Machine learning models evaluation for predicting school dropout in multicampus teaching scenarios

Francisco Alysson da Silva Sousa[1], **Vinicius Ponte Machado**[2], **Rodrigo de Melo Souza Veras**[2] **André Macedo Santana**[2]

[1]Instituto Federal do Piauí (IFPI)

[2]Departamento de Computação - Universidade Federal do Piauí (UFPI)

{webalysson,vinicius,rveras,andremacedo}@ufpi.edu.br

Abstract. *The article addresses school dropout as one of the main challenges in education. In the Federal Network of Professional Education, the wide reach of institutions adds complexity to the problem. In this context, predictive analysis can guide proactive strategies. Thus, artificial intelligence, specifically machine learning, emerges as an essential tool for educational management. The study evaluates predictive models applied to the context of multicampus technical education. Some of these models obtained significant results with harmonic means above 90%, demonstrating a balance between sensitivity and accuracy. These results suggest a relevant potential to support decision-making in combating school dropout.*

Resumo. *O artigo aborda a evasão escolar como um dos principais desafios da educação. Na Rede Federal de Educação Profissional, o amplo alcance das instituições acrescenta complexidade ao problema. Nesse contexto, a análise preditiva pode orientar estratégias proativas. Assim, a inteligência artificial, especificamente o aprendizado de máquina, surge como uma ferramenta essencial para a gestão educacional. O estudo avalia modelos preditivos aplicados ao contexto da educação técnica multicampi. Alguns desses modelos obtiveram resultados significativos com médias harmônicas acima de 90%, demonstrando equilíbrio entre sensibilidade e precisão. Esses resultados sugerem um potencial relevante para apoiar a tomada de decisão no combate à evasão escolar.*

1. Introdução

O ato de deixar de frequentar uma instituição de ensino se revela de compreensão complexa dada a diversidade de contextos sociais, regionais, culturais e socioeconômicos de comum ocorrência nesse comportamento. As várias caracterizações possíveis sinalizam como não trivial encontrar padrões na evasão escolar. Assim como denota-se não simples a formulação de direcionadas estratégias de enfrentamento, sobretudo em instituições regionalmente diversificadas, principal característica da Rede Federal de Educacação Profissional, Científica e Tecnológica (RFEPCT).

Reconhecendo limitações no monitoramento da Educação Profissional e Tecnológica (EPT), o Ministério da Educação (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 indicadores de gestão extraídos de informações validadas pelas próprias instituições integrantes [Brasil 2018].

Contudo, constata-se ainda uma carência de contribuições resultantes de estudos preditivos a partir da descentralidade dos ambientes assim constituídos. As limitações ocorrem em detrimento do potencial em subsídios que podem ser obtidos por meio das análises inferenciais multicampi. Essa demanda converge integralmente com a aplicabilidade da Aprendizagem de Máquina (AM), subárea da inteligência artificial apoiada na identificação de padrões para a predição de comportamentos [Oliveira et al. 2022].

Trabalhos analisados evidenciam em comum o auxílio da computação para se conhecer de forma prévia os discentes com tendência a evadir [Ramos et al. 2020]. No entanto, é recorrente o foco no desenvolvimento de estudos com visão institucional ampla, quando um conjunto de especificidades locais pode revelar ganhos na compreensão do problema. Consequentemente, projeta-se assim uma maior eficácia no planejamento das ações de acompanhamento dos discentes [Romero and Ventura 2020].

Nessa perspectiva, aborda-se neste trabalho a aplicação de recursos computacionais frente ao problema da evasão escolar. Dentro deste abrangente desafio, define-se como delimitação a abordagem desta adversidade no âmbito da EPT, especificamente na Educação Profissional Técnica de Nível Médio (EPTNM).

Para tanto, utilizamos informações publicadas anualmente na Plataforma Nilo Peçanha (PNP)[1], sistema oficial de estatísticas da EPT. Os microdados foram obtidos no Portal de Dados Abertos[2] do MEC e referem-se ao acompanhamento de situações das matrículas no período de 2018 a 2022. Estudos de caso foram definidos e aplicados considerando a abrangência do Instituto Federal do Piauí (IFPI).

Diante do exposto, apresenta-se como motivação contribuir com uma análise preditiva a partir das informações resultantes da iniciativa REVALIDE. Este recurso como fonte de dados proporciona um estudo escalável pois a origem contempla o monitoramento nacional EPTNM. Foram avaliadados os subconjuntos representivos de um cenário de ensino capilarizado e o mesmo ambiente como uma base única, todos em composições anuais diversificadas.

Partindo dessa instigação, o objetivo geral foi verificar estratégias para o provimento de modelos preditivos em ambientes multicampi. O método delineado apoia-se nos conceitos da Mineração de Dados Educacionais e consiste na avaliação de classificadores diante de dados regionalmente diversificados. O recorte temporal inclui bases únicas e agrupadas das edições PNP. Sobre estes conjuntos, os seguintes passos específicos foram definidos:

- Verificar atributos, por edição, quanto à equivalência e suficiência para a identificação de padrões;
- Identificar a consistência da atualização sequencial de matrículas;
- Definir estudos de caso contemplando composições diversificadas e seus reflexos nos resultados;
- Avaliar modelos, por campi e instituição, identificando métricas condizentes com a característica original em termos de distribuição por classe;
- Validar a performance preditiva dos classificadores treinados quando diante de dados novos.

[1]https://www.gov.br/mec/pt-br/pnp

[2]https://dadosabertos.mec.gov.br/pnp

2. Referencial teórico

2.1. Aprendizagem de máquina

Como forma de proporcionar verificações não limitadas a ocorrências pretéritas, estudos sobre 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 publicação em [Russel and 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 algoritmos que buscam encontrar padrões a partir de fatos ocorridos [Machado 2011]. Como contribuição direta desta habilidade, tem-se no âmbito das pesquisas um importante recurso de apoio aos especialistas, tendo em vista análises não limitadas às suas próprias conclusões [Faceli et al. 2011]

Especificamente, quanto as teorias que embasam as instruções em AM, a inferência estatística fundamenta os principais métodos, 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. Para informações não categorizadas, tem-se a possibilidade de segmentação em grupos conforme possíveis similaridades, conceito esse relacionado ao paradigma não supervisionado [Goldschmidt et al. 2015].

Dentro do potencial que a 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.

2.2. Mineração de Dados Educacionais

Quando enfatizadas questões relacionadas à educação, a minerar dados 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 et al. 2014], neste conceito há a combinação de algoritmos computacionais e métodos estatísticos. Para os autores, o objetivo não é apenas transformar dados em conhecimento, mas também aplicá-lo para a tomada de decisão.

As ocorrências em educação demandam adaptações de acordo com o contexto. [Baker et al. 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.

Nota-se portanto que a MDE contribui ao se inserir como ferramenta 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].

3. Trabalhos Relacionados

É facilmente constatando que a evasão representa um problema recorrente nas diversas formas de oferta da educação. Nessa realidade é comum encontrar trabalhos acadêmicos que se apropriam dessa investigação, notadamente com o uso de algoritmos capazes de aprender e estimar situações futuras.

Em [Dutra et al. 2022], os autores analisaram ocorrências de evasão no Instituto Federal da Paraíba (IFPB). Considerando os cursos técnicos subsequentes, buscou-se validar hipóteses como a distância de deslocamento diário. Porém, o estudo não comprova essa colocação, enquanto que se sobressaem fatores outros como a quantidade de períodos letivos cursados. O modelo testado com o algoritmo SVM gerou resultados com acurácia de 93%.

Características demográficas e socioeconômicas tiveram suas influências pesquisadas em [Souza and Cazella 2022]. Os autores elabororararam uma proposta preditiva a partir dos desempenhos em disciplinas básicas do ensino médio no Instituto Federal do Rio Grande do Sul (IFRS). Nesta investigação, os algoritmos *Decision Tree* e *Random Forest* apresentaram 90% e 80% de acurácia, respectivamente.

Ratificando o despertar por ações preventivas, [Oliveira et al. 2022] propuseram verificar atributos relevantes correlacionados com a evasão no Instituto Federal da Paraíba (IFPB). O método proposto usou dados de cursos técnicos e superiores de outros Institutos Federais a partir dos quais extraíram métricas para a avaliação dos classificadores. Destaca-se no trabalho a acurácia do algoritmo *Random Forest* (89%), superando os modelos *Multilayer Perceptron* (81%), *Decision Tree* (80%) e Naive Bayes (67%).

Nesse mesmo sentido, [Bitencourt and Ferrero 2019], abordaram também os cursos técnicos com o intuito de identificar tendências à evasão. Desenvolvido no Instituto Federal de Santa Catarina (IFSC), o estudo fez uso de amplas possibilidades de parametrização dos algoritmos. No entanto, a justificativa para escolha da configuração aplicada menciona apenas o desempenho pela acurácia, métrica não apropriada em ambientes de notório desbalancemento entre classses, como em bases educacionais.

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

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

Esta proposta avalia modelos de predição cujas as performances expressam todos os campi. Os resultados, sumarizados por médias, representam o quão eficiente se comporta um mesmo classificador nos subconjuntos por unidades de ensino. Busca-se assim melhor caracterizar a diversificação do ambiente diante da contribuição pretendida.

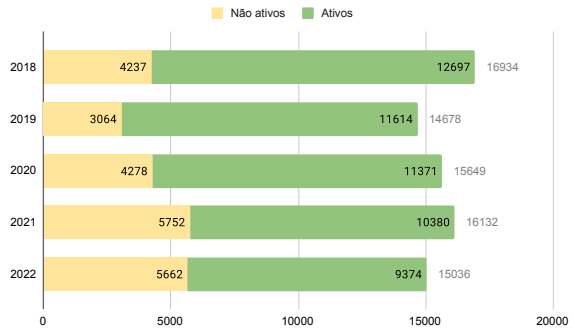
4. Ambiente Experimental

Os recortes temporais descritos a seguir delinearam quatro estudos de caso (EC). Foram experimentos avaliados enquanto composições para o provimento de modelos preditivos.

- EC1: considerou o acompanhamento conforme periodicidade de publicação da PNP que visa identificar anualmente a situação do aluno;
- EC2: contemplou dados apenas dos anos 2017, 2018 e 2019 que constam nas edições 2018, 2019 e 2020 da PNP. A primeira definição corresponde ao conceito de ano base enquanto a segunda especifica o ano de referência;
- EC3: corresponde à concatenação das edições 2018 a 2022. A inclusão do período de Pandemia de Covid-19 visa verificar implicações na identificação de padrões.
- EC4: compartilha a mesma definição de EC1, no entanto, distingui-se quanto ao particionamento de treino e teste pela técnica de validação cruzada *k-fold*.

Todas as abordagens de treinamento fizeram uso de instâncias não ativas (ativo=0) rotuladas em classes positiva, com valor 1 para evadidos, ou negativa, com valor 0 para não evadidos. A classificação e os agrupamentos mencionados constam detalhados na Figura 1(a) e sintetizados na Figura 1(b).

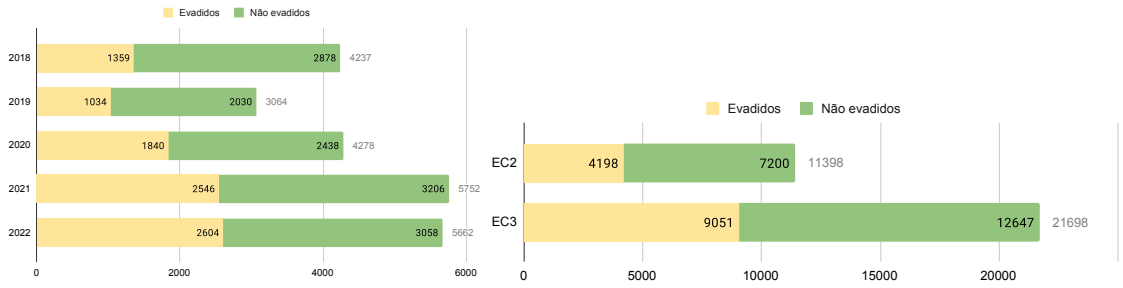
Situação	Categoria	Ativo	Evadido
Abandono Cancelada Desligada Reprovado Tranf. externa Tranf. interna	Evadido	0	1
Cursando	Em curso	1	Valor desconhecido
Concluída Integralizada	Concluínte	0	0



(a) **(b)**

Figura 1. Classificação(a) e quantiatativo(b) de matrículas por situação.

A Figura 2(a) apresenta os quantitativos das bases individuais de EC1. As composições concatenadas para EC2 e EC3 constam na figura 2(b).



(a) **(b)**

Figura 2. Composição das bases individuais(a) e concatenadas(b).

5. Metodologia

Excetuando-se a situação registrada como *em curso*, correspondente às matriculas ativas, a definição "rotuladas" está relacionada às categorias evadido ou concluinte. Com estes rótulos, formulou-se o problema de classificação binária quanto à evasão escolar no contexto analisado. A visão geral da metodologia está representada abaixo na Figura 3

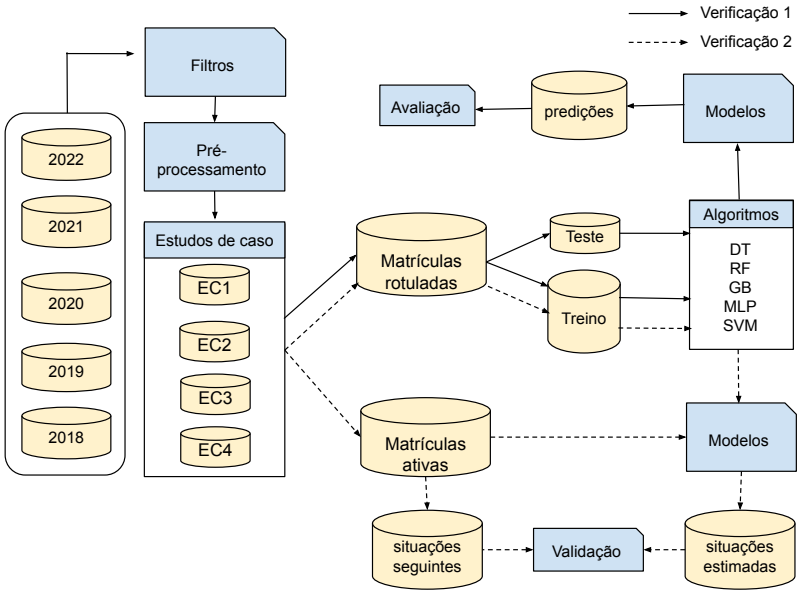


Figura 3. Metodologia de avaliação empregada.

Dessa forma, duas verificações experimentais foram definidas para o provimento dos modelos de aprendizagem de máquina.

A verificação 1 considerou somente as instâncias rotuladas para a segmentação entre treinamento e teste. No entanto, em EC1, EC2 e EC3, possíveis limitações decorrentes da execução única dos algoritmos por campi demandaram resultados mais representativos, estes buscados com a iteração *k-fold*, processamento aplicado em EC4.

A verificação 2 utilizou os subconjuntos de matrículas rotuladas e ativas. O primeiro para treinamento, e o segundo para aplicar a classificação em evadido ou não evadido no ano seguinte. Essa estratégia demandou o levantamento das situações então definidas então como "em curso", consultando-as na edições posteriores para validar as predições.

As informações comuns no monitoramento anual de matrículas pela PNP, com exceção dos identificadores únicos, foram utilizadas em todos os experimentos visando a análise comparativa. Na Tabela 1 consta a lista de atributos da base indicando os selecionados em observação ao mencionado critério ocorrência. Foram aplicados os algoritmos *Decision Tree* (DT), *Random Forest* (RF), *Gradient Boosting Classifier* (GB), *Multi Layer Perceptron* (MLP) e *Suport Vector Machine* (SVM).

Tabela 1. Atributos das edições PNP utilizadas.

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

6. Métricas de avaliação

O desbalanceamento entre as classes foi fator prepoderante na definição da metodologia de avaliação. A natural constituição majoritária por alunos não evadidos foi mantida por representar um aspecto comum em bases de dados educacionais ao termos a evasão como um comportamento excepcional.

A divergência quantitativa identificada compromete a equidade para o aprendizado dos modelos testados. Portanto, o cenário apontou claramente para a pertiência da avaliação específica, em detrimento das métricas de análise geral, como a acurácia.

Considerando esta característica na composição de todos os estudos de caso, foram usadas as métricas *recall* e *precision*, ambas compondo em igual nível de importância a métrica *F1-score* para avaliar os desempenhos dos modelos, conforme explicações a seguir:

A sensibilidade (*recall*), neste contexto em relação aos não evadidos (classe negativa), tende a vantagens em decorrência da sobreposição no mencionado desbalanceamento. Essa métrica, denota a importância da baixa ocorrência de falsos positivos (FP).

A precisão (*precision*) foi complementa a avaliação por expressar a qualidade das predições. A métrica evidencia a baixa ocorrência de falsos apontamentos, neste caso, da classe negativa.

Partindo dessas considerações, nota-se como essencial equiparar os desempenhos nas habilidades de recuperar o maior número de ocorrência da classe (*recall*) bem como quantos desses apontamentos estão realmente corretos (*precision*). Nesse sentido, a média harmônica composta por *recall* e *precision* foi usada para expressar o desempenho dos modelos.

7. Resultados

7.1. EC1

Conforme apresentado na Figura 4, observa-se que os modelos testados em EC1 apresentaram constante variação. Verificou-se nesta etapa que os desempenhos poderiam evidenciar possível relação com a composição em termos quantitativos.

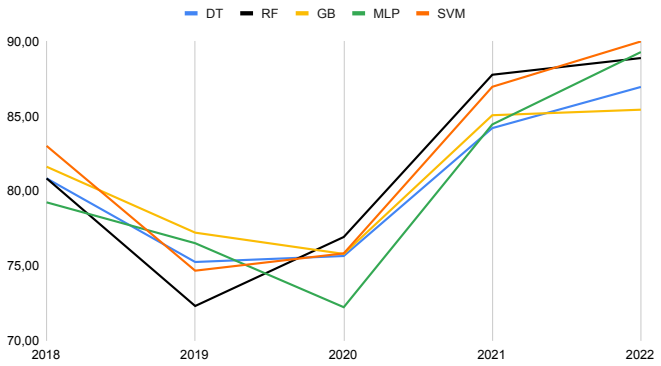


Figura 4. Evolução das médias de desempenho do estudo de caso EC1.

7.2. EC2 e EC3

Na Figura 5(a), temos a avaliação sob a perpectiva dos melhores resultados. A comparação destaca uma maior ocorrência deste registro na composição quantitativa inferior, EC2. Já em termos de desempenhos médios, Figura 5(b), 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. As análises sinalizam a distinção na identificação de padrões na relação com o volume de dados.

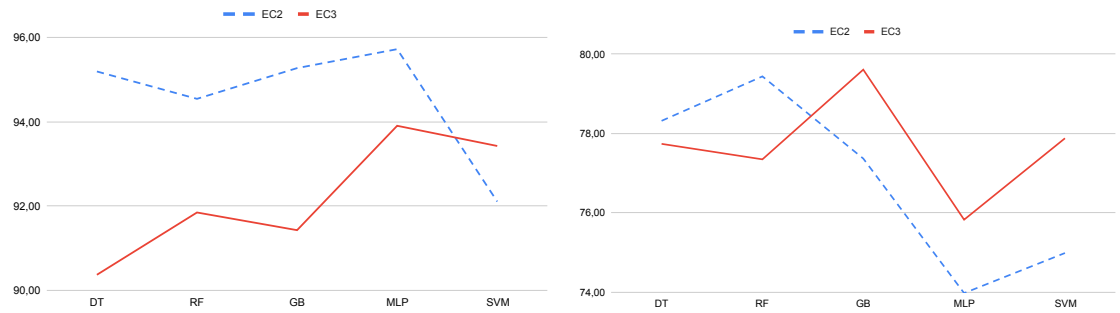


Figura 5. Melhores resultados(a) e Médias(b) dos experimentos EC2 e EC3.

Os estudos EC2 e EC3 foram verificações de hipóteses levantadas a partir de EC1 em relação a quantidade de dados. Conforme visto, o incremento de instância não implicou 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 representa cada campus.

7.3. EC4

São apresentados neste caso os valores médios extraídos com validação cruzada pela técnica *K-Fold*, com 5 (cinco) iterações em todos os subconjuntos que expressam as unidades de ensino. Conforme Figura 6, destacam-se os modelos DT e RF, não obstante à quantidade de dados disponíveis em cada ano.

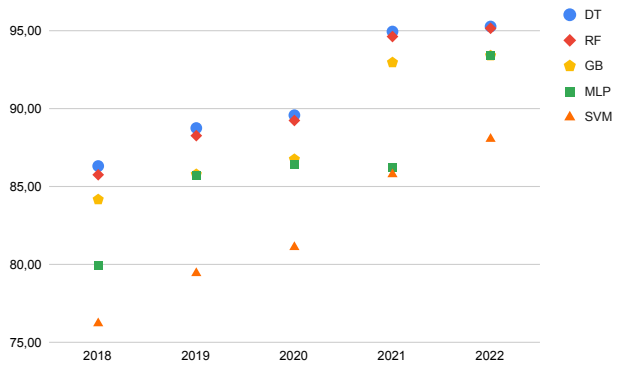


Figura 6. Médias extraídas com as bases anuais da composição de EC4.

7.4. 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. Com este intuito, as matrículas ativas em cada ano tiveram suas respectivas situações verificadas na base PNP imediatamente posterior. Ressalta-se que para as matrículas definidas como ativas da base 2022 não se procedeu a busca por situação futura. Nesse caso seriam necessários os dados de 2023, não disponíveis no momento desta operação.

A análise das habilidades específicas na Figura 7 demonstrou que RF se notabilizou por predições mais precisas, no entanto, RF se sobrassaiu na comparação quanto à sensibilidade (*recall*). Essa constatação reforça a pertinência da avaliação equiparada entre *recall* e *precision* expressa na sumarização por *F1-score*. Na representação abaixo os campi são indentificados pela nomenclatura padrão de siglas definidas pela instituição.

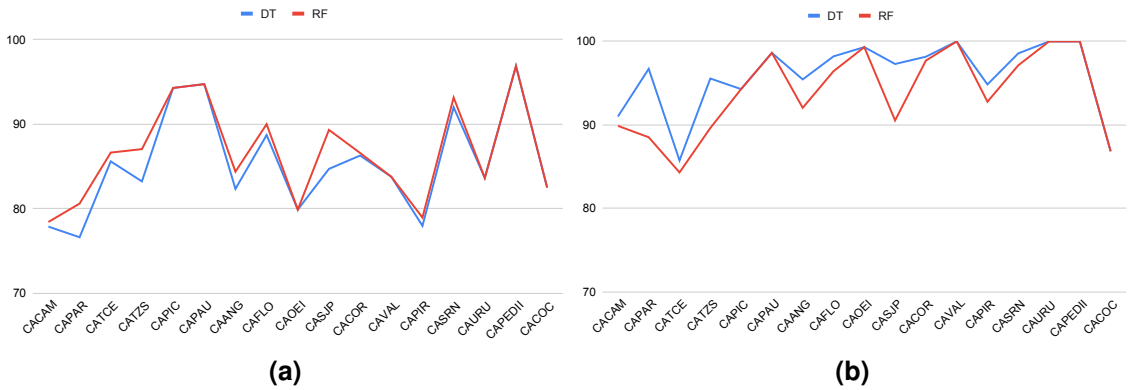


Figura 7. Comparativo da Precisão(a) e Sensibilidade(b) por campi

7.4.1. Considerações sobre a validação com alunos ativos

Em todos os testes desta etapa os desempenhos individuais e médios, por campi, apresentaram valores superiores a avaliação com a base completa do 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, evidencias que uma análise geral não contemplaria.

Na Figura 8 temos em destaque os melhores valores de *F1-score* encontrados a cada ano. Na comparação, o mesmo critério gerou a linha de tendência extraída com a base IFPI. O gráfico ressalta o potencial preditivo quando considerados os campi.

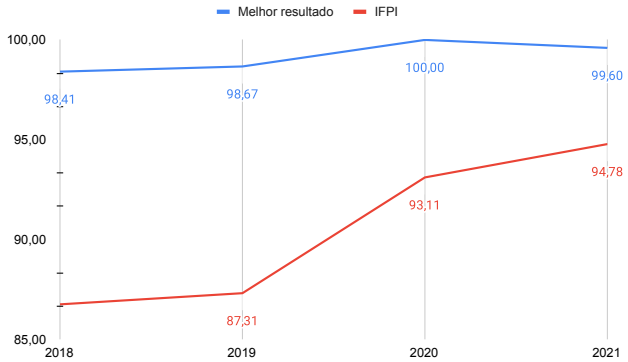
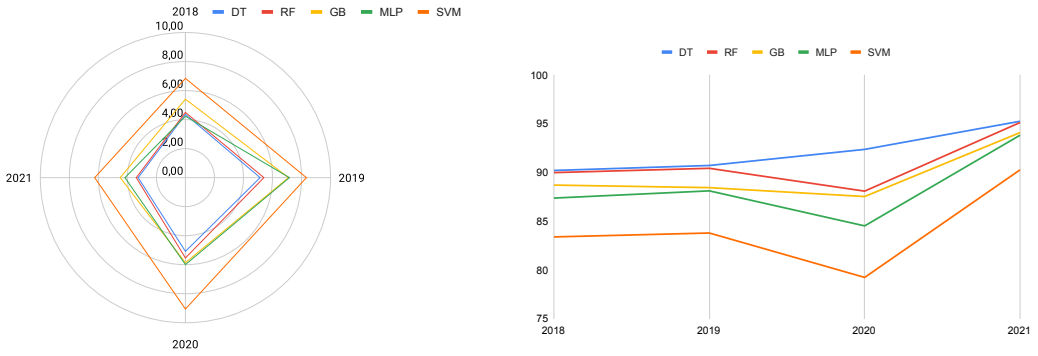


Figura 8. Comparação do melhor valor de *F1-score* a cada ano.

A Figura 9(a) denota a característica de dispersão dos resultados. Em conformidade com os destaques já mencionados para DT e RF, visualizam-se estes modelos como os que apontaram resultados menos dispersos. A Figura 9(b) apresenta as médias por *F1-score* com cada algoritmo utilizando as bases das edições anuais. Os modelos com DT e RF, novamente como destaques, ratificam as observações colocadas em torno das estruturas baseadas no conceito de árvore.



(a) **(b)**

Figura 9. Desvio padrão(a) e Médias(b) comparando modelos a cada ano.

8. Conclusão e Trabalhos Futuros

As experimentações iniciais descritas como EC1, EC2 e EC3 subsidiaram constatações como a composição melhor representativa para as predições. Neste caso, a delimitação por ano 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 estabilização. Foi possível observar uma evolução na qualidade representativa a partir dos dados de 2020. Houve progressos comprovados nas bases 2021 e 2022 para todos os algoritmos. O fato favorece a percepção de aperfeiçoamento na consistência das informações orinundas da iniciativa REVALIDE.

A evolução constatada aponta uma desvinculação entre a quantidade de instâncias e a qualidade do aprendizado. Experimentos diversificados ajudaram a constatar que os conjuntos expandidos não potencializaram a capacidade de convergência. Assim, fica evidenciado o dinamismo inerente aos fatores da evasão no aspecto temporal.

Execuções únicas em EC1, EC2 e EC3 limitaram as conclusões deste recorte. Além de resultados propensos à aleatoriedade, tem-se uma redução considerável de dados para treinamento quando procedida a segmentação única. Nesse sentido, a validação cruzada, aplicada em EC4, trouxe uma representação mais fidedigna da capacidade preditiva.

O destaque aos algoritmos DT e RF evidenciou a adequação das estruturas em árvore para as inferências. Em termos de sensibilidade e precisão, as habilidades distintas recomendariam aplicações em casos de relevância prioritária. No entanto, o objetivo demandou a consideração de ambas as métricas equiparadas na composição de *F1-score*.

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 permitem selecionar o algoritmo DT entre dos demais analisados. Considera-se a capacidade de resultados mais equilibrados a partir de métricas condizentes com o desbalanceamento.

O cenário multicampi proporcionou ainda evidenciar a importância de se considerar a composição heterogênea das instituições assim caracterizadas. As predições por campi foram mais eficientes quando comparadas com a aplicação da base completa. Reitera-se dessa forma a pertinência de consideração das especificidades locais nas instituições da Rede Federal de Educação Profissional.

Como perpectiva de aperfeiçoamento futuro, 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 poderão otimizar as estratégias de combate à evasão apoiados nesta proposta.

Referências

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de informática na educação*, 19(02):03.

Bitencourt, P. B. d. and Ferrero, C. (2019). 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*, volume 8, page 149.

Brasil (2018). Portaria nº 1, de 03 de janeiro de 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.

Dutra, J. F., de Souza, J. P. L., and de Souza Fernandes, D. Y. (2022). 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*, 59(3):1009–1027.

Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. d. L. F. d. (2011). *Inteligência artificial: uma abordagem de aprendizado de máquina*. LTC.

Goldschmidt, R., Passos, E., and Bezerra, E. (2015). *Data mining*. Elsevier Brasil.

Lopes Filho, J. A. and Silveira, I. (2021). Detecção precoce de estudantes em risco de evasão usando dados administrativos e aprendizagem de máquina. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, pages 480–495.

Machado, V. P. (2011). *Inteligência Artificial*. EDUFPI.

Oliveira, I. S., Medeiros, F. P. A., and Andrade, F. G. (2022). Seleção de atributos para classificadores de evasão escolar com dados da plataforma nilo peçanha. In *Anais do I Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 30–39. SBC.

Ramos, J. L. C., Rodrigues, R. L., Silva, J. C. S., and de Oliveira, P. L. S. (2020). Crisp-edm: uma proposta de adaptação do modelo crisp-dm para mineração de dados educacionais. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1092–1101. SBC.

Romero, C., Romero, J. R., and Ventura, S. (2014). A survey on pre-processing educational data. *Educational data mining: applications and trends*, pages 29–64.

Romero, C. and Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 10(3):e1355.

Russel, S. and Norvig, P. (2013). Inteligência artificial. tradução da terceira edição. *Rio de Janeiro, RJ: Elsevier Editora*.

Souza, V. F. d. and Cazella, S. C. (2022). Mineração de dados educacionais com algoritmos de regressão: um estudo sobre a predição do desempenho. *Revista Educar Mais*, 6:183–198.