Universidade Federal do Piauí
Centro de Ciências da Natureza
Programa de Pós-Graduação em Ciência da Computação

Avaliação de modelos preditivos baseados em aprendizagem de máquina no contexto da evasão escolar considerando um cenário multicampi

Francisco Alysson da Silva Sousa

Teresina-PI, Abril de 2024

Francisco Alysson da Silva Sousa

Avaliação de modelos preditivos baseados em aprendizagem de máquina no contexto da evasão escolar considerando um cenário multicampi

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Universidade Federal do Piauí – UFPI

Centro de Ciências da Natureza

Programa de Pós-Graduação em Ciência da Computação

Orientador: Prof. Dr. Vinicius Ponte Machado

Teresina-PI

Abril de 2024

FICHA CATALOGRÁFICA
Universidade Federal do Piauí
Biblioteca Comunitária Jornalista Carlos Castello Branco
Divisão de Representação da Informação

S725a Sousa, Francisco Alysson da Silva.
Avaliação de modelos preditivos baseados em aprendizagem de máquina no contexto da evasão escolar considerando um cenário multicampi / Francisco Alysson da Silva Sousa. – 2024.
60 f.

Dissertação (Mestrado) – Universidade Federal do Piauí, Centro de Ciências da Natureza, Programa de Pós-Graduação em Ciência da Computação, Teresina, 2024.
"Orientador: Prof. Dr. Vinicius Ponte Machado".

1. Aprendizagem da máquina. 2. Evasão escolar. 3. Predição. 4. Avaliação de classificadores. I. Machado, Vinicius Ponte. II. Título.

CDD 004.07

Bibliotecária: Francisca das Chagas Dias Leite – CRB3/1004

Francisco Alysson da Silva Sousa

Avaliação de modelos preditivos baseados em aprendizagem de máquina no contexto da evasão escolar considerando um cenário multicampi

Defesa de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da UFPI (área de concentração: Sistemas de Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência da Computação.

Trabalho aprovado. Teresina-PI, 05 de Abril de 2024:

Documento assinado digitalmente
gov.br
VINICIUS PONTE MACHADO
Data: 10/04/2024 07:37:55-0300
Verifique em https://validar.iti.gov.br

Prof. Vinicius Ponte Machado
Presidente da banca examinadora

Documento assinado digitalmente
gov.br
LUCIANO REIS COUTINHO
Data: 10/04/2024 10:54:28-0300
Verifique em https://validar.iti.gov.br

Prof. Luciano Reis Coutinho
Examinador externo à instituição

Documento assinado digitalmente
gov.br
RODRIGO DE MELO SOUZA VERAS
Data: 10/04/2024 15:51:28-0300
Verifique em https://validar.iti.gov.br

Prof. Rodrido de Melo Souza Veras
Examinador interno

Documento assinado digitalmente
gov.br
ANDRE MACEDO SANTANA
Data: 16/04/2024 12:16:21-0300
Verifique em https://validar.iti.gov.br

Prof. André Macedo Santana
Examinador interno

Teresina-PI
Abril de 2024

Aos meus pais, Maria Raimunda e Raimundo Nonato,
por sempre incentivarem os filhos na jornada da educação.

Agradecimentos

A Deus, pelo dom da vida.

Às professoras Paula e Rosiana, minhas irmãs. Ao Adriel e Dário, meus irmãos. A todos eles, por serem sempre a minha base e meus grandes incentivadores.

Aos professores do Mestrado, em especial ao grande orientador, prof. Vinicius Ponte Machado, por todos os conselhos, paciência e ensinamentos.

À minha esposa, Anny Caroliny, pelo apoio no período da minha dedicação a este trabalho e por tanto se dedicar aos nossos filhos, Lucas e Luan.

"Felizes são os que recebem a palavra de Deus e a guardam"
(Lucas 11:28)

Resumo

A evasão escolar e os diversos fatores relacionados a esse comportamento despontam como um dos grandes desafios ao pleno desenvolvimento da educação em muitos países. No Brasil, estima-se que 27% do total de alunos matriculados não concluem o percurso formativo previsto nas respectivas ofertas. Delimitando-se na proposta da Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT), especificamente na Educação Profissional Técnica de Nível Médio (EPTNM), esse estudo identificou um decréscimo no número de concluintes nos últimos cinco anos, conforme estatísticas oficiais da Plataforma Nilo Peçanha (PNP), sistema do Ministério da Educação alimentado pelas instituições integrantes da RFEPCT. Assim, a identificação de forma prévia da tendência à evasão certamente contribui como subsídio para o direcionamento de estratégias proativas visando a permanência e êxito discente. Nesse sentido, o uso da inteligência artificial, especificamente na subárea aprendizagem de máquina, apresenta-se como importante recurso de apoio à gestão educacional. É nessa perspectiva que se define a proposta deste trabalho em avaliar classificadores quanto a ocorrência da evasão no âmbito do ensino técnico multicampi. Para tanto, foram aplicados os algoritmos de *Decision Tree* (DT), *Random Forest* (RF), *Gradient Boost* (GB), *Multi-Layer Perceptron* (MLP) e *Support Vector Machine* (SVM). Os modelos utilizados foram submetidos a análises comparativas a partir de estudos de casos com dados extraídos unicamente da PNP. Os resultados dos modelos melhor avaliados (DT e RF) apresentam médias superiores a 90% quando consideradas todas as unidades da instituição em testes segmentados e agrupados, destacando ainda uma tendência de evolução nas performances que favorece o potencial escalável da proposta. Os resultados apresentados expressam a sensibilidade (*recall*) e a precisão (*precision*) com suas relevâncias equiparadas e resumidas pela métrica *F-score*.

Palavras-chaves: Aprendizagem de máquina. Evasão escolar. Predição. Avaliação de classificadores.

Abstract

School dropout and the various factors related to this behavior emerge as one of the greatest challenges to the full development of education in many countries. In Brazil, it is estimated that 27% of the total number of students enrolled do not complete the training path provided for in the respective offers. Delimited by the proposal of the Federal Network for Professional, Scientific and Technological Education (RFEPCT), specifically in Secondary Technical Professional Education (EPTNM), this study identified a decrease in the number of graduates in the last five years, according to official statistics from the Platform Nilo Peçanha (PNP), a system from the Ministry of Education fed by the institutions that are members of the RFEPCT. Thus, the prior identification of the tendency to drop out certainly contributes to support proactive strategies aimed at student retention and success. In this sense, the use of artificial intelligence, specifically in the machine learning subarea, presents itself as an important resource to support educational management. It is from this perspective that the proposal of this work is defined to evaluate classifiers regarding the occurrence of dropout in the context of multicampi technical education. To this end, the *Decision Tree* (DT), *Random Forest* (RF), *Gradient Boost* (GB), *Multi-Layer Perceptron* (MLP) and *Support Vector Machine* (SVM). The models used were subjected to comparative analyzes based on case studies with data extracted solely from the PNP. The results of the best evaluated models (DT and RF) present averages above 90% when considering all units of the institution in segmented and grouped tests, also highlighting a trend of evolution in performances that favors the scalable potential of the proposal. The results presented express sensitivity (*recall*) and precision (*precision*) with their relevance equated and summarized by the *F-score* metric.

Keywords: Machine learning. School dropout. Prediction. Classifier evaluation.

Lista de ilustrações

Figura 1 – Evolução do número de matrículas na EPT e EPTNM de 2018 a 2023. 3
Figura 2 – Evolução de indicadores da EPTNM no Piauí no período de 2018 a 2022. 3
Figura 3 – Visão geral da proposta para obtenção de resultados segmentados por campi e geral sobre a base institucional. 6
Figura 4 – Exemplo de Mineração de Dados no contexto educacional. Adaptado de (ROMERO; ROMERO; VENTURA, 2014). 8
Figura 5 – Fluxos da Mineração de Dados Educacionais. Adaptado de (AGGARWAL et al., 2015) 9
Figura 6 – Fluxo de etapas da aplicação de Aprendizagem de Máquina. Adaptado de (VIANA; SANTANA; RABêLO, 2022) 10
Figura 7 – Composição da base de dados da Plataforma Nilo Peçanha 14
Figura 8 – Representação da coleta e delimitação de dados aplicada a cada edição publicada pela PNP até o momentento deste estudo. 18
Figura 9 – Distribuição de frequência por fator de esforço do curso entre evadidos (à esquerda) e não evadidos (à direita). 21
Figura 10 – Definição do fator de esforço em níveis com base em faixa de valores . . 21
Figura 11 – Tendência de nível entre evadidos e não evadidos. 22
Figura 12 – Representação de quais bases compõem cada estudo de caso. 24
Figura 13 – Composição total incluindo alunos ativos e não ativos. 25
Figura 14 – Composição das bases individuais dos estudos de casos analisados. . . . 25
Figura 15 – Composição das bases concatenadas para EC2 e EC3 26
Figura 16 – Visão geral da metodologia 27
Figura 17 – Representação do procedimento para validação com matrículas ativas. . 28
Figura 18 – Evolução das médias de desempenho do estudo de caso EC1. 33
Figura 19 – Médias de desempenho registradas em EC2 e EC3. 34
Figura 20 – Melhores desempenhos registrados em EC2 e EC3. 35
Figura 21 – Variação das médias, por algoritmo, com as bases anuais em EC4. . . . 37
Figura 22 – Quantidade de matrículas ativas atualizadas e registros não identificados no ano seguinte. 39
Figura 23 – Quantidade de matrículas atualizadas agrupadas por classe. 39
Figura 24 – Quantidade de matrículas atualizadas por situação. 40
Figura 25 – Comparativo da precisão nas predições geradas por *Decision Tree* e *Random Forest*. 41
Figura 26 – Comparativo quanto a sensibilidade (*recall*) dos algoritmos *Decision Tree* e *Random Forest*. 42

Figura 27 – Comparativo quanto a precisão e sensibilidade dos modelos com os algoritmos DT e RF. 43
Figura 28 – Comparação das médias gerais de F1-score por modelos a cada ano. . . 45
Figura 29 – Comportamento quanto ao desvio padrão sobre as médias registradas. . 46
Figura 30 – Comparação entre os melhores valores de f1 considerando todos os campi e a instituição completa. 46

Lista de tabelas

Tabela 1 – Características gerais dos trabalhos relacionados. 13
Tabela 2 – Formato da informação dos indicadores da PNP 15
Tabela 3 – Categorias vigentes para situações de matrículas 16
Tabela 4 – Descrição das situações de matrículas relacionadas à evasão. 17
Tabela 5 – Aplicação de nomenclatura padrão aos atributos. 20
Tabela 6 – Agrupamento das situações de matrículas 20
Tabela 7 – Descrição de recursos do ambiente experimental 23
Tabela 8 – Atributos das edições PNP utilizadas. 29
Tabela 9 – Desempenho dos algoritmos considerando a média de resultados por base anual quando aplicados em todos campi (EC1). 32
Tabela 10 – Médias, por algorítmos, com bases concatenadas dos estudos de casos EC2 e EC3, considerando execução única por campi. 34
Tabela 11 – Resultados do estudo de caso EC4, ano 2018. 36
Tabela 12 – Resultados do estudo de caso EC4, ano 2019 36
Tabela 13 – Resultados do estudo de caso EC4, ano 2020. 36
Tabela 14 – Resultados do estudo de caso EC4, ano 2021. 37
Tabela 15 – Resultados do estudo de caso EC4, ano 2022. 37
Tabela 16 – 2018: Melhores resultados na predição aplicada às matrículas então ativas no ano de 2018. 41
Tabela 17 – Melhores resultados das predições com matrículas ativas em 2019. . . . 43
Tabela 18 – Melhores resultados das predições com matrículas ativas em 2020. . . . 44
Tabela 19 – Melhores resultados das predições com matrículas ativas em 2021 . . . 44
Tabela 20 – Resultados detalhados das melhores performances na validação com matrículas ativas em 2018. 55
Tabela 21 – Resultados detalhados das melhores performances na validação com matrículas ativas em 2019. 56
Tabela 22 – Resultados detalhados das melhores performances na validação com matrículas ativas em 2020.. 57
Tabela 23 – Resultados detalhados das melhores performances na validação com matrículas ativas em 2021. 58

Lista de abreviaturas e siglas

PNP	*Plataforma Nilo Peçanha*
IFPI	*Instituto Federal do Piauí*
REVALIDE	*Rede de Coleta, Validação e Disseminação das Estatísticas da Rede Federal*
EPT	*Educação Profissional e Tecnológica*
RFEPCT	*Rede Federal de Educação Profissional, Científica e Tecnológica*
EPTNM	*Educação Profissional Técnica de Nível Médio*
AM	*Aprendizagem de Máquina*
KDD	*Knowledge Discovery in Database*
MEC	*Ministério da Educação*
MDE	*Mineração de Dados Educacionais*
SISTEC	*Sistema Nacional de Informações da Educação Profissional e Tecnológica*
SIAFI	*Sistema Integrado de Administração Financeira*
SIAPE	*Sistema de Administração de Recursos Humanos*
DT	*Decision Tree*
RF	*Random Forest*
GB	*Gradient Boosting*
MLP	*Multilayer Perceptron*
SVM	*Support Vector Machine*

Sumário

1 INTRODUÇÃO 1

Introdução 1
1.1 Contexto e Motivação 2
1.2 Objetivos 4
1.3 Visão geral 5
1.4 Estrutura do trabalho 6

2 REFERENCIAL TEÓRICO 7
2.1 Descoberta de Conhecimento em Bases de Dados 7
2.2 Mineração de Dados Educacionais 8
2.3 Aprendizagem de máquina 9
2.4 Trabalhos relacionados 10

3 MATERIAIS E MÉTODOS 14
3.1 Origem e descrição dos dados 14
3.2 Etapas do processo 17
3.2.1 Coleta de dados 17
3.2.2 Pré-processamento 19
3.2.3 Normalização e Transformação 22
3.3 Ambiente Experimental 23
3.4 Metodologia 27
3.5 Métricas de avaliação 30

4 RESULTADOS E DISCUSSÕES 32
4.1 Resultados EC1 32
4.2 Resultados EC2 e EC3 33
4.3 Considerações sobre EC1, EC2 e EC3 34
4.4 Resultados EC4 35
4.5 Etapa de validação: predição de alunos ativos 38
4.5.1 Alunos ativos 2018 40
4.5.2 Alunos ativos 2019 42
4.5.3 Alunos ativos 2020 44
4.5.4 Alunos ativos 2021 44
4.6 Considerações sobre a validação com alunos ativos 45

5 CONCLUSÕES E TRABALHOS FUTUROS 47

5.1 Trabalhos futuros 48

REFERÊNCIAS 50

APÊNDICES 54

APÊNDICE A – RESULTADOS COM MATRÍCULAS ATIVAS 2018 55

APÊNDICE B – RESULTADOS COM MATRÍCULAS ATIVAS 2019 56

APÊNDICE C – RESULTADOS COM MATRÍCULAS ATIVAS 2020 57

APÊNDICE D – RESULTADOS COM MATRÍCULAS ATIVAS 2021 58

ANEXOS 59

ANEXO A – NOMENCLATURA OFICIAL DE CAMPI DO INSTITUTO FEDERAL DO PIAUÍ 60

1 Introdução

A descontinuidade discente nos diversos ciclos formativos é uma realidade em sistemas educacionais de muitos países. Observando em uma perspectiva global as análises do Programa das Nações Unidas para o Desenvolvimento (PNUD), que considera a educação com uma das dimensões para promoção do desenvolvimento humano, o Brasil apresenta nesse panorama indicadores como a média de anos de permanência na escola bem inferiores ao estimado nas projeções desejadas (FILHO; ARAÚJO, 2017a).

Na educação básica, dados do censo escolar de 2022 revelam que, apenas no ensino médio, a taxa dos que não permanecem até a conclusão supera os 19%, destacando-se o maior percentual desses desligamentos ainda no ano de início desta etapa. Cabe ressaltar que essas constatações ampliam-se ainda no ensino superior onde, embora o número de ingressantes tenha aumentado nos últimos anos, o mesmo acréscimo não se aplica quando analisados os indicadores de concluintes (BRASIL, 2023).

Esse representativo número de desligamentos, nem sempre precedidos de uma solicitação formalizada, é tema de estudos que investigam o problema em diferentes níveis e modalidades de ofertas, sobretudo no sistema público de ensino tendo em vista a otimização no dispêndio de recursos.

Dessa forma, é na definição de aluno evadido que algumas percepções não únicas são encontradas como, no entendimento do Ministério da Educação (MEC), trata-se da saída definitiva da etapa frequentada pelo aluno antes de sua conclusão, distinguindo evasão do curso ou sistema (BRASIL, 2023). Esse conceito é complementado em Moraes (2020) quando especifica ainda os casos de abandono, cancelamento, desligamento e transferências. Essas situações constituem um único grupo considerado no guia de referência.

Contudo, para além das definições, convergentes ou não, o ato de deixar de frequentar a instituição de ensino se revela de compreensão complexa dada a diversidade de contextos sociais, regionais, culturais e socioeconômicos de comum ocorrência nesse comportamento.

Na visão exposta em David e Chaym (2019), por exemplo, aspectos pedagógico e de infraestrutura são potenciais para o insucesso de permanência. Ainda em 2009, um levantamento da Fundação Getúlio Vargas apresentado em Neri et al. (2009) já verificava a relação com o mercado de trabalho, quando da necessidade de inserção prematura, em detrimento das oportunidades de elevação no nível de qualificação. Observa-se ainda a dinamicidade das variáveis associadas ao problema em questão quando constatado em Gómez e Belmonte (2020) a sobreposição dos fatores vocacionais sobre os comumente citados, como renda e necessidade de assistência.

Depreende-se destas verificações amplamente encontradas na literatura, que as várias caracterizações possíveis já sinalizam não se tratar de um trivial estudo encontrar padrões na evasão escolar, muito menos a formulação de direcionadas estratégias de enfrentamento.

Em Chiquitto e Baida (2020) percebe-se esta dificuldade quando buscaram verificar a correlação de indicadores de rendimento acadêmico com dados do atlas de desenvolvimento humano no Brasil, o objetivo foi buscar compreender a origem dos prejuízos gerados sob a perspectiva do aluno e da instituição. Já em Oliveira, Medeiros e Andrade (2022) as autoras propuseram atribuir um grau de relevância às informações comuns aos que abandonam a escola. Visou-se no estudo otimizar as intervenções preventivas baseadas nas indicações resultantes da aplicação de algoritmos. Nesse mesmo sentido, temos o uso da técnica de aprendizagem de máquina com resultados otimizados na combinação com modelos computacionais preditivos em (SOUZA; CAZELLA, 2022b).

Os trabalhos mencionados, entre outros analisados, embora com suas particularidades metodológicas, evidenciam as contribuições da tecnologia da informação para se conhecer de forma prévia os discentes com tendência a evadir. No entanto, é comum o o foco no desenvolvimento de estudos com visão ampla, quando um conjunto de especificidades pode revelar ganhos na compreensão do problema com apontamentos locais e, consequentemente, uma maior eficácia no planejamento das ações de acompanhamento (ROMERO; VENTURA, 2020).

Nessa perspectiva, abordaremos a aplicação de recursos computacionais preditivos em demandas emergentes no contexto da educação profissional. Assim, sustenta-se a delimitação deste trabalho quando considerado, dentro do abrangente desafio da evasão escolar, a verificação desta adversidade na forma articulada ao ensino médio.

Para tanto, utilizaremos dados da Plataforma Nilo Peçanha (PNP), sistema oficial de estatísticas da Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) instituído em 2018 pela Secretaria de Educação Profissional e Tecnológica (Setec). Utilizamos informações de referência anual do período de 2018 a 2022 avaliando modelos preditivos em bases por campi e intituição considerando a abrangência do Instituto Federal do Piauí (IFPI).

1.1 Contexto e Motivação

A Educação Profissional e Tecnológica (EPT) tem se destacado como a modalidade com maior crescimento em número de matrículas. Conforme a Figura 1, dados do censo escolar 2023 apontam um crescimento de 26% a partir de 2018. Deste total, 24,7% estão vinculadas à rede federal de ensino.

No âmbito da Educação Profissional Técnica de Nível Médio (EPTNM), a tendência de ampliação se mantém quando observado o mesmo período. Foram de 1.808.917 matrículas em 2018 para 2.271.607 em 2023 (BRASIL, 2024).

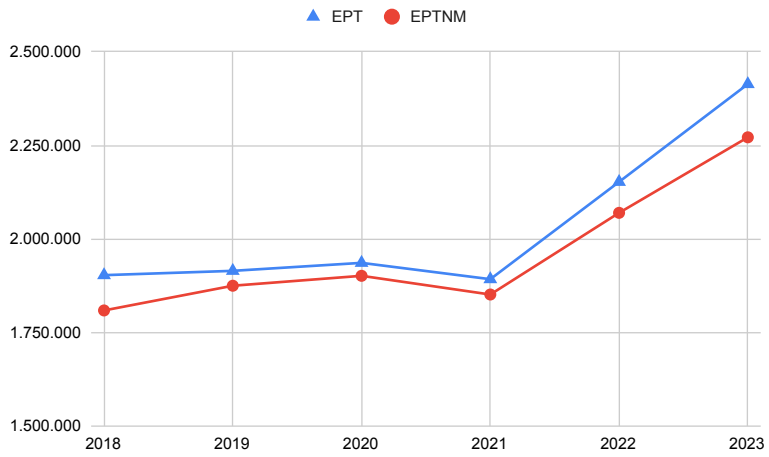

Figura 1 – Evolução do número de matrículas na EPT e EPTNM de 2018 a 2023.

Com referência à Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT), as unidades no Piauí são: os Colégios Técnicos, vinculados à Universidade Federal (CT/UFPI), e o Instituto Federal (IFPI). Para estas instituições buscou-se apresentar um panorama a partir dos dados disponíveis na PNP considerando o mesmo período da representação anterior.

No mencionado contexto estadual, os números da EPTNM, embora alinhados com a expansão nacional, revelam comportamentos que certamente instigam estudos. Na Figura 2 as informações sobre o número de vagas e ingressantes não alcançam equivalência, sinalizando dificuldades em contemplar todas as oportunidades ofertadas. Para concluíntes, a linha de tendência não converge com a expectativa a partir dos demais indicadores.

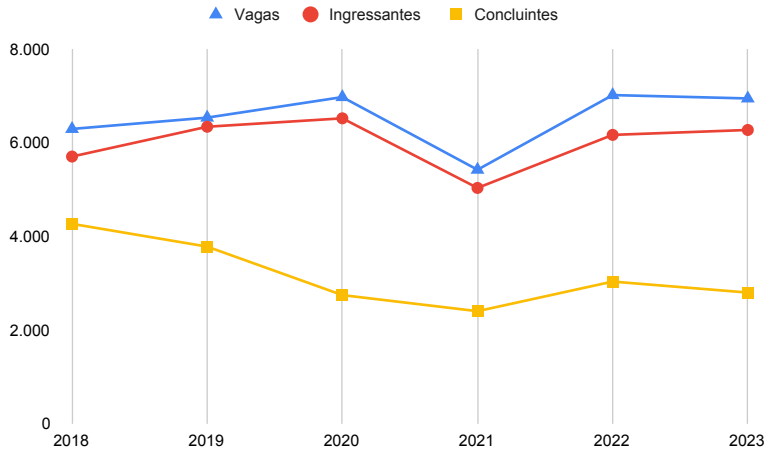

Figura 2 – Evolução de indicadores da EPTNM no Piauí no período de 2018 a 2022.

Logo, ao se analisar o levamento, observa-se facilmente um distanciamento das expectativas que norteiam os objetivos do ensino profissional. Esse notório contraponto incorpora fatores implícitos de difícil unificação e que, quando possíveis de identificação prévia, podem se tornar percepções de importantes contribuições para evitar a não conclusão como desfecho (COIMBRA; SILVA; COSTA, 2021).

Ao despertar para as investigações acadêmicas originadas a partir dessa realidade, verifica-se em destaque a ocorrência de trabalhos localizados e de abordagem geralmente limitada a certa unidade de ensino ou a cursos específicos. Diretamente relacionada à RFEPCT podemos citar trabalhos como os vistos em Assis (2020) e Pereira e Passos (2017). Estes autores contemplaram em suas pesquisas os fatores de insucesso no curso de Matemática e a eficácia da Política de Assistência Estudantil (POLAE), respectivamente, ambos direcionados aos campi centrais.

Reconhecendo limitações no monitoramento nacional da educação profissional, o MEC criou em 2018 a Rede de Coleta, Validação e Disseminação das Estatísticas da Rede Federal (REVALIDE). Essa iniciativa organiza em edições anuais uma base de dados com os principais indicadores de gestão extraídos de informações validadas pelas próprias instituições integrantes (BRASIL, 2018).

Contudo, especificamente sobre essa fonte, constata-se ainda uma escassez de estudos preditivos descentralizados. As limitações ocorrem em detrimento do potencial de contribuições que podem ser obtidas a partir das análises inferenciais multicampi. Esse contexto converge integralmente com os objetivos da Aprendizagem de Máquina (AM), área da tecnologia da informação apoiada na identificação de padrões para a predição de comportamentos a partir das observações (OLIVEIRA; MEDEIROS; ANDRADE, 2022).

Diante do exposto, apresenta-se como motivação fundamental para este trabalho contribuir com uma análise preditiva a partir das informações unificadas pela REVALIDE. São consideradas as especificidades de um cenário de ensino multicampi e o conjunto completo da instituição, auxiliando assim com subsídios às estratégias proativas de combate à evasão escolar.

1.2 Objetivos

Partindo dessa instigação, o objetivo geral é verificar a viabilidade preditiva dos dados obtidos unicamente da PNP. A estratégia delineada é a avaliação de modelos de aprendizagem de máquina treinados e testados com dados diversificados, principal característica de um ambiente multicampi.

Os classsificadores foram providos com a metodologia proposta com apoio nos conceitos e etapas do processo de Descoberta de Conhecimento em Bases de Dados (KDD).

Para tanto, foram utilizadas informações de matrículas do ensino técnico de nível médio como forma de proporcionar os conjuntos de dados necessários. Sobre estes se aplicaram os seguintes passos específicos visando o objetivo geral:

- Verificar a composição de atributos por edição quanto a equivalência e suficiência para definição de padrões;
- Identificar a consistência da atualização sequencial de matrículas no decorrer das edições;
- Definir estudos de casos que contemplem composições diversificadas e seus reflexos nos resultados;
- Avaliar o desempenho dos modelos por campi e de forma geral identificando métricas condizentes com a característica original conjunto em termos de distribuição por classe;
- Aplicar a validação da performance preditiva dos classificadores treinados quando diante de dados novos.

1.3 Visão geral

Este trabalho se fundamenta na aplicação de técnicas computacionais atreladas à verficação de padrões em dados. A abordagem construída consiste compreender acontecimentos e estimar comportamentos futuros, neste caso, relacionados ao ensino, especificamente, a evasão escolar.

Nessa perspectiva, foram extraídos microdados da PNP com referência aos anos de 2018 a 2022. A composição dos experimentos possui recorte anual delimitado ao IFPI, especificamente à EPTNM, cenário justificado pela expansão e desalinhamento de indicadores.

A base de dados resultante da delimitação foi segmentada em subconjuntos representativos da descentralidade da instituição, que por sua vez foi verificada como uma base única. Sobre estas composições foram construídos modelos preditivos de AM seguidos da avaliação e validação dos classificadores.

Cada modelo foi aplicado a todos os campi da instituição, gerando assim resultados segmentados e geral, quando considerado o conjunto único. A base completa testada visou proporcionar as comparações da visão ampla com a análise descentralizada. A Figura 3 apresenta uma visão geral desta explanação.

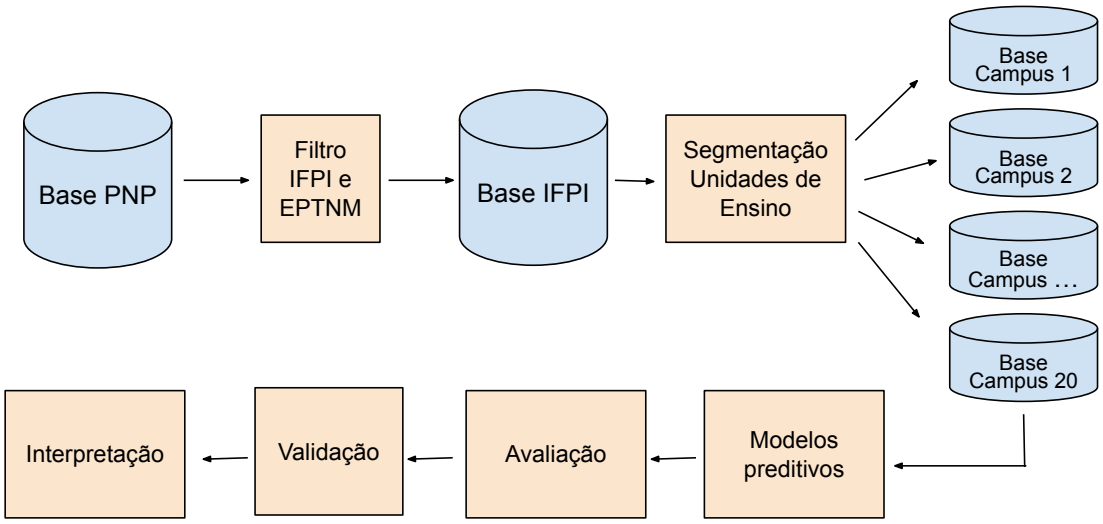

Figura 3 – Visão geral da proposta para obtenção de resultados segmentados por campi e geral sobre a base institucional.

1.4 Estrutura do trabalho

A apresentação deste trabalho segue uma estrutura organizada em 5 capítulos. O texto introdutório trouxe uma contextualização geral sobre a evasão escolar, definições relacionadas, possibilidades da tecnologia da informação frente ao problema bem como a delimitação, motivação e objetivos da proposta. O Capítulo 2 descreve os conceitos teóricos inerentes aos recursos computacionais utilizados e a literatura relacionada. A origem e o processo de preparação dos dados, a metodologia proposta e a estratégia de avaliação aplicada constam detalhados no capítulo 3. Já no capítulo 4, os resultados obtidos são apresentados seguidos das respectivas discussões. As conclusões e trabalhos futuros estão dispostas no capítulo 5.

2 Referencial Teórico

O conteúdo deste capítulo descreve os conceitos essenciais relacionados ao embasamento teórico sobre os procedimentos técnicos utilizados neste trabalho. Discorre-se aqui sobre o processo de Descoberta de Conhecimento em Bases de Dados, Mineração de Dados no contexto educacional, Aprendizagem de Máquina e as métricas de avaliação aplicadas aos modelos classificadores.

2.1 Descoberta de Conhecimento em Bases de Dados

O processo de informatização tem se ampliado constantemente nos mais diversos segmentos de atividades. Além das vantagens já inseridas como hábitos contemporâneos, os sistemas de informação proporcionam uma estruturação de dados que viabilizam processos adicionais sobre os mesmos. Em ambientes educacionais a transformação não foi diferente, ainda que sejam comuns ações estratégicas concebidas unicamente sob as experiências profissionais, reforçando o apoio da tecnologia com suporte indispensável (BAKER, 2014).

Dessa forma, o armazenamento de dados em constante expansão proporciona emergir áreas de estudos específicas a partir da aplicação métodos computacionais. Nessa perspectiva, este trabalho aplica o processo originalmente descrito como *Knowledge Discovery in Databases* (KDD) ou Descoberta de Conhecimento em Banco de Dados. O KDD pode ser definido como um fluxo iterativo e interdisciplinar que envolve a extração de informações úteis previamente desconhecidas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Em Baker (2023) constata-se a ênfase e a prevalência contemporânea do clássico método de tratamento de dados e a pertinência da aplicação das etapas descritas como seleção, pré-processamento, transformação, mineração de dados, avaliação e interpretação dos resultados. Para estes temos as seguintes definições:

- Seleção: escolha dos dados relevantes para a análise, coletando informações de diversas fontes;
- Pré-processamento: limpeza dos dados para remover ruídos, lidar com valores ausentes e preparar os dados para análise;
- Transformação: conversão em um formato apropriado para a análise, pode incluir normalização, agregação e criação de novos atributos derivados;

- Mineração de Dados: aplicação de técnicas para descobrir padrões, relações e informações ocultas, envolve algoritmos de aprendizado de máquina, análise estatística e outras abordagens;
- Avaliação: Verificação dos resultados para determinar sua relevância e qualidade. Usam-se métricas de desempenho, validação cruzada e comparação com conhecimento prévio.

Em complemento aos passos mencionados temos ainda a visualização e interpretação dos padrões descobertos e a transformação destes em conhecimento compreensível. A visualização dos resultados pode ajudar a comunicar *insights* de forma clara, o objetivo é a utilização do conhecimento, ou seja, a sua aplicação para tomar decisões, fazer previsões ou melhorar processos (ARAUJO, 2014).

2.2 Mineração de Dados Educacionais

Quando enfatizadas questões relacionadas à educação, a Mineração de Dados na composição no processo KDD se especifica com relevância na literatura. Consiste em uma estratégia de investigação sobre aspectos acadêmicos múltiplos, desde a melhoria da qualidade do ensino até a personalização da aprendizagem.

A Mineração de Dados Educacionais (MDE) apresenta potencial diversificado quanto às possibilidades de aplicação. Conforme em Romero, Romero e Ventura (2014), neste conceito há a combinação de algoritmos computacionais e métodos estatísticos. Para os autores, as etapas visualizadas na Figura 4 apresentam de forma geral o processo como forma de contribuir com o ensino.

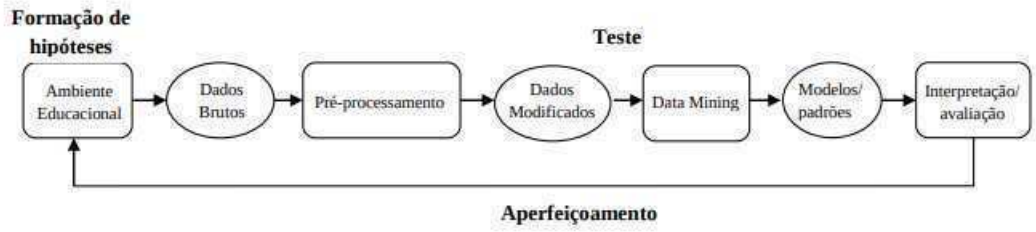

Figura 4 – Exemplo de Mineração de Dados no contexto educacional. Adaptado de (ROMERO; ROMERO; VENTURA, 2014).

Nesse modelo de mineração o objetivo não é apenas transformar dados em conhecimento, mas também aplicá-lo para a tomada de decisão. Pode-se contribuir com o ambiente educacional de modo a melhorar o ensino e potencializar a aprendizagem dos alunos.

Ratificando a consolidação da MDE, a adaptação com base em Aggarwal et al. (2015) destaca que a sequência originalmente disposta poderia ser ajustada ao fluxo demonstrado na Figura 5. Os autores enfatizam a essencialidade da etapa de processamento analítico para obtenção de resultados que favoreçam a interpretação e o *feedback*, quando necessários.

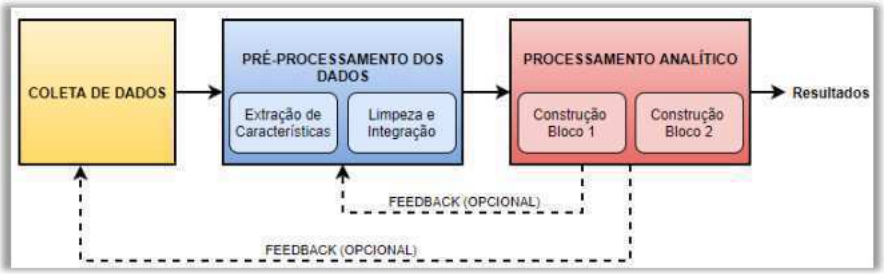

Figura 5 – Fluxos da Mineração de Dados Educacionais. Adaptado de (AGGARWAL et al., 2015)

Embora apoiada em conceitos da mineração de dados de aplicabilidade geral, as circunstâncias das ocorrências em educação demandam adaptações de acordo com o contexto. Baker, Isotani e Carvalho (2011) exemplificam que direcionada à personalização da aprendizagem, a MDE foca em resultados parciais. Para a observação dos comportamentos finais, o foco se aplica às questões da trajetória do estudante. Nesse sentido de exposição do autor, há oportunidades descritas como mineração de correlação de mineração de causas, visando as contribuições.

Em Filho e Araújo (2017b) emergem ainda outras possibilidades contemporâneas em dados educionais. Em complemento à MDE tradicional, crescem os estudos sobre *Learning Analytics* (LA). O intuito desse conceito orientado a dados é ser mais direcionado à aprendizagem. Tem-se ainda a *Academic Analytics* (AA) que também se fundamenta em dados de sistemas de ensino, porém agregando outras informações de natureza administrativa.

Constata-se portanto que a MDE tem contribuído ao se inserir como ferramenta de destaque nas indagações sobre fatores dissociativos da aprendizagem, como a evasão. Além disso, o recurso permite viabilizar análises que subsidiam interpretações otimizadas (RAMOS et al., 2020).

2.3 Aprendizagem de máquina

Como forma de proporcionar verificações não limitadas a ocorrências pretéritas, estudos sobre conjunto de dados são frequentemente complementados pela aplicação da intencionalidade preditiva. Extensamente apreciada como suporte às intervenções prévias, *Machine Learning* ou Aprendizagem de Máquina (AM) é um conceito amplamente referenciado a partir da definição em Russel e Norvig (2013). Na definição dos autores assim

se caracterizam os recursos computacionais com capacidade de assimilar comportamentos de forma automática considerando as observações disponíveis.

O aprendizado nesse contexto é efetivado por meio de de algoritmos que buscam encontrar padrões a partir da realização de experimentos com base em fatos ocorridos (MACHADO, 2011). Como contribuição direta desta habilidade tecnológica, tem-se no âmbito das pesquisas um importante recurso de apoio aos especialistas, tendo em vista a não limitação a análises condicionadas às suas próprias conclusões (FACELI et al., 2011)

Especificamente quanto as teorias que embasam as instruções das rotinas de AM, a inferência estatística fundamenta os principais métodos sendo eles definidos como supervisionados e não supervisionados. Na abordagem supervisionada, o objetivo é que seja estimada uma função com base em exemplos previamente rotulados e que seja capaz induzir um atributo alvo em novos dados. Em estudos sobre informações não categorizadas, tem-se a possibilidade de segmentação do conjunto total em grupos conforme possíveis similaridades entre eles, conceito esse relacionado ao paradigma não supervisionado (GOLDSCHMIDT; PASSOS; BEZERRA, 2015).

Dentro do potencial que o AM tem em estudos que abordam as vantagens das constatações prévias, o uso de classificadores têm revelado importantes contribuições pertinentes ao enquadramento deste trabalho, os dados educacionais.

Em uma representação recente da aplicação de AM no contexto educacional, Viana, Santana e Rabêlo (2022) expõem o fluxo de etapas que resultam no treinamento de modelos a partir de bases rotuladas, assim como a aplicação deste aprendizado visando inferir categorias com as quais que se relacionam dados novos. Conforme apresentado na figura 6, a representação dos autores facilita a compreensão.

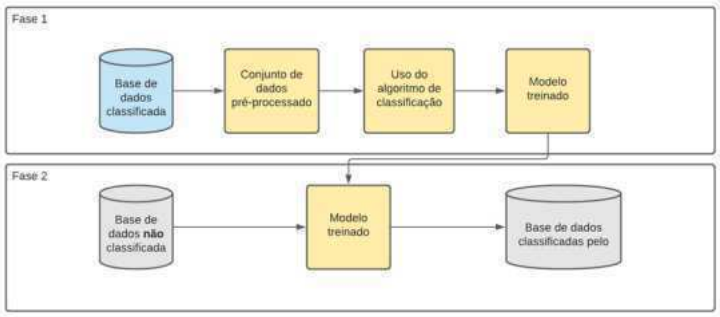

Figura 6 – Fluxo de etapas da aplicação de Aprendizagem de Máquina. Adaptado de (VIANA; SANTANA; RABêLO, 2022)

2.4 Trabalhos relacionados

Constatando que a evasão escolar é um problema recorrente nos diferentes níveis e modalidades da educação, é comum encontrar trabalhos acadêmicos que se apropriam dessa

investigação. Notadamente se aplicam tecnologias que apoiam-se em conceitos estatístico como recursos.

Tendo em vista a diversificação na composição de cenários de estudo na literatura relacionada, faz-se pertinente contemplar neste capítulo uma análise de publicações com o intuito de perceber as estratégias e aprimorar a contribuição pretendida.

Dutra, Souza e Fernandes (2022), analisaram ocorrências da evasão no Instituto Federal da Paraíba (IFPB) utilizando a MDE para construir modelos preditivos com AM. Especificamente considerando os cursos técnicos subsequentes, os autores buscaram validar hipóteses como a distância de deslocamento diário dos alunos entre os fatores, porém, o estudo não se comprova essa colocação enquanto que se sobressai a quantidade de períodos letivos cursados. O modelo testado com o algoritmo SVM gerou resultados expressivamente acurados, superando os 93% considerano a mencionada métrica.

Características demográficas e socioeconômicas tiveram suas influências pesquisadas em Souza e Cazella (2022a) ao elaborarem uma proposta preditiva a partir do desempenho em disciplinas básicas do ensino médio. No trabalho desenvolido no Instituto Federal do Rio Grande do Sul (IFRS), as pesquisadoras utilizaram os algoritmos *Decision Tree* e *Random Forest* e obtiveram 90% e 80% de acurácia, respectivamente.

Ratificando o despertar por ações preventivas, Oliveira, Medeiros e Andrade (2022) propuseram verificar atributos relevantes correlacionados com a evasão no Instituto Federal da Paraíba (IFPB) comparando com o cenário nacional. O método proposto pelos autores usou dados de cursos técnicos e superiores sobre os quais extraíram métricas para avaliar os classificadores. Foi considerada a média alcançada nos testes e assim destacaram o algoritmo *Random Forest* (89%) que superou os modelos com *Multilayer Perceptron* (81%), Árvore de Decisão (80%) e Naive Bayes (67%).

Nesse mesmo sentido, Bitencourt e Ferrero (2019) realizaram estudo também sobre cursos técnicos com o intuito de identificar estudantes com potencial de evasão. Construíram modelos preditivo baseado em árvore de decisão com precisão acima dos 86%. Desenvolvido no Instituto Federal de Santa Catarina (IFSC), o estudo faz uso de amplas possibilidades de parametrização dos algoritmos testados.

Por representar um problema que tem sido constantemente alertado em diferentes níveis de ensino, sistemas de detecção antecipada se mostram pertinentes, inclusive nos anos finais dos ensinos fundamental e médio. Essa foi a questão levantada em Filho e Silveira (2021) quando aplicaram a classificação no referido contexto. O recorte bimestral aplicado apresentou como resultado a acurácia de 94%.

Uma outra abordagem de análise temporal que se mostrou eficaz consta em Viana, Santana e Rabêlo (2022) ao definirem a semestralidade no ensino superior como escopo do método proposto. Os autores consideraram as possibilidades de fatores dinâmicos

ao longo do percurso. Coloca-se que a análise por semestre pode compor modelos mais realistas a cada momento dos cursos. Os estudos de casos utilizaram dados dos cursos de Computação da Universidade Federal do Piauí (UFPI), destacando a importância da seleção de atributos relevantes para otimização dos resultados. As predições registraram 95% de acurácia extraída com validação cruzada.

Osorio e Santacoloma (2023) também apresentam um estudo sobre modelos preditivos para alertar riscos de evasão. Além dos resultados significativos, uma constatação recorrente em dados educacionais foi evidenciada, a quantidade de instâncias por classes. Diferente da abordagem em Viana, Santana e Rabêlo (2022), o autores descreveram ganhos ao buscar equilibrar o número de registros que embasaram a instrução dos algoritmos. A proposta modelou um problema de classificação avaliado a partir da média geométrica entre precisão e recall.

Como experiências internacionais na tomada de decisão educacional baseada em dados, Rodríguez et al. (2023) propuseram no sistema de ensino do Chile uma metodologia para planejar, desenvolver e avaliar modelos de AM para predizer a evasão escolar. A proposta visou a análise da trajetória individual recente adicionando outros possíveis fatores determinados por nível escolar, familiar e extrafamiliar. Os autores defenderam a eficácia de classificadores baseados em árvores de decisão justificando serem mais adequados às variáveis categóricas.

Sorensen (2019) avaliou classificadores baseados em estruturas de árvores na composição do fluxo de regras e o algoritmo *Support Vector Machine* (SVM). Foram considerados atributos a partir de indicadores comportamentais e desempenho além de características socioeconômicas e demogŕaficas. A performance média por validação cruzada registrou 90% de classificações corretas, com destaque ao algoritmo *Random Forest*.

Depreende-se das leituras que os trabalhos analisados convergem plenamente quanto à importância das intervenções prévias. Porém, tanto em publicações recentes quanto em uma extensão temporal mais ampla, observa-se uma tendência a estudos localizados, mesmo quando procedentes de instituições de ensino com diversificada atuação regional.

Assim, esta análise de proporção institucional, considera a descentralização multicampi como característica essencial sobre a origem das informações. A proposta utiliza algoritmos recorrentes na literatura relacionada, no entanto, submetendo-os a composições variadas para fundamentar a avaliação.

As métricas são referenciadas sobre o desbalanceamento original na proporção por classes, ou seja, sem uso de recursos equiparativos, como por reamostragem. Por abordar uma estrutura de ensino de ocorrência nacional, tem-se na possibilidade de reprodução da proposta como uma das contribuições deste trabalho.

A Tabela 1 apresenta um resumo de características dos trabalhos analisados.

Tabela 1 – Características gerais dos trabalhos relacionados.

Artigo	Algoritmos	Métricas	Multicampi	Reamostragem
Dutra, Souza e Fernandes (2022)	SVM, KNN	Acurácia, Precisão, Recall	Não	Sim
Souza e Cazella (2022a)	*Decision Tree* e *Random Forest*	Acurácia e Precisão	Não	Sim
Oliveira, Medeiros e Andrade (2022)	Random Forest , MLP, *Decision Tree* e *Naive Bayes*	Acurácia, Precisão, ROC e F1-score	Não	Sim
Filho e Silveira (2021)	*Decision Jungle, Logistic Regression, Bayes Point Machine e Decision Forest*	Acurácia, Precisão, Recall e F1 Score	Não	Sim
Viana, Santana e Rabêlo (2022)	*Random Forest, Decision Tree, Extra Trees, MLP, SVM, KNN e Gaussian Naive Bayes*	Acurácia, Precisão, Recall, F1 score, Índice Kappa e curva ROC	Não	Não
Osorio e Santacoloma (2023)	*Logistic Regression*	G-mean (Média Geométrica)	Não	Sim
Rodríguez et al. (2023)	eXtreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), Categorical Boosting (CatBoost)	GM score e F1 score	Não	Sim
Sorensen (2019)	*Decision Tree, Random Forest e SVM*	Acurácia, Precisão e Recall	Não	Sim
Este trabalho	Decision Tree, Random Forest, Gradient Boost, Multilayer Perception, Support Vector Machine	Recall, Precisão, *F1-score*	Sim	Não

3 Materiais e Métodos

A apresentação deste capítulo é destinada a detalhar todo o ambiente das experimentações realizadas. A preparação das bases de dados deste estudo contemplou as etapas mencionadas do processo KDD, com destaque à Mineração de Dados onde modelos foram treinados, testados e avaliados.

3.1 Origem e descrição dos dados

Os dados utilizados neste trabalho resultam da consolidação anual realizada pela Rede de Coleta, Validação e Disseminação das Estatísticas da Rede Federal (REVALIDE) e disponibilizada na Plataforma Nilo Peçanha (PNP). Esse projeto de tratamento de informações a nível nacional é gerenciado pela Secretaria de Educação Profissional e Tecnológica (Setec), responsável por formular, planejar, coordenar, implementar, monitorar e avaliar políticas públicas da Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT).

O lançamento inicial de registros que alimentam esse sistema de dados educacionais é um passo atribuído a cada instituição da rede, por meio do Sistema Nacional de Informações da Educação Profissional e Tecnológica (SISTEC). Além do painel acadêmico, a base PNP é complementada por indicadores extraídos do Sistema Integrado de Administração Financeira (SIAFI) e indicadores de pessoal oriundos do Sistema de Administração de Recursos Humanos (SIAPE). Após a unificação das bases, um processo de qualificação e validação por diferentes agentes da RFEPCT é realizado seguindo uma série de regras de consistência para garantir a confiabilidade. O processo descrito está representado na Figura 7 logo abaixo.

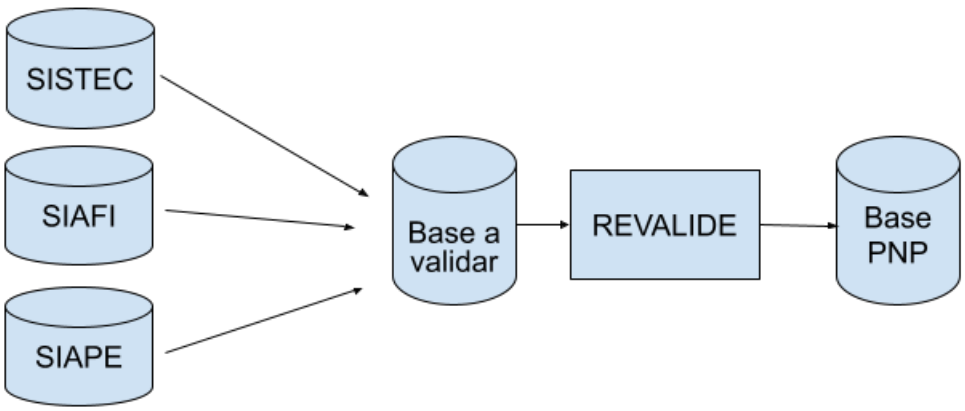

Figura 7 – Composição da base de dados da Plataforma Nilo Peçanha

Conforme apresentado em Moraes (2020), o projeto PNP é ancorado em estabelecidos princípios estatísticos. Este instrumento de apoio à gestão veio a suprir a falta de um mapeamento único e, como efeito, direcionar ações estratégicas de avaliação institucional.

Alguns aspectos de motivação dessa estrutura colaborativa podem ser destacados, como a ausência de um censo específico da EPT e as limitações do censo da educação básica. Este último, por exemplo, considera a última quarta-feira de Maio como data de referência, portanto, não inclui turmas com início no segundo semestre.

A iniciativa PNP proporciou a consolidação de um conjunto de dados condizente com as particularidades da educação profissional. Como importante contribuição, temos o avanço em avaliação de políticas a partir desta ferramenta de monitoramento (PRADO; BRITO; NUNES, 2022).

Em definições estatíticas, a caracterização descritiva de sua concepção ganha possibilidades outras quando inserida a vertente das tecnologias de reconhecimentos de padrões. O recurso inferencial agrega as oportunidades preditivas, como a finalidade deste trabalho.

O Guia de Referência Metodológica (GRM) é o instrumento que direciona todas as disseminações postas na plataforma. A padronização de conceitos e as fórmulas de cálculos permitem um diagnóstico tecnicamente respaldado, visando diminuir as decisões discricionárias (MORAES, 2020).

Dessa forma, a referida base constitui-se de informações avaliativas a nível das unidades de ensino, instituições e rede. Na versão atual constam 19 indicadores acadêmicos que sinterizam algumas variáveis. Destacamos a taxa de evasão anual, na Tabela 2, para exemplificar como os demais estão apresentados no guia de referência.

Tabela 2 – Formato da informação dos indicadores da PNP

Informação	Descrição
Taxa de evasão anual (Tea)	Percentual de matrículas que perderam o vínculo com a instituição no ano de referência sem a conclusão do curso em relação ao total de matrículas.
Modelo Matemático	$$Tea = \frac{Ev}{M} \cdot 100 \quad (3.1)$$
Evadidos(Ev)	Alunos que perderam vínculo com a instituição antes da conclusão do curso.
Matrículas(M)	Soma de todos os alunos que estiveram com matrícula ativa em pelo menos um dia no ano de referência
Polaridade	Quanto menor, melhor.
Agregação	Mìnima: curso Máxima: Rede Federal

Conforme pode ser observado no detalhamento acima, a formação da taxa de evasão anual prescinde do conhecimento das dimensões quantitativas referentes às matrículas ativas, assim como àquelas com evasão identificada. Dessa maneira, buscou-se a nível de microdados o acesso a cada instância desta composição com vistas à estruturação do banco de dados para os experimentos.

A PNP utiliza o conceito de ciclos como intervalo de referência para a definição final da situação de vínculo. Estes períodos correspondem à oferta de um curso com carga horária definida, com data de início e previsão de término.

Na edição de 2018, eram três as nomenclaturas possíveis para as situações de matrículas a cada atualização, conforme a seguinte descrição:

- Evadido: aluno que perdeu o vínculo com a instituição antes da conclusão do curso;
- Retido: aluno que permaneceu matriculado por período superior ao tempo previsto para a integralização do curso;
- Concluinte: formado ou integralizado em fase escolar, sendo:
  - Formado: aluno que concluiu com êxito todos os componentes curriculares do curso, fazendo jus à certificação;
  - Integralizado em fase escolar: concluiu a carga horária das unidades curriculares de um curso, mas ainda não pode receber a certificação por não ter concluído componentes curriculares como Estágio, Extensão obrigatória, TCC, ENADE, etc.

A partir da edição 2019, as possibilidades de registros passaram a ser organizadas em categorias, sendo: em curso, concluinte e evadidos. Os três grupos formados incluem suas respectivas situações, conforme pode ser observado na Tabela 3.

Tabela 3 – Categorias vigentes para situações de matrículas

Categoria	Situação
Em curso	Em fluxo
	Retido
Concluínte	Integralizada
	Concluída
Evadido	Cancelada
	Abandono
	Desligada
	Reprovada
	Transferência Externa
	Transferência Interna

Conforme visto, a definição vigente para considerar o aluno como evadido abrange circunstâncias distintas. A descrição de cada situação atualmente relacionada à evasão consta apresentada a seguir, na Tabela 4.

Tabela 4 – Descrição das situações de matrículas relacionadas à evasão.

Situação	Descrição
Cancelada	Aluno que solicita formalmente o cancelamento da sua matrícula antes de iniciar as atividades pedagógicas do curso.
Abandono	Mais de 25% de faltas não justificadas. Recomenda-se se modificar o status para "Abandono" somente quando não houver mais possibilidade de o aluno voltar a frequentar as aulas.
Desligada	Aluno que solicita formalmente o cancelamento da sua matrícula após iniciar as atividades pedagógicas do curso
Reprovada	O aluno finalizou o curso, porém não logrou êxito nas avaliações. Aplica-se nos casos de impossibilidade de continuação do curso.
Transferência externa	Aplica-se ao aluno que será transferido para outra Unidade de Ensino.
Transferência interna	Mudança de curso, dentro da mesma Unidade de Ensino.

3.2 Etapas do processo

As definições atribuídas aos registros discentes foram apresentadas conforme critérios regulatórios da origem dos dados deste trabalho. Nas etapas a seguir são descritos os estágios utilizados com base no processo KDD. A intenção foi viabilizar os testes experimentais e avaliá-los enquanto subsídio para contemplar os objetivos.

3.2.1 Coleta de dados

O projeto REVALIDE se respalda em dados validados pelas próprias instituições da rede para monitoramento de indicadores a nível de nacional, portanto, inclui em sua base todas as unidades da RFEPT. Embora partindo de um delineamento previamente estabe-

lecido para o cenário experimental, na fase de extração se constatou a não possibilidade de filtros. Por este motivo o conjunto completo de informações foi extraído.

Foram consideradas as bases referentes ao período de 2018 a 2022. O primeiro como ano inicial do monitoramento e o segundo representa a última edição disponível no ato deste processo. Destaca-se que as edições PNP são disponibilizadas sempre tendo como base o ano letivo anterior. Dessa maneira, o ano da publicação, ano de referência, consiste de dados do ano anterior ou ano base.

O acesso aos microdados de cada uma das bases foi realizado por meio do portal de dados abertos do Ministério da Educação. Nesse ambiente foi possível obter os arquivos com a estrutura *comma-separated-values* (CSV) contendo uma linha para cada registro de matrícula.

Nesse sentido, visando o escopo traçado para as experimentações, foram mantidos apenas os registros da instituição selecionada como estudo de caso. O recorte abrange cursos técnicos ofertados na modalidade presencial. Os respectivos filtros foram aplicados nos atributos: unidade de ensino (IFPI), tipo curso (técnico) e modalidade (presencial).

A delimitação configurada e as quantidades de instâncias a cada passo podem ser visualizadas na Figura 8. Os valores exibidos na ilustração desta etapa consideram todas as situações de matrículas.

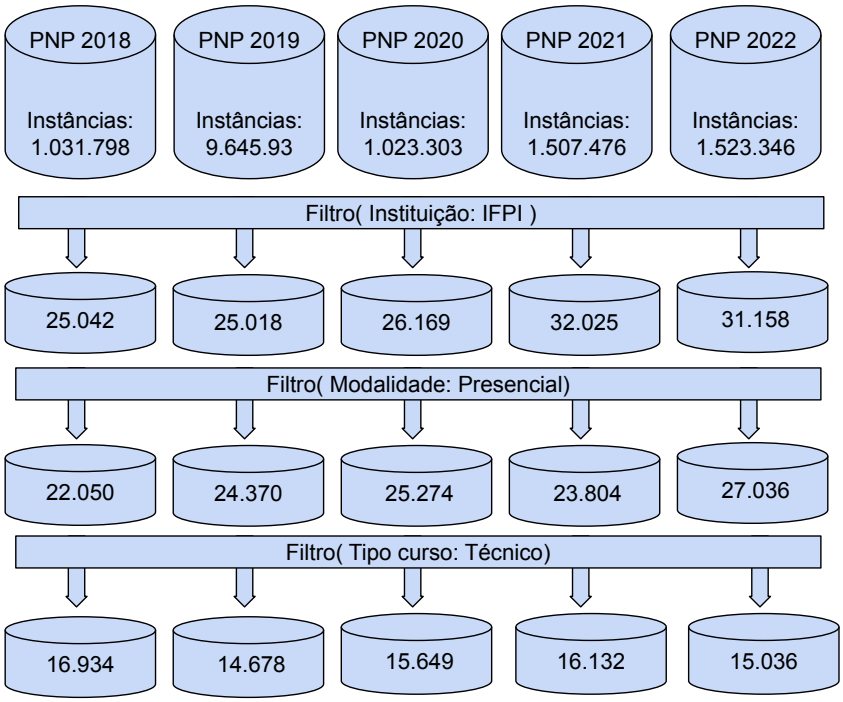

Figura 8 – Representação da coleta e delimitação de dados aplicada a cada edição publicada pela PNP até o momentento deste estudo.

3.2.2 Pré-processamento

A estratégia explicitada anteriormente considera somente o aspecto quantitativo. Iniciou-se então a etapa para tratar possíveis inconsistências a nível de valores dos atributos. Visando a qualidade dos dados, o pré-processamento pode ser caracterizado principalmente pelas fases de limpeza e transformação bem como a formação de atributos derivados, quando necessários (BITENCOURT; FERRERO, 2019).

A partir de arquivos com a separação de valores por vírgula (CSV), esta etapa se desenvolveu tecnicamente com a codificação em linguagem *Python* versão 3.8. A ferramenta de recursos adicionais a este propósito denominada *Pandas*[1] foi utilizada na versão 2.0.3. A análise inicial de caráter exploratório construída nesse ambiente proporcionou a identificação de ações necessárias, considerando o preenchimento original de cada atributo.

Uma importante verificação foi a não ocorrência de valores ausentes. Conforme destacado em HARRISON (2020), essa análise é essencial visto que a lógica dos algoritmos poderá resultar em falhas ao tentar processar campos vazios, geralmente identificados como *NaN*, um acrônimo para o termo *Not a Number*.

Em todas as bases se constatou a completude das informações, corroborando com a importância validação a qual os dados são submetidos antes da publicação. A integralidade no entanto não descarta a análise dos valores na preparação pretendida para os experimentos.

Devido a identificação originalmente designada aos atributos não seguir um padrão de nomenclatura, esta ação foi necessária para facilitar a compreensão e unificação ao longo do processo. Optou-se pela descrição em letras minúsculas com espaços substituídos pelo caracter *underscore*, em referência às recomendações de boas práticas da linguagem de programação em uso. Dessa forma, atributos como *"Mês De Ocorrencia"* na descrição original, passou a ser denominado *"mes_ocorrência"*. A identificação após aplicada esta convenção está especificada na Tabela 5.

Para além da visão exploratória que direcionou os passos iniciais do pré-processamento, demais conceitos da estatística descritivas ajudaram a verificar a pertinência da aplicação de métodos complementares no tratamento de dados. Nesse aspecto, medidas de resumo e dipersão foram usadas para avaliar atributos numéricos. A distribuição de frequências com base nas contagens de instâncias por grupos foram aplicadas às variáveis categóricas.

Uma análise especial cabe ao atributo definido como "situacao_matricula". Conforme detalhamentos do guia metodológico da PNP, publicado em Moraes (2020), o citado campo apresenta a informação base para a composição das categorias oficialmente definidas no documento. Nesse sentido, foi contruído o agrupamento de referência para o trabalho especificado na Tabela 6

[1] https://pandas.pydata.org/

Tabela 5 – Aplicação de nomenclatura padrão aos atributos.

Original	**Padrão**	**Tipo de dado**
Carga Horaria	carga_horaria	numérico
Carga Horaria Minima	carga_horaria_minima	numérico
Co Ciclo Matricula	cod_ciclo_matricula	numérico
Cor Raca	cor_raca	categórico
Dt Data Fim Previsto	dt_fim_previsto	data
Dt Data Inicio	dt_inicio_previsto	data
Dt Ocorrencia Matricula	dt_ocorrencia	data
Eixo Tecnologico	eixo_tecnologico	categórico
Fator Esforco Curso	fator_esforco_curso	numérico
Fonte de financiamento	fonte_financiamento	categórico
Mes De Ocorrencia	mes_ocorrencia	texto
Modalidade Ensino	modalidade_ensino	categórico
Nome Curso	nome_curso	texto
Renda Familiar	renda_familiar	categórico
Sg Inst	instituicao	texto
Sg Sexo	sexo	categórico
Situa___O_Matricula	situacao_matricula	categórico
Sub Eixo Tecnologico	sub_eixo	categórico
Tipo Curso	tipo_curso	categórico
Tipo Oferta	tipo_oferta	categórico
Total Inscritos	total_inscritos	numérico
Turno	turno	categórico
Unidade Ensino	unidade_ensino	texto
Vagas Ofertadas	vagas_ofertadas	numérico

Tabela 6 – Agrupamento das situações de matrículas

Situação	Categoria	Ativo	Evadido
Abandono Cancelada Desligada Reprovado Tranf. externa Tranf. interna	Evadido	0	1
Cursando	Em curso	1	Valor desconhecido
Concluída Integralizada	Concluínte	0	0

A categoria *em_curso* é atribuída apenas a alunos com matrícula ativa no ano de referência e, portanto, possue situação final ainda desconhecida. Partindo deste entendimento foram adicionados os atributos binários *ativo* e *evadido* recebendo os valores ”0” ou ”1”, sendo este último para a representação verdadeira.

O objetivo foi viabilizar um cenário de classificação a partir de matrículas rotuladas quanto à evasão. O algoritmo 1 expressa em pseudocódigo o procedimento apresentado.

Algoritmo 1 Definições atributos ativo e evadido

Entrada: categoria

se *categoria* == "*em_curso*" **então**
    ativo = 1
    evadido = *"valor desconhecido"*
senão se *categoria* == "*concluinte*" **então**
    ativo = 0
    evadido = 0
senão
    ativo = 0
    evadido = 1
fim se

O fator de esforço de um determinado curso, apresentado pelo atributo *fe_curso*, é um conceito elaborado pelo grupo de especialistas da PNP. Este indicador ajusta a contagem de matrículas para os cursos que demandam, para o desenvolvimento de suas atividades, uma menor relação de alunos por professor (BRASIL, 2018).

Compreendido no intervalo de 1.0 a 1.30, a distribuição visível na Figura 9 apontou maior ocorrência na faixa inferior a 1.15 para alunos não evadidos, enquanto acima deste ponto predominam os demais registros. Essa constatação originou o atributo fe_nivel contemplando os valores discretos *fe_baixo* e *fe_alto* em subtituição à representação contínua original.

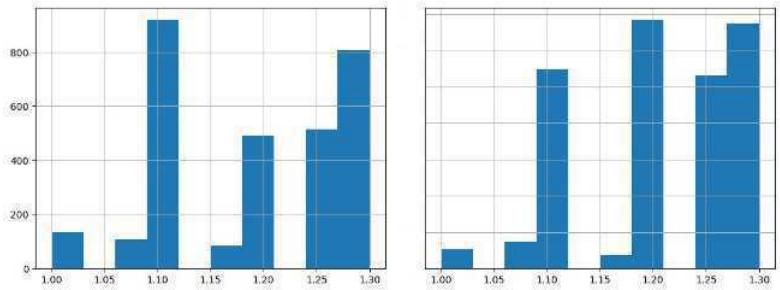

Figura 9 – Distribuição de frequência por fator de esforço do curso entre evadidos (à esquerda) e não evadidos (à direita).

A representação das faixas de valores originais em níveis por ser visualizada a seguir na Figura 10.

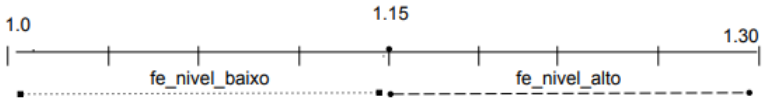

Figura 10 – Definição do fator de esforço em níveis com base em faixa de valores

A Figura 11 expressa a tendência dos níveis baixo e alto. A mesma análise corrobora com a detecção de maior porção do nível alto entre os evadidos.

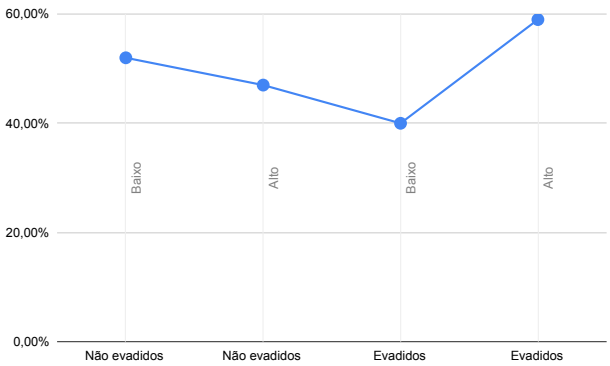

Figura 11 – Tendência de nível entre evadidos e não evadidos.

3.2.3 Normalização e Transformação

O atributo *carga_horaria* se apresenta originalmente no intervalo de 800 a 4270. Estes valores caracterizam os tipos de ofertas dos cursos técnicos que podem ocorrer nas formas integrada, concomitante ou subsequente, definições essas em relação ao ensino médio.

A distribuição original está disposta em escala altamente divergente dos demais atributos. Como consequência, podemos ter forte interferência nos resultados dos algoritmos, principalmente os que consideram conceitos de distâncias. Nesse caso foi aplicada a técnica de normalização de dados através do método *MinMax* (GÉRON, 2019).

A implementação *MinMax* considera que sendo x uma observação qualquer pertencente a uma distribuição com amplitude de valor mínimo(min) e máximo(max) conhecidos, resultará em um valor y normalizado a Equação 3.2. Por exemplo, para o valor 900 a ser normalizado, considerando os valores 800 e 4.270 como mínimo e máximo, respectivamente, temos:

$$y = \frac{(x - min)}{(max - min)} \tag{3.2}$$

$$y = \frac{900 - 800}{4.270 - 800} = \frac{100}{3.470} = 0,02 \tag{3.3}$$

Na preparação dos atributos categóricos se aplicou a representação numérica para os valores. No entanto, a base em análise apresenta composição de natureza qualitativa exclusivamente nominal, ou seja, as categorias não expressam necessariamente uma escala de ordem. Cabe destacar a importância da abordagem adequada em situações semelhantes, em Primão et al. (2022) e Dutra, Souza e Fernandes (2022), por exemplo, coloca-se que a substituição das categorias por números inteiros em sequência pode exprimir uma notação ordinal não realista a informação original.

Nesse sentido, este trabalho transformou a distinção categórica em identificação binária em todos os atributos com essa característica de composição. Com isto, os atributos faixa_etaria, cor_raca, fe_nivel, eixo_tecnologico, renda_familiar, sexo, sub_eixo, tipo_oferta e turno passaram a formar novas colunas.

3.3 Ambiente Experimental

O processamento computacional em torno do problema de classificação apresentado neste trabalho ocorreu em ambiente técnico preparado com os recursos descritos na Tabela 7. A metodologia proposta foi implementada em sua totalidade com a linguagem de programação *Python*. Nessa etapa, a aplicação prática dos conceitos de AM teve como embasamento teórico o método supervisionado no qual os rótulos conhecidos consideram a representação binária do atributo *evadido*.

Tabela 7 – Descrição de recursos do ambiente experimental

Recurso	Descrição	Licença
Google Collaboratory [2]	Ambiente de desenvolvimento online com alocação de recursos computacionais sem custos provido pela empresa Google	*Open-source*
Pandas [3]	Biblioteca que provê funções implementadas para análise e tratamento de conjunto de dados	Open-source
Scikit-learn [4]	Biblioteca com a implementação de vários algoritmos para análise preditiva de dados	*Open-source*
Numpy [5]	Pacote de recursos computacionais para processamento numérico a partir de conjuntos multidimensionais e matrizes.	*Open-source*
Matplotlib [6]	Biblioteca para construção de visualização de dados que implementa diversas possibilidades de obtenção de gráficos	*Open-source*
Seaborn [7]	Recurso baseado na biblioteca *Matplotlib* também utilizado para visualização de dados e fornece um interface de alto nível para representações gráficas.	*Open-source*

Foram então verificadas algumas definições temporais para a determinação dos intervalos a serem estabelecidos como limites na composição dos subconjuntos. Os recortes descritos a seguir, ilustrados na Figura 12, delinearam 4 (quatro) estudos de casos testados e avaliados enquanto subsídios para o provimento de modelos preditivos.

- O primeiro estudo de caso (EC1) considerou a ideia de acompanhamento anual, conforme periodicidade de publicação das edições da PNP que visa identificar a situação do aluno no mencionado intervalo de tempo;
- O segundo estudo de caso (EC2) contemplou dados dos anos 2017, 2018 e 2019 que alimentam as edições 2018, 2019 e 2020. A primeira definição corresponde ao conceito de ano base enquanto a segunda especifica o ano de referência;
- O terceiro estudo de caso (EC3) corresponde a concatenação das edições 2018 a 2022. A inclusão de dados do período de Pandemia de Covid-19 visou verificar implicações na identificação de padrões;
- O estudo de caso (EC4) compartilha a mesma definição de EC1. No entanto, distingui-se quanto ao particionamento por validação cruzada *k-fold*. Visando assim uma melhor perpepção dos modelos para a predição com matrículas ativas, estratégia aplicada como validação da proposta.

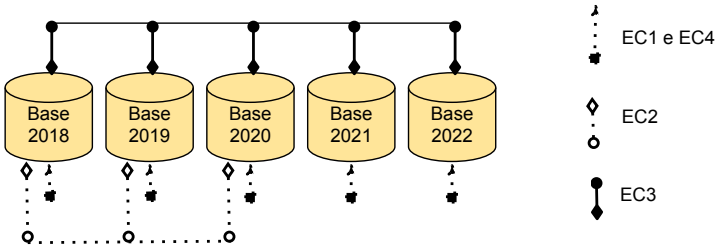

Figura 12 – Representação de quais bases compõem cada estudo de caso.

A composição dos estudos de casos não incluiu registros de alunos com situação de matrícula *"em curso"* por representarem as matrículas ativas. Dessa maneira, o uso de instâncias definidas como não ativas se justifica por ser o grupo possível de associação à classe positiva, com valor 1, ou negativa, com valor 0, sendo esta a rotulação apontada no atributo *evadido*.

Para conhecimento, os números resultantes da distinção realizada entre alunos ativos e não ativos encontram-se apresentados a seguir, na Figura 13.

[3] https://colab.research.google.com
[4] https://pandas.pydata.org/
[5] https://scikit-learn.org
[6] https://numpy.org
[7] https://matplotlib.org/

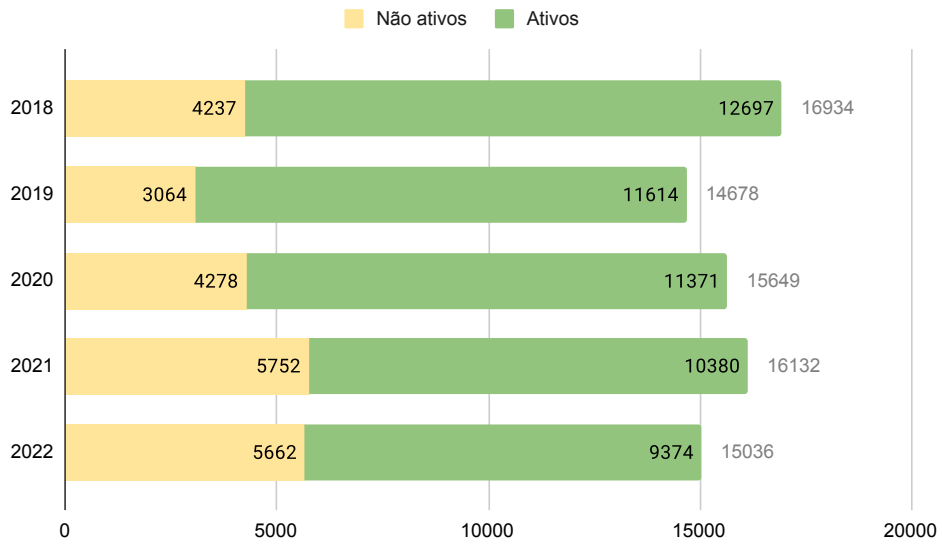

Figura 13 – Composição total incluindo alunos ativos e não ativos.

A Figura 14 logo abaixo apresenta os detalhes quantitativos das bases individuais do estudo de caso 1.

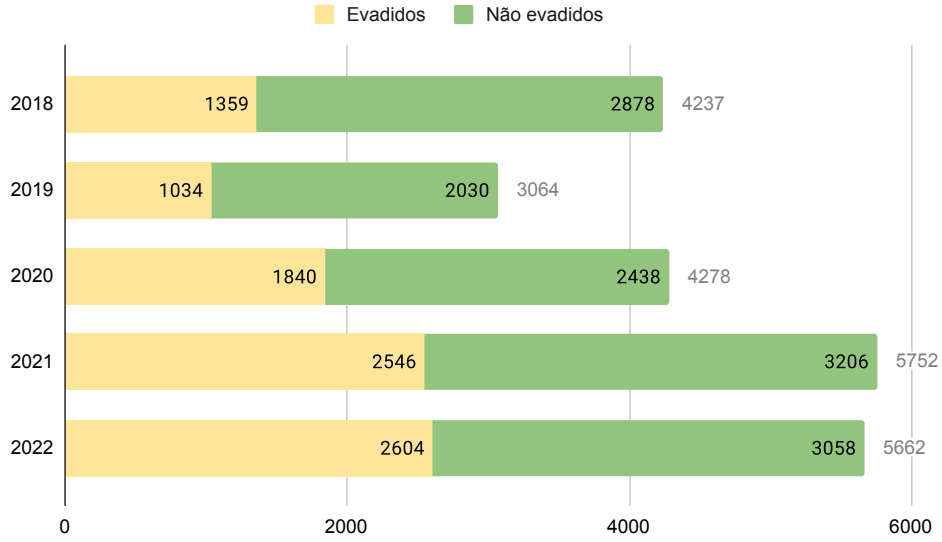

Figura 14 – Composição das bases individuais dos estudos de casos analisados.

Descrevendo os números que compõem EC1, na base 2018 são 4.237 instâncias, com 1.359 registros de evasão e 2.878 como não evadidos. Em 2019, os 1.034 evadidos e 2.030 não evadidos totalizam 3.064 instâncias. Em 2020, registaram-se 4.278 instâncias formadas por 1.840 evadidos e 2.438 não evadidos. No ano de 2021 foram contabilizadas 2.546 evasões e 3.058 não evasões resultando em 5.752 no total. Finalizando o recorte em 2022 com 5.662 instâncias sendo 2.604 evadidos e 3.058 não evadidos.

As estruturações concatenadas para os estudos de caso EC2 e EC3 constam a

seguir, na figura 15. Em detalhes, EC2 compõe-se de 11.398 registros sendo 4.198 evadidos e 7.200 não evadidos. Já em EC3 os 9.051 evadidos e 12.647 não evadidos totalizam os 21.698 registros desta composição.

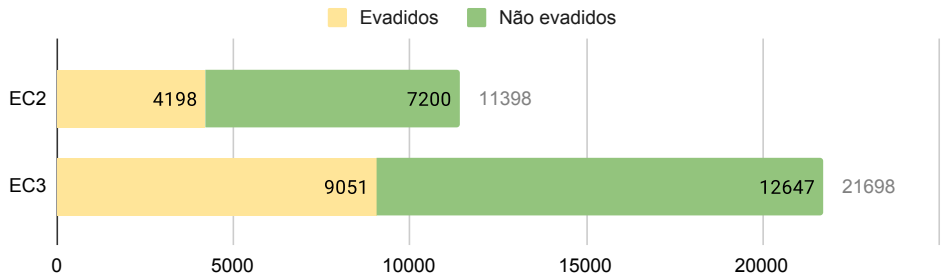

Figura 15 – Composição das bases concatenadas para EC2 e EC3

Constam em EC2 e EC3 os dados das bases selecionadas para a composição devidamente concatenados, conforme definições descritas na seção VII e expressas em notação logo abaixo. Para os casos de ocorrência de uma mesma matrícula nas bases da sequência, prevalece a última situação registrada. Os mencionados conjuntos concatenados podem ser notados como:

$$EC2 = 2018 \cup 2019 \cup 2020$$

$$EC3 = EC2 \cup 2021 \cup 2022$$

Nesses agrupamentos, os quantitativos por classes e total, resultaram em número inferior ao simples somatório da composição. O fato decorre de possíveis inconsistências sobre as quais se optou pela desconsideração de alguns regitros. Casos como a situação definida como formado seguida de vínculo ativo na base seguinte, denotam eventos pontuais de imprecisão encontrados nas informações.

Importante retomar que o ciclo anual de atualização de matrículas, proposto na PNP, consiste em aplicar a todos os registros de vínculo discente do ano base a verificação de suas respectivas situações no ano de referência. Ao monitoramento são adicionandas ainda as novas matrículas efetivadas.

Formalmente, esse procedimento de contrução da base pode ser representado como sendo $B = \{ b_1 , b_2 , b_3 , ... , b_n\}$ o conjunto contendo cada matrícula registrada do ano base, $R = \{ r_1 , r_2 , r_3 , ... , r_n\}$ representando as suas repectivas atualizações no ano de referência e $M = \{ m_1 , m_2 , m_3 , ... , m_n\}$ as novas matrículas.

Dessa maneira, cada edição PNP corresponde a um conjunto resultante das atualizações de registros anteriores incluindo as matrículas recentes. A composição pode ser expressa em notação como:

$$PNP = (B \cap R) \cup M$$

3.4 Metodologia

A partir da definição dos estudos de casos, a divisão em 2 (dois) subgrupos distinguindo matrículas ativas e rotuladas, tem como embasamento as convenções estabelecidas na PNP.

Excetuando-se a situação que esteja definida como *em curso*, correspondente às matriculas ativas, a definição "rotuladas" está relacionada às categorias evadido ou não evadido. Sobre esta representação binária se formulou o problema de classificação a ser explorado. A visão geral da metodologia está apresentada a seguir, na Figura 16

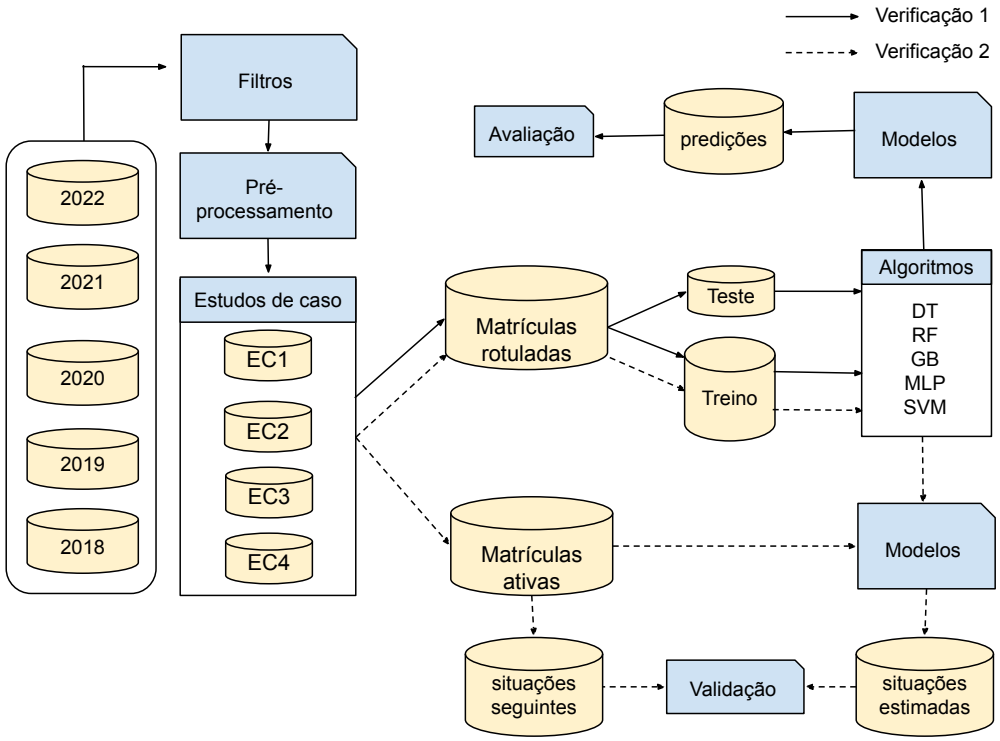

Figura 16 – Visão geral da metodologia

Duas verificações experimentais foram especificadas para o provimento dos modelos de aprendizagem de máquina. Ambas buscaram identificar a abordagem mais pertinente para estimar a situação de matrícula do aluno, seja quanto a evasão ou conclusão. Os percursos verificados podem ser descritos em detalhes da seguinte forma:

- A verificação 1 segmentou o conjunto de matrículas rotuladas entre treinamento e teste, ou seja, somente matrículas não mais ativas foram usadas em EC1, EC2 e EC3. No entanto, esta investigação inicial foi procedida com apenas uma execução dos algoritmos por cada campi. Para verificar possíveis limitações, foram buscados resultados a partir de iterações representadas por medida de resumo, processamento aplicado com validação cruzada em EC4.

- A verificação 2 utilizou todo o conjunto de dados rotulados somente para treinamento e extração de modelos. Para ampliar a percepção de eficiência da metodologia, aplicou-se a validação com os classificadores estimando a próxima situação para as matrículas atualmente ativas. Para tanto, esse cenário demandou o levantamento das situações então "em curso" buscando-as na base imediatamente posterior da PNP.

Na Figura 17 encontra-se especificada a rotina definida como validação. O fluxo apresentado resume as iterações em cada base de dados das edições PNP, de 2018 a 2021. Para estes conjuntos são extraídas as matrículas ativas, com exceção à base 2022. Como limite de dados disponíveis, para a base 2022 a consulta seguinte torna-se impossibilitada. Assim, nesta última base, apenas se pesquisam as matrículas da edição anterior. A partir de cada registro ativo é buscada a respectiva "situação seguinte"[8], verificada nos dados, e a "situação estimada"[9] pelos classificadores. Verifica-se então a coincidência para validar ou registrar o erro da predição.

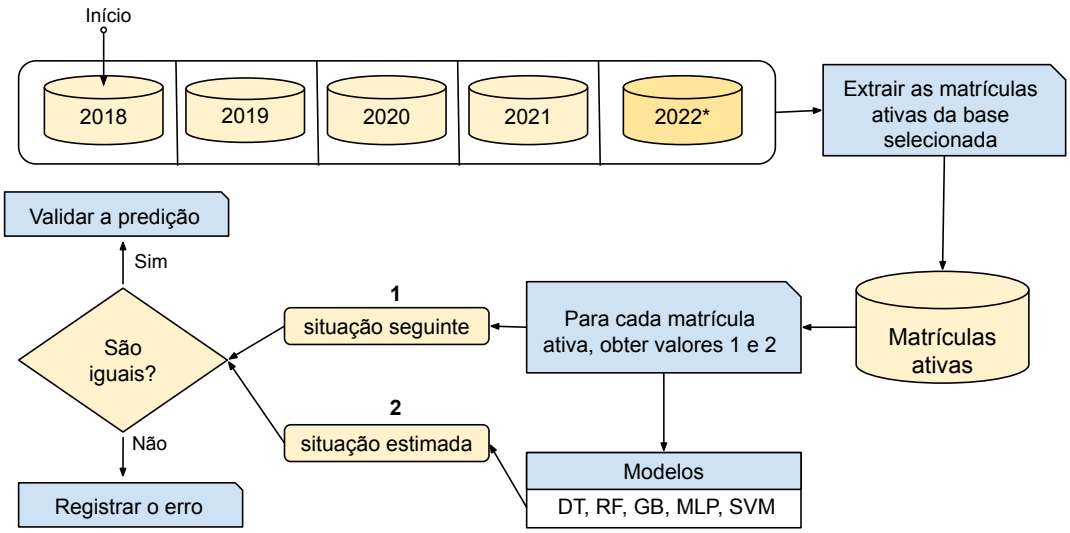

Figura 17 – Representação do procedimento para validação com matrículas ativas.

Os atributos comuns a todas as edições do monitoramento, com exceção daqueles tidos como identificadores, foram utilizados em todos os experimentos visando a análise comparativa dos modelos.

Ressalta-se que o uso do mesmo grupo de atributos , exclusivamente extraídos da PNP, justifica-se pela intencionalidade de proporcionar verificações em torno da viabilidade genérica da proposta. Dessa forma, não incluindo informações institucionalmente específicas, tem-se em vista que esta metologia seja adequadamente reproduzida com dados de outros institutos federais.

[8] Valor real da situação de matrícula no ano seguinte, conforme informações registradas.
[9] Valor predito a partir do aprendizado dos classificadores.

Na Tabela 8 consta a relação com todos os atributos da base original. Na lista são identificados aqueles que foram selecionados de acordo com o mencionado critério de continuidade no decorrer das edições analisadas. Os algoritmos *Decision Tree* (DT), *Random Forest* (RF), *Gradient Boosting Classifier* (GB), *Multi Layer Perceptron* (MLP) e *Suport Vector Machine* (SVM) foram aplicados com suas respectivas configurações da implementação padrão.

Tabela 8 – Atributos das edições PNP utilizadas.

Atributo	**Tipo de dado**	**Selecionado**
carga_horaria	numérico	Sim
carga_horaria_minima	numérico	Não
cod_ciclo_matricula	numérico	Não
cor_raca	categórico	Sim
dt_fim_previsto	data	Não
dt_inicio_previsto	data	Não
dt_ocorrencia	data	Não
eixo_tecnologico	categórico	Sim
fator_esforco_curso	numérico	Sim
fonte_financiamento	categórico	Não
mes_ocorrencia	texto	Não
modalidade_ensino	categórico	Sim
nome_curso	texto	Não
renda_familiar	categórico	Sim
instituicao	texto	Não
sexo	categórico	Sim
situacao_matricula	categórico	Sim
sub_eixo	categórico	Sim
tipo_curso	categórico	Sim
tipo_oferta	categórico	Sim
total_inscritos	numérico	Não
turno	categórico	Sim
unidade_ensino	texto	Não
vagas_ofertadas	numérico	Não

3.5 Métricas de avaliação

O desbalanceamento entre as classes foi fator prepoderante na definição da metodologia de avaliação utilizada. A natural constituição majoritária por alunos não evadidos não foi desconsiderada por representar uma tendência comum em bases de dados educacionais. Essa caracterização se define relevante no contexto em análise, considerando assim a ocorrência da evasão como um comportamento excepcional.

Entretanto, a destacada distribuição numérica divergente entre as classes compromete a equidade para o aprendizado dos modelos. Nas condições apresentadas, os estimadores de rótulos podem ser induzidos à melhor assimilação de padrões da classe sobreposta. Nesse aspecto, aponta-se, portanto, para a pertiência de avaliações mais específicas em detrimento de indicadores de descrição geral, como a acurácia.

Considerando essa característica na composição de todos os estudos de casos, foram exploradas as métricas *recall* e *precision* para análise das predições por classes, ambas compondo em igual nível de importância a métrica *F-score* para avaliar os desempenhos conforme explicações a seguir:

- A sensibilidade expressa por *recall*, neste contexto em relação aos não evadidos (classe negativa), tende a vantagens em decorrência da sobreposição mencionda. Essa métrica, como apresentada na Equação 3.4, denota a importância da baixa ocorrência de falsos positivos (FP). Avaliação essa, embora relevante, não primordial no cenário colocado.

$$Recall = \frac{TN}{TN + FP} \tag{3.4}$$

- A precisão (*precision*) complementa a avaliação por expressar a qualidade das predições. É evidenciada a baixa ocorrência de falsos apontamentos especificamente na classe negativa, como visto na Equação 3.5.

$$Precision = \frac{TN}{TN + FN} \tag{3.5}$$

Partindo dessas considerações, nota-se como essencial aos objetivos deste trabalho equiparar os desempenhos nas habilidades de recuperar o maior número de ocorrência da classe (*recall*) bem como quantos desses apontamentos estão realmente corretos (*precision*).

Especificando os rótulos, conceituamos as classes positiva e negativa representadando, respectivamente, os alunos evadidos e não evadidos. Para aferição, consideramos que:

- A métrica *recall*, no contexto da classe negativa, avalia a capacidade do modelo em apontar a não evasão entre todos desta classe, os verdadeiros negativos (TN).
- A precisão (*precision*) contribui com a avaliação ao considerar os erros ocorridos nestas predições, os falsos negativos (FN).

Assim, desenha-se a avaliação dos classificadores como um prognóstico pela perspectiva da conclusão, ou seja, os que não irão evadir (TN). A estratégia leva em consideração os ganhos na eficiência da contribuição. Diante das composições expostas, aquisição supervisionada potencializa estimadores que melhor assimilam os padrões do prosseguimento. As evasões, nesse aspecto, seriam as divergências a essa regra (FN).

Embora recorrendo a cálculos sobre a classe negativa, o método considera a distribuição original por classes com a sobreposição recorrente de não evadidos. Dispensa-se dessa forma recorrer à equivalência numérica artificial por reamostragem *oversampling*[10] ou mesmo a aleatoriedade de remoção de registros pela técnica de *undersampling*[11].

Nesse sentido, a média harmônica composta por *recall* e *precision* representada pela métrica *f1 score*, Equação 3.6, foi definida para expressar o desempenho dos modelos.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.6}$$

[10] Incremento da classe em menor representação quantitativa.
[11] Decremento da classe majoritária com a eliminação de exemplos por meio de técnicas específicas.

4 Resultados e Discussões

Neste capítulo serão expostos os resultados das experimentações definidas em cada estudo de caso. Os números são seguidos das discussões acerca do que as análises permitem concluir.

A origem dos dados a partir de uma instituição de ensino caracterizada pela descentralização ou ambiente multicampi, direcionou o agrupamento dos dados por campi. Os segmentos resultantes foram nomeados com o padrão CA + [sigla oficial], conforme nomenclaturas apresentadas no Anexo A. A capilaridade geográfica está representada nos subconjuntos submetidos às verificações em consideração às especificidades locais.

O resultados listam a avaliação de cada modelo apresentando a métrica *f1 score* como sumarização dos desempenhos. Os valores pormenorizados por testes, especificando *recall*, *precision*, as quantidades de registros e a composição das predições por classes, constam detalhados nos apêndices.

Em todos os processamentos realizados, das 20 (vinte) unidades da instituiçao, 3 (três) constam categorizadas como campus avançados. Estes não foram incluídos por representarem centros ainda em implantação e, portanto, com poucos dados disponíveis.

Inicialmente é apresentada a abordagem contemplando os estudos de casos descritos como EC1, EC2 e EC3. São avaliados modelos treinados com 80% do conjunto e testados com os demais. Nestes primeiros experimentos apenas uma execução de cada algoritmo foi realizada.

4.1 Resultados EC1

Na tabela 9, logo abaixo, constam condensados os desempenhos de cada algoritmo verificado. Este resumo contém as médias de eficiência em relação a todos os campi, a cada ano, no período de foco deste trabalho, 2018 a 2022.

Tabela 9 – Desempenho dos algoritmos considerando a média de resultados por base anual quando aplicados em todos campi (EC1).

PNP	**DT**	**RF**	**GB**	**MLP**	**SVM**
2018	80,85±0,02	80,85±0,03	81,63±0,02	79,25±0,04	83,02±0,01
2019	75,27±0,03	72,33±0,03	77,23±0,05	76,52±0,02	74,69±0,03
2020	75,67±0,02	76,94±0,01	75,81±0,03	72,25±0,01	75,84±0,04
2021	84,21±0,01	87,78±0,02	85,07±0,04	84,46±0,03	86,97±0,02
2022	86,96±0,01	88,89±0,03	85,44±0,02	89,29±0,04	90,00±0,02

A partir do comportamento das médias com as bases de cada edição PNP, podemos observar que os modelos apresentaram constante variação nos resultados preliminares, conforme apresentado graficamente na Figura 18.

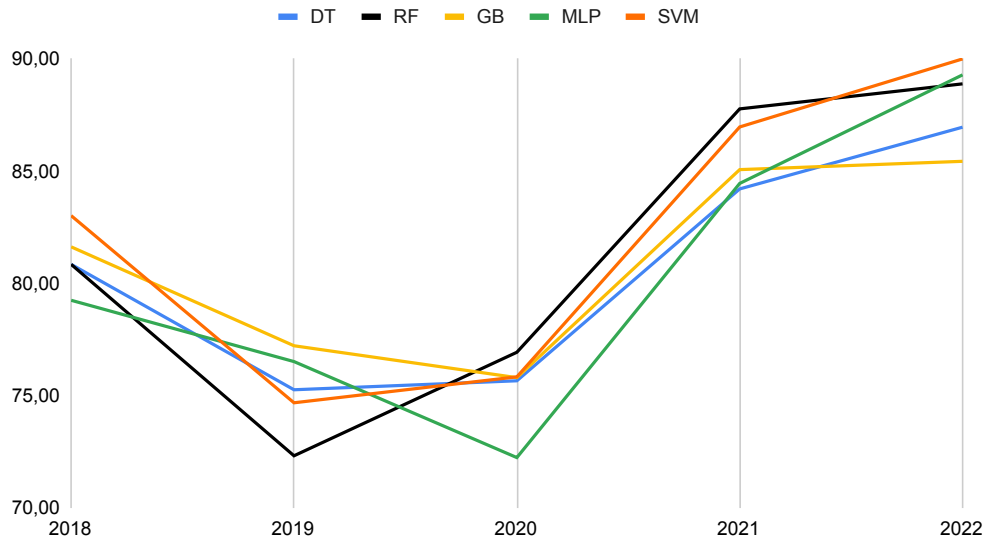

Figura 18 – Evolução das médias de desempenho do estudo de caso EC1.

Para os 3 (três) primeiros anos, a diminuição das médias foi geral. No entanto, os algoritmos SVM, RF e DT apontaram leve evolução já a partir da base 2019, enquanto MLP e GB iniciam essa tendência para 2020 em diante, ponto que demonstrou ínicio de evolução para todos.

Verificou-se nesse momento que os desempenhos poderiam evidenciar uma possível relação com a composição das bases em termos quantitativos. Conforme a variação anterior observada, o aumento ou diminuição no volume de dados impactaria diretamente nas métricas avaliativas.

4.2 Resultados EC2 e EC3

No estudo de caso EC2, houve a ampliação do número de instâncias por meio da concatenação das bases de dados 2018, 2019 e 2020. Foi possível observar que os resultados encontrados ficaram abaixo de 80%, conforme Tabela 10.

Se observamos EC1 e as médias obtidas com as bases individuais que o compõem, os valores resultantes da aplicaçãoo de EC2 ficam abaixo em relação a primeira e apenas exprimem consonância com as 2 (duas) últimas. Logo, sem que sejam constatadas melhorias que possam ser relacionadas ao incremento de dados.

O teste com maior quantidade de instâncias (21.698) foi realizado em EC3 com a junção de dados de 2018 a 2022, como exposto na Figura 15. Os resultados revelaram em termos significativos apenas a mudança de melhor performance do algoritmo GB, com 79,61%. Porém, ainda não superando o topo das médias em EC1 e EC2.

Tabela 10 – Médias, por algorítmos, com bases concatenadas dos estudos de casos EC2 e EC3, considerando execução única por campi.

	DT	**RF**	**GB**	**MLP**	**SVM**
EC2	78,32±0,03	**79,44±0,02**	77,37 ±0,04	73,98±0,03	74,99±0,02
EC3	77,74±0,02	77,35±0,01	**79,61±0,03**	75,83±0,02	77,88±0,03

4.3 Considerações sobre EC1, EC2 e EC3

A variação de eficiências compreendidas entre 73% e 79% em EC2 e 75% e 79% em EC3 atestaram que os conjuntos expandidos, visando testes para além do uso de bases únicas, não otimizaram a evolução presumida com a performances individuais.

Na Figura 20, temos a avaliação sob a perpectiva do melhor resultado alcançado por cada modelo. A comparação destaca uma maior ocorrência deste registro na composição quantitativamente inferior, EC2.

Em termos de desempenhos médios, vistos logo em sequência Figura 19, ratificam-se as performances de DT e RF no conjunto menor. A expansão de dados em EC3 aponta melhor apropriação por GB, MLP e SVM. Ambas análises sinalizam a distinção de padrões na relação com o volume de informações exploradas.

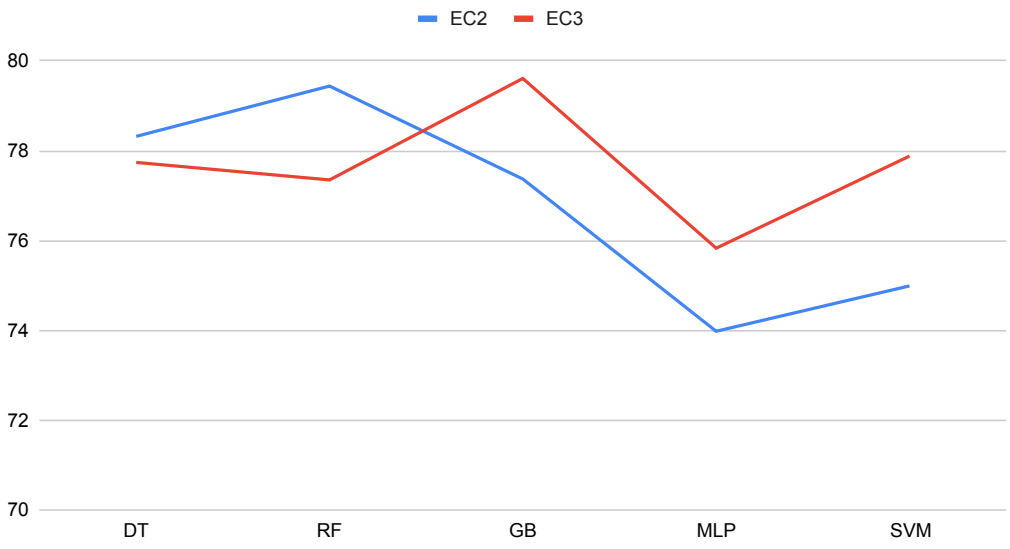

Figura 19 – Médias de desempenho registradas em EC2 e EC3.

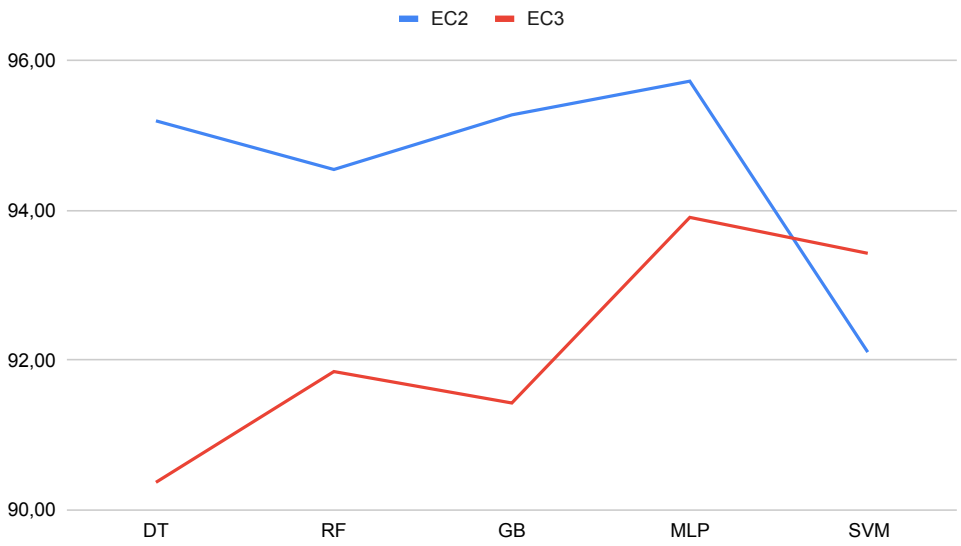

Figura 20 – Melhores desempenhos registrados em EC2 e EC3.

Os estudos de caso EC2 e EC3 foram verificações de hipóteses levantadas a partir de EC1 em relação a quantidade de dados e, conforme visto, o incremento de instância não não implicou necessariamente ganho nos resultados.

Embora tidos como forte embasamento para nortear o avanço no treinamento dos modelos, EC1, EC2 e EC3 se constituem de execuções únicas, ou seja, apenas uma iteração por subconjunto que representam os campi.

Além dos resultados limitados, a partição treino e teste realizada apenas uma vez não se mostrou pertinente aos conjuntos com poucos dados. Assim procedendo se reduziu consideravelmente o conjunto de treinamento. Essa forma de verificação sujeita-se ainda a obtenção de resultados atrelados a eventos de casualidade na seleção de amostras.

4.4 Resultados EC4

Visando contornar a aleatoriedade inerente à segmentação única para treinamento e teste, o estudo de caso EC4 apresenta resultados evidenciando as médias de desempenho, sendo estas complementadas pelos seus repesctivos desvios. Aplica-se este método como forma de melhor expressar as performances preditivas.

Os resultados de EC4 foram obtidos a partir da mesma estratégia delineada em EC1, ano a ano. No entanto, são apresentados valores médios extraídos com o processamento descrito como *cross-validation* ou validação cruzada. Utilizou-se a técnica de segmentação *K-Fold* com 5 (cinco) iterações realizadas em cada um dos subconjunto que expressam as unidades de ensino do ambiente multicampi.

A rotina *K-Fold* traduz-se como o procedimento de subdividir o conjunto em K partes iguais procedendo o treinamento em K-1 partes, com o segmento restante destinado à validação. Repete-se, portanto, o processo K vezes e ao final combinam-se os resultados para obtenção das médias.

Dessa forma foi possível obter desempenhos médios, conforme apresentado a seguir nas Tabelas 11, 12, 13, 14 e 15 com dados de 2018 a 2022, respectivamente. São apresentados os 5 (cinco) melhores resultados por modelo e campus. Para fins comparativos, apresenta-se ainda, como referência, a performance com o uso da mesma técnica quando aplicada à base completa, descrita como IFPI.

Tabela 11 – Resultados do estudo de caso EC4, ano 2018.

	DT	**RF**	**GB**	**MLP**	**SVM**
CACAM	98.27 ± 0.01	98.27 ± 0.01	98.27 ± 0.01	97.93 ± 0.01	96.61 ± 0.01
CAPAR	88.29 ± 0.01	87.83 ± 0.02	87.51 ± 0.01	82.6 ± 0.02	81.15 ± 0.01
CATCE	87.64 ± 0.01	87.41 ± 0.01	85.56 ± 0.01	85.91 ± 0.0	83.24 ± 0.01
CATZS	80.04 ± 0.01	79.87 ± 0.01	77.45 ± 0.02	76.19 ± 0.01	74.18 ± 0.01
CAPIC	85.47 ± 0.02	85.4 ± 0.02	85.33 ± 0.02	84.77 ± 0.02	82.38 ± 0.03
IFPI	84.58 ± 0.01	84.26 ± 0.01	75.17 ± 0.01	78.94 ± 0.01	75.31 ± 0.01

Tabela 12 – Resultados do estudo de caso EC4, ano 2019

	DT	RF	GB	MLP	SVM
CACAM	82.59 ± 0.03	81.06 ± 0.05	80.27 ± 0.03	79.38 ± 0.03	75.95 ± 0.03
CAOEI	91.03 ± 0.01	90.86 ± 0.01	90.7 ± 0.01	90.31 ± 0.01	87.26 ± 0.01
CAPAR	92.03 ± 0.01	91.92 ± 0.01	90.3 ± 0.01	88.01 ± 0.01	84.45 ± 0.01
CACOR	89.76 ± 0.02	89.67 ± 0.02	87.28 ± 0.02	86.46 ± 0.02	82.29 ± 0.02
CAFLO	95.36 ± 0.0	95.29 ± 0.0	93.77 ± 0.0	91.34 ± 0.01	89.1 ± 0.01
IFPI	85.6 ± 0.0	85.07 ± 0.0	74.73 ± 0.01	79.73 ± 0.01	76.91 ± 0.0

Tabela 13 – Resultados do estudo de caso EC4, ano 2020.

	DT	RF	GB	MLP	SVM
CAURU	90.14 ± 0.02	90.0 ± 0.02	90.03 ± 0.02	89.67 ± 0.02	87.85 ± 0.02
CATCE	89.2 ± 0.01	88.84 ± 0.01	84.39 ± 0.01	85.21 ± 0.0	79.82 ± 0.02
CACOR	91.99 ± 0.01	91.65 ± 0.01	90.96 ± 0.02	90.46 ± 0.01	88.23 ± 0.02
CASRN	92.8 ± 0.01	92.62 ± 0.01	90.08 ± 0.01	91.5 ± 0.02	83.88 ± 0.01
CAFLO	89.65 ± 0.01	89.2 ± 0.01	86.35 ± 0.02	79.24 ± 0.01	71.28 ± 0.03
IFPI	86.37 ± 0.0	86.0 ± 0.01	77.78 ± 0.01	81.62 ± 0.01	77.82 ± 0.01

Tabela 14 – Resultados do estudo de caso EC4, ano 2021.

	DT	RF	GB	MLP	SVM
CAPIC	99.79 ± 0.0	99.79 ± 0.0	99.79 ± 0.0	99.02 ± 0.0	98.06 ± 0.0
CATCE	96.45 ± 0.0	96.41 ± 0.0	91.76 ± 0.0	92.76 ± 0.0	89.38 ± 0.0
CASRN	97.3 ± 0.0	97.25 ± 0.0	97.19 ± 0.0	96.06 ± 0.01	93.68 ± 0.0
CACOC	98.43 ± 0.01	98.42 ± 0.01	98.41 ± 0.01	98.2 ± 0.01	98.31 ± 0.01
CAFLO	93.77 ± 0.04	92.49 ± 0.05	92.49 ± 0.05	67.9 ± 0.1	0.0 ± 0.0
IFPI	92.77 ± 0.0	92.64 ± 0.0	84.27 ± 0.0	88.46 ± 0.0	84.1 ± 0.0

Tabela 15 – Resultados do estudo de caso EC4, ano 2022.

	DT	RF	GB	MLP	SVM
CATCE	96.28 ± 0.0	96.21 ± 0.0	89.32 ± 0.0	93.27 ± 0.0	88.02 ± 0.0
CAANG	88.26 ± 0.01	87.6 ± 0.02	88.02 ± 0.02	87.35 ± 0.02	84.62 ± 0.02
CAPAU	95.0 ± 0.01	94.78 ± 0.01	91.9 ± 0.01	91.49 ± 0.01	84.04 ± 0.01
CASJP	98.17 ± 0.01	98.15 ± 0.01	98.17 ± 0.01	98.17 ± 0.01	96.2 ± 0.01
CAPIC	97.08 ± 0.01	97.04 ± 0.01	97.07 ± 0.01	97.07 ± 0.01	93.71 ± 0.01
IFPI	94.19 ± 0.0	94.11 ± 0.0	86.53 ± 0.0	91.05 ± 0.0	85.82 ± 0.0

Levando em conta que os resultados de EC4 expressam a média como resumo, coloca-se oportuno verificar o quanto esses valores encontram-se dispostos no cenário completo. Na Figura 21 são destacadas as médias por ano expressando o desempenho de cada algoritmo. A análise mostra o destaque com os modelos DT e RF, não obstante à quantidade de dados disponíveis. Foram considerados os testes com todos os campi para obter os valores apresentados.

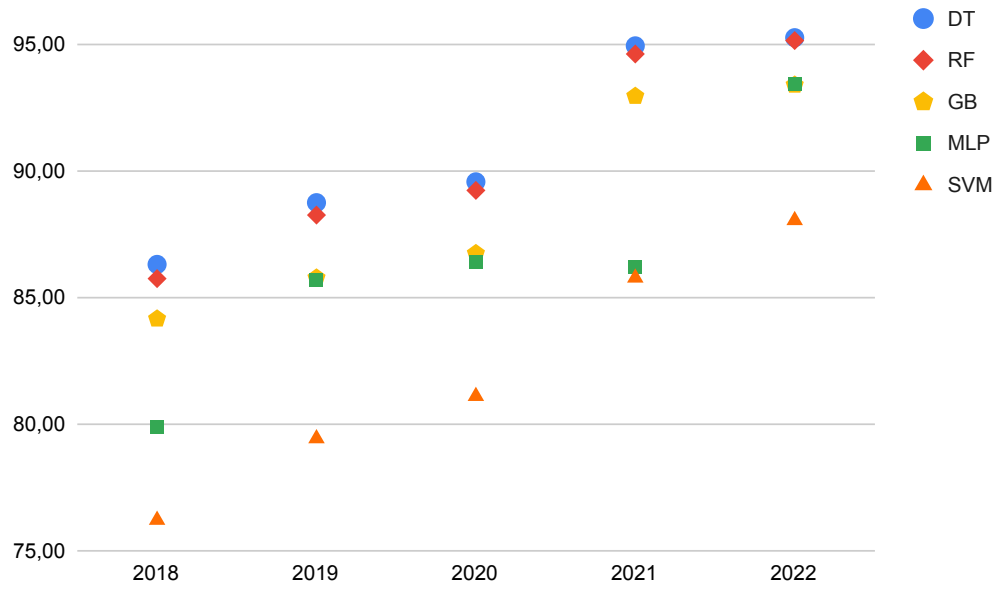

Figura 21 – Variação das médias, por algoritmo, com as bases anuais em EC4.

4.5 Etapa de validação: predição de alunos ativos

Para avaliar a habilidade de generalização dos modelos, implementou-se, como validação, o procedimento para estimar a situação de matrícula no ano seguinte. Foram considerados alvos desta ação apenas alunos na condição de vínculo ativo.

Com este intuito, as matrículas ativas em cada ano tiveram suas respectivas situações verificadas nas edições seguintes. As checagens foram armazenadas em estruturas de dados separadas para então validar a classificação.

Ressalta-se que para as matrículas definidas como ativas da base 2022 não se procedeu a busca por situação posterior. Para tal fim seriam necessários os dados de 2023, não disponíveis até o momento desta operação.

Assim, definindo como T, o conjunto das atualizações. A e B, as bases atual e seguinte, respectivamente. Para A assumindo, a cada passo, referência a 2018, 2019, 2020 e 2021. B assumindo 2019, 2020, 2021 e 2022, na mesma iteração, temos que:

$$T = (A \cap B)$$

No Algoritmo 2 estão apresentadas, em pseudocódigo, as instruções da rotina de validação. A sequência simplificada busca otimizar a compreensão do procedimento.

Algoritmo 2 Validação com matrículas ativas

```
bases ← [2018, 2019, 2020, 2021, 2022]                    ▷ Edições PNP
modelos ← [DT, RF, GB, MLP, SVM]
matriculas_ativas ← [ ]
para indice ← 0 até indice <=3 faça                        ▷ Percorra até 2021
    A ← bases[indice])
    B ← bases[indice+1])
    matriculas_ativas ← Obtenha as matrículas ativas em A
    para cada matricula em matriculas_ativas faça
        situacao_seguinte ← Busque a situação da matricula em B
        para todos os modelos faça
            situacao_estimada ← Proceda a predição para a matricula
            se situacao_seguinte == situacao_estimada então
                Valide a predição
            senão
                Registre o erro de predição
            fim se
        fim para
    fim para
fim para
```

Como exemplo, uma matrícula X, ativa em 2018 e evadida em 2019, terá sua predição validada caso esta última seja a situação estimada pelos modelos. Reitera-se que as predições se limitam a apontar as ocorrências no ano seguinte.

O número de atualizações possíveis e as matrículas não encontradas ou disjuntas na intersecção acima, constam na Figura 22. Matrículas ativas não encontradas são registros não identificados na base seguinte para checagem de suas situações, estas não foram consideradas na validação.

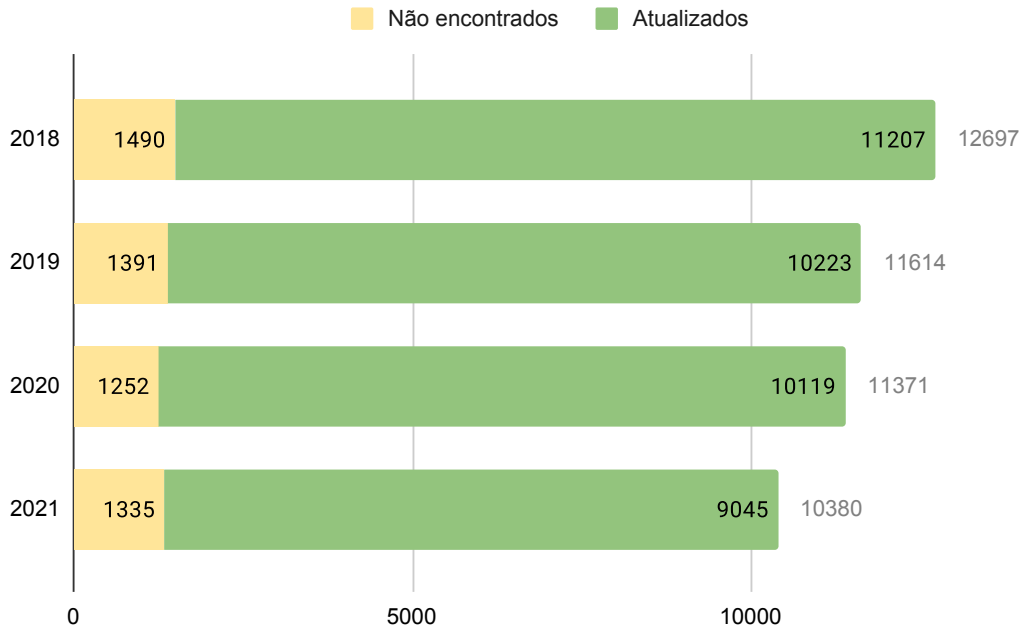

Figura 22 – Quantidade de matrículas ativas atualizadas e registros não identificados no ano seguinte.

Já na Figura 23 está especificada quantidade de evadidos e não evadidos sobre as matrículas ativas atualizadas. Faz-se referência à checagem realizada em consulta aos dados do ano seguinte. Portanto, os númereos representam registros para os quais foi possível atribuir os mencionados rótulos.

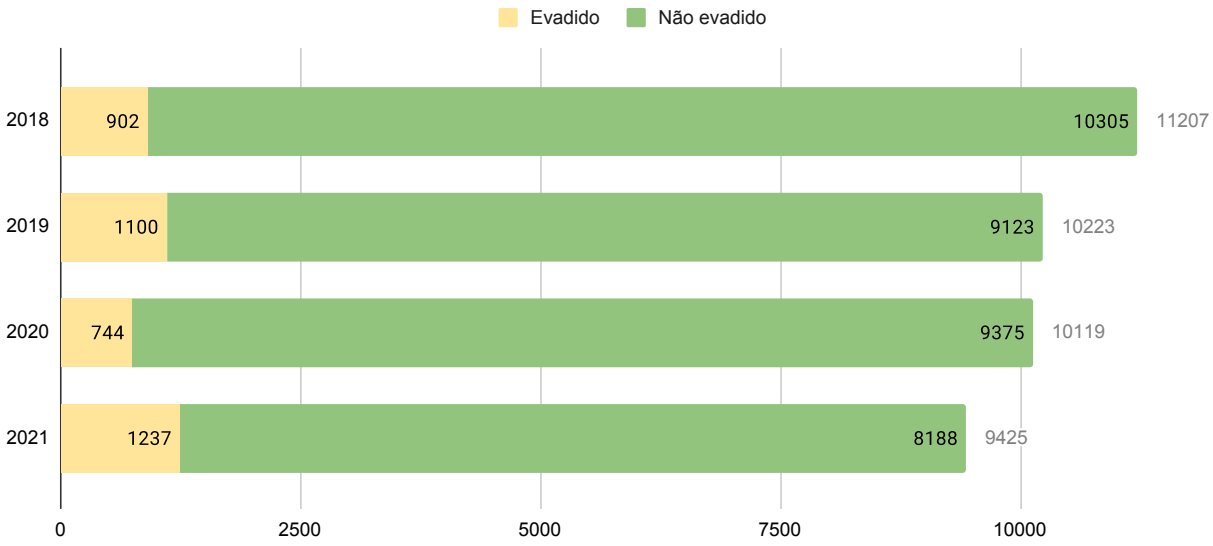

Figura 23 – Quantidade de matrículas atualizadas agrupadas por classe.

O agrumamento acima contabiliza os alunos que permaneceram ativos no ano em sequência como não evadidos. Considerando que este grupo representa a classe negativa, a composição visualizada impulsiona ainda mais a sobreposição já constatada inicialmente. Com esta forma de classificação, simples apontamentos intuitivos para a não evasão evidenciariam altas taxas de acerto, em detrimento da necessidade de estimadores treinados.

Contudo, considerou-se para a validação somente os dois primeiros grupos da Figura 24, os evadidos e não evadidos. Não foram incluídas as matrículas que permaneceram ativas tendo em vista a delimitação da inferência ao ano seguinte.

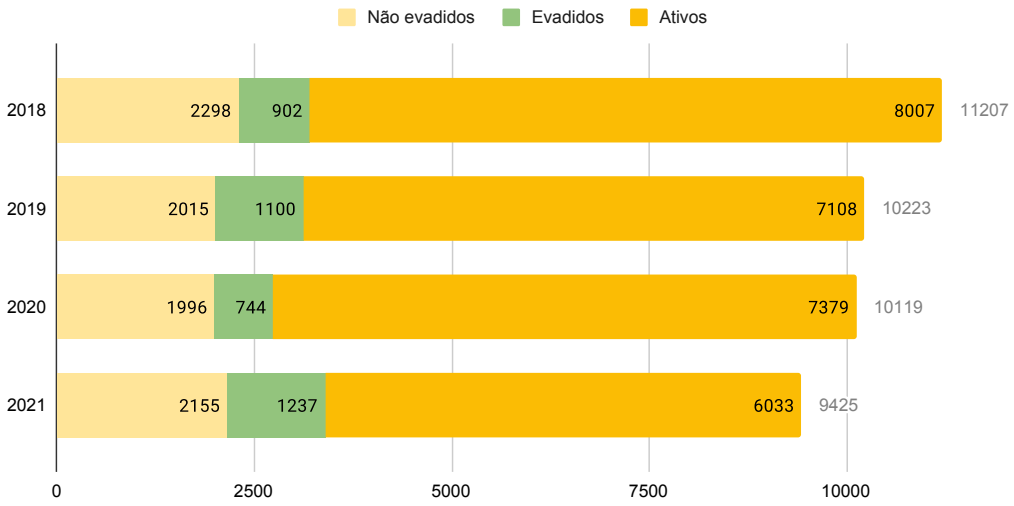

Figura 24 – Quantidade de matrículas atualizadas por situação.

Nos resultados a seguir, para a definição da quantidade de campi a listados como melhores desempenhos, tomou-se como referência o valor da métrica *F-score* na avaliação das predições diante da base completa. São exibidas as unidades de ensino para as quais registram-se performances iguais ou superiores em relação ao conjunto IFPI.

4.5.1 Alunos ativos 2018

Na Tabela 16, são listados os valores encontrados com os algoritmos que apresentaram os 2 (dois) melhores resultados no ano em verificação. A ordem em que os campi estão dispostos são equivalentes aos dois modelos.

Dos 12.697 alunos registrados como ativos na edição PNP 2018, somente 11.207 foram localizados no ano seguinte e verificados quanto a situação seguinte de matrícula. Nesta composição, 2.298 foram identificados como não evadidos. Evadidos somaram 902 e 8.007 permaceram ativos.

Tabela 16 – 2018: Melhores resultados na predição aplicada às matrículas então ativas no ano de 2018.

	DT			RF		
	pre	rec	f1	pre	rec	f1
CAPEDII	96.88	100.0	98.41	96.88	100.0	98.41
CAPAU	94.74	98.63	96.64	94.74	98.63	96.64
CASRN	92.0	98.57	95.17	93.15	97.14	95.1
CAPIC	94.29	94.29	94.29	94.29	94.29	94.29
CAFLO	88.71	98.21	93.22	90.0	96.43	93.1
CACOR	86.29	98.17	91.85	86.59	97.71	91.81
CAVAL	83.78	100.0	91.18	83.78	100.0	91.18
CAURU	83.64	100.0	91.09	83.64	100.0	91.09
CASJP	84.71	97.3	90.57	89.33	90.54	89.93
CATZS	83.23	95.56	88.97	79.89	99.32	88.55
CAOEI	79.89	99.32	88.55	87.05	89.63	88.32
CAANG	82.35	95.45	88.42	84.38	92.05	88.04
IFPI	81.72	92.43	86.75	82.93	90.36	86.48

Logo abaixo, são apresentadas nas Figuras 25 e 26 as capacidades específicas dos modelos com DT e RF. São exibidas as linhas comparativas traçadas pelos valores das métricas *precision* e *recall*, repectivamente. No Apêndice A estão detalhados os resultados especificando as predições por classe.

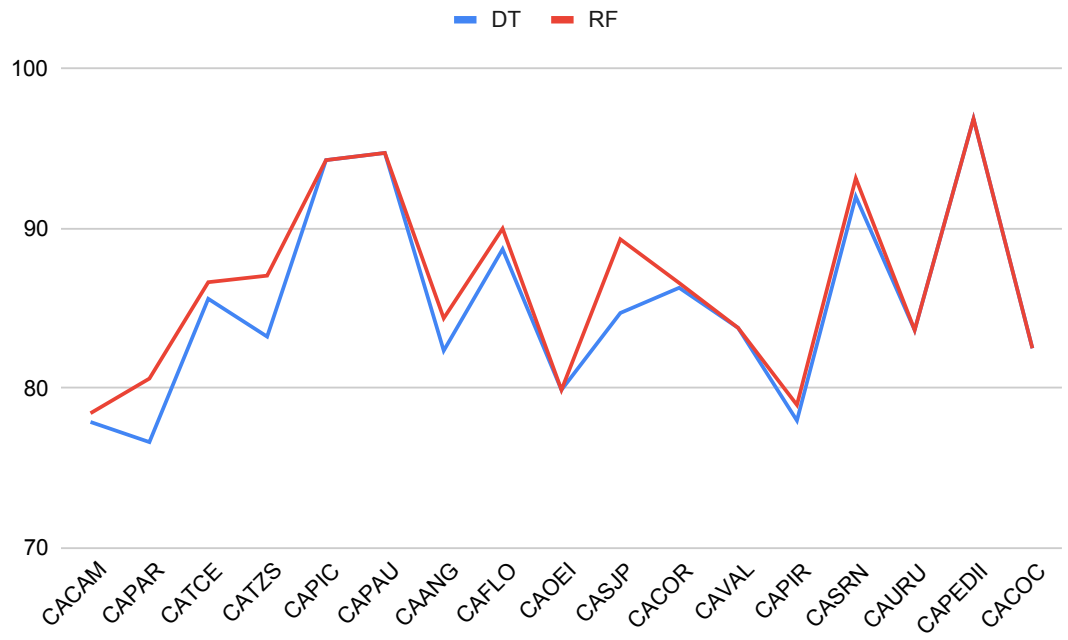

Figura 25 – Comparativo da precisão nas predições geradas por *Decision Tree* e *Random Forest*.

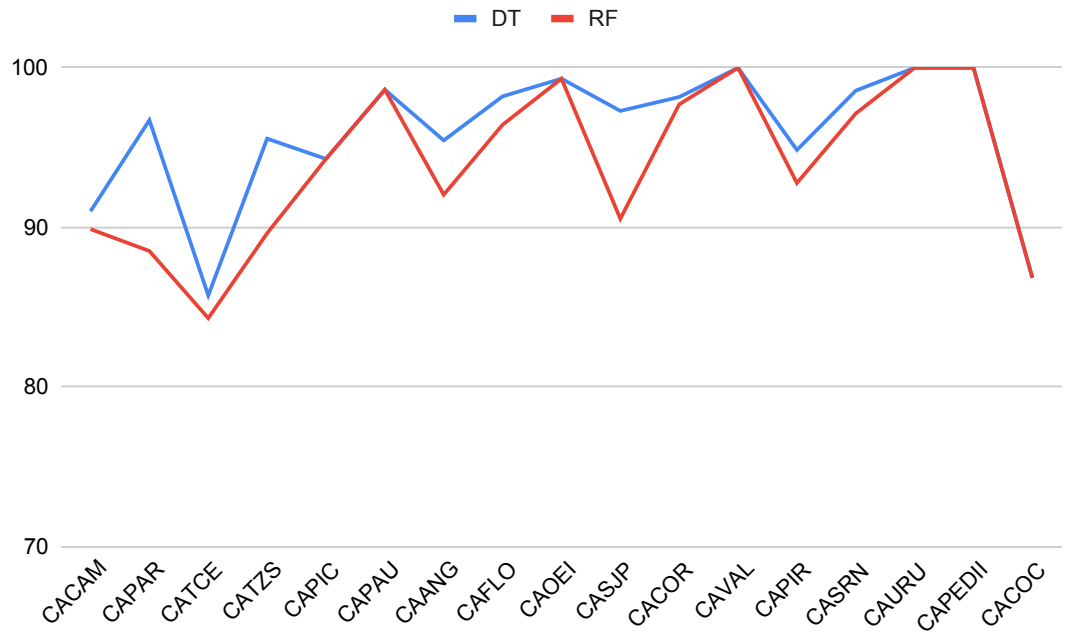

Figura 26 – Comparativo quanto a sensibilidade (*recall*) dos algoritmos *Decision Tree* e *Random Forest*.

Ainda que em destaque quando considerada a métrica de referência *F1 Score*, os algoritmos DT e RF apontaram eficiências específicamente distintas no contexto analisado. Essa constatação reforça a pertinência da consideração equiparada entre sensibilidade e precisão.

A análise das habilidades proporcionada acima demonstrou que RF notabilizou-se por predições mais precisas frente ao DT, contudo, este último se sobrassaiu na comparação feita quanto a sensibilidade (*recall*).

4.5.2 Alunos ativos 2019

Os resultados com matrículas listadas como ativas no ano 2019 constam na Tabela 17. A ordem dos campi são semelhantes, porém não idênticas para os modelos em destaque. Nesta base, em um total de 14.678 registros, 11.614 constavam como ativos e 10.223 foram atualizados. sendo 2.015 como concluíntes e 1.100 evasões. Permanceram com matrícula ativa 7.108 alunos. No Apêndice B as predições estão detalhadas por classes.

Na Figura 27, representa-se gráficamente os valores registrados para as métricas precisão e recall. Em relação ao comportamento observado no teste anterior, a análise comparativa revela não só uma otimização nas performances, mas também a manutenção das habilidades específicas de cada modelo.

Tabela 17 – Melhores resultados das predições com matrículas ativas em 2019.

	DT			RF			
	pre	rec	f1	pre	rec	f1	
CAPAU	97.37	100.0	98.67	97.37	100.0	98.67	CAPAU
CACOC	93.94	100.0	96.88	93.94	100.0	96.88	CACOC
CAPAR	91.76	100.0	95.71	92.77	98.72	95.65	CAPAR
CASJP	92.44	99.1	95.65	92.44	99.1	95.65	CASJP
CASRN	91.26	100.0	95.43	93.81	96.81	95.29	CASRN
CAPEDII	85.58	100.0	92.23	86.82	98.25	92.18	CAANG
CAANG	86.82	98.25	92.18	90.24	94.07	92.12	CAPIC
CAPIC	89.6	94.92	92.18	89.36	94.38	91.8	CAPEDII
CACAM	85.9	98.53	91.78	87.53	96.16	91.64	CATCE
CATCE	86.62	97.53	91.75	88.89	94.12	91.43	CACAM
CAURU	83.04	97.89	89.86	83.64	96.84	89.76	CAURU
CACOR	83.97	94.02	88.71	86.18	90.6	88.33	CACOR
IFPI	80.89	94.85	87.31	81.87	93.04	87.1	IFPI

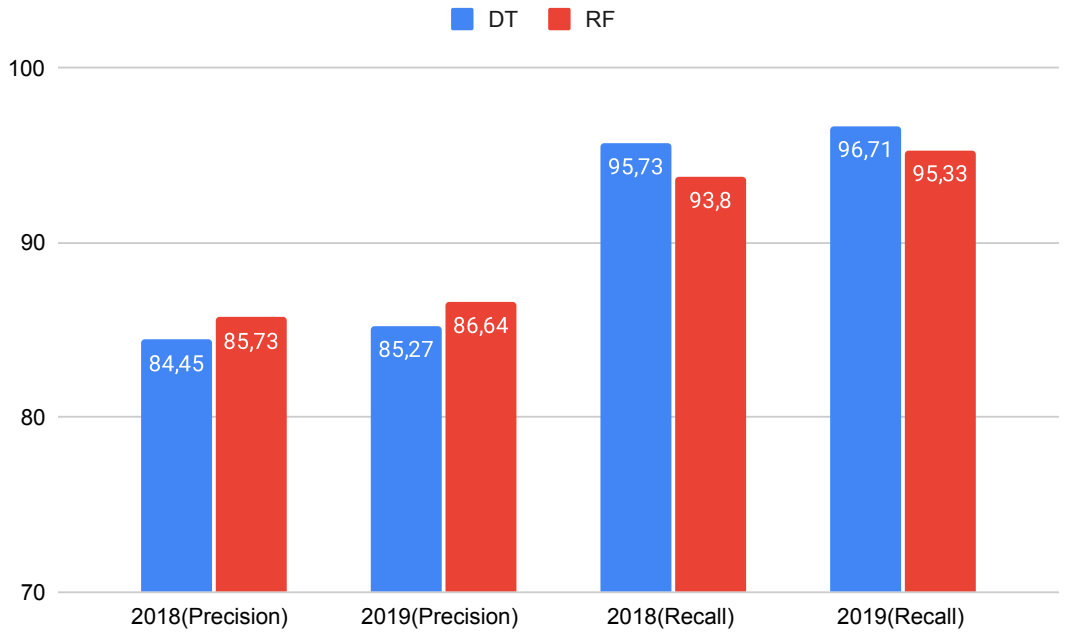

Figura 27 – Comparativo quanto a precisão e sensibilidade dos modelos com os algoritmos DT e RF.

4.5.3 Alunos ativos 2020

Na composição 2020, das 11.371 matrículas então ativas, 10.119 foram atualizadas. Neste grupo, temos identificados 1.996 concluíntes, 744 evasões e 7.379 mantendo-se em curso. Os resultados estão expostos na Tabela 18. No Apêndice C estão em detalhes os valores apresentados.

Tabela 18 – Melhores resultados das predições com matrículas ativas em 2020.

	DT			**RF**			
	pre	rec	f1	Pre	Rec	F1	
CAPIC	100.0	100.0	100.0	100.0	100.0	100.0	CAPIC
CACAM	100.0	100.0	100.0	100.0	100.0	100.0	CACAM
CAPAR	97.28	100.0	98.62	97.28	100.0	98.62	CAPAR
CASRN	94.96	100.0	97.41	94.96	100.0	97.41	CASRN
CACOC	98.15	96.36	97.25	98.15	96.36	97.25	CACOC
CATZS	94.81	99.22	96.97	96.18	97.67	96.92	CATZS
CATCE	93.88	99.49	96.6	94.17	99.15	96.59	CATCE
CASJP	93.1	100.0	96.43	93.1	100.00	96.43	CASJP
CAURU	92.59	100.0	96.15	96.00	96.00	96.00	CAURU
CAPEDII	91.67	99.00	95.19	92.45	98.00	95.15	CAPEDII
CACOR	89.47	100.0	94.44	89.47	100.0	94.44	CACOR
CAOEI	89.47	99.35	94.15	89.94	98.7	94.12	CAOEI
IFPI	88.6	98.1	93.11	89.67	96.59	93.01	IFPI

4.5.4 Alunos ativos 2021

Os resultados específicos das predições com matrículas ativas da base 2021 estão detalhados na Tabela 19. Os desempenhos sumariamente expostos constam com detalhes no Apêndice D.

Tabela 19 – Melhores resultados das predições com matrículas ativas em 2021

	DT			**RF**			
	pre	rec	f1	Pre	Rec	F1	
CAVAL	99.19	100.0	99.6	99.19	100.0	**99.6**	CAVAL
CAURU	97.47	100.0	**98.72**	97.47	100.0	98.72	CAURU
CASJP	95.74	100.0	97.83	95.74	100.0	97.83	CASJP
CACOR	95.45	100.0	97.67	95.45	100.0	97.67	CACOR
CACAM	95.29	100.0	97.59	96.39	98.77	97.56	CACAM
CACOC	94.37	100.0	97.1	94.37	100.0	97.1	CACOC
CAPIC	94.05	100.0	96.93	95.12	98.73	96.89	CAPIC
CAPAR	94.96	98.51	96.7	94.96	98.51	96.7	CAPAR
CATCE	94.79	98.51	96.62	96.8	96.28	96.54	CATCE
CAOEI	91.82	99.02	95.28	92.59	98.04	95.24	CAOEI
CAFLO	93.71	96.4	95.04	96.3	93.53	94.89	CAFLO
IFPI	91.72	98.05	94.78	93.16	96.27	94.69	IFPI

Os alunos com situação ativa em 2021, totalizaram 10.380. Para apenas 9.045 foi possível buscar a situação na edição seguinte. Foram verificados como não evadidos 2.155 e 1.237 registrados como evadidos. Continuaram ativos 6.033 alunos.

4.6 Considerações sobre a validação com alunos ativos

Validar a classificação de matrículas ativas demonstrou boa performance de forma geral. Destacam-se nessa estratégia também os modelos baseados em árvore, nesse cenário, DT e RF. O bom desempenho com essa estrutura de assimilação de regras sinaliza caracterizações sobre o grupo de variáveis preditoras. É possível estimar que as informações trabalhadas possuem potencial associativo considerável com as repectivas classes.

Observa-se ainda que em todos os testes desta etapa os desempenhos individuais e médios, por campi, apresentaram valores superiores a avaliação com a base IFPI. Esse comportamento foi verificado com regularidade em todo o período analisado, reforçando assim as especificidades da atuação geográfica ampla de uma instituição multicampi. Nota-se a adequação do cenário experimental às averiguações descentralizadas procedidas, conclusões que uma análise geral não contemplaria.

A Figura 28 apresenta as médias por *F1-score* com cada algoritmo utilizando as bases das edições anuais. Os modelos com DT seguido por RF, como destaques, ratificam as observações colocadas em torno das estruturas baseadas no conceito de árvore.

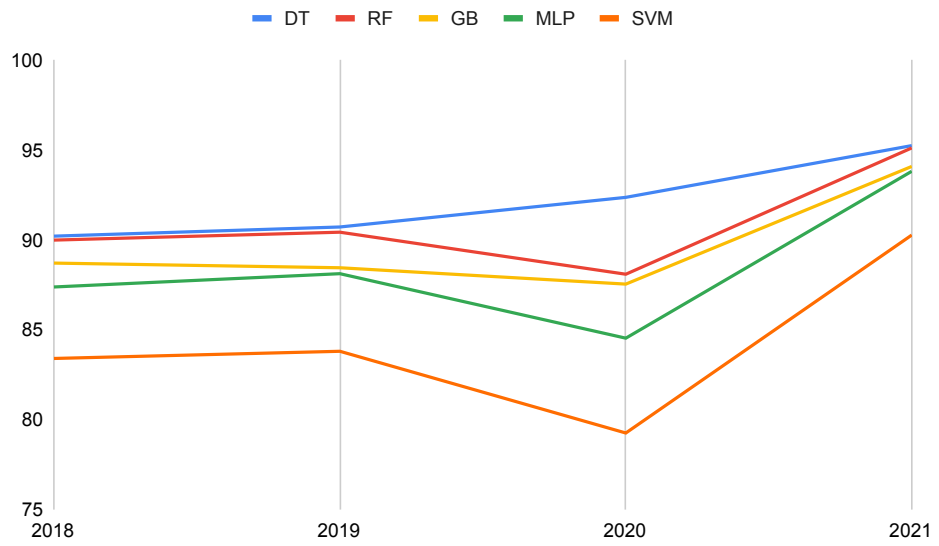

Figura 28 – Comparação das médias gerais de F1-score por modelos a cada ano.

A Figura 29 traz uma percepção geral para análise dos algoritmos a partir do conceito de desvio padrão. A representação denota a característica de dispersão do resultados encontrado com cada modelo. Em conformidade com os destaques já mencionados para DT e RF, visualizam-se estes modelos como os que possuem resultados menos variados.

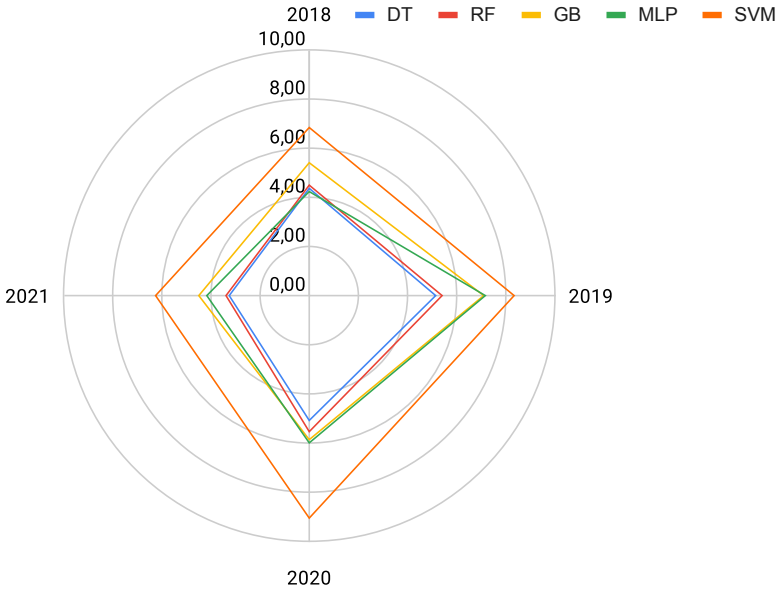

Figura 29 – Comportamento quanto ao desvio padrão sobre as médias registradas.

Ainda como análise, resume-se na Figura 30 esta etapa preditiva com ênfase ao melhor valor de *F1-score* encontrado a cada ano. Na comparação, o mesmo critério gerou a linha de tendência extraída com a base IFPI. O gráfico em destaque ressalta o potencial verificado com os conjuntos menores.

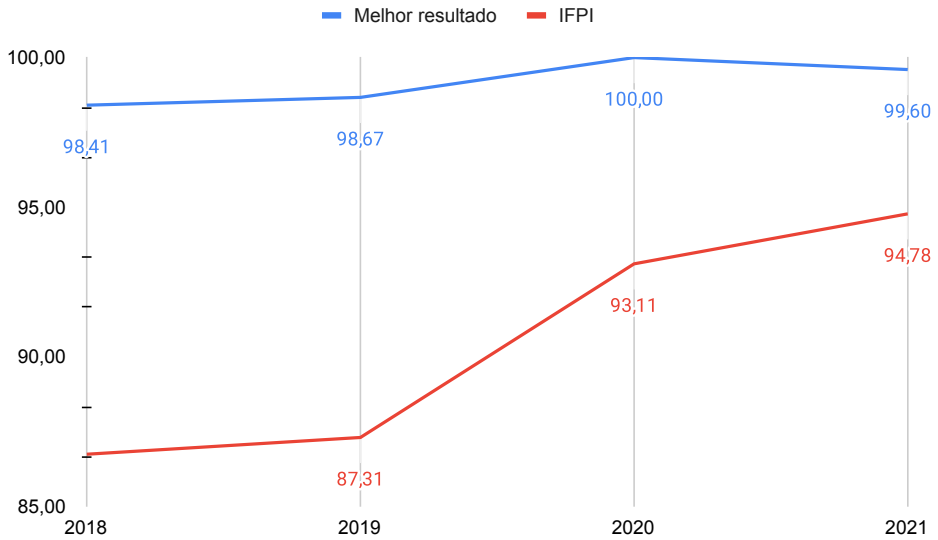

Figura 30 – Comparação entre os melhores valores de f1 considerando todos os campi e a instituição completa.

5 Conclusões e Trabalhos Futuros

O desenvolvimento deste trabalho buscou avaliar modelos preditivos baseados em aprendizagem de máquina, estes sendo providos por dados extraídos unicamente da PNP. O foco em indicadores comuns a a toda a rede de educação profissional visa a pretensão genérica para a reprodutibilidade da metodologia apresentada.

A conveniência da proposta como um recurso de alerta prévio se mostrou promissora e escalável. Ambientes de tecnologia da informação, dentro do contexto educacional analisado, podem facilmente aplicar o método para apoio às estratégias de permanência. Considera-se escalável pois a plataforma de origem dos dados contempla toda a rede.

Os estudos de casos analisados apontaram como a composição das bases de dados refletiu no desempenho dos algoritmos. Partindo de tais análises, as seguintes considerações podem ser destacadas.

Como início, retomamos os passos para contemplar o objetivo geral em torno da avaliação preditiva da PNP. As experimentações iniciais descritas como EC1, EC2 e EC3 subsidiaram constatações frente aos questionamentos específicos, como a composição melhor representativa para as predições. Neste caso, a delimitação por edição se mostrou suficiente, em detrimento das concatenações de bases.

Especificamente em EC1, os resultados mostraram que as primeiras versões da PNP pressupõem uma organização de dados ainda em estalização. Assume-se este ponto tendo em vista que nem todas as matrículas ano de referência constavam na atualização sequencial, como proposto no projeto de monitoramento contínuo.

Para além do aspecto quantitativo, foi possível observar uma evolução na qualidade representativa a partir de edição 2020. Houve progressos comprovados nas bases 2021 e 2022 para todos os algoritmos testados. Essa observação favorece a percepção de aperfeiçoamento na consistência das informações.

A evolução constatada aponta uma desvinculação entre a quantidade de instâncias e a qualidade do aprendizado. Sinaliza-se assim para a suficiência de dados por edição, estratégia que aprensentou o melhores resultados.

As 5 (cinco) edições analisadas, representando o período de 2018 a 2022, não apresentaram expansão constante em volume de dados, como esperado. Registra-se na verdade uma variação que demonstrou não possuir relação direta com a linha de tendência dos resultados.

Os experimentos diversificados ajudaram a constatar que os conjuntos expandidos não potencializaram a capacidade de convergência dos modelos. Assim fica evidenciado o

dinamismo inerente aos fatores da evasão escolar no aspecto temporal.

Testes com execuções únicas em EC1, EC2 e EC3 limitaram as conclusões deste recorte. Além de resultados propensos à aleatoriedade, tivemos a redução considerável de dados para treinamento após a segmentação única.

Nesse sentido, a validação cruzada aplicada em EC4 trouxe uma representação mais fidedigna da capacidade preditiva e, portanto, favoreceu a seleção dos modelos. A partir desta etapa, o foco foi validar a habilidade de generalização dos classificadores diante de dados novos, as matrículas ativas.

Os algoritmos *Decision Tree* e *Random Forest* foram postos em destaque na predição de matrículas ativas. O fato evidenciou a adequação das estruturas em árvore para as inferências no contexto em análise. Em termos de sensibilidade e precisão, as habilidades distintas dos dois modelos recomendariam aplicações em casos de relevância prioritária de uma outra. No entanto, o objetivo deste estudo demandou a consideração de ambas as métricas de forma equiparadas.

A avaliação por médias harmônicas superiores a 90% pode ser apontada como relevante. A afirmação considera que a métrica de resumo expressa o equilíbrio entre a sensibilidade e precisão. Há dessa forma um contraponto ao foco em indicadores isolados de ocorrência comum na literatura.

A justificada opção pelos critérios avaliativos já citados permitem selecionar o algoritmo *Decision Tree* entre dos demais analisados. Considera-se a capacidade de resultados mais equilibrados a partir de métricas condizentes com o desbalanceamento descrito nas seções anteriores.

O cenário de ensino multicampi proporcionou ainda evidenciar a composição heterogênea da instituição. Os experimentos são destacados pelas distintas performances por campi, aspecto esse de relevante consideração para ações de permanência. Ainda nessa importante contatação, podemos concluir que as predições por unidades de ensino foram mais eficientes quando comparadas a aplicação da base completa. Reitera-se dessa forma a pertinência de consideração das especificidades locais.

5.1 Trabalhos futuros

A intencionalidade genérica que norteou a metodologia proposta concentrou o processamento, até então aplicado, em definir bases com estruturas equivalentes no sentido de proporcionarem as análises comparativas. Dessa forma, temos sobreposta a exploração descritiva e preditiva das informações.

Na perpectiva de aperfeiçoamento, a intenção é complementar este estudo como uma cotribuição de apoio prescritivo, ou seja, especificar a relação entre as variáveis

utilizadas e os resultados obtidos. Pretende-se incrementar a análise e discussão expondo os fatores que indiquem predisposição do aluno à condição de permanência ou evasão.

As conclusões já alcançadas visualizam a pertinência de ações regionalmente direcionadas. Assim, a apresentação das características discentes com ênfase preditoras visa entregar subsidios aprimorados. Dessa maneira, gestores de instituições multicampi podem otimizar as estratégias de combate à evasão.

Referências

AGGARWAL, C. C. et al. *Data mining: the textbook*. [S.l.]: Springer, 2015. v. 1. Citado na página 9.

ARAUJO, F. d. *Descoberta de conhecimento em base de dados para o aprendizado da regulação médica/odontológica em operadora de plano de saúde*. Tese (Doutorado) — Dissertação de Mestrado, 2014. Citado na página 8.

ASSIS, W. R. de. Rede neural artificial: Identificação de acadêmicos em curso de licenciatura em matemática com possibilidade de desistência. *Cadernos Cajuína*, v. 5, n. 3, p. 534–544, 2020. Citado na página 4.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de informática na educação*, v. 19, n. 02, p. 03, 2011. Citado na página 9.

BAKER, R. S. Educational data mining: An advance for intelligent systems in education. *IEEE Intelligent systems*, IEEE, v. 29, n. 3, p. 78–82, 2014. Citado na página 7.

BAKER, R. S. *Big Data and Education*. Philadelphia, PA: University of Pennsylvania, 2023. A Massive Online Open Textbook (MOOT). Disponível em: <https://learninganalytics.upenn.edu/MOOT/bigdataeducation.html>. Citado na página 7.

BITENCOURT, P. B. d.; FERRERO, C. Predição de risco de evasão de alunos usando métodos de aprendizado de máquina em cursos técnicos. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2019. v. 8, n. 1, p. 149. Citado 2 vezes nas páginas 11 e 19.

BRASIL. *Portaria nº 1, de 03 de Janeiro de 2018*. Brasília, DF, 2018. Institui a Plataforma Nilo Peçanha - PNP, a Rede de Coleta, Validação e Disseminação das Estatísticas da Rede Federal de Educação Profissional, Científica e Tecnológica - REVALIDE. Disponível em: <http://pesquisa.in.gov.br/imprensa/jsp/visualiza/index.jsp?data=04/01/2018&jornal=5>. Citado 2 vezes nas páginas 4 e 21.

BRASIL. *Censo da Educaçação Escolar 2022*. 3. ed. Brasília - DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2023. Acesso em: 30 mar 2023. Citado na página 1.

BRASIL. *Censo da Educaçação Escolar 2023*. 1. ed. Brasília - DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2024. Acesso em: 30 fev 2024. Citado na página 3.

CHIQUITTO, A. G.; BAIDA, A. C. Análise quantitativa das causas da evasão escolar dos cursos técnicos de nível médio integrado dos institutos federais de educação, ciência e tecnologia. *Encontro Internacional de Gestão, Desenvolvimento e Inovação (EIGEDIN)*, v. 4, n. 1, 2020. Citado na página 2.

COIMBRA, C. L.; SILVA, L. B. e.; COSTA, N. C. D. A evasão na educação superior: definições e trajetórias. *Educação e Pesquisa*, Faculdade de Educação da

Universidade de São Paulo, v. 47, p. e228764, 2021. ISSN 1517-9702. Disponível em: <https://doi.org/10.1590/S1678-4634202147228764>. Citado na página 4.

DAVID, L. M. L.; CHAYM, C. D. Evasão universitária: um modelo para diagnóstico e gerenciamento de instituições de ensino superior. *Revista de Administração IMED*, Faculdade Meridional-IMED, v. 9, n. 1, p. 167–186, 2019. Citado na página 1.

DUTRA, J. F.; SOUZA, J. P. L. de; FERNANDES, D. Y. de S. Classificação de estudantes com potencial à evasão: aplicando mineração de dados no contexto de cursos técnicos subsequentes do ifpb. *Revista Principia-Divulgação Científica e Tecnológica do IFPB*, v. 59, n. 3, p. 1009–1027, 2022. Citado 3 vezes nas páginas 11, 13 e 22.

FACELI, K. et al. *Inteligência artificial: uma abordagem de aprendizado de máquina*. [S.l.]: LTC, 2011. Citado na página 10.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, ACM New York, NY, USA, v. 39, n. 11, p. 27–34, 1996. Citado na página 7.

FILHO, J. A. L.; SILVEIRA, I. Detecção precoce de estudantes em risco de evasão usando dados administrativos e aprendizagem de máquina. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, p. 480–495, 01 2021. Citado 2 vezes nas páginas 11 e 13.

FILHO, R. B. S.; ARAÚJO, R. M. de L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. *Educação por escrito*, v. 8, n. 1, p. 35–48, 2017. Citado na página 1.

FILHO, R. B. S.; ARAÚJO, R. M. de L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. *Educação por escrito*, v. 8, n. 1, p. 35–48, 2017. Citado na página 9.

GÉRON, A. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow*. [S.l.]: Alta Books, 2019. Citado na página 22.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data mining*. [S.l.]: Elsevier Brasil, 2015. Citado na página 10.

GÓMEZ, A. B.; BELMONTE, M. L. Evasão escolar, determinantes, políticas educacionais e itinerários subsequentes. *Research, Society and Development*, v. 9, n. 10, p. e6849109234–e6849109234, 2020. Citado na página 1.

HARRISON, M. Machine learning-guia de referência rápida. *São Paulo: Novatec Editora Ltda*, 2020. Citado na página 19.

MACHADO, V. P. *Inteligência Artificial*. [S.l.]: EDUFPI, 2011. Citado na página 10.

MORAES, G. H. *Guia de Referência Metodológica PNP*. [S.l.]: Ministério de Educação, 2020. Citado 3 vezes nas páginas 1, 15 e 19.

NERI, M. et al. Motivos da evasão escolar. 2009. Citado na página 1.

OLIVEIRA, I. S.; MEDEIROS, F. P. A.; ANDRADE, F. G. Seleção de atributos para classificadores de evasão escolar com dados da plataforma nilo peçanha. In: SBC. *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*. [S.l.], 2022. p. 30–39. Citado 4 vezes nas páginas 2, 4, 11 e 13.

OSORIO, J. K. H.; SANTACOLOMA, G. D. Predictive model to identify college students with high dropout rates. *Revista electrónica de investigación educativa*, Universidad Autónoma de Baja California, Instituto de Investigación y . . . , v. 25, 2023. Citado 2 vezes nas páginas 12 e 13.

PEREIRA, T. C. B.; PASSOS, G. d. O. Avaliação da política de assistência estudantil na educação profissional de nível técnico: análise dos indicadores de evasão e retenção no instituto federal de educação, ciência e tecnologia do piauí (ifpi) – campus teresina central. *Cadernos de Educação*, n. 57, dez. 2017. Disponível em: <https://periodicos.ufpel.edu.br/index.php/caduc/article/view/12823>. Citado na página 4.

PRADO, D. P. F.; BRITO, V. L. F. D.; NUNES, C. P. Concepções e perspectivas da plataforma nilo peçanha: regulação e emancipação. *Estudos em Avaliação Educacional*, Fundação Carlos Chagas, v. 33, 2022. Citado na página 15.

PRIMÃO, A. P. et al. Uso de algoritmos de machine learning para prever a evasão escolar no ensino superior: um estudo no instituto federal de santa catarina. 2022. Citado na página 22.

RAMOS, J. L. C. et al. Crisp-edm: uma proposta de adaptação do modelo crisp-dm para mineração de dados educacionais. In: SBC. *Anais do XXXI Simpósio Brasileiro de Informática na Educação*. [S.l.], 2020. p. 1092–1101. Citado na página 9.

RODRÍGUEZ, P. et al. A methodology to design, develop, and evaluate machine learning models for predicting dropout in school systems: the case of chile. *Education and Information Technologies*, Springer, p. 1–47, 2023. Citado 2 vezes nas páginas 12 e 13.

ROMERO, C.; ROMERO, J. R.; VENTURA, S. A survey on pre-processing educational data. *Educational data mining: applications and trends*, Springer, p. 29–64, 2014. Citado 2 vezes nas páginas 9 e 8.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, Wiley Online Library, v. 10, n. 3, p. e1355, 2020. Citado na página 2.

RUSSEL, S.; NORVIG, P. Inteligência artificial. tradução da terceira edição. *Rio de Janeiro, RJ: Elsevier Editora*, 2013. Citado na página 9.

SORENSEN, L. C. "big data" in educational administration: An application for predicting school dropout risk. *Educational Administration Quarterly*, SAGE Publications Sage CA: Los Angeles, CA, v. 55, n. 3, p. 404–446, 2019. Citado 2 vezes nas páginas 12 e 13.

SOUZA, V. F. d.; CAZELLA, S. C. Mineração de dados educacionais com algoritmos de regressão: um estudo sobre a predição do desempenho. *Revista Educar Mais*, v. 6, p. 183–198, 2022. Citado 2 vezes nas páginas 11 e 13.

SOUZA, V. F. de; CAZELLA, S. C. Mineração de dados educacionais com algoritmos de regressão: um estudo sobre a predição do desempenho. *Revista Educar Mais*, v. 6, p. 183–198, 2022. Citado na página 2.

VIANA, F.; SANTANA, A.; RABêLO, R. Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. In: *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*. Porto Alegre, RS, Brasil: SBC, 2022. p. 908–919. ISSN 0000-0000. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/22469>. Citado 5 vezes nas páginas 9, 10, 11, 12 e 13.

Apêndices

APÊNDICE A – Resultados com matrículas ativas 2018

Tabela 20 – Resultados detalhados das melhores performances na validação com matrículas ativas em 2018.

	DT						
	pre	rec	f1	tn	fp	fn	tp
CAPEDII	96.88	100.0	98.41	31.0	0.0	1.0	93.0
CAPAU	94.74	98.63	96.64	72.0	1.0	4.0	36.0
CASRN	92.0	98.57	95.17	69.0	1.0	6.0	62.0
CAPIC	94.29	94.29	94.29	33.0	2.0	2.0	36.0
CAFLO	88.71	98.21	93.22	110.0	2.0	14.0	34.0
CACOR	86.29	98.17	91.85	214.0	4.0	34.0	78.0
CAVAL	83.78	100.0	91.18	93.0	0.0	18.0	17.0
CAURU	83.64	100.0	91.09	46.0	0.0	9.0	4.0
CASJP	84.71	97.3	90.57	72.0	2.0	13.0	38.0
CATZS	83.23	95.56	88.97	129.0	6.0	26.0	69.0
CAOEI	79.89	99.32	88.55	147.0	1.0	37.0	19.0
CAANG	82.35	95.45	88.42	84.0	4.0	18.0	39.0
IFPI	81.72	92.43	86.75	2052.0	168.0	459.0	1029.0

	RF						
	pre	rec	f1	tn	fp	fn	tp
CAPEDII	96.88	100.0	98.41	31.0	0.0	1.0	93.0
CAPAU	94.74	98.63	96.64	72.0	1.0	4.0	36.0
CASRN	93.15	97.14	95.1	68.0	2.0	5.0	63.0
CAPIC	94.29	94.29	94.29	33.0	2.0	2.0	36.0
CAFLO	90.0	96.43	93.1	108.0	4.0	12.0	36.0
CACOR	86.59	97.71	91.81	213.0	5.0	33.0	79.0
CAVAL	83.78	100.0	91.18	93.0	0.0	18.0	17.0
CAURU	83.64	100.0	91.09	46.0	0.0	9.0	4.0
CASJP	89.33	90.54	89.93	67.0	7.0	8.0	43.0
CAOEI	79.89	99.32	88.55	147.0	1.0	37.0	19.0
CATZS	87.05	89.63	88.32	121.0	14.0	18.0	77.0
CAANG	84.38	92.05	88.04	81.0	7.0	15.0	42.0
IFPI	82.93	90.36	86.48	2006.0	214.0	413.0	1075.0

APÊNDICE B – Resultados com matrículas ativas 2019

Tabela 21 – Resultados detalhados das melhores performances na validação com matrículas ativas em 2019.

	DT						
	pre	rec	f1	tn	fp	fn	tp
CAPAU	97.37	100.0	98.67	74.0	0.0	2.0	9.0
CACOC	93.94	100.0	96.88	31.0	0.0	2.0	4.0
CAPAR	91.76	100.0	95.71	78.0	0.0	7.0	82.0
CASJP	92.44	99.1	95.65	110.0	1.0	9.0	29.0
CASRN	91.26	100.0	95.43	94.0	0.0	9.0	31.0
CAPEDII	85.58	100.0	92.23	89.0	0.0	15.0	49.0
CAANG	86.82	98.25	92.18	112.0	2.0	17.0	19.0
CAPIC	89.6	94.92	92.18	112.0	6.0	13.0	43.0
CACAM	85.9	98.53	91.78	67.0	1.0	11.0	11.0
CATCE	86.62	97.53	91.75	356.0	9.0	55.0	117.0
CAURU	83.04	97.89	89.86	93.0	2.0	19.0	9.0
CACOR	83.97	94.02	88.71	110.0	7.0	21.0	51.0
IFPI	80.89	94.85	87.31	2044.0	111.0	483.0	754.0

RF							
Pre	Rec	F1	tn	fp	fn	tp	
97.37	100.0	98.67	74.0	0.0	2.0	9.0	CAPAU
93.94	100.0	96.88	31.0	0.0	2.0	4.0	CACOC
92.77	98.72	95.65	77.0	1.0	6.0	83.0	CAPAR
92.44	99.1	95.65	110.0	1.0	9.0	29.0	CASJP
93.81	96.81	95.29	91.0	3.0	6.0	34.0	CASRN
86.82	98.25	92.18	112.0	2.0	17.0	19.0	CAANG
90.24	94.07	92.12	111.0	7.0	12.0	44.0	CAPIC
89.36	94.38	91.8	84.0	5.0	10.0	54.0	CAPEDII
87.53	96.16	91.64	351.0	14.0	50.0	122.0	CATCE
88.89	94.12	91.43	64.0	4.0	8.0	14.0	CACAM
83.64	96.84	89.76	92.0	3.0	18.0	10.0	CAURU
86.18	90.6	88.33	106.0	11.0	17.0	55.0	CACOR
81.87	93.04	87.1	2005.0	150.0	444.0	793.0	IFPI

APÊNDICE C – Resultados com matrículas ativas 2020

Tabela 22 – Resultados detalhados das melhores performances na validação com matrículas ativas em 2020..

DT						
pre	rec	f1	tn	fp	fn	tp
100.00	100.00	100.00	177.00	0.00	0.00	3.00
100.00	100.00	100.00	23.00	0.00	0.00	14.00
97.28	100.00	98.62	143.00	0.00	4.00	4.00
94.96	100.00	97.41	113.00	0.00	6.00	5.00
98.15	96.36	97.25	106.00	4.00	2.00	22.00
94.81	99.22	96.97	128.00	1.00	7.00	33.00
93.88	99.49	96.60	583.00	3.00	38.00	63.00
93.10	100.00	96.43	27.00	0.00	2.00	32.00
92.59	100.00	96.15	25.00	0.00	2.00	26.00
91.67	99.00	95.19	99.00	1.00	9.00	14.00
89.47	100.00	94.44	17.00	0.00	2.00	32.00
89.47	99.35	94.15	153.00	1.00	18.00	12.00
88.60	98.10	93.11	1.958.00	38.00	252.00	492.00

RF							
Pre	Rec	F1	tn	fp	fn	tp	
100	100	100	177	0	0	3	CAPIC
100	100	100	23	0	0	14	CACAM
97.28	100	98.62	143	0	4	4	CAPAR
94.96	100	97.41	113	0	6	5	CASRN
98.15	96.36	97.25	106	4	2	22	CACOC
96.18	97.67	96.92	126	3	5	35	CATZS
94.17	99.15	96.59	581	5	36	65	CATCE
93.1	100	96.43	27	0	2	32	CASJP
96	96	96	24	1	1	27	CAURU
92.45	98	95.15	98	2	8	15	CAPEDII
89.47	100	94.44	17	0	2	32	CACOR
89.94	98.7	94.12	152	2	17	13	CAOEI
89.67	96.59	93.01	1928	68	222	522	IFPI

APÊNDICE D – Resultados com matrículas ativas 2021

Tabela 23 – Resultados detalhados das melhores performances na validação com matrículas ativas em 2021.

	DT						
	pre	rec	f1	tn	fp	fn	tp
CAVAL	99.19	100.0	99.6	123.0	0.0	1.0	19.0
CAURU	97.47	100.0	98.72	193.0	0.0	5.0	16.0
CASJP	95.74	100.0	97.83	45.0	0.0	2.0	6.0
CACOR	95.45	100.0	97.67	147.0	0.0	7.0	1.0
CACAM	95.29	100.0	97.59	162.0	0.0	8.0	15.0
CACOC	94.37	100.0	97.1	67.0	0.0	4.0	6.0
CAPIC	94.05	100.0	96.93	79.0	0.0	5.0	40.0
CAPAR	94.96	98.51	96.7	132.0	2.0	7.0	45.0
CATCE	94.79	98.51	96.62	928.0	14.0	51.0	327.0
CAOEI	91.82	99.02	95.28	101.0	1.0	9.0	19.0
CAFLO	93.71	96.4	95.04	134.0	5.0	9.0	74.0
IFPI	91.72	98.05	94.78	2760.0	55.0	249.0	851.0

RF							
Pre	Rec	F1	tn	fp	fn	tp	
99.19	100.0	99.6	123.0	0.0	1.0	19.0	CAVAL
97.47	100.0	98.72	193.0	0.0	5.0	16.0	CAURU
95.74	100.0	97.83	45.0	0.0	2.0	6.0	CASJP
95.45	100.0	97.67	147.0	0.0	7.0	1.0	CACOR
96.39	98.77	97.56	160.0	2.0	6.0	17.0	CACAM
94.37	100.0	97.1	67.0	0.0	4.0	6.0	CACOC
95.12	98.73	96.89	78.0	1.0	4.0	41.0	CAPIC
94.96	98.51	96.7	132.0	2.0	7.0	45.0	CAPAR
96.8	96.28	96.54	907.0	35.0	30.0	348.0	CATCE
92.59	98.04	95.24	100.0	2.0	8.0	20.0	CAOEI
96.3	93.53	94.89	130.0	9.0	5.0	78.0	CAFLO
93.16	96.27	94.69	2710.0	105.0	199.0	901.0	IFPI

Anexos

ANEXO A – Nomenclatura oficial de campi do Instituto Federal do Piauí

Sigla	Campus
CACAM	Campus Campo Maior
CAPAR	Campus Parnaíba
CATCE	Campus Teresina Central
CATZS	Campus Teresina Zona Sul
CAPIC	Campus Picos
CAPAU	Campus Paulistana
CAANG	Campus Angical do Piauí
CAFLO	Campus Floriano
CAJFR	Campus José de Freitas
CADIR	Campus Dirceu Arcoverde
CAPIX	Campus Pio IX
CAOEI	Campus Oeiras
CASJP	Campus São João do Piauí
CACOR	Campus Corrente
CAVAL	Campus Valença
CAPIR	Campus Piripiri
CASRN	Campus São Raimundo Nonato
CAURU	Campus Uruçuí
CAPEDII	Campus Pedro II
CACOC	Campus Cocal