CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

Mineração de Dados Educacionais para Predição da Evasão em Cursos de Graduação Presenciais no Ensino Superior

Gustavo Zanini Kantorski[1], Ricardo Zimmermann Martins[1], Arthur Balejo[1], Marcio Frick[1]

[1]Universidade Federal de Santa Maria (UFSM)
97.105-900 – Santa Maria – RS – Brazil

gustavo@ufsm.br, ricardo.zimmermann@acad.ufsm.br,
arthur.balejo@acad.ufsm.br, marciofrick@ufsm.br

Abstract. *Student dropout in undergraduate's programs is an important question for academic management. This paper describes an analysis of several classification machine learning metodhs for the student predict dropout on several undergraduate's programs. We propose several machine learning methods for prediction combined to data science techniques and a business intelligence tool. Furthermore, we present experiments in 106 undergratuades programs about three periods, reaching many knowledge areas.*

Resumo. *A evasão de alunos nos cursos de graduação tem sido uma das questões relevantes para a gestão acadêmica das Instituições de Ensino Superior. Este artigo apresenta uma análise de vários métodos de aprendizagem de máquina para a predição da evasão em cursos de graduação presenciais. Várias estratégias de aprendizagem de máquina são propostas combinadas com técnicas de ciência de dados e uma ferramenta de business intelligence. Os experimentos foram realizados com os dados de 106 cursos de graduação durante três semestres, envolvendo diversas áreas de conhecimento.*

1. Introdução

A evasão no ensino superior é uma questão em constante monitoramento pelas instituições de ensino e, após a pandemia de Covid-19, as taxas de evasão tiveram um aumento, principalmente na rede pública, a qual foi mais afetada devido a suspensão das aulas presenciais [Honorato and Borges 2022]. Além disso, o abandono de estudantes em um curso de graduação implica em prejuízo para si próprio, ao não se diplomar, para o docente, que não atinge sua meta como educador, para a instituição, pelo não atendimento de sua missão e para a sociedade, pelas perdas sociais e econômicas [Silva Filho et al. 2007]. Outro fator que pode influenciar a evasão são as formas de ingresso na universidade, pois a ampliação da oferta de vagas resulta em processos seletivos que ofereçam maior ou menor facilidade de acesso, o que pode desenvolver o ingresso de alunos não confiantes sobre suas escolhas profissionais.

Analisar informações estudantis com a finalidade de estimar a possiblidade de evasão de um estudante é essencial para mitigar o processo de evasão dos cursos. Nesse contexto, a utilização de Mineração de Dados Educacionais (MDE) para classificar estudantes com maior risco de evasão é fundamental. MDE caracteriza-se pela utilização de mineração de dados ou descoberta de conhecimento em bases de dados aplicados no campo educacional com o objetivo de extrair informações significativas, padrões e

relacionamentos entre variáveis armazenadas em grande conjuntos de dados educacionais [Agrusti et al. 2019].

Este artigo apresenta um processo de análise e previsão da evasão por meio de métodos de MDE e algoritmos de aprendizado supervisionado. Assim, foram construídos diversos modelos para classificar os estudandes com maior ou menor risco de abandono dos cursos de graduação presenciais da instituição. A metodologia apresentada descreve uma solução que contempla desde a extração dos dados brutos, a análise e participação de especialistas, perpassando pela utilização de técnicas de aprendizagem de máquina até a produção de *insights*, determinação de fatores de evasão, identificação de estudantes propensos ao abandono e, finalmente, otimizando a visualização dessas informações para apoiarem os gestores no combate à evasão.

O artigo está organizado da seguinte maneira. Na Seção 2 são apresentados os trabalhos relacionados. A Seção 3 apresenta a metodologia utilizada para a realização do trabalho. A Seção 4 discute os experimentos realizados e os resultados alcançados. Finalmente, na Seção 5 são apresentadas as considerações finais e direções futuras.

2. Trabalhos Relacionados

Existem trabalhos que utilizam MDE com a finalidade compreender os fatores que causam a evasão [Sales et al. 2019, Carrano et al. 2019, Alyahyan and Düştegör 2020, Marques et al. 2020, Carneiro et al. 2022]. Entre os vários fatores encontrados nesses trabalhos destacam-se a falta de motivação dos alunos e professores, problemas pessoais e socio econômicos, insatisfação com o curso ou instituição, problemas de aprendizagem associados com metodologias de ensino e processos de avaliação, restrições do mercado de trabalho, desconhecimento prévio do curso sobre os alunos, repetências contínuas em disciplinas do curso, nível de estudo anterior a matrícula, entre outros. No trabalho de [Carneiro et al. 2022] é apresentada uma proposta para identificar quais fatores podem contribuir para a diminuição da evasão. Esses fatores são extraídos da análise das regras de aprendizagem utilizadas pelos algoritmos.

Outros trabalhos abordam a previsão de determinadas situações do processo educacional. Por exemplo, trabalhos que tratam a previsão de desempenho de estudantes [Silva et al. 2022, Costa et al. 2015], trajetórias [Êrica Carmo et al. 2022], previsão somente com dados acadêmicos e situação no curso [Santos et al. 2021] predição de reprovação em disciplinas [de Jesus et al. 2021]. Ainda, encontram-se trabalhos que auxiliam na identificação de estudantes propensos ao abandono do curso [Kantorski et al. 2016, Hortêncio Filho et al. 2020, Alyahyan and Düştegör 2020, Colpo et al. 2021, Viana et al. 2022]. Geralmente, esses trabalhos utilizam técnicas de descoberta de conhecimento em bases de dados ou mineração de dados para realizar a previsão de resultados.

Em [de Jesus et al. 2021] os autores propõem uma previsão de alunos com risco de evasão. O trabalho apresenta disciplinas que possuem mais reprovações e a predição da evasão é realizada com multilayer perceptron no curso de licenciatura em computação. Em [Hortêncio Filho et al. 2020] são utilizadas vários algoritmos para a análise da evasão, *Naïve Bayes*, árvores de decisão, *support vector machines* e vizinho mais próximo (kNN). O trabalho de [Viana et al. 2022] apresenta uma proposta de classificação de alunos entre Evadidos e Graduados para cursos de Computação e Sistemas de Informação. O tra-

balho de [Êrica Carmo et al. 2022] analisa trajetórias dos alunos evadidos e concluintes em disciplinas do curso de ciência da computação e identifica que ambos os grupos (evadidos e concluintes) percorrem as disciplinas do curso de maneira diversa, não obedecendo à matriz curricular proposta. Os trabalhos de [Alyahyan and Düştegör 2020] e [Colpo et al. 2021] abordam revisão sistemática de literatura e uso de melhores práticas encontradas em trabalhos de previsão por meio de MDE.

3. Metodologia

A metodologia proposta utiliza os conceitos de aprendizagem de máquina para determinar os estudante com maior risco de evasão. O processo de previsão foi adaptado da metodologia *CRoss Industry Standard Process for Data Mining* (CRISP-DM) [Chapman et al. 2000]. A metodologia para construção do modelo segue um ciclo virtuoso de retroalimentação buscando a melhoria contínua e o aprimoramento do processo. Para ter uma melhor compreensão da metodologia, ela foi dividida em várias etapas. A metodologia proposta compreende seis etapas: 1.entendimento da evasão, 2.compreensão dos dados, 3.preparação dos dados, 4.modelagem, 5.avaliação e 6.aplicação (disponibilização). Na Figura 1 é apresentado um resumo da metodologia desenvolvida nesse trabalho.

A primeira etapa é o entendimento da evasão e tem como finalidade identificar os objetivos sob o ponto de vista de descoberta de conhecimento em bases de dados. Esses objetivos incluem o tipo de tarefa a ser executada, por exemplo, aprendizado supervisionado, classificação e os critérios de avaliação dos modelos utilizados.

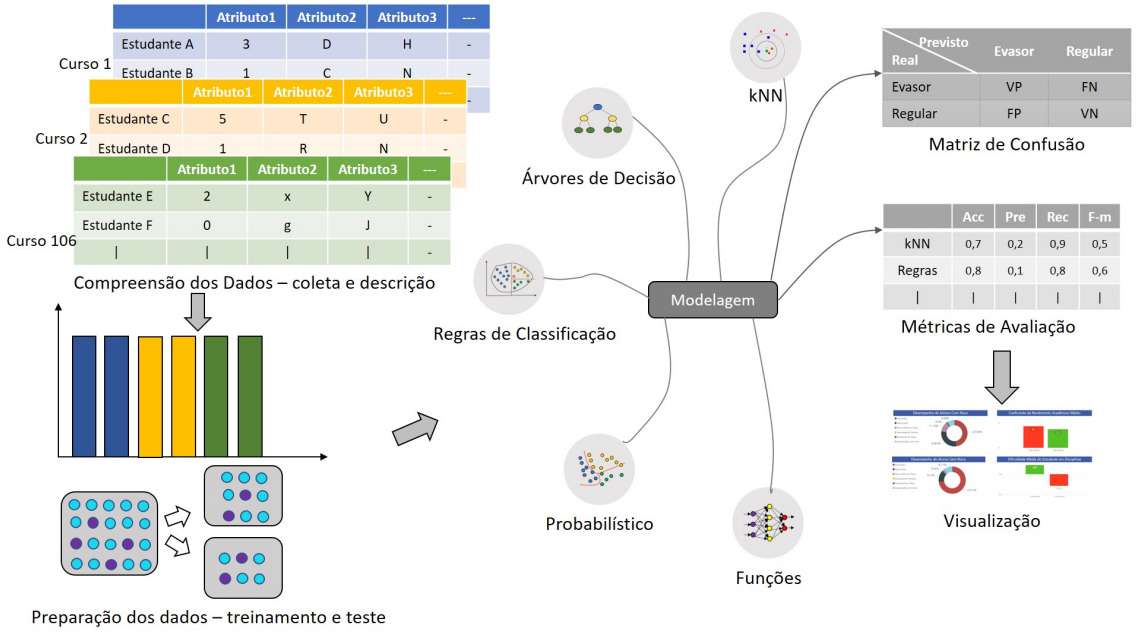

Figura 1. Diagrama da metodologia proposta

Na etapa 2, Compreensão dos dados, é realizada a coleta inicial dos dados e a descrição e exploração dos dados verificando suas propriedades e qualidade. O processo de exploração consiste na análise do dados propriamente dita e na utilização de técnicas de visualização. A descrição trata da avaliação de propriedades dos dados, tais

como, faixa de valores, números de atributos, significado de cada atributo, necessidade de normalização e a sua importância para o processo de descoberta. Nessa etapa foram coletados dados de 106 cursos de graduação presenciais, totalizando mais de 79.000 alunos de graduação entre egressos, abandonos e estudantes regularmente matriculados para o primeiro semestre de 2022.

A próxima etapa é a preparação dos dados. Nessa etapa um conjunto de dados é gerado em conformidade com os modelos de descoberta de conhecimento que serão utilizados. A geração desse conjunto envolve uma série de passos incluindo a seleção, limpeza, construção, integração e formatação dos dados. A seleção dos dados e das instâncias pode ser realizada de maneira manual ou por meio de algoritmos. A limpeza tem como finalidade melhorar a qualidade dos dados. Dentre as atividades de limpeza pode-se citar a eliminação de dados com erro, a padronização de dados, por exemplo, abreviaturas, formatos de data, valores de atributos. A construção dos dados concentra-se na normalização, se for o caso, na transformação de valores simbólicos para numéricos e na discretização dos atributos. A integração serve para combinar múltiplas fontes de dados a formatação pode realizar alterações sintáticas nos dados sem modificar o seu significado. Foram coletados 59 atributos dos cursos, sendo que após o processo de preparação, dois conjuntos de atributos foram selecionados para mineração. Um conjunto de 33 atributos baseado em seleção manual e um conjunto variável de atributos baseado no algoritmo de seleção de atributos proposto por [Hall 1998]. O conjunto variável é um subconjunto do conjunto de atributos manual (33 atributos). O conjunto de atributos é variável pois conjuntos diferentes são gerados para cursos diferentes. Como a análise foi realizada em 106 cursos de graduação tem-se a possibilidade de ter até 106 conjuntos de atributos diferentes. Na Tabela 1 são apresentados os atributos selecionados manualmente para a previsão dos cursos para o ano de 2022, primeiro semestre.

Tabela 1. Atributos da seleção manual considerados para a previsão

Atributo	Tipo	Atributo	Tipo
Período Ingresso no curso (1/2)	Categórico	Total Trancamentos Totais	Numérico
Faixa Etária	Categórico	Total Trancamentos Parciais	Numérico
Estado Civil	Categórico	Total de atividades	Numérico
Tempo no Curso (em anos)	Numérico	Média das notas das disciplinas	Numérico
Período Atual (semestre)	Numérico	Desempenho acadêmico	Categórico
Posição no Semestre	Categórico	Média de disciplinas cursadas por período	Numérico
Cotista	Booleano	Bolsa BSE	Booleano
Refeições no RU	Numérico	Bairro que reside (distância do campus)	Categórico
Uso Biblioteca (retirada obras)	Numérico	Dificuldade média do estudante	Numérico
Bolsista	Booleano	Coeficiente de rendimento acadêmico	Numérico
Morador Casa do Estudante	Booleano	Média de disciplinas aprovadas	Numérico
Monitoria	Booleano	Média de disciplinas reprovadas	Numérico
Formação Ensino Médio (pública/privada)	Categórico	Media de disciplina reprovadas por frequência	Numérico
Total Disciplinas Aprovadas	Numérico	Media de disciplinas dispensadas	Numérico
Total Disciplinas Reprovadas	Numérico	Número de projetos que participa(ou)	Numérico
Total Disciplinas Reprovadas por Frequência	Numérico	Classe (Evasão/Regular)	
Total Disciplinas Dispensadas	Numérico		

A etapa 4, Modelagem, é a descoberta de conhecimento propriamente dita. As

informações das etapas anteriores servem de insumo para esta etapa. Com a finalidade de preparar essas informações para a descoberta de conhecimento, a análise foi categorizada em estudantes egressos (formados e não formados) e estudantes regularmente matriculados. Estudantes egressos são aqueles que ingressaram e concluiram o seu curso ou abandonaram no decorrer do período. Estudantes regularmente matriculados são aqueles que no momento da análise estão matriculados e cursando alguma atividade curricular. Os atributos disponíveis na análise dos estudantes são organizados para a criação de modelos de aprendizagem de máquina para classificação. A análise preditiva envolve dois conjuntos de dados: *i.* treinamento e *ii.* teste. O conjunto de treinamento engloba os dados dos alunos egressos enquanto que o conjunto de teste compreende os dados dos alunos matriculados. Nessa etapa vários algoritmos são aplicados e várias composições de dados são construídas. Isso acontece porque nessa etapa os conjuntos de treinamento e teste necessitam de ajustes para se adaptarem a cada algoritmo de aprendizagem.

A etapa de avaliação interpreta e avalia os resultados em relação aos objetivos definidos na fase de entendimento da evasão. A avaliação compreende verificar a acurácia do processo de previsão. A avaliação determina a taxa de acerto da evasão e identifica quais as variáveis que influenciaram na previsão. A saída dessa etapa é utilizada como entrada na etapa seguinte e assim o processo entre em um ciclo virtuoso de retroalimentação em busca de constante melhoria contínua.

Finalmente, a etapa 6, aplicação, é responsável pela transformação dos resultados das etapas anteriores em referências visuais que permitam a geração de *insights* sobre esses resultados. Essa etapa permite que, por meio da visualização, seja possível identificar padrões, conexões e descobertas que não seriam percebidas. Dessa forma é possível, por exemplo, entender o impacto de atributos no processo de evasão, agregar valor na gestão do processo, mostrar relações entre variáveis e com isso tomar decisões mais acertadas. A aplicação desenvolvida nessa etapa pode ser acessada em `ufsm.br/cpd/integra`.

4. Avaliação Experimental

A metodologia apresentada na Seção 3 foi aplicada a 106 cursos de graduação presencial da Instituição em três semestres: 2022/1, 2022/2 e 2023/1. Para a avaliação experimental serão apresentados os resultados para o período 2022/1. A premissa utilizada na realização dos experimentos foi a possibilidade de ter um acerto alto das evasões. Essa premissa é importante pois implica diretamente no cálculo de avaliação dos modelos de aprendizagem. Para os 106 cursos analisados foram utilizados 06 algoritmos de aprendizagem de máquina: C4.5 [Quinlan 1993], CART [Breiman et al.], RIPPER [Cohen 1995], Naïve Bayes [John and Langley 2013], Multilayer perceptron [Bishop et al. 1995] e kNN [Aha et al. 1991]. O conjunto de treinamento total, englobando todos os cursos, possui 61.644 instâncias considerando dados a partir de 2010 até o semestre anterior à análise. O conjunto de teste é formado por 15.367 instâncias. Considerando os seis algoritmos de aprendizagem, um total de 241 modelos foram gerados para os 106 cursos analisados, gerando uma média de dois (2,27) modelos por curso. Os algoritmos foram aplicados para todos os 106 cursos de graduação com o conjunto de dados apresentado na Tabela 1. Além disso, foram aplicados no conjunto de atributos da Tabela 1, o algoritmo de seleção de atributos proposto por [Hall 1998]. Assim, existem subconjuntos variáveis de atributos para diferentes cursos.

Para realizar a avaliação, os algoritmos foram divididos em dois grupos. O pri-

meiro grupo contém os algoritmos baseados em árvores de decisão (C4.5 e CART) e regras de classificação (Ripper) e o segundo grupo composto pelos algoritmos baseados em funções (multilayer perceptron), probabilísticos (Naïve Bayes) e *lazy* (kNN - vizinho mais próximo). Essa divisão foi realizada porque o conjunto de atributos aplicado para cada grupo foi diferente. Para o primeiro grupo foi utilizado o conjunto de atributos da Tabela 1. Para o segundo grupo foi utilizado o conjunto de atributos gerado pelo algoritmo de seleção de atributos proposto em [Hall 1998], o qual é distinto para cada curso.

Os algoritmos baseados de árvores de decisão e regras de classificação utilizaram o conjunto de 33 atributos (Tabela 1). Esses algoritmos eliminam atributos fortemente relacionados no momento da divisão dos atributos. Portanto, os modelos de classificação desses algoritmos não contém todos os atributos da Tabela 1, mas um subconjunto deles. Nos experimentos realizados 142 modelos foram gerados com esses tipos de algoritmos para cursos diferentes, sendo que 75 cursos foram classificados em algoritmos de árvores de decisão, 14 cursos em algoritmos de regras de classificação e 17 cursos com algoritmos nas duas categorias. Os dez atributos mais frequentes para essas categorias aparecem na Tabela 2. De maneira geral, independente da categoria (árvores de decisão ou regras de decisão), oito dos dez atributos são comuns aos algoritmos. Ainda, nove dos dez atributos dos algoritmos de regras de classificação são dados acadêmicos (exceção do atributo *idade de ingresso no curso*) e oito dos dez atributos dos algoritmos de árvores de decisão são, igualmente, dados acadêmicos, com a exceção de *bolsista* e *idade de ingresso no curso*.

Tabela 2. Atributos mais frequentes nos modelos de árvores de decisão e regras de classificação

Seq	Atributo mais frequente para algoritmos de árvores de decisão	Número de Modelos	Atributo mais frequente para algoritmos de regras de classificação	Número de Modelos
1	Total Disciplinas Aprovadas	98	Total Disciplinas Aprovadas	49
2	Total Disciplinas Dispensadas	45	Total Disciplinas Dispensadas	26
3	Total Trancamentos Totais	37	Total Trancamentos Totais	26
4	Coeficiente de rendimento acadêmico	28	Dificuldade média do estudante	23
5	Total de atividades	22	Total de atividades	23
6	Dificuldade média do estudante	20	Coeficiente de rendimento acadêmico	21
7	Média de disciplinas aprovadas	20	Média de disciplinas aprovadas	21
8	Bolsista	17	Média das notas das disciplinas	14
9	Media de disciplinas dispensadas	16	Media de disciplinas dispensadas	13
10	Idade de ingresso no curso	15	Idade de ingresso no curso	10

Considerando os algoritmos baseados em funções, probabilísticos e *lazy* (kNN - vizinho mais próximo) o conjunto de atributos aplicados é sempre o mesmo para o mesmo curso conforme o algoritmo de seleção de atributos [Hall 1998]. Para essas categorias foram construídos 99 modelos de aprendizagem contemplando 81 cursos, sendo que 44 cursos foram classificados somente com o algoritmo Naïve Bayes, 16 cursos somente com o Mulilayer Perceptron e três com o kNN. Nenhum curso foi classificado utilizando os três algoritmos dessas categorias. A avaliação dos atributos que foram selecionados também destaca a maioria de atributos acadêmicos, tais como, disciplinas aprovadas, trancamen-

tos, coeficiente de rendimento acadêmico, entre outros.

No contexto de todas as categorias de algoritmos, as combinações que mais foram utilizadas em cursos foram algoritmos de árvores de decisão e probabilísticos (31 cursos), seguidos dealgoritmos de árvores de decisão somente (15 cursos), árvores de decisão e baseados em função (11 cursos), regras de classificação e árvores de decisão (10 cursos) e regras de classificação e probabilístico (10 cursos). Destaca-se também a utilização de três categorias - árvores de decisão, regras de classificação e probabilísticos (09 cursos).

A avaliação dos resultados da previsão foi realizada por meio das métricas de acurácia, precisão, *recall* e *f-measure*. Essas métricas foram utilizadas para avaliar a qualidade da previsão. A acurácia mede o total de acerto comparado com o total de instâncias avaliadas. A precisão avalia a fração relevante comparada com o total previsto. O *recall* avalia a fração que foi prevista comparada com o total relevante. A métrica *f-measure* combina a precisão e o *recall* em uma só medida. Seja R uma relação de todas as matrículas previstas. Seja S um conjunto de todas as matrículas que evadiram durante o processo de matrícula. As definições de precisão (P), *recall* (R) e *f-measure* (F) são dadas por: $P = \frac{|R \bigcap S|}{|R|}$, $R = \frac{|R \bigcap S|}{|S|}$, $F = \frac{(\beta^2+1)\times P \times R}{(\beta^2 \times P)+R}$, respectivamente. O valor de β é utilizado par dar ênfase maior a precisão ou *recall* ou para manter a mesma ênfase para as duas medidas. Se o valor β for igual 1, a mesma ênfase é dada a P e R. Se o valor de β for igual a 2, o *recall* e enfatizada duas vezes em relação à Precisão. Já, se o valor de β for igual a 0.5, a Precisão possui ênfase 2 vezes em relação ao *recall*.

Considerando as métricas de avaliação de modelos de aprendizagem, precisão e recall, o ideal seria ter alta precisão e alto *recall*. Nesse trabalho, a preferência é que se alcance um *recall* mais alto em compensação à precisão. Na prática, isso significa que o número de alunos previstos como maior risco seja grande (ou seja, precisão menor) mas que o número de acertos de evasão seja o maior possível (maior *recall*). Nos 106 cursos analisados, foi alcançada uma acurácia 53% com uma precisão de 2% e um *recall* de 88%. O total de alunos analisados foi 15.367 alunos matriculados em 106 cursos de graduação sendo que 7.410 foram previstos com risco de evasão. A instituição teve uma evasão de 210 alunos sendo que 184 estavam na previsão realizada. É importante salientar que esse número faz a interseção dos diversos algoritmos de aprendizagem para determinar se a matrícula tem risco de evasão. Por exemplo, se em determinado curso foram aplicados dois algoritmos de aprendizagem, nos dois algoritmos a mesma matrícula deve ter sido prevista com risco de evasão para que a matrícula seja considerada como provável evasão.

Uma análise realizada durante os experimentos foi a avaliação de cada algoritmo utilizado na previsão. A Tabela 3 apresenta os resultados encontrados. Na Tabela 3 é possível verificar que a precisão em todos os métodos é praticamente a mesma. Já o *recall* tem um valor máximo de 97% no algoritmo de árvore de decisão e o menor valor é de 87% para o algoritmo probabilístico Naïve Bayes. A média aritmética geral de *recall* foi de 92%. Ainda, a principal questão que precisa ser refinada em trabalhos futuros é o número de falsos positivos. O número de falsos positivos tem um custo considerado baixo do ponto de vista da avaliação. Isso acontece porque todo aluno tem um risco de evasão. Na prática, os estudantes classificados como prováveis evasores (futuros verdadeiros positivos ou falsos positivos) possuem características ou propriedades que os algoritmos de aprendizagem detectaram que podem levar a evasão. Essa é a razão pela qual tem-se uma precisão baixa no processo de aprendizagem. Assim, o custo de erros de classificação

causados por falsos positivos é considerado baixo, o que implica em um custo total de modelo menor do que o custo se o peso de cada rótulo de classificação fosse o mesmo. Outra informação interessante é a frequência que o algoritmo de aprendizagem é utilizado no curso. Na análise dos 106 cursos de graduação, o algoritmo que foi utilizado em mais cursos foi árvore de decisão-C4.5 - 78 cursos.

Tabela 3. Avaliação média de algoritmos aplicados na análise da evasão

Algoritmo	Cursos	VP	FP	VN	FN	Acurácia	Precisão	Recall
Árvore de Decisão -C4.5	78	143	6.821	4.242	4	0,39	0,02	0,97
Árvore de Decisão - CART	33	56	2.999	2.522	5	0,46	0,02	0,92
Multilayer Perceptron	32	53	2.843	2.747	7	0,50	0,02	0,88
Naïve Bayes	57	100	3.809	3.269	14	0,47	0,03	0,87
RIPPER	31	59	2.531	2.051	3	0,45	0,02	0,95
Vizinho mais próximo	10	19	969	648	1	0,40	0,02	0,95

Outro aspecto analisado foi a previsão de cada curso. Nessa visão temos pontos interessantes. O curso que atingiu a melhor F-measure foi o curso de Engenharia Mecânica. Nesse curso foi alcançada uma precisão de 11% e um *recall* de 100%, com a f-measure de 0,36. A Tabela 4 mostra os primeiros 10 cursos ordenados pela melhor f-measure. Em todos eles, o recall foi de 100% e a precisão acima da média geral.

Tabela 4. Top 10 Cursos com melhores valores de f-measure

Curso	VP	FP	VN	FN	Acurácia	Precisão	Recall	f-measure
Engenharia Mecânica/CS	10	84	45	0	0,39	0,11	1	0,36
Relações Públicas/FW	3	37	49	1	0,57	0,075	0,75	0,25
Tecnologia em Eletrônica Industrial	3	47	23	0	0,35	0,06	1	0,24
Letras - Espanhol	5	80	29	0	0,29	0,058	1	0,21
Educação Especial - Diurno	5	87	54	0	0,4	0,054	1	0,21
Engenharia de Telecomunicações	4	75	29	0	0,3	0,05	1	0,21
Jornalismo/FW	4	75	66	0	0,48	0,05	1	0,21
Tecnologia em Processos Químicos	3	59	26	0	0,32	0,048	1	0,17
Engenharia Sanitária e Ambiental	4	86	27	0	0,26	0,044	1	0,17

5. Conclusões

Neste trabalho foi apresentado o processo de análise e previsão da evasão em 106 cursos de graduação presenciais de uma instituição federal de ensino superior. Durante a análise da evasão foram aplicados seis algoritmos de aprendizagem de máquina, totalizando 241 modelos gerados para os 106 cursos. Os objetivos propostos para previsão da evasão foram alcançados com um recall médio de 92%. Isso mostra que a identificação de possíveis evasores é factível e pertinente em uma análise global dos cursos da instituição.

No decorrer do trabalho o foco foi não somente em acertar as evasões, mas identificar estudantes que possuem o perfil similar de alunos que abandonaram o curso. Essa informação está refletida nos números de falsos positivos. Esses casos são alunos em que processos de prevenção possam ser iniciados para evitar a evasão futura. Intervenções pedagógicas ou de outra natureza podem ser desencadeadas com esse grupo de estudantes para mitigar o abandono. Portanto, ter uma métrica de precisão baixa, ou seja, um número alto de falsos positivos abre a possibilidade de identificação de potenciais evasões futuras. O acompanhamento contínuo do estudante é fundamental para controlar a evasão. Aquele

estudante que, atualmente, foi classificado como falso positivo em um próximo período de matrícula pode ser um verdadeiro positivo.

Trabalhos futuros podem ser direcionados no aumento da métrica de precisão, ou seja, na disponibilização de uma lista menor de prováveis evasores para análise pela coordenação de curso, no intuito de reduzir o número de falsos positivos. Além disso, outros algoritmos de aprendizagem de máquina podem ser aplicados e avaliados para melhorar os resultados. Direções futuras também podem ser aplicadas na análise temporal contínua de cada período de matrícula e, na relação de dependência entre variáveis que frequentemente ocorrem juntas nos dados, além de ações de combate à evasão que podem ser recomendadas baseadas nas análises dos algoritmos.

Referências

Agrusti, F., Bonavolontà, G., and Mezzini, M. (2019). University dropout prediction through educational data mining techniques: A systematic review. *Journal of e-learning and knowledge society*, 15(3):161–182.

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, 6:37–66.

Alyahyan, E. and Düştegör, D. (2020). Predicting academic success in higher education: literature review and best practices. *International Journal of Educational Technology in Higher Education*, 17:1–21.

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford Univ. press.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. Classification and regression trees (cart). 1984. *Belmont, CA, USA: Wadsworth International Group*.

Carneiro, M. G., Dutra, B. L., Paiva, J. G. S., Gabriel, P. H. R., and Araújo, R. D. (2022). Educational data mining to support identification and prevention of academic retention and dropout: a case study in introductory programming. *Revista Brasileira de Informática na Educação*, 30:379–395.

Carrano, D., de Albergaria, E. T., Infante, C., and Rocha, L. (2019). Combinando técnicas de mineração de dados para melhorar a detecção de indicadores de evasão universitária. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 30, page 1321.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R., et al. (2000). Crisp-dm 1.0: Step-by-step data mining guide. *SPSS inc*, 9(13):1–73.

Cohen, W. W. (1995). Repeated incremental pruning to produce error reduction. In *Machine Learning Proceedings of the Twelfth International Conference ML95*.

Colpo, M., Primo, T., and Aguiar, M. (2021). Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 873–884, Porto Alegre, RS, Brasil. SBC.

Costa, F., dos Santos Silva, A. R., de Brito, D. M., and do Rêgo, T. G. (2015). Predição de sucesso de estudantes cotistas utilizando algoritmos de classificação. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 26, page 997.

de Jesus, H. O., Rodriguez, L. C., and Costa Junior, A. d. O. (2021). Predição de evasão escolar na licenciatura em computação. *Revista Brasileira de Informática na Educação*, 29:255–272.

Hall, M. A. (1998). Correlation-based feature subset selection for machine learning. *Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato*.

Honorato, G. and Borges, E. (2022). Impactos da pandemia da covid-19 para o ensino superior no brasil e experiências docentes e discentes com o ensino remoto. In *Revista Desigualdade Diversidade*, volume 22, pages 137–179. PUC-Rio.

Hortêncio Filho, F. W. B., Vinuto, T. S., and Leal, B. C. (2020). Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1132–1141. SBC.

John, G. H. and Langley, P. (2013). Estimating continuous distributions in bayesian classifiers. *arXiv preprint arXiv:1302.4964*.

Kantorski, G., Flores, E. G., Schmitt, J., Hoffmann, I., and Barbosa, F. (2016). Predição da evasão em cursos de graduação em instituições públicas. In *Brazilian Symposium on Computers in Education (SBIE)*, volume 27, page 906.

Marques, L. T., Marques, B. T., Rocha, R. S., Chaves, L., de Castro, A. F., Queiroz, P. G. G., et al. (2020). Evasão acadêmica e suas causas em cursos de bacharelado em ciência da computação: Um estudo de caso na ufersa. In *Anais do xxxi simpósio brasileiro de informática na educação*, pages 1042–1051. SBC.

Quinlan, J. R. (1993). Program for machine learning. *C4. 5*.

Sales, F., Mendes, Y., Dembogurski, B., Semaan, G., Silva, E., and Ferreira, F. (2019). Evasão no ensino básico da rede pública municipal de juiz de fora: uma análise com mineração de dados. In *Brazilian Symposium on Computers in Education (SBIE)*, volume 30, page 1371.

Santos, C. H., Martins, S., and Plastino, A. (2021). É possível prever evasão com base apenas no desempenho acadêmico? In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 792–802, Porto Alegre, RS, Brasil. SBC.

Silva, B., Pimentel, E., and Botelho, W. (2022). Predição de desempenho de estudantes: Uma revisão sistemática de literatura. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 1040–1052, Porto Alegre, RS, Brasil. SBC.

Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., and Lobo, M. B. d. C. M. (2007). A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, 37:641–659.

Viana, F. S., Santana, A. M., and Rabêlo, R. d. A. L. (2022). Avaliaçao de classificadores para prediçao de evasao no ensino superior utilizando janela semestral. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 908–919. SBC.

Êrica Carmo, Gasparini, I., and Oliveira, E. (2022). Identificação de trajetórias de aprendizagem em um curso de graduação e sua relação com a evasão escolar. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 323–333, Porto Alegre, RS, Brasil. SBC.