Previsão da Temperatura Máxima Utilizando Regressão e Técnicas de Machine Learning

Leonardo Henrique de Oliveira Matos, 10389516[1], Victor Junqueira Colombaro, 10395711[1]

[1] [1]Ciência da Computação,
Faculdade de Computação e Informática,
Universidade Presbiteriana Mackenzie,
São Paulo – SP – Brasil

10389516@mackenzista.com.br, 10395711@mackenzista.com.br

Resumo. *Este projeto tem como objetivo prever a temperatura máxima das cidades de Seattle (EUA) e Budapeste (Hungria) com base em variáveis climáticas e condições meteorológicas. Utilizando técnicas de aprendizado de máquina, foram aplicados modelos de regressão, classificação e clusterização para analisar padrões nos dados climáticos. As etapas incluíram análise exploratória de dados, preparação do dataset, implementação de modelos de aprendizado supervisionado e não supervisionado, além da avaliação do desempenho com métricas apropriadas.*

1. Introdução

O principal objetivo deste projeto é prever a temperatura máxima utilizando dados climáticos disponíveis das cidades de Seattle e Budapeste. Adicionalmente, busca-se identificar padrões climáticos relevantes nessas localidades por meio de técnicas de clusterização. Para alcançar esses objetivos, utilizamos um conjunto robusto de ferramentas, incluindo regressão linear, Random Forest e KNN para modelagem preditiva, além de K-Means para análise não supervisionada. Essa abordagem permite não apenas prever valores contínuos, mas também categorizar os dados e identificar padrões ocultos nas relações entre variáveis.

A escolha de Seattle e Budapeste como cenários de estudo foi motivada por suas diferenças climáticas significativas. Enquanto Seattle apresenta um clima temperado com alta pluviosidade, Budapeste possui características continentais com invernos rigorosos e verões quentes. Essas diferenças oferecem um desafio adicional para a generalização dos modelos preditivos, sendo fundamental avaliar o desempenho e as limitações ao aplicar os mesmos métodos em diferentes contextos climáticos.

Neste trabalho, foram implementadas técnicas de aprendizado de máquina utilizando bibliotecas amplamente reconhecidas, como `scikit-learn` e `pandas`, para análise de dados e construção de modelos. O foco também está em garantir que os dados sejam devidamente preparados, abordando questões como tratamento de valores ausentes, normalização e categorização de variáveis. Além disso, a avaliação dos modelos foi realizada por meio de métricas como precisão, recall, F1-Score e análise gráfica de resíduos.

Por fim, o estudo também explora as limitações dos modelos construídos e discute sua capacidade de generalização. Como esperado, a inclusão de variáveis adicio-

nais, como umidade e pressão atmosférica, pode melhorar significativamente o desempenho dos modelos e permitir maior aplicabilidade em regiões com padrões climáticos distintos. Este projeto, portanto, contribui para o avanço na aplicação de aprendizado de máquina em previsão climática, destacando tanto seus potenciais quanto suas limitações em cenários reais e diversos.

2. Descrição do Problema

Prever a temperatura máxima é um problema desafiador, pois depende de múltiplas variáveis climáticas que interagem de forma complexa. Modelos como regressão linear, Random Forest e KNN foram utilizados para modelar essa relação, além de clusterização com K-Means para identificar padrões.

3. Dataset

O dataset contém dados climáticos das cidades de Seattle e Budapeste, com as seguintes colunas:

- **date**: Data de registro das observações.
- **MaxTemp**: Temperatura máxima registrada no dia (em °C).
- **MinTemp**: Temperatura mínima registrada no dia (em °C).
- **weather_rain**: Indicador binário (0 ou 1) para dias com ocorrência de chuva.
- **weather_sun**: Indicador binário (0 ou 1) para dias predominantemente ensolarados.
- **weather_snow**: Indicador binário (0 ou 1) para dias com ocorrência de neve.

Preparação dos Dados:

- Tratamento de valores ausentes (`NaN`) substituindo-os pela média das colunas numéricas.
- Normalização de variáveis contínuas (**MinTemp** e **MaxTemp**) para o intervalo [0, 1] utilizando *Min-Max Scaling*.
- Categorização de **MaxTemp** em três classes: Baixa, Moderada e Alta, facilitando análises.

4. Metodologia

A metodologia adotada neste projeto foi dividida em algumas etapas para garantir a análise e modelagem dos dados climáticos. Essas etapas incluíram desde o carregamento e preparação dos dados até a aplicação e avaliação de modelos e de clusterização.

Inicialmente, os datasets das cidades de Seattle e Budapeste foram carregados e analisados para garantir sua adequação aos objetivos do projeto. Durante essa etapa, foi realizada uma análise descritiva inicial, com a geração de estatísticas básicas além de visualizações gráficas, como histogramas. Esses procedimentos permitiram identificar padrões e relações entre as variáveis e possíveis valores ausentes.

Os valores ausentes (`NaN`) foram tratados substituindo-os pela média das colunas numéricas relevantes (GéRON, 2019), como **MinTemp** e **MaxTemp**, para evitar viés nos modelos. A coluna **date** foi ignorada nesse processo, pois não possui influência direta na previsão da temperatura. Para as variáveis categóricas binárias, como **weather_rain**,

weather_snow e **weather_sun**, os valores ausentes foram preenchidos com o modo de cada coluna, garantindo consistência nos dados.

As variáveis contínuas **MinTemp** e **MaxTemp** foram normalizadas para o intervalo [0, 1] (Scikit-learn.org) utilizando o método de *Min-Max Scaling*. Essa normalização foi aplicada para alinhar a escala das variáveis, evitando que diferenças de magnitude afetassem o desempenho dos algoritmos de aprendizado de máquina.

Para facilitar a aplicação de modelos de classificação, a variável contínua **MaxTemp** foi categorizada em três classes: *Baixa*, *Moderada* e *Alta*, com base em intervalos definidos a partir da distribuição dos dados. Essa nova variável categórica, chamada **MaxTempCategory**, permitiu realizar análises qualitativas e foi utilizada no treinamento e na avaliação de modelos de classificação.

Na modelagem preditiva, foram aplicados diferentes algoritmos de aprendizado supervisionado para prever e classificar a temperatura máxima. A regressão linear foi utilizada para prever os valores contínuos de **MaxTemp**, considerando variáveis independentes como **MinTemp** e os indicadores climáticos. Modelos de classificação, como Random Forest e KNN (*K-Nearest Neighbors*), foram utilizados para prever a categoria da temperatura máxima. O número de vizinhos no KNN foi otimizado por meio de validação cruzada, e a Random Forest se destacou pela robustez em lidar com dados heterogêneos.

Além disso, foi aplicada a técnica de clusterização K-Means para explorar padrões ocultos nos dados. O método do cotovelo foi utilizado para determinar o número ideal de clusters, enquanto a interpretação visual dos agrupamentos ajudou a identificar padrões climáticos característicos de cada cidade.

5. Resultados

5.1. Modelos de Classificação

Os resultados das classificações para Seattle e Budapeste estão apresentados abaixo:

Random Forest - Seattle

Class	Precision	Recall	F1-Score
Alta	1.00	1.00	1.00
Baixa	1.00	1.00	1.00
Moderada	1.00	1.00	1.00

KNN - Seattle

Class	Precision	Recall	F1-Score
Alta	1.00	0.97	0.99
Baixa	0.99	1.00	0.99
Moderada	0.99	0.99	0.99

Random Forest - Budapeste

Class	Precision	Recall	F1-Score	Support
Alta	1.00	1.00	1.00	94
Baixa	1.00	1.00	1.00	81
Moderada	1.00	1.00	1.00	118

KNN - Budapeste

Class	Precision	Recall	F1-Score	Support
Alta	1.00	0.99	0.99	94
Baixa	1.00	1.00	1.00	81
Moderada	0.99	1.00	1.00	118

5.2. Clusterização com K-Means

Foi utilizado o método do cotovelo para determinar o número ideal de clusters, revelando agrupamentos distintos em padrões climáticos.

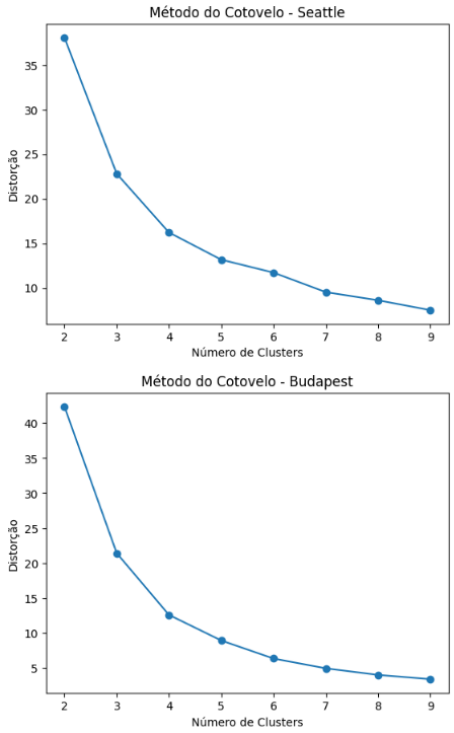

Figura 1. Clusterização

6. Conclusão

Apesar do sucesso obtido, algumas limitações foram identificadas. A generalização dos modelos para regiões com climas significativamente diferentes, como áreas tropicais, mostrou-se desafiadora devido à ausência de variáveis mais representativas. Elementos como umidade relativa, pressão atmosférica e radiação solar poderiam enriquecer o dataset e oferecer uma descrição mais completa das condições que influenciam a temperatura máxima. Essas variáveis têm o potencial de melhorar a capacidade preditiva dos modelos e permitir uma maior aplicabilidade em diferentes contextos climáticos.

Além disso, a escolha de Seattle e Budapeste como cenários de estudo proporcionou uma análise diversificada entre dois climas distintos, mas também revelou a importância de adaptar os modelos às características específicas de cada região. Essa flexibilidade será essencial para expandir a aplicação das soluções desenvolvidas neste projeto para outras localidades e condições meteorológicas.

Portanto, o trabalho não apenas alcançou os objetivos propostos, mas também lançou bases para futuras melhorias e investigações. Implementar variáveis adicionais,

explorar técnicas mais avançadas, como redes neurais, e realizar testes em outras regiões climáticas são passos promissores para ampliar o impacto e a utilidade prática deste estudo.

7. Endereço GitHub e Endereço do vídeo no YoutTube

Endereço Github: https://github.com/MatosLeo03/Projeto2-InteligenciaArtificial

Endereço Youtube:

8. Referências

- Scikit-learn: Machine Learning in Python. Disponível em: `https://scikit-learn.org`.
- GÉRON, Aurélien. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*, 2ª ed. Sebastopol: O'Reilly, 2019.
- Weather Dataset. Disponível em: `https://www.kaggle.com`.