Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus João Pessoa

Programa de Pós-Graduação em Tecnologia da Informação

RONEI DOS SANTOS OLIVEIRA

MODELO DE PREDIÇÃO DE EVASÃO ESCOLAR COM BASE EM DADOS DE AUTOAVALIAÇÃO DE CURSOS DE GRADUAÇÃO

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA

2023

Ronei dos Santos Oliveira

Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB.

Orientador: Prof. Dr. Francisco Petrônio Alencar de Medeiros

João Pessoa

2023

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Nilo Peçanha - *Campus* João Pessoa, PB.

048m Oliveira, Ronei dos Santos.

Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação / Ronei dos Santos Oliveira . – 2023.

73 f. : il.

Dissertação (Mestrado em Tecnologia da Informação) – Instituto Federal de Educação da Paraíba / Programa de Pós-Graduação em Tecnologia da Informação (PPGTI), 2023.

Orientação: Prof. D.r Francisco Petrônio Alencar de Medeiros.

1. Evasão escolar. 2. Mineração de dados educacionais. 3. Modelo preditivo. 4. Autoavaliação. 5. Educação superior. I. Título.

CDU 37.015.3(043)

Elaboração: Lucrecia Camilo de Lima, Bibliotecária – CRB 15/132

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU*

MESTRADO PROFISSIONAL EM TECNOLOGIA DA INFORMAÇÃO

RONEI DOS SANTOS OLIVEIRA

MODELO DE PREDIÇÃO DE EVASÃO ESCOLAR COM BASE EM DADOS DE AUTOAVALIAÇÃO DE CURSOS DE GRADUAÇÃO

Dissertação apresentada como requisito para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós- Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB - Campus João Pessoa.

Aprovado em 03 de agosto de 2023

Membros da Banca Examinadora:

Dr. Francisco Petrônio Alencar de Medeiros

IFPB - PPGTI

Dr. Francisco Dantas Nobre Neto

IFPB

Dr. José Jorge Lima Dias Júnior

UFPB

João Pessoa/2023

Documento assinado eletronicamente por:

- **Francisco Petronio Alencar de Medeiros**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 03/08/2023 21:45:18.
- **Francisco Dantas Nobre Neto**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 03/08/2023 22:24:49.
- **Jose Jorge Lima Dias Junior**, **PROFESSOR DE ENSINO SUPERIOR NA ÁREA DE ORIENTAÇÃO EDUCACIONAL**, em 04/08/2023 19:42:50.

Este documento foi emitido pelo SUAP em 17/07/2023. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código 449970
Verificador: 71d2010676
Código de Autenticação:

Av. Primeiro de Maio, 720, Jaguaribe, JOAO PESSOA / PB, CEP 58015-435
http://ifpb.edu.br - (83) 3612-1200

Este trabalho é dedicado às pessoas que lutam diariamente para oferecer uma educação gratuita, digna e de qualidade.

AGRADECIMENTOS

Agradeço o apoio do Instituto Federal da Paraíba (IFPB), especialmente representado pelo meu orientador Dr. Francisco Petrônio Alencar de Medeiros, que se empenhou para viabilizar este trabalho. Agradeço também aos demais colegas do Programa de Pós-Graduação em Tecnologia da Informação, que foram essenciais para obter êxito na conclusão das disciplinas cursadas e desenvolvimento deste trabalho. Por fim e não menos importante, agradeço à minha família, especialmente à minha esposa Milena Christina Cunha Soares, pela paciência e apoio incondicional.

RESUMO

A evasão escolar constitui um desafio cotidiano enfrentado pelas instituições de ensino no Brasil e no mundo. No caso específico das instituições de ensino superior brasileiras, as taxas de evasão escolar permanecem em patamares preocupantes e inaceitáveis, resultando em perdas financeiras significativas e escassez de profissionais em áreas específicas da educação. Diante desse cenário, o objetivo desta pesquisa foi desenvolver e avaliar modelos preditivos para identificar alunos com maior propensão à evasão escolar, utilizando dados de um modelo semestral de autoavaliação dos cursos de graduação da Universidade Federal da Paraíba (UFPB). A metodologia utilizada neste estudo foi a mineração de dados educacionais, com base na metodologia CRISP-EDM. Inicialmente, foi realizada uma compreensão do domínio, investigando a problemática da evasão escolar e sua relação com os dados da autoavaliação institucional. Em seguida, foi realizada uma análise exploratória dos dados da autoavaliação dos cursos da UFPB, seguida pela preparação dos dados para a tarefa de classificação. Diversas técnicas de aprendizado de máquina foram aplicadas, incluindo Árvore de Decisão (*Decision Tree* - DT), Floresta Aleatória (*Random Forest* - RF) e Máquinas de Vetores de Suporte (*Support Vector Machine* - SVM). Os modelos desenvolvidos foram avaliados usando métricas de desempenho, como acurácia, precisão, recall e medida F. Os resultados mostraram que o modelo preditivo alcançou uma acurácia de 87,97%, precisão de 91,72%, recall de 91,67% e medida F de 91,57% na identificação dos alunos com maior propensão à evasão escolar. Além disso, constatou-se que aproximadamente 59% dos alunos ativos, admitidos entre 2017 e 2021, apresentam uma maior probabilidade de abandonar seus cursos. Essas informações são relevantes para embasar decisões institucionais e orientar a implementação de políticas e ações eficazes de combate à evasão escolar, visando mitigar esse problema e alcançar melhores resultados acadêmicos. Ao adotar a abordagem da mineração de dados educacionais e a metodologia CRISP-EDM, este estudo contribui para o avanço no campo da predição de evasão escolar, fornecendo informações para a tomada de decisões e o desenvolvimento de estratégias preventivas no contexto da UFPB e de outras instituições de ensino superior.

Palavras-chaves: Evasão escolar; Mineração de dados educacionais; Modelo preditivo; Autoavaliação.

ABSTRACT

School dropout represents a daily challenge faced by educational institutions in Brazil and around the world. In the specific case of Brazilian higher education institutions, dropout rates remain at concerning and unacceptable levels, resulting in significant financial losses and shortages of professionals in specific educational fields. Given this scenario, the aim of this research was to develop and evaluate predictive models to identify students with a higher propensity for school dropout, utilizing data from a semester-based self-assessment model of undergraduate courses at the Federal University of Paraíba (UFPB). The methodology used in this study was educational data mining, based on the CRISP-DM methodology. Initially, domain understanding was achieved by investigating the issue of school dropout and its relationship with institutional self-assessment data. Subsequently, exploratory analysis of the UFPB course self-assessment data was conducted, followed by data preparation for the classification task. Various machine learning techniques were applied, including Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM). The developed models were evaluated using performance metrics such as accuracy, precision, recall, and F-measure. The results showed that the predictive model achieved an accuracy of 87.97%, precision of 91.72%, recall of 91.67%, and F-measure of 91.57% in identifying students with a higher propensity for school dropout. Additionally, it was found that approximately 59% of active students admitted between 2017 and 2021 have a higher probability of leaving their courses. This information is relevant for informing institutional decisions and guiding the implementation of effective policies and actions to combat school dropout, aiming to mitigate this issue and achieve better academic outcomes. By adopting the approach of educational data mining and the CRISP-DM methodology, this study contributes to advancements in the field of school dropout prediction, providing insights for decision-making and the development of preventive strategies within the context of UFPB and other higher education institutions.

Keywords: School dropout; Educational data mining; Predictive model; Self-assessment.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizado de Máquina
CRISP-EDM	*CRoss-Industry Standard Process for Educational Data Mining*
DT	Árvore de Decisão (*Decision Tree*)
ETL	Extração, Transformação e Carga
ETL	Extração, Transformação e Carga
FN	Falso Negativo
FP	Falso Positivo
IES	Instituições de Ensino Superior
MDE	Mineração de Dados Educacionais
RF	Floresta Aleatória (*Random Forest*)
RSL	Revisão Sistemática da Literatura
SIGAA	Sistema Integrado de Gestão de Atividades Acadêmicas
SMOTE	Synthetic Minority Oversampling Technique
SVM	Máquinas de Vetores de Suporte (*Support Vector Machine*)
UFPB	Universidade Federal da Paraíba
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1. INTRODUÇÃO

A evasão escolar é um problema antigo, complexo e presente em todos os níveis de ensino. Além disso, a evasão escolar acarreta graves prejuízos econômicos e sociais, afetando indivíduos, gestão escolar, instituições e a sociedade como um todo (PRESTES & FIALHO, 2018). No Ensino Superior, a evasão escolar tem impacto na escassez de profissionais em diversas áreas, afetando todo o ecossistema necessário (SACCARO; FRANÇA; JACINTO, 2019).

Com o objetivo de esclarecer a definição de evasão escolar no contexto da educação superior, uma comissão formada pelo Ministério da Educação (ANDIFES; ABRUEM; SESU/MEC, 1996), reconhecendo suas possíveis limitações, optou por caracterizá-la de acordo com a Tabela 1.

Tabela 1 – Definições de evasão no âmbito da educação superior.

	Definição
Evasão do curso	Quando o estudante se desliga do curso superior em situações diversas tais como: • Abandono (deixa de matricular-se); • Desistência (oficial); • Transferência ou reopção (mudança de curso); • Exclusão por norma institucional.
Evasão da instituição	Quando o estudante se desliga da instituição na qual está matriculado.
Evasão do sistema	Quando o estudante abandona de forma definitiva ou temporária o ensino superior.

Este trabalho considerará como evasão escolar apenas o fenômeno da evasão do curso, uma vez que os fenômenos da Evasão da Instituição e do Sistema possuem particularidades e nuances que não serão objeto de estudo nesta pesquisa.

Ao abordarmos os problemas decorrentes da evasão escolar nas instituições públicas brasileiras, é importante considerar que essas instituições desempenham um papel fundamental no setor produtivo, sendo responsáveis por 6 em cada 10 produções científicas realizadas no país (GAMBA & RIGHETTI, 2022) e pela maioria dos registros de patentes brasileiras, com 19 instituições entre as 25 maiores depositantes de produtos ou serviços (LOUSRHANIA, 2021). Nesse contexto, a evasão escolar vai de encontro à efetividade desse ecossistema, interrompendo a trajetória acadêmica do aluno e seu progresso rumo à conclusão do curso. Além disso, a evasão

escolar gera perdas econômicas para o Estado e para a gestão universitária, que deixam de receber os benefícios do aluno evadido.

Apesar das observações e concepções sobre a temática, compreender as razões da evasão escolar continua sendo um desafio para a maioria das instituições. Com o objetivo de obter essa compreensão, a Universidade Federal da Paraíba (UFPB) realiza, a cada semestre, por meio do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA), uma autoavaliação compulsória e anônima dos cursos de graduação, como requisito para a matrícula. O instrumento de avaliação da educação superior pelo discente, proposto por COSTA & DIAS (2020), abrange quatro dimensões: Discente, Disciplina, Docente e Curso. Na dimensão Discente, o aluno avalia seu desempenho no semestre anterior. Na dimensão Disciplina, o aluno avalia cada disciplina cursada, sua importância e dificuldade. Na dimensão Docente, o aluno pode sugerir ajustes nas disciplinas e avaliar o desempenho do docente. Na dimensão Curso, o aluno pode informar se recomendaria o curso e qual sua intenção de abandoná-lo naquele momento. Essa autoavaliação gera uma grande quantidade de dados com múltiplas perspectivas para o estudo das especificidades da instituição.

A autoavaliação das Instituições de Ensino Superior (IES) como um processo contínuo de reflexão sobre todas as ações institucionais, incluindo estrutura, ensino, pesquisa, extensão, relações internas e externas, além das atividades administrativas, fornece conteúdo necessário para orientar a gestão institucional, indo além da prestação de contas ao Ministério da Educação. A relação entre avaliação e gestão pode ser compreendida de diferentes maneiras, dependendo da missão, das razões históricas e das características de cada IES. A busca pelo conhecimento institucional e seus problemas internos é promovida por programas que propõem mudanças, incluindo medidas para reduzir o número de alunos evadidos (BAGGI & LOPES, 2011). Nesse sentido, é de extrema importância que as universidades busquem alternativas que promovam uma autoavaliação cada vez mais eficaz, refletindo as especificidades de cada IES, e que vá além de modelos genéricos de coleta e divulgação de dados, possibilitando transformações necessárias para a melhoria da qualidade educacional, incluindo mudanças na cultura acadêmica, no trabalho docente, na gestão institucional, nas definições curriculares e, acima de tudo, na estruturação.

De acordo com a revisão sistemática realizada por (SANTOS; SARAIVA; OLIVEIRA, 2021), diversos estudos de Mineração de Dados Educacionais (MDE) foram realizados nos últimos anos no contexto da evasão escolar. Dentre os tipos de dados investigados, destacam-se dados acadêmicos, socioeconômicos, demográficos, de interação social, motivação, fatores psicológicos, fatores pessoais e problemas de saúde. Embora não tenham sido identificados estudos que utilizem

dados de autoavaliação institucional, a natureza dos dados da autoavaliação dos cursos apresenta correlação com os dados atualmente estudados no contexto da evasão escolar. Com base nesse contexto, esta pesquisa pretende explorar os dados da autoavaliação dos cursos de graduação da UFPB para desenvolver um modelo preditivo de evasão escolar.

1.1. Motivação e Definição do Problema

A utilização da MDE sobre os dados do instrumento de autoavaliação dos cursos pode ser um meio importante para subsidiar a alta gestão na implementação de ações específicas que sejam determinantes para que os estudantes consigam concluir o curso no período regular estabelecido pelo projeto pedagógico, garantindo assim uma formação de qualidade que os habilite a ingressar no mercado de trabalho ou prosseguir seus estudos na academia.

Este trabalho pretende trazer contribuições nos campos educacional, científico-tecnológico e social. No campo educacional, busca fornecer insumos para a gestão institucional realizar ações que promovam maior qualidade e eficácia do ensino oferecido. Do ponto de vista científico-tecnológico, a pesquisa científica e o uso de recursos computacionais para o desenvolvimento de um modelo preditor deixam um legado nessa área. Em termos sociais, a contribuição poderá se dar na contribuição para redução dos índices de evasão escolar, permitindo que os alunos cumpram sua jornada acadêmica e ofereçam o retorno esperado pela sociedade e pela comunidade acadêmica.

A autoavaliação dos cursos de graduação também pode contribuir para os processos acadêmicos e administrativos, sendo um instrumento para corrigir metas e objetivos. No entanto, sua simples aplicação não é suficiente para a melhoria dos processos. O instrumento não pode se limitar a uma atividade de coleta e divulgação de dados, pois isso não promove as transformações necessárias para a melhoria da qualidade educacional. É preciso processar esses dados utilizando inteligência cognitiva e computacional para gerar informações de grande valor para a instituição, possibilitando mudanças na cultura acadêmica, no trabalho docente, na gestão da instituição, nas definições curriculares e, principalmente, na estruturação.

1.2. Objetivos

1.2.1. Objetivo geral

Desenvolver um modelo preditivo que utilize os dados da autoavaliação dos cursos de graduação da UFPB para identificar precocemente a evasão escolar, possibilitando a adoção de

medidas preventivas por parte dos stakeholders, como a alta gestão, coordenadores de cursos e professores.

1.2.2. Objetivos específicos

Com o propósito de atingir o objetivo central, os seguintes objetivos específicos foram definidos:

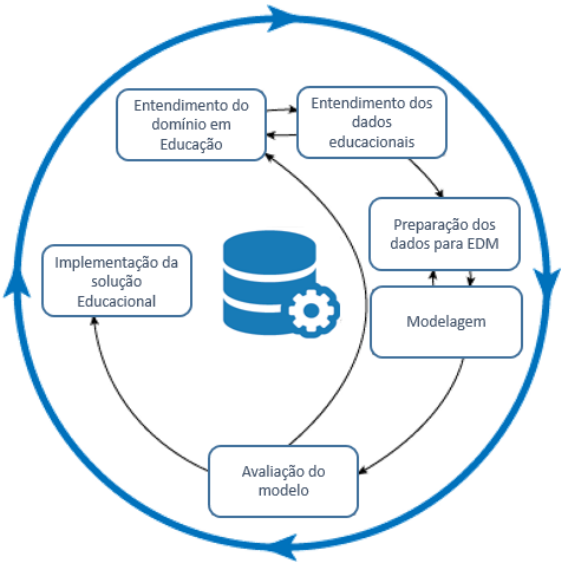
- Analisar a literatura atual em Mineração de Dados Educacionais (MDE) focada na previsão de evasão escolar.
- Explorar os dados provenientes das avaliações dos cursos da UFPB para compreender padrões relevantes.
- Desenvolver um conjunto de dados otimizado para a tarefa de classificação.
- Implementar modelos de classificação utilizando técnicas de Aprendizado de Máquina, incluindo Árvore de Decisão, Floresta Aleatória e Máquinas de Vetores de Suporte.
- Avaliar e comparar o desempenho dos modelos em relação à identificação de possíveis casos de evasão.
- Aplicar o modelo selecionado para prever evasões entre os estudantes atuais da instituição.
- Analisar o valor do conhecimento adquirido para informar futuras ações na redução da evasão escolar.

1.3. Metodologia

Para atingir os objetivos estabelecidos, foi realizado um processo de Mineração de Dados Educacionais (MDE), que envolve a aplicação de métodos de mineração de dados e aprendizado de máquina no contexto educacional, visando descobrir conhecimentos em bases de dados educacionais (SARAIVA et al., 2019). Para auxiliar nesse processo, esta pesquisa utilizou a metodologia CRISP-EDM (acrônimo de CRoss-Industry Standard Process for Educational Data Mining), uma adaptação da metodologia CRISP-DM (acrônimo de CRoss-Industry Standard Process for Data Mining) para o contexto educacional (RAMOS et al., 2020).

Cada etapa do CRISP-EDM foi desenvolvida com técnicas e abordagens adequadas ao domínio educacional em análise. Dessa forma, a pesquisa foi conduzida em seis etapas, conforme ilustrado na Figura 1.

Figura 1 – Metodologia CRISP-EDM



Fonte: Extraído de (RAMOS et al., 2020)

- **Entendimento do domínio**: Nesta fase, realizou-se um levantamento sobre a problemática da evasão escolar no contexto do ensino superior e sua relação com os dados da autoavaliação institucional da UFPB. O entendimento adequado do domínio auxiliou na definição dos objetivos da mineração e na seleção adequada das variáveis coletadas para o processo.
- **Entendimento dos dados educacionais**: Nesta etapa, foram realizadas análises e compreensões dos dados brutos da autoavaliação do curso. Foi conduzida uma análise exploratória dos dados com o objetivo de entender os caminhos a serem seguidos no processo de mineração.
- **Preparação dos dados para MDE**: Com base no entendimento do domínio e dos dados, realizou-se a preparação dos dados para o desenvolvimento da modelagem. Nessa etapa, unificaram-se dados de contextos distintos, selecionou-se um subconjunto representativo dos dados estudados, trataram-se valores nulos e *outliers* e transformaram-se os dados para otimizar o desempenho dos algoritmos de aprendizado de máquina.
- **Modelagem**: A partir dos dados previamente processados, foi realizada a etapa

elaboração dos modelos preditivos, utilizando uma abordagem de classificação supervisionada. Nesse contexto, foram selecionados algoritmos de aprendizado de máquina – Árvore de Decisão (Decision Tree - DT), Floresta Aleatória (Random Forest - RF) e Máquinas de Vetores de Suporte (Support Vector Machine - SVM) – com base no estado da arte.

- **Avaliação do modelo**: Nesta etapa, avaliou-se o desempenho dos modelos desenvolvidos com as técnicas de mineração, quando aplicados aos dados reais. Para isso, analisaram-se as métricas de Acurácia, Precisão, *Recall* e *F-Measure* dos modelos preditivos de evasão de alunos.
- **Implementação da solução educacional**: Com o modelo proposto selecionado com base nas melhores métricas, ele foi aplicado aos dados das autoavaliações dos alunos ativos na instituição. Dessa forma, observou-se que o modelo proposto pode servir como base para uma solução educacional que pode ser implantada na instituição, permitindo que os stakeholders visualizem de maneira objetiva os alunos com maior risco de evasão escolar.

Seguindo essas etapas, a pesquisa avançou de forma estruturada e direcionada, possibilitando a construção de um modelo preditivo para a evasão escolar e fornecendo informações valiosas para a tomada de decisões e ações preventivas no contexto da UFPB.

1.4. Aplicabilidade

A proposta deste estudo é desenvolver um modelo de previsão de evasão escolar a partir da análise dos dados da autoavaliação dos cursos da UFPB. Com base nesses dados, é possível identificar com alta precisão os alunos que abandonaram o curso. Assim, o modelo pode ser utilizado para detectar precocemente os alunos com maior probabilidade de desistência, permitindo que a instituição tome medidas para mitigar esse problema.

Espera-se que o modelo gerado possa ser aprimorado e se torne a base de uma solução educacional para a instituição. Validada em um ambiente real, essa solução poderá ser replicada em outras instituições de ensino superior.

1.5. Estrutura do Documento

Os capítulos seguintes estão organizados da seguinte forma:

- O Capítulo 2 aborda a fundamentação teórica necessária para compreender este trabalho, incluindo o panorama da evasão escolar no ensino superior, utilizando como base o Mapa do Ensino Superior no Brasil de 2023, conceitos e técnicas de MDE para a previsão da evasão escolar, além de revisar o estado da arte e trabalhos relacionados.
- O Capítulo 3 descreve a proposta de pesquisa, detalhando a aplicação da metodologia utilizada.
- O Capítulo 4 traz as considerações finais e indicações para trabalhos futuros.

2. REFERENCAL TEÓRICO E TRABALHOS RELACIONADOS

Este capítulo tem como propósito apresentar, de maneira clara e sucinta, todo o embasamento teórico que serviu de fundamento para o desenvolvimento deste trabalho, bem como para pesquisas relacionadas.

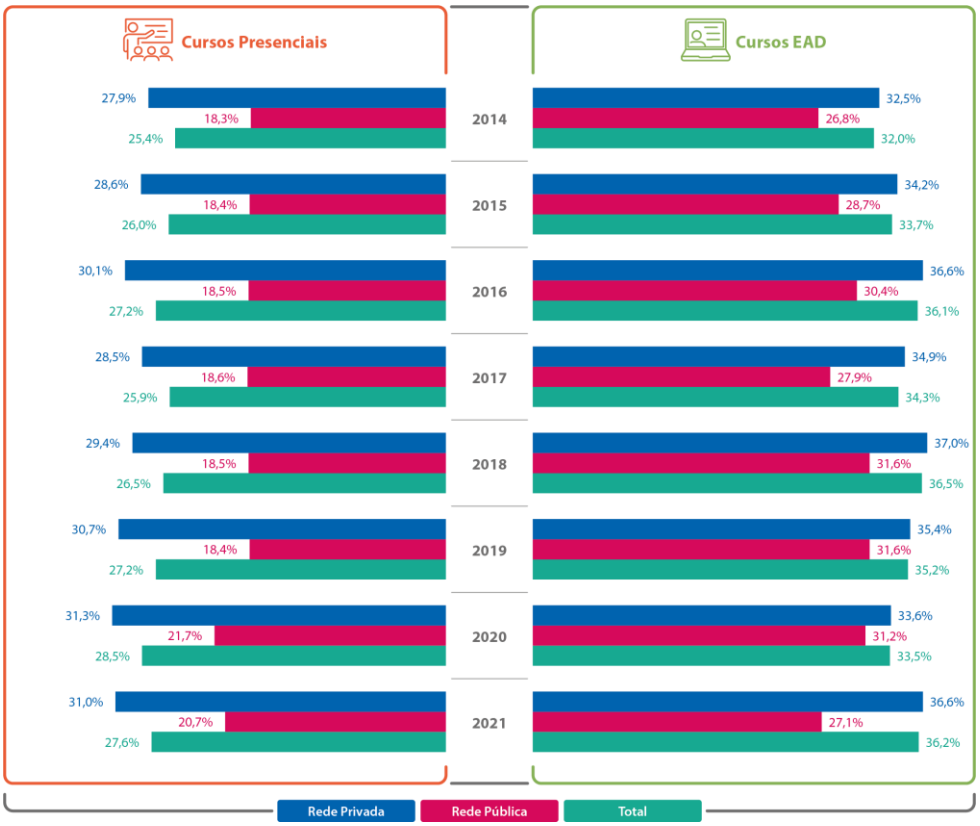
2.1. Evasão escolar na educação superior

Para entender como o fenômeno da evasão vem afetando as instituições públicas e privadas de educação superior no Brasil, é de extrema importância analisar os dados do Mapa do Ensino Superior no Brasil (2023).

2.1.1. Índices de evasão no Brasil

É possível observar que as taxas de evasão, no Brasil, na modalidade presencial, objeto de estudo desta pesquisa, apresenta um aumento nos índices nos últimos anos.

Figura 2 – Taxa de Evasão na Educação Superior no Brasil



* *Taxa de Evasão = (Matrículas Trancadas + Desvinculado Curso + Falecidos) / (Total de Matrículas + Matrículas Trancadas + Desvinculado Curso + Falecidos)

Fonte: Instituto Semesp

Esses dados trazem luz sobre a falta de efetividade das ações que estão sendo tomada pelas instituições para impedir a saída de alunos.

2.1.2. Indicadores de trajetória

Para obter uma compreensão mais precisa dos índices de evasão, é fundamental analisar os indicadores que traçam a trajetória dos estudantes nas IES, conforme destacado pelo Instituto Semesp em 2023:

- **Taxa de Permanência**: Percentual de ingressantes que estão com vínculo ativo no curso no ano de referência.
- **Taxa de Conclusão Acumulada**: Percentual de ingressantes que concluíram o curso até o ano de referência.
- **Taxa de Desistência Acumulada**: Percentual de ingressantes que desistiram do curso até o ano de referência.

Figura 3 – Indicadores de Trajetória

Fonte: Instituto Semesp

Na Figura 3, é possível observar que o Brasil segue com altas taxas de evasão. É possível acompanhar a trajetória dos estudantes, especificamente ingressantes de 2017, para checar os

percentuais de sucesso e desistência dos alunos. A taxa de conclusão é de apenas 26,3%, por exemplo, com as maiores taxas de concluintes em cursos presenciais e EAD na rede privada.

2.1.3. Evasão na educação superior da Paraíba

Outro importante contexto que devemos observar, é o do estado da Paraíba, no qual a UFPB está inserida. A representatividade da Paraíba no número total de matrículas no país é de 1,8%. Em relação ao Nordeste, esse percentual sobe para 8,8%. A Figura 4 apresenta esse quantitativo de matrículas em números.

Figura 4 – Matrículas na Educação Superior na Paraíba

		Cursos Presenciais*				Cursos EAD**			
		Matrículas				Matrículas			
Mesorregião	Municípios	Rede Privada	Rede Pública	Total	IES	Rede Privada	Rede Pública	Total	IES
Agreste Paraibano	66	12.528	24.419	36.947	14	13.103	805	13.908	48
Borborema	44	182	1.899	2.081	4	510	343	853	11
Mata Paraibana	30	31.489	28.993	60.482	25	20.671	829	21.500	56
Sertão Paraibano	83	9.356	9.214	18.570	9	8.476	536	9.012	28
Total - Estado PB	**223**	**53.555**	**64.525**	**118.080**	**43**	**42.760**	**2.513**	**45.273**	**69**

Fonte: Instituto Semesp

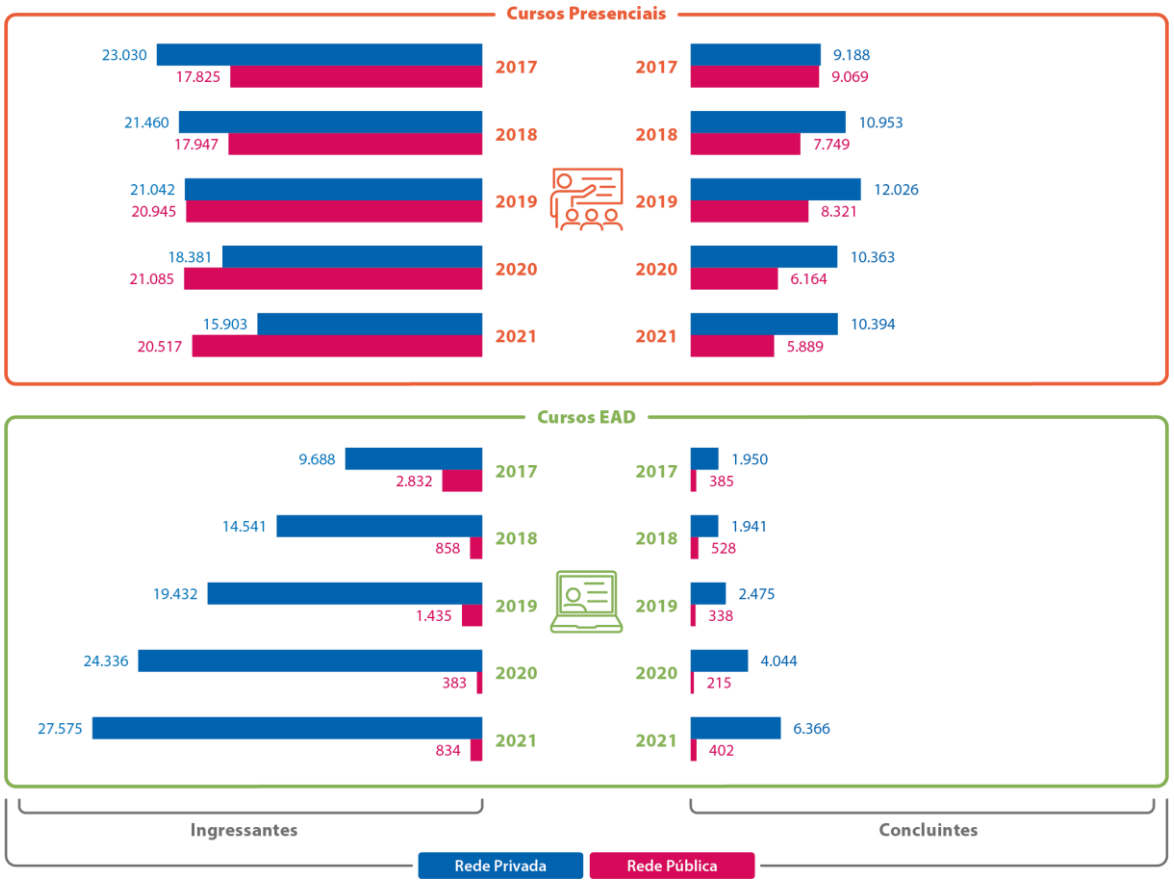
Quando analisamos os indicadores de trajetória no estado da Paraíba, Figura 5, observamos que os índices estão acima da média nacional, evidenciando um problema crítico a ser enfrentado.

Figura 5 – Indicadores de trajetória em cursos presenciais na Paraíba

Fonte: Instituto Semesp

A relação entre ingressantes e concluintes no estado, Figura 6, também apresentam números piores do que a média nacional.

Figura 6 – Ingressantes e Concluintes (Presenciais e EAD) na Paraíba



Fonte: Instituto Semesp

Os dados da educação superior na Paraíba demonstram de forma clara a necessidade de que ações efetivas precisam ser implementadas para o combate da evasão. No caso específico da UFPB, o trabalho de Prestes & Fialho (2018) demonstra que as ações realizadas pela instituição não diminuíram os índices de evasão, que vem causando prejuízos milionários para a universidade.

2.2. Mineração de Dados Educacionais

A Mineração de Dados Educacionais (MDE) tem como propósito desenvolver ou adaptar métodos e algoritmos de mineração de dados existentes, de tal modo que se possa descobrir conhecimento em bases de dados em contextos educacionais, produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem, tais como Ambientes Virtuais de Aprendizagem, Sistemas Integrados de Gestão Acadêmica, entre outros (COSTA *et al.*, 2013). A utilização de MDE pode, por exemplo, delinear o perfil dos estudantes com maiores riscos de evasão escolar (SARAIVA *et al*., 2019), assim como detectar o impacto de fatores no entorno do

estudante que podem influenciar na evasão escolar, como fatores pessoais, acadêmicos, econômicos, sociais e institucionais (ALBAN & MAURICIO, 2019).

Existem várias linhas de pesquisa na área de MDE, as quais são derivadas diretamente da área de mineração de dados. Segundo Costa *et al.* (2013), as subáreas mais demandadas dentre as propostas por (BAKER; ISOTANI; CARVALHO, 2011) são: Predição (Classificação e Regressão), Agrupamento, Mineração de Relações (Mineração de Regras de Associação, Mineração de Correlações, Mineração de Padrões Sequenciais, Mineração de Causas).

Na subárea de Predição, área de atuação desta pesquisa, a meta é desenvolver modelos que prevejam aspectos específicos dos dados, conhecidos como variáveis preditivas, através da análise e fusão dos diversos aspectos encontrados nos dados, chamados de variáveis preditoras (BAKER; ISOTANI; CARVALHO, 2011). Segundo Costa *et al.* (2013), existem dois tipos de predição: classificação e regressão. Na classificação a variável preditiva é binária ou categórica e na regressão a variável preditiva é contínua. Em ambos os casos, as variáveis preditoras podem ser categóricas ou contínuas. No caso desta pesquisa, nossa variável preditiva é do tipo binária ou categórica, ou seja, assume o valor CANCELADO quando o aluno evade do curso e CONCLUIDO quando o aluno não evadiu e concluiu a sua jornada acadêmica.

2.2.1 Tarefa de classificação e algoritmos de aprendizado de máquina

O Aprendizado de Máquina[1] (AM) tem se tornado uma das formas mais eficazes para a determinação e classificação de padrões em massas de dados nos dias de hoje (TEODORO & KAPPEL, 2020). Os sistemas de AM podem ser classificados de acordo com a quantidade e o tipo de supervisão que recebem durante o treinamento. Existem quatro categorias principais de aprendizado (GERÓN, 2019):

- **Supervisionado**: No aprendizado supervisionado, os dados de treinamento fornecidos ao algoritmo incluem as soluções desejadas, chamadas de rótulos.
- **Não supervisionado**: No aprendizado não supervisionado, os dados de treinamento não são rotulados. O sistema tenta aprender sem um professor.
- **Semissupervisionado**: Alguns algoritmos podem lidar com dados de treinamento

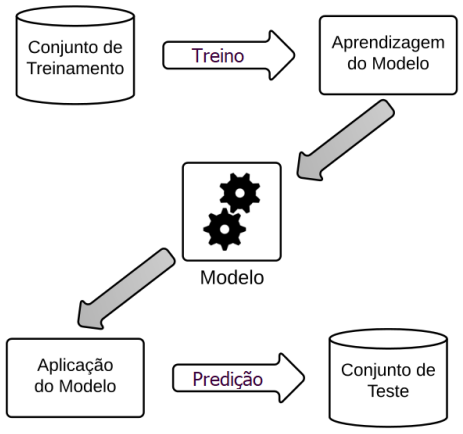
[1] Aprendizado de Máquina é um campo de pesquisa da Inteligência Artificial que estuda o desenvolvimento de métodos capazes de extrair conceitos (conhecimento) a partir de amostras de dados (MITCHELL, 1997).

parcialmente rotulados, uma grande quantidade de dados não rotulados e um pouco de dados rotulados. Isso é chamado de aprendizado semissupervisionado.

- **Por reforço**: O aprendizado por esforço atua de forma diferente dos demais. Nesse caso, o sistema de aprendizado, chamado de agente, pode observar o ambiente e selecionar ou executar ações e obter recompensas em troca — ou penalidades na forma de recompensas negativas. Ele deve aprender por si só qual é a melhor estratégia, chamada de política, para obter o maior número de recompensas ao longo do tempo. Uma política define qual ação o agente deve escolher quando está em determinada situação.

Com a utilização de técnicas e estratégias desta natureza, é possível realizar inferências direcionadas à previsão do risco de evasão escolar de alunos em instituições públicas de ensino superior no Brasil (TEODORO & KAPPEL, 2020). Este trabalho utiliza um classificador supervisionado, que tem como objetivo determinar se uma instância pertence a uma ou mais classes. A Figura 7 representa o funcionamento de um modelo classificador, que tem como entrada um conjunto de treinamento, que consiste em um conjunto de amostras (ou instâncias) de dados onde a classe já é conhecida. A partir desse conjunto de dados, o processo de aprendizagem induz um modelo classificador que em seguida é submetido a um conjunto de amostras de teste, no qual as classes são ocultas e precisam ser preditas a partir do modelo treinado.

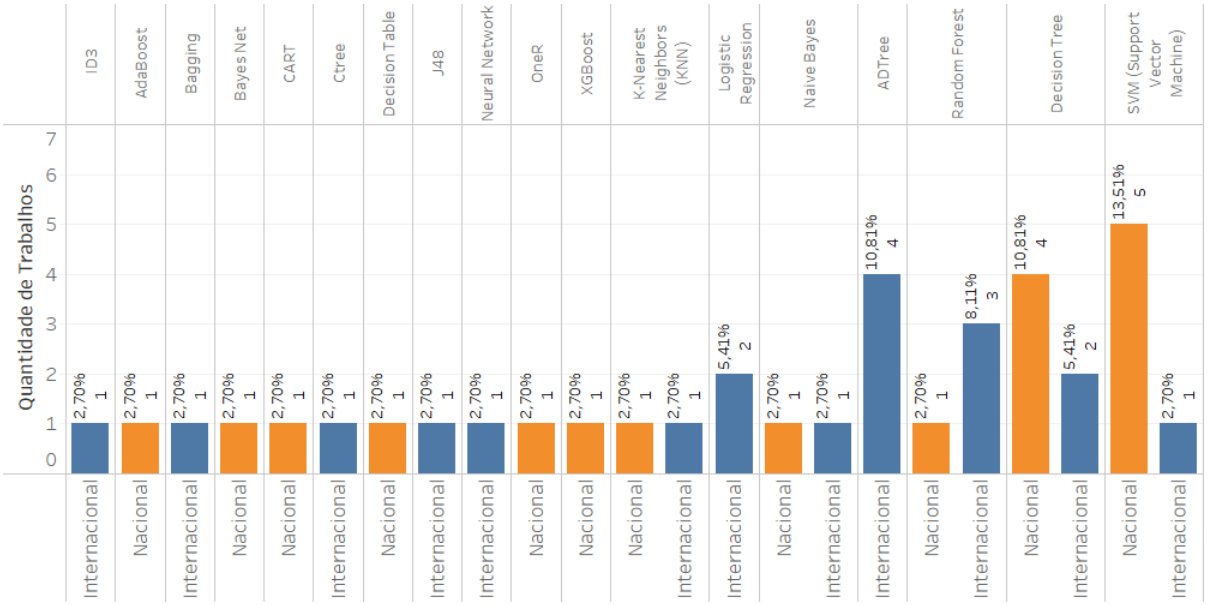
Figura 7 – Representação de um modelo classificador.



Fonte: Adaptado de (COSTA et al., 2013)

Na revisão sistemática da literatura realizada por (SANTOS; SARAIVA; OLIVEIRA, 2021), foram analisados trabalhos de MDE, nacionais e internacionais, que tratam da temática da evasão escolar em instituições e universidades. Dentre as questões de pesquisa investigadas, foram verificados quais os algoritmos de AM dentro do processo de EDM tiveram o melhor desempenho quando aplicados para realizar a predição de evasão escolar, como podemos observar na Figura 8.

Figura 8 – Algoritmos com melhor desempenho na predição de evasão escolar.



Fonte: Extraído de (SANTOS; SARAIVA; OLIVEIRA, 2021)

São inúmeros os algoritmos de AM existentes e utilizados nos diversos trabalhos investigados, porém iremos utilizar como objeto de estudo os três algoritmos com melhores desempenhos que estão presentes nos cenários nacional e internacional (SANTOS; SARAIVA; OLIVEIRA, 2021). Dessa forma os algoritmos Máquina de Vetores de Suporte (*Support Vector Machine* - SVM), Árvore de Decisão (*Decision Tree* - DT) e Floresta Aleatória (*Random Forest* - RF) serão descritos a seguir.

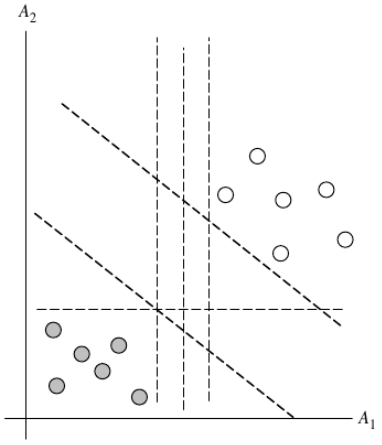
2.2.1.1 Máquina de Vetores de Suporte

Uma SVM é um modelo muito poderoso e versátil de AM capaz de realizar classificações lineares ou não lineares, de regressão e até mesmo detecção de outliers. É um dos algoritmos de AM mais populares e está presente em diversos trabalhos. As SVM são particularmente adequadas para a classificação de conjuntos de dados complexos, porém de pequeno ou médio porte (GERÓN, 2019). Dados de pequeno ou médio porte são conjuntos de dados cujo tamanho e complexidade estão em uma faixa intermediária, entre conjuntos de dados muito pequenos e dados extremamente grandes ou de alta dimensão. Embora não haja uma definição precisa para o que constitui um conjunto de dados de pequeno ou médio porte, esses termos geralmente se referem a dados que podem ser facilmente gerenciados, processados e analisados com recursos computacionais disponíveis em sistemas padrão.

Para exemplificar o funcionamento da técnica de SVM, considere os dados de treinamento apresentados na Figura 9, onde existe um número infinito de hiperplanos que podem separar as

classes. Suponha que os dados sejam relativos à autoavaliação com informações dos alunos, representados por círculos, como seu desempenho na disciplina e satisfação com o curso (variáveis preditoras). Além disso os dados rotulam cada aluno conforme seu status atual no curso (variável preditiva), alunos que não evadiram (círculos brancos) e alunos que evadiram (círculos cinzas). Intuitivamente, a meta do SVM é descobrir qual a melhor forma de separar os dois grupos de alunos.

Figura 9 – Os alunos que não evadiram (brancos) e alunos que evadiram (cinzas).



Fonte: Extraído e adaptado de (COSTA *et al*., 2013)

Nota-se que existe um número infinito de hiperplanos (linha tracejada) que podem separar as classes apresentadas (círculos brancos e círculos cinzas). Então o objetivo do SVM é encontrar qual o melhor hiperplano, ou seja, aquele que maximize a distância entre as instâncias das classes vizinhas.

2.2.1.2 Árvore de Decisão

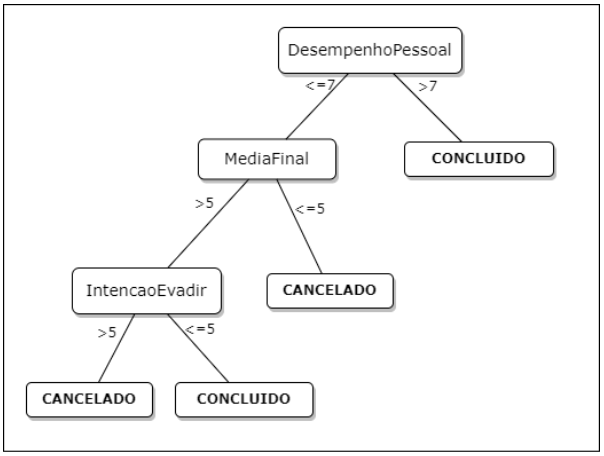
Como as SVM, as Árvores de Decisão (*Decision Tree* - DT) são algoritmos versáteis de AM que podem executar tarefas de classificação e regressão (GERÓN, 2019). Além disso, DT possuem fácil interpretação quanto às suas regras de predição (LOUPPE, 2014). Estas particularidades fazem desse modelo um algoritmo de aprendizado popular e muito difundido para a predição da evasão escolar (PEREIRA & ZAMBRANO, 2017; SUKHBAATAR; OGATA; USAGAWA, 2018; TEODORO & KAPPEL, 2020).

Árvores de Decisão para classificação são modelos estatísticos que utilizam treinamento supervisionado para predição dos dados. Ou seja, no conjunto de treinamento as variáveis preditivas Y são conhecidas. Uma DT possui uma estrutura de árvore, onde cada nó interno (não-folha), pode ser entendido como um atributo de teste, e cada nó-folha (nó-terminal) possui um rótulo de classe (HAN & KAMBER, 2000). O nó de mais alto nível numa DT é chamado de nó-raiz. Após aprendido os parâmetros do modelo, a DT irá classificar uma instância de acordo com o caminho que satisfaça

as condições desde o nó-raiz até o nó folha, ao final do processo a instância será rotulada de acordo com o nó-folha (COSTA *et al.*, 2013).

Um exemplo de DT que classifica alunos de graduação entre CANCELADO e CONCLUIDO de acordo com os atributos "Intenção de Evadir", "Média Final" e "Desempenho Pessoal" pode ser visto na Figura 10.

Figura 10 – DT que classifica CANCELADO e CONCLUIDO.



Fonte: Figura do autor

É possível observar, no exemplo, que se o desempenho pessoal do aluno for maior que 7 o aluno é considerado como CONCLUIDO, caso contrário é preciso avaliar a média final do aluno, sendo menor ou igual a 5 esse aluno é considerado CANCELADO. Caso esse aluno tenha média final maior que 5, porém ele tenha a intenção de evadir ele será classificado como CANCELADO, caso contrário será definido como CONCLUIDO.

Segundo Alvarenga Júnior (2018), dado um problema de classificação, representado por um conjunto de dados $(x_{1:n}, y_{1:n})$ em que x_i são os atributos e y_i o rótulo, uma DT particiona o espaço dos atributos de modo recursivo durante o crescimento da árvore, dividindo (*split*) uma região R_j em subregiões e atribuindo um valor de saída y_i à cada uma destas. Para o caso de um problema de classificação, o valor de saída y_i pode ser definido a partir da moda ou das probabilidades (para cada classe), tomadas com base nos pontos que habitam a respectiva R_j. Durante o crescimento da árvore, o particionamento ocorre até que cada ponto do conjunto de treinamento esteja sozinho em uma região R_j, ou que esta possua um grau máximo de pureza, ou então, até que um critério de parada ocorra. Ao final do crescimento da árvore, haverá um número de $j = 1, 2, \dots, j$ regiões disjuntas. Então a regra de predição é dada por,

$$T(x;\Theta) = \sum_{j=1}^{1} y_i \, I(x_i \, \epsilon \, R_j), \quad (1)$$

onde x é o vetor com os atributos, $\Theta = \{R_j, y_i\}_1^j$, y_i é a saída atribuída à região R_j, $I(\,)$ é uma função de indicação que retorna 1 se $x_i \, \epsilon \, R_j$, e 0 caso contrário.

2.2.1.3 Floresta Aleatória

Floresta Aleatória (*Random Forest – RF)* é um modelo baseado em árvores de decisão, que lida bem com conjunto de dados de alta dimensão (HASTIE *et al.*, 2009). Dados de alta dimensão se referem a conjuntos de dados que possuem muitas características (variáveis ou atributos) em relação ao número de observações (exemplos ou registros). Em outras palavras, são conjuntos de dados onde o número de características é substancialmente maior do que o número de pontos de dados. O algoritmo RF não é apenas utilizado para realizar tarefas de classificação, mas também para regressão, estudo de importância, seleção de variáveis, e detecção de *outlier*[2] (VERIKAS; GELZINIS; BACAUSKIENE, 2011).

Segundo Alvarenga Júnior (2018), RF são modelos do tipo *ensemble methods*[3], que combinam a predição de um conjunto de árvores de decisão, para obter uma única resposta como saída, que tende a apresentar melhor desempenho que as obtidas com cada árvore do modelo em separado, devido à redução de variância. A equação a seguir descreve a saída para este modelo,

$$\hat{f}_{fa}^{B}(x_i) = \frac{1}{B}\sum_{b=1}^{B} T(x_i, \Theta_b), \quad (2)$$

onde, B é o número total de árvores, $T(\,)$ representa a resposta de uma árvore b para um vetor de entrada x_i, e Θ_b representa os parâmetros desta árvore.

[2] Um *outlier* é um ponto de dado que é visivelmente diferente do resto. Eles representam erros na medição, má coleta de dados ou simplesmente mostram variáveis não consideradas na coleta dos dados.

[3] *Ensemble methods*, também conhecidos como aprendizado baseado em comitês ou aprendizado de múltiplos sistemas classificadores, treinam múltiplas hipóteses para resolver o mesmo problema.

RF possui algumas características desejáveis como: é tão rápido quanto outros algoritmos de *Bagging*[4] ou *Boosting*[5]; possui desempenho competitivo com outros métodos; é relativamente robusto a ruídos ou *outliers*; e é facilmente paralelizável (BREIMAN, 2001). RF possui dois processos aleatórios para a construção da floresta: A criação de conjuntos de dados *bootstrap*, que consiste em uma técnica de amostragem de dados, e o Subconjunto Aleatório de Atributos, que altera o processo de treinamento de uma árvore de decisão (PONTE; CAMINHA; FURTADO. 2020).

No primeiro processo, considere o conjunto de dados de treinamento $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$, onde x_i é um vetor de atributos da amostra i que possui dimensionalidade n, y_i é o valor alvo que se deseja prever e m é o número de amostras de treino de D. No método *bootstrap*, selecionam-se aleatoriamente amostras de D, de tal forma a construir um novo conjunto de dado B_j do mesmo tamanho de D. Essa amostragem é feita com reposição, ou seja, quando é selecionada uma nova amostra ela já pode ter sido selecionada antes. Isso fará com que essas instâncias com repetição tenham maior peso no processo de treinamento de uma determinada árvore de decisão. É importante destacar que existirão amostras fora do conjunto B_j, denominadas amostras *Out-of-Bag* (OOB). Na RF, o procedimento *bootstrap* é realizado L vezes, um para cada árvore criada, resultando nos conjuntos $B_1, B_2, \dots, B_L$ (PONTE; CAMINHA; FURTADO. 2020).

No segundo processo aleatório, Subconjunto Aleatório de Atributos, o procedimento acontece de tal maneira que, toda vez que se cria um nó na árvore, seleciona-se aleatoriamente um subconjunto dos atributos para serem candidatos na escolha do atributo daquele nó. Normalmente escolhe-se o valor de $\sqrt{n}$ para ser o tamanho do subconjunto de atributos a serem selecionados aleatoriamente.

2.2.2 Seleção de atributos

Segundo Lima (2016), a seleção de atributos (*feature selection)* é uma etapa crucial para descoberta de conhecimento em uma base de dados. Intuitivamente, pode se pensar que quanto

---

[4] O *Bagging (Bootstrap Aggregating)* é um método que gera um conjunto de dados por amostragem bootstrap dos dados originais.

5 No *Boosting,* de forma semelhante ao Bagging, cada classificador é treinado usando um conjunto de treinamento diferente.

maior o número de atributos em um conjunto de dados, maior o poder discriminatório e consequentemente maior a acurácia do classificador. Entretanto na prática esse comportamento não é verdadeiro. Diversas pesquisas nessa área indicam que muitos atributos irrelevantes podem introduzir ruídos nos dados, confundindo o algoritmo de aprendizagem e ocasionando erros na classificação. Além disso, muitos atributos desnecessários podem fazer com que os algoritmos de aprendizagem tenham dificuldade em extrair informações que sejam realmente significantes e relevantes para classificação.

Os métodos de seleção de atributos que são usados rotineiramente na classificação podem ser divididos em três categorias (GUYON *et al*., 2008; BOLÓN-CANEDO; SÁNCHEZ-MAROÑO; ALONSO-BETANZOS, 2013):

1. ***Filters***: os métodos de filtro usam a classificação de atributos como a métrica de avaliação para a seleção de atributos. Geralmente, os recursos são classificados com base em suas pontuações em vários testes estatísticos para sua correlação com a classe. Os recursos com pontuação abaixo de um determinado limite são removidos, enquanto os recursos com pontuação acima dele são selecionados. Depois que um subconjunto de recursos é selecionado, ele pode ser apresentado como uma entrada para o algoritmo classificador escolhido. Ao contrário dos outros métodos de seleção de atributos (*wrapper* e *embedded*), os métodos de filtro são independentes/separados do algoritmo do classificador.

2. ***Wrappers***: Em contraste com os métodos de filtro, os métodos *wrapper* usam o desempenho do algoritmo classificador escolhido como uma métrica para auxiliar na seleção do melhor subconjunto de atributos. Assim, os métodos *wrapper* identificam o conjunto de características de melhor desempenho para o algoritmo classificador escolhido.

3. ***Embedded***: O método incorporado é bastante semelhante aos métodos *wrappers*, pois também são usados para otimizar o desempenho do algoritmo classificador. A diferença para os métodos *wrappers* é que uma métrica de construção de modelo intrínseca é usada durante o aprendizado, ou seja, durante a etapa de treinamento, o classificador ajusta seus parâmetros internos e determina os pesos/importâncias apropriados para cada atributo, para produzir a melhor precisão de classificação. Portanto, a busca pelo subconjunto ótimo de atributos e a construção do modelo em um método embutido é combinada em uma única etapa.

2.2.3 Dados de treinamento e teste

Uma forma simples de validar a capacidade de generalização de um modelo a partir de um conjunto de dados é dividir os dados em conjuntos de treinamento e teste. Dessa forma, é possível avaliar o comportamento do modelo em dados que não foram utilizados na etapa de treinamento, estipulando a estimativa de erro do modelo no conjunto de dados desconhecidos. Esse método é denominado *Holdout*. O problema com o *Holdout* é que não é possível afirmar que o subconjunto de treinamento é representativo para o conjunto total da base de dados (GÉRON, 2019).

Para evitar "desperdiçar" muitos dados de treinamento em conjuntos de validação e gerar modelos representativos para todos os dados, uma técnica comum é utilizar a validação cruzada: o conjunto de treinamento é dividido em subconjuntos complementares e cada modelo é treinado com uma combinação diferente desses subconjuntos e validado em relação às partes restantes. Uma vez selecionados o tipo de modelo e os hiperparâmetros, um modelo final é treinado com a utilização desses hiperparâmetros no conjunto completo de treinamento e o erro generalizado é medido no conjunto de testes (GÉRON, 2019).

2.2.4 Balanceamento de dados

A maioria dos algoritmos de AM assume que os seus conjuntos de dados de treino estão balanceados, isto é, que o volume de amostras está distribuído de igual forma por cada categoria que está a ser analisada. No entanto, isto nem sempre acontece no mundo real, ou seja, pode acontecer que o número de instâncias correspondente a uma determinada classe é muito diferente do número de instâncias correspondente a outra classe. Neste caso, estamos perante um problema de desequilíbrio de classes, o que é algo bastante comum e constitui por vezes um obstáculo para a obtenção de bons índices de classificação por parte dos algoritmos (BATISTA et al., 2004).

Neste trabalho, avaliamos diferentes métodos de subamostragem e sobreamostragem para balancear a distribuição de classes nos dados de treinamento. Dois desses métodos, sobreamostragem aleatória e subamostragem aleatória, são métodos não heurísticos que foram inicialmente incluídos nesta avaliação como métodos de linha de base. A partir desses métodos, foi possível avaliar como os algoritmos se comportavam quanto as métricas de avaliação. Esses métodos são definidos como (BATISTA et al., 2004):

- **Sobreamostragem Aleatória**: a sobreamostragem aleatória é um método não heurístico que visa equilibrar a distribuição de classes por meio da replicação aleatória de exemplos de classes minoritárias.

- **Subamostragem Aleatória**: a subamostragem aleatória também é um método não heurístico que visa equilibrar a distribuição de classes por meio da eliminação aleatória de exemplos de classes majoritárias.

Para além destes, ainda é possível destacar outros métodos mais complexos, nomeadamente:

- **SMOTE** (*Synthetic Minority Over-sampling Technique*): algoritmo de sobreamostragem heurístico. Gera exemplos ou amostras artificiais das classes minoritárias através das trocas entre as instâncias que se encontram mais próximas. É um dos métodos mais atualizados atualmente, tendo servido de ponto de partida para a criação de outros métodos de sobreamostragem. De realçar que o *overfitting* é evitado e mistura o espetro da classe minoritária com o da classe maioritária (CHAWLA et al., 2002).
- **ADASYN** (*Adaptive Synthetic Over-Sampling*): este algoritmo é considerado uma extensão do SMOTE, onde existe a particularidade de serem criados mais exemplos sintéticos na área onde os exemplos da classe minoritária têm menos ocorrências (HE et al., 2008).

2.2.5 Métricas de avaliação de resultados

Para os testes que foram efetuados neste trabalho, recorrendo aos diferentes algoritmos de aprendizado de máquina, foram utilizadas várias métricas de avaliação de resultados. Quando se analisam os resultados de um algoritmo de aprendizado de máquina, em especial relacionados com sistemas de deteção de evasão escolar, todas as métricas estão relacionadas com o número de previsões:

- **Verdadeiras Positivas (VP)**: Quando um aluno CANCELADO é classificado como CANCELADO, ou seja, o aluno que evadiu do curso é classificado corretamente como evadido;
- **Verdadeiras Negativas (VN)**: Quando um aluno CONCLUIDO é classificado como CONCLUIDO, ou seja, quando um aluno que não evadiu do curso é classificado corretamente como não evadido;
- **Falsas Positivas (FP)**: Quando um aluno CONCLUIDO é classificado como CANCELADO, ou seja, o aluno que não evadiu é classificado como evadido gerando um resultado falso positivo;
- **Falsas Negativas (FN)**: Quando um aluno CANCELADO é classificado como

CONCLUIDO, ou seja, o aluno que evadiu é classificado como não evadido gerando um resultado falso negativo.

Estas previsões são também habitualmente representadas sob a forma de uma matriz denominada de matriz de confusão (Figura 11), onde o "Sim" seria o CANCELADO (evadiu do curso) e o "Não" CONCLUIDO (não evadiu do curso).

Figura 11 – Matriz de confusão.

		Detectada	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Figura do autor

Esta matriz tem como objetivo avaliar ou interpretar os resultados obtidos pelos algoritmos de classificação. Nesta matriz é obtida uma relação de verdadeiro ou falso entre o eixo dos dados previstos e o eixo dos dados reais observados.

2.2.5.1 Acurácia

A acurácia é uma métrica calculada a partir da divisão entre o número de previsões verdadeiras (VP e VN) e o número total de previsões. A acurácia é mais aconselhada para conjuntos de dados balanceados (JOSHI, 2020).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (3)$$

2.2.5.2 Precisão

O valor de Precisão é calculado a partir da divisão entre as previsões positivas verdadeiras (VP) e o total de previsões positivas (VP + FP). Um resultado elevado, indica a presença de um valor baixo de previsões positivas falsas (FP) (JOSHI, 2020).

$$Precisão = \frac{VP}{VP + FP} \quad (4)$$

2.2.5.3 Recall

A métrica *Recall* é calculada a partir da divisão das previsões positivas verdadeiras (VP) e a soma das previsões positivas verdadeiras (VP) e as previsões negativas falsas (FN). Analisando as possíveis variações de resultados, é possível verificar que um alto valor de *Recall* indica a presença de um baixo número de FN (JOSHI, 2020).

$$Recall = \frac{VP}{VP + FN} \quad (5)$$

2.2.5.4 F-Measure ou F-Score

O *F-Measure* é um valor ponderado que resulta da divisão do dobro da multiplicação entre a *Recall* e a Precisão com a soma destas duas métricas. É mais utilizada para conjuntos de dados não balanceados (JOSHI, 2020).

$$F - Measure = \frac{2 \times (Recall \times Precisão)}{Recall + Precisão} \quad (6)$$

2.3. Trabalhos Relacionados

Vários trabalhos podem ser encontrados na literatura relacionados a abordagens preditivas para identificar automaticamente estudantes que apresentam um grande risco de evasão escolar no ensino superior. Na presente análise dos trabalhos existentes relacionados à pesquisa desenvolvida neste estudo, a discussão foi organizada com base em Revisões Sistemáticas da Literatura (RSL) realizadas por outros autores, seguidas de trabalhos selecionados ad hoc.

2.3.1 Estado da arte

Nos últimos anos, foram realizadas várias Revisões Sistemáticas da Literatura (RSL) que visam identificar o estado da arte em relação ao tema da predição da evasão escolar. A maioria dessas revisões revela uma série de características em comum, como a utilização de técnicas de Mineração de Dados Educacionais (MDE) com algoritmos de classificação para identificar precocemente os alunos com maiores probabilidades de evadir.

Figura 12 – RSL por escopos 2(a), estados 2(b), modalidades 3(a) e ensino 3(b).

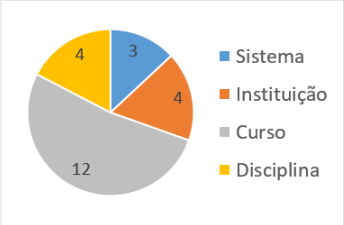


Figura 2(a). Escopos de evasão.

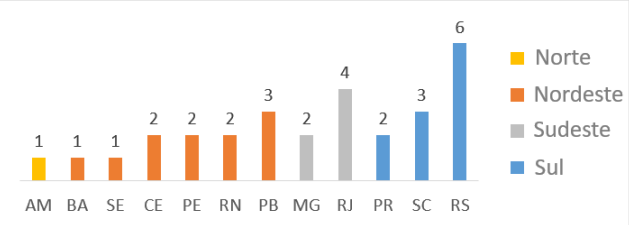


Figura 2(b). Estados dos autores.

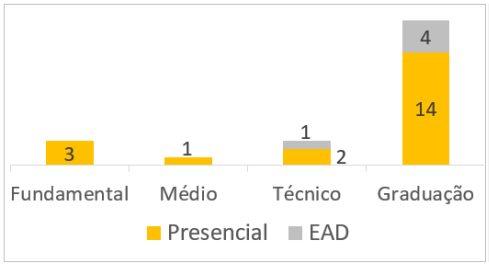


Figura 3(a). Níveis e modalidades de ensino.

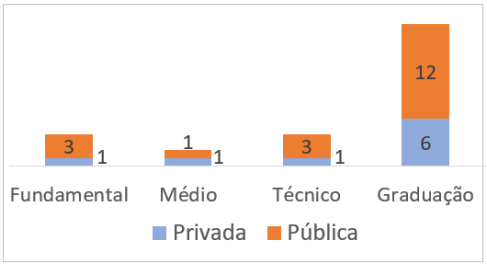


Figura 3(b). Níveis e redes de ensino.

Fonte: Extraído de (COLPO *et al.*, 2020)

O trabalho de (COLPO et al., 2020) foi considerado por apresentar o levantamento do estado da arte em trabalhos relevantes publicados no Brasil. Na Figura 12, é possível observar que uma quantidade significativa dos trabalhos analisados pela RSL é realizada na região Nordeste e trata da evasão do curso. Também é possível observar que na maioria dos trabalhos são analisados dados da graduação da rede pública de ensino superior. Dessa forma, as estatísticas obtidas pela RSL estão alinhadas com este trabalho.

Este trabalho, ao utilizar os dados do instrumento de autoavaliação da UFPB, diferencia-se do estado da arte apresentado por (COLPO et al., 2020), pois são tipos de dados que não foram identificados em nenhum trabalho da RSL. Como pode ser observado na Figura 13, dentre os trabalhos selecionados, só foram identificados os tipos de dados: Acadêmicos, Econômicos, Sociais, Interacionais, Sentimentos e Temporais.

Figura 13 – Artigos por níveis e tipos (a) ou tamanhos dos conjuntos (b) de dados.

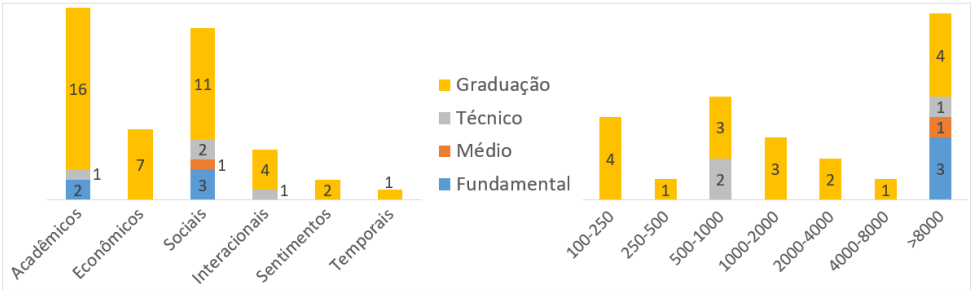


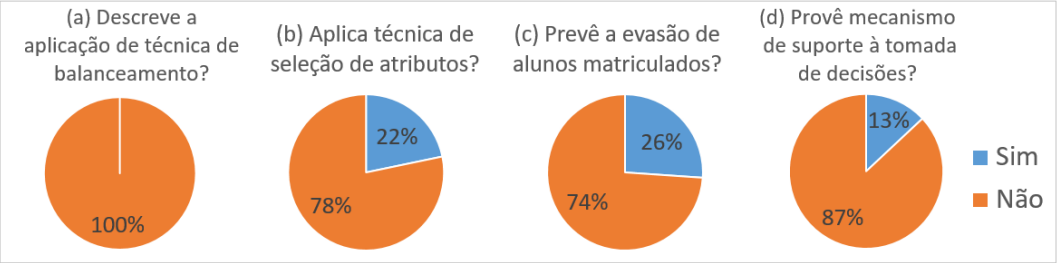
Figura 4(a). Tipos de dados e níveis.

Figura 4(b). Tamanhos dos conjuntos de dados e níveis.

Fonte: Extraído de (COLPO *et al.*, 2020)

Além de se diferenciar quanto aos tipos de dados, este trabalho apresenta características importantes no processo de desenvolvimento de um modelo de classificação, que não foram identificadas nos trabalhos analisados pela RSL ou foram identificadas apenas na minoria (Figura 14). Esta pesquisa descreve a aplicação de técnicas de balanceamento por subamostragem e sobreamostragem (Seção 3.1.4 e Seção 4.4.3), aplica técnica de seleção de atributos (Seção 4.4.1), prevê a evasão de alunos matriculados (Seção 4.6) e provê mecanismos de suporte à tomada de decisões (Seção 4.6 e Seção 5).

Figura 14 – Proporções de atendimento a determinadas características técnicas.

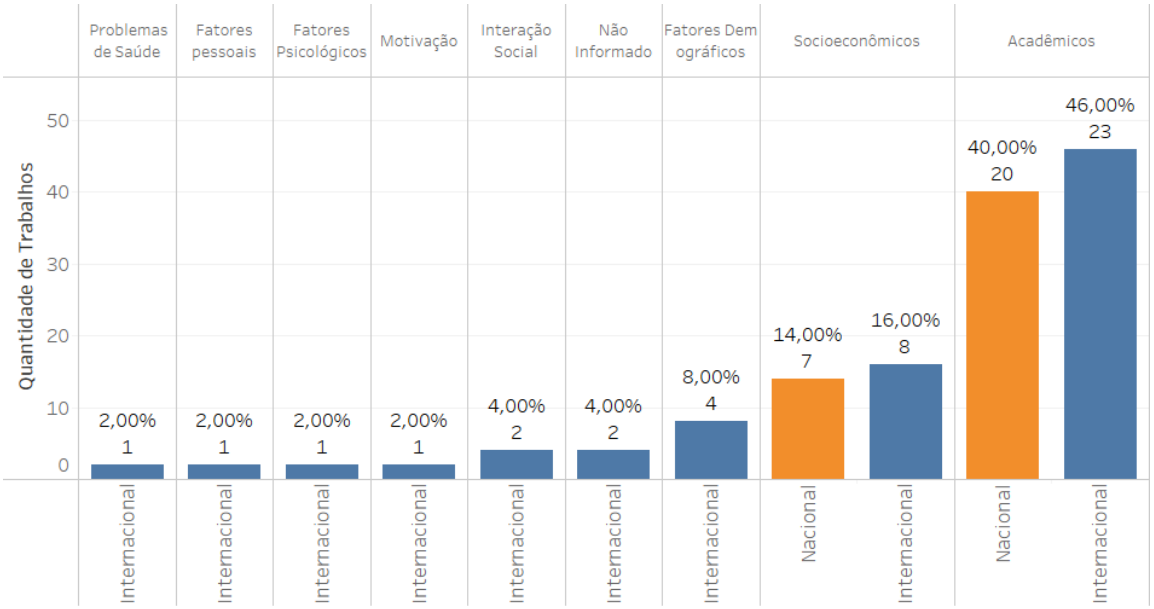


Fonte: Extraído de (COLPO *et al.*, 2020)

Na revisão realizada por (SANTOS; SARAIVA; OLIVEIRA, 2021), é realizada uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar, identificando o estado da arte em trabalhos nacionais e internacionais de relevância. Essa RSL foi determinante para as tomadas de decisões que se seguem.

A RSL realizada por (SANTOS; SARAIVA; OLIVEIRA, 2021) destaca que a ferramenta/biblioteca de aprendizado de máquina mais comumente utilizada pelos trabalhos, nos últimos anos, é a biblioteca *Scikit-learn* para a linguagem de programação Python. Por ser uma ferramenta simples e eficiente para análise preditiva de dados, acessível a todos, reutilizável em vários contextos, com código aberto e comercialmente utilizável, ela também é utilizada nesta pesquisa. A RSL também identifica que 40 algoritmos diferentes são utilizados nos diversos trabalhos analisados, porém foi verificado que os algoritmos SVM, DT e RF foram os que obtiveram o melhor desempenho na tarefa de predição de evasão escolar, conforme a Figura 8. Dessa forma, é natural que esses algoritmos sejam utilizados nesta pesquisa, visando obter o máximo desempenho na predição da evasão escolar a partir dos dados da autoavaliação da UFPB.

Figura 15 – Tipos de dados.



Fonte: Extraído de (SANTOS; SARAIVA; OLIVEIRA, 2021)

Também é importante ressaltar que nenhum dos trabalhos analisados utilizou dados de autoavaliação, sendo os tipos de dados acadêmicos e socioeconômicos predominantes nos cenários nacional e internacional (Figura 15). Essa é mais uma característica que diferencia este trabalho e destaca a contribuição original para o avanço do conhecimento na área.

2.3.2 Modelos de predição de evasão escolar na educação superior

No contexto da pesquisa de desenvolvimento de um modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação, o estudo realizado por (RAFIQ; RABBI; AHAMMAD, 2021) sobre a evasão escolar na Universidade de Bangladesh tem relevância significativa. Embora focado em uma instituição diferente, esse estudo fornece informações valiosas para abordar a preocupação com os índices crescentes de evasão escolar. A coleta de dados da base da universidade e de pesquisas com estudantes resultou na seleção de 17 variáveis relacionadas a dados acadêmicos, socioeconômicos e pessoais, que podem servir como ponto de partida para o desenvolvimento do modelo de predição na Universidade Federal da Paraíba. A abordagem de classificação supervisionada foi empregada, utilizando os algoritmos *One-vs-Rest* e *Random Forest* (RF) com a biblioteca *Scikit-learn* em Python. A utilização de 70% dos dados para treinamento e 30% para testes, juntamente com o desempenho superior do algoritmo RF, com uma precisão de 0,9745 e pontuação AUC de 0,98, destaca sua relevância para a construção do modelo de predição na instituição em questão.

O trabalho realizado por (LOTTERING; HANS; LALL, 2020) apresenta relevância significativa. Nesse estudo, foram aplicadas técnicas de Mineração de Dados Educacionais (MDE) e aprendizado de máquina para identificar alunos com risco de evasão em um curso de graduação em uma universidade de tecnologia na África do Sul. Os dados utilizados no estudo foram as notas finais obtidas pelos alunos ao longo dos anos letivos, sendo extraídas várias variáveis do banco de dados da universidade e derivadas algumas variáveis por meio do processo de Extração, Transformação e Carga (ETL) para enriquecer o conjunto de dados. No total, foram selecionados 19 atributos. Para lidar com o desbalanceamento dos dados, foi realizada uma subamostragem do conjunto de maior registro equivalente à amostra menor, resultando em um conjunto de dados reduzido com 1.156 registros. Em seguida, foram aplicados os algoritmos de aprendizado de máquina de classificação supervisionada *Naive Bayes, Decision Tree* (DT), *Support Vector Machine* (SVM), *Nearest Neighbor* e *Random Forest* (RF). Utilizando 75% dos dados para treinamento e 25% para teste, e avaliando o desempenho dos classificadores por meio das métricas de acurácia, recall, precisão e F-Measure, o algoritmo SVM obteve o melhor desempenho, alcançando uma pontuação F-Measure de 99,32%. Esses resultados destacam a relevância desse estudo para a pesquisa de desenvolvimento do modelo de predição de evasão escolar na Universidade Federal da Paraíba, fornecendo sugestões sobre técnicas de modelagem e algoritmos que podem ser explorados na construção do referido modelo.

A pesquisa realizada por (SANTOS; MARTINS; PLASTINO, 2021) é de relevância significativa. Nesse estudo, foram aplicadas técnicas de Mineração de Dados Educacionais (MDE), mais especificamente, técnicas de classificação, para investigar a viabilidade de prever a evasão ou formatura dos alunos utilizando apenas dados de desempenho acadêmico. Os dados utilizados foram obtidos de uma instituição de ensino superior brasileira, sendo selecionados 7 atributos para a análise. Com base nesses atributos, foram construídos 10 modelos de dados correspondentes aos diferentes semestres cursados, variando do 1º ao 10º semestre. Utilizando o algoritmo *Decision Tree* (DT), foram obtidos resultados satisfatórios de acurácia, variando de 79,31% (3º semestre) a 98,25% (9º semestre). Alguns modelos apresentaram uma precisão de 100% para a classe de Evasão, indicando que todas as previsões de evasão estavam corretas. No caso do modelo do 3º semestre, 80,35% das previsões de evasão foram corretas. Em relação à classe de Formatura, a precisão variou de 75,68% a 97,22%. Observou-se que alguns modelos alcançaram um recall de 100% para a classe de Formatura, indicando que todas as formaturas foram corretamente previstas. No entanto, no modelo do 3º semestre, apenas 68,57% das formaturas foram corretamente previstas. Quanto à classe de Evasão, o recall variou de 82,22% a 96,43%. É importante destacar que o poder preditivo

dos modelos aumentou à medida que os semestres avançaram, devido ao fato de que os modelos dos primeiros semestres tinham menos atributos (disciplinas) em comparação aos modelos dos semestres mais avançados. Esses resultados contribuem para a pesquisa de desenvolvimento do modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação, fornecendo visões sobre a relação entre o desempenho acadêmico dos alunos e a probabilidade de evasão ou formatura ao longo dos diferentes semestres.

O trabalho de (MANRIQUE *et al.*, 2019) desempenha um papel relevante ao explorar três abordagens distintas para a predição da evasão em uma instituição de ensino superior brasileira. A primeira abordagem consiste em um modelo de previsão baseado em recursos globais, que utiliza variáveis genéricas para representar um estudante, independentemente da instituição ou curso, permitindo a aplicação da predição em qualquer curso. A segunda abordagem adota dados específicos do curso, possibilitando a predição apenas para o curso em questão. Ambas as abordagens utilizam algoritmos de Aprendizado de Máquina, como *Gradient Boosting Tree*, *Support Vector Machine* (SVM), *Random Forest* (RF) e *Naive Bayes*. Já a terceira abordagem analisa os dados dos alunos de forma temporal ao longo do curso, empregando o algoritmo *K-Nearest Neighbors*. Os resultados obtidos indicam que o modelo baseado em recursos globais associado ao algoritmo RF alcançou os melhores resultados em termos de métricas de avaliação, como acurácia, recall, precisão e F-Measure. Essas descobertas contribuem para a pesquisa e desenvolvimento do modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação, fornecendo informações sobre diferentes abordagens e algoritmos que podem ser aplicados para melhor compreender e antecipar a evasão escolar em diversas instituições e cursos.

Este trabalho avança frente aos trabalhos relacionados, principalmente com relação ao contexto dos dados utilizados. A pesquisa de predição de evasão escolar com base em dados de um instrumento único de autoavaliação de cursos de graduação pode trazer avanços científicos significativos ao proporcionar uma compreensão mais profunda dos fatores e padrões que levam à evasão. Ao utilizar técnicas de mineração de dados e modelos preditivos, essa pesquisa pode identificar correlações ocultas e fatores de risco, fornecendo informações valiosas para o desenvolvimento de estratégias de intervenção mais eficazes. Esses avanços científicos contribuem para o campo da educação, podendo melhorar a retenção dos alunos e ajudando a desenvolver abordagens mais personalizadas e baseadas em evidências para enfrentar o desafio da evasão escolar.

3. PROPOSTA DE MODELO PREDITOR DE EVASÃO ESCOLAR

Este capítulo apresenta o processo de construção da proposta de modelo preditor que visa apoiar os atores do processo educacional no combate a evasão escolar. O processo de mineração de dados educacionais foi dividido em seis etapas, baseado na metodologia CRISP-EDM apresentada na Seção 1.3.

3.1. Fase 1: Entendimento do domínio

A primeira etapa do processo foi entender o domínio da aplicação, ou seja, qual o contexto dos dados educacionais que seriam minerados e a sua correlação com a evasão escolar. Os dados brutos obtidos são frutos da autoavaliação dos cursos de graduação que é realizada pelos alunos da UFPB a cada início de semestre. A autoavaliação é realizada com base no semestre anteriormente cursado, de forma compulsória e como pré-requisito para a obtenção de matrícula para o semestre subsequente.

O instrumento possui quatro dimensões para avaliação (COSTA & DIAS, 2020):

1. **Discente** – o discente realiza uma autoavaliação do seu desempenho para cada disciplina cursada;
    a. Por favor, dê uma nota (de 0 - muito ruim, a 10 - muito bom) para SEU desempenho pessoal na disciplina em termos de comprometimento e motivação.
2. **Disciplina** – o aluno avalia, de acordo com a sua percepção, qual o nível de importância e a dificuldade do conteúdo de cada disciplina cursada;
    a. Na sua percepção, qual o nível de importância (de 0 - sem importância, a 10 - extremamente importante) das disciplinas cursadas para o seu curso?
    b. Na sua percepção, qual o nível de dificuldade DOS CONTEÚDOS das disciplinas cursadas (de 0 - muito fácil, a 10 - muito difícil)?
3. **Docente** – é realizada uma avaliação, na visão do discente, sobre a necessidade de ajustes pelos docentes nas disciplinas ministradas e qual a satisfação geral com o desempenho dele;
    a. Considerando os itens abaixo, assinale quais deles cada professor PRECISA

AJUSTAR?

i. Cumprimento do plano de curso;

ii. Relacionamento com a turma;

iii. Comparecimento às aulas;

iv. Cumprimento do horário de início e de término das aulas;

v. Atualização dos conteúdos;

vi. Clareza na exposição dos conteúdos;

vii. Disponibilidade para atendimento fora da sala de aula;

viii. Qualidade da bibliografia;

ix. Qualidade das avaliações.

b. Por favor, aponte sua satisfação geral (de 0 - totalmente insatisfeito, a 10 - totalmente satisfeito) com o desempenho de cada professor.

4. **Curso** – o aluno pode responder se recomendaria o curso para alguém e se tem a intenção de evadir.

a. Considerando a experiência com seu curso até esse último período, a probabilidade de você recomendar esse curso para um amigo ou parente próximo é (de 0 - muito improvável, a 10 - muito provável).

b. Seu interesse em sair de curso (mudar de curso na UFPB ou para outra instituição, parar de estudar etc.) no momento atual é (de 0 - muito baixo, a 10 - muito alto).

Um aspecto relevante para o desenvolvimento do modelo é que esse instrumento começou a ser aplicado nesse formato a partir do primeiro semestre de 2017. Dentro desse contexto, os dados analisados foram restritos aos alunos matriculados a partir do mencionado semestre. Essa abordagem foi adotada devido ao fato de que alunos que não realizaram avaliações ao longo de todo o seu ciclo acadêmico poderiam gerar dados enviesados, prejudicando a eficácia do modelo. Essa constatação foi obtida ao seguir o ciclo do CRISP-EDM e realizar refinamentos no modelo.

No ano de 2021, o instrumento de autoavaliação passou por uma atualização, incorporando cinco novas perguntas que visavam aprofundar a compreensão do discente em relação ao seu curso. Essas questões abordavam aspectos como a capacidade de aprendizado, motivação, percepção sobre o valor do curso no mercado de trabalho, formação de competências profissionais e a satisfação

geral com a qualidade do curso. Embora essas novas variáveis pudessem ter uma possível relevância para o estudo da evasão, devido ao escasso histórico de dados disponíveis, optou-se por não as incluir na elaboração do modelo. Todas as demais variáveis que faziam parte do modelo de 2017 foram mantidas na versão ajustada de 2021. Essa decisão foi tomada com o intuito de manter a consistência e a comparabilidade dos resultados ao longo do tempo.

3.2. Fase 2: Entendimento dos dados educacionais

Nesta etapa, foi primordial entender os dados brutos obtidos de forma a realizar uma preparação adequada dos dados na etapa subsequente.

3.2.1 Conjunto de dados

Na tabela 2, podemos observar as informações sobre as variáveis do conjunto de dados brutos que foram utilizadas nesta etapa da metodologia.

Tabela 2 – Conjunto de dados brutos.

Variável	Descrição	Tipo de Variável
ANO	Ano de referência que está sendo avaliado.	Discreta
PERIODO	Período de referência que está sendo avaliado.	Discreta
MATRICULA	Número da matrícula do aluno.	Contínua
STATUS_DISCENTE (variável alvo)	Situação atual do aluno no momento da extração dos dados.	Categórica
CENTRO	Nome do Centro do curso do aluno.	Categórica
DEPARTAMENTO	Nome do Departamento do curso do aluno.	Categórica
CODIGO	Código da disciplina.	Contínua
DISCIPLINA	Título da disciplina.	Categórica
CODIGO_TURMA	Código da turma cursada.	Contínua
HORARIO	Horário da turma.	Categórica
LOCAL	Local da turma.	Categórica

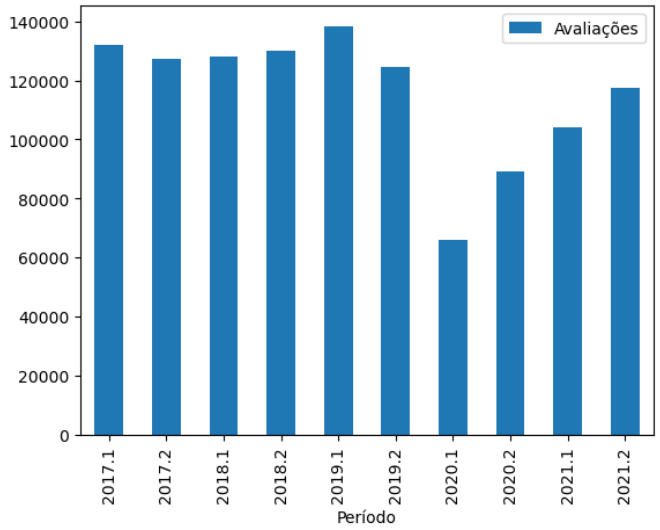
CURSO	Nome do curso do aluno.	Categórica
TURNO	Turno do curso.	Categórica
MEDIA_FINAL	Média final obtida na disciplina.	Discreta
SITUACAO_MATRICULA	Status obtido na disciplina.	Categórica
QUARTA_PROVA	Informa se a disciplina possui quarta prova.	Discreta
FALTAS	Quantidade de faltas do aluno durante o período.	Contínua
1.1.1	Nota (de 0 - muito ruim, a 10 - muito bom) para o desempenho pessoal na disciplina em termos de comprometimento e motivação.	Discreta
2.1.1	Nível de importância (de 0 - sem importância, a 10 - extremamente importante) das disciplinas cursadas.	Discreta
2.2.1	Nível de dificuldade dos conteúdos das disciplinas cursadas (de 0 - muito fácil, a 10 - muito difícil).	Discreta
3.1.1.A	Professor precisa ajustar o cumprimento do plano de curso (sim ou não).	Discreta
3.1.1.B	Professor precisa ajustar o relacionamento com a turma (sim ou não).	Discreta
3.1.1.C	Professor precisa ajustar o comparecimento às aulas (sim ou não).	Discreta
3.1.1.D	Professor precisa ajustar o cumprimento do horário de início e de término das aulas (sim ou não).	Discreta

3.1.1.E	Professor precisa ajustar a atualização dos conteúdos (sim ou não).	Discreta
3.1.1.F	Professor precisa ajustar a clareza na exposição dos conteúdos (sim ou não).	Discreta
3.1.1.G	Professor precisa ajustar a disponibilidade para atendimento fora da sala de aula (sim ou não).	Discreta
3.1.1.H	Professor precisa ajustar a qualidade da bibliografia (sim ou não).	Discreta
3.1.1.I	Professor precisa ajustar a qualidade das avaliações (sim ou não).	Discreta
3.2.1	Satisfação geral (de 0 - totalmente insatisfeito, a 10 - totalmente satisfeito) com o desempenho do professor.	Discreta
4.1.1	Probabilidade de recomendar o curso para um amigo ou parente próximo (de 0 - muito improvável, a 10 - muito provável).	Discreta
4.2.1	Interesse em sair de curso (mudar de curso na UFPB ou para outra instituição, parar de estudar etc.) atualmente (de 0 - muito baixo, a 10 - muito alto).	Discreta
OBSERVACOES	Texto livre para qualquer manifestação adicional.	Categórica
QUANTIDADE_TRANCAMENTOS	Número de trancamentos realizado no período.	Contínua

Para entender melhor os dados brutos descritos acima, foi realizada uma análise exploratória utilizando a linguagem de programação Python. Foram analisados 1.156.891 registros das

avaliações dos cursos de graduação na modalidade presencial. Na Figura 16, é possível ver a quantidade de avaliações compreendendo os anos de 2017 a 2021.

Figura 16 – Quantidade de registros de avaliações por período.



Fonte: Figura do autor

3.2.2 Variável alvo

O principal ponto de atenção dessa etapa se concentrou na variável alvo. Ela está localizada na coluna STATUS_DISCENTE e tem como possíveis valores as categorias descritas na Tabela 3.

Tabela 3 – Valores assumidos pela variável alvo.

	Definição
ATIVO	Esse status é associado ao aluno que possui vínculo em vigor com a instituição, que não se encontra com status "Trancado" ou "Formando" e que está matriculado no conjunto mínimo de componente curriculares do seu curso.
TRANCADO	É o status do aluno que realizou a suspensão de programa. Nesse caso, o discente tem vínculo “Ativo”, porém, solicitou a interrupção temporária.
CONCLUIDO	É o status do aluno que concluiu todas as pendências acadêmicas exigidas pela sua estrutura curricular, que já recebeu o grau acadêmico e teve seu diploma registrado.

ATIVO – FORMANDO	É o aluno que tem condições para a conclusão de seu curso no período atual. Ou seja, estudante que está cursando os últimos componentes curriculares para finalizar a carga horária mínima de seu curso.
ATIVO – CONCLUINTE	É o aluno que já concluiu a carga horária mínima de seu curso, porém ainda não recebeu o grau acadêmico.
CANCELADO	É a situação do aluno que teve seu vínculo finalizado como evadido, seja por desistência, insuficiência de rendimento acadêmico, decurso de prazo máximo etc.

Um fato relevante sobre a variável alvo, é que o valor dela corresponde ao status do discente no momento da extração dos dados no sistema, ou seja, o status não corresponde ao que era no momento da avaliação. Dessa forma, existem períodos que possuem avaliações de um aluno que não evadiu naquele momento, por exemplo, se um aluno cursou cinco períodos até evadir, então todas as avaliações de disciplinas dos cinco períodos terão como valor de variável alvo o status CANCELADO.

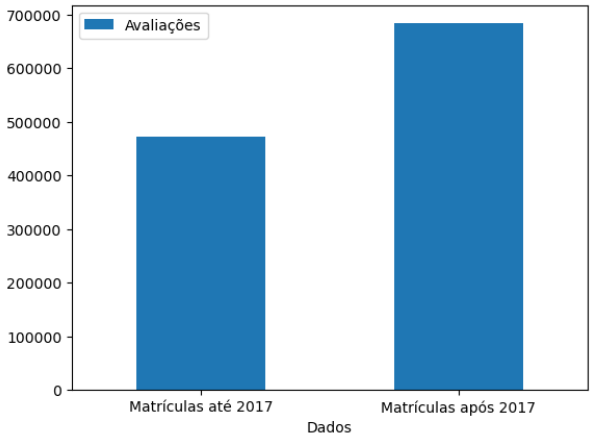
3.3. Fase 3: Preparação dos dados

Com base nas análises prévias realizadas e refinamentos após perfazer alguns ciclos do CRISP-EDM, foram realizadas algumas tomadas de decisões para gerar um subconjunto de dados capaz de performar na etapa de modelagem.

3.3.1 Filtragem

Foi realizada uma filtragem com relação ao ano de matrícula dos alunos. Como a autoavaliação só passou a ser aplicada a partir do primeiro semestre do ano de 2017, optou-se por utilizar apenas as avaliações de alunos que ingressaram nesse período em diante, pois todo o ciclo acadêmico do estudante, até o momento da extração dos dados, está representado. O mesmo não acontece com as matrículas antes de 2017, pois os registros atuais não contêm as avaliações de várias disciplinas já cursadas antes aplicação do instrumento de autoavaliação. Dessa forma, o novo conjunto de dados brutos passou a ter 683.634 registros, sendo desconsiderado 473.257 registros referentes as matrículas antes de 2017, vide Figura 17.

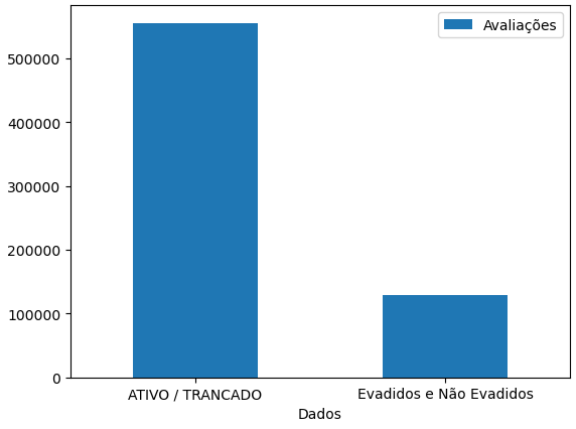
Figura 17 – Quantidade de avaliações após filtragem por ano da matrícula.



Fonte: Figura do autor

Outra filtragem que foi realizada foi com relação a nossa variável alvo STATUS_DISCENTE. Como os status ATIVO e TRANCADO são de alunos que ainda possuem vínculo com a instituição, não é possível atestar se esse aluno irá evadir ou não. Diante desse quadro, consideramos apenas os valores CONCLUIDO, ATIVO – FORMANDO, ATIVO – CONCLUINTE (não evadido) e CANCELADO (evadido), para realizar os treinamentos e testes dos modelos desenvolvidos, pois esses status são finalizadores e atestam se o aluno concluiu a sua jornada acadêmica ou saiu do curso. Com a aplicação desse filtro nosso conjunto de dados brutos passou a ter 128.235 registros (Figura 18). Os outros 555.399 registros referentes aos status não finalizadores não são utilizados na modelagem. Apesar da diferença entre os dados utilizados na modelagem e não utilizados, ainda temos uma quantidade significativa de dados para a obtenção de um modelo bastante representativo.

Figura 18 – Quantidade de avaliações após filtragem por STATUS_DISCENTE.



Fonte: Figura do autor

3.3.2 Remoção de variáveis

Algumas variáveis foram descartadas por não fazer sentido para o contexto da pesquisa. Apesar de elas possuírem informações relevantes e que poderiam levar a alguns subconjuntos de dados que podem ser estudados sob outra perspectiva, a proposta foca em um modelo genérico, independente de curso, disciplina, turno, entre outras características, que contemple todos os cursos da UFPB e sirva de baseline para pesquisas futuras. Diante desse contexto, as variáveis ANO, PERIODO, CENTRO, DEPARTAMENTO, CODIGO, DISCIPLINA, CODIGO_TURMA, HORARIO, LOCAL, CURSO, TURNO, SITUACAO_MATRICULA, QUARTA_PROVA e OBSERVACOES não foram utilizadas para a criação do modelo preditor.

3.3.3 Remoção de valores nulos e outliers

Todas as variáveis foram verificadas quanto a incidência de valores nulos e outliers. Para a variável MATRICULA não foi encontrado nenhum registro com valores nulos ou outliers. Quanto a variável STATUS_DISCENTE não foi encontrado nenhum valor nulo e foi considerado como outlier qualquer valor diferente dos apresentados na Tabela 3. Foram removidos todos os valores nulos nas variáveis MEDIA_FINAL, FALTAS, 1.1.1, 2.1.1, 2.2.1, 3.2.1, 4.1.1 e 4.2.1, para detectar os outliers foi levado em consideração que essas podem assumir valores entre 0 e 10, dessa forma qualquer valor fora desse intervalo é considerado um outlier. Com relação as variáveis 3.1.1.A, 3.1.1.B, 3.1.1.C, 3.1.1.D, 3.1.1.E, 3.1.1.F, 3.1.1.G, 3.1.1.H e 3.1.1.I, verificou-se que elas assumem o valor "X" quando o aluno assinala que o item precisar ser ajustado pelo professor e nenhum valor é atribuído (nulo) quando o aluno não assinala o item, dessa forma foi considerado como outlier qualquer valor que não fosse "X" ou nulo para essas variáveis, nesse caso os valores nulos não foram removidos. No caso da variável QUANTIDADE_TRANCAMENTOS, notou-se que eram informados a quantidade de trancamentos apenas nos casos em que ocorreram trancamentos, já nos casos que não ocorriam trancamento nenhum valor era informado (nulo), dessa forma foi considerado como outlier valores que não fossem inteiros ou nulos, aqui também não foram removidos os valores nulos. Na Tabela 4, é listado a quantidade de registros por variável que possuíam valores fora do esperado.

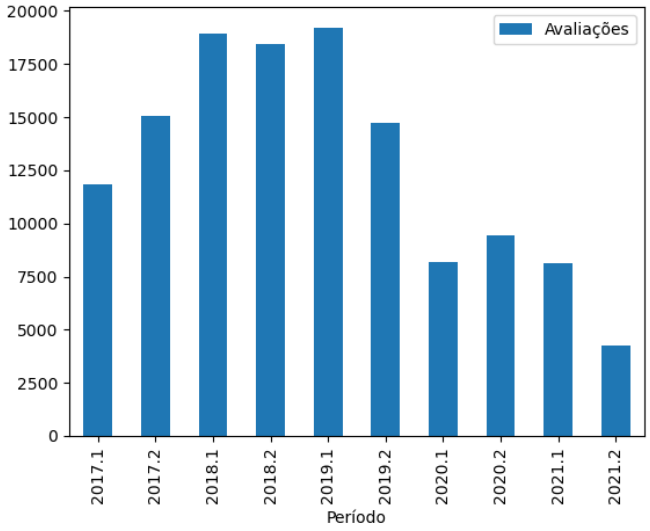
Tabela 4 – Quantidade de *outliers*.

Variável	Quantidade de *Outliers*
MATRICULA	0
STATUS_DISCENTE	0

MEDIA_FINAL	8
FALTAS	5
1.1.1	7927
2.1.1	5752
2.2.1	8610
3.1.1.A	0
3.1.1.B	0
3.1.1.C	0
3.1.1.D	0
3.1.1.E	0
3.1.1.F	0
3.1.1.G	0
3.1.1.H	0
3.1.1.I	0
3.2.1	11067
4.1.1	5260
4.2.1	12736
QUANTIDADE_TRANCAMENTOS	0

Dessa forma, após a remoção de todos os *outliers*, nosso subconjunto de dados passou a ter 120.341 registros de avaliações, como pode ser observado na Figura 19. Deve-se observar que esses dados ainda possuem inconsistências. Uma única matrícula possui vários registros nesse conjunto de dados, ou seja, para cada disciplina cursada em diferentes períodos representam um registro. No entanto, como mencionado anteriormente, todos os registros possuem como valor da sua variável alvo o status final do discente no momento da extração dos dados. Diante desse contexto foi necessário realizar transformações nesses dados.

Figura 19 – Quantidade de registros após filtragem.



Fonte: Figura do autor

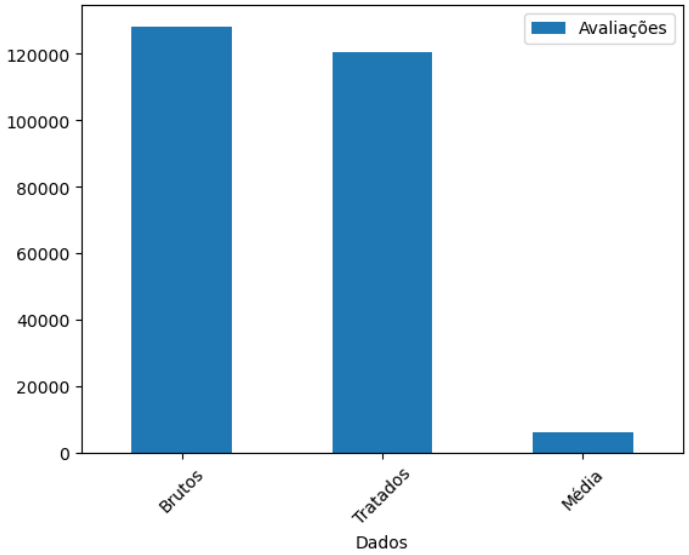
3.3.4 Transformações

A variável MATRICULA foi anonimizada para que se torna-se impossível saber qual os alunos utilizados para a elaboração do modelo. A variável alvo STATUS_DISCENTE foi transformada em valores numéricos, para que os algoritmos de aprendizado de máquina pudessem ter um melhor desempenho. Os status CONCLUIDO, ATIVO – FORMANDO e ATIVO – CONCLUINTE assumiram o valor 0 e CANCELADO assumiu o valor 1. Com relação as variáveis 3.1.1.A, 3.1.1.B, 3.1.1.C, 3.1.1.D, 3.1.1.E, 3.1.1.F, 3.1.1.G, 3.1.1.H e 3.1.1.I, o valor "X" assumiu o valor 1 e os valores nulos assumiram o valor 0. Já a variável QUANTIDADE_TRANCAMENTOS passou a ter os valores nulos representados por 0.

Concluindo as modificações mencionadas acima, surgiu a necessidade de abordar a presença de diversos registros para uma mesma matrícula, correspondendo a diferentes autoavaliações para cada disciplina por período. Considerando que a predição é executada para cada entrada no conjunto de dados, o modelo preditivo poderia deduzir situações tanto de evasão quanto de permanência para uma única matrícula. Para resolver esse desafio, uma abordagem foi adotada: uma média foi calculada para cada variável preditiva, abrangendo todas as autoavaliações associadas a uma matrícula específica. Com isso, conseguimos consolidar um único registro para cada aluno, contendo a média que encapsula todo o seu percurso acadêmico. Esse procedimento proporcionou uma base individualizada para a previsão de evasão. Adicionalmente, um passo crucial envolveu a anonimização de todas as matrículas. Essa ação foi implementada para salvaguardar a privacidade dos alunos que contribuíram para a construção do modelo de previsão.

Como resultado dessas transformações, obtivemos um subconjunto final contendo 6.138 registros. Esses registros representam a média aritmética por matrícula, proveniente dos 120.341 registros resultantes do processamento dos dados iniciais. A Figura 20 ilustra a evolução desse processo.

Figura 20 – Evolução da quantidade de registros após tratamento.



Fonte: Figura do autor

Dessa maneira, ao compilar essas etapas, conseguimos consolidar um conjunto de dados otimizado e preparado para a etapa subsequente de análise e previsão, com cada registro final representando um estudante.

3.4. Fase 4: Modelagem

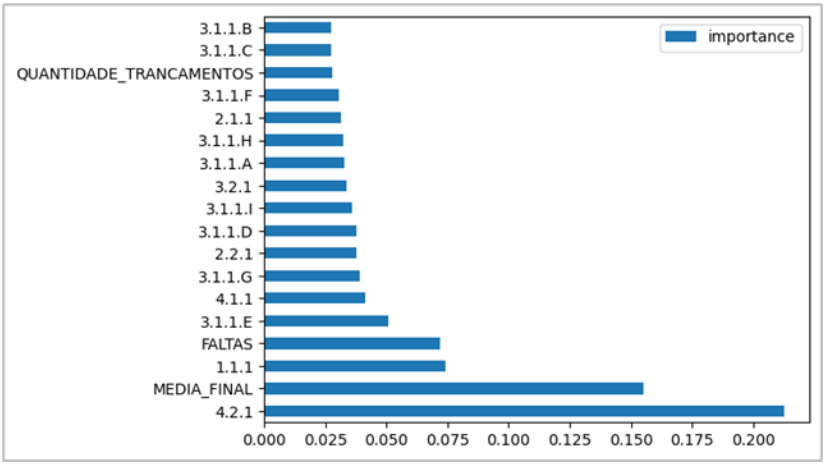
Nesta etapa, as técnicas de modelagem de dados foram estabelecidas, abrangendo um conjunto de algoritmos de aprendizado de máquina, bem como seus parâmetros correspondentes. A gama de algoritmos disponíveis para execução da classificação supervisionada é vasta; no entanto, a criação de um modelo para cada algoritmo seria de grande complexidade e custo. Diante dessa perspectiva, optou-se por escolher os algoritmos de Máquina de Vetores de Suporte (Support Vector Machine - SVM), Floresta Aleatória (Random Forest - RF) e Árvore de Decisão (Decision Tree - DT), que exibiram os resultados mais notáveis no contexto da RSL sugerida por (SANTOS; SARAIVA; OLIVEIRA, 2021). A implementação dos algoritmos selecionados se valeu da biblioteca Scikit-learn, através da linguagem de programação Python. Especificamente, o intuito foi realizar uma

avaliação comparativa dos modelos e, com base no desempenho apresentado, eleger aquele que melhor se adequa ao escopo proposto.

3.4.1 Seleção de atributos

Durante o processo de seleção de atributos, foram utilizados os métodos *Filter*, *Wrapper* e *Embedded* durante vários ciclos da metodologia CRISP-EDM e o método *Embedded* utilizando o algoritmo *Random Forest* foi o que apresentou os melhores resultados para todos os modelos desenvolvidos. Para exemplificar, a Figura 21 mostra a média de importância dos atributos selecionados utilizando o método *embedded* com o algoritmo *Random Forest*.

Figura 21 – Média de importâncias de atributos.



Fonte: Figura do autor

Como podemos observar, o atributo de maior importância corresponde a própria manifestação do aluno com relação ao seu desejo de evadir (4.2.1), no entanto, apenas essa informação seria insuficiente para prever a evasão escolar, pois, conforme Tabela 5, todos os atributos possuem algum grau de importância na tarefa de classificação.

Tabela 5 – Média de importância de atributos.

	Importância
4.2.1 Seu interesse em sair de curso (mudar de curso na UFPB ou para outra instituição, parar de estudar etc.) no momento atual é (de 0 - muito baixo, a 10 - muito alto)	0.223752
MEDIA_FINAL Média final obtida pelo aluno na disciplina avaliada	0.191523

FALTAS Total de faltas do aluno na disciplina avaliada	0.073314
1.1.1 Por favor, dê uma nota (de 0 - muito ruim, a 10 - muito bom) para SEU desempenho pessoal na disciplina em termos de comprometimento e motivação	0.071367
3.1.1.E O professor PRECISA AJUSTAR: Atualização dos conteúdos	0.052639
3.1.1.I O professor PRECISA AJUSTAR: Qualidade das avaliações	0.038201
3.1.1.G O professor PRECISA AJUSTAR: Disponibilidade para atendimento fora da sala de aula	0.033996
3.1.1.H O professor PRECISA AJUSTAR: Qualidade da bibliografia	0.033599
4.1.1 Considerando a experiência com seu curso até esse último período, a probabilidade de você recomendar esse curso para um amigo ou parente próximo é (de 0 - muito improvável, a 10 - muito provável).	0.033144
2.2.1 Na sua percepção, qual o nível de dificuldade DOS CONTEÚDOS das disciplinas cursadas (de 0 - muito fácil, a 10 - muito difícil)?	0.032224
3.1.1.D O professor PRECISA AJUSTAR: Cumprimento do horário de início e de término das aulas	0.028450
3.1.1.C O professor PRECISA AJUSTAR: Comparecimento às aulas	0.028405
3.1.1.F O professor PRECISA AJUSTAR: Clareza na exposição dos conteúdos	0.027943
3.2.1 Por favor, aponte sua satisfação geral (de 0 - totalmente insatisfeito, a 10 - totalmente satisfeito) com o desempenho de cada professor	0.027885
QUANTIDADE_TRANCAMENTOS Média de trancamentos no período de avaliação	0.026405

3.1.1.A O professor PRECISA AJUSTAR: Cumprimento do plano de curso	0.026273
2.1.1 Na sua percepção, qual o nível de importância (de 0 - sem importância, a 10 - extremamente importante) das disciplinas cursadas para o seu curso?	0.024996

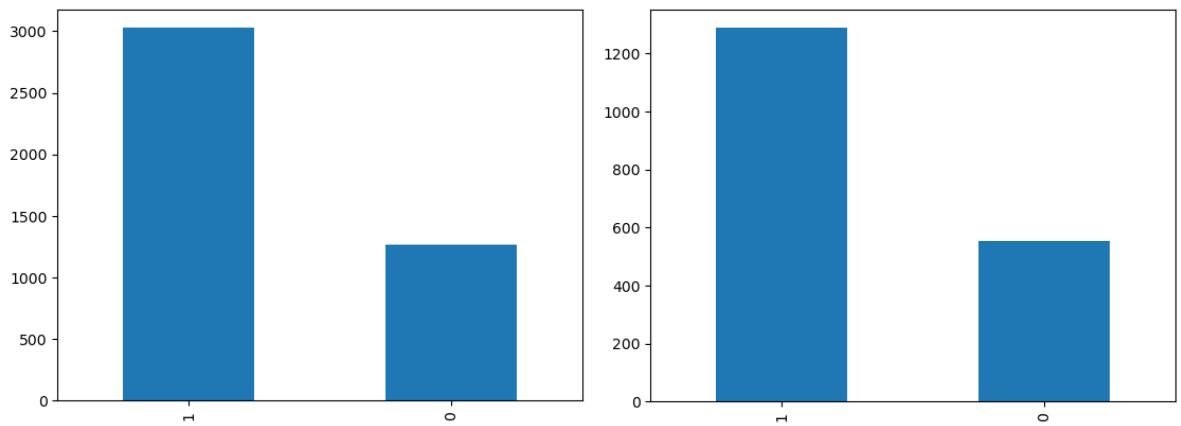
A definição dos pesos/importâncias é determinante para que o algoritmo consiga treinar o modelo de aprendizado de máquina de forma mais eficiente e eficaz. Na fase 5 da metodologia é realizada a validação do modelo treinado e constatado se a seleção de atributos refletiu em um bom desempenho do modelo classificador na predição da evasão. Esse processo é cíclico e refinado constantemente como demonstra a metodologia CRISP-EDM.

3.4.2 Separação dos dados

Para análise dos modelos desenvolvidos, foi utilizado a técnica *Holdout* no conjunto de dados, pois, a partir do recorte da avaliação é possível entender como o modelo se comporta. Para essa análise, os dados foram separados em um subconjunto aleatório de teste com 30% dos registros (1.842) e um de treinamento com 70% dos registros (4.296), como mostra a Figura 22.

Para essa análise, os dados foram separados em um subconjunto aleatório de teste com 30% dos registros (1.842) e um de treinamento com 70% dos registros (4.296), como mostra a Figura 22. Dos 4.296 exemplos do subconjunto de dados de treinamento, 3.026 correspondem a classe majoritária (CANCELADO) e 1.270 representam a classe minoritária (CONCLUIDO). Já os dados de teste são representados por 1.288 (CANCELADO) e 554 (CONCLUIDO).

Figura 22 – Dados de treinamento e teste respectivamente.



Fonte: Figura do autor

Para validar o poder de generalização do modelo desenvolvido, foi realizada uma validação cruzada dos dados, dessa forma diferentes configurações dos dados de treinamento e teste são garantidas, produzindo no total k repetições entre subconjuntos de treinamento e teste. Uma ótima alternativa de divisão de subconjuntos é definir 10 para o valor de k (GÉRON, 2019).

3.4.3 Desbalaceamento dos dados

Um ponto de atenção durante a tarefa de classificação, foi a verificação dos dados de treinamento quanto ao seu desbalanceamento. Existem mais registros de alunos evadidos (3.026) do que alunos concluídos (1.270), como podemos verificar na Figura 22.

Diante desse contexto, foram implementados modelos com os dados de treinamento balanceados através das técnicas de subamostragem (*undersampling*) e sobreamostragem (*oversampling*) utilizando a biblioteca *Imbalanced-learn* [6] e modelos utilizando os dados desbalanceados. No caso da subamostragem, foi gerada uma subamostra aleatória (*Random Undersampling*) da classe de maior frequência e foi mantida a classe de menor frequência. No caso da sobre amostragem, foi gerada uma sobre amostra replicando aleatoriamente os registros da classe de menor frequência (*Random Oversampling*), também foi gerada duas outras sobreamostras da classe de menor frequência utilizando as técnicas SMOTE e ADASYN. Com esse cenário, foi possível avaliar como os algoritmos se comportavam e qual abordagem seria mais coerente para o objetivo da pesquisa, que é detectar os alunos com maior risco de evasão escolar.

3.5. Fase 5: Avaliação dos modelos

Foram utilizadas, nesta etapa, as métricas Acurácia*,* Precisão*, Recall e F-Measure* para avaliar os modelos implementados (Seção 3.1.5). As avaliações realizadas tiveram como parâmetro o balanceamento dos dados de treinamento, sendo realizadas avaliações para dados desbalanceados e balanceados utilizando diferentes técnicas.

3.5.1 Dados desbalanceados

O primeiro modelo desenvolvido utilizou os dados de treinamento desbalanceados, dessa forma foi possível observar como os algoritmos de aprendizado de máquina se comportavam quanto ao desbalanceamento, assim essa avaliação se tornou a linha base para as avaliações posteriores.

---

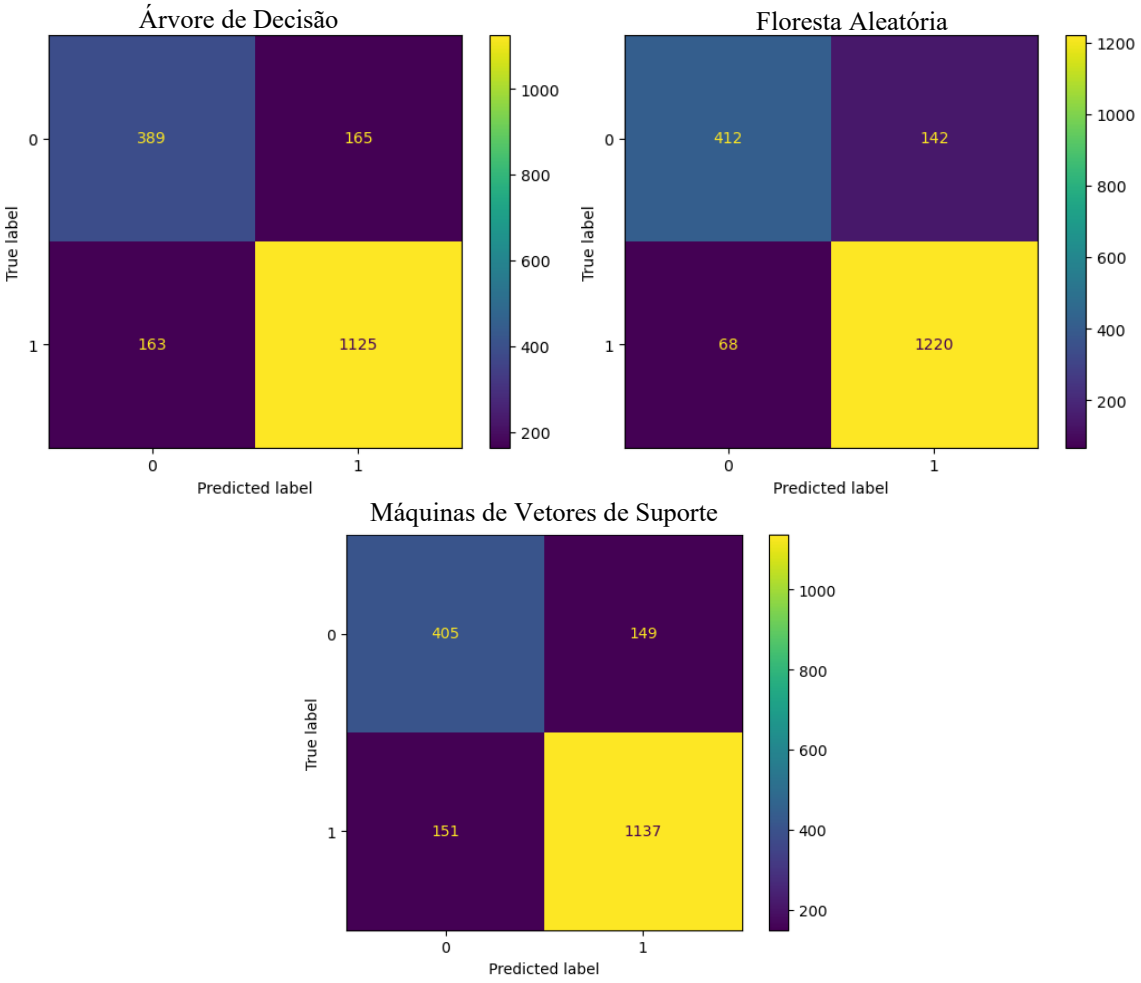
[6] https://imbalanced-learn.org/

Tabela 6 – Resultados das métricas com dados desbalanceados utilizando um conjunto de dados de treinamento (70%) e teste (30%) aleatório.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.82193268	0.87209302	0.87344720	0.87276958
Floresta Aleatória	**0.88599348**	**0.89574155**	**0.94720496**	**0.92075471**
Máquinas de Vetores de Suporte	0.83713355	0.88413685	0.88276397	0.88344988

Como podemos observar na Tabela 6, o algoritmo Floresta Aleatória obteve os melhores resultados em todas as métricas, utilizando a técnica *Holdout* para um único conjunto de dados de treinamento e teste escolhido de forma aleatória. Para entender melhor as métricas apresentadas, a Figura 23 apresenta as matrizes de confusão obtidas nos modelos utilizando um conjunto de dados de treinamento (70%) e teste (30%) aleatório.

Figura 23 – Matrizes de confusão dos modelos desbalanceados utilizando um conjunto de dados de treinamento (70%) e teste (30%) aleatório.



Fonte: Figura do autor

Podemos observar no algoritmo Floresta Aleatória, que em um universo de teste com dados reais de 1.288 alunos evadidos (CANCELADO), o modelo obteve apenas 68 Falsos Negativos (FN), atingindo um *recall* de 94,72%. O modelo de Árvore de Decisão classificou 163 como FN e apresentou um aumento de Falsos Positivos (FP), com 165 casos. Já o modelo SVM teve melhor desempenho que o modelo de Árvore de Decisão, com 151 FN e 149 FP.

Para uma validação mais precisa, foi aplicada a técnica de validação cruzada com a utilização de 10 dobras para o algoritmo *StratifiedKFold*. Neste método os dados são divididos em k=10 subconjuntos. Em seguida, o método *holdout* é repetido k vezes, de tal forma que, a cada vez, um dos k subconjuntos é usado como set de validação e os outros subconjuntos k-1 são colocados juntos para formar um set de treinamento. A média, dos k resultados das avaliações realizadas, pode ser observada na Tabela 7, sendo o algoritmo Floresta Aleatória o que obteve os melhores resultados em todas as métricas.

Tabela 7 – Média dos resultados das métricas com dados de treinamento desbalanceado utilizando a técnica de validação cruzada.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.80448055	0.86616020	0.85489494	0.86028856
Floresta Aleatória	**0.87827977**	**0.90075786**	**0.93139125**	**0.91542677**
Máquinas de Vetores de Suporte	0.83071188	0.88979089	0.86647439	0.87783455

Apesar dos resultados satisfatórios, nos quais uma média de Recall de 93,13% foi alcançada, é importante observarmos a especialização do modelo em relação à classe majoritária (CANCELADO). Como podemos observar, o modelo apresenta uma quantidade significativa de FP relativo à classe minoritária, dessa forma se tivéssemos um maior número de caso de testes para a classe minoritária (CONCLUIDO), a tendencia é que a métrica Precisão diminuísse proporcionalmente, diminuindo também a *F-Measure*. Diante do contexto, apesar de termos um modelo com baixo número de FN, poderíamos ter outro problema, que é ter um número significativo de alunos sendo classificados como com potencial de evasão escolar, gerando uma perda significativa de recursos da instituição ao concentrar esforços dos stakeholders na mitigação de evasão escolar em casos FP.

A avaliação por validação cruzada aplicada nesse modelo é a mesma aplicada para os demais modelos e é a métrica utilizada para a tomada de decisão quanto ao modelo a ser proposto.

3.5.2 Dados balanceados: subamostragem aleatória

Nesta tarefa de classificação, foram utilizados os dados de treinamento balanceados através do método de subamostragem aleatória, ou seja, um subconjunto foi selecionado aleatoriamente a partir da classe de maior frequência (CANCELADO). Dessa forma, a classe majoritária passou a ter 1.270 exemplos, igualmente a classe minoritária.

A Tabela 8 traz a avaliação dos modelos desenvolvidos com dados de treinamento balanceados pelo método de subamostragem aleatória, o modelo que apresentou o melhor resultado, utilizando a técnica *Holdout* para um conjunto de dados de treinamento (70%) e teste (30%) aleatório, foi o Floresta Aleatória.

Tabela 8 – Resultados das métricas com dados balanceados por subamostragem aleatória utilizando um conjunto de dados de treinamento (70%) e teste (30%) aleatório.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.76981541	0.89059674	0.76475155	0.82289055
Floresta Aleatória	**0.87133550**	**0.93974895**	**0.87189440**	**0.90455094**
Máquinas de vetores de suporte	0.82138979	0.93159315	0.80357142	0.86285952

Para uma melhor avaliação dos modelos, foi aplicada a técnica de validação cruzada para os modelos com os dados de treinamento balanceados por subamostragem aleatória, a Tabela 9 apresenta a média das avaliações.

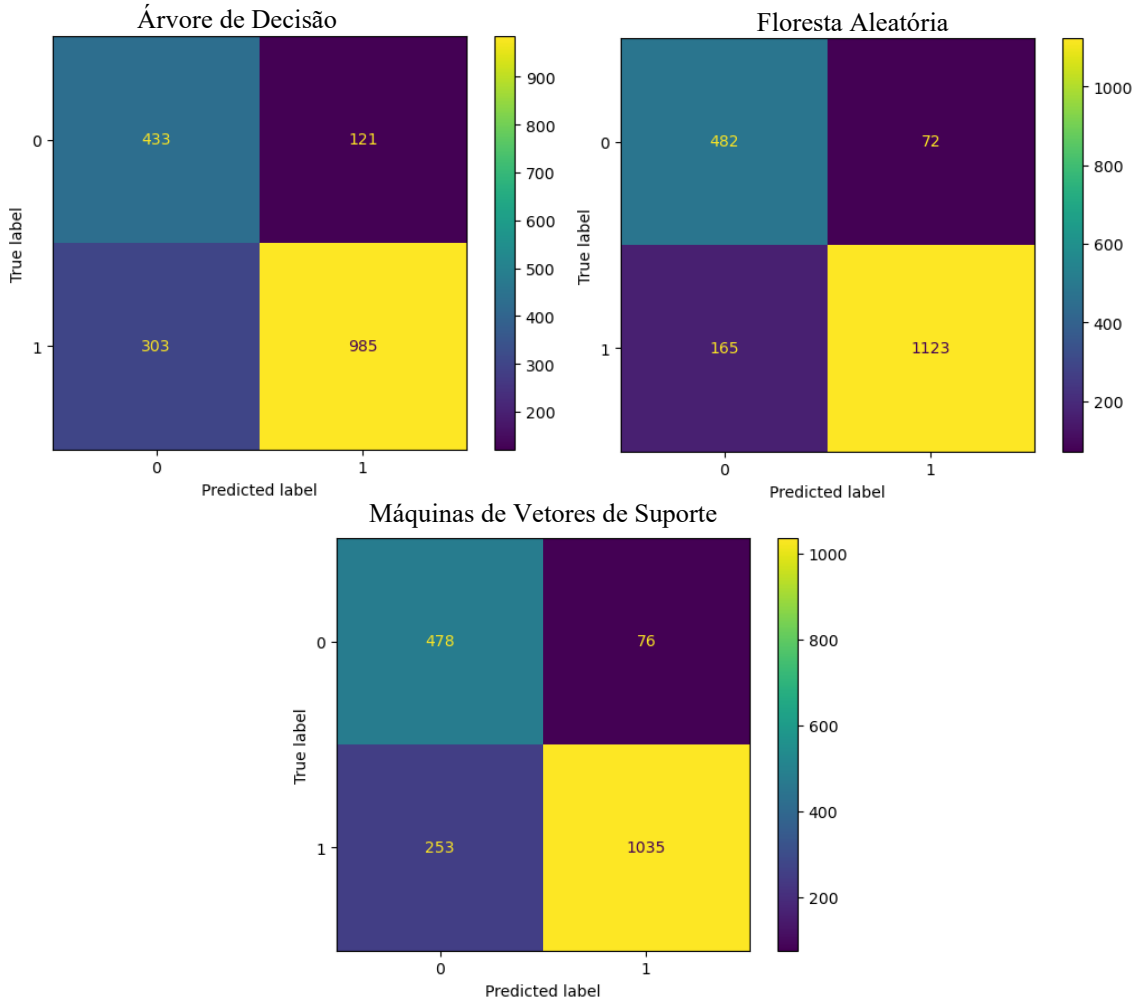
Tabela 9 – Média dos resultados das métricas com dados de treinamento balanceados por subamostragem aleatória utilizando a técnica de validação cruzada.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.79339979	0.90131798	0.78628351	0.84267739
Floresta Aleatória	**0.86068648**	**0.93532036**	**0.86417354**	**0.89483536**
Máquinas de vetores de suporte	0.80870923	0.93459041	0.77954960	0.85006802

Como podemos observar na Figura 24, ao diminuir a quantidade de exemplos da classe majoritária, ocorreu uma piora no desempenho do modelo em relação à quantidade de FN. Isso pode

ter ocorrido devido à perda de dados relevantes para o treinamento do algoritmo, o que resultou na dificuldade do algoritmo em classificar adequadamente a sobreposição de classes.

Figura 24 – Matrizes de confusão dos modelos balanceados por subamostragem aleatória utilizando um conjunto de dados de treinamento (70%) e teste (30%) aleatório.



Fonte: Figura do autor

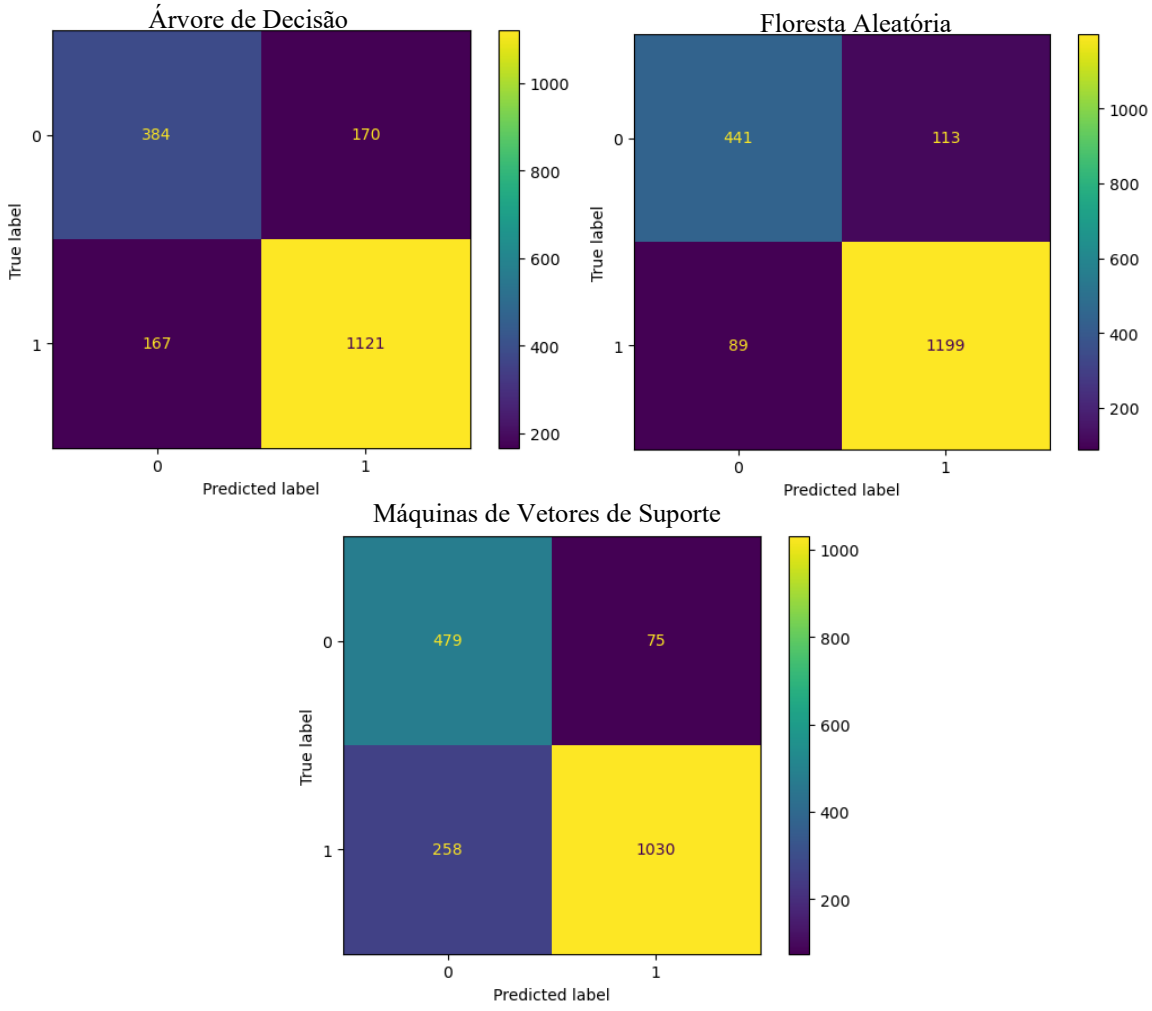
Ao analisar as métricas obtidas, observa-se uma inversão quase completa de resultados entre as métricas *Recall* e Precisão nas Tabelas 7 e 9. Esse fenômeno pode ser explicado pelos mesmos motivos que afetam a avaliação em dados de treinamento desbalanceados. Em outras palavras, devido à escassez de dados na classe majoritária, o algoritmo de AM passou a inclinar-se mais para classificar as sobreposições de classes como pertencentes à classe minoritária. Isso é o oposto do cenário anterior, em que, devido ao excesso de ajuste à classe majoritária, as sobreposições tendiam a ser classificadas como parte da classe majoritária.

Outros métodos de subamostragem como o *Edited Nearest Neighbours* (ENN) foram testados, no entanto, apresentou desempenho abaixo da subamostragem aleatória e ocorreu os mesmos problemas de classificação quanto as sobreposições de classe. Diante desse contexto, optou-se por verificar os dados balanceados através dos métodos de sobreamostragem.

3.5.3 Dados balanceados: sobreamostragem SMOTE

Utilizando a técnica de sobreamostragem SMOTE (*Synthetic Minority Oversampling Technique*), que gera novos exemplos da classe minoritária através de interpolação entre os pontos mais próximos, fez com o modelo obtivesse uma melhor performance quanto aos outros modelos.

Figura 25 – Matrizes de confusão dos modelos balanceados por sobreamostragem SMOTE utilizando um conjunto de dados de treinamento (70%) e teste (30%) aleatório.



Fonte: Figura do autor

Na matriz de confusão apresentada na Figura 25, é possível observar que o número de FP teve uma melhora significativa com relação ao modelo desbalanceado e o número de FN teve uma piora relativa, tornando esse modelo, com o algoritmo Floresta Aleatória, o mais equilibrado entre os modelos analisados. Ao analisarmos os resultados obtidos (Tabela 10), observamos que esse equilíbrio se reflete nas métricas *Recall* e Precisão, refletindo em uma equiparação das métricas quanto a classificação das sobreposições de classes que reflete em modelo mais assertivo na predição da evasão escolar.

Tabela 10 – Resultados das métricas com dados balanceados por sobreamostragem SMOTE utilizando a técnica *Holdout*.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.81704668	0.86831913	0.87034161	0.86932919
Floresta Aleatória	**0.89033659**	0.91387195	**0.93090062**	**0.92230769**
Máquinas de vetores de suporte	0.81921824	**0.93212669**	0.79968944	0.86084412

Para uma melhor avaliação dos modelos, foi aplicada a técnica de validação cruzada para os modelos com os dados de treinamento balanceados por sobreamostragem SMOTE, a Tabela 11 apresenta a média das avaliações.

Tabela 11 – Média dos resultados das métricas com dados de treinamento balanceados por sobreamostragem SMOTE utilizando a técnica de validação cruzada.

	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão	0.80578029	0.88209109	0.84863098	0.86641259
Floresta Aleatória	**0.87974371**	0.91724143	**0.91679234**	**0.91574044**
Máquinas de vetores de suporte	0.80968616	**0.93308219**	0.78580819	0.85346256

Também foram utilizados os métodos de sobreamostragem aleatória e ADASYN, no entanto, o método SMOTE foi o que apresentou os melhores resultados. O modelo Floresta Aleatória com os dados de treinamento balanceados pela técnica de sobreamostragem SMOTE também foi o modelo escolhido por esta pesquisa como o proposto para tentar mitigar o fenômeno da evasão escolar na UFPB. A seguir na Tabela 12 é possível ver um quadro resumo com todos os modelos desenvolvidos e analisados.

Tabela 12 – Quadro resumo com todos os modelos desenvolvidos e analisados com base na técnica de validação cruzada.

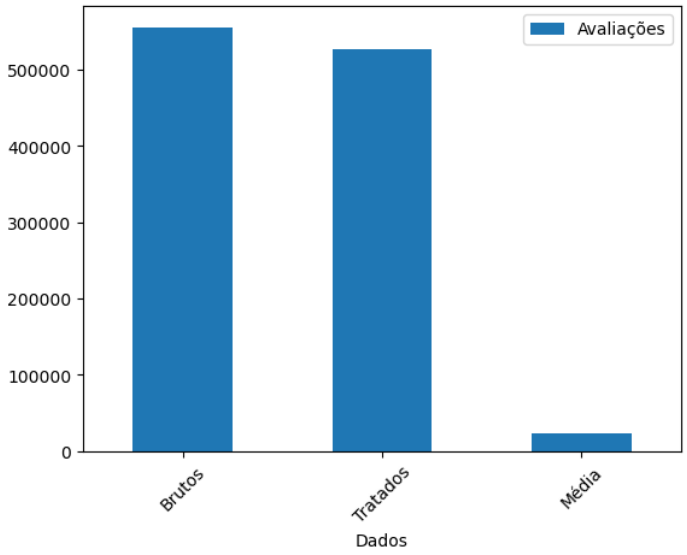
	Acurácia	**Precisão**	***Recall***	***F-Measure***
Árvore de Decisão Desbalanceado	0.80448055	0.86616020	0.85489494	0.86028856
Árvore de Decisão Balanceado por Subamostragem	0.79339979	0.90131798	0.78628351	0.84267739
Árvore de Decisão Balanceado por Sobreamostragem SMOTE	0.80578029	0.88209109	0.84863098	0.86641259
Floresta Aleatória Desbalanceado	0.87827977	0.90075786	**0.93139125**	0.91542677
Floresta Aleatória Balanceado por Subamostragem	0.86068648	**0.93532036**	0.86417354	0.89483536
Floresta Aleatória Balanceado por Sobreamostragem SMOTE	**0.87974371**	0.91724143	0.91679234	**0.91574044**
Máquinas de Vetores de Suporte Desbalanceado	0.83071188	0.88979089	0.86647439	0.87783455
Máquinas de vetores de suporte Balanceado por Subamostragem	0.80870923	0.93459041	0.77954960	0.85006802
Máquinas de vetores de suporte Balanceado por Sobreamostragem SMOTE	0.80968616	0.93308219	0.78580819	0.85346256

3.6. Fase 6: Implementação da solução educacional

Para entender o funcionamento da solução educacional, o modelo proposto foi aplicado para a base de dados de alunos ativos. Foram utilizados os 555.399 registros referentes aos status não finalizadores que não são utilizados na modelagem. A Figura 26 mostra a evolução dos dados após aplicação das mesmas etapas realizadas na modelagem.

Foi realizada a predição de evasão sobre os dados da média dos 525.694 registros tratados, resultando em 22.560 médias de autoavaliações para cada matrícula distinta. A Figura 27 apresenta a quantidade de matrículas classificadas como CANCELADA e CONCLUIDA utilizando o modelo com o algoritmo Floresta Aleatória com dados de treinamento balanceados por sobreamostragem SMOTE (Seção 4.5.3).

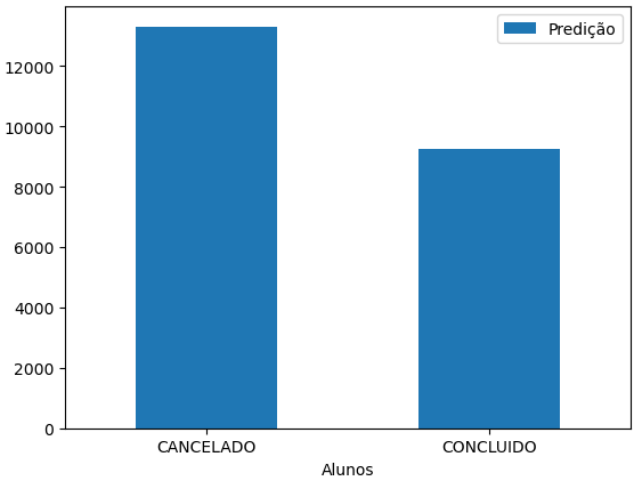
Figura 26 – Evolução da quantidade de registros de alunos ativos após tratamento.



Fonte: Figura do autor

Das 22.560 matrículas ativas de ingressantes entre os anos de 2017 e 2021, o modelo classificou 13.310 como CANCELADO e 9.250 como CONCLUIDO, ou seja, para o modelo preditor 59% dos alunos têm potencial de evasão escolar. Se compararmos com a proporção de ingressantes e concluintes no estado da Paraíba, apresentado na Seção 2.3, temos uma porcentagem totalmente compatível com o cenário atual no estado.

Figura 27 – Predição de evasão escolar para as avaliações de alunos ativos.



Fonte: Figura do autor

Com base nesses dados, é possível afirmar que o modelo proposto pode ser utilizado para implementação de soluções educacionais que auxiliem os stakeholders na tomada de decisão que resulte em intervenções para melhoria do processo educacional e mitigação da evasão escolar,

antecipando problemas antes que se tornem irreversíveis. Ao identificar estudantes em risco de evasão com base em indicadores do modelo de predição desenvolvido, as instituições podem intervir prontamente e oferecer apoio personalizado. Isso pode incluir tutorias adicionais, aconselhamento acadêmico ou a implementação de programas de orientação e suporte emocional. No entanto, a implementação efetiva dessas soluções extrapola o escopo desta pesquisa, pois ela precisa ser validada e proposta pela alta gestão da instituição.

4. CONSIDERAÇÕES E TRABALHOS FUTUROS

No presente estudo, foram apresentadas e implementadas diversas abordagens de classificação para prever a evasão escolar com base na interpretação dos dados da autoavaliação dos cursos de graduação da UFPB. Com o objetivo de alcançar esse propósito, utilizou-se a metodologia CRISP-EDM como guia para a mineração de dados. Cada uma dessas abordagens se distingue das demais tanto no equilíbrio dos dados de treinamento quanto nos algoritmos de aprendizado de máquina adotados. Para validar as diferentes abordagens de classificação, foram empregadas as métricas de Acurácia, Precisão, Recall e F-Measure. A partir dos resultados dessas métricas, foi proposto um método de predição baseado no algoritmo de Floresta Aleatória, com o balanceamento dos dados utilizando a técnica de sobreamostragem SMOTE, obtendo-se uma acurácia de 87,97%, precisão de 91,72%, Recall de 91,67% e F-Measure de 91,57%. Além disso, o método proposto foi aplicado aos dados da autoavaliação dos alunos ativos, e os resultados demonstraram compatibilidade com os índices de evasão escolar atualmente observados no estado.

De maneira geral, os modelos propostos na literatura se baseiam em variáveis que resultam do desempenho dos alunos, que, em muitos casos, só são registradas quando o aluno já está em processo de evasão. Essa observação sublinha a necessidade de uma abordagem mais proativa na identificação de sinais precoces de evasão, como apontado pelo questionário de autoavaliação. Baseado nisso, é preciso enfatizar a importância da autoavaliação pelo discente como instrumento relevante para a gestão acadêmica, este estudo destaca como esses dados muitas vezes subutilizados ou realizados apenas como um requisito burocrático podem fornecer informações cruciais para a prevenção da evasão escolar.

De acordo com o modelo preditivo, atualmente, dos 22.560 alunos ativos que ingressaram na instituição a partir de 2017 até 2021, 59% apresentam probabilidade de evasão do curso. Essa taxa elevada reflete a situação atual, conforme apontada pelo Mapa do Ensino Superior no Brasil, sublinhando a urgência de medidas direcionadas à retenção de alunos e ao aprimoramento das estratégias de apoio acadêmico.

A partir do trabalho desenvolvido, surgem novos desafios e oportunidades que direcionam a necessidade de investigações futuras. A seguir, apresentaremos uma visão geral dos trabalhos futuros que podem proporcionar informações valiosas, expandir fronteiras e abrir novas perspectivas:

- Explorar a possibilidade de identificar se as disciplinas específicas têm um impacto

significativo na evasão escolar, permitindo a criação de modelos mais precisos e direcionados. Além disso, investigar se modelos específicos por curso podem superar o desempenho do modelo genérico proposto inicialmente.

- Incorporar dados socioeconômicos, acadêmicos, culturais, entre outros, provenientes de estudos anteriores, para enriquecer o modelo preditivo. A inclusão desses dados pode resultar em resultados mais assertivos e em uma compreensão mais abrangente dos fatores que influenciam a evasão escolar.
- Criar uma interface de programação de aplicativos (API) que facilite o processamento dos dados de autoavaliação dos alunos e a aplicação do modelo preditivo. Isso permitirá a automação do processo e a fácil integração com outros sistemas.
- Coletar e incorporar dados atualizados de semestres posteriores ao modelo preditivo. Isso possibilitará uma melhoria contínua do modelo à medida que mais informações se tornarem disponíveis, tornando-o mais preciso e atualizado.
- Integrar a API desenvolvida com o Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) da instituição de ensino. Isso viabilizará a apresentação de informações relevantes e em tempo real para os stakeholders, fornecendo informações valiosas para a tomada de decisões estratégicas e implementação de ações específicas para mitigar a evasão escolar.
- Desenvolver painéis interativos que apresentem as previsões de evasão escolar para cada semestre. Esses painéis proporcionarão aos stakeholders uma visão das tendências e padrões de evasão, permitindo que eles desenvolvam políticas e aprimorem processos com base nessas informações.

Esses trabalhos futuros têm como objetivo aprimorar o modelo de predição de evasão escolar, aumentando sua precisão, incorporando dados relevantes, automatizando o processamento de dados, integrando-o a sistemas existentes e fornecendo informações acionáveis para combater a evasão escolar de maneira eficaz.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBAN, M.; MAURICIO, D. *Predicting university dropout through data mining: A Systematic Literature.* Indian Journal of Science and Technology, v. 12, n. 4, p. 1-12, 2019.

ALVARENGA JÚNIOR, W. J. Métodos de otimização hiperparamétrica: um estudo comparativo utilizando árvores de decisão e florestas aleatórias na classificação binária. 2018.

ANDIFES, A.; ABRUEM, A.; SESU/MEC, S. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas: resumo do relatório apresentado a ANDIFES, ABRUEM e SESu/MEC pela Comissão Especial. Avaliação: Revista da Avaliação da Educação Superior, [S. l.], v. 1, n. 2, 1996. Disponível em: http://periodicos.uniso.br/ojs/index.php/avaliacao/article/view/739. Acesso em: 28 mar. 2022.

BAGGI, C. A. S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. Avaliação: Revista da Avaliação da Educação Superior (Campinas), v. 16, n. 2, p. 355-374, 2011.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o brasil. Revista Brasileira de Informática na Educação (RBIE), v. 19, n. 02, p. 03, 2011.

BATISTA, G. E.; PRATI, R. C.; MONARD, M. *A study of the behavior of several methods for balancing machine learning training data*. ACM SIGKDD explorations newsletter, v. 6, n. 1, p. 20-29, 2004.

BOLÓN-CANEDO, V.; SÁNCHEZ-MAROÑO, N.; ALONSO-BETANZOS, A. *A review of feature selection methods on synthetic data*. Knowledge and information systems 34, 2013.

BREIMAN, Leo. *Random forests*. Machine learning, v. 45, n. 1, p. 5-32, 2001.

Censo da Educação Superior 2019. Instituto Nacional de Estudos e Pesquisa Educacionais Anísio Teixeira, INEP, 2020. Disponível em: https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2020/Apresentacao_Censo_da_Educacao_Superior_2019.pdf. Acesso em: 07 abr. 2022.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. *SMOTE: synthetic minority over-sampling technique*. Journal of artificial intelligence research, 321-357, 2002.

COLPO, M. P.; PRIMO, T. T.; PERNAS, A. M.; CECHINEL, C. Mineração de Dados Educacionais na Previsão de Evasão: uma RSL sob a Perspectiva do Congresso Brasileiro de Informática na Educação. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31., 2020, Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1102-1111. DOI: https://doi.org/10.5753/cbie.sbie.2020.1102.

COSTA, E; BAKER, R. S.; AMORIM, L.; MAGALHÃES, J.; MARINHO, T. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. Jornada de Atualização em Informática na Educação, v. 1, n. 1, p. 1-29, 2013.

COSTA, F. J.; DIAS, J. J. L.. Avaliação da formação superior pelo discente: proposta de um instrumento. Avaliação: Revista da Avaliação da Educação Superior (Campinas), v. 25, n. 2, p. 275–296, maio 2020. doi: https://doi.org/10.1590/S1414-4077/S1414-40772020000200003.

GAMBA, Estêvão; RIGHETTI, Sabine. Em crise, universidades federais participam de mais da metade da produção científica. Folha de São Paulo, 2022. Disponível em: <https://www1.folha.uol.com.br/educacao/2022/12/em-crise-universidades-federais-participam-de-mais-da-metade-da-producao-cientifica.shtml/>. Acesso em: 14 de mar. de 2023.

GÉRON, A. Mãos à Obra Aprendizado de Máquina com Scikit-Learn & TensorFlow: Conceitos, Ferramentas e Técnicas Para a Construção de Sistemas Inteligentes. Alta Books: Rio de Janeiro, 2019.

GUYON, I.; GUNN, S.; NIKRAVESH, M.; ZADEH, L. A. *Feature extraction: foundations and applications*. Vol. 207. Springer, 2008.

HAN, J.; KAMBER, M. *Data mining: concepts and techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. H.; FRIEDMAN, J. H. *The elements of statistical learning: data mining, inference, and prediction*. New York: springer, 2009.

HE, H.; BAI, Y.; GARCIA, E. A.; LI, S. *ADASYN: Adaptive synthetic sampling approach for imbalanced learning.* In IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322-1328, 2008.

JOSHI, A. V. Machine Learning and Artificial Intelligence. Cham: Springer, 2020.

LIMA, R. A. F. Estratégias de seleção de atributos para detecção de anomalias em transações eletrônicas. 2016.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Introdução às máquinas de vetores suporte (Support Vector Machines). Laboratório de Inteligência Computacional, ICMC/USP, São Carlos, n. 192, 2003.

LOTTERING, R.; HANS, R.; LALL, M. *A model for the identification of students at risk of dropout at a university of technology*. In: 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD). IEEE, 2020. p. 1-8.

LOUPPE, G. Understanding random forests: From theory to practice. arXiv preprint arXiv:1407.7502, 2014.

LOUSRHANIA, Larissa. Universidades públicas lideram ranking brasileiro de patentes. Rádio Agência Nacional, 2021. Disponível em: <https://agenciabrasil.ebc.com.br/radioagencia-nacional/pesquisa-e-inovacao/audio/2021-07/universidades-publicas-lideram-ranking-brasileiro-de-patentes/>. Acesso em: 14 de mar. de 2023.

MANRIQUE, R.; NUNES, B. P.; MARINO, O.; CASANOVA, M. A.; NURMIKKO-FULLER, T. *An analysis of student representation, representative features and classification algorithms to predict degree dropout*. In: Proceedings of the 9th International Conference on Learning Analytics & Knowledge. 2019. p. 401-410.

Mapa do Ensino Superior no Brasil – 13ª Edição. Instituto Semesp, 2023. Disponível em: https://www.semesp.org.br/wp-content/uploads/2023/06/mapa-do-ensino-superior-no-brasil-2023.pdf. Acesso em: 01 set. 2023.

MITCHELL, T. M. *Machine learning*. The McGraw-Hill Companies. Inc., New York, 1997.

PEREIRA, R. T.; ZAMBRANO, J. C. *Application of decision trees for detection of student dropout profiles*. In: 2017 16th IEEE international conference on machine learning and applications (ICMLA). IEEE, 2017. p. 528-531.

PONTE, Caio; CAMINHA, Carlos; FURTADO, Vasco. Otimização de Florestas Aleatórias através de ponderação de folhas em árvore de regressão. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 17., 2020, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 698-708. DOI: https://doi.org/10.5753/eniac.2020.12171.

PRESTES, E. M. T.; FIALHO, M. G. D. Evasão na educação superior e gestão institucional: o caso da Universidade Federal da Paraíba. Ensaio: Avaliação e Políticas Públicas em Educação, v. 26, p. 869-889, 2018.

RAFIQ, M. A.; RABBI, A. M.; AHAMMAD, R. *A data science approach to Predict the University Students at risk of semester dropout: Bangladeshi University Perspective*. In: 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE, 2021. p. 1350-1354.

RAMOS, J. L. C.; RODRIGUES, R. L.; SILVA, J. C. S.; OLIVEIRA, P. L. S. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31., 2020, Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1092-1101. DOI: https://doi.org/10.5753/cbie.sbie.2020.1092.

SACCARO, A.; FRANÇA, M. T. A.; JACINTO, P. A. 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. Estudos Econômicos (São Paulo), v. 49, p. 337-373, 2019.

SANTOS, C. H. D. C.; MARTINS, S. L.; PLASTINO, A. É Possível Prever Evasão com Base Apenas no Desempenho Acadêmico?. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 32., 2021, Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 792-802. DOI: https://doi.org/10.5753/sbie.2021.218105.

SANTOS, V. H. B.; SARAIVA, D. V.; OLIVEIRA, C. T. Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 32, 2021, Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 1196-1210. DOI: https://doi.org/10.5753/sbie.2021.218167.

SARAIVA, D.; PEREIRA, S.; GALLINDO, E.; BRAGA, R.; OLIVEIRA, C. Uma proposta para predição de risco de evasão de estudantes em um curso técnico em informática. In: Anais do XXVII Workshop sobre Educação em Computação. SBC, 2019. p. 319-333.

SMOLA, A. J.; BARLETT, P.; SCHOLKOPF, B.; SCHUURMANS, D. *Introduction to Large Margin Classifiers*, chapter 1, pages 1–28. 1999.

SUKHBAATAR, O.; OGATA, K.; USAGAWA, T. *Mining educational data to predict academic dropouts: a case study in blended learning course*. In: TENCON 2018-2018 IEEE region 10 conference. IEEE, 2018. p. 2205-2208.

TEODORO, L. A.; KAPPEL, M. A. A. 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 Informática na Educação (RBIE), [S.l.], v. 28, p. 838-863, nov. 2020. ISSN 2317-6121. Disponível em: <http://br-ie.org/pub/index.php/rbie/article/view/v28p838>. Acesso em: 28 mar. 2022. doi: http://dx.doi.org/10.5753/rbie.2020.28.0.838.

VERIKAS, A.; GELZINIS, A.; BACAUSKIENE, M. *Mining data with random forests: A survey and results of new tests*. Pattern recognition, v. 44, n. 2, p. 330-349, 2011.