**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №1**

**по дисциплине «Машинное обучение»**

Тема: Исследование набора данных

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

Санкт-Петербург

2024

**Цель работы.**

Выбор набора данных для дальнейшего исследования алгоритмов кластеризации и классификации. Получение практических навыков работы с инструментами анализа и визуализации данных.

**Постановка задачи.**

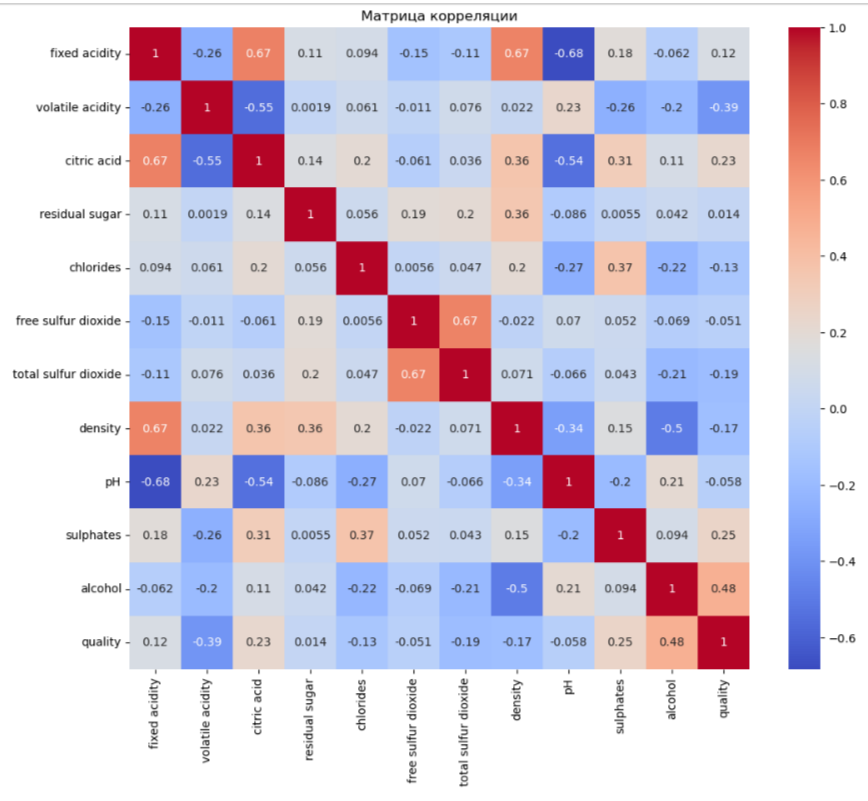
1. Создать Jupyter Notebook
2. Выбрать исследуемый набор данных из предложенных источников
3. Описать выбранный набор данных, а именно:
   1. Предметную область, источник данных, характер данных;
   2. Атрибуты, их тип, что они обозначают;
   3. Задачу анализу.
4. Для каждого атрибута:
   1. Определить среднее значение, СКО;
   2. Построить гистограмму распределения значений, определить наличие выбросов;
   3. Определить наличие пропущенных значений и их количество;
   4. Предложить вариант обработки пропущенных значений.
5. Определить:
   1. Высоко коррелированные атрибуты, характер корреляции;
   2. Не коррелированные атрибуты;
   3. Графики рассеивания.
6. Проанализировать полученные результаты.

**Выполнение работы.**

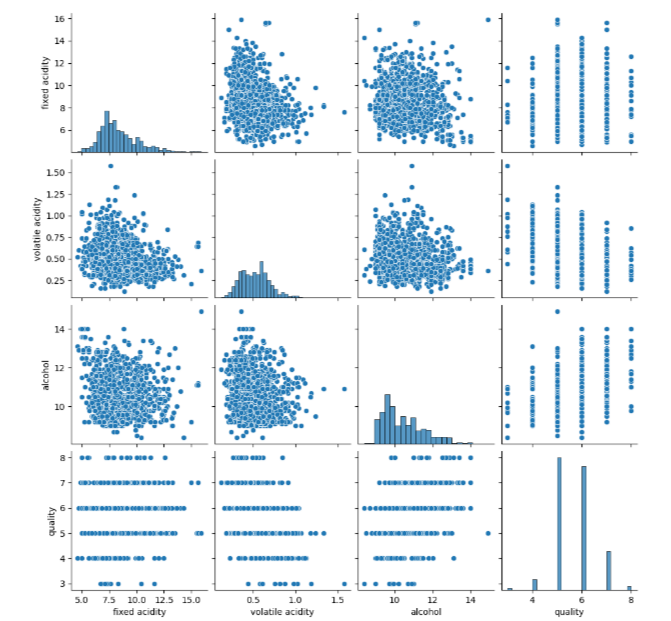
В качестве исходного набора данных был выбран публично доступный датасет "Wine Quality Data Set" из репозитория UCI Machine Learning Repository, доступный на Kaggle. Этот датасет содержит информацию о физико-химических свойствах красного вина и оценке его качества по балльной шкале. Было проведено описание датасета, включая предметную область (качество красного вина), источник данных (UCI Machine Learning Repository), характер данных (реальные данные), описание каждого атрибута (тип данных и семантическое значение) и описание задачи анализа (предсказание качества вина на основе его физико-химических свойств).

Была выполнена предобработка данных, включающая загрузку данных с помощью библиотеки Pandas. Для каждого атрибута (кроме целевого – "quality") были рассчитаны основные статистические характеристики: среднее значение и стандартное отклонение. Были построены гистограммы распределения значений для каждого атрибута с целью визуализации распределения данных и выявления возможных выбросов. В данном датасете пропущенных значений не обнаружено.

Для выявления корреляционных связей между атрибутами была рассчитана матрица корреляции с помощью метода corr() библиотеки Pandas. Полученная матрица была визуализирована в виде тепловой карты с помощью библиотеки Seaborn.



*Рис. 1. Тепловая карта корреляции*

Для наглядности были также построены графики рассеяния (pairplot) для некоторых пар атрибутов. На основе анализа матрицы корреляции и графиков рассеяния были выявлены высококоррелированные и некоррелированные атрибуты.  


*Рис. 2. Графики рассеивания*

Для обнаружения выбросов в каждом атрибуте был применен метод межквартильного размаха (IQR). Количество выбросов для каждого атрибута было определено и выведено в отчет.

**Выводы.**

В ходе работы были выбран набор данных для исследования. Получены практические навыки работы с анализом и визуализацией данных с помощью инструментов numpy, pandas, matplotlib, seaborn.