Modelos de previsão de evasão tardia na graduação de uma universidade pública

Caio Vinicius Monteiro Martins[1], Filipe Carvalho Lacerda[1], Igor Procópio do Carmo[1], Edmilson Vitorino Scovino da Silva[1], Tatiane Ornelas Martins Alves[1], Johnny Moreira Gomes[2], Ricardo Silva Campos[1,2]

[1]Instituto Metodista Granbery – Bacharelado em Sistemas de Informação

[2]Universidade Federal de Juiz de Fora – CGCO

ricardo.campos@ufjf.br

Abstract. *The students' dropout represents a critical problem for higher education institutions, because their budget is based on the number of students enrolled. Furthermore, the dropout is a disservice to job market since it causes qualified workers lack. Thus, this work proposes dropout prediction models, especially for the late dropout. A 30,000 students database was used for training supervised machine learning algorithms. Their performances are then discussed in terms of accuracy and f1-score.*

Resumo. *A evasão estudantil representa um problema crítico para as instituições de ensino superior, pois seu orçamento é baseado no número de alunos matriculados. Além disso, a evasão é um problema ao mercado de trabalho, pois causa falta de mão de obra qualificada. Assim, este trabalho propõe modelos de previsão de evasão, especialmente a evasão tardia. Um banco de dados de 30.000 alunos foi usado para treinar algoritmos de aprendizado de máquina supervisionados. Seus desempenhos são então discutidos em termos de acurácia e f1-score.*

1. Introdução

A Universidade Federal de Juiz de Fora (UFJF) possui um campus sede situado em Juiz de Fora - MG e outro campus em Governador Valadares - MG. A UFJF oferece aproximadamente 93 cursos de graduação, 45 cursos de mestrado e 24 cursos de doutorado, abrangendo todas as áreas do conhecimento. Em totalidade, a instituição possui um corpo discente superior a 20.000 alunos na modalidade presencial e cerca de 3.000 na modalidade à distância.

Desde 2003, a UFJF tem implementado um sistema próprio denominado Sistema Integrado de Gestão Acadêmica (SIGA) para gerir suas informações institucionais. O SIGA foi desenvolvido integralmente em software livre, o que contribuiu para a redução de custos e para o alinhamento estratégico com as diretrizes do Governo Federal. O SIGA é composto atualmente por 43 módulos, divididos em quatro grandes áreas ou sistemas: Acadêmico, Administrativo, Gestão de Pessoas e Sistemas de Apoio.

A evasão de universitários pode acarretar graves prejuízos para a economia, a gestão institucional e diversos aspectos sociais. A alocação de recursos financeiros destinados às universidades está diretamente relacionada à quantidade de alunos matriculados e concluintes. Quando as metas institucionais de formação de alunos não são atingidas, as instituições de ensino público são afetadas negativamente, ocasionando,

entre outros problemas, a perda de programas e planos fornecidos pelas Instituições de Ensino Superior (IES). Ademais, os docentes são prejudicados por não poderem desempenhar plenamente sua função [PRESTES, 2014].

A evasão escolar, especialmente entre estudantes em fases avançadas, é um desafio significativo enfrentado por instituições de ensino de todos os níveis, nas esferas pública e privada. De acordo com Santana (1996), a evasão representa um dos problemas mais graves do sistema educacional, pois causa desequilíbrio, desarmonia e desajuste dos objetivos educacionais. Assim, conclui-se que existe uma grande necessidade de entender a evasão do ensino superior, principalmente devido a atual escassez de profissionais qualificados no mercado de trabalho e a necessidade do país de qualificar sua população.

De Oliveira Júnior (2017) desenvolveu um método baseado em aprendizado de máquina para selecionar os atributos mais relevantes na previsão da evasão estudantil, incluindo características socioeconômicas, desempenho acadêmico, dados demográficos, entre outros. Esse método permite antecipar a probabilidade de abandono do curso pelos alunos, auxiliando na implementação de medidas efetivas de prevenção da retenção e da evasão. As previsões obtidas pelo método podem ser úteis para a tomada de decisões.

Portanto, o objetivo deste trabalho é propor modelos de previsão de evasão, em especial a evasão tardia. Serão utilizados algoritmos de aprendizagem de máquina supervisionada, com banco de dados de todos os estudantes de graduação presencial entre 2003 e 2020 da UFJF. Então, será realizada uma comparação em termos do desempenho dos diferentes algoritmos aplicados.

2. Revisão de literatura

De acordo com Vasquez *et al*. (2003), a evasão estudantil é uma condição enfrentada por um aluno que não consegue alcançar o sucesso em seu projeto educacional. Na prática, um estudante é considerado evadido quando não realiza atividades acadêmicas por três semestres consecutivos em uma instituição de ensino superior. Assim, a evasão estudantil pode ser classificada em duas perspectivas: tempo e espaço. A evasão temporal pode ser subdividida em três categorias, cada uma com suas próprias causas: evasão precoce, na qual o indivíduo foi aprovado no curso, mas não se matriculou; evasão inicial, na qual o aluno evadiu nos primeiros quatro períodos do curso; e evasão tardia, na qual o aluno evadiu a partir do quinto período.

Por outro lado, ainda segundo Vasquez *et al*. (2003) a evasão espacial se divide em: deserção institucional, que ocorre quando o estudante abandona a universidade; evasão interna, que se dá quando o estudante trocar seu programa acadêmico por outro oferecido dentro da mesma instituição de ensino; e evasão do sistema educacional, em que o estudante interrompe seus estudos.

Estudos evidenciados por Pereira (2003) apontam dois tipos de fatores determinantes para a evasão, os caracterizando como: fatores internos da instituição, que consistem em infraestrutura deficitária, acervo desatualizado, métodos de avaliação docente e deficiência didático pedagógica dos professores; e fatores externos, que consistem em dificuldades financeiras, escolha equivocada do curso, falta de base para acompanhar o curso escolhido, fato de ter sido admitido em um curso que não foi sua primeira opção e assim como uma série de outros fatores que podem se englobar nos fatores externos sejam eles os variados motivos que o estudante possa estar enfrentando.

No Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), vinculado ao Ministério da Educação, é responsável pela realização anual do censo escolar da educação superior, e por conseguinte, é responsável pelo levantamento da evasão. O INEP considera evasão como a saída antecipada, antes da conclusão, por desistência independentemente do motivo. Além disto, em seus indicadores, o censo considera como desistentes todos aqueles estudantes que encerraram seu vínculo com o curso, por transferência para outro curso, abandono ou desligamento [INEP, 2017]. Ou seja, os indicadores de evasão do censo abrangem os três tipos de evasão espacial apontados por Vasquez et al. (2003).

De acordo com Brito Júnior (2018), o impacto gerado pela evasão nos cursos de graduação é significativo, pois cria um déficit de mão de obra e consequentemente, atrapalha o desenvolvimento do mercado de trabalho. Além de causar prejuízos financeiros, pois diminui a verba arrecadada pelas universidades públicas e prejudica o fluxo de caixa das universidades privadas.

Segundo o Mapa do Ensino Superior do Instituto Semesp[1], os índices de evasão em 2019 foram de 30,7% na rede privada e 18,4% na rede pública. Entre cursos presenciais, o maior percentual de evasão é 37,6% em Sistemas de Informação, seguido de 35,9% em Administração. Em cursos à distância, as maiores evasões ocorrem em Marketing (44,7%) e Matemática Formação de Professor (44,3%).

No Brasil, já foram realizados diversos estudos para buscar entender os motivos mais frequentes que levam os alunos a evadirem de seus cursos durante sua graduação. Por exemplo, Brito Júnior (2018) desenvolveu um estudo com o uso da mineração de dados para identificar os perfis dos graduandos que evadem do curso de Sistemas de Informação, utilizando os dados da UFRN.

Couto e Santana (2017) discutiram a aplicação de técnicas de mineração de dados educacionais para identificar fatores associados à evasão em um curso de Engenharia de Produção e utilizaram técnicas de classificação e associação para identificar as variáveis mais relevantes para a previsão da evasão. Os resultados indicaram que as variáveis mais importantes para a evasão foram o desempenho acadêmico, a carga horária de trabalho e a distância entre a casa e a universidade.

Santos e Goya (2020) verificaram se os classificadores gerados por algoritmos de aprendizagem de máquina são capazes de distinguir entre alunos que concluem seus cursos ou que evadem. Assim como De Jesus *et al*. (2021), que através do aprendizado supervisionado com redes neurais artificiais, criou um modelo preditivo para prever os alunos com risco de evadir do curso de Licenciatura em Computação da Universidade do Estado do Amazonas.

Rodrigues e Gouveia (2021) utilizaram algoritmos de aprendizagem supervisionada em uma base de dados do INEP para minerar conhecimentos sobre o tempo de permanência de discentes em cursos de graduação nas IES brasileiras.

Neste artigo, utilizaremos uma base de dados de uma universidade federal para desenvolver um estudo semelhante aos citados acima, porém criaremos um modelo separado para os alunos que alcançam os períodos finais, com foco na evasão tardia. Além

[1] Disponível em: <https://www.semesp.org.br/mapa/edicao-11/brasil/evasao/> Acesso em 07/jul/2022.

disso, as outras abordagens normalmente optam por fazer uma análise limitada a um único curso de graduação ou usando um pequeno recorte de tempo, principalmente pela dificuldade em se obter e tratar os dados de diferentes fontes e cuja estrutura se altera com o passar dos anos. Este trabalho, por outro lado, lida com todos os 93 cursos de graduação de uma única IES, em um período de quase 20 anos. Pretende-se, portanto, propor modelos mais generalistas e abrangentes.

3. Métodos

Os microdados de estudantes da Universidade Federal de Juiz de Fora são públicos e foram obtidos através de uma solicitação de informação no portal governamental FalaBR[2], com o número 23546.051594/2020-15. Foram solicitados os dados de estudantes com matrículas válidas de graduação da UFJF de 2003 a 2020, na modalidade presencial. Os dados foram disponibilizados de maneira anônima para que não fossem violadas a privacidade e intimidade da comunidade acadêmica, conforme parâmetros da Lei de Acesso à Informação (nº 12.527 de 18/11/2011).

Nestes dados está representada a situação do vínculo do aluno a um curso de graduação, ou seja, a medida de evasão abrange as evasões internas, institucionais e do sistema educacional, no que se refere à espacialidade. Em relação à temporalidade, há apenas informações sobre a evasão inicial e tardia, sendo desconsideradas as informações sobre a evasão precoce.

As variáveis disponibilizadas para o desenvolvimento deste trabalho são: **Cota:** Sistema de cotas de ingresso; **Períodos**: Número de períodos cursados; **Etnia:** Amarela; branca; indígena; parda preta; outra; **GAP**: Diferença em anos entre o fim do ensino médio à entrada do ensino superior. **Grande área**; **IRA**: Índice de rendimento acadêmico, dado entre 0 e 100. **Idade de saída**: Idade ao sair do curso; **Mesma cidade** : Se o aluno faz o curso em sua cidade natal ou não; **MonTccEst**: Se o curso contém em sua grade monografia, tcc ou estágio obrigatório. **Sexo**; **Situação:** Determina se o aluno é evadido ou concluído. **Turno**: Período de realização do curso.

Portanto, o conjunto de dados será utilizado para treinar os algoritmos de aprendizagem, de forma que a partir de uma entrada qualquer na forma (cota, períodos, etnia, GAP, área, IRA, idade, mesma cidade, monttccest, sexo), será produzida uma saída com a predição categórica de “evadido” ou “concluído”.

3.1.Seleção, pré-processamento e transformação

Este projeto utiliza a linguagem de programação Python na IDE Jupyter Notebook e as bibliotecas Pandas e Scikit-learn [PEDREGOSA *et al.,* 2011].

É comum que *dataframes* contenham erros, como valores ausentes ou até mesmo valores errados e, portanto, é necessária uma limpeza. Desta forma, pode-se optar conforme a conveniência entre algumas possibilidades, como imputação de valores a partir de medidas de tendência central ou exclusão de registros.

Na coluna etnia, os valores ausentes foram substituídos por uma constante com valor “OUTRA”. Na coluna “GAP” que é referente ao tempo entre o término do aluno no ensino médio e a sua entrada no ensino superior em anos haviam números incoerentes,

[2] Disponível em: <https://falabr.cgu.gov.br> Acesso em 07/jul/2022.

devido a este fato foi adicionado um limitador para essa coluna, no qual se houvesse um “GAP” com valores muito discrepantes seriam retirados esses registros (>60), na coluna “Idade de Saída” que é referente a idade na qual o aluno terminou o curso existiam valores negativos que também foram ignorados, por fim na coluna “Prazo Ideal” que indica o número de semestres esperado para conclusão do curso pelo aluno foram retirados registros com valores iguais a zero.

Um dos fatores centrais para a compreensão do projeto em relação à metodologia e aos resultados é a noção dos grupos de acesso prioritários utilizados no ingresso à universidade: Grupo A: baixa renda, preto, pardo ou indígena (PPI), oriundo de escola pública. Grupo B: baixa renda, oriundo de escola pública. Grupo C: ampla concorrência. Grupo D: PPI, escola pública. Grupo E: escola pública.

Com base nesses parâmetros de cotas o projeto busca uma padronização em valores ordinais para esse sistema, dessa forma desenvolveu-se um sistema como uma ordem de vulnerabilidade, na qual o maior número representa um grupo que demonstra maior vulnerabilidade social e o menor grupo a ampla concorrência, dessa forma encontra-se então: Grupo A: 5; Grupo B: 4; Grupo D: 3; Grupo E: 2; Grupo C: 1. Ao classificar os grupos desta maneira os dados se tornam então tratados para serem utilizados pelos algoritmos.

Além disso, outras colunas foram adicionadas como parte do tratamento, como a coluna “Mesma Cidade” que indica se a cidade natal do aluno é a mesma onde é localizada o campus de sua graduação. A coluna “MonTccEst” foi criada como boolena e representa se o curso possui monografia, trabalho de conclusão de curso ou estágio obrigatório em seu currículo. A coluna "Percentual de Conclusão" indica o percentual de períodos que o aluno cursou em relação ao prazo ideal, com o objetivo de investigar se o número de períodos cursados influencia a evasão escolar. Finalmente, foi adicionada a coluna "Situação", com valores booleanos e indicam se o aluno está concluído ou evadido.

Outra coluna importante é a “IRA” que se refere ao índice de rendimento acadêmico, podendo ser descrito em 4 intervalos: Grupo 1: 0 a 25; Grupo 2: 25 a 50; Grupo 3: 50 a 75; Grupo 3: 75 a 100. Dessa forma será possível analisar se o IRA do aluno é impactante em sua decisão de evasão.

Para os algoritmos, é de suma importância transformar os dados categóricos em numéricos. Portanto, as variáveis categóricas nominais, como etnia e grande área, foram transformadas em dicotômicas. Após isso, obtém-se um *dataframe* que contém exclusivamente registros numéricos.

Ademais, foram descartados da base os alunos ativos, que ainda estão matriculados. A utilização destes registros não faz sentido na análise proposta, pois pretende-se apenas prever as possibilidades de evasão e conclusão.

Para finalizar o tratamento, o *dataframe* foi separado em dois. O primeiro contém todos os estudantes e é chamado de geral. O outro conjunto contém estudantes que conseguiram cursar além de 70% do número previsto de períodos, chamado de finalista.

3.2.Aprendizagem de máquina supervisionada

Existem vários algoritmos de aprendizagem de máquina supervisionada para tarefas de classificação. A seguir serão apresentados os algoritmos utilizados neste trabalho.

Árvore de decisão

O primeiro é conhecido como árvore de decisão, que representa uma função que recebe como entrada um conjunto de valores e retorna um valor de saída. Estes valores podem ser discretos ou contínuos. A técnica se baseia na divisão de dados em grupos de maneira homogênea, com o objetivo de classificá-los. A cada iteração é encontrado o atributo que gera a melhor divisão dos dados possível.

A árvore de decisão pode ser implementada de diversas maneiras. O algoritmo ID3 cria uma árvore de múltiplos caminhos por meio da busca gulosa da característica categórica que produzirá o maior ganho de informação para a classificação. O ID3 é restrito a características categóricas. Seu sucessor, o algoritmo C4.5, permitiu características numéricas e converteu árvores treinadas em conjuntos de regras se-então. Por fim, o algoritmo CART, semelhante ao C4.5, constrói árvores binárias que suportam variáveis de destino numéricas. Neste trabalho, o CART foi utilizado através da biblioteca *Scikit-learn* [PEDREGOSA *et al.,* 2011].

K-vizinhos mais próximos

O algoritmo K-vizinhos mais próximos (KNN) é uma técnica de classificação não paramétrica que se baseia na similaridade entre uma entrada e as K entradas históricas mais próximas. A classificação de uma nova entrada é determinada pela votação dos k vizinhos mais próximos, sendo atribuída a classe mais frequente entre eles. É recomendado escolher um valor ímpar de k para evitar empates em caso de número par de classes.

Neste trabalho, existem duas saídas possíveis: evadido e concluído. O parâmetro k foi escolhido empiricamente, em que o valor k=5 obteve melhores desempenhos.

Regressão logística

Segundo Gonzalez (2018), a regressão logística é uma técnica estatística que visa desenvolver um modelo a partir de um conjunto de observações para prever valores de uma variável categórica, muitas vezes binária, em função de uma ou mais variáveis independentes. Embora a variável dependente seja geralmente categórica, é possível transformá-la em uma variável dicotômica para aplicar esta técnica. Quanto às variáveis independentes, estas podem ser categóricas ou contínuas.

Florestas aleatórias

A floresta aleatória é um algoritmo de aprendizado de máquina de origem semelhante ao modelo de árvore de decisão. No entanto, em vez de usar apenas uma árvore de decisão, esse modelo utiliza um conjunto de árvores de decisão, cujas saídas são combinadas para formar uma classificação mais robusta e confiável. Cada árvore na floresta aleatória é construída a partir de uma amostra aleatória dos dados de treinamento e apresenta uma seleção aleatória de características em cada nó de decisão. Pode ser usada tanto para problemas de classificação quanto para problemas de regressão.

3.3.Protocolo de validação

Neste estudo, os dados foram pré-processados conforme a seção 3.1 e divididos em duas bases: treinamento e teste. A base de treinamento será utilizada para treinar os algoritmos.

Posteriormente, com a base de teste, todos os algoritmos são executados para obtenção das métricas de desempenho.

Foi aplicado um protocolo de validação estratificada, em que a estratificação foi feita pelos rótulos de saída. Isto significa que a seleção dos registros para compor as bases de teste e treino foi realizada de forma aleatória, mas de maneira a se manter a proporção de evadidos e concluídos da base original.

Além disso, utilizou-se uma proporção de 70% dos registros para a base de treinamento e 30% para a base de teste. Embora possa variar de acordo com o contexto, essa proporção é uma prática comum na literatura e é considerada um bom ponto de partida [KHOSHGOFTAAR, GOLAWALA, VAN HULSE, 2007].

3.4.Métricas de avaliação

A matriz de confusão é uma tabela (Tabela 1) que auxilia na visualização do desempenho de algoritmos de classificação dicotômica. Possui duas linhas e duas colunas, que relatam o número de acertos e erros. Os acertos são exemplos corretamente classificados pelo modelo, chamados de verdadeiro positivo (VP) e verdadeiro negativo (VN). Os erros são exemplos classificados equivocadamente pelo algoritmo e são chamados de falso positivo (FP) e falso negativo (FN).

Tabela 1: Matriz de confusão.

		Previsão	
		Sim	Não
Real	Sim	VP	FN
	Não	FP	VN

Uma vez obtidos os resultados de previsão a partir da base de teste, os valores de saída dos modelos são comparados à variável de saída da base de teste. A partir desta comparação, obtém-se uma porcentagem total de acertos, chamada de acurácia. Ela mede a qualidade geral das classificações e será usada neste trabalho como um dos critérios de avaliação dos modelos. Em termos da matriz de confusão, a acurácia é dada por:

$$acurácia = \frac{VP + VN}{VP + FN + FP + VN}.$$

A precisão mede a qualidade das classificações positivas. É dada pela razão entre VP e o número total de previsões positivas do modelo:

$$precisão = \frac{VP}{VP + FP}.$$

O *recall* mede a capacidade do modelo em identificar corretamente os exemplos positivos. É definido como:

$$recall = \frac{VP}{VP + FN}.$$

Exemplificando no contexto deste trabalho, a precisão da categoria concluído responde à pergunta: dentre as classificações de concluído dadas pelo algoritmo, quantas vezes ele acertou? Por outro lado, dentre todas as situações em que se esperava o valor concluído, quantas foram previstas corretamente? Esta resposta é dada pelo *recall*.

A última métrica apresentada é chamada de *F1-score*. É uma média harmônica entre a precisão e o *recall* do modelo e fornece uma visão geral do desempenho geral do modelo. Por este motivo, ela será apresentada em conjunto com a acurácia na próxima seção. É dada por:

$$F1score = 2 * \frac{precisão * recall}{precisão + recall}.$$

4.Resultados

O *dataset* disponibilizado possui as colunas descritas no Capítulo 3 e possui um total de 46914 registros totais após a limpeza dos dados, incluindo as situações de alunos concluídos, evadidos e ativos. Para o estudo ser realizado foi necessário filtrar o *dataset* mantendo apenas os concluídos e evadidos totalizando 30931 registros, uma vez que os alunos ativos não seriam úteis para o estudo, vindo da necessidade de encontrar padrões entre os alunos que evadem e que concluem e as discrepâncias entre eles.

Para realizar uma comparação entre o perfil comum do aluno que evade e o perfil do aluno que evade no final da graduação, serão feitas duas análises. Uma geral, que abrange todos os estudantes da base (geral) e outra apenas com os considerados finalistas, que já cursaram 70% do número previsto de semestres.

Desta forma, dos 30931 alunos contidos na base geral, 19265 se formaram e 11666 (37,7%) abandonaram seus cursos. Na base de finalistas, existem 22174 estudantes, dos quais 18884 concluíram e apenas 3290 (14,8%) evadiram tardiamente. Para cada uma destas bases, foi feita uma validação estratificada, em que 30% dos registros foram separados para teste. Os resultados são apresentados na Tabela 2.

Tabela 2: Métricas acurácia e *f1-score* (concluído e evadido).

	Geral			Finalistas		
Modelo	Acurácia	*F1-score*		Acurácia	*F1-score*	
		Concl.	Evad.		Concl.	Evad.
Árv. Decisão	0,90	0,92	0,87	0,88	0,93	0,58
KNN	0,90	0,92	0,86	0,89	0,94	0,48
Reg. Logística	0,90	0,92	0,87	0,91	0,93	0,58
Fl. aleatórias	0,93	0,94	0,90	0,91	0,95	0,63

O algoritmo de árvore de decisão na base geral teve acurácia total de 0,9. Ao se avaliar o *f1-score* por classes, a taxa de acerto foi de 0,92 para concluídos e 0,87 para evadidos. Na base de finalistas, a acurácia do modelo foi de 0,88, em que a classe de concluídos teve *f1-score* de 0,93 e a de evadidos, 0,58. Este padrão se repetirá nos demais algoritmos, exibidos adiante, o que pode ser explicado pela dificuldade de se formar um padrão para esta classe devido ao número pequeno de evadidos entre os finalistas. O contrário também ocorre, pois graças a um número grande de concluídos entre os finalistas, percebe-se maiores índices de acerto para esta classe.

A acurácia do KNN na base geral é de 0,90, enquanto o *f1-score* das classes de concluídos e evadidos, foram 0,92 e 0,86. Na base de finalistas, a acurácia total foi de 0,89. Entre os concluídos, obteve-se *f1-score* de 0,94 e entre os evadidos, 0,48. Ou seja, o menor índice de acerto entre os evadidos da base de finalistas se repetiu.

Na base geral, a regressão logística teve acurácia de 0,90, em que o *f1-score* da classe de concluídos foi de 0,92 e dos evadidos, 0,87. Entre os finalistas, a acurácia total foi 0,91, em que 0,93 e 0,58 são os *f1-score* entre concluídos e evadidos, respectivamente.

O algoritmo de floresta aleatória teve o melhor desempenho entre todos, com acurácia de 0,93 na base geral e f1-scores de 0,94 e 0,88. Na base de finalistas, obteve 0,91 de acurácia, sendo 0,95 de *f1-score* entre concluídos e 0,63 entre evadidos.

Ao se detalhar a última métrica, percebe-se que dentre as classificações de evadido dadas pelo algoritmo, ele acertou 78% (precisão). Sob outra perspectiva, quando o valor evadido era esperado, o algoritmo acertou a previsão em 54% (*recall)*.

5. Considerações finais

Este trabalho utilizou algoritmos de aprendizagem de máquina para prever se um estudante evade ou conclui o curso. Foi utilizada uma base com dados de todos os estudantes de graduação presencial de uma universidade federal entre 2003 e 2020. Primeiro a base foi pré-processada, em que foram tratados valores ausentes e aplicadas transformações em variáveis categóricas. Em seguida, a base foi separada em duas: uma geral com todos os registros e outra chamada de finalistas, que contém alunos que chegaram aos períodos finais de sua graduação.

Em seguida, essas bases foram utilizadas para treinar quatro algoritmos de aprendizagem de máquina supervisionados para classificação: árvores de decisão, K vizinhos mais próximos, regressão logística e florestas aleatórias. Para a base geral, os algoritmos apresentaram bons desempenhos, com acurácia e *f1-score* em torno de 90%.

Para a base de finalistas, a acurácia média dos quatro algoritmos ficou em torno de 90%. O *f1-score* para a categoria concluído foi ligeiramente mais alto, em torno de 0,94. Por outro lado, foi menor na categoria de evadidos, variando entre 0,48 e 0,63. Nesta base, as métricas de precisão e *recall* do algoritmo de florestas aleatórias foram de 78% e 54%, respectivamente. Portanto, pode-se dizer que o algoritmo geralmente está correto ao classificar como evadido, mas ele não consegue identificar todos os casos existentes. Assim, deve-se levar em consideração o contexto em que a predição será aplicada, especialmente sobre o impacto causado em se classificar estudantes possivelmente evadidos como concluídos.

Este problema pode ser explicado por dois fatores. O primeiro é a quantidade insuficiente de alunos evadidos na base de finalistas para se capturar um padrão. Outro problema que possivelmente possui maior impacto é que a evasão tardia pode ser causada por motivos variados e complexos não refletidos nos dados. É importante ressaltar que, de fato, esta pesquisa é limitada às informações que constam na base de dados disponibilizada pela instituição.

Como trabalho futuro, pretende-se analisar a importância das variáveis de entrada para cada modelo de predição proposto neste artigo, com o intuito de analisar quais características têm maior impacto no desempenho dos algoritmos.

Referências

Brito Júnior, I. (2018). Uso de mineração de dados educacionais para a classificação e identificação de perfis de evasão de graduandos em Sistemas de Informação (Monografia, Universidade Federal do Rio Grande do Norte).

Couto, D., & Santana, A. (2017). Mineração de dados educacionais aplicada à identificação de variáveis associadas à evasão e retenção. II Congresso sobre Tecnologia na Educação (pp. 333-344).

de Jesus, H. O.; Rodriguez, L. C.; Costa Junior, A. O. (2021) Predição de Evasão Escolar na Licenciatura em Computação. Revista Brasileira de Informática na Educação, v. 29, p. 255-272. [DOI: 10.5753/rbie.2021.29.0.255]

de Oliveira Júnior, J. G.; Noronha, R V.; Kaestner, C. A. A. (2017) Método de seleção de atributos aplicados na previsão da evasão de cursos de graduação. Revista de Informática Aplicada, v. 13, n. 2. [DOI: 10.13037/ria.vol13n2.206]

Gonzalez, L. A. (2018) Regressão logística e suas aplicações. Monografia (Universidade Federal do Maranhão, Campus do Bacanga). Disponível em: https://rosario.ufma.br/jspui/handle/123456789/3572 Acesso em: 17/nov/2022

INEP, INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (2017). Metodologia de Cálculo dos Indicadores de Fluxo da Educação. Brasília: Inep. Disponível em: https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2017/metodologia_indicadores_trajetoria_curso.pdf . Acesso em: 07/jul/2022

Khoshgoftaar, T. M., Golawala, M., & Van Hulse, J. (2007). An empirical study of learning from imbalanced data using random forest. In 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007) Vol. 2, pp. 310-317.

Pedregosa, F. *et al.*(2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pereira, F. C. B. et al. (2003) Determinantes da evasão de alunos e os custos ocultos para as instituições de ensino superior: uma aplicação na Universidade do Extremo Sul Catarinense. Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia de Produção.

Prestes, E. M. D. T.; Fialho, M. G. D.; Pfeiffer, D. K. (2014). A evasão no ensino superior globalizado e suas repercussões na gestão universitária. 6º Encontro Internacional da Sociedade Brasileira de Educação Comparada.

Rodrigues, E. M., & Gouveia, R. M. (2021). Técnicas de *machine learning* para predição do tempo de permanência na graduação no Âmbito do ensino superior público brasileiro. Anais do VI Congresso sobre Tecnologias na Educação (pp. 128-137). SBC.

Santana, A. P.; Perrosso, J. E. C.; Macedo, K. L. O.; Farias, S. P. D. (1996) Evasão escolar em escolas públicas municipais rurais localizadas em Montes Claros. Trabalho de Conclusão de Curso. Universidade Estadual de Montes Claros.

Santos, P.; Goya, D. (2020) Aprendizado de Máquina Aplicado à Análise de Evasão em Cursos de Sistemas de Informação. Anais Estendidos do XVI Simpósio Brasileiro de Sistemas de Informação. SBC. [DOI: 10.5753/sbsi.2020.13145]

Vasquez, J.; Castaño, E.; Gallón, S.; Gomez, K. (2003). Determinantes de la deserción estudiantil en la Universidad de Antioquia. Borradores del Cie, (04), 1-38.