RESIDÊNCIA
EM SOFTWARE
BAHIA + TECNOLOGIA + EMPREENDEDORISMO

CEPEDI

CIÊNCIA DE DADOS

ARTHUR LAGO MARTINS

JOÃO VICTOR OLIVEIRA SANTOS

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

NOVEMBRO/2024

Resumo

Este relatório técnico documenta a implementação e análise de um modelo preditivo utilizando o algoritmo k-Nearest Neighbors (kNN) aplicado a dados de influenciadores do Instagram. O objetivo foi prever a pontuação de influência com base em variáveis como número de seguidores, curtidas médias e taxa de engajamento. A metodologia incluiu análise exploratória, otimização de hiperparâmetros e validação cruzada. Os resultados destacam a eficácia do KNN para o problema, com métricas como MAE, MSE e RMSE avaliadas, além de sugestões para melhorias futuras.

Introdução

O Instagram é uma plataforma essencial para influenciadores digitais, onde métricas como número de seguidores e taxa de engajamento são cruciais para medir impacto. Este projeto utiliza o algoritmo kNN para prever a pontuação de influência com base em variáveis-chave, explorando como essas métricas se correlacionam com o impacto digital.

Os dados utilizados incluem informações como seguidores, curtidas médias, taxa de engajamento de 60 dias e país de origem. Para facilitar o uso no modelo, a variável de país foi convertida em faixas numéricas correspondentes aos continentes.

Metodologia

Análise Exploratória

A análise inicial dos dados revelou as seguintes características:

- Relação entre seguidores e curtidas médias: Os dados mostram uma correlação moderada positiva entre essas variáveis.
- Impacto da taxa de engajamento de 60 dias: Influenciadores com maior taxa de engajamento tendem a apresentar maior pontuação de influência.

Além disso, os dados foram tratados para normalização e conversão de valores textuais para numéricos.

Implementação do Algoritmo

O kNN foi configurado com as seguintes etapas:

1. **Transformação dos dados:** Variáveis como seguidores e curtidas médias foram normalizadas.
2. **Configuração inicial:** O kNN foi configurado com 5 vizinhos e distância Euclidiana.

3. **Mapeamento de continentes:** A variável country foi convertida em faixas numéricas baseadas no continente.

Validação e Ajuste de Hiperparâmetros

Utilizando o GridSearchCV, os hiperparâmetros foram otimizados para valores de k entre 3 e 11 e para métricas de distância como Euclidiana e Manhattan. A validação cruzada garantiu consistência nos resultados.

Resultados

Métricas de Desempenho

As métricas de avaliação do modelo foram:

- Modelo Inicial:
    - MAE: 5.67
    - MSE: 98.09
    - RMSE: 9.90
- Modelo Otimizado:
    - MAE: 5.50
    - MSE: 91.45
    - RMSE: 9.56

Visualizações

1. Relação entre Seguidores e Curtidas Médias:
   Gráfico de dispersão mostrou uma correlação positiva moderada.
2. Relação entre Rank e Pontuação de Influência:
   Gráfico de barras evidenciou como a pontuação de influência tende a decrescer com ranks mais altos.

Discussão

Os resultados mostraram que o kNN é eficiente para prever a pontuação de influência, especialmente com variáveis normalizadas. Limitações incluem a sensibilidade do kNN a outliers e o fato de que o modelo não considera relações complexas entre variáveis como redes neurais poderiam.

Futuras análises poderiam incluir dados de frequência de postagem e impacto de campanhas específicas para maior precisão.

Conclusão e Trabalhos Futuros

O projeto demonstrou como o kNN pode ser aplicado para prever a pontuação de influência no Instagram. Futuros trabalhos podem explorar o uso de redes neurais ou técnicas como Random Forest para capturar relações não-lineares nos dados.

Referências

Pedregosa, F. et al., *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research*.

Hastie, T., Tibshirani, R., Friedman, J., *The Elements of Statistical Learning*. Springer.

Documentação oficial do Scikit-Learn: https://scikit-learn.org