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

**ERICH BARRETO PEREIRA**

**MARIA FERNANDA CUNHA DA SILVA**

# Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

**ILHÉUS – BAHIA**

**17/11/2024**

## RESUMO

O objetivo deste projeto é implementar e avaliar o desempenho de um modelo de regressão linear para prever a taxa de engajamento de influenciadores no Instagram, com base em diversas variáveis dos dados, como número de seguidores, quantidade de curtidas e comentários. A taxa de engajamento foi escolhida como variável dependente, uma vez que é um dos principais indicadores de sucesso de um influenciador nas redes social. A metodologia do projeto envolveu três etapas principais: análise exploratória dos dados, implementação do modelo de regressão linear e otimização do modelo. Primeiramente, foi realizada uma análise para entender as variáveis e suas relações com a taxa de engajamento, tratando dados ausentes e verificando correlações. Em seguida, o modelo de regressão linear foi desenvolvido para prever a taxa de engajamento com base em variáveis como número de seguidores e curtidas. Por fim, aplicaram-se validação cruzada e ajustes de hiperparâmetros para melhorar a performance e garantir a generalização do modelo. Em relação aos resultados obtidos...



Palavras-Chave: Taxa de Engajamento. Regressão Linear. Seguidores. Instagram.

# 1 INTRODUÇÃO

O Instagram se consolidou como uma das principais plataformas de redes sociais, onde milhões de usuários compartilham conteúdos de diversos tipos, desde postagens pessoais até campanhas de marketing e publicidade. Esse cenário proporcionou o surgimento de influenciadores digitais, que possuem grande capacidade de engajamento e de impactar seus seguidores. A taxa de engajamento, que mensura o nível de interação dos seguidores com o conteúdo publicado, tornou-se um dos principais indicadores de sucesso no Instagram, sendo fundamental para empresas e marcas que buscam realizar campanhas publicitárias eficazes.

O aumento do número de influenciadores na plataforma gerou um volume significativo de dados, tornando a tarefa de entender e analisar os fatores que influenciam a taxa de engajamento mais complexa. Variáveis como número de seguidores, curtidas, comentários e outros tipos de interações estão frequentemente associados a uma maior taxa de engajamento, mas a relação entre essas variáveis nem sempre é simples e direta. Com isso, surge a necessidade de desenvolver métodos capazes de identificar padrões e prever a taxa de engajamento com base nessas informações.

O conjunto de dados utilizado para este projeto contém informações detalhadas sobre influenciadores do Instagram, com várias variáveis que podem ser usadas para prever a taxa de engajamento. Dentre essas variáveis, destacam-se o número de seguidores, que é frequentemente considerado um indicador de alcance e popularidade, e o número de curtidas e comentários, que refletem o nível de interação e o envolvimento do público com as postagens. Esses dados foram coletados de influenciadores de diferentes nichos e com diferentes níveis de popularidade, o que proporciona uma análise rica sobre o comportamento dos influenciadores em uma plataforma tão diversificada.

Este projeto tem como objetivo não apenas prever a taxa de engajamento dos influenciadores, mas também fornecer insights valiosos sobre os fatores que mais impactam o sucesso no Instagram. As técnicas aplicadas permitem uma análise mais aprofundada e fundamentada, que pode ser utilizada para otimizar estratégias de marketing e melhorar a performance de influenciadores nas redes sociais. A aplicação da regressão linear no contexto dos dados do Instagram oferece uma base sólida para

futuras pesquisas e desenvolvimento de modelos mais avançados para previsão de métricas de engajamento.

## 2 METODOLOGIA

A metodologia adotada neste projeto foi dividida em várias etapas que englobam desde a análise dos dados até a implementação e avaliação do modelo preditivo. A primeira fase envolveu a definição e preparação do problema, começando pelo acesso ao conjunto de dados, que contém informações sobre principais influenciadores do Instagram. Como proposta do projeto em armazenar o código-fonte na plataforma GitHub, a tabela de dados também foi armazenada no mesmo, a fim de manter uma única plataforma de armazenamento de dados e código, e que fosse facilmente acessado pelos usuários.

Em seguida, foi realizada uma análise exploratória dos dados, com o intuito de entender melhor a distribuição das variáveis e as relações entre elas. Essa análise incluiu a visualização de gráficos de dispersão e matrizes de correlação para identificar possíveis padrões, além de verificar a presença de valores ausentes ou outliers. A partir dessa análise, as variáveis mais relevantes para a construção do modelo foram escolhidas, e a limpeza e preparação dos dados foram feitas para garantir que os dados estivessem prontos para a modelagem. As variáveis independentes selecionadas foram, entre outras, as que possuem forte correlação com o desempenho de um influenciador nas redes sociais.

Após a preparação dos dados, implementamos o modelo de regressão linear utilizando Python e a biblioteca Scikit-learn. O modelo foi configurado para prever a taxa de engajamento com base nas variáveis selecionadas. Foram testadas diferentes combinações de variáveis independentes para ajustar os parâmetros do modelo e melhorar as previsões. A interpretação dos coeficientes obtidos foi fundamental para entender a relação entre cada variável independente e a variável dependente, indicando o impacto que cada uma delas tem sobre o engajamento.

A otimização do modelo foi um passo importante para garantir que ele fosse capaz de generalizar bem para novos dados. Para isso, foi utilizada validação cruzada, que envolveu a divisão dos dados em subconjuntos para treinamento e teste em múltiplos ciclos, assegurando a robustez do modelo. Além disso, foram aplicadas técnicas de regularização, como Lasso (L1) e Ridge (L2), que ajudam a reduzir a complexidade do modelo e evitar o *overfitting* (Baixa capacidade de generalização

para novos dados). A escolha da taxa de aprendizado e do número de iterações também foi ajustada para garantir a convergência eficiente do modelo.

A seleção de recursos também foi realizada, com base na análise de correlação entre as variáveis. Essa etapa permitiu excluir as variáveis que apresentavam baixa correlação com a variável dependente, mantendo apenas as mais significativas no modelo, o que ajudou a melhorar a precisão das previsões.

A avaliação do desempenho do modelo foi feita com base em métricas como $R^2$, Erro Quadrático Médio (MSE) e Erro Absoluto Médio (MAE), que indicam a qualidade das previsões feitas pelo modelo. Essas métricas foram calculadas em um conjunto de dados de teste para garantir que o modelo estivesse generalizando bem e não apenas ajustado aos dados de treino. Além disso, gráficos de dispersão dos resíduos e comparações entre os valores previstos e reais foram utilizados para ilustrar os resultados.

A última etapa do projeto foi realizar um relatório técnico completo, documentando todas as etapas do projeto, desde a análise exploratória até a implementação, otimização e avaliação do modelo. O relatório incluiu gráficos, tabelas e interpretações dos resultados, além de um arquivo README.md no repositório do GitHub, explicando o funcionamento do código e os principais resultados do estudo.

## 3 RESULTADOS

### a. Análise de dados

Inicialmente, foi realizado a análise da tabela e seus dados. São os seguintes dados técnicos da tabela:

- Contém 199 linhas de dados e 1 de cabeçalho.
- Possui as colunas: *rank, channel_info, influence_score, posts, followers, avg_likes, 60_day_eng_rate, new_post_avg_like, total_likes, country.*
- A coluna *rank* indica, em ordem crescente, o rank dos influenciadores de acordo com a quantidade de seguidores.
- A coluna *posts* indica a quantidade, em milhares, de postagens que tem no perfil.
- A coluna *followers* apresenta a quantidade, em milhões, de seguidores do perfil.
- A coluna *avg_likes* mostra a média, em milhares, de curtidas por postagem.
- A coluna *60_day_eng_rate* indica a taxa, em percentual, de engajamento do perfil.
- A coluna *new_post_avg_like* mostra a média, em milhares, de curtidas das novas postagens.
- A coluna *total_likes* indica a quantidade total de curtidas em bilhões e milhões.
- A coluna *country* informa o país de origem do perfil.

Na importação dos dados, foi necessário ajustar os valores de texto para valores numéricos para que pudessem ser interpretados pelo algoritmo.

Após, realizou-se a análise de correlação entre os dados e a variável dependente taxa de engajamento. A Figura 1 apresenta os gráficos de dispersão dos dados da tabela, em relação à variável dependente. É possível perceber a presença de *outliers* em todos os gráficos, características específicas de cada dado, que serão analisados de forma separada.

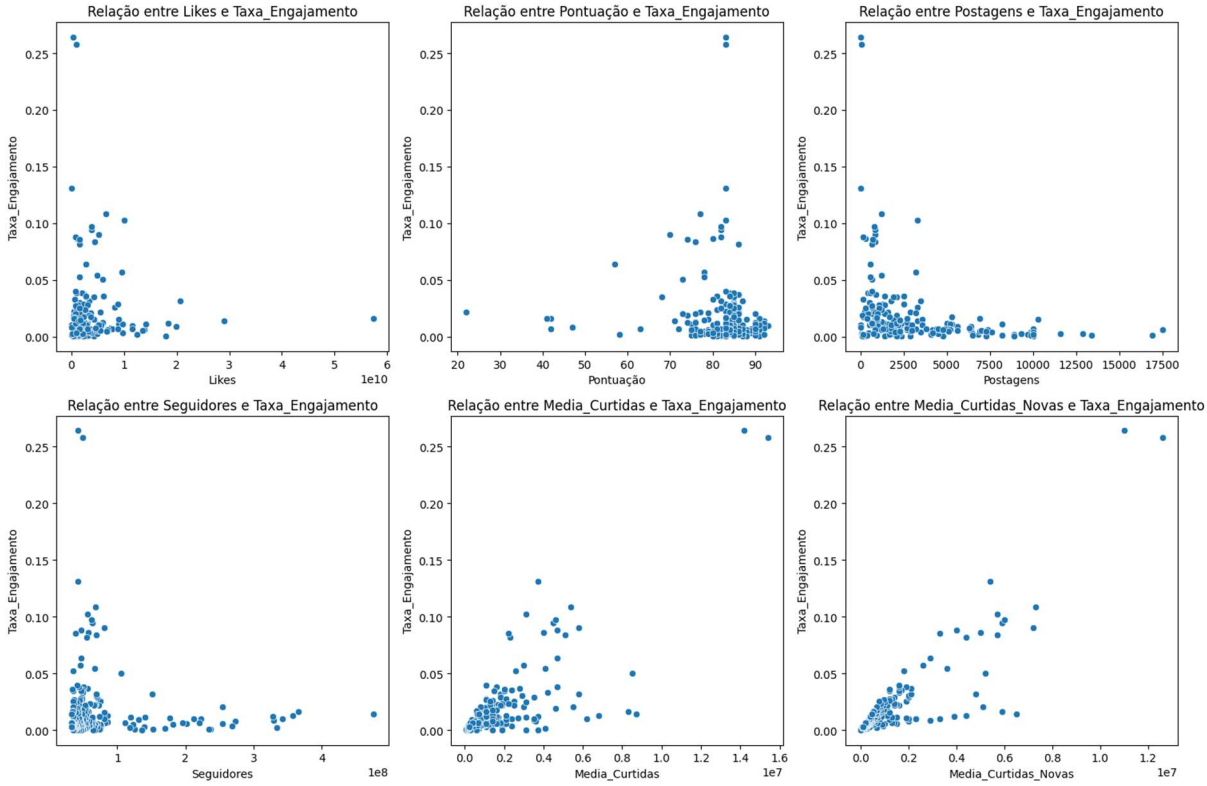


Figura 1 - Gráfico de dispersão dos dados em relação a taxa de engajamento.

- **Relação entre taxa de engajamento e curtidas:** É observável a presença de *outliers* indicando alta quantidade de curtidas, porém com taxas baixas de engajamento e perfis com alto engajamento com quantidade baixas de curtidas. Este comportamento dificulta a construção de um modelo de relação linear, sendo tendencioso para os valores extremos. De forma geral, os pontos estão concentrados em baixas quantidades de curtidas e baixas taxas de engajamento, e indicação de correlação baixa.
- **Relação entre taxa de engajamento e pontuação:** Alguns pontos possuem valores destoantes de taxa de engajamento possuindo a mesma pontuação, indicando relações não lineares na caracterização da pontuação. De forma geral, os dados estão de forma concentrada em alta quantidade de pontuação para uma taxa baixa de engajamento e baixa correlação entre pontuação e taxa de engajamento.
- **Relação entre postagens e taxa de engajamento:** Observa-se pontos de alta quantidade de postagens com baixa quantidade de taxa de engajamento e altas taxas de engajamento com baixa quantidade de postagens. De contexto geral, os pontos estão localizados em baixas quantidades de postagens e baixa quantidade de engajamento e possui baixa correlação negativa entre os dados.

- **Relação de seguidores e taxa de engajamento:** Percebe-se a dispersão dos dados entre altas quantidades de seguidores com baixa taxa de engajamento e altas taxas de engajamento com quantidade menor que seguidores. De forma geral, os dados estão localizados com baixa taxa de engajamento e baixa quantidade de seguidores, além de possuir baixa relação entre os dados.
- **Relação entre média de curtidas e taxa engajamento:** Apresenta linhas de taxa de engajamento lineares a quantidade média de curtidas, indicando a possibilidade de correlação alta e positiva. Porém, mesmo apresentando linearidade, os dados apresentam dispersão considerável e a presença de *outliers* com altos valores de engajamento e altas taxas de curtida, que tendenciarão a regressão linear aos mesmos.
- **Relação entre média de curtidas em novas postagens e a taxa de engajamento:** Apresenta-se como os dados mais lineares e com maior correlação entre os dados da tabela, a média de curtidas novas indica alta correlação positiva com a taxa de engajamento e possuindo *outliers* com alta taxa de engajamento e alta média de curtidas novas, tendenciando a regressão aos mesmos.

A Figura 2 apresenta a matriz de correlação dos dados do *dataset.*

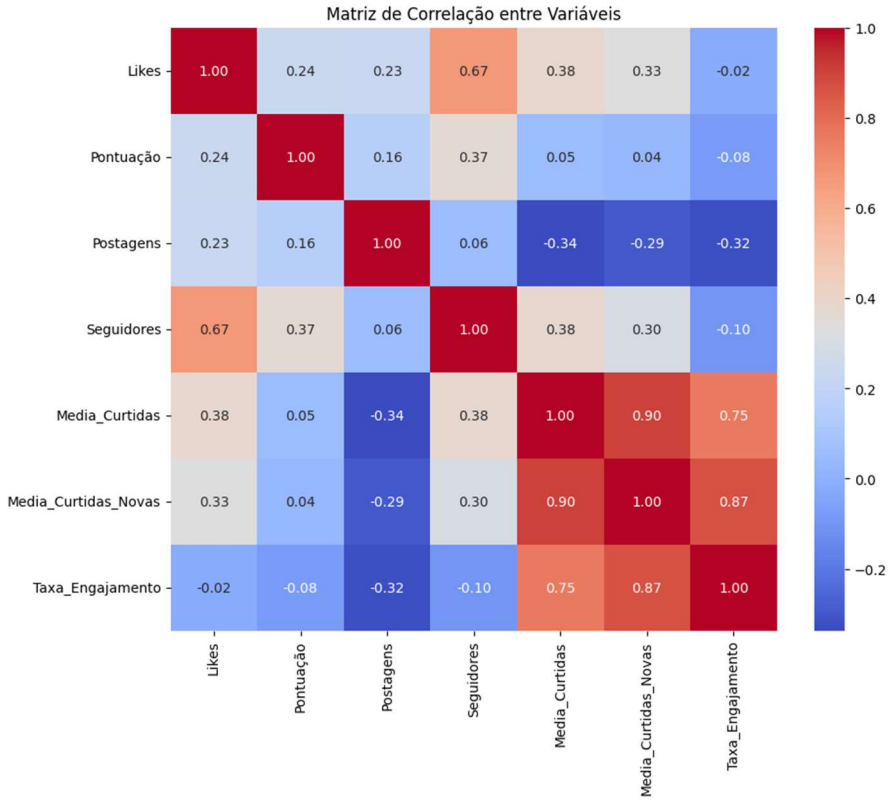


Figura 2 - Matriz de correlação dos dados.

Conforme a Figura 2, a taxa de engajamento possui alta correlação com a média de curtidas e média de curtidas novas, enquanto as outras variáveis possuem baixa correlação negativa. Um outro ponto a se destacar é a alta correlação entre a quantidade de seguidores e curtidas, indicando que a quantidade total de seguidores influencia na quantidade total de curtidas.

**b. Regressão linear entre média de curtidas e taxa de engajamento**

Utilizando a função de custo como o erro quadrático médio, conforme a Equação 1. A inserção do 2 no quociente facilita a equação do gradiente e não altera na resolução da função de custo e do gradiente.

$$MSE = \frac{1}{2m}\sum(y_{predito} - y_{real})^2 \quad (1)$$

A fim de melhorar também o desempenho do modelo, os dados foram normalizados através do método *StandardScaler* e separados os dados de treinamento e teste sendo 20% para testes e 80% para treinamento.

Utilizando a taxa de aprendizado de 0,1 e 100 interações, foi plotado o gráfico de função custo na Figura 3. É possível perceber que o valor ótimo já se aproximou do seu valor final aproximadamente na 20ª iteração e a Figura 4 apresenta o resultado da regressão linear. A Figura 4 indica a dispersão entre os dados reais e preditos, e através da regressão linear com o gradiente, obteve-se o valor de $R^2$ =0.458 e MSE = 0.000167, indicando que 45,8% dos dados preditos são reais, o que é um valor baixo para o desempenho do modelo, mesmo possuindo o erro quadrático de 0,01 %.

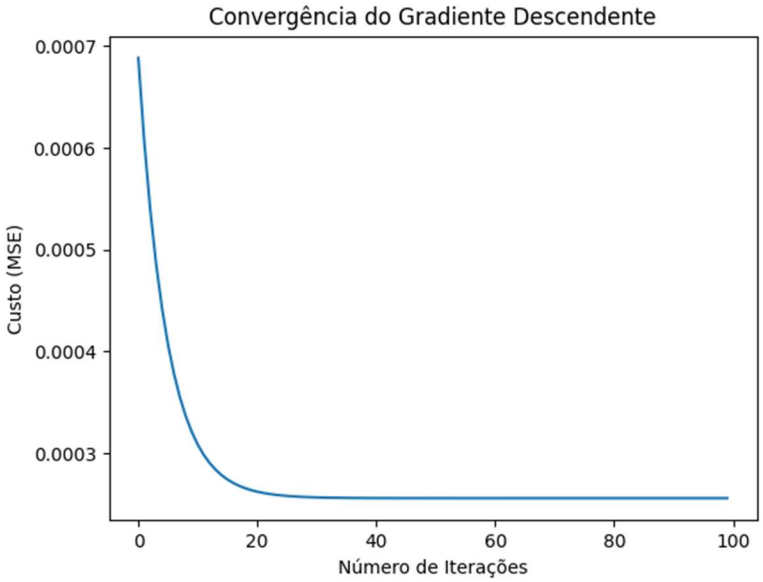


Figura 3 - Gráfico de função custo da média de curtidas e taxa de aprendizado.

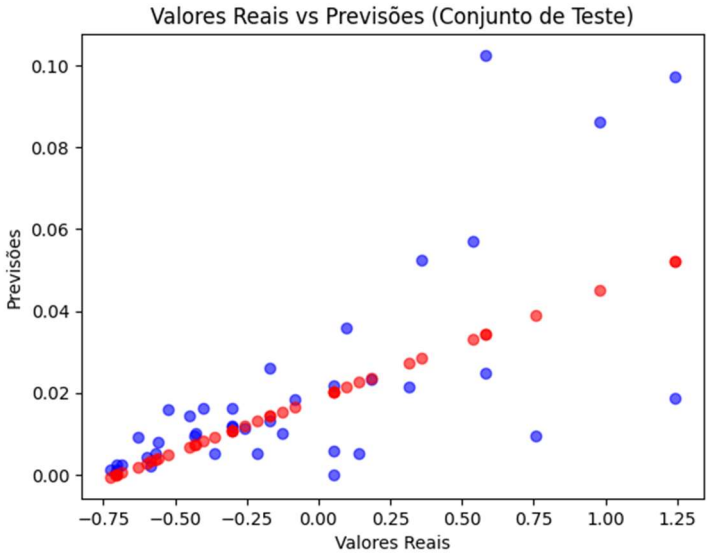


Figura 4 - Dados reais e preditos pela regressão linear com gradiente descendente entre a média de curtidas e a taxa de engajamento.

Utilizando o método dos mínimos quadrados obtemos a Figura 5, indicando os valores reais e preditos do sistema, apresentando os mesmos comportamentos da Figura 4. Para este método, obteve-se o valor de $R^2$ = 0,458 e MSE = 0,03%, sendo similar ao método do gradiente, porém com valor de MSE maior, porém baixo.

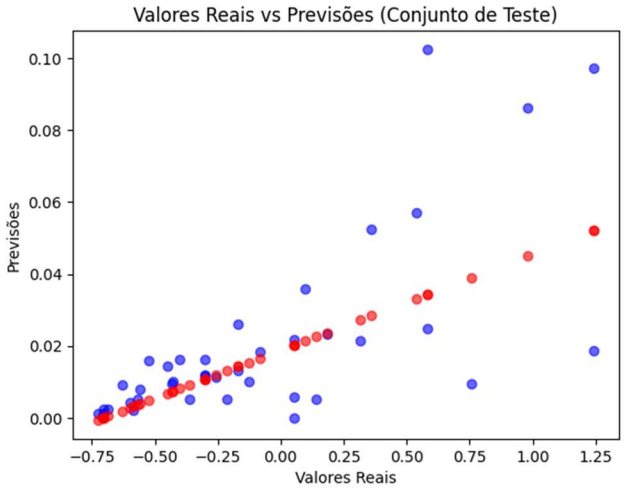


Figura 5 - Gráfico de valores reais e preditos pela regressão linear com a média de curtidas e taxa de engajamento através do método dos mínimos quadrados.

### c. Regressão linear entre média de curtidas e taxa de engajamento

Utilizando como variável independente a média de curtidas novas, foi realizado novamente utilizando os 2 métodos do projeto. Com o método do gradiente descendente, alterado o parâmetro para 0.05 de taxa de aprendizado, obteve-se a Figura 6.

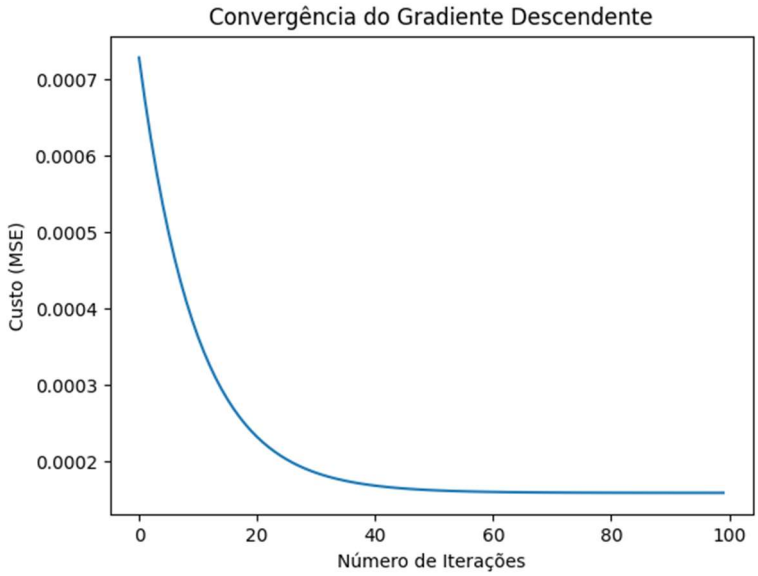


Figura 6 - Convergência da função custo no valor ótimo com a média de curtidas novas.

Como resultado, é possível perceber que o número de iterações aumentou para 40 para aproximar-se do valor mínimo da função custo, indicando maior demanda computacional para o resultado do modelo. A Figura 7 apresenta o resultado do modelo com os valores reais e de teste com os valores preditos pela regressão.

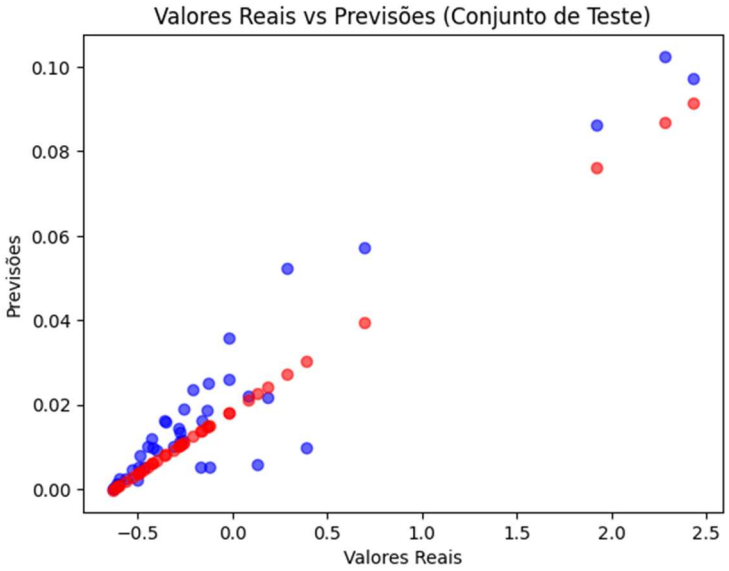


Figura 7 - Gráfico de regressão da média das curtidas novas e taxa de engajamento no método do gradiente descendente.

Conforme a Figura 7, é notável que os resultados da regressão linear se aproximam dos valores reais e o comportamento linear dos dados é mais notório. Resultado disto é o valor de $R^2 = 0.873$ e MSE = 3.919E-5, indicando baixo valor de erro quadrático e 87,3% do modelo de regressão linear atender ao valor predito.

Comparando os resultados com o método dos mínimos quadrados, a Figura 8 apresenta o gráfico dos valores reais e preditos pelo método. É notável o comportamento semelhante dos valores previstos, e os valores de $R^2 = 0,875$ e MSE = 7,723E-5 indicam a melhora do modelo com a mudança da variável independente.

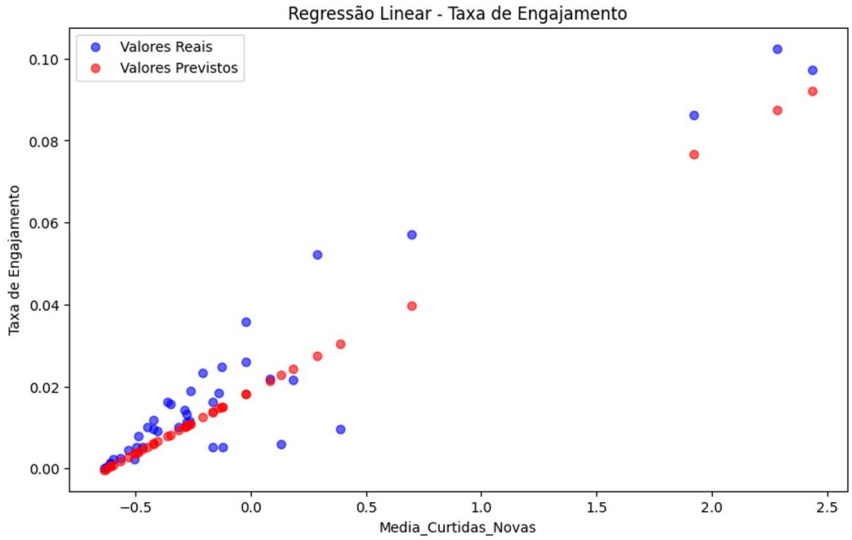

Figura 8 - Gráfico de valores reais e preditos para média de curtidas novas e a taxa de engajamento através do método dos mínimos quadrados.

**d. Regressão linear entre média de curtidas novas, seguidores e taxa de engajamento.**

Utilizando agora 2 variáveis independentes, seguidores e média de curtidas novas, foi analisado utilizando o método do gradiente descendente por conta da facilidade em alterar a taxa de aprendizado do modelo, além de novas possibilidades abordadas ao longo do projeto. Foram utilizadas estas variáveis independentes devido ao seu melhor resultado em meio a todas as variáveis. Isto pode ter sido causado devido a multicolinearidade de algumas outras variáveis e a correlação das variáveis com a variável dependente.

Usando 1000 iterações e uma taxa de aprendizado de 0,01, obteve-se a Figura 9 com a função custo do modelo. É possível perceber que aproximadamente 300 iterações, a função custo chega ao seu valor mínimo de 6,611e-5. E o resultado do modelo é apresentado na Figura 10. É observado o melhor resultado dos dados reais e preditos, sendo confirmado pelo MSE=2,975E-5 e $R^2$=0,952, sendo o melhor resultado do projeto desde então.

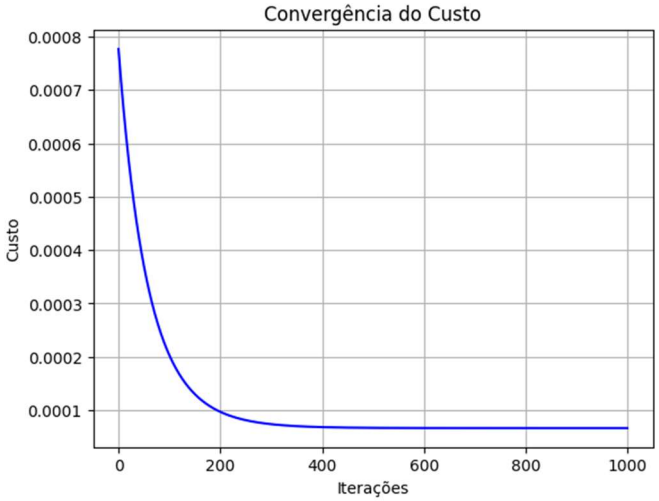


Figura 9 - Função custo do modelo com 2 variáveis independentes.

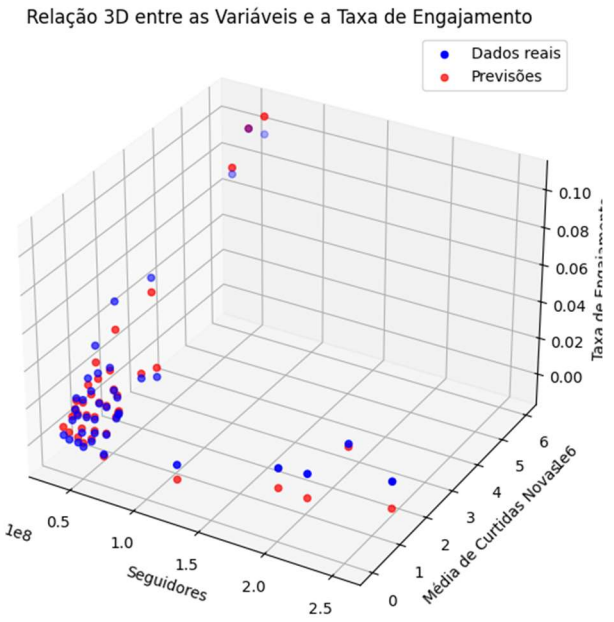


Figura 10 - Gráfico 3D de dispersão dos dados reais e preditos do modelo de 2 variáveis independentes.

**e. Regressão linear entre média de curtidas novas, seguidores e taxa de engajamento utilizando Ridge e Lasso.**

Com a proposta de melhoria do modelo, a fim de resolver problemas de *overfitting*, foi implementado ao modelo do gradiente descendente a função de Lasso e Ridge.

Para a função de Lasso, foi ajustado a função de custo para a Equação 2, sendo x o valor da variável independente e $\theta$ o valor do coeficiente encontrado para predição do modelo.

$$MSE = \frac{1}{2m}\sum(x * \theta - y_{real})^2 + \frac{\alpha}{m}\sum|\theta| \quad (2)$$

Já a função de Ridge, foi ajustado a função de custo para a Equação 3, sendo x o valor da variável independente e $\theta$ o valor do coeficiente encontrado para predição do modelo.

$$MSE = \frac{1}{2m}\sum(x * \theta - y_{real})^2 + \frac{\alpha}{2m}\sum\theta^2 \quad (2)$$

Com a inserção destes termos, espera-se que os coeficientes sofram penalizações ao introduzir os termos do absoluto ou quadrado dos coeficientes, reduzindo assim a variância e o *bias* do modelo com somente a regressão.

Assim, utilizando 0,01 para taxa de aprendizado, 1000 iterações e regularização em 0,1, obteve-se as respostas das funções custos para os modelos de Ridge e Lasso, respectivamente, na Figura 11. É observável que os valores de custos mínimos se aproximam a partir de 300 iterações do modelo, sendo o custo do Ridge menor que o Lasso.

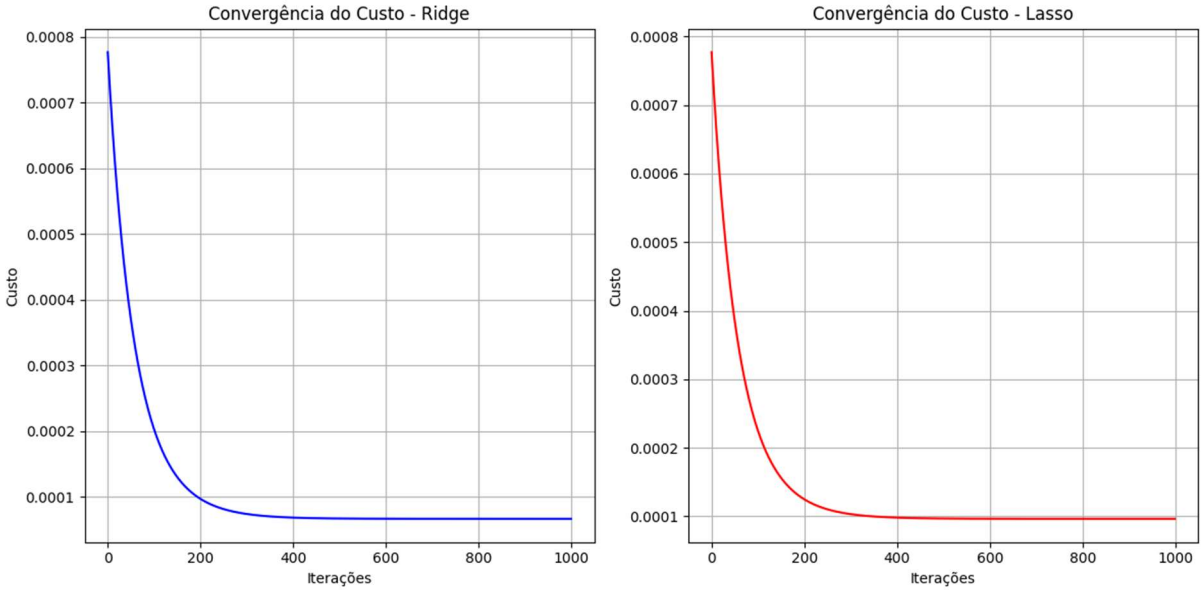


Figura 11 - Funções custo para os modelos de Ridge e Lasso.

Como resultado, obteve-se a Figura 12, com os gráficos dos modelos de Ridge e Lasso e valores reais. É possível analisar que os valores preditos estão mais próximos que a Figura 10, sendo em algumas regiões o destaque para o modelo de Lasso e outros para o Ridge. No contexto geral, o modelo de Ridge obteve $R^2 = 0,952$ e MSE=0,003%; e Lasso obteve $R^2$=0,957 e MSE=0,0027%, resultados melhores que o modelo com somente gradiente descendente.

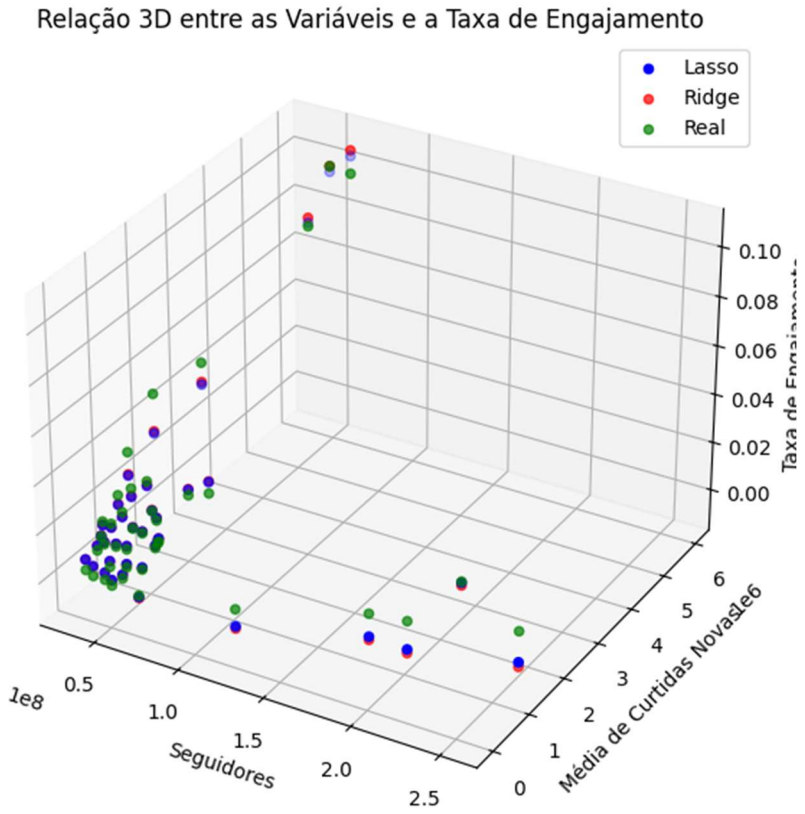


Figura 12 - Gráficos dos resultados da regressão linear com os métodos de gradiente descendente, Ridge e Lasso.

## 4 DISCUSSÃO

O presente projeto apresentou resultados satisfatórios mediante a quantidade de dados e a análise realizada. Ao analisar a matriz de correlação, é possível perceber o destaque das variáveis de média curtidas e média curtidas novas, que tiveram correlações positivas e próximas a 1. Por isso, foi realizado a regressão linear com cada variável, a fim de avaliar a servilidade dos dados. Foi constatado que a média curtidas apresentou baixo desempenho no $R^2$, mesmo possuindo MSE baixo. Já com a média curtidas novas apresentou valores de $R^2$ e MSE bons para o modelo e que poderia ser utilizado para uma regressão linear de uma variável.

A fim de obter desempenho melhor do modelo, recorreu-se para 2 variáveis independentes, e observando a matriz de correlação, percebeu-se que média curtidas e média curtidas novas tinham correlação alta também, o que atrapalharia na regressão linear e na indicação dos hiper-parâmetros. Por isso, escolheu-se as variáveis seguidores e média curtidas novas, em que obteve grande melhora nos valores de $R^2$ e MSE.

Também foi realizado a inserção dos modelos de Lasso e Ridge para validação e melhoria dos hiper-parâmetros da regressão linear, incorporando-os ao gradiente descendente e analisando as melhorias no modelo. Ainda que pouco foi o aumento do $R^2$, a redução do MSE foi significativa, retirando ainda mais o erro do modelo em novas predições.

## 5 CONCLUSÃO

Neste projeto, foi realizado a análise de dados da tabela com os principais influenciadores do Instagram, obtendo assim a visão das métricas utilizadas para a predição da taxa de engajamento dos perfis da rede social. Foram analisados os dados como quantidade de seguidores, curtidas, média de curtidas e média de curtidas novas, assim como a quantidade de posts e os países de origem.

Foi aprendido e utilizado para a análise de correlação a matriz de correlação e com auxílio dos gráficos de dispersão, encontrar as variáveis independentes que seriam mais apropriadas para uma regressão linear da taxa de engajamento da rede social. Assim, através dos métodos dos mínimos quadrados e gradiente descendente, foi realizado a predição da taxa de engajamento através dos parâmetros de média curtidas e média curtidas novas.

A fim da melhoria do modelo, foi analisado também a utilização de mais de uma variável para construção do modelo. Através de testes e análises da matriz de correlação, obteve-se as variáveis independentes seguidores e média de curtidas novas, e por meio do gradiente descendente, garantiu-se uma melhora nos resultados do modelo.

Ao final, o modelo foi reanalisado inserido as regulações de Ridge e Lasso para melhoria do modelo, a fim de equilibrar a variância e o bias, e assim, evitar o *overfitting* e *underfitting.* Foi possível obter a melhora no sistema com a redução do erro quadrático e $R^2$ em 95,6%, indicando que o sistema acertaria em 95% das predições realizadas.

## REFERÊNCIAS


TAFESSE, Wondwesen; WOOD, Bronwyn P. Followers' engagement with instagram influencers: The role of influencers' content and engagement strategy. Journal of retailing and consumer services, v. 58, p. 102303, 2021.

NANDAGIRI, Vaibhavi; PHILIP, Leena. Impact of influencers from Instagram and YouTube on their followers. **International Journal of Multidisciplinary Research and Modern Education**, v. 4, n. 1, p. 61-65, 2018.

LIMA, Ana Margarida Oliveira. **O impacto dos influenciadores digitais no instagram na decisão de compra do consumidor português**. 2022. Tese de Doutorado.

FLORIANO, Mikaela Prestes; SILVA, Andressa Hennig; KLUSENER, Monique Vigil. Influência das motivações de uso do Instagram na formação do Consumo de Status: Uma análise com consumidores brasileiros. **Organizações em Contexto**, v. 19, n. 38, p. 315-340, 2023.

SILVA, Adrielly Souza; DA COSTA, Marconi Freitas. As aparências (não) enganam: compra de serviços hoteleiros endossados por influenciadores digitais do Instagram. **ReMark-Revista Brasileira de Marketing**, v. 20, n. 1, p. 52-77, 2021.