RESIDÊNCIA EM SOFTWARE
BAHIA + TECNOLOGIA + EMPREENDEDORISMO

Relatório Técnico: Implementação e Análise do Algoritmo k-Nearest Neighbors (kNN) Aplicado ao Instagram

Grupo 8: Vitor Miranda Sousa e Wallisson da Silva Dias

Data de entrega: 17/11/2024

Resumo

Este estudo aplicou o algoritmo k-Nearest Neighbors (kNN) para prever o *influence score* de influenciadores do Instagram, métrica muito importante para o desenvolvimento de estratégias de marketing digital. A base de dados, contendo informações como número de seguidores, engajamento médio e localização geográfica, foi tratada para normalizar valores e eliminar inconsistências. Após uma análise exploratória, o kNN foi implementado com validação cruzada e ajustes iniciais de hiperparâmetros, seguidos de otimização por Otimização Bayesiana para maior eficiência. Os resultados apresentaram um Erro Absoluto Médio (MAE) de 4.6 e um Erro Quadrático Médio (RMSE) de 6.8, demonstrando boa precisão, mesmo com uma base de dados reduzida e foi identificada uma correlação fraca entre seguidores e curtidas, sugerindo comportamentos não lineares no engajamento. Conclui-se que o kNN é eficaz para prever influências em redes sociais, destacando o papel de práticas avançadas no aprimoramento de modelos preditivos.

Introdução

Atualmente, o marketing digital desempenha um papel essencial nas estratégias de negócios, especialmente no uso de influenciadores do Instagram como canais de divulgação. De acordo com estudos recentes, o sucesso de campanhas de marketing depende significativamente da escolha assertiva de influenciadores capazes de engajar e impactar o público-alvo (BRANDÃO *et al*., 2022). Nesse contexto, prever o *influence score*, métrica que reflete o impacto de um influenciador nas redes sociais, torna-se fundamental para otimizar os investimentos em marketing.

O algoritmo k-Nearest Neighbors (kNN) foi selecionado neste estudo devido à sua simplicidade e eficácia em prever valores contínuos. O kNN é amplamente reconhecido como uma abordagem robusta para problemas de regressão, especialmente em cenários com características não lineares (JAMES *et al*., 2013). Além disso, permite ajustes flexíveis de parâmetros para melhorar a performance com base nos dados disponíveis, sendo uma escolha versátil para análise preditiva em conjuntos de dados reais.

O conjunto de dados utilizado inclui informações detalhadas sobre influenciadores do Instagram, como número de postagens, seguidores, engajamento médio nos últimos 60 dias e número total de curtidas. A localização geográfica também é fornecida, possibilitando análises por continente. Esses dados, obtidos de uma fonte pública confiável, foram tratados para normalizar valores, converter sufixos numéricos (como "k" e "m") e lidar com valores ausentes, garantindo consistência para a aplicação do modelo.

Metodologia

Análise Exploratória

A análise exploratória de dados foi conduzida para compreender as características do conjunto de dados e identificar padrões ou inconsistências. Inicialmente, foram analisadas variáveis

como número de postagens, seguidores, engajamento médio e total de curtidas, avaliando suas relações com o *influence score*. Valores ausentes foram tratados com técnicas apropriadas, como substituição por médias ou medianas, dependendo da distribuição das variáveis.

Para padronização dos dados, valores representados com sufixos numéricos (como "k" para milhares e "m" para milhões) foram convertidos em números absolutos. Além disso, visualizações gráficas, como histogramas, gráficos de dispersão e mapas de calor, foram utilizadas para identificar correlações entre variáveis e possíveis outliers.

Implementação do Algoritmo kNN

Na etapa de implementação, o algoritmo kNN foi configurado para prever o *influence score* com base em variáveis numéricas selecionadas. O conjunto de dados foi dividido em treino (60%) e teste (40%). A variável country, originalmente categórica, foi transformada para uma classificação por continentes, agrupando influenciadores por regiões geográficas (África, América e Europa), a fim de identificar diferenças regionais e todas as variáveis numéricas foram normalizadas para evitar que magnitudes discrepantes impactassem negativamente o cálculo de distâncias no espaço multidimensional. Além disso, nessa mesma variável (country), havia um número alto de valores nulos que foram apagados, resultando em uma base de dados ainda menor.

Validação e Ajuste de Hiperparâmetros

O modelo foi avaliado utilizando validação cruzada k-fold com 5 divisões, garantindo maior confiabilidade e robustez nos resultados. Durante o ajuste dos hiperparâmetros, foram testados valores de k variando entre 1 e 20 para determinar o número ideal de vizinhos. Para melhorar a precisão, foram considerados dois esquemas de ponderação dos vizinhos: pesos uniformes e baseados na distância.

Após a validação inicial, foi realizada uma Otimização Bayesiana para refinar os hiperparâmetros. Essa abordagem, mais eficiente que o GridSearch, permitiu encontrar os melhores parâmetros sem testar exaustivamente todas as combinações.

Os hiperparâmetros ajustados incluíram:

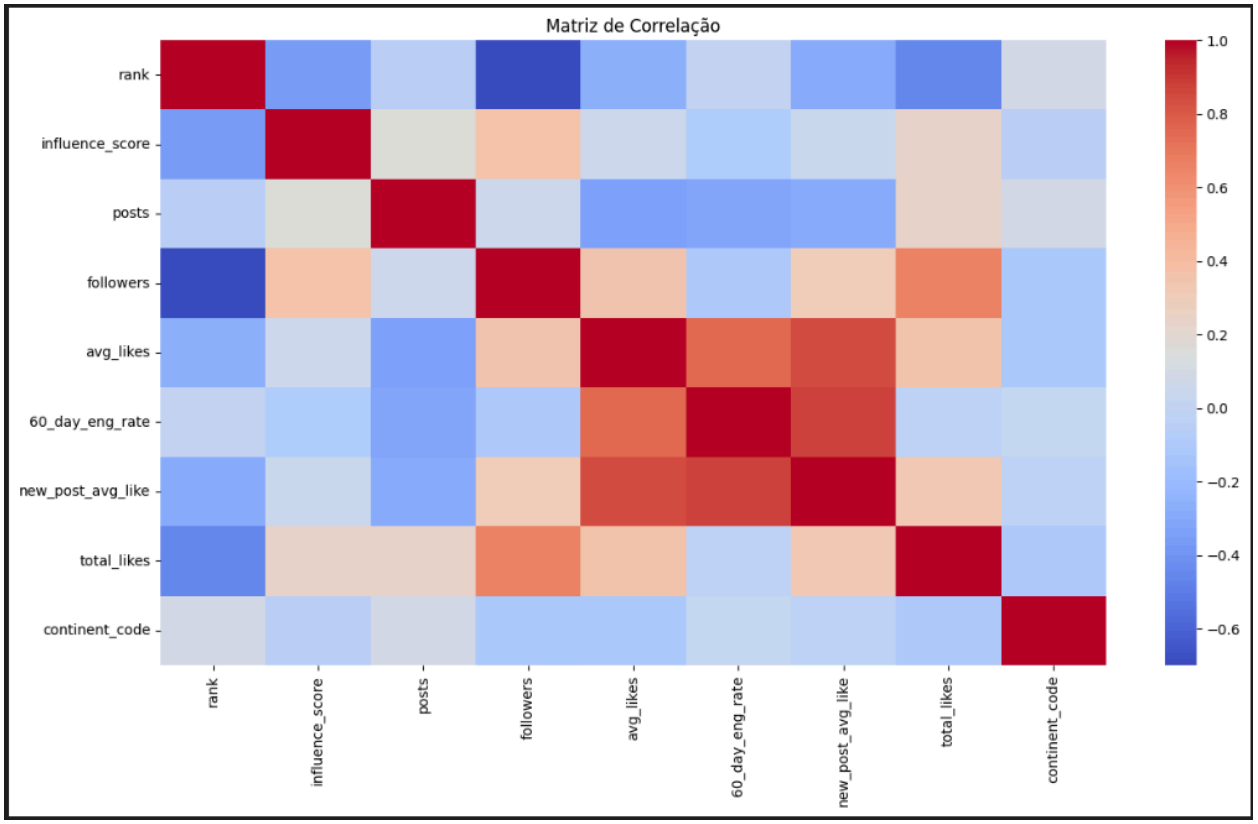
- Número de vizinhos (n_neighbors)
- Pesos das amostras (weights)
- Algoritmo de busca (algorithm)
- Tamanho da folha (leaf_size)
- Tipo de distância (p)
- Métrica (metric)

A escolha dessa abordagem visou aplicar boas práticas de Machine Learning, mesmo considerando o pequeno tamanho do conjunto de dados.

Resultados

Uma análise inicial foi realizada com base na matriz de correlação, permitindo identificar as relações entre as variáveis do conjunto de dados. A matriz de correlação é uma ferramenta importante para visualizar como as variáveis se relacionam entre si, especialmente no contexto de modelagem preditiva.

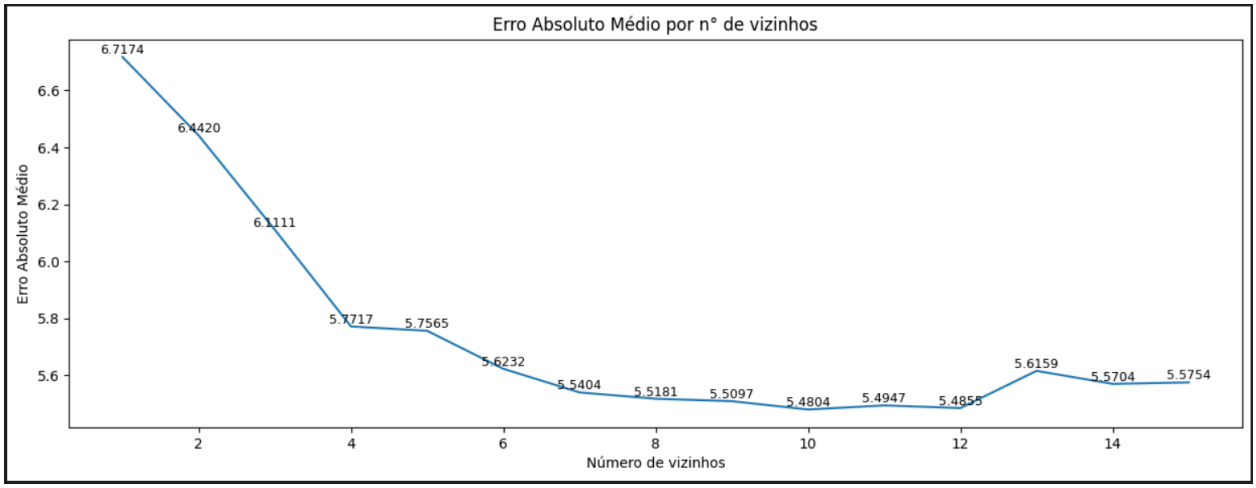
Gráfico 1: Matriz de correlação das variáveis do conjunto de dados.



A matriz de correlação ilustra as relações entre as variáveis contínuas do conjunto de dados, permitindo identificar padrões relevantes para o modelo.

O desempenho inicial do modelo k-Nearest Neighbors (kNN) foi testado considerando apenas diferentes valores para o número de vizinhos (k). Nesse teste, o menor erro absoluto médio (MAE) foi de 5.4804, com k = 10, como ilustrado na Figura 1.

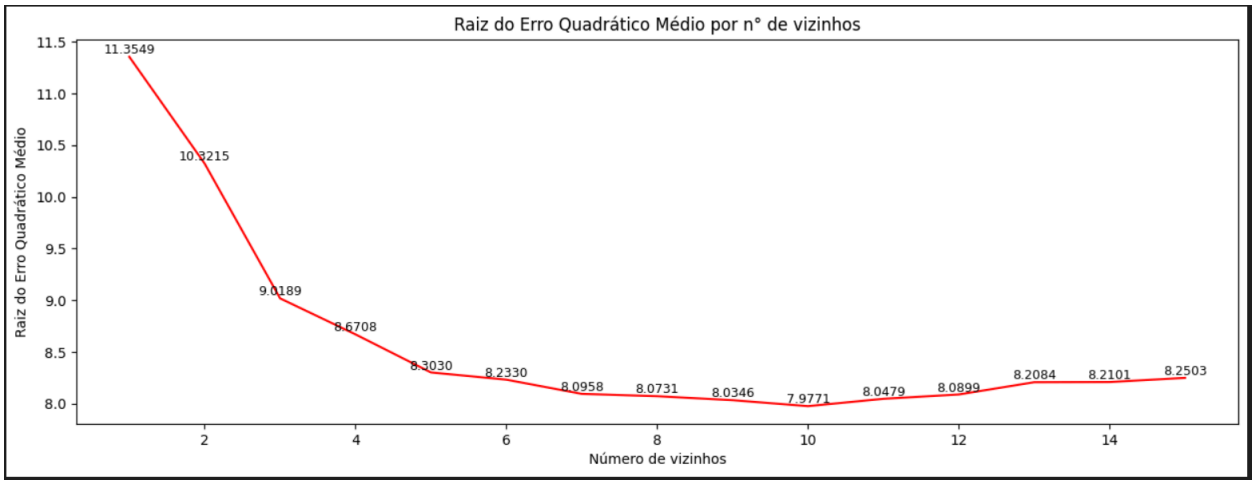
Figura 1: Erro Absoluto Médio (MAE) em função do número de vizinhos (k).



- **Melhor MAE:** 5.4804, alcançado com **k = 10**.

A análise da raiz do erro quadrático médio (RMSE) para diferentes valores de k revelou que a menor quantidade de erro também ocorreu em k = 10, conforme mostrado na Figura 2.

Figura 2: Raiz do Erro Quadrático Médio (RMSE) em função do número de vizinhos (k).



- **Melhor RMSE:** O valor de RMSE 7.9771 foi minimizado com **k = 10**, confirmando a eficácia dessa configuração para o modelo.

Esses valores indicam que **k = 10** foi a melhor escolha para o modelo, balanceando precisão e generalização.

Para aprimorar a escolha dos hiperparâmetros, foi utilizado o método de Otimização Bayesiana. Esta abordagem é mais eficiente que o GridSearch, testando combinações de hiperparâmetros de maneira inteligente, com menor custo computacional. Embora o GridSearch fosse uma alternativa viável dada a pequena base de dados, optou-se pela Otimização Bayesiana para adotar práticas mais avançadas de Machine Learning.

A Otimização Bayesiana realizou 50 iterações e identificou a seguinte configuração de hiperparâmetros como a mais eficiente:

- **n_neighbors:** 18
- **weights:** 'distance'
- **algorithm:** 'ball_tree'
- **leaf_size:** 46
- **p:** 2
- **metric:** 'manhattan'

Com essa configuração, os resultados obtidos foram:

- **Erro Absoluto Médio (MAE):** 4.645
- **Raiz do Erro Quadrático Médio (RMSE):** 6.797

Discussão

Análise Crítica dos Resultados

Os resultados obtidos a partir do algoritmo kNN trouxeram insights importantes sobre o conjunto de dados e o desempenho do modelo. A matriz de correlação (gráfico 1) mostrou uma relação fraca entre variáveis que, teoricamente, deveriam ter maior dependência, por exemplo, a relação entre followers (seguidores) e avg_likes (média de likes) apresentou uma correlação

positiva fraca, contrariando a expectativa de que o número de seguidores influencia diretamente a média de curtidas, o que sugere que apenas uma pequena parcela dos seguidores realmente interage com os posts, o que pode ser resultado de seguidores inativos ou bots.

Outro ponto relevante foi a ausência de correlação entre a taxa de engajamento dos últimos 60 dias e a taxa de engajamento geral do influenciador, o que pode ser explicado pelo fato de influenciadores muito populares manterem altos níveis de curtidas e engajamento independentemente da frequência de postagem, reforçando a ideia de que o engajamento não está necessariamente vinculado à consistência de posts, mas também à relevância do influenciador no cenário digital.

No que diz respeito à performance do modelo, o gráfico da validação inicial mostrou que o número de vizinhos ideal (k) era 10, proporcionando um erro absoluto médio (MAE) de 5.4804. No entanto, com a aplicação da Otimização Bayesiana, foi possível reduzir o MAE para 4.645 ao ajustar os hiperparâmetros de forma mais inteligente.

Limitações Encontradas

Apesar dos resultados positivos, algumas limitações foram observadas, pois conjunto de dados utilizado possui tamanho reduzido, o que pode afetar a generalização do modelo para outros contextos, essa limitação também implica que a validação e a otimização podem ter sido influenciadas por flutuações nos dados, não refletindo perfeitamente o desempenho em cenários mais amplos.

Além disso, o algoritmo kNN, por sua natureza baseada em distâncias, é sensível à escala e ao número de dimensões do conjunto de dados e apesar da normalização ter sido aplicada, a inclusão de muitas variáveis pouco correlacionadas pode ter introduzido ruído no modelo, reduzindo sua precisão.

Impacto das Escolhas no Desempenho

A escolha inicial de utilizar validação cruzada e testar valores de K foi essencial para identificar o número ideal de vizinhos. Contudo, a utilização do GridSearch para explorar combinações de hiperparâmetros apresentaria um custo computacional elevado, pois é um método que no contexto de big data é ineficiente e portanto, a substituição dessa abordagem pela Otimização Bayesiana mostrou-se uma decisão estratégica, pois otimiza a busca por configurações ideais de maneira mais eficiente.

Os resultados da Otimização Bayesiana, utilizando os hiperparâmetros n_neighbors = 18, weights = 'distance', algorithm = 'ball_tree', entre outros, trouxeram um desempenho semelhante ao modelo anterior, mas com uma abordagem mais sofisticada e alinhada às melhores práticas de Machine Learning., o que sugere que o aprendizado e a aplicabilidade do kNN podem ser aprimorados ao explorar metodologias modernas de ajuste de parâmetros.

Logo, os resultados destacam a utilidade do kNN para problemas de regressão com conjuntos de dados reais, mas também apontam limitações, como sensibilidade a dados inconsistentes e alta dependência da qualidade da parametrização. As escolhas feitas ao longo do projeto, incluindo o uso da Otimização Bayesiana, foram determinantes para a performance final do modelo.

Conclusão

O presente estudo aplicou o algoritmo k-Nearest Neighbors (kNN) para prever o influence score de influenciadores do Instagram, demonstrando a relevância de análises preditivas em contextos de marketing digital e apesar do tamanho reduzido da base de dados, o modelo apresentou um desempenho satisfatório, com um Erro Absoluto Médio (MAE) de 4.6 e um Erro Quadrático Médio (RMSE) de 6.8. Esses valores representam, respectivamente, a média das diferenças absolutas e quadráticas entre os valores reais e previstos, indicando a precisão do modelo no contexto proposto.

A análise mostrou que os valores de *influence score* têm um papel fundamental na publicidade moderna, onde empresas avaliam métricas como engajamento e alcance antes de firmar

parcerias estratégicas com influenciadores. O estudo reforça que, embora a base de dados utilizada possuísse limitações, como tamanho reduzido e distribuição potencialmente enviesada, a abordagem utilizada oferece uma modelagem robusta e alinhada às demandas do setor.

Além disso, foi discutido que essa análise poderia ser ampliada para outros objetivos preditivos, como a classificação geográfica dos influenciadores com base em suas características. No entanto, tal abordagem exigiria cuidados adicionais devido ao desbalanceamento natural das classes, já que a maioria dos usuários está concentrada em regiões como EUA e Europa e essa expansão, embora interessante do ponto de vista acadêmico, seria menos alinhada ao foco em negócios.

Em suma, este trabalho demonstrou a eficácia do kNN como uma ferramenta para análises preditivas em redes sociais e destacou a importância de abordagens mais eficientes, como a Otimização Bayesiana, na escolha de hiperparâmetros. Com bases de dados maiores e mais diversificadas, a aplicabilidade e precisão do modelo podem ser significativamente aprimoradas, ampliando seu impacto no campo do marketing digital.

Referências

BRANDÃO, A.; SILVA, M.; ALMEIDA, J. Estratégias de marketing digital no uso de influenciadores digitais. São Paulo: Editora Marketing Brasil, 2022.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An introduction to statistical learning: with applications in R. New York: Springer, 2013.