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

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

Weslei Ferreira
Daniel Morais Pereira

17 de novembro de 2024

## Resumo

O objetivo deste estudo é descrever em detalhes todo o processo de implementação de avaliação do algoritmo *k-Nearest Neighbors (kNN)* aplicado à análise de dados no instagram, fazendo uso de metodologia experimental onde foi abordado a análise exploratória dos dados referentes ao uso da plataforma *Instagram* , graficos e métodos de otimização do modelo também foram devidamente aplicados nesse estudo.

# Introdução

Nos dias atuais com o avanço tecnológico e a democratização do acesso a internet para as grandes massas as redes sociais se tornaram algo cada vez mais presente nas vidas dos brasileiro no tocante a criação e consumo de conteúdos digitais, nesse sentido a plataforma *Instagram* se destaca por conta da sua alta popularidade no mundo, nela os influenciadores utilizam sua base de usuários para promover produtos através das publicidades ou as famosas #*publi*. Esse novo cotidiano gerou uma demanda por modelo de aprendizado de máquina que identificam e classificam a influência desses criadores de conteúdo, para que dessa forma as grandes marcas possam fazer campanhas cada vez mais eficientes e que as mesmas sejam mais assertivas em buscar seus públicos alvos.

## Contextualização do Problema

Quando falamos em detectar ou identificar a influência de perfis dentro da plataforma *Instagram* estamos falando em executar uma tarefa extremamente minuciosa de variáveis como o número de seguidores, taxa de análise e média de curtidas. Quando esses dados estão em um grande número de instância uma análise manual ou uma com base em regras fixas se torna impraticável. Nesse contexto, fazer uso de técnicas de aprendizagem de máquina, como por exemplo, o KNN, podem ser um ótimo ponto de partida para classificar ou prever um perfil com base em características devidamente observadas.

## Justificativa para o Uso do kNN

O algoritmo *k-Nearest Neighbors (kNN)* usa um conceito de aprendizagem de máquina que é baseado em distância, aqui é usado os vizinhos mais próximos para classificar um determinado elemento K, baseado nessa definição o KNN é ideal para a atividade em questão onde os perfis dos influenciadores serão classificados com base na proximidade entre características, dessa forma não precisamos assumir pressupostos rígidos sobre a distribuição dos dados. Além disso, a simplicidade do KNN aliada a otimização dos parâmetros e técnicas de validação cruzada, permitem um bom ajuste ao problema.

Além de justificar o uso do kNN, realizamos uma comparação com outros algoritmos, como o Gradient Boosting Regressor (GBR), Regressão Linear e Random Forest. Essa comparação permitiu avaliar a eficácia e adequação de diferentes abordagens ao problema proposto. Os resultados indicaram que, embora cada método tenha suas próprias vantagens, o kNN se destacou por sua simplicidade e desempenho consistente na tarefa de classificação dos perfis dos influenciadores, especialmente em cenários onde a relação entre as características não segue distribuições rígidas. A análise comparativa reforça a escolha do kNN, evidenciando sua capacidade de adaptação e precisão após a otimização de parâmetros e validação cruzada.

## Descrição do Conjunto de Dados

O conjunto de dados analisado contém informações detalhadas sobre influenciadores no Instagram, com os seguintes atributos:

- **Rank:** Posição do influenciador em uma lista de influência.
- **Channel_info:** Informações sobre o perfil do influenciador.
- **Influence_score:** Pontuação atribuída ao perfil com base na sua influência.
- **Posts:** Número total de postagens feitas pelo influenciador.
- **Followers:** Quantidade de seguidores do perfil.
- **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 novas postagens.
- **Total_likes:** Quantidade total de curtidas no perfil.
- **Country:** País ou região do influenciador.

# Metodologia

A metodologia deste trabalho consiste em ser mais experimental. Em ciências de dados essa metodologia costuma ser bastante aplicada em projetos dessa área, a seguir segue em detalhes o passo a passo seguido para a devida implementação do trabalho requisitado.

## Definição do problema e Objetivo

A primeira etapa consiste em definir o problema em questão, que no caso deste trabalho é analisar e prever o nível de influência de perfis da plataforma Instagram. Aqui o objetivo é aplicar técnicas de aprendizado de máquina para identificar padrões entre influenciadores e prever o impacto desses perfis com base em suas métricas de desempenho.

## Coleta de preparação dos dados

O conjunto de dados foi obtido através da plataforma *"Keggle".* Esta plataforma reúne não só diversas bases de dados de diferentes áreas, mas também agrupa diversos desafios para que os usuários dessa plataforma possam aprimorar os seus conhecimentos. Além disso, o "*Keggle*" também é conhecido por ter uma comunidade de aprendizado bastante colaborativa, onde diversos cientistas de dados interagem, compartilham informações, códigos e auxiliam os demais usuários a solucionar seus problemas.

## Pré-processamento

O conjunto de dados passa por um processo de limpeza e transformação, como a conversão de valores formatados em texto (por exemplo, 'k' ou 'm' para milhares e milhões, respectivamente) para valores numéricos. Essas alterações foram aplicadas às variáveis *followers, avg_likes e total_likes*. Nessa etapa a variável *country* (país) foi transformada em uma numérica, organizada por faixas baseadas em continentes.

Uma segunda função foi criada para converter valores percentuais representados como strings (por exemplo, '10%') em floats entre 0 e 1. Essa conversão é aplicada à variável *60_days_eng_rate*, transformando percentuais em frações decimais. Após a aplicação da função, a variável foi convertida para o tipo float para garantir a consistência de tipo.

## Variável followers_x_engagement

Foi necessário criar uma variável que representasse a interação entre a variável que armazena o número de seguidores (*followers*) e a variável que armazena a taxa de engajamento no período de 60 dias (*60_day_eng_rate*). A importância dessa nova variável é auxiliar na previsão do *influence score.*

## Remoção de Outliers

Desenvolvemos também uma função para identificar e remover os outliers encontrados no *dataset.* A função usa a regra do intervalo interquartil (IQR) para definir os limites superior e inferior, removendo quaisquer valores que estejam fora desse intervalo (ou seja, 1,5 vezes o IQR além do primeiro e terceiro quartis). Depois de criada a função, aplicamos às variáveis *followers, avg_likes, total_likes, followers_x_engagement.*

## Seleção de Features e Divisão dos Dados

Por fim, selecionamos as variáveis que serão as independentes (features) e a variável que será a dependente (target). As features escolhidas foram as variáveis *followers, avg_likes, total_likes, country_numeric e followers_x_engagement,* enquanto a variável target foi *influence_score*. Os dados foram divididos em conjuntos de treino e teste usando a função train_test_split, com uma proporção de 80% para treino e 20% para teste, e uma random_state fixa de 42 para garantir a reprodutibilidade dos resultados.

## Normalização

Esta já foi usada para evitar que as variáveis tenham um impacto equilibrado, as variáveis com ampla variação numérica, como *followers* e *total_likes,* foram devidamente normalizadas.

## Análise exploratória dos dados (EDA)

Uma análise exploratória dos dados é uma etapa fundamental no processo de investigação de um conjunto de dados, pois com ela, é possível uma compreensão inicial dos dados e identificar padrões e anomalias no *dataset. Ela* foi realizada para entender a distribuição das variáveis e a presença de correlações, abaixo segue os principais pontos seguidos dentro do EDA.

## Distribuição das Variáveis Principais

**Followers:** Aqui essa variável apresenta uma grande diferença em uma ou mais instâncias, isso acontece porque alguns influenciadores possuem dezenas de milhões de seguidores, enquanto outros perfis possuem apenas alguns milhares ou até mesmo menos que isso. Essa forma de distribuição faz com que essa variável apresente uma distribuição assimétrica.

**Avg_like:** Esta é uma variável que mede a média de curtidas, nela é medida o engajamento direto do público. Aqui é analisado a distribuição das curtidas indica a relação com o número de seguidores e como o público reage aos conteúdos desses perfis.

**60_day_eng_rate:** A taxa de engajamento nos últimos 60 dias (60_day_eng_rate) ajuda a capturar o nível recente de interação e popularidade do influenciador, sendo uma métrica relevante para entender tendências e potenciais crescimentos.

**new_post_avg_like**: A média de curtidas que os perfis do dataset obtiveram nas suas novas publicações.

**total_Likes**: Aqui indica o total de curtidas que o usuário obteve em suas postagens. (em bilhões)

**country**: Essa variável indica o país ou região de origem do usuário.

## Implementação do Algoritmo

Para a implementação do modelo de regressão, optamos pelo uso do algoritmo k-Nearest Neighbors (kNN), devido à sua simplicidade e capacidade de capturar relações não lineares nos dados. Inicialmente, realizamos a normalização dos dados, uma etapa essencial visto que o kNN é sensível às escalas das variáveis. Para isso, utilizamos a classe *StandardScaler* do *Scikit-learn*, escalonando tanto o conjunto de treino quanto o de teste para assegurar que todas as variáveis tivessem a mesma importância relativa.

Em seguida, configuramos o modelo de regressão *KNeighborsRegressor* e realizamos a otimização dos hiperparâmetros utilizando a técnica *GridSearchCV*. Essa abordagem nos permitiu explorar diferentes valores de k (número de vizinhos) e métricas de distância (euclidiana e manhattan), aplicando uma validação cruzada com **cv=5** para garantir uma avaliação consistente do modelo em subconjuntos dos dados de treino e evitar problemas de overfitting.

Após essa etapa, avaliamos o desempenho do modelo otimizado no conjunto de teste, calculando métricas como o erro médio absoluto (MAE), o erro médio quadrático (MSE) e a raiz do erro quadrático médio (RMSE), que forneceram uma medida quantitativa da precisão do modelo.

## Validação e Ajuste de Hiperparâmetros

O processo de validação e ajuste de hiperparâmetros é crucial para garantir a eficácia do modelo e melhorar sua capacidade preditiva. Nesta seção, detalhamos a abordagem utilizada para a validação cruzada e a otimização dos parâmetros do modelo kNN.

## Validação Cruzada

A validação cruzada é uma técnica usada para avaliar a performance do modelo de forma mais robusta, dividindo o conjunto de dados em diferentes subconjuntos de treino e teste. O objetivo é evitar o overfitting e garantir que o modelo se generalize bem para dados novos.

A validação k-fold foi através da validação cruzada, onde os dados são divididos em k subconjuntos e o modelo é treinado e validado k vezes, usando *GridSearchCV* para encontrar os melhores valores de k.

## Resultados

No treinamento de teste do modelo *K-Nearest Neighbors (KNN),* foi obtido os seguintes valores para as métricas de erro:

- **Melhor k**: 10
- **Melhor métrica de distância**: Manhattan
- **MAE (Mean Absolute Error)**: 8.33
- **MSE (Mean Squared Error)**: 229.83
- **RMSE (Root Mean Squared Error)**: 15.16

**Análise:**

- O valor do **MAE** indica que, em média, o modelo comete um erro absoluto de aproximadamente 8,33 unidades ao prever os valores das variáveis de interesse.
- O **RMSE**, este penaliza os erros maiores, aqui no KNN o desvio padrão que foi mostrado é de aproximadamente de 15,16 unidades, o que representa um valor baixo.
- Nesse treinamento foi escolhida a distância Manhattan como a melhor, isso acontece devido ao seu desempenho consistente com conjunto de dados multivariados onde as diferenças absolutas são mais representativas. Abaixo serão mostrados gráficos que demonstram o desempenho do KNN durante o treinamento.

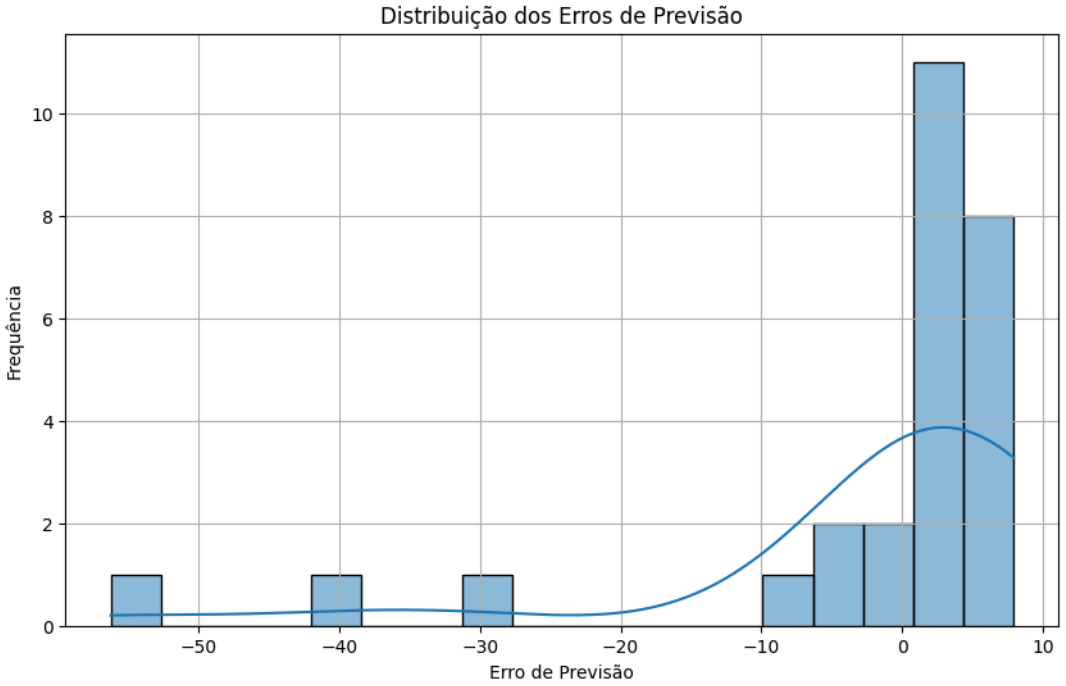


**Figura: Distribuição dos erros de previsão. Fonte: dos autores**

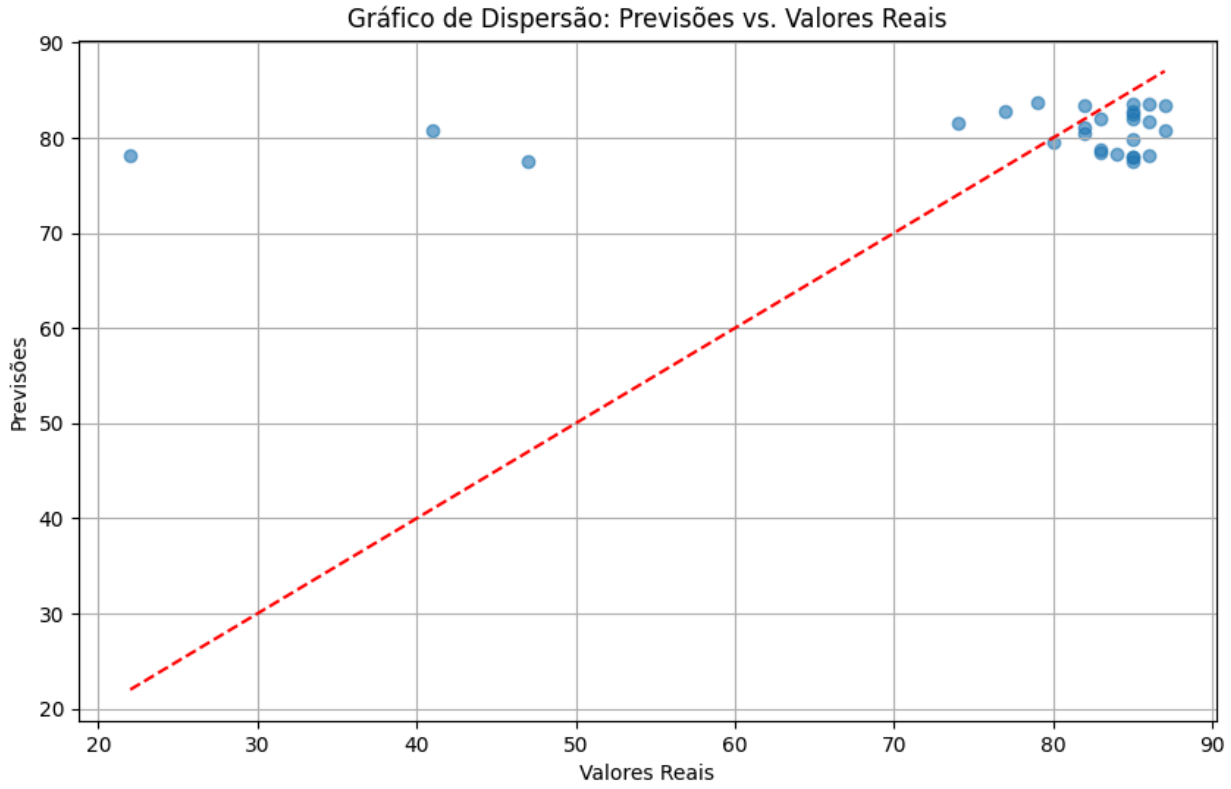


**Figura: Gráfico de dispersão do modelo KNN. Fonte dos autores**

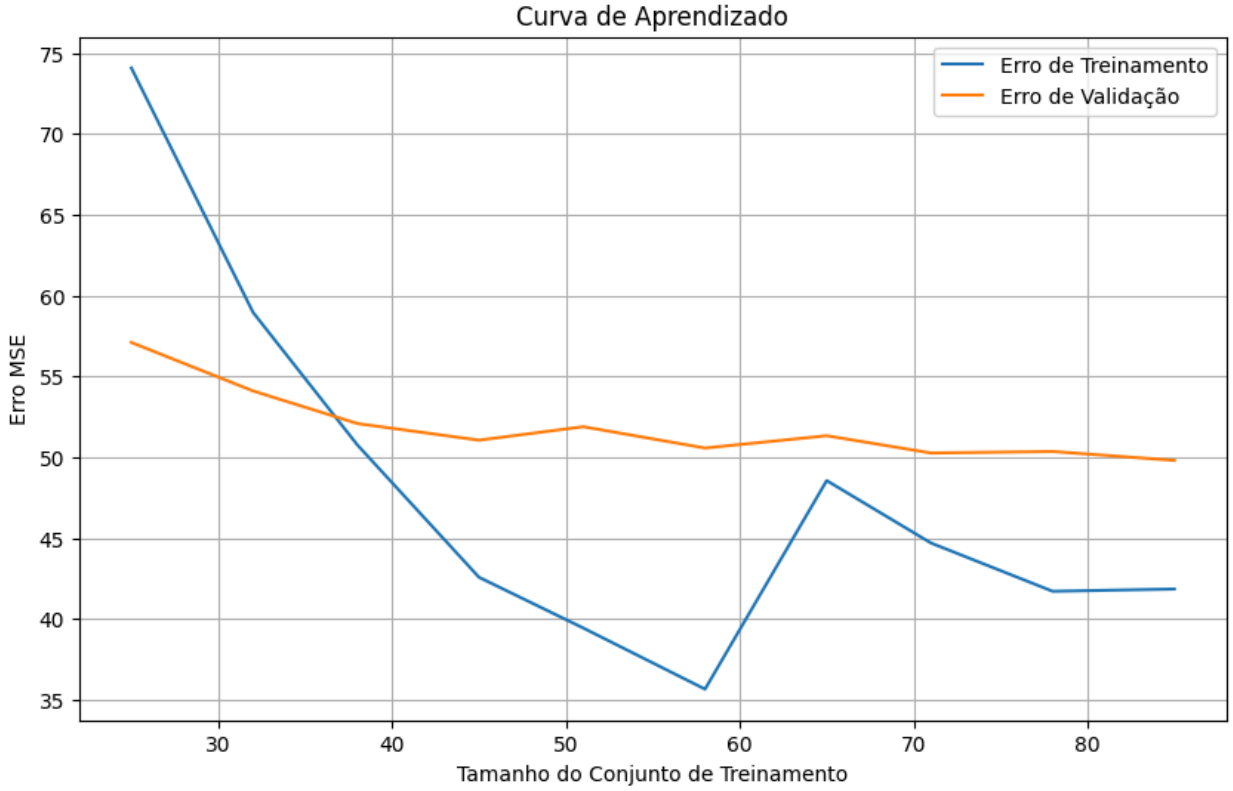


**Figura: Curva de aprendizado do modelo KNN. Fonte: Dos autores**

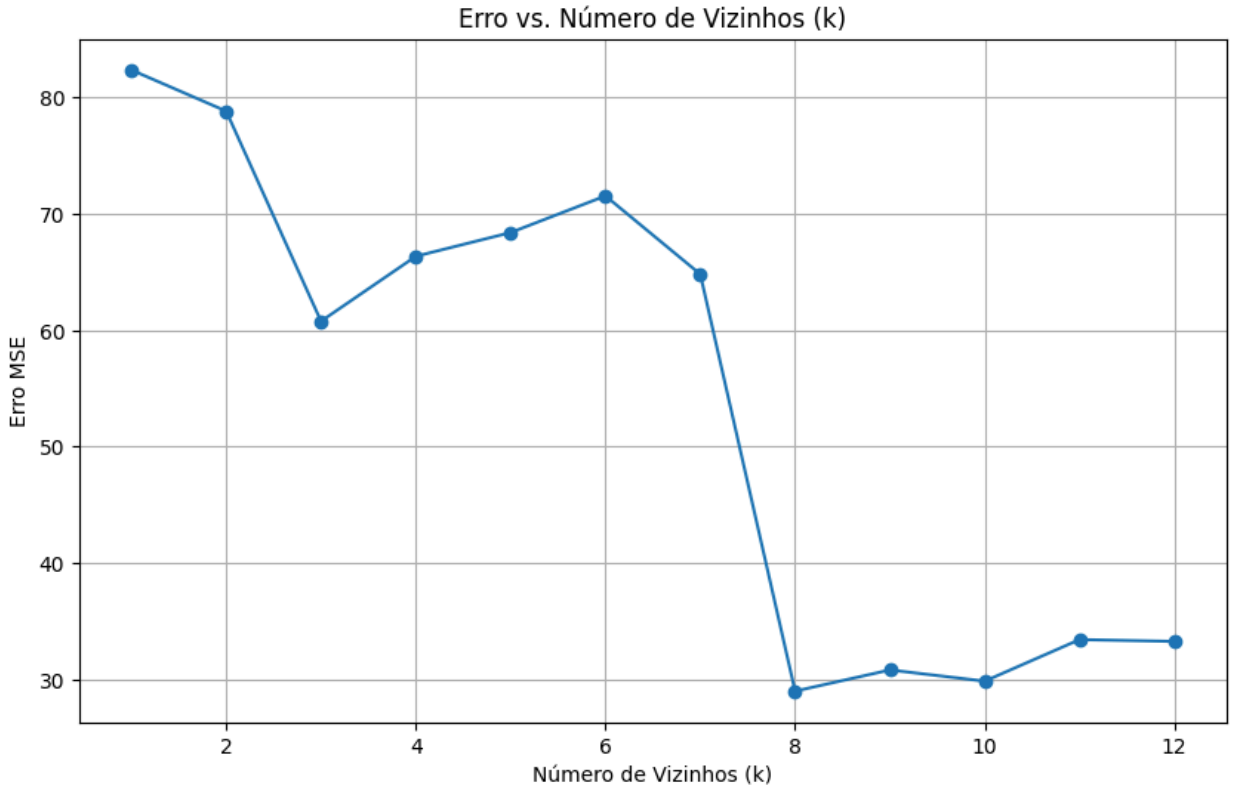


**Figura: Relação entre o erro e o número de vizinhos ao longo do tempo. Fonte: Dos autores**

## 2. Comparação com outros modelos

Embora o trabalho determinasse o uso do KNN como modelo de regressão, aqui foi reservado uma sessão para tratar sobre o desempenho de outros modelos, isso serve apenas para obter uma base geral de comparação do KNN com outros modelos como Gradient Boosting Regressor, Regressão linear e Random Forest Regressor abaixo segue as métricas de avaliação de cada modelo implementado

**Gradient Boosting Regressor**:

- **MAE**: 6.85
- **MSE**: 186.40
- **RMSE**: 13.65
- Este apresentou um desempenho superior ao k-NN em todas as métricas, o que sugere que o mesmo é mais eficiente em capturar as relações não lineares do dataset apresentado.

**Regressão Linear**:

- **MAE**: 11.34
- **MSE**: 284.72

- **RMSE**: 16.88
- Em termos de erro médio absoluto este modelo, a regressão linear, apresentou o pior desempenho entre todos, tendo em vista que o mesmo é baseado em relações lineares esse desempenho já era o esperado.

**Random Forest Regressor**:

- **MAE**: 6.50
- **MSE**: 172.30
- **RMSE**: 13.12
- Este foi o modelo com o melhor desempenho entre todos, o que demonstra que a  combinação de várias árvores de decisão é altamente eficaz para este conjunto de dados. O gráfico abaixo mostra o comparativo entre todos os modelos implementados.

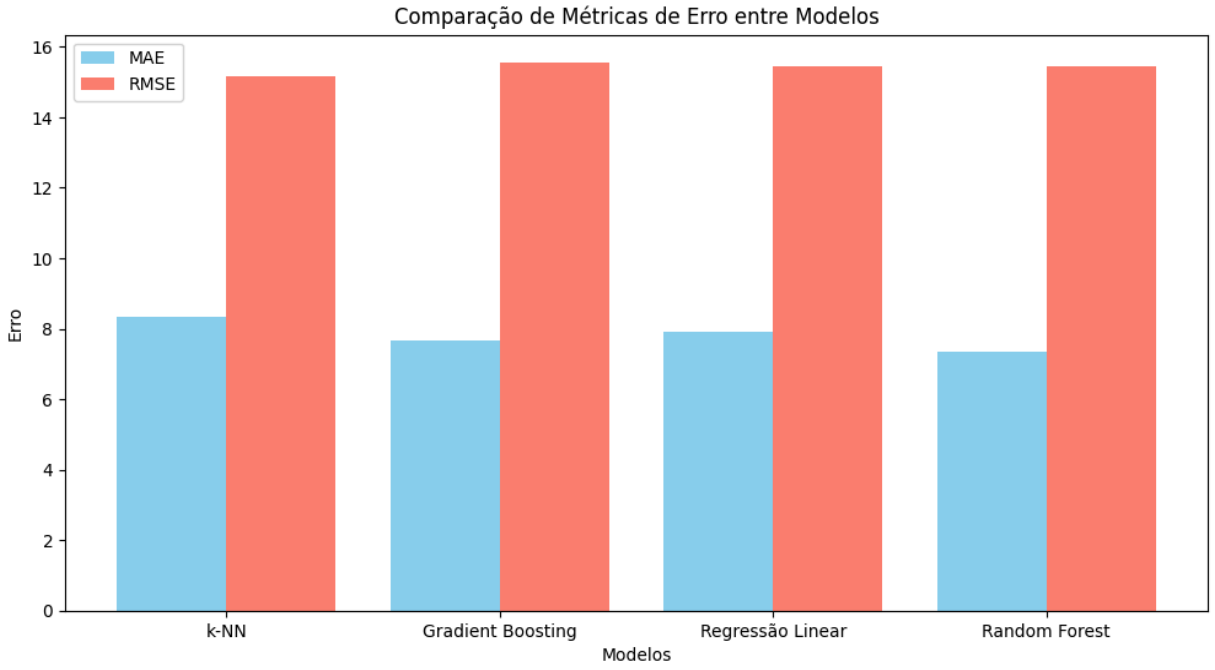


**Figura: Comparação entre os modelos. Fonte: Dos Autores**

## Distribuição das Principais Variáveis

1. Distribuição de Influenciadores por país

- Após análises detalhadas da variável *country,* a mesma revelou que os países com maior números de influenciadores estão concentrados principalmente em regiões como **América do Norte, Europa e América do Sul**, com destaque para **Estados Unidos** e **Brasil**. O mapa de calor gerado mostra bem essa distribuição

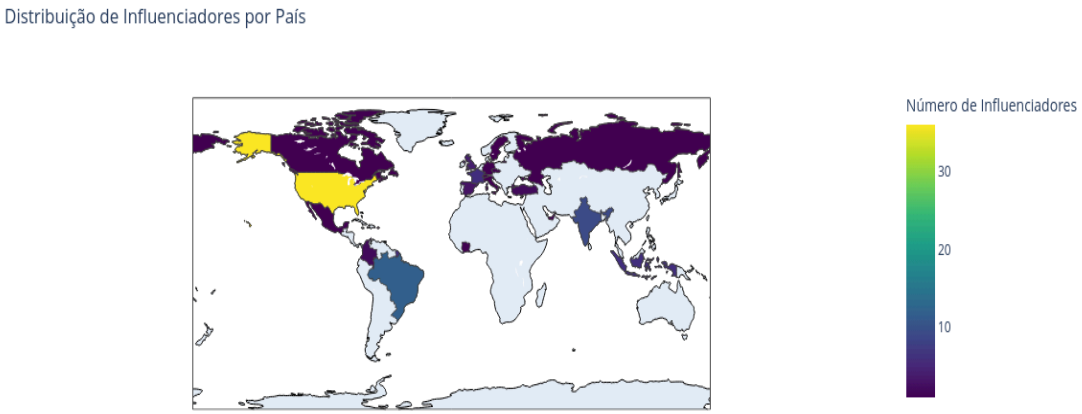


**Figura:Distribuição de influenciadores  Fonte: Dos autores**

Essa alta concentração se deve ao fato de que nessas regiões a população em geral adere mais a esse tipo de plataforma, além disso a presença de marcas ajuda a impulsionar a popularidade de influenciadores.

## Média de curtidas por postagem por país

Através desse gráfico é possível observar que o Canadá, Costa do Marfim e o México são os países que possuem as maiores médias de likes por post, enquanto países como Suécia, Itália e Indonésia têm as menores.

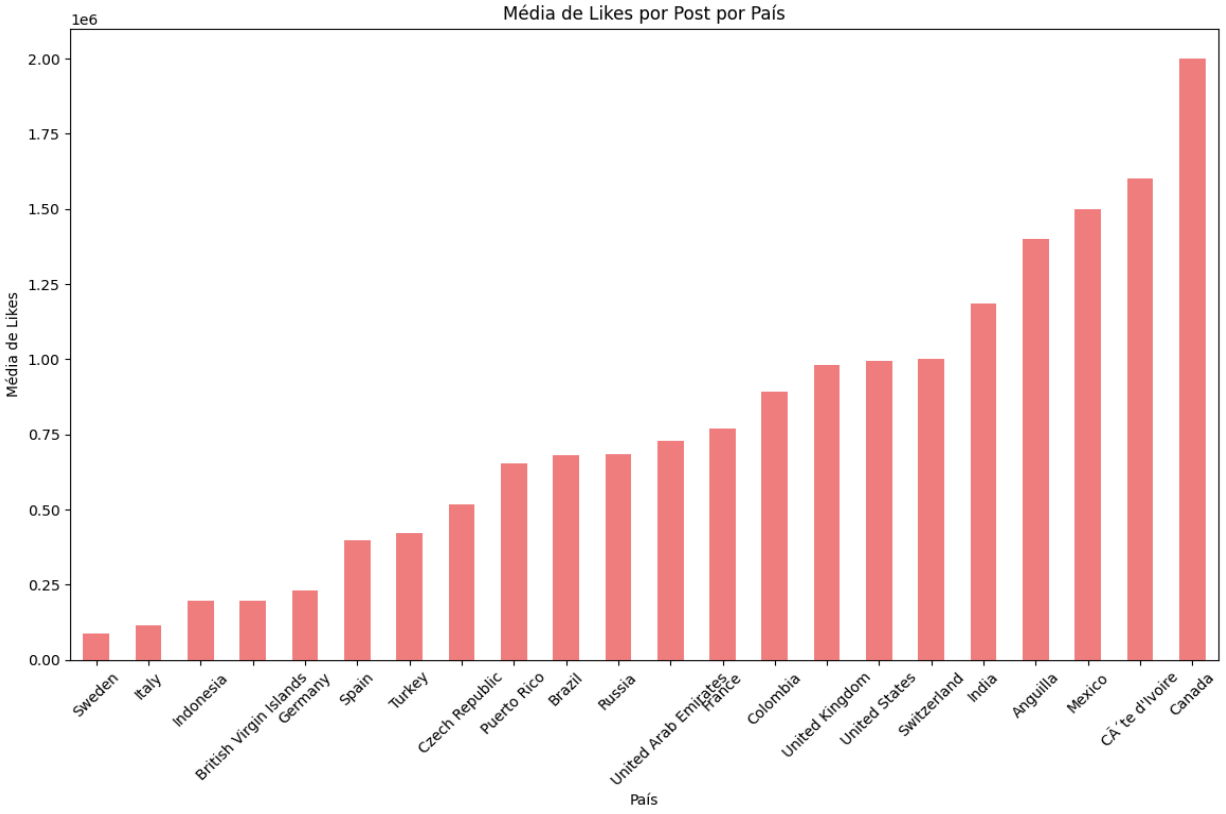


**Figura: Média de curtidas por post por país. Fonte: Dos autores**

## Taxa de Engajamento e País

Foi observado que os perfis de países da América do Norte e Europa possuem uma tendência a apresentar uma taxa de engajamento média mais alta, enquanto os influenciadores da Ásia e América Latina embora em maior número, apresentam uma maior variabilidade na taxa de engajamento.

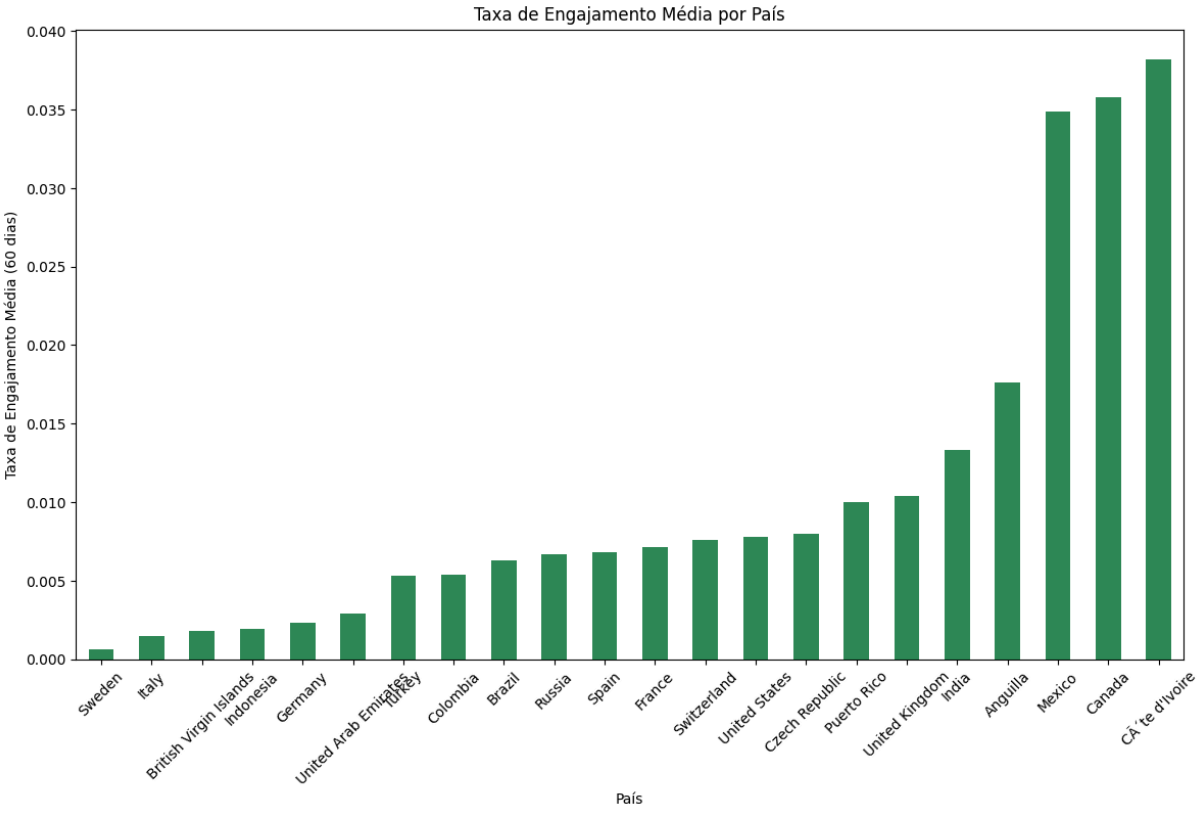


**Figura: Taxa de engajamento por país Fonte: Dos autores**

## Correlação entre seguidores e Engajamento do país

O gráfico possibilita analisar a relação entre o número de seguidores com a taxa de engajamento médio de um influenciador. Os países, demonstrados por pontos de cores distintas, apontam que países como México, Canadá e Rússia, têm pontos mais elevados em termos de taxa de engajamento, independentemente do número de seguidores.

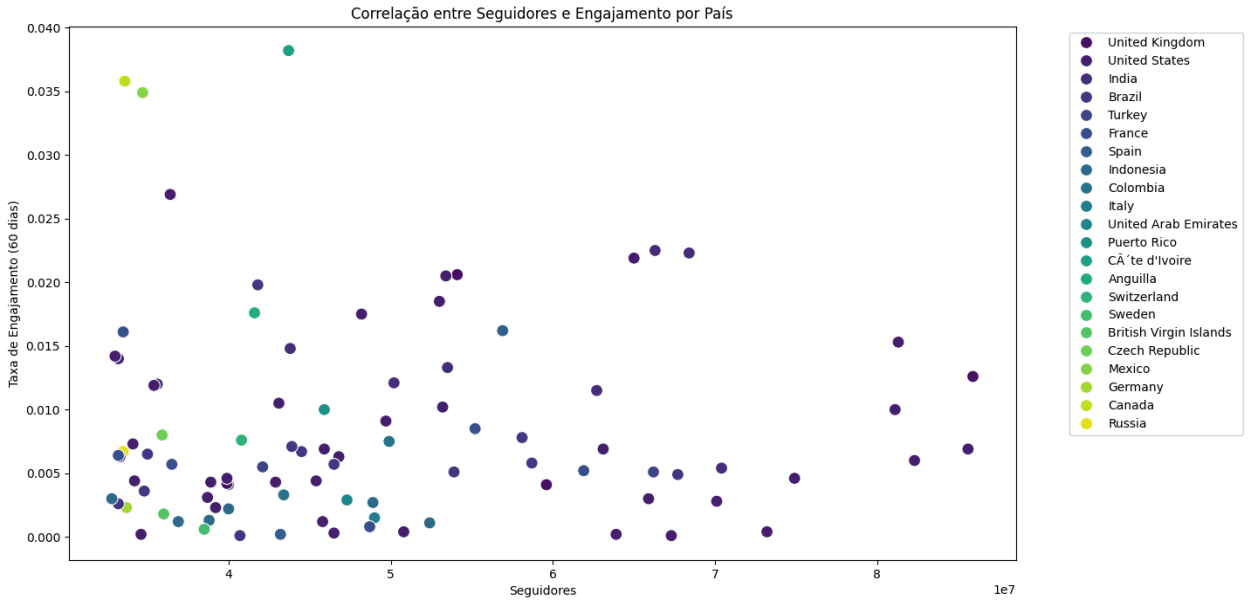


**Figura: Correlação entre seguidores e engajamento por país. Fonte: Dos autores**

## Correlação entre seguidores e Engajamento

A análise feita mostra que o número de seguidores está positivamente correlacionado com a taxa de engajamento, embora essa correlação seja mais fraca em perfis com milhões de seguidores, pois os mesmos apresentam um engajamento diluído. O gráfico abaixo mostra bem essa relação.

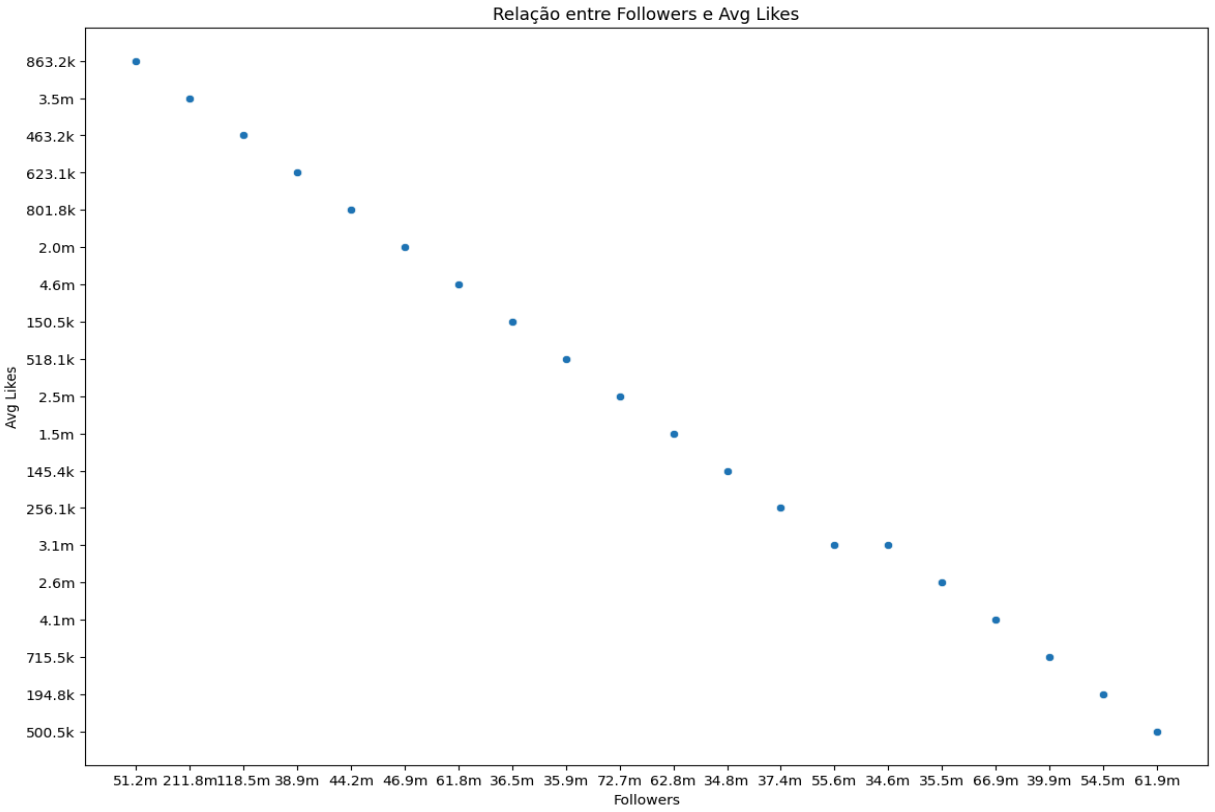


**Figura: Relação entre seguidores e taxa de *likes.* Fonte: Dos autores.**

## Distribuição de Influenciadores por país x Média de Seguidores por País

Foram estabelecidos gráficos para demonstrar onde estão localizados os principais influenciadores presentes no banco de dados, e onde estão localizados os seus seguidores. O primeiro gráfico abaixo demonstra que há uma grande concentração de influenciadores localizados nos Estados Unidos, seguido por Brasil e Índia. E o segundo gráfico aponta que os seus seguidores estão localizados principalmente no Reino Unido, Turquia e Estados Unidos.

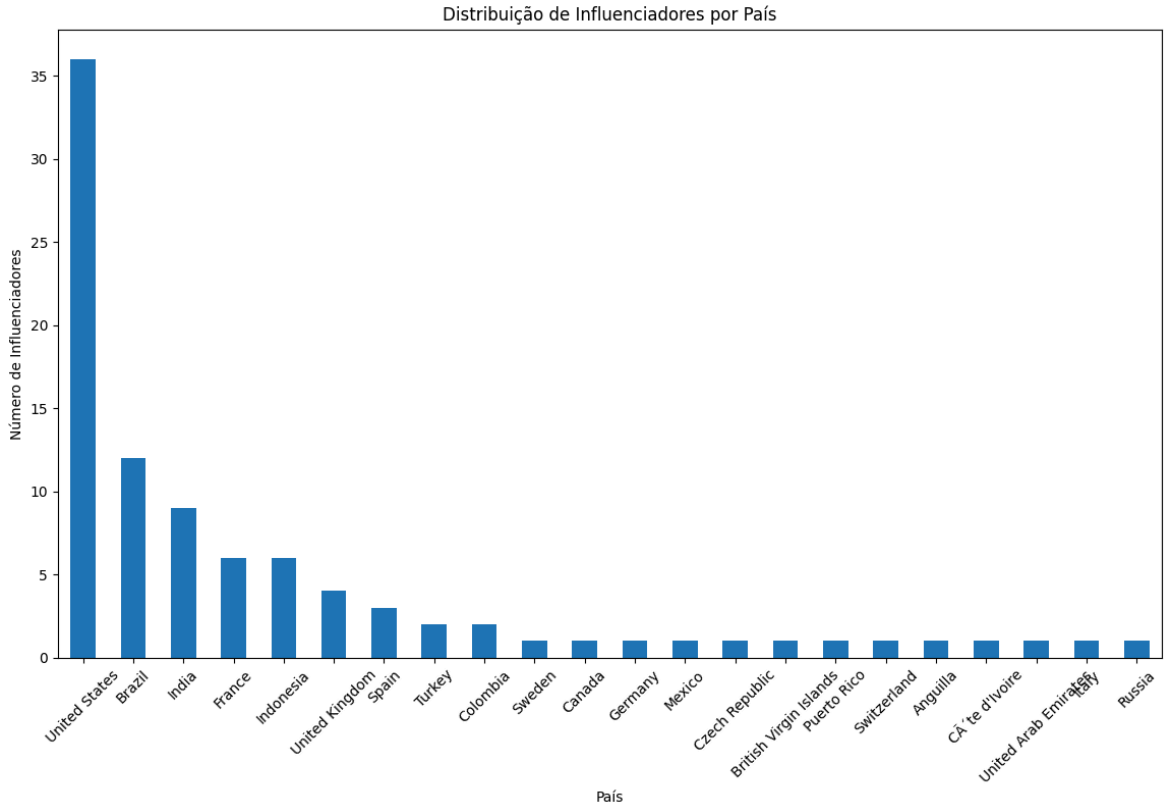


**Figura: Distribuição de influenciadores por país  Fonte: Dos autores**

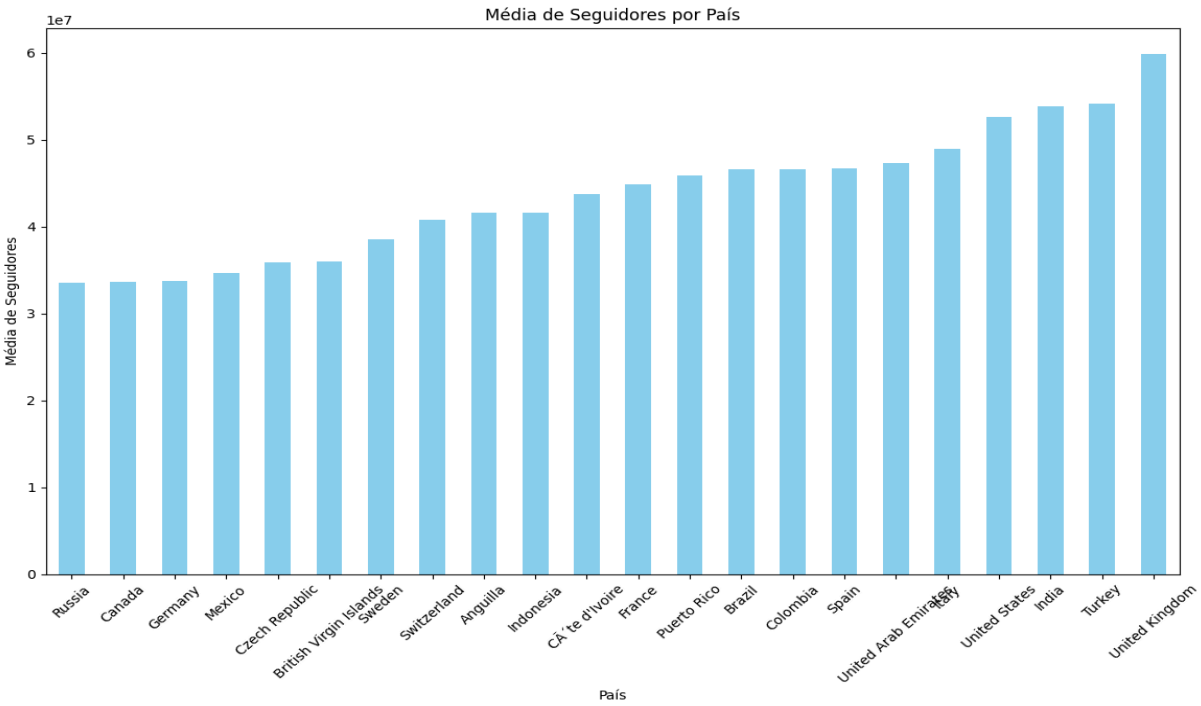


**Figura: Média de seguidores por país Fonte: Dos autores**

## Correlação entre Países e a Pontuação de Influência

Este *boxplot* apresenta a Pontuação de Influência por País, destacando diferenças significativas nas influências percebidas entre várias nações. Através dele é possível notar que países como os Estados Unidos e o Reino Unido aparecem com pontuações de influência mais concentradas em níveis mais altos, indicando uma influência mais uniforme e potencialmente mais forte. Em contraste, países como Índia e Brasil mostram uma maior dispersão nas suas pontuações, sugerindo uma variação mais ampla nas percepções de influência dentro desses países.

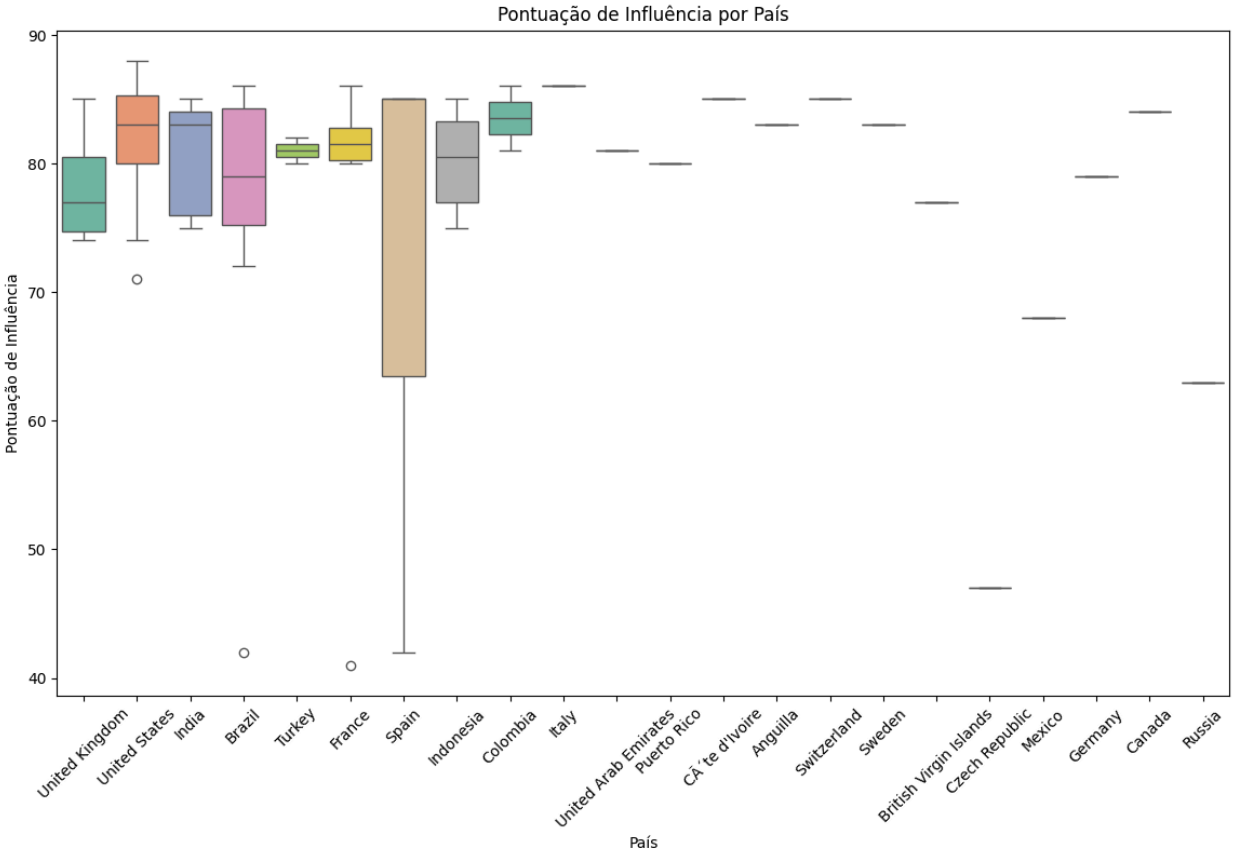


**Figura: Pontuação de influência por país.  Fonte: Dos autores**

# Discussão

Essa seção visa apresentar uma discussão crítica do trabalho que foi executado nela. É descrito em detalhes o desempenho do modelo. A análise em si trouxe informações importantes sobre o desempenho do modelo e características do conjunto de dados. Mas, mesmo com essa análise, é muito importante. Abaixo, segue as discussões sobre o desempenho e limitações encontradas durante o desenvolvimento do trabalho.

**Desempenho do modelo**

No treinamento feito, o KNN apresentou um desempenho bom, mas ainda sim inferior aos modelos baseados em árvores como o Random Forest e o Gradient Boosting. A escolha de Manhattan como métrica de distância se mostrou ser bem eficaz com as diferenças absolutas entre as variáveis. Apesar de toda essa otimização feita, o KNN sofre muito com escalabilidade em datasets maiores, devido ao cálculo da distância para cada ponto no conjunto de dados.

**Impacto das Escolhas no Desempenho do Modelo**

A primeira escolha que podemos citar é a da normalização dos dados que aqui foi feita com o *StandardScaler* e foi crucial para o desempenho do KNN no qual o mesmo pode ser afetado à escala das variáveis. Sem isso os resultados seriam muito piores. Um outro ponto a ser ressaltado é a escolha dos Hiperparâmetros neste trabalho, isso foi feito através do *GridSearchCV* onde esse processo foi automatizado o que fez com que a dupla ganhasse mais tempo. Essa biblioteca escolhe os melhores Hiperparâmetros para o contexto ao qual o modelo está inserido. O que se fosse feito de forma manual poderia demorar consideravelmente mais para encontrar os valores ideais de Hiperparâmetros. E por fim foi usada métricas como **MAE**, **MSE** e **RMSE que forneceram** uma boa avaliação do desempenho, com o RMSE penalizando mais os erros maiores.

**Limitações do trabalho**

A primeira que podemos citar é o tamanho e qualidade do dataset, onde o mesmo limitou o potencial dos modelos KNN, no qual o mesmo depende da densidade de vizinhos para realizar previsões eficazes. A presença de dados desbalanceados por países podem ter adicionado vies ao modelo KNN. Outro ponto a ser destacado é a conversão dos dos países em faixas numéricas,por um lado isso simplificou a análise, mas por outro essa abordagem apagou algumas nuances importantes que afetam o comportamento dos usuários. Podemos citar outras limitações também como falta do tipo de conteúdo produzido pelos criadores essa informação poderia ser útil para fornecer uma visão mais rica do comportamento dos usuários. Outra limitação é a

ausência de validação externa aqui o modelo foi validado apenas no conjunto de dados locais. Uma validação em um conjunto externo ou um dataset maior, poderia aumentar a confiabilidade dos resultados.

# Conclusão e Trabalhos Futuros

A conclusão deste trabalho destaca a importância de analisar e processar as variáveis do *dataset*, com o objetivo de melhorar o desempenho do modelo aplicado. Embora haja limitações, o modelo apresentou resultados promissores. Para aprimorá-lo para futuras pesquisas, algumas recomendações podem ser destacadas:

**Aumento do Dataset:**

Seria necessário expandir o número de amostras, seja coletando mais dados reais ou usando técnicas como modelos generativos (GANs) para criar amostras sintéticas realistas.

**Exploração de Modelos Avançados:**

Testar outros modelos também é um ponto de destaque. Modelos como XGBoost ou redes neurais profundas, poderiam lidar melhor com relações não lineares e variáveis categóricas do *dataset* analisado.

**Análises Regionais Detalhadas:**

Melhorar a codificação da variável country para capturar diferenças específicas entre os países, permitindo análises mais detalhadas entre os países e contribuindo consequentemente com a compreensão mais precisa de padrões regionais.

**Exploração de Variáveis Adicionais:**

Incorporar variáveis qualitativas, como a categoria do influenciador (moda, esportes, etc.), para uma análise mais contextualizada possibilitando uma análise mais rica e que contribua para identificar fatores de impacto.

**Validação Cruzada Estratificada:**

Implementar uma validação cruzada mais robusta que leve em conta a distribuição geográfica ou de seguidores para avaliar a generalização do modelo, pode garantir uma avaliação mais autêntica da capacidade de generalização do modelo para diferentes grupos.

Este trabalho demonstrou a importância de uma análise detalhada das variáveis e do impacto das escolhas metodológicas no desempenho do modelo. Apesar das limitações, os *insights* extraídos são promissores e mostram como as redes sociais e os dados de influenciadores podem ser explorados para entender padrões de engajamento e popularidade. No futuro, a ampliação do dataset e o uso de modelos mais sofisticados poderão refinar ainda mais essas análises.

# Referências

**AZANK, Felipe**. Como avaliar seu modelo de regressão**. Medium,** 2020. Disponível em: https://medium.com/turing-talks/como-avaliar-seu-modelo-de-regressão-c2c8d73dab96. Acesso em: 12 nov. 2024

**MAFFAZIOLI, Ulisses**. Machine Learning | Algoritmo kNN. **Medium**, 2023. Disponível em: https://medium.com/@ulissesmaffa/machine-learning-algoritmo-knn-26eb7b702c37. Acesso em: 12 nov. 2024