Impacto dos Dados Cadastrais, de Presença e Notas na Predição de Reprovação Estudantil: Um Estudo de Caso

Emanuel Túrmina Torres, UFSC-Joinville, emanueltt1000@gmail.com
https://orcid.org/0009-0004-4357-0035
Benjamin Grando Moreira, UFSC-Joinville, benjamin.grando@ufsc.br
https://orcid.org/0000-0002-0339-4012

Resumo: A retenção estudantil é uma preocupação no ambiente acadêmico, sendo a reprovação uma das principais causas de abandono escolar. Este estudo propõe um modelo preditivo baseado em aprendizado de máquina para antecipar o desempenho dos estudantes, demonstrando como a inclusão sucessiva de diferentes tipos de indicadores impacta a eficácia da predição. Utilizando dados do ambiente Moodle e do sistema acadêmico da Universidade, o modelo foi aplicado em uma disciplina introdutória de programação, analisando três conjuntos principais de indicadores: cadastrais, de presença e de notas. Os resultados demonstraram que a integração progressiva dos indicadores melhorou significativamente o desempenho do modelo. Inicialmente, utilizando apenas dados cadastrais, obteve-se uma acurácia de 84,3% na identificação de reprovações. A inclusão dos dados de presença elevou este índice para 90,7%, e a incorporação final dos dados de notas permitiu atingir uma sensibilidade de 96,4%.

Palavras-chave: Reprovação Estudantil, Aprendizado de Máquina, Análise Preditiva.

Impact of Demographic, Attendance, and Academic Performance Data on Student Failure Prediction: A Case Study

Abstract: Student retention is a critical concern in the academic environment, with failure being one of the main causes of school dropout. This study proposes a predictive model based on machine learning to anticipate student performance, demonstrating how the successive inclusion of different types of indicators impacts the effectiveness of the prediction. Using data from the Moodle environment and the academic system of the University, the model was applied in an introductory programming course, analyzing three main sets of indicators: demographic, attendance, and grades. The results demonstrated that the progressive integration of indicators significantly improved the model's performance. Initially, using only demographic data, an accuracy of 84.3% was achieved in identifying failures. The inclusion of attendance data raised this index to 90.7%, and the final incorporation of grade data allowed reaching a sensitivity of 96.4%.

Keywords: Student Reprobation, Machine Learning, Predictive Analysis.

1. Introdução

A retenção estudantil é questão complexa que envolve as motivações e desafios enfrentados pelos alunos. De acordo com o National Center for Educational Statistics (NCES, 2021), o baixo desempenho acadêmico e a reprovação estão entre as principais razões que levam os estudantes a abandonar os estudos.

Em um contexto específico das engenharias, cerca de 40% dos estudantes não conseguem progredir além do primeiro ano de estudos. Além disso, dentre aqueles que conseguem, aproximadamente 30% enfrentam dificuldades significativas em disciplinas fundamentais, como cálculo e programação (Boylan-Ashraf e Haughery, 2018).

A intervenção precoce, por parte das instituições de ensino, com estudantes de risco, tem demonstrado ser eficaz na redução das taxas de reprovação, com uma redução média de 13% (Ramos *et al.*, 2015). Portanto, conseguir identificar de forma antecipada

padrões e fatores que afetam o desempenho dos estudantes se tornou um campo de interesse no setor educacional (Romero e Ventura, 2010).

Dada a importância de identificar antecipadamente estudantes em risco de reprovação, propõe-se neste trabalho o desenvolvimento de modelos preditivos baseados em Aprendizado de Máquina (AM) fundamentado por dados educacionais dos estudantes.

Para uma compreensão abrangente, é importante considerar tanto dados de desempenho em disciplinas quanto dados demográficos, como etnia, nível de escolaridade, idade e gênero. Essas variáveis são obtidas por meio de pesquisas, registros acadêmicos e outras fontes externas e são consideradas como possíveis preditores para determinar o êxito acadêmico dos estudantes (Woodman, 2001). Essa abordagem holística, podendo incorporar tanto dados virtuais provenientes de um Ambiente Virtual de Aprendizagem (AVA) quanto dados demográficos, enriquece a análise e a aplicação preditiva (Cortez e Silva, 2010).

Para validar o modelo e especificar fatores que afetem a possibilidade de reprovação, é realizado um estudo de caso em turmas de uma disciplina introdutória de programação de computadores, sendo essas turmas ofertadas para diversos cursos. Os dados utilizados são coletados do AVA Moodle e do sistema acadêmico da instituição de ensino, abrangendo variáveis sociodemográficas, frequência nas aulas e notas obtidas em atividades da disciplina. É conduzida uma análise implícita, na qual diferentes algoritmos de aprendizado de máquina são comparados em quatro períodos de aplicação/conclusão na disciplina: 25%, 50%, 75% e 100%. Esses períodos correspondem à quantidade de aulas ministradas da disciplina. Para cada intervalo, examinou-se a influência dos diferentes tipos de dados, isolados e em conjunto, na antecipação correta dos estudantes reprovados.

O diferencial neste trabalho está na avaliação dos impactos que os conjuntos de dados utilizados influenciam na predição das reprovações, considerando uma contribuição apenas dos dados cadastrais do aluno (no qual se considera a previsão antes mesmo do início da disciplina), utilizando a frequência do aluno e considerando sua participação/desempenho em atividades avaliativas aplicadas.

2. Trabalhos Relacionados

A revisão de trabalhos similares tem como objetivo fornecer uma base sólida para o desenvolvimento deste estudo, identificando lacunas no conhecimento existente e destacando metodologias eficazes. Ao compreender as pesquisas anteriores, busca-se avançar na análise da predição de reprovação estudantil, e contribuir para a expansão do corpo de conhecimento da área.

Na tentativa de explorar a usabilidade de dados sociodemográficos, Kovaĉić (2010) aplicou o AM para a predição de um conjunto de 435 estudantes. Com uma acurácia máxima de 60%, o autor destaca que as informações não fornecem dados suficientes para uma predição precisa entre indivíduos bem-sucedidos e malsucedidos. Resultado similar foi obtido por Cheewaprakobkit (2013), Cortez e Silva (2010), cuja constância nos resultados esteve na faixa de 50% a 60% de acurácia e indica a incompatibilidade de uso dessa vertente de informação nos problemas relacionados à reprovação estudantil [*].

Buscando uma abordagem diferente, Almarabeh (2017) utilizou dados educacionais, como notas, presença em seminários e participação em laboratórios ao longo de um semestre como fonte de análise. Seu melhor modelo apresentou uma acurácia

[*]Embora os resultados obtidos neste trabalho tenham uma acurácia maior, a acurácia é menor em relação a utilização de outros dados.

de 91% a partir do uso dos dados do semestre.

Kaensar e Wongnin (2023) utilizaram do Moodle para conduzir uma pesquisa em quatro estágios distintos (25%, 50%, 75% e 100%) do progresso acadêmico, e obtiveram um F1-score de 81,1% ao final do curso. No entanto, um aspecto notável surgiu durante a análise: o desempenho dos algoritmos varia significativamente em diferentes estágios do período letivo, especificamente, Máquina de Vetor de Suporte obteve o melhor desempenho para o estágio inicial, Regressão Logística foi mais eficaz para os estágios intermediários, e as Árvores de Decisão demonstraram superioridade no estágio final (Kaensar e Wongnin, 2023).

Em Souza e Moreira (2024), o trabalho buscou identificar fatores que apontem a possibilidade de desistência dos alunos em uma disciplina introdutória de programação. O trabalho utilizou o relatório do ambiente Moodle de presenças e atividades avaliativas, tendo obtido, para uma turma analisada, modelos com F1-score médio de 91,8% nos estágios de conclusão da disciplina de 25%, 50%, 75% e 100%.

3. Metodologia

Neste capitulo, são descritos os métodos utilizados para alcançar o objetivo de identificar a relação das variáveis com o desempenho dos estudantes e sua influência na predição dos reprovados.

As análises empreendidas fundamentaram-se em relatórios extraídos do Moodle e do Sistema de Controle Acadêmico da instituição de ensino. Estes sistemas forneceram um conjunto abrangente de dados, abarcando variáveis cadastrais, registros de presença e distribuição de notas dos estudantes.

O relatório de dados cadastrais é composto por informações sobre o curso (alunos de um total de 8 cursos fazem a disciplina objeto deste estudo de caso), situação acadêmica, dados sociodemográficos (data de nascimento, cidade e estado de origem, ano de conclusão do ensino médio, etc.), e indicadores de desempenho escolar dos estudantes, delineando um panorama do perfil discente e fornecendo substrato para análise preditiva. Metade das turmas tem a disciplina na sua primeira fase/semestre do curso, enquanto a outra metade a disciplina é de segunda fase/semestre, ou seja, a maioria dos alunos tem contato com a disciplina em seu primeiro ano de curso, embora as altas taxas de reprovação façam com que alunos com mais tempo de curso também façam a disciplina.

O relatório de presenças registra a frequência dos alunos às aulas, detalhando a participação ou ausência em cada sessão ao longo do período letivo. O relatório é obtido através da aba de sua exportação no módulo de presenças do Moodle, sendo exportado em formato de planilha eletrônica. Destaca-se que a disciplina em questão era ministrada em dois encontros semanais e isso pode influenciar em relação a outras com apenas um encontro semanal por agregar mais dados da participação do estudante.

O relatório de notas detalha as avaliações aplicadas ao longo da disciplina que constituem a nota final do aluno. O relatório em questão apresenta atividades como Laboratórios Virtuais de Programação (LVP), questionários e as avaliações parciais e final. A disposição das colunas com as notas no relatório é definida pela estrutura que o professor define no Moodle e não indica os pesos relativos de cada atividade na nota final. Entretanto, para a disciplina em questão, os LVP e os questionários constituem a categoria de *Avaliações Parciais* e representam 40% da nota, e a *Avaliação Final* representa os 60% restantes. Assim como o relatório de presenças, o relatório de notas pode ser exportado pela aba do módulo de notas do Moodle e em formato de planilha eletrônica. Destaca-se em relação as atividades avaliativas que essas eram aplicadas periodicamente, com atividade pelo menos uma vez por semana.

Sobre os dados cadastrais, os utilizados no estudo foram o nome do curso, o sexo e cor declarados pelo estudante, ano de ingresso na graduação, se o estudante é oriundo de ensino público, o Índice de Aproveitamento Acadêmico (esse índice é uma métrica da instituição de ensino e que considera um desempenho geral nas disciplinas do curso) e a cidade de origem do estudante (foi utilizado apenas se o estudante era da cidade da universidade ou se vinha de outra cidade). A partir dos atributos disponíveis, destaca-se outros que foram criados, sendo esses:

- **Intervalo Pré-Graduação:** tempo entre a conclusão do ensino médio e o ingresso na universidade, apontando para intervalos nos estudos que podem impactar a readaptação ao ambiente acadêmico.
- **Idade de Ingresso:** comparação entre a data de nascimento do estudante e a data de ingresso na instituição. Oferece uma métrica para examinar o impacto da idade no desempenho do estudante e potencialmente significativo nos primeiros semestres.
- **Idade Cursando:** é determinada pela diferença entre a data de nascimento do estudante e a data em que o estudante está matriculado na disciplina. Similar à idade de ingresso, oferece uma métrica sobre o impacto da idade no desempenho do estudante na disciplina.
- **Diferença entre Idade de Ingresso e Cursando:** é calculada pela diferença, em anos, entre a idade do estudante no momento do ingresso na instituição e a idade enquanto cursa a disciplina. Este índice permite analisar como o momento em que o estudante decide cursar a disciplina, seja logo no primeiro ano de curso ou após alguns anos, afeta seu desempenho.
- **Reprovação:** indica se o aluno já reprovou em alguma disciplina. É importante destacar que isso significa uma reprovação em qualquer disciplina de seu histórico, e não necessariamente na disciplina alvo de análise. Este índice ajuda a entender se reprovações passadas afetam a chance de reprovação na disciplina atual.

A Tabela 1 apresenta os atributos utilizados na análise. A tabela apresenta os atributos separados em 3 categorias, que envolvem a origem dos dados. A origem dos Dados Cadastrais são os provenientes do sistema acadêmico da instituição. Os Dados de Presenças são aqueles obtidos pelo relatório de presenças do Moodle, enquanto o Dados de Notas é obtido pelo relatório de notas do Moodle.

Tabela 1. Dicionário de dados dos atributos utilizados no estudo

Origem	Atributo	Descrição
Dados Cadastrais	nomeCurso	Nome completo do curso frequentado pelo estudante.
	Sexo	Gênero do estudante.
	racaCor	Categoria de raça ou cor do estudante.
	anoIngresso	Ano de ingresso do aluno na instituição.
	ensinoPublico?	Indica se o estudante veio de escola pública.
	IAA	Índice de aproveitamento acadêmico acumulado.
	reprovou?	Indica se o aluno já reprovou em alguma disciplina.
	origemRegiao	Região de origem do aluno.
	intervaloGrad	Intervalo entre o ensino médio e ingresso do estudante na instituição.
	idadeCursando	Idade do estudante durante a disciplina.
	idadeIngresso	Idade do estudante no momento do ingresso na instituição.
	idadeDif	Diferença em anos entre o ingresso e o momento atual.
Dados de Presenças	presenca_XX	Percentual de presença nas aulas, subdividido em quartis (25, 50, 75, 100).
	faltas_Consecutivas_XX	Número de faltas consecutivas, subdividido em quartis.
Dados de Notas	LPV_XX	Média dos LPV completados, subdividido em quartis.
	Quest_XX	Média dos questionários respondidos, subdividido em quartis.
	Atv_AF	Média das atividades avaliativas completadas, para o quartil de 100%
	Atv_AF_incompletas	Atividades avaliativas finais não completadas, para o quartil de 100%

Para as análises, diversas técnicas de Aprendizado de Máquina foram

consideradas, sendo elas: Naive Bayes (NB), Gradient Busting (GB), K-Nearest Neighbors (KNN), Redes Neurais Artificiais (RNA), Máquina de Vetor de Suporte (SVM, sigla do inglês), Árvore de Decisão (AD), Random Forest (RF) e Regressão Lógistica (LR, sigla do inglês).

Os dados foram selecionados em relação ao período de tempo de aplicação da disciplina, sendo separados em períodos que representam 25%, 50%, 75% e 100% da conclusão da disciplina. A utilização de 100% do período, embora nada efetiva para uma ação antecipada para evitar a reprovação do estudante, é utilizada como um balizador para avaliação do modelo classificador.

3.1. Ferramenta da realização das análises

A análise dos dados foi feita utilizando o software Orange Datamining[†], uma ferramenta open-source de mineração de dados que oferece uma interface gráfica intuitiva. Orange permite a criação de fluxos de trabalho interativos para pré-processamento, modelagem, visualização e análise de dados.

A Figura 1 representa uma visão geral do fluxo para realizar a análise que inclui todos os conjuntos de dados. É possível observar os processos de discretização e normalização dos dados, bem como sua integração. Também é possível observar os métodos de aprendizado de máquina utilizados e sua convergência para a avaliação dos resultados, resultados esses que podem ser visualizados a partir de uma matriz de confusão.

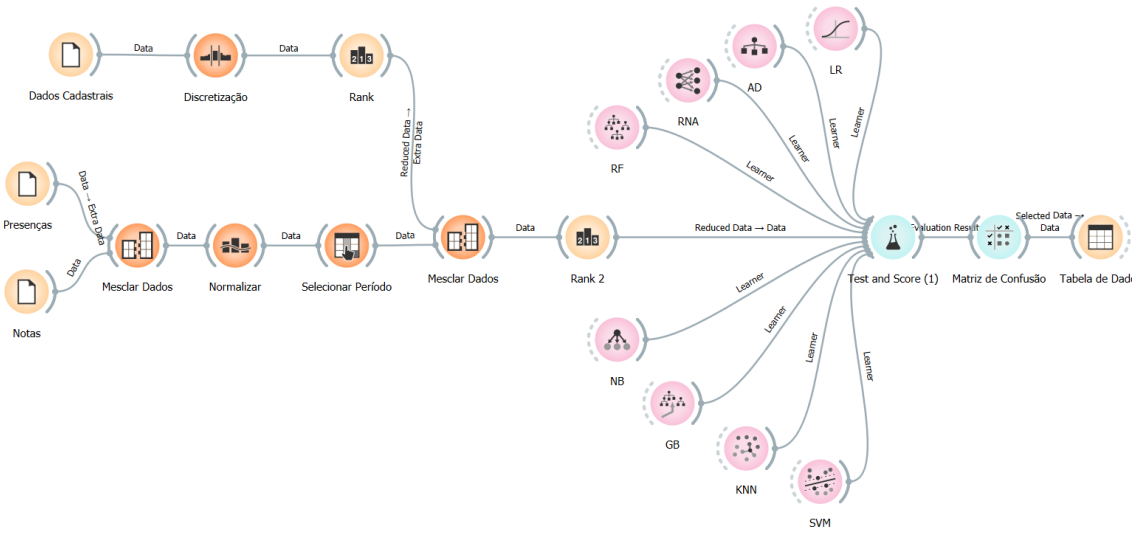

Figura 1. Fluxo para análise dos dados na Orange Datamining, com dados de todos os índices do trabalho

Desta-se o último widget (elemento mais a direita na Figura 1), que permite visualizar, de forma filtrada, os registros conforme seleção na matriz de confusão. Sendo assim, é possível visualizar todos os registros classificados na matriz de confusão (verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos), permitindo uma interpretação pontual, embora subjetiva, dos alunos e suas classificações.

4. Resultados das Análises Preditivas

Com o objetivo de compreender a influência dos diferentes tipos de dados na reprovação dos estudantes, esta seção detalha os fluxos de análise preditiva desenvolvidos

[†]Site da Orange Datamining: https://orangedatamining.com

e os resultados obtidos. A análise incide sobre os 93 estudantes de uma disciplina introdutória de programação de computadores, buscando entender a influência de cada conjunto de dados sobre os casos classificados correta e incorretamente.

Inicialmente, são descritos os resultados considerando apenas o uso dos Dados Cadastrais. Em seguida, aborda-se a inclusão dos Dados de Presença. Por fim, são adicionados os Dados de Notas e apresentada a integração de todos três conjuntos de dados.

4.1. Análise com os Dados Cadastrais

Para os índices cadastrais, as variáveis numéricas foram inicialmente discretizadas com o intuito de criar grupos com distribuições homogêneas. Cada variável é segmentada de modo que cada grupo resultante contenha um número similar de estudantes.

Os atributos disponíveis foram selecionados a partir da técnica de Seleção dos Atributos (também chamada de seleção de características, seleção de atributos ou, em inglês, feature selection). Neste trabalho os atributos foram ranqueados com base na proporção de ganho de informação (GI). Os oito índices mais relevantes foram utilizados nos algoritmos de aprendizado de máquina, e as métricas de classificação obtidas para cada abordagem (com e sem seleção de atributos) são apresentadas na Tabela 2 que mostra, além das métricas de desempenho, o algoritmo que obteve as métricas apresentadas.

Tabela 2. Métricas de Classificação de Reprovados com Dados Cadastrais

Discretização	Algoritmo	Sensibilidade	Precisão	F1-score	Acurácia
Sem seleção	NB	75,0%	72,1%	73,7%	75,0%
GI	GB	80,0%	85,2%	82,5%	84,3%

A análise dos resultados demonstra que a utilização da Seleção de Atributos é mais eficiente para identificar e classificar corretamente os dados, possibilidade já apresentada em outros trabalhos, como Xu *et al.* (2016), Souza e Moreira (2024).

É importante destacar que os algoritmos selecionados foram aqueles com maior sensibilidade aos resultados associados à reprovação dos estudantes. Para esse mesmo caso, por exemplo, o modelo KNN obteve uma sensibilidade de 93,3% para os aprovados, porém apenas 58,5% para os reprovados. Entende-se que para o problema apresentado no estudo de caso, erros classificando um estudante que será reprovado como um que será aprovado são piores do que classificar um aluno reprovado, mas ele não seria reprovado. Se o modelo classifica corretamente os aprovados, mas apresenta um alto número de falso positivos, muitos estudantes que necessitam de suporte não serão identificados e, consequentemente, não receberão a ajuda necessária. Portanto, um algoritmo com métricas gerais menos favoráveis, mas com alta sensibilidade para identificar os reprovados, é considerado mais vantajoso no estudo.

Para concluir a análise com os índices cadastrais, a matriz de confusão foi examinada com o objetivo de identificar as causas subjacentes às classificações incorretas pelos algoritmos. Nos casos de falso negativos, a maioria das ocorrências envolveu valores atípicos, onde estudantes com quatro ou mais índices associados a baixas taxas de aprovação acabaram sendo aprovados. Quanto aos falso positivos, valores atípicos representaram aproximadamente 50% das ocorrências. Para o restante, não foi possível determinar um motivo claro que justificasse a classificação incorreta.

Mesmo com métricas abaixo das encontradas em trabalhos relacionados sobre

reprovação estudantil, a sensibilidade encontrada apenas com os Dados Cadastrais indicam dificuldades que os alunos ingressantes já apresentavam antes mesmo de iniciarem na disciplina.

4.2. Análise com inclusão dos Dados de Presenças

Em relação aos dados de presenças, os valores foram submetidos a um processo de normalização, de modo que a amplitude dos mesmos se estabelecesse entre os limites de 0 e 1. Não foi realizada uma Seleção de Atributos por existirem apenas 2 atributos.

Inicialmente é avaliada a previsão relacionada apenas com os índices de presença, sendo as métricas alcançadas em cada intervalo de conclusão da disciplina detalhadas na Tabela 3.

Tabela 3. Métricas de Classificação de Reprovados com Dados de Presença

Período	Algoritmo	Sensibilidade	Precisão	F1-score	Acurácia
25%	NB	75,4%	65,3%	70,0%	70,0%
50%	GB	73,8%	77,4%	75,6%	77,9%
75%	KNN	75,4%	86,0%	80,3%	82,9%
100%	GB	76,9%	92,6%	84,0%	86,4%

Após a predição apenas com Dados de Presenças, a integração desses dados com os Dados Cadastrais foi explorada, buscando avaliar os benefícios dessa junção para a qualidade dos modelos preditivos. A hipótese subjacente é que essa abordagem possa aprimorar as métricas de classificação dos estudantes reprovados ao permitir a análise de interações entre variáveis que poderiam passar despercebidas quando examinadas isoladamente. Para essa análise, os dados foram combinados e a Seleção de Atributos foi definida com as oito variáveis de destacada relevância. A Tabela 4 apresenta as métricas de classificação obtidas a partir do uso dos Dados Cadastrais e dos Dados de Presenças.

Tabela 4. Métricas de Classificação com Dados Cadastrais e Dados de Presenças

Período	Algoritmo	Sensibilidade	Precisão	F1-score	Acurácia
25%	RNA	80,0%	83,9%	81,9%	83,6%
50%	NB	81,5%	89,8%	85,5%	87,1%
75%	RF	81,5%	91,4%	86,2%	87,9%
100%	RF	84,6%	94,8%	89,4%	90,7%

O primeiro período (conclusão de apensa 25% da disciplina) apresenta métricas similares às encontradas utilizando apenas os índices cadastrais, indicando ainda pouca influência das presenças. Entretanto, a partir do segundo período, as métricas começaram a melhorar. Quando comparadas com as métricas somente do uso dos Dados de Presenças, para 50%, 75% e 100% dos períodos, observou-se um aumento de 7,7%, 6,1% e 7,7% em sensibilidade e de 12,4%, 5,4% e 2,2% em precisão, respectivamente. Além disso, a mesclagem dos dados alterou o algoritmo de melhor desempenho para cada segmento.

Para a tabela de dados e matriz de confusão, os padrões e perfis relacionados aos índices encontrados em cada fluxo se mantiveram consistentes. No entanto, a mesclagem dos dados melhorou a classificação dos valores atípicos que anteriormente geravam classificações incorretas.

4.3. Análise com inclusão dos Dados de Notas

A seguir são apresentados os resultados da utilização dos Dados de Notas. De maneira similar ao realizado na seção anterior (com a inclusão dos Dados de Presenças),

os Dados de Notas foram normalizados entre o intervalo de 0 e 1 e não foi realizada a Seleção de Atributos pelo conjunto de atributos ser restrito a quatro elementos. A predição utilizando somente os atributos do desempenho nas atividades tem suas métricas de avaliação mostradas na Tabela 5.

Tabela 5. Métricas de Classificação de Reprovados com Dados de Nota

Período	Algoritmo	Sensibilidade	Precisão	F1-score	Acurácia
25%	NB	89,2%	85,3%	87,2%	87,9%
50%	RF	92,3%	80,0%	85,7%	85,7%
75%	SVM	93,8%	93,8%	93,8%	94,3%
100%	RNA	98,5%	95,5%	97,0%	97,1%

Os resultados da classificação utilizando apenas os atributos de desempenho em atividades avaliativas, quando comparados com os dos outros dados, são superiores. Isso possivelmente ocorre por refletirem diretamente o desempenho acadêmico do estudante, estabelecendo uma relação causal com os critérios de aprovação. As notas são indicadores consistentes do progresso do estudante e são menos suscetíveis a variações externas em comparação com presenças ou dados sociodemográficos, que são fatores mais indiretos.

A integração dos Dados de Notas, Dados Cadastrais e Dados de Presenças tem suas métricas avaliativas apresentadas na Tabela 6. Na integração de todos os tipos de dados, a Seleção de Atributos incluiu os nove melhores atributos ranqueados, dos três conjuntos de dados. Foram utilizados os nove melhores atributos por esses se destacarem dos demais no GI para classificação dos dados.

Tabela 6. Métricas de Classificação de Reprovados utilizando Dados Cadastrais, de presença e de notas

Período	Algoritmo	Sensibilidade	Precisão	F1-score	Acurácia
25%	LR	90,8%	92,2%	91,5%	92,1%
50%	SVM	92,3%	90,9%	91,6%	92,1%
75%	RNA	95,4%	92,5%	93,9%	94,3%
100%	AD	100,0%	92,9%	96,3%	96,4%

Os resultados, quando comparados com os modelos que utilizaram os índices isoladamente, revelam que a concatenação proporcionou uma melhora na sensibilidade e, principalmente, nas demais métricas de desempenho. A abordagem integrada, ao capturar a inter-relação das variáveis, apresenta-se como uma interpretação superior para a predição.

Essa melhora é melhor percebida a partir da Tabela 7, onde a sensibilidade de cada conjunto de dados é comparada. Por si só, a sensibilidade obtida considerando apenas os Dados de Notas é superior aos outros conjuntos de dados. Entretanto, ao serem integrados, houve uma melhora de 1,6%, 0% (não houve melhora na classificação), 1,6% e 1,5% para 25%, 50%, 75% e 100% do período letivo.

Finalmente, ao se considerar a aplicação dos modelos preditivos para uso efetivo, a utilização de diferentes algoritmos para períodos distintos pode ser custoso. Para exemplificar, um professor pode querer prever as probabilidades de reprovação dos seus estudantes após 33% do curso ter decorrido. Nesse contexto, torna-se difícil determinar qual algoritmo seria o mais adequado, já que não é possível identificar com precisão em que momento o SVM supera a LR em desempenho. Em relação à isso, a Tabela 8 apresenta a sensibilidade média dos quatro melhores modelos, considerando os três primeiros períodos do curso (25%, 50% e 75%).

Tabela 7. Métricas de Sensibilidade de Reprovados por Período Letivo para Cada Conjunto de Dados

Período	Todo conjunto	Notas	Presença e Cadastrais	Presença	Cadastrais
0%	-	-	-	-	80,0%
25%	90,8%	89,2%	80,0%	75,4%	-
50%	92,3%	92,3%	81,5%	73,8%	-
75%	95,4%	93,8%	81,5%	75,4%	-
100%	100,0%	98,5%	84,6%	76,9%	-

Tabela 8. Média de Sensibilidade para Cada Modelo

	LR	SVM	RNA	AD
Sensibilidade Média	90,26%	90,76%	91,13%	85,14%

O algoritmo de RNA se mostrou superior aos outros três, com uma diferença de +0,37% quando comparado com o segundo melhor modelo. Sendo assim, a fim de atender às características gerais que podem surgir da análise de diferentes turmas, propõe-se a utilização de RNA e seleção dos nove melhores atributos por GI.

5. Conclusão

Este estudo de caso investigou a eficácia da utilização de dados cadastrais, de presença e de notas para prever a reprovação estudantil em uma disciplina introdutória de programação. Os resultados demonstraram que a integração progressiva desses diferentes tipos de dados melhora significativamente o desempenho dos modelos de aprendizado de máquina na identificação de alunos em risco.

Nos trabalhos relacionados apresentados neste trabalho, mas também comum em outros trabalhos similares, o algoritmo com melhor desempenho varia com as características dos dados utilizados. Essa mudança pode ocorrer, por exemplo, quando outra turma da mesma disciplina é analisada. Por esse motivo, o trabalho não apresentou quais atributos foram mantidos a partir da Seleção de Atributos, pois essa é uma determinação a partir do processo e dados utilizados.

Embora os modelos preditivos desenvolvidos neste estudo apresentem resultados promissores, é importante ressaltar que a predição da reprovação é apenas o primeiro passo. É fundamental que as instituições de ensino utilizem essas informações para implementar estratégias de intervenção precoce e personalizada, visando oferecer suporte adicional aos alunos em risco e aumentar suas chances de sucesso acadêmico. Para isso, é importante investir na explicabilidade dos modelos.

Embora o trabalho tenha optado por privilegiar a sensibilidade (identificar ao máximo quais alunos irão reprovar), é importante também que a previsão busque diminuir o trabalho de quem irá tentar recuperar o aluno para que esse alcance a aprovação, não atuando com alunos que já teriam a aprovação. Ou seja, a precisão também é um fator importante nesse aspecto, sendo sim importante avaliar outros conjuntos de dados para melhorar o desempenho dos modelos. Por exemplo, Baessa *et al.* (2025) utilizaram o sentimento da turma em relação à disciplina como entrada para o modelo, obtendo uma melhora de 2% na classificação de reprovações.

Sugestões para trabalhos futuros incluem a investigação de outros tipos de dados que possam contribuir para a predição da reprovação, como dados de interação do aluno com o ambiente virtual de aprendizagem, dados de participação em atividades extracurriculares e dados de bem-estar emocional. Além disso, é importante realizar estudos longitudinais para avaliar o impacto das estratégias de intervenção implementadas

com base nos modelos preditivos.

Referências

Almarabeh, H. Analysis of students' performance by using different data mining classifiers. **International Journal of Modern Education and Computer Science**, v. 9, n. 8, p. 9–15, 2017.

Baessa, R. N. F. *et al.* Predição de desistência em turmas de programação utilizando sentimentos. **RENOTE**, v. 22, n. 3, p. 377–386, jan. 2025. Disponível em: ⟨https://seer.ufrgs.br/index.php/renote/article/view/145005⟩.

Boylan-Ashraf, P. C.; Haughery, J. R. Failure rates in engineering: Does it have to do with class size? In: ASEE Annual Conference & Exposition, Salt Lake City. 2018. Disponível em: ⟨https://peer.asee.org/failure-rates-in-engineering-does-it-have-to-do-with-class-size⟩. Acesso em: 19 fev. 2025.

Cheewaprakobkit, P. Study of factor analysis affecting achievements of undergraduate. In: International Multi Conference of Engineers and Computer Scientists, Hong Kong. 2013. Disponível em: ⟨https://www.iaeng.org/publication/IMECS2013/IMECS2013_pp332-336.pdf⟩. Acesso em: 19 fev. 2025.

Cortez, P.; Silva, A. Using data mining to predict secondary school student performance. **5th Annual Future Business Technology Conference**, Porto, Portugal, 2010. Disponível em: ⟨http://www3.dsi.uminho.pt/pcortez/student.pdf⟩. Acesso em: 19 fev. 2025.

Kaensar, C.; Wongnin, W. Analysis and prediction of student performance based on moodle log data using machine learning techniques. **Journal of Emerging Technologies in Learning**, v. 18, n. 10, p. 184–203, 2023.

Kovaĉić, Z. J. Early prediction of student success: Mining students enrolment data. **Informing Science and IT Education**, jun 2010. Disponível em: ⟨https://api.semanticscholar.org/CorpusID:14191196⟩. Acesso em: 19 fev. 2025.

NCES. College dropout trends in the united states: An analysis of nces data. In: . [s.n.], 2021. Disponível em: ⟨https://nces.ed.gov/fastfacts/display.asp?id=16⟩. Acesso em: 19 fev. 2025.

Ramos, V.; Wazlawick, R.; Galimberti, M.; Freitas, M.; Mariani, A. C. A comparação da realidade mundial do ensino de programação para iniciantes com a realidade nacional: Revisão sistemática da literatura em eventos brasileiros. **Simpósio Brasileiro de Informática na Educação**, v. 26, n. 1, p. 318, 2015.

Romero, C.; Ventura, S. Educational data mining: A review of the state of the art. **Systems, Man, and Cybernetics**, v. 40, p. 601–618, 2010.

Souza, B. D. de; Moreira, B. G. Detecção de desistência de estudantes em disciplinas ofertadas com apoio do ambiente moodle: uma abordagem de análise implícita. **Anais do Computer on the Beach**, v. 15, p. 064–071, 2024.

Woodman, R. **Investigation of factors that influence student retention and success rate on Open University courses in the East Anglia region**. Dissertation (Master Thesis in Computer Science) — Sheffield Hallam University, UK, 2001.

Xu, Z.; Liu, J.; Yang, Z.; An, G.; Jia, X. The impact of feature selection on defect prediction performance: An empirical comparison. In: **2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)**. [S.l.: s.n.], 2016. p. 309–320.