Отчет

Проект по дисциплине «Машинное обучение»

Международный университет Астана
Факультет информационных технологийS
Кафедра Data Science DS-23B

Датасет «winequality-white.csv »

Студенты:
Абу Олжас
Муратов Бекнур

Преподаватель: Әбдіқалық Г.Д.

Астана — 2025

Введение

Цель проекта — изучить методы машинного обучения на реальном датасете, реализовать линейную и логистическую регрессию с нуля, провести классификацию, оценить качество моделей и визуализировать результаты.

Мы использовали **датасет качества вина** (Wine Quality Dataset, Kaggle) с числовыми признаками, включающий химические характеристики вина (кислотность, сахар, pH, спирт и др.) и целевую переменную «quality».

Задачи проекта:

- ➓ Линейная регрессия для предсказания качества вина.
- ➓ Логистическая регрессия для классификации (например, хорошее/плохое качество).
- ➓ Сравнение методов классификации (логистическая регрессия vs. решающее дерево).
- ➓ Анализ метрик (accuracy, precision, recall, F1, ROC AUC).
- ➓ Демонстрация работы с интерактивными виджетами.

Данные

Источник датасета: Kaggle, Wine Quality Dataset
Ссылка: https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

Описание признаков:

- fixed acidity — постоянная кислотность
- volatile acidity — летучая кислотность
- citric acid — лимонная кислота
- residual sugar — остаточный сахар
- chlorides — хлориды
- free sulfur dioxide — свободный диоксид серы
- total sulfur dioxide — общий диоксид серы
- density — плотность
- pH — показатель pH
- sulphates — сульфаты
- alcohol — содержание алкоголя
- quality — оценка качества вина (целевой признак)

Подготовка данных:

- Проверка на пропуски — отсутствуют.
- Дубликаты — удалены.
- Стандартизация числовых признаков (MinMaxScaler или StandardScaler).
- Разделение на train/test (70/30).

Линейная регрессия (с нуля)

Метод: Реализация градиентного спуска с помощью Numpy.

- Функция потерь: MSE (среднеквадратичная ошибка)

- ➓ Параметры: скорость обучения (learning rate), количество эпох, batch size

Результаты:

- ➓ График функции потерь по эпохам (сходимость).
- ➓ Найденные коэффициенты и интерсепт.
- ➓ Scatter plot: связь признаков (например, alcohol, residual sugar) с quality, линия регрессии, доверительный интервал

Вывод: Линейная регрессия показывает базовую зависимость между химическими свойствами вина и оценкой качества, но может недооценивать нелинейные эффекты.

Логистическая регрессия (с нуля)

Метод:

- ➓ Сигмоида + log-loss
- ➓ Градиентный спуск (batch)
- ➓ Возможная регуляризация L2

Результаты:

- ➓ Коэффициенты модели
- ➓ График функции потерь
- ➓ Проблема: пустые/NaN коэффициенты при исходных данных с категориальными признаками → исправлено сменой датасета на полностью числовой.

Вывод: Логистическая регрессия хорошо работает для бинарной классификации (например, «качество ≥ 7» — хорошее, <7 — плохое).

Классификация (два подхода)

Подходы:

1. Логистическая регрессия (с нуля)
2. Decision Tree (sklearn)

Метрики:

- Accuracy, Precision, Recall, F1-score, Confusion Matrix, ROC AUC
- Сравнение моделей: дерево лучше справляется с нелинейными зависимостями, логистическая регрессия быстрее обучается и проще интерпретируется.

Вывод: Для данного датасета с числовыми признаками Decision Tree показывает более высокое качество классификации, однако логистическая регрессия остаётся хорошим базовым методом.

6. Эксперименты и анализ метрик

- Изменяли learning rate, epochs и batch size для градиентного спуска.
- Наблюдали влияние на сходимость MSE и log-loss.
- Decision Tree обучалась без гиперпараметров (базовая версия), можно улучшить с помощью max_depth, min_samples_split.
- Анализ метрик показал, что на малом learning rate сходимость медленная, на большом — возможны колебания функции потерь.

7. Интерфейс / демонстрация

- Использованы виджеты для:
  - загрузки данных
  - выбора модели (Линейная регрессия / Логистическая регрессия / Decision Tree)
  - задания гиперпараметров (learning rate, epochs, batch size)
- Графики потерь, метрик и scatter plot обновляются интерактивно.
- Титульный лист отображён в отдельной ячейке (PDF или изображение).

8. Выводы

- Мы реализовали линейную и логистическую регрессию с нуля.
- Провели классификацию с двумя подходами, сравнили метрики.

- На основе экспериментов показали влияние гиперпараметров на качество моделей.
- Понимание данных и предварительная подготовка (удаление дубликатов, стандартизация) критично важны.
- Проект демонстрирует базовые навыки работы с Python, Numpy, sklearn и визуализацией данных.