Изображение выглядит как эмблема, символ, герб, нашивка

Автоматически созданное описание

|  |
| --- |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА - Российский технологический университет»**  **РТУ МИРЭА** |

**Институт** Информационных Технологий

**Кафедра** Вычислительной Техники

**ПРАКТИЧЕСКАЯ РАБОТА №3**

**по дисциплине**

**«Проектирование интеллектуальных систем (часть 1/2)»**

Студент группы:ИКБО-04-22 \_\_Кликушин В.И.\_ *(Ф. И.О. студента)*

Преподаватель \_\_ Холмогоров В.В.\_\_

*(Ф.И.О. преподавателя)*

|  |  |  |
| --- | --- | --- |
|  |  |  |

Москва 2025

# СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc198847799)

[1 ПОСТАНОВКА ЗАДАЧИ 4](#_Toc198847800)

[2 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 5](#_Toc198847801)

[2.1 Метод локтя 5](#_Toc198847802)

[2.2 Силуэтный анализ 7](#_Toc198847803)

[2.3 Алгоритм k-means 7](#_Toc198847804)

[2.4 Оценка качества кластеризации 8](#_Toc198847805)

[2.5 Алгоритм DBSCAN 12](#_Toc198847806)

[3 ДОКУМЕНТАЦИЯ К ДАННЫМ 14](#_Toc198847807)

[3.1 Описание предметной области 14](#_Toc198847808)

[3.2 Анализ данных 14](#_Toc198847809)

[3.3 Предобработка данных 19](#_Toc198847810)

[4 ПРАКТИЧЕСКАЯ ЧАСТЬ 22](#_Toc198847811)

[4.1 Метод локтя 22](#_Toc198847812)

[4.2 Силуэтный анализ 23](#_Toc198847813)

[4.3 Алгоритм k-means 25](#_Toc198847814)

[4.4 Алгоритм DBSCAN 28](#_Toc198847815)

[ЗАКЛЮЧЕНИЕ 31](#_Toc198847816)

[СПИСОК ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ 32](#_Toc198847817)

[ПРИЛОЖЕНИЯ 33](#_Toc198847818)

# ВВЕДЕНИЕ

# 1 ПОСТАНОВКА ЗАДАЧИ

Цель работы: приобрести навыки классификации как инструмента категориального и предикативного анализа данных с контролируемым обучением.

Задачи: определить предметную область решаемой задачи, найти или сгенерировать набор данных для выбранной задачи, проведя предварительную предобработку и подготовку данных, выбрать модель классификации, провести её обучение и тестирование, определить качество модели с помощью метрик потерь, изучить алгоритмы классификации, написать программный код для реализации указанных алгоритмов, провести бинарную и/или многоклассовую классификацию данных подготовленного набора.

# 2 ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

## 2.1 Алгоритм KNN

## 2.2 Метрики классификации

# 3 ДОКУМЕНТАЦИЯ К ДАННЫМ

## 3.1 Описание предметной области

В качестве набора данных выбран широко известный Wine Dataset (данные о сортах итальянского вина, полученных в результате химического анализа образцов), хранящийся в открытом репозитории UCI Machine Learning Repository и доступный в библиотеке scikit-learn. Датасет содержит результаты аналитических измерений для красных и белых вин, произведённых тремя различными винодельческими культурами из региона северо-западной Италии. Целью сбора таких данных является изучение химических и физических свойств вина, которые позволяют не только классифицировать образцы по сорту, но и делать выводы о качестве продукта и возможных дефектах при его изготовлении.

## 3.2 Анализ данных

В исходных данных представлено 178 образцов вин. Каждый образец соответствует конкретной бутылке вина одного из трех сортов (классов), обозначенных как «Class 0», «Class 1» и «Class 2». Эти сорта соответствуют коммерческой классификации винодельческих культур:

* class 0 – сорт «Barolo»;
* class 1 – сорт «Grignolino»;
* class 2 – сорт «Barbera».

Для предобработки и анализа датасета в файле dataset\_manager.py написан класс DatasetManager. Содержание файла dataset\_manager.py представлено в Приложении А.

Исходные данные представлены в виде таблицы с 14 колонками (13 признаков и целевая метка класса). Каждый объект содержит 13 числовых признаков, характеризующих физико-химическое состояние образца, и целевое значение «target» – метку сорта (0, 1 или 2). Колонки с признаками включают:

1. Alcohol (спиртовая крепость): концентрация этанола в вине.
2. Malic acid (яблочная кислота, г/дм3): остаточное содержание яблочной кислоты после ферментации.
3. Ash (зола, г/дм3): количество минеральных веществ, оставшихся после сжигания пробы.
4. Alcalinity of ash (щелочность золы): показатель щелочности зольных составляющих.
5. Magnesium (магний, мг/дм3): концентрация ионов магния.
6. Total phenols (общие фенолы, г/дм3): суммарное содержание фенольных соединений, влияющих на цвет и вкус.
7. Flavanoids (флавоноиды, оптическая плотность): концентрация флавоноидных соединений, отвечающих за танинность и антиоксидантные свойства.
8. Nonflavanoid phenols (нефлавоноидные фенолы, оптическая плотность): другие фенольные соединения, не относящиеся к флавоноидам.
9. Proanthocyanins (проантоцианидины, оптическая плотность): полифенолы, влияющие на структуру и долговечность вина.
10. Color intensity (интенсивность цвета, оптическая плотность): показатель насыщенности цвета, получаемый спектрофотометрически.
11. Hue (оттенок): отношение определённых спектральных поглощений, характеризующее оттенок красного/фиолетового.
12. OD280/OD315 of diluted wines (отношение оптической плотности при 280 нм и 315 нм): индикатор содержания фенольных соединений при разведении.
13. Proline (пролин, мг/дм3): аминокислота, одна из наиболее представленных в вине, влияющая на вкус и аромат.

Описательная статистика признаков отображена на Рисунке 3.2.1.

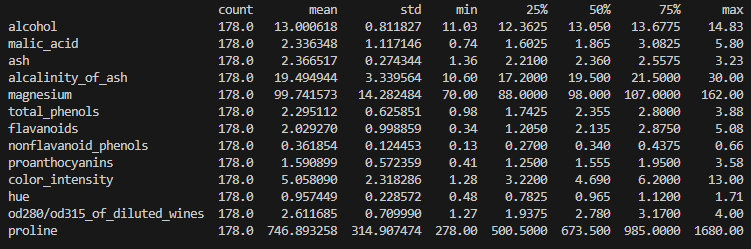


Рисунок 3.2.1 – Описательная статистика признаков

В таблице приведены стандартные метрики для каждого из 13 признаков: среднее значение (mean), стандартное отклонение (std), минимум (min), первые и третьи квартили (25% и 75%), медиана (50%) и максимум (max).

Описательные статистики показывают, что часть признаков (например, Alcohol, Magnesium) распределены относительно компактно, в то время как другие признаки (Malic acid, Proline, Proanthocyanins) обладают более широким разбросом и выраженной скошенностью. Для корректной классификации потребуется стандартизация и, возможно, дополнительная обработка выбросов.

Распределение вин по классам представлено на Рисунке 3.2.2.

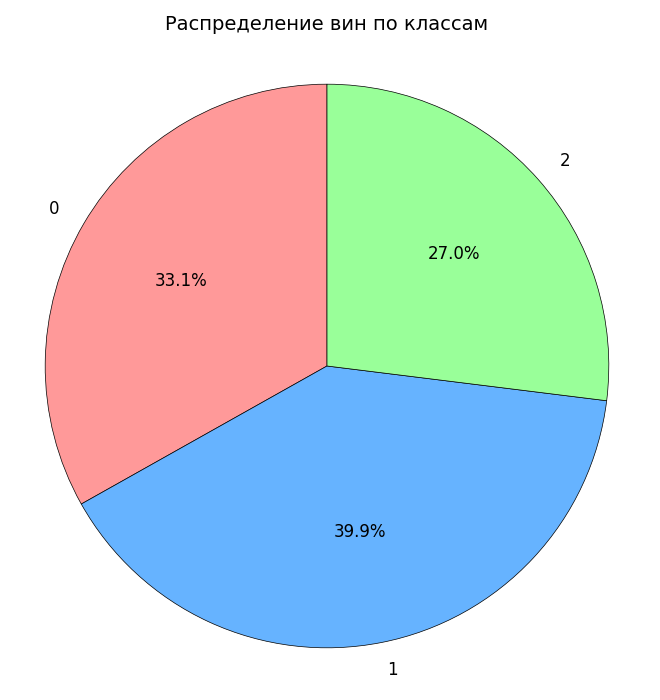


Рисунок 3.2.2 – Распределение вин по классам

Таким образом, класс 1 представлен наиболее обильно, а класс 2 – наименее. Несмотря на небольшую дисбалансировку, соотношение классов достаточно близко к равномерному.

Гистограммы распределений признаков представлены на Рисунке 3.2.3.

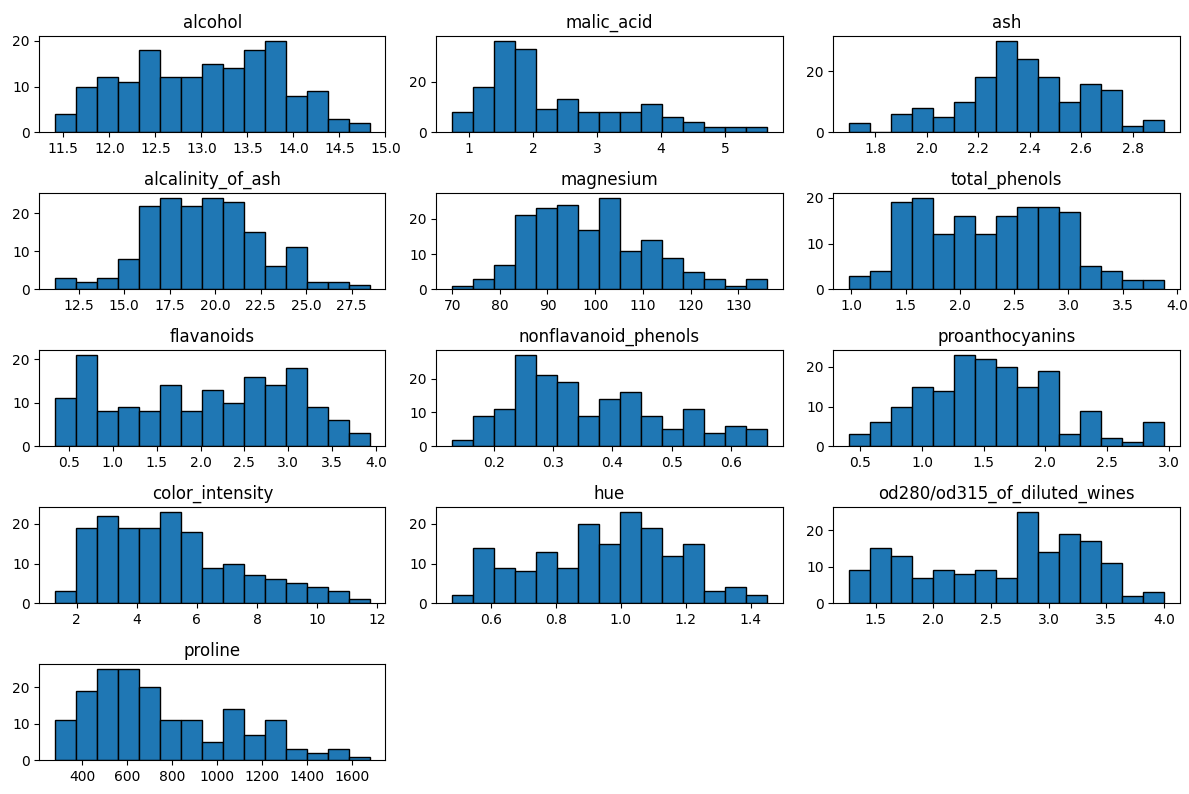


Рисунок 3.2.3 - Гистограммы распределений признаков

Практически все признаки имеют выраженную скошенность и отдельные выбросы. Это подчёркивает необходимость обработки экстремумов и обязательное масштабирование перед кластеризацией.

Матрица рассеяния для первых семи признаков представлена на Рисунке 3.2.4.

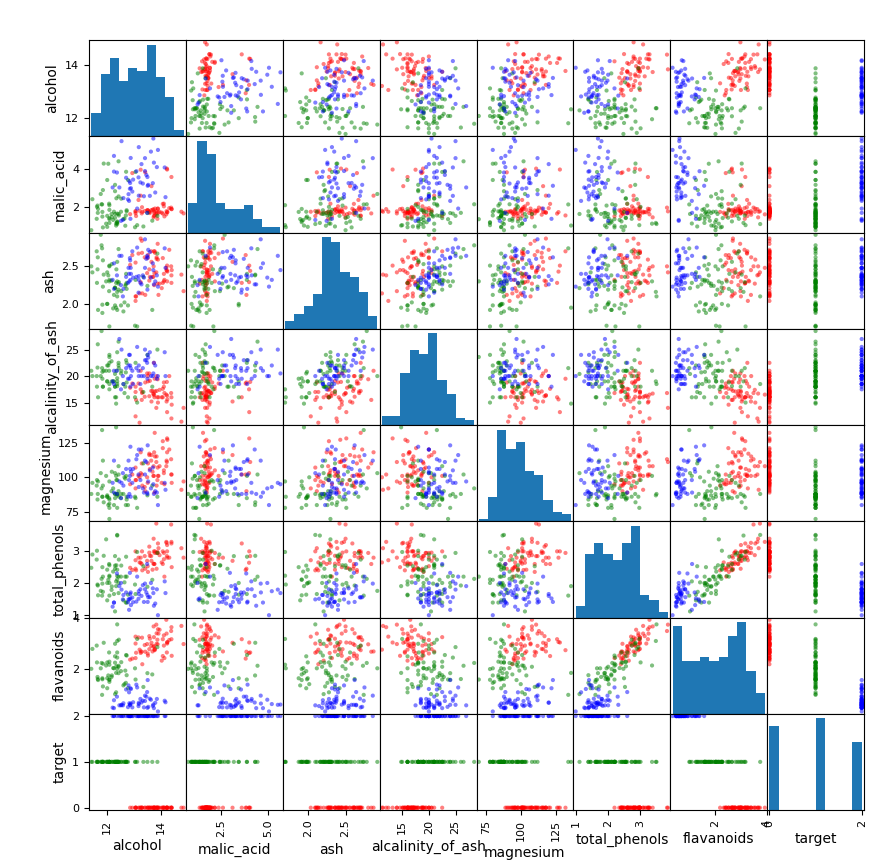


Рисунок 3.2.4 – Матрица рассеяния для первых семи признаков

На диагонали расположены гистограммы отдельных признаков (те же, что частично были на Рисунке 3.2.3, но только для первых семи). В ячейках показаны облака точек для пар признаков. Некоторые пары признаков обеспечивают достаточно чёткое различие классов. Между признаками Total phenols и Flavanoids прослеживается сильная корреляция.

Матрица корреляции признаков представлена на Рисунке 3.2.5.

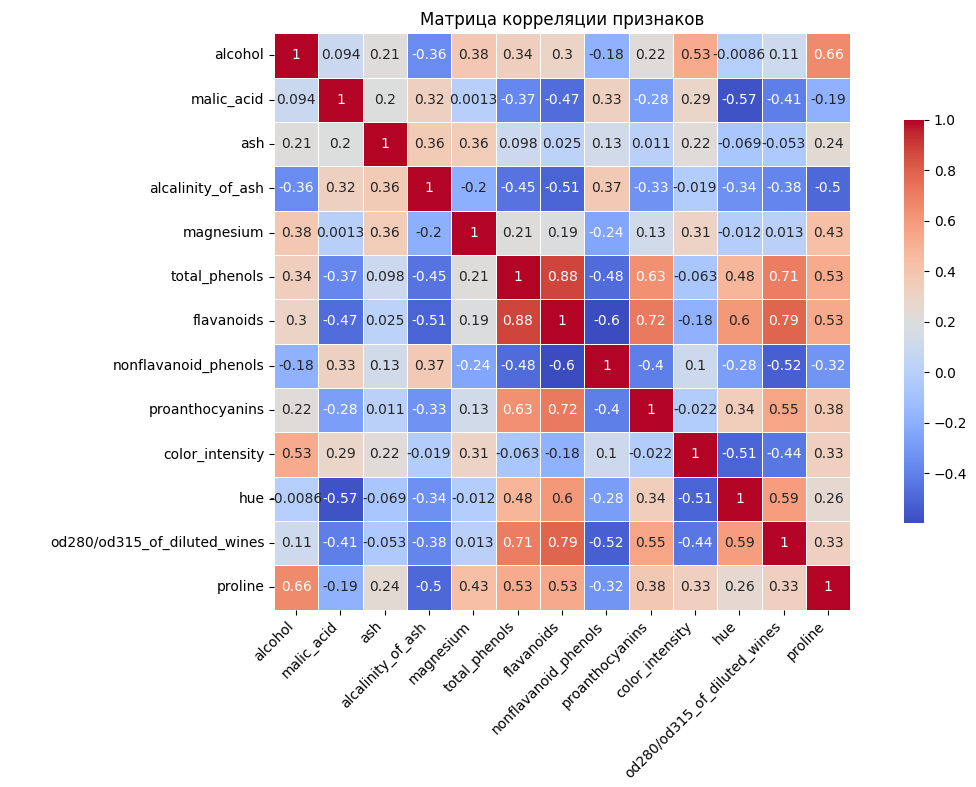


Рисунок 3.2.5 – Матрица корреляции признаков

Высокая корреляция между некоторыми признаками (Flavanoids и Total phenols) позволяет сократить размерность данных, исключив дублирующиеся признаки.

## 3.3 Предобработка данных

Реализованы следующие этапы предобработки данных:

* удаление дубликатов строк;
* удаление выбросов по Z-оценке;
* масштабирование признаков (StandardScaler);
* выбор и удаление избыточных признаков.

Удаление дубликатов строк подразумевает обнаружение и удаление полностью идентичных записей (строк) в исходном датасете. Под «идентичностью» понимается совпадение всех значений по всем признакам. В рассматриваемом датасете дубликатов не обнаружено.

Выбросы (аномальные значения) — это отдельные объекты, сильно отклоняющиеся от общей «массы» точек. Чаще всего они встречаются в признаках с широким диапазоном. Один из способов формального выявления выбросов — использовать Z-оценку.

Для каждого значения признака в образце рассчитывается величина по Формуле 3.3.1.

, (3.3.1)

где – среднее отклонение признака по всем образцам;

– стандартное отклонение признака по всем образцам.

Образец считается выбросом, если хотя бы один признак имеет .

В качестве порогового значения выбран = 3. Таким образом, удалены все образцы, в которых хотя бы один признак отклонён от среднего более чем на три стандартных отклонения. В результате удаления выбросов по Z-оценке было исключено десять строк из исходного набора данных.

Скалирование признаков — это приведение всех измеряемых величин в единый единичный масштаб. В Wine Dataset признаки измеряются в разных физических и химических единицах. Без масштабирования признаки с большим диапазоном «будут весить» значительно больше при классификации, чем признаки с узким диапазоном.

StandardScaler — один из наиболее распространённых способов стандартизации. Для каждого признака вычисляется среднее и стандартное отклонение . Каждое значение преобразуется по Формуле 3.3.2.

, (3.3.2)

В результате стандартизированный признак ​ имеет среднее 0 и стандартное отклонение 1.

На Рисунке 3.2.5 показано, что коэффициент корреляции между «Total phenols» и «Flavanoids» составляет примерно 0.86. Это значит, что эти два признака фактически несут очень близкую информацию о составе вина. Когда признаки столь сильно коррелированы, они считаются практически линейно зависимыми: наличие одного позволяет почти однозначно восстановить второй. Поэтому признак «Total phenols» исключен из набора признаков, а соответствующий столбец в датасете удален.

Данные разделены на обучающую и тестовую выборки в соотношении 80:20. Обучающая выборка содержит 134 образца, а тестовая выборка – 34 образца.

Предобработанная обучающая выборка представлена на Рисунке 3.2.6.

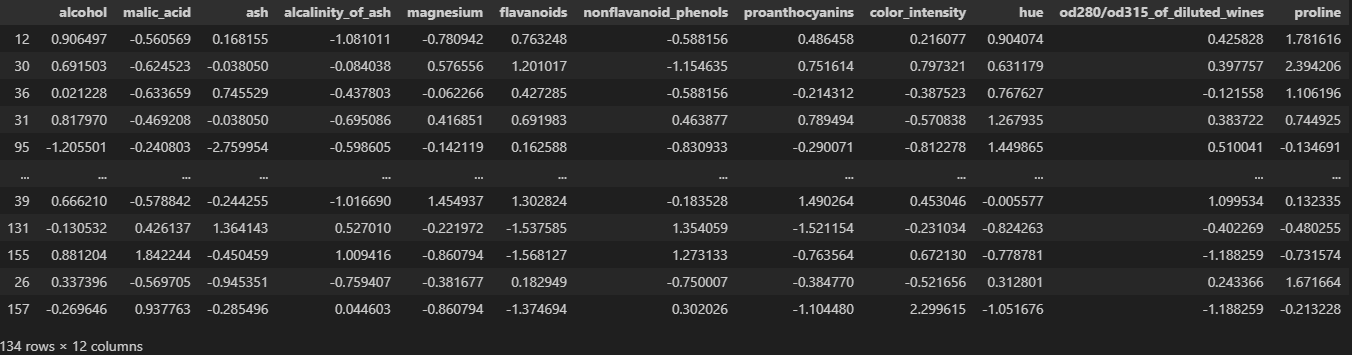


Рисунок 3.2.6 – Предобработанная обучающая выборка

На данном этапе полученный очищенный набор признаков может использоваться в алгоритмах классификации.

# 4 ПРАКТИЧЕСКАЯ ЧАСТЬ

## 4.1 Алгоритм KNN

# ЗАКЛЮЧЕНИЕ

# СПИСОК ИНФОРМАЦИОННЫХ ИСТОЧНИКОВ

1. Сорокин, А. Б. Безусловная оптимизация. [Электронный ресурс] : учебно-метод. пособие / А. Б. Сорокин, О. В. Платонова, Л. М. Железняк — М. РТУ МИРЭА , 2020.
2. Сорокин, А. Б. Введение в генетические алгоритмы: теория, расчеты и приложения. [Электронный ресурс] : учебно-метод. пособие / А. Б. Сорокин — М. МИРЭА , 2018.
3. Метод K-ближайших соседей (KNN). Принцип работы, разновидности и реализация с нуля на Python [Электронный ресурс]: Habr. URL: https://habr.com/ru/articles/801885/ (Дата обращения: 16.05.2025).
4. Машинное обучение: Классификация методом KNN. Теория и реализация. С нуля. На чистом Python [Электронный ресурс]: Habr. URL: https://habr.com/ru/articles/866636/ (Дата обращения: 17.05.2025).

# ПРИЛОЖЕНИЯ

Приложение А — Файл dataset\_manager.py для предобработки и анализа датасета.

### Приложение А

Файл dataset\_manager.py для предобработки и анализа датасета

Листинг А – Содержание файла dataset\_manager.py