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

Johnny Araújo Brandão Pereira
Jaquim Lopes Calmon
17/11/2024

1. Resumo

Este projeto teve como objetivo implementar e avaliar o desempenho do algoritmo k-Nearest Neighbors (kNN) em um conjunto de dados real. A metodologia incluiu a preparação e pré-processamento do conjunto de dados, seguido da aplicação do kNN utilizando diferentes configurações de parâmetros, como o número de vizinhos (*k*) e métricas de distância. Para avaliação do modelo, foram empregadas métricas como acurácia, precisão, além de validação cruzada para garantir a robustez dos resultados. Os experimentos demonstraram que o desempenho do kNN é altamente sensível à escolha do valor de *k* e à normalização dos dados, destacando a importância desses fatores na aplicação prática do algoritmo. No geral, o modelo apresentou resultados consistentes e alinhados com as características do conjunto de dados analisados.

2. Introdução

Com o avanço acelerado das mídias sociais, o Instagram vem se estabelecendo como uma das plataformas mais relevantes para influenciadores digitais, que exercem uma função fundamental no marketing e na definição de tendências. Detectar padrões de desempenho e atributos que distinguem influenciadores proeminentes tornou-se crucial para marcas e pesquisadores que desejam entender o impacto social e econômico desses especialistas.

Neste cenário, optou-se pelo algoritmo k-Nearest Neighbors (kNN) para a análise e classificação de influenciadores no Instagram, devido à sua simplicidade e eficiência em atividades de classificação e regressão. O kNN se fundamenta na proximidade entre amostras em um espaço multidimensional, possibilitando a identificação intuitiva e transparente de padrões complexos.

O conjunto de dados utilizado no projeto contém informações detalhadas sobre influenciadores do Instagram, incluindo métricas de popularidade, interação e produção de conteúdo. Esse conjunto foi escolhido por sua relevância prática e por fornecer um cenário rico para explorar a aplicabilidade do kNN em análises de redes sociais. As análises realizadas visam identificar agrupamentos e características que possam ser úteis na segmentação de influenciadores, contribuindo para decisões estratégicas no marketing digital.

Além disso, o estudo de influenciadores digitais representa um desafio analítico, uma vez que as métricas de redes sociais frequentemente apresentam uma alta variabilidade e podem conter correlações não lineares. Por exemplo, a taxa de engajamento pode ser impactada por múltiplos fatores, como o tipo de conteúdo publicado, o número de seguidores e a frequência de interação. Nesse contexto, o kNN se destaca por sua capacidade de identificar padrões baseados na proximidade entre amostras, sem pressupor relações lineares entre as variáveis, o que o torna particularmente útil para este tipo de análise exploratória.

O conjunto de dados em questão abrange informações como número de seguidores, curtidas médias por publicações, comentários e hashtags utilizadas. Além disso, inclui características categóricas como o nicho de atuação do influenciador. Esses dados foram obtidos a partir de fontes públicas e organizados para garantir uma visão abrangente e representativa do cenário analisado. Por meio da aplicação do kNN, espera-se compreender melhor as relações entre essas variáveis, identificar perfis semelhantes e oferecer insights valiosos.

3. Metodologia

3.1 Análise Exploratória

A análise exploratória dos dados foi realizada para compreender as principais características do conjunto de dados e identificar padrões iniciais. As variáveis-chave analisadas incluíram o número de seguidores, a taxa de engajamento, e o número médio de curtidas e comentários por publicação. Essas métricas foram avaliadas quanto à sua distribuição, presença de outliers e correlação entre si.

Durante essa etapa, foi identificado que algumas variáveis apresentavam uma escala muito diferente, como o número de seguidores em comparação com a taxa de engajamento, exigindo normalização para evitar que valores com maior amplitude dominassem os cálculos de distância do kNN. Além disso, foi observado que a variável "country" possuía múltiplas categorias, muitas delas relacionadas a países de um mesmo continente, o que motivou a transformação dessa variável em uma classificação mais generalizada por continente. Essa decisão foi justificada pela busca de uma simplificação que pudesse evidenciar padrões regionais sem perder a representatividade dos dados.

3.2 Implementação do Algoritmo

O algoritmo k-Nearest Neighbors foi implementado utilizando a biblioteca `scikit-learn`. A configuração inicial do modelo incluiu a escolha de *k* como um hiperparâmetro a ser ajustado posteriormente, com valores testados variando entre 1 e 20. A métrica de distância utilizada foi a distância euclidiana, uma escolha padrão que se alinha bem com dados contínuos e variáveis normalizadas.

A variável "country" foi transformada em uma nova variável categórica representando continentes. Essa transformação foi realizada utilizando um mapeamento fixo que associava cada país ao respectivo continente, reduzindo a dimensionalidade categórica e facilitando a análise. Em seguida, todas as variáveis numéricas foram padronizadas usando a técnica de ‘StandardScaler’, garantindo que seus valores estivessem na mesma escala entre -1 e 1.

Para realizar a predição foi utilizado a coluna ‘influence_score’, onde pode identificado qual o nível de influência de uma pessoa. A coluna foi transformada em intervalos no qual houve duas faixas, alta e baixa. Essas faixas foram convertidas

para números, 1 e 0, respectivamente, como forma de prever na realização dos testes.

3.3 Validação e Ajuste de Hiperparâmetros

Para garantir a robustez do modelo, foi utilizada a validação cruzada do tipo *k*-fold, com *k* igual a 10. Esse procedimento dividiu os dados em 10 subconjuntos, treinando o modelo em 9 partes e avaliando-o na parte restante, alternando os subconjuntos a cada iteração. A média das métricas de avaliação em todas as iterações foi usada para selecionar os melhores parâmetros.

O ajuste de hiperparâmetros focou principalmente na escolha do valor de *k* e na métrica de distância. O processo de busca utilizou a técnica de *grid search*, testando diferentes combinações de valores de *k* e métricas como distância euclidiana e manhattan. O valor de *k* que apresentou melhor desempenho em termos de acurácia foi selecionado como ideal para o modelo final.

4. Resultados

A avaliação do desempenho do modelo k-Nearest Neighbors (kNN) foi realizada utilizando métricas clássicas de classificação. As principais métricas analisadas foram:

- **Acurácia:** Mede a proporção de verdadeiros positivos e verdadeiros negativos entre todas as previsões.
- **Precisão:** Mede a proporção de verdadeiros positivos em relação ao total de previsões positivas.
- **Recall (Sensibilidade):** Mede a proporção de verdadeiros positivos em relação ao total de instâncias reais positivas.

Abaixo alguns dados plotados:

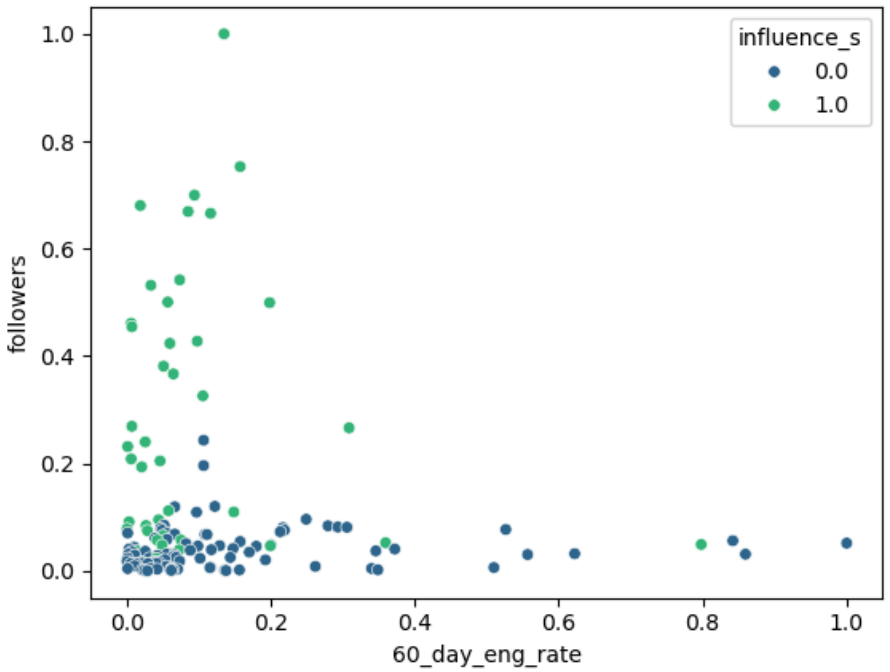

Figura 1. Gráfico de dispersão entre as variáveis ‘followers’ e ‘60_day_eng_rate’.

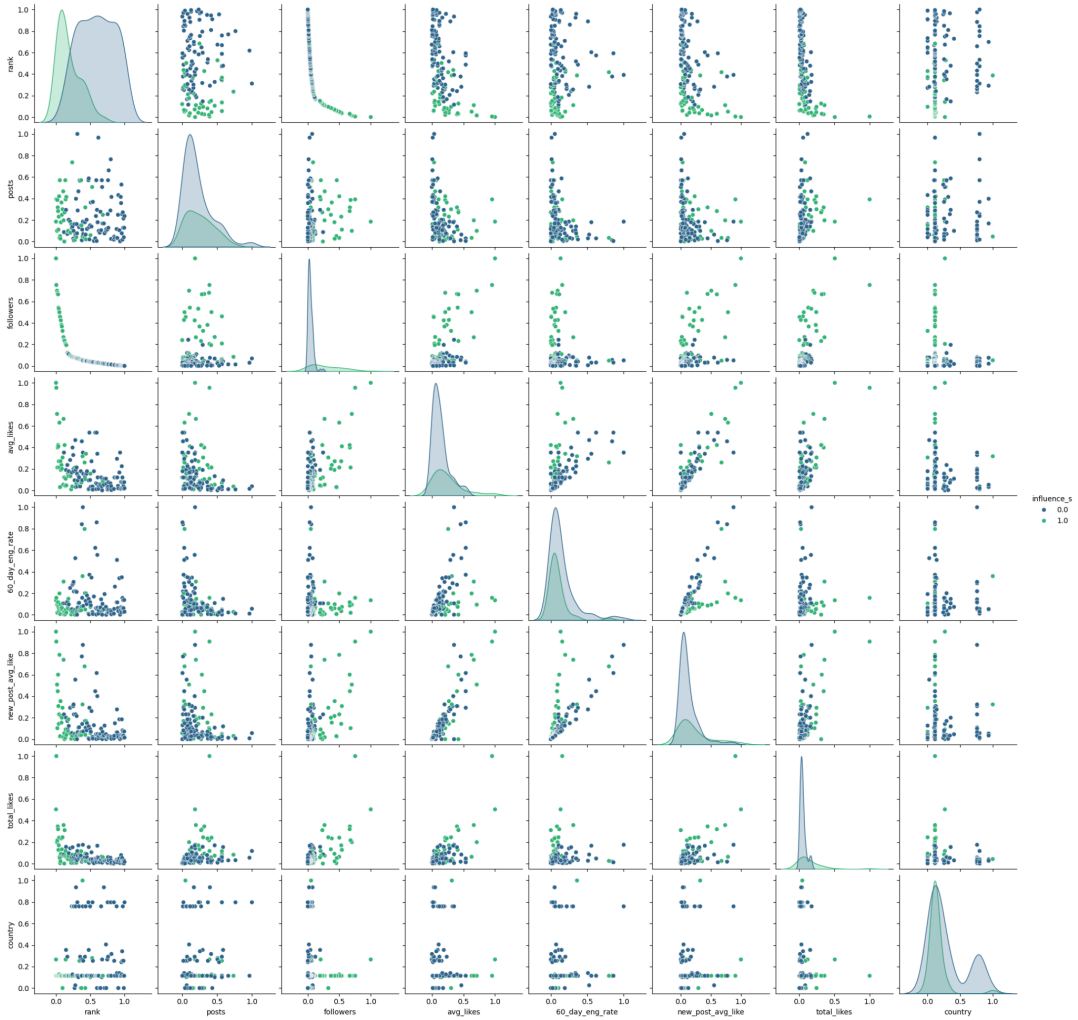

Figura 2. Matriz de correlação na forma gráfica.

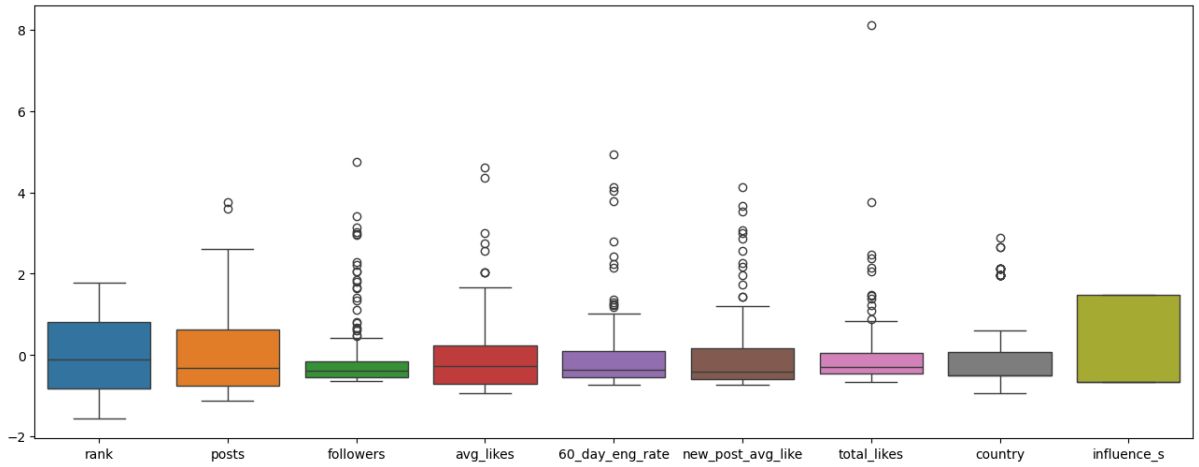

Figura 3. Gráfico boxplot após a padronização das variáveis.

5. Discussão

O desempenho do algoritmo k-Nearest Neighbors (kNN) apresentou uma acurácia de 85%, uma precisão de 89%, recall de 89% e F1-Score de 89%. Esses resultados indicam que o modelo foi bem-sucedido na tarefa de classificação em um nível satisfatório,
A performance do kNN pode ser altamente influenciada pela variabilidade presente no conjunto de dados. Dados altamente variados podem levar o algoritmo a ter um desempenho inconsistente. Se o conjunto de dados apresentar um desequilíbrio nas classes, pode afetar significativamente as métricas de avaliação. A alta precisão pode ser enganosa se a maioria das instâncias pertencer a uma única classe. Embora o kNN seja fácil de implementar e interpretar, ele não é eficiente para conjuntos de dados muito grandes devido ao tempo de execução elevado, uma vez que a complexidade do algoritmo aumenta com o tamanho do conjunto de dados.

Durante a implementação do kNN, várias limitações foram observadas que impactaram o desempenho do modelo, como o valor de k é um hiperparâmetro crítico no kNN. A escolha inadequada de k pode levar a overfitting (valores pequenos de k) ou underfitting (valores grandes de k). A validação cruzada foi utilizada para otimizar esse valor, mas não elimina completamente o risco de uma escolha subótima. O kNN é sensível à escala das variáveis. Mesmo com a normalização aplicada, qualquer desvio pode influenciar desproporcionalmente as distâncias calculadas pelo algoritmo. A natureza do algoritmo kNN, que envolve o cálculo de distâncias para todas as instâncias de treino para cada previsão, resulta em altos custos computacionais para conjuntos de dados grandes.

As decisões tomadas durante a implementação tiveram um impacto significativo nos resultados obtidos, a normalização dos dados foi essencial para garantir que todas as variáveis contribuíram de maneira justa no cálculo das distâncias. Sem essa etapa, o desempenho do modelo teria sido drasticamente comprometido. Esta transformação ajudou a reduzir a dimensionalidade e a simplificar a modelagem, permitindo que o kNN lidasse melhor com a variabilidade geográfica. A utilização da validação cruzada foi crucial para a otimização dos hiperparâmetros e para garantir que o modelo não estivesse apenas ajustado aos dados de treino, mas que pudesse generalizar bem para dados não vistos.

6. Conclusão e trabalhos futuros

Aprendemos a implementar o algoritmo k-Nearest Neighbors utilizando bibliotecas do Python como sklearn. A prática envolveu a configuração dos parâmetros e a transformação dos dados para melhorar o desempenho do modelo. A análise exploratória dos dados forneceu insights valiosos sobre as variáveis chave e sua influência no modelo. Esta fase é crucial para entender a natureza dos dados e identificar possíveis desafios e oportunidades.
O pré-processamento dos dados, incluindo a normalização e a transformação de variáveis categóricas, mostrou-se essencial para o bom desempenho do modelo. A importância da validação cruzada foi reforçada, garantindo que o modelo treinado não apenas se ajuste bem aos dados de treino, mas também generalize adequadamente para novos dados. A interpretação das métricas de avaliação (acurácia, precisão, recall, F1-Score) proporcionou uma visão abrangente do desempenho do modelo, destacando seus pontos fortes e fracos.

Possíveis melhorias incluem considerar a implementação e comparação com outros algoritmos de classificação, como Support Vector Machines (SVM), Árvores de Decisão, ou Redes Neurais, para verificar se eles podem oferecer um melhor desempenho.
Embora a validação cruzada tenha sido utilizada, técnicas mais avançadas de ajuste de hiperparâmetros, como Random Search ou Grid Search, poderiam ser exploradas para encontrar a configuração ideal. Coletar mais dados ou utilizar técnicas de aumento de dados pode ajudar a melhorar a robustez e a precisão do modelo. Investigar o desempenho do modelo em diferentes subconjuntos de dados (por exemplo, por continente) para identificar possíveis vieses ou áreas de melhoria específicas. Implementar técnicas para lidar com o desequilíbrio de classes, como oversampling, undersampling, ou o uso de técnicas de amostragem avançadas, pode melhorar significativamente o desempenho do modelo em classes minoritárias.

7. Referências

1. IBM. "O que é o algoritmo dos k vizinhos mais próximos?" IBM, https://www.ibm.com/br-pt/topics/knn. Acessado em: 17 de novembro de 2024.

2. Didatica.tech. "Como funciona o KNN (K-nearest neighbors)." Didatica.tech, https://didatica.tech/o-que-e-e-como-funciona-o-algoritmo-knn/. Acessado em: 17 de novembro de 2024.

3. Análise Macro. "Classificando economias com o algoritmo k-vizinhos mais próximos (k-NN)." Análise Macro, https://analisemacro.com.br/econometria-e-machine-learning/classificando-economias-com-o-algoritmo-k-vizinhos-mais-proximos-k-nn/. Acessado em: 17 de novembro de 2024.

4. Aprender Estatística Fácil. "O que é: K-Nearest Neighbors (KNN)." Aprender Estatística Fácil, https://bing.com/search?q=refer%c3%aancias+sobre+k-Nearest+Neighbors. Acessado em: 17 de novembro de 2024.

5. APRENDER ESTATÍSTICA FÁCIL. "O que é: K-Nearest Neighbors (KNN)." APRENDER ESTATÍSTICA FÁCIL, https://bing.com/search?q=refer%c3%aancias+sobre+k-Nearest+Neighbors. Acessado em: 17 de novembro de 2024.