PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS APLICADOS À ENGENHARIA E GESTÃO

ÁTILA CARVALHO JÚNIOR

PROPOSTA DE UM MÉTODO DE PREDIÇÃO DA EVASÃO ESTUDANTIL NO ENSINO SUPERIOR BASEADO EM UM FRAMEWORK ESCOLAR UTILIZANDO MINERAÇÃO DE DADOS

Campos dos Goytacazes/RJ

2022

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA FLUMINENSE

PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS APLICADOS À ENGENHARIA E GESTÃO

ÁTILA CARVALHO JÚNIOR

PROPOSTA DE UM MÉTODO DE PREDIÇÃO DA EVASÃO ESTUDANTIL NO ENSINO SUPERIOR BASEADO EM UM FRAMEWORK ESCOLAR UTILIZANDO MINERAÇÃO DE DADOS

Aline Pires Vieira de Vasconcelos

(Orientadora)

Jonnathan dos Santos Carvalho

(Co-orientador)

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação do Instituto Federal de Educação, Ciência e Tecnologia Fluminense, no Curso de Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão (MPSAEG), como parte dos requisitos necessários à obtenção do título de Mestre em Sistemas Aplicados à Engenharia e Gestão.

Campos dos Goytacazes/RJ

2022

Biblioteca Anton Dakitsch
CIP - Catalogação na Publicação

J95p

Júnior, Átila Carvalho
PROPOSTA DE UM MÉTODO DE PREDIÇÃO DA EVASÃO ESTUDANTIL NO ENSINO SUPERIOR BASEADO EM UM FRAMEWORK ESCOLAR UTILIZANDO MINERAÇÃO DE DADOS / Átila Carvalho Júnior - 2022.
50 f.: il. color.

Orientadora: Aline Pires Vieira Vasconcelos
Coorientador: Jonnathan dos Santos Carvalho

Dissertação (mestrado) -- Instituto Federal de Educação, Ciência e Tecnologia Fluminense, Campus Campos Centro, Curso de Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão, Campos dos Goytacazes, RJ, 2022.
Referências: f. 47 a 50.

1. Evasão Escolar. 2. Mineração de Dados. 3. Modelo Preditivo. I. Vasconcelos, Aline Pires Vieira, orient. II. Carvalho, Jonnathan dos Santos, coorient. III. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da Biblioteca Anton Dakitsch do IFF com os dados fornecidos pelo(a) autor(a).

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA FLUMINENSE

PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS APLICADOS À ENGENHARIA E GESTÃO

ÁTILA CARVALHO JÚNIOR

PROPOSTA DE UM MÉTODO DE PREDIÇÃO DA EVASÃO ESTUDANTIL NO ENSINO SUPERIOR BASEADO EM UM FRAMEWORK ESCOLAR UTILIZANDO MINERAÇÃO DE DADOS

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação do Instituto Federal de Educação, Ciência e Tecnologia Fluminense, no Curso de Mestrado Profissional em Sistemas Aplicados à Engenharia e Gestão (MPSAEG), como parte dos requisitos necessários à obtenção do título de Mestre em Sistemas Aplicados à Engenharia e Gestão.

Aprovada em 30 de Setembro de 2022.

Banca examinadora:

gov.br Documento assinado digitalmente
ALINE PIRES VIEIRA DE VASCONCELOS
Data: 18/10/2022 10:34:23-0300
Verifique em https://verificador.iti.br

Aline Pires Vieira de Vasconcelos
Doutora em Engenharia de Sistemas e Computação – IFFluminense
(Orientadora)

gov.br Documento assinado digitalmente
JONNATHAN DOS SANTOS CARVALHO
Data: 17/10/2022 19:46:57-0300
Verifique em https://verificador.iti.br

Jonnathan dos Santos Carvalho
Doutor em Computação – IFFluminense
(Co-orientador)

gov.br Documento assinado digitalmente
LUIZ GUSTAVO LOURENCO MOURA
Data: 17/10/2022 17:13:20-0300
Verifique em https://verificador.iti.br

Luiz Gustavo Lourenço Moura
Doutor em Engenharia de Sistemas e Computação – IFFluminense

gov.br Documento assinado digitalmente
MARCOS ANTONIO GUERINE RIBEIRO
Data: 17/10/2022 14:34:07-0300
Verifique em https://verificador.iti.br

Marcos Antonio Guerine Rineiro
Doutor em Computação – IFRJ

AGRADECIMENTOS

Primeiramente, gostaria de agradecer a Deus, pois me sustentou em todos os momentos difíceis que enfrentei. Durante o Mestrado surgiu uma pandemia que nos forçou a ter aulas online e encontros apenas à distância, algo que tornou ainda mais especial esta conquista.

Também sou muito grato à minha esposa Karina e meu filho Bernardo por todo apoio nas dificuldades encontradas no percurso e compreensão nos momentos em que não pude estar presente.

Agradeço imensamente aos meus pais, Átila e Adriana, que sempre me ensinaram o caminho correto e me incentivaram a estudar e cada vez mais adquirir conhecimento para crescer tanto como pessoa quanto profissionalmente.

Ao meu padrinho Adriano, que desde pequeno foi um segundo pai, contribuindo em todas as etapas da minha vida.

À minha orientadora Aline e meu coorientador Jonnathan, que tanto me ensinaram e deram todo o suporte necessário em cada dificuldade que encontrei.

RESUMO

A evasão no ensino superior é conceituada como a saída do aluno do curso de graduação de maneira definitiva ou temporária, independentemente do motivo ou causa, sem que tenha sido diplomado. O fenômeno do abandono escolar produz uma série de problemas para o sistema de educação, bem como para o governo, sendo o desperdício de recursos financeiros, sociais e humanos um dos exemplos mais graves. Portanto, a redução da taxa de evasão de alunos é um dos grandes desafios enfrentados pelas instituições públicas de ensino do Brasil na atualidade, fato que prejudica a gestão escolar e torna urgente a busca pela solução desse problema. Nesse contexto, a identificação de padrões e perfis de alunos com maior risco de evasão revela-se essencial para o desenvolvimento de um plano de mitigação voltado à redução da probabilidade de abandono do curso pelo estudante. A detecção precoce de estudantes propensos a abandonar seus cursos é crucial para o sucesso de qualquer estratégia de sucesso escolar. Uma das alternativas é utilizar técnicas de mineração de dados para criar um sistema de alerta que identifique com antecedência estudantes com risco elevado de abandono. Isto posto, o objetivo deste trabalho é desenvolver um método baseado em um *framework* escolar que torne mais simples a consulta acerca da possibilidade de um estudante evadir, a partir da utilização de um algoritmo de mineração de dados que pode ser aplicado numa plataforma Web para se tornar mais eficiente. Para tanto, é utilizada a base de dados acadêmicos de uma instituição federal de ensino com o intuito de executar as etapas de mineração para gerar um modelo de previsão de evasão. Por fim, os resultados são analisados pelos gestores escolares para realizar a tomada de decisão em relação à permanência e êxito. Como resultados, um experimento foi feito, unificando duas bases como estudo de caso para avaliar o desempenho do modelo de classificação. Diferentes algoritmos de classificação foram testados e os modelos gerados com os algoritmos Logistic Regression e Multilayer Perceptron apresentaram melhor desempenho preditivo com base nas medidas de acurácia e F1-score.

Palavras-chave: Evasão Escolar. Mineração de Dados. Modelo Preditivo.

ABSTRACT

Dropout in higher education is conceptualized as the student leaving the undergraduate course permanently or temporarily, regardless of the reason or cause, without having been graduated. The phenomenon of school dropout produces a series of problems for the education system, as well as for the government, being the waste of financial, social and human resources one of the most serious examples. Therefore, reducing the dropout rate of students is one of the great challenges faced by public educational institutions in Brazil today, a fact that impairs school management and makes it urgent to seek a solution to this problem. In this context, the identification of patterns and profiles of students with greater risk of dropout proves to be essential for the development of a mitigation plan aimed at reducing the probability of student dropping out of the course. Early detection of students likely to drop out of their courses is crucial to the success of any school success strategy. One of the alternatives is to use data mining techniques to create an early warning system that identifies students at high risk of dropping out. That said, the objective of this work is to develop a method based on a school framework that makes it simpler to consult about the possibility of a student dropping out, using a data mining algorithm that can be applied on a Web platform to make it more efficient. For this, the academic database of a federal educational institution is used in order to perform the mining steps to generate a dropout prediction model. Finally, the results are analyzed by school administrators to make a decision regarding permanence and success. As a result, an experiment was carried out, unifying two bases as a case study to evaluate the performance of the classification model. Different classification algorithms were tested and the models generated with the Logistic Regression and Multilayer Perceptron algorithms showed better predictive performance based on accuracy measures and F1-score.

Keywords: School Dropout. Machine Learning. Predictive Model.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

KDD – Knowledge Discovery in Database

MD – Mineração de Dados

MDE – Mineração de Dados Educacionais

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses

KNN – K-Nearest Neighbor

DT – Decision Tree

SVM – Support Vector Machine

RF – Random Forest

LR – Logistic Regression

MP – Multilayer Perceptron

AD – Árvore de Decisão

BPMN – Business Process Model and Notation

UML – Unified Modeling Language

WoS – Web of Science

SUMÁRIO

1. INTRODUÇÃO

1.1 Contexto

A evasão no ensino superior é conceituada como a saída do aluno do curso de graduação de maneira definitiva ou temporária, independentemente do motivo ou causa, sem que tenha sido diplomado (COSTA, 1991; MEC, 1996; SOUZA, 1999; SOUZA; OLIVEIRA; GONÇALVES, 2003; LOBO, 2012; SOUZA; DA SILVA; GESSINGER, 2016). O fenômeno do abandono escolar produz uma série de problemas para o sistema de educação, bem como para o governo, sendo o desperdício de recursos financeiros, sociais e humanos um dos exemplos mais graves (COSTA, 1991; SOUZA, 1999; SOUZA *et al*., 2003; SOUZA; DA SILVA; GESSINGER, 2016; ZAGO; PAIXÃO; PEREIRA, 2016). Portanto, a redução da taxa de evasão de alunos é um dos grandes desafios enfrentados pelas instituições públicas de ensino do Brasil, na atualidade, fato que prejudica a gestão escolar e torna urgente a busca pela solução desse problema (PRESTES; FIALHO, 2018).

Segundo a Sinopse Estatística da Educação Superior, publicada pelo INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2018), as universidades públicas brasileiras receberam um total de 2.152.752 alunos em 2017. Por outro lado, de acordo com o levantamento, somente 947.606 estudantes concluíram a graduação no mesmo ano. Uma vez que o número de alunos matriculados aumenta a cada ano (COSTA; DIAS, 2016), a desproporção na razão entrada/saída nas universidades brasileiras pode elevar ainda mais o desperdício de recursos públicos, gerando problemas econômicos, acadêmicos e sociais (FILHO *et al*., 2007; SALES; BALBY; CAJUEIRO, 2016).

A evasão é um grave problema tanto no Brasil quanto no mundo e quando o ensino superior é analisado, verifica-se que o problema atinge instituições públicas e particulares, principalmente no Brasil. Segundo o Censo da Educação Superior do Inep/MEC de 2019 foram contabilizados aproximadamente 8,6 milhões de alunos matriculados na graduação. Apesar do notável número de ingressantes, considerando o período de 2010 a 2019, apenas cerca de 41% deles obtêm o certificado de conclusão. Ou seja, a maioria entra na estatística de abandono da sala de aula. A taxa de desistência é de aproximadamente 59%. Desses, pouco mais de 75% são de instituições privadas. Por consequência, pesquisadores têm analisado constantemente os fatores de influência relacionados a esse abandono escolar.

De acordo com Prestes e Fialho (2018), a evasão acompanha a história educacional. Portanto, é necessário que ela seja um dos principais pontos de programas e planos educacionais que visam à implantação de estratégias de monitoramento e de prevenção com o intuito de buscar meios de estimular a permanência e êxito dos alunos matriculados. Nesse contexto, é necessário que as instituições de ensino superior conheçam seus indicadores de evasão. A partir de análises desses indicadores, é possível identificar causas ou perfis de alunos que possuem taxas de abandono escolar altas. Desse modo, é possível elaborar políticas, atividades e programas tendo como objetivo a permanência e êxito dos estudantes (HOFFMANN *et al.*, 2019).

Segundo Márquez-Vera *et al*. (2016), quanto antes forem identificados os estudantes propensos a evadir, maiores são as chances de sucesso na política de permanência e êxito escolar. Hoed (2016) afirma que é extremamente necessário que a instituição de ensino contenha a evasão, por estar diretamente associada à perda de recursos financeiros. A evasão de um aluno em qualquer nível da educação pode custar não só a ele e sua família, mas também pode deixar marcas no futuro da sociedade, no

crescimento da instituição e do país. Outro ponto é que estudos que permitam detectar alunos propensos a evadir fazem com que seja possível elaborar políticas mais focadas que o incentivem a permanecer e ter sucesso escolar.

Nesse sentido, a identificação de padrões e perfis de alunos com maior risco de evasão revela-se essencial para o desenvolvimento de um plano de mitigação voltado à redução da probabilidade de abandono do curso pelo estudante. Uma das alternativas é utilizar técnicas de mineração de dados no desenvolvimento de sistemas de suporte à decisão que visem alertar, com antecedência, estudantes com risco elevado de abandono (FREITAS *et al.*, 2020; NAGY E MONTOLAY, 2018).

A mineração de dados consiste na aplicação de algoritmos de descoberta de conhecimento que, sob limitações de eficiência computacional aceitáveis, identificam e extraem padrões e regras ou produzem modelos de classificação, ou preditivos, a partir de um conjunto de dados (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Nesse contexto, técnicas de classificação distintas podem ser utilizadas, com o objetivo de desenvolver modelos preditivos capazes de identificar de forma antecipada possíveis casos de evasão, visando facilitar sua mitigação (HUTAGAOL *et al.,* 2019; NAGY e EMONTOLAY, 2018; SANI *et al.,* 2020).

Seguindo esta linha, Barreto (2019) identificou características e padrões de alunos evadidos em diferentes cursos de nível superior, utilizando como estudo de caso o *campus* Campos Centro do Instituto Federal de Educação, Ciência e Tecnologia Fluminense. Dessa forma, foi possível entender os principais fatores responsáveis pelo abandono escolar dos alunos de cada curso. Os cursos de Bacharelado em Sistemas de Informação e Tecnologia em Sistemas de Telecomunicações foram identificados como os cursos com as taxas de evasão mais elevadas. Já o curso de Bacharelado em Arquitetura e Urbanismo possui uma taxa de evasão muito pequena.

Ademais, Barreto (2019) identificou os padrões existentes em relação ao perfil dos alunos que evadiram. Cada curso possui suas particularidades em relação a esse perfil, mas, o ingresso por meio de Vestibular Cota e Sisu Cota, assim como a idade, geralmente inferior a 23 anos, foram as características mais comuns presentes em diferentes cursos. Além disso, apesar das bases de dados terem ficado com 43 atributos após a realização de todos os procedimentos de limpeza, apenas nove atributos aparecerem nas árvores de decisão. São estes: forma de ingresso de matrícula, sexo, idade, coeficiente de rendimento, turno, diferença entre a conclusão do ensino médio e o início da graduação, cor, tipo de escola de origem e estado civil. Destaca-se que a forma de ingresso de matrícula apareceu como raiz em seis de um total de dez cursos, demonstrando a relevância desse atributo na análise da evasão no contexto estudado.

Os resultados podem servir de base para que os gestores escolares atuem nesses pontos, buscando evitar que a evasão ocorra de forma antecipada. Além disso, o método aplicado não é restrito ao *campus* onde a pesquisa foi aplicada. O mesmo pode ser utilizado em outros campi do Instituto Federal de Educação, Ciência e Fluminense, outros Institutos Federais ou outras instituições de ensino, sejam elas públicas ou privadas.

Existem diversos trabalhos que utilizam mineração de dados com o propósito de identificar perfis de alunos mais propensos a evadir (BARRETO, 2019; FREITAS *et al.*, 2020; CAMACHO *et al.*, 2020; HUTAGAOL *et al.,* 2019; FERNÁNDEZ-GARCÍA *et al.,* 2020; Nagy e MONTOLAY, 2018; SANI *et al.*, 2020; PERCHINUNNO *et al.,* 2019), porém em relação a geração de modelos de predição a partir de um *framework* escolar que possibilite a existência de uma plataforma colaborativa para predição de evasão no ensino superior há muito a ser pesquisado e desenvolvido. Dessa forma, este trabalho

propõe um método baseado em um *framework* escolar com o intuito de tornar possível o desenvolvimento de um modelo de classificação, ou classificador, para prever a evasão em cursos superiores de diferentes instituições de ensino. O *framework* é importante para delimitar quais são os atributos e cursos onde o modelo preditivo pode ser aplicado.

Para tanto, inicialmente, é realizada uma análise exploratória na base de dados utilizada por Barreto (2019) com o intuito de propiciar o desenvolvimento do *framework*. O *framework* consiste em um diagrama de classes que contém todos os atributos identificados na base e que, posteriormente, serão utilizados no modelo preditivo para contribuir com a identificação de alunos que têm possibilidade de evadir. Além disso, o *framework* associa os atributos dos alunos a cursos superiores que podem ser aplicados no contexto de diferentes instituições de ensino, uma vez que são atributos presentes em bases acadêmicas das instituições. Com o objetivo de apoiar o método, é previsto o desenvolvimento do protótipo de uma plataforma Web que permite acesso de gestores escolares para obtenção dos resultados do modelo de forma simplificada e, desse modo, contribuir para o estabelecimento de políticas de permanência e êxito que possam apoiar a mitigação da evasão dos alunos identificados de acordo com o modelo preditivo.

1.2 Objetivos

1.2.1 Objetivo Geral

O presente trabalho tem o objetivo de propor um método para prever a evasão escolar no ensino superior baseado em um *framework* escolar utilizando mineração de dados para gerar um modelo de classificação.

1.2.2 Objetivos Específicos

O objetivo geral desta pesquisa pode ser desdobrado nos seguintes objetivos específicos:

- Desenvolver um método de predição da evasão escolar.
- Elaborar um *framework* escolar que permita a implementação do método;
- Desenvolver um modelo de classificação utilizando mineração de dados;
- Criar o protótipo de uma plataforma Web de apoio ao método, permitindo uma execução mais eficaz do método;
- Apoiar os gestores escolares na mitigação do abandono escolar por meio do fornecimento de insumos para o desenvolvimento de políticas de permanência e êxito.

1.3. Metodologia

Os procedimentos metodológicos utilizados para o desenvolvimento desta pesquisa são apresentados na Figura 1.

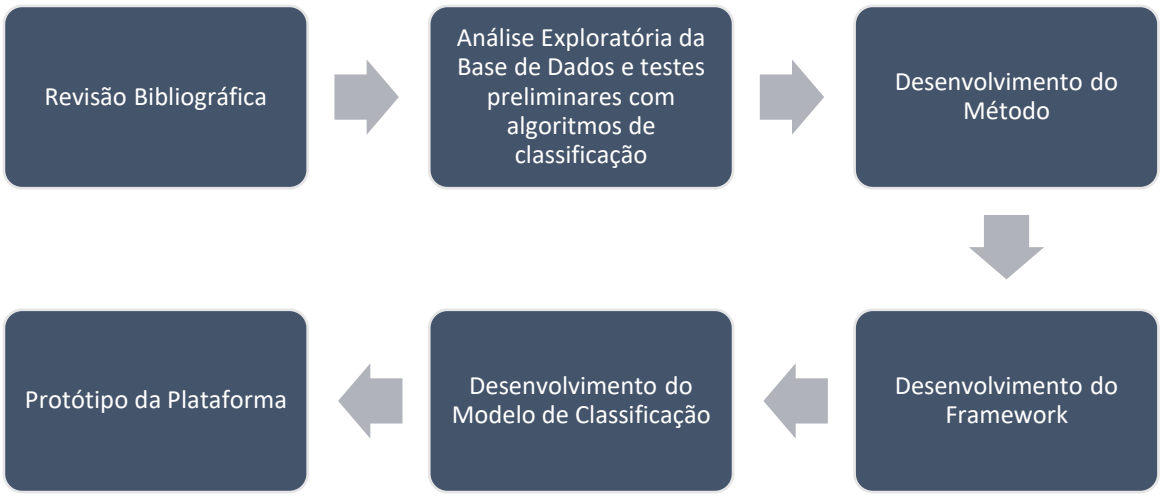

Figura 1: Etapas da metodologia.
Fonte: Autor.

1.3.1 Revisão Bibliográfica

Esta etapa consiste no estudo bibliográfico para construção da fundamentação teórica referente à abordagem proposta e é focada em conceitos e técnicas de mineração de dados e da evasão no ensino superior. A partir disso, o método PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analyses*) foi utilizado para selecionar os trabalhos mais aderentes ao tema deste projeto de dissertação (Liberati, 2009). O objetivo desta etapa é identificar na literatura trabalhos relacionados, além dos principais algoritmos de classificação utilizados nesse contexto.

1.3.2 Análise Exploratória da Base de Dados

Nesta etapa foram utilizados dados de uma instituição Federal de Ensino entre os anos de 2017 e 2019 (BARRETO, 2019), em que foram feitos alguns ajustes com a realização de testes preliminares. Os dados referentes aos anos posteriores, i.e., a partir de 2020, não foram utilizados, pois contemplam informações geradas durante a pandemia de COVID-19 e, devido ao contexto totalmente atípico, devem ser analisadas separadamente. O objetivo desta etapa é entender os dados utilizados, extraindo algumas informações preliminares, como a quantidade de registros vazios para determinados atributos, por exemplo.

1.3.3 Desenvolvimento do Método

O método é representado por um modelo de processo de negócio em BPMN (*Business Process Model and Notation*). Ele é o conjunto de etapas da proposta desta pesquisa que consiste no desenvolvimento do *framework* escolar, modelo de classificação e desenvolvimento do protótipo da plataforma.

1.3.4 Desenvolvimento do Framework

O *framework* escolar foi desenvolvido a partir de um modelo conceitual por meio de um diagrama de classes. O objetivo do *framework* é dar suporte ao método, permitindo sua implementação em diferentes instituições de ensino e cursos superiores.

1.3.5 Desenvolvimento do Modelo de Classificação

Após análise exploratória dos dados, definição do método e desenvolvimento do *framework*, esta etapa consiste na avaliação das técnicas, ou algoritmos, de classificação nos dados selecionados para identificar a que apresenta o melhor desempenho e que será utilizada para gerar o modelo de classificação final. Os modelos avaliados, bem como o modelo final, foram implementados em Python[1] e as seguintes técnicas foram testadas: *Decision Tree*, *Random Forest*, *Multilayer Perceptron*, *Support Vector Machine* e *Logistic Regression*. Estas técnicas são amplamente utilizadas em pesquisas desse tipo (BARRETO, 2019; FREITAS *et al.*, 2020; CAMACHO *et al.*, 2020; HUTAGAOL *et al.*, 2019; FERNÁNDEZ-GARCÍA *et al.*, 2020; Nagy e MONTOLAY, 2018; SANI *et al.*, 2020; PERCHINUNNO *et al.*, 2019).

1.3.6 Protótipo da plataforma

Após geração do modelo de classificação, o *framework* pode ser inserido numa plataforma com interface Web para facilitar a análise dos gestores escolares. Visto que o modelo foi implementado em Python, para facilitar a integração optou-se por utilizar uma biblioteca para desenvolvimento Web em Python chamado Django[2]. A plataforma tem o objetivo de apoiar gestores escolares, a fim de que possam desenvolver políticas de permanência e êxito, através da descoberta de alunos com possibilidade de evasão. A interface da plataforma foi desenvolvida, porém não foi integrada ao método. Portanto, há um protótipo da plataforma desenvolvido.

[1] https://www.python.org/
[2] https://www.djangoproject.com

2. REFERENCIAL TEÓRICO

Este capítulo apresenta o referencial teórico do trabalho. A Seção 2.1 apresenta um panorama geral da evasão escolar no ensino superior, descrevendo um contexto geral, situação atual e principais problemas. A Seção 2.2 apresenta conceitos de mineração de dados, iniciando por uma introdução ao assunto e, na sequência, explicando em mais detalhes a tarefa de classificação, com foco no desenvolvimento de um modelo preditivo.

2.1 Evasão Escolar no Ensino Superior

A evasão nos cursos de graduação das universidades brasileiras representa um problema complexo que atinge inúmeras instituições. Existe uma preocupação dos governos e das instituições em atenuar os índices de evasão acadêmica dos cursos por parte dos estudantes universitários (MANHÃES *et al*., 2011). Uma pesquisa realizada no Brasil mostrou que cerca de quarenta por cento (40%) dos alunos da rede pública de ensino superior está abandonando os cursos (PESSOAL, 2018). Este grave problema resulta no desperdício de dinheiro público, na não assimilação do conhecimento necessário às ciências e aos ofícios e na abdicação de uma conquista individual (SOUZA, 2008).

Além disso, é importante reforçar a relevância do problema da evasão, sob a perspectiva de que ele pode desencadear diversos outros malefícios para as instituições, tais como, problemas inerentes às áreas econômica e social, além do descumprimento da função política gerencial da instituição. Observa-se ainda que a verba para manutenção das universidades públicas está diretamente associada à quantidade de alunos com matrícula ativa (MANHÃES *et al*., 2011). Dessa forma, o orçamento universitário sofre muitas perdas, dificultando a gestão institucional, pois o número de docentes, técnicos administrativos, serviços terceirizados e a estrutura institucional continuam os mesmos, independentemente do número de alunos.

Quando a universidade não consegue manter o aluno até o final do curso, pode-se denominar de fracasso institucional, que inclui desde o docente que não conseguiu exercer o papel enquanto educador, até os programas e planos estabelecidos pelas Instituições de Ensino Superior (IES) por não cumprir a missão institucional de formar o seu alunado (DO NASCIMENTO *et al*., 2018).

Embora existam diversos fatores relacionados à evasão acadêmica, eles podem ser divididos entre fatores internos e externos à instituição. Os fatores internos são ligados ao curso, e podem ser subdivididos em: infraestrutura, corpo docente e a assistência socioeducacional. Os fatores externos à instituição estão relacionados ao aluno, e são exemplificados em: aspectos vocacionais, aspectos socioeconômicos e problemas pessoais (DA SILVA ZANATO *et al*., 2018).

Inúmeras universidades aplicam testes de vestibular com o objetivo de garantir o ingresso de estudantes com maior probabilidade de sucesso. No entanto, apesar de todos os esforços realizados pelas instituições, grande parte dos alunos selecionados passam a ser enquadrados no grupo com risco de evasão no decorrer dos cursos em que estão matriculados. Nesse contexto, diversos fatores podem influenciar a evasão, como por exemplo, questões pessoais, familiares, formação anterior, aspectos socioeconômicos, desempenho no curso em que o aluno está matriculado, entre outros (MASHILOANE E MCHUNU, 2013).

A busca por entender as causas da evasão acadêmica e a tomada de medidas preventivas está diretamente ligada às especificidades de cada instituição de ensino. No entanto, nem todas as instituições conseguem tomar medidas efetivas, pois a identificação

dos fatores que influenciam a evasão e a atribuição de uma ordem de importância para esses fatores é um trabalho complexo que está ligado à análise do conjunto de dados dos alunos (MANHÃES *et al*., 2011).

No estudo das causas da evasão escolar, uma possível saída é o uso da descoberta de conhecimento a partir de dados acadêmicos e socioeconômicos, por meio da utilização de técnicas de mineração de dados. Esse campo de estudo é responsável pelo desenvolvimento de métodos para entender de forma mais clara o comportamento dos alunos e o contexto em que estão inseridos. O principal objetivo da Mineração de Dados Educacionais (MDE) é aplicar técnicas de mineração de dados para criar modelos que preveem especificamente o abandono escolar (ROMERO e VENTURA, 2010). Portanto, a MDE nada mais é do que a aplicação de técnicas de mineração de dados em contextos educacionais.

2.2 Mineração de Dados

Antes mesmo de definir mineração de dados, é importante introduzir o processo de descoberta de conhecimento em bases de dados ou *Knowledge Discovery in Databases* (KDD). Segundo Fayyad, Pitatetsky-Shapiro e Smyth (1996), KDD é um processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis a partir de dados. Para analisar a evasão escolar a partir de dados, pode ser aplicado o processo de KDD, em que a mineração de dados se constitui em uma das suas etapas. Nesse contexto, o processo de KDD procura extrair conhecimento de dados para apoiar a análise de perfis de alunos para geração de informações e alertas que possam apoiar atividades preventivas pelos gestores escolares, embasando e agilizando as tomadas de decisões (COSTA *et al*. 2015). A Figura 2 apresenta as cinco etapas envolvidas no processo KDD, descritas a seguir.

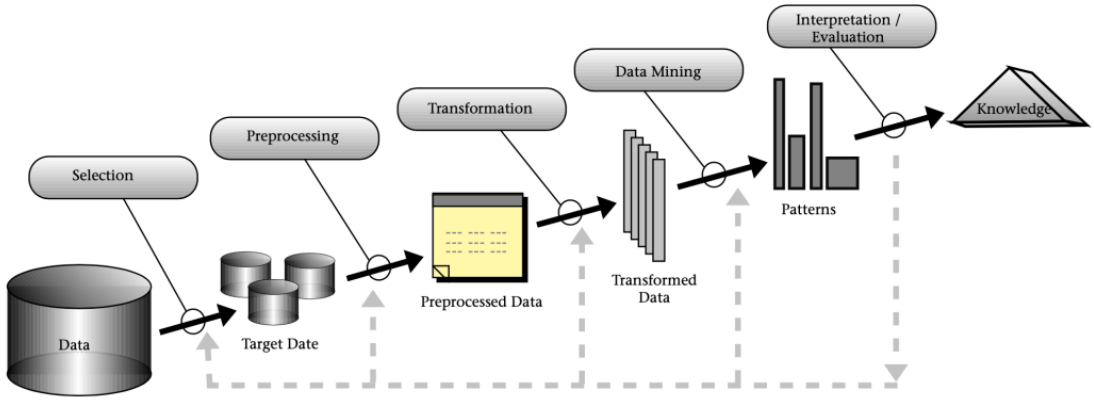

Figura 2: Etapas do processo de KDD.
Fonte: Fayyad, Pitatetsky-Shapiro e Smyth (1996)

i) Seleção: envolve a compreensão do domínio e dos objetivos da tarefa, juntamente com a coleta e seleção de dados e identificação das características, ou variáveis, que os descrevem, isto é, seus atributos.

ii) Pré-Processamento: os dados selecionados geralmente não estão em um formato apropriado para a execução das técnicas de mineração de dados, considerando que eles podem apresentar inconformidades como, por exemplo, valores faltantes para um ou mais atributos. Por isso, é necessário realizar a limpeza e tratamento dos dados antes da aplicação das técnicas.

iii) Transformação: os dados pré-processados podem ser transformados para um padrão que possam ser utilizados pelas técnicas de mineração de dados. Por exemplo, pode ser necessário transformar atributos numéricos, ou contínuos, em faixas de valores.

iv) Mineração de Dados: é a etapa que visa extrair regras e padrões ou produzir modelos preditivos a partir de dados por meio da aplicação de técnicas de mineração de dados.

v) Interpretação/Avaliação: essa etapa é onde os conhecimentos encontrados são interpretados e poderão ser usados no suporte ao processo de tomada de decisão na área de domínio da aplicação.

Como visto anteriormente, a mineração de dados é uma das etapas do processo de KDD. De acordo com Fayyad, Pitatetsky-Shapiro e Smyth (1996), a mineração de dados consiste na aplicação de algoritmos com o objetivo de extrair regras e padrões ou produzir modelos preditivos/de classificação a partir de dados, mas sofrendo com uma determinada limitação computacional. A natureza interdisciplinar da pesquisa e o desenvolvimento da mineração de dados contribuem significativamente para o sucesso de suas extensas aplicações. Para tanto, a mineração de dados incorpora muitas técnicas de outros domínios (Figura 3), como Estatística, Aprendizado de Máquina (*Machine Learning*), Reconhecimento de Padrões, Banco de Dados, Algoritmos, Computação de Alto Desempenho, entre outras (HAN; PEI; KAMBER, 2011).

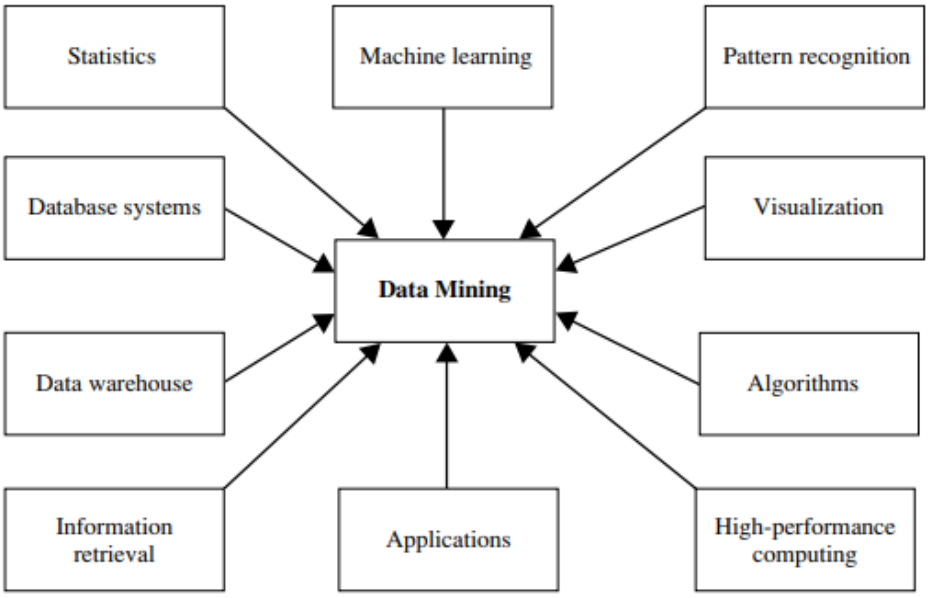

Figura 3: Técnicas adotadas pela mineração de dados.
Fonte: Han; Pei; Kamber (2011)

O estudo da mineração de dados com o objetivo de tratar o problema da evasão escolar é aplicado no campo de estudo denominado Mineração de Dados Educacionais (MDE). Segundo Baker *et al.* (2011) a MDE tem como principal objetivo o desenvolvimento de métodos para explorar conjuntos de dados coletados em ambientes educacionais, e que possui um grande potencial para melhorar a qualidade do ensino. Ainda, a MDE pelo seu enorme potencial de transformação, pode ser usada para aprender, prever e compreender o comportamento da aprendizagem, além de ajudar a desenvolver tecnologias melhores e mais inteligentes para apoiar alunos e professores (BAKER, 2014). Com a utilização da MDE é possível de forma mais eficaz que a manual,

compreender os alunos, como eles aprendem, como ocorre a aprendizagem, além de outros fatores que influenciam na aprendizagem (BAKER *et al*., 2011).

A mineração de dados pode ser aplicada com o objetivo de explicar o passado, consistindo em uma tarefa descritiva, isto é, para descrever e entender o comportamento dos dados, por meio de técnicas de clusterização (agrupamento de dados) e de extração de regras de associação, por exemplo. A mineração de dados pode ser aplicada também com o intuito de prever o futuro, isto é, prever um determinado evento ou classificar os dados como pertencendo a uma ou mais categorias, ou classes. Nesse caso, a mineração de dados consiste em uma tarefa preditiva, por meio da aplicação de técnicas de classificação (para prever um valor categórico para determinado atributo) e regressão (para prever um valor contínuo).

As tarefas de mineração descritivas fazem uso de um conjunto de dados do passado e extraem uma série de regras e padrões que precisam ser analisados por especialistas (HAN; PEI; KAMBER, 2011). Alguns exemplos são a busca pela identificação de comportamento de clientes e padrões de compras. Por outro lado, tarefas de mineração preditivas utilizam bases de dados históricas e geram modelos de classificação para fazer previsões (HAN; PEI; KAMBER, 2011), como a identificação de alunos que podem ou não evadir, gerando uma resposta direta e objetiva que não requer interpretação.

2.2.1 A Tarefa de Classificação

Segundo Kumar *et al.* (2018), a classificação é a tarefa de atribuir rótulos a instâncias de dados, ou registros, não rotulados e um classificador é usado para realizar tal tarefa. Han, Pei e Kamber (2011) afirmam que classificação é o processo de encontrar um modelo (ou função) que descreve e distingue classes ou conceitos de dados. O modelo é derivado com base na análise automática de um conjunto de dados de treinamento, ou seja, registros para os quais os rótulos de classe são conhecidos. O modelo gerado é usado para prever o rótulo da classe de registros para os quais o rótulo da classe é desconhecido.

A Figura 4 apresenta um exemplo da tarefa de classificação, onde tem-se um conjunto de dados de treinamento com uma série de registros e atributos, além do atributo classe. Um algoritmo de classificação receberá esses dados como entrada com o objetivo de gerar um modelo de classificação. Uma vez que o modelo é desenvolvido, um novo registro (um dado futuro, desconhecido) pode ser classificado pelo modelo de acordo com o atributo classe. No exemplo da Figura 4, o conjunto de dados de treinamento possui nove registros com os atributos *estado civil*, *gênero*, *renda familiar* e *idade*, além do atributo classe *evasão*. Nesse caso, o objetivo é prever se o aluno irá evadir ou não de acordo com suas características. Para tanto, um algoritmo de classificação é utilizado com o intuito de gerar o modelo preditivo. Com base neste modelo, à medida que novos registros surgem, é possível prever se o aluno vai abandonar ou não.

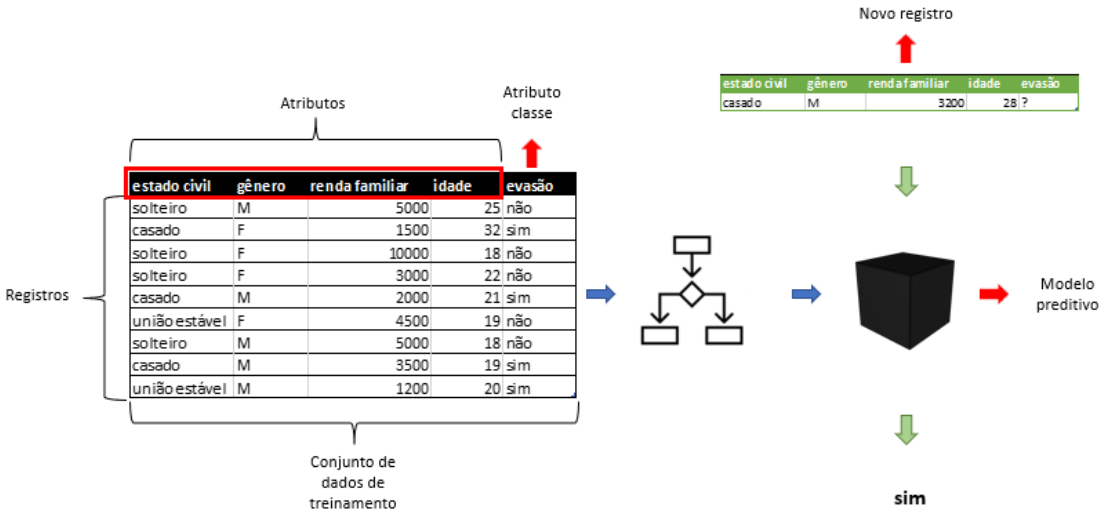

Figura 4: Modelo de classificação
Fonte: Autor.

De um modo mais formal, os dados para uma tarefa de classificação consistem em uma coleção de registros denominada conjunto de treinamento. Cada registro é caracterizado por uma tupla (*x*,*y*), em que *x* é o conjunto de valores de atributos que descrevem o registro e *y* é o rótulo da classe do respectivo registro, ou seja, o valor do atributo alvo que se deseja prever (Figura 5). O conjunto de atributos, incluindo o atributo classe, pode conter atributos categóricos, numéricos, entre outros (KUMAR *et al.*, 2018). Segundo Kumar *et al.* (2018), um modelo de classificação é uma representação abstrata da relação entre o conjunto de atributos descritivos e o atributo classe. Mais formalmente, podemos expressá-lo matematicamente como uma função alvo *f* que assume como *x* o conjunto de atributos e produz uma saída *y* correspondente ao rótulo de classe predito pela função, i.e., $y = f(x)$.

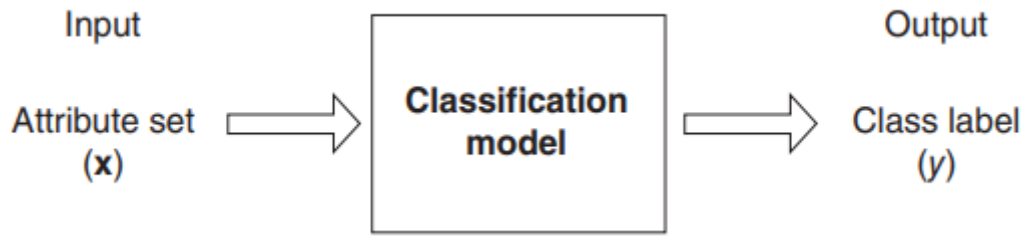

Figura 5: Ilustração de uma tarefa de classificação.
Fonte: Kumar *et al.* (2018).

A geração de um modelo de classificação corresponde as etapas de treinamento e avaliação do modelo, verificando se há um nível de confiança aceitável (HAN; PEI; KAMBER, 2011), isto é, medindo o desempenho preditivo do modelo treinado. Os algoritmos de classificação podem alcançar resultados extremamente eficientes gerando modelos treinados a partir de grandes conjuntos de dados (BISHOP, et al., apud BELÉM, et al., 2019).

Para medir o desempenho de um modelo de classificação, além do conjunto de dados usado para treinamento do modelo, é importante que haja um conjunto de dados apenas para teste, denominado conjunto de teste. Este conjunto consiste em uma coleção de registros rotulados que não foram usados para treinar o modelo, visto que o modelo

aprendido pode resultar em estimativas otimistas enganosas devido à especialização do algoritmo de aprendizado (KUMAR *et al.*, 2018).

De acordo com Kumar *et al.* (2018), existem quatro medidas essenciais usadas no cálculo de muitas métricas de avaliação. Entendê-las torna mais fácil o entendimento dessas métricas. Os registros positivos são os de interesse, ou seja, alunos evadidos no caso deste trabalho. Já os registros negativos correspondem aos registros dos alunos que não evadiram, ou seja, os alunos concluintes.

- Verdadeiro positivo ou *true positive* (TP): referem-se aos registros positivos que foram rotulados corretamente pelo classificador, em que TP representa o número de verdadeiros positivos.
- Verdadeiro negativo ou *true negative* (TN): são os registros negativos que foram rotulados corretamente pelo classificador, em que TN representa o número de verdadeiros negativos.
- Falso positivo ou *false positive* (FP): são os registros negativos que foram rotulados incorretamente como positivos, em que FP representa o número de falsos positivos.
- Falso negativo ou *false negative* (FN): são os registros positivos que foram erroneamente rotulados como negativos, em que FN representa o número de falsos negativos.

A Tabela 1 apresenta quatro métricas de avaliação e suas respectivas fórmulas, a saber: i) acurácia, ii) *recall*, iii) *precision* e iv) *F1-score*. A acurácia consiste na soma dos verdadeiros positivos (TP) com os verdadeiros negativos (TN) dividida pelo total de registros positivos e negativos, i.e., o número total de registros da base. *Recall* é uma métrica de avaliação dos positivos, onde basta-se dividir o número de positivos verdadeiros pelo total de positivos mais negativos. *Precision*, por sua vez, considera os positivos verdadeiros divididos pela soma dos positivos verdadeiros com falsos positivos. Por fim, a medida *F1-score*, multiplica precisão por *recall* e depois por dois, além de dividir pela soma de previsão por *recall*.

Medida de Avaliação	**Fórmula**
Acurácia	$\frac{TP + TN}{P + N}$
Recall	$\frac{TP}{TP + TN}$
Precision	$\frac{TP}{TP + FP}$
F1-score	$\frac{2 \times precisão \times recall}{precisão + recall}$

Tabela 1: Medidas de avaliação.
Fonte: Adaptado de Kumar *et al.* (2018).

2.3. Análise da pesquisa de Barreto (2019)

A motivação e embasamento do presente trabalho surgiu da pesquisa de Barreto (2019) tendo em vista a utilização da mesma base de dados e a detecção de perfis de evasão em cursos superiores de uma instituição de ensino por meio da mineração de dados, os quais foram analisados por gestores para melhoria da política de permanência e êxito. A proposta é avançar um pouco mais a pesquisa, com um intuito de desenvolver um modelo de classificação a partir do trabalho desenvolvido anteriormente. Barreto (2019) teve como objetivo identificar padrões nos dados de alunos evadidos de cursos superiores por meio da aplicação de técnicas de mineração de dados, utilizando como estudo de caso o *campus* Campos Centro do Instituto Federal de Educação, Ciência e Tecnologia Fluminense. Para tanto, foram realizadas as seguintes etapas: definição dos objetivos da mineração de dados; seleção de dados do sistema acadêmico; extração de dados desse sistema; pré-processamento, quando foram realizadas todas as tarefas relativas à limpeza dos dados e demais procedimentos necessários para prepará-los para a etapa de mineração; mineração de dados, quando os dados foram submetidos a algoritmos a fim de encontrar padrões e, por fim, interpretação de resultados, quando foram analisadas as informações descobertas por meio da etapa de mineração de dados.

Após o pré-processamento, a base de dados utilizada por Barreto (2019) contém 43 atributos e 923 registros. A Tabela 2 apresenta os atributos e seus respectivos domínios.

Atributo	**Domínio**
Ano_Conclusao_1_grau	Número natural
ANO_CONCLUSAO_2_GRAU	Número natural
ano_letivo_ini	Número natural
clPeriodo_Let_ini	Número formatado (Ex.: 2017/2)
COD_NACIONALIDADE	Texto
Coeficiente_Rendimento	Número decimal
Desc_Area_Procedencia_Escola_Origem	[Rural | Urbana]
Desc_Cor	[NPI | Outros]
tipo_curso	[Bacharelado | Licenciatura | Tecnologia]
desc_curso	[Arq. e Urb. | Ciênc. da Nat. | D. Gráfico | Eng. Cont. Aut. | Geografia | Letras | Manut. Ind. | Matemática | Sist. de Inf. | Sist. de Telecom.]
Desc_Estado_Civil	[Solteiro | Casado | Divorciado | Outro | Separado | União Estável | Viúvo]
Desc_Estado_Civil_Pais	[Solteiro | Casado | Divorciado | Outro | Separado | União Estável | Viúvo]
Desc_Forma_Ingresso_Matricula	[Portadores de Diploma | Sisu Ampla Concorrência | Sisu Cota | Sisu PCD | Transferência Externa | Transferência Interna | Vestibular Ampla Concorrência | Vestibular Cota]
Desc_renda_familiar	[Até 1 salário | 1 a 2 salários | 2 a 3 salários | 3 a 5 salários | 5 a 1o salários | 10 a 20 salários | Mais de 20 salários | Não declarado]
Desc_Renda_Per_Capita	[Menos de 1 salário mínimo | 1 salário mínimo | 2 salários mínimos | 3 salários mínimos | 4 salários mínimos | De 5 a 6 salários mínimos | De 7 a 10 salários mínimos | Não Declarado]
Desc_Renda_Per_Capita_SIG	[RFP <= 0,5 SM | 0,5 SM < RFP <= 1 SM | 1 SM < RFP <= 1,5 SM | 1,5 SM < RFP <= 2,5 SM | 2,5 SM < RFP <= 3 SM | RFP > 3 SM | Não Declarado]
Desc_Sit_matricula	[Cancelado | Concluído | Egresso | Evasão]
Desc_Sit_matricula_Periodo	[Aprovado | Reprovado | Egresso | Evasão]

Atributo	Domínio
Desc_Tip_Forma_Ingresso_Matricula_Periodo	[Reabertura de matrícula | Reintegração | Renovação por Aprovação | Renovação por Reprovação | Seleção]
Desc_Tipo_Escola_Origem	[Privada | Pública Municipal | Pública Estadual | Pública Federal]
Desc_Turno	[Integral | Matutino | Vespertino | Noturno]
GRAU_PARENTESCO_RESPONSAVEL	[Avó | Tia]
MAE_FALECIDA	[0 | 1]
N_FILHOS	[0 | 1 | 2 | 3 | 4]
necessidade_auditiva	[0 | 1]
necessidade_fisica	[0 | 1]
necessidade_mental	[0 | 1]
necessidade_multipla	[0 | 1]
necessidade_visual	[0 | 1]
outras_necessidades	[0 | 1]
PAI_FALECIDO	[0 | 1]
Percentual_Frequencia	Número decimal
Periodo	[0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10]
Periodo_Let	[1 | 2]
Periodo_letivo_ini	[1 | 2]
Sexo	[F | M]
TIPO_RESPONSAVEL	[A | M | O | P]
Dif_con_ens_medio_ini_grad	Número natural
ja_possuia_graduacao	[Sim | Não]
ja_possuia_pos_graduacao	[Sim | Não]
Idade	Número natural
Semestre	Número formatado (Ex.: 2017/2)
Situação	[Concluinte | Evadido]

Tabela 2: Relação de atributos de Barreto (2019)
Fonte: Barreto (2019)

A base de dados utilizada por Barreto (2019) corresponde ao período do segundo semestre de 2017 até o primeiro semestre de 2019. Foram considerados todos os cursos superiores da instituição que já haviam finalizado pelo menos um período letivo. A Figura 6 permite uma visualização do cenário de evasão no período estudado. Esse agrupamento permite uma visão ampla sobre a evasão nos cursos estudados.

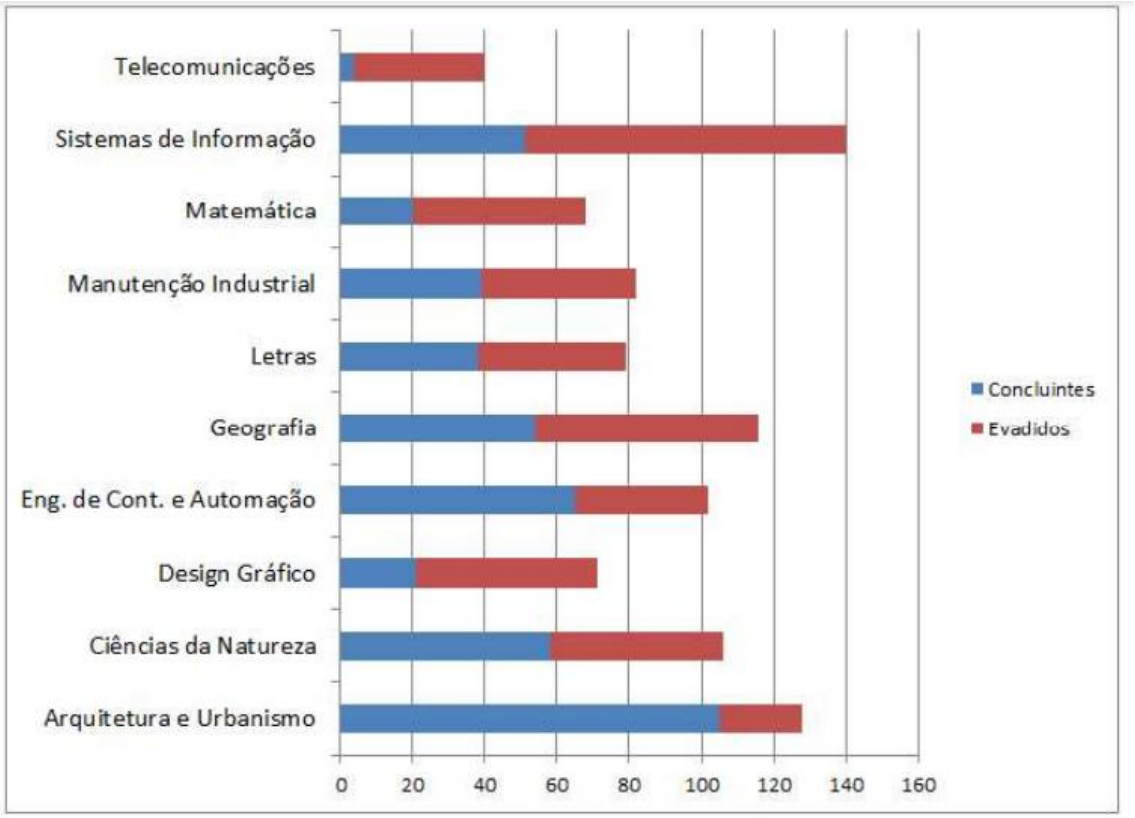

Figura 6: Demonstrativo de alunos concluintes e evadidos por curso no período 2017-2 a 2019-1.

Fonte: Barreto (2019)

A partir destas informações foram geradas árvores de decisão para analisar com um maior detalhamento o perfil dos alunos evadidos em cada curso com o objetivo de gerar políticas de combate à evasão, ou seja, ações para permanência e êxito dos estudantes. Para gerar os resultados, os cursos foram divididos em bacharelado, licenciatura e tecnologia.

2.3.1 Bacharelado

O primeiro curso analisado nesta categoria foi o Bacharelado em Arquitetura e Urbanismo. Como o número de alunos evadidos neste curso é significantemente inferior ao número total de alunos (105 concluintes e 23 evadidos), a evasão se tornou um caso raro no contexto dessa análise. Portanto, Barreto (2019) afirma que não há formação de um padrão forte para o estudo da evasão nesse curso, de modo que as características encontradas podem estar muito aderentes aos poucos alunos que evadiram.

Já o curso de Bacharelado em Engenharia de Controle e Automação possui 102 registros em sua base, dos quais 65 são concluintes e 37 evadidos, demonstrando que o valor *concluinte* para o atributo classe (atributo Situação) é majoritário. A Figura 7 apresenta a árvore de decisão gerada considerando este curso.

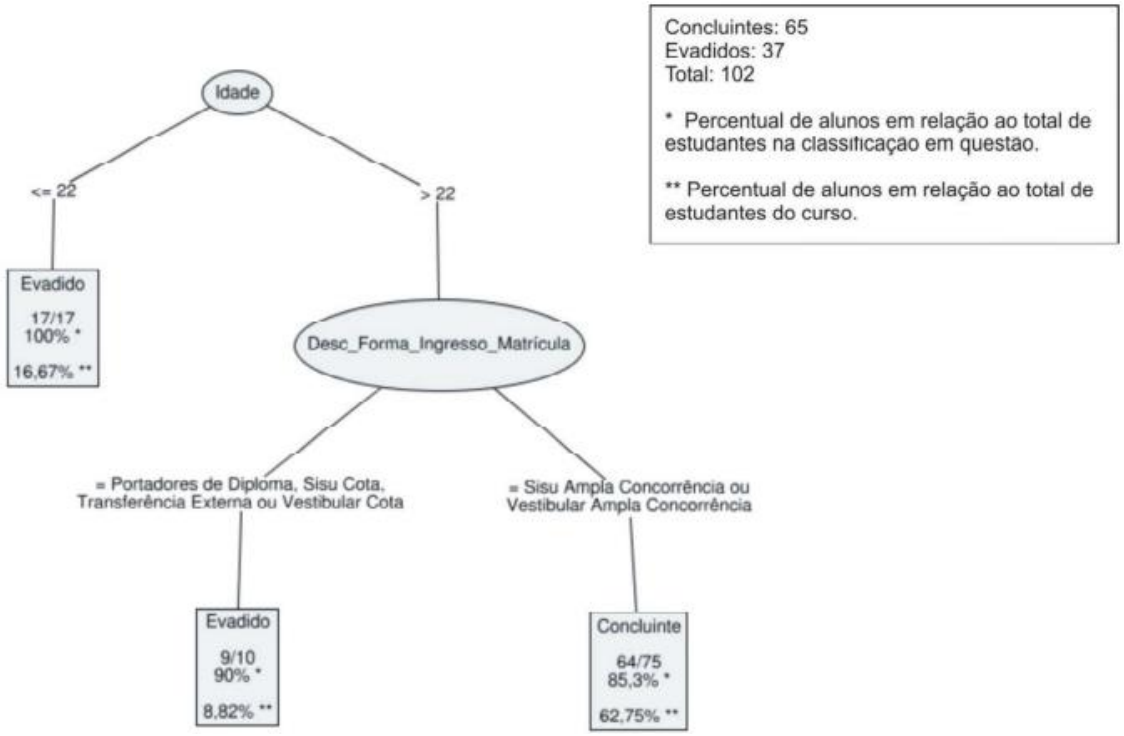

Figura 7: Árvore de decisão para o curso de Bacharelado em Engenharia de Controle e Automação.

Fonte: Barreto (2019).

Foram obtidos dois perfis de evasão com essa árvore: o primeiro apontou que 17 alunos, de um total de 17 (100% dos registros nessa situação) com idade menor ou igual a 22 anos evadiram. Em relação ao quantitativo total de alunos na base, esse valor corresponde a 16,67%. O segundo perfil encontrado apontou que nove alunos, em um total de 10, com idade maior que 22 anos e que ingressaram por meio das modalidades Portadores de Diploma, Sisu (Cota), Transferência Externa ou Vestibular (Cota) evadiram, o que representa uma taxa de evasão de 90% para os alunos com esse perfil. Já em relação ao total de alunos, esse valor corresponde a 8,82%.

2.3.2 Licenciatura

O primeiro curso de Licenciatura analisado foi Ciências da Natureza, cuja árvore de decisão é apresentada na Figura 8. Na base de dados, este curso possui 106 alunos, dos quais 58 concluíram, correspondendo a 54,7%. Além disso, foi observado o Coeficiente de Rendimento como fator determinante na evasão de alunos, ou seja, 37 alunos, em um total de 37 (100% dos casos) com Coeficiente de Rendimento menor ou igual a 4,78 evadiram. Esse valor corresponde a 34,91% do total de registros. Outro perfil de evasão também foi descoberto, porém com uma taxa bem menor de evasão: 52,9%. Nesse segundo caso, nove alunos, em um total de 17 evadiram, o que corresponde a 8,49% do total de registros. Estes são alunos que ingressaram através das modalidades Portadores de Diploma, Sisu (Ampla Concorrência), Sisu (Cota) ou Vestibular (Cota), possuindo Coeficiente de Rendimento entre 4,78 e 8.

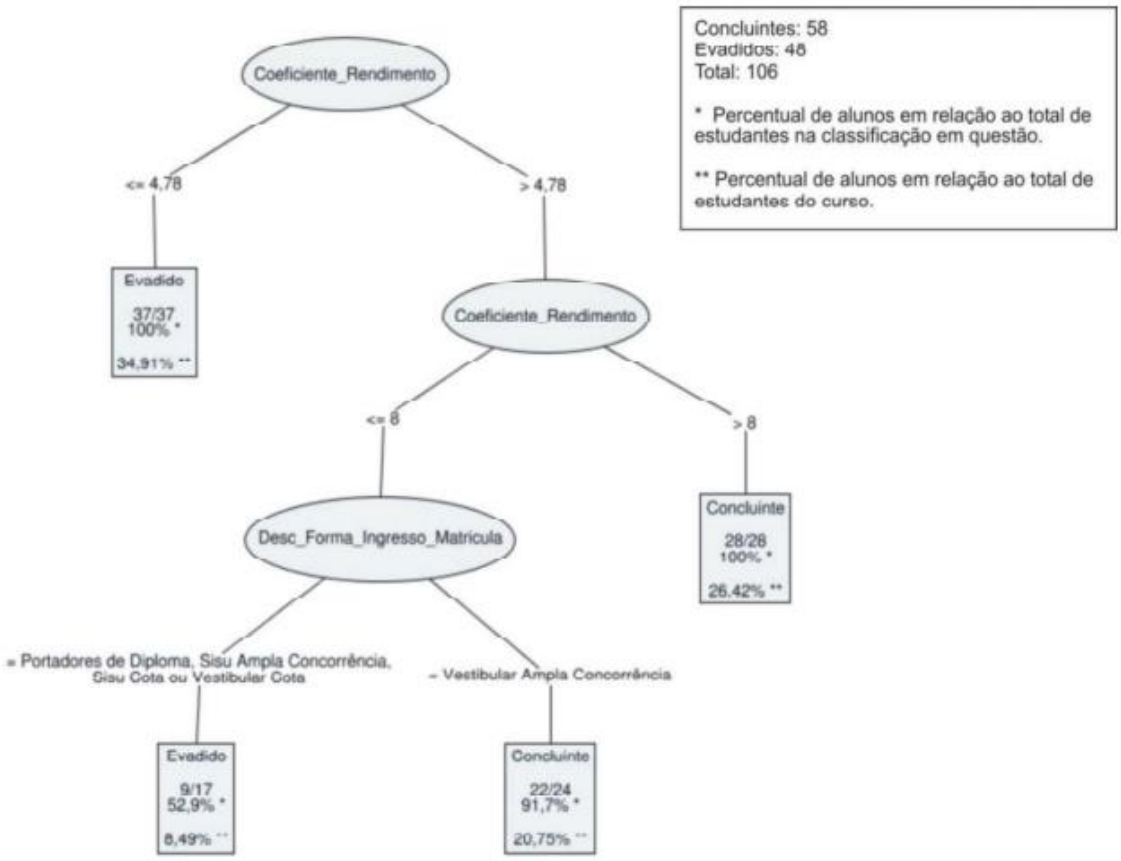

Figura 8: Árvore de decisão para o curso de Licenciatura em Ciências da Natureza.

Fonte: Barreto (2019).

Já o curso de Licenciatura em Geografia gerou três perfis de evasão, diferenciando-o dos cursos anteriores, como mostra a Figura 9. No primeiro perfil se encontram alunos com idade menor ou igual a 20 anos. 15 alunos, em um total de 15 que se enquadravam nessa situação evadiram, correspondendo a 100% dos alunos nessa situação (12,93% do total de registros da base). No segundo perfil identificado, há alunos cuja idade é maior que 20, possuindo como forma de ingresso as situações de Portadores de Diploma, Sisu (Cota) ou Vestibular (Cota). Nesse perfil, existem 15 alunos em um total de 19, correspondendo a 78,9%. Já em relação total de registros da base, esse valor corresponde a 12,93%. No terceiro perfil encontrado estão os alunos cuja idade é maior que 31 anos, possuindo como forma de ingresso as situações de Sisu (Ampla Concorrência), Transferência Externa ou Vestibular (Ampla Concorrência). Nesse último perfil, 14 de 22 alunos evadiram, o que equivale a 63,6% dos alunos que estão nessa situação, correspondendo a 12,07% do total de alunos da base desse curso.

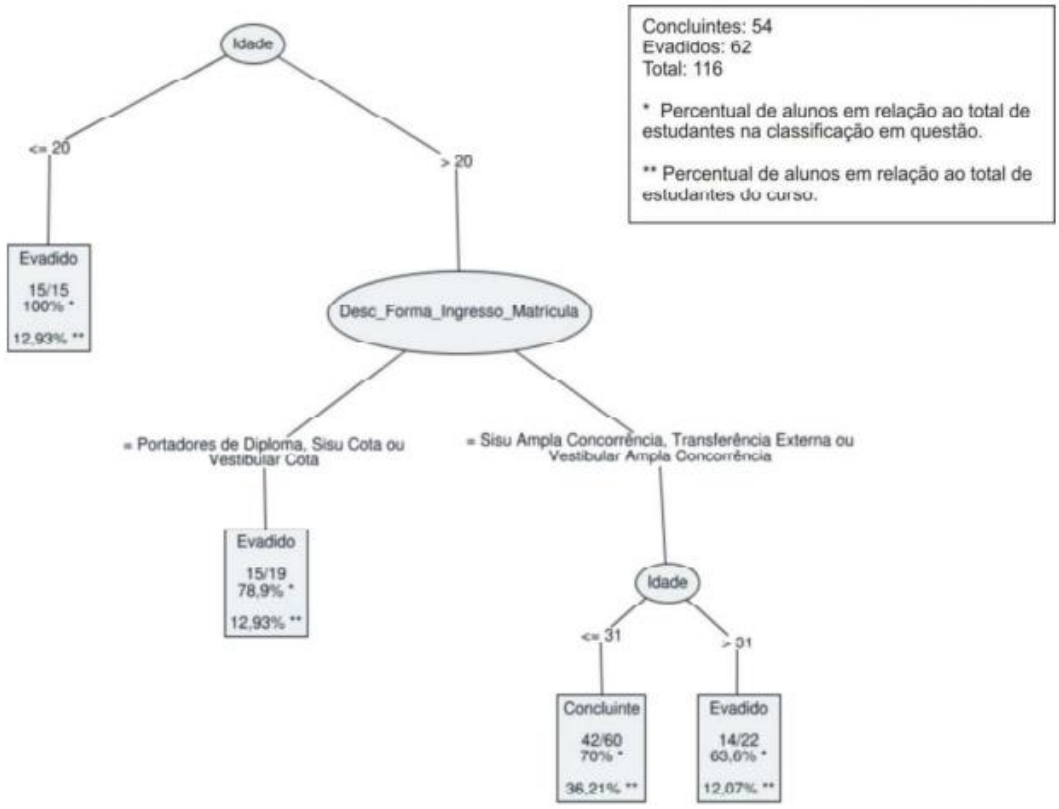

Figura 9: Árvore de decisão para o curso de Licenciatura em Geografia.

Fonte: Barreto (2019).

2.3.3 Tecnologia

Em relação ao grupo de cursos de Tecnologia, o primeiro analisado foi o de Design Gráfico. A Figura 10 apresenta a árvore de decisão gerada, onde são encontrados três perfis diferentes de evasão. O primeiro perfil de evasão está relacionado a alunos que levaram até três anos desde a conclusão do Ensino Médio até o início no curso superior, com idade menor ou igual a 22 anos. O segundo padrão se refere a alunos que levaram até três anos desde a conclusão do Ensino Médio até o início no curso superior, com idade superior a 22 anos, cuja a cor é NPI. Já o terceiro, abrange alunos que levaram mais de três anos desde a conclusão do Ensino Médio até o início no curso superior.

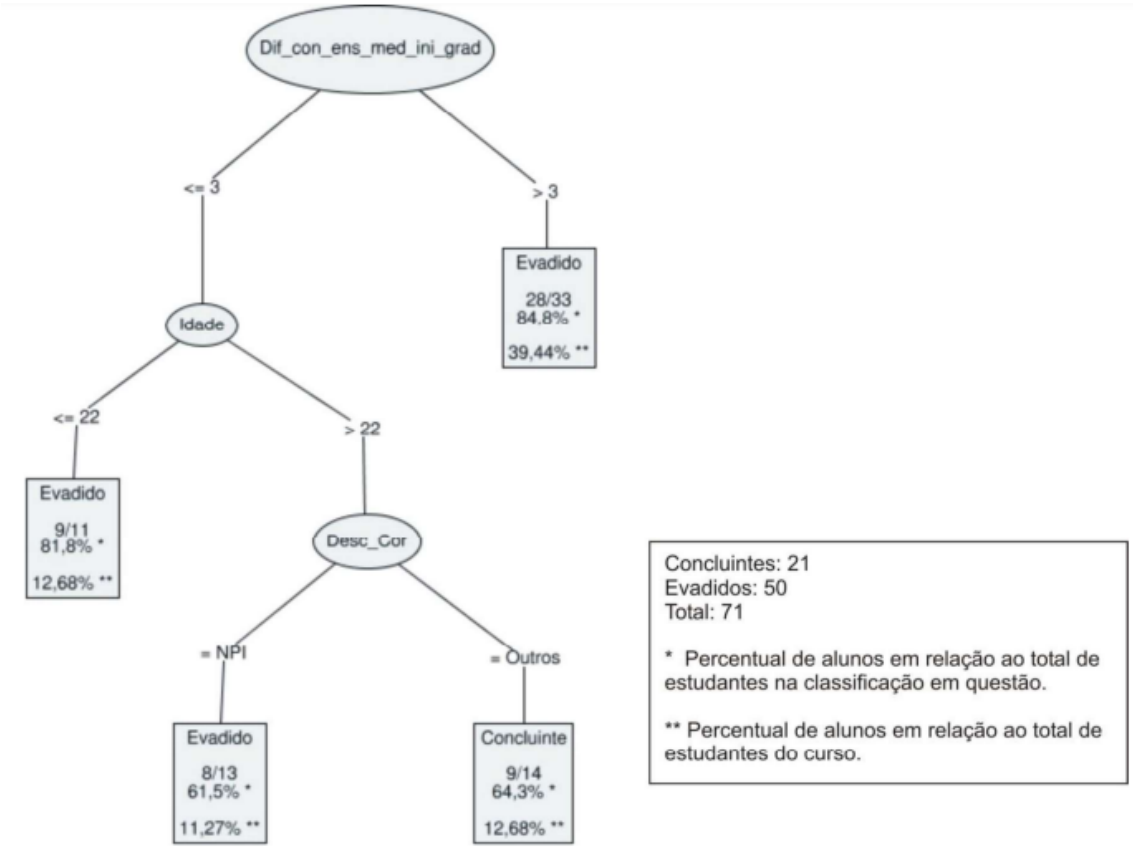

Figura 10: Árvore de decisão para o curso de Tecnólogo em Design Gráfico.

Fonte: Barreto (2019).

Em relação ao curso de Tecnologia em Manutenção Industrial, a base possui 82 registros, dos quais 43 são referentes a alunos evadidos. Para esse curso, foram descobertos três perfis de evasão (Figura 11). No primeiro perfil, encontram-se alunos que ingressaram por meio das modalidades Sisu (Cota), Transferência Interna ou Vestibular (Cota). Este perfil se destaca, pois foram identificados 15 alunos, dos quais 15 evadiram, o que corresponde a 100% dos registros nessa situação. O valor corresponde a 18,29% do total de alunos do curso. No segundo, há alunos com idade menor ou igual a 22 anos, que ingressaram por meio das modalidades Portadores de Diploma, Sisu (Ampla Concorrência), Transferência Externa ou Vestibular (Ampla Concorrência). Por fim, no terceiro, aparecem alunos que ingressaram por meio das modalidades Portadores de Diploma, Sisu (Ampla Concorrência), Transferência Externa ou Vestibular (Ampla Concorrência), com idade maior que 30 anos.

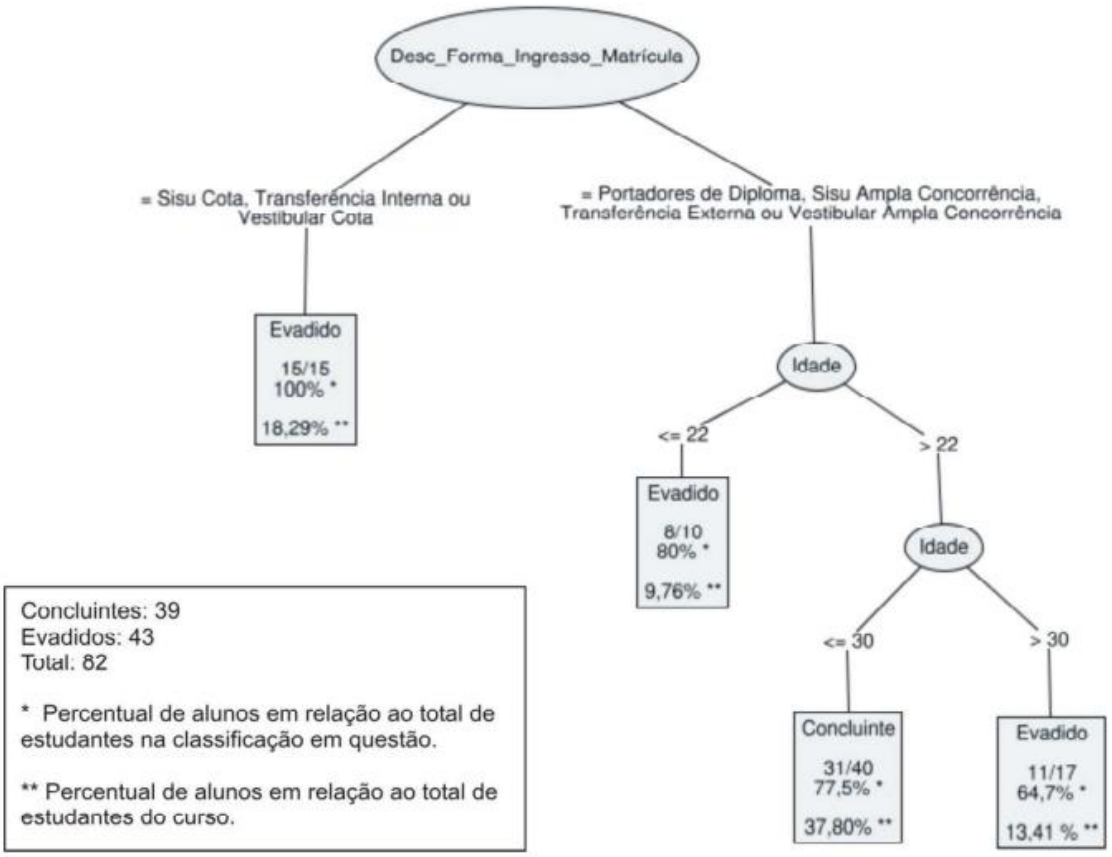

Figura 11: Árvore de decisão para o curso de Tecnologia em Manutenção Industrial.

Fonte: Barreto (2019).

No caso do curso de Tecnologia em Sistemas de Telecomunicações o número de alunos evadidos corresponde quase que à totalidade de estudantes, fazendo com que os alunos que concluem o curso se tornem um caso raro no contexto dessa análise. Por esse motivo, não há formação de um padrão forte para o estudo do sucesso nesse curso, de modo que as características encontradas podem estar muito aderentes aos poucos alunos que concluíram.

Segundo Barreto (2019), a partir dos resultados encontrados, conclui-se que a forma de ingresso de matrícula apareceu como fator determinante em seis de 10 cursos, demonstrando a relevância desse atributo na análise da evasão no contexto estudado. Entre os cursos com maiores taxas de evasão podem ser destacados o curso de Bacharelado em Sistemas de Informação, com 63,6% de evasão e o curso de Tecnologia em Sistemas de Telecomunicações, com 90%. Em relação a esses cursos, os alunos que ingressaram pelas modalidades Vestibular Cota e Sisu Cota aparecem entre os perfis que mais evadiram, estando presentes nos dois cursos mencionados.

Nesse sentido, as informações encontradas permitem diagnosticar características dos perfis dos alunos evadidos de acordo com atributos disponíveis no sistema acadêmico. No entanto, esta análise foi feita com base no passado. De forma diferente e complementar, o presente trabalha tem o objetivo de gerar um modelo de classificação da evasão com base nesses dados históricos, que busca identificar alunos em risco antes que o abandono escolar ocorra de fato.

3. REVISÃO BIBLIOGRÁFICA

3.1 Aplicação do Método PRISMA

Inicialmente, a fim de encontrar trabalhos relacionados ao tema, foi elaborada uma estratégia de busca a partir de alguns termos e tesauros (Tabela 2). Para tanto, foram utilizadas as bases de dados *Scopus* e *Web of Science*. A escolha da base *Web of Science* (WoS) justifica-se pela literatura histórica indexada, enquanto a *Scopus* privilegia a literatura recente, sendo superior em número de revistas indexadas, porém com impacto menor do que a WoS (CHADEGANI *et al.*, 2013).

Os termos inseridos na consulta nas bases de dados foram Aprendizado de Máquina, Mineração de Dados, Evasão e Ensino Superior. Os tesauros definidos de acordo com os termos selecionados podem ser visualizados na Tabela 3.

Termos	**Aprendizado de Máquina**	**Mineração de Dados**	**Evasão**	**Ensino Superior**
Tesauros	*Machine Learning*	*Data Mining*	*Evasion* *Dropout*	*Higher Education* *Higher Learning* *Tertiary Education* *University Degree* *Degree Course* *Tertiary Degree* *Graduation Course*

Tabela 3: Termos e Tesauros.
Fonte: Autor.

A metodologia utilizada para seleção de artigos foi a *Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (PRISMA). Segundo Liberati (2009), a metodologia PRISMA é uma diretriz que tem o objetivo de ajudar autores a melhorarem a qualidade do relato dos dados das pesquisas e orientar a avaliação crítica de uma revisão ou de uma meta-análise de trabalhos já publicados.

Critério	**Descrição**
Inclusão 1	Referir-se à utilização de técnicas de *machine learning* e mineração de dados.
Inclusão 2	Abordar a possibilidade de previsão da evasão escolar ou análise de desempenho dos alunos no ensino superior presencial.
Exclusão 1	Não se referir à utilização de técnicas de *machine learning* e mineração de dados.
Exclusão 2	Não abordar a possibilidade de previsão da evasão escolar ou análise de desempenho dos alunos no ensino superior presencial.
Exclusão 3	Impossibilidade de acesso ao texto completo.
Exclusão 4	Artigos não escritos em inglês ou português.
Exclusão 5	Artigos de revisão sistemática da literatura.

Tabela 4: Critérios de inclusão e exclusão.
Fonte: Autor.

A seleção de artigos realizada através do método PRISMA foi baseada em alguns critérios de inclusão e exclusão descritos na Tabela 4. Os critérios foram aplicados nas etapas de análise de título e resumo e de leitura completa do artigo. Os critérios foram definidos com o objetivo de encontrar trabalhos relacionados ao tema desta pesquisa.

A Figura 12 demonstra como foi aplicado o método PRISMA. Inicialmente, foram encontrados 16 documentos na base *Scopus* e nove documentos na base *Web of Science*, além de dois documentos oriundos de outras fontes. Posteriormente, foi necessário remover sete documentos duplicados. Havia 20 documentos encontrados após a exclusão de documentos repetidos. Em seguida, os trabalhos foram analisados de acordo com o conteúdo presente no título e no resumo. Esse procedimento resultou na remoção de nove artigos. Com isso, restaram 11 documentos que foram avaliados por meio de uma leitura completa. Após a leitura, foram retirados três artigos de acordo com os critérios de inclusão e exclusão. Dessa forma, o resultado foi de oito documentos.

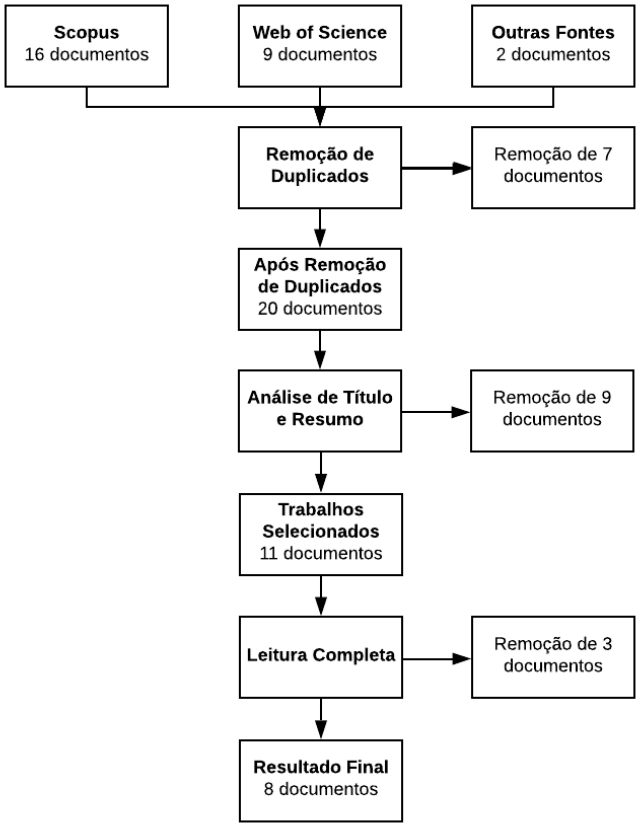

Figura 12: Aplicação do Método PRISMA.
Fonte: Autor.

3.2 Trabalhos Relacionados

Os oito trabalhos selecionados após a aplicação do método PRISMA são descritos na sequência, com exceção de Barreto (2019) que foi descrito em detalhes na Seção 2.3.

Freitas *et al.* (2020) apresentam um *framework* utilizando internet das coisas para prever o abandono escolar baseado em dados socioeconômicos usando algoritmos de classificação como árvore de decisão, *Logistic Regression*, *Support Vector Machines*, *k-nearest neighbors* (k-NN) e *multilayer perceptron.* A utilização de dados socioeconômicos torna possível identificar no ato da pré-inscrição quem são os alunos passíveis de evasão no futuro. A abordagem foi validada analisando a acurácia, F1-*score*,

recall e *precision.* Os resultados mostraram que o sistema desenvolvido obteve 99,34% de acurácia, 99,34% de F1-score e 100% de *recall* usando a árvore de decisão. Este trabalho se aproxima desta proposta de dissertação pelo fato de projetar uma plataforma que permite a descoberta de alunos com possibilidade de evasão, porém são incluídos alguns procedimentos relacionados à internet das coisas, diferenciando-o. Outro ponto importante é a base de dados utilizada. Em Freitas *et al.* (2020), foram considerados apenas dados socioeconômicos enquanto a proposta deste trabalho é incluir dados acadêmicos.

Camacho *et al.* (2020) também utilizaram internet das coisas, porém explorando tecnologias de rastreamento e um relógio inteligente com o objetivo de realizar uma análise de aprendizagem dos alunos. Nos experimentos realizados em escolas são identificados diferentes padrões de comportamento dos alunos, mostrando o progresso e participação deles. Posteriormente, a partir dos dados capturados, visando uma fase de tomada de decisão, foram utilizadas técnicas de mineração de dados para construir modelos capazes de explicar os momentos específicos em que há envolvimento do estudante. As regras obtidas, que podem ser facilmente interpretadas por um usuário que não seja especialista, ajudam o professor a observar, analisar e tomar decisões com o objetivo de fomentar o engajamento na sala de aula. Este artigo possibilita a realização de uma análise da aprendizagem dos alunos por meio de técnicas de *machine learning*, no entanto só possibilita sua implementação com a participação de um especialista, ou seja, não há um protótipo de plataforma intuitiva que permita que leigos façam a análise dos estudantes.

Hutagaol *et al.* (2019) analisam e medem a correlação entre indicadores demográficos e desempenho acadêmico para prever a possibilidade de evasão do aluno usando os classificadores k-NN, *Naïve Bayes* e árvores de decisão com o objetivo de encontrar a melhor solução de modelagem na identificação de alunos com possibilidade de abandono. Os resultados foram superiores ao combinar os algoritmos citados com o meta-classificador *Boosting*. Além disso, obtiveram a melhor acurácia de cerca de 98,82% usando este método que foi testado por validação cruzada com 10 iterações (10-*fold cross validation*). Hutagaol *et al.* (2019) utilizam algoritmos de classificação para identificar alunos com chances de evasão, porém não apresentam um método capaz de analisar o risco de abandono por meio de uma plataforma.

Fernández-García *et al.* (2020) introduzem uma abordagem que não apenas prevê o risco de evasão escolar ou desempenho dos alunos, mas também propõem medidas para ajudar alunos e instituições de ensino. O objetivo é maximizar as taxas de permanência na graduação, construindo um sistema de recomendação para auxiliar os alunos na seleção das disciplinas. Para tanto, são utilizadas técnicas de mineração de dados e algoritmos de classificação, como *Random Forest* (RF), *support vector machine* (SVM) e *logistic regression* (LR). Os autores utilizam algoritmos de classificação, obtendo bons resultados, mas não apresentam uma plataforma intuitiva para que gestores escolares possam realizar as análises, limitando apenas a especialistas.

Na pesquisa de Nagy e Montolay (2018) também são avaliados e empregados vários algoritmos de classificação para identificar alunos com risco de evasão. Apesar de não proporem medidas para solucionar o problema, apresentam uma plataforma de suporte à decisão baseada em dados para utilização de gestores escolares e demais interessados. O modelo que apresentou melhor precisão foi o SVM (*Support Vector Machine).* Mesmo apresentando uma plataforma de suporte à tomada de decisão, ela não tem o objetivo de realizar a predição da possibilidade de evasão.

Sani *et al.* (2020) desenvolvem três modelos de aprendizagem de máquina, a saber: árvores de decisão, RF (*Random Forest)* e redes neurais artificiais com o objetivo de prever o risco de evasão de alunos com renda baixa nas universidades da Malásia. Na análise de desempenho realizada, o algoritmo RF supera os outros dois modelos em termos de acurácia, precisão e *recall*. Apesar dos bons resultados alcançados, o trabalho em questão não apresenta nenhum tipo de método que permita a predição da evasão por meio de uma plataforma de apoio à gestão escolar.

Perchinunno *et al.* (2019) utilizam técnicas de mineração de dados, incluindo algoritmos de classificação. Foram utilizados métodos de classificação supervisionados para identificar o perfil dos alunos com maior probabilidade de evasão através das informações relacionadas ao desempenho dos estudantes no ensino médio. Nesta pesquisa não foi criado nenhum tipo de método com base em um *framework* que permita o desenvolvimento de um modelo preditivo para evasão escolar.

A Tabela 5 descreve os trabalhos relacionados diferenciando-os de acordo com o propósito, dados trabalhados, algoritmos utilizados e projeto de implementação de plataforma.

Trabalho	**Propósito**	**Dados Trabalhados**	**Algoritmos Utilizados**	**Plataforma**
Barreto (2019)	Identificar padrões nos dados de alunos evadidos de cursos superiores por meio da aplicação de técnicas de mineração de dados, utilizando como estudo de caso o *campus* Campos Centro do Instituto Federal de Educação, Ciência e Tecnologia Fluminense.	Dados acadêmicos e socioeconômi-cos	Árvores de Decisão e CN2 *Rule Induction*	Não implementa plataforma.
Freitas *et al.* (2020)	Realizar a automação do processo de previsão por um método capaz de obter informações que seriam difíceis e demoradas para humanos obter, contribuindo para uma previsão mais precisa.	Dados socioeconômi-cos	Árvores de Decisão, Regressão Logística, *Support Vector Machine* (SVM), KNN e *Multilayer Perceptron*	Desenvolve uma plataforma baseada na tecnologia de Internet das Coisas.
Camacho *et al.* (2020)	Utilizaram Internet das Coisas, porém explorando tecnologias de rastreamento e vestíveis com o objetivo de realizar uma análise de aprendizagem dos alunos. Nos experimentos realizados em escolas são identificados diferentes padrões de comportamento dos alunos, mostrando o progresso e participação dos mesmos.	Dados Acadêmicos	Árvore de Decisão	Não é criada uma plataforma intuitiva de predição da evasão.

Trabalho	Propósito	Dados Trabalhados	Algoritmos Utilizados	Plataforma
Hutagaol *et al.* (2019)	Analisar e medir a correlação entre indicadores demográficos e desempenho acadêmico para prever a possibilidade de o aluno evadir usando os algoritmos de classificação.	Dados acadêmicos e demográficos	KNN, *Naïve Bayes* e Árvore de Decisão	Não cria uma plataforma.
Fernández-García *et al.* (2020)	Introduzir uma abordagem que não apenas prevê o risco de evasão escolar ou desempenho dos alunos, mas também propõem medidas para ajudar alunos e instituições de ensino.	Dados acadêmicos	*Random Forest*, SVM e *Logistic regression*	Não desenvolvem uma plataforma intuitiva.
Nagy e Montolay (2018)	Empregar e avaliar vários algoritmos de aprendizado de máquina para identificar alunos com risco de evasão.	Dados acadêmicos	Árvore de Decisão, *Random Forest*, *Logistic Regression*, KNN e Naive Bayes	Proposta de uma plataforma de suporte à tomada de decisão e desenvolvimento de um *framework*.
Sani *et al.* (2020)	Desenvolver três modelos de aprendizagem de máquina com o objetivo de prever o risco de evasão de alunos com renda baixa nas universidades da Malásia.	Dados socioeconômi-cos	Árvore de Decisão, *Random Forest* e Redes neurais artificiais	Não há uma plataforma.
Perchinunno *et al.* (2019)	Implementar métodos de classificação supervisionados para identificar retrospectivamente o perfil dos alunos com maior probabilidade de evasão através das informações relacionadas ao desempenho dos estudantes no ensino médio.	Dados acadêmicos	*Random Forest*	Sem desenvolvi-mento de plataforma.

Tabela 5: Trabalhos Relacionados.
Fonte: Autor.

4. MATERIAIS E MÉTODOS

Este capítulo trata dos materiais e métodos existentes na pesquisa. Na Seção 4.1 são apresentados os métodos do desenvolvimento desta pesquisa. Na Seção 4.2 é introduzido o *framework* escolar desenvolvido a partir do método elaborado. A Seção 4.3 descreve os passos e ferramentas utilizadas no desenvolvimento do modelo de classificação. Por fim, a Seção 4.4 apresenta as tecnologias que serão usadas na criação do protótipo da plataforma.

4.1 Métodos da Pesquisa

Para uma melhor organização dos processos desenvolvidos em cada etapa desta pesquisa, foi elaborado um diagrama de processos por meio do software Bizagi. A notação selecionada foi a *Business Process Model and Notation* (BPMN) devido ao seu diferencial de apresentar um modelo para públicos-alvo distintos, a versatilidade de modelar situações diferentes no processo, além de apresentar uma grande quantidade de símbolos (ABMP, 2013). A Figura 13 apresenta os macroprocessos que representam o método proposto neste trabalho.

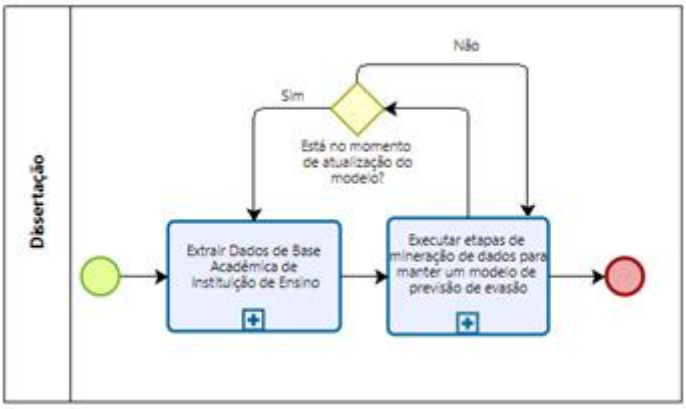

Figura 13: Macroprocessos da dissertação.
Fonte: Autor

O primeiro macroprocesso representa a extração de dados da base acadêmica de ensino. Em seguida, as etapas de mineração de dados são executadas para manter um modelo de previsão de evasão. Nesse ponto, haverá sempre o questionamento se está no momento de atualizar o modelo. Se sim, serão extraídos novos dados da base acadêmica. O modelo deve ser atualizado sempre que houver uma atualização na base de dados, em geral sem que finalizar um semestre e iniciar outro.

A seguir, serão apresentadas as tarefas existentes dentro de cada macroprocesso. O primeiro, chamado de “Extrair dados de base acadêmica de instituição de ensino”, pode ser visualizado na Figura 14.

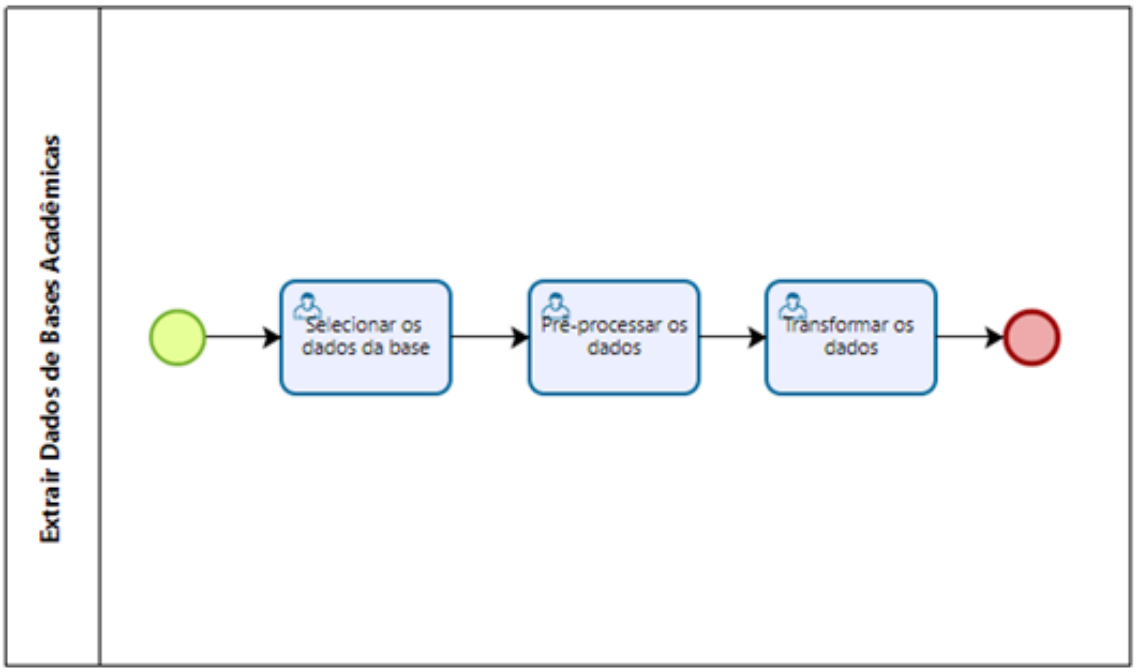

Figura 14: Primeiro Macroprocesso.
Fonte: Autor.

O processo da Figura 14 se inicia com a tarefa de selecionar dados da base. Primeiramente, a base selecionada para desenvolvimento do modelo de classificação foi extraída do trabalho de Barreto (2019) que considera dados dos semestres 2017.2, 2018.1, 2018.2 e 2019.1 dos cursos de tecnologia, bacharelado e licenciatura de uma instituição de ensino federal. Esta base de dados, inicialmente, possuía 923 registros e 44 atributos. Entretanto, foram feitas algumas modificações nas etapas de pré-processamento e transformação dos dados, como eliminação de alguns registros duplicados, exclusão de atributos considerados não relevantes e melhoria nos registros de alguns atributos para facilitar a implementação dos algoritmos.

Foram identificados 99 registros duplicados que foram excluídos da base, restando 824. Quanto aos atributos, foram descartados atributos redundantes, irrelevantes, isto é, que não influenciaram nos resultados ou que tinham uma quantidade muito pequena de registros. A relação de atributos retirados e o motivo são apresentados na Tabela 6.

Atributo	**Motivo de retirada**
ano_letivo_ini	Irrelevante
COD_NACIONALIDADE	Irrelevante
Desc_Renda_Per_Capita_SIG	Redundante
Desc_Sit_matricula	Redundante
Desc_Sit_matricula_Periodo	Redundante
Desc_Tip_Forma_Ingresso_Matricula_Periodo	Irrelevante
GRAU_PARENTESCO_RESPONSAVEL	Irrelevante
N_FILHOS	Poucos registros
necessidade_auditiva	Poucos registros
necessidade_fisica	Poucos registros
necessidade_mental	Poucos registros
necessidade_multipla	Poucos registros

Atributo	Motivo de retirada
necessidade_visual	Poucos registros
outras_necessidades	Poucos registros
PAI_FALECIDO	Poucos registros
Periodo_Let	Redundante
Periodo_letivo_ini	Redundante
TIPO_RESPONSAVEL	Poucos registros
Semestre	Irrelevante

Tabela 6: Atributos retirados.
Fonte: Autor

Por fim, após as alterações realizadas na base de dados, restaram 24 atributos, descritos na Tabela 7.

Atributo	Tipo de Dado	Domínio
Ano Conclusão 1º Grau	numérico	Valores inteiros entre 1976 a 2013
Ano Conclusão 2º Grau	numérico	Valores inteiros entre 1979 e 2018
Período Letivo Inicial	categórico	Valores entre 2002/1 e 2018/2
Coeficiente de rendimento	numérico	Valores reais entre 0 e 10
Área da Escola de Origem	categórico	Urbana ou rural
Cor	categórico	Branco, pardo, preto, indígena, amarelo ou "Não quis informar"
Tipo de Curso	categórico	Licenciatura, bacharelado ou tecnologia
Descrição do curso	categórico	Bacharelado em Engenharia de Controle e Automação, Bacharelado em Sistemas de Informação, Licenciatura em Ciências da Natureza, Licenciatura em Geografia, Licenciatura em Letras, Licenciatura em Matemática, Tecnologia em Design Gráfico, Tecnologia em Manutenção Industrial e Tecnologia em Sistemas de Telecomunicações
Estado Civil	categórico	União estável, casado, solteiro ou divorciado
Estado Civil Pais	categórico	União estável, casado, solteiro ou divorciado
Forma de ingresso	categórico	Vestibular – Ampla Concorrência, Vestibular – Cota, Transferência Externa, Transferência Interna, Portadores de Diploma, Sisu – Cota, Sisu - Ampla Concorrência e Sisu - PCD
Renda familiar	categórico	2 a 3 salários, 1 a 2 salários, 3 a 5 salários, 5 a 10 salários, 10 a 20 salários, Até 1 salário e Mais de 20 salários
Renda familiar per capita	categórico	Menos de 1 salário mínimo, 2 salários mínimos, 1 salário mínimo, 3 salários mínimos, De 5 a 6 salários mínimos, 4 salários mínimos e De 7 a 10 salários mínimos.
Tipo de Escola no 2º Grau	categórico	Pública Municipal, Pública Estadual, Pública Federal ou Privada
Turno	categórico	Noturno, matutino, vespertino e integral
Período	numérico	Valores inteiros entre 1 e 10
Mãe Falecida	categórico	Sim ou não
Pai Falecido	categórico	Sim ou não
Sexo	categórico	F ou M

Atributo	Tipo de Dado	Domínio
Diferença Entre Conclusão 2º Grau e Início do 3º Grau	numérico	Valores inteiros entre 1 e 38
Possui graduação	categórico	Sim ou Não
Possui pós graduação	categórico	Sim ou Não
Idade	numérico	Valores inteiros entre 19 e 66
Situação	categórico	Concluinte ou evadido

Tabela 7: Atributos da base de dados.
Fonte: Autor

Na Tabela 7, a coluna *Atributo* apresenta os nomes de cada atributo, a coluna *Tipo de dado* apresenta o tipo de representação do atributo, que pode ser categórico ou numérico, e a coluna *Domínio* descreve os possíveis valores de cada atributo. O atributo classe, ou seja, aquele que identifica se um aluno evadiu ou concluiu o curso, está identificado como *Situação* (o último atributo). Em relação ao atributo classe, existem 455 registros do tipo *concluinte* e 369 do tipo *evadido*.

Além da base utilizada do trabalho de Barreto (2019) com algumas modificações, foram gerados dados do semestre 2019.2 da mesma instituição de ensino federal, considerando os mesmos cursos. A base foi extraída de acordo com o *framework* desenvolvido para o método (Seção 4.2), pois ela deve ter os mesmos atributos da base de treinamento, ou seja, aqueles que estão definidos no *framework*. Com isso, as duas bases foram unificadas para utilização no estudo de caso do modelo de classificação.

O próximo macroprocesso é o de execução das atividades de mineração de dados para manter um modelo de previsão da evasão. A Figura 15 apresenta as tarefas que ocorrem dentro desta etapa.

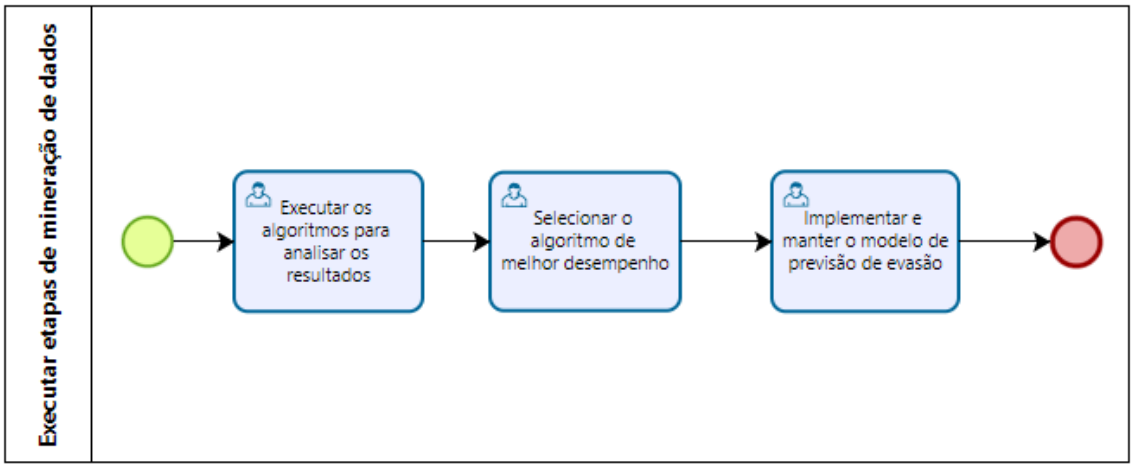

Figura 15: Segundo Macroprocesso.
Fonte: Autor.

Primeiramente, deve-se executar os algoritmos para analisar os resultados por meio do desenvolvimento de um modelo preditivo, explicado com mais detalhes na Seção 5.3. Na sequência, após testes com os algoritmos de classificação, o algoritmo que obtiver melhor desempenho preditivo, de acordo com as métricas de avaliação apresentadas na Tabela 1, deve ser selecionado para geração do modelo. Por fim, vem a tarefa de implementar e manter o modelo de previsão de evasão. Essa tarefa consiste em realimentar a base de dados de treinamento com dados de novos períodos letivos e atualizar o modelo de previsão da evasão, realizando uma nova avaliação de desempenho.

4.2 Framework Escolar

A fim de tornar o método desenvolvido nesta pesquisa mais consistente e replicável, foi elaborado um *framework* escolar, onde, ao seguir o que foi inserido nele, é possível implementar um modelo de previsão da evasão escolar. O *framework* foi desenvolvido por meio de um modelo conceitual, ou diagrama de classes, criado no software Astah[2]. O modelo conceitual tem o objetivo principal de prover um mecanismo capaz de auxiliar no processo de identificação das classes de objetos do domínio do problema utilizando a Linguagem de Modelagem Unificada (UML – *Unified Modeling Language*). Segundo Booch *et al.* (2005), UML é uma linguagem gráfica para visualização, especificação, construção e documentação de artefatos de sistemas de software. O *framework* foi gerado com base nos dados utilizados em Barreto (2019), porém considerando também a possibilidade de aplicá-los à outras instituições de ensino superior. Portanto, foram analisados tanto cursos ofertados atualmente na instituição de onde foram extraídos os dados (Barreto, 2019), quanto cursos existentes em outras instituições.

A Figura 16 apresenta o modelo conceitual que foi gerado como representação do *framework* escolar. Nele, observa-se que o curso do aluno, que pode ser das áreas de bacharelado, licenciatura ou tecnologia, de acordo com as generalizações do modelo, está vinculado a uma instituição de ensino e, além disso, está diretamente associado aos atributos dos alunos que devem ser gerados para desenvolvimento do modelo preditivo por meio da base de dados. Nota-se que cada atributo é extraído num determinado tipo de dados, como *float*, *int* e *boolean* e, para funcionamento adequado do modelo, os mesmos devem ser preservados.

Uma vez que o *framework* foi desenvolvido contemplando a possibilidade de ser aplicado em várias instituições de ensino superior, existem heranças e hierarquias de classes que dão margem a possibilidade de inclusão de novos cursos ao modelo conceitual. Observa-se na Figura 16 que a partir das superclasses Licenciatura, Tecnologia e Bacharelado podem ser acrescentados quaisquer outros cursos dentro destas modalidades de graduação, caso a instituição oferte algum curso diferente dos apresentados. Além disso, há também a superclasse Instituição que permite que sejam colocadas novas instituições no modelo desde que contenham somente os atributos citados na classe Aluno.

Os atributos existentes na classe Aluno são os mesmos apresentados na Tabela 6, cujos critérios de escolha também foram comentados anteriormente. Além disso, eles foram utilizados para a elaboração do modelo e, portanto, são fundamentais para o funcionamento do modelo de previsão da evasão. Resumidamente, o *framework* prevê a inclusão de novas instituições e cursos nas modalidades definidas, desde que os atributos sejam mantidos.

[2] www.astah.net

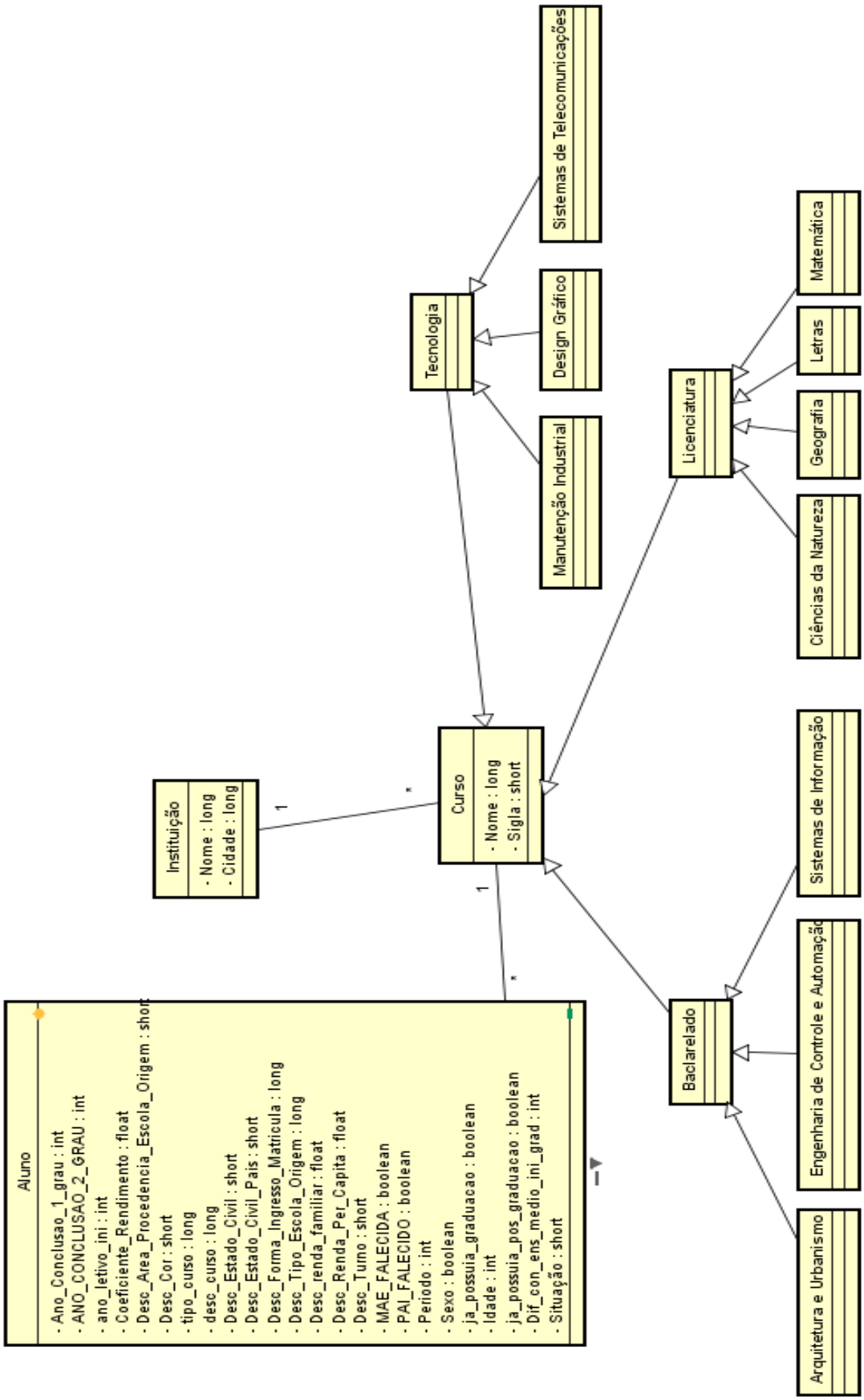

Figura 16: *Framework* escolar.
Fonte: Autor.

4.3 Modelo Preditivo

O modelo de previsão da evasão foi gerado com a biblioteca scikit-learn[3] e a linguagem de programação Python, por meio da plataforma Google Colaboratory[4]. Duas bases de dados foram unificadas e utilizadas como estudo de caso para geração e validação do modelo preditivo, onde foi realizado o experimento. Uma das bases é a mesma utilizada por Barreto (2019), porém com algumas alterações em registros e atributos, como remoção de registros repetidos e atributos não relevantes, de acordo com a análise de uma especialista no domínio. Esta base, com 24 atributos e 824 registros, contém dados de alunos do ensino superior de uma instituição de ensino federal, onde são considerados os semestres 2017.2, 2018.1, 2018.2 e 2019.1. A segunda base possui dados de cursos da mesma instituição federal, porém extraídos do semestre 2019.2. A base possui os mesmos 24 atributos e 282 registros. Com isso, unificando as duas bases, obteve-se um total de 1.106 registros.

Considerando que, para alguns registros, a base apresentava informações faltantes para alguns atributos, foi necessário resolver este problema por meio do complemento dessas informações através dos mais frequentes dentro do mesmo atributo. Para avaliar o modelo, foi utilizada a técnica de validação cruzada estratificada *k-fold*, em que a base de dados é dividida em *k* partições de (aproximadamente) mesmo tamanho. Nesta dissertação, foi adotado $k = 10$. Dessa forma, são realizadas 10 avaliações em que, a cada avaliação, uma partição é usada para teste e as nove restantes (k – 1) são usadas para treinamento. Ao final, para obter o desempenho do modelo preditivo, é calculada a média das 10 avaliações para cada métrica apresentada na Tabela 1. Os algoritmos testados para geração do modelo foram: *Decision Tree*, *Random Forest*, *Support Vector Machine*, *Logistic Regression* e *Multilayer Perceptron*, implementados nas classes DecisionTreeClassifier, RandomForestClassifier, LinearSVC, LogisticRegression e MLPClassifier da biblioteca scikit-learn, respectivamente.

Para todos os algoritmos foram utilizados os valores padrão de seus parâmetros, exceto para DecisionTreeClassifier e RandomForestClassifier, que foram definidos os valores 10 e 8, respectivamente, para o parâmetro *max_depth*, e para LogisticRegression, que foi definido o valor 10.000 para o parâmetro *max_iter*. O algoritmo utilizado está disponível no repositório *GitHub*[5].

5.4 Protótipo da Plataforma

O desenvolvimento do protótipo da plataforma é uma das etapas finais desta pesquisa. Foi desenvolvido o protótipo da tela utilizando a biblioteca de desenvolvimento Web Python chamada Django. A escolha da mesma é fruto de alguns benefícios importantes relatados em Django (2021), a saber:

- Como o modelo preditivo foi desenvolvido em Python, o fato de a plataforma Web também utilizar esta linguagem facilita integração das duas etapas para que a ferramenta funcione de forma correta;
- Possui dezenas de tarefas comuns já prontas para serem reutilizadas, como por exemplo autenticação de usuário, administração de conteúdo, mapas de site, entre outras.

[3] https://scikit-learn.org
[4] https://colab.research.google.com
[5] https://github.com/atilacarvalhojr/ModelodeClassificacao/blob/main/Algoritmo

- Django também ajuda a evitar erros de segurança comuns e proporciona escalabilidade aos sistemas, ou seja, consegue oferecer capacidade de expansão de um sistema sem perda do seu desempenho.

A figura 17 apresenta a tela da interface Web da plataforma já desenvolvida, restando a integração da mesma com o modelo preditivo criado que ficará para trabalhos futuros.

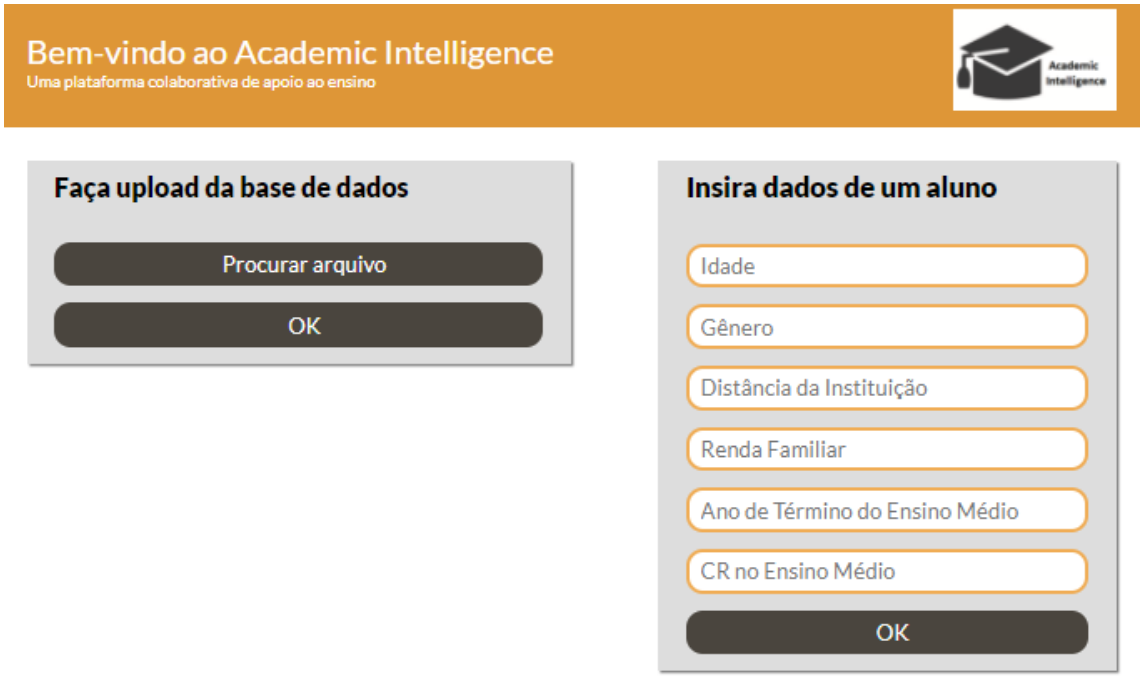

Figura 17: Protótipo da plataforma Web.
Fonte: Autor.

5. RESULTADOS

Os resultados da pesquisa foram gerados a partir do desenvolvimento de um experimento. As bases de dados de uma instituição federal foram unificadas com o intuito de serem utilizadas como estudo de caso e o modelo de classificação foi aplicado com o objetivo de identificar o algoritmo com melhor desempenho.

Para desenvolvimento da etapa de mineração de dados, foram utilizados para treinamento e validação do modelo de predição de evasão, dados do sistema acadêmico de uma instituição de ensino federal referentes aos semestres 2017.2, 2018.1, 2018.2 e 2019.1 oriundos da pesquisa de Barreto (2019) com algumas modificações e o período de 2019.2, como mencionado no Capítulo 4. A base completa contém 24 atributos e 1.106 e registros. Destes, 697 correspondem a alunos que evadiram e 517 se referem a estudantes que concluíram a graduação.

Os algoritmos de classificação identificados nos trabalhos relacionados foram avaliados a partir da base de dados selecionada para analisar os resultados e escolher o que apresentar o melhor desempenho preditivo, a saber: i) *Random Forest*, ii) *Support Vector Machine*, iii) *Logistic Regression*, iv) *Multilayer Perceptron* e v) *Decision Tree*.

A Tabela 8 apresenta os resultados em termos de acurácia e F1-score. Os resultados apontam *Logistic Regression* como o melhor classificador, com acurácia de 92,09% e 91,94% de F1-score. É interessante observar que Decision Tree, que costuma ser uma escolha comum na geração de modelos preditivos em geral, apresentou o pior desempenho, com acurácia de 90,44% e 90,30% de F1-score.

Algoritmo	Acurácia	F1-score
Logistic Regression	92,09%	91,94%
Multilayer Perceptron	91,85%	91,70%
SVM	91,10%	90,94%
Random Forest	91,02%	90,78%
Decision Tree	90,44%	90,30%

Tabela 8: Resultados em termos de acurácia e F1-score para os algoritmos avaliados.
Fonte: Autor.

A Figura 18 apresenta o gráfico da avaliação para o algoritmo com melhor desempenho geral, *Logistic Regression*, em que o eixo *y* corresponde ao valor de F1-score e o eixo *x* se refere à partição de teste avaliada. O gráfico indica que todas as 10 partições de teste avaliadas obtiveram F1-score acima de 90%, exceto a sexta avaliação, que parece ter ocorrido *overfitting*, isto é, o modelo se ajustou demais aos dados de treinamento, com F1-score de aproximadamente 96%, enquanto a avaliação na partição de teste apresentou F1-score de aproximadamente 88%.

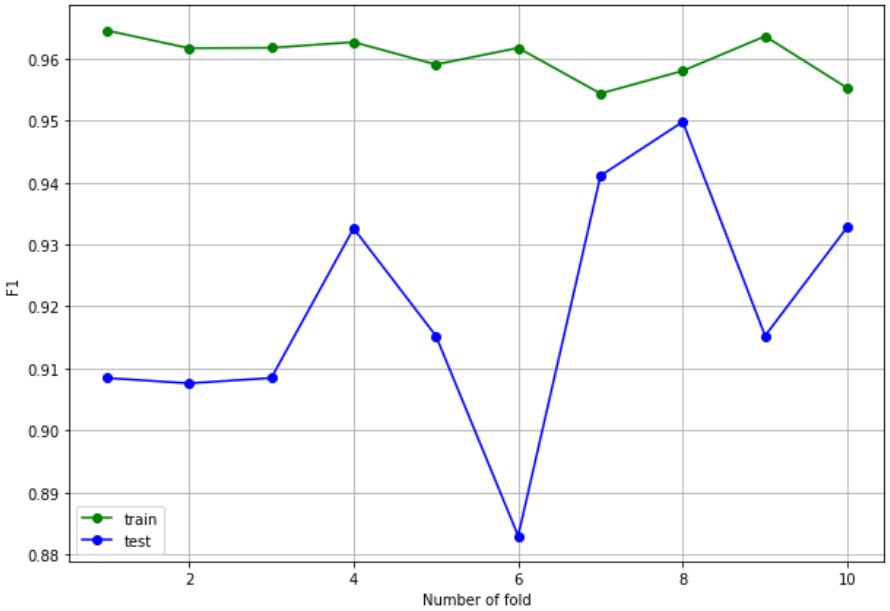

Figura 18: Avaliação do algoritmo *Logistic Regression*.
Fonte: Autor.

Conforme apresentado, a avaliação em cada partição de teste apresenta um determinado valor de F1-score. Com isso, também é possível analisar os acertos a partir da matriz de confusão. A matriz de confusão ajuda a compreender melhor a eficácia do algoritmo. Para tanto, foram geradas duas matrizes de confusão, onde uma é referente a partição 8 (Figura 19), ou seja, aquela que apresentou o melhor desempenho preditivo, e outra é referente a partição 6 (Figura 20), com pior desempenho.

Ao analisar a matriz de confusão da partição com melhor desempenho (Figura 19), observa-se que um caso que na realidade era concluinte foi classificado como evadido e cinco casos de alunos que evadiram foram classificados como concluintes.

		Valor Predito	
		Concluinte	Evadido
Real	Concluinte	51	1
	Evadido	5	64

Figura 19: Matriz de confusão da partição com melhor desempenho.
Fonte: Autor

Já a Figura 20, apresenta a matriz de confusão referente a partição que gerou o pior desempenho. Nota-se que quatro casos de alunos que concluíram foram classificados como evadidos e 10 alunos que evadiram foram classificados como concluintes.

		Valor Predito	
		Concluinte	Evadido
Real	Concluinte	47	4
	Evadido	10	60

Figura 20: Matriz de confusão da partição com pior desempenho.
Fonte: Autor

6. CONSIDERAÇÕES FINAIS

Neste trabalho, foi desenvolvido um método para previsão da evasão no ensino superior com base em um *framework* escolar utilizando mineração de dados. Para tanto, foram utilizados resultados de buscas de documentos nas bases científicas *Scopus*, *Web of Science* e outras. Como ferramenta de pesquisa, utilizando os termos mineração de dados, aprendizado de máquina, evasão e ensino superior. Após a aplicação de alguns critérios foram selecionados oito documentos.

A pesquisa apresenta como resultado um *framework* escolar utilizado para desenvolvimento do modelo preditivo que pode ser replicado em outras instituições de ensino superior desde que mantenha as modalidades de curso e atributos selecionados.

A partir dos resultados, conclui-se que o algoritmo *Logistic Regression* apresentou a melhor acurácia e *F1-score* e deve, portanto, ser utilizado para a construção do modelo preditivo proposto. Em contrapartida, o algoritmo *Decision Tree* apresentou o pior desempenho preditivo.

7. TRABALHOS FUTUROS

Como trabalhos futuros, propõe-se incorporar o método elaborado nesta pesquisa na plataforma Web que será desenvolvida, cujo protótipo já existe, de modo que os gestores escolares possam ter acesso a mesma para identificar com maior facilidade alunos propensos a evadir e realizar os testes necessários. Em relação ao modelo preditivo, propõe-se uma análise mais criteriosa da avaliação, com o objetivo de identificar os registros classificados incorretamente para refinar a qualidade do modelo final. Ademais, outro ponto a ser desenvolvido futuramente é a utilização de bases de dados de diferentes instituições para avaliar o desempenho do modelo em outros contextos estudantis.

Futuramente, podem ser acrescentados dados da assistência estudantil, gerando a possibilidade de inserir atributos socioeconômicos importantes, como saber se o aluno possui necessidades especiais ou não. Além disso, pode ser feito no futuro o planejamento e a realização de um estudo empírico, onde profissionais ligados à área de gestão escolar irão interagir com a plataforma para avaliar os processos de consulta, ou seja, verificando por meio do modelo desenvolvido os alunos com probabilidade de evasão.

Referências

ABMP BRASIL. BPM CBOK V3.0 - Guia para o Gerenciamento de Processos de Negócio Corpo Comum de Conhecimento. 1ª ed. Brasil: ABMP Brasil, 2013. v. 3.

ALPAYDIN, E., 2014. Introduction to machine learning. 3 ed. Massachusetts, MIT Press.

BAKER, R., ISOTANI, S., CARVALHO, A. (2011), Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação, 19(2), p. 3-13.

BAKER, R. S. J. (2014), Educational Data Mining: An Advance for Intelligent Systems in Education. IEEE Intelligent Systems, 29 (3), p. 78- 82.

BARRETO, D. L. Análise da Evasão em Cursos de Ensino Superior do Instituto Federal Fluminense utilizando Mineração de Dados. 2019. 86f. Dissertação (Mestrado em Sistemas Aplicados à Engenharia e Gestão) – Instituto Federal Fluminense, Campos dos Goytacazes/RJ.

BELÉM, C.; SANTOS, L.; LEITÃO A. On the Impact of Machine Learning: Architecture without Architects? In: INTERNATIONAL CONFERENCE, CAAD FUTURES 2019, 18., 2019, Daejeon, Korea. Anai. Disponível em: <http://papers.cumincad.org/data/works/att/cf2019_020.pdf>. Acesso em: 17 Ago. 2019.

BISHOP, C. M. Pattern Recognition and Machine Learning: Information Science and Statistics. Nova Iorque: Springer-Verlag, 2006.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Censo da Educação Superior. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior Acesso em: 22 mar. 2021.

BOOCH, G.; RUMBAUGH, J.; JACOBSON, I. Unified Modeling Language User Guide, The 2Nd Edition. Addison-Wesley Object Technology Series. Addison-Wesley Professional, 2005.

CHADEGANI, A. A. et al. A comparison between two main academic literature collections: Web of Science and Scopus databases. Asian Social Science, Toronto, v. 9, n. 5, p. 18-26, 2013.

COSTA, A. L. da. Evasão dos cursos de graduação da UFRGS em 1985, 1986 e 1987. Porto Alegre: UFRGS, 1991.

COSTA, S. L.; DIAS, S. M. A permanência no ensino superior e as estratégias institucionais de enfrentamento da evasão. Jornal de Políticas Educacionais, 9(17/18), 51-60, 2016.

COSTA, Susane Santos da; CAZELLA, Silvio; RIGO, Sandro José. (2015), Minerando dados sobre o desempenho de alunos de cursos de educação permanente em modalidade EAD: Um estudo de caso sobre evasão escolar na UNA-SUS. RENOTE, v. 12, n. 2.

DA SILVA, P. M. et al. Ensemble Regression Models Applied to Dropout in Higher Education. 2019 8th Brazilian Conference on Intelligent Systems (BRACIS). **Anais**... In: 2019 8TH BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS). Salvador, Brazil: IEEE, out. 2019Disponível em: <https://ieeexplore.ieee.org/document/8923655/>. Acesso em: 14 mar. 2021.

DA SILVA ZANATO, K. Y., VENTURA, T. M., and RIBEIRO, J. M. (2018). Análise da evasão de alunos da área de tecnologia da informação por meio de um banco de dados

orientado a grafos. Revista Eletrônica Argentina-Brasil de Tecnologias da Informação e da Comunicação, 1(8).

DJANGO. The web framework for perfectionists with deadlines. Disponível em: <https://www.djangoproject.com> Acesso em: 02 nov. 2021.

DO NASCIMENTO, R. L. S., DA CRUZ JÚNIOR, G. G., and de ARAÚJO FAGUNDES, R. A. (2018). Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. RENOTE, 16(1).

EMERSON, 2014, "Data Center 2025: Exploring the Possibilities". Emerson Network Power. Acesso em 4 ago. 2017.

FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

FERNANDEZ-GARCIA, A. J. et al. Creating a Recommender System to Support Higher Education Students in the Subject Enrollment Decision. IEEE Access, v. 8, p. 189069–189088, 2020.

FILHO, R. L.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. A Evasão no Ensino Superior Brasileiro. Cadernos de Pesquisa, 37(132), 641-659, 2007.

FREITAS, F. A. DA S. et al. IoT System for School Dropout Prediction Using Machine Learning Techniques Based on Socioeconomic Data. Electronics, v. 9, n. 10, p. 1613, 1 out. 2020.

GONÇALVES, T. C.; SILVA, J. C. DA; CORTES, O. A. C. Técnicas de mineração de dados: um estudo de caso da evasão no ensino superior do Instituto Federal do Maranhão. Revista Brasileira de Computação Aplicada, v. 10, n. 3, p. 11–20, 21 set. 2018.

HAN, J., PEI, J., & KAMBER, M. 3rd ed. Data mining: concepts and techniques. Elsevier, 2011.

Hoed, R. M. (2016). Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação. Universidade de Brasília, Brasília, DF. Obtido de http://repositorio.unb.br/bitstream/10482/22575/1/2016_RaphaelMagalh%C3%A3esHoed.pdf.

HEPPEN, J. B.; THERRIAULT, S. B. Developing Early Warning Systems to Identify Potential High School Dropouts. p. 13, 2008.

HOFFMANN, Ivan Londero; NUNES, Raul Ceretta; MULLER, Felipe Martins. As informações do Censo da Educação Superior na implementação da gestão do conhecimento organizacional sobre evasão. Gestão & Produção, São Paulo, v. 26, n. 2, 9 maio 2019. FapUNIFESP (SciELO).

HUTAGAOL, N.; SUHARJITO, S. Predictive Modelling of Student Dropout Using Ensemble Classifier Method in Higher Education. Advances in Science, Technology and Engineering Systems Journal, v. 4, n. 4, p. 206–211, 2019.

IFF. Apresentação. 2020. Disponível em: https://portal1.iff.edu.br/conheca-o-iffluminense/conheca-o-iffluminense>. Acesso em: 28 nov. 2021.

INEP. Sinopse Estatística da Educação Superior. Acesso em 15 de janeiro de 2021, disponível em Portal INEP: http://portal.inep.gov.br/web/guest/dados.

LIBERATI, A.; ALTMAN, D. G.; TETZLAFF, J.; et al. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. PLoS Medicine, v. 6, n. 7, p. e1000100, 2009.

LINOFF, G. S.; BERRY, M. J. Data mining techniques: for marketing, sales, and customer relationship management. [S.l.]: John Wiley & Sons, 2011.

MANHÃES, L. M. B., DA CRUZ, S. M. S., COSTA, R. J. M., ZAVALETA, J., ZIMBRÃO, G. (2011). Previsão de estudantes com risco de evasão utilizando técnicas de mineração de dados. In Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie), volume 1.

MÁRQUEZ-VERA, Carlos; CANO, Alberto; ROMERO, Cristobal; et al. Early dropout prediction using data mining: a case study with high school students. Expert Systems, v. 33, n. 1, p. 107–124, 2016.

MASHILOANE, Lebogang; MCHUNU, Mike. Mining for Marks: A Comparison of Mining Intelligence And Knowledge Exploration, Joanesburgo, p.541-552, 2013. Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-03844-5_54.

MEC/ANDIFES/ABRUEM/SESU. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. Revista Avaliação, Campinas, SP, n. 2, p. 55-65, julho 1996.

NAGAI, N. P.; CARDOSO, A. L. J. (2017). A evasão universitária: Uma análise além dos números. Revista Estudo & Debate, 24(1).

NAGY, M.; MOLONTAY, R. Predicting Dropout in Higher Education Based on Secondary School Performance. 2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES). Anais... In: 2018 IEEE 22ND INTERNATIONAL CONFERENCE ON INTELLIGENT ENGINEERING SYSTEMS (INES). Las Palmas de Gran Canaria: IEEE, jun. 2018Disponível em: < https://ieeexplore.ieee.org/document/8523888/ >. Acesso em: 14 mar. 2021

NOGARE, D., ZAVASCHI, T., 2016, Análise Preditiva com Azure Machine Learning e R. São Paulo, B2U Editora.

PERCHINUNNO, P.; BILANCIA, M.; VITALE, D. A Statistical Analysis of Factors Affecting Higher Education Dropouts. Social Indicators Research, 19 dez. 2019.

PESSOAL, D. (2018). Instituto nacional de estudos e pesquisas educacionais anísio teixeira. Disponível em: < http://ingressodiscente.univasf.edu.br/arquivos/ps_icg_2018/PS_ICG_2018_Edital_n113_Sesu_MEC _altera_edital_107.pdf >. Acesso em: 14 mar. 2021.

PRADEEP, Anjana; DAS, Smija; KIZHEKKETHOTTAM, Jubilant J. Students dropout factor prediction using EDM techniques. *In*: 2015 International Conference on Soft-Computing and Networks Security (ICSNS). Coimbatore, India: IEEE, 2015, p. 1–7. Disponível em: <http://ieeexplore.ieee.org/document/7292372/>. Acesso em: 16 maio 2020.

P.-N. TAN, M. STEINBACH, A. KARPATNE, V. KUMAR. Introduction to Data Mining, Pearson, 2nd Edition, 2018.

PRESTES, Emília Maria da Trindade; FIALHO, Maríllia Gabriella Duarte. Evasão na educação superior e gestão institucional: o caso da Universidade Federal da Paraíba.

Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 26, n. 100, p.869-889, jul. 2018. FapUNIFESP (SciELO).

SALES, A.; BALBY, L.; CAJUEIRO, A. Exploiting Academic Records for Predicting Student Drop Out: a case study in Brazilian higher education. Journal of Information and Data Management, 7(2), 166-180, 2016.

SANI, N. S. et al. Drop-Out Prediction in Higher Education Among B40 Students. International Journal of Advanced Computer Science and Applications, v. 11, n. 11, 2020.

SOUZA, C. T.; DA SILVA, C.; GESSINGER, R. M. Um estudo sobre evasão no ensino superior do Brasil nos últimos dez anos. Congressos CLABES. 2016.

SOUZA, I. M. de. Causas da evasão nos cursos de graduação da Universidade Federal de Santa Catarina. 1999. 150f. Dissertação (Mestrado em Administração) – Programa de Pós-graduação em Administração, Centro Socioeconômico, Universidade Federal de Santa Catarina, Florianópolis, 1999.

SOUZA, S. (2008). Evasão no ensino superior: um estudo utilizando a mineração de dados como ferramenta de gestão do conhecimento em um banco de dados referente à graduação de engenharia. PhD thesis, Dissertação de Mestrado. Coppe–UFRJ, Rio de Janeiro.

SOUZA, S. M. Z. L.; OLIVEIRA, R. P. de; GONÇALVES, N. G. A evasão dos alunos do programa de Pós-Graduação da FEUSP: 1990 a 2000. Avaliação: Revista de rede de avaliação institucional da educação superior. Campinas, v. 8, n. 3, p. 191-228, set. 2003.

T. HASTIE, R. TIBISHARI, J. FRIEDMAN, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, 2nd Edition, 2016.

WOODS, V., 2015, “Gartner Says It's Not Just About Big Data; It's What You Do With It: Welcome to the Algorithmic Economy”. Gartner Group. Disponível em < http://www.gartner.com/newsroom/id/3142917>. Acesso em 5 ago. 2017.