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

Nome do Residente: Fagner de Oliveira Carrena

Nome do Residente: Sthefane dos Santos Ramos

Data de Entrega: 13/11/2024

Introdução

Com o crescente uso de redes sociais, como o Instagram, a análise de dados de perfis e suas métricas tornou-se essencial para entender comportamentos de engajamento e a influência de contas populares. Esse projeto foca na implementação e análise do algoritmo *k-Nearest Neighbors* (kNN) aplicado a dados de contas influentes do Instagram, com o objetivo de explorar e identificar padrões que possam ser úteis para marcas, criadores de conteúdo e analistas de mídia.

Importância da Análise de Dados no Contexto do Instagram

O Instagram é uma das plataformas mais utilizadas globalmente, servindo tanto para comunicação pessoal quanto para estratégias de marketing digital. Através da análise de métricas como número de seguidores, curtidas médias por post e taxa de engajamento, é possível obter insights valiosos sobre o comportamento da audiência, a performance de conteúdo e o impacto de influenciadores na plataforma. Com esses dados, marcas e criadores de conteúdo podem otimizar suas estratégias para maximizar alcance e engajamento.

Justificativa para o Uso do Algoritmo k-Nearest Neighbors

O algoritmo *k-Nearest Neighbors* (kNN) foi escolhido para este projeto por sua simplicidade e eficiência em tarefas de classificação e previsão. O kNN é um algoritmo baseado em instâncias que realiza previsões considerando as instâncias mais próximas (ou vizinhas) no espaço de características. No contexto do Instagram, o kNN pode ser utilizado para identificar padrões de similaridade entre contas, classificando-as em grupos com características de engajamento e popularidade semelhantes. Além disso, o kNN é fácil de implementar e fornece

resultados interpretáveis, o que o torna adequado para o nosso objetivo de explorar dados e gerar insights acionáveis.

Resumo

Este relatório apresenta um estudo sobre a implementação e análise do algoritmo *k-Nearest Neighbors* (kNN) aplicado a dados de contas do Instagram, com o objetivo de explorar padrões de engajamento e influência. O Instagram, sendo uma das plataformas de redes sociais mais populares, oferece métricas ricas e variadas que podem ser usadas para entender o comportamento de contas influentes e as interações de seus seguidores.

Objetivo

O principal objetivo deste projeto é investigar e identificar padrões de engajamento em contas influentes do Instagram, utilizando o kNN como método de análise para classificar contas em grupos semelhantes com base em métricas como seguidores, média de curtidas e taxa de engajamento de 60 dias. O projeto visa fornecer insights úteis que possam auxiliar marcas e criadores de conteúdo a entender melhor suas audiências e otimizar suas estratégias de conteúdo.

Metodologia

A metodologia adotada incluiu a coleta e pré-processamento dos dados, seguida da aplicação do algoritmo kNN em diferentes configurações. Realizamos testes com diferentes valores de *k* e métricas de distância (Euclidiana, Manhattan e Minkowski) para identificar a configuração com melhor desempenho. Utilizamos técnicas de validação cruzada e normalização de dados para garantir a consistência e a qualidade dos resultados. Para a otimização do modelo, aplicamos o *GridSearchCV* a fim de identificar os melhores hiperparâmetros para o algoritmo.

Principais Resultados

Os resultados mostraram uma correlação moderada entre o número de seguidores e a média de curtidas, sugerindo uma relação direta, mas não determinante, entre esses fatores. Além disso,

a taxa de engajamento dos últimos 60 dias revelou-se um forte indicador do comportamento de engajamento da audiência. A otimização dos hiperparâmetros do kNN resultou em uma configuração que melhorou a acurácia do modelo, oferecendo classificações mais precisas para grupos de contas semelhantes em termos de engajamento e popularidade.

Introdução

Contextualização do Problema e Justificativa para o Uso do kNN na Análise de Influenciadores do Instagram

Com o aumento da presença digital, o Instagram se consolidou como uma das principais plataformas para influenciadores e marcas se conectarem com suas audiências. Analisar o desempenho de influenciadores requer mais do que observar o número de seguidores; é preciso entender o engajamento real de cada conta, as interações do público e como essas contas se comparam entre si em termos de influência e popularidade.

Nesse cenário, o algoritmo *k-Nearest Neighbors* (kNN) se destaca como uma ferramenta poderosa para identificar padrões de similaridade entre influenciadores, agrupando contas com características semelhantes. O kNN é ideal para essa análise, pois classifica cada influenciador em relação a um conjunto de vizinhos próximos, fornecendo insights baseados em características numéricas como seguidores, curtidas médias e taxa de engajamento. Como resultado, o kNN ajuda a identificar grupos de influenciadores com engajamento e popularidade parecidos, permitindo uma análise mais granular e prática para otimizar estratégias de marketing e parcerias.

Descrição do Conjunto de Dados de Influenciadores do Instagram

O conjunto de dados utilizado neste projeto contém informações detalhadas de contas influentes do Instagram. As principais variáveis incluem:

- **followers**: Número total de seguidores de cada influenciador, que indica o alcance potencial de suas postagens.
- **avg_likes**: Média de curtidas recebidas por postagem, fornecendo uma medida de engajamento do público.

- **60_day_eng_rate**: Taxa de engajamento calculada com base nos últimos 60 dias, representada em porcentagem, que mede a qualidade da interação entre o influenciador e sua audiência.
- **influence_score**: Uma pontuação atribuída para mensurar a influência geral de cada conta, baseada em várias métricas de engajamento.
- **country**: País de origem do influenciador, categorizado em continentes para facilitar a análise geográfica.

Esse conjunto de dados oferece uma visão ampla do comportamento dos influenciadores, permitindo uma análise robusta sobre como variáveis como engajamento e seguidores contribuem para a influência percebida no Instagram. O uso do kNN, aliado a essas variáveis, permite classificar influenciadores em grupos semelhantes, criando uma base sólida para decisões estratégicas no marketing de influência.

Metodologia

Análise Exploratória

A análise exploratória dos dados foi realizada para compreender a distribuição e as relações entre as principais variáveis do conjunto de dados. Primeiramente, foram analisadas variáveis-chave como o número de seguidores (`followers`), média de curtidas (`avg_likes`), taxa de engajamento de 60 dias (`60_day_eng_rate`) e o escore de influência (`influence_score`).

- **Distribuição de `followers` e `avg_likes`**: Um histograma foi utilizado para verificar a distribuição dessas variáveis, revelando a concentração de contas em diferentes faixas de seguidores e curtidas. Esse gráfico permitiu identificar que a maioria das contas possui uma base de seguidores moderada, enquanto apenas algumas atingem valores extremamente altos.
- **Correlação entre `followers` e `avg_likes`**: Um gráfico de dispersão entre `followers` e `avg_likes` foi gerado para verificar a relação entre o alcance potencial e o engajamento médio de cada influenciador. Observou-se uma correlação moderada, sugerindo que um número maior de seguidores não garante necessariamente um maior engajamento.

- **Relação entre `60_day_eng_rate` e `avg_likes`**: A análise entre a taxa de engajamento e a média de curtidas mostrou-se significativa, indicando que a taxa de engajamento dos últimos 60 dias é um importante preditor do comportamento do público em relação aos conteúdos.

Implementação do Algoritmo

Para a classificação e análise dos influenciadores, foi implementado o algoritmo *k-Nearest Neighbors* (kNN) com o Scikit-Learn. A implementação incluiu a configuração de diferentes valores para `k` e várias métricas de distância para testar o desempenho do modelo:

- **Configuração Inicial de `k` e Métricas de Distância**: Valores de `k` variando entre 1 e 9 foram testados para encontrar o número ideal de vizinhos. As métricas de distância utilizadas foram Euclidiana, Manhattan e Minkowski. O ajuste de `k` e das métricas permitiu observar variações de acurácia e identificar a configuração com melhor desempenho.
- **Transformação da Variável `country` para Categorização por Continente**: Como a variável `country` contém informações geográficas, foi realizada uma transformação para categorizá-la por continentes. Um mapeamento foi feito para atribuir valores numéricos a cada continente, agrupando os países em regiões específicas como América do Sul, América do Norte, Europa, Ásia, Oceania e África. Essa transformação foi útil para que o modelo pudesse identificar influenciadores com similaridade geográfica.

Validação e Ajuste de Hiperparâmetros

Para garantir a robustez do modelo, foi utilizada a validação cruzada e a otimização de hiperparâmetros com o *GridSearchCV*:

- **Validação Cruzada**: A técnica de validação cruzada com `k=5` folds foi aplicada ao modelo para avaliar a consistência dos resultados em diferentes divisões dos dados. Esse processo permitiu obter uma média de acurácia que representa a performance geral do modelo, minimizando o impacto de variações entre os conjuntos de treino e teste.

- **Otimização dos Parâmetros com GridSearchCV**: O GridSearchCV foi empregado para identificar os melhores valores de `k` e a métrica de distância que maximizassem a acurácia. O GridSearchCV avaliou uma combinação de valores de `k` (3, 5, 7, 9, 11) e métricas de distância (Euclidiana, Manhattan e Minkowski). O resultado final mostrou que a configuração ideal era `k=5` com a métrica Euclidiana, que apresentou a melhor performance entre as alternativas testadas.

Resultados

Métricas de Avaliação

A avaliação do modelo k-Nearest Neighbors (kNN) foi realizada utilizando três métricas de erro amplamente empregadas em modelos de classificação e regressão: **Erro Médio Absoluto (MAE)**, **Erro Médio Quadrático (MSE)** e **Raiz do Erro Médio Quadrático (RMSE)**. Cada uma dessas métricas fornece uma perspectiva única sobre a precisão e o desempenho do modelo:

- **MAE (Mean Absolute Error)**: Essa métrica representa a média das diferenças absolutas entre as previsões e os valores reais. Um MAE mais baixo indica que o modelo tem uma precisão elevada, uma vez que as previsões se aproximam dos valores reais sem considerar o sinal de erro. Para o modelo kNN, o MAE foi observado em um nível aceitável, sugerindo que o modelo está fazendo previsões com erro médio baixo.
- **MSE (Mean Squared Error)**: O MSE é a média dos erros ao quadrado, penalizando erros maiores de maneira mais intensa que o MAE. Um MSE elevado pode indicar a presença de algumas previsões com erros significativos. Para o kNN, o MSE apresentou um valor moderado, o que indica que, embora o modelo tenha um bom desempenho geral, algumas previsões podem estar desviando mais do que o esperado.
- **RMSE (Root Mean Squared Error)**: O RMSE, que é a raiz quadrada do MSE, facilita a interpretação dos erros, trazendo-os para a mesma unidade dos dados originais. O RMSE para o modelo mostrou-se alinhado com as previsões do MAE e MSE, validando que o modelo está desempenhando bem e que os erros são, em média, consistentes.

Essas métricas demonstram que o modelo kNN, após o ajuste de hiperparâmetros, é eficaz para o propósito de classificação de influenciadores com base em variáveis como seguidores e engajamento, fornecendo uma previsão confiável.

Visualizações

As visualizações desempenham um papel crucial para entender a distribuição dos dados e o desempenho do modelo de forma intuitiva. Abaixo estão as principais visualizações utilizadas para a análise:

- **Gráfico de Dispersão de `followers` vs `avg_likes`**: Este gráfico permitiu observar a relação entre o número de seguidores e a média de curtidas por post. A dispersão dos dados revelou uma tendência de correlação moderada, mostrando que, embora contas com mais seguidores possam receber mais curtidas, o aumento de curtidas não é diretamente proporcional ao número de seguidores.
- **Gráfico de Barras de `rank` vs `influence_score`**: O gráfico de barras foi utilizado para comparar o ranking dos influenciadores com seu escore de influência, oferecendo uma visão da distribuição de contas com diferentes níveis de influência. Esse gráfico permitiu identificar como a influência varia entre os rankings, mostrando que influenciadores de alto escalão tendem a ter escores de influência maiores.
- **Histograma de `60_day_eng_rate`**: A taxa de engajamento de 60 dias foi visualizada em um histograma, que revelou a distribuição dos influenciadores de acordo com seu nível de engajamento recente. Esse gráfico foi útil para identificar padrões de engajamento e a concentração de contas em diferentes faixas de interação com o público.
- **Gráfico de Dispersão de `60_day_eng_rate` vs `avg_likes`**: Esta visualização ajudou a entender melhor a relação entre a taxa de engajamento e a média de curtidas, destacando que contas com maior engajamento nos últimos 60 dias tendem a obter um número mais elevado de curtidas por post.

Essas visualizações fornecem um suporte visual às descobertas do modelo e ajudam a entender como variáveis-chave como seguidores, curtidas e engajamento impactam o escore de influência dos usuários no Instagram.

Discussão

Discussão Crítica dos Resultados

Os resultados obtidos com o algoritmo k-Nearest Neighbors (kNN) aplicados aos dados de influenciadores do Instagram fornecem uma visão significativa sobre os padrões de engajamento e influência na plataforma. As métricas de avaliação, especialmente o MAE, MSE e RMSE, indicaram que o modelo é capaz de realizar previsões confiáveis em relação ao engajamento e escore de influência, embora com algumas variações nos erros.

A análise das correlações revelou uma relação moderada entre o número de seguidores e a média de curtidas (`avg_likes`), sugerindo que seguidores não são o único determinante de engajamento. A taxa de engajamento de 60 dias (`60_day_eng_rate`) mostrou-se um indicador mais confiável, revelando que contas com alta taxa de engajamento recente tendem a receber mais curtidas e, consequentemente, apresentam escores de influência elevados. Esse insight confirma a importância de métricas de engajamento além dos seguidores para análise de influenciadores.

Limitações Encontradas

Apesar dos resultados positivos, o projeto apresenta algumas limitações que podem impactar a generalização dos achados:

1. **Limitação de Dados**: O conjunto de dados é restrito a influenciadores com determinada visibilidade no Instagram, o que pode limitar a aplicabilidade dos resultados para contas com menos seguidores ou com engajamento menor.
2. **Influência de Hiperparâmetros**: Embora o GridSearchCV tenha ajudado a identificar os melhores valores de `k` e métricas de distância, o modelo kNN ainda é sensível à escolha desses hiperparâmetros. Pequenas variações podem impactar significativamente a acurácia, o que destaca a necessidade de um ajuste fino e específico para cada conjunto de dados.
3. **Transformação da Variável `country`**: A transformação da variável `country` em categorias numéricas baseadas em continentes, embora útil, pode ter simplificado excessivamente a diversidade cultural e demográfica dos países. Isso limita a

capacidade do modelo de capturar diferenças culturais mais detalhadas que possam impactar o engajamento e o comportamento de seguidores.

4. **Falta de Controle de Outliers**: Certas contas com valores extremamente altos de seguidores ou curtidas não foram excluídas da análise, o que pode ter influenciado algumas métricas, especialmente o MSE, que penaliza mais erros elevados.

Impacto das Escolhas no Desempenho do Modelo

As escolhas metodológicas realizadas ao longo do projeto impactaram diretamente o desempenho do modelo e os insights gerados. A decisão de utilizar o kNN, por exemplo, proporcionou um modelo interpretável e eficaz para análise de similaridade, mas trouxe limitações em termos de sensibilidade a dados ruidosos e valores atípicos. A normalização dos dados ajudou a mitigar o impacto de variáveis com escalas distintas, melhorando a acurácia do modelo em diferentes combinações de `k` e métricas de distância.

A otimização dos hiperparâmetros com o GridSearchCV revelou-se essencial para melhorar o desempenho do modelo, permitindo uma análise comparativa entre diferentes configurações e ajudando a selecionar aquela que oferecia maior precisão. No entanto, essa abordagem exigiu um tempo de processamento adicional, o que pode ser um fator limitante em conjuntos de dados maiores.

Conclusão

Este projeto demonstrou a aplicabilidade do algoritmo k-Nearest Neighbors (kNN) para a análise de influenciadores no Instagram, proporcionando insights sobre a relação entre engajamento e popularidade das contas. Observou-se que variáveis como a taxa de engajamento de 60 dias e a média de curtidas são indicativas do comportamento de seguidores, sendo mais relevantes que o número de seguidores isoladamente. A implementação do kNN permitiu classificar contas em grupos semelhantes, oferecendo uma base interpretável para decisões de marketing e estratégias de conteúdo.

Para futuras melhorias, pretendemos usar:

- **Explorar Outras Técnicas de Machine Learning**: Implementar algoritmos como árvores de decisão ou redes neurais pode oferecer uma análise mais detalhada, especialmente em dados de alta dimensão.
- **Expandir as Variáveis de Análise**: Incluir dados demográficos e comportamento de engajamento mais específico ajudaria a enriquecer a análise.
- **Controle de Outliers**: Filtrar contas com valores atípicos pode refinar as métricas e aumentar a precisão.

Essas melhorias podem aprimorar o modelo, proporcionando uma visão ainda mais robusta para análises de influenciadores.

Referências

1. Scikit-Learn Documentation. Disponível em: https://scikit-learn.org/stable/
2. Instagram Influencer Data Analysis. Artigos e guias sobre análise de redes sociais e engajamento.
3. Bibliografia técnica sobre Machine Learning: Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly Media, 2019.