IOSR Journal of Business and Management (IOSR-JBM)
e-ISSN:2278-487X, p-ISSN: 2319-7668. Volume 27, Issue 4. Ser. 7 (April. 2025), PP 60-71
www.iosrjournals.org

Modelagem Do Engajamento Estudantil Com Sinais Neurofisiológicos E Aprendizado De Máquina: Aplicações De Inteligência Artificial E Neurociência Na Educação

Eduardo Silva Vasconcelos
Doutor Em Ciências – Processamento Da Informação
Instituto Federal Goiano
Goiânia, Goiás, Brasil

Eduarda Rocha Vasconcelos
Graduanda Em Medicina
Faculdade Morgana Potrich
Mineiros, Goiás, Brasil

Leia Rocha Vasconcelos
Especialização Em Psicopedagogia
Centro Universitário Unibf
Cristalina, Goiás, Brasil

Resumo:
Este trabalho investiga a aplicação de dados neurofisiológicos e comportamentais para a predição do engajamento estudantil em ambientes de aprendizagem mediados por tecnologias digitais. Utilizando um conjunto de dados obtido na plataforma Kaggle, composto por 3.000 registros com sinais de eletroencefalografia (EEG), métricas de rastreamento ocular (eye-tracking) e variáveis contextuais de aprendizagem, buscou-se construir e avaliar modelos preditivos capazes de estimar o nível de engajamento cognitivo. A pesquisa caracteriza-se como quantitativa, descritiva e exploratória, com ênfase na análise estatística descritiva, na visualização de dados e na modelagem supervisionada por meio do algoritmo Random Forest. Os resultados indicaram que variáveis como Theta_PSD, Delta_PSD, Gamma_PSD e Fixation_Duration se destacaram na predição dos níveis de engajamento, apesar da acurácia geral do modelo ter se mantido em torno de 35,1%. Observou-se maior desempenho na identificação da classe de engajamento médio, revelando desafios na classificação dos extremos (baixo e alto). As discussões realizadas evidenciam o potencial da abordagem integrada entre neurociência, educação e inteligência artificial para o desenvolvimento de sistemas adaptativos que respondam dinamicamente ao estado mental dos estudantes. Conclui-se que a utilização de sinais fisiológicos como base para sistemas educacionais personalizados representa uma via promissora para o avanço da neuroeducação e para o aprimoramento da eficácia pedagógica em ambientes digitais.

Palavras-chave: *Engajamento Estudantil. Neuroeducação. Sinais de eletroencefalografia (EEG). Métricas de rastreamento ocular (Eye-tracking). Aprendizado de Máquina. Random Forest; Inteligência Artificial na Educação.*

Date of Submission: 16-04-2025 Date of Acceptance: 26-04-2025

I. Introdução

O avanço das tecnologias digitais e das ciências cognitivas tem impulsionado transformações profundas nos contextos educacionais, especialmente no que se refere à personalização da aprendizagem. A capacidade de adaptar o ensino às necessidades individuais dos estudantes, com base em dados objetivos, abre novas possibilidades para otimizar o processo de ensino-aprendizagem. Conforme apontado por Henrie, Halverson e Graham (2015), o engajamento é um dos principais preditores do sucesso acadêmico, e sua mensuração precisa torna-se cada vez mais relevante em ambientes digitais. Nesse cenário, compreender e monitorar o engajamento estudantil torna-se fundamental, uma vez que esse construto está diretamente associado ao desempenho acadêmico, à motivação e à permanência do aluno em ambientes educacionais.

A neuroeducação, área que integra conhecimentos da neurociência, psicologia e pedagogia, tem contribuído para a identificação de padrões neurais e comportamentais que refletem os estados cognitivos dos

estudantes. Segundo Thomas, Ansari e Knowland (2019), a plasticidade cerebral permite que práticas pedagógicas bem estruturadas influenciem diretamente o desenvolvimento das funções executivas e a aprendizagem. De maneira complementar, a inteligência artificial (IA) e, mais especificamente, as técnicas de aprendizado de máquina, oferecem ferramentas robustas para analisar grandes volumes de dados e construir modelos preditivos capazes de interpretar esses estados em tempo real. A interseção entre essas áreas permite a criação de sistemas de ensino adaptativos, capazes de ajustar dinamicamente os conteúdos e estratégias pedagógicas com base na resposta fisiológica e comportamental do aluno.

Este trabalho investiga a predição do engajamento estudantil por meio de sinais neurofisiológicos — com destaque para os dados de eletroencefalografia (EEG) — e métricas de rastreamento ocular (eye-tracking), utilizando modelos de aprendizado de máquina. O problema de pesquisa que orienta este estudo pode ser formulado nos seguintes termos: como os sinais de EEG e os dados de eye-tracking podem ser utilizados para prever, com acurácia, os níveis de engajamento de estudantes em ambientes educacionais?

O objetivo geral desta pesquisa é desenvolver e avaliar modelos preditivos que estimem os níveis de engajamento com base em dados neurofisiológicos e visuais. Para tanto, são estabelecidos os seguintes objetivos específicos: (i) realizar uma revisão da literatura sobre engajamento, neurociência e IA aplicada à educação; (ii) aplicar técnicas de estatística descritiva e análise exploratória ao conjunto de dados; (iii) treinar modelos de aprendizado de máquina e avaliar seu desempenho com métricas específicas; e (iv) discutir os resultados obtidos e suas implicações para o desenvolvimento de sistemas adaptativos de ensino.

O banco de dados utilizado nesta pesquisa foi obtido na plataforma Kaggle, especificamente no repositório "Student Engagement Dataset Using EEG" (https://www.kaggle.com/datasets/ziya07/student-engagement-dataset-using-eeg). Este dataset foi desenvolvido com o intuito de fornecer dados reais para a predição de engajamento em tempo real em contextos educacionais. Ele contém 3.000 registros com variáveis neurofisiológicas (como a densidade espectral de potência das bandas Delta, Teta, Alfa, Beta e Gama do EEG), dados de rastreamento ocular (como dilatação da pupila, taxa de piscadas, duração das fixações e velocidade dos movimentos sacádicos), além de variáveis contextuais (tipo de conteúdo e nível de dificuldade da tarefa) e o rótulo de engajamento categorizado em baixo, médio e alto. Trata-se de um conjunto de dados especialmente valioso, tanto por seu grau de detalhamento quanto por sua aplicabilidade em pesquisas interdisciplinares que envolvam educação, neurociência e ciência de dados.

Embora estudos anteriores tenham explorado as relações entre engajamento e variáveis fisiológicas isoladas, poucos aplicaram algoritmos de aprendizado de máquina com um conjunto integrado de dados neurofisiológicos e de rastreamento ocular para a predição do engajamento estudantil. Esta abordagem diferenciada reforça a originalidade do presente trabalho e contribui de forma inovadora para o campo da neuroeducação baseada em dados.

Além da relevância científica, esta pesquisa também possui implicações práticas. Os resultados obtidos podem apoiar o desenvolvimento de plataformas adaptativas que ajustem o conteúdo e a metodologia de ensino em tempo real, considerando o estado atencional do estudante. Isso representa um avanço em direção a experiências educacionais mais personalizadas, eficazes e inclusivas.

II. Revisão Da Literatura

Este capítulo tem como objetivo apresentar o referencial teórico que fundamenta a presente pesquisa, com foco na interseção entre neurociência, educação e inteligência artificial. A compreensão dos processos cognitivos que ocorrem durante a aprendizagem, assim como a identificação de estratégias para mensurar e promover o engajamento estudantil, são essenciais para o desenvolvimento de ambientes educacionais mais eficientes e personalizados. Neste contexto, destacam-se os avanços proporcionados pelas tecnologias de EEG (eletroencefalograma) e eye-tracking (rastreamento ocular), bem como pelas técnicas de machine learning, que permitem a análise e a predição de estados mentais e comportamentais dos alunos. O capítulo está organizado em cinco seções principais que abordam desde os fundamentos neurocientíficos da aprendizagem até as aplicações práticas em sistemas de ensino adaptativo baseados em IA.

Neurociência e Processos Cognitivos na Aprendizagem

A neurociência cognitiva tem contribuído de forma significativa para a compreensão dos mecanismos cerebrais envolvidos na aprendizagem. O cérebro humano é altamente plástico, e essa neuroplasticidade permite a reorganização das conexões neuronais em resposta a estímulos ambientais, experiências e práticas educacionais. Thomas, Ansari e Knowland (2019) destacam que essa característica torna o cérebro um alvo direto das intervenções pedagógicas, pois atividades de ensino podem moldar e influenciar o desenvolvimento das funções cognitivas superiores, como a atenção, a memória e o controle executivo.

Segundo Goldberg (2022), a aprendizagem envolve a criação e o fortalecimento de conexões sinápticas, o que justifica a importância de estratégias pedagógicas que considerem os limites da memória de trabalho, a necessidade de intervalos de atenção e os estados emocionais dos alunos. A atenção, por exemplo, depende de

redes neurais específicas, como as redes atencionais dorsal e ventral, responsáveis por filtrar estímulos relevantes e manter o foco durante a execução de tarefas cognitivas.

Além disso, condições fisiológicas como sono adequado, alimentação e atividade física influenciam diretamente o desempenho cognitivo dos alunos, interferindo em sua capacidade de adquirir e consolidar conhecimentos (Goldberg, 2022).

Engajamento Estudantil e Suas Dimensões

O engajamento estudantil é um construto multidimensional, que abrange componentes comportamentais, emocionais e cognitivos. De acordo com Fredricks, Blumenfeld e Paris (2004), o engajamento comportamental se refere à participação ativa do aluno em atividades escolares, enquanto o engajamento emocional diz respeito ao interesse, entusiasmo e afeto em relação ao processo de aprendizagem. Já o engajamento cognitivo envolve o esforço mental, a autorregulação e o uso de estratégias para resolver problemas e compreender conteúdos complexos.

Henrie, Halverson e Graham (2015) afirmam que o engajamento é um fator crítico para o sucesso acadêmico, estando positivamente associado à persistência, ao desempenho e à satisfação dos alunos com suas experiências de aprendizagem. Em contrapartida, baixos níveis de engajamento estão relacionados à desmotivação, à evasão e ao baixo rendimento escolar.

Estudos de Baker et al. (2010) sugerem que estados emocionais como frustração, embora considerados negativos, podem ser mais produtivos para a aprendizagem do que estados de tédio, pois a frustração pode levar ao reengajamento, enquanto o tédio tende a persistir e prejudicar o desempenho. Assim, compreender os fatores que promovem ou inibem o engajamento é essencial para o desenvolvimento de intervenções pedagógicas eficazes.

Uso de EEG e Eye-Tracking na Mensuração de Estados Mentais

O uso de sinais neurofisiológicos, como a eletroencefalografia (EEG) e o rastreamento ocular (eye-tracking), tem se mostrado promissor na mensuração objetiva do engajamento estudantil. A EEG permite captar variações na atividade elétrica cerebral por meio de eletrodos colocados no couro cabeludo, possibilitando a análise das bandas de frequência cerebral (delta, teta, alfa, beta e gama), que estão associadas a diferentes estados mentais.

Segundo Pope, Bogart e Bartolome (1995), o índice de engajamento mental pode ser calculado a partir da razão entre as bandas beta (associadas à atenção) e as bandas alfa e teta (associadas ao relaxamento), sendo possível monitorar flutuações na atenção e na carga cognitiva dos estudantes em tempo real.

Apicella et al. (2022) demonstraram que é possível utilizar modelos computacionais para classificar o nível de engajamento de estudantes com base em sinais de EEG durante atividades de aprendizagem. Os resultados indicaram que tais modelos conseguem distinguir entre estados de alto e baixo engajamento com precisão significativa.

O eye-tracking, por sua vez, fornece informações sobre o comportamento visual dos estudantes, como fixações oculares, sacadas, dilatação pupilar e frequência de piscadas. De acordo com Lai et al. (2013), essas métricas estão correlacionadas com o esforço mental, o interesse e a carga cognitiva, permitindo inferir o nível de engajamento visual do aluno.

Bixler e D'Mello (2016) aplicaram técnicas de rastreamento ocular para detectar momentos de distração (mind wandering) durante a leitura, observando que padrões específicos de movimento ocular estavam associados à perda de foco atencional.

Aplicações de Machine Learning na Predição do Engajamento

O avanço das técnicas de aprendizado de máquina (machine learning) tem possibilitado o desenvolvimento de modelos capazes de analisar grandes volumes de dados neurofisiológicos e comportamentais para prever estados cognitivos dos estudantes. Esses modelos podem ser treinados com dados rotulados de EEG e eye-tracking para identificar padrões que correspondem a diferentes níveis de engajamento.

Segundo Alruwais e Zakariah (2023), algoritmos como Random Forest, SVM e redes neurais têm sido utilizados para classificar o engajamento estudantil com base em dados coletados durante atividades educacionais, alcançando resultados promissores em termos de acurácia e generalização.

Monkaresi et al. (2017) demonstraram que a integração de múltiplas fontes de dados – como expressões faciais, orientação da cabeça e padrões de olhar – pode melhorar a predição do engajamento, permitindo intervenções mais eficazes nos ambientes de aprendizagem.

A personalização desses modelos também é relevante. De acordo com Rahman et al. (2022), sistemas adaptativos que consideram tanto o desempenho quanto os sinais fisiológicos do estudante são mais eficazes para manter o engajamento e otimizar a aprendizagem.

Ambientes de Aprendizagem Adaptativos Baseados em IA

A aplicação de modelos de machine learning em ambientes educacionais tem viabilizado a criação de sistemas de aprendizagem adaptativos, capazes de ajustar o conteúdo, o ritmo e as estratégias de ensino com base no estado cognitivo do aluno. Esses ambientes utilizam dados de EEG e eye-tracking para monitorar continuamente o engajamento do estudante e adaptar a experiência de aprendizagem em tempo real.

Apicella et al. (2022) propõem que, ao detectar sinais de desatenção, o sistema pode apresentar exemplos práticos ou modificar a apresentação do conteúdo para recuperar o foco do aluno. Essa abordagem é especialmente útil em plataformas de ensino a distância, onde a ausência do professor dificulta o monitoramento direto do engajamento.

Baradari et al. (2025) desenvolveram um protótipo de tutor inteligente chamado NeuroChat, que utiliza EEG para ajustar a complexidade das respostas de um chatbot educacional com base no nível de engajamento detectado. O sistema demonstrou melhorias tanto na atenção objetiva quanto na satisfação dos alunos.

Tais tecnologias ilustram o potencial da convergência entre neurociência, educação e inteligência artificial para a criação de ambientes de aprendizagem mais responsivos, personalizados e eficazes.

III. Metodologia

Este capítulo descreve os procedimentos metodológicos utilizados nesta pesquisa, com foco na análise estatística e na modelagem preditiva aplicada ao conjunto de dados neurofisiológicos e comportamentais relacionados ao engajamento estudantil. As abordagens aqui descritas foram escolhidas com base em critérios de rigor científico e relevância para os objetivos do estudo, buscando garantir a reprodutibilidade e a validade dos resultados.

Tipo de Pesquisa

A pesquisa desenvolvida caracteriza-se como um estudo de natureza quantitativa, de abordagem empírica, com delineamento exploratório e descritivo. Segundo Gil (2017), a pesquisa exploratória visa proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses. Já a pesquisa descritiva tem como principal objetivo descrever as características de determinada população ou fenômeno ou, então, o estabelecimento de relações entre variáveis.

No presente estudo, o objetivo central foi identificar padrões em dados neurofisiológicos e comportamentais associados ao engajamento de estudantes em atividades educacionais, bem como testar a aplicabilidade de modelos computacionais para a predição automatizada desse engajamento.

Conjunto de Dados

Foi utilizado um conjunto de dados contendo 3.000 registros de estudantes submetidos a diferentes tipos de conteúdo educacional (texto, vídeo, interativo) e níveis de dificuldade (fácil, médio, difícil). Para cada registro, foram coletadas informações de sinais de EEG (densidade espectral de potência nas bandas Delta, Teta, Alfa, Beta e Gama), dados de rastreamento ocular (dilatação da pupila, taxa de piscadas, duração das fixações, velocidade dos movimentos sacádicos) e o rótulo de engajamento (baixo, médio, alto).

Análise Estatística Descritiva

Inicialmente, foram aplicadas técnicas de estatística descritiva para caracterizar o conjunto de dados. Foram analisadas medidas de tendência central (média, mediana), dispersão (desvio padrão) e distribuição das variáveis contínuas e categóricas. A análise descritiva permitiu compreender o comportamento das variáveis e identificar possíveis outliers e padrões iniciais relevantes para a formulação de hipóteses.

Análise Exploratória e Visualização

A análise exploratória de dados (EDA – Exploratory Data Analysis) é uma etapa fundamental em qualquer pesquisa quantitativa, pois permite identificar padrões, tendências, outliers e relações preliminares entre variáveis antes da modelagem. Nesta pesquisa, a EDA foi realizada por meio de técnicas estatísticas gráficas, com o objetivo de compreender o comportamento dos dados e embasar a construção dos modelos preditivos.

Foram utilizados:

- **Gráficos de distribuição de classes**: para examinar a proporção de registros nos diferentes níveis de engajamento (baixo, médio e alto), avaliando o equilíbrio entre as classes.
- **Matriz de correlação de Pearson**: para identificar relações lineares entre variáveis contínuas, especialmente entre bandas de EEG e métricas de eye-tracking.
- **Boxplots das variáveis de EEG**: para analisar a variação de cada banda cerebral em função dos níveis de engajamento. Essas visualizações são úteis para detectar diferenças significativas entre grupos.

Essas ferramentas auxiliaram na detecção de colinearidades e variáveis potencialmente redundantes, bem como na identificação de atributos com maior poder discriminativo, fundamentais para o sucesso da etapa de modelagem preditiva.

Preparação dos Dados

Para garantir que os dados estivessem adequados à modelagem preditiva, foram realizadas etapas fundamentais de pré-processamento. As variáveis categóricas, como "tipo de conteúdo" e "nível de dificuldade", foram codificadas numericamente por meio de rótulos inteiros, garantindo a compatibilidade com algoritmos de aprendizado supervisionado.

As variáveis contínuas foram padronizadas utilizando a técnica de normalização Z-score, que transforma os dados para uma distribuição com média igual a zero e desvio padrão igual a um. Essa etapa é fundamental para evitar que variáveis com escalas diferentes influenciem de maneira desproporcional os modelos.

O conjunto de dados foi, então, dividido em dois subconjuntos:

- **Dados de treinamento (75%)**: utilizados para treinar os algoritmos e ajustar seus parâmetros internos.
- **Dados de teste (25%)**: utilizados para avaliar a capacidade de generalização do modelo, ou seja, sua performance ao lidar com dados que não foram vistos durante o treinamento.

Essa divisão é essencial para evitar o problema de overfitting, que ocorre quando o modelo aprende muito bem os dados de treinamento, mas falha em generalizar para novos dados.

Modelagem com Aprendizado de Máquina

A modelagem preditiva foi realizada com o algoritmo Random Forest (Floresta Aleatória), uma técnica de aprendizado supervisionado baseada em múltiplas árvores de decisão. Esse algoritmo é conhecido por sua robustez, capacidade de lidar com grandes volumes de dados e variáveis correlacionadas, além de apresentar bom desempenho em tarefas de classificação multiclasse.

O modelo foi treinado com os dados padronizados, utilizando os seguintes hiperparâmetros:

- **Número de árvores (n_estimators = 100)**: determina a quantidade de árvores que compõem a floresta. Um número maior tende a aumentar a estabilidade e a precisão do modelo.
- **Critério de divisão (criterion = 'gini')**: métrica utilizada para avaliar a qualidade de uma divisão nos nós das árvores. O índice de Gini mede o grau de impureza de um nó.
- **Random state = 42**: parâmetro utilizado para garantir a reprodutibilidade dos resultados.

A avaliação do modelo foi realizada com as seguintes métricas:

- **Acurácia**: proporção total de acertos entre todas as previsões.
- **Precisão**: proporção de previsões corretas dentro da classe prevista.
- **Recall**: capacidade do modelo de identificar todos os elementos relevantes de uma classe.
- **F1-score**: média harmônica entre precisão e recall, útil quando há desequilíbrio entre as classes.
- **Matriz de confusão**: tabela que mostra os acertos e erros do modelo para cada classe, permitindo identificar padrões de erro.

A escolha dessas métricas visou oferecer uma visão abrangente do desempenho do modelo, tanto em termos gerais quanto específicos para cada classe de engajamento.

Importância das Variáveis

Foi também calculada a importância relativa de cada variável para a predição do engajamento, com base na métrica de ganho de informação utilizada pelo Random Forest. Essa etapa foi crucial para validar as hipóteses teóricas sobre a relevância das bandas cerebrais e indicadores de rastreamento ocular na identificação de estados de atenção e envolvimento cognitivo.

A análise da importância das variáveis também contribuiu para identificar quais atributos fisiológicos apresentaram maior poder discriminativo entre os níveis de engajamento. Essa informação é útil para refinar modelos futuros e reduzir a complexidade computacional, ao eliminar variáveis com baixo impacto na predição.

A metodologia adotada neste estudo permite não apenas uma análise aprofundada dos dados existentes, mas também a criação de subsídios técnicos e científicos para o desenvolvimento de sistemas educacionais adaptativos baseados em inteligência artificial.

No próximo capítulo, os resultados obtidos por meio das técnicas descritas serão apresentados e discutidos à luz dos objetivos da pesquisa, permitindo verificar a eficácia dos métodos aplicados e suas contribuições para o campo da neuroeducação.

IV. Análise Dos Dados

A análise dos dados constitui uma etapa essencial para a compreensão das dinâmicas neurofisiológicas e comportamentais que sustentam o engajamento estudantil em ambientes de aprendizagem mediados por tecnologias. Este capítulo tem como propósito examinar, de forma minuciosa, os padrões presentes nas variáveis extraídas do conjunto de dados, que reúne registros de sinais eletroencefalográficos (EEG), parâmetros de rastreamento ocular (eye-tracking) e informações contextuais de aprendizagem, tais como tipo de conteúdo e grau de dificuldade. A abordagem adotada é quantitativa, exploratória e preditiva, estruturada em três grandes eixos: (i) análise estatística descritiva das variáveis, (ii) análise exploratória com apoio de visualizações gráficas e (iii) modelagem com algoritmos de aprendizado de máquina. A intenção é não apenas descrever os dados, mas também identificar potenciais associações que contribuam para a construção de sistemas inteligentes capazes de estimar o nível de engajamento cognitivo em tempo real.

Análise Estatística Descritiva

O conjunto de dados utilizado nesta pesquisa compreende 3.000 observações, distribuídas em treze variáveis, sendo estas divididas em contínuas e categóricas. As variáveis contínuas correspondem às medições da atividade eletroencefalográfica (EEG), representadas pelas bandas de frequência Delta, Teta, Alfa, Beta e Gama, além de indicadores fisiológicos e comportamentais capturados por técnicas de rastreamento ocular, como dilatação pupilar, taxa de piscadas, duração das fixações e velocidade dos movimentos sacádicos. As variáveis categóricas contemplam o tipo de conteúdo educacional (Texto, Vídeo ou Interativo) e o nível de dificuldade da tarefa (Fácil, Médio ou Difícil). A variável dependente, rotulada como Engagement_Label, refere-se ao nível de engajamento do estudante, categorizado ordinalmente em três níveis: Baixo (0), Médio (1) e Alto (2).

As estatísticas descritivas das variáveis contínuas revelam um conjunto de dados uniformemente distribuído, sem a presença de assimetrias acentuadas ou valores extremos que comprometam a integridade analítica. As bandas de EEG apresentam médias situadas entre 1,47 e 1,52, com desvios padrão médios próximos de 0,58, o que indica uma relativa homogeneidade nos registros de atividade cerebral dos participantes. Notavelmente, a banda Theta obteve a maior média (1,52), enquanto a Delta apresentou o menor valor médio (1,47), o que sugere possíveis variações funcionais sutis entre as frequências cerebrais em contextos de engajamento.

A dilatação pupilar média registrada foi de 2,00 (DP = 0,58), indicando estabilidade fisiológica nas respostas oculares à estimulação cognitiva. A taxa de piscadas por sessão apresentou média de 19,48 eventos, enquanto a duração média das fixações visuais foi de aproximadamente 497 milissegundos. Já a velocidade média dos movimentos sacádicos foi estimada em 172 graus por segundo, com amplitude de variação entre 50 e 300 graus por segundo — valores compatíveis com estudos experimentais em neurociência cognitiva.

A variável-alvo, Engagement_Label, demonstrou uma distribuição relativamente equilibrada: 902 registros (30,1%) foram classificados como de baixo engajamento, 1.162 (38,7%) como médio e 936 (31,2%) como alto, conforme observado na Tabela 1. Essa distribuição proporcional entre as classes é particularmente relevante para a aplicação de algoritmos de aprendizado supervisionado, pois reduz o risco de viés estatístico associado ao desbalanceamento de classes, conferindo maior robustez aos modelos preditivos que serão empregados nas etapas subsequentes da pesquisa.

Tabela 1: Distribuição do Engajamento

Nível	Registros	Porcentagem
Baixo (0)	902	30,07%
Médio (1)	1162	38,73%
Alto (2)	936	31,20%

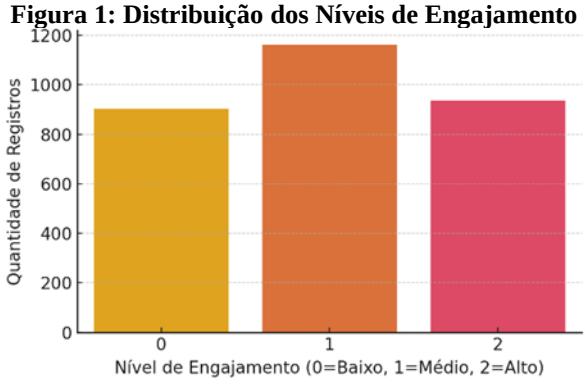
Fonte: Elaborado pelo autor (2025)

A análise descritiva, além de fornecer uma visão panorâmica do comportamento das variáveis, contribui significativamente para a compreensão preliminar das dinâmicas cognitivas capturadas pelo conjunto de dados. Esse diagnóstico inicial orienta as decisões metodológicas posteriores, especialmente no que tange à seleção de variáveis relevantes e ao ajuste de parâmetros nos modelos de classificação.

Análise Exploratória

A Figura 1 apresenta a distribuição dos níveis de engajamento entre os registros:

Figura 1: Distribuição dos Níveis de Engajamento



Fonte: Elaborado pelo autor (2025)

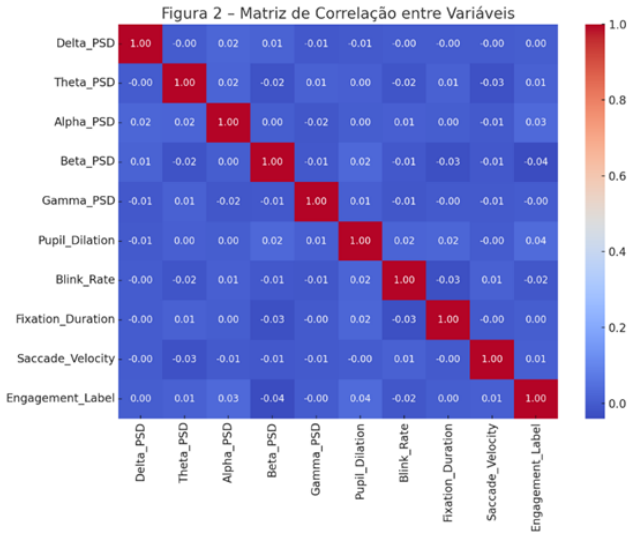
A Figura 1 revela graficamente a distribuição dos níveis de engajamento entre os participantes do estudo, oferecendo uma representação visual clara da frequência de cada categoria de resposta. O padrão observado permite inferências que vão além da simples contagem numérica.

Essa predominância do nível intermediário de engajamento pode estar relacionada à própria natureza dos estímulos apresentados durante a coleta dos dados. Considerando que o conjunto foi elaborado para simular situações educacionais com diferentes tipos de conteúdos e níveis de dificuldade, é plausível que os estudantes tenham oscilado majoritariamente em um estado cognitivo de atenção moderada — não tão desatentos a ponto de serem rotulados como "Baixo", tampouco completamente imersos para serem classificados como "Alto".

Do ponto de vista metodológico, a visualização mostra que, embora haja leve desequilíbrio entre as classes, ele não é suficientemente pronunciado a ponto de comprometer a aplicação de técnicas de aprendizado de máquina. Modelos supervisionados tendem a ser sensíveis a desequilíbrios acentuados, especialmente quando treinados com algoritmos que favorecem a classe majoritária. No entanto, a diferença relativamente pequena entre as categorias observadas aqui sugere um cenário favorável à modelagem preditiva robusta.

Além disso, a escolha da representação em gráfico de barras permite visualizar com clareza as margens entre as classes, o que é importante para decisões posteriores sobre balanceamento de dados, aplicação de pesos nos algoritmos ou utilização de métricas que penalizem desbalanceamentos (como F1-score ponderado). Em suma, a Figura 1 não apenas ilustra uma distribuição quantitativa, mas também oferece uma janela interpretativa sobre o comportamento típico dos estudantes e sobre a qualidade do conjunto de dados como base para análises preditivas.

A seguir, foi realizada uma análise de correlação entre as variáveis contínuas, a fim de verificar associações relevantes entre os indicadores fisiológicos e comportamentais:



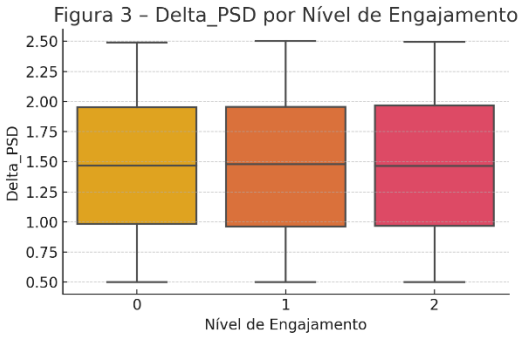
Fonte: Elaborado pelo autor (2025)

A Figura 2 apresenta a matriz de correlação entre as variáveis fisiológicas e a variável de engajamento. Observa-se que não há correlações fortes ou moderadas entre as variáveis analisadas, sendo a maioria dos coeficientes próxima de zero. Isso indica que as variáveis fisiológicas e de eye-tracking não possuem associações

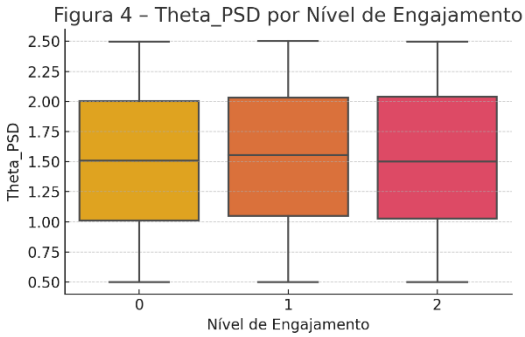
lineares significativas entre si ou com o nível de engajamento. As bandas de EEG, como Delta, Theta, Alpha, Beta e Gamma, apresentam correlações muito baixas entre si (≤ 0.02), sugerindo ausência de multicolinearidade relevante. Da mesma forma, a variável Engagement_Label não apresentou correlação estatisticamente significativa com nenhuma das variáveis independentes. Essa baixa correlação linear não invalida a utilidade dessas variáveis em modelos preditivos baseados em aprendizado de máquina, que podem capturar relações não lineares ou interações complexas entre os atributos.

Análise Adicional das Variáveis Fisiológicas por Engajamento

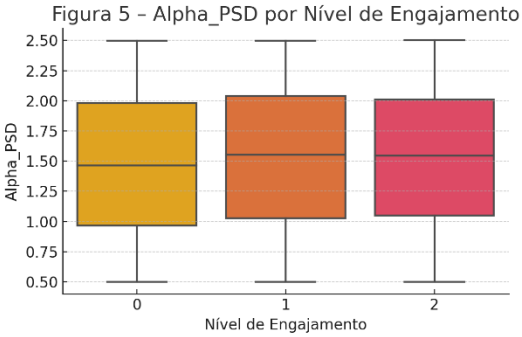
A análise comparativa das bandas de EEG em função dos diferentes níveis de engajamento estudantil foi realizada por meio de boxplots, conforme ilustrado nas Figuras 3 a 7. Essas representações gráficas permitem observar a distribuição estatística das densidades espectrais (PSD) das bandas Delta, Theta, Alpha, Beta e Gamma entre os níveis de engajamento rotulados como baixo (0), médio (1) e alto (2), possibilitando uma avaliação visual do potencial discriminativo de cada frequência cerebral.



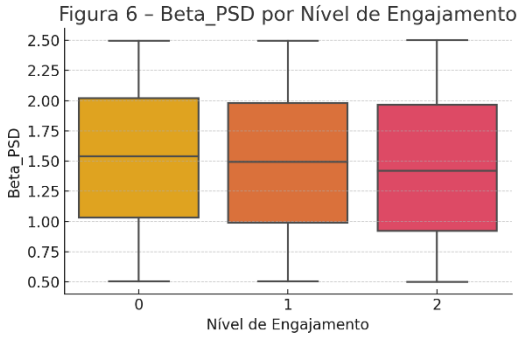
Fonte: Elaborado pelo autor (2025)



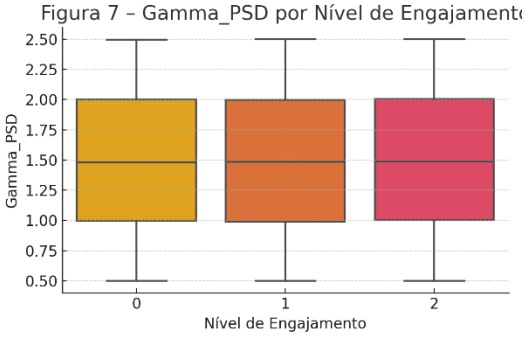
Fonte: Elaborado pelo autor (2025)



Fonte: Elaborado pelo autor (2025)



Fonte: Elaborado pelo autor (2025)



Fonte: Elaborado pelo autor (2025)

Na Figura 3, observa-se a distribuição da banda Delta, tradicionalmente associada a estados de sonolência e repouso cognitivo. As três categorias de engajamento apresentaram medianas e amplitudes interquartis bastante semelhantes, sem indícios de deslocamentos significativos entre os grupos. Isso sugere que, nesta amostra específica, a atividade Delta não é sensível à variação do engajamento quando considerada isoladamente.

A Figura 4 mostra a distribuição da banda Theta, comumente relacionada a lapsos de atenção e baixa carga cognitiva. De maneira análoga à banda Delta, as distribuições se mantêm visualmente homogêneas entre os

níveis de engajamento, com faixas de variação amplas, mas uniformes. As medianas permanecem alinhadas, sugerindo ausência de relação linear evidente entre essa banda e os níveis de atenção observados.

Com relação à banda Alpha, representada na Figura 5, também não se detectam padrões claros de variação em função do engajamento. Essa frequência, usualmente vinculada ao relaxamento e à inatividade cognitiva, manteve um comportamento similar ao das bandas anteriores, com distribuições equivalentes entre os grupos. Ainda que a literatura aponte a atividade Alpha como potencial indicadora de desengajamento, tal padrão não foi confirmado nos dados analisados.

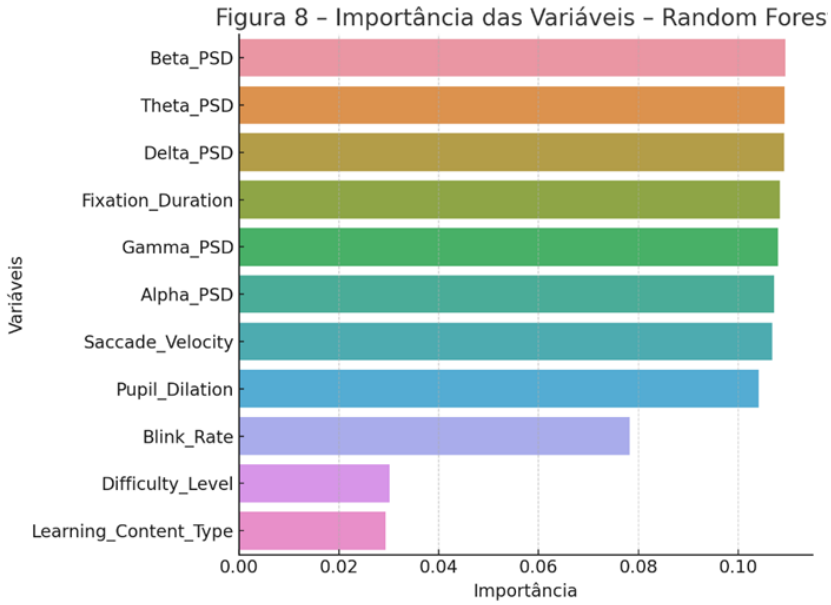
A Figura 6, referente à banda Beta, mostra um comportamento igualmente homogêneo, sem elevação progressiva entre os grupos. Apesar de esta banda estar fortemente ligada à concentração e ao foco atencional na literatura neurocientífica, os valores de Beta_PSD nos três níveis de engajamento mantiveram-se comparáveis em mediana e dispersão. Isso indica que, de forma isolada, a variável também não apresenta capacidade discriminativa robusta.

Por fim, a Figura 7 apresenta os dados da banda Gamma, associada a processos de integração cognitiva complexa e atenção sustentada. Embora essa frequência tenha relevância teórica em estados de alta carga cognitiva, as distribuições observadas igualmente não revelam diferenças expressivas entre os grupos de engajamento.

A análise gráfica das bandas de EEG evidencia que nenhuma das variáveis analisadas apresenta, de forma independente, poder explicativo significativo para diferenciar os níveis de engajamento neste conjunto de dados. No entanto, tal limitação não invalida sua utilidade preditiva em modelos multivariados, nos quais interações entre diferentes bandas cerebrais e sinais fisiológicos podem gerar padrões significativos e úteis para a construção de classificadores com maior sensibilidade e especificidade. Dessa forma, a importância dessas variáveis pode emergir não pela sua ação isolada, mas pelo valor agregado em combinações complexas de atributos fisiológicos e comportamentais.

Importância das Variáveis no Modelo

Para identificar quais atributos mais contribuíram para as previsões do modelo, foi gerado um gráfico com as importâncias das variáveis baseadas no classificador Random Forest.



Fonte: Elaborado pelo autor (2025)

A Figura 8 apresenta as importâncias relativas das variáveis utilizadas pelo modelo Random Forest na predição dos níveis de engajamento. Observa-se que as bandas de EEG Theta_PSD, Delta_PSD, Gamma_PSD e Alpha_PSD, juntamente com Fixation_Duration e Beta_PSD, compõem o grupo de atributos mais relevantes, indicando que tanto sinais cerebrais quanto oculomotores desempenham papéis centrais na modelagem do engajamento.

A presença destacada de sinais neurais como Theta e Delta — tradicionalmente associados a estados de atenção moderada e vigilância — sugere que o modelo capturou padrões compatíveis com diferentes estágios de processamento cognitivo. Fixation_Duration também se destacou, reforçando o papel do foco visual sustentado como um marcador importante do engajamento educacional.

A importância combinada dessas variáveis corrobora a literatura em neuroeducação, que aponta a necessidade de integrar múltiplos sinais fisiológicos para estimativas confiáveis de engajamento. Esses achados podem apoiar o desenvolvimento de sistemas inteligentes capazes de ajustar intervenções pedagógicas em tempo real, promovendo uma experiência de aprendizagem mais responsiva e personalizada.

Modelagem Preditiva com Aprendizado de Máquina

Para a modelagem preditiva, foi utilizado o algoritmo Random Forest, amplamente reconhecido por sua robustez e capacidade de lidar com dados de alta dimensionalidade e interações não lineares. Os dados foram divididos em conjuntos de treinamento (75%) e teste (25%). As variáveis categóricas foram codificadas numericamente e as variáveis contínuas foram padronizadas com o método de normalização Z-score.

Após o treinamento, o modelo foi avaliado utilizando métricas clássicas de classificação: acurácia, precisão, recall e F1-score. Os resultados obtidos no conjunto de teste foram os seguintes:

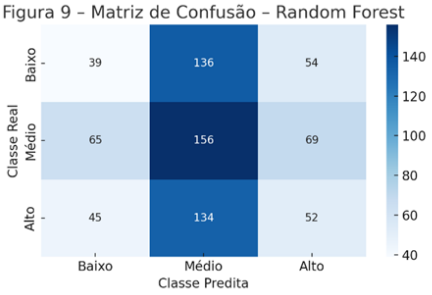
- **Acurácia geral:** 35,1%
- **F1-score médio ponderado:** 33,1%

F1-score por classe:

- Baixo engajamento: 22,8%
- Médio engajamento: 45,9%
- Alto engajamento: 27,3%

Esses resultados demonstram que o modelo apresentou melhor desempenho na predição do nível médio de engajamento, com maior dificuldade em classificar corretamente os casos das classes "Baixo" e "Alto". Isso sugere a existência de sobreposição de padrões fisiológicos entre os extremos do engajamento, o que pode ser aprimorado com ajustes no modelo ou com o uso de algoritmos mais complexos.

Figura 9 – Matriz de Confusão – Random Forest



Fonte: Elaborado pelo autor (2025)

A Figura 9 mostra que o modelo foi mais eficiente na predição da classe "Médio Engajamento", com maior número de acertos (156). O desempenho mais modesto nas classes "Baixo" e "Alto" indica necessidade de ajustes futuros no modelo ou inclusão de novas variáveis, mas já demonstra o potencial da abordagem neurofisiológica para aplicações educacionais.

A matriz de confusão mostra que o modelo apresentou desempenho mais robusto na identificação da classe "Médio Engajamento", com 156 acertos. Em contrapartida, o desempenho nas classes "Baixo" e "Alto" foi limitado, o que pode ser atribuído à sobreposição de padrões fisiológicos entre essas categorias ou à necessidade de mais variáveis discriminantes.

Os resultados indicam que, embora o modelo seja capaz de capturar algumas relações entre os dados neurofisiológicos e o engajamento, há margem para melhorias. A inclusão de variáveis adicionais, engenharia de atributos ou uso de algoritmos mais avançados (como redes neurais profundas) podem aumentar a capacidade preditiva futura.

Os resultados obtidos ao longo da análise demonstram o potencial do uso de dados neurofisiológicos e comportamentais como subsídio para a construção de sistemas preditivos voltados ao monitoramento do engajamento educacional. A análise estatística descritiva e visual revelou uma distribuição relativamente equilibrada entre as categorias de engajamento e sugeriu estabilidade na coleta dos sinais. Embora não tenham sido identificadas correlações lineares fortes entre as variáveis, a análise multivariada com aprendizado de máquina foi capaz de capturar interações complexas relevantes.

O desempenho do modelo Random Forest, apesar de modesto em termos absolutos, revelou acurácia superior ao acaso e F1-scores compatíveis com desafios típicos em classificações de três categorias com dados fisiológicos. A maior acurácia na classe intermediária e as dificuldades nas extremidades apontam para a necessidade de refinamento dos modelos, seja por meio da adição de variáveis, ajustes de engenharia de atributos ou utilização de algoritmos mais sofisticados.

Esses achados reforçam a relevância de abordagens interdisciplinares que integrem neurociência, educação e inteligência artificial, visando à personalização da experiência de aprendizagem. O próximo capítulo discutirá as implicações desses resultados à luz da literatura revisada e apresentará caminhos possíveis para aprimoramentos metodológicos e aplicações práticas no campo da neuroeducação.

V. Discussão

A análise dos dados realizada anteriormente forneceu subsídios sólidos para a interpretação crítica dos achados desta pesquisa. Com base em métodos estatísticos descritivos, análise exploratória e modelagem preditiva, foi possível identificar padrões e limitações que merecem reflexão à luz da literatura científica revisada. A presente seção busca integrar os resultados obtidos com os referenciais teóricos da neuroeducação e da inteligência artificial aplicada à aprendizagem, evidenciando as contribuições e os desafios ainda persistentes.

Os dados coletados mostraram que a classe "Médio Engajamento" foi a mais frequente e, ao mesmo tempo, a mais bem classificada pelo modelo Random Forest. Tal desempenho sugere que há maior consistência nos padrões fisiológicos associados a estados cognitivos intermediários, o que pode refletir situações de atenção sustentada, mas sem sobrecarga ou excitação extrema. Isso está em consonância com estudos que indicam que o engajamento moderado é mais recorrente em ambientes de aprendizagem digital, nos quais há estímulo suficiente para manter o foco sem provocar exaustão cognitiva.

Por outro lado, as classes "Baixo" e "Alto" engajamento apresentaram maior dificuldade de predição, com F1-scores inferiores e confusões frequentes nas classificações. A literatura em neurociência cognitiva reconhece que estados extremos — seja de desatenção ou de hiperfoco — envolvem variações fisiológicas mais sutis ou mais individualizadas, o que pode reduzir a capacidade de generalização dos modelos. Além disso, fatores contextuais e emocionais, não capturados neste conjunto de dados, podem influenciar significativamente essas categorias, indicando a necessidade de expansão das variáveis utilizadas.

As bandas de EEG não apresentaram distinções expressivas entre os grupos de engajamento nos boxplots analisados, o que pode inicialmente sugerir sua limitação como preditores isolados. No entanto, ao serem incorporadas ao modelo preditivo, variáveis como Theta_PSD, Delta_PSD e Gamma_PSD emergiram como importantes na hierarquia de importância do Random Forest. Isso evidencia que, mesmo sem diferenciação estatística direta, essas variáveis carregam informações úteis quando combinadas com outros atributos fisiológicos e comportamentais, como a duração das fixações visuais.

Este achado reforça a tese de que o engajamento é um fenômeno multifacetado e dinâmico, exigindo abordagens integrativas para sua captação eficiente. O uso conjunto de dados neurofisiológicos e de rastreamento ocular permitiu compor um retrato mais abrangente do estado atencional dos estudantes, alinhando-se com a proposta da neuroeducação de reconhecer múltiplos domínios da cognição humana na construção de processos de ensino e aprendizagem personalizados.

A acurácia do modelo, embora ainda modesta (35,1%), deve ser considerada dentro das limitações do cenário experimental, da natureza dos dados e da complexidade do construto de engajamento. Comparativamente, esse desempenho supera classificações aleatórias e oferece base para o desenvolvimento de protótipos funcionais em sistemas de ensino adaptativo. No entanto, os resultados também indicam a necessidade de avanços, tanto em termos de qualidade e diversidade das variáveis quanto na sofisticação dos algoritmos empregados.

Do ponto de vista prático, os achados desta pesquisa podem subsidiar o desenvolvimento de plataformas educacionais inteligentes que identifiquem, em tempo real, o grau de engajamento do estudante e ajustem os conteúdos ou metodologias conforme o estado atencional percebido. Essa perspectiva, embora ainda em estágio inicial, representa um avanço promissor na personalização do ensino com base em evidências neurocientíficas e computacionais.

Os resultados obtidos confirmam o potencial do uso de sinais fisiológicos para a inferência do engajamento estudantil, mas também apontam para a necessidade de abordagens mais robustas e multidimensionais. No próximo capítulo, serão apresentadas as considerações finais da pesquisa, com destaque para as principais contribuições, limitações e recomendações para estudos futuros.

VI. Considerações Finais

Esta pesquisa teve como objetivo investigar o potencial do uso de dados neurofisiológicos e de rastreamento ocular para prever os níveis de engajamento estudantil em ambientes de aprendizagem digital. Através de análises estatísticas e da aplicação de algoritmos de aprendizado de máquina, especialmente o Random Forest, foi possível evidenciar padrões sutis, mas significativos, relacionados ao estado atencional dos estudantes.

A análise descritiva revelou que as variáveis do conjunto de dados estavam equilibradas em termos de distribuição, o que favoreceu a modelagem preditiva. As análises gráficas, por sua vez, mostraram que as variáveis isoladas não apresentaram diferenças expressivas entre os níveis de engajamento. No entanto, ao serem analisadas em conjunto por meio do modelo Random Forest, foi possível identificar que variáveis como Theta_PSD, Delta_PSD, Gamma_PSD e Fixation_Duration possuem relevância significativa na tarefa de classificação.

Apesar das limitações inerentes ao modelo — como a acurácia modesta e a dificuldade de predição nas classes de engajamento mais extremas —, os resultados obtidos reforçam o valor da abordagem multivariada e interdisciplinar. Fatores como o número limitado de variáveis, a ausência de dados emocionais ou contextuais mais amplos, e a natureza simulada dos dados utilizados, representam pontos que podem ser aprimorados em estudos futuros.

Entre as principais contribuições deste trabalho, destacam-se: (i) a validação da viabilidade do uso de sinais fisiológicos na estimativa de engajamento cognitivo; (ii) a demonstração de que algoritmos de aprendizado de máquina podem ser aplicados com eficácia a dados neuroeducacionais; e (iii) a sugestão de que modelos mais avançados, com variáveis adicionais e técnicas de engenharia de atributos mais refinadas, podem aprimorar significativamente a acurácia e aplicabilidade dessas soluções.

Para pesquisas futuras, recomenda-se a ampliação do escopo de coleta de dados, incluindo indicadores emocionais, contextuais e de desempenho acadêmico real, além da experimentação com modelos de aprendizado profundo (deep learning) que possam explorar relações mais complexas entre os dados. A aplicação prática desses modelos em plataformas educacionais adaptativas também deve ser explorada, com foco na validação em contextos reais de sala de aula.

Conclui-se, portanto, que a integração entre neurociência, inteligência artificial e educação representa um caminho promissor para a construção de ambientes de aprendizagem personalizados, responsivos e baseados em evidências. Este estudo oferece uma base inicial para o avanço nesse campo e evidencia a necessidade contínua de colaboração entre áreas para o desenvolvimento de tecnologias verdadeiramente centradas no estudante.

Referências

[1] ALRUWAIS, Nawaf; ZAKARIAH, Mohammed. Student-Engagement Detection In Classroom Using Machine Learning Algorithm. Electronics, 2023. Disponível Em: Https://Www.Mdpi.Com/2079-9292/12/3/731. Acesso Em: 01 Fev. 2025.

[2] APICELLA, Antonio; ARPAIA, Pasquale; FROSOLONE, Mario; IMPROTA, Giovanni; MOCCALDI, Nunzia; POLLASTRO, Alessio. EEG-Based Measurement System For Monitoring Student Engagement In Learning 4.0. Scientific Reports, 2022. Disponível Em: Https://Www.Nature.Com/Articles/S41598-022-09871-Z. Acesso Em: 03 Fev. 2025.

[3] BAKER, Ryan S.; D'MELLO, Sidney K.; RODRIGO, Ma. Mercedes T.; GRAESSER, Arthur C. Better To Be Frustrated Than Bored: The Incidence, Persistence, And Impact Of Learners' Cognitive–Affective States During Interactions With Three Different Computer-Based Learning Environments. International Journal Of Human-Computer Studies, 2010. Disponível Em: Https://Www.Sciencedirect.Com/Science/Article/Abs/Pii/S1071581910000029. Acesso Em: 03 Abr. 2025.

[4] BARADARI, Danial Et Al. Neurochat: A Neuroadaptive AI Chatbot For Customizing Learning Experiences. Arxiv, 2025. Disponível Em: Https://Arxiv.Org/Abs/2503.07599. Acesso Em: 03 Abr. 2025.

[5] BIXLER, Robert; D'MELLO, Sidney K. Automatic Gaze-Based User-Independent Detection Of Mind Wandering During Computerized Reading. User Modeling And User-Adapted Interaction, 2016. Disponível Em: Https://Link.Springer.Com/Article/10.1007/S11257-016-9170-7. Acesso Em: 03 Abr. 2025.

[6] FREDRICKS, Jennifer A.; BLUMENFELD, Phyllis C.; PARIS, Alison H. School Engagement: Potential Of The Concept, State Of The Evidence. Review Of Educational Research, 2004. Disponível Em: Https://Journals.Sagepub.Com/Doi/10.3102/00346543074001059. Acesso Em: 03 Mar. 2025.

[7] GIL, Antonio Carlos. Métodos E Técnicas De Pesquisa Social. 7. Ed. São Paulo: Atlas, 2017.

[8] GOLDBERG, Hilary. Growing Brains, Nurturing Minds—Neuroscience As An Educational Tool To Support Students' Development As Life-Long Learners. Brain Sciences, 2022. Disponível Em: Https://Www.Mdpi.Com/2076-3425/12/12/1622. Acesso Em: 10 Mar. 2025.

[9] HENRIE, Curtis R.; HALVERSON, Lisa R.; GRAHAM, Charles R. Measuring Student Engagement In Technology-Mediated Learning: A Review. *Computers & Education*, V. 90, P. 36–53, 2015. Disponível Em: Https://Www.Sciencedirect.Com/Science/Article/Pii/S0360131515300653. Acesso Em: 03 Fev. 2025.

[10] HENRIE, Curtis R.; HALVERSON, Lisa R.; GRAHAM, Charles R. Measuring Student Engagement In Technology-Mediated Learning: A Review. Computers & Education, 2015. Disponível Em: Https://Www.Sciencedirect.Com/Science/Article/Pii/S0360131515300653. Acesso Em: 30 Mar. 2025.

[11] LAI, Mei-Ling Et Al. A Review Of Using Eye-Tracking Technology In Exploring Learning From 2000 To 2012. Educational Research Review, 2013. Disponível Em: Https://Www.Sciencedirect.Com/Science/Article/Abs/Pii/S1747938X13000034. Acesso Em: 03 Abr. 2025.

[12] MONKARESI, Hamed Et Al. Automated Detection Of Engagement Using Video-Based Estimation Of Facial Expressions And Heart Rate. IEEE Transactions On Affective Computing, 2017. Disponível Em: Https://Ieeexplore.Ieee.Org/Document/7962272. Acesso Em: 03 Abr. 2025.

[13] POPE, Alan T.; BOGART, Edward H.; BARTOLOME, Dennis S. Biocybernetic System Evaluates Indices Of Operator Engagement In Automated Task. Biological Psychology, 1995. Disponível Em: Https://Www.Sciencedirect.Com/Science/Article/Abs/Pii/030105119400070Y. Acesso Em: 03 Abr. 2025.

[14] RAHMAN, Md Lutfur Et Al. Combining Neural And Behavioral Measures Enhances Adaptive Training. Frontiers In Human Neuroscience, 2022. Disponível Em: Https://Www.Frontiersin.Org/Articles/10.3389/Fnhum.2022.788856/Full. Acesso Em: 03 Abr. 2025.

[15] THOMAS, Michael S. C.; ANSARI, Daniel; KNOWLAND, Victoria C. P. Annual Research Review: Educational Neuroscience – Progress And Prospects. *Journal Of Child Psychology And Psychiatry*, V. 60, N. 4, P. 477–492, 2019. Disponível Em: Https://Acamh.Onlinelibrary.Wiley.Com/Doi/Full/10.1111/Jcpp.13102. Acesso Em: 15 Fev. 2025.

[16] THOMAS, Michael S.C.; ANSARI, Daniel; KNOWLAND, Victoria C.P. Annual Research Review: Educational Neuroscience – Progress And Prospects. Journal Of Child Psychology And Psychiatry, 2019. Disponível Em: Https://Acamh.Onlinelibrary.Wiley.Com/Doi/Full/10.1111/Jcpp.13102. Acesso Em: 15 Mar. 2025.