Algoritmos de Mineração de Dados para Análise de Evasão na Graduação da Universidade de Brasília

Tiago de Souza Fernandes
Depto. de Ciência da Computação
Universidade de Brasília
tiagotsf2000@gmail.com

***Resumo*—A evasão no Ensino Superior é um problema que preocupa diversas instituições do mundo, e uma solução que busca identificar suas causas e casos de forma prematura é a Mineração de Dados Educacional (EDM). Este trabalho propõe aprimorar a abordagem de EDM atualmente desenvolvida na UnB, testando novas abordagens a fim de obter melhores predições. São testados: uma nova engenharia de atributos, utilizando uma escala para representar menções e gerando atributos acadêmicos como o IRA; o impacto de três informações sócio-demográficas na predição, o Sexo, a Forma de Ingresso e a Distância da Residência à UnB; e dois novos algoritmos para o processamento dos dados, o XGBoost e o SVM, definindo limiares para a otimização das métricas. Utilizando a metodologia Knowledge Discovery Process, foi obtido um classificador com sensibilidade de 88,1% e acurácia de 86,6%, treinado com os dados do curso de Engenharia Mecatrônica, utilizando o algoritmo Floresta Aleatória e de todas as melhorias propostas neste trabalho. Os atributos acadêmicos gerados foram de grande relevância para a predição pelos algoritmos, o uso da escala para representar as menções teve um impacto positivo nas métricas, e foi descoberta uma relação entre o atributo Distância da Residência à Universidade e a evasão dos alunos.**

***Index Terms*—Keywords: mineração de dados, evasão estudantil, aprendizado de máquina, educação.**

I. Introdução

A evasão é um fenômeno social complexo que se refere a interrupção no ciclo de estudos de um aluno [1]. Esse é um tema importante para as Instituições de Ensino Superior (IES), pois afeta de maneira negativa o impacto delas na sociedade [2]. Índices de evasão altos podem ser vistos como um investimento sem retorno, uma vez que esforços de profissionais e recursos financeiros são investidos, para entregar à sociedade pessoas sem a devida preparação para exercerem seus papeis com eficiência e competência [2], [3].

A evasão é resultado de uma combinação de fatores sociais, econômicos e pessoais [4], e no ensino superior em específico, existe uma dificuldade em identificar origens e causas de evasão, pois o contexto do aluno de ensino superior é complexo e sua trajetória pode tomar diversos caminhos [5]. Alguns métodos usados para entender essas causas utilizam pesquisas baseadas em perguntas e formulários, contudo são métodos custosos de serem implementados e não podem ser generalizados para outras instituições [6]. Além disso, poucas são as IES brasileiras que possuem um programa institucional profissionalizado de combate a evasão, com planejamento, acompanhamento e coleta de resultados e experiências bem-sucedidas [3].

A *Educational Data Mining* (EDM), ou Mineração de Dados Educacional [7], é uma área de pesquisa crescente nos últimos anos que utiliza dados institucionais para obter conhecimento. Mineração de dados (MD) é definida como o processo de descobrir padrões a partir de dados [8], utilizando métodos estatísticos [9] ou de algoritmos de aprendizagem de máquina [10]. O uso dessa abordagem vem tomando espaço como solução eficiente para a identificação prematura de casos e causas de evasão no ensino superior (e.g. [6], [11], [12]).

A Universidade de Brasília (UnB) reforçou no Plano de Desenvolvimento Institucional 2018-2022 (PDI) [13] a importância e o interesse que a instituição tem em mitigar os casos de evasão, entender as causas e origens do problema e propor melhores medidas institucionais de acompanhamento dos estudantes. Segundo o documento, na Universidade a evasão na graduação era de 24,54% em 2016, e a meta esperada é diminuir esse valor para 20% até o fim de 2022. A própria instituição possui algumas estratégias para tentar lidar com esse problema [13], e alguns estudos foram realizados por estudantes com o intuito de entender melhor o contexto da evasão na universidade [14]. Contudo, a UnB não possui uma ferramenta institucional que utilize a Mineração de Dados para auxiliar na tomada de decisões por parte da gestão.

Com o intuito de comprovar a viabilidade da abordagem de Mineração de Dados na Universidade, alguns estudos de cunho exploratório foram realizados, utilizando dados institucionais de cursos específicos, para a predição de casos de evasão (e.g. [15], [16], [17]). O trabalho mais recente [16] propõe um preditor treinado utilizando apenas os dados acadêmicos dos estudantes do curso de Engenharia Mecatrônica, obtendo 78,82% de acurácia e 82,58% de sensibilidade na predição com o melhor modelo.

Existem muitos caminhos para prosseguir com esse trabalho: é possível utilizar dados demográficos e sociais dos estudantes, realizar uma análise gráfica e estatística para obter entendimento amplo da base de dados, utilizar uma engenharia de atributos mais robusta e adequada ao problema, e por fim, testar novos algoritmos para o processamento dos dados, otimizando seus parâmetros e definindo melhores limiares para as predições. As melhorias apresentadas podem gerar melhores resultados ou não, e por isso devem ser testadas individualmente e em conjunto com outras mudanças, para que seja possível medir seu impacto nos resultados.

Nesse contexto, este trabalho propõe aprimorar a abordagem

de MD realizada no trabalho [16]. Os objetivos principais são testar três novos dados dos estudantes, o *sexo*, a *forma de entrada* e a *distância da residência à Universidade*, analisando o impacto nos resultados; utilizar uma nova engenharia de atributos para estruturar os dados acadêmicos, de forma a extrair novos atributos e testar o impacto destes nos resultados; e testar dois novos algoritmos, o *XGBoost* e a *Máquina de Vetores de Suporte (SVM)*, otimizando seus parâmetros, definindo limiares para a otimização das métricas e comparando os resultados obtidos com os algoritmos previamente utilizados no projeto.

Os assuntos abordados neste relatório estão organizados da seguinte forma: a Seção II possui uma revisão dos trabalhos passados, a Seção III contém uma revisão da teoria abordada no trabalho, a Seção IV apresenta a aplicação da metodologia escolhida, a Seção V mostra os resultados e métricas obtidas, e por fim, a Seção VI apresenta conclusões alcançadas e futuras melhorias para o projeto.

II. Revisão da Literatura

Com a crescente procura por soluções que identifiquem os casos e causas de evasão no ensino superior, diversos trabalhos foram realizados utilizando a Mineração de Dados Educacional, propondo diferentes modelos e abordagens em diferentes escopos para gerar um modelo preditor de evasão. Existem três pontos importantes a serem definidos em relação a predição e que variam entre os trabalhos da área: a definição de evasão, os atributos utilizados na predição e o processamento dos dados.

A definição de evasão é necessária para classificar os alunos e pode variar dependendo do escopo da base da dados. Alguns dos trabalhos correlatos definem a evasão de forma binária, *evadido* ou *não evadido*, e no geral, classificam estudantes como evadidos caso deixem de pagar as taxas do curso, comuniquem oficialmente sua saída do curso a Universidade ou sejam desligados por não alcançarem o desempenho acadêmico esperado [11], [12], [18]. Um dos trabalhos procura refinar a definição utilizando mais classes para definir evasão, como risco alto, médio e baixo [19], contudo não obteve bons resultados, dificilmente alcançando 50% de predições corretas, pois apesar de uma maior quantidade de classes parecer atrativo, possui uma grande desvantagem: a quantidade de dados por categoria diminui de forma a se tornar insuficiente para uma predição confiável [11].

Foram analisados nos trabalhos passados quais atributos foram utilizados pelo preditor para aprender a identificar os casos de evasão. Alguns dos autores afirmam que a maioria dos estudantes evade logo após o primeiro ano de curso [6], [11], [20], e utilizam dos dados acadêmicos até o primeiro ano da graduação na análise, para que seja possível predizer evasão antes que a maioria dos alunos já tenha evadido. Um dos trabalhos da um passo a frente e busca na análise, de forma dinâmica, o período mais cedo onde é possível realizar uma análise confiável [12], obtendo até 77% de acurácia analisando os dados ao fim do primeiro ano de curso.

Existe uma discordância entre os trabalhos quanto ao uso de dados demográficos e sociais na análise, enquanto alguns afirmam que esses dados possuem pelo menos uma relevância parcial em relação evasão [18], [21], outros afirmam que seu uso produz resultados menos acurados e mais enviesados [12], [6], ou até mesmo discutem sobre uma visão positiva de que alguns alunos podem aumentar suas chances de sucesso apesar de possíveis dificuldades relacionadas a informações sociodemográficas [19]. Grande parte dos trabalhos analisados utilizam do desempenho acadêmico dos estudantes como informação mais relevante para a predição (e.g. [6], [11], [12], [20]), e um dos autores também afirma que o índice de rendimento acadêmico (*IRA*) do estudante, conhecido como *GPA* em universidades do exterior, possui alta correlação com evasão [20].

Em relação ao processamento da base da dados, alguns trabalhos utilizam de classificadores conhecidos, como as *Florestas Aleatórias*, *Árvores de Decisão*, *Redes Neurais*, *Máquina de Vetores de Suporte (SVM)*, *Regressão Logística*, *XGBoost*, entre outros, e comparam os resultados obtidos por esses algoritmos, obtendo até 83% de acurácia nos experimentos [6], [11]. Outros trabalhos vão além e desenvolvem sua própria implementação do algoritmo, modificando-a para obter um controle maior do *sobre-ajuste* e das métricas utilizadas, alcançando até 77% de acurácia nos experimentos [12]. A importância das métricas de precisão e sensibilidade é discutida em um dos trabalhos, já que é desejado em um preditor de evasão minimizar o número de falsos positivos (alunos que evadiram classificados como não evadidos) [6].

No contexto da Universidade de Brasília, [16] recentemente propôs um preditor de evasão que utiliza dos dados institucionais de estudantes do curso de Engenharia Mecatrônica da UnB. No trabalho, a evasão foi definida de forma binária, foram utilizados apenas os dados de desempenho nas disciplinas cursadas, até o primeiro ano, pelos alunos, testando três modelos diferentes de engenharia de atributos e utilizando quatro algoritmos de classificação: *Árvore de Decisão*, *Floresta Aleatória*, *Redes Neurais* e *Regressão Logística*. As métricas de acurácia e sensibilidade foram analisadas, e os melhores resultados encontrados utilizaram o modelo 3 e o algoritmo de *Árvore de Decisão*, com 82.58% de sensibilidade e 78.82% de acurácia. O autor também afirma que a engenharia de atributos foi essencial para garantir os melhores resultados.

Existem algumas melhorias para trabalhos futuros, como a realização de uma análise do impacto dos dados sociodemográficos dos estudantes nos resultados, pois no trabalho apenas foram utilizados dados acadêmicos; a utilização uma variedade maior de algoritmos no processamento, visando melhores resultados da métrica sensibilidade e evitar o sobre-ajuste dos dados; e testar melhorias na engenharia de atributos utilizada, propondo novos atributos como o *IRA* do estudante.

Tendo em vista os trabalhos correlatos, para aprimorar a última abordagem explorada na Universidade de Brasília, utilizando a mesma base de dados, neste trabalho é realizada uma investigação quanto ao impacto do uso de dados sociodemográficos dos estudantes (*sexo*, *forma de entrada* e *CEP/distância da residência à Universidade*) nos resultados, são testados novos modelos de engenharia de atributos, tendo

em vista sua importância nos resultados obtidos, e o uso de dois novos algoritmos, *XGBoost* e *SVM*, pois foram utilizados em trabalhos passados e obtiveram resultados promissores [6], [12]. Assim como no trabalho de [16], é também utilizada a definição de evasão binária para a classificação dos estudantes, e são utilizados apenas os dados acadêmicos do primeiro ano de curso do aluno na análise.

III. Conceitos Teóricos

Neste relatório são utilizadas algumas ferramentas e conceitos teóricos vistos na Mineração de Dados, que serão discutidos e revisados a seguir.

A. Algoritmos de Mineração de Dados

Na Mineração de Dados são utilizados diferentes métodos para a extração de conhecimento dos dados, entre eles, os algoritmos de aprendizado de máquina, que tentam identificar padrões complexos nos dados. Os algoritmos utilizados dependem do tipo de problema que se deseja resolver: o problema da evasão pode ser visto como uma *Classificação Binária*, ou seja, onde se deseja classificar cada um dos elementos em dois grupos [22], para o propósito do trabalho, em evadido ou não-evadido. O processo utiliza um conjunto de atributos chamados *Variáveis de Entrada*, para estimar um atributo binário chamado *Variável de Saída*, esses dados são entregues aos algoritmos, para que identifiquem relações entre os dados de entrada e saída e sejam capazes de gerar sozinhos estimativas de saída para novas variáveis de entrada [22]. Existem diversos algoritmos classificadores, e dentre eles, neste trabalho, são utilizados seis, descritos a seguir.

A *Árvore de Decisão* é um algoritmo preditor baseado em um modelo hierárquico de decisões e suas consequências, possui uma estrutura simples e as escolhas feitas pelo algoritmo podem ser visualizadas e compreendidas de forma lógica. Contudo é uma estrutura suscetível a variações nos dados, ou seja, que pode generalizar mal para conjuntos de dados diferentes dos dados utilizados no treinamento do algoritmo.

A *Floresta Aleatória* é formada por uma combinação de *Árvores de Decisão* que são treinadas com subconjuntos aleatórios dos atributos presentes nos dados, e a resposta final é uma junção das respostas obtidas para cada uma das árvores. Uma grande vantagem desse algoritmo é a sua capacidade de generalizar bem para dados diferentes dos utilizados no treinamento, uma vez que ele constrói as *Árvores de Decisão* com subconjuntos menores dos dados, limitando a profundidade dessas árvores.

A *Rede Neural* é um algoritmo de inspiração biológica que busca imitar a estrutura neural de organismos inteligentes, realizando diversas conexões entre os dados de entrada e a saída, e aprendendo os melhores parâmetros para que a entrada possa ser transformada na saída por meio de funções matemáticas. Esse algoritmo é extremamente poderoso em problemas complexos com uma quantidade abundante de dados, contudo as escolhas feitas pelo algoritmo são de difícil compreensão humana, e sofrem com a falta de dados.

A *Regressão Logística* é uma técnica estatística que tenta gerar um modelo matemático, utilizando funções, para prever um conjunto de variáveis de saída utilizando um conjunto de variáveis de entrada. Contudo, uma de suas maiores limitações é não conseguir resolver problemas não lineares.

O *XGBoost* é um algoritmo com implementação semelhante a da *Floresta Aleatória*, mas utilizando um método de separação dos subconjuntos diferente, o *Gradient Boosting*, que busca melhorar o tempo de execução e o desempenho do modelo, e que normalmente supera a *Floresta Aleatória*.

O *SVM* é um algoritmo que busca separar as variáveis em hyperplanos, e funciona bem em problemas de altas dimensões, ou seja, que possui muitos atributos, e para dados pouco estruturados.

B. Métricas

As métricas são um conjunto de medidas utilizadas para analisar o resultado de um processo [22], e na Mineração de Dados são essenciais para medir a qualidade da predição dos algoritmos de classificação. Na classificação binária existem quatro possíveis resultados para uma predição: *Verdadeiro Positivo* (VP), onde a classe principal é prevista corretamente; *Falso Positivo* (FP), onde a classe principal é prevista incorretamente; *Verdadeiro Negativo* (VN), onde a classe secundária é prevista corretamente; e *Falso Negativo* (FN), onde a classe secundária é prevista incorretamente. A partir dos resultados obtidos para cada predição é possível calcular as seguintes métricas usadas neste trabalho.

A *Acurácia* é a métrica mais simples, que representa a proporção média de acertos das previsões.

$$A = \frac{VP + VN}{VP + FP + VN + FN}$$

A *Sensibilidade* é a métrica mais importante para este trabalho, pois busca responder a proporção da classe principal que foi corretamente predita, ou seja, a porcentagem de alunos evadidos que foram identificados.

$$S = \frac{VP}{VP + FN}$$

A *Precisão* é a proporção de identificações positivas que foram realmente previstas de forma correta, ou seja, a porcentagem dos alunos classificados como evadidos que evadiram de fato.

$$P = \frac{VP}{VP + FP}$$

A última métrica utilizada neste trabalho é a *F-Score*, dada pela média harmônica entre as métricas de *Sensibilidade* e *Precisão*. Para possuir um maior controle da proporção entre as duas métricas, foi utilizada uma versão mais geral da *F-Score*, a F_β, que utiliza um parâmetro β informando quantas vezes a *Sensibilidade* é considerada mais importante que a *Precisão*.

$$F_\beta = (1 + \beta^2) \cdot \frac{P \cdot S}{(P \cdot \beta^2) + S}$$

C. Melhoria do Desempenho dos Algoritmos

Os algoritmos de aprendizado de máquina funcionam a partir de um conjunto de *hiperparâmetros*, que definem as abordagens, limites e coeficientes utilizados no processamento dos dados, como a taxa de aprendizagem de uma Rede Neural, ou a profundidade máxima de uma Árvore de Decisão, por exemplo. A escolha certa de parâmetros pode gerar melhores resultados [22] e, para automatizar esse processo, é utilizada a técnica *Grid Search*, onde são realizados processamentos com diferentes combinações de parâmetros, para encontrar quais obtiveram os melhores resultados.

Um problema comum na Mineração de Dados é o *sobre-ajuste*, ou seja, quando um algoritmo se ajusta muito bem aos dados de treinamento, mas se mostra ineficaz em prever novos dados [22]. Para lidar com esse problema, é possível utilizar a técnica da *Validação Cruzada K-fold*, onde o conjunto de dados de treino é dividido em K subconjuntos, e é realizado um treinamento utilizando um dos subconjuntos para teste e os outros para treino, repetindo esse processo K vezes com um subconjunto para teste diferente a cada iteração. Dessa forma é possível obter um modelo mais confiável que generalize bem em novos dados.

D. Metodologia KDD

A Mineração de Dados envolve todo o processo desde o entendimento do problema até a obtenção de conhecimento [8]. Ao longo do tempo foram desenvolvidas metodologias para a execução desse processo, as mais utilizadas são: a *CRISP-DM* [23], a *SEMMA* [24], e a *KDD* [25], a mais antiga delas. Este trabalho utiliza a metodologia *KDD* ou *Knowledge Discovery in Databases*, por ser voltada para pesquisas científicas.

A KDD é uma metodologia de Mineração de Dados, com o objetivo principal de automatizar o processo de extração de conhecimento em base de dados, reconhecendo padrões complexos a partir da modelagem de fenômenos do mundo real [25]. O processo é iterativo e cíclico, ou seja, as etapas são feitas de forma sequencial, podendo haver retorno a etapas anteriores para implementar ideias que surgiram durante o processo. Os resultados esperados usualmente consistem de conhecimentos relevantes para auxiliarem na tomada de decisão por parte de gestores.

A metodologia KDD possui cinco etapas principais: a *Seleção e Extração dos Dados*, onde a base de dados é coletada; o *Pré-Processamento dos Dados*, que consiste na limpeza e formatação dos dados, removendo ruídos e tomando decisões a cerca de atributos com dados ausentes; a *Transformação dos Dados*, que está relacionada a engenharia de atributos, onde os dados são estruturados e organizados, de forma a gerar modelos que possuam recursos úteis para os algoritmos; a *Mineração de Dados*, que consiste no treinamento dos algoritmos com os dados transformados; e por último, a *Interpretação dos Resultados*, onde as métricas são analisadas e os conhecimentos são consolidados.

IV. Aplicação das Etapas do KDD

Nesta seção são apresentadas as etapas da metodologia adotada no trabalho, a KDD, descrevendo de forma técnica a implementação dos passos aplicados ao problema de classificação de evasão no ensino superior. Todos os processos foram implementados na linguagem de programação Python 3[1] devido a simplicidade de sua sintaxe, e a variedade de bibliotecas de manipulação e visualização de dados (NumPy[2], Pandas[3], Matplotlib[4], Seaborn[5]) e de aprendizado de máquina (Scikit-learn[6]) disponíveis. Os códigos utilizados estão disponíveis no GitHub[7].

A. Seleção e Extração dos Dados

A Universidade de Brasília possui diversas informações a cerca dos seus estudantes, como dados sociodemográficos e de desempenho nas disciplinas, contudo ainda não é possível ter acesso aos dados de todos os alunos da Universidade para o propósito desta pesquisa, apenas de cursos específicos. A base de dados utilizada neste trabalho é referente ao curso de Engenharia Mecatrônica da UnB, possui dados anonimizados de 1620 alunos que ingressaram entre 1997 a 2019, e foi extraída pelo coordenador do curso utilizando um sistema integrado ao banco de dados da Universidade.

A base de dados extraída inicialmente possui 12 atributos e 1620 alunos únicos. Cada linha da base representa o desempenho de um aluno em uma disciplina em um determinado período letivo. Dentre os atributos da base que possuem informações sobre o aluno, temos: um identificador único anonimizado, o *Sexo*, o *Tipo de Escola* que o aluno frequentou, o *CEP*, a *Forma de Ingresso*, o *Período de Ingresso*, a *Forma de Saída* e o *Período de Saída* da Universidade. Dentre os atributos referentes ao desempenho na disciplina, temos: o *Período* em que a disciplina foi cursada, o *Código da Disciplina*, o *Nome da Disciplina* e a *Menção* obtida pelo aluno. Para que seja possível identificar as disciplinas do curso e suas respectivas cargas horárias (créditos), é gerado um arquivo *JSON* relacionando o código das disciplinas às suas respectivas informações, extraídas do antigo portal acadêmico Matricula Web[8] da UnB.

B. Pré-Processamento dos Dados

Para que os dados possam ser utilizados no treinamento dos algoritmos, eles devem passar por um processo de limpeza, removendo ruídos e tratando atributos com dados ausentes.

Inicialmente são removidos os alunos ativos e que vieram a falecer, já que não podem ser classificados como *evadido* e *não-evadido*, representando 443 dos alunos da base. O atributo *Tipo de Escola* possui três valores diferentes, *"Pública"*,

[1] https://www.python.org/
[2] https://numpy.org/
[3] https://pandas.pydata.org/
[4] https://matplotlib.org/
[5] https://seaborn.pydata.org/
[6] https://scikit-learn.org/stable/
[7] https://github.com/gnramos/trj-academica
[8] https://matriculaweb.unb.br/

"Particular" e *"Não informado"*, contudo dentre os 1620 alunos, 618 não possuem esse atributo informado, representando 38% dos alunos, e pela dificuldade de lidar com um dado com muitos valores ausentes, esse atributo é removido da análise da predição. Alguns dos *CEP*s da base de dados estão incorretos, mal formatados, ou informavam o próprio *CEP* da Universidade de Brasília, provavelmente preenchido de forma errada no registro do aluno. Cerca de 123 alunos possuem o *CEP* incorreto, esses são removidos da base de dados, para facilitar o uso desse atributo.

Como são utilizados apenas os dados acadêmicos do primeiro ano da graduação, todas as disciplinas que foram realizadas após esse período são retiradas da base. Para verificar se uma disciplina foi cursada no primeiro ano, os atributos de *Período de Ingresso* e *Período* em que a disciplina foi cursada são transformados em valores numéricos, a diferença entre esses valores representa o período em que a disciplina foi cursada em relação ao ingresso no curso. As disciplinas de verão são consideradas como cursadas no primeiro semestre do mesmo ano.

As disciplinas obrigatórias são aquelas necessárias para se formar no curso, que estão presentes no componente curricular, e as não obrigatórias são disciplinas extras na formação do aluno. As disciplinas não obrigatórias são retiradas da base de dados por possuírem poucas ocorrências no primeiro ano de curso. São também removidos estudantes de outras universidades que apenas cursaram matérias de verão, assim como o atributo *Nome da Disciplina*, que possui redundância com o *Código da Disciplina*, e o atributo *Período de Saída*, por não ser utilizado. Após a limpeza dos dados, a base possui 9 atributos e 921 alunos restantes, sem dados ausentes, pronta para a etapa de transformação.

C. Transformação dos Dados

A engenharia de atributos é o processo de utilizar conhecimento do domínio para extrair atributos relevantes dos dados [8]. É nessa transformação que os dados são estruturados e organizados, de forma a gerar modelos que possuam uma estrutura correta e recursos úteis para a identificação de evasão pelos algoritmos. Neste trabalho foram gerados alguns modelos utilizando diferentes atributos e transformações nos dados, para testar quais obtiveram os melhores resultados. A seguir, são descritos os processos envolvidos na transformação dos atributos e os modelos gerados, e na Seção V são apresentados os resultados do processamento dos algoritmos para cada um dos modelos criados.

A variável de saída é o que se deseja prever, que pode ser estimada através das variáveis de entrada. Para o problema de evasão, a variável de saída é a *Forma de Saída*, que informa como o aluno saiu da Universidade, situação que se deseja estimar. Neste trabalho é utilizada a definição de evasão binária, classificando um aluno como evadido caso o atributo *Forma de Saída* informe que o aluno não se formou no curso de origem. Esse atributo possui 13 valores categóricos diferentes, desconsiderando alunos ativos e falecidos. Os únicos alunos considerados como *não-evadido* são aqueles com o valor *"Formatura"*. Todos os outros alunos são classificados como *evadido*, incluindo alunos desligados por desempenho acadêmico, como não cumprir condição ou reprovar a mesma disciplina três vezes, alunos que realizaram transferência do curso/habilitação, alunos desligados por abandono, entre outros. Um novo atributo binário é gerado utilizando essa definição de evasão e será utilizado no treinamento dos algoritmos. As variáveis de entrada neste trabalho podem ser divididas em três grupos de atributos: os *Atributos das Disciplinas*, os *Atributos Sociodemográficos* e os *Atributos Acadêmicos*, descritos a seguir.

Os *Atributos das Disciplinas* dizem respeito ao desempenho dos alunos nas disciplinas cursadas, informação importante para a classificação dos estudantes por parte dos algoritmos. Na base de dados, cada linha representa uma disciplina cursada, por um aluno em um certo período letivo. É necessário transformar essas informações em variáveis de entrada, ou seja, em atributos. Para isso, são elaboradas três transformações diferentes, com a intenção de testar o impacto de diferentes abordagens.

A primeira transformação consiste em gerar atributos binários para cada ocorrência de disciplina, período letivo e menção obtida, abordagem semelhante a realizada no trabalho [16]. Esse processo é conhecido como *One Hot Encoding*, e permite que atributos categóricos sejam transformados em atributos binários, ou seja, cada combinação de disciplina, período letivo e menção se torna um atributo binário na base de dados.

A segunda transformação não utiliza um atributo binário por menção obtida, como a primeira transformação. Nessa são gerados apenas dois atributos por disciplina/período, uma informando se o aluno foi aprovado e a outra se o aluno foi reprovado. Nesses atributos são utilizados uma escala numérica de importância da menção obtida em relação a aprovação/reprovação do aluno, inserindo um conhecimento de domínio na construção do atributo, que auxilia no aprendizado dos algoritmos. A escala consiste de um valor numérico de 1 até 3, no atributo de aprovado (MM/CC, MS e SS), onde as menções relativas a desempenhos melhores possuem um valor numérico maior, já no atributo de reprovado (MI, II e SR) as menções relativas a desempenhos piores possuem um valor numérico maior, e o valor nulo representa trancamento de disciplina, que o aluno não a cursou nesse semestre ou que ele não tenha obtido uma menção referente a uma aprovação/reprovação.

Por último, a terceira transformação busca simplificar a abordagem utilizada na segunda transformação. Em vez de utilizar dois atributos por disciplina/período, é gerado apenas um único atributo, representando o desempenho do aluno na disciplina, com valores positivos para as menções de aprovação e negativos para as de reprovação. Não ter realizado a disciplina ou ter a trancado são representados por um valor nulo. Para que disciplinas que o aluno já obteve aprovação não apresentem valores nulos nos períodos seguintes, todos os valores dos atributos representam a menção máxima obtida até esse certo período.

Os *Atributos Sociodemográficos* são gerados a partir de três dados diferentes: o *Sexo*, a *Forma de Ingresso* e o *CEP*. O *Sexo* possui dois valores categóricos: *"Feminino"* e *"Masculino"*, que são transformados em valores binários, verdadeiro e falso, respectivamente, pois grande parte dos classificadores apenas trabalham com valores numéricos e não categóricos. O atributo *Forma de Ingresso* possui 11 valores categóricos diferentes, contudo grande parte dos alunos da base ingressaram na Universidade pelo *"Vestibular"* ou *"Programa de Avaliação Seriada"* (PAS), cerca de 94%. Por isso, esse atributo é transformado em três atributos binários: o primeiro informando se o aluno ingressou pelo Vestibular, o segundo pelo PAS, e o último, por algum dos outros meios.

O atributo *CEP* é utilizado para gerar dois novos atributos: a *distância da residência à Universidade*, e um atributo binário informando se o aluno é de fora do Distrito Federal (DF). É utilizado o serviço *web* ViaCEP[9] para extrair informações dos valores de *CEP*, como o endereço, e posteriormente é utilizado o banco de dados geográfico do serviço de localização Google Maps[10], para calcular a distância do endereço até a UnB. Os alunos com mais de 50 quilômetros de distância da Universidade são considerados de fora do DF e suas distâncias são consideradas como 50 quilômetros. Como a localização da residência utilizando o serviço do Google Maps possui certa imprecisão, os valores de distância foram quantizados entre valores de 0 a 10, onde 10 representa todos os alunos considerados de fora do DF.

Três novos *Atributos Acadêmicos* foram gerados a partir dos dados de desempenho nas disciplinas. O primeiro deles é o IRA[11] (Índice de Rendimento Acadêmico), métrica utilizada pela UnB que se baseia em uma média ponderada das menções nas disciplinas com suas cargas horárias, um valor numérico que resume o desempenho acadêmico do aluno. Foram gerados atributos com o IRA parcial de cada um dos dois semestres letivos analisados. O segundo e o terceiro atributo representam a quantidade de créditos aprovados e de créditos reprovados, respectivamente, para os dois semestres separadamente, e para os dois semestres juntos.

Utilizando os atributos transformados, são gerados cinco modelos diferentes, para testar quais transformações e atributos geraram os melhores resultados. Os Modelos I, II e III utilizam, cada um, respectivamente, a primeira, a segunda e a terceira transformação dos *Atributos das Disciplinas*. O Modelo IV inclui os *Atributos Sociodemográficos* no modelo que obteve os melhores resultados dentre I, II e III. E por último, o Modelo V inclui os *Atributos Acadêmicos* ao Modelo IV.

D. Mineração de Dados

Esta etapa consiste no treinamento dos algoritmos, que utilizam os dados recém transformados para aprender a identificar e classificar alunos com risco de evasão. Para isso, os dados são divididos em dois grupos: de treino, responsável pelo aprendizado do algoritmo e que possui 70% dos dados base, e de teste, responsável por testar a capacidade de generalização do algoritmo treinado, com 30% dos dados. A divisão é feita de forma estratificada, para que a proporção de elementos *evadido* e *não-evadidos* continue a mesma nos dois grupos.

É utilizada a técnica *Grid Search* para a otimização dos hiper-parâmetros dos algoritmos, para que seja possível identificar a configuração que gera os melhores resultados. Os parâmetros testados em cada um dos algoritmos são descritos na Tabela I. O uso dos parâmetros certos é importante para evitar o sobre-ajuste, que é comum em alguns algoritmos de mineração de dados em contextos complexos e incertos como o problema da evasão. É possível destacar a escolha do parâmetro *max_depth* para os algoritmos *Floresta Aleatória* e *XGBoost*, onde foram testados apenas valores pequenos, limitando a profundidade dessas estruturas e contendo o sobre-ajuste.

Tabela I: Parâmetros do GridSearch.

	Parâmetros	Valores
Floresta Aleatória	'criterion' 'n_estimators' 'max_features' 'max_depth'	['gini', 'entropy'] [25, 50, 75, 100] ['auto', 'sqrt'] [4, 5, 7]
XGBoost	'min_child_weight' 'n_estimators' 'max_depth'	[1, 5, 10] [20, 25, 35] [2, 6, 10]
Regressão Logistica	'solver'	['liblinear', 'lbfgs']
Árvore de Decisão	'max_depth' 'criterion' 'splitter' 'min_samples_split'	[10, 25, 50, 100] ['gini', 'entropy'] ['best', 'random'] np.linspace(0.1, 1.0, 10)
Redes Neurais	'max_iter' 'solver' 'activation'	[10000] ['lbfgs'] ['identity', 'logistic', 'tanh', 'relu']
SVM	'kernel'	['linear', 'poly', 'rbf', 'sigmoid']

O treinamento é realizado utilizando a técnica da *Validação Cruzada 5-fold*, para gerar predições que generalizem bem o problema e também não sofram de sobre-ajuste. A métrica otimizada no *Grid Search* é a F_β, utilizando o parâmetro $\beta = 1,5$, indicando que a métrica sensibilidade é $1,5$ vezes mais importante que a precisão para este trabalho. Esse valor é escolhido com base no entendimento do problema da evasão, onde existe uma preferência em identificar verdadeiros positivos, ou seja, alunos evadidos que foram corretamente identificados, ao custo de uma maior quantidade de falsos negativos, alunos não-evadidos identificados como evadidos. Isso se justifica uma vez que a prioridade é identificar alunos com risco de evasão para que eles possam ser acompanhados e aconselhados, mesmo que ao custo de identificar de forma incorreta alunos que não o necessitam.

E. Interpretação dos Resultados

Após o treinamento dos algoritmos, eles são utilizados para classificar os dados de teste, e são extraídas as métricas comparando os resultados obtidos com os resultados esperados. São utilizadas as quatro métricas citadas anteriormente, a

[9]https://viacep.com.br/

[10]https://developers.google.com/maps

[11]http://www.saa.unb.br/acompanhamento-academico/17-avaliacao-de-desempenho-academico

Acurácia, a *Precisão*, a *Sensibilidade*, e por último, a $F_{1,5}$-*Score*. Para possuir um controle maior sobre os resultados, são extraídas dos classificadores probabilidades de cada um dos elementos pertencerem a classe principal (*evadido*), e a partir dessas probabilidades é escolhido o limiar de decisão para classificação, buscando a otimização da métrica $F_{1,5}$-*Score*, que prioriza a métrica sensibilidade sobre a precisão. A partir das probabilidades, é também possível gerar Curvas de Precisão-Sensibilidade, que auxiliam na visualização comparativa do desempenho dos algoritmos.

São extraídas dos algoritmos as *Feature Importances*, que informam para cada atributo, de forma geral, o quão importante ele é para a predição. A partir dessa informação é possível identificar quais atributos possuem maior impacto na evasão dos estudantes. Os testes são realizados para cada um dos modelos gerados na transformação dos dados, e é gerado um comparativo dos resultados obtidos.

V. Análise dos Resultados

Os cinco modelos de dados gerados foram estruturados de forma a incrementar gradualmente as abordagens propostas neste trabalho nos testes, para que seja possível analisar o impacto dessas melhorias ao serem inseridas no modelo, comparando as métricas obtidas pelos algoritmos. Para gerar os resultados, foi extraída, para cada um dos cinco modelos, uma tabela com as quatro métricas analisadas no trabalho, obtidas por cada um dos seis algoritmos de mineração de dados. Os resultados podem ser vistos nas Tabelas II a VI.

Tabela II: Resultados do Modelo I.

	Acurácia	Precisão	Sensibilidade	$F_{1,5}$-Score
Floresta Aleatória	81,11%	75,50%	84,44%	81,47%
XGBoost	74,92%	65,76%	89,63%	80,63%
Regressão Logistica	73,62%	66,46%	80,74%	75,73%
Árvore de Decisão	74,92%	66,48%	86,67%	79,26%
Redes Neurais	54,40%	49,06%	96,30%	74,29%
SVM	84,36%	82,71%	81,48%	**81,85%**

Tabela III: Resultados do Modelo II.

	Acurácia	Precisão	Sensibilidade	$F_{1,5}$-Score
Floresta Aleatória	83,71%	79,31%	85,19%	**83,29%**
XGBoost	80,13%	73,42%	85,93%	81,65%
Regressão Logistica	76,22%	67,61%	88,15%	80,61%
Árvore de Decisão	77,52%	70,62%	83,70%	79,19%
Redes Neurais	76,22%	67,82%	87,41%	80,27%
SVM	79,80%	72,12%	88,15%	82,51%

Os Modelos I, II e III diferem apenas pela transformação realizada nos *Atributos das Disciplinas*. É possível perceber que as métricas obtidas são maiores nos Modelos II e III, pois ao contrário do Modelo I, que gera um atributo binário para cada menção, eles utilizam uma escala para representar as menções obtidas, indicando que o uso de uma escala na transformação desses atributos possui um impacto positivo nos resultados. As métricas dos Modelos II e III são similares, mas por possuir apenas um atributo por disciplina, o Modelo III é escolhido como o melhor dentre os três modelos, e será utilizado nos testes dos modelos seguintes.

Tabela IV: Resultados do Modelo III.

	Acurácia	Precisão	Sensibilidade	$F_{1,5}$-Score
Floresta Aleatória	85,34%	83,09%	83,70%	**83,51%**
XGBoost	79,48%	71,95%	87,41%	81,99%
Regressão Logistica	84,69%	82,35%	82,96%	82,77%
Árvore de Decisão	71,34%	62,18%	88,89%	78,51%
Redes Neurais	84,69%	81,88%	83,70%	83,14%
SVM	83,06%	78,62%	84,44%	82,56%

Tabela V: Resultados do Modelo IV.

	Acurácia	Precisão	Sensibilidade	$F_{1,5}$-Score
Floresta Aleatória	85,20%	79,84%	87,29%	**84,85%**
XGBoost	84,48%	78,63%	87,29%	84,43%
Regressão Logistica	79,78%	70,39%	90,68%	83,29%
Árvore de Decisão	79,78%	72,46%	84,75%	80,55%
Redes Neurais	79,42%	69,93%	90,68%	83,09%
SVM	82,67%	75,74%	87,29%	83,37%

Tabela VI: Resultados do Modelo V.

	Acurácia	Precisão	Sensibilidade	$F_{1,5}$-Score
Floresta Aleatória	86,64%	81,89%	88,14%	**86,11%**
XGBoost	83,75%	78,29%	85,59%	83,21%
Regressão Logistica	81,95%	75,00%	86,44%	82,57%
Árvore de Decisão	83,75%	80,17%	82,20%	81,57%
Redes Neurais	80,87%	75,19%	82,20%	79,91%
SVM	85,56%	80,47%	87,29%	85,07%

Os Modelos IV e V utilizam as transformações realizadas no Modelo III, e incluem os *Atributos Sociodemográficos* nos dois modelos, e os *Atributos Acadêmicos* apenas no Modelo V. É possível perceber uma melhoria nas métricas encontradas ao inserir esses dois conjuntos de atributos, indicando que eles possuem um impacto positivo na predição de evasão, comparando com os resultados obtidos no Modelo III. O melhor resultado obtido neste trabalho utiliza o algoritmo de Floresta Aleatória treinado com os dados do Modelo V, e alcançou 86,11% de $F_{1,5}$-Score, a métrica otimizada, 88,14% de Sensibilidade, 86,64% de Acurácia e 81,89% de Precisão.

Para visualizar e comparar os resultados dos algoritmos, a partir das probabilidades extraídas foram gerados dois gráficos de Precisão-Sensibilidade, para o Modelo I, que é semelhante a abordagem do trabalho passado, e para o Modelo V, que possui

todas as melhorias propostas neste trabalho, para comparação. Cada ponto na curva de Precisão-Sensibilidade representa diferentes limiares de classificação entre os algoritmos testados, com seus respectivos valores de sensibilidade e precisão obtidos na predição ao utilizar esse limiar. Os gráficos podem ser vistos nas Figuras 1 e 2.

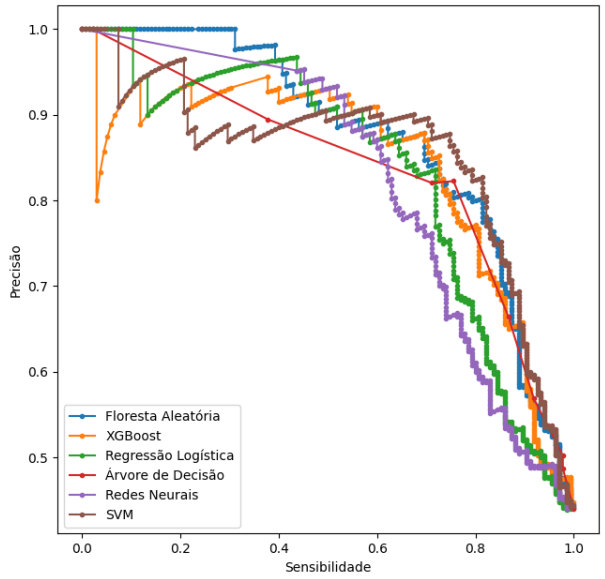

Figura 1: Curva Precisão-Sensibilidade do Modelo I.

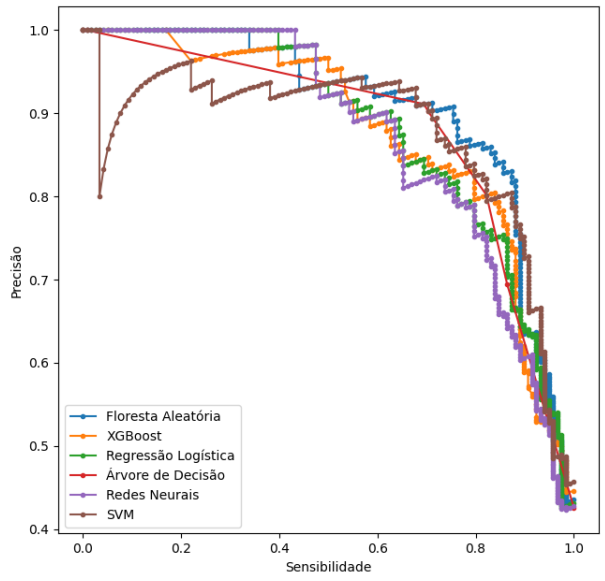

Figura 2: Curva Precisão-Sensibilidade do Modelo V.

É possível perceber que as curvas do Modelo V são mais acentuadas que as do Modelo I para todos os algoritmos. Isso significa que para valores altos da métrica de sensibilidade, os algoritmos também alcançam valores altos da métrica de precisão, em especial o algoritmo de *Floresta Aleatória*, representado pela cor azul no gráfico. O limiar de classificação utilizado para a predição é escolhido a partir de um dos pontos dessas curvas, visando um balanço entre as duas métricas, mas com um peso maior para a sensibilidade, justificando a escolha do limiar de classificação otimizando-se a métrica $F_{1,5}$-Score.

Para obter conhecimento a cerca dos padrões compreendidos pelos algoritmos, foram extraídas as *Feature Importances* para dois algoritmos que obtiveram bons resultados para o Modelo V, a *Floresta Aleatória* e o *XGBoost*. Apesar do *SVM* ter alcançado melhores resultados que o *XGBoost*, o mesmo utiliza o *Kernel RBF* (Radial), que não permite a extração dessas informações, logo, para o segundo algoritmo foi escolhido o *XGBoost*, que também alcançou bons resultados e que permite a extração das *Feature Importances*. As informações obtidas podem ser vistas nas Figuras 3 e 4, e o código das disciplinas, para compreensão dos dados, na Tabela VII.

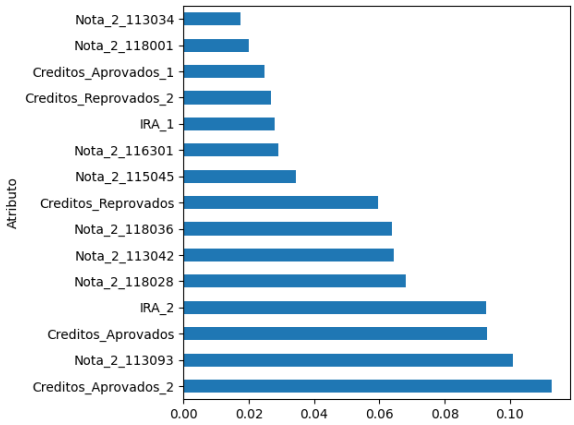

Figura 3: *Feature Importance* da *Floresta Aleatória*.

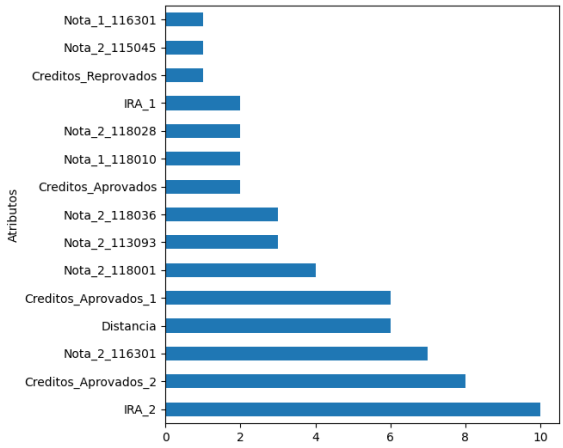

Figura 4: *Feature Importance* do *XGBoost*.

É possível perceber que os atributos mais importantes para ambos os algoritmos são *Atributos Acadêmicos*: o *Creditos_Aprovados_2*, para a *Floresta Aleatória*, que representa a quantidade de créditos aprovados no segundo semestre, e o atributo *IRA_2*, para o *XGBoost*, que representa o IRA

Tabela VII: Código das Disciplinas.

Código	Nome
113034	Cálculo 1
118001	Física 1
116301	Computação Básica
115045	Probabilidade e Estatística
118028	Física 2
113093	Álgebra Linear
118010	Física 1 Experimental
118036	Física 2 Experimental

do aluno no segundo semestre. Isso reforça a importância desses atributos para os bons resultados obtidos por parte dos algoritmos, que fazem parte da engenharia de atributos introduzida neste trabalho.

Alguns dos atributos dizem respeito ao desempenho em certas disciplinas. Para a *Floresta Aleatória*, por exemplo, o segundo atributo mais importante é o *Nota_2_113093*, informando o desempenho do aluno na disciplina de "Álgebra Linear" no segundo semestre. A importância de um atributo de disciplina não significa necessariamente que um desempenho ruim está relacionado com a evasão do aluno, é possível que, por exemplo, uma menção boa em uma disciplina aumente a probabilidade do aluno permanecer no curso, mas que não necessariamente uma menção ruim esteja relacionada a sua evasão. Independentemente, é possível utilizar dessas informações para identificar quais disciplinas estão relacionadas com a evasão de alunos, e tomar medidas para mitigar esse impacto negativo, assim como identificar disciplinas que, quando concluídas com uma boa menção, são um bom indicativo de que o aluno provavelmente não vai evadir.

Alguns dos *Atributos Sociodemográficos*, como o *Sexo* e a *Forma de Entrada*, não apareceram entre os atributos mais importantes para a predição, ou seja, indicando que nos modelos gerados, as mulheres se formam tanto quanto os homens, e que o método de entrada, seja por Vestibular, PAS, ENEM, etc, não esta relacionado com a evasão dos alunos.

A *Distância da residência à Universidade*, por outro lado, foi o quarto atributo mais importante para o algoritmo *XGBoost*, indicando que essa distância pode ter uma relação com a evasão ou permanência de alunos no curso. Foi gerado um gráfico que mostra a quantidade de alunos em relação a distância e a definição de evasão, para visualizar a distribuição desses alunos dentro do atributo distância, exibido na Figura 5.

É possível perceber que a proporção entre alunos que evadiram e que não evadiram aumenta a medida que a distância da residência à Universidade aumenta, ou seja, quanto maior a distância maior a proporção de evadidos. Uma maior distância da residência do aluno à Universidade implica em maiores dificuldades relacionadas a locomoção, como o tempo gasto, e isso pode ser uma das causas de evasão, reforçando a importância de um transporte público de qualidade na formação dos estudantes da Universidade. É possível também que o dado da distância esteja relacionado a um fator externo não considerado nessa análise, como a situação econômica do aluno, que também pode ou não estar relacionado a evasão.

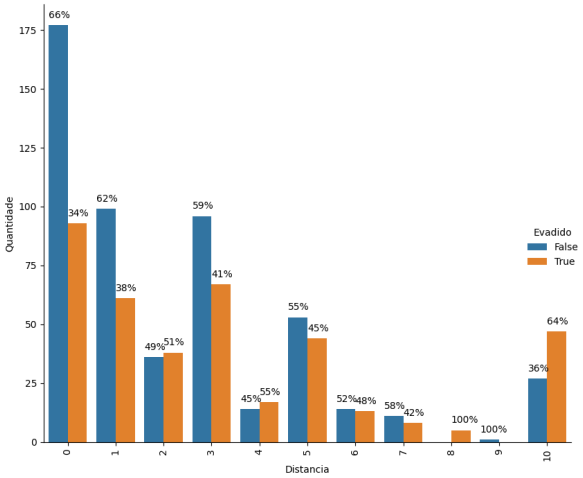

Figura 5: Gráfico Distância/Evasão.

VI. Conclusões Alcançadas e Melhorias Futuras

Dispor de uma ferramenta que oferece suporte para a predição de casos de evasão e para a identificação de suas possíveis causas, é de extremo interesse das Instituições de Ensino Superior que sofrem com índices de evasão acima do esperado, incluindo a Universidade de Brasília. A Mineração de Dados oferece uma abordagem para este problema e vem sendo desenvolvida em diversos trabalhos recentes, dentro e fora da Universidade de Brasília, com o objetivo de obter melhores predições de casos de evasão no ensino superior.

Este trabalho propôs algumas melhorias a uma abordagem sendo desenvolvida atualmente na Universidade de Brasília, incluindo a análise do impacto de dados Sociodemográficos na predição, o desenvolvimento de uma nova engenharia de atributos e o uso de dois novos algoritmos no processamento dos dados, realizando a escolha do limiar de classificação e otimizando seus parâmetros para evitar o sobre-ajuste.

Os dados Sociodemográficos *Sexo* e *Forma de Entrada* não apresentaram impacto considerável na predição de evasão pelos algoritmos, contudo a distância da residência à Universidade foi apontada pelo algoritmo *XGBoost* como o quarto atributo mais importante para a predição de evasão dos alunos. Os atributos acadêmicos e de disciplinas foram de grande importância para os algoritmos, reforçando a importância da engenharia de atributos para identificar casos de evasão com eficiência. O uso de uma escala para representar as menções resultou em um impacto positivo nas métricas obtidas, indicando que o uso da escala oferece aos algoritmos um conhecimento do domínio útil para auxiliar na predição.

Dentre os cinco modelos gerados, o que obteve as melhores métricas foi o quinto modelo, que possui todas as melhorias propostas neste trabalho. Os dois algoritmos utilizados, *XGBoost* e *SVM*, obtiveram bons resultados em todos os modelos. Contudo, o algoritmo de *Floresta Aleatória*, que foi otimizado para evitar o sobre-ajuste na predição, obteve os melhores resultados no geral. O melhor resultado obtido

neste trabalho foi utilizando o algoritmo *Floresta Aleatória*, treinado com o quinto modelo de dados, que obteve 88,14% de Sensibilidade e 86,64% de Acurácia, métricas melhores que as obtidas com o primeiro modelo, onde não são aplicadas as melhorias propostas neste trabalho, e que as obtidas em trabalhos passados [16].

Informações como as obtidas neste trabalho servem de base para a implementação de politicas por parte da gestão, para que seja possível reduzir os problemas enfrentados pelo aluno em sua formação, melhorando a qualidade e a eficiência do ensino e aprendizado na Universidade. As métricas obtidas e a relevância das informações extraídas dos algoritmos comprovam a eficiência que abordagem de Mineração de Dados Educacional possui, tanto para a identificação prematura de alunos em risco de evasão, quanto para obter conhecimento a cerca de possíveis causas relacionadas a essas evasões.

A predição de evasão é um problema complexo, vasto e pouco explorado até o momento, e este trabalho propôs apenas algumas das possíveis melhorias e abordagens para o problema. Dito isso, existem diversos caminhos para dar continuidade a essa pesquisa dentro do escopo da Universidade de Brasília. O primeiro deles é a realização de uma análise mais profunda no impacto da distância da residência à Universidade na predição de evasão, um dado gerado neste trabalho e identificado como um dos atributos mais importantes para a predição em um dos algoritmos. É possível ainda realizar uma análise encima de cada uma das disciplinas individualmente.

Outro caminho é a utilização dados de outros cursos da Universidade, para analisar o poder de generalização dos modelos desenvolvidos em dados de outros cursos, ou para gerar modelos específicos, identificando causas e possíveis problemas estruturais de outros cursos. Por último, é possível testar modelos de dados utilizando novas engenharias de atributos, visto seu impacto na predição dos algoritmos, ou utilizar novos dados sociodemográficos, como o tipo de escola do aluno no ensino médio, renda familiar, idade, entre outros.

Referências

[1] C. A. D. Santos Baggi and D. A. Lopes, "Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica," *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, vol. 16, pp. 355 – 374, 07 2011. [Online]. Available: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772011000200007&nrm=iso

[2] M. Lobo, "Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. 2012," *Associação Brasileira de Mantenedoras de Ensino Superior. Cadernos*, vol. 25, 2012.

[3] R. L. L. e. Silva Filho, P. R. Motejunas, O. Hipólito, and M. B. d. C. M. Lobo, "A evasão no ensino superior brasileiro," *Cadernos de Pesquisa*, vol. 37, pp. 641 – 659, 12 2007. [Online]. Available: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-15742007000300007&nrm=iso

[4] M. M. Braga, M. d. C. L. Peixoto, and T. F. Bogutchi, "A evasÃo no ensino superior brasileiro: O caso da ufmg," *Avaliação: Revista da Avaliação da Educação Superior*, vol. 8, no. 3, 1 1. [Online]. Available: http://periodicos.uniso.br/ojs/index.php/avaliacao/article/view/1237

[5] L. Hagedorn, "How to define retention: A new look at an old problem." 2006.

[6] L. S. Aulck, D. Nambi, N. Velagapudi, J. Blumenstock, and J. D. West, "Mining university registrar records to predict first-year undergraduate attrition," in *EDM*, 2019.

[7] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 40, pp. 601 – 618, 12 2010.

[8] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, "Appendix a - theoretical foundations," in *Data Mining (Fourth Edition)*, fourth edition ed. Morgan Kaufmann, 2017, pp. 533–552. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128042915000234

[9] R. GOLDSCHMIDT and E. PASSOS, *Data mining: um guia Prático*. Elsevier Editora, 2005. [Online]. Available: https://books.google.com.br/books?id=JJYHNrREwyEC

[10] T. M. Mitchell *et al.*, "Machine learning. 1997," *Burr Ridge, IL: McGraw Hill*, vol. 45, no. 37, pp. 870–877, 1997.

[11] A.-S. Hoffait and M. Schyns, "Early detection of university students with potential difficulties," *Decision Support Systems*, vol. 101, pp. 1–11, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923617300817

[12] F. Jiménez, A. Paoletti, G. Sánchez, and G. Sciavicco, "Predicting the risk of academic dropout with temporal multi-objective optimization," *IEEE Transactions on Learning Technologies*, vol. 12, no. 2, pp. 225–236, 2019.

[13] U. de Brasília, "Plano de desenvolvimento institucional 2018-2022 da universidade de brasília," http://www.planejamentodpo.unb.br/index.php?option=com_content&view=article&id=20&Itemid=791, acessado: 19-04-2021.

[14] R. M. HOED, "Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de computação," p. 164, 2017. [Online]. Available: https://repositorio.unb.br/handle/10482/22575

[15] L. R. S. d. ASSIS, "Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados," p. 134, 2018. [Online]. Available: https://repositorio.unb.br/handle/10482/32139

[16] W. G. M. GRAMACHO, "Algoritmos de mineração de dados para análise de evasão na graduação da universidade de brasília," p. 55, 2020. [Online]. Available: https://bdm.unb.br/handle/10483/24527

[17] G. F. SILVA, "Análise preditiva do desempenho acadêmico de alunos de graduação da unb utilizando mineração de dados," p. 52, 2017. [Online]. Available: https://bdm.unb.br/handle/10483/17910

[18] A. O. E. and D. C., "Roads to success in the belgian french community's higher education system: Predictors of dropout and degree completion at the université libre de bruxelles." *Res High Educ*, vol. 54, p. 693–723, 2013. [Online]. Available: https://link.springer.com/article/10.1007%2Fs11162-013-9290-y

[19] J. Vandamme, N. Meskens, and J. Superby, "Predicting academic performance by data mining methods," *Education Economics*, vol. 15, no. 4, pp. 405–419, 2007. [Online]. Available: https://doi.org/10.1080/09645290701409939

[20] D. Delen, "A comparative analysis of machine learning techniques for student retention management," *Decision Support Systems*, vol. 49, no. 4, pp. 498–506, 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923610001041

[21] F. Araque, C. Roldán, and A. Salguero, "Factors influencing university drop out rates," *Computers & Education*, vol. 53, no. 3, pp. 563–574, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131509000815

[22] J. Han, M. Kamber, and J. Pei, "8 - classification: Basic concepts," in *Data Mining (Third Edition)*, third edition ed., ser. The Morgan Kaufmann Series in Data Management Systems, J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 327–391. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123814791000083

[23] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, vol. 1. Springer-Verlag London, UK, 2000.

[24] "Data mining software, model development and deployment, sas enterprise miner." [Online]. Available: https://www.sas.com/en_us/software/enterprise-miner.html

[25] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI magazine*, vol. 17, no. 3, pp. 37–37, 1996.