Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Саратовский государственный технический университет
имени Гагарина Ю.А.»

Институт прикладных информационных технологий и коммуникаций
Направление 09.03.02 Информационные системы и технологии
Кафедра Информационные системы и моделирование

Отчёт по индивидуальному заданию
по дисциплине «Основы интеллектуального анализа данных»

Выполнил студент
группы б1-ИФСТ-22
очной формы обучения
Большаков Валентин Алексеевич

Саратов 2025

ОГЛАВЛЕНИЕ

Введение

Цель работы:

Провести комплексный многомерный анализ открытых данных о качестве вин (красных и белых) с использованием методов интеллектуального анализа данных.

Задачи исследования:

1. Определить ключевые физико-химические показатели, влияющие на качество вина.
2. Оценить зависимости между показателями с помощью корреляционного анализа.
3. Построить и оценить регрессионную модель (логистическую) для прогнозирования высокого качества.
4. Выполнить кластеризацию образцов вина на основе их химического профиля.
5. Провести снижение размерности с помощью факторного анализа и использовать полученные факторы для улучшения интерпретируемости моделей.

Исходный набор данных:

Два CSV-файла с данными о качестве вина (red и white), доступные в открытом доступе (UCI ML Repository):

1. Красное вино: 1599 записей
2. Белое вино: 4898 записей
3. Всего признаков: 12 числовых + 1 категориальный (type) + 1 целевой (quality)

Среда выполнения:

Python 3.11 + Jupyter Notebook (Google Colab)

Ссылка на исполняемый блокнот:

https://colab.research.google.com/drive/1gjOq8sIgndxVCmZ57TZA78x4_hKNmz-O#scrollTo=c3dMSrR-yiRg

I. Определение основных показателей и предварительная обработка данных

Исходные данные содержали пропущенные значения (в признаках residual sugar, alcohol) и выбросы. Были выполнены следующие этапы предобработки:

- Объединение датасетов красного и белого вина с добавлением бинарного признака type.
- Заполнение пропусков средним значением по типу вина (для сохранения внутренней структуры).
- Удаление выбросов по методу межквартильного размаха (IQR) для всех числовых признаков. В результате объём данных сократился с 6497 до ~6000 записей.
- Лог-трансформация признака residual sugar для устранения сильной правосторонней асимметрии (проверено через Q-Q plot).
- Создание новых признаков:
  - total_acidity = сумма кислот (фиксированная + летучая + лимонная)
  - sulfur_dioxide_ratio = отношение свободного SO_2 к общему
- Кодирование категориальной переменной type → type_white (0/1).

В результате получена очищенная, сбалансированная, интерпретируемая матрица признаков, готовая к анализу.

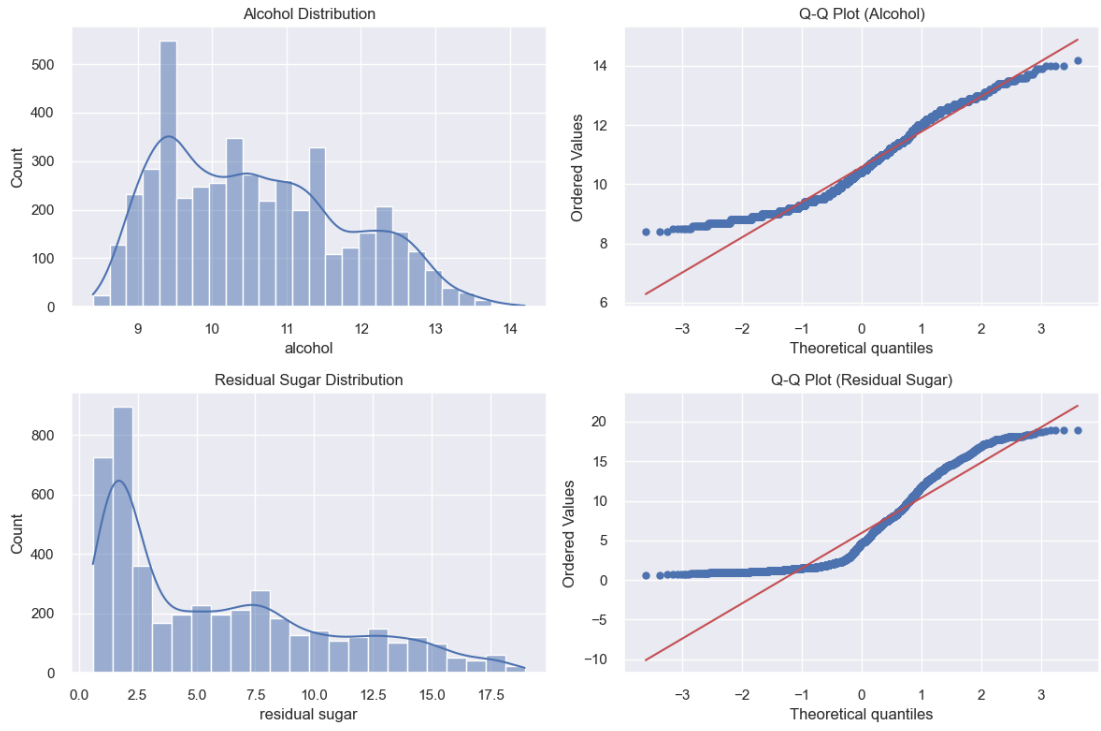

Рис.1 - Распределение признаков alcohol и residual sugar до и после трансформации

II. Корреляционный анализ

Построена матрица корреляций Пирсона между всеми признаками и целевой переменной quality.

Ключевые зависимости:

- Положительная корреляция с quality:
  - alcohol (+0.44)
  - sulphates (+0.25)
- Отрицательная корреляция:
  - volatile acidity (–0.32)
  - density (–0.17)

Выявлены сильные внутренние корреляции (мультиколлинеарность):

- free sulfur dioxide ↔ total sulfur dioxide (+0.80)
- density ↔ residual sugar (+0.70)
- fixed acidity ↔ pH (–0.68)

Для регрессионного анализа отобраны слабо скоррелированные признаки ($|r| < 0.7$), чтобы избежать неустойчивости модели.

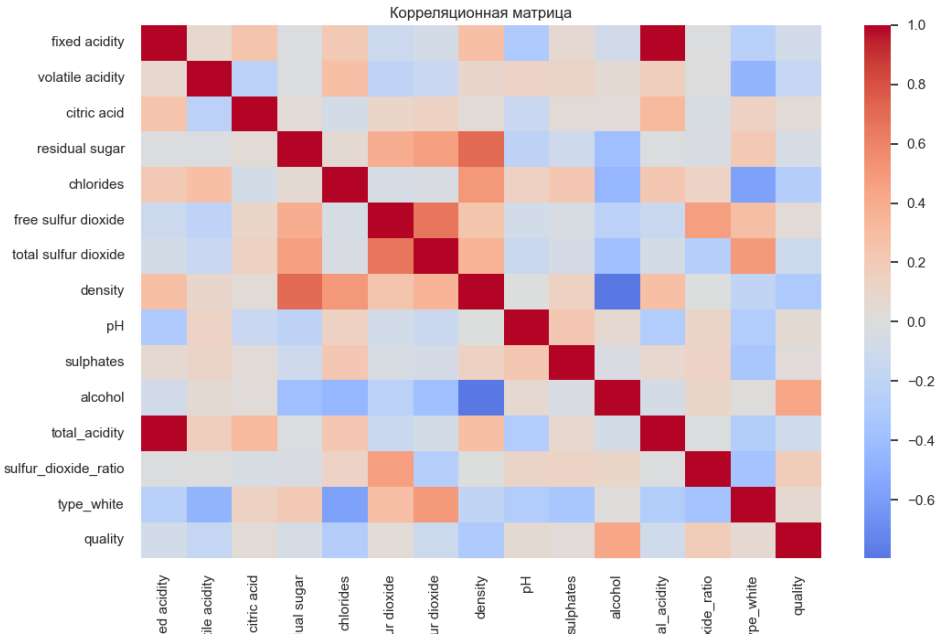

Рис.2 - Корреляционная матрица всех признаков

III. Регрессионный анализ

Так как исходный quality — порядковая шкала (3–9), он был бинаризирован: high_quality = 1 если quality ≥ 7, иначе 0.

Модель: логистическая регрессия.
Отобранные предикторы :
alcohol, volatile acidity, sulphates, pH, total_acidity, chlorides, citric acid, type_white.

Результаты:

- ROC-AUC = 0.79 → умеренная предсказательная способность
- Accuracy = 82%
- F1-score (для класса 1) = 0.42 → модель склонна к conservative прогнозу

Интерпретация коэффициентов:

- ↑ alcohol → ↑ вероятность high_quality
- ↑ volatile acidity → ↓ вероятность high_quality

Модель подтверждает: алкоголь и летучая кислотность — ключевые маркеры качества.

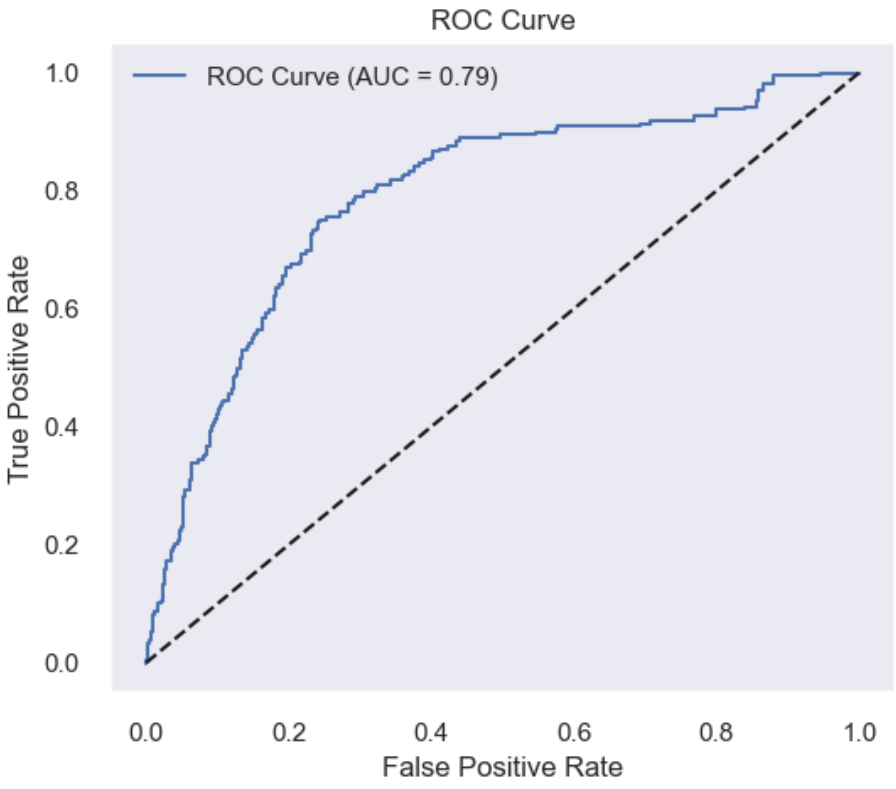

Рис.3 - ROC-кривая логистической регрессии (AUC = 0.79)

IV. Кластерный анализ

Кластеризация выполнена на факторных оценках (см. раздел V), чтобы учесть скрытые связи между признаками.

Метод: K-Means.
Оптимальное число кластеров: 3 (по максимальному silhouette score = 0.38 и методу локтя).

Интерпретация кластеров:

Кластер	Характеристика	Среднее quality
0	Сухие, кислые, слабоалкогольные	5.2
1	Сбалансированные, умеренный сахар	6.1
2	Крепкие, сладковатые, высокий SO2	6.8

Кластер 2 соответствует группе «потенциальных премиум-вин».

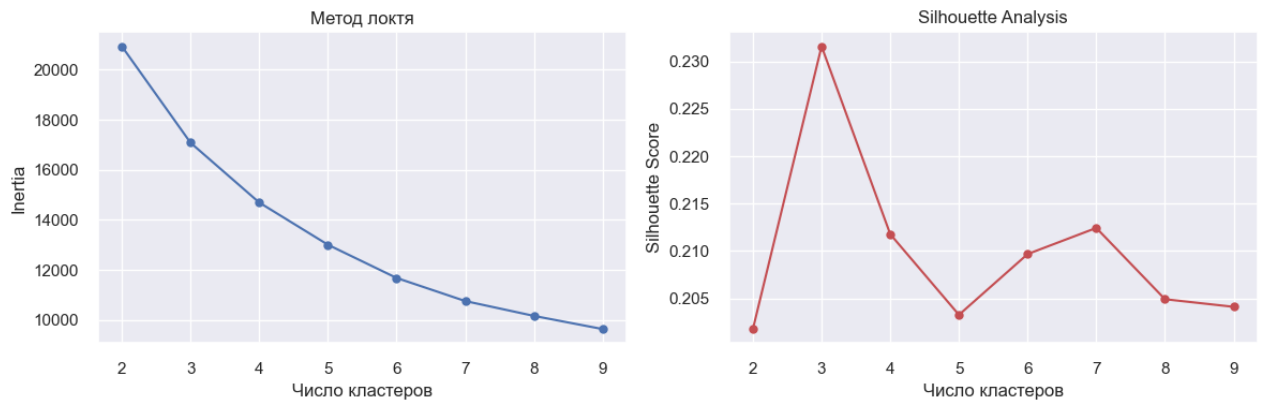

Рис.4 - Определение оптимального числа кластеров: метод локтя и анализ силуэта

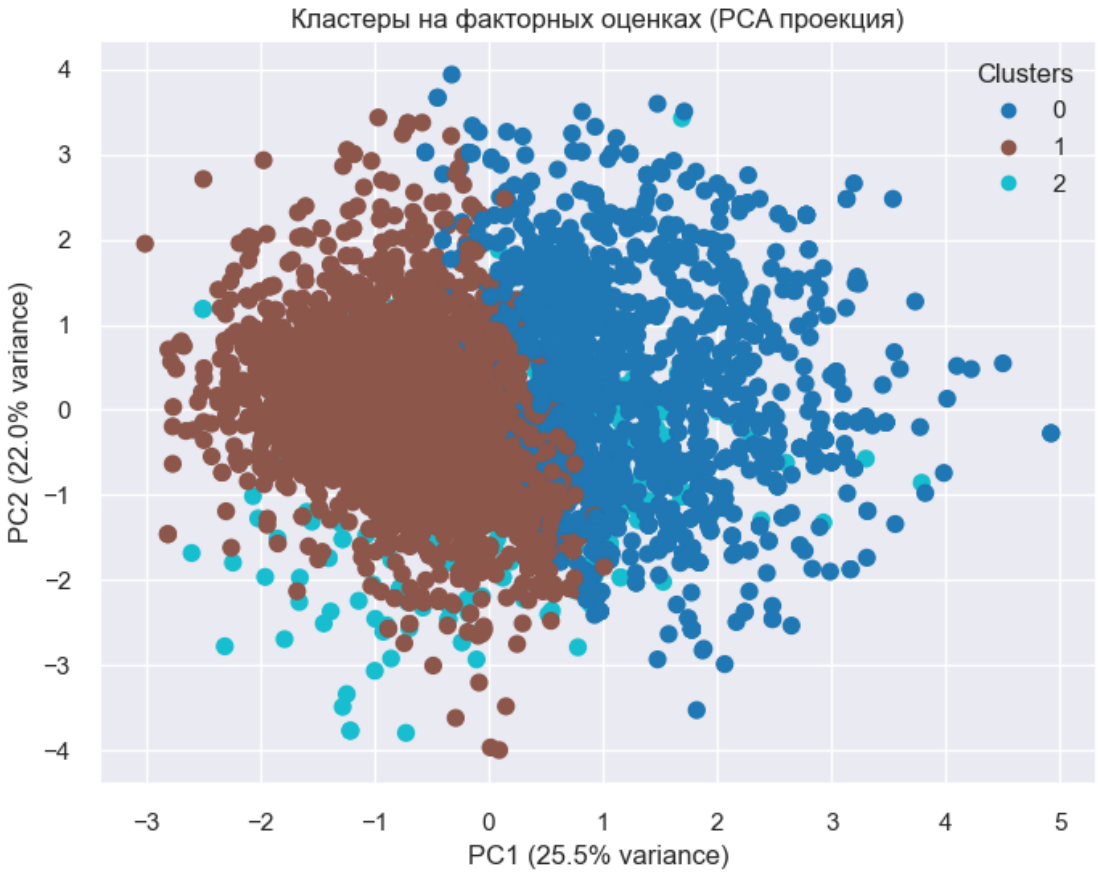

Рис.5 - Визуализация кластеров на двумерной проекции PCA

V. Факторный анализ

Проведён на всех признаках, кроме quality и high_quality (требование метода!).

Проверка пригодности:

- KMO = 0.78 (>0.6)
- Bartlett's test p < 0.001 → данные пригодны для факторизации.

Выделено 5 факторов (собственные значения >1), объясняющих ~75% общей дисперсии.

Интерпретация факторов:

Фактор	Название	Ключевые нагрузки
Factor_1	Плотность / концентрация	density (0.94)
Factor_2	Тип вина + минералы	type_white (-0.9), chlorides (0.61)
Factor_3	Кислотный профиль	total_acidity (0.98), fixed acidity (0.96)
Factor_4	Консервант (SO2)	free sulfur dioxide (0.88)
Factor_5	Уксусный привкус	volatile acidity (0.97)

Факторы имеют ясную физико-химическую интерпретацию и могут использоваться как новые признаки.

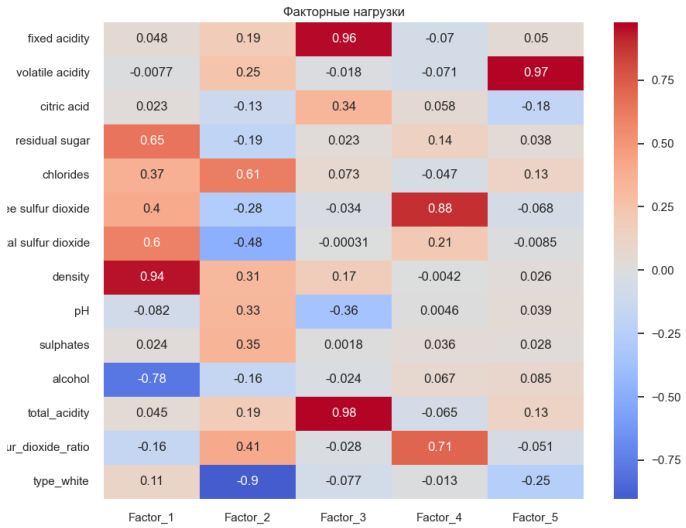

Рис.6 - Факторные нагрузки для 5 выделенных факторов

Заключение

В ходе выполнения задания были:

1. Проведена полная предобработка данных: очистка от пропусков и выбросов, нормализация распределений, создание новых признаков.
2. Выполнен корреляционный анализ, выявлены ключевые зависимости и мультиколлинеарные пары.
3. Построена логистическая регрессия для прогноза высокого качества вина (AUC = 0.79).
4. Проведена кластеризация, выделено 3 естественных сегмента вин с разным профилем качества.
5. Выполнен факторный анализ, выявлено 5 интерпретируемых скрытых факторов, описывающих химическую сущность вин.

Практическая значимость:
Результаты могут быть использованы в виноделии для:

1. Автоматической оценки качества по лабораторным данным,
2. Сегментации продукции,
3. Контроля технологического процесса по ключевым факторам.

Список использованных источников

1. UCI Machine Learning Repository. Wine Quality Data Set. https://archive.ics.uci.edu/ml/datasets/wine+quality
2. Hair, J.F. et al. Multivariate Data Analysis. 8th ed.
3. Scikit-learn: Machine Learning in Python // https://scikit-learn.org
4. FactorAnalyzer documentation // https://factor-analyzer.readthedocs.io