ПРОЕКТ

по Основам интеллектуального анализа данных и машинного обучения

НА ТЕМУ:

Моделирование климатических данных Индии с использованием линейной и логистической регрессии

Выполнили студенты 3 курса:
Романюк Дмитрий, Омирзаков Азамат

1. Цель работы

Цель проекта — исследовать задачи регрессии и бинарной классификации на реальном климатическом датасете по городам Индии 2024–2025 годов, реализовав линейную и логистическую регрессию с нуля на numpy и сравнив их качество с моделью Random Forest. В отчёте анализируется сходимость градиентного спуска, влияние гиперпараметров (learning rate, число эпох, размер батча) и качество классификации по основным метрикам.

2. Датасет: источник и описание

Происхождение данных.
Используется открытый датасет *Indian Climate Dataset 2024–2025*, размещённый на Kaggle:
источник — Kaggle, автор Ankush Narwade;
ссылка:https://www.kaggle.com/datasets/ankushnarwade/indian-climate-dataset-20242025

Краткое описание полей (главные колонки):

- `Date` — дата наблюдения.
- `City` — город (Mumbai, Delhi, Bengaluru и др.).
- `State` — штат Индии.
- `TemperatureMax C` — максимальная температура за сутки, °C.
- `TemperatureMin C` — минимальная температура за сутки, °C.
- `TemperatureAvg C` — средняя температура за сутки, °C (целевой признак для регрессии).
- `Humidity` — средняя относительная влажность, %.
- `Rainfall mm` — количество осадков, мм.
- `WindSpeed kmh` — средняя скорость ветра, км/ч.
- `AQI` — индекс качества воздуха.
- `AQICategory` — категориальное качество воздуха (Good, Satisfactory, Moderate, Poor, Very Poor).
- `Pressure hPa` — атмосферное давление, гПа.
- `CloudCover` — облачность, %.

В выборке более 7000 строк, что удовлетворяет требованию минимум 1500–2000 наблюдений и не менее 5 информативных признаков.

3. Метод и архитектуры моделей

3.1. Подготовка данных

- Удалены дубликаты, проверено отсутствие пропусков.
- Для моделей регрессии и классификации использовались числовые признаки:
  `TemperatureMax C`, `TemperatureMin C`, `Humidity`, `Rainfall mm`, `WindSpeed kmh`, `AQI`, `Pressure hPa`, `CloudCover`.
- Признаки нормализованы по формуле Z-score: $$x'_j = \frac{x_j - \mu_j}{\sigma_j}$$
- Для всех моделей данные разделены на обучающую и тестовую выборки в пропорции 80/20 с предварительным перемешиванием.
- Для моделей, реализованных с нуля, к матрице признаков добавлен bias-столбец единиц.

3.2. Линейная регрессия (регрессия температуры)

Модель. Линейная регрессия для предсказания средней температуры:

$$\hat{y} = w_0 + w_1 x_1 + \cdots + w_n x_n = \mathbf{w}^T \mathbf{x}$$

Функция потерь (MSE): $$J(\mathbf{w}) = \frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

Градиент MSE: $$\nabla J(\mathbf{w}) = \frac{2}{m} \mathbf{X}^T (\mathbf{X}\mathbf{w} - \mathbf{y})$$

Обучение (mini-batch SGD):

- размер батча B;
- на каждом шаге веса обновляются по правилу: $$\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla J(\mathbf{w})$$

где: η — скорость обучения (learning rate).

Метрики: MSE, RMSE, MAE на train и test.

3.3. Логистическая регрессия (бинарная классификация)

Постановка.
Введена бинарная целевая переменная:

- $y=1$, если `TemperatureAvg C` выше медианы по выборке — «жаркий день»;
- $y=0$ иначе — «не жаркий день».

Сигмоида: $$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Модель вероятности класса 1: $$\hat{y} = \sigma(\mathbf{w}^T \mathbf{x})$$

Логистическая функция потерь (кросс-энтропия):

$$J(\mathbf{w}) = -\frac{1}{m} \sum_{i=1}^{m} [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

Градиент лог-лосса: $$\nabla J(\mathbf{w}) = \frac{1}{m} \mathbf{X}^T (\hat{\mathbf{y}} - \mathbf{y})$$

L2-регуляризация (опционально): $$J_\lambda(\mathbf{w}) = J(\mathbf{w}) + \frac{\lambda}{2m} \sum_{j=1}^{n} w_j^2$$

Обучение (mini-batch SGD): $$\mathbf{w} \leftarrow \mathbf{w} - \eta \left(\frac{1}{B} \mathbf{X}_B^T (\hat{\mathbf{y}}_B - \mathbf{y}_B) + \frac{\lambda}{B} \mathbf{w} \right)$$

(без штрафа для w_0).

Метрики: accuracy, precision, recall, F1-score, confusion matrix, ROC AUC (по вероятностям).

3.4. Random Forest (вторая модель классификации)

В качестве второй модели классификации выбран Random Forest (реализация `RandomForestClassifier` из scikit-learn).

- Модель строит множество деревьев решений, каждое обучается на случайном подмножестве объектов и признаков.
- Предсказание — голосование деревьев (класс, набравший большинство).
- Используются те же признаки и разбиение train/test, что и для логистической регрессии.
- Метрики: accuracy, precision, recall, F1-score, confusion matrix, ROC AUC.

4. Гиперпараметры и эксперименты

4.1. Гиперпараметры регрессионных моделей

Для линейной и логистической регрессии задаются:

- learning rate η (например, 0.001, 0.01, 0.1);
- число эпох n_epochs;
- размер батча $batch\_size$;
- коэффициент L2-регуляризации λ (при необходимости).

Проводился эксперимент: изменение одного параметра при фиксированных остальных и анализ графика сходимости функции потерь.

Наблюдения, которые можно описать (подставь свои численные выводы):

- При слишком маленьком η loss убывает очень медленно.
- При умеренном η loss быстро падает и стабилизируется.
- При слишком большом η loss начинает колебаться или расти — градиентный спуск становится нестабильным.
- Увеличение числа эпох улучшает качество до определённого уровня, затем приводит к плато.
- Малый `batch_size` увеличивает шум градиента, но помогает выходить из локальных минимумов; большой батч даёт более гладкую, но дорогую по времени сходимость.

4.2. Гиперпараметры Random Forest

Основные параметры:

- число деревьев `n_estimators` (например, 100);
- максимальная глубина `max_depth` (ограниченная или `None`);
- `random_state` для воспроизводимости.

Можно кратко отметить, что увеличение числа деревьев улучшает устойчивость и качество до некоторого предела, но увеличивает время обучения.

5. Результаты

5.1. Линейная регрессия

Модель	MSE train	MSE test
Линейная регрессия	0.001254880230546737	0.0012693867392215184

ошибка на тестовой выборке мала , что подтверждает хорошую аппроксимацию средней температуры по набору признаков и отсутствие сильного переобучения.

5.2. Логистическая регрессия и Random Forest

Модель	**Accuracy test**	**Precision**	**Recall**	**F1-score**	**ROC AUC**
Логистическая регрессия	0.9965800273597811	0.995901639344262	0.9972640218878249	0.996582365003417	0.99997192908913
Random Forest	0.9945280437756497	0.995884773662551	0.9931600547195623	0.994520547945205	0.99452804377564