ИТОГОВЫЙ ОТЧЁТ ПО МАШИННОМУ ОБУЧЕНИЮ

Тема проекта: Анализ и классификация музыкальных треков Spotify
Дисциплина: Машинное обучение

Студенты:
Белов Глеб
Мубаракшин Камиль

Год: 2025

---

1. Введение

Машинное обучение является одной из ключевых областей анализа данных и широко применяется для решения задач прогнозирования и классификации. В рамках данной работы рассматривается применение базовых алгоритмов машинного обучения к реальному датасету музыкального сервиса Spotify.

Целью проекта является практическая реализация линейной регрессии и логистической регрессии с нуля, а также сравнение логистической регрессии с более сложной моделью Random Forest, реализованной с использованием библиотеки scikit-learn.

---

2. Описание датасета

Для выполнения проекта был использован реальный датасет **Spotify Music Dataset**, размещённый на платформе Kaggle.

Источник данных:
https://www.kaggle.com/datasets/solomonameh/spotify-music-dataset

Датасет содержит более 2000 музыкальных треков и включает числовые признаки, описывающие аудио-характеристики композиций.

Основные признаки датасета:

- **danceability** — танцевальность трека
- **energy** — энергетичность
- **loudness** — громкость

- **speechiness** — доля разговорной речи
- **acousticness** — акустичность
- **tempo** — темп композиции
- **popularity** — популярность трека

Датасет является структурированным, реалистичным и подходит для решения задач регрессии и классификации.

---

3. Линейная регрессия (реализация с нуля)

Линейная регрессия использовалась для решения задачи предсказания числовых значений на основе входных признаков. Модель была реализована с нуля с использованием библиотеки NumPy.

В качестве функции потерь использовалась среднеквадратичная ошибка (MSE):

$$\mathbf{MSE = (1 / n) \cdot \Sigma(y - \hat{y})^2}$$

Обучение модели осуществлялось методом градиентного спуска. В ходе экспериментов анализировалось влияние скорости обучения (learning rate) и количества эпох на процесс сходимости модели. Были получены коэффициенты линейной регрессии и построены графики изменения функции потерь.

---

4. Логистическая регрессия (реализация с нуля)

Логистическая регрессия применялась для решения задачи бинарной классификации. Модель была реализована вручную без использования готовых алгоритмов машинного обучения.

В основе модели лежит сигмоидная функция:

$$\mathbf{\sigma(z) = 1 / (1 + e^{-z})}$$

Для обучения использовалась логистическая функция потерь (log loss). Качество классификации оценивалось с помощью метрик accuracy, precision, recall, F1-score и ROC AUC.

---

5. Random Forest (сравнительная модель)

Для сравнения качества классификации была использована модель **Random Forest**, реализованная с помощью библиотеки scikit-learn.

Random Forest представляет собой ансамбль решающих деревьев и, как правило, показывает более высокую точность классификации за счёт снижения переобучения и устойчивости к шуму в данных.

---

6. Эксперименты и сравнение моделей

В ходе экспериментов было проведено сравнение логистической регрессии и Random Forest по основным метрикам качества. Результаты показали, что Random Forest демонстрирует более высокую точность классификации.

В то же время логистическая регрессия обладает высокой интерпретируемостью и простотой реализации, что делает её полезной для анализа влияния признаков на результат.

---

7. Выводы

В рамках данной работы были реализованы базовые методы машинного обучения и проведено их сравнение на реальном датасете Spotify. Экспериментальные результаты показали преимущество модели Random Forest по качеству классификации.

Полученные результаты подтверждают эффективность применения методов машинного обучения для анализа музыкальных данных и решения практических задач.