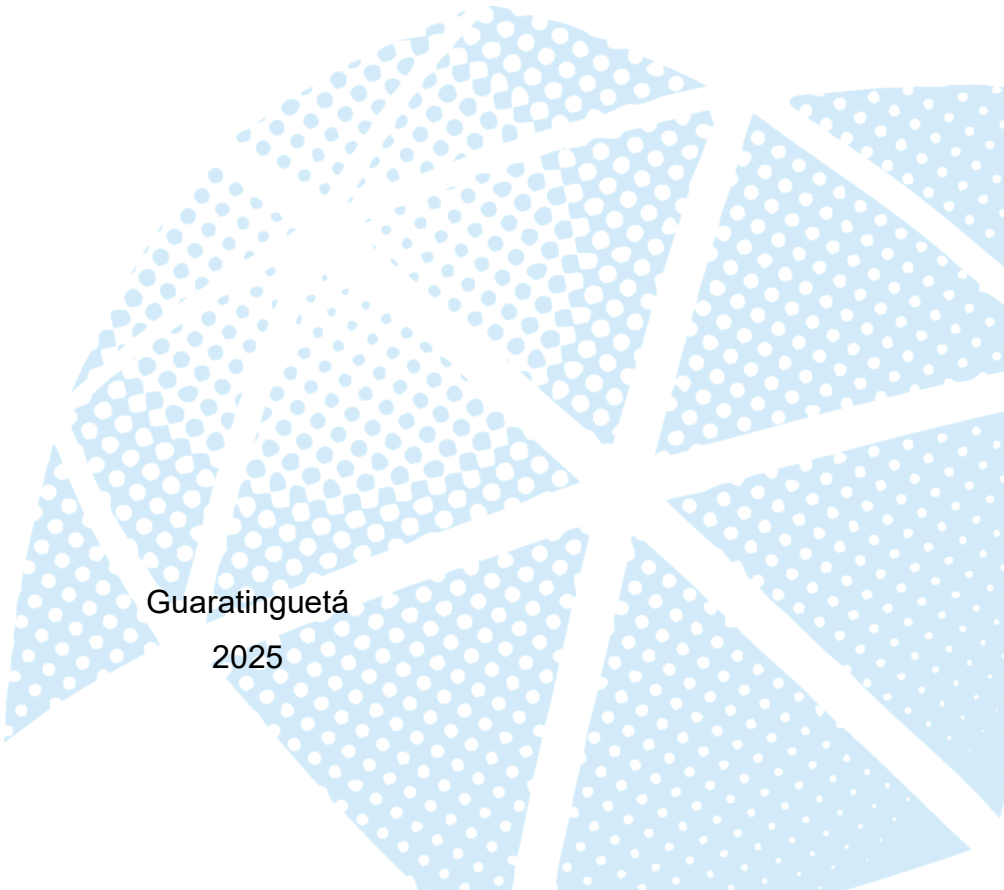
UNIVERSIDADE ESTADUAL PAULISTA – UNESP

Faculdade de Engenharia e Ciências - Campus de Guaratinguetá

Lucas Giacon Giusti

APRENDIZADO DE MÁQUINA APLICADO A SINAIS FISIOLÓGICOS PARA ANÁLISE DE ENGAJAMENTO ESTUDANTIL



Guaratinguetá

2025

Lucas Giacon Giusti

APRENDIZADO DE MÁQUINA APLICADO A SINAIS FISIOLÓGICOS PARA ANÁLISE DE ENGAJAMENTO ESTUDANTIL

Trabalho de Graduação apresentado ao Conselho de Curso de Graduação em Engenharia de Produção Mecânica da Faculdade de Engenharia e Ciências, Universidade Estadual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia de Produção Mecânica. Orientador:

Prof. Claudia Regina de Freitas

Guaratinguetá

2025

Giusti, Lucas Giacon
G538a Aprendizado de máquina aplicado a sinais fisiológicos para análise de engajamento estudantil / Lucas Giacon Giusti - Guaratinguetá, 2025.
52 f : il.
Bibliografia: f. 50-52

Trabalho de Graduação em Engenharia de Produção Mecânica – Universidade Estadual Paulista, Faculdade de Engenharia e Ciências de Guaratinguetá, 2025.

Orientadora: Profª. Drª Cláudia Regina de Freitas

1. Inteligência artificial - Aplicações educacionais. 2. Condutividade elétrica. 3. Aprendizado do computador. 4. Programas de aprendizado. I. Título.

CDU 007.52

Luciana Máximo
Bibliotecária/CRB-8 3595

IMPACTO POTENCIAL DESTA PESQUISA

O impacto esperado na sociedade deve ser redigido de forma sucinta considerando, os seguintes aspectos: o potencial científico, técnico, social, inovador, econômico, educacional e cultural; a internacionalização; a inserção local, regional e nacional; o desenvolvimento sustentável, conhecimento e temática da pesquisa.

POTENTIAL IMPACT OF THIS RESEARCH

The expected impact on society should be written succinctly, considering the following aspects: the scientific, technical, social, innovative, economic, educational and cultural potential; internationalization; local, regional and national insertion; sustainable development, knowledge and research themes.

LUCAS GIACON GIUSTI

APRENDIZADO DE MÁQUINA APLICADO A SINAIS FISIOLÓGICOS PARA ANÁLISE DE ENGAJAMENTO ESTUDANTIL

Trabalho de graduação apresentado à Universidade Estadual Paulista (UNESP), Faculdade de Engenharia e Ciências, Guaratinguetá, para obtenção do título de Grau acadêmico Bacharel em Engenharia de Produção Mecânica.

Data da defesa: 12/11/2025

Banca Examinadora:

Documento assinado digitalmente
gov.br
CLAUDIA REGINA DE FREITAS
Data: 14/11/2025 15:07:22-0300
Verifique em https://validar.iti.gov.br

_____

Profa. Dra. Claudia Regina de Freitas

UNESP - Faculdade de Engenharia e Ciências - Campus de Guaratinguetá

Documento assinado digitalmente
gov.br
JOSE ROBERTO DALE LUCHE
Data: 12/11/2025 19:53:06-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Jose Roberto Dale Luche

UNESP - Faculdade de Engenharia e Ciências - Campus de Guaratinguetá

Documento assinado digitalmente
gov.br
MESSIAS BORGES SILVA
Data: 14/11/2025 13:11:51-0300
Verifique em https://validar.iti.gov.br

_____

Prof. Dr. Messias Borges Silva

UNESP - Faculdade de Engenharia e Ciências - Campus de Guaratinguetá

“As universidades serão o que são suas bibliotecas”
(Gelfand, 1968, p. 19, tradução nossa).

RESUMO

A avalição de engajamento estudantil é um grande desafio, principalmente por ser bastante subjetiva, baseada em observações e questionários, trazendo assim um viés para essas análises. Este trabalho propõe uma análise objetiva para entender o engajamento estudantil, por meio de algoritmos de aprendizado de máquina, analisando dados fisiológicos: atividade eletrodérmica (EDA), frequência cardíaca (HR) e temperatura da pele (ST). A base de dados foi obtida a partir de um experimento com cinquenta estudantes de engenharias, submetidos a metodologias de ensino ativa e tradicional. Após o pré-processamento e a extração de variáveis derivadas (SCL, NS.SCRs, EDASymp e TVSymp), aplicaram-se três algoritmos supervisionados: Redes Neurais, Random Forest e Máquinas de Vetores de Suporte (SVM). Os modelos foram avaliados por matriz de confusão e métricas de acurácia. Os resultados mostraram que o SVM apresentou o melhor desempenho, com 98,15% de acurácia, seguido pelas Redes Neurais (95,33%) e Random Forest (89,04%). Conclui-se que o uso de técnicas de aprendizado de máquina combinadas a dados fisiológicos permite uma análise mais objetiva e precisa do engajamento estudantil, contribuindo para o desenvolvimento de ambientes educacionais inteligentes e personalizados.

Palavras-chave: aprendizado de máquina; engajamento estudantil; sinais fisiológicos; análise de dados; SVM; redes neurais; floresta randômica.

ABSTRACT

Assessing student engagement is a major challenge, mainly because it is highly subjective, based on observations and questionnaires, which introduces bias into such analyses. This study proposes an objective approach to understanding student engagement through machine learning algorithms by analyzing physiological data: electrodermal activity (EDA), heart rate (HR), and skin temperature (ST). The dataset was obtained from an experiment with fifty engineering students who were exposed to active and traditional teaching methodologies. After preprocessing and extracting derived variables (SCL, NS.SCRs, EDASymp, and TVSymp), three supervised algorithms were applied: Neural Networks, Random Forest, and Support Vector Machines (SVM). The models were evaluated using confusion matrices and accuracy metrics. The results showed that SVM achieved the best performance, with 98.15% accuracy, followed by Neural Networks (95.33%) and Random Forest (89.04%). It is concluded that the use of machine learning techniques combined with physiological data enables a more objective and precise analysis of student engagement, contributing to the development of intelligent and personalized educational environments.

Keywords: machine learning; student engagement; physiological signals; data analysis; SVM; neural networks; random forest.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A avaliação do engajamento dos alunos durante as aulas é frequentemente subjetiva e depende de métodos tradicionais como observação direta ou questionários, que podem não capturar a dinâmica completa da experiência dos estudantes. Essa lacuna evidencia a necessidade de abordagens mais robustas e baseadas em dados para uma análise mais objetiva.

Buscando mais robustez e objetividade para entender o engajamento de um indivíduo, iniciaram-se investigações sobre dados fisiológicos que refletem a atividade do Sistema Nervoso Autônomo. Sinais como a temperatura da pele (ST), frequência cardíaca (HR) e atividade eletrodérmica (EDA) emergem como potenciais indicadores objetivos das respostas psicofisiológicas associadas aos diferentes níveis de engajamento. O uso dessas variáveis combinadas com algoritmos de *Machine Learning* permite identificar diferentes tarefas cognitivas com base nas reações autonômicas dos indivíduos. (POSADA-QUINTERO; BOLKHOVSKY, 2019).

A implementação dessas técnicas em ambientes educacionais pode fornecer uma análise detalhada do engajamento e estados emocionais dos alunos durante atividades de aprendizagem, oferecendo uma nova dimensão para a personalização das metodologias de ensino.

Este estudo visa analisar os dados de um experimento realizado com estudantes de engenharias (LUCHE *et al.*, 2025). A base de dados do experimento contempla dados de HR, ST e a EDA. Além de dados subjetivos, como o engajamento do aluno e informações provenientes de observações do pesquisador durante o experimento.

As métricas de HR e EDA são objetivas e fornecem informações sobre o estado autonômico do sistema nervoso. Estudos como o de Sharma, Prakas e Kalra (2019) têm mostrado que essas medidas são indicativas de engajamento cognitivo e emocional. Por exemplo, índices como a condutância da pele (SCL), número de respostas de condutância não específica da pele (NS.SCRs), índice de simetria de EDA (EDASymp), componente de baixa frequência da variabilidade da frequência cardíaca (HRVLF) e componente normalizado de alta frequência da variabilidade da frequência cardíaca (HRVHFn) podem ser correlacionados com diferentes níveis de estresse, atenção e engajamento (POSADA-QUINTERO; BOLKHOVSKY, 2019). A análise, desses índices por meio de técnicas de *Machine Learning* pode revelar padrões complexos de comportamento e interação dos alunos.

Por meio da análise das métricas fisiológicas, este estudo utiliza abordagens de classificação de *Machine Learning*, como redes neurais, máquina de vetores de suporte (SVM) e *decision trees*, para identificar e prever estados emocionais e de engajamento dos alunos. A validação cruzada como matriz de confusão será empregada para garantir que os modelos desenvolvidos sejam robustos e generalizáveis.

1.1 OBJETIVO

1.1.1 Objetivo geral

O objetivo geral deste trabalho é desenvolver e implementar algoritmos de *Machine Learning*, a para analisar dados fisiológicos de EDA, HR e ST. A proposta consiste em utilizar essas informações para identificar padrões de engajamento e inferir estados emocionais dos alunos durante atividades de aprendizagem.

Com a finalidade de compreender como as respostas fisiológicas se relacionam com o nível de atenção, motivação e envolvimento dos estudantes, fornecendo indicadores mais objetivos e complementares às avaliações tradicionais de desempenho.

1.1.2 Objetivos específicos

- Pré-processar os dados fisiológicos coletado. (LUCHE *et al.*, 2025)
- Identificar a técnica de classificação mais eficaz para prever o comportamento por meio da análise das métricas fisiológicas: este objetivo visa determinar dentre as técnicas de aprendizado de máquinas selecionadas: redes neurais, SVM e *decision trees*, qual é a mais eficaz para prever o engajamento dos alunos durante as aulas. A eficácia será avaliada com base na precisão, sensibilidade, especificidade e praticidade dos modelos.
- Correlacionar os resultados dos modelos com os níveis de engajamento

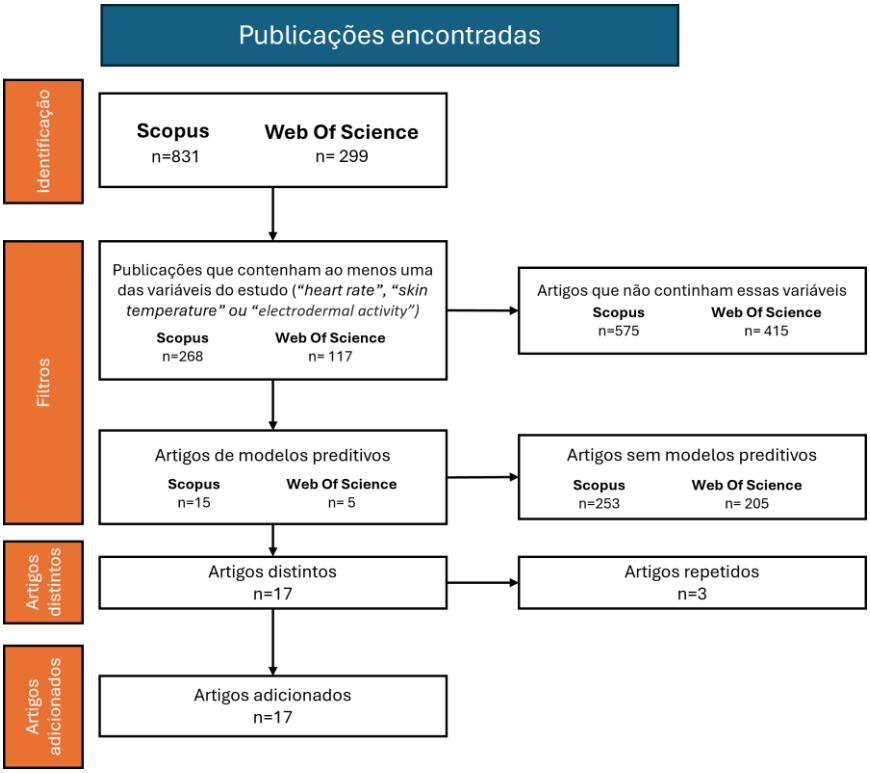
1.2 JUSTIFICATIVAS

A análise de dados fisiológicos, como ST, HR e EDA (e suas métricas derivadas que indicam atividade simpática), pode proporcionar uma compreensão aprofundada

das respostas emocionais e cognitivas dos alunos durante atividades de aprendizagem (POSADA-QUINTERO; BOLKHOVSKY, 2019). O uso de algoritmos de *Machine Learning* para interpretar esses dados pode ajudar a identificar padrões e fornecer percepções importantes para personalizar e melhorar as metodologias de ensino. A integração de medidas fisiológicas com métodos tradicionais de avaliação pode oferecer uma compreensão mais completa da experiência do usuário, ajudando a identificar pontos fortes e áreas de melhoria nas metodologias de ensino.

Para validar a importância do estudo no mundo acadêmico, realizou-se uma análise bibliográfica nas bases de dados do *Scopus e do Web of Science* buscando inicialmente os artigos que continham palavras-chave, resumos e títulos com os termos: '*Machine Learning' e* '*physiological data*', a partir de 2016 até 2024. A partir dessa pesquisa foi realizada uma revisão estruturada da literatura, cujo processo de seleção e filtragem dos artigos está ilustrado na Figura 1.

Figura 1 – Estrutura *prisma sistematic review* da revisão bibliométrica.



Fonte: Próprio autor (2025).

O processo partiu de uma amostra inicial de 831 artigos da base de dados *Scopus* e 299 da base de dados da *Web of Science*, obtida a partir da seguinte linha

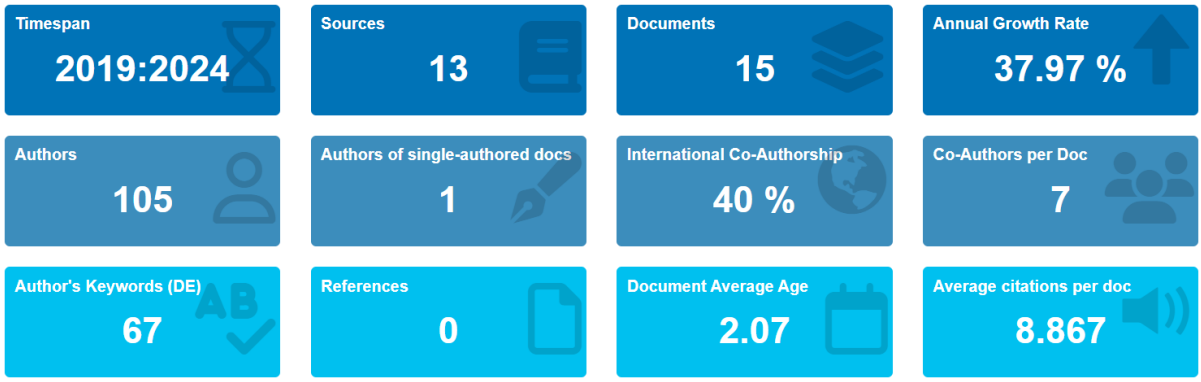
de busca: TITLE-ABS-KEY ("*machine learning*") and TITLE-ABS-KEY ("physiological data") AND (TITLE-ABS-KEY ("heart rate") or TITLE-ABS-KEY ("electrodermal activity") or TITLE-ABS-KEY ("skin temperature")) AND PUBYEAR > 2016 AND PUBYEAR < 2025.Em seguida, foram aplicados filtros para aproximar a revisão ao tema estudado,

resultando em um banco com 268 artigos na *Scopus* e 117 publicações na *Web of Science.* Adicionou-se mais uma palavra chave ao filtro, o termo *"predictive model",* reduzindo para 15 documentos na *Scopus* e 11 na *Web of Science.*

Após aplicar todos os filtros e retirar publicações duplicadas entre as bases de dados, a amostra final foi de 17 artigos. Estes trabalhos constituíram a base de referência para a análise, discussão e desenvolvimento da presente pesquisa.

Para a análise bibliométrica, foi utilizada a linguagem R em conjunto com pacotes computacionais de bibliometria (bibliometrix). A Figura 2 apresenta, de forma macro, os totais das informações obtidas com a bibliométria.

Figura 2 - Resultado da análise bibliométrica

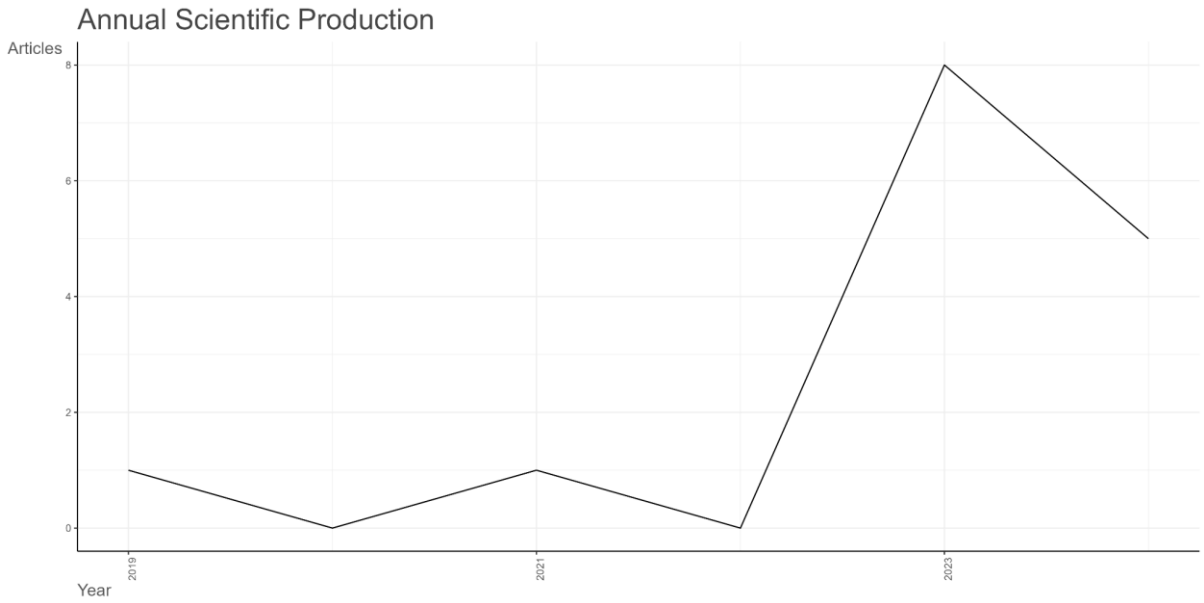


Fonte: Próprio autor (2025).

Nota-se uma tendência de crescimento significativa no número de publicações sobre o uso de *Machine Learning* com os dados fisiológicos, crescimento anual de 37,97%. Foram 13 documentos sobre o tema, com 105 autores distintos, com uma média de 8,867% de citações por documento. Dessa forma é possível observar que o tema tem apresentado uma grande relevância nos últimos anos.

A Figura 3 destaca o volume de publicações científicas anuais para tópicos de '*Machine Learning*' e '*physiological data*'.

Figura 3 - Publicações anuais entre 2019-2024

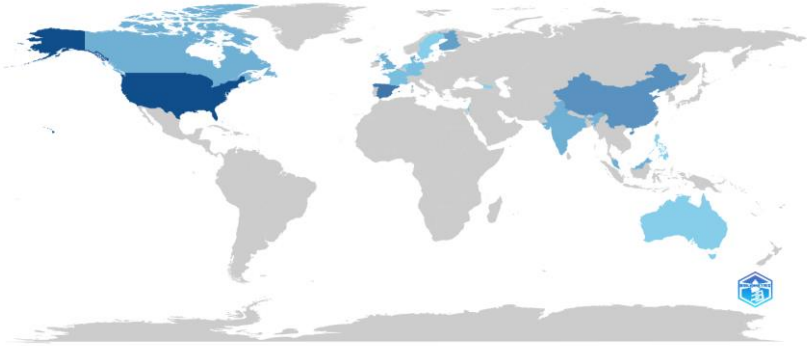


Fonte: Próprio autor (2025).

Observa-se que entre os anos de 2019 e 2022 houve um baixo volume de publicações relacionadas ao tema em questão. No entanto, a partir de 2023, o número de publicações teve um aumento considerável, chegando assim ao seu pico de publicações em um ano.

Na Figura 4 apresentam-se os países mais proeminentes em termos de produtividade nos temas, com sua contribuição diferenciada pela intensidade do tom de azul. Quanto mais escuro o tom, maior o número de publicações do país.

Figura 4 - Países mais produtivos



Fonte: Próprio autor (2025)

Nota-se na Figura 4, uma predominância de publicações nos Estados Unidos com 10 publicações, Espanha com 7 documentos e China com 5 artigos. O Brasil, no

entanto, não apresentou publicações sobre o tema, essa lacuna mostra a importância deste estudo, sendo uma oportunidade para novas frentes de pesquisas nacionais relacionadas ao tema.

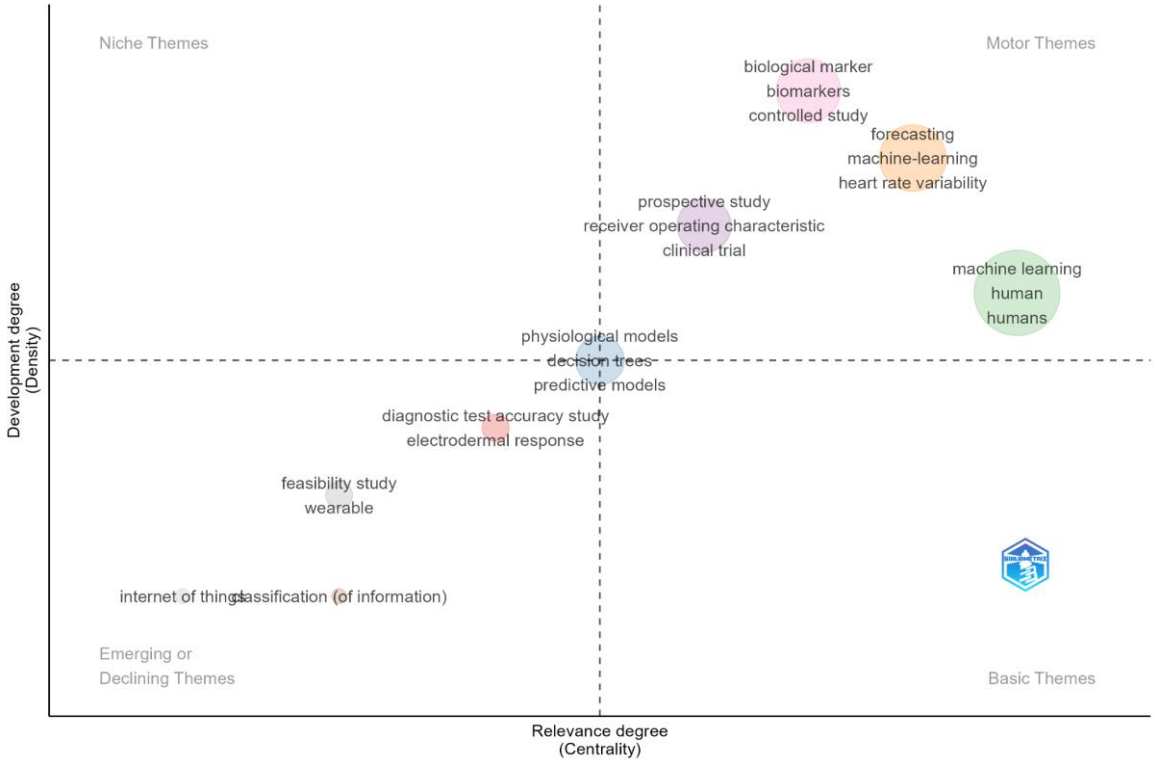
A Figura 5 ilustra a nuvem das 50 palavras mais citadas nos documentos analisados, utilizando como base os resumos dos artigos. Duas das palavras mais citadas nos são “*artificial neural network”* e “*random forest*”, dois modelos de aprendizagem de máquina que serão utilizados neste trabalho.

Figura 5 - 50 palavras mais citadas

Fonte: Próprio autor (2025).

Na Figura 6, apresenta-se o mapa temático desses artigos, que classifica os tópicos em quatro categorias distintas: temas motores, temas de nicho, temas básicos e emergentes ou em declínio. Além disso, no eixo vertical, tem-se o nível de estudo dos temas, enquanto no eixo horizontal, encontra-se seu nível de relevância.

Figura 6 - Mapa temático



Fonte: Próprio autor (2025).

Observa-se que, entre os temas motores, destacam-se aqueles relacionados a técnicas de aprendizado de máquina, com a presença de palavras-chave como *“Machine Learning”, “Neural Network”, “Random Forest” e “Predictive Model”.* Esses tópicos apresentam tanto alta relevância quanto elevado nível de desenvolvimento, o que evidencia seu papel central e consolidado nas pesquisas atuais sobre análise de dados fisiológicos e predição de engajamento.

Dadas as considerações anteriores, este estudo se justifica pela oportunidade de contribuir para uma área de pesquisa emergente e relevante, explorando uma combinação específica de técnicas e dados fisiológicos para a análise do engajamento estudantil, uma abordagem ainda pouco investigada, conforme demonstrado pela análise bibliográfica.

1.3 DELIMITAÇÃO DA PESQUISA

Este estudo baseia-se em dados reais obtidos no experimento conduzido por Luche *et al.* (2025). O objetivo é implementar modelos de *Machine Learning* para

identificar padrões de engajamento e estados emocionais de alunos em atividades de aprendizagem.

Para alcançar esse objetivo, foram utilizadas exclusivamente as medidas fisiológicas coletadas no experimento: EDA, HR e ST. O trabalho contempla as etapas de pré-processamento, extração e seleção de características relevantes dessas séries temporais, bem como a implementação, treinamento e comparação de modelos de aprendizagem supervisionada (redes neurais, Máquina de Vetores de Suporte e *decision trees*), com avaliação por meio de matriz de confusão.

O desenvolvimento foi realizado em Python, tendo como bibliotecas principais: *Pandas, NumPy, Matplotlib e Seaborn*, fundamentais para o pré-processamento e análise exploratória dos dados.

Este estudo, entretanto, não desenvolve nem testa novos dispositivos de coleta de dados, limita-se a trabalhar com o conjunto de dados já disponível (LUCHE *et al.*, 2025).

Por fim, não foi realizado o processo de validação cruzada, como o método k-fold, para verificar se o modelo estava realmente generalizando os padrões aprendidos. Sem essa etapa, a avaliação depende apenas da divisão treino-teste, o que pode limitar a confiabilidade dos resultados e aumentar o risco de overfitting.

1.4 ESTRUTURA DO TRABALHO

Este trabalho está estruturado em cinco capítulos, começando por esta introdução. O segundo capítulo será dedicado à revisão da literatura, abordando os principais conceitos teóricos que sustentam a construção de um modelo de *Machine Learning* para identificar padrões de engajamento. O terceiro capítulo descreve os métodos adotados e as ferramentas utilizadas na condução da pesquisa. No quarto capítulo, será apresentada uma análise exploratória dos dados, com o objetivo de revelar padrões, tendências e correlações significativas. Por fim, o quinto capítulo reunirá as conclusões do estudo, além de propor caminhos para investigações futuras.

2 REFERENCIAL TEÓRICO

Neste capítulo são abordados os conceitos e metodologias para o desenvolvimento do estudo. Inicialmente, serão discutidos os princípios fundamentais do aprendizado de máquina, com destaque para seus principais métodos e abordagens, serão apresentados e detalhados três algoritmos amplamente utilizados: Redes Neurais, *Random Forest* e Máquina de Vetores de Suporte (SVM), que constituem a base das técnicas empregadas neste estudo. Em seguida, será explorada a aplicação de técnicas do tema de Aprendizado de Máquina na análise de dados fisiológicos, destacando sua relevância para a identificação de padrões e para a compreensão dos estados emocionais e de engajamento em contextos educacionais.

2.1 APRENDIZADO DE MÁQUINA

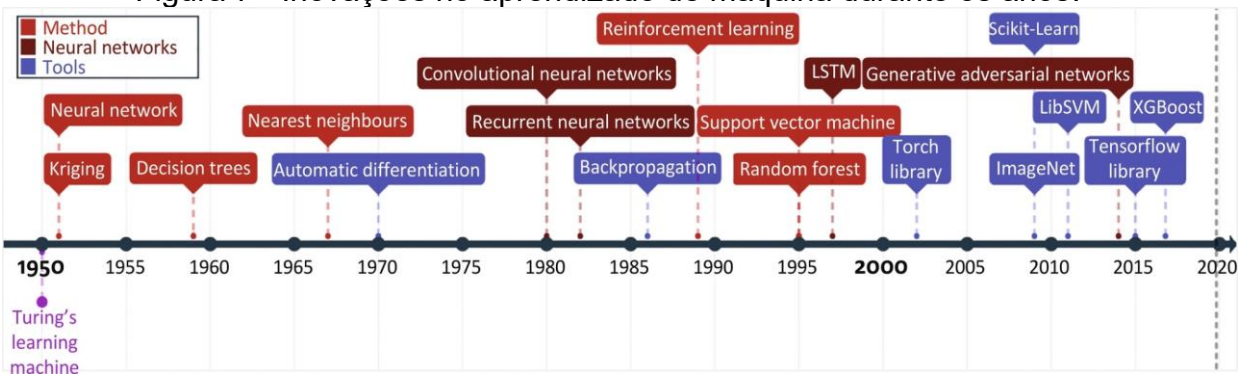
Aprendizado de máquina é um subcampo da inteligência artificial (IA), o qual se concentra no desenvolvimento de algoritmos que aprendam, a partir de uma base de dados fornecida previamente, a tomar decisões por conta própria e realizar previsões.

Os modelos de aprendizado de máquina são definidos pelas características do conjunto de dados e do problema a ser resolvido. Em outras palavras, a estrutura dos dados determina a abordagem metodológica e a escolha dos algoritmos mais adequados. O desempenho desses modelos é aprimorado a partir de um conjunto de dados de treinamento; posteriormente, utiliza-se uma parcela da base original, previamente separada e não utilizada no treinamento, para realizar os testes do algoritmo (DRAMSCH, 2020).

As décadas de 1950 e 1960 foram marcadas por um grande otimismo com o desenvolvimento de máquinas capazes de aprender a jogar jogos simples e realizar tarefas como o mapeamento de rotas. Já nessa época, métodos hoje considerados clássicos, como *k*-means, modelos de Markov e árvores de decisão, encontravam aplicações pioneiras em diversas áreas da ciência (DRAMSCH, 2020). Embora a popularização do aprendizado de máquina seja um fenômeno recente, sua história foi construída ao longo de décadas, de forma gradual. A Figura 7 apresenta a linha do tempo dessa evolução, destacando os principais marcos que fundamentaram o campo

ao longo dos anos. O histórico, mostra que a aplicação de métodos computacionais para extrair padrões de dados complexos possui uma longa tradição científica.

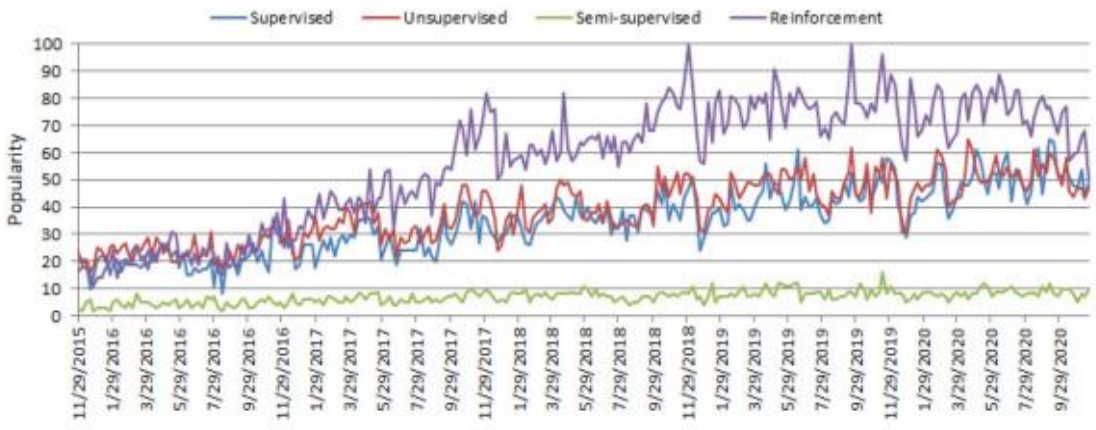
Figura 7 - Inovações no aprendizado de máquina durante os anos.



Fonte: Dramsch (2020)

O Aprendizado de Máquina, têm crescido rapidamente nos últimos anos no contexto da análise de dados e da computação, permitindo que aplicações funcionem de maneira inteligente (SARKER, 2021). A crescente popularidade dessas abordagens de aprendizado pode ser observada na Figura 8, que apresenta dados de interesse, ao longo do tempo, coletados da plataforma Google Trends.

Figura 8 – Popularidade dos Termos de Aprendizado de Máquina (2020-2025).

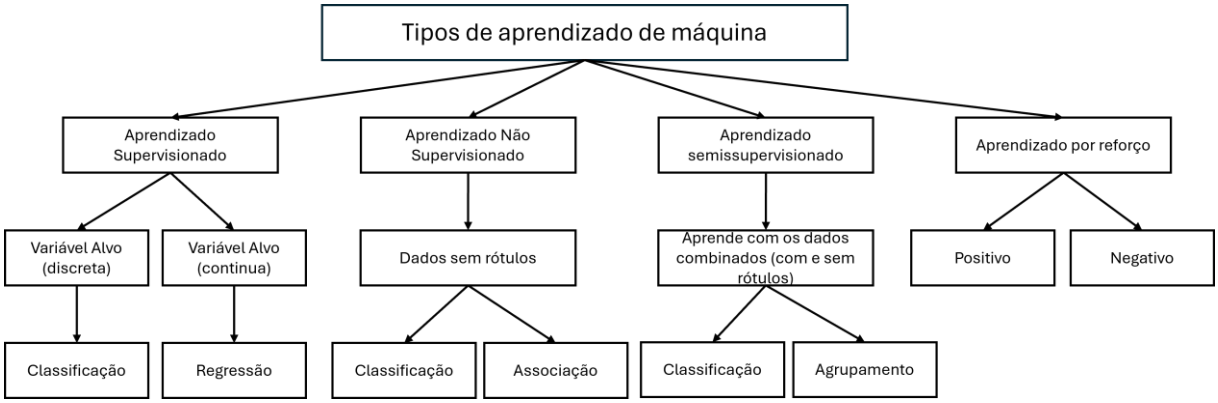


Fonte: Sarker (2021)

Conforme ilustrado, o interesse por essas técnicas tem aumentado consistentemente nos últimos cinco anos. Essas estatísticas, aliadas ao poder analítico dos modelos, motivam a aplicação do Aprendizado de Máquina neste trabalho.

Apesar do vasto potencial do Aprendizado de Máquina, a escolha do algoritmo mais adequado para um domínio particular é um desafio. Para isso, é importante compreender o problema a ser resolvido e qual será o tipo de modelo de aprendizado de máquina a ser utilizado. Na Figura 9 é possível ver os tipos de aprendizado de máquina existentes e suas ramificações:

Figura 9 - Tipos de aprendizado de máquina



Fonte: Adaptado de Sarker (2021)

2.2 TIPOS DE APRENDIZADO DE MÁQUINA

Os algoritmos de aprendizado de máquina são divididos principalmente em quatro categorias: supervisionado, não-supervisionado, semissupervisionado e aprendizado por reforço.

Segundo Sarker (2021), os algoritmos de aprendizado de máquina se dividem em quatro categorias principais: aprendizado supervisionado, que utiliza dados rotulados para treinar modelos; aprendizado não supervisionado, que identifica padrões e estruturas em dados não rotulados; aprendizado semissupervisionado, que combina uma pequena quantidade de dados rotulados com uma grande quantidade de dados não rotulados para aprimorar o desempenho do modelo; e aprendizado por reforço, que treina agentes para tomar decisões sequenciais em um ambiente, com base em recompensas ou penalidades. Este trabalho foca nos três paradigmas mais comuns: supervisionado, não-supervisionado e aprendizado por reforço

2.2.1 Aprendizado supervisionado

O aprendizado supervisionado pode ser entendido como um processo em que o modelo recebe a orientação de um “professor”. Nesse tipo de abordagem, o

algoritmo é treinado a partir de um conjunto de dados rotulados, ou seja, que contém tanto as variáveis de entrada quanto as respostas corretas. O objetivo é que o modelo identifique padrões nesses dados e, a partir deles, seja capaz de realizar previsões confiáveis em novos cenários.

Segundo Sarker (2021), as tarefas supervisionadas mais comuns são a classificação, que consiste em atribuir instâncias a categorias previamente definidas, e a regressão, que busca ajustar funções matemáticas aos dados para prever valores contínuos. Dessa forma, o aprendizado supervisionado constitui uma das técnicas mais utilizadas em *Machine Learning*, destacando-se pela sua capacidade de oferecer precisão, previsibilidade e suporte à tomada de decisão.

2.2.2 Aprendizado não supervisionado

O aprendizado não supervisionado, analisa conjuntos de dados não rotulados sem a necessidade de interferência humana. Durante o aprendizado, a máquina tenta aprender por conta própria sem a necessidade de um “professor”, ou seja, a máquina recebe o conjunto de dados sem a “resposta” correta, somente com os dados de entrada sem o rótulo, o algoritmo busca descobrir padrões ocultos, estruturas latentes e relações de proximidade. Por meio dos dados de entrada a máquina tem que aprender e criar o modelo preditivo, esses modelos são amplamente utilizados para extrair características generativas, identificar tendências, estruturas significativas, agrupamentos em resultados e para fins exploratórios. (SARKER, 2021).

Esse tipo de aprendizado é amplamente utilizado para identificação de tendências, extração de características relevantes e agrupamentos exploratórios, tendo aplicações em áreas como análise emocional via sinais fisiológicos (SHARMA; PRAKASH; KALRA, 2019).

2.2.3 Aprendizado por reforço

O aprendizado por reforço é o tipo de aprendizado mais similar a uma inteligência artificial. Durante o aprendizado o algoritmo é recompensado ou penalizado, e seu objetivo final é usar os aprendizados obtidos durante as atividades ambientais para tomar ações que aumentem a recompensa ou minimizem o risco. É uma ferramenta poderosa para treinar modelos de IA que podem ajudar a aumentar a automação ou otimizar a eficiência operacional de sistemas sofisticados, como robótica e tarefas de direção autônoma (SARKER, 2021).

Esse tipo de aprendizado tem sido amplamente aplicado em robótica, jogos e sistemas de direção autônoma, destacando-se pela sua capacidade de otimização em ambientes dinâmicos e incertos (DRAMSCH, 2020). Além disso, o aprendizado por reforço tem sido explorado em conjunto com sinais fisiológicos e sensores multimodais, possibilitando o desenvolvimento de sistemas capazes de otimizar decisões em contextos de estresse ou carga cognitiva elevada, com impacto direto em áreas como saúde, transporte e educação (LEE *et al.,* 2020).

Ao contrário do aprendizado supervisionado ou não supervisionado, o aprendizado por reforço não exige grandes volumes de dados rotulados, mas sim um mecanismo de feedback, tornando-o adequado para sistemas de automação complexos.

2.3 ALGORITMOS DE APRENDIZADO DE MÁQUINAS

Os algoritmos de aprendizado de máquina podem ser diversos, mas neste estudo serão utilizados os algoritmos de: redes neurais, *random forest* e máquinas de vetores de suporte (SVM).

2.3.1 Redes neurais

As Redes Neurais são modelos que se inspiram na estrutura e funcionamento do cérebro humano. Um modelo comum consiste em uma camada de entrada, uma ou mais camadas ocultas e uma camada de saída. Cada neurônio (ou nó) em uma camada oculta realiza uma soma ponderada das saídas da camada anterior, adiciona um viés e aplica uma função de ativação não linear (SHARMA; PRAKASH; KALRA, 2019).

Métodos de Aprendizado Profundo, os quais utilizam redes neurais com múltiplas camadas ocultas, são promissores para modelagem de séries temporais, pois podem aprender automaticamente dependências temporais e lidar com estruturas como tendências e sazonalidades (ČERTICKÝ *et al.,* 2019).

2.3.2 Random florest (Floresta aleatória)

A *Random Forest* (ou Floresta de Decisão Aleatória) é um algoritmo de aprendizado de conjunto amplamente utilizado em tarefas de classificação e regressão. Seu funcionamento baseia-se na construção de múltiplas árvores de decisão durante o processo de treinamento, cujas respostas são posteriormente

combinadas. No caso da classificação, a saída final corresponde à classe mais votada entre as árvores, enquanto, para regressão, o resultado é obtido pela média das previsões individuais. Além disso, destaca-se pela flexibilidade e simplicidade de uso, podendo lidar com variáveis de diferentes distribuições e dispensando pré-processamentos complexos, como o escalonamento de características (CHOWSHURY *et al.*, 2019).

Uma das principais vantagens da *Random Forest* é a sua capacidade de reduzir o sobreajuste (overfitting), problema recorrente em árvores de decisão isoladas. Ao agregar diversas árvores, o modelo consegue generalizar melhor para novos dados, mantendo bom equilíbrio entre precisão e robustez. Além disso, destaca-se pela flexibilidade, já que pode lidar com variáveis de diferentes distribuições e não exige pré-processamentos complexos, como o escalonamento de características (ČERTICKÝ *et al.,* 2019).

2.3.3 Máquina de vetores de suporte (SVM)

A Máquina de Vetores de Suporte (Support Vector Machine - SVM) é um algoritmo de aprendizado supervisionado utilizado tanto para classificação quanto para regressão. Em problemas com múltiplas classes, o SVM, que é inerentemente binário, pode ser adaptado classificando uma classe contra todas as outras de forma iterativa (CHOWDHURY *et al.,* 2019).

As implementações de SVM frequentemente utilizam diferentes funções de kernel para transformar os dados, permitindo a separação de classes que não são linearmente separáveis no espaço original; para problemas linearmente separáveis usam-se funções de kerner-linera (POSADA-QUINTERO; BOLKHOVSKY, 2019; ANUSHA A. S. *et al.,* 2019). Para as fronteiras de decisão não-lineares, usa-se Kernel de Função de Base Radial (RBF) ou gaussiano (ANUSHA *et al.*, 2019; FIORINI *et al.*, 2020).

Além da classificação, o SVM é utilizado em conjunto com outras técnicas como a Eliminação Recursiva de Características (SVM-RFE), um método de seleção de características que treina um modelo SVM e remove recursivamente as características menos importantes (KIM *et al.,* 2018).

3 MATERIAIS E MÉTODOS

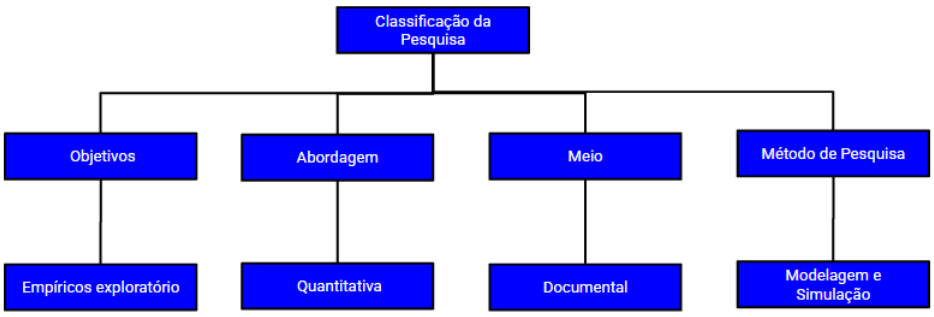
Neste estudo foi utilizada uma abordagem quantitativa, uma vez que, de acordo com Miguel *et al.* (2018), a abordagem quantitativa envolve a coleta e análise de dados numéricos, utilizando ferramentas estatísticas para interpretação dos resultados. Durante o estudo aplicou-se métricas como acurácia para mensurar o desempenho dos modelos de classificação e análises estatísticas para encontrar correlações na base de dados.

Para Gil (2002), as pesquisas descritivas têm como propósito a caracterização de fenômenos e o estabelecimento de relações entre variáveis, enquanto as pesquisas exploratórias têm como principal finalidade desenvolver, esclarecer e modificar conceitos e ideias, tendo em vista a formulação de problemas mais precisos ou hipóteses pesquisáveis para estudos posteriores. Assim, este estudo é de natureza descritiva e exploratória, pois busca encontrar padrões nos dados de fisiológicos e avalia a eficácia dos algoritmos supervisionados aplicados.

Conforme Miguel *et al.* (2018), emprega-se a modelagem e simulação quando o objetivo da pesquisa é experimentar, por meio de um modelo, um sistema real, determinando-se como esse sistema responde a modificações que lhe são propostas. Já a simulação fornece condições experimentais controladas para manipulação de variáveis. Nesse contexto, este trabalho caracteriza-se como um estudo baseado em modelagem e simulação computacional de cenários de classificação, utilizando algoritmos de aprendizado de máquina para analisar e prever padrões de engajamento estudantil.

O meio será documental utilizando dados previamente publicados, uma vez que o estudo é baseado em um banco de dados de um experimento real (LUCHE *et al.*, 2025). A Figura 10 ilustra o desenho metodológico da pesquisa:

Figura 10 – Classificação da pesquisa



Fonte: Próprio autor (2025)

3.1 BASE DE DADOS

A base de dados que foi utilizada neste trabalho foi previamente formulada a partir de um experimento realizado por Luche *et al.*(2025), envolvendo uma amostra de cinquenta estudantes universitários, com idades entre 18 a 28 anos, sendo 62% do gênero masculino e 38% do gênero feminino. A amostra foi composta por estudantes que cursavam engenharia de produção e mecânica e o ano acadêmico, nos quais os alunos estavam matriculados em seus cursos, eram diversos, porém 50% da amostra estava cursando o terceiro ano da graduação.

Durante o experimento, os participantes foram expostos a duas metodologias educacionais: métodos de aprendizagem tradicional e métodos ativos. Ao longo das atividades, foram monitorados os seguintes sinais fisiológicos de cada participante:

- Atividade Eletrodérmica EDA)
- Frequência Cardíaca (HR)
- Temperatura da Pele (ST)
- Tempo (segundos)

Os dados foram registrados continuamente durante as sessões, permitindo a associação temporal precisa entre os estímulos educacionais e as respostas fisiológicas dos alunos. Assim, a estrutura da base de dados inclui registros por aluno, contendo as medições ao longo do tempo em cada tipo de aula (tradicional e ativa).

3.2 FERRAMENTAS

A execução desta pesquisa envolveu o uso de diferentes ferramentas tecnológicas:

- Python: Linguagem de programação utilizada para todas as etapas do estudo, incluindo pré-processamento, modelagem e avaliação dos resultados.
- Bibliotecas Python:
    - Pandas e NumPy: manipulação, limpeza e transformação dos dados;
    - Matplotlib e Seaborn: análise exploratória e visualização gráfica;
    - Scikit-learn: implementação dos algoritmos de aprendizado de máquina, pré-processamento e aplicação de técnicas de validação cruzada.
- Modelos de Classificação: Foram testados três algoritmos supervisionados: Redes neurais, Máquina de Vetores de Suporte (SVM) e Floresta Randômica (*Random Forest*).

3.3 AVALIAÇÃO DOS MODELOS

Os modelos foram avaliados utilizando uma Matriz de Confusão. Essa ferramenta permite analisar os acertos e erros de um classificador a partir da comparação entre valores previstos e valores reais (REFERÊNCIA).

De acordo com Powers (2011), a matriz de confusão é composta por quatro elementos fundamentais:

- Verdadeiros Positivos (TP): número de instâncias corretamente classificadas como pertencentes à classe positiva;
- Falsos Positivos (FP): número de instâncias incorretamente classificadas como positivas;
- Verdadeiros Negativos (TN): número de instâncias corretamente classificadas como pertencentes à classe negativa;
- Falsos Negativos (FN): número de instâncias positivas que o modelo classificou de forma incorreta como negativas.

A Figura 11 apresenta um exemplo de matriz de confusão:

Figura 11 – Matriz de confusão

Fonte: Powers (2011)

A partir dessa matriz, derivam-se métricas fundamentais para a análise de desempenho de modelos supervisionados, como:

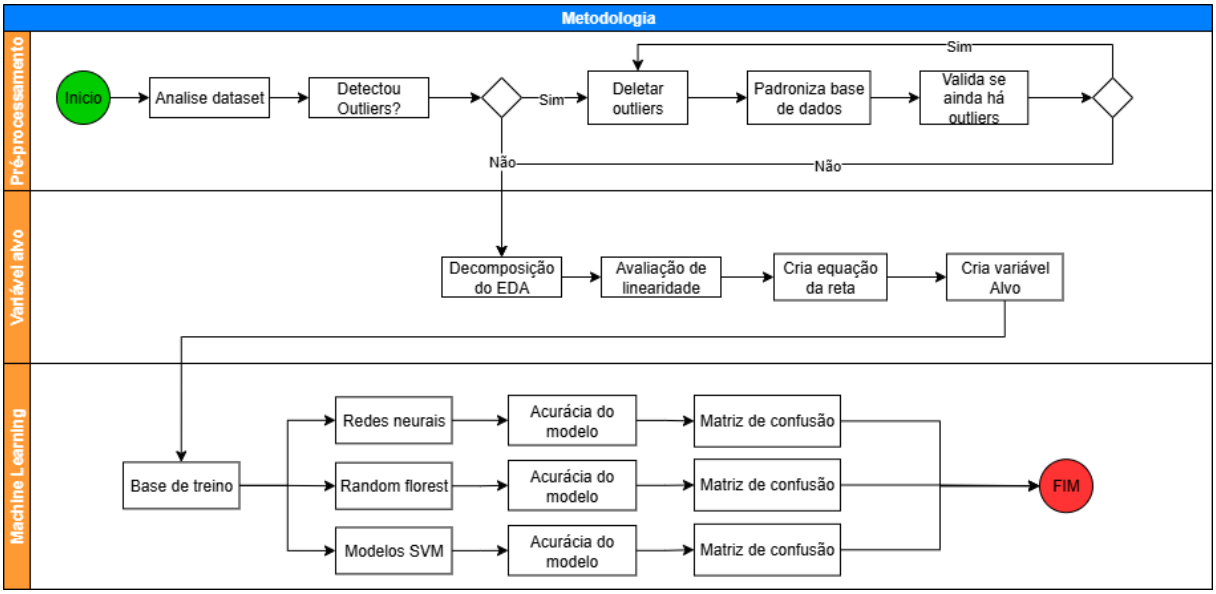
- Acurácia: proporção de predições corretas em relação ao total de instâncias.
- Precisão: proporção de acertos entre as predições positivas
- Recall (Sensibilidade): proporção de casos positivos corretamente identificados;
- Especificidade: proporção de negativos corretamente identificados;
- F1-score: média harmônica entre precisão e recall, relevante em cenários de classes desbalanceadas

Neste estudo a matriz de confusão foi a ferramenta principal para validar os algoritmos utilizados, entendendo a sua taxa de acerto e evidenciando onde os erros estão acontecendo.

4 EXPERIMENTO

O fluxograma na Figura 12, ilustra a metodologia aplicada para executar o estudo, iniciando pela etapa de análise do *dataset* e seguindo até as fases de análises e conclusões dos modelos de *machine learning.*

Figura 12 – Metodologia da pesquisa



Fonte: Próprio autor (2025)

4.1 PRÉ-PROCESSAMENTO E LIMPEZA DOS DADOS

A partir da base de dados (LUCHE *et al.*, 2025) foram realizadas etapas de pré-processamento, limpeza e transformação dos dados utilizando a linguagem de programação Python. O objetivo dessa etapa foi preparar os dados para análises subsequentes, garantindo sua consistência e qualidade.

4.1.1 Importação inicial

Na etapa inicial do desenvolvimento das análises exploratórias e o modelo de *Machine Learning,* foram importadas as bibliotecas para o processamento, manipulação e modelagem dos dados. Por meio dessas bibliotecas foi possível realizar as análises e o código dos algoritmos de aprendizado de máquina.

As principais bibliotecas utilizadas foram: pandas e numpy, matplotlib e seabor, statsmodels e scipy, sklearn e os. A última sendo uma biblioteca padrão do Python utilizada para interagir com o sistema operacional, como navegar entre diretórios e acessar arquivos CSV de forma automatizada.

A partir da importação das bibliotecas, iniciou-se o processo de importação da base de dados para o python. Os arquivos fornecidos pelo estudo anterior (LUCHE *et al.*, 2025) eram arquivos em formato csv, e cada arquivo possui os dados de um aluno, correspondente ao tipo de aula ofertada a ele (tradicional ou ativa). Para facilitar as análises futuras, os arquivos foram unificados em uma única base de dados, a qual contém informações de todos os alunos, distinguindo-os por um aluno id e por uma coluna de tipo de aula, ativa ou tradicional.

Na Figura 13 apresenta-se o código utilizado para realizar a importação do dataframe.

Figura 13 - Importando base de dados

```
import pandas as pd
from scipy.stats import zscore
import os
import pandas as pd
pasta_ativa = 'E:\giaco\TCC_Lucas_Giusti\Physiological_Data\Active'
pasta_tradicional = 'E:\giaco\TCC_Lucas_Giusti\Physiological_Data\Traditional'

lista_dfs = []

def processar_pasta(pasta, tipo_aula):
    for arquivo in os.listdir(pasta):
        if arquivo.endswith('.csv'):
            caminho_arquivo = os.path.join(pasta, arquivo)
            try:
                df = pd.read_csv(caminho_arquivo, delimiter=';')
                df['aluno_id'] = os.path.splitext(arquivo)[0]
                df['tipo_aula'] = tipo_aula
                lista_dfs.append(df)
            except Exception as e:
                print(f'Erro ao ler {arquivo}: {e}')

# Processar ambas as pastas
processar_pasta(pasta_ativa, 'aula_ativa')
processar_pasta(pasta_tradicional, 'aula_tradicional')

# Concatenar os dados
df_geral = pd.concat(lista_dfs, ignore_index=True)

# Visualizar amostra


# Salvar o arquivo unificado
df_geral.to_csv('dados_unificados.csv', index=False, sep=';')
print('Type columns:', df_geral.dtypes)
print(df_geral.head())
```

Fonte: Produção do próprio autor (2025)

A partir desse código gerou-se as análises sobre o *type* das variáveis, apresentada no Quadro 1.

Quadro 1 – Type das variáveis

Variável	*Type*
Seconds	float64
HR	float64
EDA	object
ST	object
aluno_id	object
tipo_aula	object

Fonte: Produção do próprio autor (2025)

A partir da do Quadro 1, observou-se que as colunas referentes à EDA e à ST estavam com o tipo de dado: *object*, indicando que os valores estavam armazenados como texto. Para possibilitar análises quantitativas, essas colunas foram convertidas para o tipo float64, permitindo a representação numérica com casas decimais. Por fim, na Tabela 1, é possível verificar a visualização da base de dados gerada.

Tabela 1 - Base de dados gerada

Seconds	HR	EDA	ST	aluno_id	tipo_aula
1.0	60.0	4,2	35,1	A_1	aula_ativa
2.0	78.0	4,3	35,1	A_1	aula_ativa
3.0	81.0	4,3	35,1	A_1	aula_ativa
4.0	60.0	4,3	35,1	A_1	aula_ativa
5.0	80.0	4,2	35,1	A_1	aula_ativa

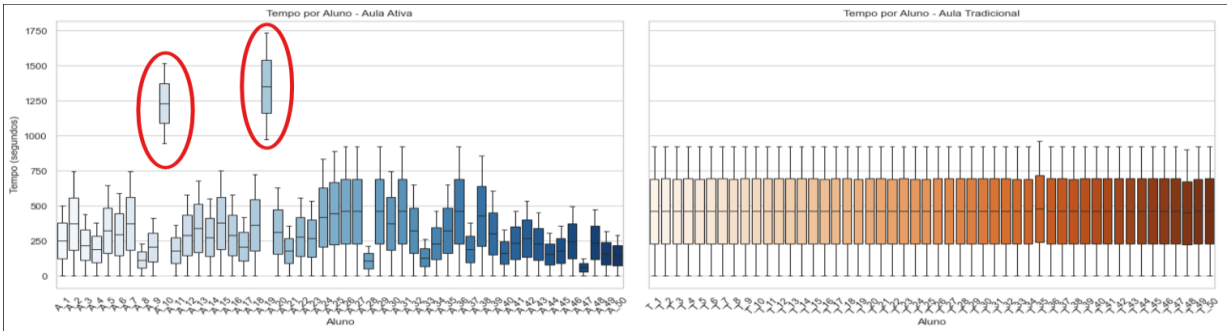
Fonte: Produção do próprio autor (2025)

4.1.2 Identificação de outliers e anomalias

Para identificar possíveis outliers e inconsistências nos dados, foram gerados gráficos do tipo boxplot para cada variável (segundos, EDA, HR e ST), segmentados por aluno e por tipo de aula (tradicional ou ativa). Essa visualização permitiu detectar valores atípicos que poderiam comprometer a análise. Durante essa etapa, foram identificadas duas principais anomalias:

Deslocamento Temporal: Os dados de tempo dos alunos A_10 e A_19 não iniciavam no segundo zero, indicando um possível erro de sincronização no início da coleta. Para corrigir essa inconsistência, os tempos desses alunos foram corrigidos, reiniciando a contagem a partir do zero. Na figura 14, apresenta-se o gráfico boxplot de segundos por alunos id e tipo de aula, é possível notar a disparidade dos dados dos alunos: A_10 e A_19.

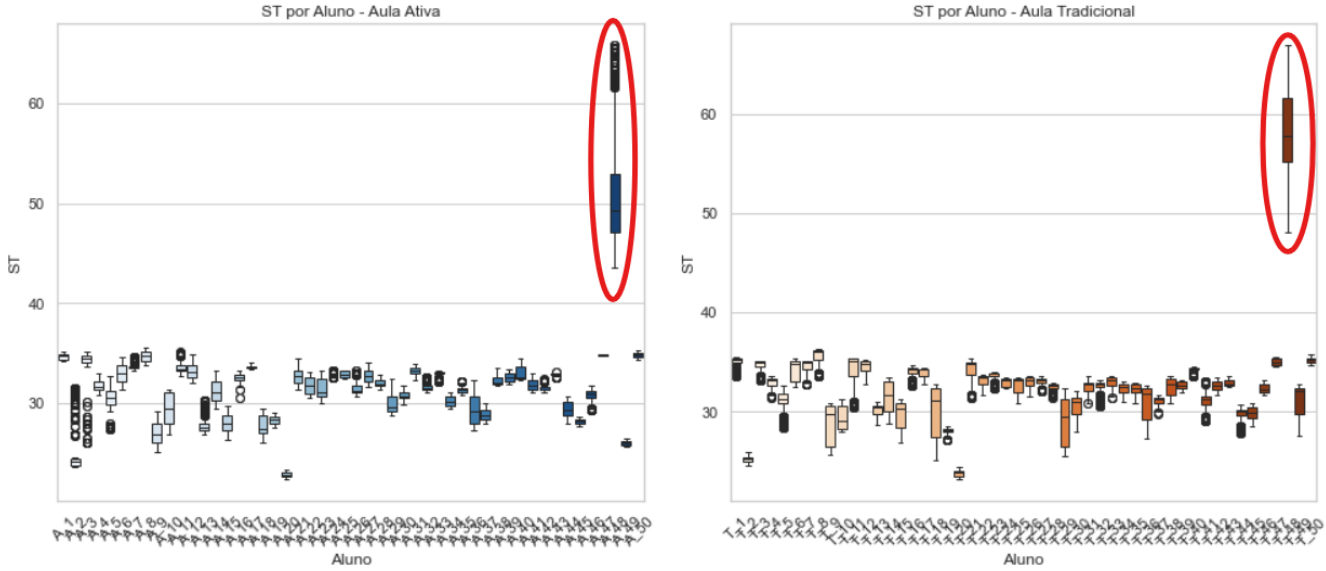
Figura 14 - Boxplot de segundos por aluno id e tipo de aula



Fonte: Produção do próprio autor (2025) (2025)

Temperatura Corporal Anômala: O aluno A_48 apresentou valores médios de temperatura corporal em torno de 58 °C, significativamente acima da faixa considerada normal para seres humanos, que varia entre 36,1 °C e 37,2 °C. Considerando que temperaturas corporais acima de 41 °C são consideradas emergências médicas, optou-se por excluir os dados desse aluno tanto nas aulas ativas quanto nas tradicionais. Na figura 15, apresenta-se o gráfico boxplot da temperatura da pele por aluno id e por tipo de aula, é possível observar como o A_48 é um *outiler* em relação aos demais.

Figura 15 - Boxplot ST (temperatura da pele) por aluno id e por tipo de aula



Fonte: Produção do próprio autor (2025) (2025)

Por fim, para remover *outliers* fisiologicamente implausíveis, foram estabelecidos intervalos de valores considerados normais para cada uma das variáveis fisiológicas, com base na literatura médica:

Frequência Cardíaca (HR): 40 a 140 batimentos por minuto (bpm). A frequência cardíaca em repouso para adultos varia entre 60 e 100 bpm, valores fora desse intervalo são considerados atípicos para o contexto do estudo (OLSHANSKY *et al.*, 2022).

Atividade Eletrodérmica (EDA): 1 a 20 microsiemens (µS) (BIOPAC, 2018). A EDA é uma medida da condutância da pele, variando conforme o estado emocional e o nível de excitação do indivíduo. Valores fora desse intervalo foram considerados outliers.

Temperatura da Pele (ST): 31 °C a 38 °C (LIU, Y. *et al.,* 2013). A temperatura da pele pode variar dependendo de fatores ambientais e fisiológicos, mas geralmente permanece dentro dessa faixa.

A aplicação desses filtros resultou na exclusão de registros que apresentavam valores fora dos intervalos estabelecidos, assegurando que a base de dados final refletisse medidas fisiológicas plausíveis para os participantes do estudo.

Nas Tabelas 2 e 3, ilustra-se a base de dados antes e após os tratamentos aplicados:

Tabela 2 - Base de dados antes da exclusão dos outliers

Cálculo	Seconds	HR	EDA	ST
count	73.056,00	73.056,00	73.056,00	73.056,00
mean	410,49	84,21	4,54	31,62
std	259,59	14,93	1,42	2,69
min	1,00	33,00	1,90	22,40
25%	188,00	78,00	3,40	30,40
50%	384,00	83,00	4,30	32,30
75%	620,00	90,00	5,30	33,30
max	961,00	145,00	9,80	36,30

Fonte: Produção do próprio autor (2025).

Tabela 3 - Base de dados após a exclusão dos outliers

Cálculo	Seconds	HR	EDA	ST
count	70.503,00	70.503,00	70.503,00	70.503,00
mean	409,97	84,54	4,59	31,90
std	259,68	14,83	1,42	2,28
min	1,00	40,00	1,90	25,00
25%	187,00	78,00	3,40	30,70
50%	383,00	83,00	4,40	32,40
75%	620,00	90,00	5,40	33,30
max	961,00	140,00	9,80	36,30

Fonte: Produção do próprio autor (2025).

O tratamento na base de dados resultou na exclusão de 2.553 linhas de dados, garantindo intervalos de máximo e mínimo mais condizentes com a realidade.

4.2 EXTRAÇÃO DE VARIÁVEIS DERIVADAS DO EDA

Com o objetivo de enriquecer a análise dos dados de atividade eletrodérmica (EDA), foram extraídas variáveis derivadas baseadas nos domínios do tempo e da frequência, conforme metodologia descrita por Posada-Quintero e Bolkhovsky (2019). Essas variáveis oferecem maior sensibilidade à identificação de respostas autonômicas do sistema nervoso simpático, particularmente relacionadas ao estresse cognitivo.

As variáveis extraídas foram:

1. Skin Conductance Level (SCL): Valor médio da condutância da pele ao longo do tempo; indica o nível geral de excitação fisiológica.
2. Skin Conductance Responses (SCRs): Variações rápidas da EDA em resposta a estímulos; refletem reações emocionais ou cognitivas pontuais.
3. NS.SCRs (Non-Specific SCRs por minuto): Quantidade de SCRs espontâneas por minuto, sem estímulo definido; mede a reatividade autonômica.
4. EDASymp: Índice espectral da EDA em frequência (0,045–0,25 Hz); avalia a atividade simpática relacionada ao estresse.
5. TVSymp: Índice espectral da EDA variável no tempo; reflete mudanças dinâmicas na ativação simpática.

4.2.1 Skin condutance level (SCL)

O SCL foi obtido mediante o sinal bruto de EDA e aplicou-se um filtro digital passa-baixa do tipo *Butterworth*, cuja função é atenuar componentes de alta frequência, permitindo a passagem apenas das frequências mais baixas, no caso deste trabalho, adotou-se um filtro de quarta ordem, com frequência de corte em 0,05 Hz. (POSADA-QUINTERO; BOLKHOVSKY, 2019).

A definição da frequência de corte leva em consideração a frequência de Nyquist, que corresponde à metade da taxa de amostragem do sinal (fs/2). Assim, a frequência de corte de 0,05 Hz foi normalizada pela frequência de *Nyquist* antes de ser aplicada ao filtro, assegurando que o algoritmo trabalhasse em termos relativos à resolução temporal do sinal coletado (POSADA-QUINTERO; BOLKHOVSKY, 2019).

A filtragem foi implementada utilizando o método *filtfilt*, que aplica o filtro duas vezes ao sinal: uma vez no sentido direto e outra no sentido inverso. O *filtfilt* garante que não haja deslocamento temporal nas características do sinal, o que é essencial para análises fisiológicas em que a temporalidade dos eventos deve ser preservada (POSADA-QUINTERO; BOLKHOVSKY, 2019).

Após essa etapa, o sinal de EDA resultante apresentava apenas as variações de baixa frequência correspondentes ao componente tônico. O SCL foi então obtido como a média desse sinal suavizado ao longo do período considerado, fornecendo uma medida representativa e estável do nível basal de condutância da pele (POSADA-QUINTERO; BOLKHOVSKY, 2019).

Figura 16 - Função para calcular o SCL

```
FUNÇÃO filtrar_scl(sinal_EDA, fs = 4):
    nyquist ← 0.5 * fs
    cutoff ← 0.05 / nyquist
    (b, a) ← filtro Butterworth passa-baixa (ordem 4, cutoff)
    sinal_filtrado ← aplicar filtfilt(b, a, sinal_EDA)
    RETORNAR sinal_filtrado

FUNÇÃO calcular_scl(sinal_EDA, fs = 4):
    scl ← filtrar_scl(sinal_EDA, fs)
    RETORNAR média(scl)
```

Fonte: Produção do próprio autor (2025).

4.2.2 Números de SCRs não específicos por minuto (NS.SCRs)

O Skin Conductance Responses não específico (NS.SCRs) foi obtido a partir da detecção de picos no sinal derivado da atividade eletrodérmica (EDA), considerando-se como resposta válida apenas aquelas cuja amplitude ultrapassava o limiar de 0,05 µS. Esse procedimento segue a recomendação de Boucsein (2012) e foi igualmente adotado por Posada-Quintero e Bolkhovsky (2019). Após a identificação dos picos, calculou-se a frequência de ocorrência das respostas por minuto, resultando no índice NS.SCRs/min, que expressa a quantidade média de respostas fásicas espontâneas ao longo do tempo. Na Figura 17 é apresentada a função implementada em Python utilizada para o cálculo desse índice:

Figura 17 - Função para calcular o NS.SCRs

```
FUNÇÃO calcular_nsscrs(sinal_EDA, fs = 4, limiar = 0.05):
    derivada ← diferença entre amostras de sinal_EDA
    picos ← detectar picos positivos em derivada com altura ≥ limiar
    duração_segundos ← tamanho(sinal_EDA) / fs
    nsscrs ← número_de_picos / (duração_segundos / 60)
    RETORNAR nsscrs
```

Fonte: Produção do próprio autor (2025).

4.2.3 EDASymp – Índice espectral de EDA

O EDASymp foi obtido a partir da análise espectral do sinal de EDA no domínio da frequência, com o objetivo de quantificar a potência associada à atividade simpática. Para isso, empregou-se o método de Welch, que consiste em uma técnica

de estimação da densidade espectral de potência a partir da divisão do sinal em segmentos, reduzindo a variância da estimativa (POSADA-QUINTERO; BOLKHOVSKY, 2019).

Utilizou-se a janela de Blackman, que minimiza o vazamento espectral, e adotou-se uma sobreposição de 50% entre os segmentos para aumentar a suavização da estimativa. A partir do espectro resultante, integrou-se a potência no intervalo de frequência entre 0,045 e 0,25 Hz, considerado sensível a respostas cognitivas (POSADA-QUINTERO; BOLKHOVSKY, 2019). O valor da integral nessa banda foi considerado o índice EDASymp, representando assim a contribuição relativa da atividade simpática no sinal de EDA. Na Figura 18 é apresentada a função implementada em Python para o cálculo desse índice:

Figura 18 - Função para calcular o EDASymp

```
FUNÇÃO calcular_tvsymp(sinal_EDA, fs = 4):
    (frequências, tempo, espectro) ← espectrograma(sinal_EDA, janela = Blackman, nperseg=128, noverlap=64)
    selecionar faixa onde 0.08 ≤ f ≤ 0.24 Hz
    potência_banda ← soma das potências nesta faixa para cada instante
    tvsymp ← média(potência_banda) / soma_total(espectro)
    RETORNAR tvsymp
```

Fonte: Produção do próprio autor (2025).

4.2.4 TVSymp – Índice espectral variável no tempo

O TVSymp corresponde a uma medida dinâmica da atividade simpática, baseada na distribuição temporal da potência espectral do sinal de EDA. Para sua obtenção, empregou-se o espectrograma, que permite representar simultaneamente as componentes de frequência e sua evolução ao longo do tempo. Essa abordagem é utilizada como aproximação do método de demodulação de frequência variável proposto por Posada-Quintero *et al.* (2019).

No cálculo, considerou-se especificamente a faixa de frequência entre 0,08 e 0,24 Hz, associada à modulação simpática (POSADA-QUINTERO; BOLKHOVSKY, 2019). Em cada instante de tempo, a potência dentro desse intervalo foi somada e, em seguida, obteve-se a média dessa potência ao longo de todo o sinal. O valor médio foi então normalizado pela potência total do espectro, resultando em um índice adimensional que reflete a proporção da atividade simpática na faixa de interesse em relação ao conteúdo espectral global. Dessa forma, o TVSymp fornece uma visão temporalmente sensível das variações simpáticas, complementando medidas

estáticas como o EDASymp. Na Figura 19 é apresentada a função implementada em Python para o cálculo do TVSymp:

Figura 19 - Função para calcular o EDASymp

```
FUNÇÃO calcular_tvsymp(sinal_EDA, fs = 4):
    (frequências, tempo, espectro) ← espectrograma(sinal_EDA, janela = Blackman, nperseg=128, noverlap=64)
    selecionar faixa onde 0.08 ≤ f ≤ 0.24 Hz
    potência_banda ← soma das potências nesta faixa para cada instante
    tvsymp ← média(potência_banda) / soma_total(espectro)
    RETORNAR tvsymp
```

Fonte: Produção do próprio autor (2025).

4.3 VALIDAÇÃO ESTATÍSTICA

A aplicação dos algoritmos foi feita por agrupamento de cada aluno e tipo de aula, garantindo as novas variáveis para cada aluno de acordo com o experimento que foi executado. A Tabela 4, apresenta a nova base de dados com as variáveis novas incluídas.

Tabela 4 - Nova base de dados com as novas variáveis

Seconds	HR	EDA	ST	aluno_id	tipo_aula	SCL	NSSCRs	EDASymp	TVSymp
1	60,00	4,20	35,10	A_1	aula_ativa	4,6275	26,6139	0,0045	0,0490
2	78,00	4,30	35,10	A_1	aula_ativa	4,6275	26,6139	0,0045	0,0490
3	81,00	4,30	35,10	A_1	aula_ativa	4,6275	26,6139	0,0045	0,0490
4	60,00	4,30	35,10	A_1	aula_ativa	4,6275	26,6139	0,0045	0,0490
5	80,00	4,20	35,10	A_1	aula_ativa	4,6275	26,6139	0,0045	0,0490

Fonte: Produção do próprio autor (2025).

A fim de validar as novas variáveis obtidas e o banco de dados após a remoção dos outliers, foram realizados alguns testes estatísticos.

4.3.1 Avaliação de linearidade

Para entender se a relação entre cada variável (HR, EDA e ST) pode ser adequadamente representada por um modelo linear, foi realizada uma avaliação de linearidade. Isso é importante porque, se a relação for de fato linear, modelos mais simples tendem a ser suficientes; se não for, é sinal de que relações mais complexas (não-lineares) ou transformações de variáveis podem ser necessárias para capturar o comportamento dos dados.

Para verificar isso, agrupou-se os dados de acordo com o aluno_id e o tipo_aula para possibilitar a análise primeiro por grupo e depois o dado global. Foi analisado o

coeficiente de determinação (R^2), uma medida estatística que indica quanto da variação total de uma variável-alvo é explicada pelo modelo usando os preditores escolhidos. Pode variar de 0 a 1, sendo que quanto mais perto do número 1, mais a variável explica o modelo.

Nesse caso, analisou-se o R^2 para um modelo linear e um polinomial, e se verificou para qual dos dois modelos, dentro de cada grupo, as variáveis mais se encaixavam, em seguida se realizou o delta, a diferença entre os R^2, considerou-se linear todo delta menor que 0,05.

Após essa etapa, foi feita uma análise global dos dados olhando o todo de cada grupo, na Tabela 5, ilustra-se essa análise global:

Tabela 5 – Resultados da análise de linearidade

Target	Total_grupos	Lineares	Nao_lineares	R2_Linear_global	R2_Poly2_global	Delta_global	Veredito_global
HR	96	91	5	0.0066	0.0166	0.0100	Linear
EDA	96	26	70	0.0148	0.0243	0.0095	Linear
ST	96	32	64	0.0415	0.0522	0.0107	Linear

Fonte: Produção do próprio autor (2025)

Com base na Tabela 5, observa-se que para HR, para a maioria dos grupos (91 de 96 grupos) apresentou comportamento linear, com um ganho médio pequeno ao incluir termos quadráticos (≈ 0,01), o que reforça a linearidade. Já para EDA e ST, existe a predominância de grupos não-lineares, no entanto, por delta ser menor que 0,05, eles foram classificados como lineares. No conjunto, a base de dados apresenta linearidade robusta para HR, mas uma tendência maior a comportamentos não-lineares em EDA e ST.

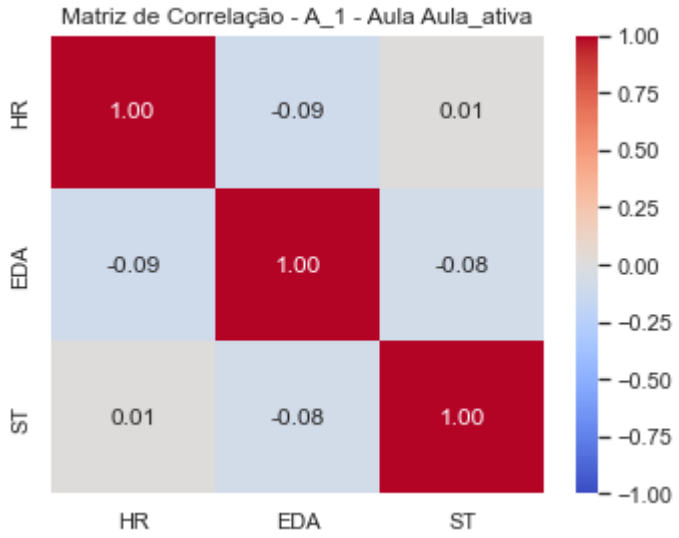
4.3.2 Corelação

A fim de concluir a etapa exploratória dos dados, foi realizada uma análise de correlação, calculando para cada aluno e tipo de aula, as correlações de Pearson entre HR, EDA e ST. Em seguida, foram obtidas as médias dessas correlações por tipo de aula, comparando padrões entre aulas ativas e tradicionais, e o desvio padrão geral, que indica a variabilidade dessas relações entre grupos.

Uma matriz de correlação, que apresenta valores entre -1 e 1 foi construída. Quanto mais próximo de -1 ou 1 um valor se encontra, maior é a correlação entre os

dados. Na Figura 20, apresenta-se a correlação das variáveis EDA, HR e ST para o aluno A_1 durante as aulas ativas:

Figura 20 - Matriz de correlação



Fonte: Produção do próprio autor (2025).

A partir do gráfico na Figura 20 conclui-se que, a correlação entre as variáveis parece ser baixa para o aluno A_1. Na Tabela 6, toda a amostra foi analisada agrupando pelo tipo de aula que foi ministrada.

Tabela 6 - Média e desvio padrão geral das correlações entre as variáveis

Variável	Média Aula Ativa	Média Aula Tradicional	Desvio Padrão
HR_EDA	-0.034996	-0.036736	0.099358
HR_ST	-0.005265	0.040577	0.103152
EDA_ST	0.043993	0.096737	0.468044

Fonte: Produção próprio autor (2025).

É possível notar que, na Tabela 6 a correlação entre as variáveis é fraca. Vale ressaltar que a correlação média na aula ativa é fraca entre as variáveis HR e EDA, levemente fraca entre as variáveis HR e ST, porém apresenta uma correlação positiva entre o EDA e ST. Durante as aulas tradicionais apresenta uma correlação fraca entre o HR e EDA, porém entre o HR e ST e entre o EDA e ST apresenta-se uma correlação positiva.

Por fim, o resultado do desvio padrão foi de baixa variabilidade entre os alunos para a correlação entre o HR e EDA e entre HR e ST. Porém, para a correlação de

EDA e ST, o desvio padrão foi de alta variabilidade, indicando que essa correlação é muito diversa entre os alunos, o que mostra que esses indicadores fisiológicos são bastantes individualizados.

4.4 CRIAÇÃO DE UMA VARIÁVEL ALVO

Para a aplicação de algoritmos de *Machine Learning,* é fundamental a definição de uma variável alvo. Neste estudo, a variável alvo representará o nível de engajamento do aluno (classificado como engajado ou não engajado) durante as aulas, inferido a partir de seus dados fisiológicos.

Como já foi apresentado na Figura 14, as aulas ativas apresentaram variabilidade no tempo de conclusão das tarefas entre os alunos. Essa variação é esperada, uma vez que as metodologias ativas frequentemente envolvem atividades cujo ritmo de desenvolvimento pode diferir consideravelmente de um estudante para outro.

Diante dessa variabilidade e da necessidade de uma métrica para avaliar o engajamento, optou-se por utilizar o tempo que cada aluno despendeu para realizar as atividades propostas durante as aulas ativas. Adotou-se o pressuposto de que, neste contexto específico, o tempo de conclusão da tarefa seria inversamente proporcional ao nível de engajamento. Ou seja, um aluno que concluísse a atividade em um tempo significativamente menor que a média da turma seria considerado mais engajado, enquanto aquele que levasse um tempo consideravelmente maior seria classificado como menos engajado. É importante ressaltar que esta é uma definição operacional de engajamento adotada para os propósitos deste estudo.

Primeiro, foram selecionadas as variáveis de interesse: HR, EDA, SCL, ST, NSSCRs, EDASymp e TVSymp.

Em seguida, aplicou-se o escalonamento pelo método z-score, que é realizado pela classe *StandardScaler* da biblioteca *scikit-learn*. Esse procedimento transforma cada variável de modo que sua média seja igual a 0 e o desvio padrão seja igual a 1.

Essa padronização é importante, uma vez que as variáveis fisiológicas possuem escalas muito diferentes, sem a normalização, variáveis com valores

absolutos maiores teriam mais influência nos cálculos subsequentes, como a regressão linear.

Portanto, ao aplicar o z-score, todas as variáveis passam a estar na mesma escala relativa, o que garante comparabilidade entre elas e evita vieses no modelo. Na Figura 21, apresenta-se o trecho de código que realiza essa transformação:

Figura 21 – Código para normalizar variáveis fisiológicas

```
# ---------------------
# 1. Normalize variáveis fisiológicas
# ---------------------
variaveis = ['HR', 'EDA', 'SCL', 'ST', 'NSSCRs', 'EDASymp', 'TVSymp']

# Copia apenas os dados para normalização
dados_fisiologicos = df_final[variaveis].copy()

# Escalonamento (z-score)
scaler = StandardScaler()
dados_normalizados = scaler.fit_transform(dados_fisiologicos)
```

Fonte: Produção do próprio autor (2025).

Considerando o tempo como uma variável independente, para a construção da variável da reta, criou-se um modelo de regressão linear considerando o tempo como variável independente (coluna Seconds), ou seja, o modelo aprende e define pesos para cada variável, a fim de criar a equação da reta, como apresentado na Figura 22:

Figura 22 - Equação da reta para a regra de engajamento

```
engajamento_rule = (
    coef[0] * df_normalizado['z_HR'] +
    coef[1] * df_normalizado['z_EDA'] +
    coef[2] * df_normalizado['z_SCL'] +
    coef[3] * df_normalizado['z_ST'] +
    coef[4] * df_normalizado['z_NSSCRs'] +
    coef[5] * df_normalizado['z_EDASymp'] +
    coef[6] * df_normalizado['z_TVSymp']
)
```

Fonte: Produção do próprio autor (2025)

Por fim, a equação da reta foi gerada:

$$y = 4.0653z_{HR} - 104.7480z_{EDA} + 118.7690z_{SCL} + 35.7774z_{ST} + 19.2611z_{NSSCRs} + 2.4647z_{EDASymp} - 48.1202z_{TVSymp} + 409.9704$$

onde:

- z_{HR}→ frequência cardíaca normalizada
- z_{EDA}→ atividade eletrodérmica
- z_{SCL}→ condutância da pele
- z_{ST}→ temperatura da pele
- z_{NSSCRs}→ número de respostas galvânicas não específicas
- $z_{EDASymp}$→ componente simpático da EDA
- z_{TVSymp}→ variabilidade simpática

Criou-se uma coluna nova no banco de dados com o score de engajamento gerado pela equação da reta, com essa regra de engajamento foi quebrada em duas faixas: engajado ou não engajado. Criando assim a variável alvo chamada: nivel_engajamento.

4.5 MODELOS DE *MACHINE LEARNING*

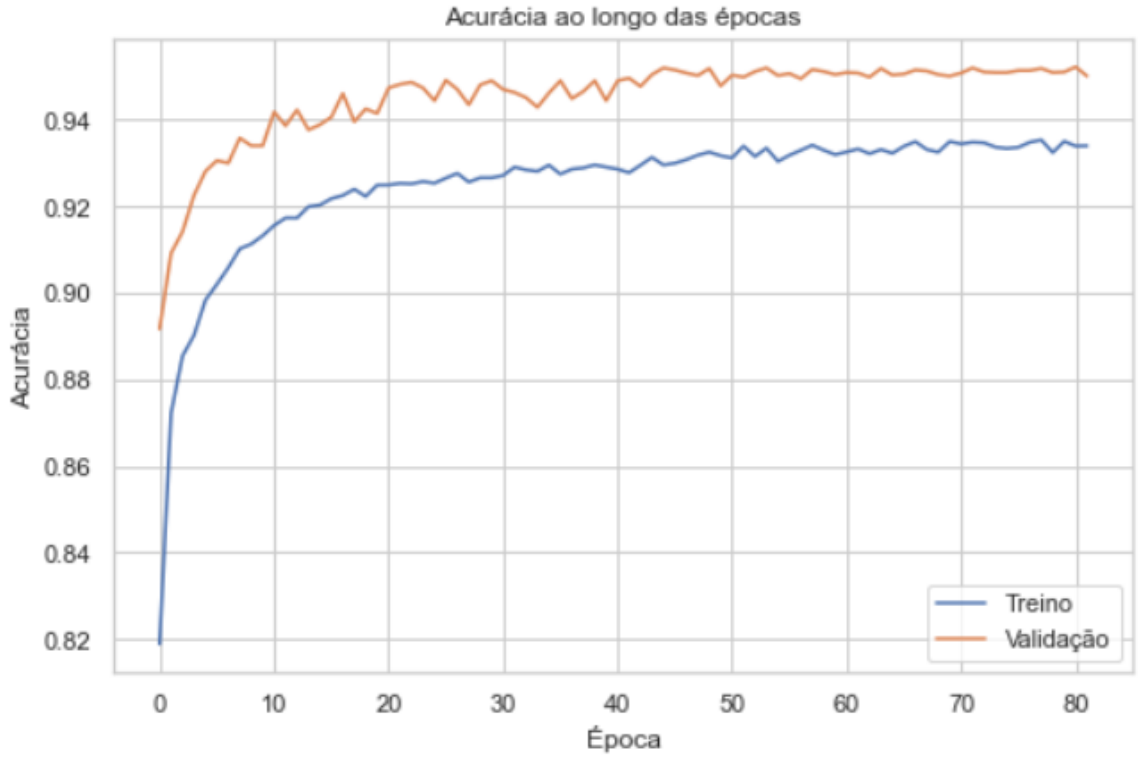
4.5.1 Redes neurais

Para a divisão dos dados, adotou-se a seguinte proporção: 70% (56402) dos dados foram destinados para a base de treino e 30% (14101) para a base de teste. Esta distribuição busca um equilíbrio entre ter dados suficientes para treinar os modelos adequadamente (base de treino) e ao mesmo tempo possuir uma quantidade significativa de dados não vistos para testar e validar o desempenho dos modelos (base de teste). Assim como foi executado no estudo de Čertický *et al.*, 2019, que dividiu os dados em doisgrupos: *subsets* de 80% para treino e 20% para teste.

4.5.2 Resultado redes neurais

Na Figura 23, apresenta-se a acurácia por épocas utilizando o algoritmo de redes neurais:

Figura 23 - Acurácia por épocas modelo de redes neurais

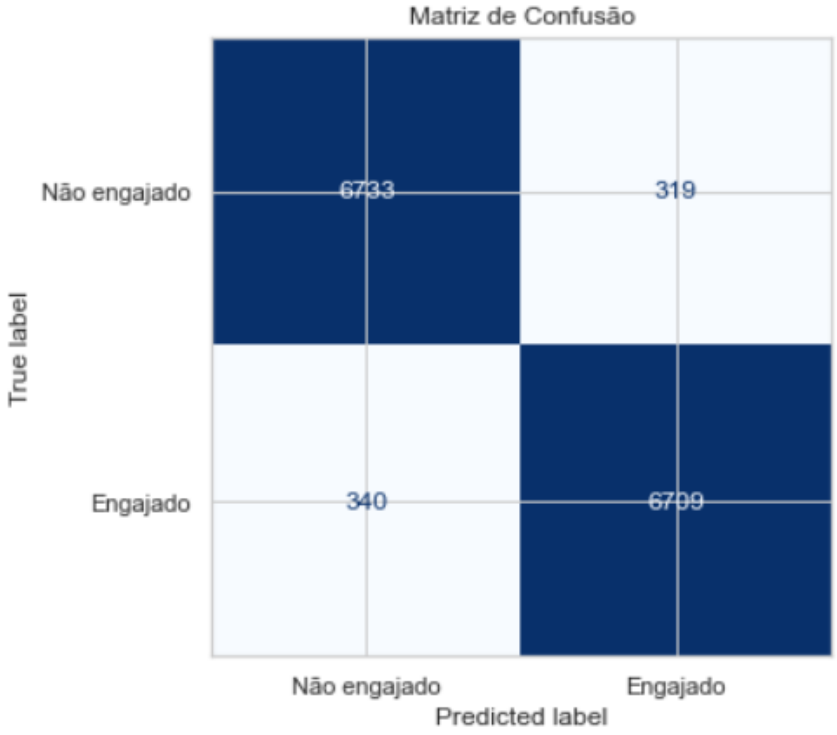


Fonte: Produção do próprio autor (2025).

O modelo apresenta alta acurácia de 95,33%, além da curva de acurácia por épocas demonstrar que o modelo não tem overfitting e generaliza bem. Desse modo pode-se afirmar que o modelo está regularizado e aprendendo os padrões reais dos dados.

A fim de entender os erros do modelo criou-se a matriz de confusão, que é apresentada na Figura 24:

Figura 24 - Acurácia por épocas modelo de redes neurais



Fonte: Produção do próprio autor (2025).

Em relação aos erros cometidos durante a classificação, observa-se:

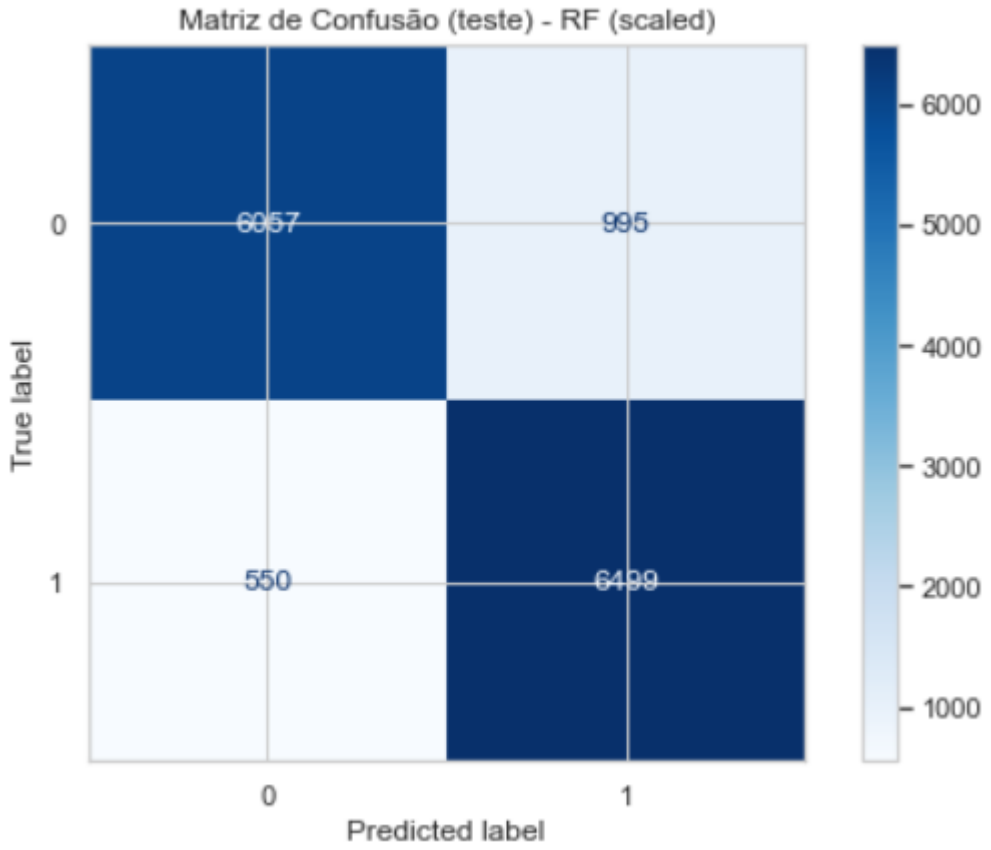
- **340 erros** ao classificar indivíduos como **não engajados**.
- **319 erros** ao classificar indivíduos como **engajados.**

A maior concentração de erros foi observada na categoria de não engajado, mas a diferença é pequena em relação aos erros para o engajado, sendo assim o modelo está consistente para as duas qualificações. Além disso, apresenta uma baixa taxa de erro, apenas 4,7% das amostras foram classificadas incorretamente.

4.5.3 *Random florest*

O modelo de floresta randômica apresenta um bom desempenho e capacidade de generalização razoável, com uma acurácia de 89,04%. Na Figura 25, apresenta-se a matriz confusão obtida utilizando o algoritmo de *random florest*:

Figura 25 - Matriz de confusão modelo de random florest



Fonte: Produção do próprio autor (2025).

Pela Figura 25 é possível notar os erros cometidos durante a classificação:

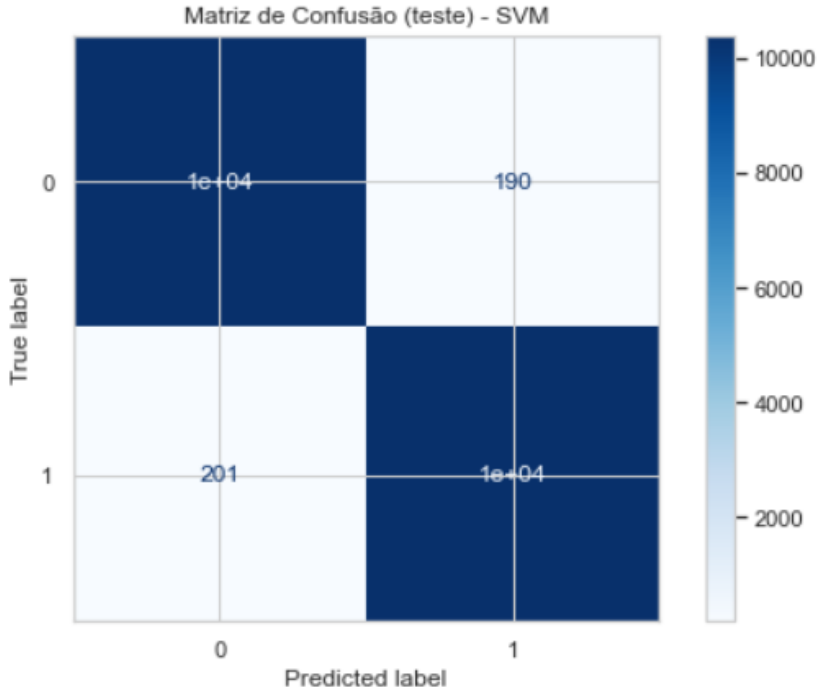
- **550 erros** ao classificar indivíduos como **não engajados**.
- **995 erros** ao classificar indivíduos como **engajados.**

Sendo assim, apesar da boa performance do modelo (89,4% de acurácia) nota-se que tende a errar mais ao classificar um indivíduo como engajado, porém acerta mais na classificação de engajados. A taxa de erro é maior do que a observada no modelo de redes neurais, sendo de 10,96%.

4.5.4 Modelo de máquina de vetores de suporte (SVM)

O modelo de SVM apresenta um ótimo desempenho e capacidade de generalização, com uma acurácia de 98,15%. Na Figura 26, apresenta-se a matriz confusão obtida utilizando o modelo:

Figura 26 - Matriz de confusão modelo SVM



Fonte: Produção do próprio autor (2025).

É possível notar que os erros cometidos durante a classificação:

- **190 erros** ao classificar indivíduos como **não engajados**.
- **201 erros** ao classificar indivíduos como **engajados.**

Sendo assim, o modelo SVM apresenta uma ótima performance e uma baixa taxa de erro (1,84%) não apresentando nenhum viés entre as classes. Sendo assim, o modelo SVM mostra-se robusto e eficaz para prever o engajamento.

5 CONCLUSÃO

Os resultados obtidos evidenciam um desempenho consistente dos algoritmos de aprendizado de máquina na classificação do nível de engajamento dos estudantes, com variações relevantes quanto à precisão e capacidade de generalização. Na Tabela 7 observa-se o resumo dos resultados levando em consideração a acurácia.

Tabela 7 – Acurácia dos modelos de aprendizado de máquina

Modelos	Acurácia
Redes neurais	95,33%
Random Florest	89,04%
SVM	98,15%

Fonte: Produção do próprio autor (2025)

A partir da Tabela 7, conclui-se que o modelo de Máquinas de Vetores de Suporte (SVM) apresentou a melhor acurácia, sendo o mais eficiente em discriminar quando um aluno é engajado ou não engajado, com uma baixa taxa de erro e erros bem equilibrados entre as categorias. O modelo de Redes Neurais também apresentou uma acurácia elevada de 95,33%, com uma taxa de erro satisfatória e equilibrada entre as categorias. Já o modelo *Random Florest* obteve a menor acurácia, 89,04% e consequentemente a maior taxa de erro, o modelo tendeu a errar mais para a categoria de engajado, logo, não apresentou um equilíbrio entre as classes.

De modo geral, os métodos aplicados foram eficazes para a tarefa de classificação proposta, sendo o modelo SVM o mais adequado para identificar padrões de engajamento a partir de dados fisiológicos, demonstrando potencial de aplicação prática em ambientes educacionais inteligentes.

REFERÊNCIAS

ANUSHA, A. S. *et al.* Electrodermal activity based pre-surgery stress detection using a wrist wearable. **IEEE Journal of Biomedical and Health Informatics**, New York, v. 23, n. 5, p. 2079-2089, 2019. DOI: 10.1109/JBHI.2019.2902128. Disponível em: https://ieeexplore.ieee.org. Acesso em: 24 ago. 2025.

BIOPAC Systems. **EDA guide**. Goleta, CA, 2018. Disponível em: https://www.biopac.com/wp-content/uploads/EDA-Guide.pdf. Acesso em: 24 ago. 2025.

CAN, Y. S. *et al.* Continuous stress detection using wearable sensors in real life: algorithmic programming contest case study. **Sensors**, Basel, v. 19, n. 8, p. 1849, 2019. DOI: 10.3390/s19081849. Disponível em: https://www.mdpi.com. Acesso em: 24 ago. 2025.

ČERTICKÝ, M. *et al.* Psychophysiological indicators for modeling user experience in interactive digital entertainment. **Sensors**, Basel, v. 19, n. 5, p. 989, 2019. DOI: 10.3390/s190508989. Disponível em: https://www.mdpi.com. Acesso em: 24 ago. 2025.

CHOWDHURY, A. K. *et al.* Prediction of relative physical activity intensity using multimodal sensing of physiological data. **Sensors**, Basel, v. 19, n. 20, p. 4509, 2019. DOI: 10.3390/s19204509. Disponível em: https://www.mdpi.com. Acesso em: 24 ago. 2025.

FIORINI, L. *et al.* Unsupervised emotional state classification through physiological parameters for social robotics applications. **Knowledge-Based Systems**, Amsterdam, v. 190, p. 105217, 2020. DOI: 10.1016/j.knosys.2019.105217. Disponível em: https://www.sciencedirect.com. Acesso em: 24 ago. 2025.

DALE LUCHE, J. R. *et al.* **Conjunto de dados fisiológicos para análise de engajamento em sala de aula**. Versão 1. [Conjunto de dados]. 2025. DOI: 10.17632/w9tcfhx9cn.1. Disponível em: https://data.mendeley.com/datasets/w9tcfhx9cn/1. Acesso em: 25 ago. 2025.

DRAMSCH, J. S. 70 years of machine learning in geoscience in review. **Advances in Geophysics**, New York, v. 61, p. 1-55, 2020. DOI: 10.1016/bs.agph.2020.08.002. Disponível em: https://www.sciencedirect.com. Acesso em: 24 ago. 2025.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2002.

GJORESKI, M. *et al.* Machine learning and end-to-end deep learning for monitoring driver distractions from physiological and visual signals. **IEEE Access**, New York, v. 8, p. 70590-70603, 2020. DOI: 10.1109/ACCESS.2020.2988711. Disponível em: https://ieeexplore.ieee.org. Acesso em: 24 ago. 2025.

HAN, J.; PEI, J.; KAMBER, M. **Data mining:** concepts and techniques. Amsterdam: Elsevier, 2011.

KIM, A. Y. *et al.* Automatic detection of major depressive disorder using electrodermal activity. **Scientific Reports**, London, v. 8, p. 17030, 2018. DOI: 10.1038/s41598-018-35147-3. Disponível em: https://www.nature.com. Acesso em: 24 ago. 2025.

LEE, H. *et al.* Distinguishing anxiety subtypes of English language learners towards augmented emotional clarity. *In:* INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE IN EDUCATION, 2020. **Proceedings** […]. Cham: Springer, 2020. p. 157-161. DOI: 10.1007/978-3-030-52240-7_12. Disponível em: https://link.springer.com. Acesso em: 24 ago. 2025.

LIMA, R.; DE NORONHA OSÓRIO, D. F.; GAMBOA, H. Heart rate variability and electrodermal activity in mental stress aloud: predicting the outcome. *In:* BIOSIGNALS, 2019. **Proceedings** […]. Cham: Springer, 2019. p. 42-51. DOI: 10.1007/978-3-030-46970-2_16. Disponível em: https://link.springer.com. Acesso em: 24 ago. 2025.

LIU, Y. *et al.* A study of human skin and surface temperatures in stable conditions. **Journal of Thermal Biology**, Oxford, v. 38, n. 6, p. 305-312, 2013. DOI: 10.1016/j.jtherbio.2012.11.016. Disponível em: https://www.sciencedirect.com. Acesso em: 24 ago. 2025.

MIGUEL, P. A. C. *et al.* **Metodologia de pesquisa em engenharia de produção e gestão de operações**. 2. ed. Rio de Janeiro: Elsevier, 2018.

NKURIKIYEZU, K.; YOKOKUBO, A.; LOPEZ, G. Importance of individual differences in physiological-based stress recognition models. *In*: INTERNATIONAL CONFERENCE ON INTELLIGENT ENVIRONMENTS, 15., 2019, Mahdia. **Proceedings** […]. Mahdia: IEEE, 2019. p. 37-43. DOI: 10.1109/IE.2019.00011. Disponível em: https://ieeexplore.ieee.org. Acesso em: 24 ago. 2025.

OLSHANSKY, B.; RICCI, F.; FEDOROWSKI, A. Importance of resting heart rate. **Trends in Cardiovascular Medicine**, Amsterdam, v. 33, n. 8, p. 502-515, 2023 Nov. DOI: 10.1016/j.tcm.2022.05.006. Disponível em: https://pubmed.ncbi.nlm.nih.gov/35623552/. Acesso em: 25 ago. 2025.

POSADA-QUINTERO, H. F.; BOLKHOVSKY, J. B. Machine learning models for the identification of cognitive tasks using autonomic reactions from heart rate variability and electrodermal activity. **Behavioral Sciences**, Chichester, v. 9, n. 4, p. 45, 2019. DOI: 10.3390/bs9040045. Disponível em: https://www.mdpi.com. Acesso em: 24 ago. 2025.

POWERS, D. M. W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. **Journal of Machine Learning Technologies**, Amsterdam, v. 2, n. 1, p. 37-63, 2011. DOI: 10.602/jmlt.v2i1.37. Disponível em: https://www.jmlt.org. Acesso em: 24 ago. 2025.

REGALIA, G. *et al.* Multimodal wrist-worn devices for seizure detection and advancing research: focus on the Empatica wristbands. **Epilepsy Research**,

Amsterdam, v. 153, p. 79-82, 2019. DOI: 10.1016/j.eplepsyres.2019.01.005. Disponível em: https://www.sciencedirect.com. Acesso em: 24 ago. 2025.

RIZWAN, A. *et al.* Non-invasive hydration level estimation in human body using galvanic skin response. **IEEE Sensors Journal**, New York, v. 20, n. 13, p. 3740-3749, 2020. DOI: 10.1109/JSEN.2020.2972944. Disponível em: https://ieeexplore.ieee.org. Acesso em: 24 ago. 2025.

SABETI, E. *et al.* Learning using concave and convex kernels: applications in predicting quality of sleep and level of fatigue in fibromyalgia. **Entropy**, Basel, v. 21, n. 5, p. 442, 2019. DOI: 10.3390/e21050442. Disponível em: https://www.mdpi.com. Acesso em: 24 ago. 2025.

SARKER, I. H. Machine learning: algorithms, real-world applications and research directions. **SN Computer Science**, Cham, v. 2, p. 160, 2021. DOI: 10.1007/s42979-021-00592-X. Disponível em: https://link.springer.com. Acesso em: 24 ago. 2025.

SHARMA, V.; PRAKASH, N. R.; KALRA, P. Audio-video emotional response mapping based upon electrodermal activity. **Biomedical Signal Processing and Control**, Amsterdam, v. 47, p. 324-333, 2019. DOI: 10.1016/j.bspc.2019.07.002. Disponível em: https://www.sciencedirect.com. Acesso em: 24 ago. 2025.