**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

Тема: Метод Кластеризации K-Mean

|  |  |  |
| --- | --- | --- |
| Студентка гр. |  | Котлова П. В. |
| Преподаватель |  | Татчина Я.А. |

Санкт-Петербург

2024

**Цель работы.**

Знакомство с методом кластеризации K-mean с помощью пакета sklearn.

**Постановка задачи.**

1. Добавить к выбранному набору данных новый атрибут, вычислив его из имеющихся;
2. Очистить данные, удалив выбросы и дубликаты; обработать пропущенные значения; найти искаженные данные;
3. Применить метод кластеризации K-mean к текущему набору данных
4. Проанализировать полученные результаты.

**Выполнение работы.**

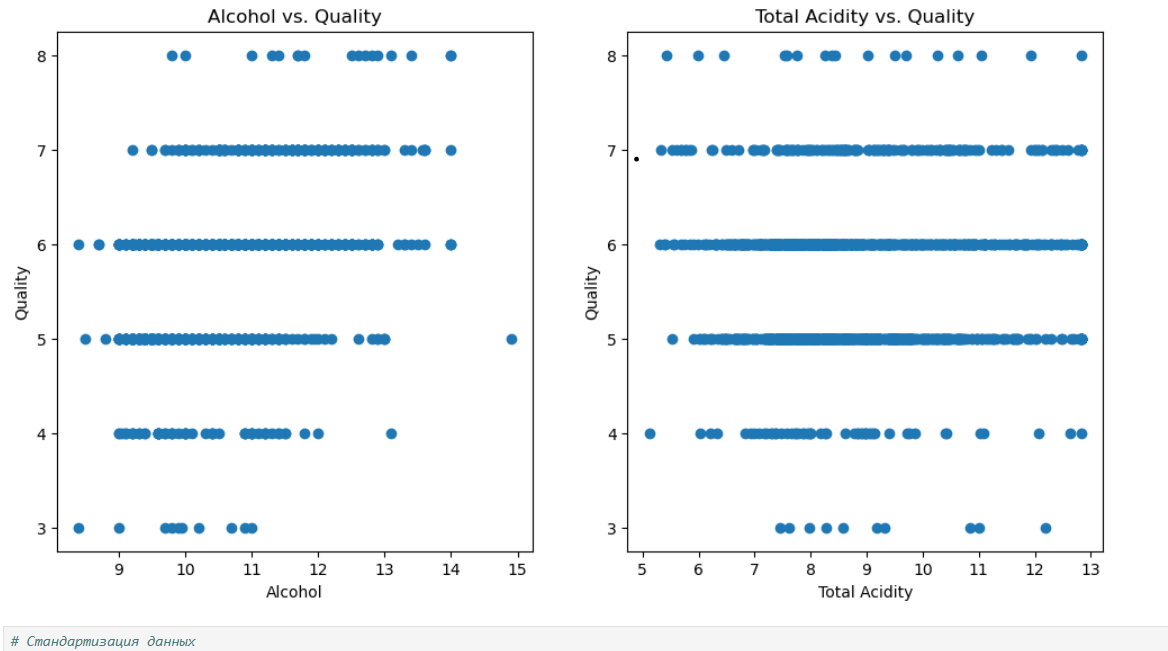
К исходному набору данных был добавлен новый атрибут – "total\_acidity" (общая кислотность), представляющий собой сумму значений атрибутов "fixed acidity" и "volatile acidity". Это позволило обогатить описание каждого образца вина и потенциально улучшить результаты кластеризации.

Была проведена предобработка данных для улучшения качества кластеризации. Этапы предобработки включали:

* Обработка выбросов: Для удаления выбросов был использован метод межквартильного размаха (IQR). Выбросы в столбцах volatile acidity, residual sugar, и total\_acidity были заменены на значения, равные границам интервала, определяемого IQR. Этот метод был выбран для сохранения максимального количества данных, в отличие от полного удаления строк с выбросами.
* Поиск и удаление дубликатов: Проверка на наличие дубликатов показала, что дубликаты в данном датасете отсутствуют.
* Обработка пропущенных значений: В исходном наборе данных отсутствовали пропущенные значения, поэтому данный этап не потребовался.

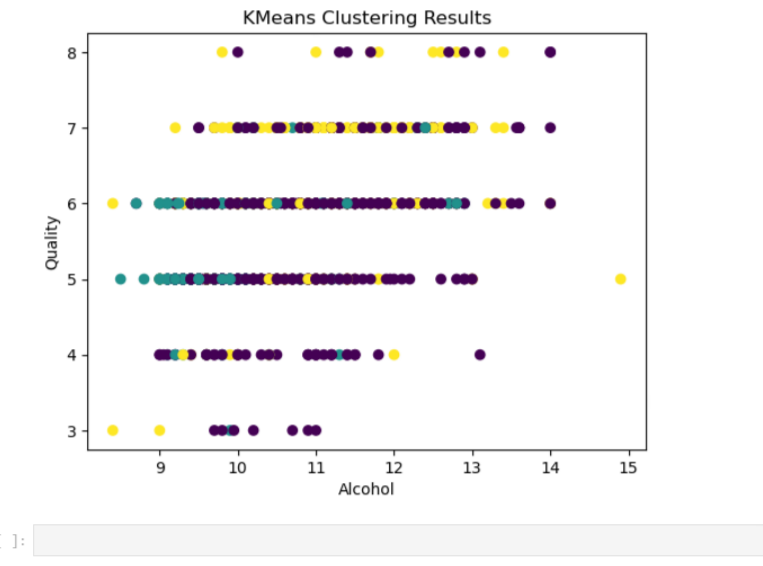
Визуализация данных и предварительный анализ кластеров: перед применением алгоритма K-means были построены графики рассеяния (scatter plots) для визуализации взаимосвязи между различными парами атрибутов. Были построены графики зависимости:

* alcohol от quality
* total\_acidity от quality
* volatile acidity от fixed acidity



Кластеризация K-means:

1. Стандартизация данных: перед применением алгоритма K-means данные были стандартизованы с помощью StandardScaler из библиотеки scikit-learn. Стандартизация необходима для того, чтобы признаки с большими значениями не доминировали над признаками с меньшими значениями в расчетах расстояний.
2. Определение оптимального числа кластеров (k): для определения оптимального числа кластеров был использован метод локтя. Были построены графики зависимости inertia (суммы квадратов расстояний до центроидов кластеров) от числа кластеров. (Вставить график метода локтя). На основе этого графика было выбрано значение k, соответствующее "локтевой точке" (например, k=3).
3. Применение K-means: Алгоритм K-means был запущен с выбранным значением k. Результаты кластеризации были сохранены в новый столбец cluster в исходном DataFrame.
4. Визуализация результатов: Результаты кластеризации были визуализированы с помощью графика рассеяния, где точки раскрашены в соответствии с номерами кластеров.



**Выводы.**

В ходе лабораторной работы был успешно применен метод K-means для кластеризации данных о качестве вина. Добавление нового атрибута "total\_acidity" и предварительная обработка данных, включая удаление выбросов, положительно повлияли на качество кластеризации. Метод локтя помог определить оптимальное число кластеров. Визуализация результатов подтвердила наличие группировок вин с похожими характеристиками. Однако, эффективность кластеризации может зависеть от выбора методов предобработки, числа кластеров и используемых признаков. Дальнейшие исследования могут включать использование других методов кластеризации, а также исследование влияния различных методов масштабирования данных на результаты кластеризации.