cap-color-g, cap-color-b, куда поместил значения компонент цвета.
5. bruises:
   Два значения – отображаем в $\{0, 1\}$.
6. odor:
   Этот признак нельзя соотнести с числом. Я закодировал его числом записей, у которых он есть: например, если число записей со значением $odor = anise$ составляет 100, то кодируем $anise$ этим числом.
7. gill-attachment:
   Два значения – отображаем в $\{0, 1\}$.
8. gill-spacing:
   Два значения – отображаем в $\{0, 1\}$.
9. gill-color:
   Поступаем аналогично 4.

10. gill-size:
    Можно соотнести с числом, подобно 2. и 3.
11. stalk-shape:
    Два значения – отображаем в $\{0, 1\}$.
12. stalk-root:
    Для этого атрибута существует записи с *пропусками*. Я применил распространённый приём кодирования с одним активным состоянием (one-hot encoding).
13. stalk-surface-above-ring и
14. stalk-surface-below-ring:
    Поступил аналогично 2.
15. stalk-color-above-ring и
16. stalk-color-below-ring:
    Поступил аналогично 4.
17. veil-type:
    Оказалось, что все записи имеют единственное значение этого признака, поэтому я его удалил.
18. veil-color:
    См. 4.
19. ring-number:
    См. 2.
20. ring-type:
    См. 6.
21. spore-print-color:
    См. 4.
22. population:
    См. 2.
23. habitat:
    См. 12.

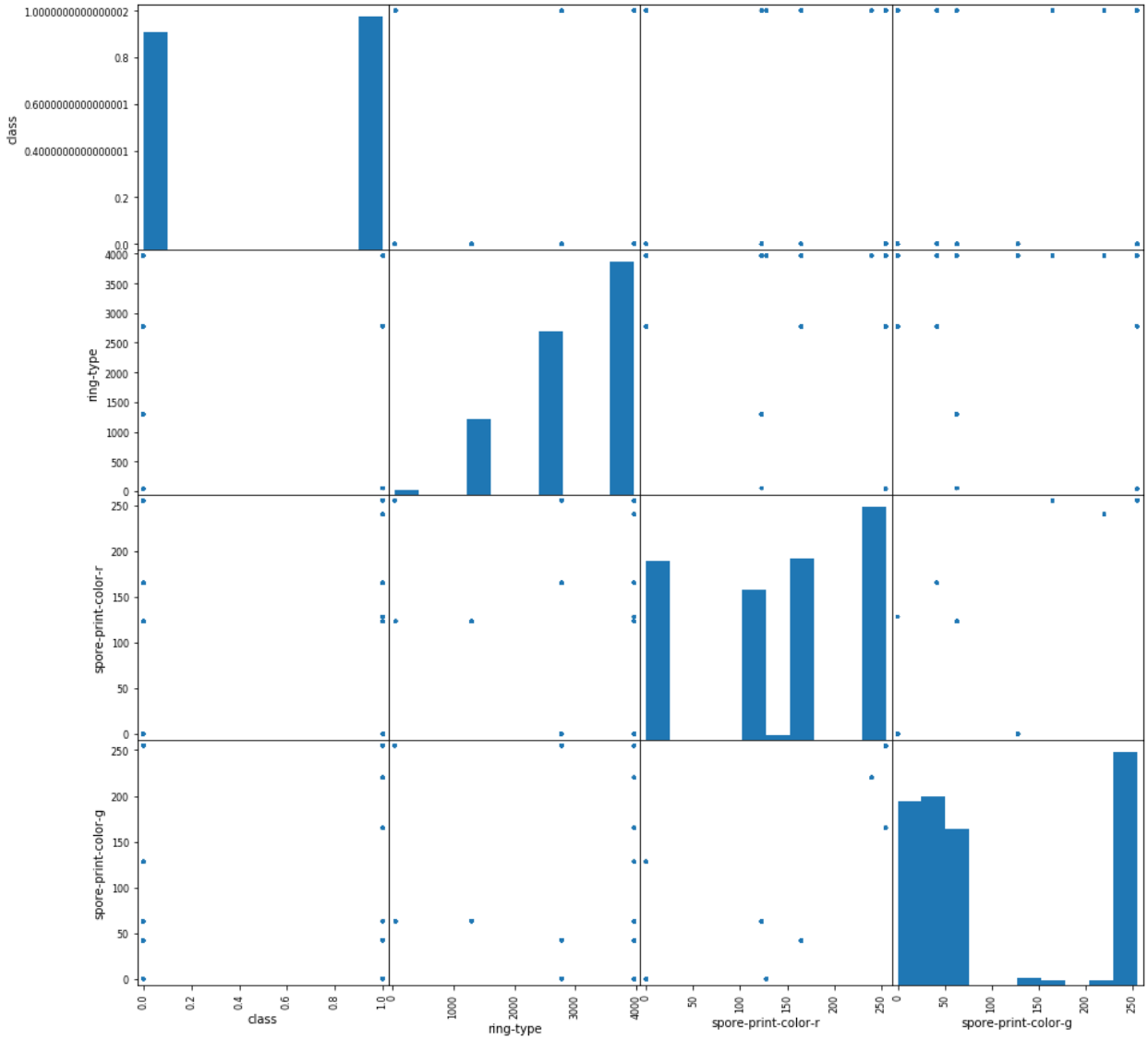
Далее, я построил гистограммы, которые показали мне тенденцию к редкости многих значений моих записей, что говорит о том, что можно исключить некоторые из атрибутов.

После я вычислил стандартный коэффициент корреляции между атрибутом class и всеми остальными. Оказалось, что на class линейным образом хорошо ($|r| \geq 0.3$) влияют 9 атрибутов:

- bruises
- odor
- gill-spacing
- gill-size
- stalk-surface-above-ring
- ring-type
- spore-print-color-r
- spore-print-color-g
- habitat=p

Данные атрибуты я решил оставить, а остальные пока исключить.

Внимательно посмотрев на графическую зависимость:



я заметил, что наличие некоторых признаков *почти наверное* говорит о принадлежности к одному из двух классов.

Далее, я решил поэкспериментировать и добавить атрибут, отвечающий за среднее значение цвета всего гриба, но оказалось, что этот атрибут хуже влияет на класс нежели значения цвета его частей.

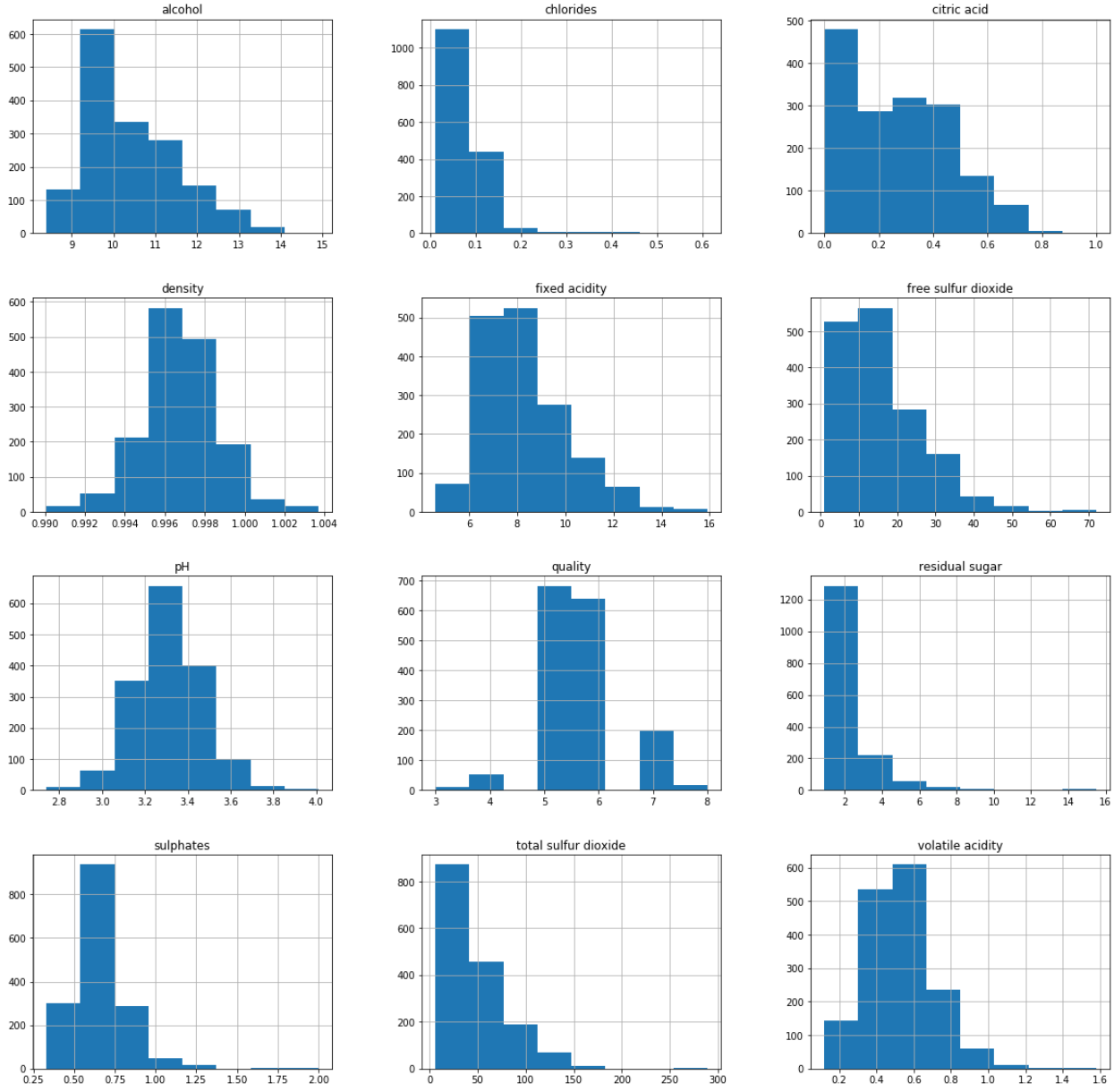
В конце я провёл разделение датасета на тренировочной и выборочный посредством метода `StratifiedShuffleSplit()` фреймворка Scikit-learn для обеспечения репрезентативности выборки по атрибуту ring-type и выполнил нормировку методом минимакса из того же фреймворка.

Датасет winequality-red

Датасет представляет собой табличный датасет без пропусков и категориальных признаков и содержит 1599 записей с 12 атрибутами о красных винах. Я решил предсказывать числовой атрибут density.

Посмотрев на гистограммы, я обнаружил *хорошую* тенденцию многих атрибутов к колоколо-образности, однако некоторые из них подвержены проблеме *медленно убывающих хвостов:* они простираются гораздо дальше вправо от медианы, чем

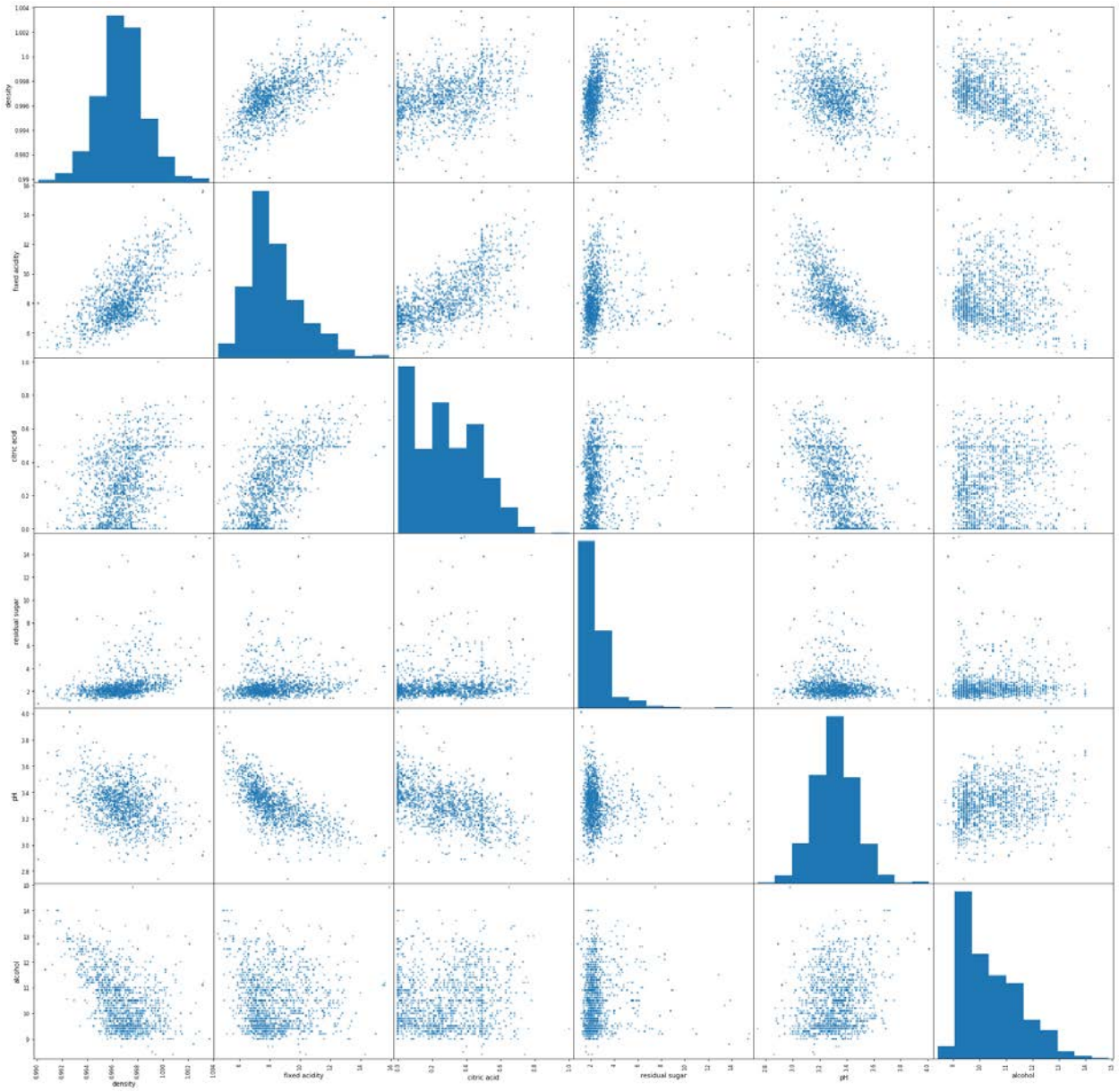
влево. Это может несколько затруднить некоторым алгоритмам МО выявлять паттерны.



Вычислив коэффициент корреляции Пирсона, я обнаружил, что атрибуты:

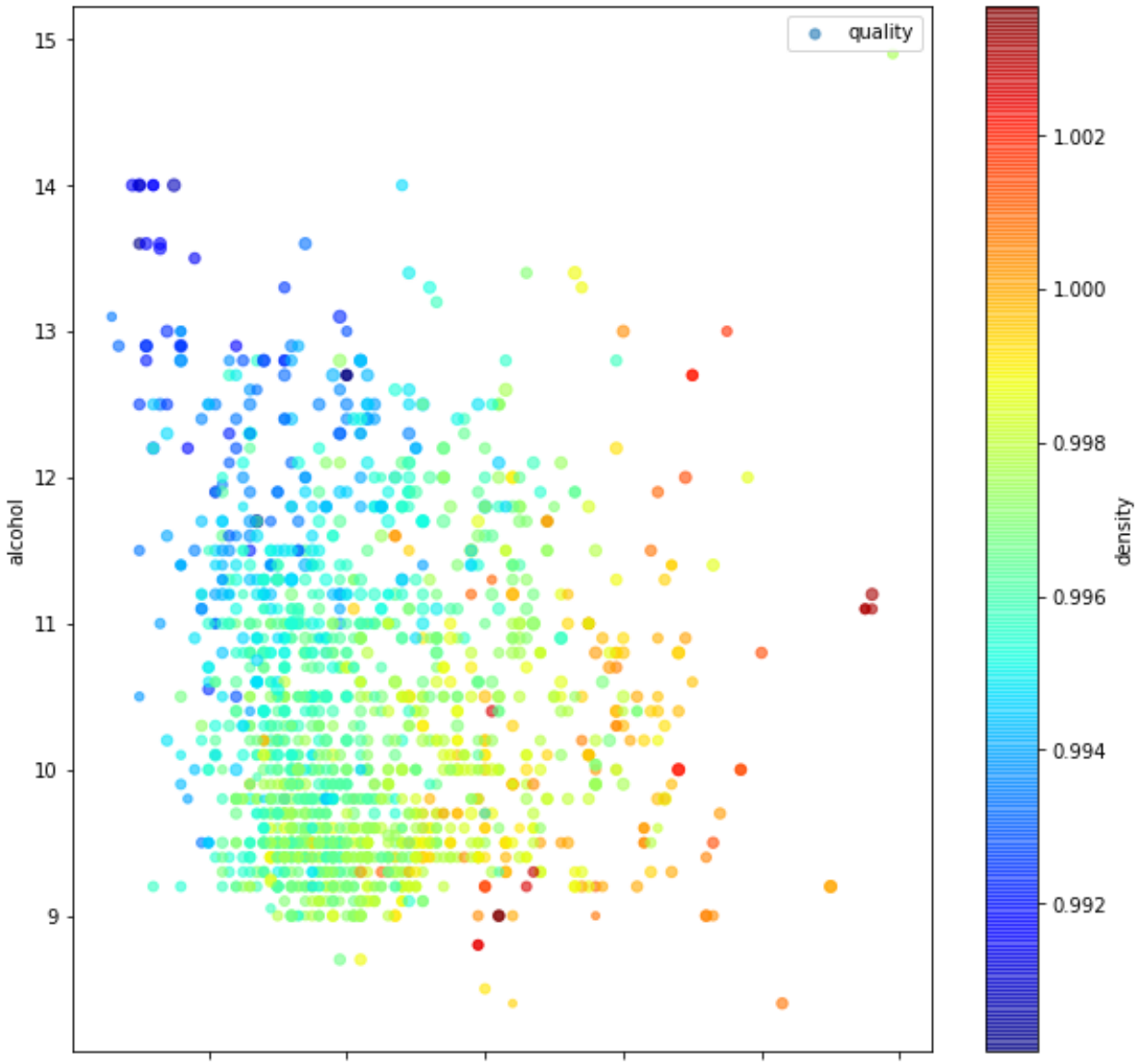
- density
- fixed acidity
- citric acid
- residual sugar
- pH
- alcohol

имеют высокое значение по модулю.



Также я проверил атрибуты citric acid и fixed acidity, total sulfur dioxide и free sulfur dioxide на взаимозаменяемость тем же способом. Оказалось, что нельзя оставить только один из них, о чём мне сообщил тот же коэффициент.

Вооружившись знанием о коэффициенте, я построил график зависимости fixed acidity от alcohol. Цвет для точек я выбрал в зависимости от предсказываемого значения density и размер в соответствии с quality. Получил график:



на котором прослеживается тенденция к «сбору» точек с одинаковыми цветами в группы, также заметно, что точки с малыми радиусами часто разбросаны вдоль границы сосредоточения «пятна».

В конце я отделил тестовую выборку от тренировочной, посредством стратификации по атрибуту quality и выполнил масштабирование по минимаксу.

Выводы

Реальные данные имеют ряд проблем, с которыми приходится сталкиваться программистам, прежде чем приступить к обучению своих моделей. Решение этих проблем зачастую нетривиально и требует некоторой «смекалки», однако существует ряд стандартных подходов к их решению.