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

Análise de dados pré-universidade para predizer a evasão de alunos ingressantes em uma instituição de ensino superior

Adair Perdomo Falcão [1], **Rosangela Villwock** [1], **Simone Aparecida Miloca** [1]

[1]Universidade Estadual do Oeste do Paraná – Unioeste
R. Universitária, 1619 - Universitário, Cascavel - PR - Brasil, CEP: 85819-110

adair.falcao, rosangela.villwock, simone.miloca@unioeste.br

Abstract. *This study presents a pre-university data classification model to predict the dropout of new students in undergraduate courses at a Brazilian Federal Institution of Higher Education. The development of effective policies and actions to reduce dropout rates is a constant challenge faced by university administrators. Early identification of these students enables the planning and implementation of more effective preventive measures. For this study, a dataset containing information from 1,086 students enrolled in 6 different courses from 2014 to 2018 was used. The obtained model showed a true positive rate of 87%, demonstrating a promising ability in the prior identification of students prone to dropping out.*

Resumo. *Este estudo apresenta um modelo de classificação de dados pré-universidade para prever a evasão de estudantes ingressantes nos cursos de graduação de uma Instituição Federal de Ensino Superior. O desenvolvimento de políticas e ações eficazes para reduzir o abandono é um desafio constante enfrentado pelos gestores universitários. A identificação precoce desses alunos possibilita o planejamento e a implementação de ações preventivas mais eficazes. Para este estudo foi utilizado um conjunto de dados contendo informações de 1.086 alunos matriculados em 6 cursos distintos de 2014 a 2018. O modelo obtido apresentou uma taxa de verdadeiro positivo de 87%, demonstrando uma capacidade promissora na identificação prévia de alunos propensos a evadir.*

1. Introdução

A evasão é uma das principais dificuldades enfrentadas pelas instituições de ensino superior (IES) [da Silva Soares 2009]. Ela é influenciada por diversos fatores, abrangendo aspectos econômicos, vocacionais e institucionais [Barosso and Falcão 2004]. Essa problemática envolve um processo complexo [Manhães 2015], resultante das circunstâncias individuais do aluno, das demandas da sociedade e das características da própria instituição de ensino.

Muitas instituições enfrentam dificuldades para desenvolver estratégias eficazes para reduzir a evasão [Manhães 2015]. Nesse contexto, surgiram diversos estudos aplicando técnicas de Mineração de Dados com o objetivo de identificar padrões e modelos capazes de detectar alunos com potencial para evadir [Colpo et al. 2020]. Quando essas técnicas são aplicadas a dados educacionais, o processo é conhecido como Mineração de Dados Educacionais (MDE) [Goldschmidt et al. 2015].

Considerando que a taxa de evasão é maior no primeiro ano e que a maioria ocorre até o terceiro período [Silva Filho et al. 2007, Davok and Bernard 2016, Oliveira 2015], o momento de prever a evasão ganha importância, uma vez que prever o risco de abandono o mais cedo possível pode permitir o planejamento e execução de ações preventivas com maior eficácia. Dessa forma, a utilização exclusiva de dados pré-universidade para prever a evasão logo no início da graduação pode ser um caminho viável para reduzir esse problema.

Este trabalho tem como objetivo construir e avaliar um modelo de classificação de dados pré-universitários para prever a evasão de alunos ingressantes nos cursos de graduação de uma instituição pública de ensino superior. Com isso, visa fornecer informações para os setores de assessoria estudantil atuarem de forma preventiva na tarefa de evitar a evasão.

2. Referencial teórico e trabalhos correlatos

A Mineração de Dados é a etapa central de um processo maior chamado Descoberta de Conhecimento em Base de Dados (do inglês *Knowledge Discovery in Databases* – KDD). Uma das definições mais aceitas foi dada por Fayyad (1996, p.40), "KDD é um processo não trivial de identificar padrões válidos, novos, potencialmente úteis e compreensíveis a partir de um conjunto de dados".

Esses padrões podem ser classificados em dois tipos: preditivos, que são padrões construídos para prever os valores de uma ou mais variáveis, e descritivos, que têm o objetivo de explicar e apresentar informações [Han and Kamber 2012].

Em termos práticos, as etapas operacionais da KDD podem ser sumarizadas em três [Goldschmidt et al. 2015]:

- Pré-processamento: É uma etapa fundamental, ela compreende a obtenção, organização e tratamento dos dados para a aplicação dos algoritmos de mineração de dados.
- Mineração de Dados: Onde ocorre a definição das tarefas e são utilizados algoritmos para buscar e identificar os padrões. As tarefas de KDD são as operações realizadas dentro da etapa de mineração de dados.
- Pós-processamento: Ocorre o tratamento das informações e do conhecimento obtido com a finalidade de avaliar utilidade e viabilizar o uso do conhecimento descoberto.

Dentro do contexto da evasão a Classificação é a tarefa de KDD mais utilizada e vem apresentando bons resultados na predição da evasão [Costa et al. 2013, Santos et al. 2021, Alban and Mauricio 2019, Agrusti et al. 2019].

Existem vários estudos com comparações de algoritmos classificadores: em Oliveira et al. (2015) foram testados os algoritmos: J48, *Multilayer Perceptron*, SMO, IBk e *Naive Bayes* no estudo da correlação da evasão de cursos de graduação com o empréstimo de livros em biblioteca. Nos dois experimentos realizados o algoritmo J48 de indução de árvore de decisão obteve a maior acurácia, com valores de 86,61% e 79,84%.

No estudo exploratório sobre o uso de classificadores para a predição de desempenho e abandono de Motta (2016, p.126), os algoritmos baseados em redes Bayesianas e

regressão logística apresentaram vantagem nos índices de acurácia. Ainda assim, o autor conclui que tais vantagens "não compensam a transparência oferecida pelas árvores de decisão".

Segundo Han e Kamber (2012) os classificadores baseados em árvore de decisão são tão populares devido à facilidade de uso. Além disso, a transparência faz com que as árvores de decisão sejam amplamente utilizadas nos processos de MDE, já que permitem melhor compreensão e interpretação dos modelos gerados. Isso é especialmente relevante na MDE, pois os dados são oriundos de um contexto complexo e rico em detalhes, em que cada informação pode ser útil na compreensão dos indivíduos envolvidos no processo de ensino-aprendizado e na compreensão do próprio processo.

A transparência deste tipo de modelo se dá pela representação na forma de uma árvore de estruturas condicionais, onde cada nó interno realiza um teste condicional com um atributo de predição e cada nó terminal (folha) possui um rótulo de classe (atributo alvo). Após definidos os parâmetros do modelo, a árvore de decisão classifica um conjunto de atributos de acordo com o caminho que satisfaz as condições, desde o nó-raiz até a folha e no final do processo esse conjunto é rotulado de acordo com o atributo (classe) da folha [Costa et al. 2013].

Um exemplo amplamente utilizado de algoritmo indutor de árvores de decisão que tem demonstrado bons resultados é o J48 [Colpo et al. 2020, Santos et al. 2021]. O J48 é uma implementação em Java do algoritmo C4.5, originalmente proposto por Quinlan em 1993 [Quinlan 1993]. Essa implementação do J48 inclui recursos adicionais, como a capacidade de lidar com valores ausentes e suporte a outras formas de poda.

3. Materiais e método

Considerando os objetivos deste estudo, trata-se de uma pesquisa exploratória e explicativa, segundo Severino (2014).

Para o processo de KDD, seguiu-se a metodologia CRISP-DM (do inglês, Cross Industry Standard Process for Data Mining). Essa metodologia apresenta um ciclo de vida recursivo com o objetivo de conduzir todo o processo repetidas vezes até gerar um modelo que apresente os resultados esperados ou aceitáveis [Goldschmidt et al. 2015]. Esse processo envolve as seguintes etapas:

- Compreensão do negócio: Tem como objetivo conhecer o contexto em que o processo de KDD será realizado, incluindo pessoas envolvidas, necessidades, recursos e condições para o sucesso.
- Compreensão dos dados: Estudo cuidadoso dos dados, avaliando qualidade, quantidade, significado e relevância dos atributos.
- Preparação dos dados: Compreende as ações de pré-processamento dos dados, incluindo seleção, limpeza e codificação para uso nos algoritmos de mineração de dados.
- Modelagem: Aplicação de algoritmos de mineração de dados, realização de testes para ajuste de parâmetros e obtenção do modelo. Pode ser necessário retornar à fase de preparação de dados.
- Avaliação: Análise quantitativa e qualitativa dos resultados, com possíveis sugestões para revisão nos processos anteriores.

- Desenvolvimento/Implantação: Planejamento e execução de ações a partir dos modelos obtidos.

3.1. Compreensão e preparação dos dados

Os dados utilizados neste trabalho referem-se aos alunos ingressantes nos cursos de Química, Física, Ciências Biológicas, Medicina Veterinária, Nutrição e Letras Português Espanhol.

Foram considerados os registros dos alunos matriculados nos anos de 2014 a 2018. A maior parte desses dados foi obtida a partir do Sistema de Gestão Acadêmica da instituição. Além disso, foram coletadas manualmente as notas do ensino médio dos alunos nas disciplinas de Matemática, Português, Química, Física, Biologia, História e Geografia.

Também foram solicitados dados do Exame Nacional do Ensino Médio (ENEM), incluindo as notas obtidas, assim como informações do questionário socioeconômico. No entanto esses dados não foram disponibilizados pela instituição.

Na preparação dos dados foram utilizadas as estratégias de segmentação do conjunto de dados, eliminação direta de atributos, redução de valores nominais, exclusão de registros, preenchimento com valores constantes e construção de novos atributos [Han and Kamber 2012, Goldschmidt et al. 2015, Bramer 2016].

A base final ficou com 1086 registros, divididos entre 566 (52%) pertencentes à classe "Sim" (evadidos) e 520 (48%) à classe "Nao" (não evadidos), apresentando um bom balanceamento entre as classes. Cada registro composto por 24 atributos, além da classe "Evadido". O Quadro 3.1 apresenta a lista de atributos e uma breve descrição de cada um deles.

3.2. Técnica, algoritmo e ferramenta

Trabalhos correlatos mostraram que a transparência dos modelos baseados em árvores de decisão superam, na maioria das vezes, as possíveis vantagens de outros métodos, especialmente dentro do contexto estudado. Deste modo, para gerar o modelo optou-se pelo algoritmo indutor de árvore de decisão J48.

A implementação do J48 está disponível na ferramenta WEKA, do projeto Waikato, neste trabalho foi utilizada a versão WEKA 3.8.6. Os resultados promissores e a ampla adoção desse algoritmo e dessa ferramenta, conforme indicado nas revisões sistemáticas realizadas por Colpo et al. (2020) e Santos et al. (2021), justificam sua aplicação neste trabalho.

Existem vários parâmetros que podem ser ajustados no algoritmo J48, alguns que alteram significativamente a estrutura final do modelo e a forma de interpretá-lo. Em termos de estrutura, o principal é o *binarySplits*. Os outros dois parâmetros mais relevantes estão relacionados aos métodos de poda *reducedErrorPruning* e *unpruned*. Considerando esses dois métodos, mais o método padrão do algoritmo, existem 3 métodos principais para a poda nesta implementação.

Outro parâmetro de grande importância, que atua diretamente no desempenho do modelo, é o *minNumObj*. Ele é utilizado para equilibrar o modelo entre a generalização e a especificidade, impactando diretamente na redução do sobreajuste (*overfitting*).

Quadro 3.1 – Atributos da base de dados

Nome do atributo	Tipo	Descrição
Ano de ingresso	Numérico	Ano de ingresso do aluno no curso
Curso	Categórico	Nome do curso
Turno	Categórico	Turno do curso
Idade no Ingresso	Numérico	Idade do aluno no primeiro dia do ano de ingresso no curso
Sexo	Categórico	Gênero autodeclarado
Raca	Categórico	Raça autodeclarada
Nacionalidade	Categórico	Nacionalidade
Origem	Categórico	País de Origem
Reg nasc BR	Categórico	Região do Brasil em que o aluno nasceu
Reg. Mor. Est.	Categórico	Região de moradia no Estado da instituição
Reg Moradia BR	Categórico	Região do Brasil em que o aluno mora
Nec Especial	Categórico	Tipo de necessidade especial
Ano Conc EM	Numérico	Ano de conclusão do ensino médio
Anos entre EM e ES	Numérico	Diferença entre o ano de conclusão do ensino médio e o ano de ingresso
Escola Publica	Booleano	Se o aluno é provindo de escola pública
Grupo no PS	Categórico	Grupo de inscrição no processo seletivo (cotas)
Modo Ingresso	Categórico	Forma de ingresso (Transferência, ENEM, Processo Seletivo, etc)
COEF BIO	Numérico	Média aritmética das notas de Biologia do ensino médio
COEF FIS	Numérico	Média aritmética das notas de Física do ensino médio
COEF GEO	Numérico	Média aritmética das notas de Geografia do ensino médio
COEF HIST	Numérico	Média aritmética das notas de História do ensino médio
COEF PORT	Numérico	Média aritmética das notas de Língua Portuguesa do ensino médio
COEF MAT	Numérico	Média aritmética das notas de Matemática do ensino médio
COEF QUI	Numérico	Média aritmética das notas de Química do ensino médio
Evadido	Categórico	Classe que indica se o aluno evadiu (Sim) ou não (Não)

Fonte: De autoria própria.

No contexto da evasão, os valores de falso negativo são de grande importância, pois quando ocorrem, o modelo indica que um aluno não irá evadir, mas ele evade. Reduzir esses valores deve ser prioridade. O Weka permite o uso de um metaclassificador chamado *Cost-Sensitive Classifier* (classificação sensível ao custo) para induzir em um classificador, como o J48, uma forma de aprendizado que leva em consideração os custos de classificações incorretas.

Para definir os melhores parâmetros, foram realizados vários experimentos utilizando validação cruzada com 5 *folds*. Os critérios mínimos adotados para a seleção do modelo foram: acurácia mínima de 70% e árvore resultante com pelo menos 3 níveis de altura. Além disso, foi avaliada a interpretabilidade das previsões do modelo e a identificação de variáveis importantes. Avaliando assim, se as previsões do modelo são úteis para os objetivos educacionais e se ela permite recomendar ações práticas e relevantes.

Inicialmente, foram conduzidos experimentos variando os métodos de poda. Para cada método, foram testados valores do parâmetro *minNumObj* de 1 a 40. Além disso, para cada valor do *minNumObj*, o peso do metaclassificador foi variado de 1 a 3, com incrementos de 0,5. No caso dos testes com o *reducedErrorPruning*, a variável *seed* foi mantida constante, enquanto o *numFolds* variou unitariamente de 2 a 5.

Após definido o método de poda, o valor para parâmetro *minNumObj* e o peso, foram realizados experimentos com os demais parâmetros. Para o *confidenceFactor*, os valores variaram de 0,5 a 0,50, com incrementos de 0,05.

Para aumentar a confiança nos resultados obtidos, o modelo gerado foi avaliado utilizando a ferramenta *Experimenter* do Weka. Por meio dessa ferramenta, o modelo foi testado em 50 execuções com o uso de validação cruzada com 5 *folds*, com variação da

semente de criação dos *folds* a cada repetição. A acurácia desse teste foi avaliada por meio do teste T-Pareado, com um nível de significância de 0,05.

O modelo final foi gerado utilizando o metaclassificador com um peso de 1,5 para Falso Negativo. Os parâmetros atribuídos ao classificador foram: *useLaplace = true, ConfidenceFactor = 0,25 e minNumObj = 6*.

O foco deste trabalho vai além da acurácia, buscando gerar modelos que possam fornecer informações novas e potencialmente úteis para auxiliar na identificação das causas da evasão e na tomada de decisões. Além disso, sendo o foco a evasão, foi dada prioridade à taxa de identificação dos alunos com tal potencial, tentando minimizar o impacto na taxa de identificação dos alunos que não evadem.

Portanto, a avaliação dos resultados obtidos não deve se basear apenas na acurácia geral do modelo. Considerando os objetivos deste trabalho, a avaliação das taxas de verdadeiro positivo e falso negativo ganha relevância, uma vez que indicam a eficácia do modelo em identificar corretamente alunos propensos à evasão.

4. Resultados e Discussão

O modelo de árvore de decisão resultante da metodologia aplicada é mostrado na Figura 1. As elipses representam nós de decisão, onde o texto interno indica o atributo de predição. A elipse inicial com o atributo ”Turno”é a raiz da árvore, quanto mais próximo da raiz mais importante (que melhor caracteriza esse conjunto de dados) é considerado o atributo. Os nós internos realizam testes condicionais que levam ao crescimento da árvore em ramos ligados a outros nós de decisão ou a folhas (nós terminais). As folhas são representadas por retângulos, cada uma com um rótulo de classe: ”Sim”para ”Evadido”e ”Não”para ”Não Evadido”. O primeiro número em cada folha é o total de instâncias ponderadas pelo peso do metaclassificador, e o segundo número é o peso das instâncias classificadas incorretamente. A árvore classifica registros/alunos ao seguir o caminho que satisfaz as condições da raiz à folha, rotulando o conjunto com o atributo (classe) da folha no final do processo.

Os resultados estatísticos do modelo estão disponíveis no Quadro 4.1. Observa-se que a acurácia geral ficou no limite do aceitável, girando em torno de 70%, confirmado pelo teste T-Pareado. Destaca-se que o valor de 87% da taxa de verdadeiro positivo pode ser considerado promissor, indicando uma boa capacidade de classificar corretamente alunos com propensão à evasão.

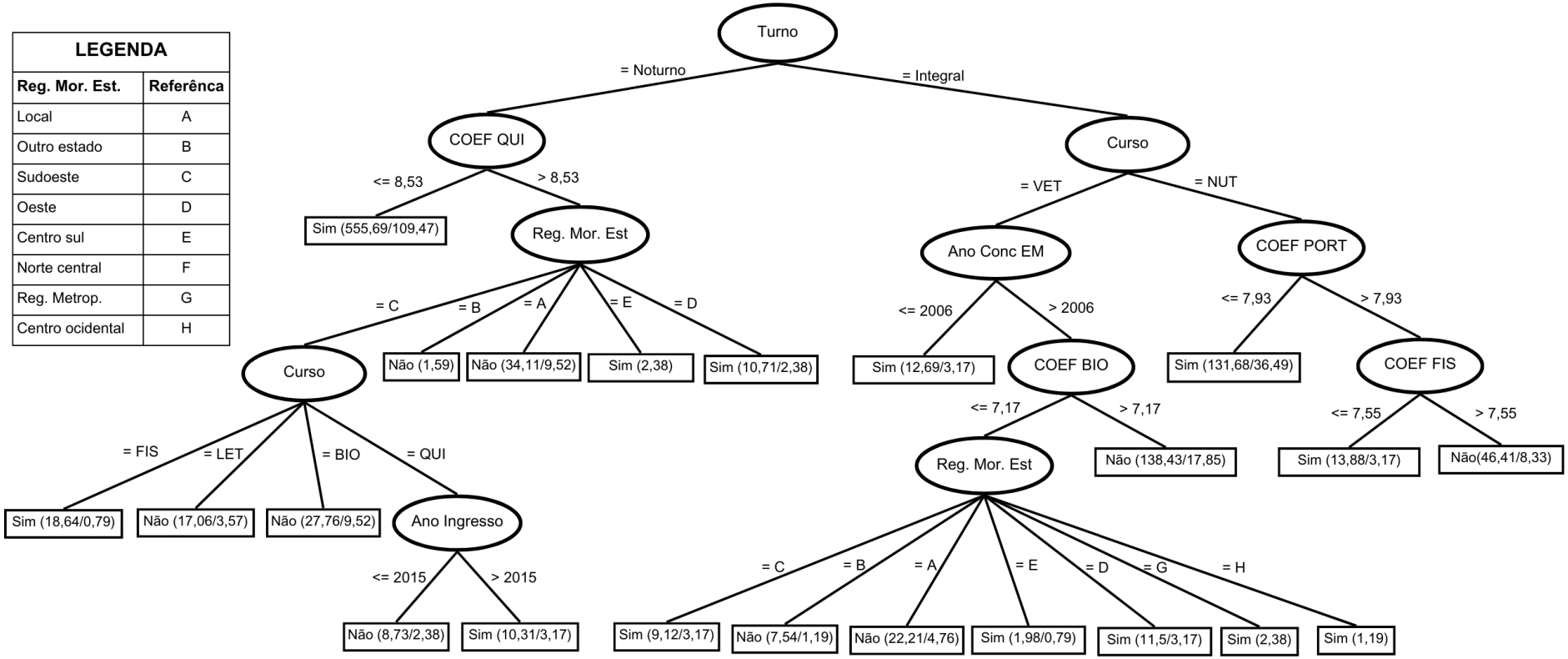
Quadro 4.1 – Estatísticas do modelo obtido calculadas pelo WEKA

Resultados estatísticos do modelo				Matriz de confusão		Valor Predito	
Classe	Sim	Não	Média Pond.			Sim	Não
Taxa VP	0,871	0,563	0,724	Valor Real	Sim	493	73
Taxa FP	0,437	0,129	0,289				
Precisão	0,685	0,801	0,740		Não	227	293
Sensibilidade	0,871	0,563	0,724				
Medida-F	0,767	0,661	0,716	Acurácia geral		72,38%	
ROC área	0,760	0,760	0,760	T-Pareado		70,43% (2.43)	

Fonte: De autoria própria.

Ao analisar a matriz de confusão, observa-se que 73 alunos foram indicados como

Figura 1. Modelo resultante na forma de árvore de decisão



Fonte: De autoria própria.

falso negativo (que o modelo errou ao predizer que não evadiria). Levando em conta que foram utilizados dados de 5 anos e de 6 cursos diferentes, distribuindo ingenuamente esses alunos pelo total de turmas nesse período, obtemos um valor aproximado de 2,4 alunos identificados incorretamente como não evadidos por turma (em um dado curso e ano). Dadas as várias situações que podem levar um aluno a evadir, bem como as limitações decorrentes do uso exclusivo de dados pré-universidade e da ausência dos dados do ENEM, tal valor pode ser considerado aceitável.

Porém, a taxa de verdadeiro negativo ficou próxima de 57%, abaixo do esperado. No entanto, dentro do contexto deste trabalho, esse resultado não é necessariamente ruim. Por um lado, essa baixa taxa significa que a instituição terá que lidar com um número maior de alunos nas intervenções para reduzir a evasão, exigindo um esforço adicional. Por outro lado, é importante lembrar que nem todos os alunos que não evadem concluem o curso sem reprovações, ou seja, alunos com perfil de evasão mas que superam as dificuldades e reprovações com empenho e dedicação. Nesse sentido, o modelo pode ser benéfico ao incluir alunos com potencial de retenção entre os evadidos, permitindo que esses alunos também sejam considerados nas medidas para redução da evasão. Isso pode resultar em um menor número de retenções e, consequentemente, em um maior número de formandos dentro do tempo previsto pelas normativas do curso

Avaliando as outras métricas, a precisão ficou um pouco abaixo para a classe "Sim", o que era esperado ao atribuir um peso maior para reduzir o índice de Falso Negativo. Por outro lado, os valores obtidos para a área abaixo da curva ROC (0,760) podem ser considerados bons, indicando que o modelo tem um bom potencial como classificador.

Na perspectiva qualitativa, do especialista de domínio, o modelo gerado apresenta algumas características interessantes. A raiz da árvore começa distinguindo o turno do curso como o atributo mais importante. Essa escolha reflete a maior evasão dos cursos noturnos e reforça o que é observado ao acompanhar o cotidiano desses alunos no ensino superior. Nesse acompanhamento é comum encontrar aqueles que apresentam dificuldades para conciliar seus estudos com outros compromissos, como trabalho, cuidar da família ou acesso ao Campus.

Considerando os cursos do turno noturno, nota-se um reflexo imediato do impacto do coeficiente de química na evasão. Isso pode ser atribuído aos cursos ofertados nesse período (Física, Química, Ciências Biológicas e Letras), uma vez que três desses cursos apresentam características semelhantes entre si e uma relevante importância de conhecimentos fundamentais de Química.

Além disso, a região de moradia desses alunos se mostrou relevante, indicando que alunos que residem localmente ou em outros estados têm uma menor tendência de evasão. Por outro lado, alunos que moram nas regiões oeste e centro-sul, próximas da instituição, apresentam uma maior tendência de evasão.

Ao analisar essa questão, os alunos que indicaram moradia em outro estado devem ter uma residência temporária no local, pois seria inviável o percurso diário para frequentar as aulas. Assim, eles podem ser considerados entre os moradores locais. Nesse caso, é evidente que a moradia local apresenta um impacto importante na vida acadêmica do aluno. Isso fica mais evidente ao comparar com alunos que possuem residência em regiões próximas (oeste e centro-sul). É razoável supor que esses alunos enfrentam diariamente

grandes deslocamentos até a universidade, o que gera fadiga e consome tempo e dinheiro, impactando seu desempenho acadêmico e contribuindo para o maior índice de evasão desses alunos.

Já os alunos que residem na região sudoeste (região onde se situa a instituição) ainda enfrentam o problema do deslocamento, mas não a ponto de ser tão decisivo para a evasão. Nesse caso, a situação depende do curso em que o aluno está matriculado. Enquanto alunos de Letras e Biologia não apresentam maior tendência a evadir nessas circunstâncias, no caso dos alunos de Física, essas dificuldades contribuem para o aumento da evasão. Além disso, para o curso de Química, a moradia não é o único fator. Aqui, também se observa a influência do ano de ingresso do aluno.

Ao considerar os cursos oferecidos no período integral, pelo modelo, nota-se que a primeira distinção ocorre entre os cursos de Nutrição e Medicina Veterinária, indicando que o perfil do aluno evasor é diferente entre eles. No caso do curso de Nutrição, as notas de Português e Física se mostraram indicadores importantes para a evasão. A relação entre o conhecimento em língua portuguesa e o desempenho acadêmico é mais clara e também é apontada por outros estudos, como o de Cabral e Tavares (2005) .

Para o curso de Medicina Veterinária, o primeiro fator considerado foi o ano de conclusão do ensino médio. Nesse caso, alunos que concluíram o EM há mais tempo e são mais velhos apresentam uma maior tendência à evasão. Isso pode ser justificado pelo tempo decorrido entre o ensino médio e o ensino superior, indicando que esses alunos podem ter enfrentado dificuldades ou indecisões durante esse período. Indica também que esses alunos podem apresentar uma maior dificuldade de adaptação ao ritmo e as demandas da universidade.

Em seguida, o coeficiente de biologia foi considerado um fator relevante, o que faz sentido, uma vez que a área de Ciências Biológicas está diretamente associada à maioria das disciplinas do curso de Medicina Veterinária. No entanto, apenas a nota de Biologia não se mostrou como um fator determinante para a evasão desses alunos. A região de moradia também foi considerada relevante para a evasão, o que segue uma avaliação similar a que foi feita para os alunos do curso noturno.

5. Conclusão

Alguns atributos do modelo se mostraram mais relevantes, como a região de moradia, o coeficiente de disciplinas específicas, o turno e o ano de conclusão do ensino médio.

A região de moradia destacou-se como um fator importante na maioria dos casos. No contexto analisado, a construção de moradia estudantil apresenta grande potencial para impactar positivamente na vida dos alunos e contribuir para a redução dos índices de evasão em todos os cursos. Além disso, a criação de áreas de lazer e descanso e a expansão de programas de suporte específicos, como auxílio transporte podem ajudar na mitigação dessas situações.

Em relação aos coeficientes das disciplinas, o modelo indicou que o coeficiente de química é um importante indicador de evasão na maioria dos cursos do período noturno. Outros coeficientes, como português e biologia, apresentaram relevância para os cursos de nutrição e veterinária, respectivamente.

Nesse sentido, as coordenações dos cursos podem identificar os ingressantes com

baixos coeficientes de disciplinas específicas e fornecer intervenções acadêmicas personalizadas, como tutorias, programas de reforço ou aconselhamento acadêmico individualizado, para ajudar os estudantes a superar dificuldades em áreas específicas do currículo. Trabalhos como o de Passos et al. (2001) e Simão et al. (2008) mostram resultados positivos e indicam, além de aspectos teóricos, caminhos práticos para a aplicação dessas formas de intervenção.

Observou-se ainda que os alunos do período noturno têm maior tendência à evasão. Em geral, esses alunos enfrentam dificuldades para conciliar os estudos com outros compromissos, como trabalho, responsabilidades familiares ou dificuldade de acesso ao campus. Permitir que parte da carga horária dos cursos seja ministrada na modalidade EAD pode ser benéfica, ajudando-os a conciliar seus compromissos com a graduação.

O ano de conclusão do ensino médio também está associado ao tempo entre o ensino médio e o ensino superior, indicando que alunos que passaram mais tempo entre essas etapas apresentam maior propensão à evasão. Esses alunos tendem a apresentar dificuldades em acompanhar o ritmo e as demandas acadêmicas da graduação.

Identificados os alunos propensos à evasão, a instituição pode oferecer programas de acolhimento e adaptação para ajudá-los a se integrarem ao ambiente educacional. Isso inclui apoio acadêmico, orientação profissional e apoio social para garantir que se sintam acolhidos e integrados à comunidade universitária. O apoio acadêmico pode ser oferecido pelos monitores, pedagogos e técnicos em assuntos educacionais. A orientação profissional e apoio social podem ser realizados pelos psicólogos e assistente sociais, já presentes no quadro da instituição.

Enfim, são muitas as possibilidades de melhorias e intervenções que podem ser implementadas para reduzir os índices de evasão. Vale destacar que antes de implementar qualquer medida, o aluno propenso à evasão deve ser identificado. Quanto mais cedo essa identificação ocorrer, maiores serão as possibilidades de intervenção e de ter sucesso em manter o aluno na instituição.

Neste sentido, este trabalho apresentou um modelo de previsão de evasão utilizando apenas dados pré-universidade, que demonstrou uma boa capacidade de identificar os alunos evadidos desde o momento de ingresso na instituição. Isso comprova o potencial positivo do uso de MDE para identificar alunos propensos à evasão, auxiliando, dessa forma, na implementação de estratégias eficazes que podem contribuir para a melhoria da experiência acadêmica dos alunos, aumentando suas chances de serem bem-sucedidos e concluírem seus cursos.

É importante ressaltar que este estudo, embora promissor, não abordou uma variedade de características, especialmente os aspectos socioeconômicos, que podem influenciar positivamente os resultados [da S. Freitas et al. 2020]. Além disso, pela melhor interpretabilidade do modelo, foi abordado apenas um algoritmo de classificação. Portanto, são necessárias mais pesquisas que explorem essa abordagem em diferentes conjuntos de dados, considerando a aplicação de múltiplos algoritmos e técnicas.

Como trabalho futuro, pretende-se ampliar a base de dados e explorar o potencial de outros algoritmos, incluindo abordagens de *ensemble*, que combinam diversos algoritmos classificadores. Isso permitirá aprofundar a nossa compreensão da evasão acadêmica e desenvolver estratégias mais abrangentes para apoiar os alunos em sua jornada.

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*, page Vol 15 No 3 (2019): Learning Aanalytics: For a dialogue between teaching pratices and educational research.

Alban, M. and Mauricio, D. (2019). Predicting university dropout trough data mining: A systematic literature. *Indian Journal of Science and Technology*, 12(4):1–12.

Barosso, M. F. and Falcão, E. B. M. (2004). Evasão universitária: o caso do instituto de física da ufrj. *IX Encontro Nacional de Pesquisa em Ensino de Física*, 9:1–14.

Bramer, M. (2016). *Principles of Data Mining*. Undergraduate Topics in Computer Science. Springer, London, 3 edition.

Cabral, A. P. and Tavares, J. (2005). Leitura/compreensão, escrita e sucesso acadêmico: um estudo de diagnóstico em quatro universidades portuguesas. *Psicologia Escolar e Educacional*, 9(2):203–213.

Colpo, M., Primo, T., Pernas, A., and Cechinel, C. (2020). Mineração de dados educacionais na previsão de evasão: uma rsl sob a perspectiva do congresso brasileiro de informática na educação. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1102–1111, Porto Alegre, RS, Brasil. SBC.

Costa, E., Baker, R., Amorim, L., Magalhães, J., and Marinho, T. (2013). Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29.

da S. Freitas, F. A., Vasconcelos, F. F. X., Peixoto, S. A., Hassan, M. M., Dewan, M. A. A., de Victor Hugo C. de Albuquerque, and Filho, P. P. R. (2020). Iot system for school dropout prediction using machine learning techniques based on socioeconomic data. *Electronics (Switzerland)*, 9(10):1–14. cited By 5.

da Silva Soares, I. (2009). Ufrj – escola politécnica – vestibular 1993-2009 – revisão histórica – vagas, evasão e retenção. *COBENGE 2009 – XXXVII Congresso Brasileiro de Educação em Engenharia, Recife, PE,*.

Davok, D. F. and Bernard, R. P. (2016). Avaliação dos índices de evasão nos cursos de graduação da universidade do estado de santa catarina - UDESC. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, 21(2):503–522.

de Albuquerque Motta, P. R. (2016). Estudo exploratório do uso de classificadores para a predição de desempenho e abandono em universidade. Master's thesis, Programa de Pós-graduação Mestrado em Ciência da Computação (INF) da Universidade Federal de Goiás, Goiânia, GO.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54.

Goldschmidt, R., Passos, E., and Bezerra, E. (2015). *Data Mining*. GEN LTC.

Han, J. and Kamber, M. (2012). *Data mining: concepts and techniques*. Kaufmann, San Francisco [u.a.].

Manhães, L. M. B. (2015). Predição do desempenho acadêmico utilizando mineração de dados educacionais. Master's thesis, Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, BR.

Oliveira, J. G. J. (2015). Identificação de padrões para a análise da evasão em cursos de graduação usando mineração de dados educacionais. Master's thesis, Pós-Graduação em Computação Aplicada da Universidade Tecnológica Federal do Paraná, Curitiba, PR.

Oliveira, J. G. J., Noronha, R. V., and Kaestner, C. A. A. (2015). Análise da correlação da evasão de cursos de graduação com o empréstimo de livros em biblioteca. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, 3(1):601.

Passos, F. J. V., Braathen, P. C., Guerreiro, M., Arruda, M. A., and Bohnenberger, J. C. (2001a). Programa de tutoria: uma experiência. *XXIX Congresso Brasileiro de Ensino de Engenharia*.

Passos, F. J. V., Guerreiro, M., Braathen, P. C., and Arruda, M. A. (2001b). Programa de tutoria: uma esperança. *XXIX Congresso Brasileiro de Ensino de Engenharia*.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Santos, V., Saraiva, D., and Oliveira, C. (2021). Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 1196–1210, Porto Alegre, RS, Brasil. SBC.

Severino, A. J. (2014). *Metodologia do trabalho científico*. Cortez Editora, São Paulo, Brasil, 1 edition.

Silva Filho, R. L. L., Motejunas, P. R., Hipólito, O., and de Carvalho Melo Lobo, M. B. (2007). A evasão no ensino superior brasileiro. *Cadernos de Pesquisa*, 37(132):641–659.

Simão, A. M. V., abd Sandra Fernandes, A. F., and Figueira, C. (2008). Tutoria no ensino superior: concepções e práticas. *Sísifo: Revista de Ciências da Educação*, 7:75–88.