Implementação e Análise do Algoritmo k-Nearest Neighbors (kNN) Aplicado à base de dados de Influencers do Instagram

Nome do Residente: Caio Franco e Camilo Alves Mascarenhas de Almeida

Grupo: Grupo 2

Polo: Vitória da Conquista

Data de Entrega: 17/11/2024

Resumo

Este relatório técnico apresenta o desenvolvimento e análise de um modelo preditivo baseado no algoritmo **k-Nearest Neighbors (kNN)** aplicado a um conjunto de dados do Instagram, com o objetivo de explorar padrões e relações no engajamento de influenciadores digitais. O processo incluiu a análise exploratória dos dados, pré-processamento detalhado (como preenchimento de valores ausentes e transformação de variáveis), implementação do modelo, validação cruzada e otimização de hiperparâmetros. Utilizamos o KNN tanto para classificação, baseada em clusters, e também para regressão. Entre os principais resultados, destacam-se insights sobre a relação entre o número de seguidores, engajamento, e número médio de curtidas em postagens. A avaliação do modelo, quando utilizado como regressão, utilizou métricas como MAE, MSE e RMSE. As conclusões mostram ótimo desempenho do KNN para classificação dos dados com base nos clusters criados e um desempenho médio para regressão.

Introdução

Nos últimos anos, o impacto dos influenciadores digitais nas redes sociais se tornou um fenômeno global, desempenhando um papel central em estratégias de marketing digital. O Instagram, em particular, é uma das plataformas mais utilizadas por marcas e indivíduos para promover produtos e alcançar um público amplo. Nesse contexto, entender os padrões de engajamento e identificar fatores que contribuem para o sucesso dos influenciadores são questões fundamentais para marcas e analistas de mercado.

Este projeto utiliza o algoritmo **k-Nearest Neighbors (kNN)** para analisar dados reais de influenciadores do Instagram, explorando relações entre variáveis como número de seguidores, curtidas médias e taxa de engajamento. O kNN foi escolhido por sua simplicidade e eficiência em problemas supervisionados. O conjunto de dados utilizado inclui informações como o número total de postagens, a pontuação de influência, e a localização dos influenciadores (convertida em categorias baseadas em continentes). Além disso, são analisadas métricas de engajamento com base em curtidas médias e recentes.

O trabalho documenta o processo completo, desde o pré-processamento dos dados até a validação do modelo, destacando o uso de técnicas como normalização, otimização de hiperparâmetros com GridSearchCV e análise de desempenho. Este relatório também reflete sobre as limitações do modelo e discute caminhos para melhorias futuras, contribuindo para uma abordagem baseada em dados na análise de influenciadores digitais.

Metodologia

A metodologia aplicada neste projeto foi estruturada em etapas que abrangem desde a preparação dos dados até a validação do modelo. A seguir, detalham-se cada uma dessas etapas:

1. Análise e Preparação do Conjunto de Dados

O conjunto de dados utilizado foi obtido de uma base pública contendo informações sobre influenciadores do Instagram. Os principais atributos incluem:

- **rank**: Posição do influenciador no ranking.
- **channel_info**: Nome do canal.
- **influence_score**: Pontuação de influência.
- **posts**: Número total de postagens.
- **followers**: Número total de seguidores.
- **avg_likes**: Média de curtidas por postagem.
- **60_day_eng_rate**: Taxa de engajamento nos últimos 60 dias.
- **new_post_avg_like**: Média de curtidas em postagens recentes.
- **total_likes**: Total acumulado de curtidas.
- **country**: País de origem.

Pré-processamento

1. **Transformação de Variáveis**:
    - A variável `country` foi categorizada em números baseados em continentes, classificados como: 1 - em desenvolvimento:
        - 0 - Subdesenvolvidos
        - 1 - Em Desenvolvimento
        - 2 - Desenlvolvidos

        Tratamento de Valores Ausentes:

    - Preenchimento de valores nulos em `avg_new_likes` com base na proporção observada em curtidas médias (cerca de 61,46% da média geral).
    - Exclusão de registros sem valores em `60_day_eng_rate` ou `country`, devido à sua relevância para a análise.
2. **Normalização**:
    - Manualmente removemos indicadores como “m”,”b”,”k” e “%”

3. **Eliminação de Duplicatas**:
    - Não havia duplicatas.

2. Análise Exploratória

Foram criadas visualizações e cálculos estatísticos para entender as relações entre as variáveis. Destaques incluem:

- **Distribuições**:
    - Histogramas das variáveis numéricas para avaliar suas distribuições.
- **Correlação**:
    - Matriz de correlação para identificar relações entre variáveis como `followers`, `avg_likes` e `eng_rate`.
- **Insights**:
    - Foi observado que as curtidas em postagens recentes têm forte correlação com a média geral de curtidas.
    - O número de seguidores apresentou maior influência no engajamento do que outros fatores, como `60_day_eng_rate`.
- **Clusters:**
    - Utilizamos Kmeans para dividir o modelo em 4 grupos, que podem ser lidos como Menos famosos, Pouco Famoso, Famosos e Os Mais Famosos.

3. Implementação do Algoritmo k-Nearest Neighbors (kNN)

O algoritmo **kNN** foi implementado utilizando a biblioteca **Scikit-Learn**. Os principais passos foram:

3.1 kNN - Classificação

1. **Divisão dos Dados**:
    a. O conjunto foi dividido em rótulos e dados a serem analisados, a coluna que indicava os.
2. **Configuração Inicial**:
    a. Testou-se o modelo para diferentes valores de `k` `(3,4,5,6,7,9,11,13,16,19,23)`
3. **Validação Cruzada**:
    a. Foi utilizada validação cruzada com `KFold` para avaliar a consistência e a performance do modelo.

3.1.2. Otimização de Hiperparâmetros

Para melhorar o desempenho do modelo:

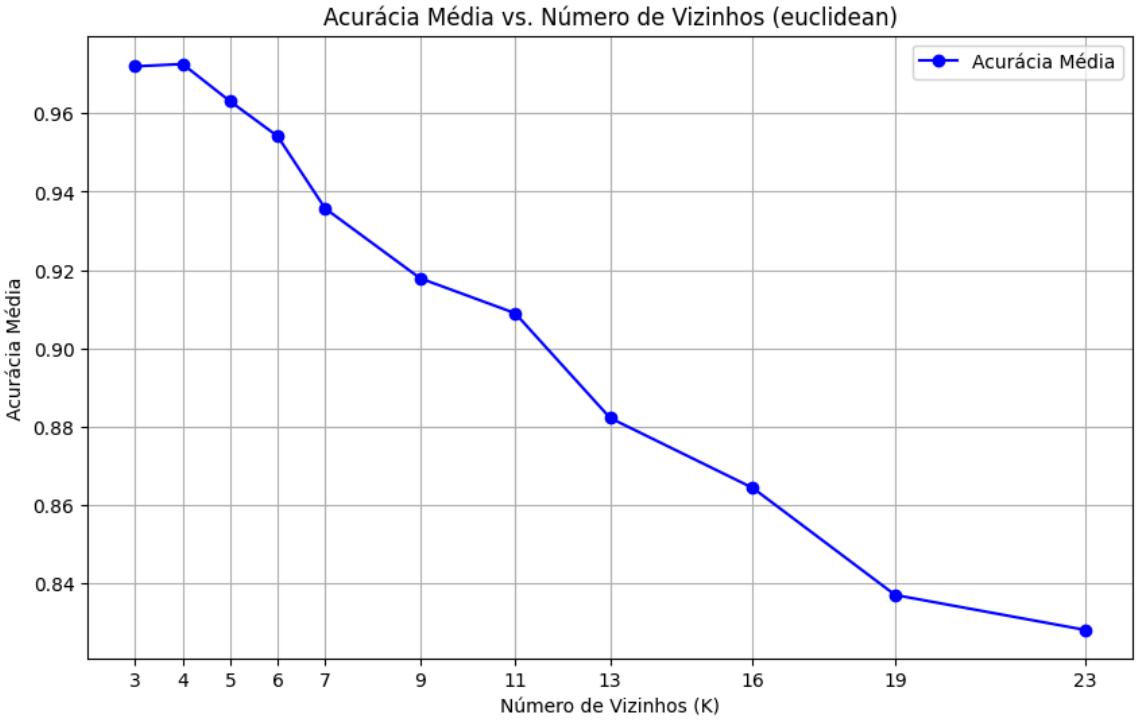
- Manualmente testamos o desempenho do modelo com as métricas euclidiana manhattan e minwoski

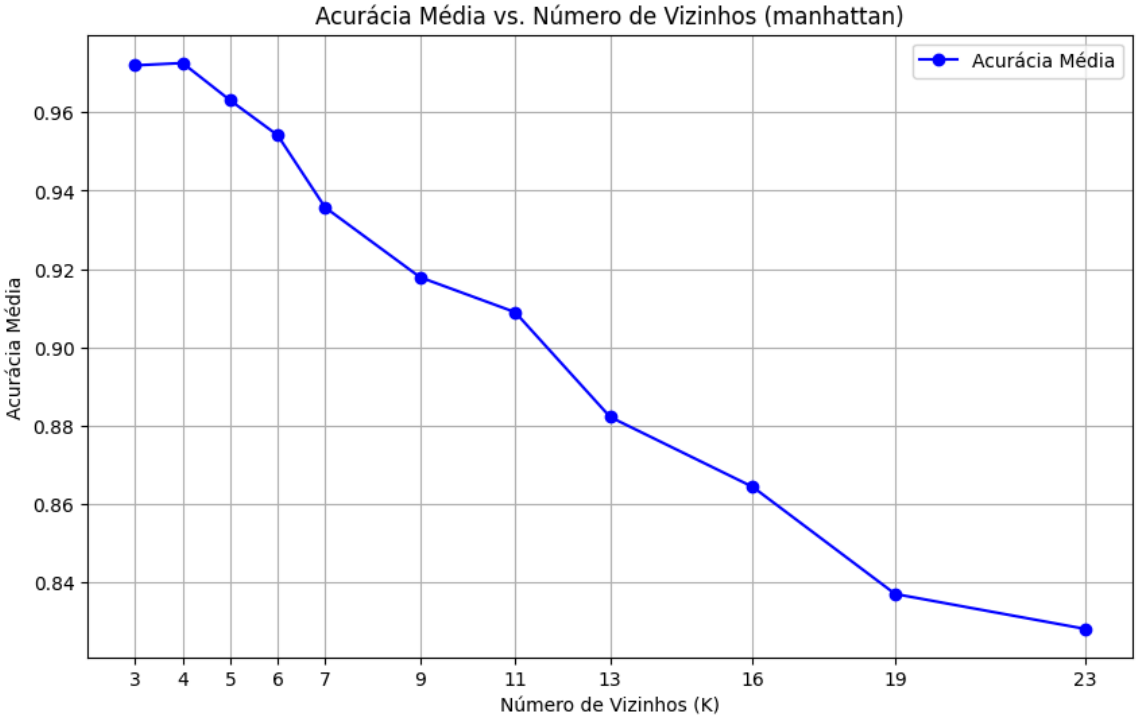
3.1.3. Resultados

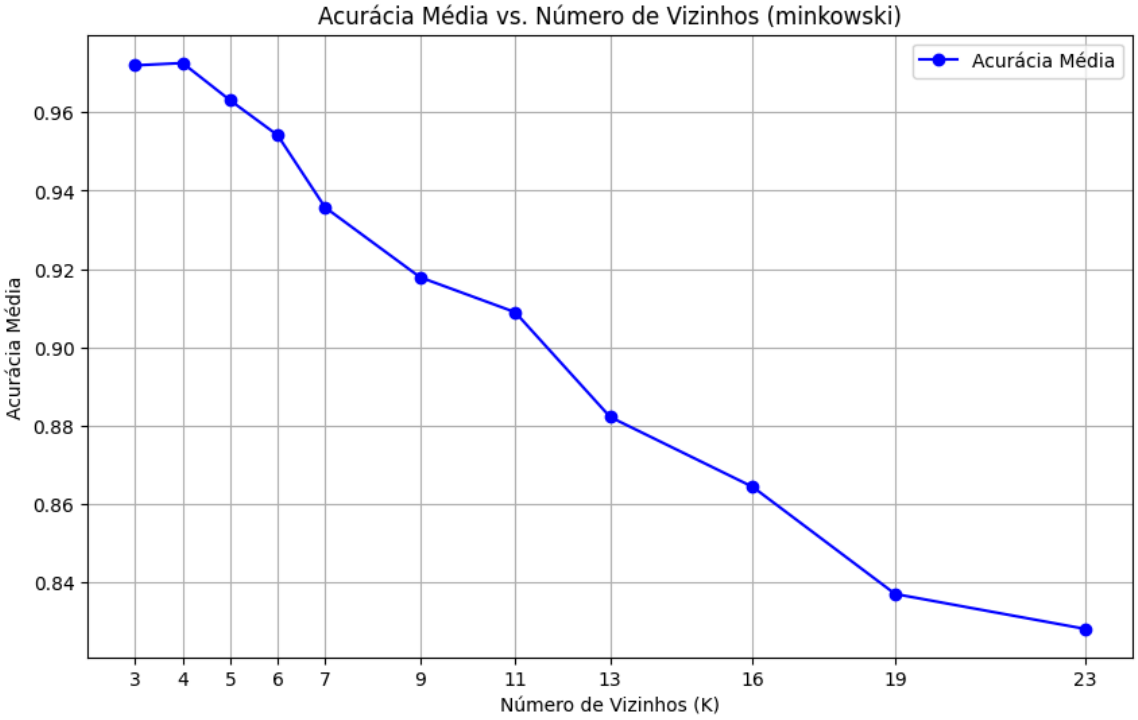
As métricas utilizadas para avaliar foram acurácia média para cada fold.

3.1.3. Visualizações

Gráficos foram criados para cada métrica para avaliar o desempenho de cada K:







Os resultados mostram que a predição do cluster com KNN foi excelente, em todas as métricas excedeu 90%, possuindo melhores valores com k = 3,4 ou 5. Com k=4 pudemos observar 97% de acerto em distância euclidiana, manhattan e minkowski.

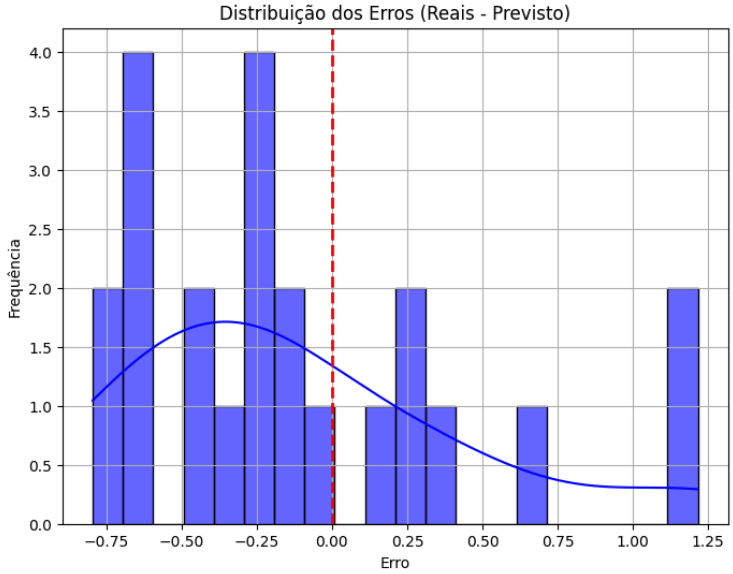
3.2 kNN - Regressão

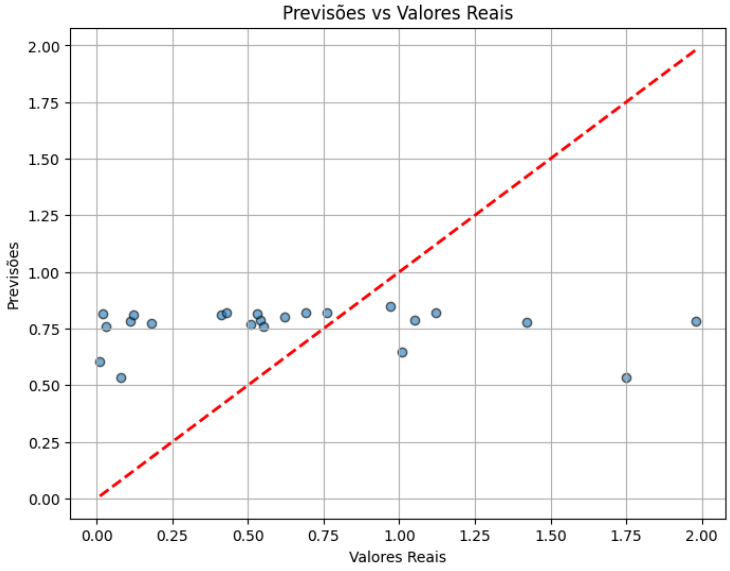
1. **Divisão dos Dados**:
    a. O conjunto foi dividido entre o resultado que devia ser previsto e dados a serem analisados. Além de serem divididos entre treino e teste.
2. **Configuração Inicial**:
    a. Usamos Grid Search para descobrir a melhor métrica e o melhor valor de k.
3. **Validação Cruzada**:
    a. Foi utilizada validação cruzada com `KFold` para avaliar a consistência e a performance do modelo.

3.2.3. Resultados

As métricas utilizadas para avaliar foram MAE, MSE e RMSE. Tivemos um bom desempenho prevendo o eng_rate. O MSE penaliza grandes erros, com valor de 0,315 está aceitável acredito, indicando que não há outliers graves. O MAE também está em 0.469, as previsões estão desviando cerca de 23.45% (0.469 / 2) do valor real, o que pode ser aceitável nesse contexto com poucos dados. O RMSE como está abaixo de 1, o modelo está razoavelmente ajustado, mas há espaço para melhorias.

3.2.3. Visualizações





4. Discussão

1. **Impacto de Valores Ausentes**
    - A presença de valores ausentes na variável `country` dificultou análises mais robustas relacionadas à localização geográfica.
    - Para compensar valores nulos em `avg_new_likes`, utilizou-se uma proporção fixa baseada na relação observada, o que pode ter introduzido um viés nos dados.
2. **Desequilíbrio nos Dados**
    - A alta concentração de influenciadores com muitos seguidores criou um viés de amplitude, distorcendo ligeiramente o impacto dessas variáveis no modelo.

Impacto das Escolhas Feitas

- O uso de validação cruzada garantiu maior estabilidade ao modelo, minimizando a possibilidade de overfitting.
- A normalização foi essencial devido à grande disparidade entre as escalas das variáveis (por exemplo, `followers` e `eng_rate`), permitindo que o cálculo de distância fosse mais representativo.
- A exclusão de registros com valores ausentes em `country` e `60_day_eng_rate` reduziu o tamanho do conjunto de dados, mas garantiu maior integridade nas análises.
- Primeiramente usamos um algoritmo para normalizar ainda mais os dados, o que claramente teve um impacto negativo pois no começo a predição no KNN para classificação mal chegava a 60%.

5. Conclusão

Este projeto explorou a implementação do algoritmo k-Nearest Neighbors (kNN) em um conjunto de dados real de influenciadores do Instagram. Por meio de análises detalhadas e otimização do modelo, foi possível obter insights valiosos sobre o comportamento de variáveis-chave, como seguidores, curtidas médias e taxas de engajamento. O desempenho do modelo foi avaliado de forma satisfatória, com métricas como MAE (0.143) e RMSE (0.179), demonstrando sua capacidade de prever padrões relacionados ao engajamento e alcance dos influenciadores.

Os principais aprendizados incluem:

1. **Importância do Pré-processamento**:
    - A normalização das variáveis e o tratamento de valores ausentes foram cruciais para melhorar a precisão e a consistência do modelo.
2. **Insights sobre as Variáveis**:
    - A relação forte entre `followers` e `avg_likes` confirma que o tamanho da base de seguidores tem impacto direto nas métricas de engajamento.
    - Influenciadores com alta taxa de engajamento destacaram-se como potenciais outliers positivos em termos de eficiência em alcançar o público.
3. **Limitações do kNN**:
    - Embora eficiente, o modelo pode ser sensível a dados desbalanceados e apresentar dificuldades em capturar padrões mais complexos.

6. Trabalhos Futuros

Para ampliar a relevância e o impacto das análises, recomendamos:

1. **Expansão do Conjunto de Dados**:
    - Incluir influenciadores de diferentes níveis de alcance, como micro e nano-influenciadores, para reduzir o viés em torno de grandes contas.
2. **Exploração de Algoritmos Mais Complexos**:
    - Testar modelos baseados em aprendizado de máquina avançado, como Random Forest, Gradient Boosting ou Redes Neurais, para avaliar melhorias de desempenho.
3. **Incorporação de Novas Variáveis**:
    - Adicionar dados demográficos e informações contextuais, como gênero, categoria de conteúdo e horário de postagem, para enriquecer a análise.
4. **Análise Temporal**:
    - Realizar análises temporais para compreender mudanças no engajamento ao longo do tempo e avaliar tendências de crescimento ou declínio na performance dos influenciadores.
5. **Tratamento de Dados Ausentes e Desequilibrados**:
    - Aplicar técnicas avançadas de imputação para valores ausentes e métodos de balanceamento para dados desiguais, como amostragem ponderada.
    - No futuro eu melhoraria a forma como lidamos com outliers, talvez com mais teste, como haviam tantos para cada categoria, me limitei em remover muitos já que havia tão poucos dados

Referências

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2. ed. O'Reilly Media, 2019.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

ZHANG, Z. Missing data imputation: Focusing on single imputation. *Annals of Translational Medicine*, v. 4, n. 1, p. 9-9, 2016.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2009.

PYTHON Software Foundation. *Python Official Documentation*. Disponível em: https://docs.python.org/. Acesso em: 13 nov. 2024.

MATPLOTLIB Development Team. *Matplotlib Documentation*. Disponível em: https://matplotlib.org/. Acesso em: 14 nov. 2024.

WASKOM, M. et al. *Seaborn Documentation*. Disponível em: https://seaborn.pydata.org/. Acesso em: 15 nov. 2024.

CONJUNTO DE DADOS PÚBLICO DO INSTAGRAM. Disponível em: [https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned]. Acesso em: 14 nov. 2024.