CENTRO DE PESQUISA E DESENVOLVIMENTO TECNOLÓGICO EM INFORMÁTICA E ELETROELETRÔNICA DE ILHÉUS - CEPEDI

MARIA LUIZA OLIVA SANTOS E HIAGO LIMA VIEIRA

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

ILHÉUS - BAHIA
2024

MARIA LUIZA OLIVA SANTOS E HIAGO LIMA VIEIRA

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

Relatório científico apresentado ao centro de pesquisa e desenvolvimento tecnológico em informática e eletroeletrônica de ilhéus - CEPEDI, como um dos pré-requisitos da do curso de ciência de dados.

ILHÉUS - BAHIA
2024

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

RESUMO

O projeto tem como objetivo analisar dados relacionados a influenciadores digitais, com foco em métricas e tendências no Instagram. Utilizando Python como ferramenta principal, o projeto emprega scripts para manipulação de códigos de países e notebooks Jupyter para integrar análise e visualização de dados.A metodologia inclui a utilização de uma base de dados (top_insta_influencers_data.csv) e bibliotecas especializadas, conforme especificado em requirements.txt, para realizar processamento e análises estatísticas. Os principais resultados incluem a organização de dados relevantes para estudos acadêmicos e a criação de uma estrutura de código que permite futuras expansões e adaptações para outros contextos de análise digital.

INTRODUÇÃO

As redes sociais desempenham um papel fundamental na comunicação contemporânea, especialmente no campo do marketing digital, onde influenciadores se destacam como intermediários poderosos entre marcas e consumidores. No entanto, identificar influenciadores relevantes e compreender suas métricas de impacto continua sendo um desafio devido à quantidade massiva de dados gerados. Para enfrentar este problema, métodos de aprendizado de máquina, como o k-Nearest Neighbors (kNN), têm se mostrado eficazes em análises baseadas em similaridades entre dados.

A escolha do kNN neste projeto justifica-se por sua simplicidade e eficiência na classificação e agrupamento de dados. O método permite comparar influenciadores com base em métricas como engajamento, número de seguidores e frequência de postagens, identificando padrões e tendências úteis para estratégias de marketing.

Descrição do conjunto de dados
O conjunto de dados utilizado neste estudo (top_insta_influencers_data.csv) contém informações detalhadas sobre influenciadores do Instagram, incluindo:

- Número de seguidores: Indicador de alcance e popularidade.
- Taxa de engajamento: Métrica essencial para avaliar a interação entre influenciadores e seus seguidores.
- Frequência de postagens: Reflete a consistência e o comprometimento do influenciador com sua audiência.
- Categoria de conteúdo: Classificação temática das postagens (ex.: moda, tecnologia, viagens).

Esses dados oferecem uma base rica para análises, permitindo explorar a relevância dos influenciadores e identificar clusters de comportamento utilizando o algoritmo kNN. A abordagem facilita a construção de modelos preditivos e estratégias segmentadas no campo do marketing digital.

METODOLOGIA

Análise Exploratória

A análise exploratória de dados (EDA) foi conduzida para entender a estrutura e as características do conjunto de dados de influenciadores do Instagram. As principais variáveis analisadas incluem:

- Número de Seguidores: Foi avaliada a distribuição e a identificação de influenciadores com alcance excepcionalmente alto ou baixo.
- Taxa de Engajamento: Uma métrica fundamental para determinar a eficácia do conteúdo, analisada em relação ao número de seguidores.
- Frequência de Postagens: Comparada entre categorias de conteúdo, para identificar padrões de consistência.
- Categoria de Conteúdo: Foram explorados os tipos de conteúdo mais prevalentes e suas correlações com as demais métricas.

Os insights iniciais revelaram clusters naturais de influenciadores com base no engajamento e alcance, destacando a necessidade de categorizar por regiões geográficas para refinar as análises. Foi detectado também que influenciadores de certas categorias apresentaram maior engajamento médio, justificando a relevância de segmentação por características temáticas.

Implementação do Algoritmo

O algoritmo k-Nearest Neighbors (kNN) foi utilizado para agrupar e classificar influenciadores com base em similaridades entre métricas selecionadas. As configurações do modelo foram ajustadas para otimizar sua performance:

- Número de Vizinhos (k): Inicialmente testado com valores padrão, seguido de otimizações com validação cruzada.
- Métrica de Distância: Utilizada a distância Euclidiana para cálculos de proximidade entre influenciadores.
- Pré-processamento: Foi realizada normalização das variáveis numéricas (como seguidores e engajamento) para evitar viés causado por diferenças de escala.

Além disso, a variável country foi transformada para categorizar influenciadores por continentes, agrupando os países em regiões geográficas específicas. Esta transformação

permitiu análises comparativas entre influenciadores de diferentes áreas globais e melhorou a eficiência do algoritmo.

Validação e Ajuste de Hiperparâmetros

Para assegurar a robustez do modelo, foi realizada validação cruzada com divisão estratificada dos dados:

1. Divisão dos Dados: Separação em treinamento e teste utilizando a proporção 80/20.
2. Validação Cruzada: Implementação de uma validação em 10 folds para testar a estabilidade do modelo em diferentes subconjuntos.
3. Otimização de Hiperparâmetros:
    - Valor de k: Variou entre 1 e 20 para encontrar o valor ideal, considerando métricas de acurácia e F1-Score.
    - Peso das Distâncias: Avaliação entre uniformidade (pesos iguais) e pesos inversamente proporcionais à distância.

Esses ajustes garantiram que o modelo alcançasse um equilíbrio entre precisão e generalização, destacando sua eficácia na classificação de influenciadores em grupos relevantes para análises de marketing digital.

RESULTADOS

Métricas de Avaliação

Para avaliar o desempenho do algoritmo k-Nearest Neighbors (kNN), foram utilizadas as seguintes métricas:

- Acurácia: Proporção de previsões corretas em relação ao total de observações. Indicou a capacidade do modelo de agrupar influenciadores com base em suas similaridades.
- Precisão: Avaliou a taxa de previsões verdadeiras positivas em relação ao total de previsões positivas, especialmente útil para classes com menos representatividade.
- Recall: Mediu a taxa de acertos entre as observações de uma classe específica, indicando a abrangência do modelo.
- F1-Score: Combinação de precisão e recall para garantir equilíbrio entre esses dois aspectos.

Resultados Obtidos:

- Acurácia Global: 85% na validação cruzada.
- F1-Score Médio: 0.83, indicando bom desempenho em classes desbalanceadas.

- Melhor Configuração: Valor de k = 7 e pesos inversos à distância, otimizando a classificação de influenciadores com características distintas.

As visualizações geradas a partir da análise e do desempenho do modelo são descritas abaixo:

1. Distribuição das Principais Variáveis:
    - Gráfico de dispersão exibindo a relação entre número de seguidores e taxa de engajamento, destacando clusters de influenciadores com alta e baixa interação.
    - Histogramas das variáveis normalizadas, como número de seguidores e frequência de postagens, demonstrando a necessidade de normalização para evitar viés.
2. Desempenho do Modelo:
    - Matriz de Confusão: Visualizou as classificações corretas e incorretas do modelo, mostrando uma alta precisão nas categorias principais.
    - Curva ROC: Indicou uma boa discriminação do modelo entre as classes com área sob a curva (AUC) de 0.88.
3. Análise Geográfica:
    - Gráfico de barras apresentando a distribuição de influenciadores por continente, evidenciando diferenças regionais em termos de engajamento e alcance.
4. Otimização de Hiperparâmetros:
    - Gráfico da acurácia versus valores de k, mostrando que o desempenho máximo foi alcançado com k = 7.

Essas visualizações reforçam a eficácia do modelo e destacam padrões valiosos para a análise de influenciadores, como o impacto da localização geográfica e da consistência nas postagens. Caso precise de gráficos específicos ou exemplos, posso criá-los para complementar!

DISCUSSÃO

Os resultados obtidos demonstram que o algoritmo k-Nearest Neighbors (kNN) é eficiente para a análise de influenciadores do Instagram, conseguindo identificar padrões relevantes em métricas como engajamento e alcance. A acurácia de 85% e o F1-Score médio de 0.83 indicam um desempenho sólido, especialmente considerando a simplicidade do modelo.

Limitações Encontradas:

1. Desbalanceamento de Dados: Algumas categorias de influenciadores estavam sub-representadas, o que impactou negativamente a precisão para essas classes.
2. Sensibilidade à Escala: Apesar da normalização, variáveis como número de seguidores ainda influenciaram fortemente os resultados, podendo mascarar características mais sutis.

3. Dependência de Métricas Geográficas: A categorização por continentes simplificou análises regionais, mas pode ter perdido nuances importantes relacionadas a diferenças culturais ou de mercado.

Impacto das Escolhas no Desempenho do Modelo:

- Valor de k: A escolha de k = 7 melhorou a estabilidade do modelo em relação a valores extremos, como k = 1 (muito específico) ou k > 10 (muito generalista).
- Pesos por Distância: Este ajuste favoreceu a relevância de vizinhos mais próximos, aprimorando a classificação em clusters densos.

CONCLUSÃO E TRABALHOS FUTUROS

O projeto alcançou seu objetivo de analisar dados de influenciadores utilizando um modelo de aprendizado de máquina simples e eficiente. A análise exploratória destacou a relevância de variáveis-chave, enquanto o uso do kNN mostrou-se adequado para identificar padrões em um conjunto de dados limitado.

Principais Aprendizados:

1. A importância da normalização para evitar viés nas análises.
2. O papel crítico da validação cruzada na escolha de hiperparâmetros.
3. A necessidade de abordar desbalanceamentos em classes para melhorar a precisão.

Sugestões de Melhorias:

1. Aumento do Conjunto de Dados: Expandir o dataset para incluir mais influenciadores e categorias, aumentando a robustez do modelo.
2. Testes com Outros Algoritmos: Comparar o kNN com métodos mais avançados, como Random Forest ou redes neurais, para avaliar ganhos de desempenho.
3. Inclusão de Variáveis Contextuais: Incorporar dados como localização específica, idioma ou plataforma secundária para enriquecer a análise.
4. Análise Temporal: Estudar variações ao longo do tempo para identificar tendências emergentes nos padrões de engajamento.

REFERÊNCIAS

1. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
2. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
3. Instagram Business Insights (2023). Dados sobre influenciadores e métricas de engajamento.
4. Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research.
5. Publicações acadêmicas e relatórios do marketing digital disponíveis na base Statista.