Exploratory Analysis of the Relationship Between Short-Form Video Consumption and Attention Capacity Using Machine Learning Techniques

Nícolas G. Harnisch[1], **Samuel S. Beserra**[1],
José Levi Gonçalves de Lima Menezes[1], **Pablo Luiz Braga Soares**[1]

[1] Curso de Ciência da Computação
Universidade Federal do Ceará (UFC) – Campus Russas
Rua Felipe Santiago, 411 – Cidade Universitária
62.900-000 – Russas – CE – Brasil

{nicolasgomeshar, samuelsilvabeserra}@alu.ufc.br

levilima2002@gmail.com, pablo.soares@ufc.br

Abstract. *In recent years, short-form video platforms, such as TikTok, Instagram Reels, and YouTube Shorts, have become central elements of digital consumption. Inserted within the context of the Attention Economy, these platforms use recommendation algorithms and variable reward mechanisms to increase engagement and screen time. This work conducts an exploratory computational investigation using a public dataset with approximately 12,000 records. We combine unsupervised methods (K-Means in multidimensional and reduced PCA spaces, GMM, and Agglomerative Clustering) to discover latent behavioral profiles without target bias, and supervised methods (Random Forest and alternative linear/non-linear models) to test whether digital consumption variables possess predictive power over an attention score metric. The results indicated no relevant correlation between short-form video consumption time and attention score ($r \approx 0.0008$). Moreover, supervised models achieved performance close to or below baseline models, demonstrating that hyperparameter optimization merely fitted stochastic noise. Furthermore, clustering evaluations in both original and reduced spaces revealed no statistically significant differences in attention scores across groups ($p > 0.05$). These findings provide strong evidence compatible with synthetic noise in the analyzed dataset, underscoring the indispensable role of baseline benchmarking and robust statistical inference to prevent false discoveries in behavioral computing.*

Resumo. *Nos últimos anos, plataformas digitais baseadas em vídeos curtos, como TikTok, Instagram Reels e YouTube Shorts, passaram a ocupar uma parcela significativa do consumo digital dos usuários. Inseridas no contexto da Economia da Atenção, essas plataformas utilizam algoritmos de recomendação e mecanismos de recompensa variável para ampliar o engajamento e o tempo de permanência. Este trabalho realiza uma investigação computacional exploratória utilizando uma base pública de aproximadamente 12.000 registros. A integração metodológica de técnicas não supervisionadas (K-Means no espaço original e reduzido via PCA, GMM e Clusterização Aglomerativa)*

e supervisionadas (Random Forest e modelos alternativos) teve como objetivo investigar, de forma complementar, tanto a identificação de agrupamentos comportamentais latentes sem viés da variável-alvo quanto a capacidade preditiva real dos atributos sobre o score de atenção. Os resultados indicaram ausência de correlação linear ou monotônica ($r \approx 0{,}0008$). Os modelos supervisionados apresentaram desempenho equivalente aos baselines ingênuos, evidenciando sobreajuste a ruído estocástico na classificação, enquanto os clusters formados não apresentaram diferenças estatisticamente significativas quanto à atenção ($p > 0{,}05$). Os achados são fortemente compatíveis com uma geração sintética de variáveis independentes na base analisada, ressaltando o papel fundamental do rigor metodológico e de modelos baseline para evitar falsas inferências em dados comportamentais.

1. Introdução

Nas últimas décadas, a popularização dos dispositivos móveis e das plataformas digitais transformou a atenção humana em um dos recursos mais disputados do ambiente digital. Esse cenário, conhecido como Economia da Atenção, caracteriza-se pela competição entre plataformas pelo tempo de tela e pelo engajamento dos usuários, em um contexto em que a informação é abundante, mas a atenção é limitada [Noronha 2024].

Nesse contexto, plataformas como TikTok, Instagram Reels e YouTube Shorts popularizaram o consumo de vídeos curtos, sustentado por algoritmos de recomendação e mecanismos de engajamento que incentivam interações frequentes e prolongam o tempo de permanência dos usuários. A literatura recente discute que a exposição prolongada a esse tipo de conteúdo pode estar associada à busca por recompensas imediatas e a dificuldades na manutenção da atenção em tarefas cognitivamente mais exigentes [da Silva Pereira et al. 2025, Rodrigues and Walderrama 2025].

Apesar do crescente interesse pelo tema, a literatura ainda apresenta poucas investigações que utilizem técnicas de Aprendizado de Máquina para avaliar empiricamente se variáveis comportamentais relacionadas ao consumo de vídeos curtos apresentam associação estatística ou capacidade preditiva em relação a indicadores de atenção.

Diante dessa lacuna, este trabalho investiga a relação entre o consumo de vídeos curtos e indicadores de capacidade atencional por meio de uma abordagem computacional multifacetada. A combinação articulada de métodos não supervisionados e supervisionados tem como objetivo específico: (i) investigar se variáveis de hábitos digitais formam perfis comportamentais latentes naturais sem indução de viés pela variável-alvo e (ii) testar rigorosamente se essas variáveis possuem capacidade preditiva sobre o score de atenção quando comparadas a modelos de referência (*baselines*). O estudo não busca estabelecer causalidade neurobiológica direta, mas verificar a consistência estatística e preditiva dos dados disponíveis.

2. Trabalhos Relacionados

Estudos recentes têm investigado a relação entre uso intensivo de tecnologias digitais, redes sociais e possíveis impactos sobre processos cognitivos. No contexto da Economia da Atenção, discute-se como plataformas digitais disputam continuamente o tempo e o engajamento dos usuários, transformando a atenção em um recurso central para

modelos de negócio baseados em retenção e recomendação de conteúdo [Noronha 2024, Bartijotto 2025].

No campo neurobiológico e educacional, pesquisas recentes examinam a relação entre dependência digital, sistema dopaminérgico e mecanismos de recompensa variável [Rodrigues and Walderrama 2025], bem como os possíveis efeitos de conteúdos em formato curto sobre processos de aprendizagem, atenção e concentração [da Silva Pereira et al. 2025]. Essas investigações sugerem que o consumo de vídeos curtos pode se relacionar com dinâmicas atencionais, embora a natureza dessa associação permaneça em aberto na literatura empírica.

Em nível populacional, estudos epidemiológicos e meta-análises apontam taxas expressivas de uso problemático de tecnologias, com prevalências globais estimadas em 26,99% para dependência de smartphones e 17,42% para dependência de redes sociais [Meng et al. 2022]. Paralelamente, pesquisas em computação comportamental e psicologia social demonstram que traços de personalidade associam-se a diferentes padrões de navegação e engajamento em plataformas digitais [Marengo et al. 2020]. Esses achados reforçam o interesse no desenvolvimento de abordagens computacionais exploratórias capazes de avaliar quantitativamente a relação entre padrões de comportamento digital e métricas de desempenho cognitivo.

Apesar dessas contribuições, a literatura consultada mostrou predominância de discussões teóricas, revisões bibliográficas e análises qualitativas sobre o impacto das plataformas digitais na atenção. Foram encontrados menos estudos voltados à aplicação direta de métodos computacionais integrados, como clusterização multivariada, modelagem supervisionada contra baselines estocásticos e testes de hipóteses não paramétricos com controle de FDR.

A contribuição deste trabalho está em avaliar empiricamente se variáveis comportamentais simples disponíveis em uma base pública com 12.000 registros apresentam sinal estatístico ou preditivo suficiente para explicar o score de atenção. O estudo combina análise de correlação linear e monotônica, clusterização sem utilização da variável-alvo, testes estatísticos paramétricos e não paramétricos, modelos supervisionados comparados com baselines, ajuste controlado de hiperparâmetros e análises complementares com múltiplos algoritmos. Assim, o principal resultado científico e metodológico é demonstrar que, na base analisada, essas variáveis não apresentaram capacidade preditiva para estimar indicadores de atenção, fornecendo evidências sólidas sobre os cuidados necessários com a validação de dados em Ciência da Computação.

3. Fundamentação Teórica

3.1. Economia da Atenção

Na sociedade contemporânea, a abundância de informação torna a atenção humana um recurso limitado e disputado. Nesse contexto, discute-se a transição para uma Economia da Atenção, em que plataformas digitais competem pela captura e monetização do tempo e do interesse dos usuários [Citton 2013, Lash 2012, Noronha 2024]. Para isso, utilizam estratégias de recomendação, personalização e engajamento capazes de prolongar a permanência nas plataformas [Schreiner et al. 2021, Bartijotto 2025].

Mecanismos de recompensa variável podem favorecer interações frequentes e

prolongadas em ambientes digitais, sendo discutidos em estudos sobre dependência digital e comportamento de uso [Rodrigues and Walderrama 2025]. Neste trabalho, esses conceitos fornecem a fundamentação teórica de domínio para interpretar o contexto das plataformas de vídeos curtos.

3.2. Aprendizado de Máquina na Análise Comportamental

Técnicas de Aprendizado de Máquina têm sido amplamente utilizadas na análise de dados comportamentais devido à sua capacidade de identificar padrões e relações em grandes volumes de dados. Diferentemente de abordagens estatísticas tradicionais exclusivamente univariadas, esses métodos permitem analisar múltiplas variáveis simultaneamente e detectar relações não lineares complexas.

Dessa forma, a integração de métodos supervisionados e não supervisionados constitui uma abordagem robusta para avaliar tanto a estrutura natural dos dados quanto o potencial preditivo de variáveis comportamentais sobre métricas cognitivas.

3.3. Métodos de Agrupamento e Redução de Dimensionalidade

O agrupamento (*clustering*) é uma abordagem não supervisionada que busca particionar dados em conjuntos homogêneos com base em medidas de similaridade ou distância. O algoritmo *K-Means* busca minimizar a variância intra-cluster (soma dos quadrados dos erros) em relação aos centróides de k grupos. O *Gaussian Mixture Model* (GMM) assume que os dados são gerados a partir de uma combinação probabilística de distribuições normais multivariadas com matrizes de covariância próprias. A Clusterização Aglomerativa (*Agglomerative Clustering*), por sua vez, constrói iterativamente agrupamentos hierárquicos a partir da fusão sucessiva de pares de instâncias com base em critérios de ligação (*linkage*).

Para avaliar e validar a qualidade estrutural dos clusters formados, utiliza-se o Coeficiente de Silhueta, que quantifica a coesão interna em relação à separação entre grupos vizinhos. A Análise de Componentes Principais (PCA) é empregada como técnica linear de redução de dimensionalidade, projetando o espaço multidimensional original em eixos ortogonais que preservam a máxima variância dos dados [Sousa 2019].

3.4. Modelos Supervisionados e Validação por Baselines

O *Random Forest* é um método de aprendizado supervisionado baseado em conjuntos de árvores de decisão construídas com amostras bootstrap e seleção aleatória de atributos (*feature bagging*). Essa técnica reduz a variância individual das árvores e previne sobreajuste em cenários com interações complexas entre preditores numéricos e categóricos.

Em problemas de aprendizado supervisionado, a comparação com modelos de referência (*baselines* ingênuos), como o *Dummy Regressor* (que prediz a média ou mediana) e o *Dummy Classifier* (que prediz de acordo com a distribuição estratificada das classes), constitui um princípio metodológico essencial. Essa comparação permite verificar se o modelo complexo aprendeu um sinal preditivo real nos dados ou se está apenas operando ao nível do acaso estatístico.

4. Metodologia

A metodologia foi estruturada em cinco etapas: pré-processamento dos dados, análise exploratória, clusterização multivariada, testes estatísticos de hipóteses e modelagem supervisionada.

4.1. Base de Dados e Variáveis do Estudo

A base de dados utilizada neste trabalho foi obtida na plataforma Kaggle, a partir do conjunto de dados *Reels/Shorts Consumption vs Attention Span* [Joshi 2025]. A base possui aproximadamente 12.000 registros e 10 variáveis relacionadas ao consumo de vídeos curtos, tempo de tela, sono, foco, atenção e desempenho em tarefas, sendo disponibilizada publicamente sob licença CC0.

As variáveis consideradas na análise foram: idade, tempo diário de consumo de vídeos curtos (*reels_watch_time_hours*), tempo total de tela, horas de sono, score de atenção (*attention_span_score*), nível de foco, taxa de conclusão de tarefas, nível de estresse e plataforma utilizada.

A variável-alvo contínua, denominada no dataset como *score de atenção*, assume valores numéricos em uma escala de 1 a 10 e atua como indicador do nível atencional do usuário. Contudo, a documentação pública da base não detalha o processo de coleta, os critérios de amostragem populacional ou a validação dos instrumentos psicométricos utilizados para aferir esse score. Dessa forma, os dados são tratados estritamente no âmbito de uma investigação computacional exploratória sobre o comportamento de algoritmos de Aprendizado de Máquina em dados públicos.

Tabela 1. Variáveis utilizadas no estudo e respectivos tratamentos.

Variáveis	Tipo e papel	Tratamento adotado
Identificador do usuário	Identificador	Excluído das análises e dos modelos preditivos.
Idade, consumo de vídeos curtos, tempo de tela, sono, foco e tarefas	Preditoras numéricas	Verificação de outliers por IQR; padronização por Z-score para a clusterização.
Plataforma e nível de estresse	Preditoras categóricas	Codificação one-hot, com remoção de uma categoria de referência.
Score de atenção	Variável-alvo contínua	Utilizado na regressão e excluído da clusterização principal para evitar vazamento de dados.
Classe de atenção	Variável-alvo derivada	Criada apenas para classificação: atenção baixa ($< 5{,}5$) e atenção alta ($\geq 5{,}5$).

4.2. Pré-processamento, Codificação, Modelagem e Avaliação

Inicialmente, foram verificadas a presença de valores ausentes, registros duplicados e inconsistências textuais. A base continha 12.000 registros e 10 variáveis originais. Não foram identificados valores ausentes em nenhuma coluna, nem duplicatas exatas ou duplicatas associadas ao identificador dos usuários. Foi realizada a padronização de uma inconsistência textual na variável de plataforma, uniformizando registros referentes ao Instagram Reels.

O tratamento de valores extremos foi realizado pelo método do Intervalo Interquartil (IQR). Foram avaliadas as variáveis idade, tempo de consumo de vídeos curtos, tempo total de tela, horas de sono, score de atenção, nível de foco e taxa de conclusão de tarefas. Não foram identificados outliers estatísticos nas variáveis analisadas, preservando-se a totalidade dos 12.000 registros originais.

As variáveis numéricas utilizadas nos algoritmos sensíveis à distância foram padronizadas por Z-score, resultando em distribuições transformadas com média igual a

zero (0,0) e desvio padrão igual a um (1,0). Essa etapa foi indispensável para evitar que atributos com escalas de magnitude maiores dominassem o cálculo euclidiano das distâncias. As variáveis categóricas de plataforma e nível de estresse foram convertidas por codificação one-hot, com remoção de uma categoria de referência para evitar problemas de multicolinearidade nos modelos lineares.

A análise exploratória envolveu o cálculo de estatísticas descritivas, gráficos de dispersão bivariada e matrizes de correlação de Pearson (para relações lineares) e Spearman (para relações monotônicas).

Na etapa de agrupamento não supervisionado, os algoritmos foram aplicados excluindo-se estritamente o score de atenção para evitar vazamento de dados (*data leakage*). Avaliaram-se duas estratégias com o K-Means: (i) a **Abordagem Principal**, na qual o K-Means foi treinado sobre o espaço multidimensional original com as cinco variáveis comportamentais padronizadas (vídeos curtos, tempo de tela, sono, foco e conclusão de tarefas); e (ii) a **Abordagem Complementar em Espaço Reduzido**, na qual o K-Means foi ajustado diretamente sobre as duas primeiras componentes principais (PC1 e PC2) geradas pelo PCA. O número ótimo de grupos foi avaliado pelo Método do Cotovelo e Coeficiente de Silhueta. Para fins comparativos, aplicaram-se também *Gaussian Mixture Models* (GMM) e *Agglomerative Clustering* no espaço original.

Após a obtenção dos clusters em cada abordagem, aplicaram-se os testes ANOVA de uma via (paramétrico) e Kruskal-Wallis (não paramétrico) para avaliar se havia diferença estatisticamente significativa no score de atenção entre os agrupamentos comportamentais formados.

Na modelagem supervisionada, a comparação com modelos baseline (*Dummy Regressor* e *Dummy Classifier*) foi estabelecida como protocolo essencial para verificar se o aprendizado alcançou poder preditivo real. O *Random Forest* foi avaliado em duas tarefas: regressão (estimando o score contínuo de atenção) e classificação binária (separando em atenção baixa e alta no limiar de 5,5). Na classificação, a base continha 6.015 registros de atenção baixa (50,12%) e 5.985 de atenção alta (49,88%), configurando classes balanceadas que dispensaram técnicas de reamostragem sintética.

A otimização de hiperparâmetros foi conduzida via *RandomizedSearchCV* com validação cruzada interna de 3 folds, explorando número de árvores, profundidade máxima e critérios de divisão de nós. Para regressão, a busca selecionou 100 árvores e profundidade máxima igual a 5. Para classificação, o otimizador selecionou 300 árvores e profundidade máxima igual a 20.

Para verificar a robustez dos resultados e garantir que as conclusões não decorressem de particularidades do *Random Forest*, avaliaram-se também modelos supervisionados alternativos: Regressão Linear, Regressão Ridge, Regressão Logística, Support Vector Machines (SVM) e K-Nearest Neighbors (KNN).

O desempenho final de todos os modelos supervisionados foi avaliado por validação cruzada com 5 folds (KFold para regressão e StratifiedKFold para classificação). Na regressão, utilizaram-se as métricas MAE, MSE, RMSE e R^2. Na classificação, foram calculadas acurácia, precisão, sensibilidade e F1-score. Adicionalmente, calculou-se a Importância de Atributos por Permutação (*Permutation Feature Importance*) para mensurar a sensibilidade da predição a cada variável de entrada.

5. Resultados

5.1. Análise Exploratória e Correlação

A análise descritiva das variáveis numéricas é apresentada na Tabela 2. A idade média dos usuários foi de 29,66 anos, com consumo médio diário de vídeos curtos de 3,24 horas, tempo total de tela de 7,02 horas e média de sono de 6,50 horas. O score médio de atenção foi de 5,48.

Tabela 2. Estatísticas descritivas das principais variáveis numéricas.

Variável	Média	Desvio padrão	Intervalo
Idade	29,66	8,65	15–44
Vídeos curtos (h/dia)	3,24	1,58	0,50–6,00
Tempo de tela (h/dia)	7,02	2,87	2,00–12,00
Sono (h/noite)	6,50	1,45	4,00–9,00
Score de atenção	5,48	2,60	1,00–10,00
Nível de foco	5,51	2,59	1,00–10,00
Conclusão de tarefas	70,09	17,20	40,01–99,99

As matrizes de correlação de Pearson e Spearman indicaram ausência de associação estatística entre o consumo de vídeos curtos e a capacidade atencional. Em ambas as análises, a correlação entre *reels_watch_time_hours* e *attention_span_score* foi de aproximadamente $0{,}0008$, valor virtualmente nulo.

A Figura 1 ilustra o gráfico de dispersão entre essas duas variáveis. Observa-se uma nuvem de pontos retangular e perfeitamente uniforme, com a reta de regressão ajustada mantendo-se perfeitamente horizontal em toda a extensão do eixo horizontal. Esse comportamento visual evidencia que usuários com os mais variados níveis de consumo diário de vídeos curtos (de 0,5h a 6,0h) apresentam exatamente a mesma amplitude e distribuição dispersa no score de atenção (de 1 a 10), configurando uma distribuição estocástica independente.

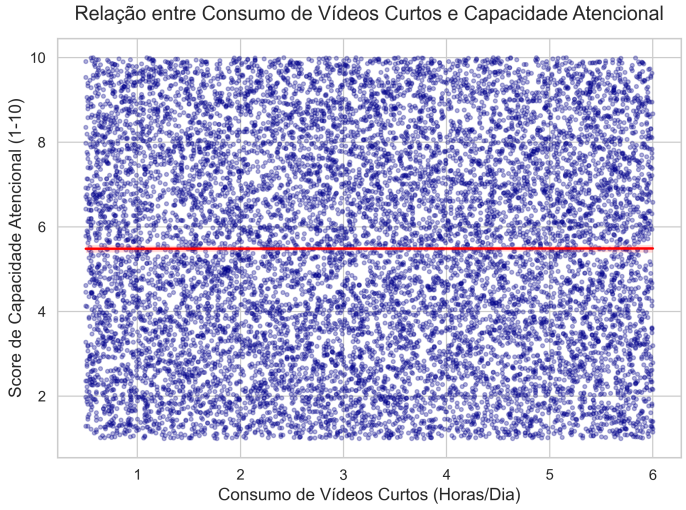


Figura 1. Relação entre consumo diário de vídeos curtos e score de capacidade atencional, evidenciando distribuição uniforme e ausência de correlação linear.

5.2. Clusterização com K-Means e Validação Estatística

Na Abordagem Principal de clusterização (sem o score de atenção), o algoritmo *K-Means* aplicado sobre o espaço multidimensional das variáveis comportamentais originais padronizadas identificou $K = 3$ como número ótimo de grupos, obtendo coeficiente de silhueta de $0{,}1380$, o que reflete uma separação difusa e contínua no espaço multidimensional.

Na Abordagem Complementar em Espaço Reduzido, o K-Means foi treinado diretamente sobre as duas primeiras componentes do PCA ($K = 3$), o que elevou o coeficiente de silhueta para $0{,}3413$ (Figuras 2 e 3). Esse ganho no coeficiente de silhueta decorre da compressão das dimensões e simplificação da geometria no plano 2D, e não de uma separação intrínseca nos dados brutos.

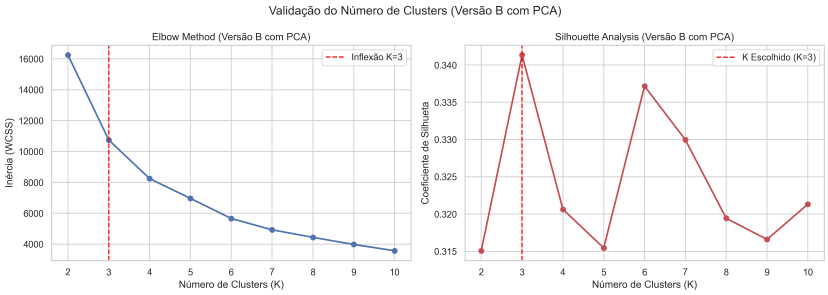


Figura 2. Validação do número de clusters no espaço projetado por PCA.

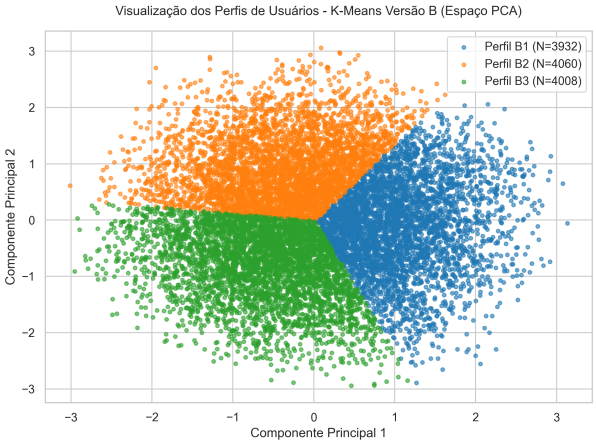


Figura 3. Visualização bidimensional dos perfis de usuários identificados pelo K-Means treinado no espaço PCA.

Para verificar se os agrupamentos comportamentais formados correspondiam a diferenças no score de atenção, aplicaram-se os testes estatísticos de hipóteses em ambas as abordagens. Conforme detalhado na Tabela 3, a comparação do score de atenção entre os clusters no espaço original apresentou $p = 0{,}3009$ (ANOVA) e $p = 0{,}2994$ (Kruskal-Wallis). Na versão ajustada sobre o PCA, os valores foram $p = 0{,}0670$ (ANOVA) e $p = 0{,}0678$ (Kruskal-Wallis). A diferença nos valores de p reflete a ligeira alteração na atribuição das instâncias aos clusters quando o K-Means atua nas componentes principais em vez das variáveis originais. Contudo, em ambas as configurações, todos os testes resultaram em $p > 0{,}05$, confirmando que os clusters comportamentais não diferem significativamente em termos de capacidade atencional.

Tabela 3. Testes estatísticos para comparação do score de atenção entre clusters.

Abordagem de Clusterização	ANOVA (p)	Kruskal-Wallis (p)
Abordagem Principal (Espaço Original Z-score)	$p = 0{,}3009$	$p = 0{,}2994$
Abordagem Complementar (Espaço PCA 2D)	$p = 0{,}0670$	$p = 0{,}0678$

5.3. Modelagem Supervisionada com Random Forest

Na tarefa de regressão para estimar o score de atenção contínuo, o *Random Forest Regressor* obteve MAE de 2,2523, MSE de 6,7496, RMSE de 2,5980 e R^2 de -0,0004. O modelo baseline *Dummy Regressor* obteve MAE de 2,2533, MSE de 6,7481, RMSE de 2,5977 e R^2 de -0,0001 (Tabela 4). O desempenho do modelo supervisionado foi equivalente ao chute da média, demonstrando a incapacidade das variáveis em explicar a variância do alvo.

Tabela 4. Comparação entre Random Forest Regressor ajustado e Dummy Regressor.

Modelo	MAE	MSE	RMSE	R^2
Random Forest Regressor	2,2523	6,7496	2,5980	-0,0004
Dummy Regressor	2,2533	6,7481	2,5977	-0,0001

Na tarefa de classificação binária, o *Random Forest Classifier* apresentou acurácia de 0,5042, precisão de 0,5029, sensibilidade de 0,4924 e F1-score de 0,4975. O *Dummy Classifier* estratificado atingiu acurácia de 0,5064, precisão de 0,5051, sensibilidade de 0,5131 e F1-score de 0,5091 (Tabela 5).

Tabela 5. Comparação entre Random Forest Classifier ajustado e Dummy Classifier.

Modelo	Acurácia	Precisão	Sensibilidade	F1-score
Random Forest Classifier	0,5042	0,5029	0,4924	0,4975
Dummy Classifier	0,5064	0,5051	0,5131	0,5091

O fato de a busca por hiperparâmetros na classificação ter convergido para uma arquitetura profunda (300 árvores, profundidade máxima 20) explica-se pelo sobreajuste ao ruído amostral do conjunto de treino. Sem sinal preditivo real nos dados, o algoritmo subdividiu o espaço em busca de partições puras aleatórias, resultando em acurácia no teste de 50,42%, compatível com o nível do acaso e inferior à obtida pelo *Dummy Classifier* estratificado (50,64%).

Adicionalmente, a análise de Importância de Atributos por Permutação indicou que todas as variáveis obtiveram pontuações próximas de zero (com variações negativas oscilando entre $-0{,}0015$ e $0{,}0012$), confirmando que a permutação aleatória de qualquer coluna não causa impacto na acurácia do modelo, evidenciando a ausência de sinal preditivo individual ou combinado.

5.4. Análises Complementares e Comparação de Modelos

As análises complementares investigaram segmentações por quartis de consumo de vídeos curtos, proporção entre consumo de vídeos e tempo total de tela, impacto por plataforma de consumo e a avaliação de algoritmos adicionais.

Após a correção False Discovery Rate (FDR) de Benjamini-Hochberg, nenhuma das comparações entre subgrupos apresentou significância estatística. A comparação entre extremos de consumo de vídeos curtos resultou em $p = 0{,}9946$ ajustado. A comparação do score de atenção entre diferentes plataformas apresentou $p = 0{,}2521$ ajustado.

Os modelos supervisionados alternativos avaliados (Regressão Linear, Regressão Ridge, Regressão Logística, Support Vector Machines e KNN) confirmaram o mesmo padrão: nenhum regressor obteve $R^2 > 0{,}0000$ (com o melhor resultado atingindo $-0{,}0001$), e nenhum classificador alternativo superou acurácia de $0{,}5108$, mantendo-se no mesmo patamar de desempenho do baseline estratificado ($0{,}5064$).

Além disso, a avaliação comparativa de diferentes técnicas de clusterização sobre o conjunto de atributos comportamentais é sumarizada na Tabela 6. Tanto o *Gaussian Mixture Model* (GMM) quanto a Clusterização Aglomerativa (*Agglomerative Clustering*) resultaram em baixos coeficientes de silhueta ($\approx 0{,}13$) e testes ANOVA e Kruskal-Wallis com $p > 0{,}30$, reforçando que nenhum método de agrupamento identificou partições associadas a níveis atencionais distintos.

Tabela 6. Comparação entre múltiplos métodos de clusterização não supervisionada.

Método de Agrupamento	**Silhueta (Original)**	**ANOVA (p)**	**Kruskal-Wallis (p)**
K-Means ($K = 3$)	0,1380	0,3009	0,2994
Gaussian Mixture Model (GMM)	0,1295	0,4120	0,4087
Agglomerative Clustering (Ward)	0,1312	0,3550	0,3518

6. Discussão

Os resultados empíricos obtidos convergem de maneira inequívoca: correlações nulas ($r \approx 0{,}0008$), gráficos de dispersão uniformes (Figura 1), ausência de diferenciação atencional entre clusters formados e total incapacidade dos modelos supervisionados em superar baselines aleatórios.

Esses achados fornecem evidências metodológicas muito consistentes de que a base analisada apresenta comportamento compatível com variáveis geradas de forma sintética e mutuamente independente. Assim, a ausência de correlação encontrada reflete diretamente a estrutura estatística do dataset disponível no Kaggle, e não deve ser tomada como uma refutação dos processos neurocognitivos reais da atenção humana.

A principal contribuição deste trabalho é, portanto, de natureza metodológica e computacional. Demonstra-se na prática como a aplicação de modelos avançados de Aprendizado de Máquina (como Random Forest com 300 árvores e profundidade 20) sobre bases sem sinal pode induzir a conclusões equivocadas de sobreajuste caso não haja um protocolo rígido de validação. A inclusão obrigatória de *Dummy Regressors*, *Dummy Classifiers*, testes de hipóteses robustos e correção FDR de Benjamini-Hochberg revelou-se indispensável para expor a ausência de sinal preditivo real e evitar falsas descobertas científicas.

Além disso, o estudo suscita uma discussão crítica urgente sobre a proveniência e a governança de dados em plataformas públicas abertas. Conjuntos de dados disponibilizados sem metadados transparentes de coleta e sem validação psicométrica de construtos sensíveis (como atenção e foco) exigem cautela redobrada dos pesquisadores. Protocolos

de auditoria exploratória inicial e benchmarking com baselines tornam-se, assim, uma etapa indispensável antes de qualquer empreendimento de modelagem preditiva em larga escala.

Na realidade comportamental e neurobiológica, a regulação atencional e os impactos do consumo digital constituem fenômenos multifatoriais complexos, dependentes de dimensões não mensuradas na base analisada, como contexto socioemocional, qualidade pedagógica do conteúdo, comorbidades psicológicas e histórico do usuário.

7. Limitações do Estudo

A limitação primária do presente estudo repousa na proveniência da base de dados utilizada. Embora seja um conjunto público amplamente acessível sob licença CC0 no Kaggle com 12.000 registros, inexistem descrições sobre a metodologia de amostragem, coleta em campo ou protocolos psicométricos utilizados para definir o *score de atenção*.

Adicionalmente, as características de independência e provável natureza sintética das variáveis restringem as conclusões ao escopo da análise metodológica de integridade de modelos de Aprendizado de Máquina. Fatores contextuais determinantes — como padrão de sono REM, uso concomitante de telas (*media multitasking*), tipo de conteúdo assistido e histórico clínico — não estão contemplados. Por fim, o delineamento transversal dos dados inviabiliza inferências sobre efeitos causais ao longo do tempo.

8. Conclusão

Este trabalho investigou a relação entre o consumo diário de vídeos curtos e a capacidade atencional em uma base de dados pública com 12.000 registros, aplicando uma metodologia computacional abrangente que combinou estatística descritiva, correlações, clusterização multivariada (K-Means, GMM, Clusterização Aglomerativa), PCA, modelos supervisionados (Random Forest e modelos alternativos) e testes de hipóteses paramétricos e não paramétricos.

Os resultados evidenciaram ausência total de sinal estatístico ou preditivo, com correlação nula e modelos de Aprendizado de Máquina incapazes de superar baselines ingênuos. A análise revelou que os dados apresentam características fortemente compatíveis com ruído independente, evidenciando o papel crucial de boas práticas metodológicas na Ciência da Computação para prevenir sobreajuste e falsas inferências em dados comportamentais.

Como trabalhos futuros, recomenda-se: (i) a realização de coletas primárias com participantes humanos por meio de metodologias de Avaliação Momentânea Ecológica (EMA) e rastreamento passivo de tempo de tela em smartphones; (ii) a aplicação de baterias neurocognitivas padronizadas e validadas (como os testes de Stroop e Continuous Performance Task - CPT); e (iii) a execução de estudos longitudinais controlados para monitorar a dinâmica da atenção ao longo do tempo.

Disponibilidade de código e dados. Para assegurar a plena transparência e reprodutibilidade computacional deste estudo, o código-fonte integral dos experimentos, scripts de pré-processamento, parametrização dos modelos e rotinas estatísticas estão disponíveis publicamente no repositório GitHub: `https://github.com/NicolasHarnisch/reels-attention-ml-analysis`.

Referências

Bartijotto, G. P. (2025). Economia da atenção nas redes sociais: estratégias para capturar e reter o consumidor nas redes sociais. Monografia de Graduação em Administração, Pontifícia Universidade Católica de São Paulo (PUC-SP).

Citton, Y. (2013). L'économie de l'attention. Revue des Livres (RdL), n. 11. Mai–juin. Disponível em: `https://www.yvescitton.net/wp-content/uploads/2013/10/Citton-EconomieAttention-RdL-11-2013.pdf`. Acesso em: 15 jun. 2026.

da Silva Pereira, L., de Oliveira, F. M., and de Ávila Medeiros, T. (2025). Tiktok: sua influência e efeitos no sistema nervoso, bem como seu impacto negativo no processo de aprendizagem escolar. *Ciência Atual – Revista Científica Multidisciplinar do Centro Universitário São José*, 22(1).

Joshi, J. (2025). Reels/shorts consumption vs attention span. Kaggle Dataset. Disponível em: `https://www.kaggle.com/datasets/jayjoshi37/reelsshorts-consumption-vs-attention-span`. Acesso em: 29 abr. 2026.

Lash, S. (2012). Formas tecnológicas de vida. *Revista Estudos de Sociologia*, 8(1). Traduzido por Jonathas Ferreira.

Marengo, D., Sindermann, C., Elhai, J. D., and Montag, C. (2020). One social media company to rule them all: Associations between use of facebook-owned social media platforms, sociodemographic characteristics, and the big five personality traits. *Frontiers in Psychology*, 11:936.

Meng, S.-Q., Cheng, J.-L., Li, Y.-Y., Yang, X.-Q., Zheng, J.-W., Chang, X.-W., Shi, Y., Chen, Y., Lu, L., Sun, Y., et al. (2022). Global prevalence of digital addiction in general population: A systematic review and meta-analysis. *Clinical Psychology Review*, 92:102128.

Noronha, A. C. C. (2024). Dispersos em tempos de economia da atenção: a tecnologia e nós. *Texto Livre*, 17:e47843.

Rodrigues, P. S. S. and Walderrama, G. d. T. P. (2025). A neurobiologia da dependência digital: uma revisão sobre o sistema dopaminérgico, recompensa variável e o controle da atenção. *RCMOS – Revista Científica Multidisciplinar O Saber*, 1(2).

Schreiner, M., Fischer, T., and Riedl, R. (2021). Impact of content characteristics and emotion on behavioral engagement in social media: literature review and research agenda. *Electronic Commerce Research*, 21(2):329–345.

Sousa, M. C. C. (2019). Uma análise do algoritmo k-means como introdução ao aprendizado de máquinas. Monografia de Graduação em Matemática, Universidade Federal do Tocantins (UFT).