RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE KNN

JOSÉ EDUARDO MENDES DA SILVA

TAÍS FARIAS RIBEIRO DE SOUZA

17 DE NOVEMBRO DE 2024

RESUMO

O seguinte relatório descreve a implementação e análise do algoritmo k-Nearest Neighbors (kNN) aplicado a dados de influenciadores do Instagram. Essa análise inclui desde a transformação dos dados até a otimização do modelo e a validação cruzada. O modelo foi avaliado com métricas de erro e discutiu-se o impacto das escolhas feitas.

1. INTRODUÇÃO

A influência de dados de redes sociais na análise de mercado tem sido amplamente estudada, especialmente no contexto do Instagram. A popularidade de uma conta pode ser medida por métricas como seguidores, curtidas médias e engajamento. Com isso, este projeto usa o kNN para prever o score de influência, explorando correlações entre atributos importantes como número de seguidores e taxa de curtidas.
O kNN foi escolhido por sua simplicidade e eficiência em problemas de regressão onde a relação entre variáveis não é linear.

2. METODOLOGIA

– Análise Exploratória:
A base de dados analisada contém informações detalhadas dos principais influenciadores do Instagram, incluindo atributos como o rank geral, o número de seguidores, o total de curtidas acumuladas, a taxa de engajamento média nos últimos 60 dias e o score de influência. Como parte do pré-processamento, houve a transformação da variável country em valores numéricos categorizados por continentes. Por exemplo, os países da América do Sul foram mapeados para a faixa de 1 a 9, enquanto os da Europa foram convertidos para valores entre 40 e 49. Essa transformação permitiu que o modelo interpretasse melhor a localização geográfica.

Para lidar com dados faltantes, foram aplicadas técnicas de imputação. Valores numéricos foram substituídos pela média, enquanto os valores não numéricos foram preenchidos com a moda. Realizou-se uma análise exploratória inicial usando gráficos como pairplots e scatterplots, que revelaram algumas correlações importantes.
Com os dados devidamente preparados, dividimos o conjunto em dados de treino e teste, na proporção de 80% para treino e 20% para teste. Para que as diferentes escalas dos atributos não distorcessem os cálculos de distância do kNN, normalizamos os dados utilizando o StandardScaler.

– Implementação do Algoritmo kNN:

O algoritmo k-Nearest Neighbors foi implementado em Python, usando a biblioteca Scikit-Learn. O inicio se deu com a utilização de uma configuração padrão e o treinamento do modelo com os dados normalizados. Foi realizada uma avaliação inicial do desempenho calculando métricas como o erro absoluto médio (MAE), o erro quadrático médio (MSE) e a raiz do erro quadrático médio (RMSE). Os resultados iniciais forneceram uma base para entender como o modelo se comportava sem ajustes.

Para validar a robustez do kNN, empregamos a técnica de validação cruzada, que divide os dados de treino em diferentes subconjuntos para garantir que o modelo não estivesse superajustado a um único conjunto de dados. O desempenho foi medido através do erro médio negativo, em que foi observada a variação nos erros para avaliar a consistência do modelo.

– Validação e Ajustes Hiperparâmetros:
Para melhorar a precisão do kNN, utilizamos o método GridSearchCV para encontrar os melhores hiperparâmetros, como o número de vizinhos mais próximos (k) e o tipo de métrica de distância (euclidiana ou de Manhattan). O processo de busca revelou os parâmetros mais adequados, que resultaram em um modelo mais otimizado e com melhor performance.

Além disso, houve a utilização da normalização dos dados, que mostrou ser fundamental para o algoritmo kNN, já que ele é altamente sensível à escala das variáveis. Se avaliou o impacto da normalização especificamente em atributos como followers, avg_likes e total_likes, observando melhorias significativas na acurácia do modelo.

3. RESULTADOS

Os resultados foram analisados usando diferentes métricas de erro. O MAE, MSE e RMSE forneceram uma visão clara de como o modelo estava se desempenhando em termos absolutos e quadráticos. A validação cruzada produziu uma média de RMSE consistente, indicando que o modelo era confiável em diferentes subconjuntos de dados. A otimização através do GridSearchCV revelou que um valor de k entre 5 e 7, combinado com a métrica de distância de Manhattan, produzia os melhores resultados.

Visualizações como scatterplots, foram utilizadas para ilustrar a relação entre seguidores e curtidas médias, o que confirmou as correlações observadas na análise exploratória inicial.

4. DISCUSSÃO

Os resultados demonstraram que o kNN é eficaz para este tipo de problema, mas também revelou algumas limitações. O desempenho do modelo é altamente influenciado pela escolha de k e pela normalização dos dados. Além disso, a categorização de países em faixas numéricas pode ter introduzido alguma imprecisão na representação geográfica. Uma discussão crítica sugere que melhorias poderiam incluir a experimentação com outras técnicas de machine learning ou o uso de métodos de seleção de características para reduzir a dimensionalidade.

5. CONCLUSÃO E TRABALHOS FUTUROS

O projeto proporcionou uma compreensão aprofundada do funcionamento do kNN e das gradações envolvidas na análise de dados de redes sociais. Para trabalhos futuros, seria interessante explorar outros algoritmos de regressão, como Random Forest ou modelos baseados em redes neurais, e comparar o desempenho com o kNN.

6. REFERÊNCIAS

JHA, Suraj. Top Instagram Influencers Data Cleaned. Disponível em: https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned. Acesso em: 17 nov. 2024.