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

Nome dos Residentes: Iaquine Santos da Silva e Renato Gomez Sousa

Data de Entrega: 17 de novembro de 2024

Resumo

Este relatório descreve a implementação do algoritmo k-Nearest Neighbors (kNN) para analisar dados de influenciadores do Instagram. O objetivo principal é prever a pontuação de influência ("influence_score") dos influenciadores com base em variáveis como número de seguidores, curtidas médias e taxa de engajamento. O conjunto de dados utilizado contém informações sobre os principais influenciadores do Instagram, incluindo dados demográficos e métricas de engajamento. O algoritmo kNN foi implementado utilizando a biblioteca scikit-learn em Python, e os hiperparâmetros foram ajustados para otimizar o desempenho preditivo. Os resultados indicam que o kNN pode ser uma ferramenta eficaz para prever a influência dos influenciadores do Instagram, com algumas ressalvas e limitações.

Introdução

O presente estudo teve como objetivo explorar a aplicação do algoritmo k-Nearest Neighbors (kNN) na análise de dados de influenciadores digitais do Instagram. O kNN, um algoritmo de aprendizado de máquina supervisionado, foi utilizado para construir um modelo preditivo para a variável "pontuação de influência". A escolha do kNN se justifica por sua simplicidade, interpretabilidade e capacidade de lidar com dados não lineares, características importantes para a análise exploratória em questão, que visa entender a relação entre as características dos influenciadores e sua pontuação de influência.

Métricas de Avaliação

Para avaliar o desempenho do modelo preditivo, utilizaremos as seguintes métricas:

- **MAE (Mean Absolute Error):** Calcula a média da diferença absoluta entre os valores reais e os valores previstos. É uma métrica fácil de interpretar, que representa o erro médio de previsão em termos da mesma unidade da variável de interesse.

- **MSE (Mean Squared Error):** Calcula a média do quadrado da diferença entre os valores reais e os valores previstos. É útil para penalizar erros maiores, uma vez que eleva ao quadrado as diferenças.

- **RMSE (Root Mean Squared Error):** É a raiz quadrada do MSE. É útil para expressar o erro em termos da mesma unidade da variável de interesse, ao mesmo tempo que mantém a penalização de erros maiores presente no MSE.

Conjunto de Dados

Os dados utilizados neste estudo foram extraídos do arquivo "top_insta_influencers_data.csv" e contêm informações sobre os principais influenciadores digitais do Instagram, incluindo:

- **rank:** Classificação do influenciador.

- **channel_info:** Informações sobre o canal do influenciador.

- **influence_score:** Pontuação de influência do influenciador.

- **posts:** Número de posts do influenciador.

- **followers:** Número de seguidores do influenciador.

- **avg_likes:** Média de curtidas por post.

- **60_day_eng_rate:** Taxa de engajamento nos últimos 60 dias.

- **new_post_avg_like:** Média de curtidas em novos posts.
- **total_likes:** Total de curtidas no canal.
- **country:** País do influenciador.

O conjunto de dados foi explorado e pré-processado para preparar os dados para o treinamento do modelo kNN, que foi utilizado para prever a variável "influence_score" com base nas outras características dos influenciadores.

Metodologia

A análise da base de dados do Instagram foi realizada utilizando o algoritmo kNN (k-Nearest Neighbors). O kNN é um algoritmo de aprendizado de máquina supervisionado que pode ser usado para resolver problemas de classificação e regressão. Neste caso, o algoritmo foi usado para prever a influência dos usuários do Instagram com base em diversas variáveis, como número de seguidores, número médio de curtidas e taxa de engajamento.

Análise Exploratória

A análise exploratória dos dados revelou insights importantes sobre os influenciadores do Instagram. As variáveis-chave analisadas incluem:

- **Número de seguidores:** Um dos principais indicadores de influência, variando de milhares a milhões.
- **Curtidas médias:** Reflete o engajamento do público com as postagens dos influenciadores.
- **Taxa de engajamento:** Uma métrica que relaciona curtidas, comentários e compartilhamentos com o número de seguidores, expressa em porcentagem.
- **País de origem:** A localização geográfica dos influenciadores, agrupada por continente para análise.

A análise exploratória se concentrou na relação entre essas variáveis e a pontuação de influência ("influence_score"), utilizando gráficos de dispersão, histogramas e tabelas de correlação.

Implementação do Algoritmo

A implementação do algoritmo kNN foi realizada com a biblioteca scikit-learn, versão 0.19.0, em Python. A transformação da variável "country" (país) para "country_code" (código do país), agrupando os países por continente, foi realizada para simplificar a análise e evitar o excesso de categorias. Os continentes foram mapeados para códigos numéricos de 1 a 7, representando América do Norte, América do Sul, Europa, Ásia, Oceania, África e Outros, respectivamente.

Validação e Ajuste de Hiperparâmetros

A validação cruzada com 5 folds foi utilizada para avaliar o desempenho do modelo e ajustar os hiperparâmetros. O método GridSearchCV foi empregado para encontrar o número ideal de vizinhos (k) dentre um intervalo de 1 a 50, considerando a pontuação de influência como métrica de avaliação.

Resultados

Implementação do Algoritmo

O algoritmo kNN foi implementado com as seguintes configurações:

- k = 5 (número de vizinhos mais próximos)
- Métrica de distância = Euclidiana

Validação e Ajuste de Hiperparâmetros

O processo de validação cruzada foi usado para avaliar o desempenho do modelo e otimizar os hiperparâmetros. Os seguintes hiperparâmetros foram ajustados:

- k (número de vizinhos mais próximos)

O melhor valor para k foi 18.

Métricas de Avaliação

Os resultados obtidos com o modelo kNN, após o ajuste de hiperparâmetros, foram:

- **MAE:** 4.64
- **MSE:** 29.91
- **RMSE:** 5.47

Discussão

Os resultados indicam que o modelo kNN pode prever a "influence_score" com razoável precisão, considerando a complexidade do problema e a influência de múltiplos fatores na pontuação de influência. As métricas de erro MAE, MSE e RMSE demonstram que o modelo apresenta um erro relativamente baixo na previsão da "influence_score". No entanto, é importante destacar que o modelo pode apresentar dificuldades em lidar com novas categorias de países, e a transformação da variável "country" para "country_code" pode resultar em perda de informação.

Após a análise dos dados e a implementação do modelo kNN, podemos observar alguns pontos interessantes. Primeiramente, a relação entre o número de seguidores e a média de curtidas por post não se mostrou linear, indicando que o crescimento de um não implica necessariamente no crescimento do outro de forma proporcional.

O impacto da taxa de engajamento em 60 dias na taxa de engajamento geral também não apresentou um padrão claro, sugerindo que a taxa de engajamento recente pode não ser um bom preditor da taxa de engajamento a longo prazo.

É importante notar que a conversão da coluna country para faixas numéricas baseadas em continentes pode ter levado a perda de informação relevante, e a utilização de técnicas como one-hot encoding poderia ser explorada para preservar a informação original dos países.

Além disso, a otimização dos hiperparâmetros do modelo kNN se mostrarou importante para a obtenção de resultados mais precisos.

A análise dos resultados por continente revelou que o modelo teve desempenho variável, o que pode indicar a necessidade de modelos específicos para cada região ou a inclusão de variáveis adicionais que capturem as características de cada continente.

Por fim, a visualização dos resultados através dos gráficos de dispersão e barras permitiu uma melhor compreensão do desempenho do modelo e das relações entre as variáveis, auxiliando na identificação de possíveis melhorias e na formulação de novas hipóteses para futuras análises.

Conclusões e Trabalhos Futuros

O projeto demonstrou a viabilidade do kNN para prever a influência dos influenciadores do Instagram, com algumas ressalvas. Para trabalhos futuros, sugere-se:

- **Aprimorar a categorização de países:** Investigar métodos alternativos para lidar com a variável "country", como a utilização de one-hot encoding ou embeddings.
- **Incorporar novas variáveis:** Incluir outras métricas de engajamento, informações sobre o tipo de conteúdo e dados demográficos mais detalhados.
- **Comparar com outros algoritmos:** Avaliar o desempenho de outros algoritmos de aprendizado de máquina, como regressão linear, árvores de decisão e redes neurais.
- **Implementar em um ambiente de produção:** Criar um sistema para coletar dados de influenciadores em tempo real e gerar previsões de influência automaticamente.

Referências

- ZHENG, Alice; CASARI, Amanda. **Feature engineering for machine learning**: principles and techniques for data scientists. 1. ed. Sebastopol, CA: O'Reilly Media, 2018. ISBN 978-1-491-95324-2.
- HAPKE, Hannes; NELSON, Catherine. **Building machine learning pipelines**: automating model life cycles with TensorFlow. 1. ed. Beijing; Boston; Farnham; Sebastopol; Tokyo: O'Reilly Media, 2020. ISBN 978-1-492-05319-4.
- AVILA, Julian; HAUCK, Trent. **Scikit-learn cookbook**: over 80 recipes for machine learning in Python with scikit-learn. 2. ed. Birmingham; Mumbai: Packt Publishing, 2017. ISBN 978-1-78728-638-2.
- AVILA, Julian; HAUCK, Trent. **Scikit-learn cookbook**: over 80 recipes for machine learning in Python with scikit-learn. 2. ed. Birmingham; Mumbai: Packt Publishing, 2017. ISBN 978-1-78728-638-2.
- DANGETI, Pratap. **Statistics for machine learning**: build supervised, unsupervised, and reinforcement learning models using both Python and R. Birmingham; Mumbai: Packt Publishing, 2017. ISBN 978-1-78829-575-8.
- PATEL, Ankur A. **Hands-on unsupervised learning using Python**: how to build applied machine learning solutions from unlabeled data. 1. ed. Beijing; Boston; Farnham; Sebastopol; Tokyo: O'Reilly Media, 2019. ISBN 978-1-492-03564-0.
- JHA, Suraj. **Top Instagram influencers data cleaned**. Disponível em: https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned. Acesso em: 17 de novembro de 2024.