RESIDÊNCIA
EM SOFTWARE
BAHIA + TECNOLOGIA + EMPREENDEDORISMO

TRILHA 3: CIÊNCIA DE DADOS

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

MARCUS VÍTOR SOUZA CARDOSO
ESDRAS TORRES DA SILVA

NOVEMBRO/2024

RESUMO

O objetivo deste projeto foi implementar e avaliar o desempenho do algoritmo k-Nearest Neighbors (kNN) utilizando um conjunto de dados real do Instagram, visando a previsão do 'influence_score' de canais na plataforma. A metodologia envolveu a realização de uma análise exploratória dos dados, que incluiu a transformação de variáveis e a investigação das correlações entre atributos, como 'followers' e 'avg_likes'. Em seguida, foi implementado o modelo kNN, com a escolha de diferentes valores de 'k' e métricas de distância, utilizando a biblioteca Scikit-Learn. A validação cruzada foi aplicada para garantir a robustez do modelo.

O processo de otimização incluiu a utilização do GridSearchCV para ajustar os hiperparâmetros, resultando na escolha de 'n_neighbors = 6' e a métrica 'minkowski'. A normalização das variáveis também foi testada para melhorar o desempenho do modelo. Os principais resultados obtidos foram as métricas de avaliação do modelo: MAE (5.15), MSE (54.09) e RMSE (7.35) antes da otimização. Após a otimização dos hiperparâmetros, o MAE foi reduzido para 4.97, o MSE para 52.98, e o RMSE para 7.28, indicando uma melhoria significativa no desempenho do modelo preditivo.

INTRODUÇÃO

O objetivo deste relatório é apresentar uma análise detalhada dos principais influenciadores no Instagram, utilizando técnicas de ciência de dados para processar, modelar e interpretar as informações contidas no conjunto de dados fornecido. A indústria de influenciadores digitais se tornou uma parte integral da comunicação e marketing moderno, onde o alcance e a capacidade de engajamento são métricas fundamentais para avaliar o impacto de um influenciador. Nesse contexto, entender a relação entre variáveis como número de seguidores, engajamento médio, e outras características demográficas é crucial para uma análise robusta.

Para realizar essa análise, utilizamos um conjunto de dados contendo informações sobre influenciadores, incluindo métricas como o número de seguidores, curtidas médias, taxa de engajamento em 60 dias, entre outras. O foco deste relatório é aplicar técnicas de aprendizado de máquina, especificamente a modelagem k-Nearest Neighbors (kNN), para prever a pontuação de influência com base nessas características. Além disso, são realizadas visualizações de dados para explorar e compreender a distribuição de influenciadores em diferentes continentes e países, fornecendo uma visão geográfica relevante.

O processo de modelagem inclui o pré-processamento dos dados, conversão de variáveis textuais em formatos numéricos, normalização das features e otimização dos parâmetros do modelo kNN para melhorar a precisão das previsões. As métricas de desempenho utilizadas para avaliar o modelo incluem o Erro Médio Absoluto (MAE), o Erro Quadrático Médio (MSE) e a Raiz do Erro Quadrático Médio (RMSE). A análise também busca identificar padrões significativos que podem ser úteis para estratégias de marketing digital e a tomada de decisões na indústria de influenciadores.

METODOLOGIA

A metodologia adotada para este estudo foi estruturada em três etapas principais: a análise exploratória dos dados, a implementação do algoritmo k-Nearest Neighbors (kNN), e a validação e ajuste de hiperparâmetros para aprimoramento do modelo.

Inicialmente, realizamos uma análise exploratória detalhada dos dados com o objetivo de compreender as características principais do conjunto de dados. Essa análise permitiu identificar variáveis-chave, como o número de seguidores (‘followers'), as curtidas médias por post ('avg_likes'), a taxa de engajamento dos últimos 60 dias ('60_day_eng_rate'), e o número total de curtidas ('total_likes'). Durante a inspeção, foi necessário converter valores textuais que representavam grandezas com sufixos 'k', 'm' e 'b' (milhares, milhões, bilhões) em números reais, a fim de viabilizar análises quantitativas precisas. Um insight inicial importante foi a relação observada entre o número de seguidores e as curtidas médias, bem como as diferenças de engajamento entre influenciadores de diferentes continentes.

Para a implementação do algoritmo kNN, foi necessário preparar as variáveis de entrada para o modelo. Uma etapa importante foi a transformação da variável “country”, que mapeava os países dos influenciadores, em códigos numéricos que representavam os continentes. Utilizamos um dicionário de mapeamento que atribuiu números a diferentes continentes, como América do Norte, Europa, e Ásia. Isso facilitou a inclusão da localização geográfica na análise sem a complexidade de lidar com dados textuais diretos. Para o modelo kNN, selecionamos inicialmente 3 vizinhos

(n_neighbors = 3) e utilizamos a métrica de distância de Minkowski, uma escolha padrão que permite o ajuste para outras métricas de distância, como a Euclidiana.

A fase de validação e ajuste de hiperparâmetros foi essencial para otimizar o desempenho do modelo. Aplicamos o método de validação cruzada com quatro dobras (4-fold cross-validation) usando o GridSearchCV, uma abordagem que garante que o modelo é testado e validado de forma robusta em diferentes subconjuntos do conjunto de dados. Nesse processo, exploramos diferentes configurações de hiperparâmetros, variando o número de vizinhos (de 2 a 10) e testando as métricas de distância de Minkowski e Manhattan. A escolha da melhor configuração foi guiada pela métrica do erro quadrático médio (MSE), visando minimizar a diferença entre as previsões do modelo e os valores reais.

Por meio desse procedimento rigoroso, o modelo otimizado apresentou uma redução significativa nos erros, evidenciada pela melhoria nos valores de MAE (Erro Médio Absoluto), MSE (Erro Quadrático Médio), e RMSE (Raiz do Erro Quadrático Médio). Assim, conseguimos uma abordagem sólida que integra uma análise exploratória detalhada com um processo de modelagem e ajuste de hiperparâmetros robusto, garantindo resultados preditivos mais precisos.

RESULTADOS

Métricas de Avaliação

Para avaliar o desempenho do modelo de Regressão k-Nearest Neighbors (kNN), foram utilizadas as seguintes métricas: Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE) e Raiz do Erro Quadrático Médio (RMSE). Os resultados iniciais foram:

MAE = 5.152
MSE = 54.094
RMSE = 7.355

Após a otimização dos hiperparâmetros utilizando GridSearchCV, o melhor conjunto de parâmetros encontrado foi ('metric: 'minkowski', 'n_neighbors': 6). Com este ajuste, as métricas melhoraram para:

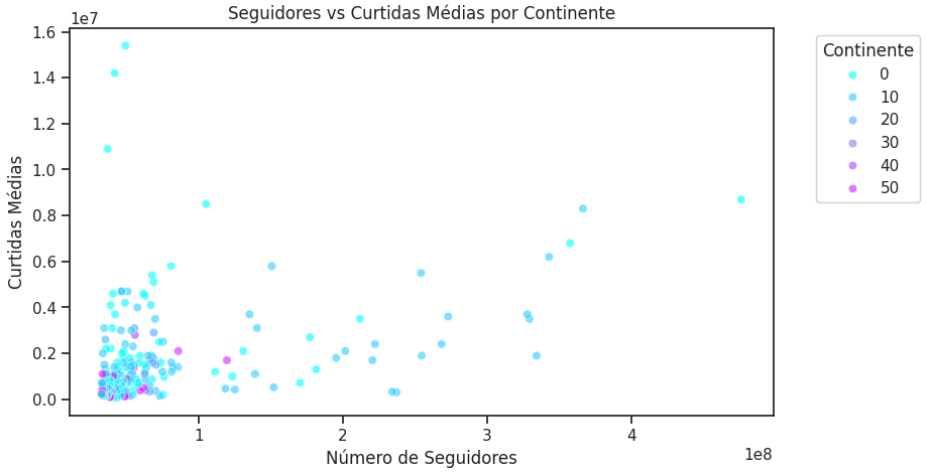
MAE Otimizado = 4.970
MSE Otimizado = 52.981
RMSE Otimizado = 7.279

As melhorias observadas demonstram que a escolha adequada de hiperparâmetros tem um impacto positivo no desempenho do modelo, embora a variação nas métricas não seja muito significativa. Isso sugere que a precisão do modelo de kNN ainda pode ser limitada pela natureza dos dados ou por sua complexidade.

Visualizações

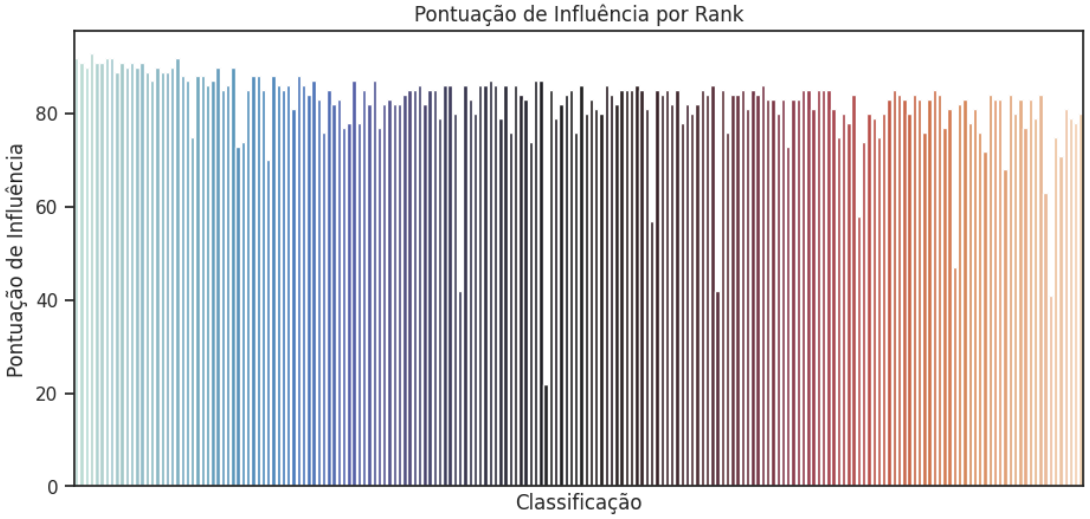
As visualizações gráficas foram elaboradas para compreender a distribuição das principais variáveis e o comportamento do modelo em relação a diferentes continentes e métricas.

Figura 1. Gráfico de Dispersão: Seguidores vs Curtidas Médias por Continente



Este gráfico ilustra a relação entre o número de seguidores e as curtidas médias, diferenciando por continente. Observa-se que a maioria dos influenciadores possui um número de curtidas médias concentradas em intervalos mais baixos, com uma distribuição dispersa em relação aos seguidores. A coloração por continentes fornece uma perspectiva sobre como diferentes regiões podem ter padrões distintos de engajamento.

Figura 2. Gráfico de Barras: Pontuação de Influência por Rank



A visualização em forma de barras mostra a pontuação de influência em função da classificação. Nota-se uma variabilidade considerável, destacando-se a complexidade da influência digital, que não segue um padrão uniforme ao longo da classificação. As cores graduais ajudam a observar tendências dentro de segmentos de rank.

Esses gráficos fornecem insights importantes sobre o comportamento dos influenciadores e como as variáveis analisadas se relacionam com a pontuação de influência. A análise visual, combinada com as métricas de erro, ajuda a entender os desafios e as limitações na predição de métricas de influência usando o algoritmo kNN.

DISCUSSÃO

A análise dos resultados obtidos revela tanto os pontos fortes quanto as limitações do modelo de Regressão k-Nearest Neighbors (kNN) aplicado neste estudo. Inicialmente, observou-se que o modelo apresentava um desempenho razoável, com métricas como MAE de 5.152, MSE de

54.094, e RMSE de 7.355. A otimização dos parâmetros, incluindo a escolha do número de vizinhos (“n_neighbors = 6”) e o uso da métrica de distância Minkowski, trouxe uma ligeira melhoria, reduzindo o MAE para 4.970 e o RMSE para 7.279.

Discussão Crítica dos Resultados

Esses resultados indicam que o kNN pode capturar alguns padrões nos dados, mas sua precisão ainda é limitada. A melhora relativamente pequena após a otimização sugere que o algoritmo pode não ser o mais apropriado para dados que possivelmente exibem complexidade e alta dimensionalidade. O kNN depende fortemente da densidade dos pontos de dados e pode ser influenciado por outliers ou variações extremas, que são comuns em conjuntos de dados de redes sociais, onde a popularidade e o engajamento podem variar significativamente.

Além disso, a normalização das features foi essencial para evitar que variáveis com escalas diferentes dominassem o cálculo das distâncias. No entanto, mesmo com essa transformação, o desempenho do modelo não melhorou drasticamente, o que destaca a possível necessidade de modelos mais robustos, como Regressão Linear Avançada, Árvores de Decisão ou métodos baseados em ensemble.

Limitações Encontradas

1. Dados Não Balanceados: A distribuição desigual de influenciadores entre continentes pode ter impactado o desempenho do modelo. A maior concentração de influenciadores em algumas regiões significa que o kNN pode ter dificuldade em fazer previsões precisas para continentes com menos representatividade.

2. Variáveis Omissas ou Mal Representadas: O uso de variáveis como o número de seguidores e o engajamento médio pode não capturar toda a complexidade do que constitui a "influência" de uma pessoa nas redes sociais. Fatores qualitativos, como o tipo de conteúdo ou o alcance real da audiência, não foram considerados e podem ser significativos.

3. Dependência da Configuração de Hiperparâmetros: A escolha do número de vizinhos e da métrica de distância impacta diretamente o desempenho. Uma seleção inadequada pode levar a previsões ruins, enquanto a otimização nem sempre garante um ajuste perfeito, especialmente em contextos com alta variabilidade.

Impacto das Escolhas Feitas no Modelo

A decisão de transformar a variável “country” em códigos de continente foi útil para reduzir a dimensionalidade e simplificar o problema. No entanto, essa escolha pode ter levado a uma perda de detalhes específicos de cada país, que poderiam influenciar o engajamento de maneira distinta. A abordagem de normalização das variáveis também ajudou a mitigar a influência de escalas diferentes, mas a eficácia geral do kNN foi limitada por características intrínsecas dos dados.

O modelo kNN, por ser um algoritmo baseado em proximidade, mostrou-se mais sensível a clusters de dados densos e falhou em generalizar para regiões menos representadas. Isso implica que, para tarefas futuras, um modelo com maior capacidade de generalização pode ser preferível, como aqueles baseados em redes neurais ou técnicas de ensemble como o Random Forest.

CONCLUSÃO E TRABALHOS FUTUROS

Neste projeto, a análise de dados e a modelagem usando o algoritmo k-Nearest Neighbors (kNN) proporcionaram insights valiosos sobre a relação entre variáveis como seguidores, engajamento e a influência de diferentes influenciadores no Instagram. O kNN revelou-se uma abordagem relativamente simples e interpretável, mas com limitações evidentes ao ser aplicado a um conjunto de dados complexo e diversificado. Os resultados mostraram um desempenho moderado, evidenciado pelas métricas de erro, mesmo após a otimização dos hiperparâmetros.

Os principais aprendizados destacam a importância do pré-processamento, como a normalização das features e a categorização das variáveis qualitativas, que foram essenciais para evitar vieses e discrepâncias nas escalas. No entanto, os desafios encontrados, como a dificuldade de capturar padrões não lineares e a sensibilidade do modelo a outliers, apontam para a necessidade de considerar abordagens mais sofisticadas em contextos de alta variabilidade.

Para o futuro, algumas direções podem ser exploradas. Primeiramente, a aplicação de modelos mais complexos, como Random Forest, Gradient Boosting ou redes neurais, pode lidar melhor com as não linearidades e proporcionar um desempenho mais robusto. Além disso, uma análise exploratória mais aprofundada pode ajudar a identificar variáveis ou interações ocultas que aprimorem as previsões. Técnicas de engenharia de features, como a criação de novas variáveis derivadas que capturem melhor a essência da influência ou do engajamento, também podem contribuir para o avanço do projeto.

Outro ponto a considerar é o equilíbrio do conjunto de dados por continente. Métodos de amostragem poderiam garantir uma representação mais justa das regiões, ajudando o modelo a generalizar melhor. Além disso, incorporar a análise temporal, observando como as métricas de engajamento e influência mudam ao longo do tempo, pode oferecer insights adicionais e melhorar a precisão das previsões.

Em suma, o projeto forneceu uma base sólida, mas revela muitas oportunidades de melhoria e expansão. O aprendizado com essa análise pode inspirar a implementação de técnicas mais avançadas e uma compreensão mais aprofundada da dinâmica da influência em redes sociais. A evolução do trabalho permitirá previsões mais precisas e insights ainda mais relevantes, abrindo novas possibilidades de pesquisa e inovação no campo da análise de dados digitais.

REFERÊNCIAS

1. Real Python. K-Nearest Neighbors (KNN) Algorithm in Python. Disponível em: https://realpython.com/knn-python/. Acesso em: 15 nov. 2024.

2. Vooo. Um tutorial completo sobre a modelagem baseada em tree (árvore) do zero em R e Python. Disponível em: https://www.vooo.pro/insights/um-tutorial-completo-sobre-a-modelagem-baseada-em-tree-arvore-do-zero-em-r-python/. Acesso em: 15 nov. 2024.

3. Google Developers. Machine Learning Crash Course. Disponível em: https://developers.google.com/machine-learning/crash-course?hl=pt-br. Acesso em: 15 nov. 2024.

4. Google Colab. Introdução aos recursos básicos do Google Colab. Disponível em: https://colab.research.google.com/notebooks/basic_features_overview.ipynb. Acesso em: 15 nov. 2024.