Adil Schuck da Silva

Desempenho Acadêmico sob a Perspectiva do Machine Learning: Uma Abordagem Analítica

Relatório Técnico elaborado conforme a ABNT NBR 10719:10, apresentado ao Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, como parte dos requisitos para a obtenção do grau de Tecnólogo em Sistemas para Internet.

Área de Concentração: Ciência de Dados

Orientador: Prof. Dr. Ricardo Alexandre Neves

São João da Boa Vista

2025

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE SÃO PAULO - CÂMPUS SÃO JOÃO DA BOA VISTA		2025

Tecnologia em Sistemas para Internet

Desempenho Acadêmico sob a Perspectiva do Machine Learning: Uma Abordagem Analítica

Adil Schuck da Silva e
Prof. Dr. Ricardo Alexandre Neves

Palavras-chave:
Desempenho Acadêmico; Aprendizado de Máquina; Análise de dados; *Ensemble learning*; Regressão logística; *Apriori*; *K-means*.

51 páginas

Ficha Catalográfica elaborada pela Biblioteca Comunitária
"Wolgran Junqueira Ferreira" do Instituto Federal de São Paulo
Câmpus São João da Boa Vista

S586d

Silva, Adil Schuck da
Desempenho acadêmico sob a perspectiva do machine learning: uma abordagem analítica / Adil Schuck da Silva; orientador: Ricardo Alexandre Neves. -- São João da Boa Vista, 2025.
52 f. : il.

Trabalho de Conclusão de Curso (Tecnologia em Sistemas para Internet) -- Instituto Federal de Educação, Ciência e Tecnologia de São Paulo, São João da Boa Vista, 2025.

1. Aprendizado de máquina. 2. Ciência de dados. 3. Comunidade acadêmica - Análise de dados. I. Neves, Ricardo Alexandre, orient. II. Título.

IFSP / SBV

CDD 006.3

ATA N.º 97/2025 - DAE-SBV/DRG/SBV/IFSP

Ata de Defesa de Trabalho de Conclusão de Curso - Graduação

Na presente data realizou-se a sessão pública de defesa do Trabalho de Conclusão de Curso intitulado **Desempenho Acadêmico sob a Perspectiva do Machine Learning: Uma Abordagem Analítica** apresentado pelo aluno **Adil Schuck da Silva (BV3031861)** do Curso **SUPERIOR EM TECNOLOGIA EM SISTEMAS PARA INTERNET,** Campus São João da Boa Vista. Os trabalhos foram iniciados às 18h pelo(a) Professor(a) presidente da banca examinadora, constituída pelos seguintes membros:

Membros	IES	Presença (Sim/Não)	Aprovação/Conceito (Quando Exigido)
Ricardo Alexandre Neves (Presidente/Orientador)	IFSP/SBV	Sim	Aprovado
Diego Cesar Valente e Silva (Examinador 1)	IFSP/SBV	Sim	Aprovado
Luiz Angelo Valota Francisco (Examinador 2)	IFSP/SBV	Sim	Aprovado

Observações:

A banca examinadora, tendo terminado a apresentação do conteúdo da monografia, passou à arguição do(a) candidato(a). Em seguida, os examinadores reuniram-se para avaliação e deram o parecer final sobre o trabalho apresentado pelo(a) aluno(a), tendo sido atribuído o seguinte resultado:

[X] Aprovado(a) [ ] Reprovado(a) Nota (quando exigido): _____

Proclamados os resultados pelo presidente da banca examinadora, foram encerrados os trabalhos e, para constar, eu lavrei a presente ata que assino juntamente com os demais membros da banca examinadora.

Campus São João da Boa Vista, 3 de dezembro de 2025

Avaliador externo: [ ]Sim [X]Não

Assinatura:

Documento assinado eletronicamente por:

- **Ricardo Alexandre Neves**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 03/12/2025 23:59:33.
- **Diego Cesar Valente e Silva**, **DIRETOR(A) GERAL - CD2 - DRG/SBV**, em 04/12/2025 13:28:07.
- **Luiz Angelo Valota Francisco**, **PROFESSOR ENS BASICO TECN TECNOLOGICO** , em 11/12/2025 10:37:19.

Este documento foi emitido pelo SUAP em 03/12/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifsp.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 1076215
Código de Autenticação: a4247ec4a4

SUMÁRIO

RESUMO

Na era digital, a crescente disponibilidade de dados no ambiente educacional tem impulsionado a busca por métodos inovadores para otimizar o processo de ensino-aprendizagem. No entanto, as ferramentas tradicionais mostram-se insuficientes para lidar com o volume e a complexidade das informações geradas. Diante da subutilização desses dados, o Aprendizado de Máquina (AM) surge como uma solução promissora. Este trabalho tem como objetivo investigar, por meio de algoritmos de aprendizado de máquina, as correlações e associações entre os fatores que influenciam a melhoria da eficiência acadêmica no Instituto Federal de São Paulo. A identificação desses fatores visa contribuir para a tomada de decisões estratégicas, gerando benefícios para a comunidade acadêmica e, consequentemente, possibilitando o aumento de recursos destinados à instituição de ensino. Para isso, foram utilizadas técnicas como estatística descritiva, Regressão Logística, *Ensemble Learning* (AdaBoost), *Apriori* e *K-means*. Optou-se por utilizar uma base de dados de Eficiência Acadêmica e uma base de dados de dados de Matriculas, ambas do MEC, com foco nos campi do Instituto Federal de São Paulo, no período de 2017 a 2024. Os resultados indicaram alta precisão nas previsões: resultados de até 99,9% para a técnica de Regressão Logística, próximas do 100% para a técnica *AdaBoost* e também próximas ao 100% para o *Apriori*. Tais resultados revelaram uma forte correlação inversa entre o Índice de Eficiência Acadêmica (IEA) e a Taxa de Evasão, bem como uma correlação direta entre o IEA e a Taxa de Conclusão, assim como a baixa presença do IEA em cursos de maior extensão. Concluiu-se, até o momento, que o combate à evasão é fundamental para elevar os índices de IEA, sendo a Taxa de Conclusão um fator essencial para fornecer *insights* relevantes à melhoria da qualidade do ensino e à tomada de decisões no ambiente educacional.

Palavras-chave: Desempenho Acadêmico; Aprendizado de Máquina; Análise de dados; *Ensemble Learning*; Regressão Logística; *Apriori*; *K-means*

1 INTRODUÇÃO

A crescente disponibilidade de dados no ambiente educacional tem impulsionado a busca por métodos inovadores para otimizar o processo de ensino-aprendizagem e, consequentemente, o desempenho acadêmico dos estudantes. Ferramentas tradicionais de análise de dados frequentemente se mostram insuficientes para lidar com o volume e a complexidade das informações geradas, evidenciando a necessidade de abordagens mais sofisticadas. Nesse contexto, como apresentado por Silva et al. (2025a), o Aprendizado de Máquina (AM) surge como uma solução promissora, oferecendo a capacidade de extrair padrões e *insights* valiosos a partir de grandes *datasets* educacionais.

Percebe-se que há uma lacuna entre a vasta quantidade de dados educacionais disponíveis, conforme destacado por Owusu-Boadu et al. (2021), e a subutilização desses dados para aprimorar efetivamente o processo de ensino, como evidenciado por Kaur, Gupta e Singla (2023). Esse aprimoramento é essencial tanto para o desenvolvimento das instituições de ensino quanto para os alunos, que muitas vezes se sentem desestimulados ao enfrentarem obstáculos no aprendizado.

A aplicação de técnicas avançadas de Aprendizado de Máquina possibilita não apenas a identificação dos fatores que influenciam o desempenho acadêmico, mas também o desenvolvimento de ferramentas de intervenção que podem auxiliar alunos, educadores e instituições a tomarem decisões mais informadas, personalizando o aprendizado e, em última instância, elevando a qualidade da educação, conforme sugerido por Silva et al. (2025a) e Duong et al. (2023).

Esta pesquisa busca identificar os principais fatores que afetam a eficiência acadêmica, utilizando uma abordagem qualitativa para relacionar os dados educacionais com os métodos de AM. A metodologia proposta será desenvolvida em etapas bem definidas, incluindo: pesquisa bibliográfica, análise de técnicas, escolha de bases de dados, aplicação das técnicas e análise dos resultados.

É importante ressaltar que a flexibilidade metodológica permitirá a adoção de técnicas alternativas caso as inicialmente aplicadas se mostrem ineficientes ou insuficientes para alcançar os resultados esperados.

1.1 Motivação

A motivação para a realização deste estudo está relacionada à necessidade de aprimorar o acompanhamento e a avaliação da eficiência acadêmica no Instituto Federal de São Paulo, incluindo o câmpus de São João da Boa Vista, especialmente diante da

limitação de instrumentos que permitam explorar de forma sistemática e aprofundada os dados educacionais disponíveis. O crescimento no volume de informações acadêmicas exige abordagens analíticas mais avançadas, como o uso de técnicas de aprendizado de máquina, capazes de identificar padrões e gerar conhecimento relevante para apoiar a tomada de decisão na gestão escolar. Ao promover uma análise mais qualificada dos indicadores acadêmicos, o trabalho busca contribuir para a melhoria contínua da qualidade do ensino e do processo de aprendizagem, beneficiando diretamente a comunidade acadêmica. Como resultado do fortalecimento dos indicadores institucionais e do desempenho acadêmico, espera-se, de forma consequente, a ampliação dos recursos financeiros destinados ao câmpus de São João da Boa Vista, possibilitando investimentos em infraestrutura, materiais e equipamentos educacionais, o que reforça e sustenta as melhorias alcançadas no contexto educacional.

1.2 Objetivos

Este trabalho tem por objetivo investigar, por meio de algoritmos de aprendizado de máquina, as correlações e associações do desenvolvimento de um mecanismo que por meio deste será possível identificar os pontos de melhoria no câmpus de São João da Boa Vista - visando atuar em ações de planejamento pedagógico e de gestão escolar -, identificando os fatores que influenciam os índices de qualidade considerados neste câmpus do Instituto Federal de São Paulo, para que seja possível atuar na melhoria da qualidade do ensino e do processo de aprendizagem.

Para alcançar o objetivo geral, os seguintes objetivos específicos serão desenvolvidos:

- Levantar e pré-processar os dados acadêmicos que compõem o escopo deste estudo, preparando-os para a aplicação dos algoritmos.
- Aplicar modelos de aprendizado de máquina para predizer a eficiência acadêmica e identificar correlações entre as variáveis.
- Avaliar o desempenho e a acurácia dos modelos aplicados por meio de métricas de performance estatísticas.
- Identificar e hierarquizar as variáveis de maior impacto no desempenho acadêmico, com base nos resultados do modelo mais eficaz.

2 CONSIDERAÇÕES GERAIS

Este Capítulo apresenta uma análise geral sobre o tema pesquisado, fornecendo a base teórica necessária para a compreensão do estudo. Inicialmente, é realizada uma pesquisa bibliográfica que aborda conceitos fundamentais relacionados à Mineração de Dados, Aprendizado de Máquina e suas aplicações no contexto educacional, com ênfase na predição de desempenho acadêmico. São discutidas técnicas, métodos e ferramentas relevantes, além de aspectos relacionados ao processamento e à análise de grandes volumes de dados educacionais. Em seguida, são apresentados trabalhos correlatos que contribuem para o embasamento científico deste estudo, destacando as abordagens mais recentes e os principais resultados encontrados na literatura.

2.1 Pesquisa Bibliográfica

De acordo com Raju et al. (2020), a Mineração de Dados (MD) é uma das melhores técnicas para o reconhecimento de padrões em grandes quantidades de dados, tendo como primeiro passo observar os dados sob diferentes perspectivas e encontrar as informações mais valiosas da forma mais direta possível (Kamal et al., 2022). Além disso, segundo Al-Alawi et al. (2023), a MD possui grande potencial, principalmente por identificar informações ocultas dentro de grandes bancos de dados.

A Mineração de Dados Educacionais (MDE) é uma área de pesquisa dentro da Mineração de Dados que foca em extrair correlações a partir de grandes volumes de dados acadêmicos, descobrindo padrões, tendências e associações. Dentro da MDE, a predição de notas é um dos principais fatores considerados, uma vez que as notas são indicadores diretos da capacidade dos alunos de aprender os conteúdos apresentados no ambiente educacional (Sun et al., 2023).

Os dados educacionais podem ser extraídos de plataformas de aprendizagem, sistemas acadêmicos e relatórios de professores, geralmente contendo diferentes níveis de importância das informações, aspecto que deve ser analisado com base nas propriedades dos dados (Raju et al., 2020).

A pesquisa de Czibula et al. (2022) apresenta o desenvolvimento de um *framework* baseado no Aprendizado de Máquina (AM) aplicado ao contexto da MDE, denominado IntelliDaM, que atua na mineração de dados de performance acadêmica e oferece três tipos de componentes de análise: análise e seleção de conteúdos, análise de dados não supervisionados e modelos preditivos baseados em dados supervisionados. No desenvolvimento do IntelliDaM, é destacado que um de seus principais focos é a predição da performance dos

estudantes, fator essencial para a eficácia do *framework* proposto.

Segundo Zhang et al. (2024), graças à presença da internet em todos os âmbitos sociais atuais, todas as ações dos estudantes geram dados comportamentais que podem ser utilizados por outra área que examina as atividades acadêmicas dos alunos: a Análise de Aprendizagem (AA). Essa área utiliza dados coletados de diferentes fontes e busca correlações que conectam a performance e as ações dos estudantes, tendo como núcleo o monitoramento do desempenho ao longo do curso. O estudo também destaca que a rede local do câmpus analisado gera um grande acúmulo de dados diariamente, refletindo informações relevantes sobre a linha de pensamento dos estudantes, suas emoções e suas dinâmicas comportamentais.

De acordo com Kaur, Gupta e Singla (2023), ferramentas e técnicas básicas de análise de dados não possuem a capacidade necessária para processar grandes volumes de dados acadêmicos. Dessa forma, é apresentado um método de discretização de dados, que consiste em dividir os dados contínuos em pequenas partes ou categorias durante o pré-processamento, facilitando o entendimento das distribuições de valores e dos domínios analisados. O método de discretização está inserido na área da Mineração de Dados Educacionais (MDE). Desenvolvendo esse pensamento sobre o processamento de dados, Sihare e Gupta (2024) afirma que analisar grandes quantidades de informações é uma tarefa difícil para o ser humano e, por isso, seu estudo tem como objetivo utilizar a tecnologia para automatizar métodos convencionais de processamento de dados, de ensino pelos professores e de compreensão pelos alunos.

O processamento de dados é uma etapa crucial para prever a performance dos estudantes em ambientes educacionais (Alhakami; Alsubait; Aljarallah, 2020). Segundo Al-Alawi et al. (2023), para obter sucesso na previsão de fatores que afetam o desempenho acadêmico, é essencial identificar, estudar e analisar os problemas, visando melhorar a qualidade do ensino, levando em consideração diferentes tipos de informações educacionais, culturais, demográficas e psicológicas (Contreas-Bravo; Nieves-Pimiento; Guerrero, 2023). O estudo de (Silva et al., 2025a) comprovou que alunos que participaram de processos de tutoria obtiveram, em média, um aumento de 0,3 pontos em suas notas finais, demonstrando a importância da aplicação de metodologias adequadas após a mineração dos dados.

As diferenças no desempenho entre modelos de processamento de dados também são evidenciadas no estudo de Silva et al. (2025b), no qual fica claro que escolher o modelo adequado, com base no contexto, é essencial para obter os resultados desejados. Esse processamento pode ser realizado com o auxílio de diversos algoritmos e métodos presentes no âmbito do Aprendizado de Máquina, como, por exemplo:

- **Regressão Logística:** A predição da performance dos estudantes também está associada ao método de Regressão Logística (RL), citada por Silva et al. (2025a) como

um modelo linear que se destaca pela capacidade de estimar probabilidades. O artigo de Sagala et al. (2022) aplicou a RL no contexto do curso de Ciências da Computação da Universidade Bina Nusantara, na Indonésia, no período de 2010 a 2020. O estudo concluiu que, em 70% dos trabalhos correlatos analisados sobre RL, o principal indicador acadêmico utilizado é o índice de rendimento acadêmico, enquanto menos de 1% das pesquisas utilizaram dados provenientes de informações demográficas e atividades estudantis. Diferentemente desses trabalhos, o estudo de Sagala et al. (2022) considera diversos fatores que influenciam a performance estudantil, como o emprego dos pais ou responsáveis pelo aluno e o gênero do estudante. A RL obteve uma precisão de 91% em suas predições (Sagala et al., 2022).

Outros trabalhos, como o de Maheshwari et al. (2024), demonstram que a RL também é útil para a previsão da taxa de evasão dos estudantes, fator diretamente relacionado à sua performance, que pode ser mitigado com previsões bem executadas e com programas de apoio ao aprendizado. Esse estudo também destaca outros fatores que influenciam a predição, como conquistas educacionais e ambições dos próprios alunos, indicando que a ausência de influências culturais e econômicas significativas pode impactar negativamente os estudantes.

Assim como Sagala et al. (2022), o artigo de Lawanont e Timtong (2022) apresentou um estudo realizado na Alemanha, demonstrando que, apesar dos benefícios da digitalização das mídias e dos conteúdos, ainda existem obstáculos a serem superados para que se consiga implementar uma transformação digital eficaz na educação de qualidade. O estudo constatou que a automatização desses processos é tão importante nas universidades quanto no setor privado. Esses trabalhos consideram dados como gênero, faixa etária e status escolar, e apontam que um dos fatores mais relevantes é o retorno que as instituições podem oferecer aos alunos com menores índices de rendimento acadêmico. Os resultados de Lawanont e Timtong (2022) indicam que as notas finais apresentam correlação com os fatores analisados. A partir dessa perspectiva, as instituições podem oferecer retornos personalizados aos estudantes, com base na análise de seus fatores individuais. Caso o aluno responda positivamente a esse retorno, é possível esperar uma melhora em seu desempenho acadêmico.

- ***Ensemble Learning:*** O *Ensemble Learning* (EL) é uma técnica que combina diversos algoritmos de aprendizado com o objetivo de construir um modelo de predição com desempenho superior ao de algoritmos individuais. Esse método geralmente utiliza dados provenientes de indicadores acadêmicos, como o Índice de Rendimento Acadêmico Semestral, o Índice de Rendimento Acadêmico Acumulado e o desempenho acadêmico geral. Segundo Devkishan, Singh e Bharti (2024), um modelo de EL proposto em sua pesquisa supera outros modelos de predição baseados em EL, com destaque para o seu próprio modelo e para o algoritmo Random Forest.

O Random Forest é considerado um dos principais modelos de predição, com alta precisão, junto de modelos como o Support Vector Machine (SVM) e a RL, seguidos por algoritmos que apresentam desempenho consideravelmente inferior, como o Naive Bayes, as Árvores de Decisão e o KNN. De acordo com Wiradinata et al. (2021), o algoritmo SVM possui uma acurácia ligeiramente superior à do algoritmo Naive Bayes, alcançando resultados de 86,82% de precisão. O desempenho acadêmico pode ser mais bem previsto utilizando modelos de aprendizado de máquina de última geração, especialmente abordagens baseadas em *ensemble*, em comparação com modelos mais convencionais, conforme destacado por Maheshwari et al. (2024).

Outro algoritmo bastante citado é o *XG Boost*, que, segundo Sihare e Gupta (2024), apresentou precisão de 97%. O *XG Boost* foi comparado com outros algoritmos, como o AdaBoost e o MLP, e obteve os melhores resultados nos testes realizados.

- ***Apriori e K-means:*** O estudo de Li e Zhang (2024) consiste no desenvolvimento de um modelo de monitoramento da qualidade de ensino, utilizando os algoritmos *Apriori* e *K-means*, onde escolha do *K-means* é justificada pela sua eficiência, escalabilidade e natureza não supervisionada, o que se mostra ideal para explorar a estrutura intrínseca de grandes volumes de dados de ensino, revelando potenciais agrupamentos. O *Apriori*, que é uma técnica clássica de mineração de regras de associação utilizada para descobrir padrões frequentes e regras de associação entre os dados, é introduzido após o *K-means* para minerar as regras de associação entre os indicadores acadêmicos, identificando problemas e correlações no processo de ensino. Este estudo obteve uma precisão média de 93,79% e uma eficiência de 96,07%, concluindo que a aplicação desses algoritmos melhora o monitoramento do ensino.

Assim como abordado nos estudos sobre RL, é importante observar que fatores externos influenciam o processamento e os resultados dos dados analisados. Por exemplo, o estudo de Alhakami, Alsubait e Aljarallah (2020) demonstra que a idade dos alunos interfere em seu desempenho: na Universidade de Umm Al-Qura, a maioria dos alunos com idade entre 30 e 39 anos obteve notas médias, enquanto a maioria dos alunos entre 16 e 22 anos alcançou notas excelentes. Técnicas e ferramentas devem considerar todos os fatores relevantes na predição do desempenho dos alunos, uma vez que essas variáveis podem impactar significativamente os resultados em diferentes contextos (Contreas-Bravo; Nieves-Pimiento; Guerrero, 2023).

De acordo com Owusu-Boadu et al. (2021), o avanço da tecnologia trouxe um aumento exponencial na coleta de dados educacionais, o que sugere que as plataformas educacionais podem ser uma peça muito importante na análise e previsão do desempenho acadêmico. O crescimento dos Ambientes Virtuais (AV) durante a pandemia global da COVID-19 é exposto pelo estudo de Baessa et al. (2024), no qual é destacada a importância

dessas ferramentas de Ensino a Distância (EaD). Um modelo de ensino diretamente relacionado aos AV que ganhou muita popularidade após a pandemia foram os cursos online, com destaque para o *Massive Online Open Course* (MOOC), modelo que apresenta cursos abertos ao público com o intuito de alcançar um grande número de pessoas e que armazena, de maneira completa, os comportamentos de professores e alunos, gerando informações contínuas sobre o aprendizado (Guo et al., 2021).

O estudo foi aplicado em 1.426 alunos na plataforma Superstar da Universidade de Ciência e Tecnologia de Wuhan, onde, graças ao MOOC, foram obtidas informações sobre os estudantes do curso de Ciências da Computação, como, por exemplo, que os alunos tendem a estudar no período entre as 8 horas da manhã e o meio-dia. Também é destacado que, utilizando o MOOC, existe a possibilidade de aplicação de algoritmos de previsão para melhorar a qualidade do aprendizado.

Duong et al. (2023) sugere a aplicação de um sistema de dois avisos aos alunos: o primeiro no início do semestre e o segundo antes das provas finais. O primeiro teria como objetivo atualizar os alunos em relação ao seu desempenho atual, permitindo que se ajustem desde o começo do semestre; o segundo aviso teria como finalidade atualizar as pendências restantes, visando o sucesso dos alunos na conclusão do semestre. Segundo o estudo, esses avisos podem ajudar tanto os alunos quanto os professores na tarefa do aprendizado, atuando o mais cedo possível na raiz do problema.

Diversos estudos mencionam fatores que influenciam diretamente o desempenho dos estudantes e de algoritmos que tentam prever esse desempenho, como, por exemplo, (Owusu-Boadu et al., 2021), que constata a importância do engajamento dos alunos nas atividades em classe, onde a maior parte dos alunos que participam obtém desempenho melhor que os demais. Estudos sobre o assunto, como mencionado por Zhang et al. (2024), devem sempre se basear em estudos empíricos e científicos, destacando a importância de usar os dados da maneira mais eficiente possível. Fatores culturais, demográficos e psicológicos possuem relação direta com o desempenho acadêmico dos alunos, de modo que mesmo fatores menos considerados, como idade e gênero, possuem relevância (Al-Alawi et al., 2023). A pesquisa de Maheshwari et al. (2024) acentua também outros fatores que influenciam a predição, como conquistas educacionais e ambições dos próprios alunos.

O estudo de Silva et al. (2025a) aprofunda-se nas variáveis de tutoria e apoio parental, destacando o impacto, pois alunos que passaram pela tutoria apresentaram 0,3 pontos a mais na média final, enquanto o apoio parental diminuiu em 22% o risco de reprovação ao final do semestre. É citado também no estudo a importância da presença do aluno, onde 83% dos alunos que obtiveram mais de 15 faltas no semestre apresentaram notas abaixo de 2,0. De acordo com Silva et al. (2025b), um dos principais desafios do ensino superior no Brasil é a evasão escolar, que está associada a desigualdades econômicas e falhas estruturais no ensino, de modo que compreender e prever os fatores que influenciam

a evasão se mostra essencial para o desenvolvimento de estratégias de retenção do aluno.

Assim como existem categorias de dados que influenciam de maneira mais assertiva, os bancos de dados geralmente possuem alta densidade de dados e, quase em todos os casos, atributos que não têm função importante na previsão, destacando a necessidade de uma seleção cautelosa dos dados mais relevantes (Khan et al., 2021).

Um ponto a ser considerado é a região geográfica e o contexto cultural do estudo, fatores que podem interferir na diversidade dos dados coletados, como destacado por (Zhang et al., 2024). A diversidade dos dados pode ser prejudicada por limitações na coleta, como o exemplo de Silva et al. (2025b), que restringiu sua coleta a apenas um curso, limitando a generalização do modelo da pesquisa.

Com base nessas informações, fica claro que a MDE e o processamento dos dados devem considerar informações-chave que possuem maior influência no resultado da pesquisa, como, por exemplo, alguns indicadores acadêmicos (GPA, SGPA, desempenho acadêmico), citados por Sathe e Adamuthe (2021), que também indica que fatores de confiança podem aumentar drasticamente a performance de um aluno. Benevento e Meirelles (2023) destaca a Previsão do Desempenho dos Alunos (PDA) como importante para a melhoria do desempenho estudantil.

Algo a ser destacado também é o Índice Geral de Cursos (IGC), que aponta um indicador obtido a partir da média das notas dos cursos superiores de cada instituição de ensino superior brasileira (Rodrigues et al., 2023), onde é apresentado o nível de qualidade e, consequentemente, a dificuldade nos cursos superiores no Brasil, observando-se que cursos mais bem qualificados tendem a possuir maior taxa de reprovação de alunos e menor Índice de Rendimento Acadêmico Semestral (SGPA).

Um tópico que não pode ser deixado de lado é a Inteligência Artificial (IA), que, segundo Yousuf, Wahid e Khan (2023), recentemente se tornou uma ferramenta-chave que pode, e vai, revolucionar diversos setores. Ferramenta atualmente presente em nosso dia a dia, a IA se mostra eficaz em aprimorar a performance dos métodos e modelos já estudados, sendo especialmente relevante em temas como a previsão de desempenho acadêmico, pois, ao identificar precocemente possíveis dificuldades, é possível prevenir ou reduzir a taxa de evasão dos alunos (Almasri et al., 2022).

A IA também é apontada como uma ferramenta que, na educação, pode ser aplicada de diversas maneiras, como o uso de chatbots de suporte ao aluno e sistemas de tutoria que se adaptam à aprendizagem de cada estudante, mas que possui também obstáculos significativos em sua implementação, desde questões éticas até a falta de formação dos docentes, que podem encontrar dificuldades na aplicação dessas ferramentas (Ribeiro, 2024).

Dentro do âmbito da IA, o estudo de Benevento e Meirelles (2023) mostra a

aplicação do Generative Pre-trained Transformer (GPT), ferramenta desenvolvida pela OpenAI em 2018 que ganhou muita popularidade principalmente na segunda metade de 2024, combinada com o Aprendizado de Máquina para prever e melhorar o desempenho de alunos em tarefas educacionais, destacando a melhora não somente do desempenho dos alunos, como também do ensino dos professores. O estudo também ressalta que é possível utilizar o GPT para melhorar o aprendizado, com sugestões para alunos que ficaram abaixo do esperado ao longo do período. Apesar dos benefícios citados do GPT, é necessário ressaltar também os problemas envolvidos nessa ferramenta, que pode sofrer com deficiências de viés, graças às fontes de dados que a abastecem, as quais muitas vezes não são checadas e contêm preconceitos e desigualdades (Benevento; Meirelles, 2023).

2.2 Trabalhos Correlatos

Durante a pesquisa bibliográfica, foram identificados alguns trabalhos correlatos que possuem maior semelhança com a abordagem escolhida, como, por exemplo, o de Benevento e Meirelles (2023), que trata do uso combinado de aprendizado de máquina e GPT para tentar alcançar uma melhora no desempenho dos alunos, analisando ferramentas de AM como a Regressão Logística e as SVM, onde conclui que ambas são ferramentas importantes e relevantes no tópico da análise e *feedback* de desempenho acadêmico, destacando que, principalmente o GPT, possui deficiências e falhas consideráveis.

De maneira semelhante, o trabalho de Sihare e Gupta (2024) analisa métodos de predição de performance estudantil usando aprendizado de máquina, afirmando que analisar grandes quantidades de dados é difícil para o ser humano e, por isso, seu estudo tem como objetivo usar a tecnologia para automatizar métodos convencionais de processamento dos dados, destacando principalmente a ferramenta *XG Boost*, que obteve uma precisão de 97% em seu teste.

Assim como os dois trabalhos citados anteriormente, deve ser destacado o trabalho de Lima, Ávila e Gilaverte (2014), que consiste na aplicação de algoritmos de mineração de dados utilizando a linguagem Java e as *Application Programming Interface* (APIs) disponíveis na ferramenta *WEKA* com o intuito de identificação de padrões para análise de evasão no Instituto Federal de Educação, Ciência e Tecnologia do Sul de Minas, Câmpus Pouso Alegre. Este trabalho possui relevância direta para o desenvolvimento do presente relatório técnico, principalmente considerando a utilização da ferramenta *WEKA* e a proposta tomada em sua aplicação. Este trabalho obteve um índice de acerto de 85%.

Por fim, o trabalho de Al-Alawi et al. (2023) aplica o uso do aprendizado de máquina para prever fatores que afetam a performance acadêmica, utilizando aprendizado supervisionado, área do aprendizado de máquina que lida com redes neurais e árvores de decisão, concluindo que o processamento dos dados é essencial para obter sucesso no

objetivo de prever fatores que influenciam a performance acadêmica. Foram escolhidos os algoritmos J48, *Random Forest*, *Random Tree*, Naïve Bayes, K-NN, SVM e ANN. Destes, deve ser destacado o J48 que obteve 82,35% de precisão nos testes.

Após o estudo e a pesquisa dos trabalhos escolhidos, o próximo capítulo aborda a metodologia deste trabalho, incluindo os tópicos de pesquisa bibliográfica, escolha das técnicas, escolha e análise das bases de dados, pré-processamento dos dados e aplicação das técnicas.

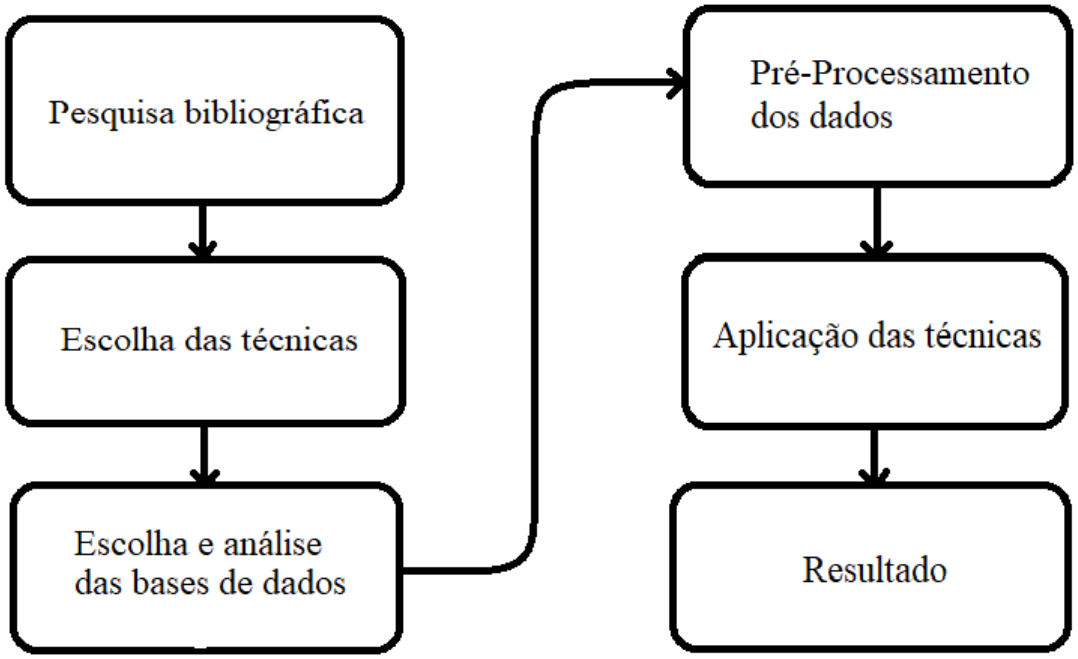
3 METODOLOGIA

Este Capítulo apresenta, de forma detalhada e estruturada, a metodologia adotada para o desenvolvimento deste trabalho. Inicialmente, é apresentada uma visão geral das etapas envolvidas, seguida da descrição dos procedimentos de pesquisa bibliográfica, tais como: seleção das técnicas, escolha e análise das bases de dados, pré-processamento dos dados, aplicação das técnicas escolhidas e análise dos resultados.

3.1 Visão Geral

A metodologia do trabalho será estruturada conforme representado na Figura 1, contemplando os seguintes tópicos: (1) pesquisa bibliográfica, (2) escolha das técnicas, (3) seleção e análise das bases de dados, (4) pré-processamento dos dados, (5) aplicação das técnicas e (6) análise dos resultados.

Figura 1 – Planejamento do estudo



Fonte: Elaborada pelo autor

3.1.1 Pesquisa Bibliográfica

Durante a pesquisa bibliográfica, etapa que tem como objetivo realizar a busca e seleção de trabalhos relevantes para auxiliar na elaboração do desenvolvimento e funda-

mentação teórica, utilizou-se plataformas de pesquisa, tais como IEEE Xplore e Capes. Para tanto, foram elaboradas *strings* de busca com palavras-chave nas línguas portuguesa e inglesa, relacionadas ao interesse da pesquisa, por exemplo, *"machine learning"*, "desempenho acadêmico", "predição" e *"technique"*.

Destaca-se também que, nas bases IEEE Xplore e Capes, as *strings* de busca foram aplicadas somente na língua inglesa, em busca de materiais mais específicos e delimitados dentro do contexto, diferentemente das buscas no Google Acadêmico, onde a pesquisa foi mais abrangente. Na Tabela 1, são exibidas a lista das *strings* de busca, incluindo a base escolhida e seus respectivos resultados.

Tabela 1 – *Strings* de busca

Fonte	**String**	**Resultados**
IEEE Xplore	"academic performance data"AND "machine learning"AND "algorithm"AND "technique"AND "data analysis"AND "prediction"	1
IEEE Xplore	("academic performance data"OR "student performance") AND ("machine learning"OR "ML") AND ("algorithm"OR "model") AND "technique"AND "data analysis" AND "prediction"	10
IEEE Xplore	("academic achievement"OR "academic performance data") AND "machine learning"AND "prediction"AND "algorithm" AND ("technique"OR "strategy") AND "data analysis"	5
IEEE Xplore	("academic performance data"OR "education data") AND "machine learning"AND "algorithm" AND "data analysis"	12
Capes	"academic performance data"AND "machine learning" AND "algorithm"AND "technique"AND "data analysis"AND "prediction"	1
Capes	"academic performance data"AND "machine learning" AND "data analysis"	2
Capes	("academic performance"AND "data") AND "machine learning" AND ("algorithm"OR "classifier") AND ("technique"OR "method")	163
Google Acadêmico	"dados de desempenho acadêmico"AND "aprendizado de máquina" AND algoritmo AND técnica AND "análise de dados"AND predição	12
Google Acadêmico	(("desempenho acadêmico"AND "dados") AND "aprendizado de máquina" ("algoritmo"OR "classificador") AND ("técnica"OR "método")) -evasão	177

Fonte: Elaborada pelo autor

Das bases de dados pesquisadas, foram escolhidos trinta e um trabalhos para compor a pesquisa bibliográfica, com o intuito de estabelecer uma base sólida de informações e referenciais necessários para o desenvolvimento deste relatório técnico. Dentre os escolhidos, vinte e dois artigos estão na língua inglesa e oito na língua portuguesa, todos listados

na Tabela 2, que apresenta a fonte do artigo, seu título e a língua em que o trabalho foi escrito.

Tabela 2 – Lista de artigos pesquisados

Fonte	**Título**	**Língua**
IEEE Xplore	Descriptive Statistical Analysis and Discretization of Academic Data for Machine Learning Techniques	Inglesa
IEEE Xplore	A University Student Performance Prediction Model and Experiment Based on Multi-Fusion and Attention Mechanism	Inglesa
IEEE Xplore	Educational Data Mining: A Comprehensive Study	Inglesa
IEEE Xplore	Ensemble Learning for Student Performance Assessment: Identifying and Analyzing Significant Affecting Factors in Higher Education	Inglesa
IEEE Xplore	Predicting Computer Science Student's Performance using Logistic Regression	Inglesa
IEEE Xplore	IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining	Inglesa
IEEE Xplore	Comparative Analysis of Machine Learning Models in Predicting Academic Outcomes: Insights and Implications for Educational Data Analytics	Inglesa
IEEE Xplore	An Implementation of Support Vector Machine Classification for Developer Academy Acceptance Prediction Model	Inglesa
IEEE Xplore	Explainable Artificial Intelligence Models using Students' Academic Record Data, Tree Family Classifiers, and K-means Clustering to Predict Students' Performance	Inglesa
IEEE Xplore	Smart Education Using Machine Learning for Outcome Prediction in Engineering Course	Inglesa
IEEE Xplore	Machine-Learning based MOOC learning data analysis	Inglesa
Capes	Evaluation of Machine Learning Methods for Prediction Student Performance	Inglesa
Capes	Data Mining for Student Advising	Inglesa
Capes	Prediction of University-Level Academic Performance through Machine Learning Mechanisms and Supervised Methods	Inglesa

Continua na próxima página

Tabela 2 – Continuação

Fonte	**Título**	**Língua**
Capes	Comparative Study of Supervised Algorithms for Prediction of Students' Performance	Inglesa
Capes	Academic performance warning system based on data driven for higher education	Inglesa
Capes	Using machine learning to predict factors affecting academic performance: the case of college students on academic probation	Inglesa
Capes	Metaheuristics Method for Classification and Prediction of Student Performance Using Machine Learning Predictors	Inglesa
Capes	Academic Performance Modelling with Machine Learning Based on Cognitive and Non-Cognitive Features	Inglesa
Capes	A Conceptual Framework to Aid Attribute Selection in Machine Learning Student Performance Prediction Models	Inglesa
Capes	Simulation-Based Machine Learning for Predicting Academic Performance Using Big Data	Inglesa
Capes	Exploring the Effectiveness of AI Algorithms in Predicting and Enhancing Student Engagement in an E-Learning	Inglesa
Google Acadêmico	Utilização de aprendizado de máquina na predição de desempenho acadêmico	Portuguesa
Google Acadêmico	Modelagem preditiva para sucesso acadêmico: um estudo de caso em um curso de ciência da computação	Portuguesa
Google Acadêmico	Estudo de caso com análise de dados para a detecção da desistência de estudantes em disciplinas ofertadas com apoio do ambiente MOODLE	Portuguesa
Google Acadêmico	O impacto da inteligência artificial na educação: oportunidades e desafios nas escolas	Portuguesa
Google Acadêmico	Mineração de dados educacionais para apoio à gestão acadêmica na formulação de prognóstico de perfil de aluno ingressante em cursos superiores	Portuguesa
Google Acadêmico	Seleção de um modelo de aprendizado de máquina para previsão de desempenho academico a partir do Método SAPEVO-M-NC	Portuguesa
Google Acadêmico	Prever e melhorar o desempenho dos alunos com o uso combinado de aprendizagem de máquina e GPT	Portuguesa

Continua na próxima página

Tabela 2 – Continuação

Fonte	Título	Língua
Google Acadêmico	Aprendizado de máquina para agrupamento e associação de dados do ensino superior público brasileiro	Portuguesa
Google Acadêmico	Implementação de um Modelo para Previsão de Evasão Escolar no IFSULDEMINAS	Portuguesa

Fonte: Elaborada pelo autor

3.1.2 Principais Indicadores Acadêmicos

A pesquisa bibliográfica destaca um conjunto de indicadores acadêmicos presentes nos artigos pesquisados, sendo que estes possuem relação com o desempenho acadêmico dos alunos e da instituição. Na Tabela 3 estão relacionados tais indicadores, bem como a frequência em que eles foram citados, ao londo entre dos artigos pesquisados.

Tabela 3 – Indicadores acadêmicos presentes na pesquisa bibliográfica

Tipo de indicador	Quantidade de aparições
Dados Institucionais Simples	21
Desempenho acadêmico	10
SGPA (Índice de rendimento acadêmico semestral)	2
CGPA (Índice de Rendimento Acadêmico Acumulado)	1
IRA (Indíce de rendimento acadêmico)	1
IGC (Índice geral de cursos)	1
Taxa de evasão	1
Taxa anual de evasão	1
Índice de verticalização	1
Nota média final	1

Fonte: Elaborada pelo autor

3.1.2.1 Características dos Indicadores

De acordo com os indicadores encontrados na pesquisa bibliográfica, vale ressaltar o Índice de Eficiência Acadêmica (IEA) e a Relação Aluno Professor (RAP). Ambos indicadores possuem grande relevância no foco do desempenho acadêmico, e podem ser estudados no Guia de Referência Metodológica da Plataforma Nilo Peçanha, onde são apresentadas todas as variáveis que formam o índice e suas respectivas equações. A seguir estarão presentes os cálculos do IEA e do RAP:

- IEA:

$$I_{EA}[\%] = C_{Ciclo} + \left[\left(\frac{C_{Ciclo}}{C_{Ciclo} + E_{Ciclo}}\right) \times R_{Ciclo}\right] \times 100 \tag{3.1}$$

O Índice de Eficiência Acadêmica, é um indicador que mede a proporção de alunos que concluem o curso com sucesso dentro do período previsto, incluindo uma projeção de alunos que poderão concluir. O cálculo do I_{EA} é detalhado na Equação 3.1, cujas variáveis são definidas como:

- $C_{Ciclo}[\%]$: Percentual de Concluintes em relação às matrículas vinculadas aos ciclos concluídos no ano anterior ao Ano de Referência.
- $E_{Ciclo}[\%]$: Percentual de Evadidos em relação às matrículas vinculadas aos ciclos concluídos no ano anterior ao Ano de Referência.
- $R_{Ciclo}[\%]$: Percentual de matriculados que são classificados como **Retidos** por terem ultrapassado o período previsto para integralização do curso (acréscimo de um ano) em relação às matrículas vinculadas aos ciclos concluídos no anterior ao Ano de Referência.

- RAP:

$$RAP = \frac{(Meq_{CG} \times FCG) + (Meq_{DC})}{DEq} \tag{3.2}$$

A Relação Aluno Professor, é um indicador que mede a relação entre a quantidade de matrículas equivalentes e a quantidade de docentes efetivos, ponderados pelo Regime de Trabalho (RT). O cálculo do RAP é apresentado na Equação 3.2, onde os termos são definidos como:

- Meq_{CG}: Matrículas Equivalentes em Cursos de Graduação. Quantidade de matrículas que estiveram ativas em pelo menos um dia no ano de referência em Cursos de Graduação, ponderada pelos fatores de equivalência previstos.
- Meq_{DC}: Matrículas Equivalentes nos Demais Cursos (Exceto Graduação). Quantidade de matrículas que estiveram ativas em pelo menos um dia no ano de referência em todos os cursos, exceto os Cursos de Graduação, ponderada pelos fatores de equivalência.
- FCG: Fator de Correção de Graduação. Fator que considera que a meta de alunos por professor é diferente para cursos técnicos e de graduação. O valor estabelecido é $FCG = 20/18 = 1,111$.
- DEq: Docentes Equivalentes. Quantidade de professores efetivos que atuam no Regime de Trabalho (RT) 20h multiplicados por 0,5, somado à quantidade de professores efetivos que atuam nos RT 40h e RDE.

3.1.3 Escolha das Técnicas

Todas as técnicas foram implementadas utilizando a ferramenta *WEKA* (*Waikato Environment for Knowledge Analysis*), um software de aprendizado de máquina e mineração de dados desenvolvido pela Universidade de Waikato, na Nova Zelândia. Além disso, a linguagem *Python*[1] foi empregada de forma complementar nos trabalhos de pré-processamento do *dataset*, pois oferece bibliotecas e recursos necessários para a aplicação das técnicas selecionadas.

3.1.4 Escolha e Análise das Bases de Dados

A base de dados foi escolhida através do portal de dados abertos da Plataforma Nilo Peçanha (MORAES; ALMEIDA; ALVES, 2018), buscando conjuntos que contenham dados importantes para o cálculo dos indicadores acadêmicos, como o IEA, além de percentuais de conclusão, evasão, situação de matrícula e também a localização dessas matrículas. No site da Plataforma Nilo Peçanha, estão presentes diversos conjuntos de dados como os Microdados de Eficiência Acadêmica, os Microdados de Matrículas e os Microdados dos Servidores da rede pública.

3.1.5 Pré-Processamento dos Dados

O pré-processamento dos dados foi realizado por meio da aplicação de técnicas voltadas a garantir a consistência e a qualidade da base de dados. Nesta fase do trabalho, destacam-se as seguintes técnicas aplicadas: tratamento de valores ausentes, filtragem do escopo da base, padronização dos nomes das colunas e formatação dos números. A seguir, são apresentados exemplos práticos dessas etapas.

- **Tratamento dos valores ausentes:** Consiste em remover dados nulos presentes em linhas ou colunas da base de dados, visando obter um processamento mais consistente.

  Leitura da base e tratamento dos valores:

```
1 import pandas as pd
2
3 df = pd.read_csv('EficienciaAcademica.csv', sep=';')
4
5 df = df.dropna()
```

[1] O *Python* é uma linguagem de programação de alto nível, lançada oficialmente em 1991, que ganhou grande popularidade nos últimos anos. Foi escolhido por ser amplamente utilizado no tratamento e visualização de dados, além de oferecer pacotes que facilitam a integração com o *WEKA* (Python Software Foundation, 2024).

- **Definição do escopo da base:** Etapa que irá definir o recorte desejado da base para aplicação das técnicas, nesse caso, filtrando a base para utilizar somente os Câmpus do Instituto Federal de São Paulo, foco deste trabalho.

  Filtragem do escopo:

```
instituicao = "Instituto Federal de São Paulo"

dff = df[(df['Instituição (Nome)'] == instituicao)]
```

- **Alteração dos nomes:** Após o tratamento dos dados, serão alterados os nomes das colunas que não possuem claro entendimento, notando também, que acentos e símbolos serão excluídos para melhor entendimento no software do *WEKA*.

  Alteração dos nomes:

```
novos_nomes = {
    'Região': 'regiao',
    'Organização Acadêmica': 'organizacao academica',
    'Instituição (Nome)': 'instituicao nome',
    'Eficiência Acadêmica | Concluídos' : 'concluidos',
    'Eficiência Acadêmica | Concluídos %' : 'concluidos %',
    'Eficiência Acadêmica | Índice de Eficiência Acadêmica %' : '
       indice de eficiencia academica %',
    'Eficiência Acadêmica | Número de Evadidos' : 'numero de
       evadidos',
    'Eficiência Acadêmica | Retidos' : 'retidos',
    'Eficiência Acadêmica | Retidos %' : 'retidos %',
    'Eficiência Acadêmica | Taxa de Evasão %' : 'taxa de evasao %'
}

dff.rename(columns=novos_nomes, inplace=True)
```

- **Formatação:** Por fim, etapa que irá substituir com pontos os números não inteiros escritos com vírgula. Etapa que também tem como intuito excluir os símbolos de porcentagem para evitar problemas no reconhecimento da tabela de formato csv pelo software da ferramenta *WEKA*.

  Formatação dos números:

```
for col in [
    'concluidos %',
    'indice de eficiencia academica %',
    'retidos %',
    'taxa de evasao %'
]:
    dff[col] = dff[col].str.replace('%', '', regex=False).str.
       replace(',', '.', regex=False).astype(float)
```

```
8
9 dff
```

Após as etapas iniciais, aplica-se técnicas distintas de pré-processamento, de acordo com o algoritmo utilizado. Para a Regressão Logística, foi empregada a normalização da escala; para o método de *Ensemble Learning* e para o algoritmo *K-means*, foi aplicada a discretização dos dados. Para o *Apriori*, além da normalização da escala, os atributos foram convertidos em nominais e alguns atributos irrelevantes passaram por etapa de filtragem. O *K-means* também irá ter atributos não numéricos convertidos em binários para melhor aplicação. Todas as técnicas serão implementadas por meio da ferramenta *WEKA*.

No caso da Regressão Logística e do algoritmo *Apriori*, a normalização será realizada utilizando o método *StandardScaler*, que ajusta os dados para que as variáveis sejam representadas em uma escala comum. Esse processo é fundamental para evitar que valores discrepantes (*outliers*) tenham influência excessiva sobre os resultados.

Para a conversão dos atributos nos algoritmos selecionados, serão utilizados métodos como o *NumerictoNominal*, *NominaltoBinary* ou *StringtoNominal*, todos disponibilizados pela ferramenta *WEKA*.

3.1.6 Aplicação das Técnicas

A aplicação das técnicas escolhidas, de acordo com a literatura, será dividida em duas etapas, combinando o pré-processamento dos dados utilizando *Python* e o uso da ferramenta *WEKA* para aplicação dos algoritmos de aprendizado de máquina.

3.1.6.1 Utilização da Ferramenta *WEKA*

A ferramenta *WEKA* utilizada neste trabalho é a versão 3.8.6, instalada em um sistema operacional Windows 10. A seguir, são descritos alguns procedimentos básicos a serem adotados, frente ao uso da *WEKA*, para a aplicação dos algoritmos de Regressão Logística, *Ensemble Learning*, *Apriori* e *K-means*:

- **Regressão Logística:**

  Após carregar no *WEKA* a base de dados previamente filtrada, sem valores nulos e limitada ao escopo desejado, o primeiro passo será aplicar o método `Standardize` — equivalente ao `StandardScaler` — com o objetivo de normalizar os dados para a aplicação adequada da Regressão Logística. A regressão logística será executada no *WEKA* por meio da seguinte linha de comando:

```
1 LinearRegression -S 0 -R 1.0E-8 -num-decimal-places 4 -
      batch-size 500
```

Nesta configuração, o algoritmo utiliza todos os atributos disponíveis (parâmetro `-S 0`), aplica regularização com coeficiente `1.0E-8` (`-R`), apresenta os coeficientes com quatro casas decimais (`-num-decimal-places 4`) e processa 500 instâncias por lote (`-batch-size 500`).

- ***Ensemble Learning*:**

  O *Ensemble Learning* será realizado por meio do algoritmo **AdaBoost**, escolhido por sua relevância na pesquisa bibliográfica e por estar disponível no *WEKA*. Após carregar a base de dados, a etapa inicial será o pré-processamento, que inclui a **discretização** dos dados para categorização.

  A discretização será aplicada especialmente nas colunas **Taxa de Evasão** e **Índice de Eficiência Acadêmica**, classificando os valores em categorias como baixo, médio e alto. Isso permite que o algoritmo avalie a relação entre esses grupos de dados.

  A classificação com o AdaBoost será executada utilizando a linha de comando:

```
1 AdaBoostM1 -P 100 -S 1 -I 10 -W weka.classifiers.trees.
      DecisionStump -batch-size 500
```

  Nesta configuração, o parâmetro `-batch-size 500` define o número de instâncias processadas por lote, `-I 10` especifica 10 iterações, `-S 1` determina a semente para geração de números aleatórios, e `-W` indica o uso do classificador base `DecisionStump`.

- ***Apriori*:**

  Assim como na Regressão Logística, o primeiro passo será aplicar o método Standardize — equivalente ao StandardScaler — com o objetivo de normalizar os dados para a aplicação adequada do *Apriori*, seguido pela conversão de todos os atributos para Nominal, através dos métodos *NumerictoNominal* e *StringtoNominal* presentes na ferramenta *WEKA*. O algoritmo *Apriori* será testado utilizando diferentes configurações, alterando principalmente os parâmetros de confiança e suporte máximo, com intuito de encontrar a combinação que traga os resultados mais relevantes, sendo executado no *WEKA* por meio da seguinte linha de comando:

```
1 Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0
      -c -1
```

  Nesta configuração, o algoritmo *Apriori* é ajustado para gerar no máximo 10 regras (-N 10), utiliza o método padrão de seleção de regras (-T 0), exige confiança mínima de 0,9 (-C 0.9), define o incremento mínimo de suporte em 0,05 (-D 0.05), permite

suporte máximo de 1.0 (-U 1.0), estabelece o suporte mínimo inicial em 0.1 (-M 0.1), usa um ajuste automático de suporte com valor -1.0 (-S -1.0) e considera todos os atributos disponíveis na base de dados (-c -1).

Lembrando que serão aplicados diferentes configurações no parâmetro de confiança (-C) e no parâmetro de suporte máximo (-U).

- ***K-means*:**

  Para aplicar o *K-means*, será utilizado o *dataset* discretizado, da mesma maneira que na aplicação do *AdaBoost*, também utilizando dos métodos presentes na ferramenta *WEKA* como o filtro *NominaltoBinary* e o *Standardize*, visando deixar o conjunto mais homogêneo e transformando atributos nominais que não seriam interpretados pelo algoritmo em binário, para que assim o algoritmo consiga interpretar o conjunto como um todo. O algoritmo será aplicado utilizando as configurações padrão através da linha de comando:

```
1   SimpleKMeans -init 0 -max-candidates 100 -periodic-
       pruning 10000 -min-density 2.0 -t1 -1.25 -t2 -1.0 -N
        2 -A "weka.core.EuclideanDistance -R first-last" -I
        500 -num-slots 1 -S 10
```

  Nesta configuração, o algoritmo SimpleKMeans é ajustado para inicializar os centróides utilizando o método padrão (-init 0), considera até 100 candidatos na busca de melhores centróides iniciais (-max-candidates 100) e realiza poda periódica a cada 10.000 iterações para otimizar o desempenho (-periodic-pruning 10000). O parâmetro de densidade mínima é definido como 2.0 (-min-density 2.0), enquanto os limiares de poda são estabelecidos em -1.25 (-t1 -1.25) e -1.0 (-t2 -1.0).

  O algoritmo é configurado para formar 2 clusters (-N 2), utilizando a distância euclidiana como medida de similaridade entre as instâncias (-A "weka.core.EuclideanDistance -R first-last"), considerando todos os atributos disponíveis no *dataset*. O número máximo de iterações é definido em 500 (-I 500), o processamento é realizado em apenas uma thread (-num-slots 1) e a semente aleatória utilizada para inicialização é igual a 10 (-S 10), garantindo reprodutibilidade dos resultados.

4 RESULTADOS

Este Capítulo apresenta os resultados obtidos através da aplicação das técnicas de análise estatística descritiva e de Aprendizado de Máquina selecionadas para análise do Índice de Eficiência Acadêmica (IEA) nos campi do Instituto Federal de São Paulo e uma análise mais profunda, utilizando microdados de matriculas focada nos Câmpus de São João da Boa Vista do IFSP. Os dados analisados compreendem o período de 2017 a 2024, e os resultados são organizados em três seções principais: caracterização dos dados, desempenho dos algoritmos aplicados e análise das correlações identificadas.

4.1 Processo de escolha das técnicas

As técnicas selecionadas para este trabalho (discretização de dados, regressão logística, *Ensemble Learning*, algoritmo de *Apriori* e algoritmo *K-means*) foram escolhidas com base na pesquisa bibliográfica realizada. A decisão de usá-las foi norteadada pelos resultados promissores demonstrados nas abordagens selecionadas.

A discretização é uma técnica de pré-processamento que categoriza os dados para otimizar a análise, enquanto a regressão logística, o *Ensemble Learning*, o *Apriori* e o *K-means* são modelos de Aprendizado de Máquina reconhecidos por sua eficiência.

Como apresentado na pesquisa bibliográfica, estudos (Sagala et al., 2022) indicam que a regressão logística possui aproximadamente 91% de precisão, e algoritmos de *Ensemble Learning* chegam a 86% e 97% de precisão utilizando o SVM e o *XG Boost*.

O algoritmo *Apriori* trabalha com base em associações e busca relações frequentes em bases de dados. Porém, o *K-means* é um algoritmo de agrupamento (*clustering*), que separa os dados em conjuntos que possuem maior similaridade.

4.2 Processo de escolha e análise das bases de dados

As bases de dados escolhidas para a aplicação das técnicas foram a de Microdados de Eficiência Acadêmica e a de Microdados de Matrículas (MORAES; ALMEIDA; ALVES, 2018), disponível no portal de dados abertos da Plataforma Nilo Peçanha, vinculada ao Ministério da Educação (MEC). Tais microdados contém dados no intervalo entre os anos de 2017 e 2024, de Institutos Federais, Cefet's e Colégios Pedro II de todo o Brasil, incluindo os estados de Minas Gerais, Rio de Janeiro, Rio Grande do Sul, Santa Catarina, Bahia, São Paulo, Ceará, Maranhão, Goiás, Rio Grande do Norte, Paraná, Pernambuco, Piauí, Paraíba, Espírito Santo, Pará, Mato Grosso, Alagoas, Amazonas, Tocantins, Distrito

Federal, Mato Grosso do Sul, Rondônia, Sergipe, Roraima e Amapá. A base inclui também Universidades Federais, porém com apenas 3% dos dados.

A fase inicial dos testes concentrou-se na análise dos resultados obtidos a partir dos Microdados de Eficiência Acadêmica. Posteriormente, foi realizada uma tentativa de agregar a este *dataset* um conjunto de dados suplementares, fornecido pela direção do Câmpus do IFSP de São João da Boa Vista. No entanto, esta agregação não se mostrou viável para a análise, pois o segundo *dataset* apresentava uma alta incidência de dados faltantes que comprometeria os resultados. Sendo assim, nesta etapa, considerou-se apenas o *dataset* original de Microdados de Eficiência Acadêmica.

A segunda etapa dos testes consistiu na combinação dos Microdados de Matrículas com os Microdados de Eficiência Acadêmica. O objetivo central desta agregação foi obter resultados mais significativos e consistentes na aplicação dos algoritmos.

A base de dados combinada enriqueceu a análise, pois acrescentou informações detalhadas de cada matrícula (como curso, carga horária, faixa etária, renda familiar e situação da matrícula) aos indicadores acadêmicos existentes (índice de eficiência, conclusão, evasão e retenção de alunos).

É fundamental destacar que, para o escopo deste trabalho, os dados de matrícula foram filtrados e restritos especificamente ao Câmpus São João da Boa Vista do Instituto Federal de São Paulo.

Vale ressaltar que o uso das bases da Plataforma Nilo Peçanha foi concretizado já que as bases internas do Câmpus presentes no sistema do SUAP não foram disponibilizadas a tempo para execução dos testes.

4.3 Aplicação das Técnicas

Após a aplicação das técnicas, de acordo com o roteiro presente na metodologia, os resultados foram obtidos e interpretados a partir da caracterização e de métodos de estatística descritiva, buscando encontrar resultados empíricos relacionados aos indicadores e as propostas do trabalho.

4.4 Caracterização dos Dados e Estatística Descritiva

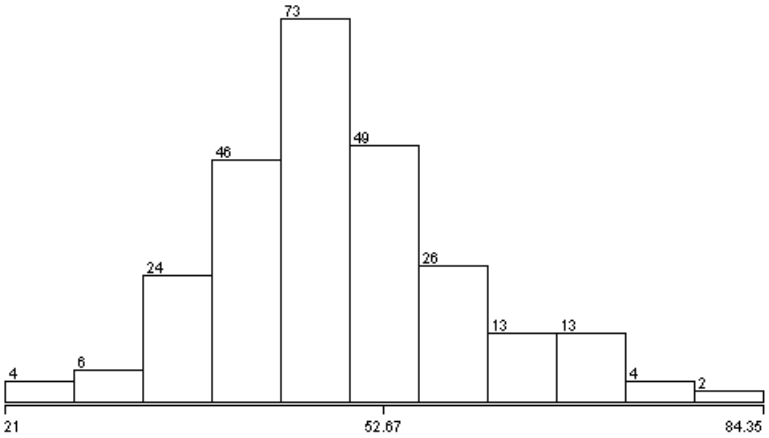
A base de dados utilizada neste estudo foi filtrada para incluir exclusivamente informações dos campi do Instituto Federal de São Paulo, resultando em um conjunto de 260 instâncias para análise. A seleção das variáveis foi baseada na aplicação dos algoritmos de Regressão Logística e *AdaBoost*, que identificaram três variáveis com maior poder preditivo em relação ao IEA: o próprio Índice de Eficiência Acadêmica, a Taxa de

Conclusão e a Taxa de Evasão.

4.4.1 Análise Descritiva do Índice de Eficiência Acadêmica

O Índice de Eficiência Acadêmica apresentou distribuição com características específicas que merecem destaque. A Figura 2 ilustra a distribuição dos valores de IEA ao longo do período analisado, enquanto a Tabela 4 apresenta as medidas de tendência central e dispersão calculadas.

Figura 2 – Distribuição do Índice de Eficiência Acadêmica



Fonte: Elaborada pelo autor

Tabela 4 – Medidas de Tendência Central e Dispersão do IEA

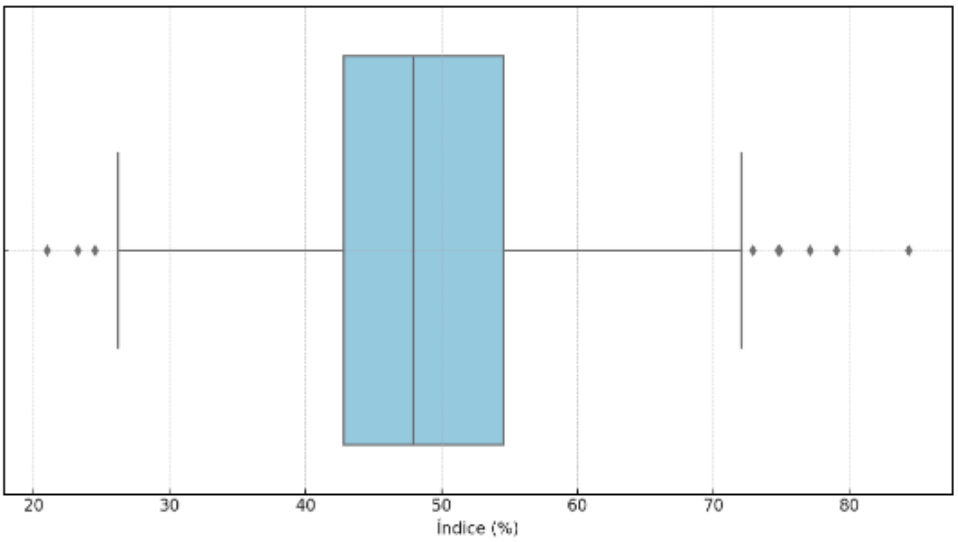
Estatística	**Valor**
Valor Mínimo	21,00
Valor Máximo	85,35
Média	49,05
Mediana	48,63
Moda	43,09
Desvio Padrão	10,44
Amplitude	63,35

Fonte: Elaborada pelo autor

Os resultados da análise descritiva revelam que o IEA apresenta uma distribuição aproximadamente normal, com média de 49,05 e mediana de 48,63, indicando baixa assimetria. O desvio padrão de 10,44 sugere variabilidade moderada entre os campi analisados. A amplitude de 63,35 pontos, com valores variando entre 21,00 e 85,35, demonstra heterogeneidade significativa no desempenho acadêmico entre as diferentes unidades.

Para identificação de valores atípicos, foi elaborado o gráfico *boxplot* apresentado na Figura 3. Esta análise é fundamental para compreender a distribuição dos dados e identificar campi com desempenho significativamente diferente da média.

Figura 3 – *Boxplot* do Índice de Eficiência Acadêmica



Fonte: Elaborada pelo autor

O *boxplot* revela que a maioria dos valores de IEA concentra-se na faixa entre 40% e 60%, com os quartis bem definidos. Valores inferiores a 26 ou superiores a 72 são considerados *outliers*, representando campi com desempenho excepcionalmente baixo ou alto, respectivamente. Esta informação é relevante para identificar unidades que necessitam de atenção especial ou que podem servir como modelo de boas práticas.

4.5 Desempenho dos Algoritmos nos Microdados de Eficiência Acadêmica

Para avaliar as relações entre as variáveis selecionadas e o IEA, foram aplicados quatro algoritmos de Aprendizado de Máquina: Regressão Logística, *AdaBoost*, *Apriori* e o *K-means*. A escolha destes algoritmos baseou-se em suas características complementares, sendo a Regressão Logística um método linear interpretável, o *AdaBoost* um método *ensemble* que pode capturar relações não-lineares complexas, o algoritmo *Apriori* que realiza associações com base nos dados analisados e o *K-means* que divide os dados em agrupamentos correspondentes.

4.5.1 Resultados da Regressão Logística

A aplicação da Regressão Logística produziu resultados satisfatórios, conforme apresentado na Tabela 5. O algoritmo foi implementado utilizando a ferramenta *WEKA*,

que facilitou a visualização e interpretação dos resultados.

Tabela 5 – Métricas de Desempenho da Regressão Logística

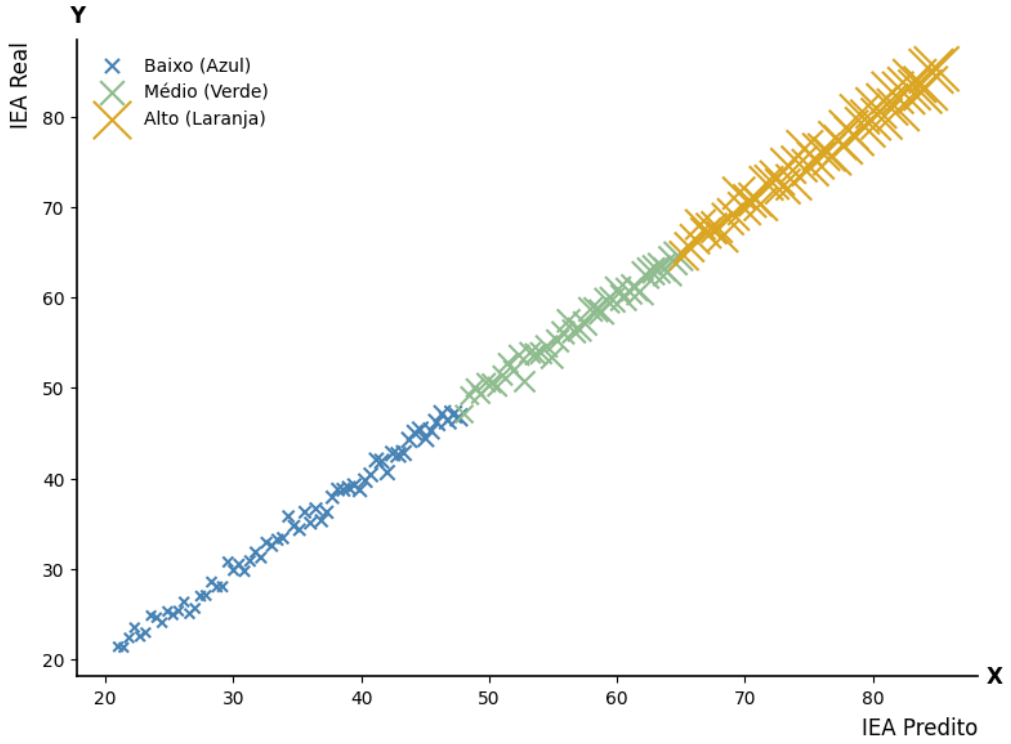
Métrica	**Valor**
Coeficiente de Correlação	0,9974
Erro Absoluto Médio	0,4611
Raiz do Erro Quadrático Médio	0,7482
Erro Absoluto Relativo	5,7124%
Raiz do Erro Quadrático Relativo	7,1441%
Número Total de Instâncias	260

Fonte: Elaborada pelo autor

O coeficiente de correlação de 0,9974 indica uma correlação forte entre os valores reais e preditos do IEA, demonstrando que o modelo consegue explicar aproximadamente 99,74% da variabilidade dos dados. O erro absoluto médio de 0,4611 representa uma diferença média de menos de meio ponto entre os valores reais e preditos, o que é considerado excelente para este tipo de análise.

A Figura 4 apresenta a relação entre os eixos X e Y, após a aplicação do *StandartScaler*, evidenciando a alta precisão do modelo. A proximidade dos pontos à linha diagonal ideal confirma a qualidade das predições realizadas.

Figura 4 – Relação entre IEA Real e Predito - Regressão Logística



Fonte: Elaborada pelo autor

4.5.2 Resultados do *AdaBoost*

O algoritmo *AdaBoost*, pertencente à família de métodos *ensemble*, apresentou desempenho satisfatório, considerando os resultados do algoritmo de regressão logística, conforme demonstrado na Tabela 6. Este algoritmo utiliza múltiplos classificadores fracos para construir um classificador forte, sendo particularmente eficaz na identificação de padrões complexos nos dados.

Tabela 6 – Métricas de Desempenho do *AdaBoost*

Métrica	**Valor**
Instâncias Classificadas Corretamente	258 (99,23%)
Instâncias Classificadas Incorretamente	2 (0,77%)
Estatística Kappa	0,9885
Erro Absoluto Médio	0,0148
Raiz do Erro Quadrático Médio	0,0722
Erro Absoluto Relativo	3,3259%
Raiz do Erro Quadrático Relativo	15,3131%
Número Total de Instâncias	260

Fonte: Elaborada pelo autor

O *AdaBoost* classificou corretamente 258 das 260 instâncias (99,23% de precisão), com apenas 2 classificações incorretas. A estatística *Kappa* de 0,9885 indica concordância quase perfeita entre as classificações preditas e reais, superando significativamente o que seria esperado por acaso. O erro absoluto médio extremamente baixo (0,0148) demonstra a capacidade superior deste algoritmo em capturar as nuances dos dados.

4.5.3 Resultados do *Apriori*

A aplicação do *Apriori* foi realizada através de uma série de testes com configurações diferentes, alterando parâmetros de confiança e suporte máximo do algoritmo. Através dos testes realizados as principais associações presentes, na maior parte das configurações, serão apresentadas na Tabela 7.

A síntese da aplicação do algoritmo revela um alto *Lift* [1] em seus resultados, o que indica a força das associações em relação a variáveis independentes, de modo que por exemplo a primeira associação é 2,64 vezes mais forte do que o comum. Esses resultados expõem uma alta aparição de alunos que possuem um alto IEA na categoria de baixa evasão.

[1] O **Lift** é uma métrica chave em mineração de regras de associação que mede a força da associação entre o antecedente e o consequente. É definido como a razão da Confiança de uma regra sobre o Suporte do consequente: $\text{Lift}(\text{A} \rightarrow \text{B}) = \frac{\text{Confidence}(\text{A} \rightarrow \text{B})}{\text{Support}(\text{B})}$. Um valor de Lift > 1 indica associação positiva, Lift $= 1$ indica independência, e Lift < 1 indica associação negativa. (Han; Kamber; Pei, 2012)

Tabela 7 – Melhores regras encontradas

Regra	Associação	Conf.	Lift	Lev.	Conv.
1	IEA_cat=Alta ⇒ evasao_cat=Baixa	0.89	2.64	0.18	5.26
2	evasao_cat=Baixa ⇒ IEA_cat=Alta	0.89	2.64	0.18	5.26
3	IEA_cat=Baixa ⇒ evasao_cat=Alta	0.83	2.47	0.16	3.62
4	evasao_cat=Alta ⇒ IEA_cat=Baixa	0.83	2.47	0.16	3.62

4.5.4 Resultados do *K-means*

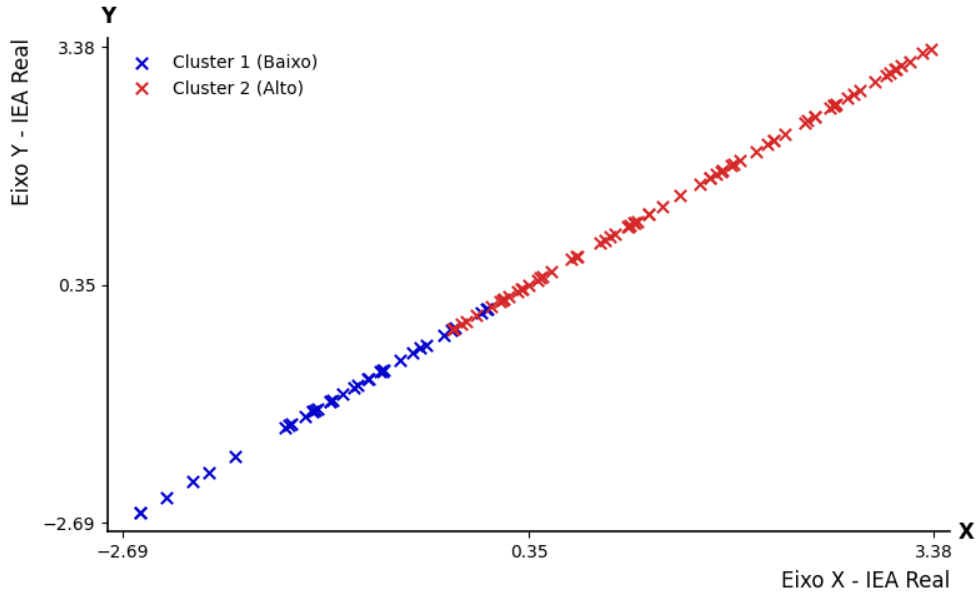
O algoritmo *K-means* foi aplicado para realizar o agrupamento de dados, de modo que fiquem expostas as principais categorias presentes nos dados e ajudar na visualização dessas categorias.

A aplicação foi executada seguindo as configurações padrões da ferramenta *WEKA*, executada em duas etapas distintas, sendo elas o agrupamento em duas partes e o agrupamento em três partes.

- **Dois Agrupamentos:**

  A visualização dos dois agrupamentos pode ser observada na Figura 5, onde foi traçado um gráfico que relaciona o IEA em eixo X e Y, de modo que fique clara a aplicação dos agrupamentos.

Figura 5 – Visualização dos dois grupos



Fonte: Elaborada pelo autor

Os pontos marcados em azul apresentam os dados presentes no grupo de menor IEA (Grupo 0), da mesma maneira que os pontos marcados em vermelho apresentam o grupo de maior IEA (Grupo 1).

Os resultados também apresentam as informações desses agrupamentos, onde são expostos os seguintes dados:

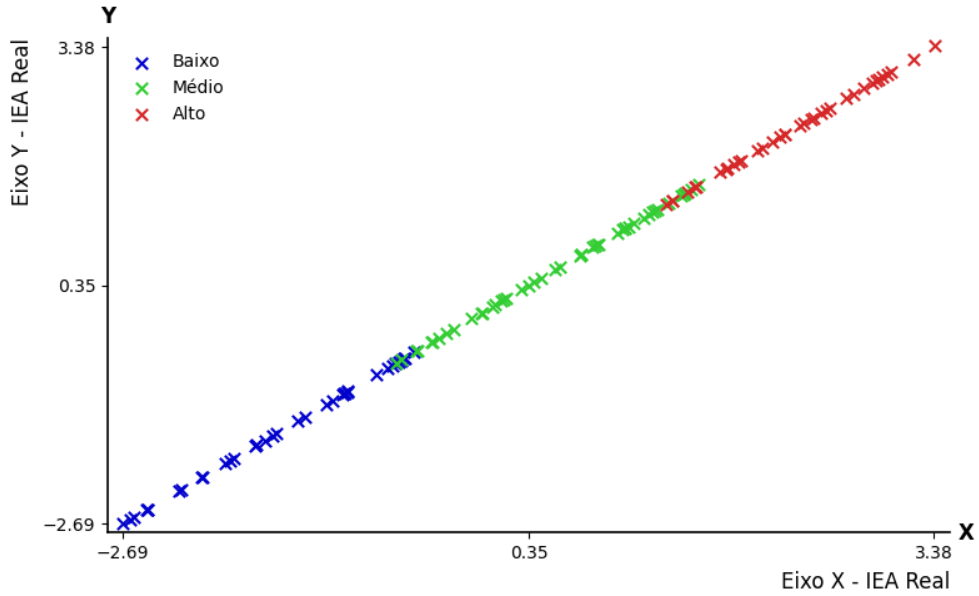
- **Grupo 0 (39%)**: Possuem alta taxa de evasão, IEA na categoria baixa, menor percentual de concluintes e são o perfil de menor desempenho.
- **Grupo 1 (61%)**: Possuem grande IEA e menor taxa de evasão, compondo o grupo com maior proporção de concluintes e melhor desempenho.

Esses resultados revelam que apesar de menor a quantidade de alunos em situação de baixo IEA, ainda representa uma parcela importante de quase dois quintos de matriculados que não atingem níveis agradáveis de retenção e IEA.

- **Três Agrupamentos:**

  Assim como nos dois agrupamentos, a visualização em três agrupamentos será observada na Figura 6, que também apresenta um gráfico composto pela relação do IEA em seu eixo X e Y.

Figura 6 – Visualização dos três grupos



Fonte: Elaborada pelo autor

As partes deste agrupamento são representados pelos pontos azuis de menor IEA (Grupo 0), pelos pontos verdes com um IEA mediano (Grupo 1) e pelos pontos vermelhos que possuem um grande IEA (Grupo 2).

Os resultados apresentam as seguintes informações para cada uma das categorias:

- **Grupo 0 (33%)**: Perfil acadêmico de menor porcentagem de concluintes e IEA, com uma alta evasão. Compõe o grupo de pior desempenho.

- **Grupo 1 (33%)**: Perfil acadêmico que possui valores próximos de uma média geral, taxas de evasão e IEA medianas e valores intermediários, compondo um grupo que não se situa de maneira critica, mas que pode melhorar.
- **Grupo 2 (33%)**: Perfil acadêmico de maior desempenho, possuindo a maior porcentagem de concluintes e menor porcentagem e número de evadidos, compondo um grupo de alto desempenho e baixa evasão. Algo que pode servir como modelo de boas práticas.

Após a interpretação dos três agrupamentos, vale ressaltar a consistência na divisão dos agrupamentos, onde cada uma das faixas de desempenho possuem quantidade similar de alunos.

Ao relacionar as duas maneiras diferentes de agrupamento, a presença de uma grande quantidade de alunos em situação mediana fica clara, de modo que, diferente do apresentado na aplicação de dois grupos, existem quantidades semelhantes de alunos que possuem grande desempenho e pouco desempenho, adicionando os alunos que possuem desempenho mediano, algo que deve ser visado também na aplicação de medidas que busquem a melhora do IEA.

4.6 Desempenho dos Algoritmos nas Bases de Microdados Combinadas

Após a aplicação dos algoritmos selecionados na base de Microdados de Eficiência Acadêmica, foram também aplicados em uma base de dados feita em uma junção dos Microdados de Eficiência Acadêmica e Microdados de Matriculas, de modo que essas aplicações retornassem resultados complementares aos já obtidos.

Essa base de dados nova foi obtida adicionando as colunas de evadidos, completados e índice de eficiência acadêmica na base de matriculas. Todos os dados dessa nova base foram utilizados com o escopo do Câmpus São João da Boa Vista em um período entre 2017 até 2023. A aplicação dos algoritmos foi mantida em padrões parecidos com as aplicações prévias, visando obter mais consistência na relação dos resultados novos com os já obtidos previamente.

4.6.1 Resultados da Regressão Logística

A aplicação da Regressão Logística na nova base será apresentada na Tabela 8.

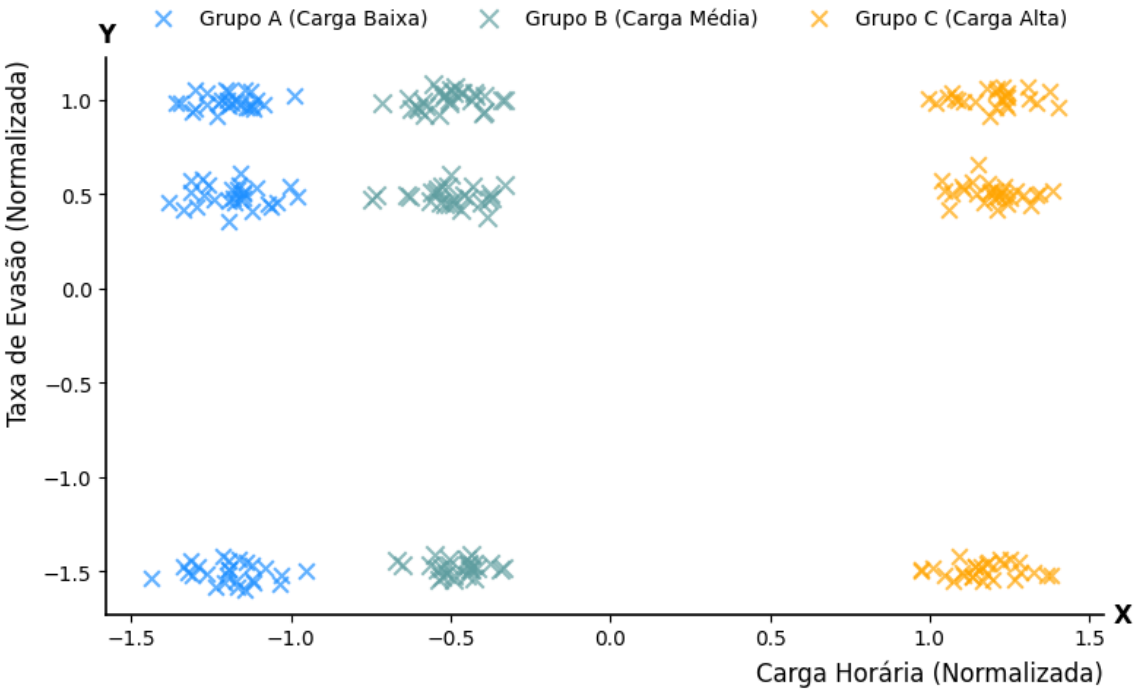
Tabela 8 – Métricas de Desempenho da Regressão Logística

Métrica	**Valor**
Coeficiente de Correlação	0,9998
Erro Absoluto Médio	0,0107
Raiz do Erro Quadrático Médio	0,019
Erro Absoluto Relativo	1,1779%
Raiz do Erro Quadrático Relativo	1,8971%
Número Total de Instâncias	4237

Fonte: Elaborada pelo autor

Estes novos testes obtiveram precisão ainda maior que os primeiros, apresentados pelo coeficiente de correlação e pelas demais estatísticas obtidas. Através desse novo teste foi obtido o gráfico presente na Figura 7.

Figura 7 – Relação entre a Carga Horária e Taxa de evasão - Regressão Logística



Fonte: Elaborada pelo autor

Observa-se, no gráfico da Figura 7, que a taxa de evasão é mais elevada nos cenários em que há maior concentração de matrículas associadas a cargas horárias mais altas.

4.6.2 Resultados do *AdaBoost*

Já na nova aplicação do *AdaBoost*, foi realizada utilizando a coluna de categorização do IEA como coluna principal na utilização do algoritmo. A tabela 9 mostra o desempenho dessa aplicação.

Tabela 9 – Métricas de Desempenho do *AdaBoost* - IEA categorizado

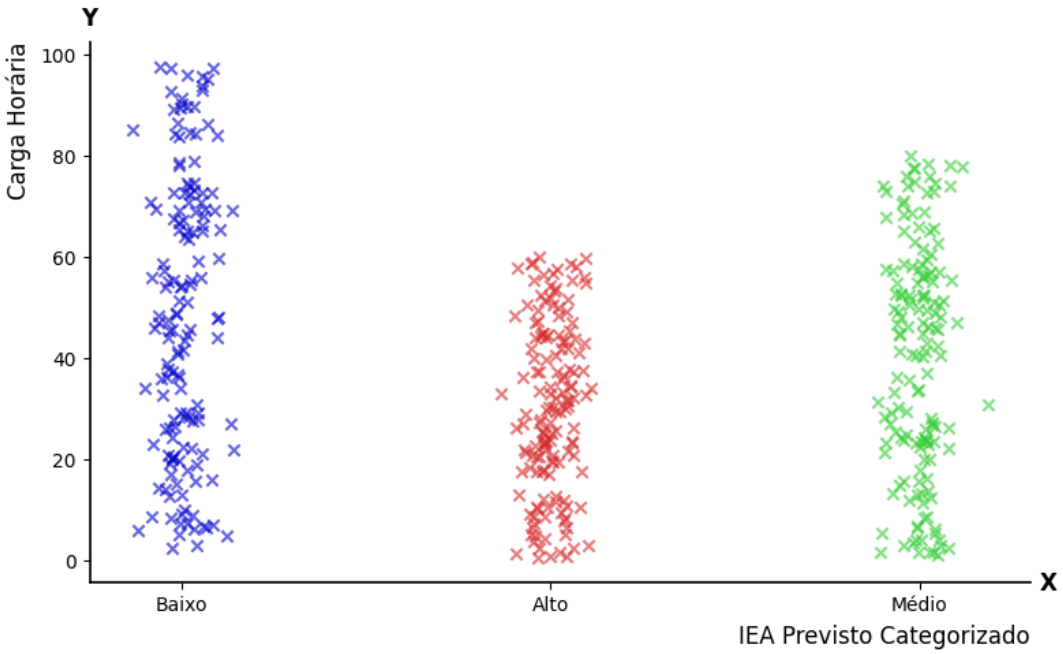
Métrica	**Valor**
Instâncias Classificadas Corretamente	4237 (100%)
Instâncias Classificadas Incorretamente	0 (0%)
Estatística Kappa	1
Erro Absoluto Médio	0
Raiz do Erro Quadrático Médio	0
Erro Absoluto Relativo	0%
Raiz do Erro Quadrático Relativo	0%
Número Total de Instâncias	4237

Fonte: Elaborada pelo autor

Como apresentado nas métricas de desempenho da aplicação utilizando o IEA categorizado como coluna principal, o algoritmo obteve 100% de precisão, principalmente porque na combinação dos *datasets*, esses atributos de IEA categorizada referentes ao Câmpus foram distribuídos para os dados de matrículas, o que preenche o *dataset* com muitas informações relacionadas a essa coluna e aumenta a capacidade de associação do algoritmo.

Os resultados dessa aplicação trouxeram algumas informações, principalmente a que será demonstrada na Figura 8.

Figura 8 – Relação entre a previsão do IEA e Carga horária - *AdaBoost*



Fonte: Elaborada pelo autor

Neste gráfico, da Figura 8, está apresentada a relação entre a previsão das categorias de IEA no eixo X e a carga horária das matriculas no eixo Y, de modo que na coluna azul

estão os pontos de baixo IEA, na coluna vermelha estão os pontos de alto IEA e na coluna verde os pontos de médio IEA. Ainda neste gráfico, percebe-se a presença de um IEA alto em cursos de carga horária média e pequena e em cursos de carga horária alta, exaltando a excessiva presença de um IEA baixo em cargas horárias altas, onde há praticamente apenas a presença do IEA baixo.

4.6.3 Resultados do *Apriori*

O algoritmo de *Apriori* foi aplicado utilizando as configurações padrão da ferramenta *WEKA*, assim como nos testes prévios, porém, diferente dos outros algoritmos, não foram encontradas informações que possam complementar a compreensão dos resultados, de todo modo, a Tabela 10 apresenta as principais regras encontradas na aplicação.

Tabela 10 – Melhores regras encontradas

Regra	**Associação**	**Conf.**	**Lift**	**Lev.**	**Conv.**
1	TaxEvas_P $\leq 9 \rightarrow$ Retidos_P 35–40	1	3	0	875
2	Retidos_P 35–40 $\rightarrow$ TaxEvas_P ≤ 9	1	3	0	875
3	Concl_P 49–51 $\rightarrow$ TaxEvas_P 42–47	1	2	0	670
4	Sexo=M; TaxEvas_P $\leq 9 \rightarrow$ Retidos_P 35–40	1	3	0	537
5	Sexo=M; Retidos_P 35–40 $\rightarrow$ TaxEvas_P ≤ 9	1	3	0	537
6	Raca=Branca; Concl_P 49–51 $\rightarrow$ TaxEvas_P 42–47	1	2	0	398
7	Sit=Em curso; Concl_P 49–51 $\rightarrow$ TaxEvas_P 42–47	1	2	0	397
8	Raca=Branca; TaxEvas_P $\leq 9 \rightarrow$ Retidos_P 35–40	1	3	0	486
9	Raca=Branca; Retidos_P 35–40 $\rightarrow$ TaxEvas_P ≤ 9	1	3	0	486
10	Concl_P 41–42 $\rightarrow$ IEA_P ≤ 10	1	3	0	458

- **Regras 1 e 2:** Estas regras mostram um padrão bidirecional indicando que estudantes com taxa de evasão prevista menor ou igual a 9% tendem a apresentar taxa de retenção entre 35% e 40%, e vice-versa.
- **Regra 3:** Estudantes com porcentagem de conclusão entre 49% e 51% apresentam evasão prevista relativamente alta (42–47%).
- **Regras 4 e 5:** Para estudantes do sexo masculino, observa-se o mesmo padrão identificado nas Regras 1 e 2: homens com baixa evasão prevista tendem a estar na faixa de retenção de 35–40%, e aqueles com retenção nessa faixa também tendem a apresentar baixa evasão prevista.
- **Regra 6:** Estudantes brancos com conclusão entre 49% e 51% tendem também a apresentar evasão prevista elevada (42–47%).
- **Regra 7:** Estudantes que ainda estão “em curso”, com conclusão entre 49% e 51%, apresentam igualmente evasão prevista alta (42–47%).

- **Regras 8 e 9:** Entre estudantes brancos, novamente surge a mesma estrutura de associação: baixa evasão prevista relaciona-se com retenção entre 35–40%, e essa retenção também aparece associada de volta à baixa evasão.
- **Regra 10:** Estudantes com conclusão entre 41% e 42% tendem a apresentar IEA baixo.

A terceira regra mostrou-se particularmente relevante, pois indica que alunos com taxa de conclusão na média tendem a apresentar elevados índices de evasão. As demais regras, por sua vez, revelaram-se menos expressivas, uma vez que corroboram resultados já conhecidos no contexto institucional analisado. Ainda assim, essas regras não devem ser desconsideradas, pois evidenciam que, embora o IEA exerça influência sobre a conclusão dos cursos, há casos de conclusão acadêmica mesmo entre estudantes com IEA baixo.

4.6.4 Resultados do *K-means*

Nessa nova base, o *K-means* foi aplicado no agrupamento em quatro grupos, buscando identificar categorias diferentes de alunos e então descobrir padrões incomuns nesses agrupamentos. Os resultados apresentam as seguintes informações para cada um dos quatro grupos identificados:

- **Grupo 0 (39%)**: Este grupo reúne alunos majoritariamente em situação de curso ativo, com faixa etária entre 20 e 24 anos e predominância de cursos da área de Informação e Comunicação. Trata-se de um perfil com alta carga horária, percentuais elevados de retidos e taxas de evasão também altas (41–47%). Esse conjunto caracteriza um grupo com baixa eficiência acadêmica e necessidade de acompanhamento contínuo para evitar evasão futura.
- **Grupo 1 (16%)**: Composto principalmente por alunos jovens (15 a 19 anos), matriculados em cursos de curta duração da área de Multimeios Didáticos. Apresenta carga horária significativamente menor (1000–1500 horas), menores índices de retidos (35–40%) e a menor taxa de evasão entre todos os grupos ($< 9\%$). Esses estudantes possuem perfil com maior probabilidade de progressão no curso, compondo um grupo de desempenho favorável e baixa evasão.
- **Grupo 2 (23%)**: Grupo composto em sua maioria por alunos em cursos de Gestão e Negócios, com carga horária média (1500–1900 horas). Embora mantenham situação ativa e apresentem desempenho moderado, possuem retidos elevados (acima de 57%) e taxas de evasão também altas (47–52%). Esse grupo representa um grupo de risco acadêmico, com desempenho mediano, mas com forte tendência à evasão se não houver intervenção.

- **Grupo 3 (22%)**: Formado predominantemente por alunos evadidos, com carga horária extremamente baixa (400 horas) e perfis concentrados em cursos de Formação Inicial e Continuada (FIC) e de Desenvolvimento Social. Apresenta as maiores porcentagens de concluintes e eficiência acadêmica final (IEA $>$ 52), mas isso ocorre porque grande parte dos que permanecem até o final concluem rapidamente. Representa o grupo crítico do ponto de vista da permanência, com evasão estruturada e já concretizada.

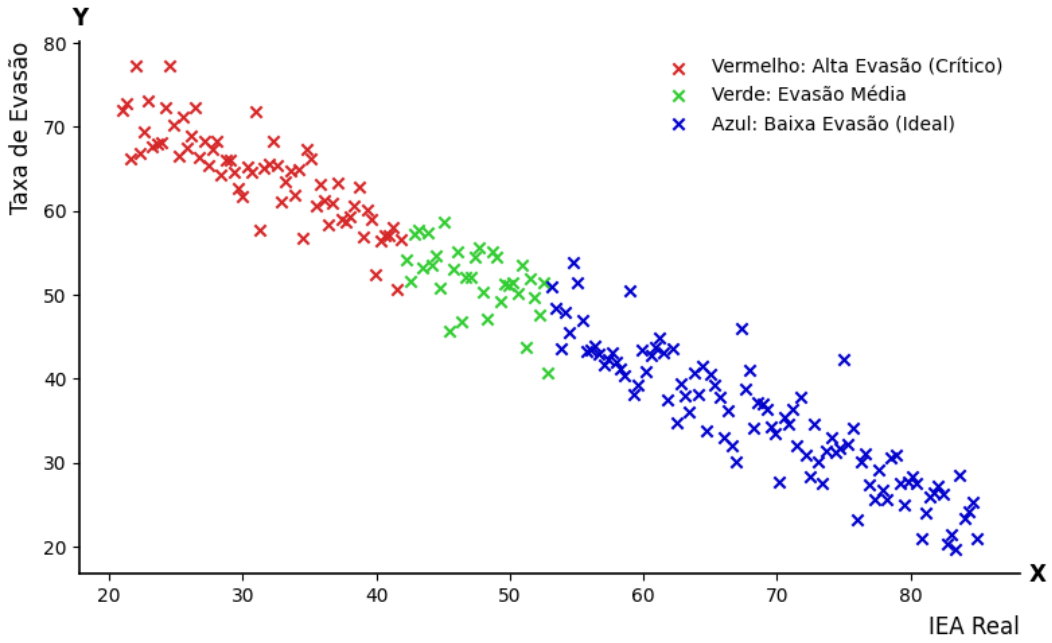
4.7 Análise das Correlações Identificadas

A aplicação dos algoritmos permitiu identificar e quantificar as correlações entre o IEA e as demais variáveis estudadas. Esta análise é fundamental para compreender os fatores que mais influenciam a eficiência acadêmica nos campi do IFSP.

4.7.1 Correlação entre IEA e Taxa de Evasão

A análise revelou uma correlação negativa forte entre o IEA e a Taxa de Evasão, conforme ilustrado na Figura 9. Esta relação inversa era esperada teoricamente, mas sua quantificação precisa fornece insights valiosos para a gestão acadêmica.

Figura 9 – Correlação entre IEA e Taxa de Evasão



Fonte: Elaborada pelo autor

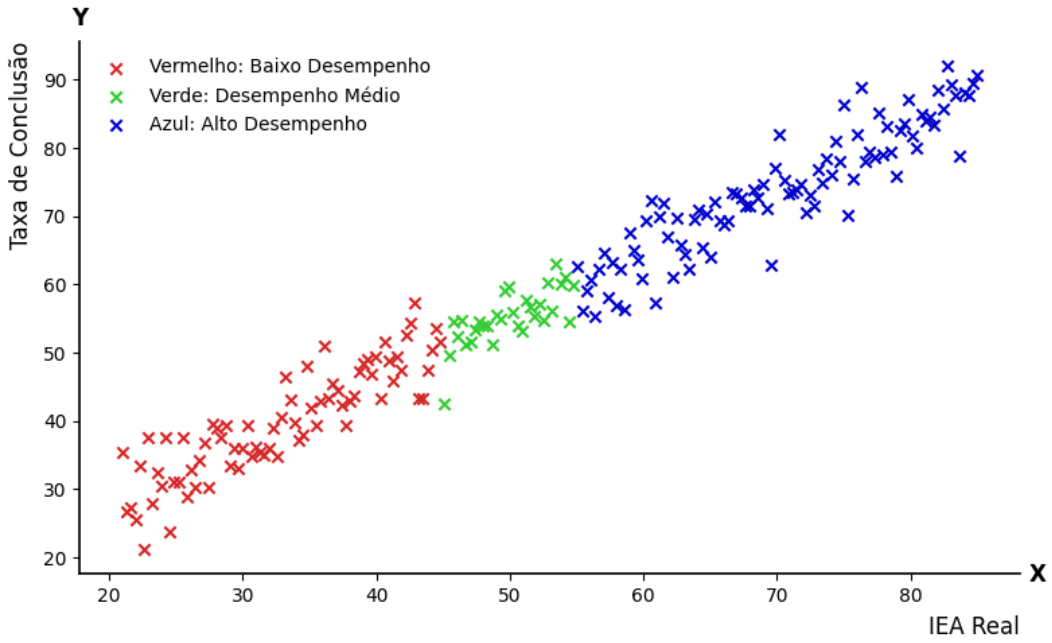
No gráfico da Figura 9 está demonstrado que campi com maiores taxas de evasão tendem a apresentar menores valores de IEA. Esta correlação negativa sugere que estratégias

eficazes de retenção de estudantes podem ter impacto direto na melhoria da eficiência acadêmica institucional.

4.7.2 Correlação entre IEA e Taxa de Conclusão

Em contraste com a relação anterior, o IEA apresenta correlação positiva com a Taxa de Conclusão, conforme mostrado na Figura 10. Esta relação direta confirma que campi com melhor desempenho acadêmico conseguem formar mais estudantes com sucesso.

Figura 10 – Correlação entre IEA e Taxa de Conclusão



Fonte: Elaborada pelo autor

A correlação positiva observada, na Figura 10, indica que investimentos em melhorias que elevem o IEA tendem a resultar em maiores taxas de conclusão, criando um ciclo virtuoso de melhoria da qualidade educacional.

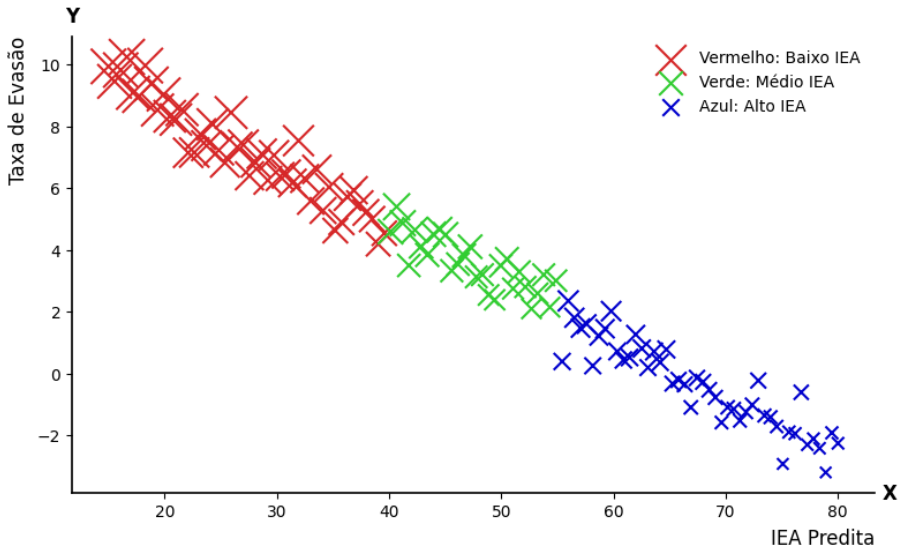
4.7.3 Correlação entre Carga Horária e Índice de Eficiência Acadêmica

Como apresentado como uma sugestão na Figura 7 da aplicação das bases unidas na regressão logística e na Figura 8, presente na aplicação do *AdaBoost* também nas bases unidas, a carga horária dos cursos tem uma relação direta com o IEA nesses mesmos cursos, de modo que, por fatores terceiros como por exemplo a complexidade desses cursos mais longos, ou até mesmo o interesse dos alunos que varia ao longo do tempo, esses cursos devem ser priorizados em medidas de retenção.

4.7.4 Modelo Preditivo para IEA baseado na Taxa de Evasão

Utilizando a Regressão Logística, foi desenvolvido um modelo preditivo que permite estimar o IEA com base na Taxa de Evasão, conforme apresentado na Figura 11. Este modelo tem aplicação prática importante para o planejamento e avaliação de políticas educacionais.

Figura 11 – Modelo Preditivo: IEA em função da Taxa de Evasão



Fonte: Elaborada pelo autor

O modelo, conforme apresentado na Figura 11, confirma matematicamente a relação inversa entre Taxa de Evasão e IEA, fornecendo uma ferramenta quantitativa para gestores educacionais estimarem o impacto de políticas de retenção no desempenho institucional.

4.7.5 Síntese Crítica dos Resultados Obtidos

Este capítulo demonstrou a aplicação bem-sucedida de técnicas de AM, as quais envolveram a análise do IEA e as Matrículas no âmbito do Instituto Federal de São Paulo. Os resultados obtidos são robustos e confirmam correlações importantes, mas requerem uma síntese que pondere tanto os seus pontos fortes quanto as suas limitações.

4.7.5.1 Pontos Positivos

- Alta Precisão Preditiva: Os modelos de Aprendizado Supervisionado, notadamente a Regressão Logística e o *AdaBoost*, alcançaram uma alta precisão. O *AdaBoost* classificou corretamente 99,23% das instâncias na primeira base e 100% na base combinada, enquanto a Regressão Logística apresentou coeficientes de correlação de até 0,9998 na base de dados combinada. Esta performance atesta a adequação dos

algoritmos escolhidos para prever a eficiência acadêmica e a qualidade dos dados utilizados.

- Validação de Correlações Chave: As correlações identificadas confirmaram relações teoricamente esperadas:

  1. O IEA correlaciona-se negativamente com a Taxa de Evasão;
  2. O IEA correlaciona-se positivamente com a Taxa de Conclusão.

- Descoberta de Padrões e Agrupamentos: Os algoritmos não supervisionados agregaram valor analítico, sendo que o *Apriori* revelou regras de associação com alta confiança (entre 0,83 e 1,00), como a forte correlação entre "IEA cat Alta" e "evasao cat Baixa" (*Lift* de 2,64). Por outro lado, o *K-means* permitiu a identificação e caracterização de perfis de alunos com alto, médio e baixo desempenho (Grupos 0, 1 e 2), fornecendo *insights* práticos para a intervenção pedagógica.

- Impacto da Carga Horária: O *AdaBoost* e a Regressão Logística destacaram que cursos de maior carga horária estão fortemente associados a um IEA baixo e alta taxa de evasão, um achado crucial para a priorização de ações de retenção.

4.7.5.2 Pontos Negativos e Considerações: Limitações e Desafios

- Risco de *Overfitting* (Base Combinada): A precisão de 100% do *AdaBoost* e o coeficiente de correlação de 0,9998 da Regressão Logística na base combinada, embora impressionantes, levantam a preocupação sobre um potencial *overfitting*, ou seja, sobreajuste. A alta acurácia pode ser um reflexo da distribuição de atributos de IEA categorizada do *dataset* de Eficiência Acadêmica para os Microdados de Matrículas, conforme observado na discussão do *AdaBoost*. Modelos com precisão tão elevada podem falhar em generalizar para novos dados não vistos.

- Regras Limitadas do *Apriori* (Base Combinada): Na aplicação na base de dados combinados, as regras encontradas pelo *Apriori* foram consideradas menos relevantes para complementar a compreensão dos resultados, oferecendo associações mais óbvias (ex: conclusão média → evasão média).

- Interpretabilidade (Black Box): Embora a Regressão Logística seja interpretável, os modelos de *Ensemble Learning* (*AdaBoost*) e o agrupamento *K-means* (em sua aplicação mais complexa) são, por natureza, menos transparentes.

Em síntese, os algoritmos de AM aplicados demonstraram ser ferramentas muito eficazes e altamente precisas para a análise e predição do IEA. Os resultados validam a

importância de políticas de combate à evasão e de apoio a cursos de maior carga horária como pilares centrais para a melhoria da eficiência acadêmica institucional. No entanto, o sucesso dos modelos de alta precisão deve ser interpretado com cautela, considerando o risco de sobreajuste e a necessidade de futuras validações em bases de dados internas (SUAP) para garantir a generalização e a aplicabilidade prática dos *insights* gerados, conforme sugerido em trabalhos futuros.

5 CONCLUSÕES

Neste capítulo são apresentadas as conclusões, considerando os resultados obtidos e também as discussões dos resultados apresentados para o estabelecimento do objetivo de investigar, por meio da aplicação de algoritmos de Aprendizado de Máquina, as correlações e associações entre fatores que influenciam a eficiência acadêmica no âmbito do Instituto Federal de São Paulo.

A metodologia adotada, que combinou modelos supervisionados e não supervisionados aplicados a dados oficiais do MEC, mostrou-se adequada para atender a esse objetivo, permitindo uma análise consistente e alinhada com o problema proposto.

De forma geral, os resultados confirmam que a eficiência acadêmica está fortemente associada a fatores como evasão, conclusão e características estruturais dos cursos, especialmente a carga horária. Essas relações, que já eram esperadas do ponto de vista teórico e institucional, foram validadas pelos modelos aplicados, organizando e hierarquizando os fatores de maior impacto de maneira objetiva e baseada em dados.

Sob a perspectiva da gestão educacional, os achados indicam que ações voltadas à redução da evasão e ao acompanhamento de cursos com maior carga horária devem ser tratadas como prioridades estratégicas. A identificação de perfis distintos de alunos e de padrões recorrentes de desempenho reforça a importância de substituir abordagens genéricas por intervenções mais direcionadas, com base em evidências quantitativas.

Conclusivamente, os resultados demonstraram que os algoritmos de Aprendizado de Máquina têm potencial para serem utilizados como ferramentas de apoio à gestão acadêmica, especialmente no monitoramento contínuo da eficiência institucional. Ainda assim, é necessário considerar as limitações observadas, como o risco de sobreajuste em bases combinadas e a menor interpretabilidade de alguns modelos, o que reforça a necessidade de validações adicionais antes de sua aplicação em ambiente operacional.

5.1 Recomendações e Trabalhos Futuros

Com base nos resultados obtidos, esta seção apresenta recomendações para a gestão do câmpus e indica possibilidades de continuidade do trabalho, considerando tanto os achados quanto as limitações identificadas.

5.1.1 Recomendações de Ações para a Gestão do Câmpus

- As recomendações propostas têm como foco principal o combate à evasão e o apoio a cursos e perfis de alunos com maior risco acadêmico, conforme os padrões identificados pelos modelos aplicados.
- Como primeira ação, recomenda-se a utilização de modelos preditivos para a criação de um sistema de alerta proativo de evasão, permitindo a identificação antecipada de estudantes com maior probabilidade de abandono. Essa abordagem possibilita a realização de intervenções mais rápidas e direcionadas, antes que o desengajamento acadêmico se consolide.
- Outra recomendação é o direcionamento prioritário de ações de apoio pedagógico para cursos com maior carga horária, que apresentaram maior associação com baixos níveis de eficiência acadêmica. Medidas como tutoria, monitoria e acompanhamento contínuo podem contribuir para reduzir os fatores de risco identificados.
- Também se destaca a importância de personalizar as intervenções com base nos perfis de alunos identificados pelos métodos de agrupamento, permitindo que a gestão ofereça ações diferenciadas de acordo com o nível de desempenho e risco acadêmico de cada grupo.
- Por fim, recomenda-se a utilização dos padrões e associações identificados como subsídio para a criação ou validação de regras de negócio e indicadores dentro do sistema SUAP, de modo a transformar os resultados analíticos em instrumentos operacionais de apoio à gestão.

5.1.2 Trabalhos Futuros

Como continuidade deste trabalho, destaca-se a necessidade de validar os modelos em bases de dados internas do SUAP, de forma a mitigar o risco de sobreajuste e garantir a generalização dos resultados em um contexto real de aplicação.

Além disso, sugere-se a inclusão de variáveis comportamentais e de processo, como frequência, participação em atividades acadêmicas e uso de recursos institucionais, com o objetivo de ampliar a interpretabilidade dos modelos e identificar fatores mais imediatos relacionados ao desempenho dos alunos.

Por fim, propõe-se o desenvolvimento de uma ferramenta de visualização gerencial, na forma de um dashboard, que integre os modelos preditivos e os resultados de agrupamento, permitindo que a gestão acompanhe de forma contínua os indicadores de eficiência acadêmica e os riscos associados à evasão.

REFERÊNCIAS

AL-ALAWI, L. et al. Using machine learning to predict factors affecting academic performance: the case of college students on academic probation. **Education and Information Technologies**, Springer, v. 28, n. 10, p. 12407–12432, 2023. 11, 12, 15, 17

ALHAKAMI, H.; ALSUBAIT, T.; ALJARALLAH, A. Data mining for student advising. **International Journal of Advanced Computer Science and Applications**, Science and Information (SAI) Organization Limited, v. 11, n. 3, p. 526–532, 2020. 12, 14

ALMASRI, A. et al. Explainable artificial intelligence models using students' academic record data, tree family classifiers, and k-means clustering to predict students' performance. In: IEEE. **2022 10th International Conference on Smart Grid (icSmartGrid)**. [S.l.], 2022. p. 46–51. 16

BAESSA, R. N. F. et al. Estudo de caso com análise de dados para a detecção da desistência de estudantes em disciplinas ofertadas com apoio do ambiente moodle e sentimentos coletados ativamente por questionários. **Repositório Institucional UFSC**, Joinville, SC., 2024. 14

BENEVENTO, M.; MEIRELLES, F. d. S. Prever e melhorar o desempenho dos alunos com o uso combinado de aprendizagem de máquina e gpt. **Regae: Revista de Gestão e Avaliação Educacional**, Autores mantém os direitos autorais e concedem ao periódico o direito de . . . , v. 12, n. 21, 2023. 16, 17

CONTREAS-BRAVO, L. E.; NIEVES-PIMIENTO, N.; GUERRERO, K. G. Prediction of university-level academic performance through machine learning mechanisms and supervised methods. **Ingeniería**, Universidad Distrital Francisco José de Caldas, v. 28, n. 1, 2023. 12, 14

CZIBULA, G. et al. Intellidam: A machine learning-based framework for enhancing the performance of decision-making processes. a case study for educational data mining. **IEEE Access**, IEEE, v. 10, p. 80651–80666, 2022. 11

DEVKISHAN, T. S.; SINGH, S. K.; BHARTI, A. K. Ensemble learning for student performance assessment: Identifying and analyzing significant affecting factors in higher education. In: IEEE. **2024 7th International Conference on Contemporary Computing and Informatics (IC3I)**. [S.l.], 2024. v. 7, p. 271–277. 13

DUONG, H. T.-H. et al. Academic performance warning system based on data driven for higher education. **Neural Computing and Applications**, Springer, v. 35, n. 8, p. 5819–5837, 2023. 9, 15

GUO, H. et al. Machine-learning based mooc learning data analysis. In: IEEE. **2021 7th IEEE Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing,(HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS)**. [S.l.], 2021. p. 63–68. 15

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3rd. ed. [S.l.]: Morgan Kaufmann, 2012. 35

KAMAL, M. et al. Metaheuristics method for classification and prediction of student performance using machine learning predictors. **Mathematical Problems in Engineering**, Wiley Online Library, v. 2022, n. 1, p. 2581951, 2022. 11

KAUR, B.; GUPTA, A.; SINGLA, R. K. Descriptive statistical analysis and discretization of academic data for machine learning techniques. In: IEEE. **2023 10th International Conference on Computing for Sustainable Global Development (INDIACom)**. [S.l.], 2023. p. 1494–1499. 9, 12

KHAN, I. et al. A conceptual framework to aid attribute selection in machine learning student performance prediction models. **International Journal of Interactive Mobile Technologies**, v. 15, n. 15, 2021. 16

LAWANONT, W.; TIMTONG, A. Smart education using machine learning for outcome prediction in engineering course. In: IEEE. **2022 14th International Conference on Knowledge and Smart Technology (KST)**. [S.l.], 2022. p. 63–68. 13

LI, Y.; ZHANG, H. Big data technology for teaching quality monitoring and improvement in higher education-joint k-means clustering algorithm and apriori algorithm. **Systems and Soft Computing**, Elsevier, v. 6, p. 200125, 2024. 14

LIMA, G. S. d.; ÁVILA, P. M. d.; GILAVERTE, R. Implementação de um modelo para previsão de evasão escolar no ifsuldeminas. 2014. 17

MAHESHWARI, A. et al. Comparative analysis of machine learning models in predicting academic outcomes: insights and implications for educational data analytics. In: IEEE. **2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)**. [S.l.], 2024. p. 1–7. 13, 14, 15

MORAES, G. H.; ALMEIDA, S.; ALVES, T. Plataforma nilo peçanha: guia de referência metodológica. **Brasília/DF: Editora Evobiz**, v. 101, 2018. 25, 30

OWUSU-BOADU, B. et al. Academic performance modelling with machine learning based on cognitive and non-cognitive features. **Applied Computer Systems**, Sciendo, v. 26, n. 2, p. 122–131, 2021. 9, 14, 15

Python Software Foundation. **Python Official Website**. 2024. Accessed: 2025-06-25. Disponível em: <https://www.python.org/>. 25

RAJU, R. et al. Educational data mining: A comprehensive study. In: IEEE. **2020 International Conference on System, Computation, Automation and Networking (ICSCAN)**. [S.l.], 2020. p. 1–5. 11

RIBEIRO, M. V. M. O impacto da inteligência artificial na educação: oportunidades e desafios nas escolas. **REVISTA DELOS**, v. 17, n. 61, p. e2309–e2309, 2024. 16

RODRIGUES, E. M. et al. Aprendizado de maquina para agrupamento e associacao de dados do ensino superior publico brasileiro. **Revista de Sistemas e Computação-RSC**, v. 13, n. 1, 2023. 16

SAGALA, T. N. et al. Predicting computer science student's performance using logistic regression. In: IEEE. **2022 5th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)**. [S.l.], 2022. p. 817–821. 13, 30

SATHE, M. T.; ADAMUTHE, A. C. Comparative study of supervised algorithms for prediction of students' performance. **International Journal of Modern Education and Computer Science**, Modern Education and Computer Science Press, v. 13, n. 1, p. 1, 2021. 16

SIHARE, M.; GUPTA, R. K. Evaluation of machine learning methods for prediction student performance. **International Journal for Research in Applied Science & Engineering Technology (IJRASET)**, v. 12, n. 1, p. 534, January 2024. ISSN 2321-9653. SJ Impact Factor: 7.538, IC Value: 45.98. Disponível em: <https://www.ijraset.com>. 12, 14, 17

SILVA, G. H. et al. Utilização de aprendizado de máquina na predição de desempenho acadêmico. Insitituto Federal de Educação, Ciência e Tecnologia de Goiás, 2025. 9, 12, 15

SILVA, R. C. C. da et al. Modelagem preditiva para sucesso acadêmico: um estudo de caso em um curso de ciência da computação. **REVISTA DELOS**, v. 18, n. 63, p. e3586–e3586, 2025. 12, 15, 16

SUN, D. et al. A university student performance prediction model and experiment based on multi-feature fusion and attention mechanism. **IEEE Access**, IEEE, v. 11, p. 112307–112319, 2023. 11

WIRADINATA, T. et al. An implementation of support vector machine classification for developer academy acceptance prediction model. In: IEEE. **2021 2nd International Conference on Innovative and Creative Information Technology (ICITech)**. [S.l.], 2021. p. 110–116. 14

YOUSUF, E.; WAHID, A.; KHAN, M. Exploring the effectiveness of ai algorithms in predicting and enhancing student engagement in an e-learning. **International journal on recent and innovation trends in computing and communication**, v. 11, n. 10, p. 23–29, 2023. 16

ZHANG, C. et al. Simulation-based machine learning for predicting academic performance using big data. **International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)**, IGI Global, v. 16, n. 1, p. 1–20, 2024. 12, 15, 16