UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"

ESTATÍSTICA E PROBABILIDADE: ESTUDO E ANÁLISE DOS DADOS METEOROLÓGICOS DE MANAUS

Alunos: Murilo Brandão e Leonardo de Lima

Seção 1:Murillo
Seção 2: Leonardo

Sorocaba/SP

2025

Sumário

Obs: A Seção 1 corresponde à parte do trabalho desenvolvida por Murillo, enquanto a Seção 2 apresenta as contribuições realizadas por Leonardo, mantendo a mesma estrutura de tópicos, mas com abordagens e parâmetros distintos.

Introdução

A análise de dados por meio de técnicas estatísticas e computacionais tem se consolidado como uma ferramenta essencial na compreensão de fenômenos complexos em diversas áreas do conhecimento. No contexto ambiental, o tratamento e a interpretação de dados meteorológicos permitem não apenas a descrição de padrões climáticos, mas também a construção de modelos preditivos com potencial aplicação em planejamento urbano, agricultura, saúde pública e gestão de riscos.

Este trabalho tem como objetivo principal aplicar métodos de análise estatística e modelagem preditiva a um conjunto de dados meteorológicos, explorando diferentes abordagens para a previsão e classificação de variáveis climáticas. A proposta envolve a avaliação comparativa entre estratégias estatísticas e de aprendizado de máquina, buscando identificar seus pontos fortes, limitações e aplicabilidades em cenários reais.

Ao longo do desenvolvimento, são consideradas tanto a qualidade estatística dos modelos quanto aspectos como robustez, interpretabilidade e eficiência computacional. A partir disso, espera-se construir uma análise crítica que vá além da precisão dos resultados, envolvendo também a compreensão do comportamento dos dados e das implicações das escolhas metodológicas realizadas.

Descrição e justificativa do dataset

O presente estudo utiliza um conjunto de dados meteorológicos da cidade de Manaus, disponibilizado pelo Instituto Nacional de Meteorologia (INMET), uma fonte oficial e amplamente reconhecida por sua confiabilidade e abrangência. A base contempla um recorte temporal contínuo com registros diários compreendidos entre os anos de 1995 e 2025, totalizando 10.969 observações.

Cada linha da base representa um dia específico e inclui variáveis essenciais para a análise climática: data da medição, insolação total diária (em horas), precipitação total (em milímetros), temperatura máxima e mínima diária (em °C), umidade relativa do ar (média diária, em %), e velocidade média diária do vento (em m/s). As variáveis são todas de natureza contínua, o que favorece a aplicação de técnicas estatísticas e modelos preditivos tanto de regressão quanto de classificação binária, após a devida transformação de alguns atributos.

A escolha deste dataset se justifica por três razões principais. Primeiramente, a cidade de Manaus apresenta um regime climático equatorial úmido, com padrões de temperatura e precipitação que se destacam por sua regularidade e intensidade, oferecendo um cenário interessante para análise. Em segundo lugar, a série histórica extensa e contínua permite observar comportamentos sazonais, padrões de longo prazo e fenômenos extremos. Por fim, o conjunto de dados atende aos requisitos técnicos do projeto, oferecendo variáveis relevantes, bem documentadas e com qualidade suficiente para suportar diferentes abordagens metodológicas de análise estatística e modelagem preditiva.

1. **Seção**

Modelagem preditiva

A etapa de modelagem preditiva tem como objetivo aplicar algoritmos estatísticos e computacionais para estimar ou classificar valores de variáveis-alvo com base em padrões encontrados nos dados disponíveis. A partir do conjunto de dados meteorológicos previamente descrito, foram selecionadas variáveis climáticas relevantes como preditoras, e definidas duas abordagens complementares: uma voltada para a previsão de valores contínuos e outra para a classificação de eventos binários

Regressão linear

Entre os modelos aplicados na tarefa de previsão de variáveis contínuas, utilizamos a regressão linear como abordagem inicial. Esse modelo foi empregado especificamente para estimar a temperatura máxima diária com base em variáveis climáticas como insolação, precipitação, umidade, temperatura mínima e velocidade do vento.
Os modelos adotados na tarefa de regressão foram escolhidos considerando as características do conjunto de dados meteorológicos, que apresenta variações naturais nas variáveis climáticas e relações não necessariamente lineares entre elas sendo tais modelos os seguintes abaixo:

- ***Árvore de Regressão (Decision Tree Regressor)*** foi utilizada por ser um modelo simples e capaz de lidar com dados que possuem estrutura hierárquica ou regras de decisão claras, o que é comum em dados ambientais.

- ***Random Forest Regressor*** foi incluído por ser uma extensão mais robusta da árvore de decisão, adequada para reduzir variações nos dados e melhorar a generalização em séries históricas como as de clima.

- ***XGBoost Regressor*** foi utilizado por ser eficiente em capturar padrões complexos com precisão, especialmente útil em dados contínuos como temperatura, que apresentam flutuações sazonais e dependência entre variáveis.

Regressão logística

Na tarefa de classificação binária, foi adotado um modelo supervisionado com o objetivo de prever a ocorrência de eventos de chuva extrema com base em variáveis meteorológicas. Considerando que o conjunto de dados possui estrutura bem definida e relações lineares claras entre os atributos climáticos e a variável-alvo, o modelo escolhido se mostrou adequado para separar as classes com alta precisão. O modelo utilizado é apresentado a seguir:

- **Regressão Logística (Logistic Regression)** foi utilizada para classificar a variável *chuva_extrema*, definida como 1 quando a precipitação diária ultrapassa 20 mm e 0 caso contrário. Como o problema envolve uma decisão binária, esse modelo foi escolhido por ser direto e eficiente para esse tipo de classificação.

Avaliação e Benchmarking

A avaliação dos modelos foi feita com base em três métricas, conforme exigido: uma métrica de erro e duas de desempenho. Para a tarefa de regressão, foram aplicadas as métricas RMSE, R^2 e Explained Variance. Já na tarefa de classificação binária, foram utilizadas F1-Score, AUC-ROC e Acurácia.

Os resultados obtidos foram organizados em gráficos para facilitar a comparação entre os modelos e destacar diferenças de desempenho em termos estatísticos. As comparações a seguir apresentam, de forma visual, a performance dos modelos aplicados ao conjunto de dados meteorológicos.

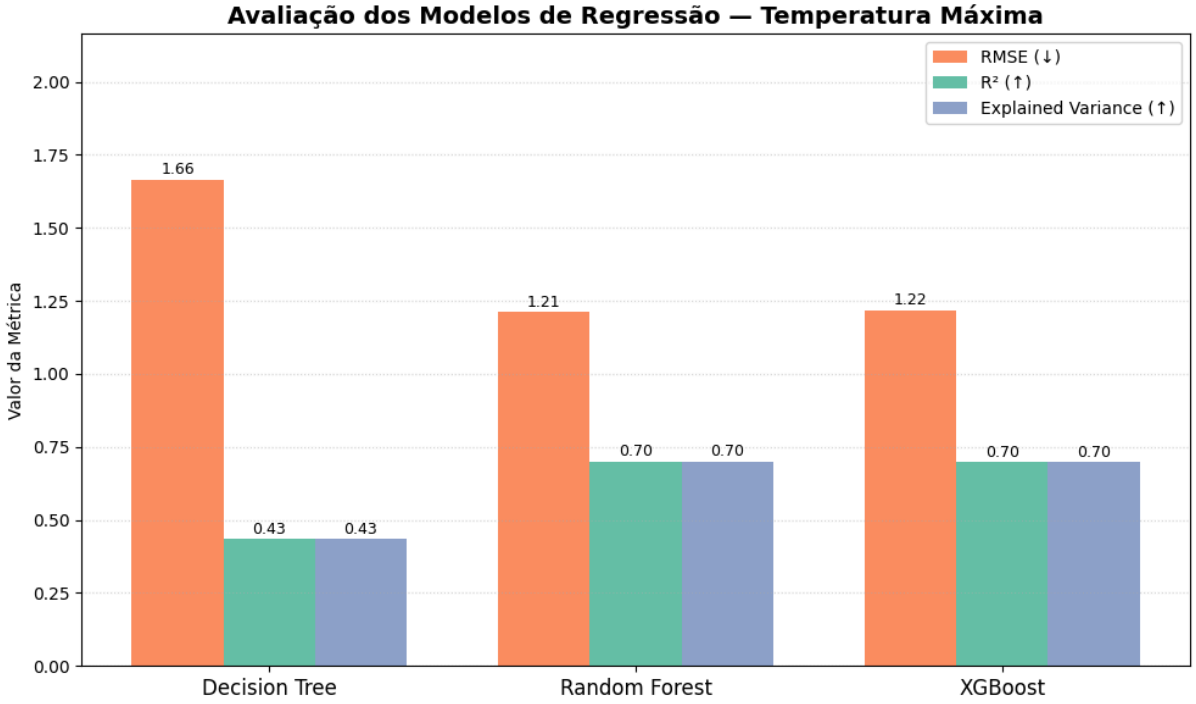
Métricas regressão linear

Para a regressão linear, foram definidas, conforme solicitado, uma métrica de erro e duas métricas de desempenho para avaliação dos modelos. As métricas utilizadas foram:

- **RMSE (Root Mean Squared Error)**: avalia o erro médio das previsões, penalizando desvios maiores;
- **R^2 (Coeficiente de Determinação)**: indica a proporção da variabilidade da variável dependente explicada pelo modelo;
- **Explained Variance Score**: mede a variância explicada pelas previsões, refletindo a consistência do modelo.

Desempenhos dos modelos

Para avaliar os modelos aplicados na previsão da temperatura máxima diária, foram utilizadas três métricas: RMSE, R² e Explained Variance Score. Cada métrica oferece uma perspectiva complementar sobre a performance dos modelos. O gráfico a seguir apresenta a comparação entre os três modelos adotados para regressão linear:



Fazendo uma análise crítica de cada modelo isolado observamos que:

Decision Tree Regressor - Apresentou o pior desempenho entre os três modelos. Com um RMSE de 1.66, mostrou maior erro médio nas previsões. Os valores de R² e Explained Variance (ambos próximos de 0.43) indicam que o modelo teve dificuldade em capturar a variabilidade da temperatura máxima, sendo pouco eficaz para essa tarefa.

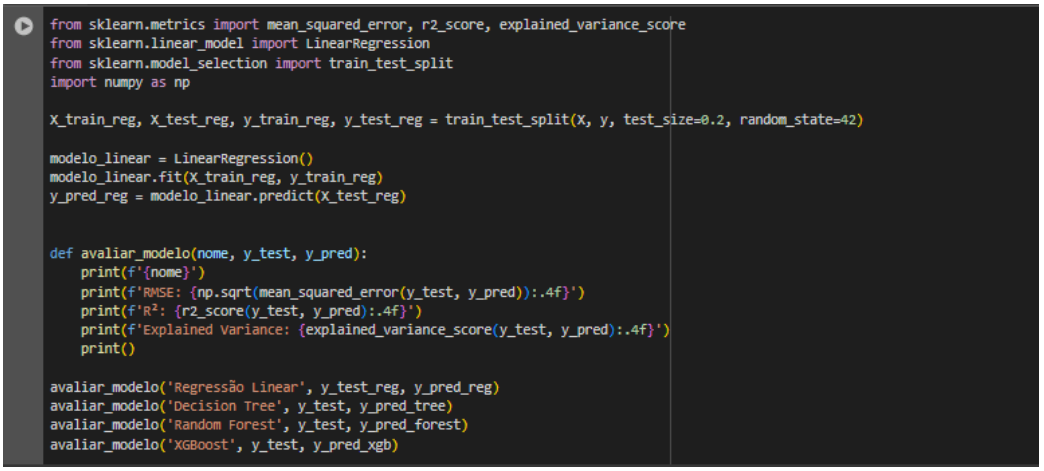
Random Forest Regressor - Foi o modelo com melhor desempenho geral. Apresentou o menor RMSE (1.21) e os maiores valores de R² e Explained Variance (ambos em torno de 0.70), indicando que além de reduzir o erro, conseguiu explicar bem a variabilidade da variável-alvo. Se destacou pela estabilidade e consistência das previsões.

XGBoost Regressor - Teve desempenho muito próximo ao do Random Forest, com RMSE de 1.22 e R² e variância explicada também em torno de 0.70. Apesar da pequena diferença no erro, seu desempenho geral foi equilibrado e sólido, mostrando-se uma boa alternativa ao Random Forest.

Entre os três modelos testados, o Random Forest Regressor se destacou como o mais eficaz na previsão da temperatura máxima, por apresentar o menor erro e maior capacidade de explicação da variabilidade dos dados. O XGBoost obteve desempenho semelhante, enquanto a árvore de decisão simples foi significativamente inferior em todos os critérios.

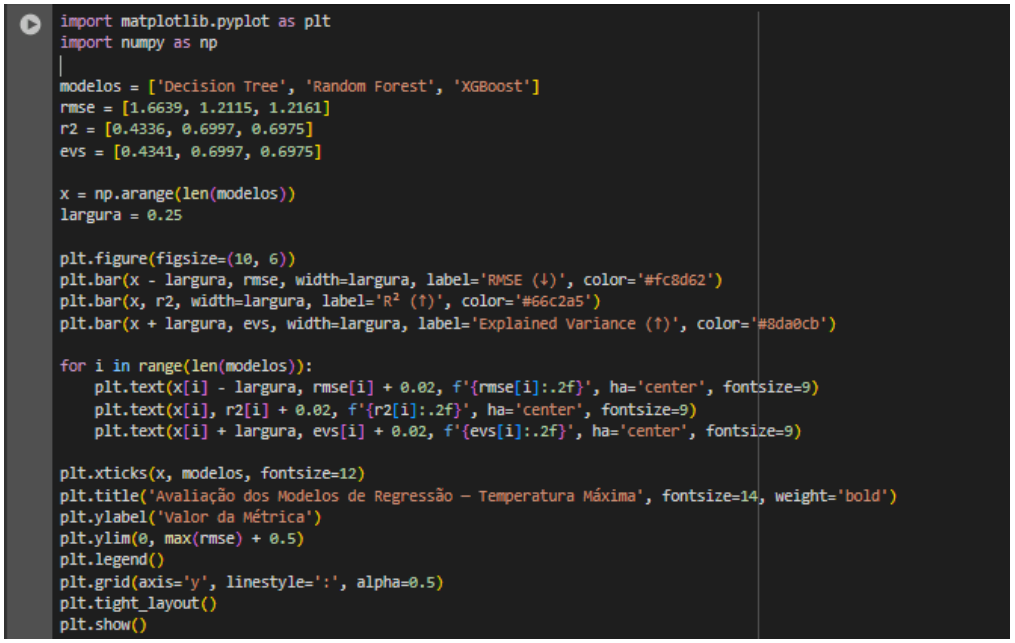
Estrutura do código

A avaliação dos modelos foi feita utilizando uma única função em Python, que recebe como parâmetros o nome do modelo, os valores reais e os valores previstos. A partir disso, são calculadas as métricas RMSE, R² e Explained Variance. Os modelos foram previamente treinados e testados com divisão 80/20, e as previsões foram comparadas com os valores reais do conjunto de testes. A biblioteca *scikit-learn* foi utilizada para todas as métricas, e o *numpy* para calcular a raiz quadrada do erro quadrático médio. Essa estrutura permitiu a padronização da análise de desempenho entre todos os modelos testados. Segue o código:



```python
from sklearn.metrics import mean_squared_error, r2_score, explained_variance_score
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import numpy as np

X_train_reg, X_test_reg, y_train_reg, y_test_reg = train_test_split(X, y, test_size=0.2, random_state=42)

modelo_linear = LinearRegression()
modelo_linear.fit(X_train_reg, y_train_reg)
y_pred_reg = modelo_linear.predict(X_test_reg)


def avaliar_modelo(nome, y_test, y_pred):
    print(f'{nome}')
    print(f'RMSE: {np.sqrt(mean_squared_error(y_test, y_pred)):.4f}')
    print(f'R²: {r2_score(y_test, y_pred):.4f}')
    print(f'Explained Variance: {explained_variance_score(y_test, y_pred):.4f}')
    print()

avaliar_modelo('Regressão Linear', y_test_reg, y_pred_reg)
avaliar_modelo('Decision Tree', y_test, y_pred_tree)
avaliar_modelo('Random Forest', y_test, y_pred_forest)
avaliar_modelo('XGBoost', y_test, y_pred_xgb)
```

O gráfico de desempenho comparativo foi gerado utilizando a biblioteca *matplotlib*, com os valores de RMSE, R² e Explained Variance.O código abaixo apresenta a estrutura utilizada para criar a visualização final, permitindo uma comparação direta entre os modelos testados quanto ao seu erro médio e capacidade explicativa.



```python
import matplotlib.pyplot as plt
import numpy as np

modelos = ['Decision Tree', 'Random Forest', 'XGBoost']
rmse = [1.6639, 1.2115, 1.2161]
r2 = [0.4336, 0.6997, 0.6975]
evs = [0.4341, 0.6997, 0.6975]

x = np.arange(len(modelos))
largura = 0.25

plt.figure(figsize=(10, 6))
plt.bar(x - largura, rmse, width=largura, label='RMSE (↓)', color='#fc8d62')
plt.bar(x, r2, width=largura, label='R² (↑)', color='#66c2a5')
plt.bar(x + largura, evs, width=largura, label='Explained Variance (↑)', color='#8da0cb')

for i in range(len(modelos)):
    plt.text(x[i] - largura, rmse[i] + 0.02, f'{rmse[i]:.2f}', ha='center', fontsize=9)
    plt.text(x[i], r2[i] + 0.02, f'{r2[i]:.2f}', ha='center', fontsize=9)
    plt.text(x[i] + largura, evs[i] + 0.02, f'{evs[i]:.2f}', ha='center', fontsize=9)

plt.xticks(x, modelos, fontsize=12)
plt.title('Avaliação dos Modelos de Regressão – Temperatura Máxima', fontsize=14, weight='bold')
plt.ylabel('Valor da Métrica')
plt.ylim(0, max(rmse) + 0.5)
plt.legend()
plt.grid(axis='y', linestyle=':', alpha=0.5)
plt.tight_layout()
plt.show()
```

Métricas regressão logística

Para a regressão logística, foram definidas, uma métrica de desempenho voltada ao equilíbrio entre classes e duas métricas adicionais para avaliar a qualidade geral da classificação. As métricas utilizadas foram:

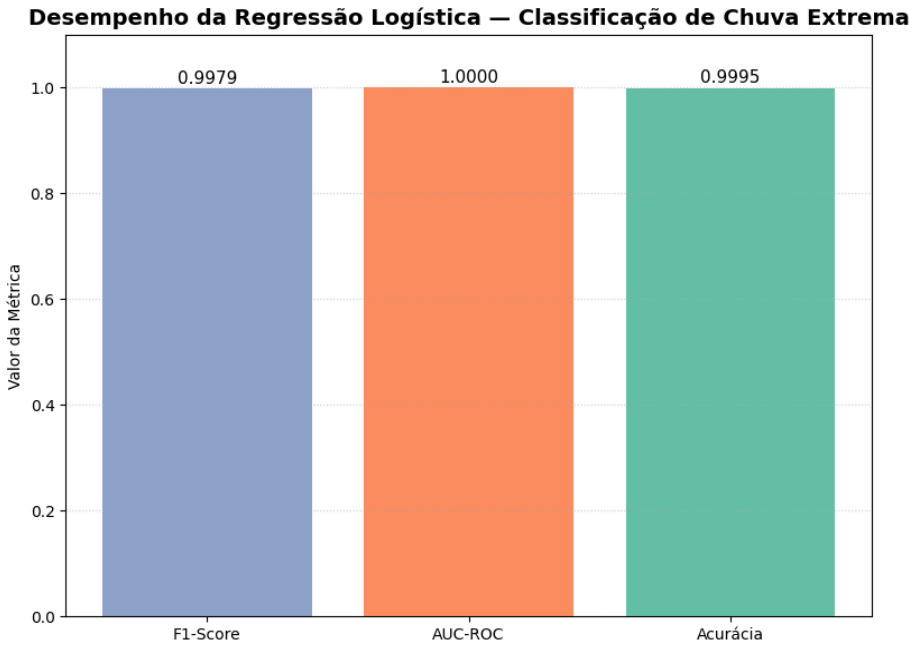
- **F1-Score:** mede o equilíbrio entre **precisão** e **revocação (recall)**, sendo especialmente útil em problemas com classes desbalanceadas, como a ocorrência de chuva extrema;

- **AUC-ROC (Área sob a Curva ROC):** avalia a capacidade do modelo de distinguir corretamente entre as duas classes (chuva extrema e não extrema), considerando diferentes limiares de decisão;

- **Acurácia:** indica a proporção total de previsões corretas (positivas e negativas) realizadas pelo modelo, sendo útil como medida geral de acerto quando as classes estão relativamente equilibradas.

Desempenho do modelo

O desempenho da regressão logística foi avaliado por meio das três métricas citadas anteriormente. Os resultados obtidos demonstram que o modelo foi altamente eficaz na tarefa de classificação da variável chuva_extrema, que representa a ocorrência de chuva intensa (precipitação superior a 20 mm).



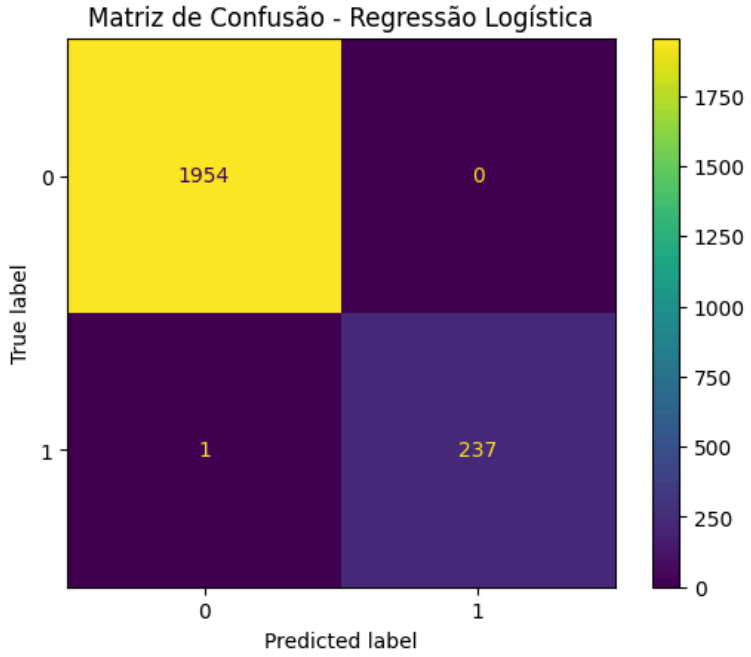
A análise das métricas revela um desempenho excepcional do modelo de regressão logística. O F1-Score de 0.9979 indica que o modelo mantém um forte equilíbrio entre precisão e recall, o que é fundamental em situações com leve desbalanceamento de classes, como neste caso, onde dias com chuva extrema são minoria.

O valor de AUC-ROC igual a 1.0 demonstra que o modelo possui capacidade perfeita de separação entre as classes, sendo capaz de identificar corretamente tanto os dias com quanto sem ocorrência de chuva forte. Além disso, a acurácia de 99,95% confirma que o modelo obteve uma altíssima taxa de acertos globais.

Antes da avaliação do modelo, o conjunto de dados foi dividido em 80% para treinamento e 20% para teste, totalizando aproximadamente 2.192 registros reservados exclusivamente para validação do desempenho do classificador. Essa divisão segue uma prática padrão em ciência de dados, permitindo verificar a capacidade de generalização do modelo em dados não vistos. Dos 2.192 exemplos presentes no conjunto de teste:

- 1954 dias foram corretamente classificados como não tendo chuva extrema (classe 0);
- 237 dias foram corretamente classificados como tendo chuva extrema (classe 1);
- Apenas 1 dia foi classificado erroneamente como sem chuva extrema, quando na verdade houve (falso negativo);
- Nenhum falso positivo foi registrado, ou seja, o modelo nunca previu chuva extrema quando ela não ocorreu.

Essa distribuição confirma que o modelo foi extremamente confiável, com baixíssimo índice de erro, alta sensibilidade à classe minoritária e total ausência de alarmes falsos para eventos de chuva intensa. Essa precisão é especialmente relevante em aplicações reais de previsão climática, nas quais alertas indevidos ou omissões podem causar impactos significativos.



.

Em síntese, o modelo de regressão logística se mostrou altamente eficaz, mesmo com um modelo relativamente simples e um conjunto de dados com leve desequilíbrio de classes.

Estrutura do código

O código da regressão logística foi implementado em Python com a biblioteca scikit-learn. As variáveis preditoras e a variável-alvo (chuva_extrema) foram definidas, e os dados foram divididos em 80% para treino e 20% para teste com *train_test_split.*

O modelo *LogisticRegression* foi treinado com *fit()*, e as previsões foram geradas com *predict()* e *predict_proba()*. As métricas F1-Score e AUC-ROC foram calculadas sobre o conjunto de testes para avaliar o desempenho da classificação binária.

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import f1_score, roc_auc_score, accuracy_score

X_clf = X_reg  # Mesmas features
y_clf = df['chuva_extrema']

X_train_clf, X_test_clf, y_train_clf, y_test_clf = train_test_split(X_clf, y_clf, test_size=0.2, random_state=42)

modelo_log = LogisticRegression(max_iter=1000)
modelo_log.fit(X_train_clf, y_train_clf)

y_pred_log = modelo_log.predict(X_test_clf)
y_proba_log = modelo_log.predict_proba(X_test_clf)[:, 1]

print("\nRegressão Logística:")
print("F1 Score:", f1_score(y_test_clf, y_pred_log))
print("AUC-ROC:", roc_auc_score(y_test_clf, y_proba_log))
print("Acurácia:", accuracy_score(y_test_clf, y_pred_log))
```

2. Seção

Modelagem Preditiva

Foi realizado uma regressão linear com os modelos KNN Regressor e Lasso Regression para prever a precipitação diária com base nas outras features : ‘insolação’, ‘temp_max’, ‘temp_min’, ‘vento’, ‘umidade’.

Regressão Linear

- **KNN Regressor** foi utilizado por se tratar de um modelo que é baseado na ideia de que amostras parecidas, tendem a ter saídas semelhantes. E também é sensível a padrões locais, o que é útil em dados meteorológicos podem estar associados a condições similares.

- **Lasso Regression** foi usado por ser um modelo que realiza uma seleção de variáveis automaticamente, fazendo com que coeficientes menos relevantes sejam zero. É útil para entender quais variáveis tem um impacto mais significativo, como a precipitação.

Regressão Logística

- **MLPClassifier** foi utilizado por ser um modelo extremamente eficiente, que pode capturar relações não lineares e problemas de difícil análise, como previsão de chuva com base em várias condições atmosféricas.

Avaliação e Benchmarking

Métricas Regressão Linear

Para a regressão linear, foram utilizadas, uma métrica de erro e duas métricas de desempenho para avaliação dos modelos. As métricas utilizadas foram:

- **RMSE (Root Mean Squared Error)**: avalia o erro médio das previsões, penalizando desvios maiores;
- **R^2 (Coeficiente de Determinação)**: indica a proporção da variabilidade da variável dependente explicada pelo modelo;
- **Explained Variance Score**: mede a variância explicada pelas previsões, refletindo a consistência do modelo.

Desempenho dos modelos

KNN Regressor

- RMSE (Root Mean Squared Error): 14.17 mm
- R^2 (Coeficiente de Determinação): 0.0089
- Explained Variance: 0.009

Esses resultados indicam um desempenho bastante limitado do modelo. O valor de R^2 é próximo de zero, indicando que o KNN foi incapaz de explicar a variabilidade da precipitação com base nas features fornecidas. O RMSE com 14.17, também confirma que as previsões do modelo se distanciaram significativamente dos valores reais.

O KNN é sensível a outliers, pois cada vizinho influencia diretamente o resultado da predição. Como a feature 'precipitação' possui dias com valores extremos, pode ter afetado a performance do modelo. Além disso, como a precipitação depende de fatores complexos e possivelmente, o KNN não conseguiu capturar essas relações de forma eficaz.

Lasso Regression

- RMSE (Root Mean Squared Error) : 14.17 mm: Em média, o erro da previsão é de 14.17 mm de precipitação, um valor relativamente alto dependendo do intervalo da variável.
- R^2 (Coeficiente de Determinação) : 0.135: O modelo explica apenas 13,5% da variabilidade dos dados de precipitação.
- Explained Variance : 13,6%: Coerente com o R^2, também indica que a maior parte da variação não é explicada pelo modelo.

Lasso Regression é um modelo de fácil interpretação, e é útil para verificar as variáveis relevantes. Porém se observa que seu poder preditivo não é significativo.

Métricas Regressão Logística

- **F1-Score:** mede o equilíbrio entre precisão e revocação (recall), sendo especialmente útil em problemas com classes desbalanceadas, como a ocorrência de chuva extrema;

- **AUC-ROC (Área sob a Curva ROC):** avalia a capacidade do modelo de distinguir corretamente entre as duas classes (chuva extrema e não extrema), considerando diferentes limiares de decisão;

- **Acurácia:** indica a proporção total de previsões corretas (positivas e negativas) realizadas pelo modelo, sendo útil como medida geral de acerto quando as classes estão relativamente equilibradas.

Desempenho do modelo

MLPClassifier

- F1-Score : 0.201
- AUC-ROC : 0.797
- Acurácia : 0.801

O modelo MLPClassifier foi treinado para prever dias com chuva forte. A acurácia obtida foi de 80,1%, porém o F1-score de apenas 0,20 indicando que o modelo tem baixa capacidade de prever corretamente os casos positivos, que são as chuvas fortes. A métrica AUC-ROC de 0,797 revela que o modelo conseguiu aprender padrões de separação entre as classes, mas sofre com o desbalanceamento dos dados.