Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus João Pessoa

Programa de Pós-Graduação em Tecnologia da Informação

Nível Mestrado Profissional

ISLEIMAR DE SOUZA OLIVEIRA

ANÁLISE DE DADOS APLICADA À EVASÃO ESCOLAR: UM ESTUDO DE CASO DO IFPB

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA

2023

Isleimar de Souza Oliveira

ANÁLISE DE DADOS APLICADA À EVASÃO ESCOLAR: Um estudo de caso do IFPB

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
Coorientador: Prof. Dr. Fabio Gomes de Andrade

João Pessoa

2023

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

048a Oliveira, Isleimar de Souza.

Análise de dados aplicada à evasão escolar : um estudo de caso do IFPB / Isleimar de Souza Oliveira. – 2023.
136 f. : il.
Dissertação (Mestrado em Tecnologia da Informação) – Instituto Federal de Educação da Paraíba / Programa de Pós-Graduação em Tecnologia da Informação, 2023.
Orientação: Prof. D.r Francisco Petrônio Alencar de Medeiros.
Coorientação: Prof. D.r Fabio Gomes de Andrade.

1. Predição de evasão escolar. 2. Seleção de atributos preditivos. 3. Mineração de dados. 4. Alunos – IFPB. 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

ISLEIMAR DE SOUZA OLIVEIRA

ANÁLISE DE DADOS APLICADA À EVASÃO ESCOLAR: Um estudo de caso do IFPB

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 02 de agosto de 2023

Membros da Banca Examinadora:

Dr. Francisco Petrônio Alencar de Medeiros

IFPB - PPGTI

Dr. Diego Ernesto Rosa Pessoa

IFPB - PPGTI

Dr. Fábio Gomes de Andrade

IFPB

Dra. Thaís Gaudêncio do Rego

UFPB

gov.br
Documento assinado digitalmente
THAIS GAUDENCIO DO REGO
Data: 30/08/2023 07:43:15-0300
Verifique em https://validar.iti.gov.br

João Pessoa/2023

Documento assinado eletronicamente por:

- **Francisco Petronio Alencar de Medeiros**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 02/08/2023 18:26:04.
- **Diego Ernesto Rosa Pessoa**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 02/08/2023 20:34:24.
- **Fabio Gomes de Andrade**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 06/08/2023 15:54:43.

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 Verificador: 449757
Código de Autenticação: d6b6ef3ae7

Dedico este trabalho à minha mãe Eugênia de Souza Oliveira (in memoriam), mulher guerreira e de fibra cujo empenho em me educar sempre veio em primeiro lugar. Sei que, apesar de não estar presente fisicamente, ilumina os meus passos e orienta as minhas decisões, aqui estão os resultados dos seus esforços.

AGRADECIMENTOS

Agradeço a todos os professores que me acompanharam durante essa jornada, em especial ao Prof. Dr. Francisco Petrônio Alencar de Medeirose ao Prof. Dr. Fabio Gomes de Andrade, pelo suporte e orientação fornecidos durante todo o processo de realização do mestrado. Sem a orientação de vocês, eu não teria conseguido alcançar meus objetivos, vocês foram incansáveis em sua dedicação, fornecendo feedbacks construtivos e apoiando-me em cada etapa do processo. Agradeço pelo tempo dedicado e pela disponibilidade em responder minhas dúvidas e pela ajuda na resolução dos desafios encontrados ao longo do caminho.

RESUMO

O desafio global da evasão escolar impõe dificuldades às instituições educacionais, causando impactos negativos em alunos, comunidades e nos próprios estabelecimentos. Para contrapor essa questão, soluções preventivas direcionadas aos alunos em risco de evasão se tornam vitais, demandando a identificação antecipada de características vinculadas a esse fenômeno. Neste contexto, esta pesquisa propõe uma abordagem preditiva da evasão escolar, empregando dados educacionais do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba (IFPB). O objetivo primordial é explorar a interrelação entre diferentes características dos alunos, como gênero, renda familiar, idade, turno e origem escolar, para identificar os determinantes da evasão escolar em distintos contextos. A metodologia adotada engloba a análise das bases de dados do Sistema Unificado de Administração Pública (SUAP) e da Plataforma Nilo Peçanha (PNP) relacionadas aos cursos do IFPB. O enfoque consiste na determinação de atributos para a previsão da evasão, empregando algoritmos de classificação, tais como Árvore de Decisão, Floresta Aleatória, Naive Bayes, Multilayer Perceptron e SVM. A avaliação dos resultados é conduzida por meio da métrica F1-Score. Foram conduzidos testes exaustivos, considerando diferentes seletores de atributos e quantidades variáveis de características selecionadas. Os resultados alcançados revelam insights sobre padrões e tendências da evasão escolar em cada agrupamento de cursos e características dos alunos. O conjunto de dados do SUAP demonstrou resultados mais substanciais na métrica F1-Score, variando entre 0,84 e 0,98, enquanto o conjunto da PNP registrou oscilações entre 0,58 e 0,94. Essa disparidade ressalta a necessidade de abordagens distintas para cada contexto. Com base nas conclusões, as análises fornecem uma visão ampla dos padrões e tendências da evasão escolar em diversos agrupamentos de cursos e características dos alunos. Tais análises podem servir como base para futuras pesquisas, contribuindo para o aperfeiçoamento contínuo das estratégias de prevenção e enfrentamento da evasão escolar. A compreensão aprofundada dos fatores que influenciam a evasão em cenários diversos promove o desenvolvimento de políticas educacionais mais embasadas, fomentando uma educação inclusiva e acessível para todos os alunos.

Palavras-chaves: Predição de evasão escolar; seleção de atributos preditores; mineração de dados.

ABSTRACT

The global challenge of school dropout poses difficulties for educational institutions, causing negative impacts on students, communities, and the institutions themselves. To address this issue, preventive solutions targeted at students at risk of dropping out become vital, requiring early identification of characteristics linked to this phenomenon. In this context, this research proposes a predictive approach to school dropout, using educational data from the Federal Institute of Education, Science, and Technology of Paraíba (IFPB). The primary objective is to explore the interrelation between different student characteristics such as gender, family income, age, schedule, and school origin to identify determinants of dropout in various contexts. The adopted methodology encompasses the analysis of databases from the Unified Public Administration System (SUAP) and the Nilo Peçanha Platform (PNP) related to IFPB courses. The focus lies in attribute determination for dropout prediction, employing classification algorithms such as Decision Trees, Random Forests, Naive Bayes, Multilayer Perceptron, and SVM. Result evaluation is conducted using the F1-Score metric. Exhaustive tests were performed, considering different attribute selectors and varying quantities of selected features. The attained results unveil insights into patterns and trends of school dropout in each course grouping and student characteristics. The SUAP dataset exhibited more substantial F1-Score metrics ranging from 0.84 to 0.98, while the PNP dataset showed fluctuations between 0.58 and 0.94. This disparity underscores the necessity for distinct approaches to each context. Based on the conclusions, the analyses provide a broad view of school dropout patterns and trends across various course groupings and student characteristics. Such analyses can serve as a foundation for future research, contributing to the continuous enhancement of prevention and intervention strategies for school dropout. A deep understanding of factors influencing dropout across diverse scenarios promotes the development of well-informed educational policies, fostering inclusive and accessible education for all students.

Key-words: School dropout prediction; selection of predictive attributes; data mining.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

ACM	*Association for Computing Machinery*
AUC-ROC	*Area Under the Curve - Receiver Operating Characteristic Curve* (Área Sob a Curva - Curva Característica de Operação do Receptor)
AVA	Ambiente Virtual de Aprendizagem
BN	*Bayesian network* (Rede Bayesiana)
CE	Critérios de Exclusão
CENSUP	Censo da Educação Superior
CI	Critérios de Inclusão
COINTI-RE	Coordenação de Inovação de Tecnologia da Informação da Reitoria do IFPB
CSV	*Comma-Separated Values* (Valores Separados Por Virgula)
DM	*Data Mining* (Mineração de Dados)
DT	*Decision Tree* (Árvore de Decisão)
EDM	*Educational Data Mining* (Mineração de Dados Educacionais)
ETL	*Extract, Transform and Load* (Extrair, Transformar e Carregar)
FIC	Formação Inicial e Continuada
FN	Falso Negativo
FP	Falso Positivo
IA	Inteligência Artificia
IEEE	*Institute of Electrical and Electronics Engineers*
IEs	Instituições de Ensino
IFPB	Instituto Federal de Ciência e Tecnologia da Paraíba
IFRN	Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte
IFs	Instituições Federais
INEP	Instituto Nacional de Estudo e Pesquisa Educacional Anísio Teixeira

KDD	*Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bancos de Dados)
KNN	*K-nearest neighbors* (K-Vizinhos Mais Próximos)
MDE	Mineração de Dados Educacionais
ML	*Machine Learning* (Aprendizado de Máquina)
MLP	*Multilayer Perceptron* (Perceptron Multicamadas)
NB	*Naive Bayes* (Baías Ingénuas)
PNP	Plataforma Nilo Peçanha
QP	Questões de Pesquisa
RFP	Renda Familiar *Per Capita*
RBIE	Revista Brasileira de Informática na Educação
RF	*Random Forest*
RL	Regressão Logística
RMSE	*Root Mean Square Error* (Raiz do Erro Quadrático Médio)
ROC	*Receiver Operating Characteristic Curve* (Curva característica de operação do receptor)
RSL	Revisão Sistemática da Literatura
SBIE	Simpósio Brasileiro de Informática na Educação
SIGA	Sistemas de Gestão Acadêmica
SMOTE	*Synthetic Minority Oversampling Technique* (Técnica de Sobreamostragem Minoritária Sintética)
STI	Sistema Tutor Inteligentes
SUAP	Sistema Unificado de Administração Pública
SVM	*Support Vector Machine* (Máquina de Vetores de Suporte)
TFP	Taxa de Falsos Positivos
TI	Tecnologias da Informação
TIC	Tecnologias da Informação e Comunicação

TVP	Taxa de Verdadeiro Positivos
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

A educação tem papel fundamental no processo de transformação da realidade cotidiana do aluno, propiciando igualdade social e desenvolvimento, melhorando os aspectos moral, intelectual e material, tanto para o indivíduo, quanto para o ambiente que o cerca (OLIVEIRA et al., 2013). As Instituições Federais (IFs) cumprem essa função social de educar, garantindo a oferta de cursos aos alunos. A evasão escolar é um dos grandes desafios, afetando não apenas os próprios alunos, mas suas comunidades, e isso se reflete inclusive na redução de recursos destinados aos próprios campi, pois há uma relação entre esses recursos e o número de alunos matriculados (PRIM; FÁVERO, 2013).

Segundo os dados fornecidos pelo Instituto Nacional de Estudo e Pesquisa Educacional Anísio Teixeira (INEP), em sua Sinopse Estatística da Educação Superior do ano de 2019, do total de 8.603.824 alunos vinculados aos cursos de graduação presencial e à distância no Brasil, excetuando os 1.229 alunos falecidos, um total de 3.744.522 alunos evadiram de seus cursos, o que representa aproximadamente 43,5% dos alunos matriculados (INEP. . . , 2021).

A alta taxa de evasão escolar é um problema que afeta instituições públicas e privadas de ensino no Brasil e no mundo. Pesquisas realizadas para identificar as causas da evasão escolar mostram que questões pessoais, institucionais ou gerais contribuem de forma relevante para que alunos decidam por abandonar seus estudos (AQUINO, 1997). Uma forma de combater a evasão escolar e, consequentemente, os problemas causados por ela, consiste em identificar de forma antecipada os perfis de alunos com risco potencial de evadir, o que auxilia gestores de políticas educacionais na tomada de decisão e na elaboração de planos para evitar o abandono escolar (ROMERO; VENTURA, 2010).

Nos últimos anos, vários estudos têm sido desenvolvidos na área de Tecnologia da Informação (TI) para reconhecer características que contribuem para o problema da evasão escolar, tendo a Mineração de Dados contribuído com resultados promissores (KAUR; SINGH; JOSAN, 2015). Existem trabalhos que descrevem um sistema capaz de relacionar as dimensões utilizadas em modelos preditivos (HAN; KAMBER; PEI, 2011), outros utilizam técnicas de aprendizagem de máquina para melhorar a performance acadêmica dos estudantes (THAI-NGHE; BUSCHE; SCHMIDT-THIEME, 2009), ou aplicam modelos de predição para prever o desempenho acadêmico dos alunos (RAMASWAMI; BHASKARAN, 2010).

A Mineração de Dados tem como objetivo realizar a extração de conhecimento a partir da análise de um conjunto de dados volumoso, mediante técnicas estatísticas e algoritmos que buscam por padrões presentes nos dados (PAZ; CAZELLA, 2017). A aplicação da Mineração de Dados no contexto da Educação é conhecida como Mineração de Dados Educacionais (MDE) (ROMERO; VENTURA, 2010).

Apesar de existirem atributos pré-definidos para modelos de classificação de evasão escolar, como idade, gênero, fator socioeconômico, notas em disciplinas específicas, entre outros, esses atributos são genéricos e não apresentam os melhores resultados para todos os programas dos cursos, pois o conjunto de áreas de conhecimento das disciplinas mudam de acordo com as matriz dos cursos, necessitando variar os atributos para contextos específicos (SIEBRA; SANTOS; LINO, 2020). Dessa forma, não é possível afirmar que a utilização desse conjunto de características garante os melhores resultados para os conjuntos de atributos em modelos preditivos aplicados aos cursos disponibilizados em todas as Instituições de Ensino (IEs). Isso acontece porque, mesmo que alguns fatores possam ser decisivos na permanência de um aluno em um curso, não obrigatoriamente esses fatores são relevantes para todos os alunos em todos os cursos.

Uma forma de construir ferramentas que auxiliem os gestores das IEs a executarem ações que favoreçam a conclusão dos alunos dos cursos está associada à identificação de características e perfis de alunos com propensão para evadir. A redução dos índices de evasão propicia o desenvolvimento pessoal e comunitário para o aluno, melhorando a eficiência dos recursos investidos. A seleção de atributos como forma de identificar os atributos mais relevantes, para o conjunto de dados produzidos no processo educacional, constitui o primeiro passo para construção de ferramentas que auxiliem na redução da evasão escolar.

A Plataforma Nilo Peçanha (PNP) é um importante recurso, que reúne dados acadêmicos e estatísticas oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica. Por meio da PNP, é possível acessar os *Microdados*, que englobam informações detalhadas sobre o corpo docente, discente, técnico-administrativo e os gastos financeiros das unidades da Rede Federal. Esses Microdados fornecem uma visão abrangente e aprofundada dos diversos aspectos relacionados às instituições educacionais, permitindo uma análise minuciosa das características e dinâmicas presentes na Rede Federal. Esses Microdados são um conjunto valioso de informações acadêmicas que podem ser exploradas para análises e estudos relacionados ao desempenho e à gestão educacional. Desde o ano de 2017, a Plataforma Nilo Peçanha passou a armazenar e disponibilizá-los, fornecendo uma visão abrangente e detalhada sobre a educação profissional, científica e tecnológica nas instituições da Rede Federal. Essas informações são coletadas, validadas e organizadas, permitindo a extração de conhecimentos e a realização de análises mais aprofundadas sobre a evasão escolar, desempenho acadêmico, distribuição de recursos, entre outros aspectos relevantes para a gestão educacional.

Ao utilizar os dados da Plataforma Nilo Peçanha, pesquisadores, gestores educacionais e profissionais da área têm acesso a informações detalhadas e atualizadas, que permitem a realização de estudos e a implementação de políticas mais embasadas e efetivas. Esses dados acadêmicos são valiosos recursos para a tomada de decisões no contexto educacional, contribuindo para o desenvolvimento e aprimoramento das instituições de ensino da Rede Federal, bem como para o avanço do campo de análise de dados educacionais.

Esse conjunto de informações disponibilizadas pela Plataforma Nilo Peçanha é relevante para instituições como o Instituto Federal de Ciência e Tecnologia da Paraíba (IFPB), que oferece cursos de educação profissional e tecnológica, educação básica, superior e pós-graduação, nos níveis técnico (integrado, concomitante e subsequente), bacharelado, licenciatura, tecnologia, *lato sensu*, *stricto sensu*. Utilizando esses dados como alicerce, o IFPB reforça sua habilidade em formar e qualificar cidadãos, alavancando o desenvolvimento econômico, social e humano em diversas regiões do estado da Paraíba. Por meio do entendimento aprofundado dos padrões de evasão e dos fatores que influenciam esse fenômeno, a instituição está melhor equipada para direcionar esforços em estratégias personalizadas de intervenção, garantindo que os alunos possam concluir suas trajetórias educacionais de forma bem-sucedida. O IFPB possui 188 cursos de educação profissional e tecnológica, educação básica, superior e pós-graduação, nas modalidades de educação presencial e a distância e ainda educação de jovens e adultos, ofertando educação profissional e tecnológica nos diferentes níveis e modalidades de ensino, formando e qualificando cidadãos para atuação profissional, promovendo o desenvolvimento econômico, social e humano, tendo alcance de abrangência local, regional e nacional.

O IFPB faz uso da plataforma Sistema Unificado de Administração Pública (SUAP), desenvolvida pelo Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN), composta por módulos que gerenciam diversos processos necessários ao funcionamento do Instituto. Entre esses módulos existem os responsáveis pela gestão dos Processos Administrativos, Gestão de Pessoas, Tecnologia da Informação, Central de Serviços e o módulo de Gestão Acadêmica, responsável por gerenciar as informações relacionadas à vida acadêmica dos alunos. Dessa forma, o SUAP possui dados que podem ser utilizados em processos de Mineração de Dados Educacionais (MDE), para identificar perfis de alunos com risco de evadir.

Segundo informações disponíveis no SUAP, o IFPB possui (em 2020) um total de 136.312 registros de alunos, com matrícula em 486 cursos, e 21 *Campi* distribuídos em todo o território paraibano, desde João Pessoa, no litoral até Cajazeiras, no alto sertão paraibano, conforme apresentado no Plano de Desenvolvimento Institucional do IFPB (IFPB..., 2020). Essa variedade em cursos e a extensão geográfica presente no IFPB propicia uma diversidade de contextos e características, que podem ser aproveitados para análises dos dados educacionais utilizando diversas configurações de agrupamentos dos dados.

1.1 Justificativa

A evasão escolar é um desafio enfrentado por instituições de ensino em todo o mundo, com consequências negativas para os alunos, a comunidade e a própria instituição. A fim de lidar com esse problema, é fundamental adotar medidas preventivas que visem a identificação precoce dos alunos em risco de evasão, assim como a implementação de estratégias de intervenção adequadas. Nesse contexto, a análise de dados educacionais e a aplicação de técnicas de mineração de dados e aprendizado de máquina podem fornecer *insights* para a identificação de fatores de

risco e a tomada de decisões embasadas.

A presente pesquisa propõe uma abordagem para a predição de evasão escolar, por meio da análise de dados educacionais do Instituto Federal de Educação Ciência e Tecnologia da Paraíba (IFPB). Ao explorar a base de dados do módulo de controle acadêmico do Sistema Unificado de Administração Pública (SUAP), bem como dados disponíveis no portal de dados abertos da Plataforma Nilo Peçanha (PNP), será possível identificar os atributos mais relevantes para a predição de evasão. A escolha do IFPB como estudo de caso é relevante, uma vez que é uma instituição de ensino com um grande número de alunos e uma diversidade de cursos, tanto técnicos quanto superiores. Ao utilizar uma gama de atributos, a pesquisa busca fornecer uma análise abrangente e significativa para diferentes contextos educacionais.

Além disso, a utilização de algoritmos de classificação, como Árvore de Decisão, Floresta Aleatória, *Naive Bayes*, *Multilayer Perceptron* e SVM (em inglês, *Support Vector Machine*), permitirá a comparação e a seleção dos modelos mais eficazes para a predição de evasão. As métricas de desempenho *F1-Score* foram empregadas para avaliar a precisão e a eficácia dos modelos propostos. Os resultados esperados deste estudo têm o potencial de contribuir para a compreensão dos fatores que influenciam a evasão escolar no contexto do IFPB, bem como para o desenvolvimento de estratégias de intervenção mais efetivas. Essas descobertas poderão ser aplicadas também em outras instituições de ensino, auxiliando no desenvolvimento de políticas e programas de combate à evasão escolar em âmbito nacional.

Portanto, a presente pesquisa justifica-se pela importância de abordar o problema da evasão escolar. Além disso, ressalta-se a relevância da aplicação de técnicas de mineração de dados e aprendizado de máquina para a identificação de fatores de risco. A pesquisa também atende à necessidade de fornecer subsídios para a implementação de medidas preventivas e intervenções adequadas, com o objetivo de melhorar o desempenho acadêmico e a permanência dos alunos nas instituições de ensino.

1.2 Delimitação do trabalho

Esta pesquisa se concentrou especificamente em analisar o fenômeno da evasão escolar entre os estudantes do IFPB, abrangendo tanto os cursos técnicos, quanto os cursos superiores. O foco foi compreender os fatores que influenciam a evasão escolar nessa instituição específica. A análise foi direcionada para identificar os alunos em risco de evasão, buscando identificar os atributos mais relevantes que permitem prever esse comportamento com maior precisão. No que diz respeito aos algoritmos de classificação, este trabalho se concentra na utilização de Árvore de Decisão, Floresta Aleatória, *Naive Bayes*, *Multilayer Perceptron* e SVM como modelos de predição. A seleção desses algoritmos foi fundamentada na sua reconhecida relevância e eficácia em problemas de classificação, como apontado pela pesquisa bibliográfica realizada durante a execução do trabalho. Além disso, a métrica de desempenho utilizada para avaliar os modelos foi

a métrica *F1-Score*, que é amplamente aceita e utilizada para avaliar o desempenho de modelos de classificação.

Vale ressaltar que este trabalho não aborda a implementação de estratégias de intervenção para prevenir a evasão escolar identificada. O objetivo principal é fornecer uma análise robusta dos dados educacionais, identificando os atributos mais relevantes e avaliando o desempenho dos modelos de predição propostos.

Por fim, é importante destacar que os resultados e conclusões obtidos neste estudo são aplicáveis especificamente ao contexto do IFPB, embora as metodologias e abordagens utilizadas possam ser relevantes e adaptáveis para outras instituições de ensino, que enfrentam desafios semelhantes relacionados à evasão escolar. Portanto, esta pesquisa delimita-se à análise dos dados educacionais do IFPB, à identificação de atributos relevantes para a predição de evasão escolar, ao uso de algoritmos de classificação selecionados e à avaliação de desempenho por meio da métrica *F1-Score*.

1.3 Objetivos

O objetivo geral deste trabalho é analisar os dados educacionais do Instituto Federal de Educação Ciência e Tecnologia da Paraíba, especificamente utilizando as bases da Plataforma Nilo Peçanha e do Sistema Unificado de Administração Pública, com o intuito de identificar os atributos relevantes na predição de evasão escolar através da análise de dados educacionais do IFPB, utilizando as bases da plataforma Nilo Peçanha e SUAP.

1.3.1 Objetivos específicos

Para alcançar esse objetivo geral, os seguintes objetivos específicos foram estabelecidos:

- Realizar uma revisão sistemática da literatura sobre técnicas de mineração de dados aplicadas à predição de evasão escolar, a fim de obter um panorama atualizado das abordagens e metodologias utilizadas nessa área.
- Coletar e pré-processar os dados educacionais do IFPB, incluindo a base de dados do módulo de controle acadêmico do Sistema Unificado de Administração Pública (SUAP) e os dados disponíveis no portal de dados abertos da PNP.
- Realizar análises quantitativas dos dados coletados, visando identificar padrões e tendências relacionados à evasão escolar no contexto do IFPB.
- Utilizar técnicas de seleção de atributos para identificar os atributos mais relevantes na predição de evasão escolar, com base nos dados educacionais analisados.

- Aplicar os algoritmos de classificação, incluindo Árvore de Decisão, Floresta Aleatória, *Naive Bayes*, *Multilayer Perceptron* e SVM, para avaliar o desempenho desses modelos na predição de evasão escolar.
- Utilizar uma análise para identificação de solução única, onde o mesmo algoritmo de classificação, ferramenta de seleção de atributos e quantidade de atributos serão aplicados, visando a predição de evasão escolar e utilizando a métrica *F1-Score*.
- Analisar os resultados obtidos comparando os desempenhos alcançados pelas bases de dados da PNP e do SUAP, visando identificar possíveis variações nos resultados para diferentes contextos educacionais.

1.4 Estrutura do Documento

Além da parte introdutória, este documento está organizado da seguinte forma. O segundo capítulo aborda a fundamentação teórica, descrevendo as tecnologias utilizadas no desenvolvimento do trabalho. No terceiro capítulo é realizada a Revisão Sistemática da Literatura. No quarto capítulo é descrita a solução proposta neste trabalho, junto com a metodologia utilizada. No quinto capítulo são expostos os resultados alcançados. Por fim, o sexto capítulo apresenta a conclusão e as considerações finais.

2 FUNDAMENTAÇÃO TEÓRICA

A complexidade do problema da evasão na educação prejudica a identificação de suas causas, inclusive até mesmo de sua definição, pois, as múltiplas formas de interpretação não permitem chegar a uma definição precisa de "evasão e abandono escolar", uma vez que essa requer uma compreensão das relações entre os motivos de ingresso e a trajetória dos permanecentes, dos desistentes e egressos desse público (FILHO; ARAÚJO, 2017). As inovações no âmbito das Tecnologias da Informação e Comunicação (TIC), aplicadas aos processos de ensino e aprendizagem em Ambiente Virtual de Aprendizagem (AVA), Sistema Tutor Inteligentes (STI) e Sistemas de Gestão Acadêmica (SIGA), possibilitam registrar dados em maior quantidade e, assim, extrair informações que auxiliam na identificação dos fatores que interferem na permanência dos alunos (COSTA et al., 2012). Todo esse volume de dados tem ensejado a realização de pesquisas com o propósito de descobrir padrões relacionados à evasão no contexto educacional (GOTTARDO; KAESTNER; NORONHA, 2012).

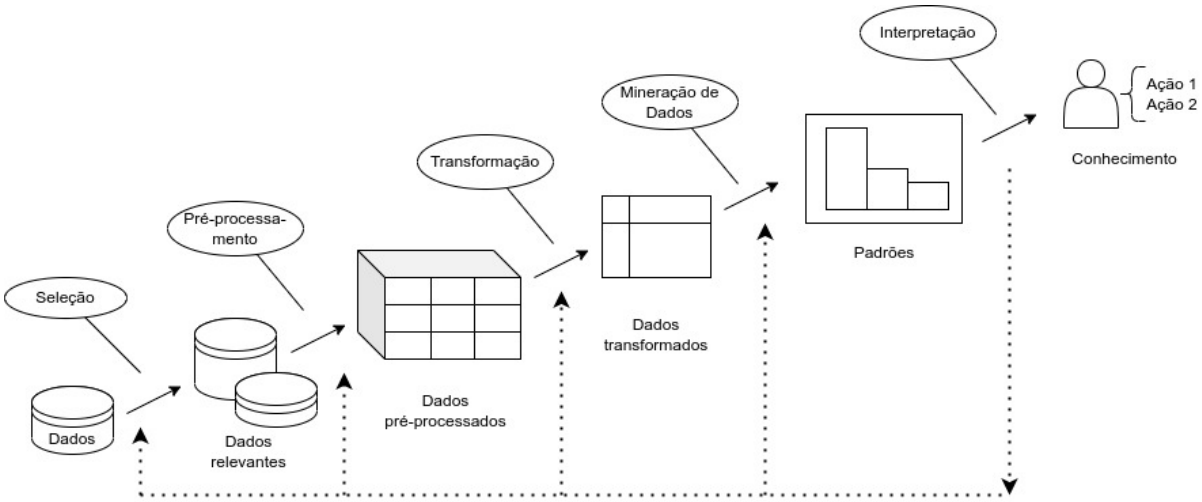
De acordo com Marr (2015), o aumento exponencial da geração de dados é um fenômeno amplamente reconhecido na era contemporânea. O autor observa que, em todo o mundo, o volume de informações armazenadas em uma variedade diversificada de processos tem apresentado um crescimento significativo, dobrando a cada período aproximado de 20 meses. Essa rápida expansão da quantidade de dados é resultado da proliferação de tecnologias digitais, sistemas de informação e dispositivos conectados. Essa tendência tem implicações profundas no cenário empresarial e na sociedade em geral, uma vez que esse acúmulo constante de dados oferece oportunidades sem precedentes para a obtenção de *insights* valiosos e embasados em análises, direcionando, assim, as decisões e melhorando a performance em diversos domínios (MARR, 2015).Porém, o crescente aumento da quantidade de dados gerados pode, paradoxalmente, prejudicar a proporção de dados analisados devido, entre outras coisas, à diversidade existente nos formatos destes dados (GAMA et al., 2019).

2.1 Descoberta de Conhecimento em Banco de Dados

O crescente aumento da quantidade de dados disponíveis atualmente torna a sua análise manual impraticável. Como forma de resolver esse problema, Fayyad, Piatetsky-Shapiro e Smyth (1996) descrevem um modelo de processo de Descoberta de Conhecimento em Banco Dados (do inglês *Knowledge Discovery in Databases*, KDD). A Figura 1 apresenta as etapas que compõem o processo de KDD. O KDD tem o propósito de extrair informações em grandes conjuntos de dados buscando padrões explicáveis nestes, permitindo a sua interpretação e extrapolação para eventos futuros. O processo de KDD extrai conhecimento implícito, a partir de um conjunto de dados volumoso, possibilitando obter *insights* que auxiliam em tomadas de decisões (BAKER et

al., 2011).

Figura 1 – Visão geral das etapas que compõem o processo de KDD.



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

No entanto, é importante ressaltar que no processo de KDD, é possível ocorrer a necessidade de retornar a etapas anteriores, em função das descobertas realizadas em cada uma delas (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Isso significa que as etapas do processo são iterativas e interativas, permitindo uma abordagem flexível e adaptável, à medida que novas informações são reveladas. A seguir, serão apresentadas e descritas em detalhes as etapas que compõem o processo de KDD:

- **Etapa de Seleção de Dados**: esta é a primeira etapa do processo de KDD. Nela são realizadas a identificação da origem das informações e seleção das características que serão utilizadas na etapa de Mineração de Dados, de acordo com a identificação do objetivo, a partir do ponto de vista do cliente;

- **Pré-processamento dos Dados**: nesta etapa tem-se como objetivo assegurar a qualidade dos dados selecionados, que serão utilizados nas etapas seguintes. Para tanto, são aplicadas técnicas de análise, com o propósito de identificar valores ausentes, valores incorretos, *outliers*, ou qualquer tipo de inconsistência presente nos dados, que possa distorcer os resultados. A estratégia mais básica utilizada, quando detectada alguma inconsistência nos dados, é a exclusão da instância. Porém, essa abordagem pode gerar problemas nos casos de base de dados pequena, ou mesmo quando grandes quantidades são excluídas, gerando distorções nestes dados. Para esses casos, técnicas mais elaboradas são usadas com o intuito de corrigir, ou até mesmo atribuir valores ausentes, aplicando análises estáticas, ou mesmo predição, com base nos dados presentes, mantendo assim, maior quantidade possível de instâncias;

- **Transformação dos Dados**: na etapa de transformação tem-se como objetivo converter os dados em modelos que facilitam a sua análise. Algumas das transformações possíveis nessa etapa são a discretização, na qual variáveis contínuas são convertidas em categóricas; a redução de dimensionalidade, quando variáveis são combinadas para reduzir a quantidade de variáveis e, por fim, a criação de novas variáveis, através de análise de agrupamentos;
- **Mineração de Dados**: essa etapa é o cerne do processo do KDD, consistindo basicamente na aplicação de modelos, métodos estatísticos, critérios e algoritmos de busca para extrair padrões inerentes aos dados e relacionados ao interesse da pesquisa. Essa etapa é descrita com mais detalhes na seção 2.2;
- **Interpretação dos Resultados**: nesta etapa são apresentadas as descobertas obtidas, para que seja possível determinar as melhores estratégias para a tomada de decisão. Após essa etapa, também é possível decidir por mudanças na abordagem das etapas anteriores e, se necessário, refazer as partes do processo.

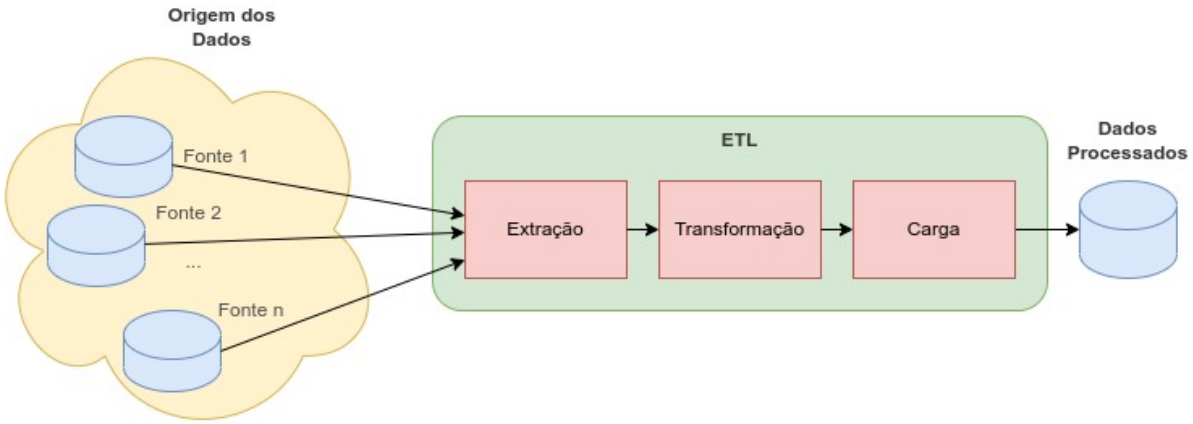
Todo o início do KDD está associado, em geral, a uma grande base de dados relacionada ao objetivo analisado. A sua saída é direcionada à tomada de decisão, em decorrência do conhecimento adquirido e, por se tratar de um processo cíclico, este será redefinido e melhorado após a análise de sua saída (MAIMON; ROKACH, 2010).

2.2 Extração, Transformação e Carga dos Dados

Após definir a origem dos dados, a próxima etapa do KDD é a aquisição desses dados, que será realizada por meio do processo de Extração, Transformação e Carga (ETL, do inglês *Extract, Transform, and Load*). O ETL é composto por três etapas distintas, conforme ilustrado na Figura 2. Na primeira etapa, ocorre a extração dos dados, podendo ser provenientes de diversas fontes. Em seguida, temos a etapa de transformação, na qual os dados são modificados utilizando técnicas de higienização, padronização, filtragem e criação de dicionário. Por fim, na etapa de carga, os dados são armazenados de forma persistente em uma base de dados (FERREIRA et al., 2010).

Essa etapa de ETL desempenha um papel crucial na preparação dos dados para análise posterior. Ao extrair os dados de diferentes fontes, transformá-los para garantir sua qualidade e consistência, e carregá-los em uma estrutura adequada, é possível obter dados limpos, integrados e prontos para serem explorados nas etapas subsequentes do processo de KDD. A qualidade e a eficácia da etapa de ETL têm um impacto significativo na qualidade dos *insights* e descobertas obtidos no processo de análise de dados. Portanto, uma execução cuidadosa e precisa do ETL é fundamental para buscar resultados mais confiáveis e úteis na descoberta de conhecimento (FERREIRA et al., 2010).

Figura 2 – Extração, Transformação e Carga.



Fonte: Adaptado de Ferreira et al. (2010).

2.3 Mineração de Dados

Ao longo da história, o ser humano sempre se dedicou à análise de dados com o intuito de descobrir padrões e obter informações. No entanto, com o aumento exponencial do volume e da velocidade de geração de dados nos dias atuais, a análise manual exclusiva tornou-se impraticável. Diante desse cenário, surgiram processos e técnicas que são capazes de auxiliar na análise desse imenso volume de dados. Um dos principais processos para extrair informações de conjuntos de dados extensos é a Mineração de Dados (ou *Data Mining*, DM), que desempenha um papel fundamental como parte do processo de Descoberta de Conhecimento em Bancos de Dados. No entanto, a abrangência da Mineração de Dados vai além, extrapolando os limites de áreas como banco de dados, estatística, aprendizado de máquina, computação natural, visualização de dados, recuperação de informações, processamento de imagens e sinais, entre outras (CASTRO; FERRARI, 2016).

Com a aplicação da Mineração de Dados, é possível explorar grandes conjuntos de dados em busca de padrões, tendências e relações ocultas, que podem fornecer informações valiosas para a tomada de decisões e a descoberta de conhecimento. Essa abordagem baseia-se em algoritmos e técnicas avançadas que permitem a identificação de informações relevantes e úteis em meio a uma vasta quantidade de dados. Assim, a Mineração de Dados pode desempenhar um papel crucial na análise eficaz e eficiente de conjuntos de dados cada vez maiores e mais complexos.

A finalidade básica da DM pode ser descrita como a capacidade de detectar relações que sejam úteis. Dessa forma, a sua utilização é amplamente associada à pesquisa científica, como também ao avanço das empresas que visam impulsionar a sua competitividade, devido à vantagem que o conhecimento garante. As tarefas básicas da DM são comumente classificadas em duas categorias: descritiva e preditiva. A categoria preditiva propõe a construção de modelos capazes de prever propriedades ou tendências dos conjuntos de dados. A categoria descritiva

tem como objetivo extrair informações concisas destes (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Tanto nas tarefas preditivas, quanto nas descritivas, a mineração de dados faz uso das técnicas de aprendizagem de máquina.

Quando aplicada no contexto da educação, a Mineração de Dados é conhecida como Mineração de Dados Educacionais (ou *Educational Data Mining*, EDM). Nesse contexto, os dados são provenientes de sistemas educacionais e o objetivo da aplicação é explorar informações sobre comportamentos, avaliações e conquistas dos alunos, bem como configurações e domínio dos conteúdos (COSTA et al., 2012). A MDE pode ser usada para responder a perguntas como:

> O que pode prever o sucesso dos alunos? Qual sequência de cenários é mais eficiente para um aluno específico? Quais são as ações dos alunos que indicam o progresso da aprendizagem? Quais são as características de um ambiente de aprendizagem que permite uma melhor aprendizagem? (PEÑA-AYALA, 2014)

A MDE é estruturada de acordo com o usuário final, seja ele aluno, educador, administrador ou pesquisador (ROMERO; VENTURA, 2010). Essa abordagem possui características distintas que a tornam relevante e adaptável às necessidades específicas de cada usuário. Algumas dessas características são:

- **Aluno**: nesse caso, a EDM visa apoiar reflexões e fornecer *feedbacks* adaptativos que melhoram o desempenho e o processo de aprendizagem;
- **Educador**: nesse caso, a EDM visa fornecer suporte para o processo de aprendizagem e possibilitar uma auto avaliação dos métodos de ensino, para melhorar o desempenho docente;
- **Pesquisador**: nesse caso, a EDM visa desenvolver e comparar técnicas de mineração de dados para avaliar problemas educacionais específicos, com o propósito de tornar mais eficaz o processo de ensino e aprendizagem.

Em determinados cenários, os benefícios da aplicação da MDE se estendem a mais de um usuário. Nesse contexto, Peña-Ayala (2014) identificou diferentes objetivos gerais que podem ser associados ao processo de EDM, abrangendo uma variedade de finalidades e necessidades:

- **Modelagem de aluno**: o objetivo é realizar uma pesquisa detalhada das características e estados dos alunos relacionados ao seu conhecimento, habilidades, motivações, experiências e progresso de aprendizagem.
- **Prever o desempenho dos alunos e os resultados de aprendizagem**: o objetivo é prever as notas finais de um aluno, ou outros tipos de resultados de aprendizagem, com base nos dados das atividades do curso.

- **Gerando recomendação**: o objetivo é recomendar aos alunos qual conteúdo é o mais adequado para eles no momento atual;
- **Analisar o comportamento do aluno**: o objetivo consiste em classificar o aluno em um grupo relacionado às três áreas discutidas anteriormente (modelos de alunos, previsão, geração de recomendação);
- **Comunicação aos interessados**: o objetivo é ajudar os administradores e educadores do curso a analisarem as atividades dos alunos e as informações de uso nos cursos;
- **Análise da estrutura do domínio**: o objetivo é determinar a estrutura de domínio e aperfeiçoar modelos de domínio utilizando a capacidade de prever o desempenho do aluno;
- **Manter e melhorar os cursos**: está relacionado com os dois objetivos anteriores, visando melhorar os cursos usando informações sobre a aprendizagem dos alunos;
- **Apoio pedagógico**: o objetivo é estudar os efeitos de diferentes tipos de apoios pedagógicos fornecidos por softwares de aprendizagem;
- **Avanço científico**: o objetivo é avançar o conhecimento científico sobre a aprendizagem e os alunos, através da construção de modelos do aluno, do domínio e do suporte pedagógico.

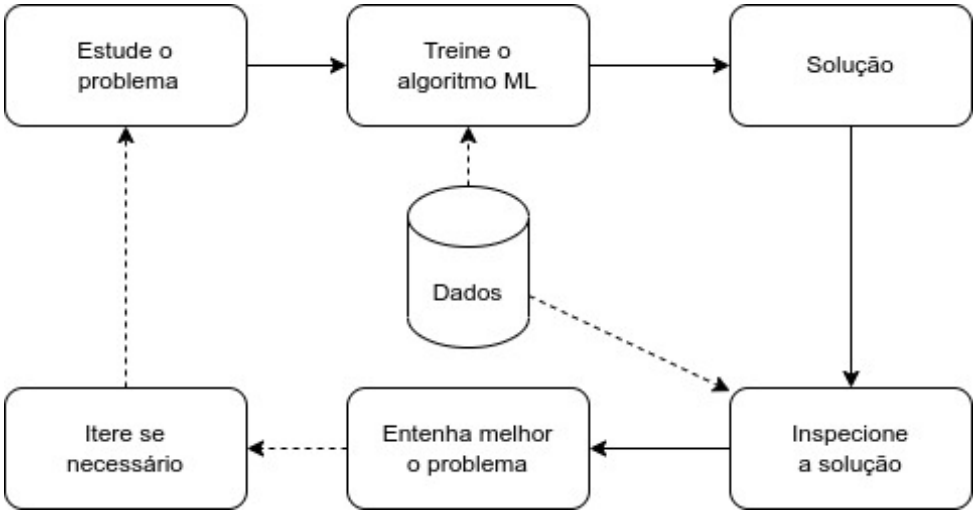
Neste trabalho, a aplicação da MDE terá como foco principal a previsão da evasão escolar, estando relacionado assim com "*Prever o desempenho dos alunos e os resultados de aprendizagem*". Será empregada uma abordagem analítica para explorar os dados educacionais, buscando identificar padrões e tendências, que possam contribuir para uma melhor compreensão do progresso acadêmico dos alunos e para identificar características relacionadas a evasão escolar.

2.4 Aprendizagem de Máquina

A Aprendizagem de Máquina (no inglês *Machine Learning*, ML) permite que programas de computador modifiquem seu comportamento sem a necessidade de serem explicitamente programados, usando como base resultados alcançados anteriormente. Tais comportamentos são guiados pela análise amostral dos dados, tendo flexibilidade nas tomadas de decisão, sem a necessidade de se seguir regras inflexíveis e instruções predefinidas. Essa flexibilidade permite que a solução seja constantemente revista e adequada a novas configurações que se façam necessárias (GERON, 2019). A Figura 3 mostra o fluxo do processo de treinamento da Aprendizagem de Máquina, onde os dados são utilizados, tanto na etapa de treinamento, quanto para avaliar os resultados fornecidos pelo modelo.

Dada a sua flexibilidade, a ML tem sido intensamente utilizada na Mineração de Dados para extrair informações de modo automatizado, a partir de bases de dados volumosas, suscitando

Figura 3 – Diagrama de treinamento da Aprendizagem de Máquina.



Fonte: Adaptado de Geron (2019)

também, grande interesse da área de Inteligência Artificial (IA) (RUSSEL; NORVIG, 2013; ARTERO, 2009). A IA visa entender e descrever a inteligência humana em suas mais diversas abordagens de interpretação, ao mesmo tempo que busca replicá-la em agentes inteligentes capazes de processar raciocínios e comportamentos similares aos realizados pelos seres humanos. De acordo com Russel e Norvig (2013), um agente inteligente é tudo que é capaz de perceber seu ambiente, por meio de sensores, e de agir sobre esse ambiente, por intermédio de atuadores, como também sistemas dotados de Inteligência Artificial capazes de que executar funções, mapear sequências de percepções do ambiente e planejar ações com alcance em ambientes desconhecidos. Para Geron (2019), a ML pode ser classificado de acordo com a supervisão, possuindo quatro categorias principais sendo elas:

- **Aprendizagem supervisionada**: nesta abordagem são apresentados alguns exemplos de pares de entrada e saída, e o agente constrói a associação entre os dados (entradas) e os rótulos (saídas esperadas), estendendo essa associação a outros conjuntos de dados não fornecidos previamente. As tarefas principais da aprendizagem supervisionada são a classificação e a regressão. Enquanto a classificação busca associar itens a uma classe, de acordo com as características observadas, a regressão tenta compreender a relação entre conjunto de dados para prever valores do atributo alvo;
- **Aprendizagem não supervisionada**: nessa categoria o agente aprende os padrões mesmo sem ser fornecido nenhum *feedback* explícito. Essa técnica é mais utilizada em processos de clusterização, onde grupos são formados de acordo com características compartilhadas entres as instâncias;
- **Aprendizagem semi-supervisionada**: para essa categoria os dados são parcialmente rotulados, possuindo poucos dados rotulados, em oposição a uma maioria de dados não

rotulados;

- **Aprendizagem por reforço**: nessa categoria o agente interage repetidamente com um ambiente e aprende a partir de uma série de reforços, recompensas ou punições. O algoritmo chega à solução mais otimizada para o problema por meio de repetidos erros e tentativas.

Os dados desempenham um papel fundamental na Mineração de Dados, sendo o recurso principal para o processo. No entanto, é importante reconhecer que vários desafios podem surgir em relação à qualidade dos dados utilizados no método de treinamento, impactando diretamente nos resultados obtidos. A seguir, destacam-se alguns problemas frequentemente observados na etapa de treinamento dos modelos de aprendizagem de máquina, relacionados à qualidade dos dados disponibilizados:

- **Quantidade insuficiente**: o processo de ML requer uma grande quantidade de dados para que a maioria dos algoritmos funcionem corretamente, mesmo para problemas simples. Desse modo, quanto maior a complexidade do problema, como, por exemplo, no reconhecimento de imagem ou de fala, a quantidade de dados necessita ser ainda maior;
- **Dados não representativos**: é crucial que os dados utilizados no treinamento tenham a maior representatividade possível em relação ao conjunto todo, inclusive para os casos novos. Mesmo amostras muito grandes podem não ser representativas se a forma de selecionar as instâncias para o treinamento for falha;
- **Baixa qualidade dos dados**: os dados podem conter erros, *outliers* ou ruídos. A presença de dados de baixa qualidade prejudica significativamente a eficiência de detecção de padrões. Por essa razão, a etapa de limpeza e tratamento dos dados é tão importante nas etapas anteriores à extração do conhecimento;
- **Características irrelevantes**: a eficiência dos sistemas de ML estão relacionadas diretamente à relevância dos dados utilizados. Por consequência, o processo de seleção, ou descoberta de atributos, é muito importante para os algoritmos de ML. Na seção 2.5 são detalhadas algumas técnicas de descobertas de atributos;
- **Sobreajuste dos dados**: no contexto da ML, ocorre o sobreajuste quando um modelo apresenta um desempenho satisfatório nos dados de treinamento, porém falha em generalizar esses resultados para novos dados não vistos anteriormente. Esse fenômeno frequentemente ocorre quando o modelo se ajusta excessivamente aos ruídos presentes nos dados de treinamento, resultando em um desempenho insatisfatório ao lidar com exemplos não familiares, prejudicando a capacidade de generalização do modelo;
- **Subajuste dos dados**: é o oposto do sobreajuste, e ocorre quando o modelo é muito simples para a aprendizagem da estrutura dos dados, sendo impreciso nos exemplos de

treinamento. As principais opções para resolver esse problema são a seleção de um modelo mais robusto, com melhores características, e a redução de eventuais restrições no modelo.

2.5 Seleção de Atributos

Muitas vezes, o processo de Mineração de Dados trabalha com conjuntos de alta dimensionalidade. O número excessivo de dimensões no conjunto de dados leva ao aumento da complexidade do modelo, dificultando a sua interpretação, aumentando o tempo no processo de treinamento ou impactando na precisão das previsões (LIU; MOTODA, 2007). Técnicas de seleção de atributos (no inglês, *Feature Selection*) eliminam características irrelevantes, reduzindo a complexidade do modelo resultante. O objetivo final é a obtenção de um modelo mais simples possível, que seja mais rápido de processar e não prejudique a qualidade preditiva. Assim, a seleção de características não significa necessariamente reduzir o tempo de treinamento, uma vez que algumas técnicas podem inclusive aumentar o tempo total do treinamento, mas sim reduzir o tempo de pontuação do modelo. Ainda segundo Liu e Motoda (2007), existem três tipos de técnicas de seleção de características , sendo elas:

- ***Filter***: essa técnica remove as características que não apresentam utilidade para o modelo, como também remove informações redundantes. Essa técnica apresenta baixo custo de processamento, porém, não leva em consideração o algoritmo que está sendo empregado. Dessa forma, pode não ser capaz de selecionar as melhores características para o modelo;
- ***Wrapper***: essa técnica permite experimentar subconjuntos de características junto ao algoritmo que será empregado no modelo, evitando assim, eliminar características relevantes para o mesmo. Entretanto, essa técnica requer um maior processamento;
- ***Embedded***: esse método utiliza a seleção de características no processo de treinamento do modelo, sendo em alguns casos parte integrante de alguns algoritmos de classificação (como na Árvore de Decisão e Floresta Aleatória) em suas etapas de seleção de atributos. Eles não são tão custosos quanto a técnica *Wrapper*, em quesito de processamento, nem tão baratos em comparação ao *Filter*, atingindo o equilíbrio entre o gasto computacional e a qualidade do resultado. O algoritmo Árvore de Decisão, apresentado na seção 2.9, utiliza técnicas de seleção de características em seu algoritmo de classificação.

Além das técnicas já mencionadas, existem outras abordagens relevantes para a seleção de atributos, cada uma com suas peculiaridades e vantagens. Algumas destas ferramentas são abordagens específicas de seleção de atributos que merecem destaque em processos de Mineração de Dados Educacionais:

- **KBest e Chi2:** Estas técnicas também fazem parte da categoria "Filter". O KBest realiza uma seleção dos melhores atributos baseados em medidas de relevância, como a análise

de variância (ANOVA). O Chi2 (Qui-quadrado) é uma medida estatística que avalia a independência entre variáveis categóricas. Ambas são aplicadas antes do treinamento do modelo e têm como foco selecionar atributos que possuam maior relação com a variável alvo, eliminando aqueles que apresentam pouca influência no processo de classificação.

- **Wrappers com Gradient Boosting e Logistic Regression:** A técnica *Wrapper* é refinada com o uso de algoritmos específicos como o *Gradient Boosting* e a Regressão Logística. Estes métodos de seleção de atributos realizam uma exploração mais profunda dos subconjuntos de características, avaliando de forma iterativa a eficácia de cada conjunto por meio do treinamento e validação do modelo. O *Wrapper* com *Gradient Boosting* e *Logistic Regression* é mais computacionalmente intenso em relação ao *Filter*, mas pode resultar em conjuntos de atributos altamente relevantes e adaptados ao algoritmo de classificação escolhido.

- **Embedded (do Random Forest):** Esta técnica, conhecida como *Embedded*, é caracterizada por incluir o processo de seleção de características diretamente no treinamento do modelo. No contexto de algoritmos de classificação como a Árvore de Decisão e a Floresta Aleatória, a seleção de atributos é realizada de forma integrada ao processo de construção das árvores. Isso otimiza a relevância dos atributos selecionados para a tarefa de classificação específica, sem demandar o processamento intensivo associado às abordagens *Wrapper*. A Árvore de Decisão, por exemplo, incorpora técnicas de seleção de características em sua própria estrutura, permitindo a escolha de atributos mais informativos para as divisões de nodos.

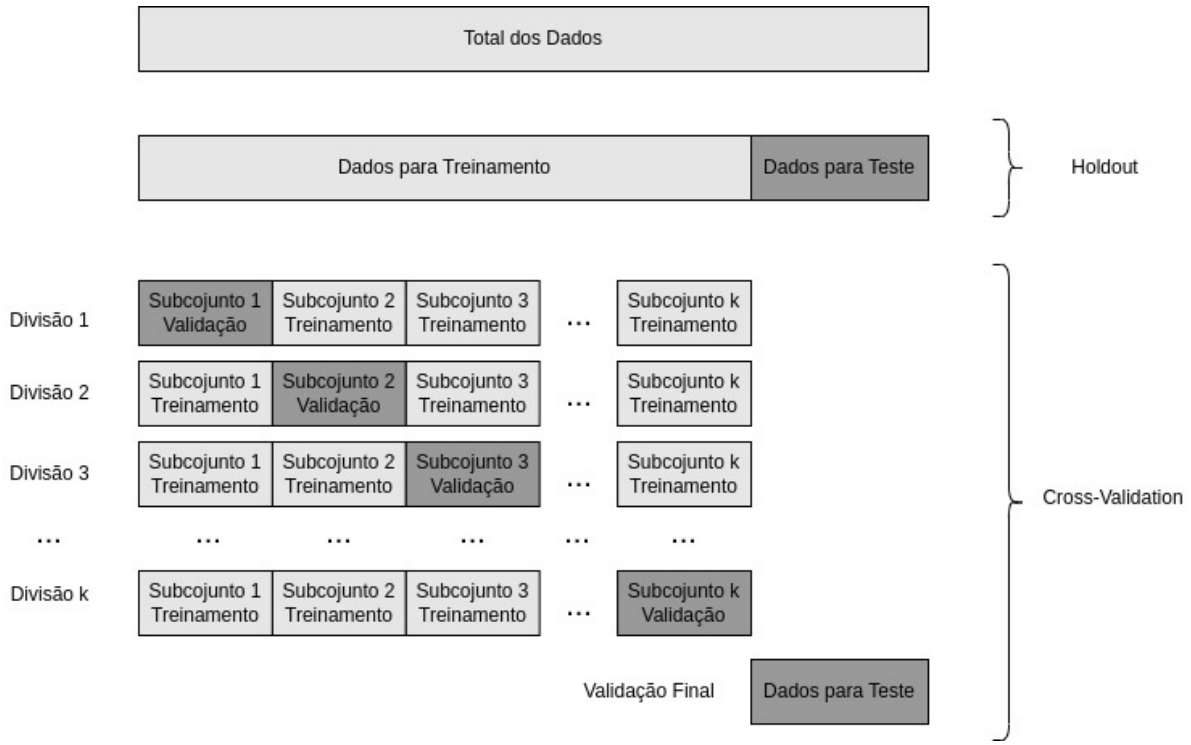
Essas ferramentas de seleção de atributos desempenham um papel crucial na otimização do desempenho dos modelos de predição de evasão escolar. Cada abordagem oferece uma abordagem única para identificar os atributos mais relevantes, balanceando considerações de eficiência computacional e precisão do modelo.

2.6 Treinamento e Teste

Uma abordagem simples para avaliar a capacidade de generalização de um modelo é realizar a divisão dos dados em conjuntos de treinamento e teste. Essa divisão permite avaliar como o modelo se comporta em dados que não foram utilizados durante o treinamento, fornecendo uma estimativa do erro do modelo, em relação a conjuntos de dados desconhecidos. Esse método é conhecido como *Holdout*. No entanto, uma limitação do *Holdout* é que não há garantia de que o subconjunto de treinamento seja totalmente representativo do conjunto de dados completo (GERON, 2019). Uma abordagem alternativa para a separação dos dados de treinamento e teste é a técnica da validação cruzada (*Cross-Validation*). Essa técnica divide os dados em k subconjuntos e utiliza cada um deles como conjunto de teste, enquanto os demais atuam como conjunto

de treinamento. Diferentes combinações de treinamento e teste são realizadas, resultando em *k* repetições do processo de treinamento e teste, como ilustrado na Figura 4. Uma configuração comum é adotar *k=10* subconjuntos (GERON, 2019), oferecendo uma boa alternativa para a divisão dos dados.

Figura 4 – Processos de Treinamento e Validação dos Dados.

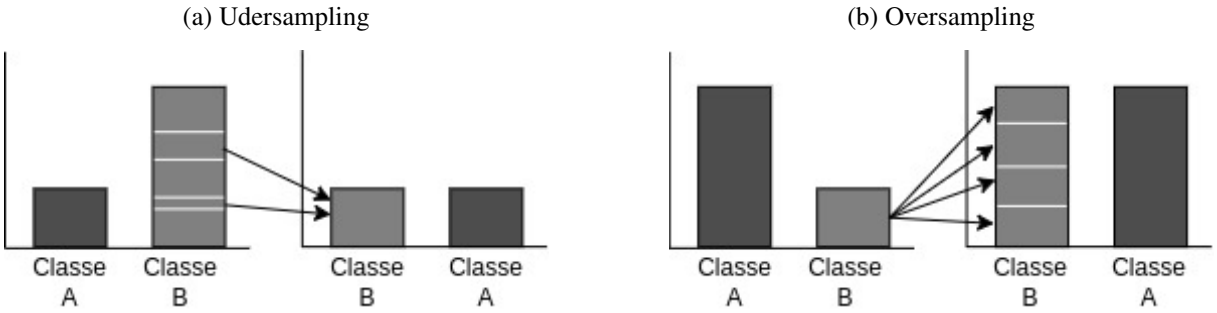


Fonte: Adaptado de Geron (2019).

Quando um conjunto de dados apresenta um número significativamente maior de instâncias em uma classe em comparação com outra, esse conjunto é considerado desbalanceado ou desequilibrado. A classe com maior quantidade de instâncias é denominada classe majoritária, enquanto a classe com menor quantidade é chamada de classe minoritária. Em muitos problemas de classificação, é comum lidar com conjuntos de dados desbalanceados. Embora alguns modelos preditivos sejam eficazes nesse contexto, há desafios quando o modelo não consegue lidar adequadamente com o desequilíbrio, resultando em problemas durante o treinamento e possíveis viéses de previsão para a classe majoritária. A principal solução utilizada para lidar com o problema do desbalanceamento é o balanceamento dos dados, o que pode ser feito por meio de técnicas de subamostragem e sobreamostragem (HE; GARCIA, 2009). A subamostragem (em inglês *undersampling*) consiste em equilibrar as classes através da redução da quantidade de instâncias das classes majoritárias, conforme ilustrado na Figura 5a. As técnicas de retirar instâncias da classe majoritárias podem ser agrupadas em duas categorias:

- **Retirada aleatória**: nessa técnica as instâncias majoritárias são retiradas de forma aleatória. Esse método é simples e requer pouco processamento, mas pode acarretar em uma perda

Figura 5 – Balanceamento de cargas



Fonte: Adaptado de He e Garcia (2009)

grave de informações relevantes, causando prejuízo na qualidade da previsão;

- **União de instâncias**: nessa técnica duas ou mais instâncias são concatenadas em uma única instância, o que acarreta em uma perda menor de informações. Porém, o custo de processamento é bem mais elevado que o da técnica de retirada aleatória pois, para realizar a junção das instâncias, são utilizados métodos de clusterização para calcular a semelhança entre as instâncias que farão parte do processo de união.

A sobreamostragem (em inglês, *oversampling*) consiste em equilibrar as classes através da inclusão de instâncias da classe minoritária, conforme ilustrado na Figura 5b. Existem duas técnicas básicas que podem ser utilizadas quando se pretende incluir instâncias na classe minoritária (CHAWLA et al., 2002), sendo elas:

- **Cópia simples**: Nessa técnica utiliza-se apenas a repetição das instâncias minoritárias para equiparar a classe majoritária. Alguns problemas de viés podem ser gerados quando o modelo tende a classificar as instâncias como pertencentes às classes minoritárias pela repetição de dados exatamente iguais;
- **Criação de novas instâncias**: Certos algoritmos, como o SMOTE (em inglês, *Synthetic Minority Oversampling Technique*), empregam abordagens estatísticas para gerar novas instâncias minoritárias. Essas instâncias recém-criadas não são simples cópias das instâncias originais; em vez disso, representam exemplos mais abrangentes das características já presentes. Isso amplia a diversidade dos dados da classe minoritária e ajuda a mitigar o desequilíbrio de classes, aprimorando a capacidade do modelo de aprendizado de lidar com instâncias minoritárias de maneira mais eficaz.

2.7 Tarefas de Classificação

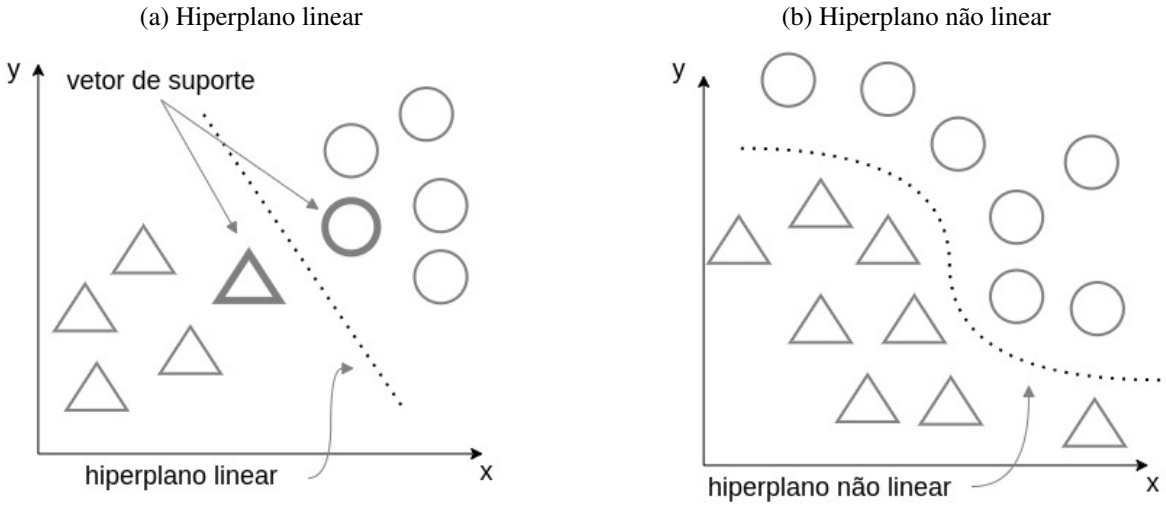
Como descrito na seção 2.4, a tarefa de classificação está associada à Aprendizagem Supervisionada e tem como objetivo verificar se uma determinada instância pertence ou não a

uma classe. A seguir são apresentados alguns dos principais modelos de classificação utilizados no aprendizagem de máquina supervisionada.

2.8 Máquina de Vetor de Suporte

O algoritmo Máquina de Vetores de Suporte (em inglês, *Support Vector Machine*, SVM) é um dos modelos de Aprendizagem de Máquina supervisionada mais populares. O SVM é capaz de realizar classificações lineares ou não lineares, regressão e até detecção de *outliers*. É especialmente adequado para a classificação de conjuntos de dados complexos, ainda que de pequeno ou médio porte. No SVM, cada ponto de dados é representado como um ponto em um espaço n-dimensional (onde 'n' é o número de características disponíveis), e a disposição desses pontos é determinada pelos valores das características. O hiperplano é a fronteira que melhor separa as classes, como exemplificado na Figura 6 (GERON, 2019).

Figura 6 – Representação dos hiperplanos no SVM



Fonte: Adaptado de Geron (2019)

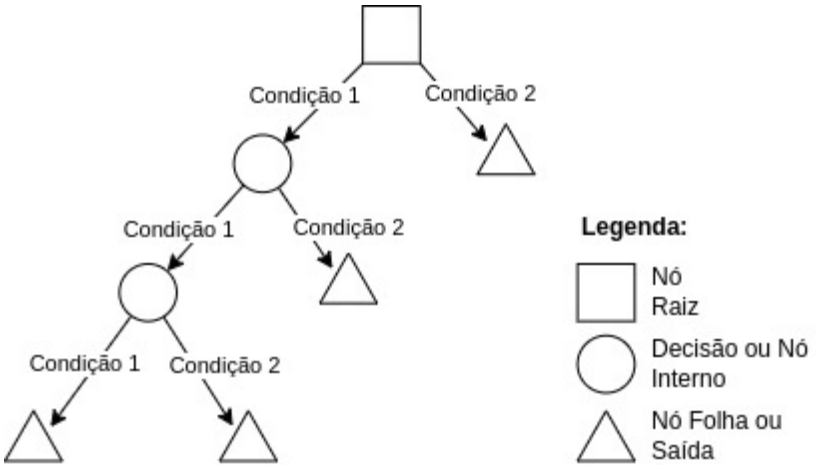
Os "Vetores de Suporte" são os elementos que possuem coordenadas individuais mais próximas e desempenham um papel crucial na definição do hiperplano, que é a fronteira que separa as classes, com a maior distância possível dos vetores de suporte. Em certas situações, um hiperplano linear (ilustrado na Figura 6a) é capaz de estabelecer uma fronteira entre as classes, tornando o processamento mais simples. No entanto, há casos em que a fronteira não pode ser descrita por uma equação linear, exigindo a descoberta de um polinômio que melhor se ajuste a essa fronteira (Figura 6b). Nessas circunstâncias, o processamento se torna mais complexo e dispendioso. Em algumas situações, devido à complexidade do polinômio e à natureza dos dados, esse modelo pode se tornar inviável para uso prático (GERON, 2019).

2.9 Árvores de Decisão

Árvores de decisão (em inglês *Decision Tree*, DT) correspondem a um classificador estruturado em formato de uma árvore binária, tendo como entrada um vetor de valores de atributos, de modo que, através de seus diversos caminhos possíveis de decisão, selecionado de acordo com os valores fornecidos para os atributos, a árvore retorne uma saída única que reflete o valor da decisão. Os dados utilizados em uma DT podem ser tanto valores numéricos, quanto categóricos (RUSSEL; NORVIG, 2013). O algoritmo de árvore de decisão é extremamente recomendado quando se pretende entender e interpretar as decisões, por ser possível acompanhar a trajetória para uma previsão (GRUS, 2016).

A estrutura genérica de uma DT é apresentada na Figura 7, na qual podem ser observados três tipos de nós. O "nó raiz" representa o início da árvore, a partir do qual, de acordo com a análise dos valores dos atributos, o resultado será direcionado para caminhos diversos, podendo conter "nós internos", ou "nós folhas". Cada "nó interno" realiza uma análise semelhante à realizada no "nó raiz", direcionando o processo de classificação para outro "nó interno", ou para um "nó folha". Cada "nó folha" é responsável por retornar o resultado final da árvore de decisão, cujo valor está associado a um atributo alvo ou a um valor de regressão (RUSSEL; NORVIG, 2013).

Figura 7 – Estrutura genérica de uma árvore de decisão.



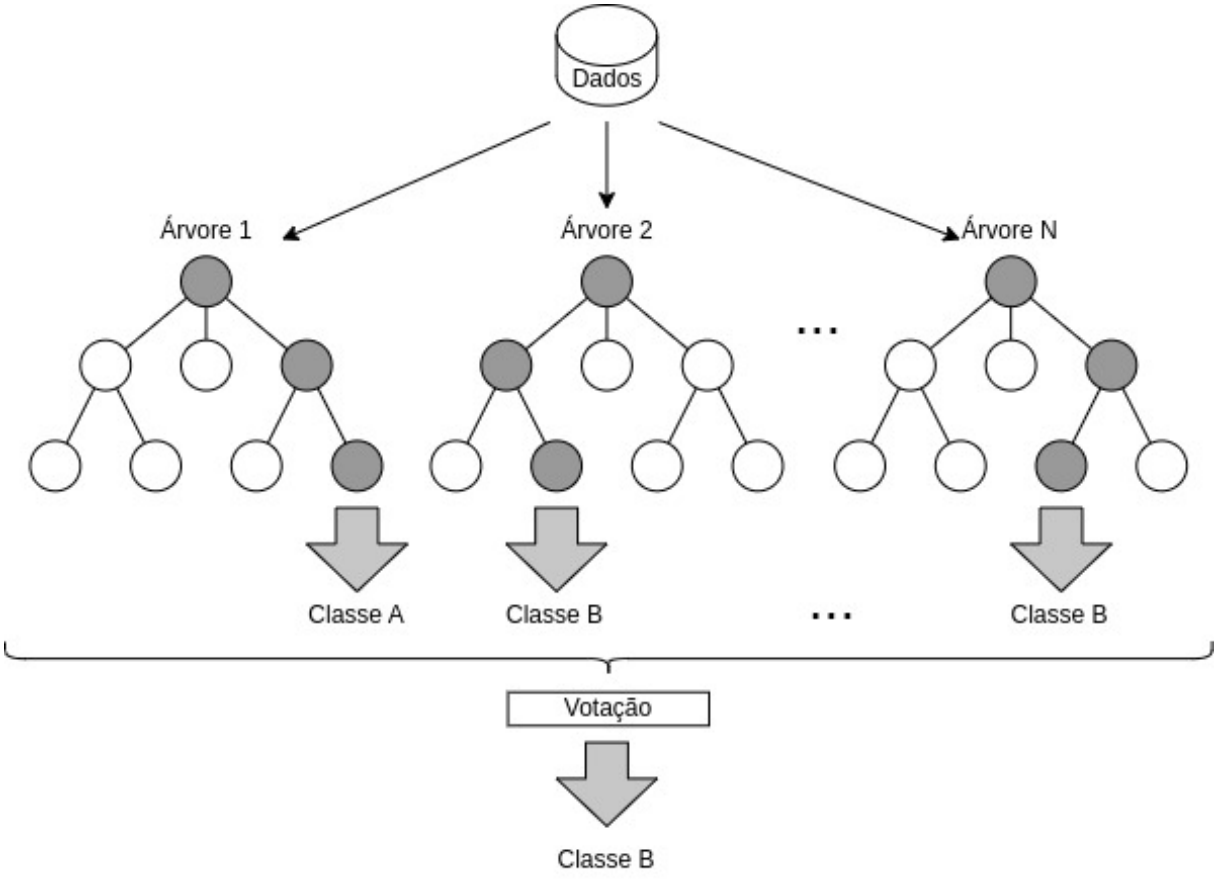
Fonte: Adaptador de Russel e Norvig (2013).

2.10 Floresta Aleatória

A Floresta Aleatória (do inglês *Random Forest*, RF) combina múltiplas árvores de decisão para realizar tarefas de classificação e regressão. É uma técnica popular e eficaz devido à sua capacidade de lidar com conjuntos de dados complexos e realizar previsões precisas. O funcionamento da Floresta Aleatória, conforme ilustrado na Figura 8, baseia-se na criação de um conjunto de árvores de decisão, onde cada árvore é construída a partir de uma amostra aleatória

do conjunto de dados de treinamento. Cada árvore é treinada de forma independente, dividindo o conjunto de dados em diferentes subconjuntos e aplicando critérios de divisão para tomar decisões em cada nó (ARTERO, 2009).

Figura 8 – Floresta Aleatória .



Fonte: Adaptado de Geron (2019).

Durante o processo de treinamento, a Floresta Aleatória utiliza uma técnica conhecida como bagging (*bootstrap aggregating*), que consiste em criar várias amostras de treinamento por meio de reamostragem com reposição. Essa abordagem permite que cada árvore seja treinada em diferentes subconjuntos do conjunto de dados original, aumentando a diversidade e reduzindo a correlação entre as árvores. Além disso, em cada divisão de nó, a Floresta Aleatória seleciona aleatoriamente um subconjunto de atributos para considerar como candidatos para a divisão. Essa aleatoriedade adicionada ao processo de construção das árvores ajuda a reduzir o viés e a variância do modelo, melhorando a generalização e a capacidade de lidar com sobreajuste (GERON, 2019).

Uma vez que todas as árvores são construídas, a Floresta Aleatória realiza a classificação por maioria de votos (no caso de classificação), ou a média das previsões (no caso de regressão) feitas por cada árvore individual. Essa agregação de resultados das árvores ajuda a obter uma previsão final mais robusta e confiável. A Floresta Aleatória apresenta várias vantagens, como a capacidade de lidar com dados ausentes e variáveis categóricas, a robustez em relação a *outliers*

e a capacidade de avaliar a importância relativa dos atributos para a tarefa de classificação ou regressão. Além disso, ela é eficiente em termos computacionais e pode ser paralelizada para acelerar o processo de treinamento em conjuntos de dados volumosos (RUSSEL; NORVIG, 2013).

2.11 Naive Bayes

O *Naive Bayes* (NB) é um algoritmo baseado no Teorema de Bayes, que utiliza a probabilidade condicional para realizar tarefas de classificação. É uma técnica simples e eficiente, amplamente utilizada em problemas de processamento de linguagem natural, filtragem de spam, diagnóstico médico, entre outros. O funcionamento do *Naive Bayes* é fundamentado no Teorema de Bayes, que relaciona a probabilidade condicional de um evento ocorrer, dado que outro evento ocorreu, com a probabilidade desses eventos individualmente (IZBICKI; SANTOS, 2020), conforme apresentado na Equação 1.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \tag{1}$$

A equação do Teorema de Bayes pode ser expressa como:

- **P(A|B)** é a probabilidade do evento A ocorrer, dado que o evento B ocorreu;
- **P(B|A)** é a probabilidade do evento B ocorrer, dado que o evento A ocorreu;
- **P(A)** é a probabilidade do evento A ocorrer;
- **P(B)** é a probabilidade do evento B ocorrer.

No contexto do *Naive Bayes*, considera-se que os atributos são independentes entre si, dada a classe. Essa suposição simplificadora é conhecida como "ingenuidade de Bayes"(*naive assumption*). Com base nessa suposição, pode-se calcular a probabilidade de um exemplo pertencer a uma determinada classe utilizando a fórmula do Teorema de Bayes. O algoritmo *Naive Bayes* atribui a classe mais provável a um exemplo, com base na avaliação da probabilidade condicional de cada classe, dada a evidência dos atributos observados. Essa probabilidade condicional é calculada usando a suposição de independência e a distribuição de probabilidades dos atributos para cada classe. As principais variantes do *Naive Bayes* incluem o *Naive Bayes* Gaussiano, *Naive Bayes* Multinomial e *Naive Bayes* Bernoulli, que se adaptam a diferentes tipos de dados (GRUS, 2016).

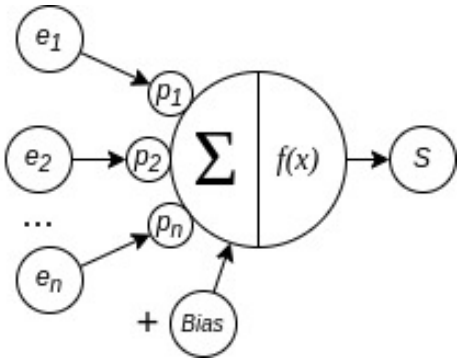
Uma das vantagens do *Naive Bayes* é sua eficiência computacional, uma vez que o cálculo das probabilidades pode ser realizado de forma rápida e simples. Além disso, o algoritmo

é robusto em relação a dados ausentes e pode lidar com conjuntos de dados de alta dimensionalidade. No entanto, a suposição de independência entre os atributos nem sempre é válida em todos os problemas, o que pode levar a uma perda de precisão nas previsões (GERON, 2019).

2.12 Redes Neurais Artificiais

Uma Rede Neural Artificial (RNA) corresponde a um modelo preditivo matemático inspirado na estrutura neural de organismos inteligentes, assemelhando-se ao funcionamento do cérebro, onde cada único neurônio teria como representação a rede neural mais simples denominada perceptron. A Figura 9 representa o modelo de um neurônio artificial (perceptron), de forma que ela possui n entradas de valor e_i, onde cada e_i está associado a um peso sináptico específico p_i. O processamento é realizado através da somatória do produto do valor de cada entrada por seu peso específico, somado junto a um valor *bias* atribuído, a fim de facilitar a aproximação da função. Esse valor é então submetido a uma função de ativação que retorna o valor de saída S do perceptron (GRUS, 2016).

Figura 9 – Neurônio Artificial .

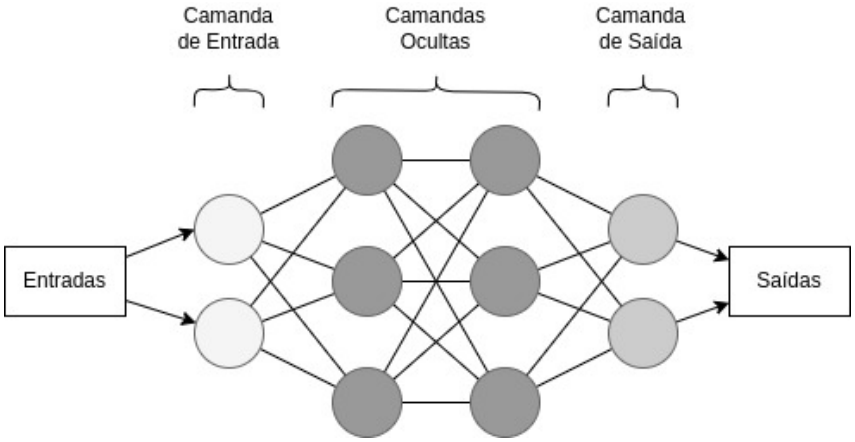


Fonte: Adaptado de Grus (2016).

A associação de diversas redes perceptron possibilita a resolução de problemas muito complexos, como na utilização de visão computacional, ou processamento de sons. Cada camada da RNA é classificada, conforme sua posição, como camada de entrada, oculta ou saída. Essa formação é denominada *Multilayer Perceptron* (MLP) (ARTERO, 2009) e é ilustrada na Figura 10. Essa configuração de conexão entre os neurônios funciona como uma rede progressiva, onde cada neurônio da camada anterior conecta-se aos neurônios da camada posterior. Os neurônios da "camada de entrada" são responsáveis por receber os valores de entrada. As saídas desses neurônios estão conectadas à primeira "camada oculta" que, por sua vez, está conectada à próxima "camada oculta", seguindo assim até a última camada, que está conectada à "camada de saída", responsável por fornecer os resultados da rede.

Quando uma rede neural é iniciada, são atribuídos valores aleatórios aos pesos sinápticos de cada neurônio. O processo de treinamento da rede consiste em fornecer os valores de entrada e observar o valor de saída. Quando o valor de saída difere do valor esperado, os pesos sinápticos

Figura 10 – Multilayer Perceptron .



Fonte: Adaptado de Geron (2019).

são corrigidos através de retropropagação, conhecida como *backpropagation*, que corrige os valores dos pesos pela diferença entre o valor esperado e o obtido (GERON, 2019).

2.13 Medidas de Desempenho dos Classificadores

Para saber o quão bom é um modelo de predição é preciso definir uma métrica ou medida de desempenho. Existem diversas técnicas para avaliar o desempenho dos modelos de classificação, mas de forma geral, aquele que é capaz de produzir mais resultados corretos em relação aos resultados errados é o que apresentará o melhor desempenho (GERON, 2019). A seguir são apresentadas as principais medidas de desempenho dos modelos de classificação.

2.14 Matriz de Confusão

A Matriz de Confusão é uma matriz que contém a distribuição dos resultados previstos pelo classificador em comparação com a classificação real das instâncias, na qual são distribuídos entre resultados verdadeiros e os resultados falsos (GRUS, 2016). A matriz de confusão é aplicada em situações nas quais se deseja verificar se um conjunto de instâncias pertencem ou não a uma determinada classe. O Quadro 1 apresenta a Matriz de Confusão e as quatro configurações possíveis.

Quadro 1 – Matriz de Confusão.

		Prevista	
		Positivo	**Negativo**
Real	**Positivo**	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Negativo	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Adaptado de Grus (2016).

A seguir são listados e descritos os resultados possíveis alcançados na matriz de confusão:

- **Verdadeiro Positivo (VP)**: representa os casos em que o modelo classificou corretamente uma instância como pertencente à classe positiva. Em outras palavras, o modelo identificou corretamente uma condição ou evento de interesse;
- **Verdadeiro Negativo (VN)**: refere-se aos casos em que o modelo classificou corretamente uma instância como não pertencente à classe positiva. Isso significa que o modelo identificou corretamente que uma condição ou evento indesejado não está presente;
- **Falso Negativo (FN)**: ocorre quando o modelo classifica erroneamente uma instância como não pertencente à classe positiva, quando na verdade deveria ter sido classificada como positiva. Isso significa que o modelo não conseguiu identificar corretamente uma condição ou evento de interesse, resultando em uma falha na detecção;
- **Falso Positivo (FP)**: quando o modelo classifica erroneamente uma instância como pertencente à classe positiva, quando na verdade deveria ter sido classificada como negativa. Isso significa que o modelo identificou incorretamente uma condição ou evento de interesse que não estava presente.

A diagonal principal da matriz de confusão (destacada no Quadro 1) apresenta a quantidade de instâncias classificadas corretamente, enquanto que a diagonal secundária apresenta a quantidade de instâncias classificadas incorretamente. Diversas medidas de desempenho são baseadas nos valores calculados na Matriz de Confusão.

2.15 Cálculos das Medidas de Desempenho

A seguir serão apresentadas as principais métricas utilizadas para avaliar o desempenho de modelos de ML. Essas medidas fornecem informações sobre a qualidade das previsões realizadas pelos modelos e permitem comparar diferentes abordagens. A análise dessas medidas é essencial para compreender o desempenho dos modelos e tomar decisões informadas na seleção e ajuste dos algoritmos de classificação.

Acurácia (em inglês, *Accuracy*) é a medida de desempenho que avalia a proporção entre o número de instâncias classificadas corretamente, em relação ao total de instâncias. É uma métrica básica que fornece uma visão geral da precisão do classificador, sendo calculada com base na Equação 2. A acurácia é uma ótima medida de desempenho para dados bem balanceados. Porém, em um universo de dados desbalanceados, quando alguma classe apresenta muitas instâncias em relação às demais, é possível que a acurácia não seja suficiente para uma boa medida de desempenho (GERON, 2019).

$$Accuracy = \frac{VN + VP}{VP + FN + VN + FP} \tag{2}$$

É importante considerar outras métricas, como precisão, revocação e *F1-score*, além da acurácia, para ter uma avaliação mais completa do desempenho do classificador, especialmente em casos de desequilíbrio de classes (GRUS, 2016).

A precisão (em inglês *precision*), também conhecida como valor preditivo positivo, é a medida de desempenho que retorna a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias classificadas como positivas (soma de verdadeiros positivos e falsos positivos) (GERON, 2019). A precisão mede a capacidade do classificador de evitar a classificação incorreta de instâncias negativas como positivas, sendo calculada por meio da Equação 3.

$$Precision = \frac{VP}{VP + FP} \tag{3}$$

A precisão é importante em situações em que a classificação correta de instâncias negativas é crucial, tendo como prioridade evitar falsos positivos. Em algumas situações, a precisão pode não ser adequada, especialmente quando há desequilíbrio entre as classes, ou quando os custos associados aos diferentes tipos de erros de classificação são diferentes. Por exemplo, em um sistema de detecção de doenças graves, um falso negativo (classificar erroneamente um paciente doente como saudável) pode ter consequências muito mais graves do que um falso positivo (classificar erroneamente um paciente saudável como doente) (GERON, 2019).

Nesse caso, a precisão não leva em consideração os custos associados aos diferentes tipos de erros. Para contornar este problema, pode ser necessário utilizar em conjunto outras medidas de desempenho, como a revocação (em inglês, *Recall*), que também é conhecida como como taxa de verdadeiro positivo, ou sensibilidade. A revocação retorna a proporção de instâncias corretamente classificadas como positivas, em relação ao total de instâncias verdadeiramente positivas (soma de verdadeiros positivos e falsos negativos) (GRUS, 2016), conforme apresentado na Equação 4.

$$Recall = \frac{VP}{VP + FN} \tag{4}$$

A revocação é especialmente importante em situações em que a identificação correta de instâncias positivas é crucial, como em um sistema de detecção de câncer, onde é essencial identificar corretamente todos os casos positivos (verdadeiros positivos) para garantir um tratamento oportuno. Em resumo, a revocação é uma métrica que avalia a capacidade do classificador de identificar corretamente as instâncias positivas, complementando a métrica de precisão, que mede a capacidade de evitar a classificação incorreta de instâncias negativas como positivas.

A revocação é uma métrica útil para medir a capacidade do classificador de encontrar todas as instâncias positivas, mas não leva em consideração os falsos positivos. Portanto, é importante considerar outras métricas, como a especificidade (em inglês *Specificity*), que é responsável por avaliar o quanto um classificador é capaz de identificar instâncias erroneamente

classificadas como pertencentes à classe, sendo utilizada na medida de desempenho da área abaixo da Curva das Características Operacionais do Receptor (em inglês, *Receiver Operating Characteristic Curve*, ROC) (GERON, 2019). A métrica especificidade é calculada conforme apresentado na Equação 5.

$$Specificity = \frac{VN}{VN + FP} \tag{5}$$

É comum combinar a precisão e a revocação em uma única medida de desempenho, principalmente quando necessário comparar dois classificadores. Essa métrica é chamada de pontuação *F1-Score*, que corresponde à média harmônica da precisão com a revocação (GERON, 2019), tendo seu valor calculado conforme apresentado na Equação 6.

$$F1 - Score = \frac{TP}{TP + \frac{FN+FP}{2}} \tag{6}$$

A pontuação *F1-Score* favorece os classificadores que apresentam valores similares de precisão e revocação. Porém, em alguns casos é necessário dar prioridade à precisão, quando for necessário garantir a mínima existência de falsos positivos, mesmo que alguns reais positivos não sejam identificados. Por outro lado, em outros casos, a revocação é mais importante quando é necessário garantir um mínimo de falsos negativos, mesmo que alguns reais negativos sejam classificados como pertencentes a classe (GERON, 2019). Assim posto, é necessário analisar com cuidado o contexto de uso do classificador, para selecionar a melhor métrica para o caso em estudo.

No contexto da aplicação de classificadores no contexto da evasão escolar, é necessário levar em consideração que tanto a presença de muitos *FN* é prejudicial, pois existirão muitos casos de alunos que poderiam ser beneficiados com programas para permanência no cursos, e por não serem identificados como possíveis de evasão, ficariam de fora. Da mesma forma, a presença de muitos *FP* acarretaria em um excessivo gasto de recursos caros, por serem utilizados com alunos que não teriam, de fato, predisposição para evadir.

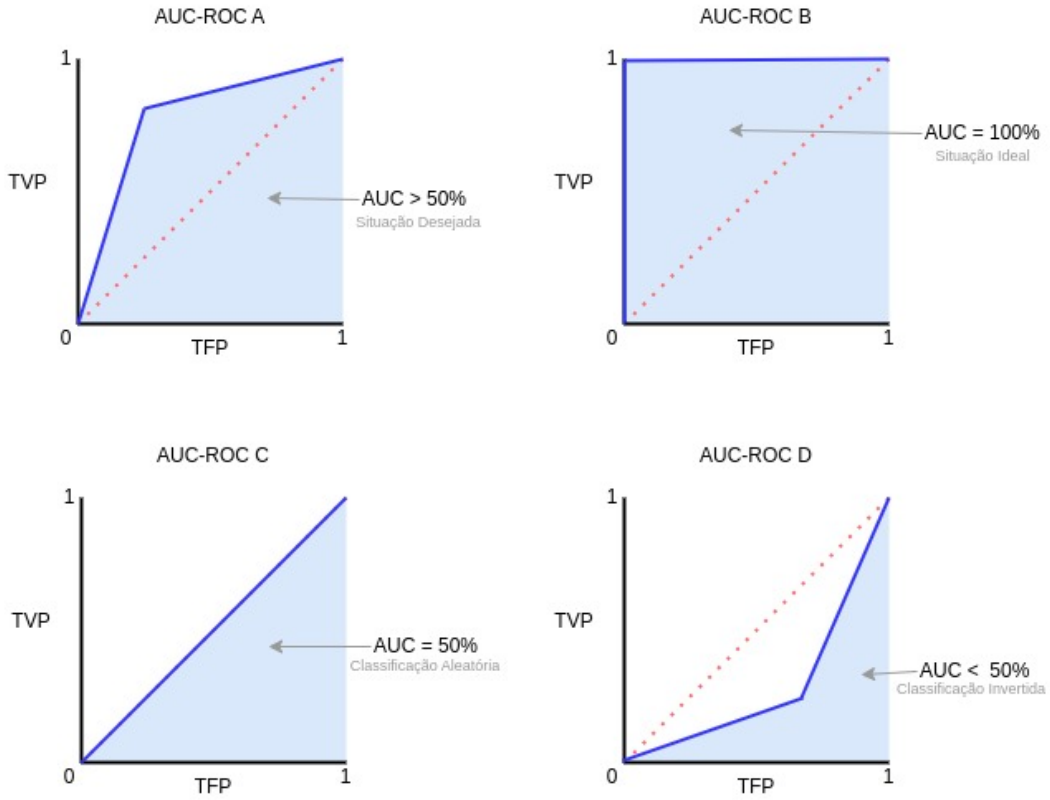
Desta forma, para este trabalho, foi estabelecida a necessidade da utilização da métrica de pontuação *F1-Score*. No entanto, não se abre mão de recorrer a métricas como a acurácia e a precisão. Isso ocorre porque a análise de todas essas métricas reforça a eficiência dos recursos investidos nos programas desenvolvidos para reduzir os casos de evasão escolar.

A ROC é um gráfico que representa a performance que um modelo têm em distinguir diferentes classes, sendo construído em função da sensibilidade, dado pelo valor da Taxa de Verdadeiro Positivos (TVP) e a fração dos falsos positivos (1 - *specificity*) dado pela Taxa de Falsos Positivos (TFP) (GERON, 2019). O valor da Área Abaixo da Curva (em inglês, *Area Under the Curve - Receiver Operating Characteristic Curve*, AUC-ROC) corresponde à métrica

de desempenho AUC-ROC. Na Figura 11 são apresentadas algumas das configurações que o gráfico pode apresentar, tendo as possibilidades descritas a seguir:

- **AUC > 50%**: quanto maior o valor da AUC, melhor será o desempenho apresentado pelo classificador;
- **100%**: quando o AUC apresenta valor de 100%, significa que o classificador conseguiu relacionar corretamente todas as instâncias, tanto as verdadeiras, quanto as falsas. Essa situação é muito pouco provável de acontecer em classificações com conjunto de dados reais;
- **AUC = 50%**: quanto mais próximo ao valor de 50%, pior é o classificador, tendo essa classificação características de uma total aleatoriedade, não representando corretamente as instâncias em suas devidas classes;
- **AUC < 50%**: o AUC apresenta valores inferiores a 50% quando há uma inversão das classes de aprendizagem em relação às de validação. Essa situação não pode acontecer em uma configuração correta dos dados no processo de classificação.

Figura 11 – Gráfico da Curva ROC .



Fonte: Adaptado de Geron (2019).

A métrica AUC-ROC é particularmente adequada para a aplicação em classificadores no contexto da evasão escolar. Isso se deve ao fato de que, durante seu cálculo, ela enfatiza tanto

o maior valor da taxa de *VP* quanto a menor ocorrência de *FP*. Portanto, essa métrica é outra ferramenta estabelecida para validar os resultados apresentados neste trabalho.

Outra medida de desempenho é a Raiz do Erro Quadrático Médio (em inglês *Root Mean Square Error* RMSE). Essa é uma métrica amplamente utilizada na área de aprendizado de máquina para avaliar o desempenho de modelos de regressão. Ela é especialmente útil quando estamos lidando com problemas de previsão, nos quais queremos estimar um valor contínuo ou quantitativo (GERON, 2019). O cálculo do RMSE é mostrado na Equação 7.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{7}$$

O RMSE mede a diferença média entre os valores previstos pelo modelo e os valores reais observados. Ele é calculado pela raiz quadrada da média dos quadrados dos erros. Essa métrica é interessante porque penaliza de forma mais significativa os erros maiores, tornando-se uma medida mais sensível a desvios consideráveis entre os valores previstos e os valores reais.

2.16 SUAP e dados de interesse da pesquisa

O Sistema Unificado da Administração Pública (SUAP) foi criado em 2006 pela equipe de desenvolvimento do Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte (IFRN) para a gestão dos processos administrativos e acadêmicos dos Institutos Federais. O SUAP vem sendo utilizado por diversos Institutos Federais em todo o Brasil, através de convênios firmados com o IFRN de forma colaborativa entre todas as instituições parceiras. O sistema possibilita o gerenciamento dos dados produzidos por diversas áreas integradas através de módulos responsáveis pela gestão de suas funcionalidades, e cada módulo é relacionado a uma área específica. Os módulos atuais disponibilizados pelo SUAP são (IFRN. . . , 2022):

- **Gestão de Pessoas**: módulo da área de Gestão de Pessoas, que é responsável pelas funcionalidades de gestão de editais de remoção, contracheques, competições desportivas, férias, digitalização de pastas funcionais, impressões de carteiras funcionais e crachás, indicadores de pessoal, entre outros;
- **Ponto Eletrônico**: módulo responsável pelo acompanhamento do ponto eletrônico por meio da autenticação por biometria;
- **Protocolo**: módulo responsável pela tramitação de documentos entre as unidades;
- **Patrimônio**: módulo responsável pela gestão patrimonial;
- **Almoxarifado**: módulo responsável pelo controle e distribuição de materiais de consumo;
- **Planejamento**: módulo responsável pela gestão e controle de objetivos, metas e ações;

- **Contratos**: módulo responsável pelo controle de medições, execução e fiscalização dos contratos celebrados pela instituição;
- **Convênios**: módulo de gestão de convênios;
- **Catálogo de Materiais**: módulo de controle de materiais utilizados pelo almoxarifado;
- **Compras**: módulo para levantamento de necessidades de compras pelos campi.
- **Chaves**: módulo responsável pela gestão dos empréstimos de chaves;
- **Gestão de Projetos de Extensão**: módulo responsável pela gestão, acompanhamento e emissão de relatórios de projetos de extensão;
- **Controle de Acesso de Visitantes**: módulo responsável pelo controle de acesso de visitantes;
- **Gestão de Cursos e Concursos**: módulo responsável pelo acompanhamento dos trabalhos para a execução de cursos e concursos;
- **Gestão Acadêmica**: módulo responsável pela gestão de todas as atividades de ensino da instituição, dividido entre áreas de ensino, documentação e gestão;
- **Indicadores de Gestão**: módulo responsável pelos cálculos dos indicadores de gestão institucional normatizados pelo Tribunal de Contas da União ;
- **Autoavaliação**: módulo responsável pela gestão do processo de autoavaliação institucional;
- **Consulta pública PDI**: módulo responsável pela gestão da consulta pública da comunidade institucional para a construção do Plano de Desenvolvimento Institucional;
- **Gestão de Programas Sociais e Bolsas de Trabalho**: módulo responsável pela gestão do Programa de Assistência Social da Instituição;
- **Currículos e Grupos CNPQ Lattes**: módulo de importação de currículos para a Plataforma Lattes, bem como de informações de Grupos de Pesquisa;
- **Clipping**: módulo responsável pela gestão de clipping para o setor de Comunicação Social;
- **Central de Serviços de TI**: módulo responsável pela gestão da Central de Serviços de TI;
- **Sistema Gestor de Concursos**: módulo responsável pela gestão de processo para realização de Concursos Públicos e Processos Seletivos para Discentes.

O módulo de gestão acadêmica é responsável por gerenciar as informações relacionadas aos processos educacionais de todos os alunos dos cursos disponibilizados pelo IFPB. Possui dados sobre matrizes curriculares, informações pessoais, institucionais e socioeconômicas, bem como todos os aspectos relacionados à permanência dos alunos nos cursos, tais como períodos cursados, disciplinas matriculadas, reprovadas e concluídas, além das respectivas notas.

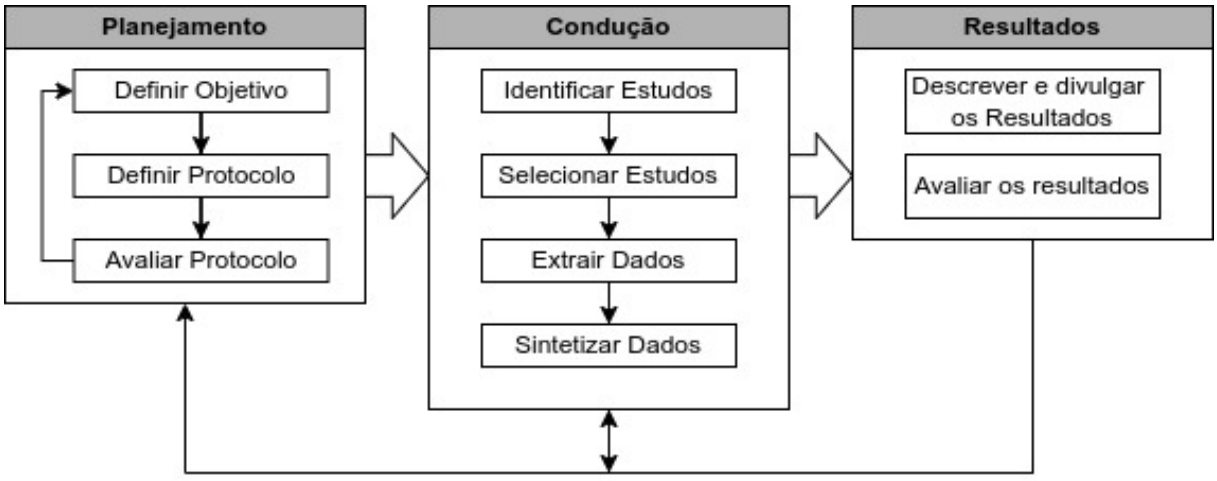
3 ESTADO DA ARTE SOBRE SELEÇÃO DE ATRIBUTOS PARA PREDIÇÃO DE EVASÃO ESCOLAR

A Revisão Sistemática da Literatura (RSL) é um método de pesquisa que tem como objetivo identificar e analisar estudos relevantes relacionados a uma questão de pesquisa específica ou fenômeno de interesse. Essa abordagem possibilita uma análise abrangente e sistemática dos estudos primários que contribuem para a revisão. Os estudos primários desempenham um papel fundamental na realização de uma revisão sistemática, sendo selecionados com base em critérios pré-definidos e fornecendo informações essenciais para responder à questão de pesquisa em foco (KITCHENHAM, 2004; PICHETH, 2007). No presente trabalho, foram estabelecidas diferentes etapas para a execução da RSL, garantindo uma abordagem metódica na busca, seleção e análise dos estudos relevantes. A seguir, são apresentadas essas etapas:

- **Planejamento**: tem como objetivo a execução da pesquisa, sendo composta pelas atividades principais de definição do objetivo, preparação do protocolo que norteará a RSL;
- **Condução**: nesta etapa são identificados os estudos através da aplicação da estratégia de busca e selecionados conforme o protocolo definido na etapa de planejamento;
- **Resultados**: a última etapa do processo de RSL está relacionada com a documentação e descrição dos resultados, bem como, elaboração das respostas para as questões de pesquisa.

Na Figura 12 é apresentado o protocolo de pesquisa seguido neste trabalho, com as fases e atividades realizadas em cada uma das etapas da RSL.

Figura 12 – Atividades da Revisão Sistemática da Literatura .



Fonte: Dados do Autor.

3.1 Planejamento

O protocolo de pesquisa aplicado neste trabalho tem como objetivo identificar tecnologias, modelos e orientações para o processo de seleção de atributos utilizados em algoritmos para predição da evasão escolar. Como forma de identificar as metodologias aplicadas nos trabalhos selecionados, foram definidas as Questões de Pesquisa (QP), as quais são apresentadas no Quadro 2.

Quadro 2 – Questões de pesquisa.

Questão	**Descrição**
QP1	Qual o objetivo do estudo e origem dos dados utilizados?
QP2	Quais métodos utilizados na etapa de pré-processamentos dos dados ?
QP3	Quais estratégias foram utilizadas para seleção de atributos?
QP4	Quais tecnologias foram utilizadas na predição?

Fonte: Dados do Autor.

Foram realizadas pesquisas nas bibliotecas digitais *Association for Computing Machinery* (ACM)[1], *Institute of Electrical and Electronics Engineers* (IEEE)[2], Revista Brasileira de Informática na Educação (RBIE)[3], Simpósio Brasileiro de Informática na Educação (SBIE)[4], ScienceDirect[5], Scopus[6] e Web of Science[7] acerca de trabalhos relacionados à seleção de atributos utilizados na predição de evasão escolar, desenvolvidos nos últimos 10 anos. A consulta se deu sobre os termos "seleção de atributos", "predição", "classificação" e "evasão escolar", através das strings de busca adaptadas ao idioma da biblioteca digital, conforme apresentados no Quadro 3.

Em seguida, como forma de selecionar os estudos relevantes, foram estipulados critérios de inclusão (CI) e exclusão (CE):

- **CI1**: foram selecionados os artigos cujo foco principal é a classificação relacionada à seleção de atributos para preditores de evasão escolar;
- **CI2**: foram selecionados os artigos que abordam os algoritmos de classificação relacionados à seleção de atributos para preditores de evasão escolar, mesmo não sendo o foco principal do estudo;
- **CE1**: foram descartados os estudos primários que não satisfazem a nenhum critério de inclusão;

[1] https://dl.acm.org/
[2] https://ieeexplore.ieee.org/Xplore/home.jsp
[3] https://sol.sbc.org.br/journals/index.php/rbie
[4] https://ceie.sbc.org.br/evento/2021/SBIE.html
[5] https://www.sciencedirect.com
[6] https://www.scopus.com/home.uri
[7] https://www.webofscience.com/wos/woscc/basic-search

Quadro 3 – Strings de busca.

Idioma	Bibliotecas Digital	String de Busca
Inglês	ACM, IEEE, ScienceDirect, Scopus e Web of Science	((“attribute classification” OR “attribute selection” OR “feature classification” OR “feature selection”) AND “dropout” AND (“education’ OR ‘school’ OR ‘college’ OR ‘university’) AND ‘prediction’)
Português	RBIE, SBIE e outras conferências e revistas que publicam na SOL (SBC Open Library)	((“classificação de atributos” OR “seleção de atributos” OR “classificação de características” OR “seleção de características” OR “classificação de features” OR “seleção de features”) AND “evasão” AND (“educação’ OR ‘escola’ OR ‘faculdade’ OR ‘Universidade’) AND ‘prediction’)

Fonte: Dados do Autor.

- **CE2**: foram descartados os estudos que não estão nos idiomas inglês ou português;
- **CE3**: foram descartados os estudos produzidos anteriormente a 2011, em virtude do critério de limite de 10 anos da produção;
- **CE4**: foram descartadas as publicações duplicadas;
- **CE5**: foram descartadas as publicações que consistiam em republicações de artigos previamente apresentados em conferências, revistas, congressos, entre outros;
- **CE6**: foram descartadas as publicações que não ofereciam acesso gratuito à sua versão completa;
- **CE7**: foram descartadas as publicações que não apresentam informações básicas completas nas bases de dados (título, autor, ano de publicação, fonte e resumo).

3.2 Condução

Depois do planejamento e definição do protocolo da RSL foi iniciada a fase de condução dos estudos, cujas atividades são descritas nas seções a seguir. Após a definição do texto a ser usado nas consultas e dos critérios de inclusão e exclusão de artigos, foram realizadas as consultas nas bibliotecas digitais mencionadas. Após essas consultas, foram localizados 83 artigos, sendo 21 (vinte um) na ACM, 01 (um) na IEEE, 02 (dois) na SBIE, 52 (cinquenta e dois) na ScienceDirect, 05 (cinco) na Scopus, 02 (dois) na Web of Science e 0 (zero) na RBIE. Com a aplicação da estratégia dos critérios de exclusão, foram descartados 65 trabalhos conforme apresentados na Tabela 1.

Após a aplicação de todos os passos da estratégia de seleção de trabalhos, foram selecionados um total de 17 (dezessete) artigos para a etapa de análise e extração de informações para responder as questões da pesquisa. O Quadro 4 descreve os artigos selecionados. Para cada

Tabela 1 – Quantidade de trabalhos excluídos.

Critério de Exclusão	**Biblioteca Digital**	**Quantidade**
CE1	ACM	17
	ScienceDirect	44
	Scopus	1
CE3	SBIE	1
CE4	Web Of Science	2
TOTAL		65

Fonte: Dados do Autor.

um desses artigos, são apresentados um identificador, a biblioteca a partir da qual o artigo foi encontrado, o título do artigo, a lista de autores e o ano da sua publicação.

A análise dos artigos selecionados visa responder às Questões de Pesquisa relevantes para a execução deste trabalho. Após a análise dos 17 (dezessete) artigos selecionados, foi possível elaborar as respostas para as questões de pesquisa definidas neste trabalho.

3.2.1 QP1 - Qual o objetivo do estudo e origem dos dados utilizados?

Em uma análise mais abrangente, os trabalhos têm como objetivo aplicar algoritmos em conjuntos de dados para a predição relacionada ao contexto educação. O trabalho de Dwan, Oliveira e Fernandes (2017) propõe um modelo de classificação de zonas de aprendizagem dos alunos na disciplina de Introdução de Programação de Computadores da Universidade Federal do Amazonas (UFAM), utilizando dados extraídos dos resultados das listas de exercícios aplicadas dentro do Ambiente de Correção Automática de Código.

Os trabalhos de Meca et al. (2020), Ma et al. (2017), Zhang e Wu (2019), Chanlekha e Niramitranon (2018), Dwan, Oliveira e Fernandes (2017), Thammasiri et al. (2014), Lara et al. (2014), Hu, Lo e Shih (2014), Miguéis et al. (2018), Costa et al. (2017), Nandeshwar, Menzies e Nelson (2011), Chango, Cerezo e Romero (2021), Hershkovitz e Nachmias (2011), Regha e Rani (2015), Siebra, Santos e Lino (2020) e Urbina-Najera, Camino-Hampshire e Barbosa (2020) têm como objetivo prever o desempenho dos alunos e os resultados alcançados nos cursos.

Os trabalhos de Chanlekha e Niramitranon (2018) e Hu, Lo e Shih (2014), além da predição do desempenho acadêmico dos alunos, possuem o objetivo de construir sistemas de notificação e comunicação com os interessados no processo de aprendizagem. Já os estudos de Ma et al. (2017), Zhang e Wu (2019), Silva et al. (2019), Chango, Cerezo e Romero (2021), Regha e Rani (2015) e Urbina-Najera, Camino-Hampshire e Barbosa (2020) são voltados ao processos de seleção de atributos relevantes para algoritmos de aprendizagem de máquina, utilizando técnicas de *Filter*, *Wrapper* e *Embedded*.

Com base na origem dos dados utilizados no trabalho, foi identificado que os trabalhos de Meca et al. (2020), Chanlekha e Niramitranon (2018), Thammasiri et al. (2014), Miguéis

Quadro 4 – Resultado da busca dos artigos e identificador.

ID	Base	Título e Autores
T01	ACM	Early Warning Methodology for Dropping out of University Degrees. Meca et al. (2020)
T02	ACM	Improving Prediction of Student Performance Based on Multiple Feature Selection Approaches. Ma et al. (2017)
T03	ACM	Research and Application of Grade Prediction Model Based on Decision Tree Algorithm. Zhang e Wu (2019)
T04	ACM	Student Performance Prediction Model for Early-Identification of at-Risk Students in Traditional Classroom Settings. Chanlekha e Niramitranon (2018)
T05	IEEE	Ensemble Regression Models Applied to Dropout in Higher Education. Silva et al. (2019)
T06	SBIE	Predição de Zona de Aprendizagem de Alunos de Introdução à Programação em Ambientes de Correção Automática de Código. Dwan, Oliveira e Fernandes (2017)
T07	Science Direct	A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. Thammasiri et al. (2014)
T08	Science Direct	A system for knowledge discovery in e-learning environments within the European Higher Education Area – Application to student data from Open University of Madrid, UDIMA. Lara et al. (2014)
T09	Science Direct	Developing early warning systems to predict students' online learning performance. Hu, Lo e Shih (2014)
T10	Science Direct	Early segmentation of students according to their academic performance: A predictive modelling approach. Miguéis et al. (2018)
T11	Science Direct	Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Costa et al. (2017)
T12	Science Direct	Learning patterns of university student retention. Nandeshwar, Menzies e Nelson (2011)
T13	Science Direct	Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. Chango, Cerezo e Romero (2021)
T14	Science Direct	Online persistence in higher education web-supported courses. Hershkovitz e Nachmias (2011)
T15	Scopus	A novel clustering based feature selection for classifying student performance. Regha e Rani (2015)
T16	Scopus	A self-adjusting approach for temporal dropout prediction of E-learning students. Siebra, Santos e Lino (2020)
T17	Scopus	University dropout: Prevention patterns through the application of educational data mining. Urbina-Najera, Camino-Hampshire e Barbosa (2020)

Fonte: Dados do Autor.

et al. (2018), Costa et al. (2017), Chango, Cerezo e Romero (2021) e Regha e Rani (2015) utilizaram dados oriundos de bases de dados acadêmicas. Os trabalhos de Ma et al. (2017), Lara et al. (2014), Hu, Lo e Shih (2014), Chango, Cerezo e Romero (2021), Hershkovitz e Nachmias

(2011) e Siebra, Santos e Lino (2020) utilizaram dados extraídos de um Ambiente Virtual de Aprendizagem (AVA). Zhang e Wu (2019), Chanlekha e Niramitranon (2018), Dwan, Oliveira e Fernandes (2017) e Chango, Cerezo e Romero (2021) utilizaram em seus trabalhos dados de sistemas de aprendizagem direcionados a conteúdos específicos. Por outro lado, Silva et al. (2019) e Nandeshwar, Menzies e Nelson (2011) utilizaram dados oriundos de censos da educação, enquanto Urbina-Najera, Camino-Hampshire e Barbosa (2020) utilizaram dados retirados de questionários respondidos pelos alunos.

3.2.2 QP2 - Quais foram os métodos utilizados na etapa de pré-processamento dos dados?

A etapa de pré-processamento contempla as etapas de higienização, padronização e filtragem dos dados que serão utilizados no trabalho. Os registros com valores nulos ou *outliers* foram removidos nos trabalhos de Meca et al. (2020), Zhang e Wu (2019), Silva et al. (2019), Dwan, Oliveira e Fernandes (2017) e Thammasiri et al. (2014). Também foram removidos os registros cujo atributo alvo não possuía os estados desejáveis para a classificação binária de evasão ou conclusão, como aconteceu no trabalho de Meca et al. (2020). Nos trabalhos de Meca et al. (2020), Ma et al. (2017), Zhang e Wu (2019), Chanlekha e Niramitranon (2018) e Silva et al. (2019) foram imputados valores para os dados ausentes que puderam ser estipulados.

Algumas transformações foram realizadas para adequar os dados aos algoritmos utilizados. Nos trabalhos de Zhang e Wu (2019), Chanlekha e Niramitranon (2018), Miguéis et al. (2018), Costa et al. (2017) e Chango, Cerezo e Romero (2021), as variáveis contínuas foram convertidas em variáveis categóricas. Nos trabalhos de Dwan, Oliveira e Fernandes (2017) e Chango, Cerezo e Romero (2021) houve a etapa de normalização dos dados, nas quais os valores dos atribuídos foram convertidos em valores numéricos contínuos dentro de intervalo entre zero e um. De forma similar, nos trabalhos de Lara et al. (2014), Hu, Lo e Shih (2014), Nandeshwar, Menzies e Nelson (2011) e Hershkovitz e Nachmias (2011), as variáveis categóricas foram convertidas em variáveis numéricas inteiras.

No trabalho de Urbina-Najera, Camino-Hampshire e Barbosa (2020), os registros foram filtrados de forma a selecionar uma quantidade menor de registros aplicando técnicas de seleção de amostras, mantendo uma confiança de 97,5%, e erro amostral de 5%. Os trabalhos de Regha e Rani (2015) e Siebra, Santos e Lino (2020) não descreveram as etapas de pré-processamento dos dados.

3.2.3 QP3 - Quais estratégias foram utilizadas para seleção de atributos?

Os trabalhos de Meca et al. (2020), Ma et al. (2017), Silva et al. (2019), Dwan, Oliveira e Fernandes (2017), Lara et al. (2014), Costa et al. (2017), Nandeshwar, Menzies e Nelson (2011), Chango, Cerezo e Romero (2021) e Urbina-Najera, Camino-Hampshire e Barbosa (2020) utilizaram técnicas de seleção de atributos baseada em *Filter*. Os trabalhos de Chango, Cerezo e Romero (2021), Regha e Rani (2015) e Siebra, Santos e Lino (2020) utilizaram soluções baseadas

em técnicas de *Wrapper*. Já os trabalhos de Chango, Cerezo e Romero (2021), Hershkovitz e Nachmias (2011), Siebra, Santos e Lino (2020), Chanlekha e Niramitranon (2018) e Hu, Lo e Shih (2014) utilizaram soluções *Embedded*, quando as técnicas de seleção estão embarcadas nos próprios algoritmos de aprendizagem de máquina.

Os trabalhos de Zhang e Wu (2019), Thammasiri et al. (2014), Lara et al. (2014), Miguéis et al. (2018) e Siebra, Santos e Lino (2020) utilizaram os atributos apontados pelas literaturas que apresentam maior relevância para a permanência ou evasão dos alunos. O trabalho de Siebra, Santos e Lino (2020) também utilizaram o processo de seleção manual de atributos, repetindo o processo de treinamento e validação selecionando diversos atributos diferentes, comparando, em seguintes, aqueles que apresentaram melhores resultados.

3.2.4 QP4 - Quais tecnologias foram utilizadas na predição?

Foi utilizado o algoritmo de Árvore de Decisão nos trabalhos de Meca et al. (2020), Zhang e Wu (2019), Chanlekha e Niramitranon (2018), Dwan, Oliveira e Fernandes (2017), Thammasiri et al. (2014), Hu, Lo e Shih (2014), Miguéis et al. (2018), Costa et al. (2017), Nandeshwar, Menzies e Nelson (2011), Chango, Cerezo e Romero (2021), Hershkovitz e Nachmias (2011), Regha e Rani (2015) e Urbina-Najera, Camino-Hampshire e Barbosa (2020). Nos trabalhos de Chanlekha e Niramitranon (2018), Silva et al. (2019), Dwan, Oliveira e Fernandes (2017) e Miguéis et al. (2018) foi utilizado o algoritmo de Floresta Aleatória.

Nos trabalhos de Ma et al. (2017), Chanlekha e Niramitranon (2018), Dwan, Oliveira e Fernandes (2017), Thammasiri et al. (2014) e Costa et al. (2017) foi utilizado o algoritmo de Máquina de Vetor de Suporte. O algoritmo de *Naive Bayes* foi utilizado nos trabalhos de Ma et al. (2017), Chanlekha e Niramitranon (2018), Miguéis et al. (2018), Costa et al. (2017) e Nandeshwar, Menzies e Nelson (2011). O algoritmo K-Vizinhos Mais Próximo (em inglês, *K-Nearest Neighbors*, KNN) foi utilizado nos trabalhos de Ma et al. (2017), Dwan, Oliveira e Fernandes (2017) e Lara et al. (2014). Nos trabalhos de Ma et al. (2017) e Nandeshwar, Menzies e Nelson (2011) foi utilizado o algoritmo de BN. Em Dwan, Oliveira e Fernandes (2017) também foi utilizado o algoritmo AdaBoosting. Redes Neurais Artificiais foram utilizadas nos trabalhos de Chanlekha e Niramitranon (2018), Thammasiri et al. (2014) e Costa et al. (2017).

3.3 Resultados - RSL

Nesta seção são apresentados os resultados da análise dos trabalhos selecionados na etapa de Revisão Sistemática da Literatura. Após a análise foi possível extrair informações sobre as origens dos dados utilizados em cada trabalho. O Quadro 5 apresenta a origem do dado utilizado, bem como a lista de IDs dos trabalhos e quantidades.

Ao analisar a origem dos dados foi possível constatar que a maior quantidade dos trabalhos analisados é proveniente de Base de Dados Acadêmica, tendo sete trabalho identificados, em

Quadro 5 – Origens de dados utilizadas nos Trabalhos.

Origens dos Dados	**IDs dos Trabalhos**	**Quant.**
Base de Dados Acadêmicos	T01, T04, T07, T10, T11, T13, T15	7
Ambiente Virtual de Aprendizagem	T02, T03, T04, T06, T08, T09, T13, T14, T16	9
Dados do Censos da Educação	T05, T12	2
Questionário	T17	1

Fonte: Dados do Autor.

seguida Ambiente Virtual de Aprendizagem, com nove trabalhos, Dados do Censo da Educação, com dois trabalhos e, por fim Questionário, com apenas um trabalho identificado. Após a análise, foi possível extrair informações sobre as ferramentas de seleção de atributos utilizadas em cada trabalho. O Quadro 6 apresenta o tipo de tecnologia de seleção de atributos utilizado, bem como a lista de IDs dos trabalhos e quantidade.

Quadro 6 – Técnicas de Seleção de Features utilizadas nos Trabalhos.

Seleção de Feature	**IDs dos Trabalhos**	**Quant.**
Filter	T01, T02, T05, T06, T08, T11, T12, T13, T17	9
Embedded	T04, T09, T13, T14, T16	5
Variáveis Sugeridas pela Literatura	T03, T07, T08, T10, T16	5
Wrapper	T02, T09, T13, T14	4
Manual	T16	1

Fonte: Dados do Autor.

Ao analisar as técnicas utilizadas nos trabalhos, é possível verificar a maior quantidade de trabalhos utilizando solução com tecnologia de *Filter* nove trabalhos, seguido de *Embedded* e os atributos indicados pela legislação com cinco trabalhos cada. O *Wrapper* foi utilizado por quatro trabalhos e um trabalho utilizou uma técnica manual de seleção de atributos. Dos 17 (dezessete) trabalhos, 4 (quatro) utilizaram mais de uma técnica de seleção de atributos. Com a análise dos trabalhos pesquisados, foi possível avaliar e construir o Quadro 7 com a incidência dos algoritmos mais utilizados.

Após a contagem, foi possível verificar que os algoritmos mais utilizados são aqueles baseados em árvore de decisão, com o total de treze trabalhos. Os algoritmos SVM e *Naive Bayes* foram utilizados em cinco trabalhos cada, Floresta Aleatória foi utilizado em quatro trabalhos, RNA e K-NN foram utilizado em três trabalhos cada e, por fim, BN e o algoritmo baseado em regras foram utilizados em dois trabalhos cada.

3.4 Trabalhos relacionados

Nesta seção, são apresentados trabalhos selecionados na etapa de revisão sistemática da literatura, que buscam identificar atributos relacionados à evasão escolar, para serem utilizados

Quadro 7 – Algoritmos de aprendizagem de máquina utilizados nos trabalhos.

Algoritmos	**IDs dos Trabalhos**	**Quant.**
Árvore de Decisão	T01, T03, T04, T06, T07, T09, T10, T11, T12, T13, T14, T15, T17	13
Naive Bayes	T02, T04, T10, T11, T12	5
SVM	T02, T04, T06, T07, T10	5
Floresta Aleatória	T04, T05, T06, T10	4
RNA	T04, T07, T11	3
K-NN	T02, T06, T08	3
Rede Bayes	T02, T12	2
Baseado em Regras	T13, T16	2

Fonte: Dados do Autor.

em algoritmos de classificação.

O estudo conduzido por Ma et al. (2017) teve como objetivo essencial identificar os atributos mais significativos que impactam o desempenho dos alunos. Para isso, foram empregados diversos algoritmos de classificação, incluindo Regressão Logística (RL), SVM, NB, K-NN e BN. A base de dados utilizada foi obtida da plataforma edX, uma provedora americana de cursos online em larga escala (MOOC) criada pela Harvard e pelo MIT. O trabalho aplicou técnicas de seleção de atributos, tanto de *Filter* quanto de *Wrapper*, para aprimorar a identificação dos fatores mais relevantes. Os resultados obtidos pela pesquisa indicaram que a utilização de técnicas de seleção de atributos resultou em melhorias na acurácia, exceto no caso do algoritmo BN, que mostrou piores resultados ao aplicar essas técnicas. Isso demonstra a importância de uma abordagem adaptada a cada algoritmo e contexto específico.

O estudo realizado por Silva et al. (2019) emprega técnicas de Mineração de Dados Educacionais com o objetivo de identificar os fatores vinculados à evasão escolar nas instituições de ensino superior do Brasil. A pesquisa utiliza indicadores educacionais oriundos do Banco de Dados do Censo da Educação Superior (CENSUP) e dos Indicadores de Fluxo da Educação Superior, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) no ano de 2013. Para preparar os dados, é realizada uma limpeza, removendo instâncias com valores em branco. A fim de evitar o sobreajuste, o estudo emprega os algoritmos de classificação RF e LR, adotando o método de *Bagging* para reduzir a variância e escolhendo o resultado com a maior frequência entre vários estimadores. A etapa de seleção de atributos faz uso da abordagem *Filter*. Os modelos que combinam diferentes algoritmos demonstram resultados mais promissores. Quanto aos atributos analisados, os principais fatores correlacionados à evasão e permanência nos cursos são: taxa de conclusão de disciplina, estudo noturno e a taxa de permanência de alunos no curso.

O trabalho de Nandeshwar, Menzies e Nelson (2011) utilizou mineração de dados para encontrar padrões de retenção de alunos em universidades americanas, do primeiro período. Os

dados foram extraídos do sistema de informação estudantil, selecionado os atributos relacionados a informações demográficas, acadêmicas e de ajuda financeira dos calouros do primeiro ano. Os algoritmos utilizados foram DT, NB e BN. Os resultados apresentados pelo trabalho apontam que os seis principais atributos que afetam a retenção estão relacionados à ajuda financeira.

O trabalho de Regha e Rani (2015) apresenta uma técnica para seleção de atributos, chamada *Non-negative Matrix Factorization Clustering Based Feature Selection* (NMFCBFS) que propõe uma redução dos atributos irrelevantes e redundantes. O processo utiliza o algoritmo de classificação DT e baseia-se na análise sucessiva do ganho de precisão do classificador entre diversos subconjuntos de atributos, criados a partir do conjunto de dados. Os resultados do trabalho apresentaram melhora nas métricas alcançadas, inclusive com redução do tempo, quando comparado com os dados sem a utilização da ferramenta de seleção de atributos.

O trabalho de Urbina-Najera, Camino-Hampshire e Barbosa (2020) visa identificar quais os principais fatores que influenciam a evasão nas Instituições de Ensino Superior (IES) públicas e privadas na cidade de Puebla no México. O algoritmo de classificação utilizado foi a DT o método de seleção de atributos empregado foi o *Wrapper*. Como resultados do trabalho, foram identificadas cinco principais causas relacionadas à evasão escolar universitárias, sendo elas: falta de orientação, ambiente estudantil inadequado, falta de acompanhamento acadêmico, baixa qualidade educacional e mau atendimento em geral.

Após a conclusão da revisão sistemática da literatura, ficou evidente a importância e a relevância dos estudos existentes sobre a evasão escolar. No entanto, também foi identificado um espaço para aprimoramentos e abordagens mais abrangentes. É nesse contexto que este trabalho se destaca, trazendo uma perspectiva inovadora e enriquecedora para a análise desse problema complexo. Este trabalho tem como objetivo analisar a evasão escolar dentro do IFPB, considerando diferentes perspectivas, explorando diversos agrupamentos de cursos e características dos alunos, tais como sexo, renda familiar, faixa etária, turno e origem da escola, entre outros atributos. A proposta é ir além dos trabalhos acadêmicos selecionados na etapa de revisão sistemática da literatura, realizando uma análise mais aprofundada dos atributos relacionados à evasão escolar.

Uma das principais contribuições desta pesquisa é a investigação da relação entre a evasão escolar e diferentes agrupamentos de cursos. Ao analisar os dados em grupos como Bacharelado, Especialização, Licenciatura, Mestrado, Qualificação Profissional (FIC) e Tecnologia, é possível identificar padrões específicos de evasão em cada modalidade. Isso proporciona uma compreensão mais refinada das dinâmicas que afetam a evasão em cada tipo de curso, permitindo a formulação de estratégias mais direcionadas para prevenção e intervenção.

Outro aspecto relevante deste trabalho é a busca por uma solução única que possa ser aplicada em diferentes grupos de cursos, tanto na base de dados da PNP, quanto no SUAP. Para isso, foram realizados testes, avaliando diferentes algoritmos de classificação, seletores de atributos e quantidades ideais de atributos. O objetivo foi encontrar uma solução que apresentasse

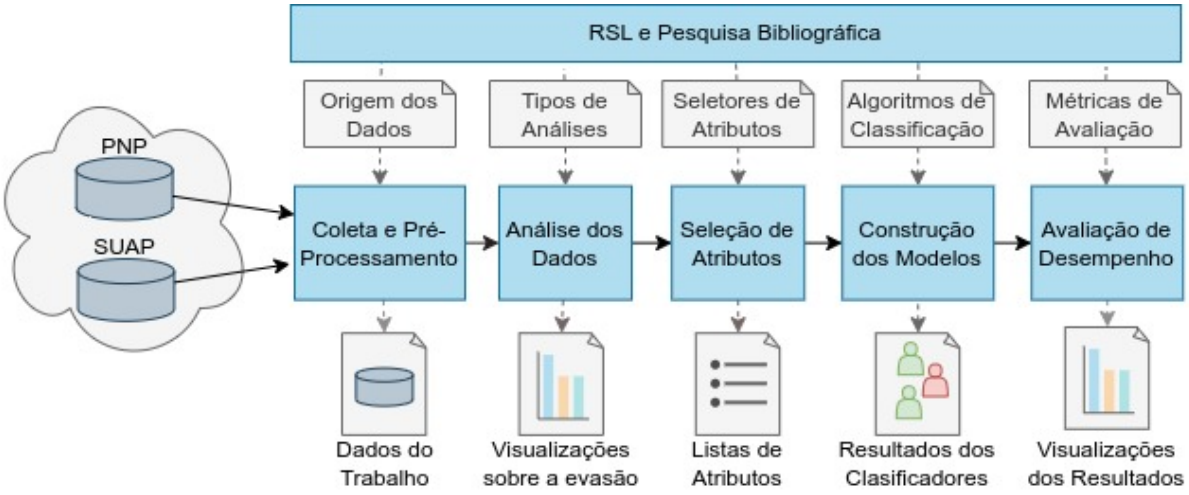
resultados consistentes e promissores em diferentes conjuntos de dados, garantindo assim a generalização dos resultados e a aplicabilidade prática do modelo desenvolvido.

Assim, este trabalho representa uma contribuição significativa para o campo da análise da evasão escolar. Ao considerar diversos agrupamentos de cursos e características dos alunos, ele fornece uma visão mais abrangente e detalhada desse fenômeno complexo. Os resultados obtidos permitem uma compreensão mais precisa dos fatores que influenciam a evasão, fornecendo subsídios para o desenvolvimento de estratégias preventivas e intervencionistas mais eficientes. Além disso, a busca por uma solução única, que possa ser aplicada em diferentes conjuntos de dados, amplia o alcance e a utilidade prática deste estudo, abrindo portas para futuras pesquisas e aplicações na área.

4 METODOLOGIA

Este capítulo apresenta a metodologia aplicada na execução do trabalho. Essa metodologia foi composta por seis atividades: revisão sistemática da literatura, coleta e pré-processamento dos dados, análise dos dados, seleção de atributos, construção dos modelos de predição e avaliação de desempenho. A Figura 13 apresenta a sequência de execução dessas atividades, bem como as etapas e os artefatos produzidos em cada uma delas. As próximas seções detalham a execução de cada atividade desenvolvida.

Figura 13 – Etapas para execução da proposta do trabalho.



Fonte: Dados do autor.

4.1 Revisão Sistemática da Literatura

Foi conduzida uma revisão sistemática da literatura detalhada no Capítulo 3, com o objetivo de identificar estudos relacionados à predição de evasão escolar utilizando técnicas de mineração de dados e ML. As etapas da revisão sistemática incluíram a seleção de estudos com base nos critérios estabelecidos, a extração de informações relevantes dos estudos selecionados e a análise dos resultados, com o objetivo de reunir materiais semelhantes de vários autores e produzir uma análise estatística sobre os trabalhos, com o objetivo de responder questões sobre formas de utilizar a mineração de dados, com foco em identificar atributos relacionados à evasão escolar. Foram selecionados 17 trabalhos relevantes que abordavam essa temática.

Dos 17 trabalhos analisados, observou-se que a origem dos dados utilizados nos estudos foi um aspecto importante. Dentre os trabalhos analisados, sete trabalhos tinham como origem os dados “Base de Dados Acadêmicos”, indicando que esses estudos utilizaram informações provenientes dos sistemas de controle acadêmico de instituições de ensino específicas. Outros nove trabalhos utilizaram dados provenientes de “Ambientes Virtuais de Aprendizagem”, que

são plataformas online utilizadas para o suporte às atividades de ensino, sistemas de gestão de aprendizagem e plataformas de ensino online.

Com base nesses resultados, verificou-se que a maioria dos trabalhos analisados utilizou dados oriundos de bases de dados acadêmicos, como sistemas de controle acadêmico e ambientes virtuais de aprendizagem. Esses dados fornecem informações relevantes sobre o desempenho acadêmico dos alunos, como o histórico educacional, os dados socioeconômicos, e outros aspectos que podem ser úteis para a predição da evasão escolar. Considerando-se essa análise e a disponibilidade de dados no contexto do IFPB, para a implementação deste trabalho, foi definida a utilização dos dados acadêmicos disponibilizados pela PNP e do módulo de controle acadêmico do SUAP. Essa escolha permitiu a utilização de informações detalhadas sobre os alunos, suas atividades acadêmicas e demais aspectos relevantes para a predição de evasão escolar no IFPB.

Além da origem dos dados, a revisão sistemática da literatura também permitiu analisar as soluções de ferramentas de seleção de atributos mais utilizados nos estudos. Dos 17 trabalhos analisados, observou-se que nove utilizaram soluções baseadas em *Filter* para seleção de atributos. Essa abordagem consiste em avaliar as características dos atributos independentemente do algoritmo de classificação utilizado. Além disso, quatro trabalhos utilizaram soluções baseadas em *Wrapper*, que envolvem a avaliação dos atributos com base no desempenho de um algoritmo de classificação específico. Por fim, cinco trabalhos utilizaram soluções baseadas em *Embedded*, que são algoritmos de classificação que incorporam a seleção de atributos diretamente durante o processo de treinamento.

Quanto aos algoritmos de classificação mais utilizados nos trabalhos analisados, observou-se a seguinte distribuição: 13 trabalhos utilizaram DT como técnica de classificação. Essa técnica é amplamente utilizada devido à sua capacidade de gerar regras de decisão claras e interpretáveis. Além disso, cinco trabalhos utilizaram NB, um algoritmo probabilístico simples e eficaz para a classificação de dados. Outros cinco trabalhos utilizaram o algoritmo SVM, conhecido por sua capacidade de lidar com problemas de classificação complexos. Também foram identificados quatro trabalhos que utilizaram RF como algoritmo de classificação, que é uma técnica que combina múltiplas árvores de decisão. Por fim, três trabalhos utilizaram MLP.

As informações obtidas por meio da revisão sistemática da literatura foram fundamentais para embasar a seleção dos algoritmos de classificação e das ferramentas de seleção de atributos neste trabalho, com foco na predição de evasão escolar. Com base nessa revisão, foram escolhidas soluções baseadas em técnicas de seleção de atributos *Filter*, *Wrapper* e *Embedded*, consideradas eficazes e relevantes em estudos anteriores na área. Quanto aos algoritmos de classificação, optou-se por utilizar DT, NB, SVM, RF e MLP, que também apresentaram comprovação de eficácia em trabalhos relacionados. Essas escolhas foram embasadas nas evidências encontradas na revisão sistemática da literatura, proporcionando uma abordagem mais sólida e confiável para a análise preditiva da evasão escolar.

4.2 Coleta e Processamento dos Dados

Os dados utilizados neste trabalho foram coletados a partir de duas fontes principais. A primeira fonte consiste nos *Microdados de Matrícula* disponíveis na PNP, que são dados abertos referentes aos alunos do IFPB. Esses microdados representam o estado dos alunos no ano anterior, por exemplo, os dados de 2018 são referentes à situação dos alunos em 2017. Esses dados foram extraídos utilizando um *script* desenvolvido em Python, que acessa os dados e realiza todo o processo de Extração, Transformação e Carga em base de dados.

A segunda fonte de dados é o módulo de gestão acadêmica do SUAP do IFPB. Esses dados foram solicitados ao setor de Coordenação de Inovação de Tecnologia da Informação da Reitoria do IFPB (COINTI-RE). Eles contêm um total de 63 atributos relacionados ao ambiente acadêmico, características socioeconômicas e dados pessoais do aluno. Para preservar a privacidade dos alunos, foram tomadas precauções para garantir a anonimização dos dados, evitando informações que pudessem identificá-los.

Após a coleta dos dados, uma fase de pré-processamento foi conduzida para adequadamente preparar os dados para a análise subsequente. Durante esse processo, diversas etapas foram empreendidas, incluindo a limpeza dos dados para a remoção de ruídos e valores atípicos. Ademais, foram empregadas estratégias para tratar valores ausentes, garantindo a integridade dos conjuntos de dados. Além disso, procedimentos de normalização e transformação foram aplicados quando necessário, buscando assimilar os dados a uma forma mais homogênea para análise. Simultaneamente, foi realizada uma exploração inicial dos dados, a qual teve como finalidade a identificação de potenciais inconsistências ou padrões significantes. Esse processo de preparação é fundamental para buscar a qualidade e a confiabilidade das análises subsequentes.

Dessa forma, as fontes de dados utilizadas neste trabalho abrangem tanto os microdados da PNP, quanto os dados do módulo acadêmico do SUAP. Essa combinação de fontes proporcionou uma visão abrangente e detalhada das informações relacionadas aos alunos do IFPB, permitindo uma análise mais precisa e embasada para a predição de evasão escolar.

4.2.1 Dados da Plataforma Nilo Peçanha

A PNP publica anualmente, desde 2018, microdados sobre as instituições de ensino que compõem a rede federal (como, por exemplo, os institutos e as universidades federais). Dentre os dados disponibilizados pela plataforma, foram utilizados nesta pesquisa os *Microdados de Matrícula*, que possuem informações relacionadas às matrículas de alunos, cursos e instituições. As linhas desse arquivo contêm o registro da situação acadêmica de cada aluno da rede federal de educação. Para cada aluno são disponibilizadas uma série de informações, como o sexo, a faixa etária, a renda, a idade, o tipo de curso, o eixo tecnológico, a modalidade de ensino, a quantidade de vagas ofertadas pelo curso, o turno do curso, a instituição de ensino, o município, a unidade federativa e a situação do aluno no curso (*em curso*, *evadidos* e *concluintes*).

Os microdados disponibilizados em cada ano se referem à situação dos alunos no ano anterior. Por exemplo, os microdados de 2018 são referentes à situação dos alunos em 2017, enquanto que os correspondentes ao ano de 2018 são representados no conjunto de dados disponibilizado em 2019 e, assim, sucessivamente. Durante a execução deste trabalho, foram usados os microdados da PNP disponibilizados nos anos de 2019, 2020, 2021 e 2022. Os dados disponibilizados em 2018 foram descartados porque apresentam uma quantidade muito reduzida de campos quando comparados com os dados dos demais anos. Assim, a sua utilização produziria uma banco de dados com muitos registros contendo campos vazios, o que prejudicaria a sua análise.

Uma vez definida a origem dos dados, a próxima etapa consistiu no desenvolvimento da ferramenta ETL, que é responsável por ler os dados obtidos na plataforma PNP, transformá-los para o esquema lógico proposto e carregá-los no banco de dados. Os dados da PNP foram extraídos, descompactados e armazenados localmente, com a utilização de um arquivo de *script* escrito na linguagem de programação Python[1] (versão 3.8.10). Todos os dados utilizados neste trabalho foram disponibilizados pela plataforma no formato CSV (*Comma-Separated Values*).

Os dados importados foram transformados e convertidos para o esquema lógico definido para o armazenamento. Essa foi a etapa mais dispendiosa do processo de ETL, devido ao grande número de registros existentes em cada arquivo e ao esforço para analisar a integridade dos dados, pois estes necessitam passar por diversos processos de higienização, padronização, filtragem e por fim, a geração de dicionário.

Após a higienização dos dados, foi possível definir um dicionário destes, o qual é apresentado na Tabela 2. Na etapa de criação desse dicionário, os campos de valores relacionados que apresentam nomes diferentes foram modificados para que todas as instâncias pudessem ser acessadas utilizando o mesmo rótulo. Dessa forma, o dicionário foi construído como forma de padronizar os nomes dos atributos para serem utilizados na integração das bases de dados selecionadas.

Tabela 2 – Dicionário dos dados da PNP.

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
ano	Número Inteiro	Ano de Carga dos Dados na Plataforma. Não disponível na Base da PNP, este campo é gerado após a coleta na etapa de transformação dos dados.	2019, 2020, 2021 e 2022.

[1] Welcome to Python.org Disponível em <https://www.python.org/>. Acesso em: 10 out. 2021.

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
carga horaria	Número Inteiro	Carga horária do ciclo de matrícula declarada no SISTEC[2] e validada na PNP pela instituição.	..., 1003, 1004, 1005, ...
carga minima	Número Inteiro	Carga horária mínima do curso de acordo com o art. 5º, §2º, da Portaria SETEC[3] nº 51, de 21 de novembro de 2018.	..., 200, 210, 240, ...
categoria situacao	Texto	Agrupa as situações de matrícula em três categorias: Concluintes, Em curso e Evadidos. O agrupamento se dá de acordo com as relações: Concluintes (Concluída, Integralizada) Em curso (Em Curso) Evadidos (Desligada, Cancelada, Abandono, Reprovada, Transf ext, Transf int)	Concluintes, Em curso e Evadidos.
codigo ciclo matricula	Texto	Código gerado no SISTEC que identifica de maneira única cada ciclo de matrícula.	..., 822608, 822611, 822612, ...
codigo municipio dv	Texto	Código IGBE[4] do município, com dígito verificador, onde está instalada a unidade de ensino.	..., 1100205, 1100304, 1100320, ...
codigo unidade ensino sistec	Número Inteiro	Código gerado no SISTEC que identifica de maneira única cada unidade de ensino.	..., 12739, 12877, 12973, ...
cod matricula	Texto	Código da matrícula no SISTEC.	..., 100000978, 100000980, 100000982, ...
cor raca	Texto	Cor/Raça do aluno declarado na PNP pela instituição. Opções: Amarela, Branca, Indígena, Parda, Preta e Não Declarada.	Amarela, Branca, Indígena, Parda, Preta e Não Declarada.

[2] SISTEC - Sistema Nacional de Informações da Educação Profissional e Tecnológica
[3] SETEC - Secretaria de Educação Profissional e Tecnológica
[4] IBGE - Instituto Brasileiro de Geografia e Estatística

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
data matricula	Data	Mês/Ano em que a matrícula foi efetivada na instituição. Não se refere à data em que a matrícula foi cadastrada no sistema. O campo foi validado/atualizado pela instituição na PNP. A referência para data é sempre o dia 1º.	..., 01/01/2014, 01/01/2015, 01/01/2016, ...
eixo tecnologico	Texto	Eixo tecnológico do curso associado pela instituição na PNP.	..., Infraestrutura, Propedêutico, Segurança, ...
faixa etaria	Texto	Agrupamento baseado na idade dos estudantes. Os grupos de faixa etária são: Menor de 14 anos, 15 a 19 anos, 20 a 24 anos, 25 a 29 anos, 30 a 34 anos, 35 a 39 anos, 40 a 44 anos, 50 a 54 anos, 55 a 59 anos e Maior de 60 anos.	Menor de 14 anos, 15 a 19 anos, 20 a 24 anos, 25 a 29 anos, 30 a 34 anos, 35 a 39 anos, 40 a 44 anos, 45 a 49 anos, 50 a 54 anos, 55 a 59 anos e Maior de 60 anos.
fator esforco curso	Número Real	Ajusta a contagem de matrículas-equivalentes para cursos que demandem, para o desenvolvimento de suas atividades, uma menor relação aluno por professor, conforme valores relacionados no Anexo II da Portaria SETEC nº 51, de 21 de novembro de 2018 e o nome de curso associado na PNP pela instituição.	(1.0), (1.001), (1.002), ...
fim ciclo	Data	Data prevista para o final do ciclo de matrícula validado/atualizado pela instituição na PNP.	..., 01/01/2018, 01/01/2019, 01/01/2020, ...
fonte financiamento	Texto	Informa se a matrícula é financiada por recursos próprios (Sem Programa Associado) ou pelos Programas UAB[5] ou PRONATEC-Rede[6] e-Tec Brasil.	Aprenda Mais, E-TEC, Outros MOOC, Recursos Orçamentários, Sem Programa Associado e UAB

[5] UAB - Universidade Aberta do Brasil

[6] PRONATEC - Programa Nacional de Acesso ao Ensino Técnico e Emprego

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
inicio ciclo	Data	Data de início do ciclo de matrícula validado/atualizado pela instituição na PNP.	..., 01/01/2011, 01/01/2012 e 01/01/2013.
instituicao	Texto	Sigla da Instituição na PNP. Cada escola técnica vinculada às Universidades Federais foi compreendida como uma Instituição.	..., IFNMG, IFPA, IFPB, ...
mes ocorrencia	Data	Mês/Ano em que a situação da matrícula efetivamente mudou. Não se refere à data em que a situação da matrícula foi atualizada no sistema. Para as matrículas que não tiveram situação alterada, corresponde ao mês de ocorrência da matrícula. A referência para data é sempre o dia 1º. Campo validado/atualizado pela instituição na PNP.	..., 01/01/2010, 01/01/2011, 01/01/2012, ...
modalidade ensino	Texto	Classificação para identificar ensino presencial ou ensino a distância.	Educação Presencial e Educação a Distância.
municipio	Texto	Nome do município onde está instalada a unidade de ensino.	..., Caicó, Cajazeiras, Camaquã, ...
nome curso	Texto	Nomenclatura adotada para padronização dos nomes de cursos de acordo com: Resoluções CNE[7], CNCT[8], CNCST[9], Guia FIC[10] , dentre outros. Os nomes dos cursos cadastrados no SISTEC foram associados às opções disponíveis na PNP. Aqueles que não possuíam correspondência foram associados à nomenclatura composta pelo "Tipo de Curso – Eixo Tecnológico".	..., Análise e Desenvolvimento de Sistemas, Apicultor, Aquicultor, ...

[7] CNE - Conselho Nacional de Educação
[8] CNCT - Catálogo Nacional de Cursos Técnicos
[9] CNCST - Catálogo Nacional de Cursos Superiores de Tecnologia
[10] FIC - Formação Inicial e Continuada

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
regiao	Texto	Região Geográfica do país onde está instalada a instituição.	Região Centro-Oeste, Região Nordeste, Região Norte, Região Sudeste e Região Sul.
renda familiar	Texto	Faixa de Renda Per Capita Familiar (RFP) do aluno, declara na PNP pela instituição (Opções: 0<RFP<=0,5; 0,5<RFP<=1; 1<RFP<=1,5; 1,5<RFP<=2,5; 2,5<RFP<=3,5; RFP>3,5; Não declarada).	"0<RFP<=0,5", "0,5<RFP<=1,0", "1,0<RFP<=1,5", "1,5<RFP<=2,5", "2,5<RFP<=3,5", "RFP>3,5" e "NÃO DECLARADA".
sexo	Texto	Informa o sexo do estudante constante no SISTEC.	Feminino, Masculino e S/I.
sub eixo tecnologico	Texto	Categorização própria da PNP para distinguir cursos de um mesmo Eixo Tecnológico em suas diferentes áreas de concentração.	..., Design, Elétrica, Estética, ...
tipo curso	Texto	Categorização transversal utilizada para diferenciar os cursos da EPCT[11] em seus diferentes níveis e graus. Opções: Educação Infantil, Ensino Fundamental I, Ensino Fundamental II, Ensino Médio, Qualificação Profissional (FIC), Técnico, Tecnologia, Licenciatura, Bacharelado, Especialização (Lato Sensu), Mestrado Profissional, Mestrado, Doutorado Profissional e Doutorado.	Educação Infantil, Ensino Fundamental I, Ensino Fundamental II, Ensino Médio, Qualificação Profissional (FIC), Técnico, Tecnologia, Licenciatura, Bacharelado, Especialização (Lato Sensu), Mestrado Profissional, Mestrado, Doutorado Profissional e Doutorado

[11] EPCT - Educação Profissional, Científica E Tecnológica

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
tipo oferta	Texto	Categorização transversal utilizada para diferenciar as formas de oferta dos Cursos Técnicos e de Qualificação Profissional (FIC). Opções: Integrado, Subsequente, Concomitante, PROEJA[12] – Concomitante e PROEJA – Integrado.	Concomitante, Integrado, Não se aplica, PROEJA - Concomitante, PROEJA - Integrado e Subsequente
total inscritos	Número Inteiro	Corresponde aos candidatos participantes dos processos seletivos que concorreram às vagas disponibilizadas para a fase inicial de um ciclo de matrícula em determinado curso, em suas diversas formas de seleção.	..., 40, 41, 42, ...
turno	Texto	Período de tempo determinado em que o aluno cursa a maior parte das aulas. Opções: matutino, vespertino, noturno ou integral. Não se aplica aos cursos com Modalidade de Ensino a Distância.	Matutino, Vespertino, Noturno, Integral e Não se aplica.
uf	Texto	Unidade da Federação onde está instalada a instituição.	..., PA, PB, PE, ...
unidade ensino	Texto	Nome da unidade de ensino a qual a matrícula está vinculada.	..., Campus Caicó, Campus Cajazeiras, Campus Camaquã, ...
vagas ofertadas	Número Inteiro	Corresponde ao número de vagas disponibilizadas para novas matrículas no início do ciclo de um curso, considerando todas as formas de ingresso disponibilizadas (vestibular, sorteio, SISU[13] ou outras formas de ingresso) no ano de referência.	..., 40, 41, 42, ...

Fonte: Dados do Autor.

Após a inclusão das bases de dados selecionadas e a aplicação dos nomes de campo de

[12] PROEJA - Programa Nacional de Integração da Educação Profissional com a Educação Básica na Modalidade de Educação de Jovens e Adultos

[13] SISU - Sistema de Seleção Unificada

acordo com o dicionário, foi gerado o conjunto de dados final. No entanto, vale ressaltar que a base de dados original não continha o campo *ano*. Para solucionar essa questão, foi realizado um procedimento em que o valor do campo *ano* foi incluído com base no período da base de dados da qual os registros foram extraídos. Por exemplo, ao extrair os registros da base de dados referente a 2020, foi atribuído o valor "2020" para o campo *ano* em todos os registros desse conjunto de dados. Dessa forma, a inclusão do campo *ano* permitiu a identificação do ano correspondente a cada conjunto de registros e facilitou a análise temporal dos dados.

Outros campos apresentam formatos diferentes nas duas bases de dados, quando referindo-se às mesmas informações dos estudantes. Por exemplo, o campo sexo, na base de 2019 apresenta os valor *F* e *M*, enquanto que na base de 2020 os valores são *Masculino* e *Feminino*. Assim, para realizar a integração das bases de dados, fez-se necessário transformar os valores desses campos para um mesmo formato, visando à padronização da base de dados final.

A análise exploratória dos dados mostrou que os dados apresentavam algumas imprecisões, como a presença de registros com valores nulos. Esses registros foram removidos da base de dados final. No campo *idade* foi verificada a existência de registros incongruentes, com por exemplo, valores negativos ou superiores a 12.000. Nesses casos, os registros foram excluídos quando fora do intervalo de 10 a 90 anos. Em seguida a remoção dos registros que apresentavam inconsistências, foi realizada uma filtragem no campo *categoria_situacao*, com o propósito de selecionar apenas os registros que apresentam os valores *Concluintes* e *Evadidos*. Os registros referentes com o valor *Em Curso* para esse campo foram descartados porque não possuem um dos estados desejados para o objetivo do trabalho, que é identificar se o aluno concluiu ou evadiu do curso.

Após a etapa de extração e transformação dos dados da Plataforma Nilo Peçanha, a próxima etapa do processo de ETL consistiu na carga dos dados em um banco de dados SQLite[14] versão 3.39.4. A escolha do banco de dados SQLite foi feita levando-se em consideração a sua simplicidade, eficiência e portabilidade. Para realizar a carga dos dados, foi criado um banco de dados vazio e estruturado, de acordo com o esquema definido no dicionário para armazenar os dados da PNP. Esse esquema foi projetado levando em consideração as necessidades de armazenamento e consulta dos dados relevantes para a predição de evasão escolar. Em seguida, os dados transformados e preparados na etapa anterior foram inseridos no banco de dados. Isso envolveu a criação de tabelas correspondentes às entidades e atributos presentes nos dados da PNP, bem como a inserção dos registros correspondentes a cada aluno. Após a conclusão da carga dos dados, o banco de dados ficou pronto para ser utilizado nas etapas subsequentes da pesquisa.

[14] SQLite. Disponível em <https://www.sqlite.org/>. Acesso em 04 set 2022.

4.2.2 Dados do SUAP

Para o processamento dos dados do SUAP, foi realizado o processo de extração, transformação e carga dos dados acadêmicos dos alunos obtidos do SUAP, por meio da solicitação à Coordenação de Inovação de Tecnologia da Informação da Reitoria (COINTI-RE) do IFPB. O objetivo principal desse processo foi coletar informações relevantes para a utilização em classificadores de evasão escolar. É importante destacar que durante esse processo foi mantida a anonimidade dos alunos, uma vez que não foi solicitado qualquer dado que pudesse identificá-los.

Inicialmente, foram obtidos os dados acadêmicos do campus de Campina Grande, que eram compostos por um total de 860 registros, sendo 529 registros do Curso Superior de Tecnologia em Telemática e 331 registros do Curso Técnico em Informática Subsequente ao Ensino Médio. Em seguida, foram obtidos os dados do campus Cajazeiras, que tinham 534 registros, sendo 325 do Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas e 209 do Curso Técnico em Informática Integrado ao Ensino Médio.

Por fim, foram obtidos os dados do campus João Pessoa. Sobre esse campus, foram disponibilizados um total de 2.439 registros, dos quais 775 eram referentes ao Curso Superior de Bacharelado em Engenharia Elétrica, 662 ao Curso Superior de Tecnologia em Sistemas para Internet, 546 ao Curso Técnico em Eletrotécnica Integrado ao Ensino Médio e 456 ao Curso Técnico em Eletrotécnica Subsequente ao Ensino Médio.

Após a obtenção dos dados, foi realizado o processo de transformação, que envolveu a limpeza e padronização dos dados, a criação de variáveis adicionais, entre outros procedimentos necessários para garantir a qualidade e a consistência dos dados. Após a limpeza dos dados, foi possível estabelecer um dicionário de dados, conforme mostrado na Tabela 3, que foi utilizado durante a execução do projeto. Durante a fase de elaboração do dicionário, os campos de valores relacionados, que possuíam denominações distintas, foram ajustados, de modo que todas as instâncias pudessem ser acessadas utilizando uma mesma identificação. Dessa forma, o dicionário foi construído com o intuito de padronizar os nomes dos atributos para facilitar a integração das bases de dados selecionadas.

Tabela 3 – Dicionário dos dados da SUAP.

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
aluno ano ingresso	Texto	Ano em que a matrícula foi efetivada na instituição.	..., 2010, 2011, 2021, ...

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
aluno cota mec	Texto	Classificação para identificar se a matrícula por alunos, oriundos de escola pública, que realizada na por ter sido contemplado pela Cota MEC conforme as seguintes categorias: com renda inferior a 1,5 Salário Mínimo (SM), declarado Preto, Pardo ou Indigena (PPI) e Pessoa com Deficiência (PCD). O aluno pode ser contemplado a mais de uma cota.	"Nao se aplica", "Oriundo de escola publica", "Oriundo de escola publica, com renda inferior a 1,5 SM", "Oriundo de escola publica, com renda inferior a 1,5 SM, declarado PCD", "Oriundo de escola publica, com renda inferior a 1,5 SM, declarado PPI", "Oriundo de escola publica, com renda inferior a 1,5 SM, declarado PPI, declarado PCD", "Oriundo de escola publica, declarado PCD", "Oriundo de escola publica, declarado PPI" e "Oriundo de escola publica, declarado PPI, declarado PCD".
aluno cota sistec	Texto	Descrição da cota em que o aluno ingressou no curso.	"Escola Publica", "Nao se aplica" e "Necessidades Especiais".
aluno data matricula	Data e Hora	Data e hora em que a matrícula foi efetivada na instituição, se refere à data em que a matrícula foi cadastrada no sistema.	..., "2007-03-21 00:00:00", "2008-07-21 00:00:00", "2009-06-16 00:00:00", ...
aluno forma ingresso	Texto	Informa o processo seletivo onde o aluno foi selecionado para realizar a matrícula no curso.	..., PSCS, PSCT, PSE, ...
aluno notas selecao	Texto	Informa a nota alcançada pelo aluno no processo seletivo, as informações podem variar entre uma única nota ou o conjunto de notas correspondente aos componentes curriculares avaliados no processo.	..., 426,12, "Enem C.N.T.:592.1;C.H.T.: 719.7; L.C.T.: 615.5; M.T.:681.0; RED.:400.0", "C.N.T 581.6; C.H.T. 516.8; L.C.T. 594.2; M.T. 733.4; RED. 820.0.", ...

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
aluno periodo ingresso	Número Inteiro	Corresponde ao número do período letivo de ingresso do aluno, podendo ser 1 para o 1º período ou 2 para o 2º período.	1 e 2.
aluno situacao sistemica	Texto	Classificação para identificar se a matrícula do aluno foi realizada diretamente no SUAP ou precisou realizar a migração da matrícula do sistema Q-Acadêmico para o SUAP.	"Matriculado no SUAP" e "Migrado do Q-Academico para o SUAP"
aluno turno	Texto	Período de tempo determinado em que o aluno cursa a maior parte das aulas. Opções: diurno, matutino, noturno e vespertino. Não se aplica aos cursos com Modalidade de Ensino a Distância.	Diurno, Matutino, Noturno e Vespertino.
aluno vinculo	Texto	Classificação para identificar se o aluno é um estudante de outra instituição, nesse caso é "Especial" ou é um estudante "Regular" quando possui matrícula em curso da própria instituição.	Especial e Regular.
categoria situacao	Texto	Agrupa as situações de matrícula em três categorias: Concluintes, Em curso e Evadidos. O agrupamento se dá de acordo com as relações: Concluintes (Concluída, Integralizada) Em curso (Em Curso) Evadidos (Desligada, Cancelada, Abandono, Reprovada, Transf ext, Transf int)	Concluintes, "Em curso" e Evadidos.
curso campus	Texto	Identificação do Campus onde o aluno possui a matrícula.	CAMPUS-CG, CAMPUS-CZ e CAMPUS-JP.
curso codigo	Texto	Código gerado pelo SUAP que identifica de maneira única o curso.	013, 121, 201, 23, 37, 61, 72 e 93.

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
curso codigo matriz	Texto	Código gerado pelo SUAP que identifica de maneira única a matriz do curso.	..., 169, 182, 190, ...
curso descricao	Texto	Nomes dos cursos.	"Curso Superior de Bacharelado em Engenharia Eletrica", "Curso Superior de Tecnologia em Analise e Desenvolvimento de Sistemas", "Curso Superior de Tecnologia em Sistemas para Internet", "Curso Superior de Tecnologia em Telematica", "Curso Tecnico em Eletrotecnica Integrado ao Ensino Medio", "Curso Tecnico em Eletrotecnica Subsequente ao Ensino Medio", "Curso Tecnico em Informatica Integrado ao Ensino Medio" e "Curso Tecnico em Informatica Subsequente ao Ensino Medio".
curso descricao matriz	Texto	Nomenclatura adotada para identificação do curso, a modalidade de ensino o Campus e a matriz.	..., "Tecnico em Eletrotecnica Subsequente - Joao Pessoa", "Tecnico em Informatica Integrado - Cajazeiras", "Tecnico em Informatica Integrado - Cajazeiras (2020.1)", ...
curso modalidade	Texto	Categorização transversal utilizada para identificar a modalidade dos Cursos. Opções: Bacharelado, Integrado, Subsequente e Tecnologia.	Bacharelado, Integrado, Subsequente e Tecnologia.
curso nivel ensino	Texto	Categorização transversal utilizada para identificar o nível dos Cursos. Opções: Graduacao e Medio.	Graduacao e Medio.

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
endereco cidade	Texto	Nome da cidade onde reside o aluno.	..., “Caicara - PB”, "Cajazeiras - PB”, "Caldas Brandao - PB”, ...
escola anterior ano conclusao	Texto	Ano de conclusão da habilitação anterior do estudante.	..., 1995, 1996, 1997, ...
escola anterior cidade	Texto	Nome da cidade onde o aluno estudou anteriormente ao curso atual.	..., “Jequie - BA”, "Joao Pessoa - PB”, "Joaquim Nabuco - PE”, ...
escola anterior nivel ensino	Texto	Nível de habilitação anterior do estudante.	-, Fundamental, Graduacao, Medio e Pos-graduacao.
escola anterior nome	Texto	Nome da instituição de ensino onde o aluno estudou anteriormente ao curso atual.	..., “CENTRO DE ENSINO CORUJINHA”, "CENTRO DE ENSINO DECISAO”, "CENTRO DE ENSINO EDUCA NEXUS”, ...
escola anterior tipo	Texto	Categorização transversal utilizada para identificar o tipo de instituição de ensino onde o aluno estudou anteriormente. Opções: Privada e Publica.	Privada e Publica.
faixa etaria	Texto	Agrupamento baseado na idade dos estudantes. Os grupos de faixa etária são: Menor de 14 anos, 15 a 19 anos, 20 a 24 anos, 25 a 29 anos, 30 a 34 anos, 35 a 39 anos, 40 a 44 anos, 50 a 54 anos, 55 a 59 anos e Maior de 60 anos.	"Menor de 14 anos”, "15 a 19 anos”, "20 a 24 anos”, "25 a 29 anos”, "30 a 34 anos”, "35 a 39 anos”, "40 a 44 anos”, "45 a 49 anos”, "50 a 54 anos”, "55 a 59 anos” e "Maior de 60 anos".
pessoal cor raca	Texto	Cor/Raça do aluno declarado no momento da matrícula na instituição. Opções: Amarela, Branca, Indígena, Parda, Preta e Não Declarada.	Amarela, Branca, Indigena, Parda, Preta e "Nao declarada".

Nome do Campo	**Tipo**	**Descrição**	**Exemplo dos Dados**
pessoal estado civil	Texto	Estado civil declarado no momento da matrícula na instituição. Opções: -, Casado, Divorciado, Solteiro, “Uniao Estavel” e Viuvo.	-, Casado, Divorciado, Solteiro, "Uniao Estavel” e Viuvo.
pessoal nacionalidade	Texto	Nacionalidade do aluno.	Brasileira, "Brasileira - Nascido no exterior” e Estrangeira.
pessoal naturalidade	Texto	Naturalidade do aluno.	..., “Caico - RN”, "Cajazeiras - PB”, "Camacari - BA”, ...
pessoal sexo	Texto	Informa o sexo do estudante.	Feminino e Masculino.

Fonte: Dados do Autor.

Por fim, os dados foram carregados em um banco de dados SQLite, que serviu como fonte de dados para as etapas seguintes do projeto. Assim como aconteceu com os dados da PNP, foram selecionados apenas os registros que possuíam os valores *Concluintes* ou *Evadidos*. Dessa forma, foi possível direcionar a análise para os casos de conclusão e evasão, que são de interesse principal para o treinamento dos algoritmos de classificação neste trabalho.

4.3 Análise dos Dados

Nesta etapa, os dados obtidos da PNP e do SUAP foram analisados. Durante esse processo, foram realizadas análises descritivas em diversos agrupamentos dos dados, com o objetivo de investigar o problema da evasão escolar. As análises foram realizadas com base na comparação do campo *categoria_situação* (que pode assumir os valores *Concluídos* ou *Evadidos*), a fim de se identificar possíveis padrões e *insights* relacionados à evasão. Para os dados da PNP, os seguintes subconjuntos foram criados com base na categoria de situação:

- **Todos os cursos**: inclui todos os registros dos cursos do IFPB presentes nos dados da PNP;
- **Bacharelado**: agrupa os registros dos cursos de graduação do tipo bacharelado;
- **Especialização**: contém os registros dos cursos de pós-graduação do tipo lato sensu;
- **Licenciatura**: inclui os registros dos cursos de graduação do tipo licenciatura;
- **Mestrado**: agrupa os registros dos cursos de pós-graduação do tipo mestrado;

- **Qualificação Profissional (FIC)**: contém os registros dos cursos de formação inicial e continuada (FIC) de curta duração;
- **Tecnologia**: inclui os registros dos cursos de graduação do tipo tecnologia;
- **Técnico**: agrupa os registros dos cursos técnicos de nível médio.

Para os dados do SUAP, os seguintes subconjuntos foram criados com base na categoria de situação:

- **Todos os cursos**: inclui todos os registros de cursos do IFPB presentes nos dados do SUAP;
- **Bacharelado**: agrupa os registros de cursos de graduação do tipo bacharelado;
- **Tecnologia**: contém os registros de cursos de graduação do tipo tecnologia;
- **Integrado**: inclui os registros de cursos técnicos integrados ao ensino médio.
- **Subsequente**: agrupa os registros de cursos técnicos subsequentes ao ensino médio.

Essa segmentação dos dados permitiu uma análise mais específica da evasão escolar em cada subconjunto de cursos. As análises descritivas foram utilizadas para obter estatísticas resumidas sobre a evasão em cada subconjunto, como a proporção de alunos evadidos e concluídos. Durante a análise dos dados da PNP, foi identificada a quantidade de alunos concluídos e evadidos, com base nos seguintes campos: "*ano*", "*sexo*", "*cor_raca*", "*faixa_etaria*", "*renda_familiar*", e "*turno*". Por outro lado, durante a análise dos dados do SUAP, os dados foram avaliados com base nos seguintes campos: "*turno*", "*aluno_cota_sistec*", "*aluno_cota_mec*", "*endereco_zona_residencial*", "*escola_anterior_tipo*", "*faixa_etaria*", "*pessoal_cor_raca*" e "*pessoal_estado_civil*". Para cada um desses campos, foram realizadas análises descritivas, a fim de investigar possíveis associações e padrões relacionados à evasão escolar.

As análises descritivas permitiram obter uma visão geral da distribuição dos dados e identificar características predominantes em cada campo. Para cada análise, foram calculadas estatísticas como contagem e proporção, dependendo da natureza dos dados em questão. Essa análise abrangente dos campos, nos conjuntos de dados da PNP e do SUAP, permitiu a identificação de características demográficas, socioeconômicas e acadêmicas, que podem estar relacionadas à evasão escolar. Esses *insights* podem contribuir para a compreensão dos fatores de risco e a implementação de medidas preventivas e de intervenção eficazes.

4.4 Predição da Evasão Escolar

A etapa de predição de evasão escolar adotada neste trabalho consistiu em realizar uma série de procedimentos para definir os seletores de atributos, a quantidade de campos

e as ferramentas de classificação a serem utilizadas para cada conjunto de dados, tanto o da PNP quanto o do SUAP. Após os conjuntos de dados PNP e SUAP serem obtidos, a partir de suas respectivas fontes, e realizado um processo de ETL para construir a base de dados geral do projeto, houve a separação dos grupos dentro do conjunto de dados da PNP, utilizando o campo "*tipo_curso*", e do conjunto de dados do SUAP, utilizando o campo "*curso_modalidade*". Essa separação foi realizada para analisar os padrões e tendências da evasão escolar em cada agrupamento específico.

Para a seleção de atributos, foram adotadas diferentes soluções, como o seletor *Kbest*, *Chi2*, *Wrapper GB*, *Wrapper LR* e *Embedded* do *Random Forest*. Esses seletores foram utilizados para identificar os atributos mais relevantes em cada conjunto de dados. Além disso, foram definidas as quantidades de atributos selecionadas, variando entre 1 e 31, a fim de investigar a influência da quantidade de campos nas análises e predições da evasão escolar. A etapa seguinte envolveu a classificação dos dados utilizando os algoritmos DT, RF, NB, MLP e SVM. A métrica utilizada para avaliar o desempenho dos classificadores foi a *F1 Score*. Todos os resultados obtidos foram armazenados na base de dados de resultados, permitindo a identificação dos melhores para cada grupo. No caso da PNP, foi realizado o agrupamento pelos campos "*tipo_curso*", enquanto no SUAP foi realizado o agrupamento pelo campo "*curso_modalidade*".

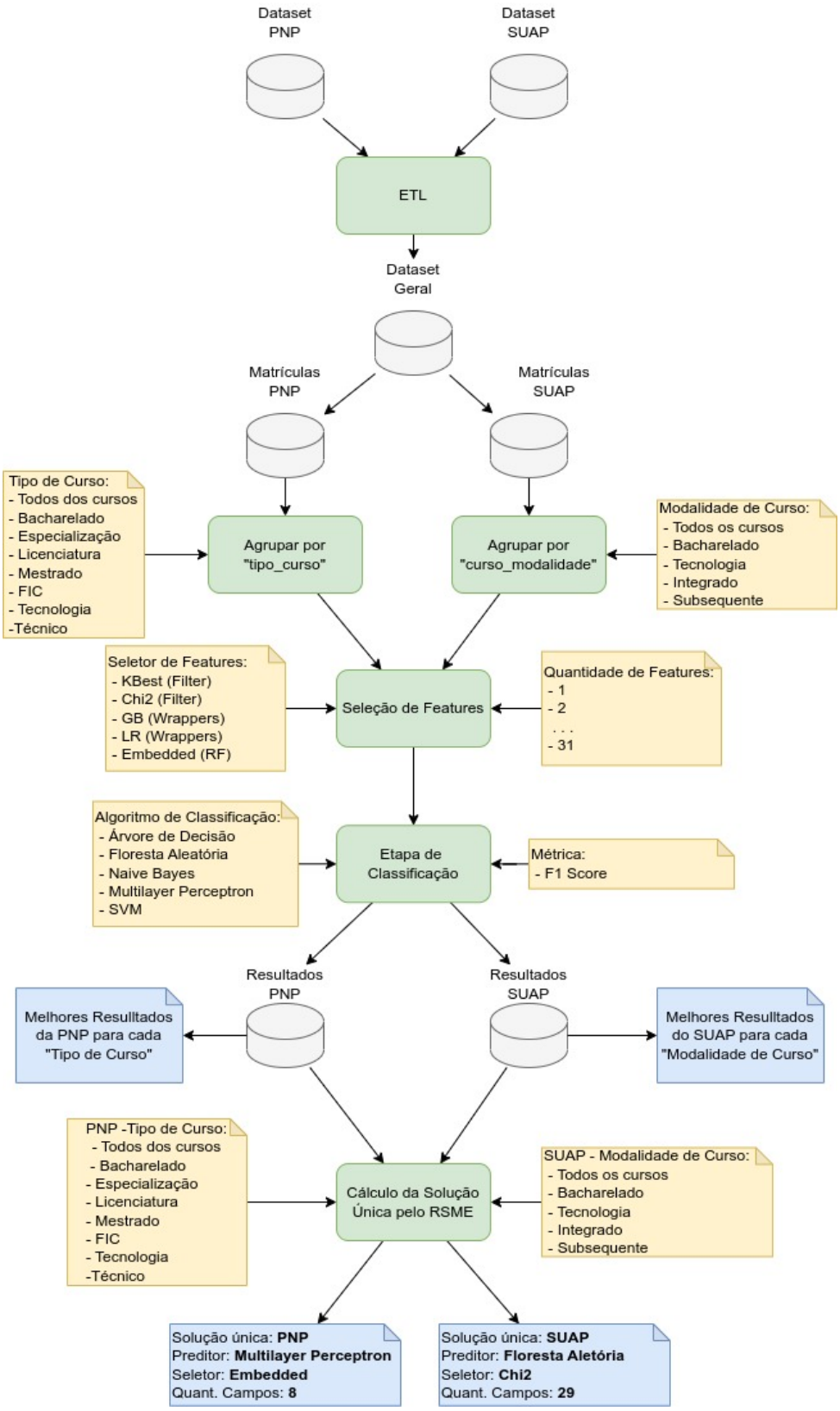
Após isso, foi calculada a solução única para os dados da PNP e outra para os dados do SUAP, utilizando o agrupamento por "*tipo_curso*" para a PNP e "*curso_modalidade*" para o SUAP. Essas soluções únicas buscaram identificar o melhor conjunto de atributos e algoritmos de classificação para cada grupo. Por fim, foram identificados os resultados das soluções únicas. Para o grupo da PNP, o melhor desempenho foi obtido com o algoritmo MLP em conjunto com o seletor *Embedded* e utilizando oito campos. Já para o grupo do SUAP, a melhor solução foi encontrada com o algoritmo RF em conjunto com o seletor *Chi2* e utilizando 29 campos. Uma representação gráfica dessas etapas e resultados está apresentada na Figura 14.

Através dessa metodologia, foi possível identificar as melhores combinações de atributos e algoritmos de classificação para prever a evasão escolar, nos diferentes agrupamentos de cursos e características dos alunos nos conjuntos de dados da PNP e SUAP. Os processos descritos acima serão apresentados e detalhados, de forma mais ampla, nas seções subsequentes deste trabalho. Nas próximas seções, serão fornecidos mais detalhes sobre a seleção de atributos, construção dos modelos de predição e avaliação de desempenho. Essas informações complementares permitirão uma compreensão mais abrangente e aprofundada de todo o processo metodológico adotado neste estudo.

4.5 Seleção de Atributos

Durante a etapa de seleção de atributos foram utilizadas cinco soluções distintas: *KBest*, *Chi2*, *Wrappers Gradiente Boosting*, *Wrappers Logistic Regression* e *Embedded* do *Random*

Figura 14 – Diagrama das etapas e predição da evasão.



Fonte: Dados do autor.

Forest. Essas soluções foram escolhidas com base na revisão sistemática da literatura, que identificou que elas são as mais utilizadas nos trabalhos selecionados.

Para a implementação dessas técnicas, foi utilizado o pacote "*scikit-learn*"[15] na versão 1.1.2, uma biblioteca de código aberto amplamente reconhecida no campo da ML. No caso do *KBest*, utilizou-se a classe "*SelectKBest*" para selecionar os melhores atributos. O teste Qui-quadrado (Chi2) foi aplicado por meio da função "*chi2*" do "*scikit-learn*". Para as abordagens Wrapper, a biblioteca foi empregada com os métodos RFE (*Recursive Feature Elimination*) com os algoritmos L*ogistic Regression* e *Gradient Boosting Classifier*. A técnica Ensemble foi implementada com o uso do algoritmo "*RandomForestClassifier*". Essas bibliotecas facilitaram a implementação eficiente das técnicas, auxiliando na identificação das características mais relevantes para a previsão da evasão escolar e no aprimoramento dos modelos de classificação resultantes.

A seleção de atributos foi realizada por meio de iterações, começando com a seleção de apenas um atributo e aumentando gradativamente até o total de atributos presentes nos conjuntos de dados, que é igual a 31. As duas soluções baseadas em filtro, *KBest* e *Chi2*, são amplamente empregadas na seleção de atributos. O *KBest* seleciona os *k* melhores atributos com base em um critério estatístico, enquanto o *Chi2* utiliza o teste qui-quadrado para medir a dependência entre cada atributo e a classe-alvo.

As soluções baseadas em *Wrappers*, *Wrappers Gradiente Boosting* e *Wrappers Logistic Regression*, envolvem o treinamento iterativo de modelos com diferentes conjuntos de atributos. Esses *Wrappers* avaliam o desempenho do modelo em cada iteração, selecionando gradualmente os atributos mais relevantes.

A solução *Embedded* do *Random Forest* combina características de *Filter* e *Wrapper*. Nessa abordagem, o algoritmo *Random Forest* é utilizado para estimar a importância de cada atributo durante o processo de treinamento. Os atributos com maior importância são selecionados como relevantes para o modelo. A escolha dessas cinco soluções de seleção de atributos foi motivada pelo fato de serem as mais frequentemente mencionadas nos trabalhos selecionados durante a revisão sistemática da literatura. Essas soluções demonstraram eficácia em diversos estudos anteriores, contribuindo para a seleção dos atributos mais relevantes para a predição da evasão escolar.

Ao empregar essas soluções de seleção de atributos e realizar iterações com diferentes quantidades de atributos selecionados, desde 1 até o total de 31 atributos presentes nos conjuntos de dados, buscou-se identificar quais características dos alunos têm maior influência na evasão escolar e qual a quantidade ideal de atributos para cada conjunto de dados. Essa análise abrangente permitiu compreender a importância relativa de cada atributo e sua contribuição para o processo de predição. Além da iteração com diferentes quantidades de atributos, o processo de seleção foi

[15] scikit-learn Machine Learning in Python. Disponível em <https://scikit-learn.org/stable/>. Acesso em: 05 ago. 2022

executado em diversos conjuntos e subconjuntos de dados, utilizando atributos específicos para agrupar as informações. Essa abordagem permitiu uma análise mais detalhada e segmentada dos conjuntos de dados da PNP e do SUAP.

Para a análise dos dados da PNP, os conjuntos foram agrupados com base no atributo "*tipo_curso*". Essa categorização permitiu identificar padrões e características específicas relacionadas a cada tipo de curso, auxiliando na compreensão dos fatores que podem influenciar a evasão escolar em diferentes contextos educacionais.

Da mesma forma, para os dados do SUAP, os conjuntos foram agrupados utilizando o atributo "*curso_modalidade*". Essa abordagem proporcionou uma visão mais detalhada das características socioeconômicas e acadêmicas dos alunos, levando em consideração as particularidades de cada modalidade de curso. Dessa forma, ao realizar a seleção de atributos nos diferentes grupos e subgrupos de dados, foi possível identificar os fatores mais relevantes para a evasão escolar em cada tipo de curso ou modalidade. Essa análise granular possibilita o desenvolvimento de estratégias de intervenção personalizadas, voltadas para as necessidades específicas de cada grupo de alunos.

Para avaliar o resultado da seleção de atributos, foi utilizado um método baseado no cálculo do RMSE, selecionando o melhor seletor em cada conjunto de dados, passando por uma análise individual, na qual foram testados diferentes seletores. Para cada seletor, foram geradas previsões utilizando os modelos de classificação e calculado o RMSE correspondente. O objetivo era identificar o seletor que apresentasse o menor erro, indicando a melhor seleção de atributos para o conjunto de dados em questão. Após a avaliação em cada subconjunto de dados, o seletor que obteve o menor RMSE foi selecionado como solução para as próximas etapas do processo. A escolha do seletor com base no menor erro RMSE garante que os atributos selecionadas possuam maior poder de explicação e relevância para a saída, contribuindo para a precisão do modelo.

Esses agrupamentos por tipo de curso ou modalidade permitem uma análise mais focada nos diferentes contextos educacionais presentes nos conjuntos de dados. Ao realizar a seleção de atributos em cada um desses subconjuntos, foi possível identificar os fatores mais relevantes para a evasão escolar em cada tipo de curso ou modalidade, além da iteração com diferentes quantidades de atributos selecionados e a análise segmentada dos dados com base nos atributos de agrupamento nos proporciona uma compreensão mais aprofundada dos fatores, que influenciam a evasão escolar em cada contexto educacional.

4.6 Construção dos Modelos de Predição

Nesta seção, é descrito o processo de construção dos modelos foi utilizado o pacote "*scikit-learn*" para os cinco classificadores distintos: DT, RF, NB, MLP e SVM. Essa seleção foi baseada na Revisão Sistemática da Literatura, que destacou a frequente utilização desses classificadores em trabalhos relacionados ao tema. O processo de construção dos modelos foi

realizado de forma iterativa e abrangente. Inicialmente, cada classificador foi implementado e treinado utilizando o conjunto de dados selecionados. Em seguida, foram ajustados os parâmetros e configurações específicas de cada algoritmo, visando otimizar o desempenho dos modelos.

Para o algoritmo DT, foi empregado o "*DecisionTreeClassifier*", sendo adotadas a profundidade máxima da árvore (*max_depth*) definida como 5 e critério de divisão *splitter* como "*best*". A algoritmo RF utilizou o "*RandomForestClassifier*", com a configuração do número de estimadores (*n_estimators*) definido como 100, profundidade máxima das árvores (*max_depth*) igual a 10. Para o NB, foi utilizado o algoritmo "*ComplementNB*", com as Configurações padrão do algoritmo "*ComplementNB*". A arquitetura de rede neural artificial MLP foi usado "*MLPClassifier*", com a configuração de uma camada oculta com 100 neurônios, função de ativação *ReLU*, otimizador Adam e máximo de 200 iterações. No caso do algoritmo SVM, foi configurado o kernel linear padrão do algoritmo "*SVC*". A fim de tratar o desbalanceamento das classes, foi aplicado o método SMOTE. Quanto à divisão dos conjuntos de treinamento e teste, a estratégia de *cross-validation* com 10 conjuntos foi adotada para uma avaliação mais robusta dos modelos. Essas configurações, embasadas no referencial teórico, visaram a obtenção de resultados representativos e confiáveis na avaliação do desempenho dos modelos de classificação.

Posteriormente, para aprimorar a análise e obter uma configuração mais robusta, o processo foi repetido 31 vezes para cada classificador. Em cada repetição, selecionamos um atributo adicional do conjunto de dados utilizando cinco soluções de seleção de atributos distintas: *KBest*, *Chi2*, *Wrappers Gradiente Boosting*, *Wrappers Logistic Regression* e *Embedded* do *Random Forest*. Essas ferramentas de seleção de atributos foram escolhidas com base na sua eficácia e ampla utilização também identificada na RSL. Isso resultou em um total de 155 repetições, explorando diferentes combinações de atributos selecionados pelas ferramentas de seleção. Cada repetição permitiu treinar e avaliar os modelos com um conjunto de atributos único, buscando identificar quais atributos têm maior relevância na predição de evasão escolar.

Durante cada repetição, os modelos foram avaliados utilizando a métricas *F1-Score*, com o objetivo de avaliar o desempenho dos modelos, permitindo analisar sua capacidade de prever corretamente as classes de evasão escolar. Além disso, é importante ressaltar que todo o processo descrito acima foi executado considerando diversos conjuntos e subconjuntos dos dados. Para essa segmentação, foram utilizados dois atributos-chave: "*tipo_curso*" para agrupar os dados da PNP e "*curso_modalidade*" para agrupar os dados do SUAP.

Na construção dos modelos de previsão, os conjuntos de dados foram segmentados com base em atributos como "*tipo_curso*" e "*curso_modalidade*". Isso permitiu uma análise detalhada do desempenho dos modelos em diferentes contextos acadêmicos. Para a PNP, a categorização por "*tipo_curso*" revelou padrões específicos para cada tipo, como bacharelado, especialização, licenciatura, entre outros. O mesmo aconteceu com o conjunto de dados do SUAP, agrupado por "*curso_modalidade*", destacando particularidades em modalidades como bacharelado, tecnologia, integrado e subsequente. Essa abordagem possibilitou a identificação

precisa de padrões e tendências relacionados à evasão em cada tipo de curso e modalidade.

Todos os resultados obtidos para as métricas avaliadas em cada repetição foram persistidos no banco de dados. Ao persistir os resultados das métricas avaliadas no banco de dados, foi garantido que as análises posteriores possam levar em consideração não apenas o desempenho geral dos modelos, mas também a variação de resultados entre os diferentes conjuntos e subconjuntos de dados. Essa abordagem permite armazenar e organizar os dados para análises posteriores, facilitando a identificação da melhor configuração possível dos modelos de predição. O objetivo dessa abordagem iterativa e persistência dos resultados é realizar análises detalhadas e comparativas entre as diferentes configurações dos modelos, com o intuito de identificar a combinação mais eficaz de atributos e classificadores para prever casos de evasão escolar.

4.7 Avaliação de Desempenho

A utilização das métricas acurácia, precisão, revocação, especificidade, AUC-ROC e *F1-Score* proporciona uma avaliação mais completa e precisa do desempenho dos modelos de predição de evasão escolar, permitindo uma compreensão aprofundada e embasada para o desenvolvimento de estratégias de prevenção e intervenção eficazes. Cada resultado obtido para as métricas mencionadas foi persistido no banco de dados, garantindo que todas as informações relevantes estejam disponíveis para análises posteriores. Essa abordagem permite a comparação e a identificação da melhor configuração de atributos e classificadores com base nas métricas de desempenho. Porém como forma de avaliar os resultados dos preditores dentro do contexto de delimitação do trabalho e considerando a necessidade de escolher uma única métrica para avaliar o desempenho dos modelos de predição de evasão escolar, uma métrica aceitável é o *F1-Score*.

5 RESULTADOS

Neste capítulo, são apresentados os resultados obtidos a partir da análise dos dados educacionais do IFPB e da aplicação dos modelos de predição de evasão escolar. Os resultados são organizados em três seções principais: análise dos dados, resultados da seleção de atributos e resultados dos classificadores.

5.1 Análise dos Dados

Nesta seção, são apresentados os resultados da análise dos dados educacionais do IFPB. Essa análise descritiva permite identificar algumas diferenças na taxa de conclusão dos alunos, variando de acordo com o tipo de curso, ano de análise, sexo dos estudantes, cor/raça, faixa etária, renda familiar e turno das aulas. Tais informações podem ser úteis para um melhor entendimento dos desafios e oportunidades enfrentados pelos alunos em diferentes áreas e contribuir para a implementação de ações voltadas para a melhoria da taxa de conclusão e do sucesso acadêmico em cada contexto específico.

5.1.1 Análise dos Dados da PNP em Relação ao Ano

A análise descritiva dos grupos de tipo de curso do IFPB, com base nos dados da PNP, revelou informações importantes sobre as taxas de aprovação em cada categoria, agrupadas em função dos anos. Os resultados estão apresentados na Tabela 4.

Tabela 4 – Análise da PNP para Tipo de Curso em Relação ao Ano.

Tipo de Curso	Ano	Concluintes	Evadidos	Total
Bacharelado	2019	98 (37,98%)	160 (62,02%)	258 (9,92%)
	2020	134 (44,22%)	169 (55,78%)	303 (11,64%)
	2021	157 (57,51%)	116 (42,49%)	273 (10,49%)
	2022	205 (11,60%)	1.563 (88,40%)	1.768 (67,95%)
		594 (22,83%)	**2.008 (77,17%)**	**2.602 (6,62%)**
Especialização (Lato Sensu)	2019	47 (58,02%)	34 (41,98%)	81 (9,98%)
	2020	146 (76,04%)	46 (23,96%)	192 (23,65%)
	2021	38 (50,00%)	38 (50,00%)	76 (9,36%)
	2022	355 (76,67%)	108 (23,33%)	463 (57,02%)
		586 (72,17%)	**226 (27,83%)**	**812 (2,07%)**
Licenciatura	2019	149 (26,37%)	416 (73,63%)	565 (15,86%)
	2020	89 (14,50%)	525 (85,50%)	614 (17,24%)
	2021	129 (21,18%)	480 (78,82%)	609 (17,10%)

Tipo de Curso	Ano	Concluintes	Evadidos	Total
	2022	198 (11,16%)	1.576 (88,84%)	1.774 (49,80%)
		565 (15,86%)	**2.997 (84,14%)**	**3.562 (9,07%)**
Mestrado	2019	7 (77,78%)	2 (22,22%)	9 (19,15%)
	2020	14 (87,50%)	2 (12,50%)	16 (34,04%)
	2021	10 (90,91%)	1 (9,09%)	11 (23,40%)
	2022	11 (100,00%)	0 (0,00%)	11 (23,40%)
		42 (89,36%)	**5 (10,64%)**	**47 (0,12%)**
Mestrado Profissional	2019	0 (0,00%)	1 (100,00%)	1 (1,96%)
	2020	0 (0,00%)	1 (100,00%)	1 (1,96%)
	2021	11 (78,57%)	3 (21,43%)	14 (27,45%)
	2022	29 (82,86%)	6 (17,14%)	35 (68,63%)
		40 (78,43%)	**11 (21,57%)**	**51 (0,13%)**
Qualificação Profissional (FIC)	2019	268 (59,03%)	186 (40,97%)	454 (4,80%)
	2020	896 (62,79%)	531 (37,21%)	1.427 (15,08%)
	2021	288 (42,35%)	392 (57,65%)	680 (7,18%)
	2022	2.737 (39,64%)	4.167 (60,36%)	6.904 (72,94%)
		4.189 (44,26%)	**5.276 (55,74%)**	**9.465 (24,10%)**
Tecnologia	2019	376 (29,58%)	895 (70,42%)	1.271 (15,81%)
	2020	530 (35,26%)	973 (64,74%)	1.503 (18,70%)
	2021	243 (39,45%)	373 (60,55%)	616 (7,66%)
	2022	422 (9,08%)	4.226 (90,92%)	4.648 (57,83%)
		1.571 (19,54%)	**6.467 (80,46%)**	**8.038 (20,46%)**
Técnico	2019	1.432 (40,66%)	2.090 (59,34%)	3.522 (23,96%)
	2020	2.089 (50,75%)	2.027 (49,25%)	4.116 (28,00%)
	2021	1.593 (55,31%)	1.287 (44,69%)	2.880 (19,59%)
	2022	2.755 (65,88%)	1.427 (34,12%)	4.182 (28,45%)
		7.869 (53,53%)	**6.831 (46,47%)**	**14.700 (37,43%)**
Todos os Tipos de Curso	2019	2.377 (38,58%)	3.784 (61,42%)	6.161 (15,69%)
	2020	3.898 (47,70%)	4.274 (52,30%)	8.172 (20,81%)
	2021	2.469 (47,86%)	2.690 (52,14%)	5.159 (13,13%)
	2022	6.712 (33,92%)	13.073 (66,08%)	19.785 (50,37%)
		15.456 (39,35%)	**23.821 (60,65%)**	**39277**

Fonte: Dados do Autor.

Ao analisar os diferentes grupos de tipos de cursos, é possível observar que a taxa de aprovação varia consideravelmente. Os cursos de Licenciatura apresentam a maior taxa de evasão, com 84,14% dos alunos abandonando o curso. Em seguida, os cursos de Tecnologia, com uma taxa de evasão de 80,46%, e os cursos de Bacharelado, com 77,17%. Por outro lado, os tipos

de curso com as menores taxas de evasão são Especialização (*Lato Sensu*), com 27,83% de abandono, seguido pelo Mestrado Profissional, com 21,57%, e Mestrado, com 10,64%.

Existem também os tipos de curso com taxas de evasão próxima a 50%. Os cursos de Qualificação Profissional (FIC) possuem uma taxa de evasão de 55,74%, enquanto os cursos Técnicos apresentam uma taxa de evasão de 46,47%. Considerando todos os grupos de tipos de cursos juntos, a taxa de evasão média é de 60,65%. Esses resultados evidenciam diferenças significativas nas taxas de evasão entre os grupos de tipo de curso do IFPB. Essas informações podem ser relevantes para a identificação de áreas que necessitam de maior atenção e intervenção, buscando melhorar as taxas de aprovação e reduzir a evasão escolar em determinados cursos.

Quando analisada a quantidade de registros em relação ao ano, é possível observar um aumento considerável no número de registros quando comparados os anos de 2021 e 2022. Esse crescimento é observado em quase todos os grupos de tipos de curso, com exceção ao grupo Mestrado que manteve os 11 (onze) registros. Quando contabilizado a diferença total de registros entre 2021 (que era 5.159 registros), para 2022 (que passou a ser 19.785), isso representa um acréscimo superior a 283%.

Algumas possibilidades podem ser cogitadas para explicar esse aumento no número de registros entre os anos de 2021 e 2022. É possível que tenha havido um aumento no número de alunos matriculados nas Instituições Federais entre estes anos, como também é possível que tenha havido sub-registro nos anos anteriores, onde em 2022 essa diferença pode ter sido corrigida. Ao considerarmos a análise dos grupos de tipo de curso em função do ano, observa-se algumas tendências relacionadas à taxa de conclusão dos alunos nos cursos. É importante ressaltar que, junto com o aumento no número de registros nos anos de 2021 e 2022, foram identificadas variações significativas na taxa de evasão.

Ao analisar as taxas de evasão por tipo de curso, verifica-se algumas tendências preocupantes. Os cursos de Bacharelado apresentaram um aumento significativo na taxa de evasão, passando de 42,49% em 2021 para 88,40% em 2022. Da mesma forma, os cursos de Licenciatura tiveram um acréscimo na taxa de evasão, aumentando de 78,82% em 2021 para 88,84% em 2022. No entanto, alguns cursos apresentaram redução na taxa de evasão. Destaca-se o grupo de Especialização (*Lato Sensu*), que reduziu consideravelmente a taxa de evasão, passando de 50,00% em 2021 para 23,33% em 2022. O curso de Mestrado também registrou uma queda, indo de 9,09% em 2021 para 0,00% em 2022, embora deva ser levado em consideração que o número de registros é pequeno para esse grupo.

Em relação ao Mestrado Profissional, houve uma pequena queda na taxa de evasão, passando de 21,43% em 2021 para 17,14% em 2022. Já os cursos de Qualificação Profissional (FIC) e Técnico apresentaram taxas de evasão relativamente estáveis, com uma pequena variação de 57,65% para 60,36% e de 44,69% para 34,12%, respectivamente.

5.1.2 Análise dos Dados da PNP em Relação ao Sexo

A análise descritiva dos grupos de tipo de curso do IFPB, utilizando os dados da PNP, trouxe informações significativas sobre as taxas de aprovação em cada categoria, classificadas conforme o campo sexo dos estudantes. Os resultados podem ser encontrados na Tabela 5.

Tabela 5 – Análise da PNP para Tipo de Curso em Relação ao Sexo.

Tipo de Curso	Sexo	Concluintes	Evadidos	Total
Bacharelado	Feminino	256 (26,64%)	705 (73,36%)	961 (36,93%)
	Masculino	338 (20,60%)	1.303 (79,40%)	1.641 (63,07%)
		594 (22,83%)	**2.008 (77,17%)**	**2.602 (6,62%)**
Especialização (Lato Sensu)	Feminino	397 (76,20%)	124 (23,80%)	521 (64,16%)
	Masculino	189 (64,95%)	102 (35,05%)	291 (35,84%)
		586 (72,17%)	**226 (27,83%)**	**812 (2,07%)**
Licenciatura	Feminino	301 (16,65%)	1.507 (83,35%)	1.808 (50,76%)
	Masculino	264 (15,05%)	1.490 (84,95%)	1.754 (49,24%)
		565 (15,86%)	**2.997 (84,14%)**	**3.562 (9,07%)**
Mestrado	Feminino	11 (84,62%)	2 (15,38%)	13 (27,66%)
	Masculino	31 (91,18%)	3 (8,82%)	34 (72,34%)
		42 (89,36%)	**5 (10,64%)**	**47 (0,12%)**
Mestrado Profissional	Feminino	22 (88,00%)	3 (12,00%)	25 (49,02%)
	Masculino	18 (69,23%)	8 (30,77%)	26 (50,98%)
		40 (78,43%)	**11 (21,57%)**	**51 (0,13%)**
Qualificação Profissional (FIC)	Feminino	2.497 (45,33%)	3.012 (54,67%)	5.509 (58,20%)
	Masculino	1.692 (42,77%)	2.264 (57,23%)	3.956 (41,80%)
		4.189 (44,26%)	**5.276 (55,74%)**	**9.465 (24,10%)**
Tecnologia	Feminino	580 (22,43%)	2.006 (77,57%)	2.586 (32,17%)
	Masculino	991 (18,18%)	4.461 (81,82%)	5.452 (67,83%)
		1.571 (19,54%)	**6.467 (80,46%)**	**8.038 (20,46%)**
Técnico	Feminino	3.702 (53,86%)	3.171 (46,14%)	6.873 (46,76%)
	Masculino	4.167 (53,24%)	3.660 (46,76%)	7.827 (53,24%)
		7.869 (53,53%)	**6.831 (46,47%)**	**14.700 (37,43%)**
Todos os Tipos de Curso	Feminino	7.766 (42,45%)	10.530 (57,55%)	18.296 (46,58%)
	Masculino	7.690 (36,65%)	13.291 (63,35%)	20.981 (53,42%)
		15.456 (39,35%)	**23.821 (60,65%)**	**39277**

Fonte: Dados do Autor.

Ao analisar os grupos de tipo de curso do IFPB em função do sexo dos alunos, observam-se algumas diferenças significativas nas taxas de conclusão. No grupo de Bacharelado, embora a maioria dos alunos seja do sexo masculino (63,07%), a taxa de conclusão para o sexo feminino

(26,64%) é ligeiramente superior à taxa para o sexo masculino (20,60%). No Mestrado Profissional, o grupo apresenta uma divisão quase igual entre os sexos feminino (49,02%) e masculino (50,98%). Nesse caso, as alunas apresentam uma taxa de conclusão mais elevada (88,00%) em comparação aos alunos do sexo masculino (69,23%).

No grupo de Tecnologia, a maioria dos alunos é do sexo masculino (67,83%), porém tanto os alunos do sexo feminino (22,43%), quanto os do sexo masculino (18,18%) apresentam taxas de conclusão relativamente baixas. Quando considerados todos os cursos em conjunto, constata-se que as alunas representam 46,58% do total de alunos, enquanto os alunos do sexo masculino correspondem a 53,42%. Nesse cenário, as alunas apresentam uma taxa de conclusão superior (42,5%) em relação aos alunos do sexo masculino (36,65%).

O grupo Especialização (*Lato Sensu*) tem uma maior proporção de alunas (64,16%), e as alunas também apresentam uma taxa de conclusão mais elevada (75,20%), em comparação aos alunos do sexo masculino (35,84%), que possuem uma taxa de conclusão de 64,95%. O grupo Licenciatura possui uma distribuição quase equilibrada entre alunas (50,76%) e alunos do sexo masculino (49,24%). No entanto, tanto as alunas (16,65%), quanto os alunos do sexo masculino (15,05%) apresentam baixas taxas de conclusão. O grupo Mestrado possui uma maioria de alunos do sexo masculino (72,34%), e as alunas apresentam uma taxa de conclusão menor (84,62%), em comparação aos alunos do sexo masculino (91,18%).

O grupo Qualificação Profissional (FIC) tem uma maior proporção de alunas (58,20%), e os alunos do sexo masculino (42,77%) possuem uma taxa de conclusão um pouco inferior às alunas (45,33%). O grupo Técnico possui uma distribuição quase equilibrada entre alunas (46.76%) e alunos do sexo masculino (53,24%). Ambos os grupos apresentam taxas de conclusão semelhantes, em torno de 53%

5.1.3 Análise dos Dados da PNP em Relação à Cor/Raça

A análise descritiva dos grupos de tipo de curso do IFPB, com base nos dados da PNP, revelou informações importantes sobre as taxas de aprovação em cada categoria, agrupadas em função da cor/raça declarada pelos os alunos. Os resultados estão apresentados na Tabela 6. Ao analisar os grupos em função da Cor/Raça nos diferentes tipos de curso do IFPB, observam-se as diferenças nas taxas de conclusão entre os grupos raciais em cada tipo de curso.

Tabela 6 – Análise da PNP para Tipo de Curso em Relação à Cor/Raça.

Tipo de Curso	Cor/Raça	Concluintes	Evadidos	Total
Bacharelado	Amarela	8 (21,62%)	29 (78,38%)	37 (1,42%)
	Branca	230 (23,78%)	737 (76,22%)	967 (37,16%)
	Indígena	0 (0,00%)	3 (100,00%)	3 (0,12%)
	Parda	232 (22,12%)	817 (77,88%)	1.049 (40,32%)
	Preta	26 (16,56%)	131 (83,44%)	157 (6,03%)

Tipo de Curso	Cor/Raça	Concluintes	Evadidos	Total
	Não Declarada	98 (25,19%)	291 (74,81%)	389 (14,95%)
		594 (22,83%)	**2.008 (77,17%)**	**2.602 (6,62%)**
Especialização (Lato Sensu)	Amarela	9 (69,23%)	4 (30,77%)	13 (1,60%)
	Branca	213 (72,45%)	81 (27,55%)	294 (36,21%)
	Indígena	2 (100,00%)	0 (0,00%)	2 (0,25%)
	Parda	253 (70,47%)	106 (29,53%)	359 (44,21%)
	Preta	37 (77,08%)	11 (22,92%)	48 (5,91%)
	Não Declarada	72 (75,00%)	24 (25,00%)	96 (11,82%)
		586 (72,17%)	**226 (27,83%)**	**812 (2,07%)**
Licenciatura	Amarela	7 (15,56%)	38 (84,44%)	45 (1,26%)
	Branca	172 (16,12%)	895 (83,88%)	1.067 (29,96%)
	Indígena	0 (0,00%)	2 (100,00%)	2 (0,06%)
	Parda	235 (14,11%)	1.431 (85,89%)	1.666 (46,77%)
	Preta	24 (8,51%)	258 (91,49%)	282 (7,92%)
	Não Declarada	127 (25,40%)	373 (74,60%)	500 (14,04%)
		565 (15,86%)	**2.997 (84,14%)**	**3.562 (9,07%)**
Mestrado	Branca	21 (91,30%)	2 (8,70%)	23 (48,94%)
	Parda	6 (85,71%)	1 (14,29%)	7 (14,89%)
	Preta	1 (100,00%)	0 (0,00%)	1 (2,13%)
	Não Declarada	14 (87,50%)	2 (12,50%)	16 (34,04%)
		42 (89,36%)	**5 (10,64%)**	**47 (0,12%)**
Mestrado Profissional	Amarela	1 (100,00%)	0 (0,00%)	1 (1,96%)
	Branca	14 (77,78%)	4 (22,22%)	18 (35,29%)
	Parda	12 (75,00%)	4 (25,00%)	16 (31,37%)
	Preta	1 (50,00%)	1 (50,00%)	2 (3,92%)
	Não Declarada	12 (85,71%)	2 (14,29%)	14 (27,45%)
		40 (78,43%)	**11 (21,57%)**	**51 (0,13%)**
Qualificação Profissional (FIC)	Amarela	74 (42,29%)	101 (57,71%)	175 (1,85%)
	Branca	1.061 (43,52%)	1.377 (56,48%)	2.438 (25,76%)
	Indígena	10 (35,71%)	18 (64,29%)	28 (0,30%)
	Parda	1.912 (43,75%)	2.458 (56,25%)	4.370 (46,17%)
	Preta	289 (36,49%)	503 (63,51%)	792 (8,37%)
	Não Declarada	843 (50,72%)	819 (49,28%)	1.662 (17,56%)
		4.189 (44,26%)	**5.276 (55,74%)**	**9.465 (24,10%)**
Tecnologia	Amarela	25 (21,55%)	91 (78,45%)	116 (1,44%)
	Branca	489 (18,12%)	2.209 (81,88%)	2.698 (33,57%)
	Indígena	3 (13,64%)	19 (86,36%)	22 (0,27%)
	Parda	681 (18,56%)	2.989 (81,44%)	3.670 (45,66%)

Tipo de Curso	Cor/Raça	Concluintes	Evadidos	Total
	Preta	105 (17,36%)	500 (82,64%)	605 (7,53%)
	Não Declarada	268 (28,91%)	659 (71,09%)	927 (11,53%)
		1.571 (19,54%)	**6.467 (80,46%)**	**8.038 (20,46%)**
Técnico	Amarela	90 (50,56%)	88 (49,44%)	178 (1,21%)
	Branca	2.492 (57,39%)	1.850 (42,61%)	4.342 (29,54%)
	Indígena	31 (62,00%)	19 (38,00%)	50 (0,34%)
	Parda	3.849 (51,33%)	3.649 (48,67%)	7.498 (51,01%)
	Preta	530 (47,92%)	576 (52,08%)	1.106 (7,52%)
	Não Declarada	877 (57,47%)	649 (42,53%)	1.526 (10,38%)
		7.869 (53,53%)	**6.831 (46,47%)**	**14.700 (37,43%)**
Todos os Tipos de Curso	Amarela	214 (37,88%)	351 (62,12%)	565 (1,44%)
	Branca	4.692 (39,60%)	7.155 (60,40%)	11.847 (30,16%)
	Indígena	46 (42,99%)	61 (57,01%)	107 (0,27%)
	Parda	7.180 (38,53%)	11.455 (61,47%)	18.635 (47,45%)
	Preta	1.013 (33,85%)	1.980 (66,15%)	2.993 (7,62%)
	Não Declarada	2.311 (45,05%)	2.819 (54,95%)	5.130 (13,06%)
		15.456 (39,35%)	**23.821 (60,65%)**	**39.277**

Fonte: Dados do Autor.

No Grupo Bacharelado, os alunos da cor/raça Branca têm a maior taxa de conclusão, alcançando 23,78%, e representam a maior proporção do grupo (37,16%). Os alunos das cor/raça Parda e Amarela também apresentam taxas de conclusão relativamente altas, com 22,12% e 21,62%, respectivamente. Por outro lado, os alunos das cor/raça Preta e Indígena têm as taxas de conclusão mais baixas, com 16,56% e 0,00%, respectivamente. Além disso, eles representam uma proporção menor do grupo, com 6,03% e 0,12% respectivamente. Os alunos que não declararam sua cor/raça têm uma taxa de conclusão de 25,19%, superando a média geral do grupo. No Grupo Especialização (*Lato Sensu*), novamente, os alunos da cor/raça Branca têm a maior taxa de conclusão, atingindo 72,45% e representam a maior proporção do grupo (36,21%). Os alunos das cor/raça Parda e Amarela também apresentam taxas de conclusão relativamente altas, com 70,47% e 69,23%, respectivamente. Os alunos da cor/raça Preta têm uma taxa de conclusão de 77,08%, superando a média geral do grupo. Os alunos indígenas e os que não declararam sua cor/raça também concluíram seus cursos, com taxas de 100% e 75,00% respectivamente.

Esses padrões se repetem em outros grupos, como Licenciatura, Mestrado, Mestrado Profissional, Qualificação Profissional (FIC), Tecnologia e Técnico, onde as taxas de conclusão e as proporções de cada cor/raça variam de acordo com o grupo. No geral, a cor/raça Branca tende a ter taxas de conclusão mais altas, seguida pela Parda e Amarela, enquanto as cor/raça Preta e Indígena apresentam taxas de conclusão mais baixas.

Ao analisar todos os tipos de curso é possível observar que os alunos da cor/raça Branca têm a maior taxa de conclusão, com 39,60%, seguidos pelos alunos Pardos, com 38,53%. Os alunos das cor/raça Preta têm a taxa de conclusão mais baixa, com 33,85%, enquanto os alunos Indígenas têm uma taxa de 42,99%. Os alunos que não declararam sua cor/raça têm uma taxa de conclusão de 45,05%.

5.1.4 Análise dos Dados da PNP em Relação à Faixa Etária

Através da análise descritiva dos grupos de tipo de curso do IFPB, com base nos dados obtidos da PNP, foram obtidas informações importantes acerca das taxas de aprovação em cada categoria, organizadas de acordo com o campo faixa etária dos discentes. Os resultados estão disponíveis na Tabela 7.

Tabela 7 – Análise da PNP para Tipo de Curso em Relação à Faixa Etária.

Tipo de Curso	Faixa Etária	Concluintes	Evadidos	Total
Bacharelado	15 a 19 anos	0 (0,00%)	127 (100,00%)	127 (4,88%)
	20 a 24 anos	194 (32,55%)	402 (67,45%)	596 (22,91%)
	25 a 29 anos	264 (32,04%)	560 (67,96%)	824 (31,67%)
	30 a 34 anos	69 (13,64%)	437 (86,36%)	506 (19,45%)
	35 a 39 anos	27 (10,19%)	238 (89,81%)	265 (10,18%)
	40 a 44 anos	14 (10,37%)	121 (89,63%)	135 (5,19%)
	45 a 49 anos	15 (17,65%)	70 (82,35%)	85 (3,27%)
	50 a 54 anos	7 (18,42%)	31 (81,58%)	38 (1,46%)
	55 a 59 anos	4 (19,05%)	17 (80,95%)	21 (0,81%)
	> 60 anos	0 (0,00%)	5 (100,00%)	5 (0,19%)
		594 (22,83%)	**2.008 (77,17%)**	**2.602 (6,62%)**
Especialização (Lato Sensu)	< 14 anos	1 (100,00%)	0 (0,00%)	1 (0,12%)
	20 a 24 anos	8 (47,06%)	9 (52,94%)	17 (2,09%)
	25 a 29 anos	173 (73,00%)	64 (27,00%)	237 (29,19%)
	30 a 34 anos	165 (75,00%)	55 (25,00%)	220 (27,09%)
	35 a 39 anos	112 (71,34%)	45 (28,66%)	157 (19,33%)
	40 a 44 anos	60 (68,97%)	27 (31,03%)	87 (10,71%)
	45 a 49 anos	30 (71,43%)	12 (28,57%)	42 (5,17%)
	50 a 54 anos	22 (66,67%)	11 (33,33%)	33 (4,06%)
	55 a 59 anos	12 (80,00%)	3 (20,00%)	15 (1,85%)
	> 60 anos	3 (100,00%)	0 (0,00%)	3 (0,37%)
		586 (72,17%)	**226 (27,83%)**	**812 (2,07%)**
Licenciatura	15 a 19 anos	0 (0,00%)	104 (100,00%)	104 (2,92%)
	20 a 24 anos	122 (18,91%)	523 (81,09%)	645 (18,11%)
	25 a 29 anos	137 (18,22%)	615 (81,78%)	752 (21,11%)

Tipo de Curso	Faixa Etária	Concluintes	Evadidos	Total
	30 a 34 anos	101 (14,77%)	583 (85,23%)	684 (19,20%)
	35 a 39 anos	91 (15,85%)	483 (84,15%)	574 (16,11%)
	40 a 44 anos	61 (15,64%)	329 (84,36%)	390 (10,95%)
	45 a 49 anos	28 (13,86%)	174 (86,14%)	202 (5,67%)
	50 a 54 anos	15 (12,50%)	105 (87,50%)	120 (3,37%)
	55 a 59 anos	6 (10,00%)	54 (90,00%)	60 (1,68%)
	> 60 anos	4 (12,90%)	27 (87,10%)	31 (0,87%)
		565 (15,86%)	**2.997 (84,14%)**	**3.562 (9,07%)**
Mestrado	20 a 24 anos	2 (66,67%)	1 (33,33%)	3 (6,38%)
	25 a 29 anos	23 (100,00%)	0 (0,00%)	23 (48,94%)
	30 a 34 anos	8 (80,00%)	2 (20,00%)	10 (21,28%)
	35 a 39 anos	3 (75,00%)	1 (25,00%)	4 (8,51%)
	40 a 44 anos	2 (66,67%)	1 (33,33%)	3 (6,38%)
	45 a 49 anos	3 (100,00%)	0 (0,00%)	3 (6,38%)
	55 a 59 anos	1 (100,00%)	0 (0,00%)	1 (2,13%)
		42 (89,36%)	**5 (10,64%)**	**47 (0,12%)**
Mestrado Profissional	20 a 24 anos	0 (0,00%)	2 (100,00%)	2 (3,92%)
	25 a 29 anos	3 (75,00%)	1 (25,00%)	4 (7,84%)
	30 a 34 anos	14 (77,78%)	4 (22,22%)	18 (35,29%)
	35 a 39 anos	9 (90,00%)	1 (10,00%)	10 (19,61%)
	40 a 44 anos	3 (100,00%)	0 (0,00%)	3 (5,88%)
	45 a 49 anos	7 (77,78%)	2 (22,22%)	9 (17,65%)
	50 a 54 anos	3 (100,00%)	0 (0,00%)	3 (5,88%)
	55 a 59 anos	0 (0,00%)	1 (100,00%)	1 (1,96%)
	> 60 anos	1 (100,00%)	0 (0,00%)	1 (1,96%)
		40 (78,43%)	**11 (21,57%)**	**51 (0,13%)**
Qualificação Profissional (FIC)	< 14 anos	132 (86,84%)	20 (13,16%)	152 (1,61%)
	15 a 19 anos	703 (49,47%)	718 (50,53%)	1.421 (15,01%)
	20 a 24 anos	679 (38,19%)	1.099 (61,81%)	1.778 (18,78%)
	25 a 29 anos	674 (38,49%)	1.077 (61,51%)	1.751 (18,50%)
	30 a 34 anos	610 (42,13%)	838 (57,87%)	1.448 (15,30%)
	35 a 39 anos	537 (47,73%)	588 (52,27%)	1.125 (11,89%)
	40 a 44 anos	352 (46,56%)	404 (53,44%)	756 (7,99%)
	45 a 49 anos	227 (48,09%)	245 (51,91%)	472 (4,99%)
	50 a 54 anos	151 (50,84%)	146 (49,16%)	297 (3,14%)
	55 a 59 anos	74 (44,31%)	93 (55,69%)	167 (1,76%)
	> 60 anos	50 (51,02%)	48 (48,98%)	98 (1,04%)
		4.189 (44,26%)	**5.276 (55,74%)**	**9.465 (24,10%)**

Tipo de Curso	Faixa Etária	Concluintes	Evadidos	Total
Tecnologia	< 14 anos	0 (0,00%)	2 (100,00%)	2 (0,02%)
	15 a 19 anos	2 (0,82%)	242 (99,18%)	244 (3,04%)
	20 a 24 anos	531 (26,89%)	1.444 (73,11%)	1.975 (24,57%)
	25 a 29 anos	479 (23,13%)	1.592 (76,87%)	2.071 (25,77%)
	30 a 34 anos	262 (16,22%)	1.353 (83,78%)	1.615 (20,09%)
	35 a 39 anos	127 (12,60%)	881 (87,40%)	1.008 (12,54%)
	40 a 44 anos	86 (15,30%)	476 (84,70%)	562 (6,99%)
	45 a 49 anos	35 (13,11%)	232 (86,89%)	267 (3,32%)
	50 a 54 anos	24 (14,72%)	139 (85,28%)	163 (2,03%)
	55 a 59 anos	21 (24,71%)	64 (75,29%)	85 (1,06%)
	> 60 anos	4 (8,70%)	42 (91,30%)	46 (0,57%)
		1.571 (19,54%)	**6.467 (80,46%)**	**8.038 (20,46%)**
Técnico	< 14 anos	2 (22,22%)	7 (77,78%)	9 (0,06%)
	15 a 19 anos	3.917 (62,81%)	2.319 (37,19%)	6.236 (42,42%)
	20 a 24 anos	2.498 (54,43%)	2.091 (45,57%)	4.589 (31,22%)
	25 a 29 anos	542 (36,80%)	931 (63,20%)	1.473 (10,02%)
	30 a 34 anos	324 (37,24%)	546 (62,76%)	870 (5,92%)
	35 a 39 anos	228 (35,35%)	417 (64,65%)	645 (4,39%)
	40 a 44 anos	140 (35,62%)	253 (64,38%)	393 (2,67%)
	45 a 49 anos	88 (38,94%)	138 (61,06%)	226 (1,54%)
	50 a 54 anos	73 (50,34%)	72 (49,66%)	145 (0,99%)
	55 a 59 anos	36 (46,75%)	41 (53,25%)	77 (0,52%)
	> 60 anos	21 (56,76%)	16 (43,24%)	37 (0,25%)
		7.869 (53,53%)	**6.831 (46,47%)**	**14.700 (37,43%)**
Todos os Tipos de Curso	< 14 anos	135 (82,32%)	29 (17,68%)	164 (0,42%)
	15 a 19 anos	4.622 (56,84%)	3.510 (43,16%)	8.132 (20,70%)
	20 a 24 anos	4.034 (42,00%)	5.571 (58,00%)	9.605 (24,45%)
	25 a 29 anos	2.295 (32,17%)	4.840 (67,83%)	7.135 (18,17%)
	30 a 34 anos	1.553 (28,91%)	3.818 (71,09%)	5.371 (13,67%)
	35 a 39 anos	1.134 (29,94%)	2.654 (70,06%)	3.788 (9,64%)
	40 a 44 anos	718 (30,83%)	1.611 (69,17%)	2.329 (5,93%)
	45 a 49 anos	433 (33,15%)	873 (66,85%)	1.306 (3,33%)
	50 a 54 anos	295 (36,92%)	504 (63,08%)	799 (2,03%)
	55 a 59 anos	154 (36,07%)	273 (63,93%)	427 (1,09%)
	> 60 anos	83 (37,56%)	138 (62,44%)	221 (0,56%)
		15.456 (39,35%)	**23.821 (60,65%)**	**39.277**

Fonte: Dados do Autor.

No caso dos cursos de Bacharelado, observa-se que a faixa etária de 20 a 24 anos apresenta alta taxa de evasão, com 67,45% dos estudantes abandonando os cursos. Essa faixa etária também tem a maior quantidade de evadidos em números absolutos, com 402 alunos. Por outro lado, as faixas etárias dos extremos de 15 a 19 anos e Maior de 60 anos registram 100% de evasão, indicando que nenhum estudante concluiu o curso nessas faixas etárias. Nos cursos de Especialização (*Lato Sensu*), a faixa etária de 25 a 29 anos apresenta a maior evasão absoluta, com 64 estudantes abandonando os cursos. No entanto, a faixa etária de 20 a 24 anos registra a maior taxa de evasão, com 64%.

Na categoria de Licenciatura, identifica-se que a faixa etária de 20 a 24 anos tem a maior evasão absoluta, com 523 estudantes evadidos dos cursos. Porém essa faixa etária registra um dos menores percentuais de evasão, com 81,09% alunos, mesmo sendo um valor alto para taxa de evasão escolar. Os estudantes com 15 a 19 anos apresentam a maior taxa de evasão, com 100,00%. No caso dos cursos de Mestrado, as faixas etárias de 25 a 29 anos, 45 a 49 anos e 55 a 59 anos registram uma taxa de evasão de 0%, indicando que nenhum estudante abandonou os cursos nessas faixas etárias. Já as faixas etárias de 20 a 24 anos e 40 a 44 anos tem a maior taxa de evasão, com 33,33% dos estudantes abandonando os cursos.

Nos cursos de Mestrado Profissional, a faixa etária de 30 a 34 anos apresenta a maior taxa de evasão, com 4 estudantes abandonando os cursos. No entanto, as faixas etárias 20 a 24 anos e 55 a 59 anos tem a maior taxa de evasão, com 100,00% de abandono dos cursos. Porém este grupo apresenta uma proporção pequena, em relação ao conjunto de dados, com apenas 0,13%, fazendo com que estes resultados possam não ser confiáveis, para serem usados nas tomadas de decisões.

Para a categoria de Qualificação Profissional (FIC), os estudantes menores de 14 anos têm a menor taxa de evasão, com apenas 13,16% abandonando os cursos. Já a faixa etária de 20 a 24 anos registra a maior evasão absoluta, com 1.099 estudantes deixando os cursos. Nos cursos de Tecnologia, constata-se que a faixa etária de 25 a 29 anos apresenta a maior evasão absoluta, com 1.592 dos estudantes abandonando os cursos. No entanto, essa faixa etária apresenta uma das menores taxas de evasão relativas, com 73,11%.

Nos cursos técnicos, constata-se que a faixa etária de 15 a 19 anos apresenta a maior evasão, com 2.319 estudantes abandonando os cursos. Porém, essa faixa etária também registra a menor taxa de evasão com 37,19%. Por outro lado, os estudantes menores de 14 anos têm a maior taxa de evasão, com 77,78% e apenas 7 evasões absolutas.

5.1.5 Análise dos Dados da PNP em Relação à Renda Familiar

A análise descritiva dos grupos de tipo de curso do IFPB, com base nos dados da PNP, revelou informações importantes sobre as taxas de aprovação em cada categoria, agrupadas em função da renda familiar per capita (RFP) dos alunos. Os resultados estão apresentados na Tabela 8. A RFP é comparada em relação a quantidade de salários mínimos por membro familiar.

Tabela 8 – Análise da PNP para Tipo de Curso em Relação à Renda Familiar.

Tipo de Curso	Renda Familiar	Concluintes	Evadidos	Total
Bacharelado	0<RFP<=0,5	214 (34,68%)	403 (65,32%)	617 (23,71%)
	0,5<RFP<=1,0	124 (30,92%)	277 (69,08%)	401 (15,41%)
	1,0<RFP<=1,5	74 (39,36%)	114 (60,64%)	188 (7,23%)
	1,5<RFP<=2,5	37 (33,04%)	75 (66,96%)	112 (4,30%)
	2,5<RFP<=3,5	11 (26,83%)	30 (73,17%)	41 (1,58%)
	RFP>3,5	9 (20,45%)	35 (79,55%)	44 (1,69%)
	Não Declarada	125 (10,43%)	1.074 (89,57%)	1.199 (46,08%)
		594 (22,83%)	**2.008 (77,17%)**	**2.602 (6,62%)**
Especialização (Lato Sensu)	0<RFP<=0,5	137 (74,86%)	46 (25,14%)	183 (22,54%)
	0,5<RFP<=1,0	165 (80,49%)	40 (19,51%)	205 (25,25%)
	1,0<RFP<=1,5	50 (68,49%)	23 (31,51%)	73 (8,99%)
	1,5<RFP<=2,5	42 (75,00%)	14 (25,00%)	56 (6,90%)
	2,5<RFP<=3,5	15 (78,95%)	4 (21,05%)	19 (2,34%)
	RFP>3,5	17 (73,91%)	6 (26,09%)	23 (2,83%)
	Não Declarada	160 (63,24%)	93 (36,76%)	253 (31,16%)
		586 (72,17%)	**226 (27,83%)**	**812 (2,07%)**
Licenciatura	0<RFP<=0,5	247 (22,54%)	849 (77,46%)	1.096 (30,77%)
	0,5<RFP<=1,0	99 (20,33%)	388 (79,67%)	487 (13,67%)
	1,0<RFP<=1,5	36 (20,81%)	137 (79,19%)	173 (4,86%)
	1,5<RFP<=2,5	18 (18,37%)	80 (81,63%)	98 (2,75%)
	2,5<RFP<=3,5	7 (17,07%)	34 (82,93%)	41 (1,15%)
	RFP>3,5	5 (12,20%)	36 (87,80%)	41 (1,15%)
	Não Declarada	153 (9,41%)	1.473 (90,59%)	1.626 (45,65%)
		565 (15,86%)	**2.997 (84,14%)**	**3.562 (9,07%)**
Mestrado	0<RFP<=0,5	10 (90,91%)	1 (9,09%)	11 (23,40%)
	0,5<RFP<=1,0	7 (87,50%)	1 (12,50%)	8 (17,02%)
	1,0<RFP<=1,5	1 (100,00%)	0 (0,00%)	1 (2,13%)
	1,5<RFP<=2,5	4 (66,67%)	2 (33,33%)	6 (12,77%)
	2,5<RFP<=3,5	1 (100,00%)	0 (0,00%)	1 (2,13%)
	Não Declarada	19 (95,00%)	1 (5,00%)	20 (42,55%)
		42 (89,36%)	**5 (10,64%)**	**47 (0,12%)**
Mestrado Profissional	0<RFP<=0,5	2 (50,00%)	2 (50,00%)	4 (7,84%)
	0,5<RFP<=1,0	3 (100,00%)	0 (0,00%)	3 (5,88%)
	1,0<RFP<=1,5	3 (75,00%)	1 (25,00%)	4 (7,84%)
	1,5<RFP<=2,5	4 (50,00%)	4 (50,00%)	8 (15,69%)
	2,5<RFP<=3,5	5 (100,00%)	0 (0,00%)	5 (9,80%)
	RFP>3,5	4 (57,14%)	3 (42,86%)	7 (13,73%)

Tipo de Curso	Renda Familiar	Concluintes	Evadidos	Total
	Não Declarada	19 (95,00%)	1 (5,00%)	20 (39,22%)
		40 (78,43%)	**11 (21,57%)**	**51 (0,13%)**
Qualificação Profissional (FIC)	0<RFP<=0,5	851 (48,57%)	901 (51,43%)	1.752 (18,51%)
	0,5<RFP<=1,0	365 (47,90%)	397 (52,10%)	762 (8,05%)
	1,0<RFP<=1,5	179 (49,72%)	181 (50,28%)	360 (3,80%)
	1,5<RFP<=2,5	137 (58,55%)	97 (41,45%)	234 (2,47%)
	2,5<RFP<=3,5	30 (51,72%)	28 (48,28%)	58 (0,61%)
	RFP>3,5	87 (57,24%)	65 (42,76%)	152 (1,61%)
	Não Declarada	2.540 (41,32%)	3.607 (58,68%)	6.147 (64,94%)
		4.189 (44,26%)	**5.276 (55,74%)**	**9.465 (24,10%)**
Tecnologia	0<RFP<=0,5	614 (28,36%)	1.551 (71,64%)	2.165 (26,93%)
	0,5<RFP<=1,0	308 (27,65%)	806 (72,35%)	1.114 (13,86%)
	1,0<RFP<=1,5	127 (30,09%)	295 (69,91%)	422 (5,25%)
	1,5<RFP<=2,5	83 (33,88%)	162 (66,12%)	245 (3,05%)
	2,5<RFP<=3,5	38 (41,30%)	54 (58,70%)	92 (1,14%)
	RFP>3,5	30 (36,14%)	53 (63,86%)	83 (1,03%)
	Não Declarada	371 (9,47%)	3.546 (90,53%)	3.917 (48,73%)
		1.571 (19,54%)	**6.467 (80,46%)**	**8.038 (20,46%)**
Técnico	0<RFP<=0,5	4.650 (60,15%)	3.081 (39,85%)	7.731 (52,59%)
	0,5<RFP<=1,0	1.089 (63,91%)	615 (36,09%)	1.704 (11,59%)
	1,0<RFP<=1,5	323 (65,92%)	167 (34,08%)	490 (3,33%)
	1,5<RFP<=2,5	118 (58,71%)	83 (41,29%)	201 (1,37%)
	2,5<RFP<=3,5	39 (68,42%)	18 (31,58%)	57 (0,39%)
	RFP>3,5	68 (65,38%)	36 (34,62%)	104 (0,71%)
	Não Declarada	1.582 (35,85%)	2.831 (64,15%)	4.413 (30,02%)
		7.869 (53,53%)	**6.831 (46,47%)**	**14.700 (37,43%)**
Todos os Tipos de Curso	0<RFP<=0,5	6.725 (49,60%)	6.834 (50,40%)	13.559 (34,52%)
	0,5<RFP<=1,0	2.160 (46,11%)	2.524 (53,89%)	4.684 (11,93%)
	1,0<RFP<=1,5	793 (46,35%)	918 (53,65%)	1.711 (4,36%)
	1,5<RFP<=2,5	443 (46,15%)	517 (53,85%)	960 (2,44%)
	2,5<RFP<=3,5	146 (46,50%)	168 (53,50%)	314 (0,80%)
	RFP>3,5	220 (48,46%)	234 (51,54%)	454 (1,16%)
	Não Declarada	4.969 (28,24%)	12.626 (71,76%)	17.595 (44,80%)
		15.456 (39,35%)	**23.821 (60,65%)**	**39.277**

Fonte: Dados do Autor.

No grupo de cursos de Bacharelado, observa-se que a taxas de evasão e altas em todos os níveis de renda. No grupo com faixa de renda “0<RFP<=0,5”, a taxa de evasão é de 65,32%,

o que indica que a maioria dos estudantes nessa faixa de renda não concluíram os cursos. Essa taxa se mantém elevada nos grupos com faixas de renda “0,5<RFP<=1,0” (69,08%), “1,0<RFP<=1,5” (60,64%), “1,5<RFP<=2,5” (66,96%), “2,5<RFP<=3,5” (73,17%) e “RFP>3,5” (79,55%). Isso mostra que os estudantes de todos os níveis de renda nesses grupos enfrentam desafios significativos para concluir o bacharelado.

Analisando os cursos do tipo Especialização (*Lato Sensu*), observa-se taxas de evasão relativamente mais baixas (27,83%) em comparação aos demais grupos, indicando uma maior taxa de conclusão desses cursos. No grupo com faixa de renda “0<RFP<=0,5”, a taxa de evasão é de 25,14%, o que sugere que uma parcela significativa dos estudantes nessa faixa de renda não concluiu a especialização. Essa taxa diminui ainda mais nos grupos com faixas de renda “0,5<RFP<=1,0” (19,51%), “1,0<RFP<=1,5” (31,51%), “1,5<RFP<=2,5” (25,00%), “2,5<RFP<=3,5” (21,05%) e “RFP>3,5” (26,09%). Esses números apontam para uma tendência de maior sucesso na conclusão dos cursos de especialização, independentemente do nível de renda.

No grupo de cursos de “Licenciatura”, verifica-se a maior taxa de evasão (84,14%), indicando um desafio significativo na conclusão desses cursos. Para o grupo com faixa de renda “0<RFP<=0,5”, a taxa de evasão é de 77,46%. Esse número é bastante alto e sugere que a maioria dos estudantes nessa faixa de renda não consegue concluir a licenciatura. Essa tendência se mantém nos grupos com faixas de renda “0,5<RFP<=1,0” (79,67%), “1,0<RFP<=1,5” (79,19%), “1,5<RFP<=2,5” (81,63%), “2,5<RFP<=3,5” (82,93%) e “RFP>3,5” (87,80%). Esses dados revelam uma dificuldade generalizada na conclusão dos cursos de licenciatura, independentemente do nível de renda.

No grupo de cursos de “Mestrado”, os dados revelam uma baixa taxa de evasão escolar no geral, porém por ser um grupo pequeno, com apenas 47 registros, os valores podem não representar uma tendência real. No caso dos cursos de “Mestrado Profissional”, verifica-se que as maiores taxas de evasão (50,00%) estão concentradas nos grupos com faixas de renda “0<RFP<=0,5” e “1,5<RFP<=2,5”. Essas taxas de evasão são bastante elevadas, o que pode indicar dificuldades enfrentadas pelos estudantes com menor renda em concluir esses cursos. Por outro lado, as maiores taxas de conclusão (100,00%) ocorrem nos grupos com faixas de renda “1,0<RFP<=1,5” e “2,5<RFP<=3,5”. Essa situação também pode ser atribuída à pequena quantidade de registros desses grupos na base de dados, representando apenas 0,13% do total.

Já para os cursos de “Qualificação Profissional (FIC)”, observa-se que as maiores taxas de evasão ocorrem nos grupos com menor renda. Os grupos com faixas de renda “0<RFP<=0,5” (51,43%), “0,5<RFP<=1,0” (52,10%) e “1,0<RFP<=1,5” (50,28%) apresentam as maiores taxas de evasão. Isso sugere que os estudantes de baixa renda podem enfrentar mais desafios para concluir esses cursos de qualificação profissional. Por outro lado, as menores taxas de evasão são observadas nos grupos com faixas de renda mais altas, como “1,5<RFP<=2,5” (41,45%), “2,5<RFP<=3,5” (48,28%) e “RFP>3,5” (42,76%). Isso pode indicar que os estudantes com

maior renda têm mais recursos disponíveis e, portanto, enfrentam menos dificuldades para concluir esses cursos.

5.1.6 Análise dos Dados da PNP em Relação ao Turno

Ao realizar a análise descritiva dos grupos de tipo de curso do IFPB, utilizando os dados fornecidos pela PNP, foram identificadas informações de destaque sobre as taxas de aprovação em cada categoria, agrupadas com base no campo Turno dos cursos dos alunos. Os resultados são expostos na Tabela 9.

Tabela 9 – Análise da PNP para Tipo de Curso em Relação ao Turno.

Tipo de Curso	Turno	Concluintes	Evadidos	Total
Bacharelado	Integral	549 (24,50%)	1.692 (75,50%)	2.241 (86,13%)
	Matutino	0 (0,00%)	14 (100,00%)	14 (0,54%)
	Noturno	0 (0,00%)	9 (100,00%)	9 (0,35%)
	Não se Aplica	45 (13,31%)	293 (86,69%)	338 (12,99%)
		594 (22,83%)	**2.008 (77,17%)**	**2.602 (6,62%)**
Especialização (Lato Sensu)	Integral	59 (56,73%)	45 (43,27%)	104 (12,81%)
	Noturno	91 (60,26%)	60 (39,74%)	151 (18,60%)
	Vespertino	10 (52,63%)	9 (47,37%)	19 (2,34%)
	Não se Aplica	426 (79,18%)	112 (20,82%)	538 (66,26%)
		586 (72,17%)	**226 (27,83%)**	**812 (2,07%)**
Licenciatura	Integral	112 (58,64%)	79 (41,36%)	191 (5,36%)
	Noturno	153 (23,94%)	486 (76,06%)	639 (17,94%)
	Vespertino	48 (9,50%)	457 (90,50%)	505 (14,18%)
	Não se Aplica	252 (11,32%)	1.975 (88,68%)	2.227 (62,52%)
		565 (15,86%)	**2.997 (84,14%)**	**3.562 (9,07%)**
Mestrado	Integral	42 (89,36%)	5 (10,64%)	47 (100,00%)
Mestrado Profissional	Integral	3 (50,00%)	3 (50,00%)	6 (11,76%)
	Noturno	3 (50,00%)	3 (50,00%)	6 (11,76%)
	Vespertino	6 (100,00%)	0 (0,00%)	6 (11,76%)
	Não se Aplica	28 (84,85%)	5 (15,15%)	33 (64,71%)
		40 (78,43%)	**11 (21,57%)**	**51 (0,13%)**
Qualificação Profissional (FIC)	Matutino	282 (70,68%)	117 (29,32%)	399 (4,22%)
	Noturno	877 (62,73%)	521 (37,27%)	1.398 (14,77%)
	Vespertino	562 (65,50%)	296 (34,50%)	858 (9,06%)
	Não se Aplica	2.468 (36,24%)	4.342 (63,76%)	6.810 (71,95%)
		4.189 (44,26%)	**5.276 (55,74%)**	**9.465 (24,10%)**
Tecnologia	Integral	259 (25,07%)	774 (74,93%)	1.033 (12,85%)
	Matutino	316 (12,90%)	2.133 (87,10%)	2.449 (30,47%)

Tipo de Curso	Turno	Concluintes	Evadidos	Total
	Noturno	729 (20,79%)	2.778 (79,21%)	3.507 (43,63%)
	Vespertino	267 (25,45%)	782 (74,55%)	1.049 (13,05%)
		1.571 (19,54%)	**6.467 (80,46%)**	**8.038 (20,46%)**
Técnico	Integral	4.307 (69,21%)	1.916 (30,79%)	6.223 (42,33%)
	Matutino	742 (62,99%)	436 (37,01%)	1.178 (8,01%)
	Noturno	1.735 (35,61%)	3.137 (64,39%)	4.872 (33,14%)
	Vespertino	914 (48,03%)	989 (51,97%)	1.903 (12,95%)
	Não se Aplica	171 (32,63%)	353 (67,37%)	524 (3,56%)
		7.869 (53,53%)	**6.831 (46,47%)**	**14.700 (37,43%)**
Todos os Tipos de Curso	Integral	5.331 (54,15%)	4.514 (45,85%)	9.845 (25,07%)
	Matutino	1.340 (33,17%)	2.700 (66,83%)	4.040 (10,29%)
	Noturno	3.588 (33,91%)	6.994 (66,09%)	10.582 (26,94%)
	Vespertino	1.807 (41,64%)	2.533 (58,36%)	4.340 (11,05%)
	Não se Aplica	3.390 (32,38%)	7.080 (67,62%)	10.470 (26,66%)
		15.456 (39,35%)	**23.821 (60,65%)**	**39.277**

Fonte: Dados do Autor.

No grupo de “Bacharelado”, os cursos apresentam altas taxas de evasão em todos os turnos. O turno integral registra uma taxa de evasão de 75,50%, enquanto os turnos matutino e noturno têm uma taxa de evasão de 100,00%. Essa taxa de evasão de 100,00% é afetada pela pequena quantidade de registros (23). Porém a taxa geral de 77,17% indica que a evasão é um desafio significativo nesse grupo, independentemente do turno das aulas. No grupo de “Especialização (*Lato Sensu*)”, as taxas de evasão variam de acordo com o turno. O turno integral apresenta uma taxa de evasão de 43,27%, o turno noturno registra 39,74% e o turno vespertino tem uma taxa de 47,37%. Esses resultados sugerem que o turno noturno é o que apresenta a menor taxa de evasão nesse grupo, enquanto o turno vespertino tem uma taxa ligeiramente mais alta.

No grupo de “Licenciatura”, as taxas de evasão são significativamente altas em todos os turnos. O turno integral registra uma taxa de evasão de 41,36%, o turno noturno tem uma taxa de 76,06% e o turno vespertino apresenta a maior taxa de evasão, atingindo 90,50%. Isso indica que a evasão é um desafio particularmente significativo nos cursos de licenciatura, com uma taxa ainda mais alta no turno vespertino. No grupo de “Mestrado”, há apenas o turno integral com a taxa de evasão de 10,64%, que representa a taxa de evasão do próprio grupo. No grupo de “Mestrado Profissional”, os dados mostram uma taxa de evasão de 50,00%, tanto no turno integral, quanto no turno noturno. No entanto, o turno vespertino registra uma taxa de evasão de 0,00%, o que indica que não houve evasão nesse turno. É importante notar que o número de registros geral é pequeno, o que pode afetar os resultados.

No grupo de "Qualificação Profissional (FIC)", as taxas de evasão variam entre os turnos. O turno matutino registra uma taxa de evasão de 29,32%, o turno noturno tem uma taxa de 37,27% e o turno vespertino apresenta uma taxa de 34,50%. Essas diferenças são relativamente pequenas e sugerem que a evasão nos cursos de qualificação profissional não varia significativamente, de acordo com o turno das aulas. No grupo de "Tecnologia", os cursos apresentam taxas de evasão relativamente altas em todos os turnos (80,46%). O turno integral registra uma taxa de evasão de 74,93%, o turno matutino tem uma taxa de 87,10%, o turno noturno registra 79,21% e o turno vespertino apresenta uma taxa de 74,55%. Isso indica que a evasão é um desafio comum nos cursos de tecnologia, independentemente do turno das aulas.

No grupo de "Técnico", as taxas de evasão também apresentam variações significativas, de acordo com o turno das aulas. O turno integral registra uma taxa de evasão de 30,79%, enquanto o turno matutino tem uma taxa de 37,01%. No entanto, as taxas de evasão aumentam consideravelmente nos turnos noturno e vespertino, com taxas de 64,39% e 51,97%, respectivamente. Isso sugere que a evasão é um desafio mais significativo nos cursos técnicos quando as aulas ocorrem à noite ou à tarde.

5.1.7 Análise dos Dados do SUAP em Relação ao Turno

Mediante a análise descritiva dos grupos de modalidade de curso do IFPB, a partir dos dados fornecidos pelo SUAP, foram obtidas informações cruciais acerca das taxas de aprovação em cada categoria, agrupadas segundo o campo turno dos cursos dos estudantes. Os resultados são apresentados na Tabela 10.

Tabela 10 – Análise do SUAP para Modalidade do Curso em Relação ao Turno.

Modalidade do Curso	Turno	Concluintes	Evadidos	Total
Bacharelado	Diurno	95 (23,51%)	309 (76,49%)	404 (100,00%)
		95 (23,51%)	**309 (76,49%)**	**404 (20,98%)**
Integrado	Diurno	58 (55,24%)	47 (44,76%)	105 (25,74%)
	Matutino	100 (62,11%)	61 (37,89%)	161 (39,46%)
	Vespertino	68 (47,89%)	74 (52,11%)	142 (34,80%)
		226 (55,39%)	**182 (44,61%)**	**408 (21,18%)**
Subsequente	Noturno	59 (16,95%)	289 (83,05%)	348 (100,00%)
		59 (16,95%)	**289 (83,05%)**	**348 (18,07%)**
Tecnologia	Diurno	76 (18,54%)	334 (81,46%)	410 (53,52%)
	Vespertino	60 (16,85%)	296 (83,15%)	356 (46,48%)
		136 (17,75%)	**630 (82,25%)**	**766 (39,77%)**
Todas as Modalidades de Curso	Diurno	229 (24,92%)	690 (75,08%)	919 (47,72%)
	Matutino	100 (62,11%)	61 (37,89%)	161 (8,36%)
	Noturno	59 (16,95%)	289 (83,05%)	348 (18,07%)

Modalidade do Curso	Turno	Concluintes	Evadidos	Total
	Vespertino	128 (25,70%)	370 (74,30%)	498 (25,86%)
		516 (26,79%)	**1.410 (73,21%)**	**1.926**

Fonte: Dados do Autor.

No grupo de cursos Bacharelado, que representa 20,98% de todos os registros, encontramos uma taxa de evasão de 76,49%, sendo diurno o único turno disponível. Isso indica que os cursos de bacharelado enfrentam um desafio significativo em relação à evasão, com uma proporção substancial de alunos abandonando seus estudos. No grupo de cursos Integrado, que representa 21,18% dos registros, a taxa geral de evasão é de 44,61%. As taxas de evasão variam de acordo com o turno das aulas, no turno diurno, a taxa de evasão é de 44,76%, enquanto no turno matutino é de 37,89% e no turno vespertino é de 52,11%. Esses números sugerem que os cursos integrados apresentam um padrão de evasão mais elevado no turno vespertino, seguido pelo turno diurno e pelo turno matutino.

Para a modalidade de curso Subsequente, que representa 18,07% dos registros, identifica-se uma taxa de evasão considerável de 83,05%, tendo apenas a modalidade de turno noturno. Isso indica que os cursos subsequentes enfrentam um desafio significativo em relação à evasão, com a maioria dos alunos abandonando seus estudos. No grupo de cursos Tecnologia, que representa 39,77% dos registros, a taxa geral de evasão é alta com 82,25%. Tanto o turno diurno, quanto o turno vespertino, apresentam altas taxas de evasão. No turno diurno, a taxa de evasão é de 81,46%, e no turno vespertino é de 83,15%. Isso sugere que os cursos de tecnologia enfrentam desafios consistentes de evasão.

Ao considerar todas as modalidades de curso, em conjunto, a taxa de evasão é de 73,21%. Quando analisado por turnos o diurno registra uma taxa de evasão de 75,08%, o turno matutino apresenta uma taxa de 37,89%, o turno noturno tem uma taxa de 83,05%, e o turno vespertino registra uma taxa de 74,30%. Isso indica que, com exceção ao turno Matutino, a evasão escolar é um desafio significativo em todas as modalidades de curso.

5.1.8 Análise dos Dados do SUAP em Relação à Cota SISTEC

Através da análise descritiva dos grupos de modalidade de curso do IFPB, a partir dos dados provenientes do SUAP, foram identificadas informações importantes sobre as taxas de aprovação em cada categoria, organizadas em função do campo cota SISTEC dos discentes. Os resultados são exibidos na Tabela 11.

Tabela 11 – Análise do SUAP para Modalidade do Curso em Relação à Cota SISTEC.

Modalidade do Curso	Cota SISTEC	Concluintes	Evadidos	Total
	Cor/Raça	4 (7,27%)	51 (92,73%)	55 (13,61%)

Bacharelado

Modalidade do Curso	Cota SISTEC	Concluintes	Evadidos	Total
	Escola Pública	6 (10,91%)	49 (89,09%)	55 (13,61%)
	Nec. Especiais	0 (0,00%)	18 (100,00%)	18 (4,46%)
	Não Se Aplica	85 (38,29%)	137 (61,71%)	222 (54,95%)
	S/I	0 (0,00%)	54 (100,00%)	54 (13,37%)
		95 (23,51%)	**309 (76,49%)**	**404 (20,98%)**
Integrado	Cor/Raça	22 (52,38%)	20 (47,62%)	42 (10,29%)
	Escola Pública	52 (46,43%)	60 (53,57%)	112 (27,45%)
	Neces. Especiais	1 (20,00%)	4 (80,00%)	5 (1,23%)
	Não Se Aplica	150 (65,50%)	79 (34,50%)	229 (56,13%)
	S/I	1 (5,00%)	19 (95,00%)	20 (4,90%)
		226 (55,39%)	**182 (44,61%)**	**408 (21,18%)**
Subsequente	Cor/Raça	9 (11,69%)	68 (88,31%)	77 (22,13%)
	Escola Pública	15 (24,59%)	46 (75,41%)	61 (17,53%)
	Nec. Especiais	1 (9,09%)	10 (90,91%)	11 (3,16%)
	Não Se Aplica	34 (18,38%)	151 (81,62%)	185 (53,16%)
	S/I	0 (0,00%)	14 (100,00%)	14 (4,02%)
		59 (16,95%)	**289 (83,05%)**	**348 (18,07%)**
Tecnologia	Cor/Raça	5 (5,00%)	95 (95,00%)	100 (13,05%)
	Escola Pública	14 (9,09%)	140 (90,91%)	154 (20,10%)
	Nec. Especiais	0 (0,00%)	18 (100,00%)	18 (2,35%)
	Não Se Aplica	117 (29,25%)	283 (70,75%)	400 (52,22%)
	S/I	0 (0,00%)	94 (100,00%)	94 (12,27%)
		136 (17,75%)	**630 (82,25%)**	**766 (39,77%)**
Todas as Modalidades de Curso	Cor/Raça	40 (14,60%)	234 (85,40%)	274 (14,23%)
	Escola Pública	87 (22,77%)	295 (77,23%)	382 (19,83%)
	Nec. Especiais	2 (3,85%)	50 (96,15%)	52 (2,70%)
	Não Se Aplica	386 (37,26%)	650 (62,74%)	1.036 (53,79%)
	S/I	1 (0,55%)	181 (99,45%)	182 (9,45%)
		516 (26,79%)	**1.410 (73,21%)**	**1.926**

Fonte: Dados do Autor.

No grupo de modalidade de cursos Bacharelado, observa-se que a taxa de evasão varia de acordo com o tipo de cota. Para a cota “Cor/Raça”, a taxa de evasão é de 92,73%. Para a cota “Escola Pública”, a taxa de evasão é de 89,09%. Já para a cota “Necessidades Especiais”, a taxa de evasão é de 100%. Por outro lado, para aqueles que não possuem cotas, a taxa de evasão é de 61,71%. Esses dados indicam que os alunos beneficiados pelas cotas, especialmente os que se enquadram na cota “Necessidades Especiais”, enfrentam maiores desafios em relação à

evasão, enquanto aqueles sem cotas apresentam uma taxa relativamente menor. No grupo de cursos “Integrado”, também observa-se variações na taxa de evasão de acordo com o tipo de cota. Para a cota “Cor/Raça”, a taxa de evasão é de 47,62%. Para a cota “Escola Pública”, a taxa de evasão é de 53,57%. Já para a cota “Necessidades Especiais”, a taxa de evasão é de 80%. Para aqueles que não possuem cotas, a taxa de evasão é de 34,50%.

Para o grupo de cursos “Subsequente”, novamente observa-se variações na taxa de evasão, de acordo com o tipo de cota. Para a cota “Cor/Raça”, a taxa de evasão é de 88,31%. Para a cota “Escola Pública”, a taxa de evasão é de 75,41%. Já para a cota “Necessidades Especiais”, a taxa de evasão é de 90,91%. Para aqueles que não possuem cotas, a taxa de evasão é de 81,62%. No grupo de cursos “Tecnologia”, a taxa de evasão também varia de acordo com o tipo de cota. Para a cota “Cor/Raça”, a taxa de evasão é de 95%. Para a cota “Escola Pública”, a taxa de evasão é de 90,91%. Já para a cota “Necessidades Especiais”, a taxa de evasão é de 100%. Para aqueles que não possuem cotas, a taxa de evasão é de 70,75%.

Considerando todas as modalidades de curso em conjunto, observa-se que as taxas de evasão também variam de acordo com o tipo de cota. A cota “Cor/Raça” apresenta uma taxa de evasão de 85,40%, enquanto a cota “Escola Pública” possui uma taxa de evasão de 77,23%. Para a cota “Necessidades Especiais”, a taxa de evasão é de 96,15%. Já para aqueles que não possuem cotas, a taxa de evasão é de 62,74%. Esses resultados indicam que, independentemente da modalidade do curso, as cotas têm um impacto significativo nas taxas de evasão, especialmente para os alunos com necessidades especiais.

5.1.9 Análise dos Dados do SUAP em Relação à Cota MEC

A análise dos grupos de modalidade de curso do IFPB, em função da cota MEC dos alunos, revela algumas tendências interessantes. A Tabela 12 a seguir Para melhor compreensão dos dados apresentados, é importante esclarecer o significado das siglas utilizadas. No contexto da análise da evasão escolar no IFPB, algumas siglas representam diferentes categorias e critérios de classificação dos estudantes. A sigla “EEP” refere-se a “Egresso de Escola Pública”, indicando que o aluno concluiu o ensino fundamental ou médio em instituições públicas. “RENDA” representa a condição de ter uma renda familiar inferior a 1,5 salários mínimos. A sigla “PPI” abrange a categoria de estudantes autodeclarados como “Pretos, Pardos ou Indígenas”. Por fim, “PCD” refere-se a “Pessoa com Deficiência”, abrangendo aqueles que possuem algum tipo de limitação física, sensorial ou intelectual. Para melhor compreensão dos dados apresentados, é importante esclarecer o significado das siglas utilizadas. No contexto da análise da evasão escolar no IFPB, algumas siglas representam diferentes categorias e critérios de classificação dos estudantes. A sigla “EEP” refere-se a “Egresso de Escola Pública”, indicando que o aluno concluiu o ensino fundamental ou médio em instituições públicas. “RENDA” representa a condição de ter uma renda familiar inferior a 1,5 salários mínimos. A sigla “PPI” abrange abrange a categoria de estudantes autodeclarados como “Pretos, Pardos ou Indígenas”. Por fim,

"PCD" refere-se a "Pessoa com Deficiência", abrangendo aqueles que possuem algum tipo de limitação física, sensorial ou intelectual.

Tabela 12 – Análise do SUAP para Modalidade do Curso em Relação à Cota MEC.

Modalidade do Curso	Cota MEC	Concluintes	Evadidos	Total
Bacharelado	EEP	2 (11,11%)	16 (88,89%)	18 (4,46%)
	EEP, RENDA	2 (8,00%)	23 (92,00%)	25 (6,19%)
	EEP, RENDA, PPI	1 (3,23%)	30 (96,77%)	31 (7,67%)
	EEP, RENDA, PPI, PCD	0 (0,00%)	3 (100,00%)	3 (0,74%)
	EEP, PCD	0 (0,00%)	3 (100,00%)	3 (0,74%)
	EEP, PPI	4 (14,29%)	24 (85,71%)	28 (6,93%)
	EEP, PPI, PCD	0 (0,00%)	5 (100,00%)	5 (1,24%)
	Não Se Aplica	86 (36,29%)	151 (63,71%)	237 (58,66%)
	S/I	0 (0,00%)	54 (100,00%)	54 (13,37%)
		95 (23,51%)	**309 (76,49%)**	**404 (20,98%)**
Integrado	EEP	14 (56,00%)	11 (44,00%)	25 (6,13%)
	EEP, RENDA	20 (60,61%)	13 (39,39%)	33 (8,09%)
	EEP, RENDA, PPI	23 (40,35%)	34 (59,65%)	57 (13,97%)
	EEP, PPI	18 (46,15%)	21 (53,85%)	39 (9,56%)
	Não Se Aplica	150 (64,10%)	84 (35,90%)	234 (57,35%)
	S/I	1 (5,00%)	19 (95,00%)	20 (4,90%)
		226 (55,39%)	**182 (44,61%)**	**408 (21,18%)**
Subsequente	EEP	0 (0,00%)	18 (100,00%)	18 (5,17%)
	EEP, RENDA	7 (29,17%)	17 (70,83%)	24 (6,90%)
	EEP, RENDA, PPI	8 (17,39%)	38 (82,61%)	46 (13,22%)
	EEP, RENDA, PPI, PCD	0 (0,00%)	2 (100,00%)	2 (0,57%)
	EEP, PPI	7 (14,89%)	40 (85,11%)	47 (13,51%)
	EEP, PPI, PCD	1 (50,00%)	1 (50,00%)	2 (0,57%)
	Não Se Aplica	36 (18,46%)	159 (81,54%)	195 (56,03%)
	S/I	0 (0,00%)	14 (100,00%)	14 (4,02%)
		59 (16,95%)	**289 (83,05%)**	**348 (18,07%)**
Tecnologia	EEP	5 (11,11%)	40 (88,89%)	45 (5,87%)
	EEP, RENDA	4 (8,89%)	41 (91,11%)	45 (5,87%)
	EEP, RENDA, PPI	3 (3,75%)	77 (96,25%)	80 (10,44%)
	EEP, RENDA, PPI, PCD	0 (0,00%)	3 (100,00%)	3 (0,39%)

Modalidade do Curso	**Cota MEC**	**Concluintes**	**Evadidos**	**Total**
	EEP, PCD	0 (0,00%)	2 (100,00%)	2 (0,26%)
	EEP, PPI	5 (7,04%)	66 (92,96%)	71 (9,27%)
	EEP, PPI, PCD	0 (0,00%)	3 (100,00%)	3 (0,39%)
	Não Se Aplica	119 (28,13%)	304 (71,87%)	423 (55,22%)
	S/I	0 (0,00%)	94 (100,00%)	94 (12,27%)
		136 (17,75%)	**630 (82,25%)**	**766 (39,77%)**
Todas as Modalidades de Curso	EEP	21 (19,81%)	85 (80,19%)	106 (5,50%)
	EEP, RENDA	33 (25,98%)	94 (74,02%)	127 (6,59%)
	EEP, RENDA, PPI	35 (16,36%)	179 (83,64%)	214 (11,11%)
	EEP, RENDA, PPI, PCD	0 (0,00%)	8 (100,00%)	8 (0,42%)
	EEP, PCD	0 (0,00%)	5 (100,00%)	5 (0,26%)
	EEP, PPI	34 (18,38%)	151 (81,62%)	185 (9,61%)
	EEP, PPI, PCD	1 (10,00%)	9 (90,00%)	10 (0,52%)
	Não Se Aplica	391 (35,90%)	698 (64,10%)	1.089 (56,54%)
	S/I	1 (0,55%)	181 (99,45%)	182 (9,45%)
		516 (26,79%)	**1.410 (73,21%)**	**1.926**

Fonte: Dados do Autor.

No caso dos cursos de Bacharelado, verifica-se que a taxa de evasão é alta para os alunos que se enquadram em diferentes cotas. Os estudantes que pertencem às cotas “EEP” e “RENDA” apresentam taxas de evasão de 88,89% e 92,00%, respectivamente. A taxa de evasão é ainda maior para aqueles que possuem combinações de cotas, como “EEP, RENDA, PPI, PCD”, com uma taxa de evasão de 100,00%. Nota-se também que os alunos que não se enquadram em nenhuma cota possuem uma taxa de evasão de 63,71%.

Para o caso dos cursos Integrados, a taxa de evasão é menor em comparação com os cursos de Bacharelado. Os estudantes que se enquadram apenas na cota “EEP” apresentam uma taxa de evasão de 44,00%, enquanto aqueles que possuem a combinação de “EEP, RENDA” possuem uma taxa de evasão de 39,39%. Os dados indicam que a presença da cota “EEP” tende a reduzir a taxa de evasão nessa modalidade. Para os cursos Subsequentes, observa-se uma alta taxa de evasão para os alunos que se enquadram em diferentes cotas. Os estudantes que pertencem à cota “EEP” possuem uma taxa de evasão de 100,00%, enquanto aqueles que possuem a combinação de “EEP, RENDA, PPI, PCD” também apresentam uma taxa de evasão de 100,00%. É importante ressaltar que a cota “EEP, PPI, PCD” possui uma taxa de evasão menor (50,00%), sugerindo que a presença de múltiplas cotas pode ter um efeito mitigador no abandono escolar.

No caso dos cursos de Tecnologia, é observada uma alta taxa de evasão para os alunos que se enquadram nas diferentes combinações de cotas. Aqueles que possuem apenas a cota "EEP" apresentam uma taxa de evasão de 88,89%, enquanto os alunos que possuem a combinação de "EEP, RENDA, PPI, PCD" têm uma taxa de evasão de 100,00%. Mais uma vez, os dados sugerem que a presença de múltiplas cotas não é suficiente para reduzir a taxa de evasão nessa modalidade.

Ao considerar todas as modalidades de curso em conjunto, os dados mostram que as taxas de evasão são relativamente altas para os alunos que se enquadram em diferentes combinações de cotas. A presença da cota "EEP, PCD" resulta em uma taxa de evasão de 100,00%. No entanto, é interessante notar que os alunos que não se enquadram em nenhuma cota possuem uma taxa de evasão de 64,10%, o que indica que a ausência de cotas pode estar associada a uma maior continuidade nos estudos.

5.1.10 Análise dos Dados do SUAP em Relação à Zona de Residência

Ao realizar a análise descritiva dos grupos de modalidade de curso do IFPB, com base nos dados coletados do SUAP, foram obtidas informações relevantes acerca das taxas de aprovação em cada categoria, classificadas de acordo com o campo Zona de residência dos alunos. Os resultados estão disponibilizados na Tabela 13.

Tabela 13 – Análise do SUAP para Modalidade do Curso em Relação à Zona.

Modalidade do Curso	**Zona**	**Concluintes**	**Evadidos**	**Total**
Bacharelado	Urbana	94 (27,73%)	245 (72,27%)	339 (83,91%)
	Rural	0 (0,00%)	10 (100,00%)	10 (2,48%)
	S/I	1 (1,82%)	54 (98,18%)	55 (13,61%)
		95 (23,51%)	**309 (76,49%)**	**404 (20,98%)**
Integrado	Urbana	212 (58,56%)	150 (41,44%)	362 (88,73%)
	Rural	13 (52,00%)	12 (48,00%)	25 (6,13%)
	S/I	1 (4,76%)	20 (95,24%)	21 (5,15%)
		226 (55,39%)	**182 (44,61%)**	**408 (21,18%)**
Subsequente	Urbana	56 (17,50%)	264 (82,50%)	320 (91,95%)
	Rural	2 (15,38%)	11 (84,62%)	13 (3,74%)
	S/I	1 (6,67%)	14 (93,33%)	15 (4,31%)
		59 (16,95%)	**289 (83,05%)**	**348 (18,07%)**
Tecnologia	Urbana	131 (20,28%)	515 (79,72%)	646 (84,33%)
	Rural	5 (20,83%)	19 (79,17%)	24 (3,13%)
	S/I	0 (0,00%)	96 (100,00%)	96 (12,53%)
		136 (17,75%)	**630 (82,25%)**	**766 (39,77%)**

Modalidade do Curso	Zona	Concluintes	Evadidos	Total
Todas as Modalidades de Curso	Urbana	493 (29,57%)	1.174 (70,43%)	1.667 (86,55%)
	Rural	20 (27,78%)	52 (72,22%)	72 (3,74%)
	S/I	3 (1,60%)	184 (98,40%)	187 (9,71%)
		516 (26,79%)	**1.410 (73,21%)**	**1.926**

Fonte: Dados do Autor.

No caso dos cursos de bacharelado, verifica-se que a taxa de evasão é significativamente alta, tanto para os estudantes que residem em áreas urbanas (72,27%), quanto para aqueles que residem em áreas rurais (100,00%). Isso indica um desafio considerável na retenção de alunos em ambos os contextos. No que diz respeito aos cursos integrados, percebemos uma diferença um pouco menos acentuada. Os estudantes da zona urbana apresentam uma taxa de evasão de 41,44%, enquanto aqueles da zona rural têm uma taxa de evasão ligeiramente mais elevada, atingindo 48,00%. Embora a diferença não seja tão significativa, é importante considerar estratégias de apoio específicas para alunos que residem em áreas rurais, a fim de minimizar a evasão nesse grupo.

No caso dos cursos subsequentes, tanto os alunos da zona urbana, quanto os da zona rural enfrentam desafios consideráveis em relação à evasão escolar. A taxa de evasão para alunos da zona urbana é de 82,50%, enquanto aqueles da zona rural apresentam uma taxa ainda maior, atingindo 84,62%. Quanto aos cursos de tecnologia, observa-se uma relativa uniformidade nas taxas de evasão entre alunos da zona urbana (79,72%) e alunos da zona rural (79,17%). Embora ambas as taxas sejam consideráveis, a falta de discrepância significativa indica que a zona de residência pode ter menos influência sobre a evasão nessa modalidade.

5.1.11 Análise dos Dados do SUAP em Relação à origem da Escola

A análise descritiva dos grupos de modalidade de curso do IFPB, com base nos dados do SUAP, revelou informações relevantes sobre as taxas de aprovação em cada categoria, agrupadas de acordo com o campo origem da escola dos alunos. Os resultados são apresentados na Tabela 14.

Tabela 14 – Análise do SUAP para Modalidade do Curso em Relação à Escola de Origem.

Modalidade do Curso	Origem Escola	Concluintes	Evadidos	Total
Bacharelado	Pública	54 (21,26%)	200 (78,74%)	254 (62,87%)
	Privada	41 (27,33%)	109 (72,67%)	150 (37,13%)
		95 (23,51%)	**309 (76,49%)**	**404 (20,98%)**
Integrado	Pública	141 (54,86%)	116 (45,14%)	257 (62,99%)
	Privada	85 (56,29%)	66 (43,71%)	151 (37,01%)

Modalidade do Curso	Origem Escola	Concluintes	Evadidos	Total
		226 (55,39%)	**182 (44,61%)**	**408 (21,18%)**
Subsequente	Pública	42 (14,24%)	253 (85,76%)	295 (84,77%)
	Privada	17 (32,08%)	36 (67,92%)	53 (15,23%)
		59 (16,95%)	**289 (83,05%)**	**348 (18,07%)**
Tecnologia	Pública	85 (15,32%)	470 (84,68%)	555 (72,45%)
	Privada	51 (24,17%)	160 (75,83%)	211 (27,55%)
		136 (17,75%)	**630 (82,25%)**	**766 (39,77%)**
Todas as Modalidades de Curso	Pública	322 (23,66%)	1.039 (76,34%)	1.361 (70,66%)
	Privada	194 (34,34%)	371 (65,66%)	565 (29,34%)
		516 (26,79%)	**1.410 (73,21%)**	**1.926**

Fonte: Dados do Autor.

No caso do Bacharelado, observa-se que a taxa de evasão é de 78,74% para alunos provenientes de escolas públicas, enquanto para aqueles provenientes de escolas privadas, a taxa é ligeiramente menor, com 72,67%. Essa diferença pode sugerir que estudantes oriundos de escolas públicas enfrentam desafios adicionais ao longo de sua jornada acadêmica, o que pode contribuir para uma maior taxa de evasão.

No Integrado, por outro lado, tanto alunos de escolas públicas, quanto privadas apresentam taxas de evasão relativamente baixas, sendo 45,14% e 43,71%, respectivamente. Isso indica que, nesse contexto, a origem da escola parece ter um impacto semelhante na evasão, independentemente de ser pública ou privada. Já na modalidade Subsequente, os dados revelam uma diferença mais significativa. Os alunos provenientes de escolas públicas apresentam uma taxa de evasão consideravelmente mais alta, atingindo 85,76%. Por outro lado, aqueles provenientes de escolas privadas possuem uma taxa de evasão de 67,92%. Isso sugere que estudantes que frequentaram escolas públicas podem enfrentar desafios adicionais nessa modalidade específica, impactando sua permanência e conclusão dos cursos.

No caso da modalidade Tecnologia, a diferença entre as taxas de evasão dos alunos provenientes de escolas públicas (84,68%) e privadas (75,83%) também indica um impacto relevante da origem da escola. Esses números indicam que estudantes vindos de escolas públicas podem enfrentar obstáculos que afetam sua permanência e conclusão dos cursos tecnológicos no IFPB. Ao considerar todas as modalidades de curso, a taxa de evasão para alunos provenientes de escolas públicas é de 76,34%, enquanto para aqueles provenientes de escolas privadas é de 65,66%. Essa análise abrangente confirma que a origem da escola pode influenciar a evasão escolar no IFPB, sendo mais prevalente entre os alunos oriundos de escolas públicas.

5.1.12 Análise dos Dados do SUAP em Relação à Faixa Etária

A análise descritiva dos grupos de modalidade de curso do IFPB, utilizando os dados do SUAP, trouxe à tona informações significativas sobre as taxas de aprovação em cada categoria, classificadas conforme o campo faixa etária dos estudantes. Os resultados podem ser encontrados na Tabela 15.

Tabela 15 – Análise do SUAP para Modalidade do Curso em Relação à Faixa Etária.

Modalidade do Curso	Faixa Etária	Concluintes	Evadidos	Total
Bacharelado	15 a 19 anos	0 (0,00%)	10 (100,00%)	10 (2,48%)
	20 a 24 anos	7 (5,51%)	120 (94,49%)	127 (31,44%)
	25 a 29 anos	63 (38,41%)	101 (61,59%)	164 (40,59%)
	30 a 34 anos	21 (38,89%)	33 (61,11%)	54 (13,37%)
	35 a 39 anos	3 (11,54%)	23 (88,46%)	26 (6,44%)
	40 a 44 anos	0 (0,00%)	13 (100,00%)	13 (3,22%)
	45 a 49 anos	1 (14,29%)	6 (85,71%)	7 (1,73%)
	55 a 59 anos	0 (0,00%)	2 (100,00%)	2 (0,50%)
	> 60 anos	0 (0,00%)	1 (100,00%)	1 (0,25%)
		95 (23,51%)	**309 (76,49%)**	**404 (20,98%)**
Integrado	< 14 anos	0 (0,00%)	1 (100,00%)	1 (0,25%)
	15 a 19 anos	69 (46,00%)	81 (54,00%)	150 (36,76%)
	20 a 24 anos	156 (61,90%)	96 (38,10%)	252 (61,76%)
	25 a 29 anos	1 (25,00%)	3 (75,00%)	4 (0,98%)
	30 a 34 anos	0 (0,00%)	1 (100,00%)	1 (0,25%)
		226 (55,39%)	**182 (44,61%)**	**408 (21,18%)**
Subsequente	15 a 19 anos	0 (0,00%)	6 (100,00%)	6 (1,72%)
	20 a 24 anos	9 (12,50%)	63 (87,50%)	72 (20,69%)
	25 a 29 anos	16 (17,78%)	74 (82,22%)	90 (25,86%)
	30 a 34 anos	14 (21,88%)	50 (78,13%)	64 (18,39%)
	35 a 39 anos	8 (14,81%)	46 (85,19%)	54 (15,52%)
	40 a 44 anos	1 (2,86%)	34 (97,14%)	35 (10,06%)
	45 a 49 anos	7 (35,00%)	13 (65,00%)	20 (5,75%)
	50 a 54 anos	4 (57,14%)	3 (42,86%)	7 (2,01%)
		59 (16,95%)	**289 (83,05%)**	**348 (18,07%)**
Tecnologia	< 14 anos	0 (0,00%)	2 (100,00%)	2 (0,26%)
	15 a 19 anos	0 (0,00%)	11 (100,00%)	11 (1,44%)
	20 a 24 anos	18 (9,63%)	169 (90,37%)	187 (24,41%)
	25 a 29 anos	55 (22,73%)	187 (77,27%)	242 (31,59%)
	30 a 34 anos	42 (25,30%)	124 (74,70%)	166 (21,67%)

Modalidade do Curso	**Faixa Etária**	**Concluintes**	**Evadidos**	**Total**
	35 a 39 anos	9 (9,47%)	86 (90,53%)	95 (12,40%)
	40 a 44 anos	8 (20,51%)	31 (79,49%)	39 (5,09%)
	45 a 49 anos	2 (15,38%)	11 (84,62%)	13 (1,70%)
	50 a 54 anos	1 (16,67%)	5 (83,33%)	6 (0,78%)
	55 a 59 anos	0 (0,00%)	3 (100,00%)	3 (0,39%)
	> 60 anos	1 (50,00%)	1 (50,00%)	2 (0,26%)
		136 (17,75%)	**630 (82,25%)**	**766 (39,77%)**
Todas as Modalidades de Curso	< 14 anos	0 (0,00%)	3 (100,00%)	3 (0,16%)
	15 a 19 anos	69 (38,98%)	108 (61,02%)	177 (9,19%)
	20 a 24 anos	190 (29,78%)	448 (70,22%)	638 (33,13%)
	25 a 29 anos	135 (27,00%)	365 (73,00%)	500 (25,96%)
	30 a 34 anos	77 (27,02%)	208 (72,98%)	285 (14,80%)
	35 a 39 anos	20 (11,43%)	155 (88,57%)	175 (9,09%)
	40 a 44 anos	9 (10,34%)	78 (89,66%)	87 (4,52%)
	45 a 49 anos	10 (25,00%)	30 (75,00%)	40 (2,08%)
	50 a 54 anos	5 (38,46%)	8 (61,54%)	13 (0,67%)
	55 a 59 anos	0 (0,00%)	5 (100,00%)	5 (0,26%)
	> 60 anos	1 (33,33%)	2 (66,67%)	3 (0,16%)
		516 (26,79%)	**1.410 (73,21%)**	**1.926**

Fonte: Dados do Autor.

A análise dos dados de evasão escolar no IFPB, na modalidade de Bacharelado, em relação à faixa etária dos alunos aponta alguns pontos relevantes. Destaca-se a taxa de evasão de 100% para alunos com idades de 15 a 19 anos e de 40 a 44 anos. Alunos na faixa etária de 20 a 24 anos apresentam uma alta taxa de evasão de 94,49%. Já para os alunos com idades entre 25 e 29 anos, a taxa de evasão é de 61,59%. Esses resultados indicam que os alunos mais jovens e mais velhos enfrentam desafios significativos que contribuem para a evasão escolar. Medidas de suporte e acompanhamento específicas para essas faixas etárias podem ser necessárias para melhorar os índices de retenção e conclusão dos cursos.

No caso da modalidade Integrado, é importante ressaltar que a faixa etária é mais restrita, pois se trata de um curso integrado ao ensino médio. A análise dos dados revela que alunos menores de 14 anos apresentam uma taxa de evasão de 100%, o que indica um desafio particular nessa faixa etária. Para os alunos de 15 a 19 anos, a taxa de evasão é de 54%, demonstrando que ainda há uma parcela significativa de evasão neste grupo. Por outro lado, os alunos entre 20 e 24 anos apresentam uma taxa de evasão mais baixa, de 38,10%. Embora a quantidade de registros para as faixas de 25 a 29 anos e 30 a 34 anos seja menor, observa-se que a taxa de evasão é mais alta para esses grupos, atingindo 75% e 100%, respectivamente.

Na modalidade Subsequente, a análise dos dados em relação à faixa etária revela alguns pontos. A faixa etária de 15 a 19 anos apresenta uma taxa de evasão de 100%, indicando um cenário desafiador para esses estudantes. Já para os alunos de 20 a 24 anos, a taxa de evasão é de 87,50%, evidenciando a importância de estratégias de apoio para esse grupo. Os alunos entre 25 e 29 anos também apresentam uma taxa de evasão significativa, atingindo 82,22%. As faixas de 30 a 34 anos e 35 a 39 anos registram taxas de evasão de 78,13% e 85,19%, respectivamente. É interessante notar que a faixa etária de 50 a 54 anos apresenta uma taxa de evasão menor, de 42,86%. No entanto, é importante considerar que a porcentagem de registros para essa faixa etária é menor em comparação às outras.

Ao analisar a modalidade Tecnologia em relação à faixa etária dos alunos, podemos destacar algumas informações. Observa-se que a presença de alunos menores de 14 anos é muito baixa, representando apenas 0,26% dos registros, porém todos eles apresentaram evasão, o que evidencia a dificuldade dessa faixa etária na modalidade. Os alunos de 15 a 19 anos também apresentam uma taxa de evasão de 100%, mesmo com uma porcentagem de registros baixa, de 1,44%. Já os estudantes entre 20 e 24 anos representam a maior porcentagem de registros, com 24,41%, e uma taxa de evasão de 90,37%. As faixas etárias de 25 a 29 anos e 30 a 34 anos registram taxas de evasão de 77,27% e 74,70%, respectivamente. Vale ressaltar que a faixa etária de 55 a 59 anos apresenta 100% de evasão, mas com uma representatividade baixa nos registros, de apenas 0,39%.

Ao analisar todas as modalidades de curso em relação à faixa etária dos alunos, podemos destacar alguns padrões. A presença de alunos menores de 14 anos é bastante baixa, representando apenas 0,16% dos registros, e todos eles apresentaram evasão. A faixa etária mais representativa é a de 20 a 24 anos, com 33,13% dos registros, e uma taxa de evasão de 70,22%. Os alunos entre 25 e 29 anos e os de 30 a 34 anos apresentam taxas de evasão próximas, com 73,00% e 72,98%, respectivamente. É importante notar que a faixa etária de 35 a 39 anos registra uma taxa de evasão mais alta, chegando a 88,57%, assim como a faixa de 40 a 44 anos, com 89,66% de evasão. A faixa etária de 55 a 59 anos e os menores de 14 anos têm taxa de evasão de 100%, mas com baixa representatividade nos registros. Já os alunos acima de 60 anos apresentam uma taxa de evasão de 66,67%.

5.1.13 Análise dos Dados do SUAP em Relação à Cor/Raça

Através da análise descritiva dos grupos de modalidade de curso do IFPB, com base nos dados disponibilizados pelo SUAP, foram obtidas informações importantes acerca das taxas de aprovação em cada categoria, organizadas de acordo com o campo cor/raça dos discentes. Os resultados estão disponíveis na Tabela 16.

Tabela 16 – Análise do SUAP para Modalidade do Curso em Relação à Cor/Raça.

Modalidade do Curso	Cor/Raça	Concluintes	Evadidos	Total
Bacharelado	Amarela	1 (33,33%)	2 (66,67%)	3 (0,74%)
	Branca	38 (23,03%)	127 (76,97%)	165 (40,84%)
	Parda	34 (18,99%)	145 (81,01%)	179 (44,31%)
	Preta	8 (25,00%)	24 (75,00%)	32 (7,92%)
	Não Declarada	14 (56,00%)	11 (44,00%)	25 (6,19%)
		95 (23,51%)	**309 (76,49%)**	**404 (20,98%)**
Integrado	Amarela	1 (50,00%)	1 (50,00%)	2 (0,49%)
	Branca	102 (63,35%)	59 (36,65%)	161 (39,46%)
	Parda	109 (51,66%)	102 (48,34%)	211 (51,72%)
	Preta	8 (34,78%)	15 (65,22%)	23 (5,64%)
	Não Declarada	6 (54,55%)	5 (45,45%)	11 (2,70%)
		226 (55,39%)	**182 (44,61%)**	**408 (21,18%)**
Subsequente	Amarela	0 (0,00%)	7 (100,00%)	7 (2,01%)
	Branca	17 (19,77%)	69 (80,23%)	86 (24,71%)
	Indigena	0 (0,00%)	1 (100,00%)	1 (0,29%)
	Parda	35 (18,04%)	159 (81,96%)	194 (55,75%)
	Preta	4 (9,09%)	40 (90,91%)	44 (12,64%)
	Não Declarada	3 (18,75%)	13 (81,25%)	16 (4,60%)
		59 (16,95%)	**289 (83,05%)**	**348 (18,07%)**
Tecnologia	Amarela	1 (14,29%)	6 (85,71%)	7 (0,91%)
	Branca	75 (23,15%)	249 (76,85%)	324 (42,30%)
	Indigena	0 (0,00%)	3 (100,00%)	3 (0,39%)
	Parda	46 (13,33%)	299 (86,67%)	345 (45,04%)
	Preta	9 (15,52%)	49 (84,48%)	58 (7,57%)
	Não Declarada	5 (17,24%)	24 (82,76%)	29 (3,79%)
		136 (17,75%)	**630 (82,25%)**	**766 (39,77%)**
Todas as Modalidades de Curso	Amarela	3 (15,79%)	16 (84,21%)	19 (0,99%)
	Branca	232 (31,52%)	504 (68,48%)	736 (38,21%)
	Indigena	0 (0,00%)	4 (100,00%)	4 (0,21%)
	Parda	224 (24,11%)	705 (75,89%)	929 (48,23%)
	Preta	29 (18,47%)	128 (81,53%)	157 (8,15%)
	Não Declarada	28 (34,57%)	53 (65,43%)	81 (4,21%)
		516 (26,79%)	**1.410 (73,21%)**	**1.926**

Fonte: Dados do Autor.

Ao analisar a evasão escolar no IFPB agrupada pela cor/raça declarada pelos alunos em cada modalidade de curso, observa-se alguns padrões interessantes. No curso de Bacharelado, a maioria dos registros é composta por alunos brancos (40,84%), seguidos por alunos pardos (44,31%). As taxas de evasão para esses grupos são de 76,97% e 81,01%, respectivamente. Alunos pretos representam uma porcentagem menor (7,92%), mas apresentam uma taxa de evasão de 75,00%. Já os alunos de cor/raça amarela têm a menor representatividade (0,74%) e uma taxa de evasão de 66,67%.

No curso integrado, novamente encontramos a predominância de alunos brancos (39,46%) e pardos (51,72%), com taxas de evasão de 36,65% e 48,34%, respectivamente. Alunos pretos têm uma representatividade menor (5,64%), porém, apresentam uma taxa de evasão de 65,22%. Os alunos de cor/raça amarela têm uma representatividade ainda menor (0,49%) e uma taxa de evasão de 50,00%. Na modalidade subsequente, os alunos pardos representam a maioria dos registros (55,75%) e têm uma taxa de evasão de 81,96%. Alunos brancos representam 24,71% dos registros, com uma taxa de evasão de 80,23%. Os alunos pretos (12,64%) têm uma taxa de evasão mais alta, chegando a 90,91%. Alunos de cor/raça amarela e indígena têm uma representatividade menor (2,01% e 0,29%, respectivamente) e apresentam uma taxa de evasão de 100,00%.

No curso de Tecnologia, os alunos pardos representam a maior porcentagem de registros (45,04%) e têm uma taxa de evasão de 86,67%. Alunos brancos representam 42,30% dos registros, com uma taxa de evasão de 76,85%. Alunos de cor/raça amarela têm uma taxa de evasão de 85,71%, enquanto alunos pretos apresentam uma taxa de evasão de 84,48%. Alunos indígenas têm uma representatividade menor (0,39%) e taxa de evasão de 100,00%.

Ao considerar todas as modalidades de curso, os alunos pardos também são a maioria em termos de registros (48,23%) e têm uma taxa de evasão de 75,89%. Alunos brancos representam 38,21% dos registros, com uma taxa de evasão de 68,48%. Alunos de cor/raça amarela e pretos têm uma taxa de evasão de 84,21% e 81,53%, respectivamente. Os alunos indígenas têm a menor representatividade (0,21%) e uma taxa de evasão de 100,00%.

5.1.14 Análise dos Dados do SUAP em Relação ao Estado Civil

Ao realizar a análise descritiva dos grupos de modalidade de curso do IFPB, utilizando os dados fornecidos pelo SUAP, foram identificadas informações de destaque sobre as taxas de aprovação em cada categoria, agrupadas com base no campo estado civil dos alunos. Os resultados são expostos na Tabela 17.

Tabela 17 – Análise do SUAP para Modalidade do Curso em Relação ao Estado Civil.

Modalidade do Curso	Estado Civil	Concluintes	Evadidos	Total
Bacharelado	Casado	7 (16,67%)	35 (83,33%)	42 (10,40%)

Modalidade do Curso	Estado Civil	Concluintes	Evadidos	Total
	Divorciado	0 (0,00%)	2 (100,00%)	2 (0,50%)
	Solteiro	88 (28,76%)	218 (71,24%)	306 (75,74%)
	S/I	0 (0,00%)	54 (100,00%)	54 (13,37%)
		95 (23,51%)	**309 (76,49%)**	**404 (20,98%)**
Integrado	Casado	2 (100,00%)	0 (0,00%)	2 (0,49%)
	Solteiro	223 (57,92%)	162 (42,08%)	385 (94,36%)
	União Estável	0 (0,00%)	1 (100,00%)	1 (0,25%)
	S/I	1 (5,00%)	19 (95,00%)	20 (4,90%)
		226 (55,39%)	**182 (44,61%)**	**408 (21,18%)**
Subsequente	Casado	15 (19,23%)	63 (80,77%)	78 (22,41%)
	Divorciado	3 (25,00%)	9 (75,00%)	12 (3,45%)
	Solteiro	41 (17,15%)	198 (82,85%)	239 (68,68%)
	União Estável	0 (0,00%)	4 (100,00%)	4 (1,15%)
	Viúvo	0 (0,00%)	1 (100,00%)	1 (0,29%)
	S/I	0 (0,00%)	14 (100,00%)	14 (4,02%)
		59 (16,95%)	**289 (83,05%)**	**348 (18,07%)**
Tecnologia	Casado	22 (27,85%)	57 (72,15%)	79 (10,31%)
	Divorciado	0 (0,00%)	7 (100,00%)	7 (0,91%)
	Solteiro	113 (19,48%)	467 (80,52%)	580 (75,72%)
	União Estável	1 (16,67%)	5 (83,33%)	6 (0,78%)
	S/I	0 (0,00%)	94 (100,00%)	94 (12,27%)
		136 (17,75%)	**630 (82,25%)**	**766 (39,77%)**
Todas as Modalidades de Curso	Casado	46 (22,89%)	155 (77,11%)	201 (10,44%)
	Divorciado	3 (14,29%)	18 (85,71%)	21 (1,09%)
	Solteiro	465 (30,79%)	1.045 (69,21%)	1.510 (78,40%)
	União Estável	1 (9,09%)	10 (90,91%)	11 (0,57%)
	Viúvo	0 (0,00%)	1 (100,00%)	1 (0,05%)
	S/I	1 (0,55%)	181 (99,45%)	182 (9,45%)
		516 (26,79%)	**1.410 (73,21%)**	**1.926**

Fonte: Dados do Autor.

No curso de Bacharelado, a maioria dos alunos declarou-se solteira e apresentou uma taxa de evasão considerável (71,24%). Por outro lado, a representatividade dos alunos casados é menor (10,40%), mas eles enfrentam uma taxa de evasão ainda mais alta (83,33%). No curso Integrado, a maioria dos alunos é solteiro (94,36%) e apresenta uma taxa de evasão relativamente baixa (42,08%). Além disso, não há registros de evasão entre os alunos casados nessa modalidade.

No curso Subsequente, os alunos solteiros têm a maior representatividade (68,68%) e

também enfrentam uma taxa de evasão significativa (82,85%). Os alunos casados e em união estável também apresentam taxas de evasão consideráveis. No curso de Tecnologia, os alunos solteiros têm a maior representatividade (75,72%) e também uma taxa de evasão considerável (80,52%). Além disso, os alunos casados (80,77%) e em união estável (100,00%) também enfrentam taxas de evasão significativas.

5.2 Resultados das Seleção de Atributos

Na etapa de seleção de atributos sobre evasão escolar no IFPB, foram realizadas análises para identificar quais atributos apresentam maior relevância na predição da evasão escolar. O objetivo era identificar a solução de seleção de atributos que apresenta os melhores resultados para todos os subconjuntos de dados analisados, e desse modo, selecionar as variáveis mais significativas que contribuem para a precisão dos modelos preditivos.

5.2.1 Resultados das Seleção de Features da PNP

No processo de seleção de ferramentas para atributos no conjunto de dados do IFPB proveniente da base da PNP, foram empregados os métodos *Chi2*, *Embedded*, *KBest*, *Wrappers Gradiente Boosting* (*Wrappers GB*) e *Wrappers Logistic Regression* (*Wrappers LR*). Esses métodos foram aplicados a diversas agrupações dos dados com base no campo "*tipo_curso*". O objetivo era determinar qual ferramenta ofereceria os resultados mais robustos em termos de desempenho e precisão.

Além do processo de escolha da ferramenta de seleção de atributos, foi realizada uma etapa adicional para determinar a quantidade ideal de atributos a serem selecionados em cada conjunto de dados. Foram repetidos os testes, variando de 1 até 31 atributos, em cada conjunto de dados, utilizando o seletor selecionado anteriormente.

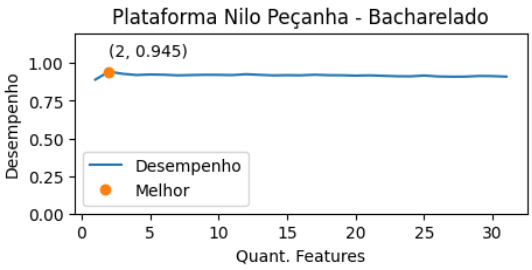
Cada análise foi repetida com todos os conjuntos de atributos selecionados e ferramentas de classificação utilizando a métrica *F1-Score* para analisar a quantidade ideal de atributos. O objetivo dessa etapa foi encontrar o número ótimo de atributos que oferecesse um bom desempenho na previsão da evasão escolar, evitando a inclusão de atributos desnecessários que poderiam afetar negativamente o desempenho do modelo ou atribuir uma carga desnecessária ao modelo. Na Figura 15 é apresentado os resultados obtidos com a variação da quantidade de atributos.

Os seletores e a quantidade de atributos que apresentaram os melhores resultados para cada grupo foram:

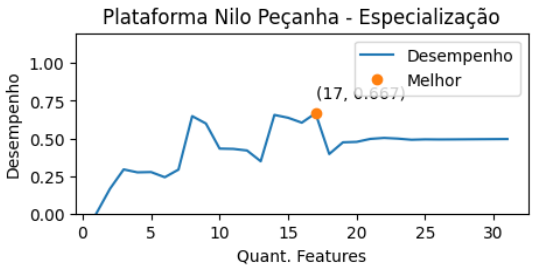
- **Bacharelado**: *Wrappers GB* com 2 atributos.
- **Especialização**: *Wrappers* LR com 17 atributos.

Figura 15 – Número de Atributos dados da PNP

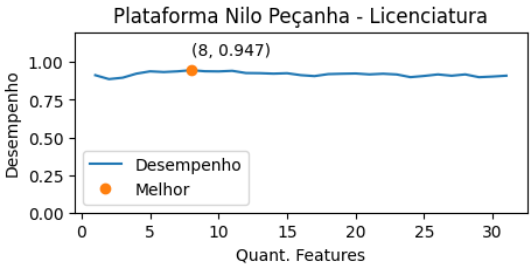
(a) Árvore de Decisão com *Wrappers* GB



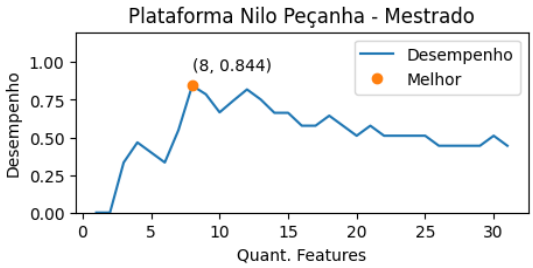
(b) *Naive Bayes* com *Wrappers* LR



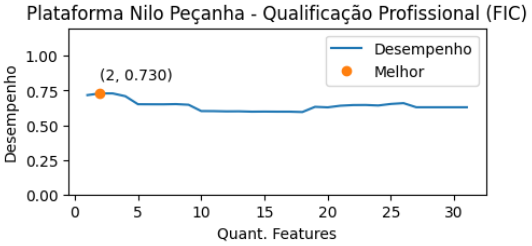
(c) *Multilayer Perceptron* com *Embedded*



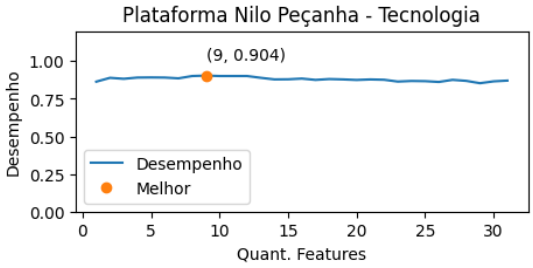
(d) *Multilayer Perceptron* com *Embedded*



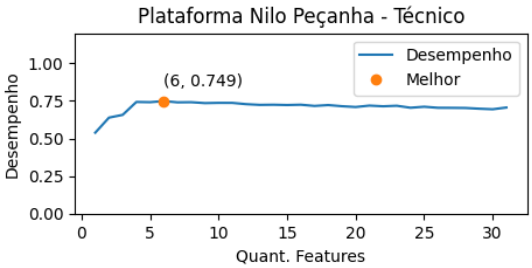
(e) SVM com *Embedded*



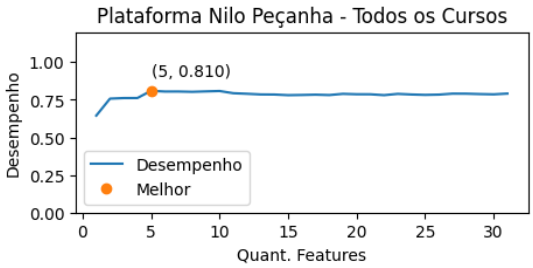
(f) Floresta Aleatória com *Wrappers* GB



(g) Floresta Aleatória com *Wrappers* GB



(h) Floresta Aleatória com *Wrappers* LR



Fonte: Dados do Autor.

- **Licenciatura**: *Embedded* com 8 atributos.
- **Mestrado**: *Embedded* com 8 atributos.
- **Qualificação Profissional (FIC)**: *Embedded* com 2 atributos.
- **Tecnologia**: *Wrappers GB* com 9 atributos.
- **Mestrado**: *Wrappers GB* com 6 atributos.
- **Todos os Cursos**: *Wrappers LR* com 5 atributos.

Após a conclusão dos testes, procedeu-se ao cálculo dos valores do RMSE utilizando a métrica *F1-Score* como referência. A análise dos resultados permitiu identificar a quantidade ideal de atributos fornecida por um único seletor, otimizando o desempenho do modelo para todos os grupos no conjunto de dados da PNP. Considerando todas as modalidades de curso, o seletor *Embedded* mostrou os melhores resultados ao utilizar oito atributos, sendo selecionado como a solução única de seletor e quantidade de atributos para as próximas etapas do processo de modelagem.

Os atributos selecionados em cada grupo "*tipo_curso*" são apresentados nas tabelas Tabela 18 e Tabela 19, ordenados por relevância, conforme identificado pelo seletor *Embedded*.

Tabela 18 – Resultado da Seleção de Atributos pelo Tipo de Curso na PNP - Parte 1/2.

#	Todos os Cursos	Bacharelado	Especialização	Licenciatura
1	cod_matricula	cod_matricula	cod_matricula	cod_matricula
2	mes_ocorrencia	mes_ocorrencia	mes_ocorrencia	mes_ocorrencia
3	carga_horaria	codigo_ciclo_ matricula	fonte_financiamento	codigo_ciclo_ matricula
4	codigo_ciclo_ matricula	ano	codigo_municipio_dv	municipio
5	nome_curso	fim_ciclo	codigo_ciclo_ matricula	renda_familiar
6	total_inscritos	renda_familiar	faixa_etaria	fim_ciclo
7	faixa_etaria	faixa_etaria	renda_familiar	total_inscritos
8	carga_minima	municipio	unidade_ensino	carga_horaria
9	renda_familiar	fator_esforco_curso	cor_raca	vagas_ofertadas
10	tipo_oferta	data_matricula	municipio	codigo_municipio_dv
11	unidade_ensino	total_inscritos	data_matricula	faixa_etaria
12	fim_ciclo	inicio_ciclo	total_inscritos	ano
13	ano	carga_horaria	inicio_ciclo	turno
14	cor_raca	vagas_ofertadas	fim_ciclo	data_matricula

#	Todos os Cursos	Bacharelado	Especialização	Licenciatura
15	vagas_ofertadas	cor_raca	ano	fator_esforco_curso
16	inicio_ciclo	eixo_tecnologico	sexo	inicio_ciclo
17	fator_esforco_curso	sexo	codigo_unidade_ ensino_sistec	unidade_ensino
18	data_matricula	codigo_municipio_dv	carga_horaria	cor_raca
19	codigo_unidade_ ensino_sistec	codigo_unidade_ ensino_sistec	vagas_ofertadas	codigo_unidade_ ensino_sistec
20	municipio	unidade_ensino	nome_curso	nome_curso
21	sub_eixo_tecnologico	fonte_financiamento	sub_eixo_tecnologico	fonte_financiamento
22	sexo	sub_eixo_tecnologico	turno	sexo
23	tipo_curso	tipo_oferta	eixo_tecnologico	tipo_oferta
24	codigo_municipio_dv	nome_curso	tipo_oferta	sub_eixo_tecnologico
25	turno	turno	modalidade_ensino	modalidade_ensino
26	eixo_tecnologico	carga_minima	fator_esforco_curso	regiao
27	fonte_financiamento	modalidade_ensino	carga_minima	eixo_tecnologico
28	modalidade_ensino	tipo_curso	regiao	uf
29	regiao	regiao	tipo_curso	tipo_curso
30	uf	uf	uf	carga_minima
31	instituicao	instituicao	instituicao	instituicao

Tabela 19 – Resultado da Seleção de Features pelo Tipo de Curso na PNP - Parte 2/2.

#	Mestrado	FIC	Tecnologia	Técnico
1	cod_matricula	cod_matricula	cod_matricula	cod_matricula
2	mes_ocorrencia	codigo_ciclo_ matricula	mes_ocorrencia	mes_ocorrencia
3	total_inscritos	faixa_etaria	codigo_ciclo_ matricula	carga_horaria
4	codigo_ciclo_ matricula	inicio_ciclo	total_inscritos	codigo_ciclo_ matricula
5	fim_ciclo	mes_ocorrencia	carga_horaria	total_inscritos
6	faixa_etaria	cor_raca	renda_familiar	fim_ciclo
7	ano	vagas_ofertadas	fim_ciclo	ano
8	vagas_ofertadas	renda_familiar	ano	tipo_oferta
9	inicio_ciclo	nome_curso	fator_esforco_curso	renda_familiar
10	renda_familiar	total_inscritos	municipio	fator_esforco_curso
11	sub_eixo_tecnologico	sexo	faixa_etaria	inicio_ciclo

#	Mestrado	FIC	Tecnologia	Técnico
12	data_matricula	fim_ciclo	inicio_ciclo	faixa_etaria
13	nome_curso	sub_eixo_tecnologico	nome_curso	data_matricula
14	cor_raca	data_matricula	data_matricula	turno
15	carga_horaria	carga_horaria	vagas_ofertadas	codigo_unidade_ ensino_sistec
16	fonte_financiamento	codigo_unidade_ ensino_sistec	eixo_tecnologico	cor_raca
17	eixo_tecnologico	eixo_tecnologico	cor_raca	nome_curso
18	sexo	unidade_ensino	codigo_unidade_ ensino_sistec	municipio
19	municipio	modalidade_ensino	unidade_ensino	vagas_ofertadas
20	tipo_oferta	municipio	sub_eixo_tecnologico	unidade_ensino
21	turno	codigo_municipio_dv	fonte_financiamento	sub_eixo_tecnologico
22	modalidade_ensino	turno	codigo_municipio_dv	codigo_municipio_dv
23	tipo_curso	carga_minima	sexo	sexo
24	codigo_unidade_ ensino_sistec	tipo_oferta	carga_minima	fonte_financiamento
25	codigo_municipio_dv	ano	turno	eixo_tecnologico
26	unidade_ensino	fonte_financiamento	tipo_oferta	carga_minima
27	uf	fator_esforco_curso	modalidade_ensino	modalidade_ensino
28	regiao	tipo_curso	tipo_curso	regiao
29	carga_minima	uf	uf	tipo_curso
30	fator_esforco_curso	regiao	regiao	uf
31	instituicao	instituicao	instituicao	instituicao

Fonte: Dados do Autor.

A análise comparativa dos atributos mais e menos relevantes, com base no agrupamento por Tipo de Curso na PNP, revela um quadro intrigante e esclarecedor sobre os fatores que podem afetar a evasão escolar em contextos acadêmicos diversos.

Inicialmente, chamou a atenção o fato de o atributo “cod_matricula” estar entre os mais relevantes em todos os tipos de cursos. A princípio, esse resultado pode parecer estranho, já que o código de matrícula deveria ser um valor único, arbitrário e aparentemente aleatório. No entanto, essa relevância pode ser justificada pela hipótese de que os números de matrícula podem, de alguma forma, conter informações ocultas sobre os alunos, como o período de ingresso ou alguma característica do sistema de matrícula que influencie a evasão. Além disso, ao considerarmos que todas as análises são restritas a uma única instituição de ensino (IFPB), faz sentido que atributos relacionados à instituição (“*instituicao*”), unidade federativa (“*uf*”) e região geográfica

(“*regiao*”) apresentem baixa relevância. Afinal, dentro do contexto do IFPB, esses atributos não apresentariam variações significativas que pudessem influenciar a evasão.

No tocante aos atributos mais relevantes, é notável que características específicas de cada tipo de curso são identificadas como fatores influentes na evasão. Por exemplo, em cursos de bacharelado, variáveis como “codigo_ciclo_matricula” e “fim_ciclo” mostraram-se importantes, o que pode indicar que a progressão do ciclo e a conclusão do mesmo afetam a permanência dos alunos. Por outro lado, em cursos de especialização, a “fonte_financiamento” e o “codigo_municipio_dv” demonstraram relevância, sugerindo que fatores financeiros e locais podem ser determinantes nesse cenário.

Também é válido destacar que, ao considerar a abordagem dos algoritmos de classificação, certos modelos apresentaram melhor desempenho na predição da evasão de acordo com o tipo de curso. Essa variação pode indicar que diferentes tipos de cursos possuem relações distintas entre seus atributos e o fenômeno da evasão, o que reforça a importância de um enfoque específico para cada contexto acadêmico. De maneira geral, essa análise comparativa reforça a complexidade do problema da evasão escolar e a necessidade de uma abordagem minuciosa e personalizada para lidar com essa questão. O entendimento dos atributos mais e menos relevantes em diferentes contextos acadêmicos pode fornecer *insights* para o desenvolvimento de estratégias de prevenção e intervenção mais direcionadas, contribuindo para a melhoria da retenção dos alunos no ambiente educacional. Em suma, a análise comparativa dos atributos mais e menos relevantes fornece uma base para uma abordagem na compreensão e enfrentamento da evasão escolar dentro do IFPB. A utilização dessas informações na formulação de estratégias educacionais pode promover uma maior retenção de alunos, contribuindo para um ambiente educacional mais inclusivo e bem-sucedido.

5.2.2 Resultados das Seleção de Atributos do SUAP

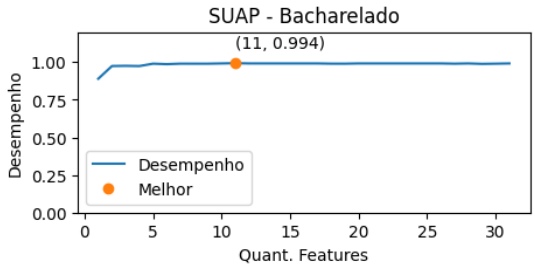
No processo de escolha da ferramenta de seleção de atributos para os dados do IFPB da base de dados do SUAP, foram utilizados os métodos de seleção de atributos *Chi2*, *Embedded*, *KBest*, *Wrappers Gradiente Boosting*(*Wrappers GB*) e *Wrappers Logistic Regression*(*Wrappers LR*) em todos os diferentes agrupamentos dos dados a partir do campo “*curso_modalidade*”, com o objetivo de identificar a ferramenta que proporcionasse os melhores resultados em termos de desempenho e precisão. Além do processo de escolha da ferramenta de seleção de atributos, foi realizada uma etapa adicional para determinar a quantidade ideal de atributos a serem selecionados em cada conjunto de dados. Foram repetidos os testes, variando de 1 até 31 atributos, em cada conjunto de dados, utilizando o seletor selecionado anteriormente.

Cada análise foi repetida também com todos os conjuntos de atributos selecionados e ferramentas de classificação utilizando a métrica *F1-Score* para analisar a quantidade ideal de atributos. O objetivo dessa etapa foi encontrar o número ótimo de atributos que oferecesse um bom desempenho na previsão da evasão escolar, evitando a inclusão de atributos desnecessários

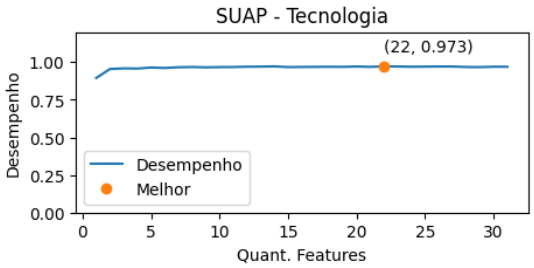
que poderiam afetar negativamente o desempenho do modelo. Na Figura 16 são apresentados os resultados obtidos com a variação da quantidade de atributos.

Figura 16 – Número de Atributos dados do SUAP

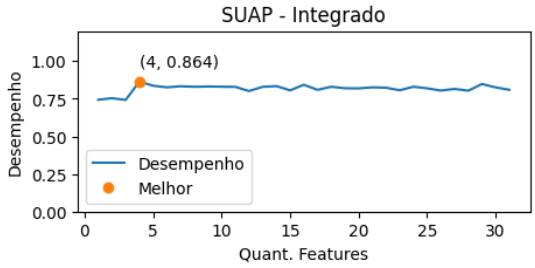
(a) Floresta Aleatória com *Wrappers* GB



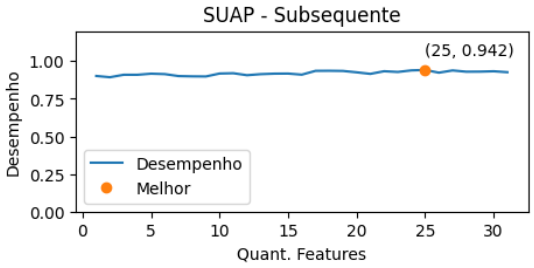
(b) Floresta Aleatória com *Embedded*



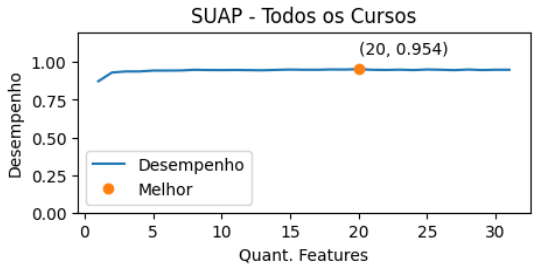
(c) Floresta Aleatória com *Chi2*



(d) Floresta Aleatória com *Chi2*



(e) Floresta Aleatória com *Embedded*



Fonte: Dados do Autor.

Os seletores e a quantidade de atributos que apresentaram os melhores resultados para cada grupo foram:

- **Bacharelado**: *Wrappers GB* com 11 atributos;
- **Tecnologia**: *Embedded* com 22 atributos;
- **Integrado**: *Chi2* com 4 atributos;
- **Subsequente**: *Chi2* com 25 atributos;
- **Todos os Cursos**: *Embedded* com 20 atributos.

Após a realização dos testes, procedeu-se ao cálculo dos valores do RMSE utilizando a métrica *F1-Score* como referência. Através dessa análise, foi possível determinar a quantidade

ideal de atributos fornecidos por um único seletor, otimizando o desempenho do modelo em todos os grupos do conjunto de dados do SUAP. Considerando todas as modalidades de curso, o seletor *Wrappers GB* apresentou os melhores resultados ao utilizar 4 atributos, sendo selecionado como a solução única de seletor e quantidade de atributos para as próximas etapas do processo de modelagem.

As tabelas Tabela 20 e Tabela 21 apresentam os atributos selecionados em cada grupo "*curso_modalidade*", ordenados por relevância, conforme determinado pelo seletor *Wrappers GB*.

Tabela 20 – Resultado da Seleção de Features pela Modalidade de Curso no SUAP - Parte 1/2

#	Todos os Cursos	Bacharelado	Tecnologia
1	escola_anterior_nome	escola_anterior_nome	escola_anterior_nome
2	curso_codigo_matriz	aluno_ano_ingresso	aluno_ano_ingresso
3	escola_anterior_cidade	aluno_cota_sistec	escola_anterior_cidade
4	aluno_ano_ingresso	aluno_cota_mec	escola_anterior_nivel_ ensino
5	escola_anterior_ano_ conclusao	curso_descricao_matriz	num_origem
6	escola_anterior_nivel_ ensino	pessoal_estado_civil	escola_anterior_ano_ conclusao
7	curso_descricao	num_origem	endereco_cidade
8	aluno_periodo_ingresso	escola_anterior_cidade	aluno_cota_sistec
9	aluno_cota_sistec	escola_anterior_nivel_ ensino	curso_codigo_matriz
10	num_origem	pessoal_cor_raca	aluno_periodo_ingresso
11	curso_descricao_matriz	escola_anterior_ano_ conclusao	aluno_forma_ingresso
12	pessoal_naturalidade	aluno_forma_ingresso	pessoal_idade
13	curso_nivel_ensino	pessoal_naturalidade	curso_campus
14	aluno_forma_ingresso	pessoal_idade	pessoal_naturalidade
15	pessoal_idade	endereco_cidade	pessoal_cor_raca
16	aluno_cota_mec	curso_codigo_matriz	endereco_zona_residencial
17	endereco_cidade	aluno_periodo_ingresso	curso_descricao_matriz
18	curso_codigo	endereco_zona_residencial	escola_anterior_tipo
19	pessoal_cor_raca	pessoal_sexo	curso_codigo
20	aluno_turno	aluno_programa	aluno_cota_mec
21	curso_campus	ano	curso_descricao
22	curso_modalidade	curso_modalidade	aluno_turno

#	Todos os Cursos	Bacharelado	Tecnologia
23	endereco_zona_residencial	curso_nivel_ensino	aluno_programa
24	escola_anterior_tipo	aluno_linha_pesquisa	pessoal_sexo
25	pessoal_estado_civil	escola_anterior_tipo	curso_modalidade
26	pessoal_sexo	curso_campus	curso_nivel_ensino
27	aluno_polo	aluno_turno	aluno_linha_pesquisa
28	aluno_linha_pesquisa	curso_descricao	aluno_polo
29	pessoal_nacionalidade	aluno_polo	pessoal_estado_civil
30	ano	pessoal_nacionalidade	ano
31	aluno_programa	curso_codigo	pessoal_nacionalidade

Fonte: Dados do Autor.

Tabela 21 – Resultado da Seleção de Features pela Modalidade de Curso no SUAP - Parte 2/2

#	Integrado	Subsequente
1	escola_anterior_nome	escola_anterior_nome
2	escola_anterior_ano_conclusao	escola_anterior_cidade
3	aluno_ano_ingresso	aluno_ano_ingresso
4	aluno_cota_mec	num_origem
5	escola_anterior_cidade	pessoal_idade
6	pessoal_naturalidade	aluno_forma_ingresso
7	escola_anterior_tipo	aluno_periodo_ingresso
8	num_origem	escola_anterior_ano_conclusao
9	aluno_cota_sistec	endereco_cidade
10	pessoal_idade	pessoal_naturalidade
11	endereco_cidade	aluno_cota_mec
12	aluno_turno	escola_anterior_tipo
13	pessoal_cor_raca	curso_codigo_matriz
14	pessoal_sexo	pessoal_cor_raca
15	aluno_forma_ingresso	pessoal_estado_civil
16	endereco_zona_residencial	aluno_cota_sistec
17	escola_anterior_nivel_ensino	curso_descricao
18	curso_descricao_matriz	pessoal_sexo
19	curso_codigo_matriz	ano
20	aluno_programa	curso_modalidade
21	ano	endereco_zona_residencial
22	curso_modalidade	curso_nivel_ensino
23	pessoal_nacionalidade	curso_campus

#	Integrado	Subsequente
24	curso_campus	escola_anterior_nivel_ensino
25	curso_nivel_ensino	aluno_turno
26	curso_descricao	aluno_polo
27	aluno_linha_pesquisa	pessoal_nacionalidade
28	aluno_polo	curso_descricao_matriz
29	pessoal_estado_civil	aluno_linha_pesquisa
30	aluno_periodo_ingresso	aluno_programa
31	curso_codigo	curso_codigo

Fonte: Dados do Autor.

A análise comparativa dos atributos mais relevantes e menos relevantes, considerando a modalidade de curso no contexto do SUAP, oferece *insights* para compreender as particularidades da evasão escolar em diferentes cenários acadêmicos.

Os atributos mais relevantes que emergiram dessa análise, tais como “escola_anterior_nome", “curso_codigo_matriz", “aluno_ano_ingresso” e “escola_anterior_cidade", sugerem que a origem dos alunos, suas trajetórias educacionais prévias e os primeiros anos de estudo têm um impacto significativo na evasão. Essa descoberta pode indicar que o suporte específico a alunos que ingressam com características particulares ou de determinadas escolas anteriores pode ser crucial para melhorar a retenção. As características dos atributos mais relevantes também variam com a modalidade do curso. No caso dos bacharelados, por exemplo, o atributo “aluno_cota_mec” torna-se relevante, sugerindo a importância de abordar questões de inclusão e equidade para essa modalidade. Já para cursos de tecnologia, “num_origem” emerge como relevante, indicando a relevância do local de origem dos alunos.

Por outro lado, os atributos menos relevantes, como “aluno_programa", “ano” e “aluno_polo", parecem ter menos influência na evasão. Isso sugere que fatores como programa de estudo e ano de ingresso, que podem não ser diretamente ligados à evasão, não devem ser priorizados nas estratégias de prevenção. Ao se concentrar em atributos relevantes, o processo de tomada de decisão pode se tornar mais eficaz e direcionado. Por exemplo, programas de suporte acadêmico e psicossocial poderiam ser desenvolvidos com base em dados como a origem dos alunos e suas escolas anteriores, visando melhorar a adaptação e a permanência dos estudantes.

Em resumo, a análise comparativa dos atributos mais e menos relevantes ressalta a importância de adotar uma abordagem mais personalizada na prevenção da evasão escolar. Isso envolve identificar as características únicas de cada modalidade de curso e os fatores específicos que influenciam a trajetória dos alunos. Ao entender os determinantes da evasão de forma mais detalhada, as instituições de ensino podem tomar medidas mais eficazes para melhorar a retenção e o sucesso dos estudantes.

5.3 Resultados dos Classificadores

Serão apresentados os resultados dos modelos de predição utilizando os algoritmos de classificação selecionados: DT, RF, NB, MLP e SVM. Serão discutidas o desempenho obtido, como a métrica *F1-Score*, para cada modelo. Será identificado o modelo com melhor desempenho na predição de evasão escolar no contexto do IFPB.

5.3.1 Resultados dos Classificadores da PNP

Nesta seção são apresentados os resultados do trabalho realizado no processo de predição da evasão escolar no IFPB utilizando os dados da PNP. O objetivo foi identificar o algoritmo de classificação, ferramenta de seleção de atributos e a quantidade ideal de atributos que possa ser utilizado para cada tipo de curso oferecido pela instituição, além de identificar uma solução única de seletor de atributos, quantidade de atributos e algoritmo de classificação, que pode ser utilizado em todos os agrupamentos de dados obtendo o melhor desempenho.

Inicialmente, os dados foram agrupados com base no campo "*tipo_curso*", obtendo os seguintes grupos: Bacharelado, Especialização, Licenciatura, Mestrado, Qualificação Profissional (FIC), Tecnologia e um grupo que engloba todos os cursos. Em seguida, o processo de predição foi repetido para cada grupo, utilizando diferentes seletores de atributos: *Chi2*, *Embedded*, *KBest*, *Wrappers GB* e *Wrappers LR*. Além disso, a quantidade de atributos selecionados foi incrementada, testando desde 1 até o máximo de 31 atributos selecionados pelos respectivos seletores.

Repetido esse processo com os oito grupos a partir do campo "*tipo_curso*", cinco algoritmos de classificação, cinco seletores de atributos e 31 conjuntos de atributos. Ao todo, foram realizados 6.200 (seis mil e duzentos) treinamentos e testes. Foi realizado o cálculo do desempenho do classificador utilizando a métrica *F1 Score* em uma validação cruzada (*Cross-Validation*) para cada grupo. O cálculo RSME foi empregado para avaliar o melhor desempenho na utilização de um único preditor, seletor e quantidade de campos.

Como exemplo, considerando que o algoritmo de classificação Árvore de Decisão, utilizando a ferramenta de seleção de atributos *Chi2*, 15 atributos e os grupos por "*tipo_curso*" "Todos os Cursos", "Bacharelado", "Especialização", "Licenciatura", "Mestrado", "Qualificação Profissional (FIC)", "Tecnologia" e "Técnico", retornaram os seguintes desempenhos: 0,78; 0,92; 0,57; 0,91; 0,56; 0,56; 0,88 e 0,65, respectivamente. Utilizando o cálculo do RSME apresentado na Equação 8, onde, d é o desempenho do classificador do grupo e n é a quantidade de grupos de "*tipo_curso*", teremos o RMSE de 0,76.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(1-d_i)^2}{n}} \tag{8}$$

Esse cálculo é repetido para todos os algoritmos de classificação, seletores de atributos

e quantidade de atributos. Os menores valores do RSME para cada conjunto de algoritmos de classificação (Preditor) e ferramenta de seleção de atributos (Seletor) foram selecionados. Na Tabela 22 são apresentados esses valores, bem como a quantidade de campos utilizada.

Tabela 22 – Resultado RSME dos Classificadores da PNP

Preditor	**Seletor**	**Quant. Campos**	**RSME**
DT	Chi2	15	0,76
	Embedded	5	0,74
	KBest	21	0,75
	Wrappers GB	10	0,72
	Wrappers LR	5	0,69
MLP	Chi2	15	0,75
	Embedded	**8**	**0,60**
	KBest	13	0,70
	Wrappers GB	13	0,63
	Wrappers LR	12	0,66
NB	Chi2	11	1,04
	Embedded	8	0,89
	KBest	11	1,00
	Wrappers GB	4	0,85
	Wrappers LR	9	0,89
RF	Chi2	9	0,80
	Embedded	4	0,70
	KBest	7	0,79
	Wrappers GB	7	0,68
	Wrappers LR	5	0,71
SVM	Chi2	15	0,83
	Embedded	16	0,81
	KBest	13	0,83
	Wrappers GB	22	0,81
	Wrappers LR	10	0,80

Fonte: Dados do Autor.

Comparando os resultados da Tabela 22, é possível identificar a configuração que apresenta o menor valor RSME, que utiliza o algoritmo de classificação **MLP** com **oito** atributos, selecionados pela ferramenta de seleção ***Embedded***. Essa solução única não corresponde necessariamente ao melhor resultado de desempenho, apenas à configuração que apresentou o menor erro comparando todos os conjuntos analisados. Na Tabela 23 é possível ver a comparação entre o desempenho da solução selecionada e o melhor desempenho encontrado nas configurações testadas.

É possível notar que, eventualmente a configuração da solução pode coincidir com o melhor resultado em algum grupo. Esse fato ocorreu nos grupos “Licenciatura” e “Mestrado” que obtiveram, respectivamente, os desempenhos 0,95 e 0,84. Ao analisar os resultados, observa-se que alguns grupos apresentaram resultados excelentes, com altos valores de métricas de

Tabela 23 – Comparação com os Melhores Resultados da PNP

Curso	Desempenho da Solução Única	Melhores Resultados			
		Classificador	Seletor	Q.C.	Desempenho
Bacharelado	0,93	DT	*Wrappers* GB	2	0,94
Especialização	0,58	NB	*Wrappers* LR	17	0,67
Licenciatura	0,95	MLP	*Embedded*	8	0,95
Mestrado	0,84	MLP	*Embedded*	8	0,844
Q.P. (FIC)	0,59	SVM	*Embedded*	2	0,73
Tecnologia	0,90	RF	*Wrappers* GB	9	0,90
Técnico	0,71	RF	*Wrappers* GB	6	0,75
Todos os Cursos	0,80	RF	*Wrappers* LR	5	0,81

Fonte: Dados do Autor.

desempenho. Outros grupos apresentaram resultados inferiores, com métricas de desempenho mais baixas. Isso pode indicar a complexidade e a variabilidade dos dados, exigindo uma abordagem mais personalizada e adaptada.

Os grupos que já apresentavam desempenho baixo, produziram resultados bem inferiores quando utilizada a solução única (MLP, oito atributos e *Embedded*), como os casos dos grupos Especialização (0,58) e FIC (0,59). Nestes casos os resultados podem comprometer a eficiência das predições. Os melhores desempenhos dos grupos foram apresentados por Licenciatura (0,94) e Bacharelado (0,93) , com resultado acima de 0,90. Na sequência um pouco abaixo estão Tecnologia (0,90), Mestrado (0,84) e a predição de todos os cursos (0,80).

Através da repetição dos processos de seleção de algoritmos, seleção de atributos e determinação da quantidade ideal de atributos, foi obtido um conjunto robusto de resultados que permite identificar as melhores abordagens para cada grupo de dados. Os resultados obtidos demonstram a importância de uma abordagem criteriosa na escolha dos componentes do modelo preditivo.

5.3.2 Resultados dos Classificadores do SUAP

Nesta seção são apresentados os resultados do trabalho realizado no processo de predição da evasão escolar no IFPB, utilizando os dados do SUAP. O objetivo foi identificar o algoritmo de classificação, ferramenta de seleção de atributos e a quantidade ideal de atributos que possam ser utilizados para cada tipo de curso oferecido pela instituição, além de identificar uma solução única de seletor de atributos, quantidade de atributos e algoritmo de classificação que podem ser utilizados em todos os agrupamentos de dados, obtendo o melhor desempenho.

Inicialmente, os dados foram agrupados com base no campo "*curso_modalidade*", obtendo os seguintes grupos: Bacharelado, Tecnologia, Integrado, Subsequente e um grupo que engloba todos os cursos. Em seguida, o processo de predição para cada grupo foi repetido, utilizando diferentes seletores de atributos: *Chi2*, *Embedded*, *KBest*, *Wrappers GB* e *Wrappers*

LR. Além disso, a quantidade de atributos selecionados foi incrementada, testando desde 1 até o máximo de 31 atributos selecionados pelos respectivos seletores. Repetido esse processo com os cinco grupos formados a apartir do campo "*curso_modalidade*", cinco algoritmos de classificação, cinco seletores de atributos e 31 conjuntos de atributos. Ao todo, foram realizados 3.875 (três mil, oitocentos e setenta e cinco) treinamentos e testes.

Foi realizado o cálculo do desempenho do classificador utilizando a métrica *F1 Score* em uma validação cruzada (*Cross-Validation*) para cada grupo. O cálculo RSME foi empregado para avaliar o melhor desempenho na utilização de um único preditor, seletor e quantidade de campos. Esse cálculo é repetido para todos os algoritmos de classificação, seletores de atributos e quantidade de atributos.

Os menores valores do RSME foram selecionados para cada conjunto de algoritmos de classificação (Preditor) e ferramenta de seleção de atributos (Seletor). Na Tabela 24 são apresentados esses valores, bem como a quantidade de campos utilizada.

Tabela 24 – Resultado RSME dos Classificadores da SUAP

Preditor	**Seletor**	**Quant. Campos**	**RSME**
DT	*Chi2*	29	0,18
	Embedded	16	0,18
	KBest	5	0,17
	Wrappers GB	5	0,18
	Wrappers LR	9	0,17
MLP	*Chi2*	11	0,17
	Embedded	8	0,17
	KBest	5	0,17
	Wrappers GB	8	0,17
	Wrappers LR	14	0,16
NB	*Chi2*	31	0,47
	Embedded	4	0,44
	KBest	3	0,43
	Wrappers GB	6	0,42
	Wrappers LR	4	0,36
RF	***Chi2***	**29**	**0,14**
	Embedded	21	0,14
	KBest	29	0,14
	Wrappers GB	23	0,14
	Wrappers LR	29	0,14
SVM	*Chi2*	25	0,21
	Embedded	16	0,20
	KBest	26	0,21
	Wrappers GB	17	0,20
	Wrappers LR	13	0,20

Fonte: Dados do Autor.

Comparando os resultados da Tabela 24, é possível identificar a configuração que apre-

senta o menor valor de RSME. A configuração que possui o menor erro é a que utiliza o algoritmo de classificação **Floresta Aleatória** com **29** atributos, selecionados pela ferramenta de seleção ***Chi2***. Essa solução não corresponde necessariamente ao melhor resultado de desempenho, apenas à configuração que apresentou o menor erro comparando todos os conjuntos analisados. Na Tabela 25 é possível ver a comparação entre o desempenho da solução selecionada e o melhor desempenho encontrado nas configurações testadas.

Tabela 25 – Comparação com os Melhores Resultados da SUAP

Curso	**Desempenho da Solução Única**	**Melhores Resultados**			
		Classificador	**Seletor**	**Q.C.**	**Desempenho**
Bacharelado	0,99	RF	*Wrappers* GB	11	0,99
Tecnologia	0,97	RF	*Embedded*	22	0,97
Integrado	0,85	RF	*Chi2*	4	0,86
Subsequente	0,93	RF	*Chi2*	25	0,94
Todos os Cursos	0,95	RF	*Embedded*	20	0,95

Fonte: Dados do Autor.

Os melhores desempenhos foram Bacharelado (0,99), Tecnologia (0,97), todos os cursos juntos (0,95) e Subsequente (0,93), com resultado acima de 0,93. Mesmo o curso Integrado (0,85) que apresentou o pior resultado está acima de 0,84. Os classificadores utilizando o conjunto de dados do SUAP apresentaram excelentes resultados, com desempenho superior aos da PNP. Esses resultados promissores representam um avanço significativo na prevenção da evasão escolar no IFPB. Com a capacidade de construir classificadores mais robustos e precisos, podemos direcionar esforços e recursos para o apoio personalizado aos alunos que mais precisam, intervindo precocemente e aumentando as chances de sucesso acadêmico.

5.4 Discussão dos Resultados

A análise dos dados da PNP e dos preditores de evasão escolar proporcionou compreensões valiosas sobre o fenômeno da evasão no IFPB. Através dos resultados obtidos, observa-se que diferentes grupos de cursos apresentaram comportamentos distintos em relação à evasão, demonstrando a importância de considerar as características específicas de cada modalidade de curso, ao desenvolver estratégias de prevenção e intervenção.

Alguns grupos de cursos obtiveram resultados com altos valores de métricas de desempenho, como: Licenciatura (0,95), Bacharelado (0,93), Tecnologia (0,90) e Mestrado (0,84). Esses valores indicam que os preditores utilizados foram capazes de identificar de forma eficaz os alunos em risco de evasão. Isso oferece oportunidades significativas para a criação de análises mais precisas e direcionadas, permitindo uma intervenção mais efetiva e personalizada. No entanto, é necessário mencionar que alguns grupos de cursos apresentaram resultados inferiores com métricas de desempenho mais baixas, como: FIC (0,59) e Especialização (0,58). Isso pode

ser atribuído a diversas razões, como a complexidade dos dados, a influência de fatores externos não considerados nos preditores, ou a falta de correlação clara entre as variáveis analisadas e a evasão. Essas discrepâncias ressaltam a importância de uma abordagem cuidadosa e contínua na análise da evasão escolar, buscando aprimorar constantemente os modelos e preditores utilizados.

Apesar das variações nos resultados, a análise dos dados da PNP fornece uma base sólida para a compreensão da evasão escolar no IFPB. Esses resultados podem servir como ponto de partida para a implementação de ações preventivas e estratégias de retenção de alunos, visando melhorar os índices de conclusão e sucesso acadêmico. É fundamental que essas análises sejam combinadas com outras informações e conhecimentos contextuais para uma compreensão abrangente e precisa da evasão escolar.

Em suma, a análise dos dados da PNP sobre a evasão escolar no IFPB oferece informações importantes que podem direcionar ações e políticas educacionais mais eficientes. Ao considerar as particularidades de cada grupo de cursos, é possível desenvolver intervenções mais direcionadas e adaptadas, visando melhorar a retenção dos alunos e promover uma educação de qualidade. No entanto, é necessário continuar aprimorando os modelos de análise e preditores utilizados, bem como considerar outras fontes de dados e informações para obter uma visão abrangente do problema da evasão escolar.

A análise dos dados do SUAP referentes à evasão escolar no IFPB revelou resultados superiores até mesmo aos obtidos com os dados da PNP. Esses resultados promissores abrem novas perspectivas para a construção de classificadores mais robustos e eficazes na identificação de alunos em risco de evasão. Diferentes grupos de cursos do SUAP foram investigados, e em todos eles, os classificadores obtiveram desempenho notável, como: Bacharelado (0,99), Tecnologia (0,97), Integrado (0,85) e Subsequente (0,93). Tendo valores elevados de métricas de avaliação, *F1-Score*, e um baixo valor do RMSE (0,14) para utilização de uma solução única. Esses resultados destacam a capacidade dos preditores selecionados em identificar padrões relevantes nos dados e fornecer previsões precisas sobre a evasão escolar. A utilização de diferentes seletores de atributos e quantidades de atributos permitiu explorar diversas combinações, resultando em um processo de seleção otimizado. A análise minuciosa dos dados, combinada com os algoritmos de classificação apropriados, possibilitou obter modelos preditivos mais eficientes, capazes de identificar com precisão os alunos com maior probabilidade de evasão.

Ao compararmos este estudo com os principais trabalhos pesquisados na revisão sistemática da literatura, notamos que ele oferece uma perspectiva abrangente e aprofundada sobre a análise da evasão escolar. Enquanto cada trabalho anterior abordou aspectos específicos desse desafio complexo, esta dissertação unifica e amplia essas abordagens, resultando em uma contribuição significativa para o campo. O estudo conduzido por Ma et al. (2017) teve como foco a identificação de atributos influentes para o desempenho em cursos online, utilizando seleção de atributos e vários algoritmos de classificação. Embora tenham obtido melhorias na acurácia, exceto para o algoritmo BN, este trabalho transcende essa abordagem, incorporando agrupamen-

tos específicos de cursos e características dos alunos. Essa estratégia permite identificar padrões distintos de evasão em diferentes modalidades, o que pode informar estratégias mais eficazes de prevenção. Silva et al. (2019) centraram-se na análise da evasão em instituições de ensino superior brasileiras, utilizando indicadores educacionais. Esta dissertação se alinha a essa abordagem, mas expande as possibilidades de análise, investigando vários agrupamentos de cursos e atributos dos alunos, visando uma visão mais completa. Enquanto Silva et al. se concentraram na limpeza de dados e em algoritmos RF e LR, esta pesquisa avança ao incorporar uma seleção otimizada de atributos e diversos algoritmos de classificação, visando a generalização dos resultados.

O trabalho de Nandeshwar, Menzies e Nelson (2011) enfocou a retenção de alunos em universidades americanas, destacando a relevância dos atributos ligados à ajuda financeira. Enquanto esse foco é valioso, esta dissertação amplia o horizonte, explorando não apenas os atributos financeiros, mas também características demográficas, acadêmicas e outras, em diferentes grupos de cursos. Além disso, enquanto eles aplicaram DT, NB e BN, esta pesquisa investigou diversos algoritmos de classificação, visando a eficácia em contextos variados. O estudo de Regha e Rani (2015) apresentou uma técnica para seleção de atributos, enfatizando a redução de atributos irrelevantes e redundantes. Similarmente, esta dissertação também considerou a seleção otimizada de atributos, mas expandiu o escopo, aplicando diferentes técnicas de seleção e diversos algoritmos de classificação em um contexto mais diversificado. O estudo de Urbina-Najera, Camino-Hampshire e Barbosa (2020) buscou identificar causas de evasão em Instituições de Ensino Superior no México, utilizando DT e o método *Wrapper*. Enquanto esta dissertação segue esse caminho, ela vai além, considerando diferentes grupos de cursos e atributos dos alunos, resultando em uma análise mais refinada e detalhada.

Este trabalho de dissertação se alinha com os achados dos estudos revisados na literatura sobre a evasão escolar, identificando a importância e a relevância das pesquisas já realizadas nesse campo. Entretanto, também reconhece a oportunidade de expandir e aprimorar abordagens para compreender mais profundamente esse problema complexo. No contexto das contribuições desses trabalhos, esta dissertação se destaca por trazer uma abordagem abrangente na análise da evasão escolar no âmbito do IFPB. Enquanto muitos estudos focam em análises restrita, este trabalho busca uma compreensão mais abrangente, explorando diversas perspectivas, como diferentes agrupamentos de cursos, algortimos e seletores de características dos alunos.

Uma das principais distinções deste estudo é a exploração da relação entre evasão e diferentes categorias de cursos, como Bacharelado, Especialização, Licenciatura, Mestrado, Qualificação Profissional (FIC) e Tecnologia. Ao investigar padrões específicos de evasão em cada modalidade, o trabalho vai além das análises generalizadas, permitindo a formulação de estratégias mais direcionadas para mitigação desse fenômeno. Outro aspecto relevante é a busca por uma solução única e aplicável a diferentes grupos de cursos, tanto na base de dados da PNP quanto no SUAP. Enquanto outros estudos podem ter se concentrado em análises específicas para cada conjunto de dados, esta dissertação procura encontrar um modelo que seja consistente

e promissor em múltiplos contextos, visando à aplicação prática.

Essa pesquisa representa, portanto, uma contribuição significativa na análise da evasão escolar. Ao considerar várias dimensões, incluindo diferentes agrupamentos de cursos e atributos dos alunos, este estudo oferece uma compreensão mais completa e detalhada desse problema. Os resultados obtidos fornecem uma perspectiva sobre os fatores que influenciam a evasão, propiciando a formulação de estratégias para prevenção e intervenção. Adicionalmente, a busca por uma solução unificada que abrange múltiplos conjuntos de dados expande a relevância prática deste estudo, abrindo portas para futuras investigações e aplicações na área da evasão escolar. Esses resultados proporcionam uma base para a implementação de ações preventivas e estratégias de manutenção dos alunos no IFPB. Com base nas informações obtidas, é possível direcionar esforços e recursos de maneira mais efetiva, visando reduzir a evasão e aumentar os índices de conclusão e sucesso acadêmico.

Dessa forma, este estudo se destaca por sua abordagem, ao explorar agrupamentos específicos de cursos e atributos, além da seleção otimizada de atributos e diversos algoritmos de classificação, esta dissertação enriquece a compreensão e abordagem da evasão escolar. Em resumo, os resultados obtidos com os dados do SUAP indicam que é possível construir classificadores eficazes para prever a evasão escolar no IFPB. Essa análise oferece uma visão clara do problema e fornece subsídios para a implementação de medidas assertivas e personalizadas. A partir desses resultados promissores, o IFPB pode adotar estratégias mais eficientes para reduzir a evasão e promover uma educação de qualidade, garantindo o sucesso acadêmico e a formação dos seus estudantes.

6 CONCLUSÃO

Neste trabalho, foi realizada uma análise abrangente dos dados de evasão escolar no IFPB, utilizando as bases de dados da PNP e do SUAP. Foram aplicadas técnicas de seleção de atributos, avaliação de algoritmos de classificação e predição da evasão escolar em diferentes grupos de cursos. Os resultados obtidos demonstraram a eficácia da abordagem adotada, permitindo a identificação de padrões e a construção de modelos preditivos precisos. Tanto os dados da PNP, quanto os do SUAP, revelaram informações valiosas sobre a evasão escolar, possibilitando a compreensão dos fatores que influenciam esse fenômeno e fornecendo compreensões importantes para a tomada de decisões.

Foi realizada uma revisão sistemática da literatura para obter um panorama atualizado das técnicas de mineração de dados aplicadas à predição de evasão escolar. Os dados educacionais do IFPB, incluindo a base de dados do módulo de controle acadêmico do SUAP e os dados disponíveis no portal de dados abertos da Plataforma Nilo Peçanha, foram coletados e pré-processados. As análises quantitativas dos dados coletados foram conduzidas para identificar padrões e tendências relacionados à evasão escolar no contexto do IFPB.

Foram aplicadas técnicas de seleção de atributos para identificar os mais relevantes na predição de evasão escolar, com base nos dados educacionais analisados. Algoritmos de classificação, como DT, RF, NB, MLP e SVM, foram implementados e o desempenho desses modelos na predição de evasão escolar foi avaliado. A métrica *F1-Score* foi utilizada para avaliar a precisão e a eficácia dos modelos propostos.

No caso dos dados da PNP, observou-se que alguns grupos apresentaram resultados excelentes, acima de 0,94, enquanto outros obtiveram desempenho inferior. No entanto, mesmo com essas variações, é possível desenvolver análises satisfatórias para tratar do tema da evasão escolar, direcionando estratégias preventivas e de retenção de alunos de forma mais efetiva.

Já os dados do SUAP demonstraram resultados ainda mais promissores, superando os obtidos com a base da PNP. Os classificadores utilizados nessa análise mostraram-se eficientes na identificação de alunos em risco de evasão, possibilitando a implementação de ações preventivas e intervenções personalizadas. Esses resultados destacam a importância de considerar as particularidades de cada grupo de cursos ao desenvolver estratégias de prevenção e retenção.

Em conclusão, este trabalho evidenciou a relevância da análise de dados e da aplicação de técnicas de predição na área da educação, especificamente no contexto da evasão escolar. A utilização de algoritmos de classificação, seleção de atributos e quantidades ideais de atributos possibilitou a construção de modelos preditivos precisos, capazes de auxiliar na tomada de decisões e no desenvolvimento de estratégias para combater a evasão e promover o sucesso acadêmico. As descobertas deste estudo têm o potencial de trazer impactos positivos para o IFPB.

Os resultados obtidos permitem que a instituição adote medidas preventivas e eficazes visando a redução da evasão e o aumento das taxas de conclusão dos cursos. Além disso, as conclusões alcançadas contribuem para o avanço do campo de análise de dados educacionais, fornecendo *insights* valiosos, tanto para pesquisadores, como para profissionais da área.

É importante ressaltar que a evasão escolar é um desafio complexo e multifacetado, envolvendo uma série de fatores socioeconômicos, pessoais e institucionais. Portanto, o uso de abordagens analíticas, como as realizadas neste estudo, é fundamental para compreender e enfrentar esse problema de maneira efetiva. Dessa forma, a análise de dados e os modelos preditivos desenvolvidos neste trabalho têm o potencial de contribuir significativamente para a melhoria do ensino e aprendizagem no IFPB, proporcionando um ambiente acadêmico mais inclusivo, orientado para o sucesso dos alunos e promovendo a formação de profissionais qualificados.

6.1 Trabalhos Futuros

Existem diversas oportunidades para futuros trabalhos relacionados à análise de dados e predição da evasão escolar no IFPB. Com base nos resultados e nas descobertas deste estudo, algumas direções para pesquisas futuras podem ser exploradas. Primeiramente, é possível analisar padrões de evasão por período, investigando se existem diferenças significativas nas taxas de evasão entre os períodos letivos (semestres, trimestres, etc.) e identificar possíveis padrões sazonais ou flutuações que possam afetar a permanência dos alunos. Além disso, é interessante considerar a incorporação de dados adicionais, como informações socioeconômicas e dados de desempenho acadêmico dos alunos, para obter uma visão mais completa dos fatores que influenciam a evasão escolar. Outra perspectiva de pesquisa é a análise de tendências temporais, investigando possíveis mudanças nas taxas de evasão ao longo dos anos e compreendendo o impacto de políticas e intervenções implementadas. Além disso, é relevante realizar análises específicas nos dados do SUAP para os demais cursos do IFPB, levando em conta toda a base de dados disponível no SUAP. Por fim, é fundamental avaliar a eficácia das intervenções implementadas para reduzir a evasão escolar, analisando os resultados de estratégias de retenção de alunos e contribuindo para a tomada de decisões institucionais. Em suma, a investigação contínua nessa área pode fornecer ferramentas valiosas, direcionadas a promover o sucesso acadêmico dos estudantes do IFPB.

REFERÊNCIAS

AQUINO, J. G. O mal-estar na escola contemporânea: erro e fracasso em questão. In: ____. São Paulo/SP: Summus, 1997. Citado na página 17.

ARTERO, A. O. *Inteligência Artificial: teórica e prática*. São Paulo: Livraria da Física, 2009. ISBN 8578610296. Citado 3 vezes nas páginas 29, 37 e 39.

CASTRO, L. D.; FERRARI, D. *Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações*. [S.l.: s.n.], 2016. ISBN 9788547200985. Citado na página 26.

CHANGO, W.; CEREZO, R.; ROMERO, C. Multi-source and multimodal data fusion for predicting academic performance in blended learning university courses. *Comput. Electr. Eng.*, v. 89, p. 106908, 2021. Disponível em: <https://doi.org/10.1016/j.compeleceng.2020.106908>. Citado 4 vezes nas páginas 50, 51, 52 e 53.

CHANLEKHA, H.; NIRAMITRANON, J. Student performance prediction model for early-identification of at-risk students in traditional classroom settings. In: CHBEIR, R. et al. (Ed.). *Proceedings of the 10th International Conference on Management of Digital EcoSystems, MEDES 2018, Tokyo, Japan, September 25-28, 2018*. ACM, 2018. p. 239–245. Disponível em: <https://doi.org/10.1145/3281375.3281403>. Citado 4 vezes nas páginas 50, 51, 52 e 53.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado na página 34.

COSTA, E. et al. Mineração de dados educacionais: Conceitos, técnicas, ferramentas e aplicações. In: . [S.l.: s.n.], 2012. v. 1, n. 1, p. 1–29. Citado 2 vezes nas páginas 23 e 27.

COSTA, E. de B. et al. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Comput. Hum. Behav.*, v. 73, p. 247–256, 2017. Disponível em: <https://doi.org/10.1016/j.chb.2017.01.047>. Citado 4 vezes nas páginas 50, 51, 52 e 53.

DWAN, F.; OLIVEIRA, E.; FERNANDES, D. Predição de zona de aprendizagem de alunos de introdução à programação em ambientes de correção automática de código. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2017. v. 28, n. 1, p. 1507. Citado 4 vezes nas páginas 50, 51, 52 e 53.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, v. 17, n. 3, p. 37, Mar. 1996. Disponível em: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>. Citado 3 vezes nas páginas 23, 24 e 27.

FERREIRA, J. T. A. et al. O processo etl em sistemas data warehouse. In: BARBOSA, L.; CORREIA, M. P. (Ed.). *INForum 2010: actas do II Simposio de Informatica, Braga, 2010*. Braga: Universidade do Minho, 2010. ISBN 978-989-96863-0-4. Artigo em ata de conferência. Disponível em: <https://hdl.handle.net/1822/11435>. Citado 2 vezes nas páginas 25 e 26.

FILHO, R. B. S.; ARAÚJO, R. M. d. L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. *Educação Por Escrito*, v. 8, n. 1, p. 35–48, jun.

2017. Disponível em: <https://revistaseletronicas.pucrs.br/ojs/index.php/porescrito/article/view/24527>. Citado na página 23.

GAMA, J. et al. *Extração de conhecimento de dados: data mining*. 3. ed. Portugal: Edições Silabo, 2019. 436 p. Citado na página 23.

GERON, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. 2nd. ed. [S.l.]: O'Reilly Media, Inc., 2019. ISBN 1492032646. Citado 13 vezes nas páginas 28, 29, 32, 33, 35, 37, 39, 40, 41, 42, 43, 44 e 45.

GOTTARDO, E.; KAESTNER, C.; NORONHA, R. V. Previsao de desempenho de estudantes em cursos ead utilizando mineraçao de dados: uma estratégia baseada em séries temporais. In: *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*. [S.l.: s.n.], 2012. v. 23, n. 1. Citado na página 23.

GRUS, J. *Data Science do Zero*. Rio de Janeiro: Alta Books, 2016. Citado 5 vezes nas páginas 36, 38, 39, 40 e 42.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3rd. ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011. ISBN 0123814790. Citado na página 17.

HE, H.; GARCIA, E. A. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, Ieee, v. 21, n. 9, p. 1263–1284, 2009. Citado 2 vezes nas páginas 33 e 34.

HERSHKOVITZ, A.; NACHMIAS, R. Online persistence in higher education web-supported courses. *Internet and Higher Education*, Elsevier BV, v. 14, n. 2, p. 98–106, mar 2011. ISSN 1096-7516. Citado 4 vezes nas páginas 50, 51, 52 e 53.

HU, Y.-H.; LO, C.-L.; SHIH, S.-P. Developing early warning systems to predict students' online learning performance. *Computers in Human Behavior*, Elsevier, v. 36, p. 469–478, 2014. Citado 4 vezes nas páginas 50, 51, 52 e 53.

IFPB - Plano de Desenvolvimento Institucional 2020-2024. 2020. Disponível em: <https://suap.ifpb.edu.br/>. Acesso em: 02 Abr. de 2022. Citado na página 19.

IFRN - Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte. 2022. Portal do IFPB. Acesso em: 05 Fev. de 2022. Disponível em: <https://suap.ifpb.edu.br/>. Citado na página 45.

INEP - Instituto Nacional de Estudo e Pesquisa Educacional Anísio Teixeira. 2021. Sinopse Estatística da Educação Superior. Acesso em: 22 nov. 2021. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados>. Citado na página 17.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.]: Rafael Izbicki, 2020. Citado na página 38.

KAUR, P.; SINGH, M.; JOSAN, G. S. Classification and prediction based data mining algorithms to predict slow learners in education sector. *Procedia Computer Science*, v. 57, p. 500–508, 2015. ISSN 1877-0509. 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015). Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050915019018>. Citado na página 17.

KITCHENHAM, B. Procedures for performing systematic reviews. *Keele, UK, Keele University*, Citeseer, v. 33, n. 2004, p. 1–26, 2004. Citado na página 47.

LARA, J. A. et al. A system for knowledge discovery in e-learning environments within the european higher education area–application to student data from open university of madrid, udima. *Computers & Education*, Elsevier, v. 72, p. 23–36, 2014. Citado 4 vezes nas páginas 50, 51, 52 e 53.

LIU, H.; MOTODA, H. *Computational methods of feature selection*. [S.l.]: CRC press, 2007. Citado na página 31.

MA, C. et al. Improving prediction of student performance based on multiple feature selection approaches. In: *Proceedings of the 2017 1st International Conference on E-Education, E-Business and E-Technology*. [S.l.: s.n.], 2017. p. 36–41. Citado 5 vezes nas páginas 50, 51, 52, 53 e 55.

MAIMON, O.; ROKACH, L. *Data Mining and Knowledge Discovery Handbook*. [S.l.]: Springer Science & Business Media, 2010. v. 14. Citado na página 25.

MARR, B. *Big Data: Using SMART big data, analytics and metrics to make better decisions and improve performance*. [S.l.]: John Wiley & Sons, 2015. Citado na página 23.

MECA, I. et al. Early warning methodology for dropping out of university degrees. In: *Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality*. [S.l.: s.n.], 2020. p. 245–249. Citado 4 vezes nas páginas 50, 51, 52 e 53.

MIGUÉIS, V. L. et al. Early segmentation of students according to their academic performance: A predictive modelling approach. *Decision Support Systems*, Elsevier, v. 115, p. 36–51, 2018. Citado 4 vezes nas páginas 50, 51, 52 e 53.

NANDESHWAR, A.; MENZIES, T.; NELSON, A. Learning patterns of university student retention. *Expert Systems with Applications*, Elsevier, v. 38, n. 12, p. 14984–14996, 2011. Citado 5 vezes nas páginas 50, 51, 52, 53 e 55.

OLIVEIRA, T. et al. Escola, conhecimento e formação de pessoas: considerações históricas. *Políticas Educativas – PolEd*, v. 6, n. 2, abr. 2013. Disponível em: <https://seer.ufrgs.br/index.php/Poled/article/view/45662>. Citado na página 17.

PAZ, F.; CAZELLA, S. Identificando o perfil de evasão de alunos de graduação através da mineração de dados educacionais: um estudo de caso de uma universidade comunitária. In: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. [S.l.: s.n.], 2017. v. 6, n. 1, p. 624. Citado na página 17.

PEÑA-AYALA, A. Educational data mining. *Studies in Computational Intelligence*, Springer, v. 524, 2014. Citado na página 27.

PICHETH, F. M. *PeArte: um ambiente colaborativo para a formação do pesquisador que atua no ensino superior por meio da participação em pesquisas do tipo estado da arte*. Tese (Doutorado) — Pontifícia Universidade Católica do Paraná, 2007. Citado na página 47.

PRIM, A. L.; FÁVERO, J. D. Motivos da evasão escolar nos cursos de ensino superior de uma faculdade na cidade de blumenau. *Revista e-TECH: Tecnologias para Competitividade Industrial - ISSN - 1983-1838*, p. 53–72, dez. 2013. Disponível em: <https://etech.emnuvens.com.br/revista-cientifica/article/view/382>. Citado na página 17.

RAMASWAMI, M.; BHASKARAN, R. A chaid based performance prediction model in educational data mining. *arXiv: Learning*, Feb 2010. Citado na página 17.

REGHA, R. S.; RANI, R. U. A novel clustering based feature selection for classifying student performance. *Indian Journal of Science and Technology*, Indian Society for Education and Environment, v. 8, p. 135, 2015. Citado 5 vezes nas páginas 50, 51, 52, 53 e 56.

ROMERO, C.; VENTURA, S. Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, v. 40, n. 6, p. 601–618, 2010. Citado 2 vezes nas páginas 17 e 27.

RUSSEL, S. J.; NORVIG, P. *Inteligência Artificial. [Sl]*. [S.l.]: Rio de Janeiro: Elsevier Editora Ltda, 2013. Citado 3 vezes nas páginas 29, 36 e 38.

SIEBRA, C. A.; SANTOS, R. N.; LINO, N. C. A self-adjusting approach for temporal dropout prediction of e-learning students. *International Journal of Distance Education Technologies (IJDET)*, IGI Global, v. 18, n. 2, p. 19–33, 2020. Citado 5 vezes nas páginas 18, 50, 51, 52 e 53.

SILVA, P. M. D. et al. Ensemble regression models applied to dropout in higher education. In: IEEE. *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.], 2019. p. 120–125. Citado 5 vezes nas páginas 50, 51, 52, 53 e 55.

THAI-NGHE, N.; BUSCHE, A.; SCHMIDT-THIEME, L. Improving academic performance prediction by dealing with class imbalance. In: *Ninth International Conference on Intelligent Systems Design and Applications, ISDA 2009, Pisa, Italy , November 30-December 2, 2009*. IEEE Computer Society, 2009. p. 878–883. Disponível em: <https://doi.org/10.1109/ISDA.2009.15>. Citado na página 17.

THAMMASIRI, D. et al. A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition. *Expert Systems with Applications*, Elsevier, v. 41, n. 2, p. 321–330, 2014. Citado 4 vezes nas páginas 50, 51, 52 e 53.

URBINA-NAJERA, A.; CAMINO-HAMPSHIRE, J.; BARBOSA, R. C. University dropout: Prevention patterns through the application of educational data mining/desercion escolar universitaria: Patrones para prevenirla aplicando mineria de datos educativa. *RELIEVE: Revista Electronica de Investigacion y Evaluacion Educativa*, Interuniversity Association of Pedagogical Research, v. 26, n. 1, p. 1a–1a, 2020. Citado 5 vezes nas páginas 50, 51, 52, 53 e 56.

ZHANG, Y.; WU, B. Research and application of grade prediction model based on decision tree algorithm. In: *Proceedings of the ACM Turing Celebration Conference-China*. [S.l.: s.n.], 2019. p. 1–6. Citado 4 vezes nas páginas 50, 51, 52 e 53.