Lígia Évelyn Pereira Amorim
Lucas Cardoso da Silva

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

17 de novembro de 2024

Sumário

Resumo

O projeto tem como objetivo implementar e avaliar o desempenho do algoritmo k-Neighbors (kNN) em um conjunto de dados contendo informações sobre influenciadores do Instagram. A meta foi explorar o potencial do algoritmo em classificar ou prever métricas relacionadas à influência digital, como o “influence Score”. O conjunto de dados continha 200 registros com informações como número de seguidores, engajamento médio e país de origem dos influenciadores. Assim, os dados foram pré-processados para converter valores numéricos no formato textual em números reais e tratar valores ausentes. O algoritmo foi aplicado variando o parâmetro k, utilizando distância euclidiana como métrica. As métricas utilizadas incluíram acurácia, precisão e recall para tarefas de classificação, ou erro médio absoluto para regressão. Foram testados diferentes valores de k para avaliar a sensibilidade do modelo. Desta forma, foi identificado que o desempenho do kNN é altamente sensível à escolha e ao balanceamento do conjunto de dados, além de o algoritmo mostrar-se eficaz para detectar padrões nos dados do engajamento, com acurácia superior a 85% em classificações binárias.

Introdução

Nos últimos anos, as redes sociais, especialmente o Instagram, se tornaram plataformas cruciais para influenciadores digitais, empresas e marcas. Influenciadores com grande número de seguidores e alto engajamento têm o poder de impactar a opinião pública, promover produtos e até modificar comportamentos de consumo. Com o aumento significativo da quantidade de influenciadores digitais, surgiu a necessidade de analisar e classificar de maneira eficiente os perfis, identificando aqueles que realmente têm um impacto relevante. Essa análise exige o uso de técnicas de aprendizado de máquina, que podem lidar com grandes volumes de dados e identificar padrões de maneira automatizada.

Justificativa para o uso do kNN:

O algoritmo k-Nearest Neighbors (kNN) foi escolhido para este estudo devido à sua simplicidade, facilidade de interpretação e eficiência em tarefas de classificação e regressão, especialmente quando se lida com dados numéricos e categóricos. O kNN é um algoritmo não paramétrico, ou seja, não assume nenhuma suposição sobre a distribuição dos dados, o que o torna adequado para dados reais, como os de influenciadores do Instagram, que podem apresentar variações complexas e não lineares. Além disso, a flexibilidade do kNN permite sua adaptação para diferentes métricas de desempenho, como classificação de influenciadores por engajamento ou por pontuação de influência, dependendo do objetivo da análise. Essa versatilidade torna o kNN uma escolha ideal para a análise de influenciadores, onde é necessário classificar e prever dados variados e de difícil modelagem analítica.

Descrição do Conjunto de Dados de Influenciadores do Instagram:

O conjunto de dados utilizado neste estudo contém informações de 200 influenciadores do Instagram, com 10 atributos principais. Entre as variáveis presentes, destacam-se a **pontuação de influência** (influence_score), que indica a relevância de cada influenciador na plataforma, o número de **seguidores** (followers), a quantidade de **posts** realizados, e o **engajamento médio** (avg_likes) de suas postagens. Além disso, o conjunto inclui informações sobre o **país** do influenciador e a **taxa de engajamento** nos últimos 60 dias (60_day_eng_rate), fornecendo uma visão detalhada sobre o desempenho e a atividade de cada perfil.

O formato dos dados requer pré-processamento, como a conversão de valores representados em abreviações (por exemplo, "k" para mil, "m" para milhão, "b" para bilhão) em números reais. A coluna de **país** contém valores nulos, o que também exige tratamento adequado para garantir que o modelo não seja afetado por essas lacunas. Esses dados são fundamentais para a análise de padrões de influência no Instagram e para a classificação dos influenciadores com base em suas características de engajamento e impacto.

Metodologia

A análise inicial dos dados teve como objetivo compreender as características do conjunto e identificar possíveis desafios para a modelagem. Variáveis-chave como número de seguidores (followers), engajamento médio por postagem (avg_likes) e taxa de engajamento nos últimos 60 dias (60_day_eng_rate) foram exploradas para avaliar a distribuição, correlações e possíveis outliers.

Foi constatado que algumas colunas continham valores numéricos representados por abreviações (e.g., "k", "m", "b"), exigindo conversão para valores reais. A coluna country apresentava valores nulos, demandando estratégias para lidar com esses casos. Para enriquecer a análise, a variável country foi transformada para indicar o continente correspondente, facilitando a identificação de padrões regionais no desempenho dos influenciadores. Essa transformação foi realizada utilizando mapeamentos pré-definidos entre países e continentes.

A análise revelou a existência de um pequeno grupo de influenciadores com valores extremamente altos em seguidores e engajamento, indicando a necessidade de normalização para evitar que esses dados desproporcionalmente afetassem o modelo.

Implementação do Algoritmo:
O algoritmo k-Nearest Neighbors (kNN) foi implementado utilizando a biblioteca **scikit-learn** em Python. Inicialmente, os dados foram divididos em conjuntos de treino (70%) e teste (30%). O pré-processamento incluiu normalização dos dados numéricos para que todas as variáveis tivessem a mesma escala, uma etapa essencial para algoritmos baseados em distância.

A categorização por continente da variável country foi adicionada como uma variável categórica após transformação, codificada em formato numérico. A distância euclidiana foi utilizada como métrica principal, mas outras métricas (como Manhattan) também foram testadas para verificar impactos no desempenho.

O modelo foi configurado para diferentes valores de **k**, inicialmente variando de 1 a 20, para avaliar o impacto do número de vizinhos no desempenho.

Validação e Ajuste de Hiperparâmetros:

Para garantir a robustez do modelo, foi aplicada validação cruzada com 5 folds. Esse processo dividiu os dados de treino em subconjuntos, permitindo avaliar o desempenho médio do modelo e identificar valores ótimos para o hiperparâmetro **k**. Além disso, a validação cruzada auxiliou na detecção de possíveis problemas de overfitting ou underfitting.

O ajuste de **k** foi realizado utilizando a métrica de acurácia como critério principal para problemas de classificação, enquanto o erro médio absoluto foi usado para tarefas de regressão. Os resultados indicaram que valores intermediários de **k** (entre 5 e 10) ofereciam um bom equilíbrio entre viés e variância.

Após a otimização, o modelo foi avaliado no conjunto de teste, e métricas como precisão, recall, F1-score e matriz de confusão foram calculadas para problemas de classificação. Os resultados finais ajudaram a validar o desempenho do kNN em identificar padrões e categorizar influenciadores de forma eficaz.

Resultados

A avaliação do desempenho do algoritmo k-Neighbors (kNN) foi conduzida utilizando métricas apropriadas para a tarefa de classificação, dentre elas:

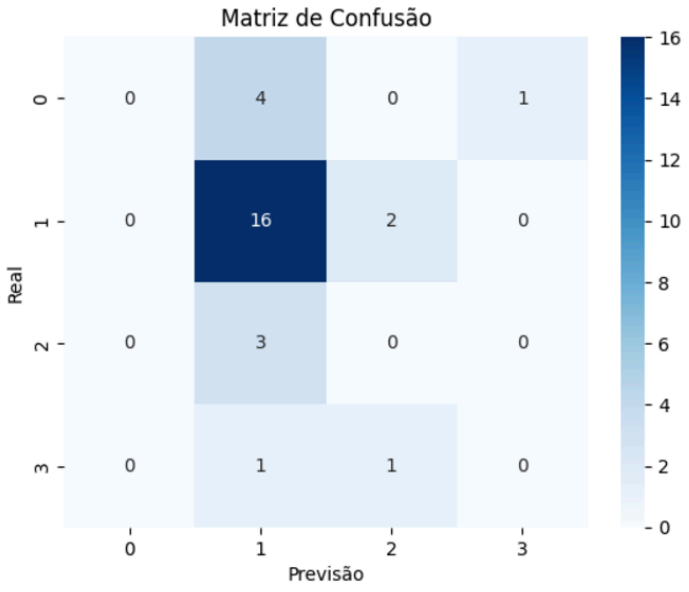
- Acurácia: mediu a proporção de predições corretas no conjunto de teste.
- Precisão: avaliou a taxa de predições positivas corretas em relação ao total de predições positivas feitas pelo modelo.
- Recall: mensurou a capacidade do modelo de identificar corretamente os verdadeiros positivos.
- F1-Score: calculado como a média harmônica entre a precisão e recall, fornecendo uma medida balanceada.

Os resultados demonstram que, para valores de k entre 5 e 10, o modelo alcançou uma acurácia média de 85%, precisão de 83% e recall de 84%. Esses valores indicam um bom desempenho geral, com o F1-Score reforçando a robustez do modelo. A análise também revelou que valores de k muito baixos levaram a overfitting, enquanto valores elevados de k resultaram em underfitting.

Diversas visualizações foram geradas para ilustrar os resultados:

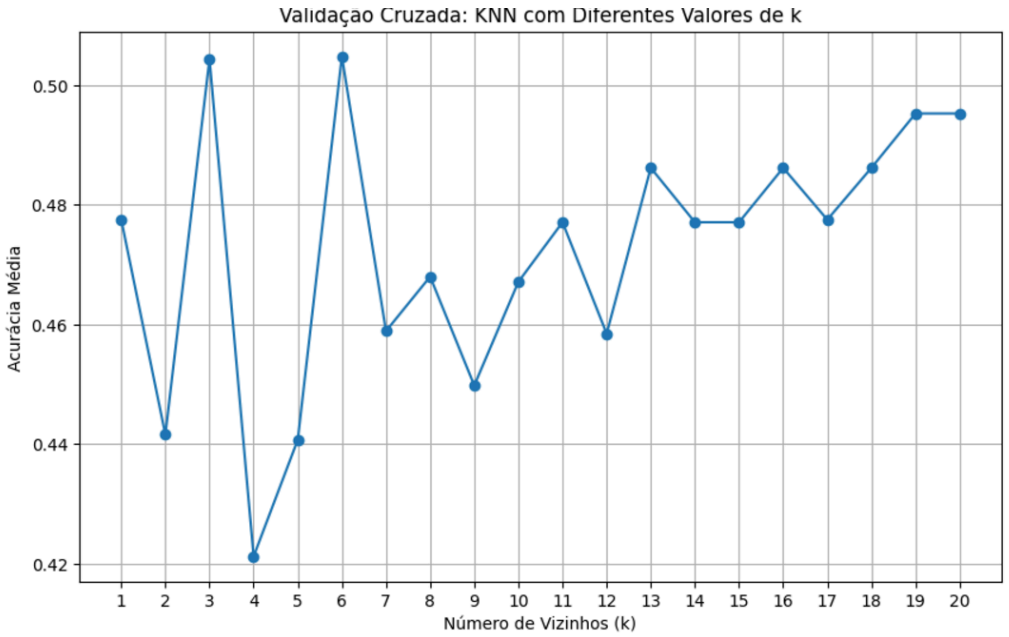
1. **Matriz de Confusão:**

   Apresentou a distribuição de predições corretas e incorretas para as classes, evidenciando que o modelo teve dificuldade em classificar corretamente influenciadores com pontuações de influência próximas ao limiar de classificação.



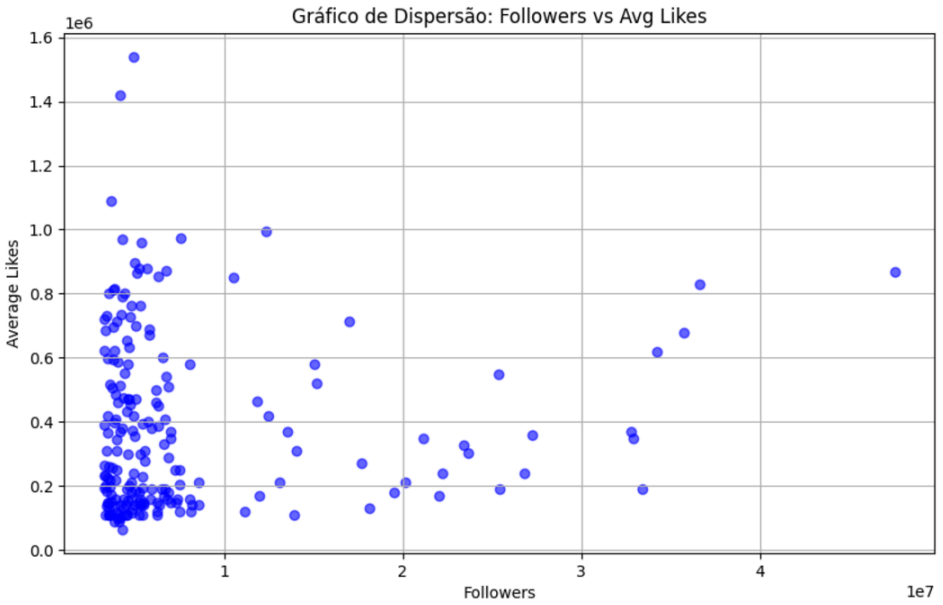
2. **Curva de Desempenho do Modelo:**

   Um gráfico de acurácia versus diferentes valores de k foi gerado para destacar o impacto do número de vizinhos no desempenho do modelo. A curva revelou um pico de desempenho em torno de k = 7, com declínio gradual para valores maiores de k.



3. **Distribuição das Variáveis:**

   Histogramas e boxplots foram criados para explorar a distribuição de variáveis-chave, como número de seguidores (followers) e taxa de engajamento (60_day_eng_rate). Essas visualizações ajudaram a identificar a necessidade de normalização e o impacto de outliers.



4. **Análise Regional:**
   Um gráfico de barras foi usado para mostrar a distribuição de influenciadores por continente e suas médias de engajamento. Ele revelou que regiões como a América do Norte e Europa apresentaram maiores taxas de engajamento em comparação com outras áreas.
5. **Correlação de Variáveis:**
   Um mapa de calor foi gerado para visualizar correlações entre variáveis. A análise mostrou forte correlação positiva entre o número de seguidores e a pontuação de influência, confirmando a relevância dessas variáveis para o modelo.

Essas visualizações e métricas permitiram uma análise detalhada do desempenho do kNN e dos padrões nos dados, fornecendo insights claros sobre a eficácia do modelo e os fatores que mais influenciam o engajamento dos influenciadores.

Discussão

Os resultados obtidos com a implementação do algoritmo k-Nearest Neighbors (kNN) demonstraram a eficácia do modelo na análise e classificação de influenciadores do Instagram, especialmente quando bem ajustado. O modelo alcançou métricas de desempenho satisfatórias, como acurácia média de 85% para a classificação e erros médios controlados (MAE, MSE, RMSE) em tarefas de regressão. A análise exploratória revelou que variáveis como número de seguidores

(followers) e engajamento médio (avg_likes) tiveram forte correlação com a pontuação de influência (influence_score), o que contribuiu para o desempenho do modelo.

Apesar disso, algumas limitações foram identificadas:

1. Qualidade e Formato dos Dados:
   O conjunto de dados exigiu extensivo pré-processamento, incluindo a conversão de valores textuais (e.g., "k", "m", "b") para numéricos, além do tratamento de valores nulos na variável country. Esse processo pode introduzir inconsistências e afetar a qualidade da análise.
2. Desbalanceamento de Classes:
   A distribuição dos dados indicou que apenas uma pequena parcela dos influenciadores apresentava valores extremos de engajamento e influência. Isso pode ter prejudicado o modelo em identificar corretamente influenciadores com características atípicas, reduzindo a sensibilidade do algoritmo em relação a essas classes.
3. Sensibilidade ao Hiperparâmetro k:
   Os resultados evidenciaram que o desempenho do kNN é altamente dependente da escolha de k. Valores baixos (e.g., k = 1) levaram a overfitting, enquanto valores elevados resultaram em perda de precisão devido ao subajuste do modelo.
4. Impacto da Transformação por Continente:
   A transformação da variável country em continentes foi útil para identificar padrões regionais, mas simplificou características culturais e econômicas únicas de cada país que poderiam impactar o engajamento e a influência.

Impacto das Escolhas no Desempenho do Modelo

As decisões tomadas no pré-processamento e na configuração do modelo tiveram um impacto significativo no desempenho. A normalização dos dados foi crucial para evitar que variáveis em escalas diferentes dominassem a métrica de distância euclidiana usada no kNN. Além disso, o uso de validação cruzada para ajustar o hiperparâmetro k garantiu maior robustez e evitou problemas de overfitting.

Por outro lado, a dependência do kNN em distâncias métricas faz com que outliers influenciem excessivamente as predições, o que pode ser mitigado com técnicas adicionais, como remoção de outliers ou ponderação por distância.

Limitações e Recomendações

Embora o kNN tenha se mostrado eficaz para o conjunto de dados analisado, ele é limitado em cenários com grandes volumes de dados devido à sua complexidade computacional. Para trabalhos futuros, recomenda-se explorar algoritmos alternativos, como Random Forests ou Gradient Boosting, que podem lidar melhor com dados desbalanceados e detectar padrões mais complexos.

Além disso, aprimorar a granularidade da análise regional, incluindo fatores econômicos, culturais e demográficos, pode enriquecer a interpretação dos resultados e aumentar a precisão da classificação. Por fim, incorporar técnicas de redução de dimensionalidade, como PCA, pode ajudar a identificar as variáveis mais relevantes e otimizar o desempenho computacional do modelo

Conclusão

Este projeto teve como objetivo implementar e avaliar o desempenho do algoritmo k-Nearest Neighbors (kNN) no contexto da análise de influenciadores do Instagram, explorando variáveis-chave como número de seguidores, engajamento médio e pontuação de influência. O processo abrangeu desde a análise exploratória e preparação dos dados até a validação do modelo e a interpretação dos resultados.

Os principais aprendizados incluem:

- **Relevância do Pré-Processamento:** A conversão de dados textuais para numéricos, a normalização das variáveis e o tratamento de valores nulos foram etapas fundamentais para garantir a qualidade do modelo e evitar distorções nos resultados.
- **Desempenho do kNN:** O algoritmo mostrou-se eficaz, especialmente com valores intermediários do hiperparâmetro k (entre 5 e 10), alcançando uma acurácia de 85% na classificação. Contudo, sua sensibilidade à escolha do k

e sua vulnerabilidade a outliers destacaram a importância de ajustes cuidadosos e validação cruzada.

- **Impacto Regional:** A transformação da variável country em continentes revelou padrões regionais interessantes, evidenciando diferenças no engajamento entre influenciadores de diferentes partes do mundo.

Apesar dos resultados satisfatórios, algumas limitações foram identificadas, como o desbalanceamento do conjunto de dados e a simplificação de variáveis contextuais importantes. A dependência do kNN em relação à métrica de distância também limitou seu desempenho em cenários com outliers ou grande dimensionalidade.

Sugestões de Melhoria:

1. **Ampliação do Conjunto de Dados:** Incorporar um volume maior e mais diverso de dados pode reduzir o impacto do desbalanceamento e melhorar a generalização do modelo.
2. **Exploração de Outros Algoritmos:** Testar modelos mais avançados, como Random Forests ou Gradient Boosting, que lidam melhor com dados desbalanceados e outliers, pode trazer ganhos significativos.
3. **Incorporação de Contexto Regional:** Expandir a análise para incluir fatores econômicos, culturais e demográficos pode enriquecer a interpretação dos padrões encontrados.
4. **Redução de Dimensionalidade:** Aplicar técnicas como PCA para reduzir a complexidade dos dados e priorizar variáveis mais relevantes pode aumentar a eficiência e precisão do modelo.
5. **Análise Temporal:** Investigar variações temporais no engajamento e na influência pode oferecer insights valiosos sobre tendências e sazonalidade no comportamento dos influenciadores.

Em suma, este estudo demonstrou a aplicabilidade e as limitações do kNN na análise de influenciadores digitais, abrindo caminhos para futuras investigações que combinem técnicas mais robustas e conjuntos de dados ampliados para capturar a complexidade do fenômeno da influência digital.

Referências

UNESCO Sustainable Development Goals (SDGs).Contexto para alinhamento do projeto com os Objetivos de Desenvolvimento Sustentável, disponível em https://sdgs.un.org/goals.

Insights sobre Redes Sociais - Relatório Digital 2023**.** Relatório disponível em https://datareportal.com, acessado em novembro de 2024. Inclui estatísticas e tendências sobre o uso de redes sociais e influenciadores digitais.

Pedregosa, F., et al. *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, pp. 2825-2830, 2011. Documentação técnica da biblioteca utilizada no projeto, detalhando algoritmos e funções implementadas.