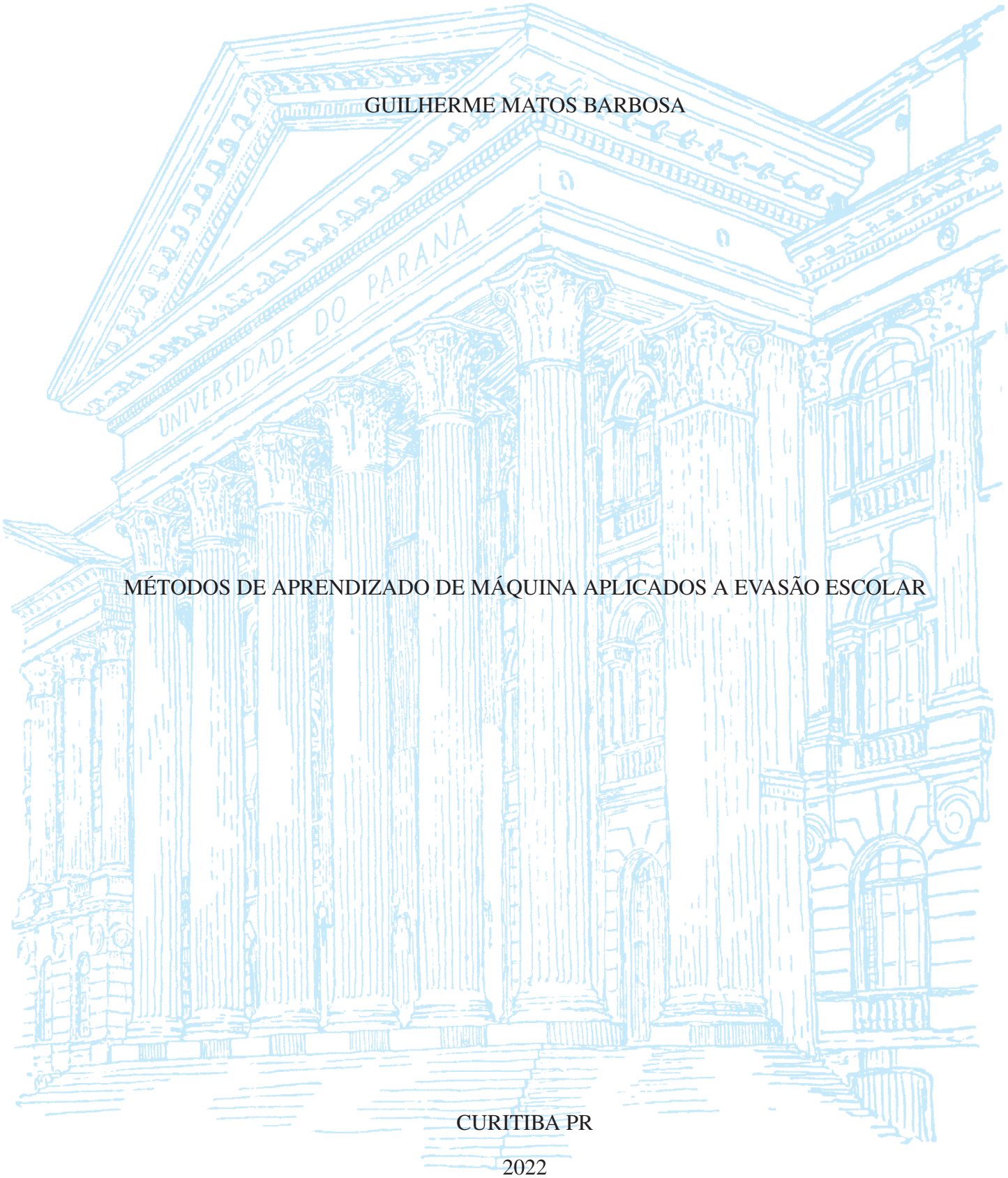
UNIVERSIDADE FEDERAL DO PARANÁ

GUILHERME MATOS BARBOSA



MÉTODOS DE APRENDIZADO DE MÁQUINA APLICADOS A EVASÃO ESCOLAR

CURITIBA PR

2022

GUILHERME MATOS BARBOSA

MÉTODOS DE APRENDIZADO DE MÁQUINA APLICADOS A EVASÃO ESCOLAR

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Informática no Programa de Pós-Graduação em Informática, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: Prof. Dr. Wagner Hugo Bonat.

CURITIBA PR

2022

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA CIÊNCIA E TECNOLOGIA

Barbosa, Guilherme Matos
Métodos de aprendizado de máquina aplicados a evasão escolar. / Guilherme Matos Barbosa. – Curitiba, 2022.
1 recurso on-line : PDF

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Ciências Exatas, Programa de Pós-Graduação em Informática.

Orientador: Prof. Dr. Wagner Hugo Bonat.

1. Mineração de dados (Computação). 2. Aprendizagem. 3. Evasão escolar. I. Bonat., Wagner Hugo. II. Universidade Federal do Paraná. Programa de Pós-Graduação em Informática. III. Título.

Bibliotecária: Roseny Rivelini Morciani CRB-9/1585

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação INFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da Dissertação de Mestrado de **GUILHERME MATOS BARBOSA** intitulada: **MÉTODOS DE APRENDIZADO DE MÁQUINA APLICADOS A EVASÃO ESCOLAR**, sob orientação do Prof. Dr. WAGNER HUGO BONAT, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 22 de Junho de 2022.

Assinatura Eletrônica
27/06/2022 10:07:02.0
WAGNER HUGO BONAT
Presidente da Banca Examinadora

Assinatura Eletrônica
27/06/2022 10:50:54.0
PAULO JUSTINIANO RIBEIRO JUNIOR
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
27/06/2022 10:03:26.0
LUIZ EDUARDO SOARES DE OLIVEIRA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Documento assinado eletronicamente de acordo com o disposto na legislação federal Decreto 8539 de 08 de outubro de 2015.
Gerado e autenticado pelo SIGA-UFPR, com a seguinte identificação única: 199602
Para autenticar este documento/assinatura, acesse https://www.prppg.ufpr.br/siga/visitante/autenticacaoassinaturas.jsp e insira o codigo 199602

A meu amor Danielle, aos meus pais Jandir e Sandra, a minha irmã Gabriella, aos meus amigos e todos que sempre me deram suporte nos meus estudos, trabalhos e projetos.

AGRADECIMENTOS

Minha maior gratidão a Deus, meus pais, pelo dom da vida e pela oportunidade de vivenciar toda essa trajetória acadêmica, a minha irmã, Gabriella, que sempre teve o meu lado me dando apoio, essa vitória é inteiramente de vocês!

Ao meu anjo da guarda por sempre ter me guiado para o melhor caminho e a Nossa Senhora Aparecida por interceder sempre que precisei.

Ao meu orientador, Professor Doutor Wagner Hugo Bonat, pela dedicação e compromisso para realização da pesquisa.

Aos meus amigos Andrea e Ricardo pelo incentivo para realizar o mestrado.

A Universidade Federal do Paraná, por oferecer toda estrutura necessária ao desenvolvimento dessa pesquisa.

Agradeço também a Danielle, meu amor, sempre a disposição de ajudar e me consolar nos momentos difíceis e a sorrir e comemorar nos momentos de vitória.

RESUMO

A evasão escolar pode ser definida como a descontinuação de um estudante no seu ensino, podendo ocorrer de diferentes formas e, esse fenômeno, está cada vez mais presente no cenário do ensino superior, inclusive na UFPR (Universidade Federal do Paraná). Torna-se importante então buscar maneiras de reduzir a taxa de evasão e a mineração de dados e as técnicas de aprendizagem de máquina permitem identificar padrões e gerar modelos computacionais que podem predizer se um aluno será um evasor. Na UFPR existe o SIGA (Sistema de Gestão Acadêmica), um sistema integrado que contempla informações sobre matrícula, cursos, professores, disciplinas, frequências, notas, entre outras. Este trabalho, utilizando dados do SIGA, tem como objetivo usar modelos de classificação para predizer quais alunos estão em risco de evasão, assim como identificar os atributos mais determinantes. Para isso, foram aplicadas os algoritmos de regressão logística, árvore de decisão, *k-Nearest Neighbours*, *Support Vector Machine* e *random forest*.A regressão logística, a árvore de decisão e o *random forest* permitiram identificar que as variáveis mais significativas foram categoria de cota, forma de ingresso, setor de estudo, e considerando o primeiro semestre: índice de rendimento , carga curricular, número de reprovações por frequência e notas. Os melhores resultados de predição foram obtidos pelos algoritmos *random forest* com AUC de 0,863 e acurácia de 0,734 e o SVM com AUC de 0,847 e acurácia de 0,741.

Palavras-chave: Evasão escolar. Mineração de dados. Aprendizagem de máquina.

ABSTRACT

School dropout can be defined as the discontinuation of a student in his or her education, and it can occur in different ways, and this phenomenon is increasingly present in the higher education scenario, including at UFPR (Federal University of Paraná). It becomes important then to look for ways to reduce the dropout rate and data mining and machine learning techniques allow to identify patterns and generate computational models that can predict if a student will be a dropout. At UFPR there is SIGA (Academic Management System), an integrated system that includes information about enrollment, courses, professors, subjects, frequencies, grades, among others. This work, using data from SIGA, aims to use classification models to predict which students are at risk of dropping out, as well as to identify the most determinant attributes. To do this, the algorithms logistic regression, decision tree, k-Nearest Neighbours, Support Vector Machine and random forest were applied.The logistic regression, decision tree and random forest allowed us to identify that the most significant variables were quota category, entrance form, study sector, and considering the first semester: performance index , curricular load, number of failures by frequency and grades. The best prediction results were obtained by the algorithms random forest with AUC of 0.863 and accuracy of 0.734 and the SVM with AUC of 0.847 and accuracy of 0.741.

Keywords: School dropout. Data mining. Machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ACRÔNIMOS

AUC	Area Under the ROC Curve
AM	Aprendizagem de Máquina
AD	Árvores de Decisão
CES	Censo da Educação Superior
CRISP-DM	Cross Industry Standard Process for Data Mining
CEUPB	Comissão Especial de Estudos sobre Evasão nas Universidades Públicas Brasileiras
ENEM	Exame Nacional do Ensino Médio
IES	Instituições de Ensino Superior
IA	Inteligência Artificial
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KNN	k-Nearest Neighbors
MEC	Ministério da Educação
NB	Naive Bayes
PNP	Plataforma Nilo Peçanha
PROGRAD	Pró-Reitoria de Graduação e Educação Profissional
PSE	Processo Seletivo Estendido
REUNI	Programa de Planos de Reestruturação e Expansão das Universidades Federais
ROC	Receiver Operating Characteristic
SEMMA	Sample, Explore, Modify, Model, and Assess
SIE	Sistema de Informações para o Ensino
SIGA	Sistema de Gestão Acadêmica
SVM	Máquina de Vetores de Suporte
SQL	Structured Query Language
UFPR	Universidade Federal do Paraná
UnB	Universidade de Brasília
UFPel	Universidade Federal de Pelotas

SUMÁRIO

1 INTRODUÇÃO

A evasão escolar é a interrupção do ciclo de estudos de um estudante junto a uma instituição de ensino. É um fenômeno que possui vários aspectos complexos, com diferentes características, com incidência internacional e tem forte impacto nas instituições (Santos e Guimarães, 2019).

Deve-se observar a evasão sob dois panoramas para um melhor entendimento. Um deles é a média anual de evasão, que descobre quantos alunos evadiram em um certo período calculando o percentual de alunos matriculados em um curso que, ainda não formado, não efetivou a matrícula no semestre/ano seguinte. Outro é a evasão total que após a entrada do aluno na instituição não finalizou a formação em um número de anos assim verificando a taxa de formandos na totalidade (Silva et al., 2007).

O INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira) destaca que em 2017, 16,41% dos alunos se desvincularam de suas instituições. Quando considerado apenas nível de ensino superior esse valor atinge 11,56%, acarretando uma grande perda de investimentos tanto para a instituição quanto para o aluno evadido (Costa, 2021).

A taxa crescente de evadidos tornou-se um dos maiores entraves da administração das instituições de ensino que vêm buscando diversas maneiras de identificar de maneira precoce e combater de maneira mais eficaz o aumento deste índice (Teodoro e Kappel, 2020).

A problemática da evasão torna-se ainda mais laboriosa, pois os dados no ensino superior não são emitidos de maneira oficial devido à dificuldade de caracterizar o aluno evadido que pode ter diversas faces: aluno que apenas abandonou seu curso e instituição; aluno que deixou o curso mas permaneceu na instituição, porém agora em outro curso; aluno vinculado em mais de uma instituição que pode evadir-se de uma permanecendo ativo nas restantes; aluno que evadiu de uma disciplina (Assis, 2017).

A Plataforma Nilo Peçanha (PNP), que tem a finalidade de unir e apresentar estatísticas oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica, mostra que no ano base de 2018 o número de evadidos no ensino superior chegou a 14%, no ano base de 2019 o índice foi de aproximadamente 13%, correspondente principalmente à evasão por abandono (7,35%), seguido de desligamento (4,83%), cancelamento (0,26%), transferência externa (0,24%), transferência interna (0,03%) e reprovação (0,02%) (PNP, 2018, 2019).

Para minimizar o baixo desempenho e evitar que os alunos descontinuem seus estudos é necessário o desenvolvimento de novas ferramentas. Diante disso, a comunidade científica vem buscando formas de criar modelos conceituais que possam descrever adequadamente a evasão escolar, pois eles dão suporte para melhor compreender, caracterizar, explicar e até prever esse importante fenômeno. Um exemplo disso é a área de ciência de dados aplicada a dados educacionais, que vem proporcionando desenvolver modelos quantitativos fortemente baseados em dados para caracterização do fenômeno da evasão escolar e caracterizar os fatores que contribuem substancialmente para o fenômeno e consequentemente servir de ferramenta de predição (Barros, 2020).

A ciência de dados pode ser definida como um processo que utiliza técnicas estatísticas e computacionais para resolver o problema da descoberta de conhecimentos valiosos em grandes base de dados, para isso utiliza-se de algoritmos capazes de vasculhar as bases de dados com a finalidade de extrair modelos que descrevam padrões interessantes escondidos em uma montanha de dados (Corrêa, 2019).

Este conjunto de métodos objetiva apoiar decisões baseadas em dados, não só entendendo os dados passados mas também realizando análises de forma preditiva, por exemplo, utilizando técnicas de mineração de dados, como reconhecimento de padrões e análise exploratória, e aprendizagem de máquina (Escovedo e Koshiyama, 2021).

Na literatura há diversos trabalhos baseados em ciência de dados e inteligência artificial, geralmente focados nos modelos de aprendizagem para predição de desempenho ou evasão do aluno. Essas pesquisas conseguem lidar com um grande volume de dados e outras restrições como valores ausentes, dependência, correlação, desbalanceamento de dados, alta dimensionalidade, normalidade e relações não-lineares, produzindo melhores precisões na predição (Barros, 2020).

A mineração de dados e aprendizagem de máquina (AM) vem sendo empregada como uma ferramenta de combate à evasão. Através das técnicas da AM, redes neurais, por exemplo, é possível identificar padrões que podem ser utilizados para gerar um modelo computacional propiciando assim classificar se um aluno é provável concluinte ou provável evasor baseada em suas características e nas do curso em que está inserido. Outras técnicas permitem realizar uma avaliação interna do modelo facilitando identificar os atributos mais determinantes na classificação como, por exemplo, árvores de decisão e *random forest* (Teodoro e Kappel, 2020).

Considerando esta problemática o objetivo da pesquisa é utilizar os modelos de classificação para identificar padrões característicos de alunos com maior tendência de evadir, assim como identificar e quantificar os atributos mais determinantes utilizando os dados coletados do Sistema de Gestão Acadêmica da Universidade Federal do Paraná (SIGA/UFPR).

1.1 OBJETIVO GERAL

O presente trabalho tem como objetivo construir modelos que possibilitem identificar e estimar os fatores de risco que estão associados à evasão de alunos de graduação da Universidade Federal do Paraná - UFPR.

1.2 OBJETIVOS ESPECÍFICOS

Para atingir o objetivo principal, faz-se necessário obter alguns objetivos mais específicos, que são:

- Realizar levantamento bibliográfico para conhecer as possíveis causas da evasão, os modelos e teorias relacionadas à questão;
- Verificar os registros na base de dados quanto às informações disponíveis;
- Selecionar, relacionar e correlacionar as características mais relevantes para compor o banco de dados das amostras;
- Realizar estatística descritiva das variáveis utilizadas no estudo;
- Preparar os conjuntos de dados para treinamento e teste dos modelos;
- Treinar, testar e validar os modelos;
- Analisar os resultados obtidos com o modelos propostos.

1.3 ORGANIZAÇÃO DO TRABALHO

No capítulo 2 consta toda a fundamentação teórica necessária para o desenvolvimento da ideia principal da pesquisa, com a apresentação dos conceitos sobre a evasão no ensino superior, o Sistema Integrado de Gestão Acadêmica, à mineração de dados, aprendizado de máquina e sobre as técnicas de classificação que serão utilizadas.

No terceiro capítulo são apresentados os trabalhos que se relacionam diretamente ao contexto abordado nesta pesquisa. No capítulo quatro, é apresentado a metodologia da pesquisa e a forma como ela foi conduzida, bem como quais foram os dados usados, os experimentos conduzidos e o que eles representam.

O capítulo cinco consiste nos resultados obtidos e por fim o capítulo 6 contém as conclusões sobre a pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Para que seja possível compreender esse trabalho, alguns conceitos precisam ser explorados de forma mais detalhada. Este capítulo aborda uma revisão de literatura a respeito da evasão no ensino superior, além de uma breve introdução ao Sistema Integrado de Gestão Acadêmica (SIGA), à mineração de dados, aprendizado de máquina e sobre as técnicas de classificação que serão utilizadas.

2.1 EVASÃO NO ENSINO SUPERIOR

Em 1995, o Ministério da Educação (MEC) iniciou pesquisas relacionadas à evasão através de um seminário realizado pela agenda governamental, neste evento se originou a Comissão Especial de Estudos sobre Evasão nas Universidades Públicas Brasileiras (CEUPB) com o objetivo de desenvolver estudo sobre os desempenhos das Instituições de Ensino Superior Públicas, Federais e Estaduais (Soares, 2020). Para o Governo Federal o estudo sobre políticas de combate a evasão é de grande importância, pois a alta taxa de evasão significa desperdício de recursos (Pinheiro et al., 2018).

Assis (2017), define a evasão como o desligamento do aluno durante seu trajeto escolar sem que seja por óbito ou pela conclusão de seu curso. Esta condição é considerada uma das maiores problemáticas nas instituições de ensino superior e apenas em 2000 começou a aumentar as pesquisas sobre o assunto no Brasil (Baker et al., 2011; Soares, 2020).

Dada a importância da problemática, a evasão é o foco de estudo em vários países e pode ser caracterizada em 3 níveis: evasão do curso, da instituição e do sistema. A evasão do curso consiste no aluno abandonar seu curso de origem, a institucional, quando o aluno desliga-se da instituição que estava matriculado e evasão do sistema é quando o discente abandona o ensino superior por completo (Martins, 2018).

Em algumas pesquisas, quando os dados eram disponibilizados de maneira agrupada por curso pelo INEP, a evasão era estabelecida através de um método de estimativa utilizando o número de alunos matriculados, concluintes e ingressantes em cada ano. Outras vertentes têm pesquisas que utilizam dados internos da instituição para calcular a evasão e para entender os fatores da desistência utilizam questionários ou entrevistas com os alunos evadidos. Mais recentemente, com métodos quantitativos, estudos vêm sendo realizados utilizando técnicas de estatística como análise de sobrevivência, redes neurais e análise de cluster (Marques, 2020).

Duas categorias são regularmente encontradas nas pesquisas e publicações quando se trata das causas de evasão: quando o discente se desvincula por não conseguir acompanhar os conteúdos das aulas por falta de base para os conteúdos e a outra trata da evasão ocasionada pela falta de programas por parte da instituição para ajudar os alunos que ingressaram através de cotas, o que dificulta o aluno continuar vinculado a instituição (Santos e Guimarães, 2019).

Os fatores que podem ser citados e que levam a evasão de acordo com Marques (2020) são a falta de comprometimento do discente com o curso, a falta de estrutura familiar, falta de envolvimento em atividades acadêmicas, baixo desempenho escolar, Instituições de Ensino Superior (IES) com condições precárias além da desmotivação que o aluno pode ter em relação à perspectiva da profissão. Nos últimos anos, o número de vagas em IES aumentaram consideravelmente, porém o percentual de formados não segue esta regra e este fato pode ser justificado pela alta taxa de evasão.

O Programa de Planos de Reestruturação e Expansão das Universidades Federais (REUNI) e as instituições de ensino superior, em seus contratos de gestão, visam manter a margem de evasão em 10% nos cursos de graduação. Entretanto, esse número ainda está longe de ser alcançado pelas universidades (Santos e Guimarães, 2019).

A evasão é algo que gera prejuízos tanto na esfera pública quanto privada. No setor público reflete como uma perda de receita investida na formação de mão de obra, que não se tornará qualificada, e nas instituições privadas acarreta principalmente em um menor lucro, além disso, há desgaste emocional do aluno evadido (Hoffmann et al., 2019).

Quando se analisa a evasão no cenário de instituições públicas deve ser considerado o agravante social, visto que além das perdas econômicas os alunos evadidos indiretamente impedem outros alunos de adentrarem no sistema ao ocuparem vagas mas não concluírem o curso (Branco, 2020).

Em muitos casos mais de um fator interfere para que o aluno se desvincule da instituição de ensino. A base de dados das instituições possibilitam observar algumas das causas e motivações dos estudantes a abandonar a universidade. Esses dados apresentam muitas características de discentes, docentes e da instituição permitindo adquirir conhecimentos valiosos no combate à evasão (Andriola et al., 2006; Saccaro et al., 2019).

2.2 SIGA

O SIGA (Sistema Integrado de Gestão Acadêmica) trata-se de um sistema desenvolvido pela UFPR, através da colaboração entre a equipe técnica e a comunidade acadêmica, científica e extensionista.

Seu principal objetivo é ter um sistema integrado e baseado nos processos e necessidades da instituição no qual docentes, discentes e técnicos administrativos tenham acesso a todas as bases e sistemas da UFPR em um único acesso. Com o passar do tempo foram adicionados no SIGA novos módulos que contemplam informações acadêmicas importantes como as pesquisas realizadas na universidade e as atividades de extensão.

A idealização deste sistema teve início em 2009, sendo primariamente para gerenciar a pós-graduação que não possuía nenhuma ferramenta de gestão. Em 2014, o sistema finalmente começou a ser utilizado até que em 2015 todos os cursos da pós-graduação contavam com o SIGA.

Em 2019 o SIGA passou por um projeto piloto da Pró-reitoria de Graduação e Educação Profissional (Prograd) no qual 12 cursos da UFPR, sendo dez de graduação e dois técnicos, foram escolhidos para terem o Portal do Aluno e do Sistema de Informações para o Ensino (SIE) substituídos, após esse período avaliaram e recomendaram ajustes e modificações, proporcionando mais tarde que todos os cursos de graduação fossem gerenciados pelo SIGA, possibilitando a extração de indicadores confiáveis das atividades dos discentes e dos docentes e das ações realizadas por departamentos e setores.

Desde o início de seu desenvolvimento e sua implantação o SIGA vem em constante atualização para atender melhor às necessidades da UFPR, além de permitir o registro de mais informações acadêmicas e melhoria de sua utilização (SIGAUFPR, 2021).

2.3 MINERAÇÃO DE DADOS

Em diferentes áreas do conhecimento como educação, indústria, ciência e engenharia, a coleta e o armazenamento de dados vem ocorrendo em um ritmo acelerado e isso criou a necessidade de técnicas que consigam traduzir ou extrair conhecimento a partir de tais dados,

podendo assim gerar informações para analisar problemas cotidianos como a evasão escolar em uma instituição (Assis, 2017; Marquesone, 2018; Costa, 2021).

A mineração de dados é uma alternativa para analisar um grande volume de registros, utilizando técnicas estatísticas, matemáticas e de aprendizado de máquina (AM), é possível extrair informações úteis e padrões desconhecidos (Marquesone, 2018). Foram desenvolvidos diversos padrões na busca por procedimentos e técnicas para extrair conhecimento de grandes volumes de dados, tanto na academia como nas indústrias (Assis, 2017).

Dos processos encontrados na literatura pode-se citar duas que se destacaram: SEMMA (*Sample, Explore, Modify, Model, and Assess*) e CRISP-DM (*Cross Industry Standard Process for Data Mining*) (Marquesone, 2018).

Esses padrões seguem um processo sistemático que possibilita obter sucesso ao realizar a análise de dados. Em geral, as etapas são (Hoed, 2016; Marquesone, 2018; Melo, 2019):

- entendimento do negócio: aqui são definidas as perguntas, o objetivo da análise de dados e o plano a ser seguido;
- entendimento dos dados: etapa utilizada para coletar e explorar os dados, aumentando a compreensão sobre sua estrutura, atributos e contexto;
- preparação dos dados: após a análise exploratória, inicia-se o processo de limpeza, filtragem, estruturação, redução e integração dos dados;
- modelagem dos dados: envolve as tarefas de seleção dos dados, definição e construção do modelo. Nessa fase que é selecionado os algoritmos de AM para a solucionar o problema;
- validação do modelo: os resultados gerados pelo modelo são avaliados, para verificar se a precisão obtida está satisfatória e coesa;
- utilização do modelo: após serem validados, os resultados dos modelos são utilizados e monitorados.

2.4 APRENDIZADO DE MÁQUINA

A inteligência artificial (IA) é uma ciência ampla que objetiva simular habilidades humanas e tem o aprendizado de máquinas (AM) como um subcampo. AM é um método de análise de dados que automatiza o desenvolvimento de modelos analíticos a partir de dados históricos para tomar decisões, identificar padrões e fazer previsões com intervenção humana mínima (Souza, 2020).

AM não é uma ciência nova, mas atualmente vem ganhando espaço como ferramenta para analisar o volume crescente de dados em diferentes âmbitos. Seus algoritmos aprendem com os cálculos anteriores para reconhecer padrões, produzir decisões, resultados confiáveis e reproduzíveis. O aspecto iterativo do aprendizado de máquina é importante pois conforme os modelos são expostos a novos dados são capazes de se adaptar de forma independente (Escovedo e Koshiyama, 2021).

Podem ser descritas 3 categorias de AM: Aprendizado não supervisionada, supervisionada e semi-supervisionado. Na modalidade não supervisionada o algoritmo deve agrupar os dados baseando-se em características semelhantes, sem a necessidade da variável que se deseja prever. Na supervisionada, o algoritmo busca associações entre as variáveis preditoras e a que deseja prever em um conjunto de dados. Na semi-supervisionado utiliza uma mescla do

aprendizado não supervisionado com aprendizado supervisionado tendo maior utilização com grande volume de dados (Souza, 2020; ENAP, 2020).

A aprendizagem supervisionada baseia-se em modelos preditivos e dentre as tarefas mais comuns temos a classificação que analisa dados de entrada, rotula e prevê uma categoria como por exemplo a aprovação ou reprovação de um aluno, e a regressão que também analisa dados de entrada mas com a diferença que determina um valor numérico (contínuo ou discreto) como saída (Souza, 2020).

Nesse trabalho serão utilizados algoritmos para classificação, como: regressão logística e árvore de decisão. Segundo Assis (2017) esses algoritmos permitem interpretar o modelo através dos coeficientes gerados para cada atributo que indicam seu peso na classificação da classe de estudo. Outros algoritmos testados são *k-Nearest Neighbours* (kNN) e *Support Vector Machine* (SVM).

2.5 CLASSIFICAÇÃO

A classificação é uma ação inerente a espécie humana, desde um simples filtro de e-mails, quanto classificações mais complexas como a identificação de objetos celestes em imagens telescópicas. Realizar esse processo manualmente é viável em conjuntos de dados pequenos e simples com apenas alguns atributos, conjuntos de dados maiores e mais complexos exigem uma solução automatizada (Tan et al., 2005).

Os dados, para serem classificados, consistem em uma coleção de instâncias (registros). Cada uma destas instâncias é caracterizada pela tupla (x, y), onde x é o conjunto de atributos, valores que descrevem a instância, e y é o rótulo da classe da instância. O conjunto de atributos x pode conter atributos de qualquer tipo, enquanto o rótulo da classe y deve ser categórico. A figura 2.1. ilustra a ideia geral por trás da classificação (Tan et al., 2005; Escovedo e Koshiyama, 2021).

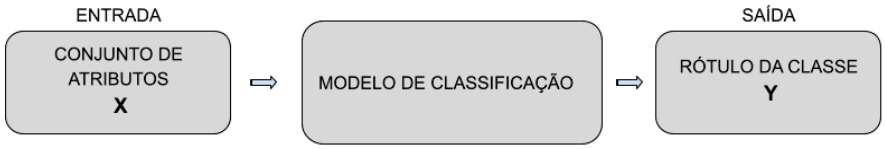


Figura 2.1: Ideia geral da classificação. Adaptado de (Tan et al., 2005)

Um modelo de classificação trata-se de uma representação abstrata do relacionamento entre o conjunto de atributos e o rótulo da classe podendo ser representado de muitas maneiras: como uma árvore, uma tabela de probabilidade, um vetor de parâmetros com valor real ou então, podemos expressá-lo matematicamente como uma função alvo f que toma como entrada o atributo conjunto x e produz uma saída correspondente à classe prevista rótulo. O modelo classifica corretamente uma instância (x, y) se f (x) = y (Tan et al., 2005; Escovedo e Koshiyama, 2021).

O modelo de classificação é usado como um modelo preditivo para classificar instâncias não rotuladas anteriormente. Para ser considerado um bom modelo de classificação é necessário fornecer previsões precisas com um tempo de resposta rápido. Além disso, serve como um modelo descritivo para identificar quais características distinguem instâncias de diferentes classes. Isso é particularmente útil para aplicações críticas, como evasão escolar, onde é insuficiente ter um modelo que faça uma previsão sem justificar como chega a tal decisão (Tan et al., 2005).

Para criação do modelo de classificação podem ser utilizados diferentes algoritmos, que podem ser: regressão logística, árvores de decisão, *naive bayes*, *k-nearest neighbors*, máquina de vetores de suporte.

Esses algoritmos foram escolhidos por aparecerem no levantamento bibliográfico em diferentes trabalhos e segundo Assis (2017) permitem interpretar o modelo através dos coeficientes gerados para cada atributo que indicam seu peso na classificação da classe de estudo como é o caso da regressão logística e árvore de decisão.

Estes algoritmos são avaliados utilizando métricas como acurácia, precisão, sensibilidade (ou *recall*), Curva ROC (*Receiver Operating Characteristic*) e AUC (*Area Under the ROC Curve*) (Escovedo e Koshiyama, 2021).

2.6 REGRESSÃO LOGÍSTICA

A regressão logística tem aplicabilidade em problemas de classificação. Esse algoritmo é utilizado principalmente para prever variáveis de valores discretos, isto é, valores binários, sim/não, verdadeiro/falso, a partir de um grupo de variáveis independentes (Delen, 2011; Escovedo e Koshiyama, 2021).

De acordo com Alghamdi et al. (2017) a regressão logística é um classificador estatístico linear que fornece a probabilidade de prever a classe rotulada do tipo categórico usando vários atributos. O classificador de modelo de previsão mede o relacionamento entre os atributos e a classe rotulada.

Em uma equação representando a regressão logística os valores de entrada (x) são combinados linearmente usando pesos ou valores de coeficiente para prever um valor de saída (y) (Escovedo e Koshiyama, 2021).

Como a variável resposta é binária, ela não pode ser modelada diretamente por regressão linear. Portanto, em vez de prever uma estimativa pontual do evento, ele constrói o modelo para prever as hipóteses de sua ocorrência (Delen, 2011). Internamente é calculado a probabilidade de um evento acontecer através do ajuste dos dados a uma função *logit* (Escovedo e Koshiyama, 2021). Está função é descrita por 2.1:

$$logit(p) = log\left(\frac{p}{1-p}\right) = log(p) - log(1-p) \tag{2.1}$$

Como o valor de saída está entre 0 e 1, como a probabilidade prevê, o algoritmo utiliza a função logística, também denominada de sigmóide (dada pela equação 2.2), que fornece uma curva em forma de 'S' que pode pegar qualquer número com valor real e mapeá-lo para um valor entre 0 e 1 (Escovedo e Koshiyama, 2021). A figura 2.2 ilustra uma transformação logística entre -6 e 6.

$$f(x) = \left(\frac{1}{1+e^{-x}}\right) \tag{2.2}$$

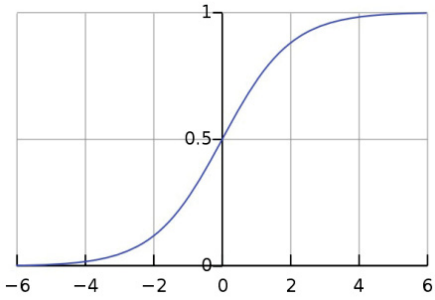


Figura 2.2: Exemplo de transformação logística. Fonte: (Escovedo e Koshiyama, 2021).

Embora a regressão logística seja uma ferramenta poderosa, ela assume que a variável resposta (o log de uma razão de probabilidades (odds)) é linear nos coeficientes das variáveis preditoras (Delen, 2011).

Essa categoria de regressão pode ter seus coeficientes estimados por dados de treinamento utilizando método de estimação de máxima verossimilhança. Nesse método busca-se valores dos coeficientes para reduzir os erros das probabilidades preditas. Os melhores coeficientes resultam em um modelo cujos valores serão muito próximos de 1 para classe padrão e próximo de 0 para a outra. Após estimados os coeficientes basta aplicar a equação resultante para realizar as predições (Escovedo e Koshiyama, 2021).

2.7 ÁRVORE DE DECISÃO

Árvore de decisão é um dos modelos preditivos mais básicos e tem como inspiração a forma humana de tomada de decisões. As árvores podem ser usadas em problemas de classificação ou de regressão (Escovedo e Koshiyama, 2021).

A figura 2.3 ilustra uma árvore de decisão.

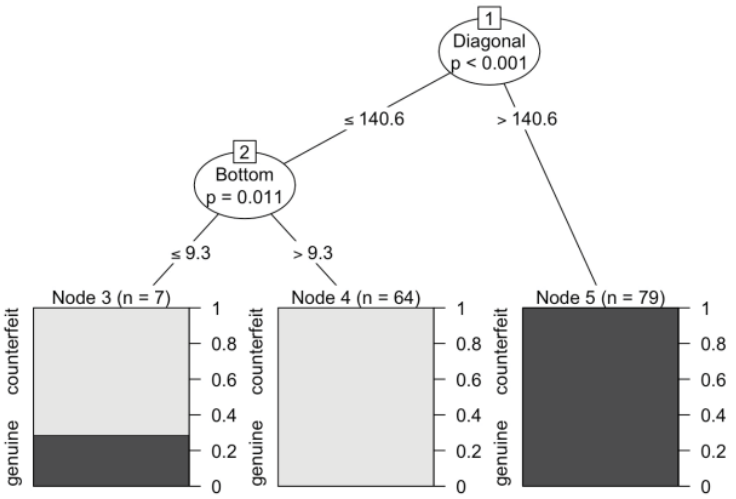


Figura 2.3: Exemplo de Árvore de Decisão. Fonte: (Escovedo e Koshiyama, 2021).

Com a árvore de decisão é possível gerar modelos preditivos e descritivos. Trata-se de uma técnica não paramétrica que classifica uma população em segmentos, como ramos, formando assim uma árvore invertida e consequentemente prevê uma variável de destino. Este método é utilizado principalmente para classificação e tem como vantagens a possibilidade de manipular um grande e complexo conjunto de dados (Pérez et al., 2018).

É um método que associa acurácia e interpretabilidade, apresenta a informação de maneira mais fácil, visualmente falando. Basicamente cada nó interno da árvore representa uma decisão sobre um atributo que no que lhe concerne determina como os dados estão particionados nos nós subsequentes (Escovedo e Koshiyama, 2021).

Um atributo é selecionado para servir de raiz da árvore e os ramos são criados a partir de cada valor do atributo selecionado avaliando uma regra predefinida, uma condição ou operação matemática, as folhas da árvore representam o valor previsto. O trajeto desde a raiz até à folha corresponde a uma regra de classificação (Souza, 2020; Assis, 2017).

São exemplos de algoritmos de árvore de decisão CHAID, CTree, C4.5, CART e Hoeffding Tree. Esses diferentes algoritmos são parecidos entre si e são embasados em uma ação de divisão e conquista, com uma abordagem recursiva. A principal diferença é como as

variáveis são selecionadas e o critério de particionamento e de parada para o crescimento da árvore (Escovedo e Koshiyama, 2021).

2.8 KNN (K-NEAREST NEIGHBOURS)

A técnica kNN vem do inglês *k-Nearest Neighbours*, que significa k vizinhos mais próximos, é um dos algoritmos disponíveis para aprendizagem supervisionada, tanto para problemas de classificação como regressão. É um algoritmo simples de entender e não paramétrico, uma vez que não assume premissas sobre a distribuição dos dados (Escovedo e Koshiyama, 2021).

Segundo Wu et al. (2007) o kNN é um método iterativo simples que detecta um grupo de k objetos utilizados na fase de treinamento e que mais se aproxima a um novo dado que está sendo testado e classifica essa nova instância conforme a classificação que mais ocorre nessa vizinhança.

Em muitos casos é útil considerar mais de um vizinho, de modo que a técnica é mais habitualmente referida como classificação de k-vizinhos mais próximos, em que k-vizinhos mais próximos são usados para determinar a classe (Cunningham e Delany, 2007).

A ideia básica é mostrada na figura 2.4, que descreve um classificador de 3 vizinhos mais próximos em um problema de duas classes em um espaço de recursos bidimensional.

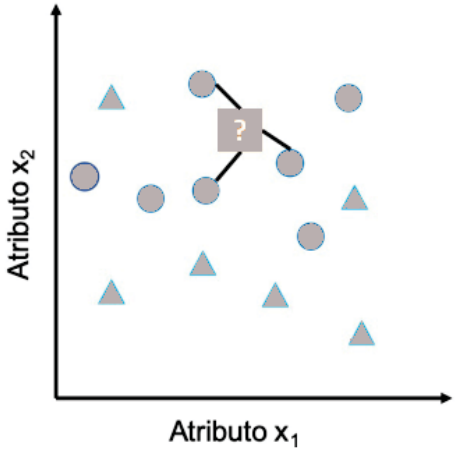


Figura 2.4: Exemplo do funcionamento do KNN. Fonte: (Escovedo e Koshiyama, 2021).

Observando a figura 2.4, quando um novo registro deve ser classificado, ele é comparado a todos os registros do conjunto de treinamento, já armazenados inicialmente, de acordo com alguma métrica de distância, a decisão neste exemplo é direta, todos os vizinhos mais próximos são da mesma classe.

2.9 SVM (SUPPORT VECTOR MACHINE)

Support Vector Machine (SVM), ou máquina de vetor de suporte, é um algoritmo de classificação. O seu funcionamento consiste em encontrar a melhor fronteira que separa as classes. Para isso, utiliza os vetores de suporte, os exemplos mais próximos à borda de separação (Pinheiro et al., 2018).

O SVM é um dos algoritmos mais robusto e preciso entre todos os métodos conhecidos e tem como vantagens: o bom fundamento teórico, necessidade de poucos exemplos para a fase de treinamento e é insensível ao número de atributos (Wu et al., 2007).

Embora o treinamento dos modelos de SVM costume ser lento trata-se de um dos algoritmos mais eficientes, exige poucos ajustes, tende a apresentar boa acurácia e conseguem modelar fronteiras de decisão complexas e não lineares (Escovedo e Koshiyama, 2021).

O funcionamento do SVM ocorre por uma função matemática criada na fase de treinamento capaz de distinguir as possíveis classes. Em uma tarefa de aprendizado de duas classes para um conjunto de dados linearmente separável, uma função de classificação linear corresponde a um hiperplano de separação f(x) que passa pelo meio das duas classes (Wu et al., 2007).

Como existem muitos desses hiperplanos lineares, o que o SVM garante adicionalmente é que a melhor função é encontrada maximizando a margem entre às duas classes, que corresponde à distância mais curta entre os pontos de dados mais próximos a um ponto no hiperplano (Wu et al., 2007).

2.10 SVM (SUPPORT VECTOR MACHINE)

O Random Forest baseia-se em duas técnicas: bagging ou boostrap aggregting e árvore de decisão. Este algoritmo é utilizado principalmente em problemas de classificação e regressão. Esse algoritmo compõe-se de diversas árvores de decisão não correlatas que moldarão o modelo final. Trata-se de um tratamento ensemble, isto é, mesclar diferentes métodos para a construção de um modelo final, que por sua vez será um algoritmo mais complexo e robusto, que requer um tempo computacional maior e tem melhores resultados. No random forest, inicialmente, deve-se criar amostras aleatórias do banco de dados, com o método de reamostragem bootstrap, para a seleção das amostras com reposição dos elementos e, baseado nesse princípio, para cada amostra criada é modelada uma árvore de decisão. Neste algoritmo, para divisão de um nó, as variáveis são selecionadas aleatoriamente e testadas para apurar qual será mais adequada para a separação do nó.

2.11 RANDOM FOREST

O *random forest* baseia-se em duas técnicas: *bagging* ou *boostrap aggregting* e árvore de decisão. Este algoritmo é utilizado principalmente em problemas de classificação e regressão (Silva, 2022).

Esse algoritmo compõe-se de diversas árvores de decisão não correlatas que moldarão o modelo final. Trata-se de um tratamento *ensemble*, isto é, mesclar diferentes métodos para a construção de um modelo final, que por sua vez será um algoritmo mais complexo e robusto, que requer um tempo computacional maior e tem melhores resultados (Teodoro e Kappel, 2020).

No *random forest*, inicialmente, deve-se criar amostras aleatórias do banco de dados, com o método de reamostragem *bootstrap*, para a seleção das amostras com reposição dos elementos e, baseado nesse princípio, para cada amostra criada é modelada uma árvore de decisão (Teodoro e Kappel, 2020).

Neste algoritmo, para divisão de um nó, as variáveis são selecionadas aleatoriamente e testadas para apurar qual será mais adequada para a separação do nó (Silva, 2022).

2.12 MÉTRICAS DE AVALIAÇÃO

2.12.1 Matriz de confusão

A matriz de confusão é um teste para comparar qual classificação o algoritmo previu com a classificação real da instância, fornece um detalhamento do desempenho do modelo. Indica

nas colunas as classes reais e nas linhas são identificadas as classes preditas (Costa, 2021). Os termos utilizados na composição de uma matriz de confusão são:

- Verdadeiro Positivo (VP): número de exemplos positivos classificados corretamente;
- Falso Negativo (FN): número de exemplos negativos classificados incorretamente;
- Falso Positivo (FP): número de exemplos positivos classificados incorretamente;
- Verdadeiro Negativo (VN): número de exemplos negativos classificados corretamente.

2.12.2 Acurácia

A acurácia representa a taxa de acertos em relação ao total de amostras (Costa, 2021). A Equação 2.3 mostra como calcular a acurácia:

$$Acuracia = \left(\frac{VP + VN}{VP + VN + FP + FN} \right) \tag{2.3}$$

Quanto maior a proporção de erros (ou seja, quanto maior FP e FN), menor será o valor da Acurácia. Uma acurácia alta significa mais acertos por parte do modelo.

2.12.3 Precisão

A precisão representa a taxa de objetos positivos classificados corretamente, podendo ser um falso positivo ou um verdadeiro positivo. Mostra entre aqueles sendo classificados como de uma determinada classe, quantos realmente são (Costa, 2021). A Equação 2.4 mostra como calcular a precisão:

$$Precisao = \left(\frac{VP}{VP + FP} \right) \tag{2.4}$$

2.12.4 Sensibilidade

A sensibilidade, ou *recall*, representa a taxa de verdadeiros positivos corretamente classificados, é a capacidade de se identificar corretamente os indivíduos que apresentam a característica de interesse (Escovedo e Koshiyama, 2021). A Equação 2.5 mostra como calcular:

$$Sensibilidade = \left(\frac{VP}{VP + FN} \right) \tag{2.5}$$

2.12.5 F1-Score

É a média harmônica entre as medidas de precisão e sensibilidade (Costa, 2021). A Equação 2.6 mostra como calcular:

$$F = 2 * \left(\frac{Precisao * Sensibilidade}{Precisao + Sensibilidade} \right) \tag{2.6}$$

Em classificação binária essa métrica oferece uma boa capacidade de avaliação. Possibilita que seja realizada a análise de todo o quadro com a mesma importância e não apenas de um de seus aspectos.

2.12.6 Especificidade

A especificidade representa a taxa de falsos positivos corretamente classificados, é a capacidade em identificar corretamente os indivíduos que não apresentam a condição de interesse (Escovedo e Koshiyama, 2021). A Equação 2.7 mostra como calcular:

$$Especificidade = \left(\frac{VN}{VN + FP} \right) \tag{2.7}$$

2.12.7 Curva ROC e AUC

A curva ROC (*Receiver Operating Characteristic*) é um gráfico que mostra o desempenho dos modelos, representamos a sensibilidade no eixo y e 1 - especificidade no eixo x. Contrasta com os benefícios da classificação correta e o custo da classificação incorreta. Quanto mais a curva estiver próxima do canto superior esquerdo, melhor o classificador (Escovedo e Koshiyama, 2021).

Para resumir a qualidade mensurada pela curva, é comum a utilização da métrica AUC (*Area Under the Curve*), um algoritmo que calcula a área sob a curva ROC, podendo assim comparar os classificadores utilizando um único escalar, a área abaixo da curva ROC. Quanto mais alto é o valor do AUC melhor é a capacidade do modelo de classificação em distinguir entre as classes positivas e negativas (Teodoro e Kappel, 2020).

3 ESTADO DA ARTE

Este capítulo apresenta os trabalhos relacionados a esta pesquisa, que classificam estudantes como em risco de evasão em universidades utilizando mineração de dados e aprendizagem de máquina.

Em 2009, Dekker et al., desenvolveu um estudo experimental na busca de um modelo preditivo para evasão no curso de Engenharia Elétrica da Universidade de Tecnologia de Eindhoven, na Holanda. A taxa de evasão após o primeiro ano era em torno de 40%. Neste trabalho, o aluno evasor era aquele que após 3 anos da data de seu ingresso não havia conseguido concluir com sucesso as matérias do primeiro ano. Para isso, os autores analisaram dados de antes do ingresso na universidade e dados de desempenho durante a universidade, e treinaram diferentes algoritmos: regressão logística, OneR, algoritmos de árvores de decisão (CART, C4.5), classificador bayesiano, algoritmo baseado em aprendizagem de regras e florestas aleatórias. Os algoritmos que obtiveram maior acurácia foram os de árvores de decisão (CART, C4.5 e Florestas Aleatórias) e o algoritmo de regressão logística, com valores em torno de 80%.

Delen (2011) utilizou dados com informações financeiras, acadêmicas e demográficas de mais de 25 mil estudantes de uma universidade pública dos Estados Unidos. Destes 25 mil, 19 mil retornavam após o primeiro ano, gerando uma taxa de evasão em torno de 21% após o primeiro ano de graduação. A partir disso, o autor utilizou histórico do ensino médio e do primeiro ano de curso na universidade para treinar classificadores e obteve como resultado acurácia de 81% em redes neurais, 78% com a árvore de decisão e 74% com a regressão logística. No trabalho também é constatado que os fatores relacionados para a classificação da evasão são, principalmente, o desempenho acadêmico deste aluno, tanto no presente quanto no passado.

Em seu trabalho, Aulck et al. (2019), destaca que a taxa de evasão nas instituições de bacharelado nos Estados Unidos, após o primeiro ano, é de 30%. Para seu trabalho, os autores analisaram dados de mais de 66 mil estudantes da Universidade de Washington, desde o desempenho escolar até características demográficas através de regressão logística, florestas aleatórias, máquina de vetores de suporte (SVM), árvores com gradiente de reforço e *k-Nearest Neighbors* (KNN) e assim classificar os alunos como evadidos e concluintes. Para este trabalho a definição aceita para aluno evasor é a que determina que o aluno é evasor se em 6 anos após a data de ingresso ele não concluiu sua graduação. O classificador com maior acurácia foi o de regressão logística, com 83.2% de acurácia.

Maksimova et al. (2021) fez um estudo de caso de aprendizado de máquina educacional objetivando prever a taxa de evasão dos alunos do primeiro ano de ciência da computação na Virumaa College of Tallinn University of Technology além de definir os fatores que influenciam essa taxa de evasão. Para isso foram utilizados dois conjuntos de dados: histórico dos alunos e dados do sistema de informação TalTech. Os conjuntos foram submetidos às técnicas de redes neurais, árvore de decisão, SVM, regressão logística e *naive bayes*. A partir dos dados da pré matrícula dos alunos a acurácia das técnicas atingiu em torno de 70%. Quando acrescentados os dados dos alunos após o primeiro semestre de estudos a acurácia chega a 90%.

Hoed (2016) analisou a evasão nas graduações dos cursos relacionados a computação a partir de dados do INEP e dados fornecidos pela UnB (Universidade de Brasília) através de estudos quantitativos em que se aplicou análise de sobrevivência e mineração de regras de associação via algoritmo Apriori e questionários aplicados aos alunos evadidos para identificar as causas de evasão.

Como resultados do trabalho, Hoed (2016) obteve indícios que a relação candidatos/vagas é inversamente proporcional à evasão, que cursos da área de ciências exatas possuem maiores valores de alunos evadidos e que o sexo, a forma de ingresso e se o aluno é cotista ou não também interferem de forma direta na taxa de evasão.

Assis (2017) utilizou a mineração de dados para traçar o perfil dos alunos evasores relacionando os dados de curso área de estudo e instituição. O autor combinou informações do censo da educação superior e do ENEM para criar 5 modelos de algoritmos classificatórios: naive bayes, redes neurais, regressão logística e dois algoritmos de árvores de decisão (CART e C5.0).

No trabalho de Assis (2017) o CART foi o algoritmo que se sobressaiu quando comparado aos outros classificadores, ao atingir a sensibilidade, neste caso classificar os alunos evadidos, de 84%, no restante dos testes os resultados não foram estatisticamente significativos quando comparados os algoritmos. O autor identificou como principais características de propensão à evasão as seguintes:

- aluno ingressar no primeiro semestre;
- aluno possuir vínculo com mais de uma instituição de ensino;
- alunos que obtiveram notas acima da média no ENEM;
- já ter concluído o ensino médio antes de realizar o ENEM.

Além disso, o autor desenvolveu um pacote para o R que possibilita treinar novos classificadores de evasão que permite ser utilizado em qualquer instituição de ensino superior a fim de determinar quais os alunos com uma maior probabilidade de evadir.

Pinheiro et al. (2018) através das técnicas de AM, *Naive Bayes* (NB), *Support Vector Machines* (SVMs) e Árvores de Decisão (AD), procuraram identificar precocemente quais alunos são mais propensos a evadir. Os dados utilizados foram retirados do Sistema Acadêmico do Instituto Federal de Educação Ciência e Tecnologia do Maranhão, que contém informações socioeconômicas e acadêmicas de seus alunos.

Os atributos utilizados por Pinheiro et al. (2018), são cinco referentes ao momento da matrícula: estado civil, forma de ingresso, turno do curso, sigla do curso e sexo, além de outras 12 referentes a vida acadêmica do aluno como, por exemplo, coeficiente de rendimento, média de reprovações por falta e nota, média das notas e presença.

O trabalho foi dividido em 3 partes, em uma foi feito o processamento utilizando apenas os dados do momento da matrícula, em outra com os dados da vida acadêmica e a terceira parte considerando tanto os atributos da matrícula quanto os acadêmicos. Os resultados obtidos mostram que os algoritmos da AM conseguem prever de forma eficaz os casos de evasão, a classificação teve mais relevância com informações acadêmicas mas em geral quando os dados foram combinados com os da matrícula foi obtido um melhor desempenho (Pinheiro et al., 2018).

Melo (2019), utilizou a mineração de dados através de redes neurais artificiais para prever a possibilidade de um aluno abandonar o ensino superior. Com os dados do Sistema Acadêmico da Universidade Federal do Triângulo Mineiro a ferramenta criada foi capaz de identificar 63,8% dos evadidos, e conseguiu identificar alunos predispostos a evadir com 36 dias de antecedência do fato ocorrer, em média, possibilitando assim a instituição atuar de forma preventiva.

Na última década o ensino superior do Brasil teve sua taxa de evasão acima de 20%, e os problemas decorrentes da evasão não atingem apenas as instituições, mas também os alunos. Mesmo sabendo da taxa de evasão, a coordenação das instituições apresenta dificuldades em

prever quais serão os evasores. Moreira (2020), analisou a trajetória dos alunos do curso de ciência da computação da UFPR durante o período de 2011 a 2019, para isso, utilizou 2 técnicas de aprendizado de máquina, árvore de decisão e regressão logística, com a finalidade de prever a evasão. Com os dados do primeiro ano foi possível prever com uma acurácia de 74% e sensibilidade de 85% utilizando a árvore de decisão e acurácia de 75% e sensibilidade de 85% com a regressão logística.

Teodoro e Kappel (2020) , destacam os preocupantes dados de evasão no ensino superior e ressaltam a importância de identificar perfis com maior probabilidade de evasão para que as instituições então tracem planos em busca de redução nessa taxa.

Considerando essa problemática os autores (Teodoro e Kappel, 2020), visaram identificar os padrões característicos e os atributos mais determinantes nestes padrões dos alunos com um maior potencial de desvincular-se das instituições de ensino superior aplicando cinco técnicas de aprendizado de máquina (*Naive Bayes, K-Nearest Neighbors*, Árvores de Decisão, *Random Forest* e Redes Neurais) nos dados do INEP.

A técnica com melhor resultado foi a *Random Forest* com uma taxa de acerto de cerca de 80% além de fornecer dados que tornam possível a geração de um relatório sobre quais atributos foram os mais importantes para as suas previsões. Como características determinantes, Teodoro e Kappel (2020) citam: idade, participação em atividades extracurriculares e a carga horária do curso em que o aluno está inserido.

No trabalho de Costa (2021) os três primeiros semestres cursados foram objeto de estudo para criação de modelos de predição do risco de um aluno evadir-se do curso de ciência da computação e engenharias da Universidade Federal de Pelotas (UFPel). Para isso foi utilizado conjunto de dados com um total de 22 atributos. Os dados foram extraídos do próprio sistema acadêmico da universidade e aplicados em cinco algoritmos. Para os dados do curso da ciência da computação o algoritmo com maior precisão foi o modelo de regressão logística (90,16%) e para os dados das doze engenharias o modelo de floresta aleatória foi o mais preciso com (83,4%).

4 METODOLOGIA

Este capítulo apresenta os procedimentos metodológicos da pesquisa com o intuito de atingir os objetivos geral e específicos propostos.

A figura 4.1 é um esquema com todas as etapas do processo realizado que se inicia com a caracterização de um problema ou de uma ideia seguida da aplicação das técnicas de mineração e avaliação dos resultados.

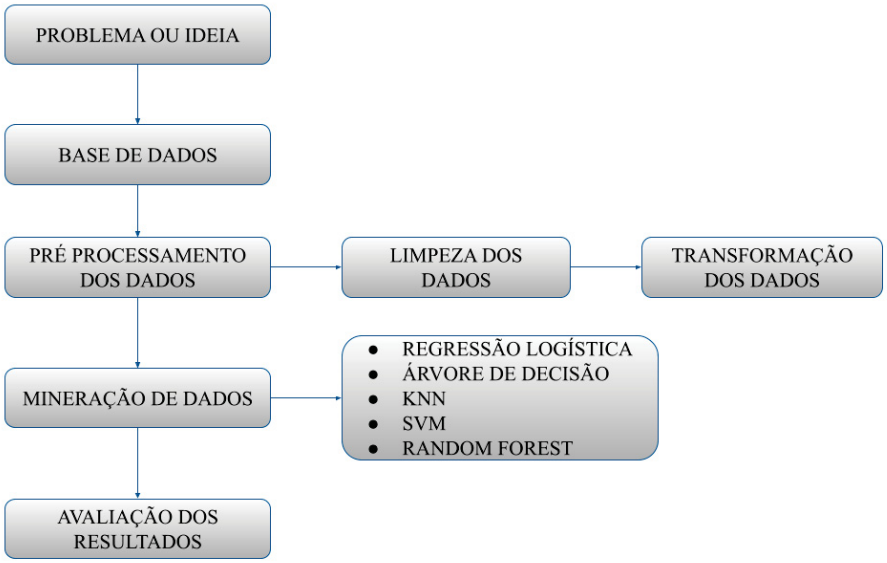


Figura 4.1: Etapas do Processo.

A primeira etapa de um projeto pode ser uma problemática descoberta, neste trabalho trata-se da evasão no ensino superior. Com o objetivo bem definido, obtém-se o conjunto de dados, que, a partir dele, podem ser realizadas operações de extração e pré-processamento dos dados: fase em que é realizada a limpeza e transformação dos dados com a finalidade de prepará-los para os modelos que serão construídos (Corrêa, 2019; Escovedo e Koshiyama, 2021).

Com os dados selecionados, limpos e transformados é realizada então a mineração de dados, onde são aplicadas as técnicas de mineração e aprendizagem de máquina. Nesse trabalho foram empregados a regressão logística, árvore de decisão, kNN, SVM e *random forest*. Como resultado haverá a extração de um modelo representando padrões de relacionamento identificados no conjunto de dados (Corrêa, 2019; Escovedo e Koshiyama, 2021).

Na etapa de avaliação dos resultados é realizado a interpretação do modelo extraído como, por exemplo, se o modelo de classificação gerado consegue prever com elevada acurácia a probabilidade de um estudante evadir da instituição ou analisar se ainda é preciso melhorar de alguma forma (Corrêa, 2019; Escovedo e Koshiyama, 2021).

4.1 ENTENDIMENTO DOS DADOS

Os dados utilizados são da base de dados do SIGA, esses dados tem um volume crescente e incluem: matrícula, cursos, professores, disciplinas, frequências, notas, situação do aluno no curso, caso tenha evadido qual forma de evasão entre outras informações e possibilita realizar

diversas investigações em busca de padrões e informações relacionadas com a condição do discente.

As informações foram coletadas através de uma consulta em SQL (*Structured Query Language*) diretamente da base de dados em julho de 2021 e foi realizado o levantamento de algumas informações para atualização em dezembro de 2021. Foram levantadas 67 variáveis, as informações sensíveis não foram selecionadas impossibilitando assim a identificação de qualquer aluno. A principal variável de interesse neste trabalho é a que indica a situação do aluno no curso.

Os dados analisados neste estudo são referentes a 36.401 registros, sendo que cada registro é a última informação de cada aluno em cada curso e estão distribuídos conforme figura 4.2, dessa maneira cada estudante possui no máximo um vínculo com cada curso.

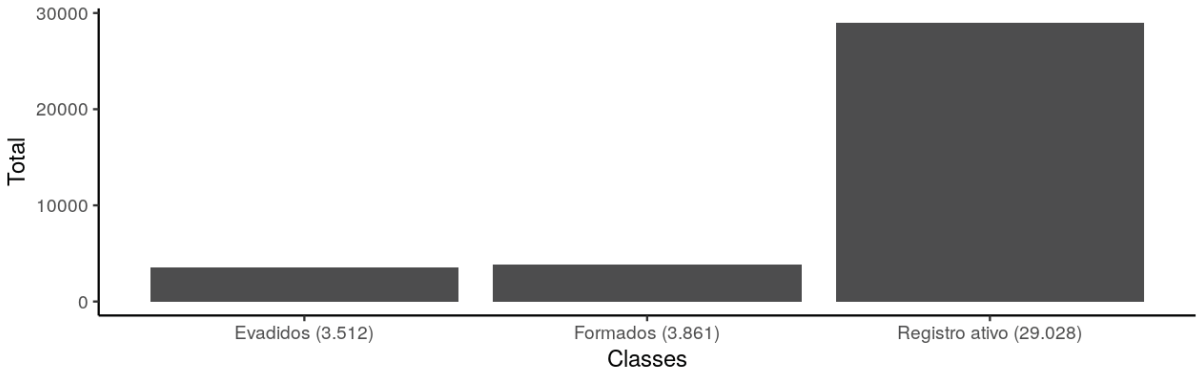


Figura 4.2: Distribuição dos Alunos por Situação.

4.2 PRÉ-PROCESSAMENTO

Primeiramente, na variável que indica qual a forma de evasão foi identificado classificações de alunos que fugiam do objetivo da pesquisa ou que teriam pouca interferência no resultado, como, por exemplo, alunos com as seguintes formas de evasão: “Desistência Vestibular/PROVAR”, “Falecimento”, “Mudança de Campus/Habilitação interna”, “Não Confirmação de Vaga”, “Término de intercâmbio” e para a maioria dos que estavam como “Término de Registro Temporário”, restando dados de alunos que estavam com a forma de evasão conforme a tabela 4.1.

Os registros que se encontravam como “Término de Registro Temporário” precisaram ser melhor analisados, pois em alguns cursos como, por exemplo, estatística os alunos inscritos no Processo Seletivo Estendido (PSE) entram com um GRR (código de identificação do aluno) temporário e aqueles que permanecem recebem outro permanente e com isso um mesmo aluno acabou aparecendo em duas observações, uma com situação evadido e outra como aluno ativo, assim a observação com situação evadido foi retirada.

Foi realizada uma etapa de seleção de variáveis, em que foram removidas aquelas que não fazem sentido para o contexto, que apresentavam muitos valores nulos ou inválidos ou apresentaram baixíssima variabilidade entre os valores.

Para corrigir problemas relacionados a dados faltantes ou *outliers* as variáveis foram categorizadas tendo como base o cálculo do InfoValue (IV) estatística que permite categorizar e avaliar o poder de discriminação de uma variável, de modo que cada categoria de cada variável tenha no mínimo 5% de representatividade em relação ao todo. A exemplo temos a categorização da variável raça apresentada na tabela 4.2.

Tabela 4.1: Formas de evasão

Formas de evasão	**Frequência absoluta**
Abandono	1148
Cancelamento a Pedido do Calouro	423
Cancelamento Administrativo	10
Cancelamento Judicial	19
Cancelamento Pedido	829
Decisão Administrativa	2
Desistência	1
Formatura	3861
Jubilamento	2
Novo Vestibular	140
Reopção	417
Término de Registro Temporário	36
Transferência	2
Transferência Externa	12

Tabela 4.2: Categorização da Variável Raça

	Branca	**Não Informada**	**Outras**	**Total**
Não Evento	2.527	675	659	3.861
Evento	998	1.710	332	3.040
%Pop	51%	34,6%	14,4%	100%
% de Não Eventos	65,4%	17,5%	17,1%	–
% de Eventos	32,83%	56,25%	10,92%	–

Outro ponto foi a criação de uma nova variável para saber a idade do aluno ao ingressar no curso através da data de nascimento do aluno e da data de entrada no curso.

Os dados dos alunos que ainda estão com o registro ativo foram apartados da modelagem, desta forma, após o tratamento do conjunto de dados, para modelagem foram consideradas 15 variáveis conforme a tabela 4.3, referente aos 6.901 alunos sendo 3861 formados e 3040 evadidos.

Tabela 4.3: Variáveis Testadas no Estudo

Variável	**Descrição**
status	Situação do aluno no curso.
sexo	Sexo do aluno.
racaCor	Raça/Cor do aluno.
cota	Descrição do tipo de cota do aluno.
turno	Turno do curso.
periodoIngresso	Período de Ingresso.
formaIngresso	Forma de Ingresso.
setor	Setor que o aluno pertence.
idadeIngresso	Idade de Ingresso no Curso.
paisNascimento	País de Nascimento.
estadoNascimento	Estado de Nascimento.
ira	Índice de Rendimento Acadêmico no 1° semestre.
reprovacoesNota	Quantidade de Reprovações por Nota no 1° semestre.
reprovacoesFrequencia	Quantidade de Reprovações por Frequência no 1° semestre.
chCurriculo	Carga Curricular cursada no 1° semestre.

4.3 ANÁLISE EXPLORATÓRIA DOS DADOS

Antes de proceder com a modelagem preditiva propriamente dita, realizou-se uma análise descritiva para entender melhor o comportamento da variável de interesse “Evasão” em relação às demais variáveis explicativas.

4.3.1 Quantidade de Reprovações por Nota dos Alunos (1° semestre) por Situação

A partir da tabela 4.4 observa-se que o número médio de reprovações por nota dos alunos evadidos é um mais que o dobro em relação à média dos formados. Já a variabilidade é maior entre os formandos frente para os evadidos.

Tabela 4.4: Estatísticas Descritivas da Quantidade de Reprovações por Nota dos Alunos por Situação.

Estatísticas	**Evadidos**	**Formados**
Mínimo	0,00	0,00
Média	0,90	0,45
Máximo	9,00	9,00
Desvio Padrão	1,45	0,96
Coeficiente de Variação	1,61	2,13

4.3.2 Quantidade de Reprovações por Frequência dos Alunos (1° semestre) por Situação

Observa-se a partir da tabela 4.5 que o número médio de reprovações por frequência dos alunos evadidos é bem maior em relação à média dos formados. Já a variabilidade é maior entre os formandos frente para os evadidos.

Tabela 4.5: Estatísticas Descritivas da Quantidade de Reprovações por Frequência dos Alunos por Situação.

Estatísticas	Evadidos	Formados
Mínimo	0,00	0,00
Média	0,55	0,11
Máximo	9,00	7,00
Desvio Padrão	1,33	0,48
Coeficiente de Variação	2,42	4,4917

4.3.3 Índice de Rendimento Acadêmico dos Alunos (1° semestre) por Situação

Observa-se a partir da tabela 4.6 que a média do Índice de Rendimento Acadêmico dos alunos evadidos é inferior à média dos alunos formados. Em relação à variabilidade os alunos evadidos apresentam maior variação em comparação aos alunos formados.

Tabela 4.6: Estatísticas Descritivas do Índice de Rendimento Acadêmico dos Alunos por Situação.

Estatísticas	Evadidos	Formados
Mínimo	0,00	0,00
Média	0,46	0,75
Máximo	1,00	1,00
Desvio Padrão	0,33	0,14
Coeficiente de Variação	0,71	0,19

4.3.4 Carga Horária Cursada (1° semestre) por Situação

Observa-se a partir da tabela 4.7 que a média da carga horária cursada pelos alunos evadidos é inferior à média dos alunos formados. Em relação à variabilidade os alunos evadidos apresentam maior variação em comparação aos alunos formados.

Tabela 4.7: Estatísticas Descritivas da Carga Horária Cursada dos Alunos por Situação.

Estatísticas	Evadidos	Formados
Mínimo	0,00	0,00
Média	257,85	399,59
Máximo	945	468
Desvio Padrão	179,66	960
Coeficiente de Variação	0,70	0,38

4.3.5 Sexo dos Alunos por Situação

A partir da figura 4.3 observa-se que alunos do sexo masculino apresentam taxa de evasão 47,6% relativamente maior que a dos alunos do sexo feminino onde o percentual de evadidos é de 42,2%.

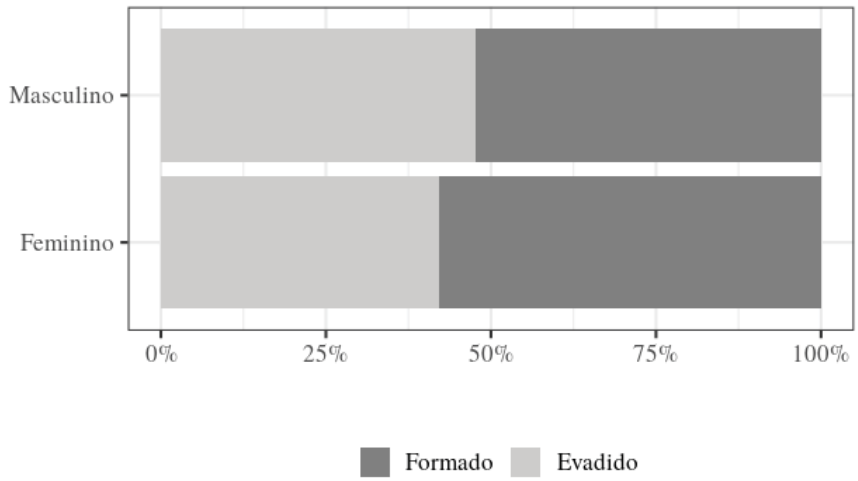


Figura 4.3: Distribuição do Sexo dos Alunos por Situação.

4.3.6 Raça/Cor dos Alunos por Situação

Observa-se a partir da figura 4.4 que alunos onde a raça não foi informada apresenta maior número de evasão 76,9%, a classe outras, sendo a categorização da cor/raça amarela, indígena, parda e preta, apresenta um percentual de evasão de aproximadamente 33,5%, já na classe de alunos autodeclarados brancos apresentou o menor percentual de evadidos 27,4% (998).

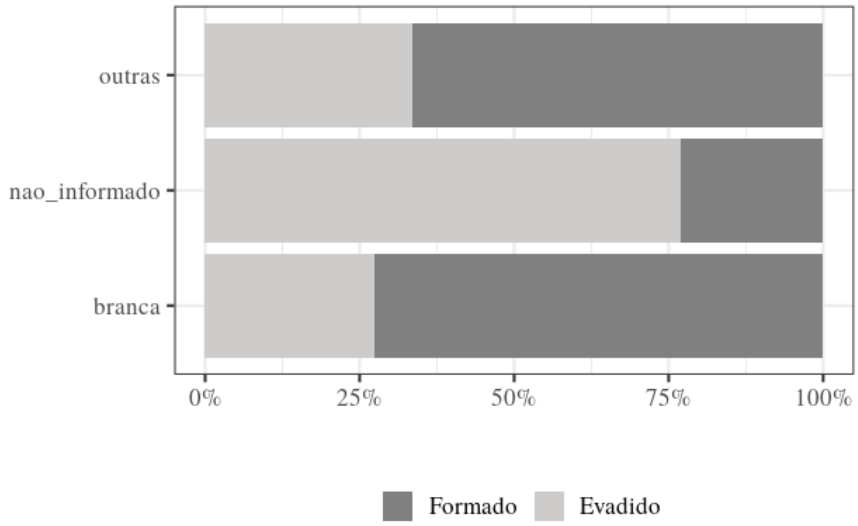


Figura 4.4: Distribuição da Cor/Raça dos Alunos por Situação.

4.3.7 Cotas dos Alunos por Situação

A partir da figura 4.5 observa-se que o maior percentual de alunos evadidos se encontra na classe categorizada que incluem a cota de Portador de Deficiência e a cota Racial 63,9%, seguido pela classe Ampla concorrência este percentual cai para 44,5%, já alunos inclusos nas cotas de Escola Pública são os que apresentam o menor percentual de evasão 19,5%.

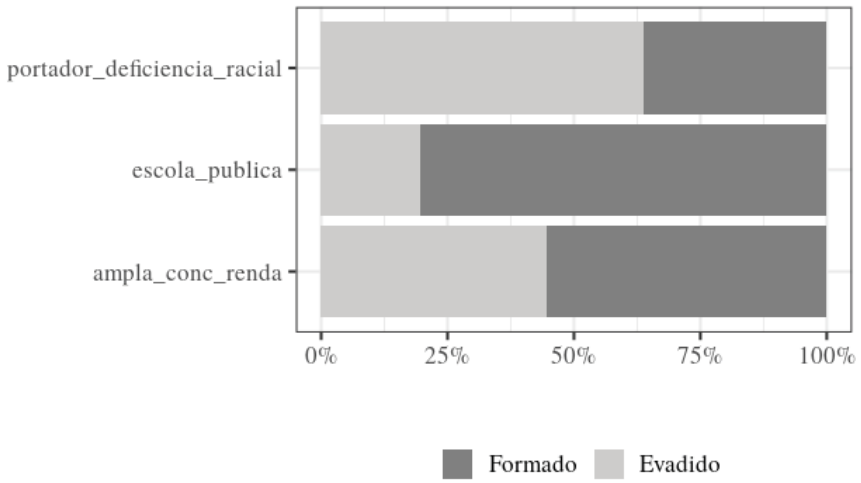


Figura 4.5: Distribuição da Cota dos Alunos por Situação.

4.3.8 Turno dos Alunos por Situação

Observa-se a partir da figura 4.6 que os alunos matriculados nos turnos matutino/NSA Vespertino e noturno apresentam maior percentual de evasão 51% e 46,9% respectivamente, seguido pelo integral 40,6%.

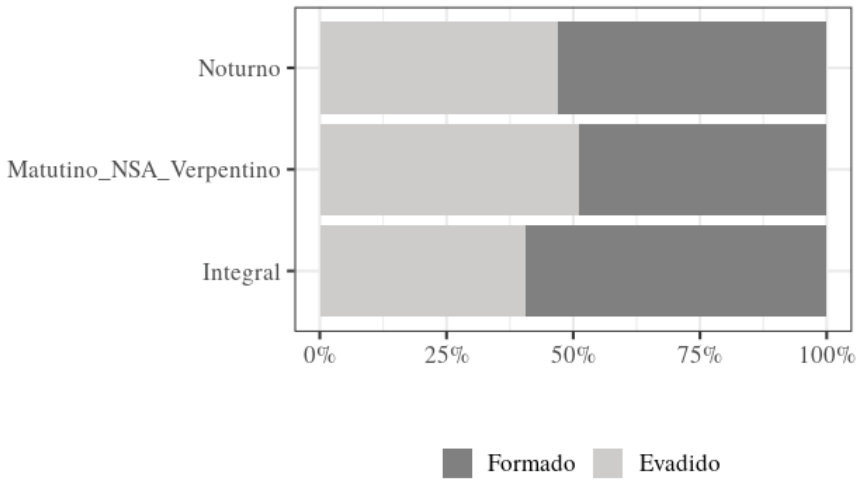


Figura 4.6: Distribuição do Turno dos Alunos por Situação.

4.3.9 Período de Ingresso dos Alunos por Situação

A partir da figura 4.7 observa-se que os alunos que ingressaram no primeiro semestre do ano letivo apresentam maior taxa de evasão 45,8%, seguido pelos alunos ingressantes no segundo semestre e período anual com taxas de 41,4% e 38,6% respectivamente.

4.3.10 Forma de Ingresso dos Alunos por Situação

Observa-se a partir da imagem 4.8 que os alunos que ingressaram por meio do SISU apresentam maior taxa de evasão 57,3% seguido pela classe categorizada outros e alunos ingressantes por vestibular com percentual de evadidos de respectivamente 42,6% e 41,5%.

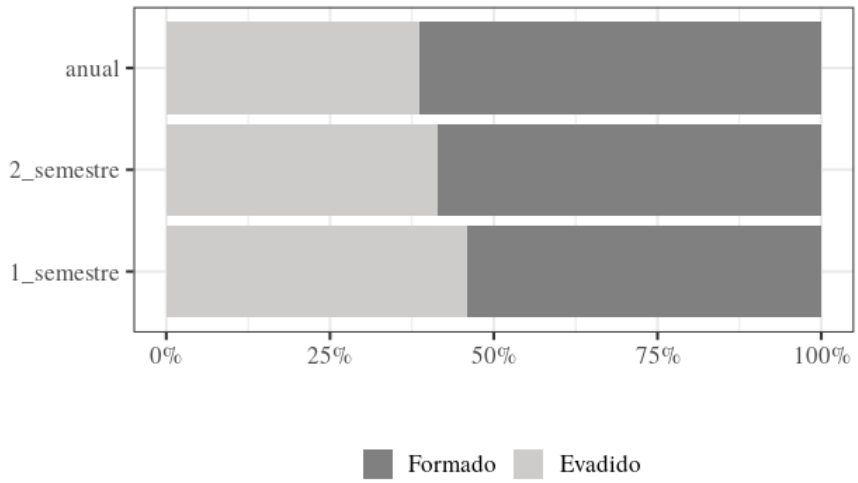


Figura 4.7: Distribuição do Período de Ingresso dos Alunos por Situação.

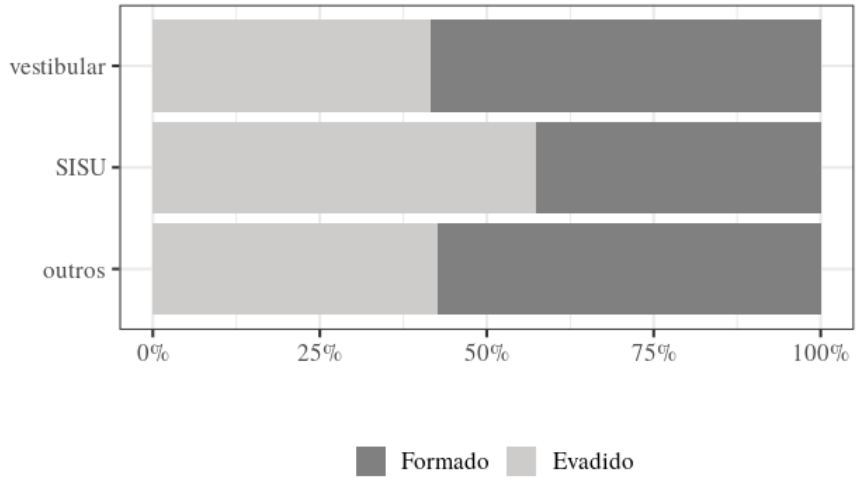


Figura 4.8: Distribuição da Forma de Ingresso dos Alunos por Situação.

4.3.11 Setor dos Alunos por Situação

Observa-se a partir da imagem 4.9 que os setores em que há maior índice de alunos evadidos é o de ciências exatas com uma taxa de 68,8% seguido pelos setores de ciências humanas e agrárias com aproximadamente 55,00% de evasão, já o setor em que há o menor percentual de evadidos é o de ciências da saúde 25,7%.

4.3.12 Idade de Ingresso dos Alunos por Situação

A partir da imagem 4.10 observa-se que alunos que ingressam com idade acima de 30 anos apresentam maior percentual de evadidos 56,3%, seguido pelos alunos com idades entre 22 e 30 anos e de 20 a 21 anos com respectivamente 50,1% e 44,4%, os alunos com até 18 anos são os que apresentam o menor percentual de evasão 39,5%.

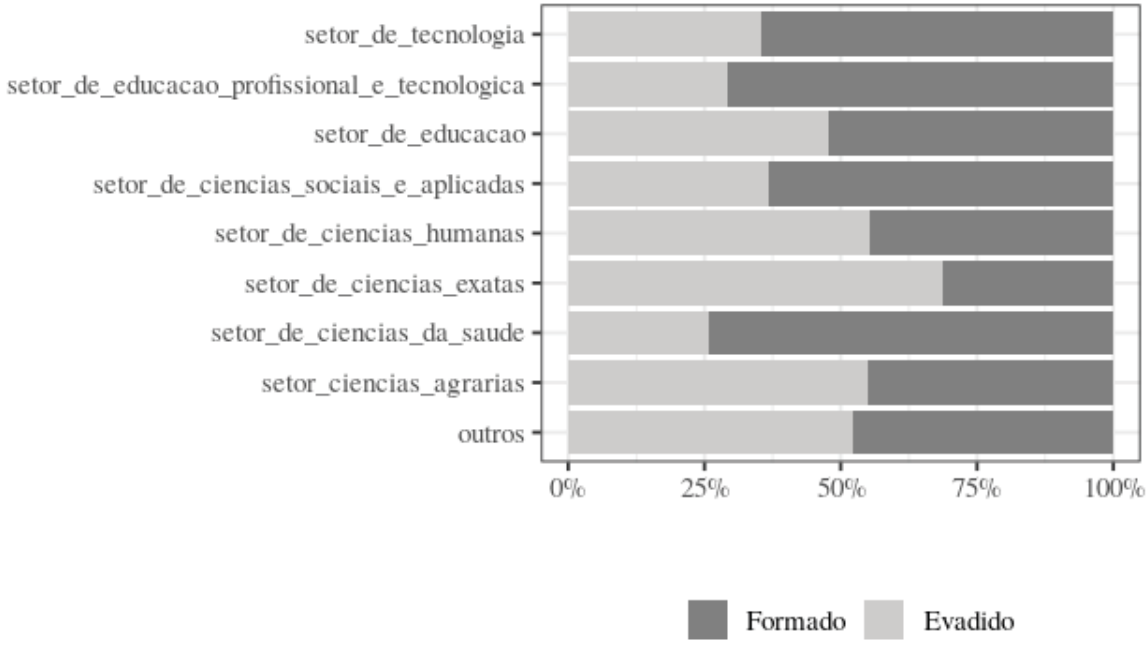


Figura 4.9: Distribuição do Setor dos Alunos por Situação.

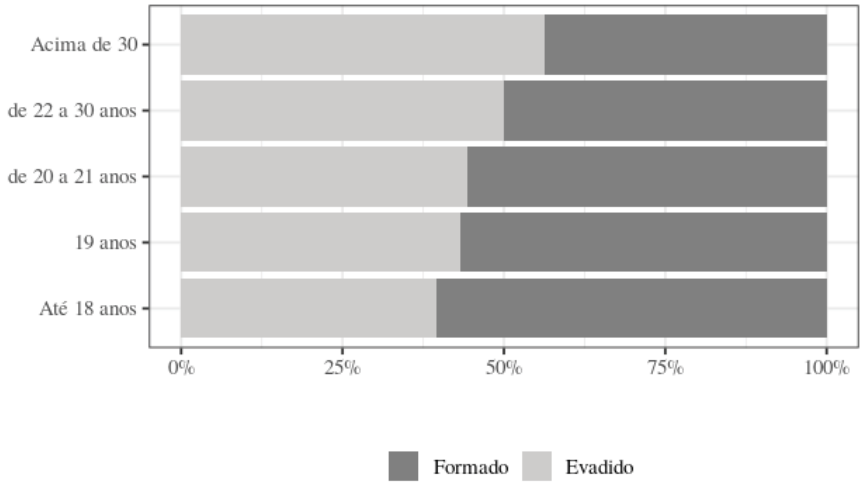


Figura 4.10: Distribuição da Idade de Ingresso de Nascimento dos Alunos por Situação.

4.3.13 País dos Alunos por Situação

Observa-se a partir da imagem 4.11 que alunos onde o país não foi informada/outros apresentam maior número de evasão 69,4% e o Brasil apresentou o menor percentual de evadidos 40,8%.

4.3.14 Estado de Nascimento dos Alunos por Situação

Observa-se a partir da imagem 4.12 que os alunos que não informaram o seu estado de nascimento apresentam maior evasão 69,8% seguido por alunos do Paraná e Santa Catarina com 41,3% cada, já alunos dos estados de São Paulo apresentaram o menor número de evasão 36%.

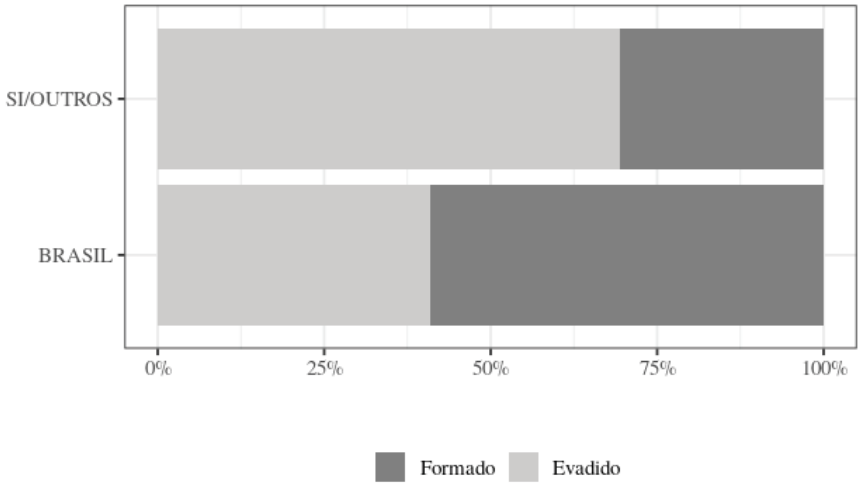


Figura 4.11: Distribuição do País dos Alunos por Situação.

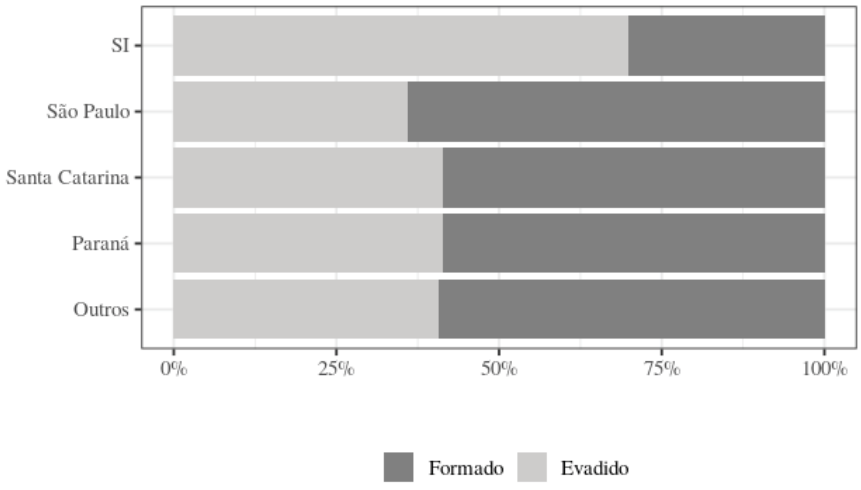


Figura 4.12: Distribuição do Estado de Nascimento dos Alunos por Situação.

5 RESULTADOS

Nesta etapa foram desenvolvidos modelos empíricos, para estimar a propensão a evasão dos alunos utilizando as técnicas de modelagem de regressão logística, árvore de decisão, kNN, SVM e *random forest* indicadas em problemas de classificação. Para tal, foram utilizados como atributos preditores as informações coletadas conforme apresentadas na tabela 4.3.

5.1 ETAPAS DE MODELAGEM

Após a análise exploratória, apresentada na seção anterior, para o ajuste dos modelos, o próximo passo verificou, individualmente, a existência de associação entre as variáveis independentes e a variável dependente. Neste caso, foi utilizado o teste qui-quadrado (x^2) de associação para variáveis qualitativas e para verificar a diferença de médias entre duas amostras independentes, para as variáveis quantitativas, utilizou-se o teste t de Student considerando significativo se valor-p < 0,05.

Com a análise bivariada tem-se uma ideia das principais covariáveis que apresentam diferenças entre os grupos. A partir disso, objetivou-se analisar o efeito conjunto das covariáveis.

No processo de modelagem, após a composição e tratamento do conjunto de dados, o mesmo foi dividido em duas partes, sendo 70% da base para treino e 30% para teste. No conjunto de dados de treino foi realizada a seleção de atributos e ajuste dos parâmetros dos algoritmos (*tunning*). A estratégia definida de treinamento foi com validação cruzada 10-fold e 3 repetições.

Após o processo de treino foi realizado o teste de cada um dos algoritmos selecionados, com ponto de corte de 0.8, assim gerou-se a matriz de confusão.

Testou-se inicialmente o ajuste de uma regressão logística, retirando as variáveis raça/cor, país de nascimento e estado de nascimento pois a categoria que mais influencia é a sem informação, com resultado apresentado na Tabela 5.1 resumidos na saída da ANOVA.

Tabela 5.1: Modelo com todas as variáveis.

	Deviance	**Resid. Df**	**Resid. Dev**	**Pr(>Chi)**
sexo	14.12	4899	6726.7	< 0,001
cota	246.63	4897	6480.1	< 0,001
turno	27.30	4895	6452.8	< 0,001
periodoIngresso	3.72	4893	6449.1	0.1553772
formaingresso	69.92	4891	6379.1	< 0,001
setor	269.59	4883	6109.6	< 0,001
idadeIngresso	54.10	4879	6055.5	< 0,001
ira	1215.57	4878	4839.9	< 0,001
reprovacoesNota	15.74	4875	4576.7	< 0,001
reprovacoesFrequencia	2.09	4876	4592.4	< 0,001
chCurriculo	235.36	4877	4604.5	< 0,001

O método adotado para encontrar a melhor equação de regressão foi o *stepwise* adotando nível de significância de 10% para inclusão de covariáveis no modelo. Assim, observa-se na Tabela 5.2 que as variáveis que se mostraram significativas no modelo final foram: categoria de cota, forma de ingresso, setor de estudo, ira do primeiro semestre, carga curricular cursada no primeiro semestre, número de reprovações por frequência e nota no primeiro semestre.

Tabela 5.2: Modelo com todas as variáveis.

	Coeficiente	**Erro-Padrão**	**Valor-p**
(Intercept)	2.5650884	0.2761282	< 0,001
sexoM	0.1448284	0.0763244	0.057757
cotaEscolaPublica	-1.3479755	0.1562441	< 0,001
cotaDeficienciaRacial	0.7378800	0.1044759	< 0,001
turnoMatutinoNSAVerpentino	0.2632975	0.1211817	0.029799
turnoNoturno	0.1759581	0.1030234	0.087647
periodoIngresso2Semestre	-0.1275878	0.1152353	0.268210
periodoIngressoAnual	0.8822294	0.1953519	< 0,001
formaIngressoSISU	0.7621937	0.1416800	< 0,001
formaIngressoVestibular	0.5602150	0.1206330	< 0,001
setorOutros	0.5686255	0.1571835	< 0,001
setorAgrarias	0.2747818	0.1969959	0.163058
setorExatas	0.6848994	0.1964334	< 0,001
setorHumanas	0.3387230	0.1952018	0.082697
setorSociaisAplicadas	-0.1619223	0.2105076	0.441775
setorEducacao	0.7383781	0.3204544	0.021214
setorSept	-0.7605499	0.2076707	< 0,001
setorTecnologia	-0.1677885	0.1629601	0.303184
idadeIngressoAcimaDe30	0.3139929	0.1578034	0.046616
idadeIngressoAte18Anos	-0.1352145	0.1042440	0.194598
idadeIngressoDe20a21Anos	0.0143393	0.1166992	0.902207
idadeIngressoDe22a30Anos	0.1263414	0.1155254	0.274120
ira	-3.4675850	0.2651378	< 0,001
reprovacoesNota	0.1561075	0.0393783	< 0,001
reprovacoesFrequencia	0.2356747	0.0575543	< 0,001
chCurriculo	-0.0048522	0.0003225	< 0,001

Para verificar se os resíduos *deviance* se parecem com uma distribuição normal para a distribuição da variável resposta foi realizado o diagnóstico por meio do gráfico de quantis, com envelope simulado, onde os resíduos *deviance* são plotados contra os quantis da distribuição Binomial (PETTERLE et al., 2017). Na Figura 5.1 verifica-se que não há ocorrência de afastamentos da variável resposta para a distribuição binomial, visto que os resíduos *deviance* encontram-se no envelope simulado.

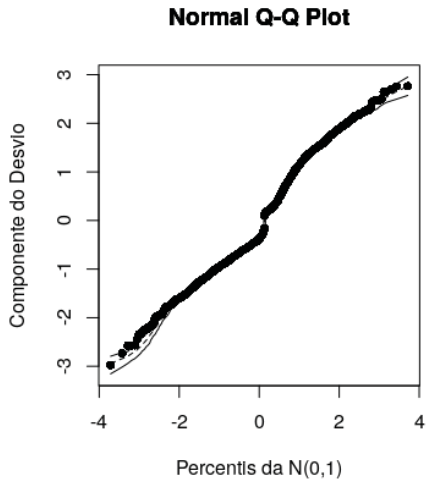


Figura 5.1: Quantis da Distribuição Normal Padrão.

Para realizar o teste de bondade de ajuste foi utilizado o Hosmer-Lemeshow, que deixa as observações em ordem crescente em termos das probabilidades preditas para o evento de interesse (Y = 1) e depois as divide em g grupos com tamanhos aproximadamente iguais. A estatística do teste segue cerca de distribuição qui-quadrado com (g-2) graus de liberdade. A hipótese nula associada ao teste, assume que o modelo apresenta um bom ajuste, diferentemente da hipótese alternativa que considera o ajuste insatisfatório (PETTERLE et al., 2017).

O resultado do teste de bondade foi X^2=11.744 e p-value=0.163 com g = 10 indicando a não rejeição da hipótese nula. Logo, pode-se concluir que o modelo apresenta ajuste satisfatório.

As estimativas do coeficiente apresentadas na tabela 5.2 diz se a presença de determinada característica aumenta (se positiva) ou diminui (se negativa) a probabilidade de um de o aluno evadir ou não.

5.2 COMPARAÇÃO DE MODELOS

Para efeito de validação, procedeu-se com a modelagem alternativa utilizando mais 5 modelos de classificação, sendo: árvore de classificação (CART, CTree), kNN, SVM e *random forest*. A utilização de diferentes técnicas de classificação teve como objetivo alcançar o melhor resultado possível para a previsão de evasão.

Da figura 5.2 até 5.7 apresentam-se as matrizes de confusão dos algoritmos testados. Para este estudo a classe evasão é a positiva, enquanto a classe formado é a negativa.

Através das matrizes de confusão é possível observar o comportamento geral da predição e da confiança de cada algoritmo, no entanto, apenas estes dados não sugerem necessariamente qual terá uma maior confiança para predição perante as instâncias utilizadas para treino e teste.

Esta análise tem como função avaliar a confiabilidade dos algoritmos para predizer se um aluno será evasor. Deduz-se das matrizes de confusão que os algoritmos possui uma

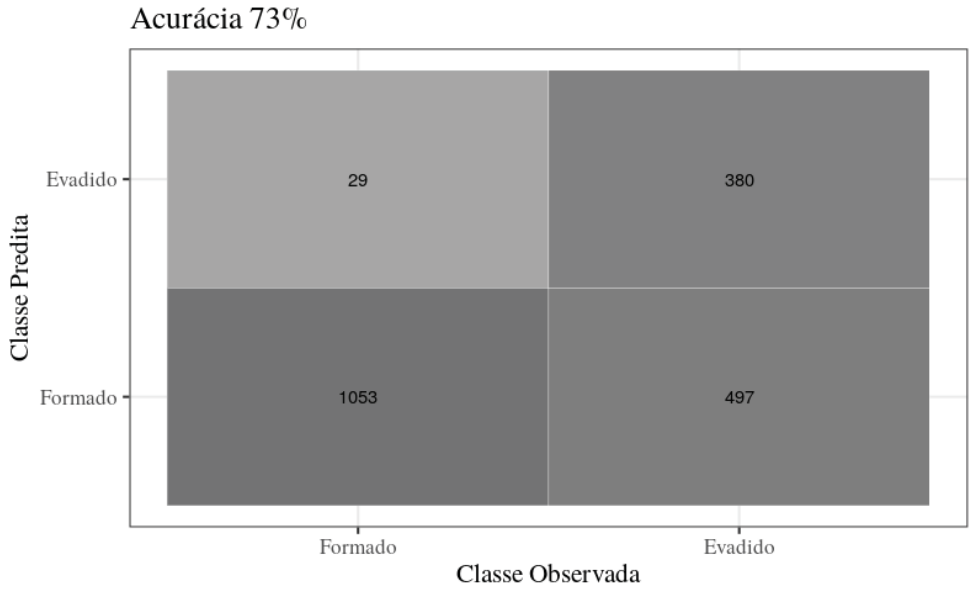


Figura 5.2: Matriz de confusão - Regressão Logística.

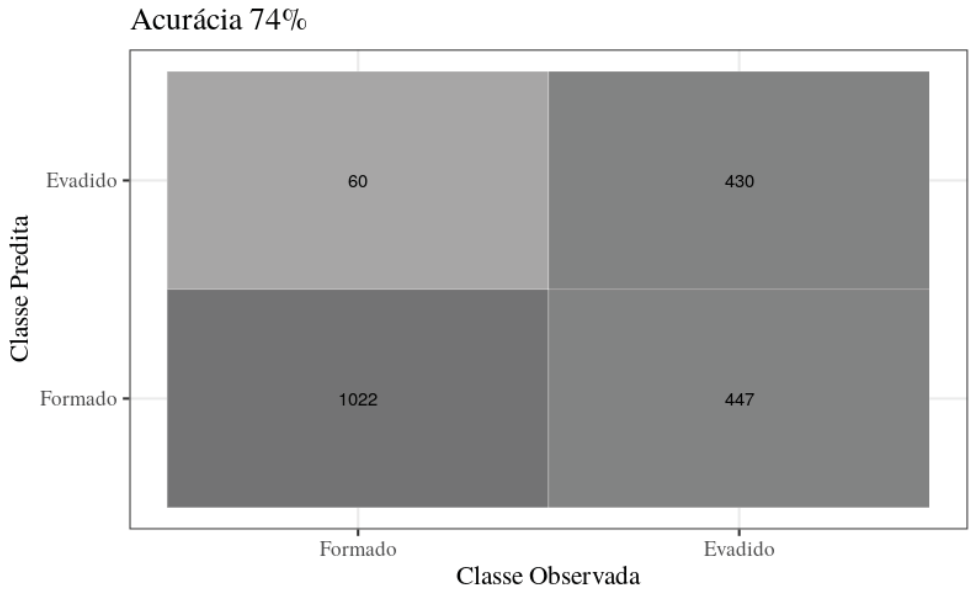


Figura 5.3: Matriz de confusão - CART.

boa predição de evadidos. Identificar os melhores algoritmos é possível avaliando apenas as quantidades de:

- Melhor VP: CART (430);
- Melhor VN: Random Forest (1059);
- Melhor FP: Random Forest (23);
- Melhor FN: CART (447).

Entretanto, é importante ressaltar que as demais métricas de qualidade devem ser verificadas. A tabela 5.3 compara as métricas dos modelos criados, ordenada por AUC.

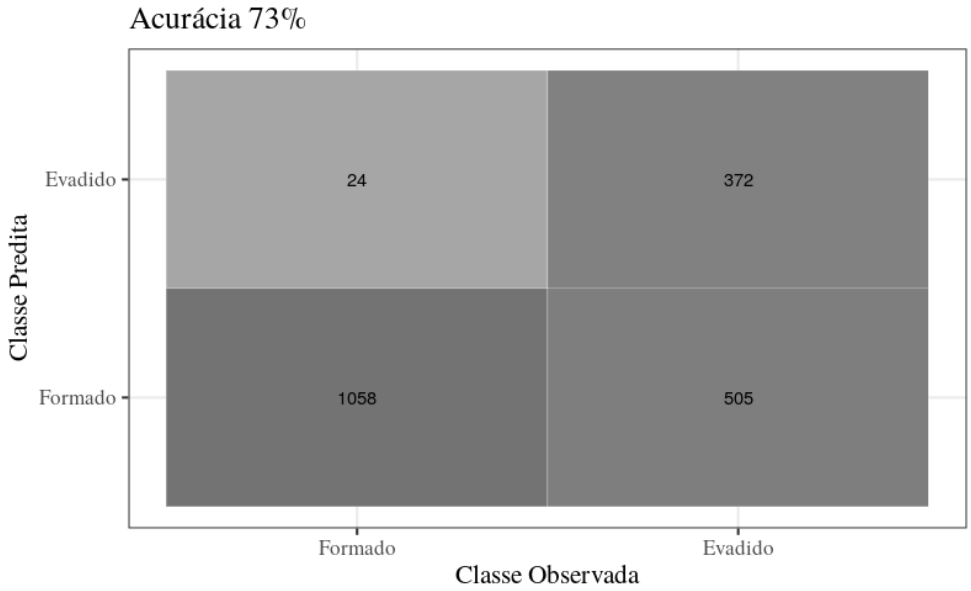


Figura 5.4: Matriz de confusão - CTree.

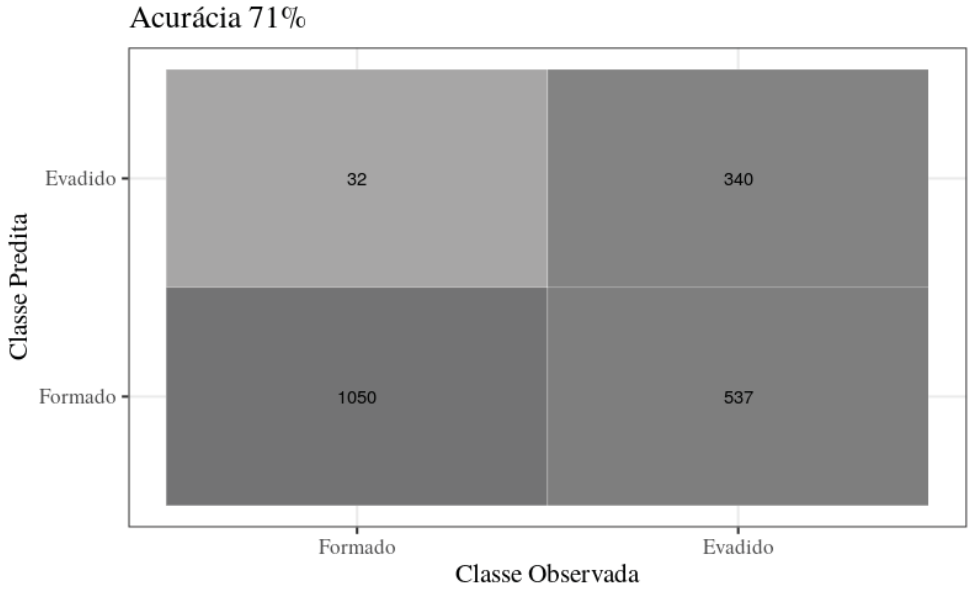


Figura 5.5: Matriz de confusão - KNN.

Tabela 5.3: comparação dos modelos do experimento.

	Random Forest	**SVM**	**CTree**	**Reg. Logística**	**KNN**	**CART**
AUC	0,863	0,847	0,845	0,841	0,84	0,797
Acurácia	0,734	0,741	0,73	0,731	0,71	0,741
Sensibilidade	0,979	0,959	0,978	0,973	0,97	0,945
Especificidade	0,432	0,472	0,424	0,433	0,388	0,49
Precisão	0,68	0,692	0,677	0,679	0,662	0,696
F1-Score	0,803	0,804	0,8	0,8	0,787	0,801

A AUC que foi a métrica utilizada para encontrar o melhor modelo nessa fase de comparação todos obtiveram um valor expressivo entre 0,797 até 0,863, onde o *random forest* e o SVM foram os melhores algoritmos.

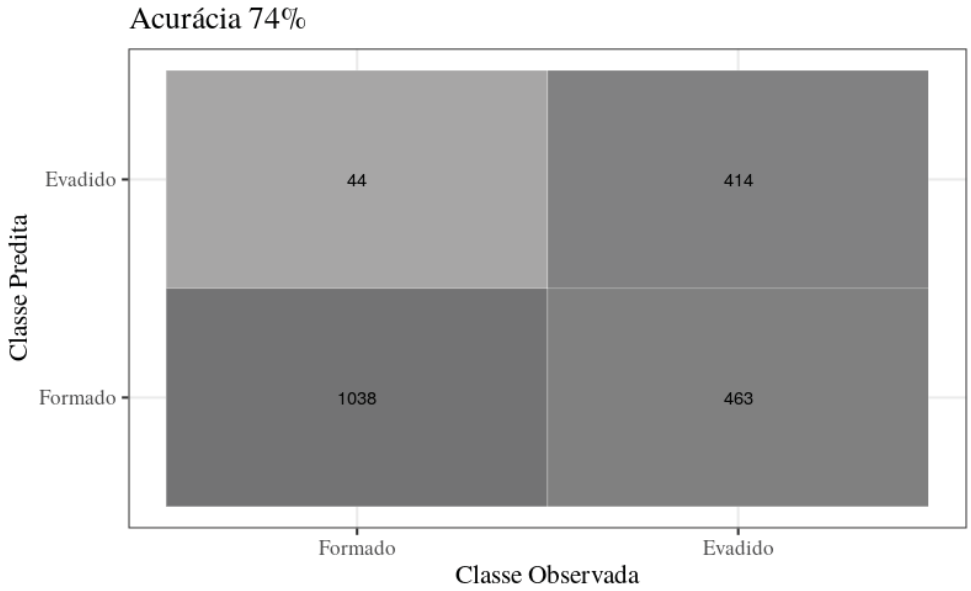


Figura 5.6: Matriz de confusão - SVM.

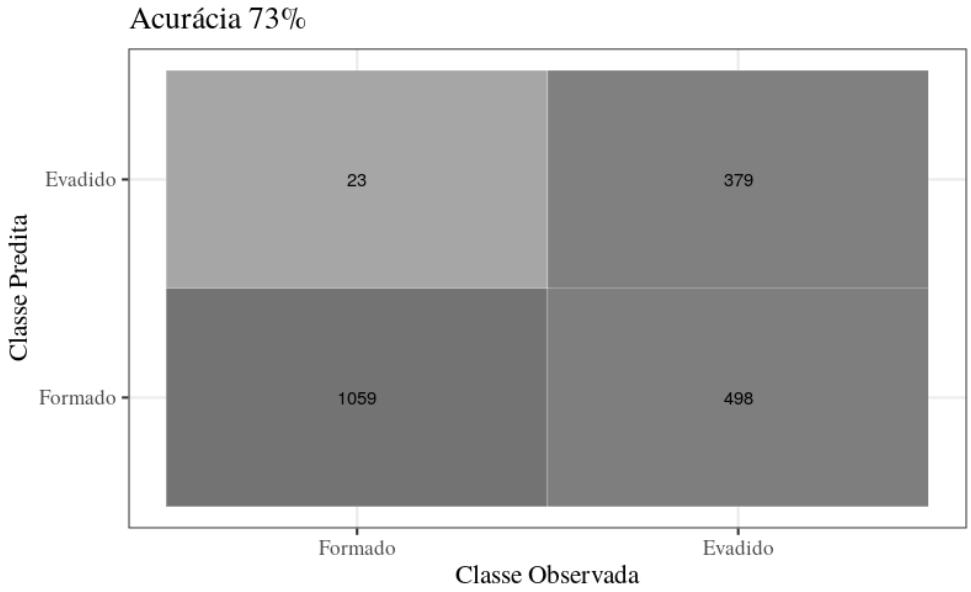


Figura 5.7: Matriz de confusão - *Random Forest*.

Observa-se também que a acurácia dos modelos se mantiveram entre 71% até 74%, o que pode-se considerar um bom resultado comparando com os trabalhos citados nessa pesquisa que na maioria se manteve na mesma porcentagem de acerto.

A métrica sensibilidade que neste trabalho é de grande valia, tendo em vista que é mais problemático deixar de acompanhar os alunos com probabilidade de evadir do que acompanhar os com probabilidade de formar obtiveram valores acima de 0,94. Alcançando uma taxa melhor do que os citados nos trabalhos de Moreira (2020) e Assis (2017) que ficaram entre 84% e 85% respectivamente.

Analisando os resultados dos modelos nas predições realizadas, percebe-se que, apesar da especificidades serem valores baixos, foi alcançado bons resultados nas métricas calculadas,

indicando que os modelos provavelmente apresentarão bons resultados preditivos em dados não vistos.

Além de verificar as métricas da tabela 5.3, plotou-se a curva ROC, como mostra a Figura 5.8, que corrobora com os dados mostrados pelas demais métricas. Visualmente pode-se perceber uma pequena vantagem do algoritmo *random forest*, como resume o valor da área abaixo da curva (AUC), frente aos demais.

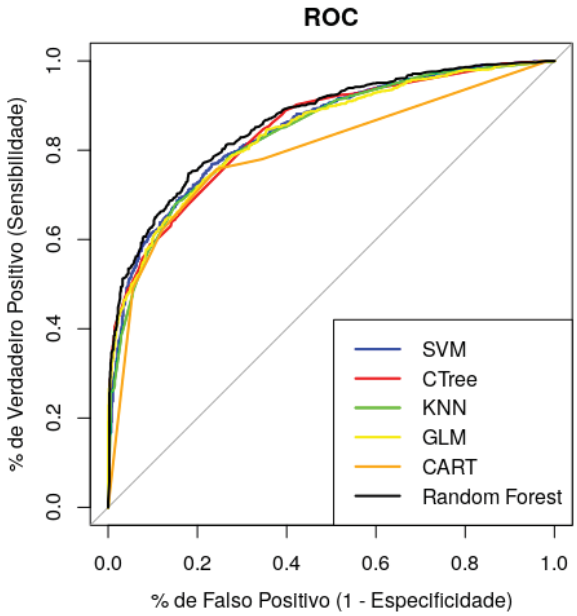


Figura 5.8: Curva ROC.

Os modelos de árvore de decisão e *random forest* contam com a vantagem de permitir a avaliação direta das características mais importantes para a classificação de um estudante. A fim de discutir sobre os fatores que, potencialmente, mais influenciam na distinção entre alunos evadidos e formados, e comparar com as variáveis que se mostraram significativas no modelo final da regressão logística, foi gerado a árvore de decisão do modelo CART, Figura 5.9, e também foi gerado gráfico da importância das variáveis para o modelo *random forest*, Figura 5.10.

Ao se analisar a Figura 5.9, percebe-se que a variável índice de rendimento acadêmico (ira) ficou na raiz da árvore de decisão e que o valor de corte para a evasão é menor a 0.48, esse é o mesmo valor alcançado no trabalho de Moreira (2020), que também usou dados da UFPR, quando analisado os dados do primeiro ano.

As outras variáveis que obtiveram maior ganho de informação na árvore de decisão foram setor, cota e carga horária cursada.

O modelo *random forest* não tem uma interpretação clara como na árvore de decisão, no entanto é possível calcular uma medida de importância de cada variável no modelo final. Essa medida é baseada na redução da soma de quadrados dos resíduos de cada divisão e te permite uma análise de quais variáveis são mais importantes na classificação.

Através dos modelos de regressão logística, árvore de decisão e *random forest* pode-se analisar que as variáveis mais importantes para os modelos permaneceram as mesmas. É possível também comparar que as variáveis mais importantes encontradas como o índice de desempenho acadêmico, por exemplo, corroboram com as encontradas nas pesquisas de Delen (2011) e Pinheiro et al. (2018).

A forma de ingresso citado na pesquisa de Teodoro e Kappel (2020) trata-se de uma variável importante, uma vez que os alunos ingressantes pelo SISU tendem a evadir mais do que

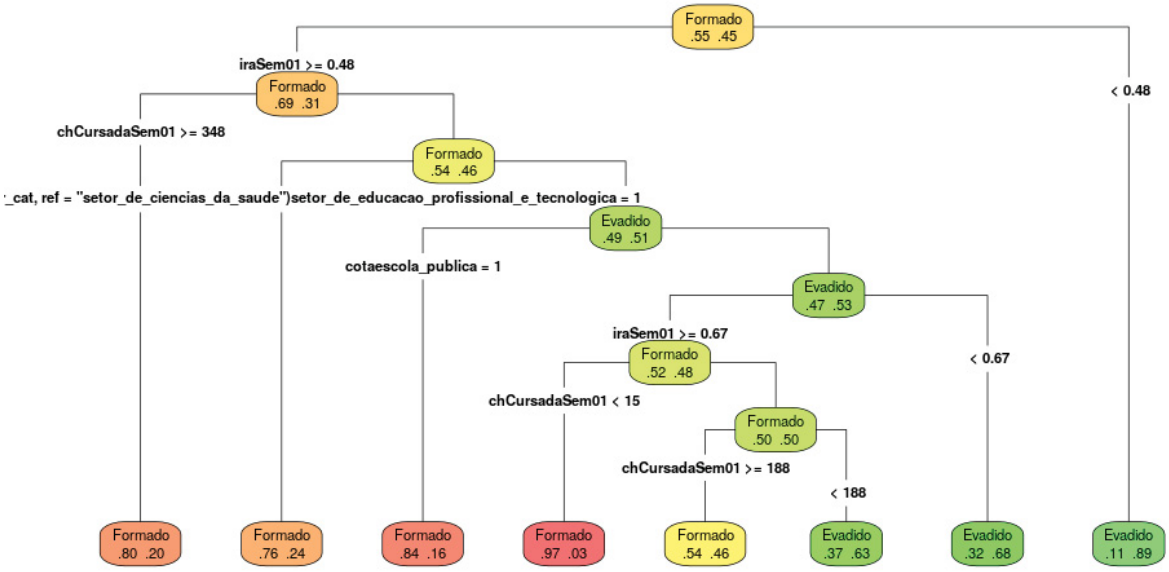


Figura 5.9: Árvore de decisão - CART.

Figura 5.10: Importância das variáveis no modelo *random forest*.

os ingressantes através do vestibular. Esse fato pode ser justificado pelo fato de muitos discentes acabam sendo chamados inicialmente para uma segunda opção e, quando reclassificados para o curso de primeira opção, desistem da vaga anterior.

A partir da maioria dos resultados obtidos e do respaldo dos resultados dos trabalhos citados, é possível evidenciar as principais características dos alunos evasores da UFPR. Além disso, é possível constatar que o *random forest* pode ser uma ferramenta acertiva na identificação de alunos com potencial de evasão e, com isso, os gestores podem agir de maneira precoce para prevenir esse abandono.

6 CONCLUSÃO

A evasão tem impacto em toda a comunidade escolar e não é fácil de ser combatida. As vagas desocupadas no sistema de ensino geram um desperdício de recursos, sendo assim, identificar padrões de um aluno que possivelmente não será um concluinte pode ser uma forma de enfrentar essa problemática.

As tecnologias de aprendizado de máquina estão crescendo consideravelmente nos últimos anos, principalmente no meio educacional. Nesse sentido, o principal objetivo tem sido gerar conhecimento a partir dos milhares de registros dos bancos de dados de sistemas e dar subsídios para pesquisas, que, no que lhe concerne, podem ser voltadas para as políticas de combate à evasão.

A UFPR possui um robusto banco de dados, do Sistema de Gestão Acadêmica da Universidade Federal do Paraná (SIGA/UFPR), em que é possível acessar informações de todo o tipo, como as informações socioeconômicas e vivência acadêmica dos discentes, por exemplo, podendo ser considerada uma ótima fonte de informações para estudos.

Neste trabalho com o modelo de regressão logística foi possível verificar quais as principais causas de evasão e também as variáveis que mais contribuem e apresentam maior risco para tal evento.

Observou-se, por exemplo, com relação às cotas, que, em geral, ingressar na universidade por cota de escola pública é fator de proteção em relação à ampla concorrência. Que quanto maior o IRA menor probabilidade de evasão. Já um aluno que ingressou na universidade pelo SISU é fator de risco em relação ao vestibular ou que quanto maior o número de reprovações por nota ou frequência maior chance de evadir.

Também foi possível comprovar que as técnicas de AM podem ser utilizadas de forma satisfatória para a mineração de dados, mostrando que a regressão logística, árvore de decisão, kNN, SVM e *random forest* são classificadores que possibilitam que a tarefa de identificar os alunos passíveis de evasão possa ser realizada de forma automática e que a gestão possa assim que finalizado o primeiro semestre tomar decisões a partir deste conhecimento gerado, sendo possível traçar estratégias de combate à evasão proporcionando uma mudança organizacional significativa.

Como trabalhos futuros espera-se aplicar outras técnicas de aprendizado de máquina, como métodos *ensemble* para comparar com o melhor classificador deste trabalho. Além de adicionar mais dados sobre o histórico acadêmico dos alunos, elencar perfis de alto risco, assim como, avaliar evolução temporal.

REFERÊNCIAS

Alghamdi, M., Al-Mallah, M., Keteyian, S., Brawner, C., Ehrman, J. e Sakr, S. (2017). Predicting diabetes mellitus using smote and ensemble machine learning approach: The henry ford exercise testing (fit) project. *Public Library of Science*, 12(7).

Andriola, W. B., Andriola, C. G. e Moura, C. P. (2006). Opiniões de docentes e de coordenadores acerca do fenômeno da evasão discente dos cursos de graduação da universidade federal do ceará (ufc). *Revista Ensaio: aval. pol. públ. Educ.*, 14(52):365–382.

Assis, L. R. S. (2017). Perfil de evasão no ensino superior brasileiro: uma abordagem de mineração de dados. Dissertação de Mestrado, Mestrado Profissional em Computação Aplicada - Universidade de Brasília, Brasília.

Aulck, L. S., Nambi, D., Velagapudi, N., Blumenstock, J. e West, J. D. (2019). Mining university registrar records to predict first-year undergraduate attrition. Em *Proceedings of The 12th International Conference on Educational Data Mining*, página 9–18.

Baker, R. S. J., Isotani, S. e Carvalho, A. M. J. B. (2011). Mineração de dados educacionais; oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19.2(36):3–11.

Barros, T. M. (2020). *Um processo orientado a dados para geração de modelo de predição de evasão escolar*. Tese de doutorado, UFRN - UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE, Natal, - Brasil. 137 pgs.

Branco, U. V. C. (2020). Ensino superior público e privado na paraíba nos últimos 15 anos: reflexões sobre o acesso, a permanência e a conclusão. `https://doi.org/10.1590/S1414-40772020000100004`. Acessado em 06/04/2021.

Corrêa, E. (2019). *Pandas Python - Data Wraling para Ciência de Dados*. Casa do Código.

Costa, A. G. (2021). Aplicação de técnicas de mineração de dados e learning analytics para predição de evasão de alunos nos cursos de ciência da computação e engenharias da ufpel. Dissertação de Mestrado, Mestrado em Ciência da Computação - Universidade Federal de Pelotas, Pelotas.

Cunningham, P. e Delany, S. J. (2007). Featureless similarity. Relatório técnico, University College Dublin. School of Computer Science and Informatics, Dublin.

Dekker, G. W., Pechenizkiy, M. e Vleeshouwers, J. M. (2009). Predicting students drop out: A case study. Em *International Working Group on Educational Data Mining*, Cordoba.

Delen, D. (2011). Predicting student attrition with data mining methods. Em *Journal of College Student Retention: Research, Theory & Practice*, página 17–35.

ENAP (2020). Análise de dados em linguagem r. `https://www.escolavirtual.gov.br/curso/325`. Acessado em 13/09/2020.

Escovedo, T. e Koshiyama, A. (2021). *Introdução a Data Science - Algoritmos de Machine Learning e Métodos de Análise*. Casa do Código.

Hoed, R. M. (2016). Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de computação. Dissertação de Mestrado, Mestrado Profissional em Computação Aplicada - Universidade de Brasília, Brasília.

Hoffmann, I. L., Nunes, R. C. M. e Martins, F. (2019). As informações do censo da educação superior na implementação da gestão do conhecimento organizacional sobre evasão. *Gestão & Produção*, 26(2):1–14.

Maksimova, N., Pentel, A. e Dunajeva, O. (2021). Predicting first-year computer science students drop-out with machine learning methods: A case study. Em *International Conference on Interactive Collaborative Learning*, página 719–726.

Marques, F. T. (2020). A volta aos estudos dos alunos evadidos do ensino superior brasileiro. *Cadernos de Pesquisa*, 50(178):1061–1077.

Marquesone, R. (2018). *Big Data - Técnicas e tecnologias para extração de valor dos dados*. Casa do Código.

Martins, T. S. (2018). Evasão universitária no ensino à distância: Análise dos fatores influenciadores. *Revista Estudos e Pesquisas em Administração*, 2(2).

Melo, A. L. (2019). Uso da técnica de mineração de dados como uma ferramenta de gestão da evasão no ensino superior. Dissertação de Mestrado, Mestrado Profissional em Inovação Tecnológica - Universidade Federal do Triângulo Mineiro, Uberaba.

Moreira, F. J. R. (2020). Aprendizagem de máquina na predição da evasão no ensino superior. `https://hdl.handle.net/1884/71062`. Acessado em 10/07/2021.

PETTERLE, R. R., BARBOZA, L. A. S. e CARVALHO, M. (2017). Fatores de risco associados À nefrolitÍase recorrente via regressÃo logÍstica binÁria. *Journal of Biometrics*, 35(2):348–360. r.

Pinheiro, M. A. L., Silva, J. C. e d. Souza, B. F. (2018). Aprendizado de máquina aplicado à análise de evasão no ensino superior. Em *IX Computer on the Beach*, páginas 512–521, Florianópolis.

PNP (2018). Plataforma nilo peçanha. `http://plataformanilopecanha.mec.gov.br/2019.html`. Acessado em 10/10/2021.

PNP (2019). Plataforma nilo peçanha. `http://plataformanilopecanha.mec.gov.br/2020.html`. Acessado em 10/10/2021.

Pérez, A., Grandón, E. E., Caniupán, M. e Vargas, G. (2018). Comparative analysis of prediction techniques to determine student dropout: Logistic regression vs decision trees. Em *37th International Conference of the Chilean Computer Science Society (SCCC)*, páginas 1–8, Santiago, Chile.

Saccaro, A., Franca, M. T. A. e Jacinto, P. A. (2019). Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. *Estud. Econ.*, 49(2):337–373.

Santos, G. L. e Guimarães, P. B. V. (2019). *Governo Digital - uma abordagem interdisciplinar na gestão da educacão superior*, páginas 9–22. Editora Motres.

SIGAUFPR (2021). Sistema integrado de gestão acadêmica. `https://siga.ufpr.br/portal/`. Acessado em 03/04/2021.

Silva, A. R. (2022). Modelos de machine learning utilizando o pacote caret. `https://statplace.com.br/blog/modelos-de-machine-learning-utilizando-o-pacote-caret/`. Acessado em 03/09/2022.

Silva, R. L. L., Motejunas, P. R., Hipólito, O. e Lobo, M. B. C. M. (2007). A evasão no ensino superior brasileiro. Em *Scielo Brasil*, páginas 641–659.

Soares, L. C. C. P. (2020). Soluções tecnológicas para o problema da evasão universitária, sob a óptica de ferramentas de inteligência artificial. Dissertação de Mestrado, Programa de Pós-Graduação em Propriedade Intelectual e Transferência de Tecnologia para a Inovação - Universidade Federal do Tocantins, Palmas.

Souza, A. M. (2020). Machine learning e a evasão escolar - análise preditiva no suporte à tomada de decisão. Dissertação de Mestrado, Mestrado em Sistemas de Informação e Gestão do Conhecimento - Universidade FUMEC, Belo Horizonte.

Tan, P. N., Steinbach, M., Karpatne, A. e Kumar, V. (2005). Introduction to data mining (second edition). `https://www-users.cse.umn.edu/~kumar001/dmbook/index.php`. Acessado em 15/10/2021.

Teodoro, L. A. e Kappel, M. A. A. (2020). Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no brasil. *Revista Brasileira de Informatica na Educação- RBIE*, 28:838–863.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., P. S. Yu, Z. H. Z., Steinbach, M., Hand, D. J. e Steinberg, D. (2007). Top 10 algorithms in data mining. Em *Knowledge and Information Systems volume*, página 14(1):1–37.