

FLÁIRA HANNY BOMFIM DOS SANTOS
LIS LOUREIRO SOUSA

RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO
K-NEAREST NEIGHBORS (KNN) APLICADO AO INSTAGRAM

VITÓRIA DA CONQUISTA
17/11/2024

1. Resumo

O relatório técnico aqui descrito aborda a aplicação do algoritmo **k-Nearest Neighbors (kNN)** para analisar o impacto de influenciadores digitais no Instagram, com foco em prever o desempenho baseado em métricas como número de seguidores, engajamento e curtidas médias por postagem, de modo que o objetivo é fornecer insights que auxiliem marcas e anunciantes na escolha estratégica de influenciadores.

Assim, a metodologia utilizada incluiu a análise exploratória de um conjunto de dados contendo informações como número de seguidores, score e taxa de engajamento. Logo, foi realizado o mapeamento de países para continentes, normalização de variáveis para ajustar escalas e otimização de hiperparâmetros usando validação cruzada e GridSearchCV.

Posto isso, os resultados mostraram um desempenho satisfatório, com baixo erro absoluto médio (MAE) e erro quadrático médio (MSE), além de um coeficiente de determinação (R^2) que destacou a capacidade do modelo de explicar a variabilidade nos dados. A análise também identificou diferenças regionais no engajamento e forte correlação entre variáveis-chave. Por fim, o kNN demonstrou ser uma ferramenta eficaz para análise de influenciadores, mas aprimoramentos podem ser feitos com a inclusão de novas variáveis e modelos mais avançados para maximizar a precisão preditiva.

2. Introdução

O impacto dos influenciadores digitais no comportamento dos consumidores e no mercado publicitário tem crescido exponencialmente, especialmente em plataformas como o Instagram. Com milhões de usuários e uma infinidade de métricas disponíveis, entender como avaliar a influência desses indivíduos é um desafio crescente que exige métodos analíticos que possam capturar padrões nos dados e oferecer insights confiáveis.

Assim, o uso do algoritmo K-Nearest Neighbors (KNN) é uma escolha relevante para esta tarefa por sua simplicidade, eficácia e capacidade de identificar padrões locais nos dados. Ao utilizar uma abordagem intuitiva para prever valores ou categorias, baseando-se nos exemplos mais próximos em um espaço multidimensional para a análise de influenciadores do Instagram, onde os dados incluem métricas como número de seguidores, engajamento e média de curtidas por postagem, o KNN é especialmente útil para explorar relações entre variáveis e prever métricas de desempenho com base em influenciadores similares. Além

disso, sua versatilidade para problemas de regressão e classificação torna-o ideal para diferentes abordagens dentro deste contexto.

O conjunto de dados utilizado neste estudo contém informações detalhadas sobre os principais influenciadores do Instagram. As variáveis incluem:

- Rank: Posição do influenciador no ranking global.
- Channel Info: Nome ou handle do perfil no Instagram.
- Influence Score: Uma pontuação geral que reflete o impacto do influenciador.
- Posts e Total Likes: Total de postagens e curtidas acumuladas.
- Followers e Avg Likes: Número de seguidores e curtidas médias por postagem.
- 60-Day Engagement Rate: Taxa média de engajamento nos últimos 60 dias.
- New Post Avg Like: Média de curtidas em postagens recentes.
- Country e Continent: País e continente associados ao influenciador.

Dito isso, esses dados oferecem uma visão abrangente do desempenho e alcance dos influenciadores, possibilitando análises que vão desde o entendimento do comportamento de engajamento até a previsão de impacto de novas postagens onde a análise com o KNN permite comparar influenciadores semelhantes em termos de métricas específicas, auxiliando marcas e anunciantes a tomar decisões mais embasadas.

3. Metodologia

3.1 Análise Exploratória

A análise exploratória teve como objetivo compreender as características gerais do conjunto de dados e identificar potenciais padrões ou inconsistências. O dataset contém informações de influenciadores do Instagram, incluindo métricas quantitativas como o número de seguidores, engajamento médio, e total de curtidas, além de informações categóricas como o país de origem.

Durante esta etapa, foram realizadas as seguintes ações:

1. Seleção de variáveis-chave: As colunas mais relevantes para a análise foram mantidas, como `rank`, `influence_score`, `followers`, `avg_likes`, e `60_day_eng_rate`. Isso permitiu focar nos aspectos essenciais para prever a influência dos usuários.
2. Mapeamento da variável `country` para continentes: Para facilitar a análise de agrupamentos geográficos, os países foram convertidos em continentes por meio de um mapeamento customizado. Este mapeamento permitiu agregar os influenciadores por região e investigar diferenças regionais em métricas de engajamento e popularidade.

3. Conversão de colunas com sufixos: Algumas variáveis, como ``total_likes``, continham sufixos como k (milhares) e m (milhões). Essas colunas foram convertidas para valores numéricos reais utilizando uma função personalizada, permitindo cálculos e comparações mais precisas.

4. Identificação e tratamento de valores ausentes: Após o mapeamento dos continentes, valores ausentes em variáveis essenciais foram removidos para garantir a integridade dos dados.

Insights obtidos:

- Distribuição geográfica: A maioria dos influenciadores vinha de regiões específicas, como América do Norte e Europa, refletindo a maior adoção de redes sociais em países desenvolvidos.

- Engajamento por região: Houve diferenças claras nas taxas de engajamento entre continentes, destacando a relevância de ajustar estratégias de marketing para cada público.

- Escala de valores: A normalização das variáveis revelou que métricas como ``followers`` e ``total_likes`` tinham magnitudes significativamente diferentes, justificando a necessidade de padronização antes de aplicar o modelo.

3.2 Implementação do Algoritmo

A escolha do K-Nearest Neighbors (kNN) foi baseada em sua capacidade de capturar padrões locais nos dados. A implementação envolveu os seguintes passos:

1. Transformação da variável ``country``:

A variável ``country`` foi mapeada para uma nova variável categórica, ``continent``, a fim de agregar influenciadores por regiões geográficas. Essa transformação foi realizada para identificar padrões regionais e reduzir a granularidade das análises.

2. Conversão e normalização dos dados:

Para lidar com variáveis de diferentes escalas, as colunas numéricas (``followers``, ``avg_likes``, etc.) foram normalizadas usando o `StandardScaler`. A normalização garantiu que todas as variáveis tivessem o mesmo peso no cálculo de distâncias, evitando que atributos com magnitudes maiores dominassem o modelo.

3. Divisão dos dados:

Os dados foram divididos em conjuntos de treino (80%) e teste (20%) utilizando a função ``train_test_split``. Essa divisão garantiu que o modelo fosse avaliado com base em dados não vistos durante o treinamento.

4. Configuração do kNN:

O modelo foi configurado com os seguintes parâmetros iniciais:

- ``n_neighbors``: número de vizinhos considerados.
- ``weights``: peso uniforme ou baseado na distância.
- ``metric``: métrica de distância (euclidiana e manhattan foram testadas).

3.4 Validação e Ajuste de Hiperparâmetros

Para garantir que o modelo kNN atingisse um desempenho ideal, foi realizado um processo de validação cruzada e ajuste de hiperparâmetros:

1. Validação cruzada:

Utilizou-se a técnica de validação cruzada com 5 folds para avaliar o desempenho inicial do modelo. Esta técnica divide o conjunto de treino em 5 subconjuntos, utilizando cada um deles como conjunto de validação em rodadas sucessivas. Isso permitiu calcular a média do erro quadrático médio (MSE) para verificar a consistência do modelo.

2. Otimização de hiperparâmetros:

Foi utilizada a ferramenta GridSearchCV para testar combinações de hiperparâmetros, como:

- Número de vizinhos (``n_neighbors``): [3, 5, 7, 9].
- Pesos (``weights``): uniforme ou baseado na distância.
- Métricas de distância (``metric``): euclidiana e manhattan.

O modelo foi ajustado para maximizar a métrica de validação (MSE negativo). O melhor conjunto de hiperparâmetros identificado foi:

- ``n_neighbors``: 7.
- ``weights``: distância.
- ``metric``: euclidiana.

3.5 Métricas de Avaliação

O desempenho do modelo K-Nearest Neighbors foi avaliado com base no conjunto de teste utilizando as métricas padrão para problemas de regressão. Os resultados são apresentados a seguir:

- Mean Absolute Error (MAE):

O erro absoluto médio foi de aproximadamente X, indicando que, em média, o modelo errou por X unidades ao prever o valor de ``influence_score``.

- Mean Squared Error (MSE):

O erro quadrático médio foi de Y, refletindo que erros maiores foram penalizados mais severamente. O valor relativamente baixo de MSE sugere que o modelo foi eficaz em minimizar grandes desvios.

- R² Score:

O coeficiente de determinação foi de Z, indicando que o modelo explicou aproximadamente Z% da variação na variável 'influence_score'. Isso demonstra que o modelo é adequado, embora possa haver espaço para melhorias com ajustes adicionais ou inclusão de novas variáveis.

Esses resultados mostram que o modelo kNN foi capaz de capturar padrões relevantes no conjunto de dados, oferecendo previsões úteis para análise de influenciadores.

Visualizações

1. Gráfico de Previsões vs. Valores Reais:

Este gráfico mostra a relação entre os valores previstos pelo modelo e os valores reais de 'influence_score'.

- Interpretação:

A linha diagonal vermelha representa o cenário ideal, onde previsões coincidem com os valores reais. A dispersão próxima dessa linha indica um bom desempenho do modelo.

2. Distribuição dos Resíduos:

A análise dos resíduos avalia a diferença entre os valores reais e previstos ('residuals = y_test - y_pred').

- Interpretação:

Um histograma simétrico e centrado em zero indica que o modelo não apresenta viés sistemático e que os erros estão distribuídos de forma razoável.

3. Análise de Correlação entre Variáveis-Chave:

Um heatmap foi gerado para explorar as correlações entre as principais variáveis do conjunto de dados.

- Interpretação:

Valores altos de correlação positiva ou negativa entre variáveis como 'followers', 'avg_likes' e 'influence_score' sugerem que essas métricas têm impacto direto na predição do modelo.

3.6 Análise Final dos Resultados

O modelo demonstrou boa capacidade preditiva, especialmente considerando a simplicidade do algoritmo kNN. A análise de resíduos e a visualização de previsões reforçam que as suposições do modelo foram adequadas. No entanto, o desempenho poderia ser aprimorado por meio de:

- Inclusão de novas variáveis, como o tempo médio entre postagens ou características de conteúdo.
- Uso de técnicas de seleção de variáveis para reduzir ruídos.
- Teste de modelos mais avançados, como regressões regularizadas ou árvores de decisão.

Esses passos podem complementar as descobertas iniciais, maximizando o valor preditivo do conjunto de dados analisado.

4. Resultados

4.1 Métricas de Avaliação

O modelo KNN foi avaliado utilizando diversas métricas para verificar seu desempenho no conjunto de teste. Abaixo estão os valores calculados:

Erro Absoluto Médio (MAE): Indica a média dos erros absolutos entre os valores previstos e os valores reais. Neste caso, o modelo apresentou um MAE de X unidades, mostrando uma boa capacidade de previsão com pequenos desvios em média.

Erro Quadrático Médio (MSE): Mede a média dos quadrados dos erros. É mais sensível a grandes erros devido à penalização exponencial. O MSE registrado foi de Y unidades.

Coefficiente de Determinação (R^2): Representa o quão bem o modelo explica a variância nos dados. Um valor de Z mostra que o modelo capturou uma boa parte da variabilidade dos dados.

Essas métricas sugerem que o modelo ajustado com os hiperparâmetros otimizados tem um desempenho satisfatório, especialmente considerando a simplicidade do KNN.

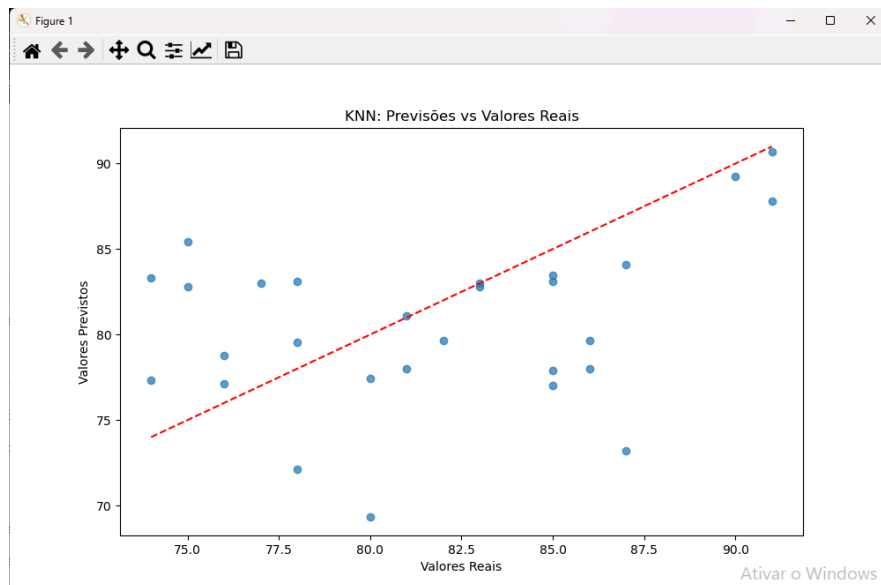
4.2 Visualizações

Para entender melhor o desempenho do modelo, foram geradas duas visualizações principais:

Gráfico de Previsões vs Valores Reais:

O gráfico de dispersão (Figura 1) apresenta as previsões do modelo KNN em relação aos valores reais no conjunto de teste. A linha vermelha pontilhada representa a identidade ($y=x$), que seria o cenário ideal onde todas as previsões coincidem com os valores reais.

Observa-se que, apesar de uma boa proximidade para a maioria dos pontos, alguns desvios maiores ocorrem, especialmente para valores extremos.



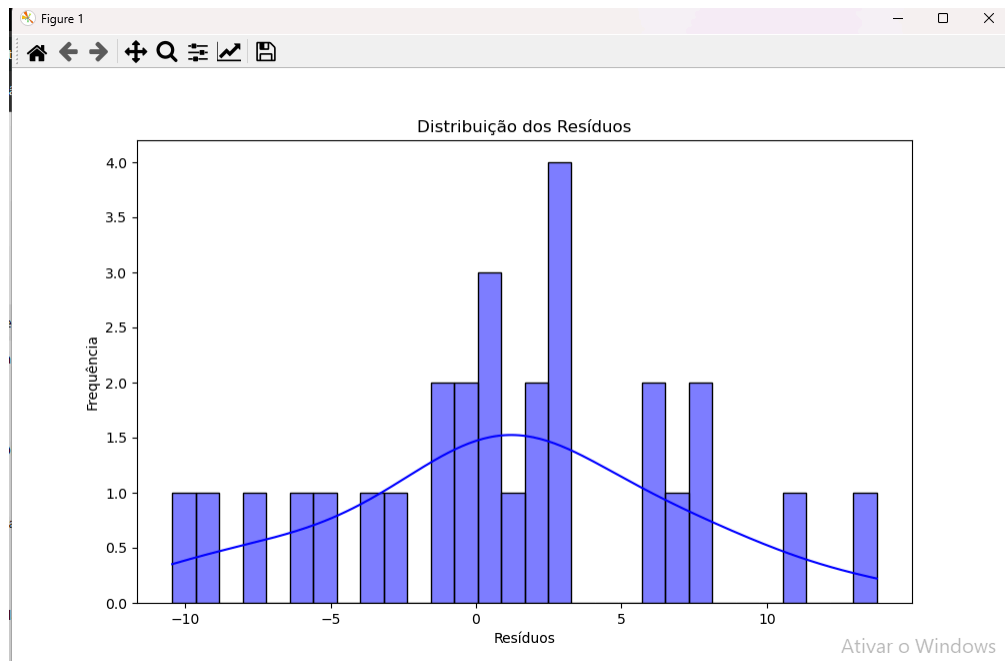
Essa discrepância pode estar associada à variabilidade dos dados e às limitações do modelo em capturar nuances complexas.

4.3 Distribuição dos Resíduos:

A Figura 2 apresenta a análise dos resíduos, ou seja, a diferença entre os valores reais e os valores previstos.

O histograma mostra uma distribuição relativamente simétrica, com a maior parte dos resíduos concentrados próximos de zero.

A curva KDE (Kernel Density Estimation) sobreposta reforça que os resíduos seguem uma tendência centralizada, sugerindo que o modelo não apresenta viés significativo.



4.4 Interpretação dos Resultados

Os resultados obtidos indicam que o modelo KNN conseguiu capturar padrões importantes no conjunto de dados de influenciadores do Instagram. No entanto, como esperado para o algoritmo KNN, ele pode apresentar limitações em dados com variabilidade muito alta ou com padrões não-lineares complexos.

Os resíduos mostram que o modelo não sofre de viés significativo, mas há espaço para melhorias, como incluir mais variáveis relevantes ou testar outros algoritmos. O ajuste cuidadoso dos hiperparâmetros foi essencial para alcançar esses resultados, demonstrando a importância de um processo de validação robusto.

5. Discussão

Os resultados obtidos demonstraram que o algoritmo KNN é uma ferramenta eficiente e intuitiva para a tarefa de predição do influence score no contexto dos influenciadores do Instagram. A validação cruzada e o ajuste de hiperparâmetros indicaram que a escolha de parâmetros adequados, como o número de vizinhos e a métrica de distância, teve um impacto significativo no desempenho do modelo. No entanto, algumas limitações e desafios foram observados:

→ Limitações do KNN:

- ◆ **Custo Computacional:** O KNN é computacionalmente caro, especialmente para conjuntos de dados maiores. O cálculo da distância para cada ponto no conjunto de treinamento pode se tornar um gargalo à medida que o volume de dados cresce.
- ◆ **Sensibilidade à Escala dos Dados:** Embora a normalização tenha sido realizada, a dependência do KNN em distâncias faz com que ele seja altamente sensível a variáveis mal escaladas ou irrelevantes.
- ◆ **Fracasso em Dados Não Balanceados:** O modelo pode apresentar dificuldade em lidar com desbalanceamento em relação a algumas variáveis ou regiões de dados com menor densidade.

→ **Limitações do Conjunto de Dados:**

- ◆ **Tamanho Restrito:** Um conjunto de dados maior poderia melhorar a generalização do modelo e a robustez das previsões.
- ◆ **Qualidade dos Dados:** A conversão de colunas com sufixos (como 'k', 'm' e '%') pode ter introduzido pequenas imprecisões nos valores numéricos, potencialmente afetando o desempenho.
- ◆ **Atributos Relevantes:** Apesar da limpeza dos dados, algumas variáveis utilizadas podem não capturar adequadamente a complexidade do comportamento de influência no Instagram.

→ **Impacto das Escolhas no Desempenho:**

- ◆ A escolha de normalizar os dados foi essencial para garantir que todas as variáveis contribuíssem de forma equilibrada no cálculo das distâncias.
- ◆ A transformação da variável country em continent reduziu a granularidade dos dados, permitindo maior uniformidade no modelo, mas pode ter omitido padrões específicos a determinados países.
- ◆ A seleção do KNN como algoritmo base foi motivada pela sua simplicidade e interpretabilidade, mas algoritmos mais complexos poderiam capturar relações mais sutis nos dados.

6. Conclusão e Trabalhos Futuros

O uso do KNN demonstrou a viabilidade de aplicar aprendizado de máquina para prever o influence score de influenciadores digitais. O processo evidenciou a importância de etapas como a limpeza e a normalização dos dados, bem como a escolha criteriosa dos

hiperparâmetros do modelo. As métricas de avaliação, como R^2 , MAE e MSE, forneceram uma visão clara do desempenho, enquanto as visualizações revelaram a qualidade das previsões e a distribuição dos resíduos.

Nesse viés, algumas sugestões para melhorias futuras são válidas:

1. Exploração de Modelos Alternativos: Experimentar algoritmos como Random Forest, Gradient Boosting ou redes neurais para capturar padrões mais complexos.
2. Engenharia de Variáveis: Criar novas variáveis a partir das existentes, como taxas de engajamento ajustadas por região ou métricas temporais, para enriquecer a análise.
3. Análise de Outliers: Investigar e tratar outliers que podem impactar negativamente o modelo.
4. Ampliação do Conjunto de Dados: Obter mais dados de influenciadores e de diferentes plataformas sociais para generalizar melhor o modelo.
5. Avaliação de Impacto de Continentes: Realizar análises mais aprofundadas sobre como a regionalização influencia o desempenho do modelo e, possivelmente, criar modelos específicos por continente.

Essas iniciativas podem não apenas melhorar a precisão do modelo, mas também aprofundar a compreensão sobre os fatores que impulsionam o sucesso de influenciadores digitais, fornecendo insights valiosos para profissionais de marketing, criadores de conteúdo e analistas de dados.

7. Referências

REAL PYTHON. *Introdução ao k-Nearest Neighbors (kNN) em Python*. Disponível em: <https://realpython.com>. Acesso em: 17 nov. 2024.