UFRRJ

INSTITUTO DE CIÊNCIAS EXATAS

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

DISSERTAÇÃO

O USO DA INTELIGÊNCIA ARTIFICIAL PARA AUXILIAR NA PREDIÇÃO DE ABANDONO DE ALUNOS EM UMA ESCOLA DE ENSINO REGULAR

Vanessa Alexandre Silva

2025

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

O USO DA INTELIGÊNCIA ARTIFICIAL PARA AUXILIAR NA PREDIÇÃO DE ABANDONO DE ALUNOS EM UMA ESCOLA DE ENSINO REGULAR

VANESSA ALEXANDRE SILVA

Sob orientação de
Marcelo Dib Cruz

e co-orientação de
Ronaldo Malheiros Gregório

Dissertação submetida como requisito parcial para obtenção do grau de **Mestre** no Programa de Pós-Graduação em Modelagem Matemática e Computacional, Área de Concentração em Modelagem Matemática e Computacional.

Seropédica, RJ, Brasil
Fevereiro de 2025

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
com os dados fornecidos pelo(a) autor(a)

S381m

Silva, Vanessa Alexandre, 1983-
MESTRE / Vanessa Alexandre Silva. - RIO DE JANEIRO, 2025.
58 f.

Orientadora: Marcelo Dib Cruz.
Coorientadora: Ronaldo Malheiros Gregório.
Dissertação(Mestrado). -- Universidade Federal Rural do Rio de Janeiro, PROGRAMA DE PÓS GRADUAÇÃO EM MODELAGEM MATEMÁTICA COMPUTACIONAL, 2025.

1. PREDIÇÃO DE ABANDONO ESCOLAR. 2. INTELIGÊNCIA ARTIFICIAL. 3. APRENDIZADO DE MAQUINA. 4. APRENDIZADO SUPERVISIONADO. 5. REDES NEURAIS. I. Cruz, Marcelo Dib, 1967-, orient. II. Gregório, Ronaldo Malheiros, 1978-, coorient. III Universidade Federal Rural do Rio de Janeiro. PROGRAMA DE PÓS GRADUAÇÃO EM MODELAGEM MATEMÁTICA COMPUTACIONAL. IV. Título.

MINISTÉRIO DA EDUCAÇÃO

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO

PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM MATEMÁTICA E COMPUTACIONAL

VANESSA ALEXANDRE SILVA

Dissertação submetida como requisito parcial para a obtenção de grau de Mestra, no Programa de Pós-Graduação em Modelagem Matemática e Computacional PPGMMC, área de Concentração em Modelagem Matemática e Computacional.

DISSERTAÇÃO APROVADA EM 25/02/2025

Marcelo Dib Cruz Drº (Orientador, Presidente da Banca-UFRRJ)

Carlos Andres Reyna Vera Tudela Drº (membro interno-UFRRJ)

Ronaldo Malheiros Gregório Drº (membro interno-UFRRJ)

Lucídio dos Anjos Formiga Cabral Drº (UFPB-Externo à Instituição)

ATA Nº ata/2025 - ICE (12.28.01.23)
(Nº do Documento: 743)

(Nº do Protocolo: NÃO PROTOCOLADO)

(Assinado digitalmente em 21/03/2025 16:31)
CARLOS ANDRES REYNA VERA TUDELA
COORDENADOR CURS/POS-GRADUACAO
PPGMMC (12.28.01.00.00.00.61)
Matrícula: ###336#3

(Assinado digitalmente em 22/03/2025 09:53)
MARCELO DIB CRUZ
PROFESSOR DO MAGISTERIO SUPERIOR
DCOMP (11.39.97)
Matrícula: ###680#1

(Assinado digitalmente em 24/03/2025 08:29)
RONALDO MALHEIROS GREGORIO
PROFESSOR DO MAGISTERIO SUPERIOR
DeptTL/IM (12.28.01.00.00.90)
Matrícula: ###696#7

(Assinado digitalmente em 23/03/2025 19:41)
LUCIDIO DOS ANJOS FORMIGA CABRAL
ASSINANTE EXTERNO
CPF: ###.###.883-##

*Visualize o documento original em https://sipac.ufrrj.br/documentos/ informando seu número: **743**, ano: **2025**, tipo: **ATA**, data de emissão: **21/03/2025** e o código de verificação: **f8212db689***

Agradecimentos

Gostaria de expressar minha gratidão aos respeitados professores do Programa de Pós-Graduação em Modelagem Matemática e Computacional da UFRRJ (PPGMMC-UFRRJ), em especial aos meus orientadores, Marcelo Dib Cruz e Ronaldo Malheiros Gregório, cuja orientação e apoio foram fundamentais para a realização deste trabalho.

Agradeço também ao meu marido, Jaime Urtado Alves, e à minha mãe, Rejane Alexandre Silva, pelo apoio incondicional e incentivo durante toda a minha jornada acadêmica. Suas palavras de encorajamento e compreensão foram essenciais para que eu pudesse superar os desafios enfrentados.

Não poderia deixar de mencionar meu colega de turma, Rodrigo Cabral de Freitas, pela parceria e colaboração.

Adicionalmente, registro meu sincero agradecimento ao apoio financeiro recebido.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

A todos, muito obrigada.

"Educação não transforma o mundo. Educação muda as pessoas. Pessoas transformam o mundo."Paulo Freire

RESUMO

ALEXANDRE SILVA, Vanessa. **O USO DA INTELIGNCIA ARTIFICIAL PARA AUXILIAR NA PREDIÇÃO DE ABANDONO DE ALUNOS EM UMA ESCOLA DE ENSINO REGULAR**. 2025. 47f. Dissertação (Mestrado em Modelagem Matemática e Computacional). Instituto de Ciências Exatas, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2025.

A evasão escolar é um desafio para os sistemas educacionais em todo o mundo[10]. Compreender e prever os fatores que levam os alunos a abandonar os seus estudos é fundamental para o desenvolvimento de estratégias eficazes de prevenção e intervenção. A Inteligência Artificial (IA), especialmente por meio de técnicas de Aprendizado de Máquina, tem a capacidade de analisar dados do passado para prever eventos futuros [1]. Entre os diversos algoritmos utilizados para essa finalidade, destacam-se as Redes Neurais Artificiais, Árvore de Decisão, K-Nearest Neighbor(KNN) e Support Vector Machine (SVM).

O objetivo deste estudo é desenvolver um modelo preditivo, utilizando vários métodos, para identificar alunos que estejam em risco de abandonar a escola. Para isso, será necessário investigar as causas subjacentes à evasão escolar. Como parte deste processo, foi elaborado um questionário direcionado aos alunos, com o intuito de coletar informações relevantes, uma vez que as escolas geralmente carecem de dados detalhados que permitam a identificação precisa desse problema.

A pesquisa foi realizada com a colaboração de estudantes da Educação de Jovens e Adultos do Novo Ensino Médio (EJANEM), que em algum momento de suas trajetórias escolares interromperam os estudos, bem como com a participação de alunos do ensino regular, tanto do ensino médio quanto do ensino fundamental. A participação desses grupos permitiu uma análise das diversas razões que podem conduzir à evasão escolar.

Palavras-chave: Evasão Escolar, Inteligência Artificial, Aprendizado de Máquina, Redes Neurais, Modelo de Classificação, Aprendizado Supervisionado.

ABSTRACT

ALEXANDRE SILVA, Vanessa. **THE USE OF ARTIFICIAL INTELLIGENCE TO ASSIST IN THE PREDICTION OF STUDENT DROPOUT IN A REGULAR EDUCATION SCHOOL**. 2025. 47p. Dissertation (Master in Mathematical and Computational Modeling). Instituto de Ciências Exatas, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2025.

School dropout is a significant challenge for educational systems worldwide. Understanding and predicting the factors that lead students to abandon their studies is essential for developing effective prevention and intervention strategies. Artificial Intelligence (AI), particularly through Machine Learning techniques, has the capability to analyze historical data to predict future events. Among the various algorithms used for this purpose, Artificial Neural Networks stand out.

The objective of this study is to develop a predictive model using Artificial Neural Networks to identify students at risk of dropping out of school. To achieve this, it is necessary to investigate the underlying causes of school dropout. As part of this process, a questionnaire will be developed for students to collect relevant information, given that schools often lack detailed data necessary for accurately identifying this issue.

The research will be conducted with the collaboration of students from the Education for Youth and Adults in the New High School (EJANEM) program, who at some point in their educational journeys interrupted their studies, as well as with the participation of regular education students, both from high school and elementary school. The involvement of these groups will enable a comprehensive and in-depth analysis of the various reasons that may lead to school dropout

Keywords: School Dropout, Artificial Intelligence, Machine Learning, Classification Model, Supervised Learning, Neural Networks..

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

A evasão escolar é um desafio que afeta os sistemas educacionais em todo o mundo [10]. Diversos fatores contribuem para este fenômeno, incluindo questões familiares, necessidade de contribuir financeiramente com as despesas da família, gravidez na adolescência e falta de perspectivas [7][8][11][12]. Compreender as causas do abandono escolar é essencial para a criação de estratégias eficazes de prevenção e intervenção [11].

Nesse cenário, a inteligência artificial (IA) tem se destacado como uma ferramenta inovadora para a solução de problemas complexos. Em particular, o aprendizado de máquina permite que sistemas aprendam com dados do passado para prever situações futuras [1]. Modelos como Redes Neurais, inspirados no funcionamento do cérebro humano, têm demonstrado alta eficácia na análise e predição de padrões [5][6]. Além disso, outros algoritmos supervisionados, como árvores de decisão, k-Nearest Neighbors (k-NN), e Máquinas de Vetores de Suporte (SVM), também oferece soluções robustas para solucionar problemas em diversas áreas do conhecimento.

Neste estudo, a inteligência artificial foi utilizada para enfrentar o desafio da evasão escolar, com o desenvolvimento de um modelo preditivo capaz de identificar alunos em risco de abandono.

Inicialmente a pesquisa foi realizada em uma escola da rede estadual de esino do Rio de Janeiro. começou utilizando dados existentes na unidade, pois a escola recentemente havia participado de uma pesquisa referente à merenda escolar, nesses dados incluía informações gerais sobre os alunos, como registros de presença, notas e características socioeconômicas. O que foi fundamental para a implementação inicial do código e para a análise preliminar dos resultados do modelo de previsão de abandono. Contudo, verificou-se que tais dados não eram suficientes para obter resultados precisos. O que levou a necessidade de uma mudança de estratégia e obteção de novos dados. Essa nova amostra, a princípio, contava com 268 formulários válidos, e foi posteriormente ampliada para 397 formulários, abrangendo escolas de diferentes municípios. Os dados obtidos foram analisados por meio de Redes Neurais e outros algoritmos supervisionados, permitindo a comparação entre métodos e a conquista de resultados mais precisos.

Com os resultados deste trabalho pretende-se que sejam oferecidos subsídios para a criação de ferramentas preditivas que auxiliem gestores educacionais na formulação de estratégias de intervenção direcionadas a estudantes em risco. Assim, este estudo não apenas pode contribuir para a redução da evasão escolar, como também pode promover um ambiente educacional mais inclusivo.

1.1 Motivação

A pesquisa visa contribuir de maneira significativa para o aprimoramento da gestão educacional, promovendo a permanência dos alunos e, consequentemente, sua progressão dentro da unidade escolar. Adicionalmente, este estudo pretende fornecer dados substanciais que poderão ser utilizados no desenvolvimento de estratégias mais eficazes para a retenção de alunos, visando a criação de um ambiente escolar mais acolhedor e inclusivo. Tais esforços buscam contribuir para o desenvolvimento de uma sociedade mais equitativa. As melhorias decorrentes beneficiam não apenas os estudantes em nível individual, mas também geram um impacto positivo no fortalecimento do sistema educacional como um todo.

1.2 Objetivo

A abordagem metodológica concentra-se na aplicação de Redes Neurais Artificiais, Árvores de Decisão, K-Nearest Neighbor (KNN) e Support Vector Machine (SVM), técnicas que permitem analisar diversos fatores que influenciam a evasão escolar. Estes métodos são capazes de identificar padrões complexos e estabelecer correlações significativas entre variáveis que podem não ser imediatamente evidentes em análises tradicionais.

Com essa abordagem, é possível prever quais alunos estão em maior risco de abandonar a escola, permitindo que as instituições educacionais intervenham de forma antecipada. Por meio de medidas preventivas e personalizadas, as escolas podem agir antes que o abandono ocorra, ajustando suas estratégias de suporte de acordo com as necessidades individuais de cada aluno. Além disso, esta abordagem não se limita a identificar o risco de evasão, mas também visa criar um ambiente escolar mais inclusivo. Ao reconhecer de forma precoce os desafios que os alunos enfrentam, as escolas podem oferecer intervenções específicas, como apoio acadêmico, ou dentro do possível, social e emocional, que estão diretamente ligados aos fatores que contribuem para a evasão [7] [8] [11] [9]. Esta combinação de tecnologia e personalização tem o potencial de melhorar significativamente as taxas de retenção escolar [9] [10] e, ao mesmo tempo, proporcionar um percurso académico mais justo e adaptado às necessidades de cada estudante.

1.3 Justificativa

Este estudo se justifica pela necessidade de desenvolver estratégias inovadoras que possam auxiliar na mitigação do problema da evasão escolar através de tecnicas de algorítmos de iteligência artificial .

Além do impacto direto na vida pessoal dos alunos, a evasão escolar também resulta em consequências para a sociedade. A baixa escolaridade está frequentemente associada a menores oportunidades de emprego, perpetuando o aumento da desigualdade social. Assim, a predição desse fenômeno não apenas auxilia no enfrentamento imediato do problema, mas também serve de embasamento para o desenvolvimento de políticas públicas .

Outro aspecto que deve ser considerado desta pesquisa é sua contribuição para o aprimoramento das práticas pedagógicas. A adoção de ferramentas tecnológicas pode facilitar o acompanhamento contínuo do desempenho estudantil, permitindo ajustes rápidos e direcionados, considerando as particulariedades de cada aluno.

Uma vez que identificado um possivél caso de abandono algumas estratégias podem ser criadas, como:

- A orientação sobre programas sociais disponíveis para famílias em situação de vulnerabilidade, garantindo que os alunos e seus responsáveis tenham acesso a benefícios ao quias ele tem direito;
- Acompanhamento psicopedagógico;
- Cadastramento em políticas de transporte gratuito (passes estudantis), que podem reduzir problemas de locomoção;
- Programas de mentoria;
- Grupos de apoio podem ser implementados para oferecer um ambiente mais inclusivo;
- Estratégias pedagógicas adaptadas, caso o aluno possua alguma necessidade específica (cuidador, mediador, interprete e outros);

Assim, esta pesquisa se insere em um contexto de inovação educacional, propondo soluções baseadas em inteligência artificial aliadas a políticas pedagógicas que possam tornar o sistema educacional mais equitativo, garantindo melhores oportunidades para todos os estudantes.

1.4 Organização da Dissertação

O Capítulo 2 apresenta uma seleção de referências bibliográficas relevantes ao tema evasão escolar. Neste segmento, são analisadas as causas que podem levar um aluno a interromper seus estudos, considerando variáveis como aspectos socioeconômicos e comportamentais dos alunos. A revisão abrange uma variedade de estudos, incluindo diferentes abordagens, discutidas por diversos autores, o que proporciona uma compreensão detalhada do problema. No Capítulo 3, são revisitados conceitos fundamentais já estabelecidos na literatura, essenciais para fundamentar a metodologia utilizada neste trabalho. O Capítulo 4 descreve o processo adotado na pesquisa, enfatizando como os temas abordados no Capítulo 3 foram integrados na formulação da metodologia. Ainda no Capítulo 4, são apresentados os resultados obtidos a partir da aplicação da metodologia proposta. Por fim, no Capítulo 5, são apresentadas as considerações finais da pesquisa, acompanhadas de sugestões para futuras investigações.

2 Evasão Escolar

A evasão escolar é um problema complexo e amplo que afeta negativamente o desenvolvimento educacional e social em diversos contextos. No Brasil, as causas desse fenômeno são amplamente discutidas na literatura, destacando-se fatores como questões econômicas, dificuldades familiares, gravidez precoce e a falta de motivação ou perspectivas por parte dos alunos [7][8][11][12].

Esse problema não apenas compromete a trajetória educacional dos jovens, mas também impacta o desenvolvimento socioeconômico do país, criando desafios consideráveis para o mercado de trabalho e a inclusão social [10].

Por sua vez, análisar de forma sistemática as causas da evasão pode ajudar na criação de políticas públicas mais eficazes e estratégias pedagógicas que promovam a permanência do aluno na unidade escolar [9].

Nesse contexto, compreender as razões que levam ao abandono escolar é fundamental para planejar intervenções adequadas, garantindo o apoio necessário para que estudantes em situação de vulnerabilidade possam permanecer na escola e alcançar melhores perspectivas futuras.

2.1 Revisão Bibliográfica

A obra “Evasão Escolar no Brasil: Uma Perspectiva Multidimensional"[7], apresenta uma análise sobre o fenômeno complexo da evasão escolar no contexto brasileiro. A autora, oferece uma visão aprofundada e multidimensional desse desafio. Ainda no contexto da evasão escolar[7], explora uma variedade de fatores que contribuem para a evasão, incluindo questões socioeconômicas, culturais, educacionais e psicológicas. Ao longo da obra, a autora baseia suas análises em dados empíricos e estudos de caso, oferecendo uma base sólida para suas conclusões. Além disso, ela examina as políticas públicas educacionais implementadas no Brasil e sua eficácia na redução da evasão, contribuindo para o debate sobre possíveis soluções. Uma contribuição do trabalbho é a ênfase na importância da abordagem preventiva, destacando a necessidade de intervenções não apenas no nível escolar, mas também em níveis mais amplos da sociedade. O autor propõe perspectivas inovadoras e estratégias integradas para enfrentar o desafio, reconhecendo a diversidade de contextos regionais e socioeconômicos no Brasil.

O artigo [8], oferece uma análise aprofundada das causas subjacentes à evasão escolar no contexto do Ensino Fundamental. Os autores buscam identificar fatores determinantes e propor soluções valiosos para o entendimento e enfrentamento desse desafio. O estudo inicia com uma revisão crítica da literatura existente sobre evasão escolar, fornecendo um contexto sólido

para a pesquisa. [8] adota uma abordagem metodológica robusta, utilizando análise de dados quantitativos e qualitativos para fundamentar suas conclusões. Os resultados apresentados indicam uma variedade de fatores que contribuem para a evasão escolar no Ensino Fundamental. Os autores exploram não apenas causas óbvias, como problemas socioeconômicos, mas também fatores mais sutis relacionados ao ambiente escolar, metodologias de ensino e dinâmicas entre alunos e professores.

O Uso de Inteligência Artificial para Predição de Evasão na Rede Doctum de Ensino [9], busca abordar o problema da evasão em uma instituição particular de ensino superior, através do desenvolvimento de uma metodologia que antecipa os casos de abandono escolar. Visando a criação de estratégias de retenção de alunos, a fim de diminuir as taxas de evasão. A abordagem adotada envolve a implementação de um sistema de gerenciamento de dados na instituição de ensino Rede Doctum, empregando a ferramenta Weka, utilizado para realizar tarefa de mineração de dados e um algoritmo de redes neurais artificiais. O objetivo é classificar alunos entre evadidos e não evadidos, gerando um panorama do percentual de evasão na instituição e identificando grupos de risco. Para fazer a predição o trabalho considera relevantes os dados que correspondem ao perfil dos alunos da instituição, como: inadimplência, curso, idade, valor da mensalidade, entre outros. Os resultados obtidos na fase de treinamento revelaram um percentual de acerto em torno de 95%, diminuindo para 93% durante a fase de confirmação

O artigo [10] "Pode a inteligência artificial apoiar ações contra evasão escolar Universitária?", trata a evasão escolar como um problema global e demanda um critério de investigação profunda para compreensão de seus impactos negativos . Este trabalho propõe a utilização de Mineração de Dados Educacionais, empregando técnicas de Aprendizado de Máquina, com o objetivo de identificar variáveis cruciais na caracterização do perfil de estudantes em risco de evasão. O estudo utiliza diversas técnicas de Aprendizado de Máquina, incluindo Máquina de Vetores de Suporte, Gradient Boosting Machine, Floresta Aleatória e um comitê de máquina. A aplicação dessas técnicas conta com o registro de 1.429 estudantes de cursos superiores em um campus do IFMG, no período de 2013 a 2019. Com os resultados obtidos no artigo é possível observar uma superioridade de desempenho do comitê de máquina em relação às demais técnicas.

Evasão e abandono escolar na educação básica no Brasil: fatores, causas e possíveis consequências [11], visa contribuir para o debate em torno da evasão e abandono escolar na educação brasileira, enfocando considerações importantes sobre suas diversas dimensões. Destaca-se a complexidade do tema, evidenciando que diferentes interpretações coexistem, dificultando a obtenção de uma definição precisa. Um desafio adicional é a falta de um conceito claro, reconhecido até mesmo por órgãos oficiais de educação. A persistência da evasão e do abandono escolar é influenciada por uma interação de fatores internos e externos. O artigo cita que Brasil apresenta a terceira maior taxa de abandono escolar entre os 100 países com maior Índice de Desenvolvimento Humano (IDH). Isso sublinha a urgência de abordar essas questões de maneira eficaz para melhorar o sistema educacional. Segundo o artigo o processo de evasão é delineado como sendo complexo, dinâmico e acumulativo, muitas vezes equiparado à expulsão escolar. Apesar das metas estabelecidas pela Constituição Federal de 1988 para a universalização do ensino fundamental e a "erradicação"do analfabetismo, esses objetivos ainda não foram totalmente alcançados.Na leitura do autor a evasão é caracterizada como um "ato solitário"e pode ser interpretada como um reflexo do fracasso das relações sociais. A ausência de uma origem claramente definida implica que a evasão não tem uma conclusão intrínseca. Diversos fatores, como o envolvimento com drogas, repetidas reprovações, prostituição, falta de apoio da família e da escola, são apontados como possíveis catalisadores para a saída do educando do ambiente escolar.

Na monografia, intitulada "A problemática da evasão escolar: uma revisão bibliográfica integrativa"[12], concentra-se na problemática persistente da evasão escolar no Brasil, com ênfase no ensino médio. A educação no país enfrenta inúmeros desafios, sendo a evasão escolar um dos temas mais questionados e pesquisados. Esta revisão bibliográfica integrativa propôs analisar e identificar as causas e consequências da evasão escolar no ensino médio, com o objetivo de refletir sobre possíveis intervenções para o problema.

Como metodologia, a pesquisa utilizou fontes como o Google Acadêmico, Periódicos, Scielo e revistas eletrônicas, selecionando artigos e pesquisas que abordam a evasão escolar no ensino. O trabalho aborda a importância do direito à educação, define o que é fracasso e evasão escolar, e discorre sobre suas principais causas e consequências. Segundo os autores revisados, esses fatores estão relacionados à escola, aos alunos, à família e às desigualdades sociais. A fundamentação teórica baseou-se em quatro artigos selecionados para análise e comparação, resultando em uma síntese de dados que revelou resultados semelhantes. Entre os destaques, surgiram fatores socioeconômicos, falta de estrutura escolar e ausência de políticas públicas inclusivas como elementos que desencadeiam consequências graves, como a exclusão social.

A presente revisão bibliográfica percorreu obras e estudos fundamentados no campo da evasão escolar.

Referências utilizadas na Revisão Bibliográfica

Título	**Autores**	**Ano**	**Revista**	**Metodologia**
Evasão Escolar no Brasil: Uma Perspectiva Multidimensional	Carvalho, M. M. C.	2014	-	Revisão teórica
Análise das Causas da Evasão Escolar no Ensino Fundamental	Mendes, L. S.; Souza, F. C.	2021	Educação em Foco	Estudo de caso
O uso de IA para predição de evasão na rede Doctum	Dutra, R. M.	2015	-	Modelos de IA
IA no combate à evasão escolar universitária	Bitencourt, W. A.; Silva, D. M.; Xavier, G. C.	2021	Ensaio	Revisão bibliográfica
Evasão e abandono escolar na educação básica	Filho, R. B. S.; Araújo, R. M. de L.	2017	Educação Por Escrito	Estudo empírico
A problemática da evasão escolar	Lino, E. R. O.	2020	-	Revisão bibliográfica

Tabela 2.1 – Resumo das referências utilizadas na Revisão Bibliográfica

3 Revisão Teórica

Este capítulo apresenta uma introdução breve à base teórica que sustenta os temas abordados neste estudo. A seção explora as principais abordagens e contribuições da literatura especializada relacionadas à inteligência artificial e ao aprendizado de máquina.

3.1 Inteligência Artificial (IA)

A inteligência artificial teve suas primeiras incursões realizadas por Warren McCulloch e Walter Pitts em 1943, os quais propuseram um modelo baseado no funcionamento dos neurônios do cérebro humano. A inteligência artificial refere-se à capacidade das máquinas de realizar tarefas de maneira semelhante aos seres humanos. De acordo com Russell [1], é possível categorizar as definições de IA em quatro estratégias:

- Processos de Pensamento e Raciocínio Humano;
- Comportamento humano;
- Sucesso em termos de fidelidade ao desempenho de tarefas relacionadas ao comportamento humano;
- Racionalidade, relacionada ao estudo dos projetos de agentes inteligentes.

Ainda de acordo com [1], em 1950, Turing desenvolveu um teste conhecido como o "Teste de Turing". O teste consiste em uma série de perguntas, onde um interlocutor humano precisa determinar se a resposta foi fornecida por uma máquina ou por um ser humano. Se o interlocutor não puder identificar a origem da resposta, considera-se que o teste foi bem sucedido. No campo da Aprendizagem de Máquina, qualquer agente é capaz de aprender por meio da observação do mundo. A aprendizagem é contínua, e inclui desde tarefas simples até o desenvolvimento de conhecimentos mais complexos. Esse conceito de aprendizagem foi expandido para o estudo da inteligência artificial, onde, a partir de um conjunto de pares de entradas e saídas, é possível aprender uma função que prevê a saída para novas entradas. Os agentes também podem aprimorar o aprendizado com base em dados. Para otimizar o desenvolvimento dessa técnica, quatro fatores devem ser considerados:

- Qual componente deve ser aprimorado;
- Qual conhecimento prévio o agente possui;

- Qual representação é utilizada tanto para os dados quanto para os componentes;
- Qual feedback está disponível para o aprendizado.

3.2 Aprendizagem supervisionada

Após a revisão de algumas referências disponíveis na literatura [1, 3, 6, 5], podemos definir um **problema de aprendizagem supervisionada** como aquele em que se deseja aprender uma função desconhecida f que mapeia um conjunto de entradas para suas respectivas saídas:

$$Y = f(X). \tag{3.1}$$

O objetivo é estimar ou aprender a função f a partir de um conjunto de dados rotulados. Esse conjunto de dados consiste em pares de entrada e saída, denotados por (X, Y), onde:

- X representa a variável de entrada;
- Y é a variável de saída correspondente.

Se a variável Y assume um número finito de valores distintos, o problema de aprendizagem é classificado como um problema de **classificação**.

Para resolver esse problema, utilizamos uma função hipotética, denotada por h, que serve como uma aproximação da função verdadeira e desconhecida f:

$$h(X) \approx f(X). \tag{3.2}$$

Durante o treinamento, o modelo é ajustado de forma a minimizar a diferença entre suas previsões e os valores reais presentes nos dados de treinamento.

A **distribuição condicional de probabilidade**, também chamada de distribuição probabilística condicional, modela a probabilidade de uma variável aleatória Y ocorrer, dado que outra variável aleatória X já ocorreu. A notação padrão para essa relação é:

$$P(Y|X), \tag{3.3}$$

que se lê como "a probabilidade de Y dado X".

O objetivo final do modelo é generalizar bem para novos dados, de modo que suas previsões sejam precisas mesmo em situações não observadas durante o treinamento.

3.3 Métodos de Classificação

Árvores de Decisão

As árvores de decisão são algoritmos de classificação baseados em regras hierárquicas. O algoritmo segmenta o espaço de características em regiões menores, associando cada uma a uma classe específica. Cada nó interno da árvore representa uma decisão baseada em uma característica do conjunto de dados, enquanto os nós folha representam as classes finais.

As árvores de decisão são amplamente utilizadas devido à sua facilidade de interpretação e eficácia em uma variedade de problemas [22].

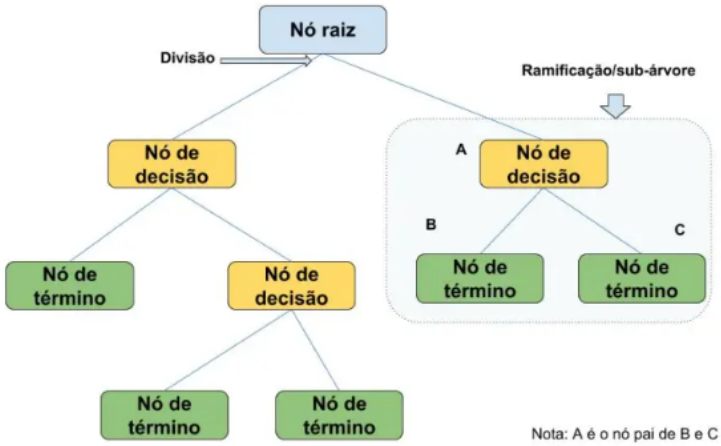


Figura 3.1 – Exemplo de Árvore de Decisão.
Fonte: https://vooo.pro/insights/um-tutorial-completo-sobre-a-modelagem-baseada-em-tree-arvore-do-zero-em-r-python/Acesso em: 18 mar. 2025.

K-Nearest Neighbors(K-NN)

O K-NN é um algoritmo de classificação, que classifica um ponto de dados com base na classe da maioria dos seus k vizinhos mais próximos. A medida de proximidade pode ser definida usando diferentes métricas, como distância euclidiana. Este método é intuitivo e simples, embora possa ser sensível a pontos fora da curva e dimensionamento de características [5].

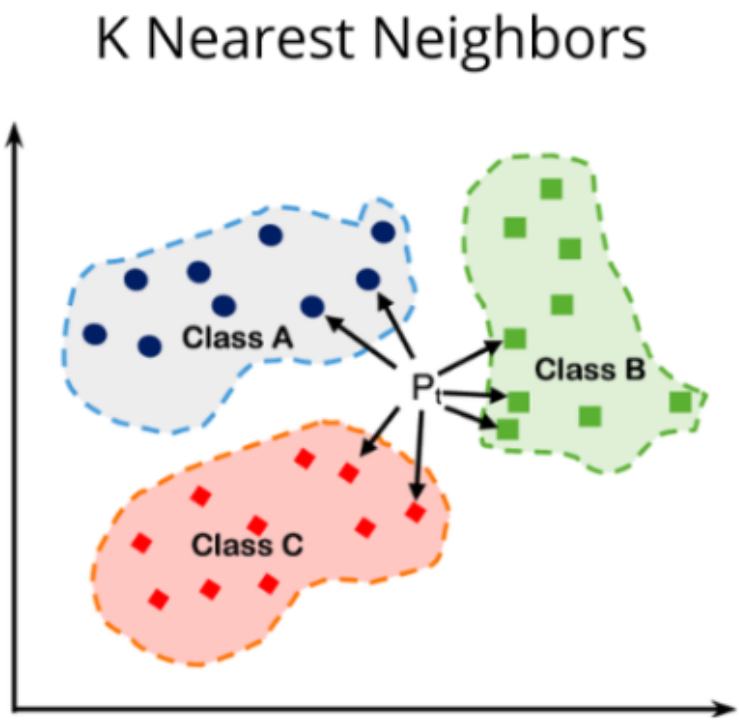


Figura 3.2 – K-Nearest Neighbors (K-NN).
Fonte: https://medium.com/@sachinsoni600517/k-nearest-neighbours-introduction-to-machine-learning-algorithms-9dbc9d9fb3b2. Acesso em: 26 nov. 2024.

Máquinas de Vetores de Suporte (SVMs)

Uma SVM (Support Vector Machine) é um algoritmo de aprendizado de máquina supervisionado utilizado para classificação e regressão. A ideia central é encontrar o melhor hiperplano que separa os dados em diferentes classes, maximizando a margem entre essas classes.

Como funcionam?

- Mapeamento dos dados: Os dados são mapeados para um espaço de alta dimensão, onde se busca um hiperplano que os separe da melhor forma.
- Maximização da margem: O algoritmo busca o hiperplano que maximiza a distância entre os pontos mais próximos de cada classe, chamados de vetores de suporte.
- Classificação: Novos dados são classificados de acordo com o lado do hiperplano em que se encontram. [21][5].

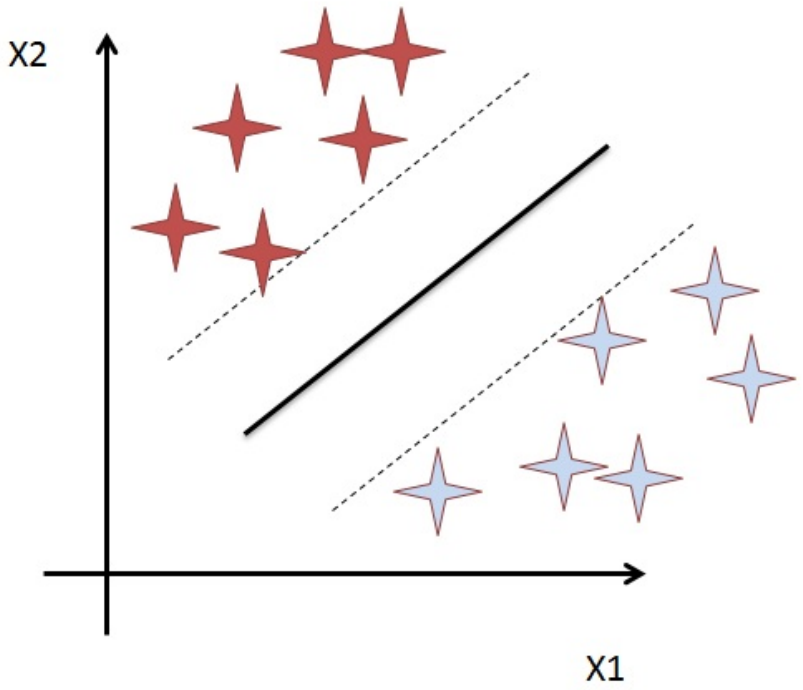


Figura 3.3 – Maquina de Vetores de Suporte.
Fonte: https://pt.wikipedia.org/wiki/Máquina_de_vetores_de_suporte. Acesso em: 26 nov. 2024.

Redes Neurais Artificiais

são um componente central do Aprendizado de Máquina, inspiradas no funcionamento do cérebro humano. Uma RNA consiste em unidades interconectadas, chamadas neurônios, organizadas em camadas [5] [6] [24].

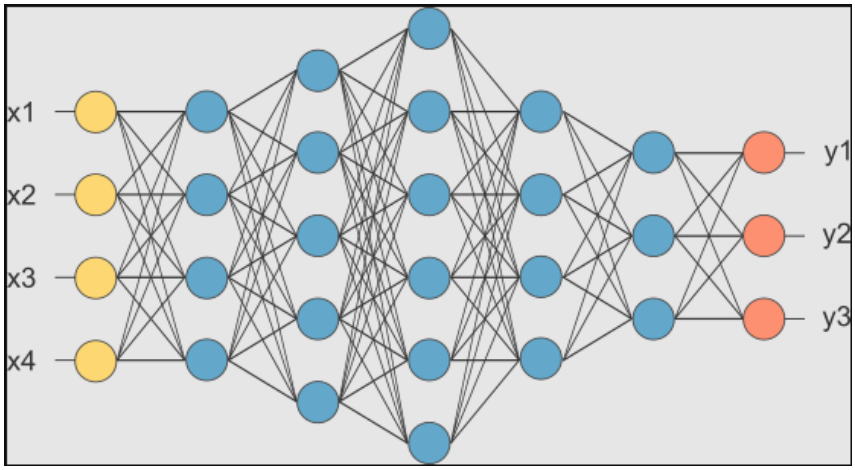


Figura 3.4 – Redes Neurais Artificiais.
Fonte: https://www.oficinadanet.com.br/tecnologia/25007-o-que-sao-as-redes-neurais-artificiais. Acesso em: 26 nov. 2024.

Neurônios Artificiais

- Entradas (Inputs): Cada neurônio recebe várias entradas numéricas.
- Pesos (Weights): As entradas são multiplicadas por pesos para ajustar sua importância.
- Função de Ativação: A soma ponderada das entradas passa por uma função que determina a saída do neurônio.

Camadas

- Camada de Entrada: Recebe os dados iniciais.
- Camadas Ocultas: Processam informações e aprendem padrões.
- Camada de Saída: Produz o resultado final da rede.

Função de Ativação

- Sigmoid: Mapeia valores para o intervalo (0, 1), útil em problemas de classificação binária.
- ReLU (Rectified Linear Unit): Retorna zero para valores negativos e linear para positivos, comum em camadas ocultas.
- Softmax: Usada na camada de saída para problemas de classificação multiclasse, convertendo valores em probabilidades.

Propagação para Frente (Forward Propagation)

- Os dados passam pela rede, aplicando multiplicação de pesos, somas ponderadas e funções de ativação.

Função de Perda

- Avalia o desempenho comparando as saídas previstas com os valores reais.

Retropropagação

- Calcula o gradiente da função de perda em relação aos pesos da rede.

Otimização

- Gradiente Descendente: Ajusta os pesos para minimizar a perda.
- Taxa de Aprendizado: Controla o tamanho dos ajustes.

Iteração

- O processo de propagação, cálculo da perda, retropropagação e ajuste de pesos é repetido até a rede atingir um desempenho satisfatório.

Regularização

- Técnicas como dropout ou regularização L1/L2 evitam overfitting.

Generalização

- A rede prevê com precisão em novos exemplos, generalizando para dados não vistos.

4

A Utilização do Aprendizado de Máquina na Predição do Abandono Escolar

O estudo teve início com o objetivo de realizar um levantamento dos alunos que evadiram ao longo de um período de 10 anos em uma escola da rede estadual de ensino, localizada no bairro da Posse, no município de Nova Iguaçu, Rio de Janeiro. O objetivo dessa etapa era identificar padrões de evasão escolar ao longo do tempo. No entanto, observou-se que havia uma insuficiência de informações disponíveis, e as poucas informações existentes ainda contavam com registros incompletos ou desatualizados, principalmente no que diz respeito aos alunos que abandonaram a escola, o que impôs limitações significativas a proposta inicial, exigindo a adoção de uma abordagem metodológica alternativa.

O EJANEM é uma modalidade de ensino que atende jovens e adultos, essa modalidade apresenta características únicas que os tornam interessante para análise da evasão escolar. Esses alunos em algum momento de suas trajetórias, por alguma razão precisaram iterromper seus estudos. Entender as razões que os levaram a tomar essa decisão foi fundamental para o desenvolvimento da pesquisa. Os dados necessários foram coletados por meio de questionários referentes a época em que ocorreu a evasão, e incluíram informações sobre idade, gênero, escolaridade anterior, motivos para evasão e desempenho acadêmico. Para obter resultados preliminares que orientassem o refinamento do modelo preditivo, foi utilizada uma base de dados pré-existente, fornecida por um grupo de pesquisadores que havia conduzido uma pesquisa sobre a merenda escolar. Essa base de dados incluía informações sobre desempenho acadêmico, condições socioeconômicas, entre outras variáveis relevantes.

Posteriormente foi realizada a organização e o cadastramento das informações em sistemas de bancos de dados apropriado. Em seguida, foi realizada a adequação das colunas de atributos, considerando que os dados coletados no EJANEN não coincidiam integralmente com aqueles previamente disponibilizados referentes aos alunos do ensino regular. Para a obtenção dos resultados preliminares, tornou-se necessário utilizar apenas as colunas de dados que apresentavam correspondência entre os dois conjuntos de informações.

Para o desenvolvimento deste modelo inicial, foram considerados os seuintes atributos:

- Série;
- Sexo atribuído ao nascimento;
- Idade;
- Exerce alguma atividade além de estudar;

- Renda familiar;
- Quantas pessoas compõem o núcleo familiar;
- É responsável por alguém do seu núcleo familiar;
- O bairro onde mora possui casos de violência que afetam sua frequência escolar;
- A oferta de merenda afeta sua frequência escolar;
- Seu núcleo familiar afeta sua frequência escolar.

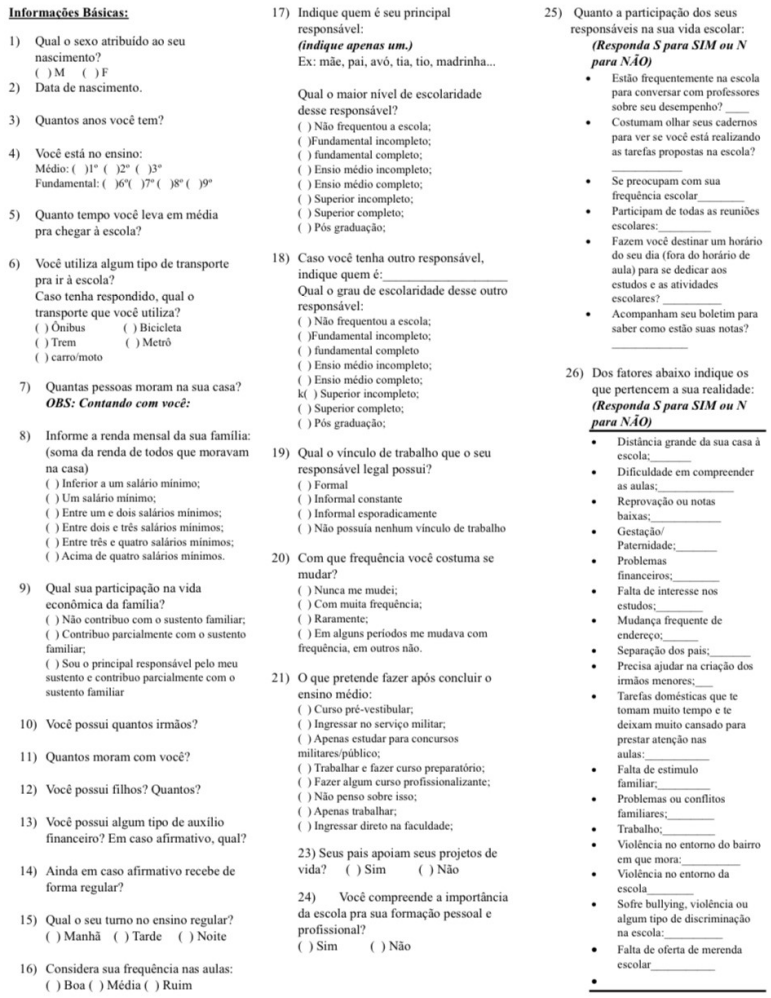
Com essa primeira amostra foi possível obter os primeiros resultados que foram fundamentais para validar modelo que resultou em uma acurácia de 79%. Para melhorar o desempenho do estudo, foi necessário ampliar a amostra, e para isso elaborou-se um formulário fundamentado em uma revisão bibliográfica [7, 8, 11, 12, 14, 15, 17]. A revisão da literatura permitiu identificar fatores significativos, o que auxiliou na construção de um instrumento de coleta de dados capaz de captar informações detalhadas sobre as causas associadas à evasão escolar. O formulário também foi projetado para coletar dados sociodemográficos, acadêmicos e contextuais, cosiderando fatores relevantes para o problema em questão e oferecendo uma visão panorâmica das variáveis que podem influenciar o abandono escolar. A pesquisa incluiu também a participação dos alunos do EJANEM (Educação de Jovens e Adultos do Novo Ensino Médio), considerando que muitos desses estudantes haviam frequentado o ensino regular anteriormente, mas interromperam seus estudos por diversos motivos. Esse grupo proporcionou uma fonte riquíssima de informações que contribuíram significativamente para alimentar o modelo, permitindo a geração de dados passados para uma previsão futura. A coleta de dados junto a esse público foi essencial para uma compreensão mais profunda das razões que os levaram a interromper a trajetória escolar. Inicialmente, o formulário foi disponibilizado online, com o intuito de facilitar o acesso e a conveniência na resposta. Contudo, a adesão foi pequena, principalmente devido à dificuldade de compreensão das perguntas por parte dos alunos e à falta de engajamento, o que comprometeu a representatividade dos dados coletados. Para superar essas limitações, optou-se pela aplicação de formulários impressos em todas as turmas dos quatro módulos do EJANEM. Essa abordagem presencial permitiu maior alcance, contemplando aproximadamente oito turmas do EJANEM e 12 turmas do ensino regular. A presença de um mediador durante a aplicação do formulário garantiu que dúvidas fossem sanadas imediatamente, resultando em uma maior precisão e qualidade dos dados coletados. O novo formulário contendo uma série de atributos detalhados, incluindo variáveis como:

- Tipo de ensino (regular ou EJANEM);
- Sexo atribuído ao nascimento;
- Data de nascimento;
- Idade;
- Série escolar;
- Escolaridade dos responsáveis;
- Tempo de deslocamento até a escola;
- Tipo de transporte utilizado;

- Número de moradores na residência;
- Renda familiar mensal;
- Participação na vida econômica da família;
- Quantidade de irmãos;
- Quantidade de irmãos que residem na mesma casa;
- Presença de filhos e, em caso afirmativo, o número de filhos;
- Recebimento de algum tipo de auxílio financeiro, especificando o tipo e se o recebe de forma regular;
- Turno em que estuda;
- Frequência escolar;
- Identificação dos responsáveis legais;
- Tipo de atividade econômica realizada pelos responsáveis (formal ou informal);
- Frequência de mudanças de endereço;
- Existência de planos para o futuro;
- Apoio dos responsáveis em relação a esses planos;
- Compreensão da importância dos estudos para o desenvolvimento pessoal e profissional;
- Participação dos responsáveis na vida acadêmica (reuniões, verificação de frequências e notas);
- Incentivo dos responsáveis aos estudos;
- Dificuldades nos estudos, como notas baixas, dificuldade de acompanhar o conteúdo das disciplinas, falta de interesse pela escola ou histórico de reprovações;
- Responsabilidade na criação de irmãos menores;
- Interferência das tarefas domésticas na realização das atividades escolares;
- Falta de estímulo por parte dos familiares;
- Condição de estar empregado;
- Problemas de violência no entorno do bairro onde reside;
- Problemas de violência no entorno do bairro onde está localizada a escola, ou mesmo dentro da unidade escolar;
- Experiência de discriminação dentro da escola;
- Vivência de bullying;
- Influência da oferta de merenda escolar na melhoria da experiência do aluno na escola.

Este conjunto de atributos foi cuidadosamente selecionado para permitir uma análise detalhada dos fatores que podem influenciar a evasão escolar, fornecendo uma base sólida para o desenvolvimento de estratégias de intervenção mais eficazes. O novo formulário pode ser observado na imagem abaixo

Ensino Regular:



Informações Básicas:

1) Qual o sexo atribuído ao seu nascimento?
( ) M ( ) F
2) Data de nascimento.
3) Quantos anos você tem?
4) Você está no ensino:
Médio: ( )1° ( )2° ( )3°
Fundamental: ( )6°( )7° ( )8° ( )9°
5) Quanto tempo você leva em média pra chegar à escola?
6) Você utiliza algum tipo de transporte pra ir à escola?
Caso tenha respondido, qual o transporte que você utiliza?
( ) Ônibus ( ) Bicicleta
( ) Trem ( ) Metrô
( ) carro/moto
7) Quantas pessoas moram na sua casa?
OBS: Contando com você:
8) Informe a renda mensal da sua família: (soma da renda de todos que moravam na casa)
( ) Inferior a um salário mínimo;
( ) Um salário mínimo;
( ) Entre um e dois salários mínimos;
( ) Entre dois e três salários mínimos;
( ) Entre três e quatro salários mínimos;
( ) Acima de quatro salários mínimos.
9) Qual sua participação na vida econômica da família?
( ) Não contribuo com o sustento familiar;
( ) Contribuo parcialmente com o sustento familiar;
( ) Sou o principal responsável pelo meu sustento e contribuo parcialmente com o sustento familiar
10) Você possui quantos irmãos?
11) Quantos moram com você?
12) Você possui filhos? Quantos?
13) Você possui algum tipo de auxílio financeiro? Em caso afirmativo, qual?
14) Ainda em caso afirmativo recebe de forma regular?
15) Qual o seu turno no ensino regular?
( ) Manhã ( ) Tarde ( ) Noite
16) Considera sua frequência nas aulas:
( ) Boa ( ) Média ( ) Ruim
17) Indique quem é seu principal responsável:
(indique apenas um.)
Ex: mãe, pai, avó, tia, tio, madrinha...

Qual o maior nível de escolaridade desse responsável?
( ) Não frequentou a escola;
( )Fundamental incompleto;
( ) fundamental completo;
( ) Ensio médio incompleto;
( ) Ensio médio completo;
( ) Superior incompleto;
( ) Superior completo;
( ) Pós graduação;
18) Caso você tenha outro responsável, indique quem é:_____
Qual o grau de escolaridade desse outro responsável:
( ) Não frequentou a escola;
( )Fundamental incompleto;
( ) fundamental completo
( ) Ensio médio incompleto;
( ) Ensio médio completo;
k( ) Superior incompleto;
( ) Superior completo;
( ) Pós graduação;
19) Qual o vínculo de trabalho que o seu responsável legal possui?
( ) Formal
( ) Informal constante
( ) Informal esporadicamente
( ) Não possuia nenhum vinculo de trabalho
20) Com que frequência você costuma se mudar?
( ) Nunca me mudei;
( ) Com muita frequência;
( ) Raramente;
( ) Em alguns períodos me mudava com frequência, em outros não.
21) O que pretende fazer após concluir o ensino médio:
( ) Curso pré-vestibular;
( ) Ingressar no serviço militar;
( ) Apenas estudar para concursos militares/público;
( ) Trabalhar e fazer curso preparatório;
( ) Fazer algum curso profissionalizante;
( ) Não penso sobre isso;
( ) Apenas trabalhar;
( ) Ingressar direto na faculdade;

23) Seus pais apoiam seus projetos de vida? ( ) Sim ( ) Não

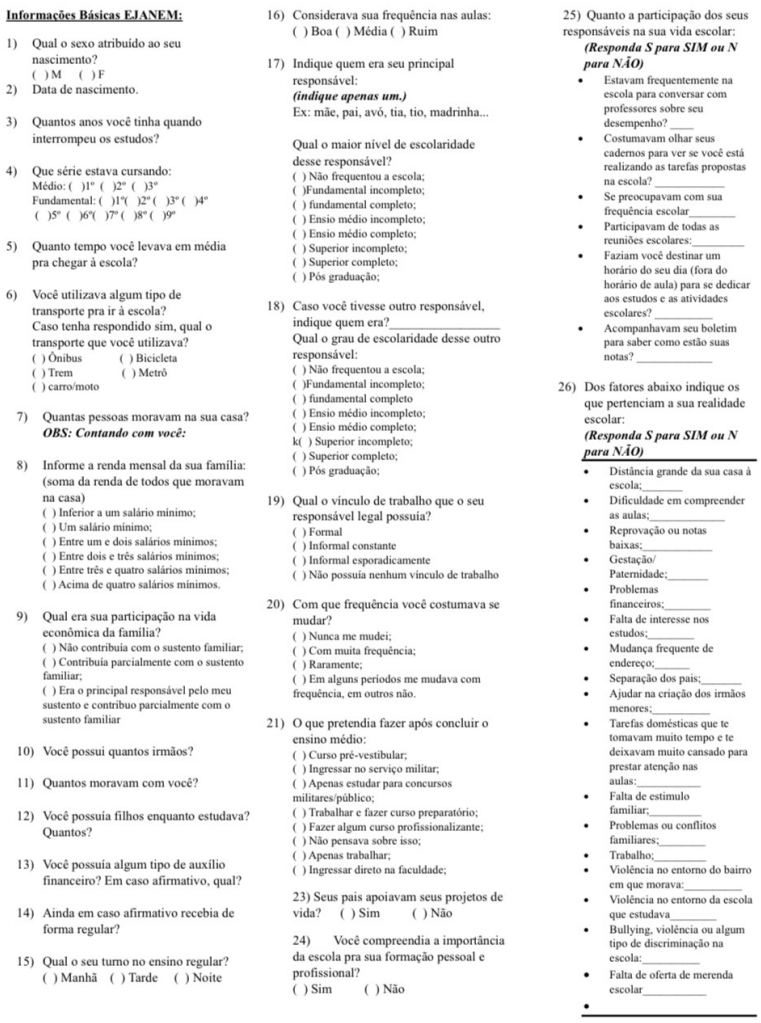
24) Você compreende a importância da escola pra sua formação pessoal e profissional?
( ) Sim ( ) Não

25) Quanto a participação dos seus responsáveis na sua vida escolar:
(Responda S para SIM ou N para NÃO)
- Estão frequentemente na escola para conversar com professores sobre seu desempenho? ____
- Costumam olhar seus cadernos para ver se você está realizando as tarefas propostas na escola? _____
- Se preocupam com sua frequência escolar_____
- Participam de todas as reuniões escolares:_____
- Fazem você destinar um horário do seu dia (fora do horário de aula) para se dedicar aos estudos e as atividades escolares? _____
- Acompanham seu boletim para saber como estão suas notas? _____

26) Dos fatores abaixo indique os que pertencem a sua realidade:
(Responda S para SIM ou N para NÃO)
- Distância grande da sua casa à escola;_____
- Dificuldade em compreender as aulas;_____
- Reprovação ou notas baixas;_____
- Gestação/ Paternidade;_____
- Problemas financeiros;_____
- Falta de interesse nos estudos;_____
- Mudança frequente de endereço;_____
- Separação dos pais;_____
- Precisa ajudar na criação dos irmãos menores;____
- Tarefas domésticas que te tomam muito tempo e te deixam muito cansado para prestar atenção nas aulas:_____
- Falta de estimulo familiar;_____
- Problemas ou conflitos familiares;_____
- Trabalho;_____
- Violência no entorno do bairro em que mora:_____
- Violência no entorno da escola_____
- Sofre bullying, violência ou algum tipo de discriminação na escola:_____
- Falta de oferta de merenda escolar_____
-

Figura 4.1 – Formulário aplicado para os alunos do ensino regular.
Fonte: autor

EJANEM



Informações Básicas EJANEM:

1) Qual o sexo atribuído ao seu nascimento?
( ) M ( ) F
2) Data de nascimento.

3) Quantos anos você tinha quando interrompeu os estudos?

4) Que série estava cursando:
Médio: ( )1° ( )2° ( )3°
Fundamental: ( )1°( )2° ( )3° ( )4°
( )5° ( )6°( )7° ( )8° ( )9°

5) Quanto tempo você levava em média pra chegar à escola?

6) Você utilizava algum tipo de transporte pra ir à escola?
Caso tenha respondido sim, qual o transporte que você utilizava?
( ) Ônibus ( ) Bicicleta
( ) Trem ( ) Metrô
( ) carro/moto

7) Quantas pessoas moravam na sua casa?
OBS: Contando com você:

8) Informe a renda mensal da sua família: (soma da renda de todos que moravam na casa)
( ) Inferior a um salário mínimo;
( ) Um salário mínimo;
( ) Entre um e dois salários mínimos;
( ) Entre dois e três salários mínimos;
( ) Entre três e quatro salários mínimos;
( ) Acima de quatro salários mínimos.

9) Qual era sua participação na vida econômica da família?
( ) Não contribuía com o sustento familiar;
( ) Contribuía parcialmente com o sustento familiar;
( ) Era o principal responsável pelo meu sustento e contribuo parcialmente com o sustento familiar

10) Você possui quantos irmãos?

11) Quantos moravam com você?

12) Você possuía filhos enquanto estudava? Quantos?

13) Você possuía algum tipo de auxílio financeiro? Em caso afirmativo, qual?

14) Ainda em caso afirmativo recebia de forma regular?

15) Qual o seu turno no ensino regular?
( ) Manhã ( ) Tarde ( ) Noite

16) Considerava sua frequência nas aulas:
( ) Boa ( ) Média ( ) Ruim

17) Indique quem era seu principal responsável:
(indique apenas um.)
Ex: mãe, pai, avó, tia, tio, madrinha...

Qual o maior nível de escolaridade desse responsável?
( ) Não frequentou a escola;
( )Fundamental incompleto;
( ) fundamental completo;
( ) Ensio médio incompleto;
( ) Ensio médio completo;
( ) Superior incompleto;
( ) Superior completo;
( ) Pós graduação;

18) Caso você tivesse outro responsável, indique quem era?_____
Qual o grau de escolaridade desse outro responsável:
( ) Não frequentou a escola;
( )Fundamental incompleto;
( ) fundamental completo
( ) Ensio médio incompleto;
( ) Ensio médio completo;
k( ) Superior incompleto;
( ) Superior completo;
( ) Pós graduação;

19) Qual o vínculo de trabalho que o seu responsável legal possuía?
( ) Formal
( ) Informal constante
( ) Informal esporadicamente
( ) Não possuía nenhum vínculo de trabalho

20) Com que frequência você costumava se mudar?
( ) Nunca me mudei;
( ) Com muita frequência;
( ) Raramente;
( ) Em alguns períodos me mudava com frequência, em outros não.

21) O que pretendia fazer após concluir o ensino médio:
( ) Curso pré-vestibular;
( ) Ingressar no serviço militar;
( ) Apenas estudar para concursos militares/público;
( ) Trabalhar e fazer curso preparatório;
( ) Fazer algum curso profissionalizante;
( ) Não pensava sobre isso;
( ) Apenas trabalhar;
( ) Ingressar direto na faculdade;

23) Seus pais apoiavam seus projetos de vida? ( ) Sim ( ) Não

24) Você compreendia a importância da escola pra sua formação pessoal e profissional?
( ) Sim ( ) Não

25) Quanto a participação dos seus responsáveis na sua vida escolar:
(Responda S para SIM ou N para NÃO)
- Estavam frequentemente na escola para conversar com professores sobre seu desempenho? ____
- Costumavam olhar seus cadernos para ver se você está realizando as tarefas propostas na escola? _____
- Se preocupavam com sua frequência escolar_____
- Participavam de todas as reuniões escolares:_____
- Faziam você destinar um horário do seu dia (fora do horário de aula) para se dedicar aos estudos e as atividades escolares? _____
- Acompanhavam seu boletim para saber como estão suas notas? _____

26) Dos fatores abaixo indique os que pertenciam a sua realidade escolar:
(Responda S para SIM ou N para NÃO)
- Distância grande da sua casa à escola;_____
- Dificuldade em compreender as aulas;_____
- Reprovação ou notas baixas;_____
- Gestação/ Paternidade;_____
- Problemas financeiros;_____
- Falta de interesse nos estudos;_____
- Mudança frequente de endereço;_____
- Separação dos pais;_____
- Ajudar na criação dos irmãos menores;_____
- Tarefas domésticas que te tomavam muito tempo e te deixavam muito cansado para prestar atenção nas aulas:_____
- Falta de estimulo familiar;_____
- Problemas ou conflitos familiares;_____
- Trabalho;_____
- Violência no entorno do bairro em que morava:_____
- Violência no entorno da escola que estudava_____
- Bullying, violência ou algum tipo de discriminação na escola:_____
- Falta de oferta de merenda escolar_____
-

Figura 4.2 – Formulário aplicado para os alunos do EJANEM.
Fonte: autor

A evasão escolar é resultado de um conjunto de fatores interligados, como condições socioeconômicas, falta de apoio familiar, dificuldades acadêmicas, violência e sobrecarga doméstica. As referências citadas oferecem uma base sólida para compreender essas causas.

Condições Socioeconômicas e Fatores Demográficos

Itens: Tipo de ensino, escolaridade dos responsáveis, renda familiar mensal, tempo de deslocamento, tipo de transporte, número de moradores na residência, e frequência de mudanças de endereço.

Referências:

- [7] - Aborda como condições socioeconômicas (renda, transporte) influenciam na permanência escolar.

- [8] - Fatores como dificuldades financeiras e distâncias afetam diretamente a evasão escolar, sobretudo no ensino fundamental.
- [11] - Exploram o impacto das desigualdades sociais e barreiras estruturais no abandono escolar.

Apoio Familiar e Participação na Vida Acadêmica

Itens: Participação dos responsáveis na vida acadêmica, incentivo aos estudos, apoio aos planos futuros, e compreensão sobre a importância dos estudos.
Referências:

- [7] - Ressalta a importância do suporte familiar e o papel dos responsáveis em apoiar e acompanhar os estudos.
- [12] - Discute como o incentivo familiar e a presença ativa dos responsáveis são cruciais para evitar o abandono.
- [11] - Analisam como a falta de acompanhamento dos responsáveis pode levar ao desinteresse e evasão.

Dificuldades Acadêmicas e Interesses

Itens: Notas baixas, dificuldade de acompanhar conteúdos, falta de interesse na escola, histórico de reprovações.
Referências:

- [8] - Identificam que o desinteresse e as dificuldades acadêmicas são fatores primários na evasão.
- [11] - Dificuldades em acompanhar o ritmo escolar estão associadas ao histórico de reprovações e eventual abandono.

Questões Sociais e Comunitárias

Itens: Problemas de violência no entorno ou dentro da escola, discriminação, bullying, e falta de estímulo por parte de familiares.
Referências:

- [7] - Analisa como a violência, tanto no ambiente escolar quanto comunitário, desencoraja a permanência dos alunos.
- [11]- Destacam o impacto do bullying e da discriminação no bem-estar emocional e na decisão de abandono escolar.
- [12] - A violência social e escolar aparece como fator recorrente em estudos sobre evasão.

Carga Doméstica e Trabalho Infantil

Itens: Participação na vida econômica da família, ajuda na criação dos irmãos, tarefas domésticas impeditivas, e se já trabalham.
Referências:

- [7] - Discorre sobre a sobrecarga de responsabilidades domésticas e o trabalho infantil como barreiras para a continuidade escolar.
- [8] - Relatam a interferência das tarefas domésticas e do trabalho na regularidade da frequência escolar, principalmente em contextos socioeconômicos mais baixos.

Oferta Escolar e Experiência do Aluno

Itens: Oferta de merenda escolar, condições da escola, e turno de estudo.
Referências:

- [7] - Discute como políticas públicas, como a merenda escolar, ajudam a atrair e manter alunos na escola.
- [11] - Ressaltam a importância de uma escola acolhedora e do suporte em questões básicas como alimentação.

Vale salientar que, além das revisões bibliográficas, foram considerados fatores relevantes ao abandono com base em apontamentos feitos pelas percepções e práticas dos profissionais da educação que atuam nas escolas onde a pesquisa foi realizada.

Uma vez que o formulário estava elaborado e fundamentado em uma revisão bibliografiga rigorosa, a pesquisa foi realizada em ambas as modalidades, regular e EJANEM, contou com a participação de 176 alunos do EJANEM e 201 alunos do ensino regular após o descarte de formulários em branco ou preenchido de forma inadequada, totalizou-se um número de 267 formulários somando as duas modalidades.

Com o novo formulário e os resultados gerados por ele, a pesquisa foi ampliada para incluir outras três escolas de ensino regular situadas nas regiões de Dom Bosco e Km 32, no município de Nova Iguaçu, e em Engenheiro Pedreira, no município de Japeri. A inclusão dessas instituições visou a comparação de contextos regionais e socioeconômicos distintos, enriquecendo a análise dos fatores que contribuem para o problema. Também, foram incorporados dados de alunos do EJANEN (Educação de Jovens e Adultos do novo ensiono médio) das regiões de Seropédica e do Rio de Janeiro, ampliando o escopo da pesquisa para diferentes localidades. Nas novas unidades de ensino, a aplicação da pesquisa foi realizada em turmas selecionadas, a fim de não comprometer a logística e o funcionamento regular das escolas. Essa estratégia permitiu ampliar o número de amostras válidas de 267 para 397 amostras válidas. Para transformar todas as informações coletadas acima em um algoritmo capaz de prever o abandono escolar utilizando redes neurais artificiais, foi necessário seguir uma série de etapas que abrangem desde a preparação dos dados até a construção e treinamento do modelo preditivo [13, 14, 15, 16, 17, 18].

Abaixo, descrevo essas etapas:

Preparação dos Dados: A primeira etapa consiste na preparação dos dados coletados para que possam ser utilizados em um modelo de rede neural. Isso envolve várias etapas posteriores:

Limpeza dos Dados: Remoção de entradas duplicadas, tratamento de valores ausentes e correção de inconsistências nos dados. Por exemplo, se algum formulário estiver incompleto ou contiver informações contraditórias, essas entradas devem ser ajustadas ou excluídas.

Normalização e Padronização: Como as redes neurais artificiais são sensíveis à escala dos dados, é necessário normalizar ou padronizar os valores dos atributos, de modo que todos fiquem dentro de uma mesma faixa (por exemplo, entre 0 e 1). Isso ajuda a melhorar a convergência do algoritmo durante o treinamento

Codificação de Variáveis Categóricas: Atributos como "tipo de transporte"ou "turno de estudo"são categóricos e precisaramm ser convertidos em um formato numérico que a rede neural possa entender.

Veja abaixo os detalhes:

Sexo

- M = 0
- F = 1
- Outros = 2

data de nascimento

- dd/mm/aaaa

Escolaridade:

Fundamental

- 1 - 6°
- 2 - 7°
- 3 - 8°
- 4 - 9°

Médio

- 5 - 1°
- 6 - 2°
- 7 - 3°
- 8 - Antes do 6º ano

Deslocamento (tempo em minutos):

transporte

- 1 - Ônibus
- 2 - Trem
- 3 - Carro/ moto
- 4 - Bicicleta
- 5 - Metro
- 6 - Caminhando

Série que o aluno se encontra (ensino regular), ou a série em que ocorreu a evasão(alunos do EJANEM)

- 1° ano = 1
- 2° ano = 2
- 3° ano = 3
- antes do 1° ano do ensino médio = 4
- 0 significa que não sabe ou não preencheu

Renda familiar

- 1 - Inferior a 1 salário mínimo
- 2 - 1 salário mínimo
- 3 - Entre 1 e 2
- 4 - Entre 2 e 3
- 5 - Entre 3 e 4
- 6 - Acima de 4

Participação na vida ecomica da família

- 1 - Não contribui
- 2 - Contribui
- 3 - É o principal responsável pelo sustento filhos
- 1 - sim
- 2 - não

Quantidade filhos: Quantidade exata
Recebe auxílio financeiro

- 1 - Sim
- 2 - Não

Turno

- 1 - Manhã
- 2 - Tarde
- 3 - Noite

frequência escolar

- 1 - Boa
- 2 - Média
- 3 - Ruim

Primeiro responsável

- 1 - Mãe
- 2 - Pai
- 3 - Avó
- 4 - Avô
- 5 - Tia
- 6 - Tio
- 7 - Madrinha
- 8 - Padrinho
- 9 - Madrasta
- 10 - Padrasto
- 11 - irmã
- 12 - irmão
- 13 - Não tenho
- 14 - o próprio
- 15 - Prima

Escolaridade desse responsável

- 1 - Não frequentou escola
- 2 - Fundamental incompleto
- 3 - Fundamental completo
- 4 - Ensino médio incompleto
- 5 - Ensino médio completo
- 6 - Superior incompleto
- 7 - Superior completo
- 8 - Pós graduação

Segundo responsável

- 1 - Mãe
- 2 - Pai
- 3 - Avó
- 4 - Avô
- 5 - Tia
- 6 - Tio
- 7 - Madrinha
- 8 - Padrinho
- 9 - Madrasta
- 10 - Padrasto
- 11 - irmã
- 12 - irmão
- 13 - Não tenho
- 14 - o próprio
- 15 - Prima

Escolaridade desse responsável

- 1 - Não frequentou escola
- 2 - Fundamental incompleto
- 3 - Fundamental completo

- 4 - Ensino médio incompleto
- 5 - Ensino médio completo
- 6 - Superior incompleto
- 7 - Superior completo
- 8 - Pós graduação

Tipo de atividade realizada pelo responsável

- 1 - Formal
- 2 - Informal constante
- 3 - Informal esporádico
- 4 - Não possui

Mudança frequente de endereço

- 1 - Nunca se mudou
- 2 - Com muita frequência
- 3 - Raramente
- 4 - Em alguns períodos me mudava com frequência, em outros não

Planos para o futuro

- 1 - Curso pré-preparatório
- 2 - Serviço Militar
- 3 - Estudar para concursos
- 4 - Trabalhar e fazer curso pré-preparatório
- 5 - Curso profissionalizante
- 6 - Não penso sobre isso
- 7 - Apenas trabalhar
- 8 - Ingressar direto na faculdade

Encontra apoio dos responsáveis para realização desses palnos

- 1 - Sim

- 2 - Não

Compreende a importância da escola na sua formação

- 1 - Sim
- 2 - Não

Quanto a participação que os responsáveis desempenham na vida escola, o aluno deveria responder sim ou não para os questioamentos abaixo:

- 1- Sim
- 2- Não
- Frequentar conversas com os professores sobre o desempenho do aluno =
- Olhar os cadernos do aluno
- Preocupação com a frequência do aluno
- Participar das reuniões escolares
- Incentivam a dedicação do aluno aos estudos fora de sala de aula
- Acompanham boletim

Quanto aos fatores abaixo, os aluos devem responser sim ou não se pertencem a sua realidade

- 1- Sim
- 2- Não
- Distância da casa até a escola
- Dificuldade de compreender as aulas
- Reprovação ou notas baixas
- Gestação
- Problemas financeiros
- Falta de interesse
- Mudança de endereço
- Separação dos pais
- Ajuda na criação dos irmãos
- Tarefas domésticas que te tomam muito tempo
- Falta de estímulo

- Problemas familiares
- Trabalho
- Violência no bairro em que mora
- Violência no entorno da escola
- Sofre discriminação na escola
- Falta de merenda escolar

Divisão dos Dados: Após a preparação dos dados, eles são divididos em conjuntos de treino e teste:

Conjunto de Treino: Representa a maior parte dos dados (70%) e é utilizado para treinar a rede neural.

Conjunto de Teste: Mantido separado do treinamento, é usado para avaliar o desempenho final do modelo, garantindo que ele seja capaz de generalizar para novos dados. Esse conjunto representa 30% dos dados.

Construção da Rede Neural: Com os dados prontos, a próxima etapa é a construção da rede neural

Escolha da Arquitetura: A rede neural será composta por camadas de neurónios. Cada camada é conectada à próxima através de pesos que são ajustados durante o treinamento. A arquitetura básica inclui uma camada de entrada (correspondente ao número de atributos), uma ou mais camadas ocultas (responsáveis por capturar padrões complexos) e uma camada de saída (que fornecerá a predição final).

Função de Ativação: Cada neurónio utiliza uma função de ativação, como a ReLU (Rectified Linear Unit) ou a sigmoide, para introduzir não-linearidades no modelo, o que é essencial para capturar relações complexas nos dados.

Treinamento da Rede Neural: Durante o treinamento, o modelo ajusta os pesos das conexões entre os neurónios para minimizar o erro entre as predições do modelo e os valores reais:

Função de Custo: O erro é quantificado através de uma função de custo, como a entropia cruzada, que mede a discrepância entre as predições da rede e os dados reais de abandono escolar.

Otimização Um algoritmo de otimização, como o gradiente descendente, é utilizado para ajustar os pesos da rede de modo a minimizar a função de custo. O processo de ajuste ocorre iterativamente, com cada iteração denominada "época".

Regularização: Para evitar o sobreajuste, técnicas de regularização, como dropout (onde uma fração de neurónios é desativada aleatoriamente durante o treinamento) podem ser aplicadas.

4.1 Avaliação e Validação do Modelo

Após o treinamento, o modelo é avaliado utilizando o conjunto de teste para verificar seu desempenho em dados que não foram vistos durante o treinamento. Métricas como a acurácia, são utilizadas para quantificar a eficácia do modelo. Durante a fase de avaliação e validação do modelo, é essencial medir a frequência com que o modelo gera falsos positivos e falsos negativos. Isso é feito utilizando métricas como a matriz de confusão, que contabiliza corretamente

as predições positivas e negativas, bem como os erros (falsos positivos e falsos negativos). Durante o treinamento da rede neural, técnicas como ajuste de hiperparâmetros e regularização são utilizadas para minimizar tanto os falsos positivos quanto os falsos negativos, aumentando a precisão e a robustez do modelo. Mas vale ressaltar que a presença de falsos positivos e falsos negativos no desempenho de um modelo preditivo é inevitável, mas a meta é minimizar esses erros para garantir que o modelo seja uma ferramenta confiável na identificação de alunos em risco de abandono escolar. Avaliar e ajustar o modelo com foco nesses aspectos é fundamental para o desenvolvimento de um sistema preditivo eficaz e para a implementação de estratégias educacionais que realmente façam a diferença [19] [20] [6].

4.2 Validação Cruzada

A validação cruzada é uma técnica essencial para testar a qualidade de modelos de aprendizado de máquina. Seu objetivo principal é verificar o desempenho do modelo de forma mais confiável, reduzindo o risco de avaliações enviesadas que podem ocorrer devido a uma divisão específica dos dados [19][5].

Considere um conjunto de dados que deseja usar para treinar e testar um modelo. Dividir esses dados uma única vez em dois grupos (treinamento e teste) pode gerar resultados dependentes dessa divisão específica, o que não é ideal. A validação cruzada resolve esse problema ao dividir os dados em várias partes. Cada parte é usada alternadamente para testar o modelo, enquanto as demais são usadas para treiná-lo. Isso garante que todas as partes dos dados sejam usadas tanto para o treinamento quanto para o teste.

Neste estudo, utilizamos a função cross_val_score da biblioteca Scikitlearn para realizar a validação cruzada de forma prática e automatizada. Aqui está como ela funciona:

Divisão dos dados: Com o parâmetro cv=3, o conjunto de dados foi dividido em três partes iguais. Em cada iteração:

- Duas partes são usadas para treinar o modelo.
- Uma parte é usada para testar o modelo. Esse processo se repete três vezes, alternando a parte usada como teste.

Execução e avaliação: A função aplica o modelo definido em cada uma dessas três divisões. Durante o processo, o modelo é ajustado aos dados de treinamento e testado na parte reservada para avaliação.

Resultados armazenados: Os resultados das avaliações (como a acurácia ou outras métricas de desempenho) são armazenados em uma variável chamada scores. Essa variável contém os valores de cada iteração, permitindo calcular, por exemplo, a média e o desvio padrão, para obter uma visão mais abrangente do desempenho do modelo.

A validação cruzada oferece vantagens como minimizar o impacto de particionamentos específicos do conjunto de dados na avaliação do modelo, o uso mais eficiente dos dados e uma análise mais robusta do desempenho do modelo.

Implementação e Utilização: Uma vez validado, o modelo pode ser implementado em sistemas educacionais para prever o abandono escolar em tempo real. Dados novos de estudantes podem ser inseridos no modelo, que gerará uma probabilidade de abandono escolar. Esses resultados podem então ser utilizados para tomar decisões proativas, como fornecer suporte adicional aos alunos em risco.

Aprimoramento Contínuo: É essencial monitorar continuamente o desempenho do modelo e atualizá-lo com novos dados conforme necessário, para assegurar que ele mantenha

sua eficácia ao longo do tempo. A escolha das redes neurais como uma opção preferencial pode ser justificada por diversas razões. Redes neurais, especialmente as profundas, possuem uma capacidade superior para modelar e capturar complexidades nos dados. A arquitetura flexível das redes neurais permite a criação de modelos com múltiplas camadas de abstração, possibilitando a aprendizagem de representações hierárquicas dos dados [13] [4].

4.3 Implementação do Algoritmo de Redes Neurais e outros Algortmos Supervisionados

Os resultados obtidos na pesquisa demonstram uma análise detalhada da eficácia de diferentes modelos de aprendizado supervisionado na previsão da evasão escolar. Inicialmente, foi realizado um estudo com aproximadamente 200 alunos da modalidade EJANEM, utilizando informações já existentes na escola sobre aproximadamente 300 alunos do ensino regular. Devido a discrepâncias nos atributos, foi necessário restringir o uso a algumas colunas do dataset. Esta primeira abordagem resultou em uma acurácia de 79%, apesar da limitação no número de atributos.

Uma nova pesquisa foi realizada contando com 176 alunos do EJANEM e 201 do ensino regular, totalizando 267 formulários válidos. Posteriormente essa amostra foi expandida para outras escolas e outros municípios, totalizando uma quantidade de 397 formulários válidos. Os dados foram divididos em 70% para treinamento e 30% para teste do modelo. O processo de implementação foi conduzido em Python, utilizando bibliotecas como pandas, numpy, seaborn e scikit-learn (sklearn). A categorização das variáveis foi realizada através da conversão de dados categóricos em numéricos, e os valores ausentes foram preenchidos com -1, garantindo a integridade dos dados para análise [6]. Após o pré-processamento e a transformação dos dados, alguns modelos de classificação foram analisados.

Com o avanço de técnicas de regularização, otimização e a disponibilidade de poder computacional, as redes neurais têm demonstrado um excelente desempenho em uma ampla gama de tarefas, especialmente em problemas com grandes volumes de dados e alta dimensionalidade. Além de sua capacidade de lidar com a complexidade dos dados e a eficácia em extrair padrões complexos, o que pode, em alguns casos, superar as limitações de métodos tradicionais de aprendizado supervisionado [4][5].

No entanto, para fins de comparação, o modelo também foi ampliado para incluir diversos algoritmos de aprendizado supervisionado, como árvores de decisão, k-Nearest Neighbors (k-NN) e Máquinas de Vetores de Suporte (SVM). Cada um desses métodos apresenta características distintas que influenciam a sua aplicabilidade e desempenho em diferentes cenários.

Árvores de decisão são conhecidas por sua simplicidade e fácil interpretação, permitindo a visualização clara das regras de decisão e o entendimento das razões por trás das previsões [22]. No entanto, elas podem sofrer com overfitting, especialmente em datasets com muitas variáveis e instâncias.

O k-Nearest Neighbors (k-NN) é um método baseado em instâncias que realiza classificações ou regressões com base nas características dos vizinhos mais próximos. Embora seja intuitivo e eficaz em situações onde a estrutura dos dados é bem definida, o kNN pode ser computacionalmente intenso e sofrer com a questão da dimensionalidade em datasets grandes e complexos [5].

Máquinas de Vetores de Suporte (SVM) são valorizadas por sua capacidade de lidar com problemas de alta dimensionalidade e por sua eficácia em encontrar a margem ótima de separação entre classes. Entretanto, a escolha adequada do kernel e a otimização de parâmetros são cruciais para o desempenho do modelo [23].

Para avaliar o desempenho de modelos de classificação em aprendizado de máquina é necessário observar uma ferramenta conhecida como matriz de confusão. Ela é usada para verificar quantas vezes o modelo acerta ou erra em cada classe de uma classificação, apresentando as previsões do modelo em comparação com os resultados reais [19] [20]. Aqui estão os componentes principais de uma matriz de confusão para um modelo de classificação binária (com duas classes: positivo e negativo):
1. Verdadeiros Positivos (VP): Número de instâncias corretamente classificadas como positivas. Exemplo: O modelo prevê que um aluno irá evadir, e isso realmente ocorre.
2. Falsos Positivos (FP): Número de instâncias incorretamente classificadas como positivas. Exemplo: O modelo prevê que um aluno irá evadir, mas ele não evade. Esse tipo de erro é também chamado de erro tipo I.
3. Verdadeiros Negativos (VN): Número de instâncias corretamente classificadas como negativas. Exemplo: O modelo prevê que um aluno não irá evadir, e ele realmente não evade.
4. Falsos Negativos (FN): Número de instâncias incorretamente classificadas como negativas. Exemplo: O modelo prevê que um aluno não irá evadir, mas ele evade. Esse erro é conhecido como erro tipo II.
Essas informações são organizadas em uma matriz 2x2, assim:

$$\mathbf{A} = \begin{bmatrix} VP & FP \\ FN & VN \end{bmatrix}$$

Para analisar o desempenho com base na matriz de confusão fornecida e na acurácia do modelo, vamos detalhar as métricas derivadas da matriz e a interpretação geral dos resultados.
Métricas Derivadas:
Acurácia: A acurácia é a proporção de classificações corretas (tanto positivas quanto negativas) em relação ao total de amostras.

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{Total de Amostras}}$$

Precisão (para a classe positiva): A precisão mede a proporção de verdadeiros positivos em relação ao total de casos classificados como positivos.

$$\text{Precisão} = \frac{\text{VP}}{\text{VP} + \text{FP}}$$

Recall (ou Sensibilidade) (para a classe positiva): O recall mede a proporção de verdadeiros positivos em relação ao total de casos que realmente pertencem à classe positiva.

$$\text{Recall} = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

Especificidade (para a classe negativa): A especificidade mede a proporção de verdadeiros negativos em relação ao total de casos que realmente pertencem à classe negativa.

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}}$$

O F1-Score é uma métrica harmônica que considera tanto a precisão quanto o recall de um modelo de classificação. É particularmente útil em casos onde há um desbalanceamento entre as classes.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precisão} \cdot \text{Recall}}{\text{Precisão} + \text{Recall}} \tag{4.1}$$

A acurácia da validação cruzada é uma estimativa mais robusta do desempenho de um modelo, evitando o overfitting. Ela é calculada como a média das acurácias obtidas em cada fold da validação cruzada.

$$\text{Acurácia da Validação Cruzada} = \frac{1}{k} \sum_{i=1}^{k} \text{Acurácia}_i \tag{4.2}$$

onde: * k: número de folds * Acurácia_i: acurácia obtida no i-ésimo fold
Com base os cálculos acima chegamos nos seguintes resultados.
A seguir podemos observar o desempenho de cada um dos modelos:

4.4 Resultados de Modelos de Classificação

Este resultados foram obtidos por diferentes modelos de classificação: Árvore de Decisão, Redes Neurais, KNN, SVM. Para cada método, são fornecidas as métricas de desempenho calculadas a partir da matriz de confusão, bem como uma tabela comparativa com os principais indicadores. O atributo turno não foi considerado, pois muitos alunos do EJANEM preencheram com 3º turno, que mostrava uma relação direta com o abandono.

Os dados utilizados neste estudo foram coletados em diferentes amostras de alunos para prever o abandono escolar. As amostras variam em tamanho e diversidade socioeconômica, com a primeira amostra conta com um preenchimento melhor e mais variedade de séries, enquanto a amostra subsequente possue dados adicionais, porém com menor variedade. Este documento apresenta um comparativo dos modelos treinados com as duas amostras e discute os impactos da qualidade e diversidade dos dados nos resultados.

Características das Amostras

- **Amostra 1:** A primeira amostra foi composta por 268 dados e coletada em uma única escola pública. Esta escola atende um público de ampla diversidade econômica, conforme exemplificado no gráfico abaixo.

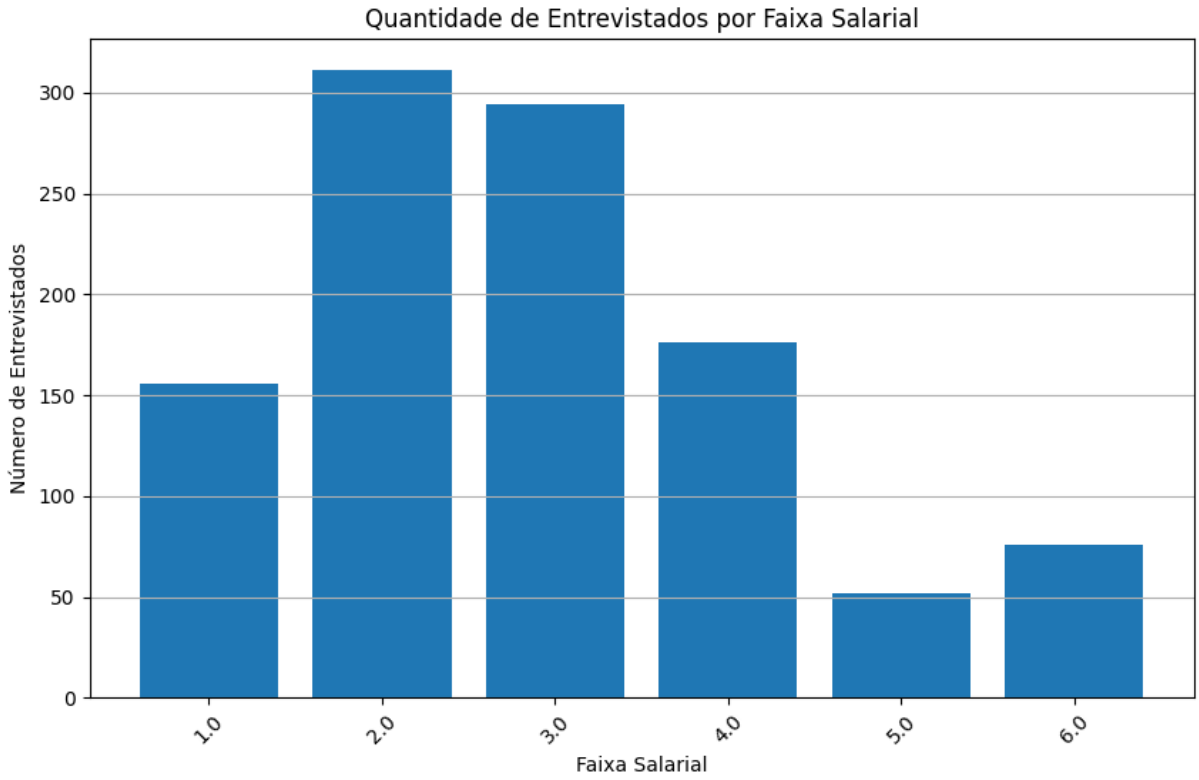


Figura 4.3 – Quantidade de Alunos por Faixa Salarial.
Fonte: autor

Além disso, a escola está localizada em uma região central, o que a torna acessível para estudantes de várias regiões adjacentes, incluindo zonas urbanas e rurais. Evidenciando a diversidade de contextos socioeconômicos e socioculturais dos alunos como é possivel observar na figura abaixo.

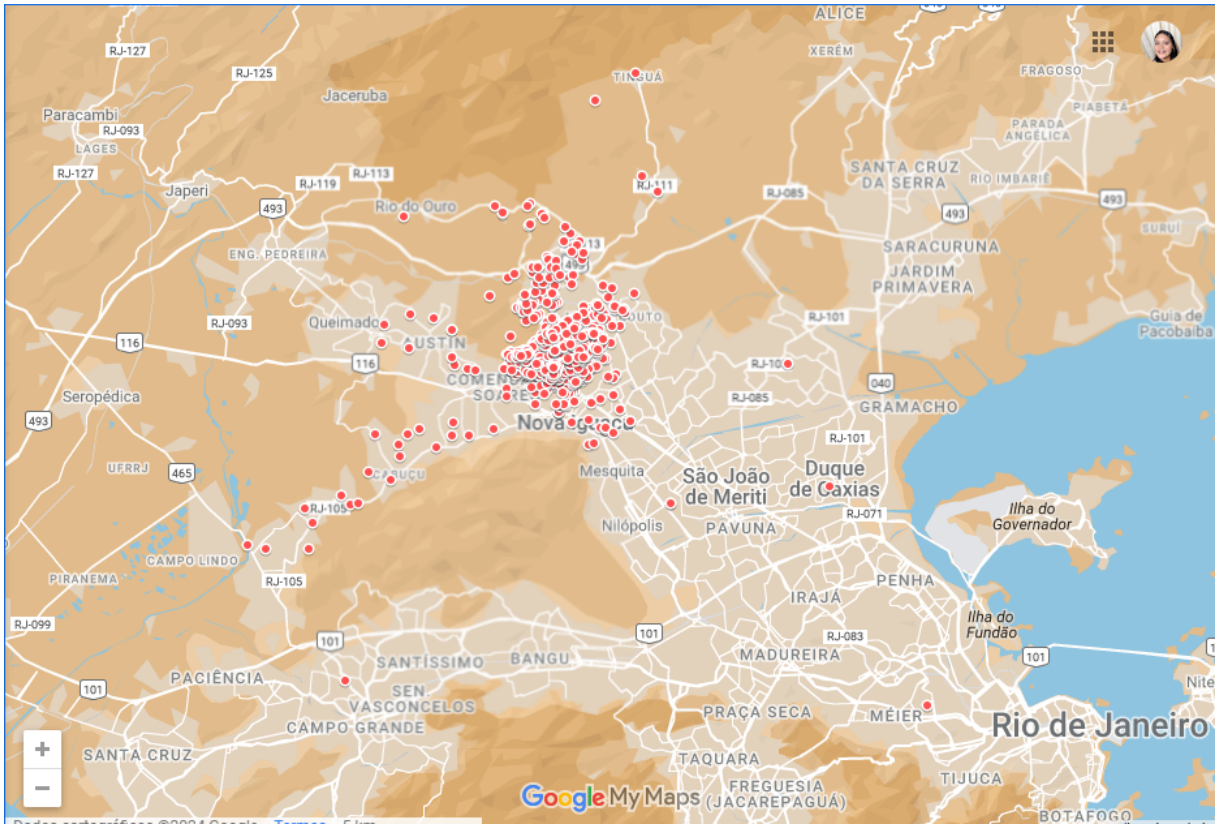


Figura 4.4 – Mapa de Análise Espacial.
Fonte: autor

Outro diferencial importante desta amostra é a maior variedade de séries escolares contempladas, abrangendo alunos de diferentes idades e estágios educacionais. Adicionalmente, o preenchimento mais completo dos dados nesta amostra resultou em informações precisas e confiáveis, com menos ruídos em comparação a outra amostra analisadas.

- **Amostra 2:** 397 formulários, adicionando 129 formulários válidos à primeira amostra. Esses dados adicioais possuem uma menor variedade turmas analisadas e apresentam menor completude no preenchimento.

Resultados dos Modelos - Amostra 1

Redes Neurais

$$\text{Matriz de Confusão} = \begin{bmatrix} 47 & 5 \\ 6 & 18 \end{bmatrix} \tag{4.3}$$

VP = 47, FP = 5, VN = 18, FN = 6.
Acurácia: 0.8553
Acurácia da Validação Cruzada: 0.8526

Árvore de Decisão

$$\text{Matriz de Confusão} = \begin{bmatrix} 42 & 10 \\ 4 & 20 \end{bmatrix} \tag{4.4}$$

VP = 42, FP = 10, VN = 20, FN = 4.
Acurácia: 0.8158
Acurácia da Validação Cruzada: 0.8184

KNN

$$\text{Matriz de Confusão} = \begin{bmatrix} 49 & 3 \\ 21 & 3 \end{bmatrix} \tag{4.5}$$

VP = 49, FP = 3, VN = 3, FN = 21.
Acurácia: 0.6842
Acurácia da Validação Cruzada: 0.7711

SVM

$$\text{Matriz de Confusão} = \begin{bmatrix} 46 & 6 \\ 6 & 18 \end{bmatrix} \tag{4.6}$$

VP = 46, FP = 6, VN = 18, FN = 6.
Acurácia: 0.8421
Acurácia da Validação Cruzada: 0.8553

Resultados dos Modelos - Amostra 2

Redes Neurais

$$\text{Matriz de Confusão} = \begin{bmatrix} 47 & 5 \\ 1 & 24 \end{bmatrix} \tag{4.7}$$

Verdadeiros Positivos (VP) = 47, Falsos Positivos (FP) = 5, Verdadeiros Negativos (VN) = 24, Falsos Negativos (FN) = 1.
Acurácia: 0.9221
Acurácia da Validação Cruzada: 0.8427

Árvore de Decisão

$$\text{Matriz de Confusão} = \begin{bmatrix} 46 & 6 \\ 7 & 18 \end{bmatrix} \tag{4.8}$$

VP = 46, FP = 6, VN = 18, FN = 7.
Acurácia: 0.8312
Acurácia da Validação Cruzada: 0.7902

KNN

$$\text{Matriz de Confusão} = \begin{bmatrix} 50 & 2 \\ 13 & 12 \end{bmatrix} \tag{4.9}$$

VP = 50, FP = 2, VN = 12, FN = 13.
Acurácia: 0.8052
Acurácia da Validação Cruzada: 0.7717

SVM

$$\text{Matriz de Confusão} = \begin{bmatrix} 47 & 5 \\ 6 & 19 \end{bmatrix} \tag{4.10}$$

VP = 47, FP = 5, VN = 19, FN = 6.
Acurácia: 0.8571
Acurácia da Validação Cruzada: 0.8611

Tabela Comparativa das Métricas

A tabela a seguir apresenta uma comparação entre os modelos avaliados (Redes Neurais, Árvore de Decisão, KNN e SVM) em termos das principais métricas de desempenho, incluindo Recall, Precisão, Especificidade, F1-Score, Acurácia e Acurácia da Validação Cruzada para diferentes tamanhos de amostras.

Modelo	Amostra	Acurácia	Val. Cruzada	Precisão	Recall	Especificidade	F1-Score
Redes Neurais	Amostra 1	0.8553	0.8526	0.9038	0.8868	0.7826	0.8952
	Amostra 2	0.9221	0.8427	0.9038	0.9792	0.8276	0.9400
Árvore de Decisão	Amostra 1	0.8158	0.8184	0.8077	0.9130	0.6667	0.8571
	Amostra 2	0.8312	0.7902	0.8846	0.8679	0.7500	0.8762
KNN	Amostra 1	0.6842	0.7711	0.9423	0.7000	0.5000	0.8033
	Amostra 2	0.8052	0.7717	0.9615	0.7937	0.8571	0.8696
SVM	Amostra 1	0.8421	0.8553	0.8846	0.8846	0.7500	0.8846
	Amostra 2	0.8571	0.8611	0.9038	0.8868	0.7917	0.8952

Tabela 4.1 – Métricas de desempenho dos modelos para as duas amostras.

Análise de Desempenho dos Modelos

Redes Neurais

- **Amostra 1:** O modelo obteve uma Acurácia de 85,53% e uma Acurácia de Validação Cruzada de 85,26%, indicando uma boa generalização. A Precisão de 90% e o Recall de 89% mostram um equilíbrio adequado na identificação de casos positivos e negativos. A Especificidade de 78% sugere que alguns negativos foram classificados incorretamente, enquanto o F1-score de 90% confirma um bom equilíbrio entre precisão e recall.
- **Amostra 2:** A Acurácia aumentou para 92,21%, mas a Acurácia de Validação Cruzada caiu para 84,27%, indicando um leve sobreajuste ao novo conjunto de dados. O Recall

aumentou significativamente para 98%, indicando uma excelente identificação de verdadeiros positivos, mas a Especificidade caiu para 83%, sugerindo um aumento nos falsos positivos. A Precisão manteve-se em 90%, e o F1-score foi o mais alto entre os modelos, 94%, reforçando sua consistência.

Árvore de Decisão

- **Amostra 1:** A Acurácia foi de 81,58%, com uma Acurácia de Validação Cruzada de 81,84%, mostrando boa estabilidade. O Recall de 91% sugere que o modelo identifica corretamente a maioria dos verdadeiros positivos, mas a Precisão de 81% indica um número considerável de falsos positivos. A Especificidade de 67% reforça essa dificuldade em distinguir os verdadeiros negativos, enquanto o F1-score de 0,86% mostra um equilíbrio razoável do modelo.

- **Amostra 2:** O modelo melhorou a Precisão para 88%, mas o Recall caiu para 87%, sugerindo um equilíbrio maior entre as classificações. A Acurácia aumentou para 83,12%, mas a Acurácia de Validação Cruzada caiu para 79,02%, sugerindo maior sensibilidade ao ruído dos novos dados. A Especificidade caiu para 75%, mostrando que houve um leve aumento nos falsos positivos.

k-Nearest Neighbors (K-NN)

- **Amostra 1:** O modelo teve a menor Acurácia (68,42%) entre os modelos, com uma Acurácia de Validação Cruzada de 77,11%. A Precisão de 94% foi alta, mas o Recall de 70% mostrou dificuldades em identificar corretamente positivos. A baixa Especificidade de 50% indica que muitos negativos foram classificados erroneamente, resultando em um F1-score de 80%.

- **Amostra 2:** O desempenho melhorou, com a Acurácia subindo para 80,52% e a Acurácia de Validação Cruzada para 77,17%. A Precisão aumentou para 96% e o Recall para 79%, resultando em um F1-score de 87%. A Especificidade de 85% sugere uma melhor distinção entre positivos e negativos. Esse comportamento pode ser explicado pelo menor número de turmas na segunda amostra, reduzindo a variabilidade dos dados e favorecendo um modelo baseado em proximidade.

SVM (Support Vector Machines)

- **Amostra 1:** A Acurácia foi de 84,21%, com uma Acurácia de Validação Cruzada de 85,53%. A Precisão e o Recall equilibrados em 88% mostram um bom desempenho geral. A Especificidade de 75% indica que alguns negativos foram classificados incorretamente, mas o F1-score de 88% confirma um desempenho estável.

- **Amostra 2:** O modelo manteve uma Acurácia alta (85,71%), com uma Acurácia de Validação Cruzada de 86,11%. A Precisão de 90% e o Recall de 89% mostram um excelente equilíbrio na classificação. A Especificidade caiu para 79%, indicando um pequeno aumento de falsos positivos, mas o F1-score de 90% reforça sua consistência do modelo.

Influência das Amostras no Desempenho dos Modelos

A primeira amostra, composta por 268 registros, apresentava maior diversidade socioeconômica e completude dos dados, permitindo um melhor treinamento dos modelos. Essa característica favoreceu modelos mais estruturados como SVM e Redes Neurais, que apresentaram desempenho mais equilibrado.

A segunda amostra adicionou 129 novos registros, totalizando 397, mas com menor variedade de turmas e maior incompletude nos dados. Isso impactou diretamente a Especificidade de alguns modelos, resultando em mais falsos positivos. Modelos como KNN, que dependem da distribuição dos dados, se beneficiaram dessa menor variabilidade, enquanto modelos mais robustos, como SVM e Redes Neurais, mantiveram um desempenho estável e confiável.

4.5 Identificação de Padrões de Abandono Escolar por Idade

O gráfico abaixo representa a quantidade de abandono por idade. Essa informação revela padrões importantes sobre a dinâmica da evasão escolar ao longo do tempo. A identificação de faixas etárias mais vulneráveis pode subsidiar políticas públicas e intervenções específicas voltadas à retenção de estudantes.

Figura 4.5 – Gráfico referente a quantidade de abandono por idade.
Fonte: autor

O gráfico mostra uma tendência clara de aumento na taxa de abandono a partir dos 13 anos, com uma elevação contínua até alcançar o pico que ocorre entre 17 e 18 anos, marcando esta faixa etária como a de maior evasão escolar. A partir dos 19 anos, observa-se uma redução na taxa de abandono, evidenciando uma diminuição na evasão escolar após essa idade.

Análise de Correlação

A correlação é uma ferramenta estatística utilizada para medir o grau de relação linear entre diferentes variáveis em um conjunto de dados. O coeficiente de correlação varia de -1 a 1, onde valores próximos a 1 ou -1 indicam uma correlação forte (positiva ou negativa, respectivamente), enquanto valores próximos a 0 indicam pouca ou nenhuma correlação.

Na Amostra 1, observou-se uma forte correlação entre as variáveis. Isso facilita o trabalho dos modelos de aprendizado, pois os padrões mais evidentes nas relações entre variáveis podem ser diretamente utilizados para melhorar a precisão das previsões.

Por outro lado, na Amostra 2, a correlação entre as variáveis foi consideravelmente mais fraca, dificultando a extração de padrões claros e aumentando os desafios para os modelos de aprendizado. Apesar disso, a robustez de alguns modelos, como Redes Neurais, pode permitir a identificação de padrões complexos mesmo em cenários com baixa correlação.

4.6 Avaliação de Atributos por Série

Vamos nos concentrar na análise dos padrões de abandono escolar por série, buscando identificar se um aluno irá abandonar a escola e, em caso positivo, em qual etapa educacional isso ocorrerá. Isso corresponde a um problema de classificação multiclasse.

Nesse caso, a análise considera nove classes distintas que representam os diferentes momentos em que o abandono pode ocorrer, além de uma classe específica para alunos que permanecem na escola.

As classes representam as séries em que os alunos abandonaram os estudos. Por exemplo:

Classes do Problema de Abandono Escolar:

- **Classe 1:** Alunos que abandonaram antes do 6º ano do ensino fundamental.
- **Classe 2:** Alunos que abandonaram no 6º ano do ensino fundamental.
- **Classe 3:** Alunos que abandonaram no 7º ano do ensino fundamental.
- **Classe 4:** Alunos que abandonaram no 8º ano do ensino fundamental.
- **Classe 5:** Alunos que abandonaram no 9º ano do ensino fundamental.
- **Classe 6:** Alunos que abandonaram no 1º ano do ensino médio.
- **Classe 7:** Alunos que abandonaram no 2º ano do ensino médio.
- **Classe 8:** Alunos que abandonaram no 3º ano do ensino médio.
- **Classe 9:** Alunos que não abandonaram a escola.

A matriz confusa apresentada é de dimensão 9×9, correspondendo às 9 classes do problema, que representam os diferentes momentos de abandono ou a ausência de abandono. Cada entrada na matriz fornece informações importates sobre o desempenho do modelo em prever a classe correta para cada caso.

Matrizes de Confusão e Acurácias

Redes Neurais

Amostra 1

$$\text{Matriz de Confusão} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 4 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 15 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \end{bmatrix} \tag{4.11}$$

Acurácia do modelo: 0.375
Acurácia média da validação cruzada: 0.3649

Amostra 2

$$\text{Matriz de Confusão} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 3 & 0 & 2 & 0 & 0 \\ 0 & 1 & 0 & 0 & 5 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 12 & 2 & 7 & 1 & 0 \\ 0 & 0 & 0 & 0 & 3 & 3 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 3 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 2 & 0 & 0 & 0 \end{bmatrix} \tag{4.12}$$

Acurácia do modelo: 0.3117
Acurácia média da validação cruzada: 0.3543

Árvore de Decisão

Amostra 1

$$\text{Matriz de Confusão} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 6 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 3 & 0 & 2 \\ 2 & 0 & 0 & 0 & 4 & 0 & 9 & 1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{4.13}$$

Acurácia do modelo: 0.2292
Acurácia média da validação cruzada: 0.3139

Amostra 2

$$\text{Matriz de Confusão} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 2 & 1 & 3 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 6 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 15 & 2 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 1 & 1 & 0 & 4 \\ 0 & 0 & 6 & 0 & 7 & 3 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \end{bmatrix} \tag{4.14}$$

Acurácia do modelo: 0.2727
Acurácia média da validação cruzada: 0.3148

k-Nearest Neighbors (KNN)

Amostra 1

$$\text{Matriz de Confusão} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 1 & 1 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 1 & 13 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix} \tag{4.15}$$

Acurácia do modelo: 0.3125
Acurácia média da validação cruzada: 0.3399

Amostra 2

$$\text{Matriz de Confusão} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 & 1 & 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 & 4 & 0 & 2 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 14 & 0 & 6 & 0 & 0 \\ 0 & 1 & 1 & 0 & 3 & 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 10 & 2 & 6 & 0 & 0 \\ 0 & 0 & 2 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 2 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.16}$$

Acurácia do modelo: 0.3377
Acurácia média da validação cruzada: 0.2914

Support Vector Machines (SVM)

Amostra 1

$$\text{Matriz de Confusão} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 7 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 16 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 0 \end{bmatrix} \tag{4.17}$$

Acurácia do modelo: 0.3333
Acurácia média da validação cruzada: 0.3525

Amostra 2

$$\text{Matriz de Confusão} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 3 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 5 & 0 & 3 & 0 & 0 \\ 0 & 1 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 1 & 13 & 1 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 1 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 & 7 & 3 & 4 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 2 & 1 & 0 & 0 & 0 \end{bmatrix} \tag{4.18}$$

Acurácia do modelo: 0.2597
Acurácia média da validação cruzada: 0.3204

Análise Comparativa dos Modelos de Aprendizado de Máquina

Nesta seção, comparamos o desempenho dos quatro modelos de aprendizado de máquina (Redes Neurais, Árvore de Decisão, k-Nearest Neighbors e Support Vector Machines) em duas amostras diferentes. A avaliação é baseada nas métricas de acurácia, acurácia da validação cruzada e matrizes de confusão.

Redes Neurais

- **Amostra 1**: A acurácia do modelo foi de 37,5%, com uma acurácia da validação cruzada de 36,49%. A matriz de confusão mostra dificuldades em prever corretamente as classes,

resultando em uma performance limitada.

- **Amostra 2**: O desempenho caiu para uma acurácia de 31,17%, com a validação cruzada ainda menor (35,43%). A inclusão de novos dados pode ter introduzido ruídos, reduzindo a capacidade preditiva do modelo.

Árvore de Decisão

- **Amostra 1**: Com acurácia de 22,92% e uma acurácia da validação cruzada de 31,39%, a árvore de decisão teve dificuldades em aprender padrões consistentes, provavelmente devido à alta variabilidade da amostra.

- **Amostra 2**: A acurácia aumentou para 27,27%, enquanto a acurácia da validação cruzada permaneceu praticamente inalterada (31,48%). Isso indica que, embora o modelo tenha melhorado sua adaptação aos novos dados, a generalização ainda é um desafio.

k-Nearest Neighbors (KNN)

- **Amostra 1**: A acurácia foi de 31,25%, com uma acurácia da validação cruzada de 33,99%. O modelo se mostrou sensível à distribuição dos dados e teve dificuldades em classificar corretamente as instâncias.

- **Amostra 2**: O desempenho foi ligeiramente superior, com uma acurácia de 33,77% e uma validação cruzada de 29,14%. O KNN parece ter sido menos afetado pela adição de novos dados, mas ainda apresenta dificuldades em capturar padrões robustos.

Support Vector Machines (SVM)

- **Amostra 1**: Com uma acurácia de 33,33% e uma validação cruzada de 35,25%, o SVM apresentou um desempenho relativamente melhor que os outros modelos nesta amostra, indicando uma melhor separação das classes.

- **Amostra 2**: A acurácia caiu para 25,97%, com uma validação cruzada de 32,04%. Isso sugere que a complexidade adicional dos novos dados impactou negativamente a capacidade do modelo de generalizar corretamente.

Tabela Comparativa das Métricas dos Modelos

Modelo	Amostra	Acurácia	Val. Cruzada
Redes Neurais	Amostra 1	0.3750	0.3649
	Amostra 2	0.3117	0.3543
Árvore de Decisão	Amostra 1	0.2292	0.3139
	Amostra 2	0.2727	0.3148
k-Nearest Neighbors	Amostra 1	0.3125	0.3399
	Amostra 2	0.3377	0.2914
Support Vector Machines	Amostra 1	0.3333	0.3525
	Amostra 2	0.2597	0.3204

Tabela 4.2 – Comparação das métricas de Acurácia e Validação Cruzada entre os diferentes modelos de aprendizado de máquina.

Análise dos Modelos nas Amostras 1 e 2

A análise comparativa dos modelos nas Amostras 1 e 2 evidencia o impacto da qualidade dos dados no desempenho dos algoritmos de aprendizado de máquina. A Amostra 1, caracterizada por uma maior completude e diversidade de dados, resultou em desempenhos relativamente superiores para todos os modelos. Em contrapartida, a Amostra 2, embora maior em volume, apresentou maior ruído e menor variedade de turmas, levando a uma redução significativa na acurácia e na validação cruzada de todos os modelos.

Os modelos SVM e KNN foram os mais afetados pela degradação da qualidade dos dados, registrando as maiores quedas de desempenho na Amostra 2. A Árvore de Decisão, embora tenha apresentado desempenhos inferiores de forma geral, mostrou uma menor variação entre as amostras. Já as Redes Neurais, apesar de sofrerem impacto negativo, mantiveram um desempenho relativamente estável.

Esses resultados reforçam a importância de trabalhar com bases de dados de alta qualidade na construção de modelos robustos. A escolha do modelo mais adequado deve considerar não apenas o volume de dados disponíveis, mas também sua completude e diversidade, uma vez que ruídos e desbalanceamentos podem comprometer a capacidade preditiva dos algoritmos.

4.7 Propostas para Projetos Futuros: Construção de Subdatasets

Para aprofundar a compreensão sobre os fatores que influenciam a evasão escolar, uma abordagem promissora para estudos futuros é a criação de subdatasets. Essa estratégia consiste na seleção de subconjuntos de atributos, permitindo a análise segmentada de diferentes fatores de risco.

Ao adotar essa estratégia, algumas possibilidades incluem:

- **Perfis socioeconômicos:** Analisar especificamente o impacto de variáveis como renda familiar, acesso a programas sociais e disponibilidade de transporte escolar.
- **Desempenho acadêmico:** Avaliar como notas, frequência e participação em atividades extracurriculares influenciam a evasão.

- **Fatores geográficos:** Investigar como aspectos como violência no entorno da escola ou da residência do aluno, bem como a distância até a instituição de ensino, podem impactar a permanência estudantil.

A experimentação com subdatasets pode aprimorar os modelos preditivos e viabilizar informações mais precisas para distintos perfis de alunos, permitindo a criação de estratégias personalizadas para a redução da evasão escolar.

5 Conclusão

Conclusão

A análise dos dados e a avaliação dos diferentes modelos de aprendizado supervisionado aplicados à previsão da evasão escolar indicam que as **Redes Neurais** se destacam como uma excelente escolha entre os métodos testados. Com uma acurácia de aproximadamente 85% na amostra 1, chegando a alcancar 92% de acurácia na amostra 2, elas demonstraram uma capacidade robusta de prever corretamente a evasão, apresentando uma boa combinação de precisão e recall. Esses resultados são especialmente relevantes em contextos educacionais, onde a identificação precoce de alunos em risco é essencial para a implementação de intervenções eficazes.

Além do ótimo desempenho, as Redes Neurais têm a capacidade de lidar com interações complexas entre variáveis, capturando padrões sutis nos dados. Apesar de exigir mais recursos computacionais e um maior volume de dados para o treinamento, os resultados obtidos justificam plenamente o investimento.

A **Árvore de Decisão** também apresentou resultados sólidos e consistentes. Com acurácia elevada (81,58% e 83,12% nas amostras 1 e 2 respectivamente), o modelo demonstrou um bom equilíbrio entre recall e precisão, alcançando valores de precisão de 80,77% e 88,46% e recall de 91,3% e 86,79% nas duas amostras. Esses números indicam que a Árvore de Decisão foi eficaz tanto na identificação de casos de abandono quanto em evitar falsos positivos.

Aqui está uma versão revisada do texto, com melhorias na fluidez e clareza:

O modelo **KNN** apresentou o pior desempenho entre os modelos avaliados. Na primeira amostra, obteve uma acurácia de apenas 68,42% e um recall de 70%, o que indica uma baixa sensibilidade.

Na segunda amostra, o modelo mostrou uma melhora significativa. Embora a acurácia da validação cruzada tenha se mantido em torno de 77%, o recall apresentou uma leve melhora. Além disso, tanto a precisão quanto o F1-score apresentaram ótimos resultados.

O **SVM** se mostrou um modelo equilibrado e com bom resultado, demonstrando ótimo desempenho em ambos os cenários. Com acurácias de 84,21% na Amostra 1 e 85,71% na Amostra 2, além de um recall elevado e precisão consistente, o SVM manteve-se estável mesmo com o aumento do ruído e da complexidade dos dados. Esse comportamento demonstra sua eficácia em lidar com cenários desafiadores, tornando-o uma ótima escolha para problemas que envolvem padrões complexos.

A pesquisa também análisou os modelos de aprendizado de máquina testados por série. E revelou diferenças no desempenho de cada abordagem.

Em geral, todos os resultados apresentaram desempenho ruins. Diante desse cenário as **Redes Neurais** apresentaram os melhores resultados, especialmente na Amostra 1, onde ob-

tiveram uma acurácia de 0.3750 e uma validação cruzada de 0.3649. Esse desempenho pode ser atribuído à capacidade das Redes Neurais de capturar padrões mais complexos nos dados mais diversos. No entanto, na Amostra 2, onde os dados são menos diversos e mais incompletos, a acurácia caiu para 0.3117, indicando que pra esse tipo de abordagem o modelo depende fortemente da qualidade e da diversidade dos dados para um bom desempenho.

A **Árvore de Decisão** apresentou os piores desempenhos entre os modelos analisados. Sua acurácia na Amostra 1 foi de apenas 0.2292, e na Amostra 2 subiu levemente para 0.2727. A validação cruzada se manteve próxima em ambas as amostras, mas em valores baixos (0.3139 e 0.3148), sugerindo que o modelo pode estar simplificando demais os padrões dos dados e, consequentemente, não generaliza bem.

O modelo **k-Nearest Neighbors (kNN)** demonstrou um desempenho também considerado insatisfatório, com acurácia de 0.3125 na Amostra 1 e 0.3377 na Amostra 2. A validação cruzada variou consideravelmente entre as amostras, sendo 0.3399 na Amostra 1 e 0.2914 na Amostra 2.

O modelo **SVM** apresentou um comportamento instável, com uma acurácia de 0.3333 na Amostra 1 e uma queda para 0.2597 na Amostra 2. Essa variação sugere que para esse tipo de predição o SVM pode ser sensível a diferenças na distribuição dos dados, especialmente quando estes são mais escassos ou incompletos. Apesar disso, sua validação cruzada foi relativamente consistente entre as amostras (0.3525 e 0.3204), indicando que o modelo ainda possui uma capacidade razoável de generalização.

Outro aspecto relevante da pesquisa é a necessidade de considerar o contexto socioeconômico e cultural dos alunos. A variação na taxa de evasão por idade e localização indica que esse fenômeno é influenciado por múltiplos fatores. O uso de aprendizado de máquina não apenas aprimora a capacidade preditiva, mas também possibilita investigações mais profundas sobre as razões que levam os alunos a abandonarem os estudos.

A inclusão de novos dados, embora não tenha melhorado substancialmente a acurácia geral dos modelos mais simples, foi essencial para testar sua robustez em diferentes cenários. Essa análise reforça a importância de integrar informações de diversas fontes para enriquecer as previsões e ampliar a compreensão sobre a evasão escolar.

Por fim, Os resultados obtidos evidenciam a superioridade das Redes Neurais na predição da evasão escolar, esse modelo demonstrou uma capacidade excepcional de identificar padrões mesmo em um conjunto de dados menos diversos e estruturados. Essa performance destaca a robustez das Redes Neurais em comparação com outras abordagens, como Árvores de Decisão, KNN e SVM, que, embora tenham apresentado desempenhos consistentes, não atingiram o mesmo nível de precisão. Assim, a utilização de Redes Neurais se mostra uma ferramenta altamente eficaz para a análise preditiva da evasão, oferecendo maior confiabilidade na identificação de alunos em risco.

Futuros trabalhos deverão explorar técnicas como balanceamento de classes, otimização de hiperparâmetros, criação de subdatasets e integração de dados mais diversificados para melhorar ainda mais o desempenho. Será igualmente essencial monitorar o impacto da inserção de novos dados e ajustar os modelos para otimizar a precisão das previsões.

As descobertas desta pesquisa oferecem suporte para gestores e educadores na luta contra a evasão escolar, ajudando a garantir que mais alunos permaneçam na escola e completem sua formação.

6

Referências bibliográficas

1 RUSSEL, S.; NORVIG, P. Inteligência Artificial. 3. ed. Rio de Janeiro: Editora Elsevier, 2013.

2 BISHOP, Christopher M. Neural Networks for Pattern Recognition. Oxford University Press, 1995.

3 MITCHELL, T. M. Machine Learning. McGraw-Hill, 1997.

4 BENGIO, Y.; COURVILLE, A.; VINCENT, P. Representation Learning: A Review and New Perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 35, n. 8, p. 1798–1828, 2013.

5 HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2009.

6 GÉRON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, 2019.

7 CARVALHO, Marta Maria Chagas de. Evasão Escolar no Brasil: Uma Perspectiva Multidimensional. 2014.

8 MENDES, L. S.; SOUZA, F. C. Análise das Causas da Evasão Escolar no Ensino Fundamental. Educação em Foco, v. 25, n. 2, p. 89-102, 2021. DOI: 10.5752/P.2317-788X.2021v25n2p89.

9 DUTRA, Renan Martins. O uso de inteligência artificial para predição de evasão na rede Doctum de ensino. 2015. Trabalho de Conclusão de Curso (Ciência da Computação) - Instituto Doctum de Educação e Tecnologia, Faculdades Integradas de Caratinga – FIC, Caratinga.

10 BITENCOURT, W. A.; SILVA, D. M.; XAVIER, G. C. Pode a inteligência artificial apoiar ações contra evasão escolar universitária? Ensaio: Avaliação e Políticas Públicas em Educação, 2021. SciELO Brasil.

11 FILHO, R. B. S.; ARAÚJO, R. M. DE 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, 2017.

12 LINO, Ellen Rízia Oliveira. A problemática da evasão escolar: uma revisão bibliográfica integrativa. 2020. Monografia (Graduação em Ciências Biológicas - Licenciatura) - Pontifícia Universidade Católica de Goiás, Escola de Ciências Agrárias e Biológicas, Curso de Ciências Biológicas Licenciatura, Goiânia.

13 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. MIT Press, 2016.

14 BOWERS, A. J.; SPROTT, R. Examining the multiple predictors of student dropout: A survival analysis approach. Journal of Educational Research, v. 105, n. 3, p. 176–195, 2012. DOI: 10.1080/00220671.2011.552075.

15 KOTSIANTIS, S. B.; PIERRAKEAS, C.; PINTELAS, P. Predicting students' performance in distance learning using machine learning techniques. Applied Artificial Intelligence, v. 18, n. 5, p. 411-426, 2004. DOI: 10.1080/08839510490442058.

16 CHOLLET, F. Deep Learning with Python. Manning Publications, 2018.

17 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. DOI: 10.1109/TSMCC.2010.2053532.

18 BAKER, R. S. J. d. Data mining for education. In: MCGAW, B.; PETERSON, P.; BAKER, E. (Eds.). International Encyclopedia of Education. 3. ed. Elsevier, 2010, p. 112-118.

19 POWERS, D. M. W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, v. 2, n. 1, p. 37-63, 2011.

20 FAWCETT, T. An introduction to ROC analysis. Pattern Recognition Letters, v. 27, n. 8, p. 861-874, 2006. DOI: 10.1016/j.patrec.2005.10.010.

21 DOMINGOS, P. A few useful things to know about machine learning. Communications of the ACM, v. 55, n. 10, p. 78-87, 2012.

22 QUINLAN, J. R. Induction of Decision Trees. Machine Learning, v. 1, n. 1, p. 81-106, 1986.

23 CORTES, C.; VAPNIK, V. Support-Vector Networks. Machine Learning, v. 20, n. 3, p. 273-297, 1995.

24 LECUN, Y.; BENGIO, Y.; HINTON, G. Deep Learning. Nature, v. 521, n. 7553, p. 436-444, 2015