

Detecção de desistência de estudantes em disciplinas ofertadas com apoio do ambiente MOODLE: uma abordagem de análise implícita

Bernardo Dalfovo de Souza
bernardo.dalfovo@gmail.com
Universidade Federal de Santa Catarina
Joinville, Santa Catarina, Brasil

Benjamin Grando Moreira
benjamin.grando@ufsc.br
Universidade Federal de Santa Catarina
Joinville, Santa Catarina, Brasil

RESUMO

School dropout is a matter of great relevance due to its significant social, economic, and personal repercussions. This study focuses on the issue of course attrition in higher education, a significant contributing factor to dropout rates. Here, we present the analysis of 216 results, obtained from combining nine different preprocessing and three sampling methods, as well as eight machine learning algorithms, based on student attendance, grades, and completion of evaluative activities, extracted from the MOODLE learning environment. The results of this study, obtained through implicit analysis, reveal a combination of methods capable of identifying student attrition with a recall of 95.31% and accuracy of 92.67% in just 25% of semester's duration.

KEYWORDS

Previsão de desistência, modelos estatísticos, análise implícita

1 INTRODUÇÃO

Durante a realização de um curso de graduação, percebe-se que diversos alunos não conseguem completar o cronograma de determinadas disciplinas por vários motivos, como falta de motivação e dificuldade em compreender o conteúdo. A não conclusão da matéria, conhecida como *desistência*, caracteriza um tipo de fracasso escolar e configura um prejuízo pessoal e profissional para o aluno, além de impactar os resultados sobre a produtividade das universidades e sociedade [12]. De acordo com Mallada [9], quatro domínios podem ser impactados no que diz respeito ao estudante:

- (1) **Aspecto Psicológico:** Esta esfera pode manifestar-se por meio de sintomas que incluem depressão, ansiedade e até mesmo a consideração de ideias suicidas;
- (2) **Aspecto Físico:** A evasão escolar também pode ter efeitos físicos adversos, tais como alterações no padrão de sono e o desenvolvimento de quadros de hipertensão, correlacionados ao estresse e à pressão associados à não conclusão dos estudos;
- (3) **Aspecto Escolar:** Estudantes que enfrentam dificuldades na conclusão de seus cursos podem desenvolver atitudes negativas em relação às tarefas acadêmicas e experimentar uma diminuição no seu rendimento acadêmico;
- (4) **Aspecto Interpessoal:** A decisão de abandonar o ensino superior pode levar à desmotivação, à redução da qualidade de vida e até mesmo ao surgimento de estados de irritabilidade, influenciando negativamente o bem-estar global do aluno.

Essas implicações enfatizam a importância de abordar a questão da evasão escolar de forma holística, considerando não apenas seus impactos acadêmicos, mas também as ramificações significativas no âmbito psicológico e social dos estudantes [9].

O Ministério da Educação (MEC) divide a definição de evasão educacional em três tipos: evasão de curso, que ocorre quando um estudante abandona um curso específico de graduação; evasão institucional, que se refere ao abandono da instituição de ensino pelo aluno; e evasão do sistema, quando o aluno deixa de participar do ensino superior como um todo [2]. Enquanto a literatura da área utiliza o conceito de evasão relativa aos cursos de graduação, ao realizar trabalhos de análise de dados para previsão da evasão, o presente estudo se concentra na análise da não conclusão de disciplinas em um curso de graduação de modalidade presencial, tratando essa situação como desistência ou abandono [10, 11, 15].

O trabalho busca identificar indicadores que possam antecipar o potencial abandono de uma disciplina, utilizando dados obtidos de atividades no ambiente virtual de ensino Modular Object-Oriented Dynamic Learning Environment (MOODLE). Os dados utilizados foram dados exportáveis pelo MOODLE (relatório de presenças, notas e conclusões de atividades) pois várias instituições optam por evitar a instalação de complementos (*plugins*) em suas instâncias do sistema. O trabalho aborda questões relacionadas às melhores práticas de avaliação, incluindo a aplicação de métricas de avaliação, e diferentes métodos de classificação desenvolvidos para abordar esse tipo específico de problema. Destacam-se dois aspectos que diferenciam este estudo em relação ao que é convencional em trabalhos similares:

- (1) **Enfoque na desistência em disciplinas:** Concentra-se na análise da desistência em disciplinas específicas, diferenciando-se de abordagens que objetivam prever a evasão de curso ou reprovação em disciplinas em geral;
- (2) **Utilização exclusiva de dados da disciplina:** Utilizam-se exclusivamente dados da disciplina sob análise, o que representa uma diferença em relação a abordagens que podem considerar informações mais amplas.

Estas características específicas ressaltam a singularidade e a relevância do presente estudo no contexto da análise da desistência em disciplinas acadêmicas.

2 FUNDAMENTAÇÃO TEÓRICA

De acordo com Pinto [14], apenas no ano de 2019 a quantidade de alunos que evadiram cursos de graduação, em universidades públicas chegou a 203.784, resultando em um custo contábil de R\$ 8.974.795.155,90 às instituições públicas de ensino. Além do prejuízo

contábil às universidades causado pela evasão, o discente também arca com um custo econômico, muitas vezes sem perceber. A autora descreve esse custo econômico como os potenciais salários renunciados pelos alunos evadidos, e resulta em uma média brasileira de R\$ 21.277,59 para cada evadido, representando cerca de um ano de trabalho, de acordo com o salário mínimo da época [14].

A Ciência de Dados Educacionais (CDE) aplica análise de dados à educação e integra habilidades técnicas e sociais para entender práticas educacionais em diferentes contextos de aprendizagem. Originária dos anos 2000, a CDE evoluiu a partir de conferências sobre Mineração de Dados Educacionais (MDE) e Analítica de Aprendizagem, fortalecida por estudos em Inteligência Artificial na Educação [7].

Ao analisar dados, diversos modelos e técnicas podem ser empregados, dividindo as formas de análise em análise exploratória, explícita e implícita. Na análise explícita, as informações desejadas são diretamente acessíveis nos dados, mas requerem operações específicas para sua extração e destaque. Técnicas de análise explícita incluem a elaboração de predicados lógicos para filtrar subconjuntos de dados, identificação e tratamento de *outliers* que podem distorcer os resultados analíticos, e estratificação para resumir intervalos de variáveis numéricas [1]. Entre os procedimentos e técnicas para auxiliar na compreensão de dados, a CDE faz uso de modelos matemáticos construídos para comunicar *insights* a educadores, alunos, gestores, designers instrucionais e outras partes interessadas [7].

Por outro lado, a análise implícita aborda objetivos como classificação, regressão, agrupamento e associação. Esses objetivos são comumente alcançados com o auxílio de técnicas de aprendizado de máquina, incluindo algoritmos como árvores de decisão e redes neurais artificiais. Essas abordagens são fundamentais para interpretar grandes volumes de dados e identificar padrões não evidentes [3, 10]. Os resultados apresentados neste trabalho são obtidos a partir de análise implícita.

2.1 Trabalhos similares

Nesta seção, serão apresentados trabalhos que trazem uma aplicação similar à proposta deste estudo. Foram selecionados trabalhos nacionais e internacionais para discussões que abordam os objetivos deste estudo, incluindo os resultados obtidos, técnicas utilizadas, métricas de avaliação dos resultados e o grau de antecipação alcançado por meio da predição.

Em Queiroga et al. [15], foi analisada a evasão em um curso de graduação de dois anos utilizando a contagem de interações dos estudantes no ambiente virtual MOODLE como a principal fonte de informação para gerar e comparar modelos analíticos. Os autores utilizaram *Bayes Net* (BN), *Simple Logistic* (SL), *Multilayer Perceptron* (MLP), *Random Forest* (RF) e J48, sendo que o modelo RF atingiu acurácia de 99% a partir da quinta semana (5% do tempo total do curso).

Burgos et al. [3] exploraram a evasão acadêmica usando uma metodologia incremental com base nas atividades dos estudantes ao longo de um semestre. Foi aplicada a regressão logística (RL) para prever a evasão e compararam-na com outras técnicas de aprendizado de máquina, encontrando uma precisão de 98,95%, sensibilidade de 96,73%, especificidade de 97,14% e acurácia de 97,13% na semana média de evasão. O modelo analisa atividades realizadas

pelos alunos e pode identificar a evasão, em média, 1,6 semanas antes de ocorrer.

No estudo de Gottardo, Kaestner e Noronha [8], o ambiente MOODLE é utilizado com o objetivo de prever desempenho dos estudantes, sem considerar a evasão ou desistência. São aplicados os algoritmos RF e MLP, sendo que o RF alcança acurácia média de 77,4%, enquanto o MLP apresenta acurácia média de 80,1%. Não são mencionadas tentativas de antecipar o desempenho acadêmico usando modelos com uma base de dados reduzida.

No estudo de Viana, Santana e Rabêlo [16] a avaliação se concentra na evasão do curso de graduação como um todo e a coleta de dados foi realizada por meio de atributos sociais e acadêmicos. RF, MLP, *Support Vector Machine* (SVM), *Árvore de Decisão* (AD), *Extra Trees* (ET), *K-Nearest Neighbors* (kNN) e *Gaussian Naive Bayes* (GNB) foram os algoritmos de classificação avaliados, sendo o melhor resultado obtido a partir de RF, com precisão média de 91,55% e sensibilidade de 92%. Não foram realizadas tentativas de antecipar a previsão.

Em Garcia et al. [4], os autores criaram um ambiente em que os professores pudessem prever a probabilidade de um aluno ser aprovado ou reprovado no início do semestre e a coleta de dados se deu a partir de informações acadêmicas passadas e pessoais. Dos algoritmos de aprendizado de máquina avaliados, que incluem *Naive Bayes* (NB), *Instance Based Learner* (IBK), JRip, J48, RF e MLP, RF teve melhor resultado, com sensibilidade de 93% e uma acurácia de 81,43% para as turmas em busca da graduação em Matemática, e uma sensibilidade de 80,5% e uma acurácia de 74,01% para o conjunto de alunos do curso de Computação.

O estudo de Manhães et al. [10] prevê a evasão de alunos em um curso de graduação de 5 anos com dados acadêmicos do primeiro semestre da graduação. Os algoritmos de classificação utilizados foram: *One Rule* (OneR), JRip, *Decision Table* (DT), *Simple Cart* (SC), J48, RF, SL, MLP, NB e BN. Embora alguns algoritmos tenham alcançado altas acurácias, como o SL com 82,29%, eles também apresentaram altas taxas de falsos positivos (36%), levando os pesquisadores a considerá-los inadequados devido ao erro crítico de classificar alunos em risco de evasão como não em risco. Algoritmos como MLP e RF, com acurácias ligeiramente inferiores (74,31% e 80,21%, respectivamente), demonstraram taxas de falsos positivos mais baixas, com 27% e 29%, respectivamente, tornando-os mais apropriados para a tarefa de identificação de alunos em risco.

A partir dos trabalhos similares é possível perceber que os problemas tratados são, em sua maioria, relacionados à evasão de curso e não à desistência de disciplina. Normalmente, leva-se em consideração o comportamento em vários semestres, bem como dados além daqueles obtidos em disciplinas, como a presença e notas. A maioria dos trabalhos avaliou os resultados não considerando apenas a acurácia, mas também a precisão e sensibilidade. O uso da precisão e sensibilidade é apropriado, uma vez que a evasão e desistência são situações que deveriam ser atípicas, e a acurácia poderia mascarar os resultados obtidos.

Além das métricas de avaliação, é importante considerar a identificação dos alunos desistentes de forma antecipada. Ou seja, a classificação não é útil se considerar todo o desempenho do estudante no semestre, sendo necessário identificar o quanto antes para agir a fim de evitar a desistência. Nesse sentido, apenas os trabalhos

de Queiroga et al. [15], Burgos et al. [3] e Garcia et al. [4] consideraram ou indicaram suas classificações levando em consideração essa avaliação da classificação antecipada.

3 MATERIAIS E MÉTODOS

Para a condução do estudo, um professor, devidamente informado sobre a realização da pesquisa, configurou o ambiente virtual do MOODLE de forma a ampliar a coleta de dados, aplicando avaliações contínuas, como questionários, atividades e fóruns, ao longo do semestre com frequência quase semanal. Ao término do período, o professor procedeu à identificação manual dos alunos que desistiram e dos que permaneceram engajados em duas turmas distintas.

A primeira turma, que será chamada de Turma 1, possui 41 alunos matriculados (dos quais dez são desistentes) e um total de 25 atividades realizadas. Esse conjunto representa uma turma da disciplina de Modelagem de Sistemas do primeiro semestre de 2023.

A segunda turma, que será referida como Turma 2, foi cursada por 86 alunos (dos quais 32 são desistentes), com um total de 24 atividades realizadas durante o semestre. Os dados dessa turma são referentes à disciplina de Programação 1, também do primeiro semestre de 2023.

A primeira turma é ofertada para os cursos de Engenharia Mecatrônica, Bacharelado em Ciência e Tecnologia, e Engenharia de Transportes e Logística e a segunda para os cursos de Engenharia Aeroespacial, Engenharia Automotiva, Engenharia Ferroviária e Metroviária, e Engenharia Civil de Infraestrutura, no Centro Tecnológico de Joinville da Universidade Federal de Santa Catarina. Ambas as disciplinas têm uma carga horária de 72 horas-aula no semestre e são ministradas em dois encontros semanais.

As análises foram baseadas em três relatórios essenciais do MOODLE: presença, notas e conclusão de atividades marcadas como importantes. A escolha desses elementos do ambiente se deu pela facilidade de acesso e porque não exigem *plugins* adicionais ou acesso ao banco de dados, evitando restrições de instituições usuárias do MOODLE. Assim, os meios propostos permitem uma aplicação mais simples e menos burocrática.

3.1 Sobre o relatório de presenças

O MOODLE oferece um *plugin* para gerenciamento de presenças¹, o que facilita o controle de frequência no sistema. A Figura 1 ilustra como esse recurso pode ser configurado, permitindo ajustes em vários elementos, como a pontuação para cada sessão de presença registrada. De maneira geral, a presença é classificada de três formas: alunos que alcançam a pontuação máxima por sessão são considerados presentes; aqueles com pontos intermediários são rotulados como justificados ou atrasados; e aqueles sem pontos são marcados como faltantes ou ausentes. No relatório exportado, cada categoria é representada por uma fração, na qual o denominador é a maior pontuação possível e o numerador indica a quantidade de pontos obtida.

A Figura 2 apresenta um segmento do relatório de presenças do MOODLE, formatado como uma planilha eletrônica. A descrição que acompanha cada tipo de registro de presença pode variar conforme a configuração estabelecida pelo professor. Contudo, no fim

¹Link para o *plugin* de presenças do MOODLE: https://moodle.org/plugins/mod_attendance

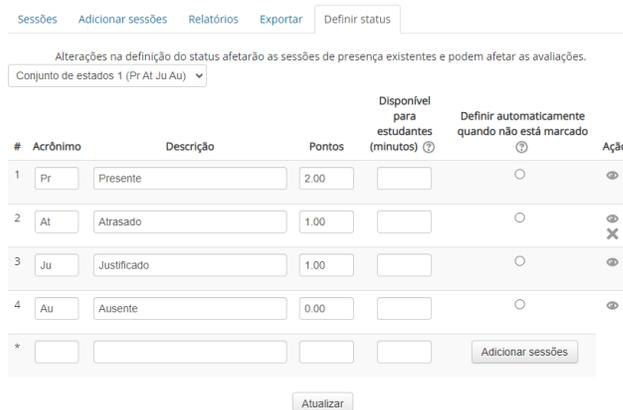


Figura 1: Configuração de presenças no *plugin* de presenças do MOODLE

de cada descrição, há sempre uma indicação entre parênteses que mostra o registro de pontos para aquele dia. Na figura exemplificada, as presenças são categorizadas como completa (2/2), justificada (1/2) ou ausência (0/2). Há também situações em que a pontuação é indicada por um ponto de interrogação, que significa que a presença ainda não foi registrada.

4	ID do Est.	Sobrenon	Nome	7/03/2023	8/03/2023
5	1	11	111	Au (0/2)	Ju (1/2)
6	2	12	112	Au (0/2)	Ju (1/2)
7	3	13	113	Pr (2/2)	Pr (2/2)

Figura 2: Parte do relatório de presenças em formato de planilha eletrônica

Para análise de dados de presença, alguns problemas foram identificados quanto aos dados/marcação dessas presenças:

- Professor pode não realizar a presença em determinada data (ou lançar as presenças obtidas de outra forma);
- O professor pode atribuir presença para a turma toda; e
- Existe a possibilidade dos alunos se auto atribuírem presenças (o professor pode habilitar essa opção), mas o professor precisa confirmar as ausências e pode não realizar tal ação.

3.2 Sobre o relatório de notas

O relatório de notas engloba avaliações individuais, categorias de avaliações e a nota global do curso, proporcionando uma visão abrangente do progresso do aluno. A Figura 3 exibe um exemplo de relatório de notas, que demonstra os variados tipos de atividades que podem ser utilizados para a atribuição de notas. Além disso, podem existir atividades registradas que, embora estejam no sistema, não tenham sido disponibilizadas ou avaliadas.

O relatório não detalha as ponderações associadas a cada atividade na composição final da nota do curso, tampouco fornece informações sobre as datas de realização das atividades ou se o aluno completou uma atividade que ainda não foi avaliada pelo professor. Ressalta-se que a ordem de colunas é definida a partir da organização realizada pelo professor, portanto, sem padrão estabelecido.

- *activities_complete_majority*: porcentagem de atividades completas pelo aluno quando a taxa de conclusão da turma na atividade é superior a 50%;
- *activities_complete_minority*: porcentagem de atividades completas pelo aluno quando a taxa de conclusão da turma na atividade é inferior a 50%;
- *activities_incomplete*: porcentagem de atividades incompletas;
- *activities_incomplete_majority*: porcentagem de atividades incompletas pelo aluno quando a taxa de conclusão da turma na atividade é superior a 50%;
- *activities_incomplete_minority*: porcentagem de atividades incompletas pelo aluno quando a taxa de conclusão da turma na atividade é inferior a 50%;
- *attendance_below_mean*: porcentagem de aulas em que a pontuação de presença do aluno foi abaixo da média de pontuações da turma;
- *missing*: porcentagem de faltas;
- *partial_presence*: porcentagem de presença parcial;
- *sequencial_missing_X*: quantidade de faltas sequenciais, compostas por X faltas maiores que um, no período.

Primeiramente, uma análise com todos os atributos foi realizada, porém, com o uso do módulo de pré-processamento, os atributos mais importantes foram selecionados com o objetivo de melhorar as métricas obtidas. A seleção dos atributos mais relevantes é detalhada na Seção 4.1.

4 RESULTADOS

A partir da disponibilização de dados de duas turmas reais, onde os alunos foram classificados manualmente pelo docente em desistentes ou não desistentes, torna-se possível calcular diferentes métricas de avaliação a partir da comparação entre os resultados previstos com a classificação real. Como não houve contato com os dados reais, a análise do classificador automático não possui viés.

Para avaliar a aplicabilidade prática do programa na prevenção da desistência escolar, as métricas foram analisadas utilizando quatro diferentes proporções do período letivo em duas turmas distintas. O uso de 100% dos dados de cada aluno (considerando todo o período do aluno na disciplina), embora não seja prático para intervenções imediatas, serve como referência para a melhor métrica possível, pois considera o desempenho completo do aluno durante o semestre. Também foram adotadas abordagens com 75% e 50% dos dados de cada aluno, que visam medir o desempenho da classificação em momentos que antes da conclusão da disciplina. Além disso, a análise com apenas 25% dos dados foi realizada para testar a capacidade do sistema de identificar sinais precoces de desistência, proporcionando uma janela de oportunidade para intervenções mais no início da disciplina.

Com o objetivo de avaliar e comparar os resultados obtidos pela análise explícita, apresentada no estudo de Souza e Moreira [6], e outros trabalhos similares, uma análise a partir dos 22 atributos apresentados na Seção 3.4 foi conduzida, apesar de que seis desses representam atividades marcadas como importantes, que não foram utilizadas para os conjuntos de dados analisados.

As turmas foram analisadas separadamente e foram utilizados três diferentes técnicas de divisão de dados para avaliação dos

resultados obtidos com os classificadores: método *holdout* com 33% dos dados para teste e 67% para treino (HO), mantendo a mesma proporção de desistentes em ambos os conjuntos; subamostragem aleatória com 67% dos dados para treinamento (SA); e validação cruzada com 10 divisões (VC). A seleção de características dos dados foi realizada com os cinco e dez melhores atributos selecionados por quatro técnicas de avaliação: análise de variância (ANOVA), chi-quadrado (Chi2), ganho de informação (GI) e relação de ganho (RG). Com relação às métricas de avaliação utilizadas, essas foram: acurácia, sensibilidade, precisão e pontuação F_1 .

4.1 Seleção dos atributos

Segundo Pimentel, Lobato e Jacob [13], a literatura dissemina diversas definições para a tarefa de seleção de características (também chamada de seleção de atributos ou, em inglês, *feature selection*), estabelecendo critérios de seleção para avaliar e encontrar subconjuntos de atributos adequados para uma análise. Para Mendes e de Jesus [5], um conjunto de dados pode apresentar um grande número de atributos, sendo a seleção de características uma abordagem para a seleção de atributos com algoritmos que detectam a relevância e correlações entre essas características. Alguns desses atributos podem ser considerados como ruídos, representando pouca influência para os algoritmos de aprendizado de máquina.

Com o uso dos métodos de avaliação de atributos, pôde-se verificar quais são as características mais importantes para a análise realizada. A Figura 6 representa os 10 atributos com pontuação mais alta para 100% dos dados em cada um dos métodos de seleção de características para a Turma 1. Adicionalmente, atributos que se repetem entre métodos foram codificados em cores. Percebe-se que *grades_between_0_2_5* é classificado entre os três mais relevantes atributos para todos os métodos e nota-se a relevância de *activities_complete_majority*, *activities_incomplete_majority* e *grades_average*, todos posicionados entre os 5 melhores em todos os casos.

Rank	Ganho de Informação	Relação de Ganho	Chi2	ANOVA
1	activities_complete_majority	activities_complete_majority	grades_between_0_2_5	activities_incomplete
2	activities_incomplete_majority	activities_incomplete_majority	activities_incomplete_majority	grades_between_0_2_5
3	grades_between_0_2_5	grades_between_0_2_5	activities_complete_majority	activities_complete_majority
4	grades_average	activities_incomplete	grades_below_5	activities_incomplete_majority
5	activities_incomplete	grades_average	grades_average	grades_average
6	grades_below_5	activities_complete_minority	grades_below_mean	grades_below_5
7	grades_between_5_7_5	activities_incomplete_minority	partial_presence	grades_below_mean
8	partial_presence	grades_below_5	grades_between_5_7_5	grades_between_7_5_10
9	grades_below_mean	partial_presence	activities_incomplete	attendance_below_mean
10	attendance_below_mean	grades_between_5_7_5	attendance_below_mean	missing

Figura 6: Seleção de atributos para a Turma 1

A categoria de atributos que mais aparece entre os 10 atributos mais importantes é a de notas, com 47,5% do total de atributos, seguido pela categoria de atividades concluídas, com 35% dos atributos selecionados, e, por fim, presença representa 17,5% das características escolhidas. Para a Turma 2, diferentemente da Turma 1, apenas *grades_between_0_2_5* é consistentemente considerado como um dos quatro atributos mais relevantes, seguido por *grades_between_7_5_10*, como pode-se perceber na Figura 7, que mostra os 10 atributos com maior pontuação em cada método

de seleção de atributos para a Turma 2, novamente codificados em cores, quando se repetem entre colunas.

Rank	Ganho de Informação	Relação de Ganho	Chi2	ANOVA
1	grades_between_7_5_10	activities_complete_minority	grades_between_0_2_5	activities_complete_majority
2	grades_between_0_2_5	activities_incomplete_minority	grades_below_mean	activities_incomplete_majority
3	grades_below_5	grades_between_7_5_10	grades_between_7_5_10	activities_incomplete
4	grades_below_mean	grades_between_0_2_5	activities_complete_majority	grades_between_0_2_5
5	activities_complete_majority	grades_below_5	grades_average	grades_average
6	activities_incomplete	activities_incomplete	activities_incomplete_majority	grades_below_5
7	activities_incomplete_majority	grades_below_mean	activities_complete_minority	grades_below_mean
8	grades_average	activities_complete_majority	grades_below_5	grades_between_7_5_10
9	activities_complete_minority	activities_incomplete_majority	activities_incomplete	activities_complete_minority
10	activities_incomplete_minority	grades_average	activities_incomplete_minority	activities_incomplete_minority

Figura 7: Seleção de atributos para a Turma 2

Nota-se a distribuição de 50% para as categorias de notas e atividades concluídas igualmente para todos os métodos, o que também significa que a categoria de presença não foi selecionada entre as 10 mais relevantes nenhuma vez para os quatro métodos de seleção. Essas diferenças mostram as particularidades de cada turma e aluno matriculado. Por serem consistentemente considerados entre os dez atributos mais relevantes em ambas as turmas, *grades_between_0_2_5*, *activities_incomplete_majority*, *activities_complete_majority*, *grades_average*, *grades_below_5* e *activities_incomplete* são apontados como preditores fortes, enquanto que *partial_presence*, *attendance_below_mean* e *grades_between_5_7_5* são específicos para a Turma 1 e *activities_incomplete_minority* e *activities_complete_minority* são específicos para a Turma 2.

4.2 Processamento dos dados

Foram empregados oito algoritmos de aprendizado de máquina que são compostos por AD, KNN, RF, NB, MLP, RL, GB e SVM. Estes algoritmos foram combinados com quatro diferentes métodos de avaliação de atributos (GI, RG, Chi2 e ANOVA) e aplicados tanto com a seleção de cinco e dez atributos quanto sem nenhuma seleção de características. Além disso, três técnicas de divisão do conjunto de dados foram utilizadas, que, ao fim da abordagem, gerou um total de 216 resultados distintos com quatro métricas (sensibilidade, precisão, acurácia e pontuação F_1) calculadas para cada.

A métrica de sensibilidade foi escolhida como o principal critério de avaliação do programa. Neste trabalho, essa métrica sinaliza alunos desistentes que são erroneamente classificados como não desistentes. Em outras palavras, ela reflete a capacidade de identificar corretamente aqueles alunos que, apesar de apresentarem sinais de desistência, poderiam não ser detectados como tal.

A partir de uma análise preliminar dos resultados obtidos a partir de HO, percebeu-se que os subconjuntos de treinamento e teste gerados por essa técnica de divisão de dados não representava precisamente o conjunto total de dados. Dessa forma, tal método foi desconsiderado das análises e as comparações utilizarão apenas SA e VC.

4.3 Resultados obtidos - métricas e discussões

A seguir, são apresentadas as principais combinações obtidas para cada divisão do banco de dados, das quais apenas as métricas obtidas

com divisão de conjunto por método *holdout* foram suprimidas. Os resultados serão apresentados primeiramente para a Turma 1 e, em seguida, para a Turma 2. Ao fim da seção, uma síntese dos principais resultados é elaborada para facilitar o entendimento.

4.3.1 Turma 1. As análises com 100%, 75%, 50% e 25% do conjunto de dados para a Turma 1 serão descritas separadamente nesta seção. Devido à quantidade de possíveis combinações, as métricas para cada proporção de dados foram primeiramente organizadas em função da sensibilidade, métrica de maior interesse do presente trabalho, e em seguida, a partir da precisão. Caso haja um empate entre duas combinações em relação à sensibilidade, a precisão é usada como critério de desempate. Caso o empate se mantenha, ambas as combinações serão apresentadas. Uma solução alternativa foi utilizada, na qual a combinação de maior acurácia é apresentada.

Para o conjunto completo de dados (100%), NB apresentou a maior sensibilidade, com **90,00%**. Suas métricas de precisão, acurácia e pontuação F_1 foram de 52,94%, 78,05% e 66,67%, respectivamente. Para atingir tais métricas, o método de avaliação utilizado foi VC sem qualquer seleção de atributos. Como uma opção alternativa, focada na acurácia, GB, a partir de VC e chi-quadrado com os cinco melhores atributos (Chi2-5) obteve métricas de 76,00%, 97,44%, **93,66%** e 85,39% para sensibilidade, precisão, acurácia e pontuação F_1 , respectivamente.

Com 75% do total de dados, a combinação escolhida com ênfase na sensibilidade foi atingida a partir de NB e SA, com chi-quadrado com os dez melhores atributos (Chi2-10) para a seleção de atributos. Suas métricas são **87,43%**, 49,16%, 75,43% e 62,93%, para sensibilidade, precisão, acurácia e pontuação F_1 , respectivamente. Como solução alternativa, priorizando a acurácia, GB com método de seleção de atributos por ganho de informação com os cinco melhores atributos (GI-5) e avaliação por VC, apresentou sensibilidade de 70,00%, precisão de 74,47%, acurácia de **86,83%** e pontuação F_1 de 72,16%.

Para uma divisão de 50% da base de dados, NB combinado a análise de variância com os dez melhores atributos (ANOVA-10) e avaliação a partir de VC obteve sensibilidade, precisão, acurácia e pontuação F_1 de **86,00%**, 51,50%, 76,83% e 64,42%, respectivamente, que foi definida como a melhor escolha para uma combinação focada em sensibilidade. Alternativamente, com a maior acurácia, LR com características selecionadas por GI-5 ou relação de ganho com os cinco melhores atributos (RG-5) e avaliado com VC obteve métricas de 64,00%, 69,57%, **84,39%** e 66,67%, na mesma ordem da solução principal.

Por fim, com 25% do total dos dados disponíveis e ao utilizar Chi2-10, NB foi escolhida por sua métrica de sensibilidade, avaliado pelo método VC. Suas métricas apresentaram valores de **88,00%**, 56,41%, 80,49% e 68,75%, para sensibilidade, precisão, acurácia e pontuação F_1 , respectivamente. Ao invés de escolher a melhor combinação a partir da sensibilidade e focar na acurácia como métrica de importância, o uso de RF com seleção de atributos do método GI-5, avaliado por VC, apresenta sensibilidade de 66,00%, precisão de 75,86%, acurácia de **86,59%** e pontuação F_1 de 70,59%. Com todas as proporções de dados analisadas para a Turma 1, partiu-se para a escolha de combinações para a Turma 2.

4.3.2 Turma 2. Da mesma forma que realizado para a Turma 1, as métricas foram analisadas para cada proporção do conjunto de

dados total para a Turma 2. Priorizou-se a métrica de sensibilidade, mas os métodos com acurácia mais alta também são apresentados como uma solução alternativa, na qual se deseja obter resultados mais corretos, do que garantir a detecção da desistência.

Inicialmente todos os dados (100% do conjunto) foram utilizados para gerar as combinações e percebeu-se que o uso de MLP com a escolha de atributos por GI-5 e avaliação por VC apresentava as maiores métricas tanto para sensibilidade quanto para a acurácia, com **100,00%** e **98,14%**, respectivamente. Adicionalmente, obteve precisão de 95,24% e pontuação F_1 de 97,56%.

Em seguida, houve a redução para 75% do total, onde a composição de NB com Chi2-10 e SA apresentaram valores de sensibilidade, precisão, acurácia e pontuação F_1 de **99,54%**, 89,23%, 95,34% e 94,10%, respectivamente. Priorizando a métrica de acurácia, SVM com análise de variância com os cinco melhores atributos (ANOVA-5) e avaliação por VC atingem sensibilidade de 98,44%, precisão de 96,04%, acurácia de **97,91%** e pontuação F_1 de 97,22%.

Ao usar metade da base de dados (50%), NB se destaca ao ser combinado com ANOVA-5 e avaliado por VC, gerando resultados com métricas de sensibilidade, precisão, acurácia e pontuação F_1 de **100,00%**, 86,96%, 94,42% e 93,02%, respectivamente. Por outro lado, a escolha de SVM com ANOVA-5 e VC, apresenta a maior acurácia, com **97,21%**, além de sensibilidade de 97,19%, precisão de 95,40% e pontuação F_1 de 96,28%.

Finalmente, com apenas 25% dos dados disponíveis, NB com seleção de características por RG-5 e avaliação por VC apresentou sensibilidade, precisão, acurácia e pontuação F_1 de **97,19%**, 86,87%, 93,49% e 91,74%, respectivamente, sendo a combinação escolhida para a solução principal. Como alternativa, focada em acurácia, MLP com RG-5 e avaliação por SA apresentou sensibilidade de 95,25%, precisão de 92,01%, acurácia de **95,28%** e pontuação F_1 de 93,60%.

4.3.3 Síntese dos resultados. A Tabela 1 apresenta a síntese dos resultados atingidos para a Turma 1 e Turma 2. Pode-se verificar a principal combinação de método de avaliação de resultados, técnica de escolha de atributos e algoritmo de aprendizado de máquina para cada proporção de dados e a métrica de sensibilidade atingida.

Inferiu-se, a partir da análise da Tabela 1, que em relação ao total de combinações:

- 87,5% usam o algoritmo de aprendizado de máquina *Naive Bayes*;
- 75% usam o método de avaliação de validação cruzada *10-fold*;
- 50% usam os 10 atributos mais bem pontuados pela técnica de seleção de atributos;
- 37,5% usam o cálculo chi-quadrado para pontuar atributos.

A fim de atender às características gerais das turmas analisadas, foi criado um modelo com os aspectos mais comuns de todas as combinações, a partir do qual realizou-se uma nova análise com as quatro proporções de dados usadas ao longo do trabalho. Dessa forma, pode-se visualizar na Tabela 2, a síntese dos resultados alcançados com o método de validação cruzada *10-fold*, a seleção de atributos Chi2-10 e o algoritmo NB para ambas as turmas.

Como analisado na Seção 2.1, o estudo realizado por Burgos et al. [3] apresenta objetivo similar ao presente trabalho, visto que pretende prever a evasão ao longo de um semestre de 20 semanas.

Os autores atingem, com seu melhor modelo, 96,73% de sensibilidade, 98,35% de precisão e 97,13% de acurácia com 50% dos dados totais. Na mesma proporção de dados, os resultados alcançados no presente estudo com um modelo geral de análise implícita atingem sensibilidade 2,65% superior, porém precisão e acurácia inferiores por 12,4% e 3,41%, respectivamente. O melhor modelo treinado com 50% dos dados do trabalho atual atinge métricas de sensibilidade 0,46% superior, precisão 3,55% inferior e acurácia 0,08% superior. Dessa forma, o presente trabalho é capaz de detectar uma maior proporção de desistentes, mas incorretamente classifica alunos não desistentes como desistentes em maior quantidade.

5 CONCLUSÕES

O presente artigo apresentou diferentes combinações de análise implícita com o objetivo de identificar a desistência de alunos em disciplinas ofertadas pelo MOODLE. Diferente de outros trabalhos similares, o trabalho se baseia apenas em dados obtidos de três relatórios simples de serem obtidos do ambiente MOODLE (relatório de presença, relatório de notas e relatório de atividades concluídas), cujo diferencial permite uma utilização simplificada, reduzindo a necessidade de se integrar dados de várias fontes e possibilita uma aplicação dos métodos³ em diferentes instituições.

A partir dos resultados apresentados na Seção 4.3, percebe-se a predominância do algoritmo de aprendizado de máquina *Naive Bayes*, ainda que não tenha os maiores índices para todas as métricas. Considera-se, para o objetivo do presente artigo, mais grave não classificar corretamente um aluno que desiste, do que incorretamente classificar um aluno não desistente. Visto que, ainda que o segundo caso possa causar uma intervenção desnecessária por parte do docente, o primeiro caso pode causar na desistência do aluno da matéria. Dessa forma, ainda que *Naive Bayes* não possua os melhores índices de precisão e acurácia em todas as instâncias, sua sensibilidade é, na maioria dos casos, superior a de outros métodos, com índices entre 86% e 90% para a Turma 1 e entre 97,19% e 100% para a Turma 2.

Na análise implícita da Turma 1, similarmente ao que ocorre para sua análise explícita apresentada por Souza e Moreira [6] entre 25% e 50% do total de dados, ao aumentar a base de dados, a métrica de sensibilidade apresenta decréscimo, de forma que as métricas com maiores valores são encontradas a partir do uso de apenas 25% dos dados. Fato que não é analisado nos resultados da Turma 2, que apresentam acréscimo a cada aumento da proporção de dados. Comparada à explícita, as métricas da implícita apresentam variações médias de -3% para a sensibilidade e 8,59% para a precisão da Turma 1. Em contraste, para a Turma 2 a sensibilidade varia em média -0,62% e a precisão 12,71%. Assim, percebe-se um aumento da precisão e diminuição da sensibilidade, na qual esta é mais acentuada para a primeira turma.

Uma conclusão que pode ser realizada, a partir da comparação dos resultados atingidos entre a Turma 1 e a Turma 2, é a possibilidade do uso de *Naive Bayes* para turmas maiores, como é o caso da Turma 2, e algoritmos mais assertivos, como os que são apresentados como combinações alternativas para cada divisão de dados nas

³Os códigos desenvolvidos para o presente trabalho podem ser encontrados em <https://github.com/bernardodalfovo/tcc>, embora os dados em si não sejam disponibilizados.

Tabela 1: Síntese de principais resultados obtidos para a Turma 1 e Turma 2

Turma	Proporção	Avaliação	Seleção de Atributos	Algoritmo	Sensibilidade
1	100%	VC	-	NB	90,00%
1	75%	SA	Chi2-10	NB	87,43%
1	50%	VC	ANOVA-10	NB	86,00%
1	25%	VC	Chi2-10	NB	88,00%
2	100%	VC	GI-5	MLP	100,00%
2	75%	SA	Chi2-10	NB	99,54%
2	50%	VC	ANOVA-5	NB	100,00%
2	25%	VC	RG-5	NB	97,19%

Tabela 2: Síntese dos resultados alcançados a partir de um modelo geral para a Turma 1 e Turma 2

Turma	Proporção de dados	Sensibilidade	Precisão	Acurácia	Pontuação F_1
1	100%	80,00%	53,33%	78,05%	64,00%
1	75%	82,00%	51,90%	77,07%	63,57%
1	50%	82,00%	48,81%	74,63%	61,19%
1	25%	87,00%	55,06%	79,51%	67,44%
2	100%	100,00%	88,89%	95,35%	94,12%
2	75%	100,00%	86,49%	94,19%	92,75%
2	50%	99,38%	85,95%	93,72%	92,17%
2	25%	95,31%	86,40%	92,67%	90,64%

subseções da Seção 4.3, no quesito acurácia e precisão para turmas menores, caso da Turma 1.

As turmas analisadas podem facilitar a detecção das desistências em relação a outros trabalhos, uma vez que analisou dados de turmas com dois encontros semanais e atividades aplicadas periodicamente (praticamente de forma semanal). Isso aumenta a quantidade de dados disponíveis para observar o comportamento dos discentes e perceber, antecipadamente, situações, como a de desistência.

Como trabalhos futuros, pode-se realizar a combinação das soluções primárias disponibilizadas na Seção 4.3, com as soluções alternativas, apresentadas em sequência da solução principal. De maneira a realizar a interseção de algoritmos com alta sensibilidade com algoritmos de alta precisão e acurácia, e, como resultado esperado, encontrar um modelo que junte os principais benefícios de cada técnica.

REFERÊNCIAS

- [1] Fernando Amaral. 2016. *Introdução à ciência de dados*. Alta Books Editora, Rio de Janeiro.
- [2] Brasil. Ministério da Educação. 1996. *Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas*. Brasil. Ministério da Educação, Brasília. <http://www.dominiopublico.gov.br/download/texto/me001613.pdf>
- [3] Concepción Burgos, María L. Campanario, David la Peña, Juan A. Lara, David Lizcano, and María A. Martínez. 2018. Data mining for modeling students' performance: a tutoring action plan to prevent academic dropout. *Computers & Electrical Engineering* 66 (Feb. 2018), 541–556.
- [4] Léo Manoel Lopes da Silva Garcia, Daiany Francisca Lara, Raquel Salcedo Gomes, and Sílvio Cezar Cazella. 2022. Mineração de Dados Educacionais na Predição do Desempenho Acadêmico: um prognóstico a partir do percurso curricular realizado. In: XXXIII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. *Anais [...]*. (2022). <https://doi.org/10.5753/sbie.2022.225221>
- [5] Rudson Franciso da Silva Mendes and João Victor Ribeiro de Jesus. 2021. Exploração de modelos de aprendizado de máquina e seleção de atributos para Employee Attrition. *Anais do Computer on the Beach* 12 (2021), 267–272.
- [6] Bernardo Dalfovo de Souza and Benjamin Grando Moreira. 2024. Detecção de desistência de estudantes em disciplinas ofertadas com apoio do ambiente Moodle: uma abordagem de análise explícita. *Revista Ibérica de Sistemas e Tecnologias de Informação* E66 (Feb. 2024), 487–500.
- [7] Andrea Cristina Filatro. 2020. *Data science na educação*. Saraiva Educação S.A., São Paulo.
- [8] Ernani Gottardo, Celso Antônio Alves Kaestner, and Robinson Vida Noronha. 2014. Estimativa de desempenho acadêmico de estudantes: análise da aplicação de técnicas de mineração de dados em cursos a distância. *Revista Brasileira de Informática na Educação* 22, 1 (2014), 45–56.
- [9] Francisco Javier Ribaya Mallada. 2011. La gestión del absentismo escolar. *Anuario Jurídico y Económico Escurialense* 44 (2011), 579–596.
- [10] Laci Mary Barbosa Manhães, Sérgio Manuel Serra Cruz, Raimundo J. Macário Costa, Jorge Zavaleta, and Geraldo Zimbrão. 2011. Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In: XXII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. *Anais [...]*. (2011). <http://ojs.sector3.com.br/index.php/sbie/article/download/1585/1350>
- [11] Alana Marques Morais. 2018. *Abordagem avaliativa multidimensional para previsão da evasão do discente em cursos online*. Tese (Doutorado em Ciência da Computação). Centro de Engenharia Elétrica e Informática, Universidade Federal de Campina Grande.
- [12] Nathália Prochnow Nagai and André Luís Jankovski Cardoso. 2017. A evasão universitária: uma análise além dos números. *Revista Estudo & Debate* 24, 1 (2017), 193–215.
- [13] Cláudia Patrícia S Pimentel, Fábio Lobato, and Antonio Jacob Jr. 2023. Aplicação de técnicas de aprendizagem de máquina com seleção de variáveis na previsão de receitas públicas de 8 capitais. *Anais do Computer on the Beach* 14 (2023), 264–271.
- [14] Silvana Correa Pinto. 2021. *Os custos da evasão de discentes das universidades brasileiras na modalidade de ensino presencial: uma perspectiva de custos contábeis e custos econômicos*. Dissertação (Mestrado em Ciências Contábeis). Unidade Acadêmica de Pesquisa e Pós-Graduação, Universidade do Vale do Rio dos Sinos.
- [15] Emanuel M. Queiroga, João L. Lopes, Ricardo M. Araujo, and Cristian Cechinel. 2018. Modelo de predição da evasão de estudantes em cursos técnicos a distância a partir da contagem de interações. *Revista Thema* 15, 2 (2018), 425–438.
- [16] Filipe Soares Viana, André Macêdo Santana, and Ricardo de Andrade Lira Rabêlo. 2022. Avaliação de Classificadores para Predição de Evasão no Ensino Superior Utilizando Janela Semestral. In: XXXIII SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO. *Anais [...]*. (2022). <https://doi.org/10.5753/sbie.2022.224764>