INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO PIAUÍ
CAMPUS CORRENTE
ANÁLISE E DESENVOLVIMENTO DE SISTEMAS

FRANCISCO LEITE DO NASCIMENTO

A INFLUÊNCIA DOS FATORES SOCIOECONÔMICOS NO DESEMPENHO ESCOLAR DOS ALUNOS DO SAEB EM ESCOLAS PÚBLICAS DO ESTADO DO PIAUÍ: UMA ANÁLISE COM MINERAÇÃO DE DADOS.

CORRENTE
2024

FRANCISCO LEITE DO NASCIMENTO

A INFLUÊNCIA DOS FATORES SOCIOECONÔMICOS NO DESEMPENHO ESCOLAR DOS ALUNOS DO SAEB EM ESCOLAS PÚBLICAS DO ESTADO DO PIAUÍ: UMA ANÁLISE COM MINERAÇÃO DE DADOS.

Trabalho de Conclusão de Curso apresentado como exigência parcial para obtenção do diploma do Curso de Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Corrente.

Orientador: Prof. Me. Carlos Estevão Bastos Sousa.

Coorientador: Prof. Me. Tulio Vidal Rolim.

CORRENTE

2024

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD

Nascimento, Francisco Leite do

N244i A influencia dos fatores socioeconômicos no desempenho escolar dos alunos do SAEB em escolas públicas do estado do Piauí: : uma análise com mineração de dados / Francisco Leite do Nascimento. - 2024.

43 p.

Trabalho de Conclusão de Curso (Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia do Piauí, Campus Corrente, 2024.

Orientador : Prof Me. Carlos Estevão Bastos Sousa..

Coorientador : Prof Me. Tulio Vidal Rolim.

1. Analise e Desenvolvimento de Sistemas - estudo e ensino. 2. Sistema de Avaliação da Educação Básica (SAEB). 3. algoritmos K-Nearest Neighbors (KNN) e Random Forest. 4. Desempenho educacional. I.Título.

CDD - 004

Elaborado por Tefischer Huanderson Soares e Sousa CRB 3/1251

LISTA DE FIGURAS

Figura 1 - Ranking de incidência da pobreza nacional……10
Figura 2 - Processo KDD……14
Figura 3 - Representação da árvore de decisão……15
Figura 4 - Pipeline da metodologia……..23
Figura 5 - Contagem de respostas por idioma…….25
Figura 6 - Variação de importância de variáveis…….28
Figura 7 - Distribuição de respostas para variável TX_RESP_Q07……....29
Figura 8 - Distribuição de respostas para variável TX_RESP_Q08……30
Figura 9 - Distribuição de respostas para variável TX_RESP_Q20e……..31
Figura 10 - Distribuição de respostas para variável TX_RESP_Q02……..31
Figura 11 - Distribuição de respostas para variável TX_RESP_Q14……..32
Figura 12 - Variação de relevância de variáveis……..33
Figura 13 - Distribuição de opções para variável TX_RESP_Q07……..34
Figura 14 - Distribuição de opções para variável TX_RESP_Q08……..35
Figura 15 - Distribuição de opções para variável TX_RESP_Q02……..35
Figura 16 - Distribuição de opções para variável TX_RESP_Q14……..36
Figura 17 - Distribuição de opções para variável TX_RESP_Q20e……37

LISTA DE QUADROS

Quadro 1 - Ranking das cinco variáveis mais significativas……29

LISTA DE SIGLAS

IBGE - Instituto Brasileiro de Geografia e Estatística.

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

INSE - Índice de Nível Socioeconômico.

KNN - *K-Nearest-Neighbors*.

MD - Mineração de Dados.

SAEB - Sistema de Avaliação da Educação Básica.

SLR - *Stepwise Logistic Regression.*

A INFLUÊNCIA DOS FATORES SOCIOECONÔMICOS NO DESEMPENHO ESCOLAR DOS ALUNOS DO SAEB EM ESCOLAS PÚBLICAS DO ESTADO DO PIAUÍ: UMA ANÁLISE COM MINERAÇÃO DE DADOS.

Francisco Leite do Nascimento

RESUMO

A educação é um pilar fundamental para o desenvolvimento individual e social de um país. No entanto, o desempenho escolar no Brasil e, em especial no estado do Piauí, ainda apresenta disparidades significativas, especialmente quando se consideram fatores socioeconômicos. Diante disso, é proposta uma análise das interações entre esses e o desempenho acadêmico dos estudantes do 3º ano do Ensino Médio em escolas públicas do estado, especificamente nas disciplinas de Língua Portuguesa e Matemática. Para essa investigação, foram utilizados dados provenientes do Sistema de Avaliação da Educação Básica (SAEB) de 2021. A metodologia adotada compreendeu uma fase de análise exploratória, pré-processamento e modelagem dos dados, sendo estas etapas realizadas com os algoritmos *K-Nearest Neighbors* (KNN) e Random Forest. Os resultados obtidos indicam que os fatores socioeconômicos mais influentes no desempenho dos alunos nessas duas disciplinas estão diretamente ligados ao nível de escolaridade dos pais ou responsáveis. Portanto, é possível concluir que o maior envolvimento e suporte educacional por parte da família exercem um impacto significativo no aprimoramento do desempenho acadêmico dos alunos. Adicionalmente, destaca-se que este estudo proporciona uma visão mais abrangente das dinâmicas que influenciam o desempenho escolar, destacando a importância da colaboração entre família e escola no processo educacional.

Palavras-chaves: Desempenho Escolar, fatores socioeconômicos, mineração de dados, Piauí, SAEB.

ABSTRACT

Education is a fundamental pillar for the individual and social development of a country. However, school performance in Brazil and, especially in the state of Piauí, still presents significant disparities, especially when socioeconomic factors are considered. Given this, an analysis of the interactions between these and the academic performance of students in the 3rd year of high school in public schools in the state is proposed, specifically in the subjects of Portuguese Language and Mathematics. For this investigation, data from the 2021 Basic Education Assessment System (SAEB) were used. The methodology adopted comprised a phase of exploratory analysis, pre-processing, and data modeling, with these steps being carried out using the K-Nearest Neighbors algorithms. (KNN) and Random Forest. The results obtained indicate that the most influential socioeconomic factors on student performance in these two subjects are directly linked to the level of education of parents or guardians. Therefore, it is possible to conclude that greater involvement and educational support on the part of the family have a significant impact on the improvement of students' academic performance. Additionally, it is noteworthy that this study provides a more comprehensive view of the dynamics that influence school performance, highlighting the importance of collaboration between family and school in the educational process.

Keywords: data mining, socioeconomic factors, SAEB, School performance, Piauí.

SUMÁRIO

1 INTRODUÇÃO 10
1.1 OBJETIVO 11
1.1.1 OBJETIVO GERAL 11
1.1.2 OBJETIVOS ESPECÍFICOS 12
1.3 ORGANIZAÇÃO DO TRABALHO 12
2 FUNDAMENTAÇÃO TEÓRICA 14
2.1 SAEB 14
2.2 TRABALHOS RELACIONADOS 14
2.3 MINERAÇÃO DE DADOS 18
2.4 SCIKIT-LEARN 18
2.4.1 RANDOM FOREST 19
2.5 PYOD 20
2.5.1 KNN 21
2.8 NORMALIZAÇÃO DE DADOS 22
3 PROCEDIMENTOS METODOLÓGICOS 24
3.1 SELEÇÃO E COLETA DE DADOS 25
3.2 PREPARAÇÃO E MODELAGEM 26
3.3 IMPLEMENTAÇÃO DO MODELO 27
4 RESULTADOS E DISCUSSÕES 29
4.1 PROFICIÊNCIA EM LÍNGUA PORTUGUESA 29
4.2 PROFICIÊNCIA EM MATEMÁTICA 33
5 CONCLUSÕES E SUGESTÕES DE TRABALHOS FUTUROS 39
REFERÊNCIAS 41

1 INTRODUÇÃO

No cenário educacional brasileiro, o Sistema de Avaliação da Educação Básica (SAEB) desempenha um papel crucial na análise e melhoria do sistema educacional do país, pois tal ferramenta é empregada pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) para avaliar o desempenho da educação básica no país para poder mensurar o Índice de Desenvolvimento da Educação Básica (IDEB). No SAEB é aplicado um questionário socioeconômico para obter informações do contexto socioeconômico dos estudantes. Esses dados, de acordo com INEP, são fundamentais para apresentar o contexto dos resultados das avaliações, interpretar o impacto de fatores sociais e econômicos no desempenho dos alunos e calcular o Índice de Ideb.

De acordo com Mapa de pobreza e desigualdade feito pelo Instituto Brasileiro de Geografia e Estatística (IBGE) no ano de 2023 o estado do Piauí, como é apresentado na Figura 1, tem uma incidência de pobreza e desigualdade de 53,11 %, ou seja, cerca de mais da metade da população do estado vive em situação de carência ocupando a 5º posição no ranking nacional.

Figura 1: Ranking de incidência da pobreza nacional.

Fonte: IBGE (2023).

De acordo com Silva (2022), o fato apresentado é um dos motivos contribuintes para as dificuldades educacionais enfrentadas pela população. Sendo assim, deve-se considerar o contexto socioeconômico dos alunos que participam do SAEB, pois esse fator é considerado como predominante no fracasso acadêmico.

Neste contexto cabe um estudo que esclareça a relação socioeconômica e os resultados no SAEB dos alunos do estado do Piauí. Diante disso, neste trabalho é proposto realizar tal estudo por meio de técnicas de mineração de dados que, de acordo com Soares et al (2023), contém recursos que permitem examinar variados conjuntos de informações a fim de identificar padrões e evidências, viabilizando a descoberta de conhecimentos que não estão acessíveis em ferramentas convencionais de gerenciamento de dados. Dessa forma, Ramos et al (2020), destacam que a utilização da mineração de dados na educação, abrange seu potencial para pesquisa, análise de dados de alunos e previsão de desempenho e abordagem de questões relacionadas ao sucesso do aluno, como por exemplo as condições sociais e econômicas.

Desta forma, este estudo se torna viável pelo fato de ter a necessidade de compreender e lidar com as diferenças econômicas e sociais que afetam diretamente o desempenho educacional dos estudantes no estado do Piauí, conforme evidenciado pelos dados sobre carência e desigualdade fornecidos pelo IBGE. Dado que o contexto econômico e social é reconhecido como um dos principais fatores que determinam o sucesso ou fracasso acadêmico, é crucial investigar como esses elementos impactam os resultados alcançados no SAEB. Ao

empregar técnicas mineração de análise de dados, como proposto neste estudo, é viável explorar extensas informações para identificar padrões e relações que anteriormente não eram acessíveis por métodos tradicionais. A utilização dessas abordagens na área educacional, como ressaltado por vários autores, proporciona oportunidades significativas para compreender de forma mais aprofundada as condições sociais e econômicas dos estudantes, antecipar seu desempenho e, assim, contribuir para o desenvolvimento de estratégias mais eficazes para aprimorar a educação. Portanto, este estudo é relevante ao oferecer uma análise detalhada da ligação entre o contexto socioeconômico e os resultados no SAEB, visando fornecer perspectivas valiosas para a elaboração de políticas educacionais mais justas e eficazes no estado do Piauí.

1.1 OBJETIVO

Com base nas abordagens anteriores, foram delineados uma série de objetivos visados por este trabalho, os quais são apresentados a seguir.

1.1.1 Objetivo Geral

A partir das observações anteriores sobre o SAEB em relação aos fatores socioeconômicos e à mineração de dados, este estudo tem como objetivo realizar uma análise abrangente das relações entre esses fatores e o desempenho acadêmico dos alunos nas escolas públicas do estado do Piauí, utilizando técnicas de mineração de dados. O objetivo central é identificar as principais variáveis socioeconômicas que influenciam os resultados dos alunos nessa avaliação.

1.1.2 Objetivos Específicos

- Coletar dados abertos acerca do resultado do SAEB 2021, para assim selecionar os dados referentes aos alunos da 3ª série da rede pública de ensino do Piauí;
- Pré-processar os dados coletados, implementando técnicas de limpeza de valores nulos, tratamentos de anomalias com o algoritmo KNN e, por fim, aplicar técnicas de normalização dos dados;
- Implementar modelos usando os dados obtidos para treinamento e testes, sendo modelos com variáveis alvo diferentes, ou seja, um modelo com proficiência em Língua Portuguesa e outro proficiência em Matemática como variável alvo;
- Implementar novos modelos em ambas as proficiências em aplicando redução de atributos a fim de reduzir variáveis para o estudo, eliminado sempre as consideradas menos relevantes para e pelo modelo por meio do método de importância de atributos.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho está estruturado em cinco capítulos principais, delineados para fornecer uma compreensão abrangente das relações entre fatores socioeconômicos e o desempenho acadêmico dos alunos na avaliação do SAEB em escolas públicas do estado do Piauí. A seguir, é apresentada uma breve visão geral de cada capítulo:

No Capítulo 1, são apresentados um texto introdutório em que há, entre outros, a justificativa deste trabalho; o objetivo geral e os objetivos específicos, delineando claramente o propósito e os focos da pesquisa.

No Capítulo 2, é apresentada a fundamentação teórica, isto é, uma abordagem acerca dos conceitos essenciais, começando com revisão dos trabalhos relacionados para situar o estudo no contexto existente, logo depois uma breve abordagem acerca do SAEB, em seguida a conceituação de mineração de dados, seguida de uma discussão sobre a biblioteca Scikit-Learn, com ênfase em no algoritmo Random Forest, e da mesma forma acontece com biblioteca PyOD com foco no algoritmo *K-Nearest Neighbors* (KNN), e por fim, apresenta-se a exploração da importância da normalização de dados.

No Capítulo 3, são apresentados os procedimentos metodológicos deste trabalho, no qual são detalhadas as etapas seguidas durante o desenvolvimento do estudo, desde a seleção e coleta de dados até a preparação e modelagem dos mesmos, culminando na implementação do modelo proposto. Esta seção oferece uma visão transparente do processo adotado na pesquisa.

No Capítulo 4, são explicitados os resultados e discussões de forma a fornecer uma análise aprofundada dos dados obtidos, com uma ênfase especial na proficiência em língua portuguesa e em matemática, as áreas de interesse específicas do estudo. Aqui, são discutidas as descobertas, interpretadas de acordo com o contexto teórico apresentado anteriormente e relacionadas aos objetivos estabelecidos, fornecendo insights valiosos sobre a eficácia do modelo proposto.

Por fim, no Capítulo 5, são apresentadas as conclusões e elaboradas com base nos resultados obtidos, destacando as principais descobertas e contribuições do estudo. Além disso, são oferecidas sugestões para trabalhos futuros, identificando lacunas no conhecimento ou áreas que merecem investigação adicional, visando enriquecer ainda mais a compreensão do tema e promover avanços na área.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo são abordados o Sistema de Avaliação da Educação Básica (SAEB), os Trabalhos Relacionados, e uma descrição essencial dos conceitos relacionados à mineração de dados, suas técnicas e demais tecnologias utilizadas neste trabalho.

2.1 SAEB

O SAEB, de acordo com Inês (2016), surge em um momento crucial da história brasileira, durante o processo de redemocratização do país em 1988 e em meio a importantes reformas políticas, sociais e econômicas. Sua institucionalização, ocorrida em 1994, marca uma mudança significativa no modo como as políticas educacionais são formuladas, implementadas e gerenciadas, refletindo a necessidade de avaliação contínua e participativa na área da educação.

Desta forma Pereira et al.(2022), destacam que o SAEB foi criado para servir como um sistema de avaliação externa que fornece uma referência para os municípios avaliarem a qualidade da educação em suas respectivas áreas. O SAEB atua como um termômetro, alertando os municípios a continuarem investindo em educação para manter e melhorar seus padrões educacionais. Também visa fornecer conhecimento sobre a estrutura de avaliação, as habilidades que os alunos precisam dominar e como os municípios podem utilizar os resultados para melhorar seus sistemas educacionais internos.

2.2 TRABALHOS RELACIONADOS

Na pesquisa realizada por Brito et al. (2022) onde o objetivo foi investigar a relação entre o índice de nível socioeconômico e a proficiência em Matemática no SAEB, das edições de 2019 a 2021, de estudantes do 9º ano nas escolas públicas do estado de Pernambuco, por meio da mineração de dados. Uma abordagem dos autores foi a formulação de duas questões de pesquisa específicas: (1) existe alguma diferença estatisticamente comprovada entre as médias de proficiência em matemática dos estudantes de acordo com o seu nível socioeconômico? e (2) quais fatores socioeconômicos são mais relevantes para a diferença no nível socioeconômico?

No procedimento metodológico abordando o primeiro questionamento proposto pelos autores, foram usadas técnicas de estatística descritiva e inferencial utilizando o software RStudio na versão 4.1.0 e os pacotes tidyverse, ggpubr e statix. Como teste para verificar a diferença entre os dados observados foi o Kruskal-Wallis, teste escolhido por ser apropriado para comparar duas ou mais distribuições de uma variável sendo observada em duas ou mais amostras distintas.

Foram utilizadas duas bases de dados, as bases do SAEB de 2011 a 2019 e do Índice de Nível Socioeconômico (INSE) estabelecido nos anos de 2011, 2015 e 2019. Foram selecionadas variáveis das duas bases, sendo feitas filtragens de escolas públicas estaduais e municipais do estado de Pernambuco no SAEB e com a eliminação dos valores nulos e por fim houve a aplicação do teste.

Seguindo a abordagem do estudo, comprovou-se uma diferença estatisticamente significativa na proficiência em matemática entre estudantes de diferentes origens socioeconômicas, indicando que fatores socioeconômicos desempenham um papel no desempenho dos alunos, sendo encontrado algumas diferenças entre escolas municipais e estaduais, mas todas mostraram estarem nesse segmento de resultado.

Prosseguindo para o segundo questionamento, no qual os autores visam identificar quais fatores socioeconômicos são mais relevantes para a diferença no nível socioeconômico, foi adotado o método *Stepwise Logistic Regression* (SLR), que realiza uma seleção automatizada do conjunto de variáveis independentes que formam a melhor construção possível do modelo de predição para uma variável dependente. Os dados utilizados são oriundos da base de dados do SAEB do ano de 2019, sendo usado apenas os dados da tabela dos alunos do 9º ano, que resultou na criação de uma amostra de 13.411 alunos da rede estadual e 25.547 da rede municipal.

Com os dados todos pré-processados, o algoritmo SLR foi aplicado duas vezes para a rede estadual e outra para a rede municipal. Com um total de 49 perguntas do questionário socioeconômico do SAEB utilizadas como variáveis independentes, enquanto a variável dependente estando com valores binários referentes a proficiência em matemática, onde a variável binária representava se os alunos estavam acima ou abaixo da média do estado do Pernambuco, considerando médias diferentes da rede estadual e municipal. E por fim os autores aplicaram o algoritmo Odds Ratio para identificar as cinco características socioeconômicas mais determinantes na classificação se o aluno estava acima ou abaixo da média.

Após as etapas descritas no parágrafo anterior, a cerca a aplicação do algoritmo Odds Ratio, dentre as demais variáveis essas cinco variáveis foram identificadas como as mais determinantes na classificação se o aluno estava acima ou abaixo da média em matemática, relacionadas às seguintes questões: "Fora da escola em dias de aula, quanto tempo você usa para lazer?", "Você já foi reprovado?", "Na sua casa tem TV a cabo?", "Na sua casa tem Freezer (independente ou segunda porta da geladeira)?" e "Alguma vez você abandonou a escola deixando de frequentá-la até o final do ano escolar?".

Araújo et al (2022) propõe um trabalho como o objetivo analisar a Proficiência em Matemática dos municípios do Ceará, utilizando a Mineração de Dados Educacionais (MDE) e dois conjuntos de dados: os Indicadores Educacionais do INEP e os resultados da Avaliação de Matemática do SPAECE/2019 para a 3ª série do Ensino Médio na rede pública de ensino do estado. Através do processo de KDD, a pesquisa extraiu informações dessas fontes de dados para criar um Modelo de Conhecimento que pudesse prever e classificar o desempenho dos municípios na Avaliação de Matemática do SPAECE, utilizando os indicadores educacionais como base.

No processo metodológico foi utilizado o KDD, que utiliza técnicas de mineração de dados para extrair conhecimento útil de conjuntos de dados. No estudo em questão, foram empregadas técnicas de mineração de dados, especificamente a construção de um modelo de Árvore de Decisão, para analisar e classificar Indicadores de Educação (IE) relevantes para o problema investigado.

O modelo de Árvore de Decisão mencionado no parágrafo anterior foi construído usando a ferramenta WEKA, que é um conjunto de implementações de vários algoritmos de mineração de dados, incluindo o algoritmo J48 para construção de árvore de decisão.

A pesquisa integrou dados de diferentes fontes, incluindo os Indicadores Educacionais formulados pelo INEP e dados de Proficiência em Matemática nos municípios do Ceará, para desenvolver o modelo de conhecimento. O estudo se concentrou na análise de vários atributos preditivos, como taxa de distorção idade-série, complexidade gerencial, média diária de horas de aula e adequação do treinamento de professores, para entender seu impacto na proficiência em matemática.

Na fase de resultados da pesquisa, um modelo bem-sucedido foi desenvolvido com sucesso, atingindo uma precisão de 85,86 %, no qual o mesmo identificou os indicadores educacionais relevantes como a alta taxa de distorção idade-série que sugere que há um número significativo de estudantes que não estão progredindo no ritmo esperado e que os pais ou responsáveis dos alunos concluíram ou não o ensino fundamental e/ou concluíram o ensino médio para abordar a questão da proficiência em matemática no estado do Ceará.

Soares et al. (2023) realizaram uma pesquisa com o objetivo de investigar quais os fatores que influenciam a qualidade da educação no estado do Maranhão, usando técnicas de mineração de dados aplicadas nas bases fornecidas pelo INEP, especificamente na base do IDEB e do SAEB. A pesquisa teve como objetivo identificar e analisar as principais variáveis que influenciam o desempenho das escolas públicas estaduais, com foco no ensino médio.

Os autores buscaram realizar uma análise exploratória dos dados do Ensino Médio, com o propósito de encontrar as variáveis com maior influência na nota do IDEB e das escolas públicas do estado. O objetivo geral da pesquisa foi criar evidências que expliquem o baixo rendimento escolar no estado do Maranhão.

Para a pesquisa foi adotada a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM), na qual é uma abordagem amplamente utilizada em pesquisas de mineração de dados, devido a sua flexibilidade de adaptação de categoria de negócios e por se tratar de uma metodologia iterativa, ou seja, é executada por etapas onde pode-se voltar para uma etapa anterior se necessário. Essa metodologia utiliza de 6 etapas:

- Entendimento do negócio: compreensão dos objetivos da organização e definição dos objetivos específicos do projeto de mineração de dados;
- Entendimento dos dados: coleta e exploração dos dados disponíveis para o projeto;
- Preparação dos dados: limpeza, transformação e formatação dos dados para análise;
- Modelagem: seleção e aplicação de técnicas de mineração de dados para extrair padrões ou relações nos dados;

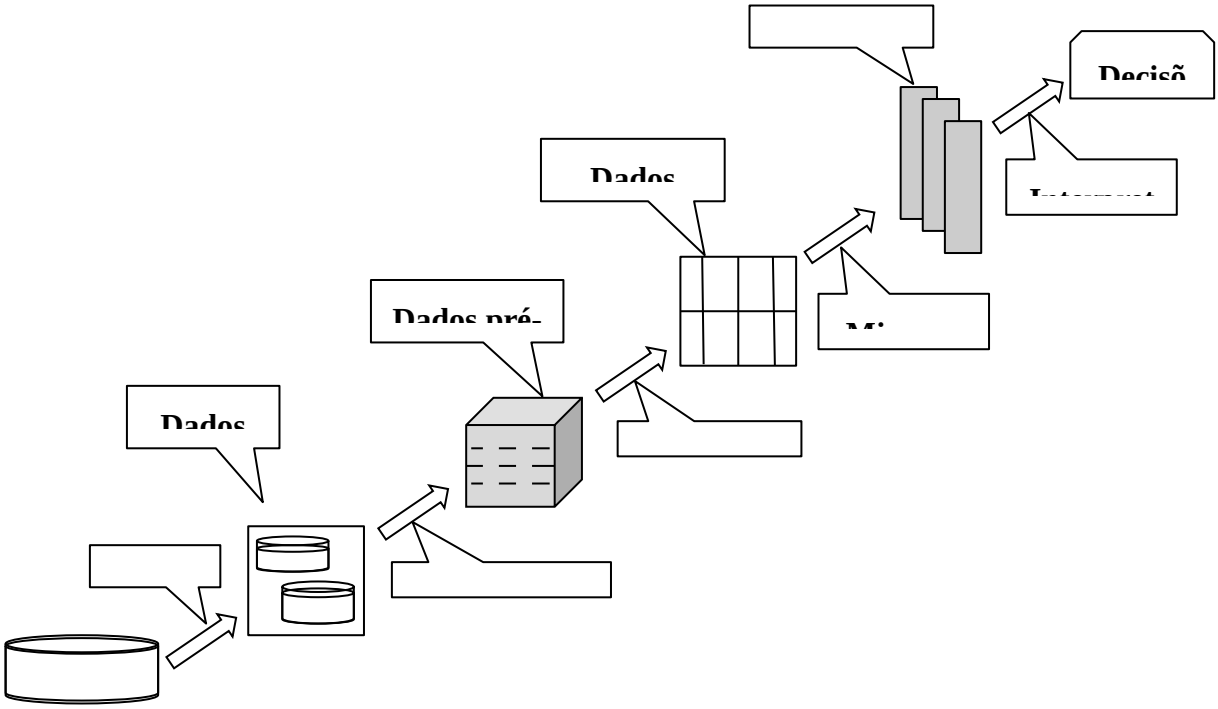
- Avaliação: avaliação do desempenho do modelo gerado e verificar se atende aos objetivos do projeto;
- Distribuição: amostragem dos resultados da análise dos dados, realizada através de relatórios, imagens, gráficos e ilustrações.

Desta forma, após a implementação dessa metodologia os pesquisadores conseguiram identificar que o nível de escolaridade dos pais ou responsáveis e o acesso à tecnologia, como computadores e internet, influenciam significativamente o desempenho educacional no Maranhão. Outro fator identificado foi o envolvimento dos pais na educação, incluindo a participação em reuniões escolares, a discussão de assuntos escolares com os filhos, o incentivo ao estudo, a conclusão do dever de casa e a motivação da frequência escolar. Assim demonstrando que as principais variáveis estão relacionadas às condições socioeconômicas.

2.3 MINERAÇÃO DE DADOS

De acordo com Fayyad et al. (1996) a Mineração de Dados (MD) é o processo de extrair padrões e conhecimento de grandes conjuntos de dados. Está intimamente relacionado à descoberta de conhecimento em bancos de dados, isto é, *Knowledge Discovery in Databases* (KDD), e envolve técnicas de aprendizado de máquina, estatísticas e bancos de dados, na Figura 2 é apresentado o processo KDD e onde a MD se encaixa neste processo. As técnicas de MD são usadas para descobrir padrões, relacionamentos e esclarecimentos ocultos dos dados, que podem então ser usados para várias aplicações.

Figura 2: Processo KDD.



Fonte: Adaptado de Fayyad et al. (1996).

O processo delineado na Figura 2 representa uma abordagem sistemática do KDD, no qual os dados são submetidos a várias etapas. Inicialmente, ocorre a seleção dos dados relevantes do conjunto disponível, seguida pelo pré-processamento para preparação adequada. Em seguida, os dados são transformados para facilitar a mineração em busca de padrões interpretáveis, que, por sua vez, fornecem padrões para embasar a tomada de

decisões.Goldschmidt et al. (2005) destaca que no decorrer da etapa de Mineração de Dados, acontece a busca de conhecimentos úteis no contexto da aplicação do KDD. Sendo considerada a principal etapa deste processo, pois nesse estágio, diversas técnicas e algoritmos, tais como Redes Neurais, Algoritmos Genéticos, Modelos Estatísticos e Probabilísticos, podem ser empregados. A seleção da técnica apropriada está intrinsecamente ligada ao tipo específico de tarefa de KDD a ser executada.

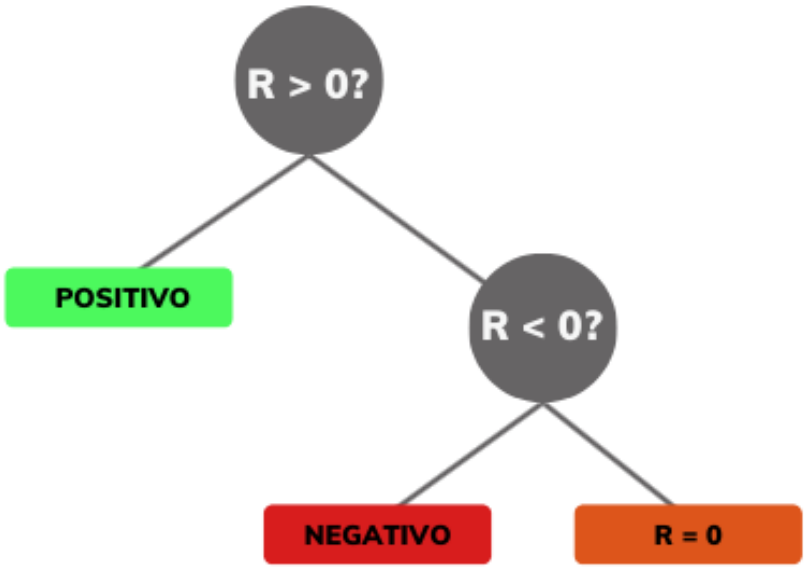
2.4 SCIKIT-LEARN

De acordo com a documentação da Scikit-learn (2023), esta é uma biblioteca de aprendizado de máquina para Python que oferece ferramentas eficientes e fáceis de usar para análise preditiva de dados. Ela funciona bem com outras bibliotecas Python, como NumPy e SciPy, e inclui uma ampla variedade de algoritmos de *Machine Learning* para tarefas como classificação, regressão, agrupamento e redução de dimensionalidade. Além disso, a biblioteca oferece ferramentas para pré-processamento de dados e avaliação de modelos, tornando-a uma escolha popular entre iniciantes e profissionais experientes.

2.4.1 *Random Forest*

Conforme Breiman (2001) e Biau (2012) a técnica de *Random Forest* é um método de conjunto baseado em árvores de decisão que avalia continuamente se o rótulo atribuído aos recursos atende a critérios específicos. Por exemplo, em um conjunto de números, se o recurso for maior que 0, será classificado como positivo; se for menor que 0, será classificado como negativo; caso contrário, será classificado como 0. O contexto descrito é apresentado na Figura 3.

Figura 3: Representação da árvore de decisão.



Fonte: Autor (2024).

A eficácia do algoritmo é evidente em tarefas de aprendizado supervisionado, em que utiliza conjuntos conhecidos de entradas e saídas para ajustar os pesos do modelo até que esteja precisamente ajustado (SOTNIKOV et al., 2023).

O algoritmo *Random Forest*, existente na biblioteca Scikit-learn, possui dois parâmetros cruciais que merecem atenção. O primeiro é *n_estimators*, que define o número de árvores de decisão no modelo em que aumentar esse valor pode aprimorar a precisão, mas acarretar em maior tempo de treinamento e consumo de recursos. O segundo é *feature_importances_*, que avalia a importância de cada variável no modelo, atribuindo valores numéricos que indicam a contribuição relativa para a precisão das previsões. Variáveis com valores mais elevados são consideradas mais importantes, facilitando a seleção de recursos relevantes.

De acordo com Yu et al. (2020), a ideia principal da avaliação da importância das características com base no Random Forest é calcular o valor de contribuição de cada característica em cada árvore e, finalmente, comparar o valor médio de contribuição das características. O Índice *Gini* e as taxas de erro de dados *out-of-bag* são dois métodos comuns de cálculo das contribuições. A função *feature_importances_* na biblioteca Sklearn do Python analisa a importância das características baseando-se no Índice *Gini* para ordenar as características. Além disso, o valor da importância das características após a operação é normalizado. A fórmula do cálculo é apresentada na equação 1.

$$Gini(p) = \sum_{k=1}^{K} P_K(1 - P_K) \quad (1)$$

Na equação apresentada, $Gini(p)$ representa o índice Gini para a probabilidade p em um determinado contexto, em que K é o número total de classes ou categorias, e para cada uma delas, calcula-se $P_K(1 - P_K)$, que é a probabilidade de uma amostra pertencer à classe K multiplicada pela probabilidade de não pertencer à K. Isso é somado para todas as classes para obter índice de Gini para o contexto p. A fórmula avalia a impureza do conjunto de dados, sendo que valores mais baixos indicam uma distribuição mais pura das classes.

2.5 PYOD

De acordo com a documentação da PyOD (2023), a biblioteca, também conhecida como Python Outlier Detection, é uma ferramenta especializada em detectar valores discrepantes em conjuntos de dados usando técnicas de *Machine Learning*.

Os algoritmos disponíveis na PyOD incluem métodos supervisionados e não supervisionados, permitindo que os usuários escolham a abordagem mais adequada para suas necessidades específicas.

Adicionalmente, destaca-se que a PyOD foi concebida com a finalidade de ser interoperável com outras bibliotecas amplamente utilizadas de aprendizado de máquina em Python, a exemplo do Scikit-learn. Essa característica propicia uma integração facilitada em fluxos de trabalho preexistentes. No contexto de manipulação de conjuntos de dados suscetíveis à presença de valores atípicos ou anômalos, a consideração da PyOD pode revelar-se uma opção significativa para análises e implementações de cunho científico.

2.5.1 KNN

De acordo com Verdier e Rosa (2011) a técnica conhecida como k vizinhos mais próximos (*K-Nearest Neighbors* ou KNN), representa um método de classificação supervisionada. Nesse método, a classificação de um novo objeto é determinada ao analisar as distâncias entre esse objeto e as amostras mais próximas em um espaço de características específicas. Dessa forma, amostras que não possuem rótulos são classificadas com base em sua semelhança com as amostras presentes em um conjunto de treinamento. O KNN, como regra

de classificação supervisionada, opera com base na similaridade, onde a observação tende a ser atribuída à mesma classe.

Dessa forma, na biblioteca PyOD, o KNN é utilizado para identificar valores atípicos, isto é, aqueles que se destacam significativamente dos demais dados, com o objetivo de detectar pontos de dados anormais. Para calcular a anomalia ou a magnitude da discrepância de um ponto de dados em relação aos demais pontos, emprega-se o KNN que usa como medida de distância. Esse algoritmo tem o atributo *n_neighbors* que permite ajustar o equilíbrio entre a sensibilidade e a precisão do modelo KNN na detecção de anomalias.

Um outro atributo que deve ser destacado é o *metric* do KNN, atributo este que determina qual o cálculo de distância a ser utilizado no modelo, alguns exemplos são Minkowski, Euclidiana e Manhattan. Na biblioteca PyOD por padrão é utilizado o Minkowski. Conforme Wang (2023), a distância de Minkowski não é uma medida única, mas sim um conjunto de possíveis medidas de distância. Para dois pontos $x(x1, x2, \dots)$ e $y(y1, y2, \dots)$ em um espaço com n dimensões, a fórmula geral da distância de Minkowski pode ser expressa na equação 2:

$$d_{xy} =^{p} \sqrt{\sum_{k-i}^{n} (x_k - y_k)^p} \quad (2)$$

Na equação fornecida d_{xy} representa a distância Minkowski entre dois pontos, e p é um parâmetro variável que determina a métrica do espaço, pois quando $p = 1$, é referida como distância de Manhattan, e quando $p = 2$, é conhecida como distância Euclidiana. E por fim Σ indica uma soma sobre todos os elementos dos vetores x_k e y_k.

2.8 NORMALIZAÇÃO DE DADOS

De acordo com Han e Kamber (2001), a normalização de um atributo ocorre quando seus valores são ajustados de modo a situarem-se dentro de uma faixa específica, geralmente de 0,0 a 1,0. Esse processo é particularmente benéfico em algoritmos de classificação que envolvem redes neurais, bem como em técnicas que utilizam medidas de distância, como classificação do vizinho mais próximo e agrupamento.

No caso do uso do algoritmo de retropropagação em redes neurais para tarefas de classificação, a normalização dos valores de entrada para cada atributo, conforme observado nas amostras de treinamento, contribuirá para acelerar a fase de aprendizado. No contexto de métodos baseados em distância, a normalização desempenha um papel crucial ao evitar que atributos com intervalos inicialmente amplos dominem sobre aqueles com intervalos inicialmente mais estreitos. Existem diversas técnicas de normalização, na Equação 3 é apresentado o método minmax.

$$Xc = \frac{X - Xmin}{Xmax - Xmin} \quad (3)$$

Na equação fornecida é descrito o cálculo para normalização de valores em um conjunto de dados de por meio do método minmax. Aqui estão os elementos fundamentais dessa equação:

- Xc: Este é o valor resultante após a normalização, indicando onde o valor original X se posiciona dentro do intervalo normalizado;
- X: Este é o valor original a ser utilizado para a normalização;
- $Xmin$: Representa o valor mínimo observado no conjunto de dados original;
- $Xmax$: Representa o valor máximo observado no conjunto de dados original.

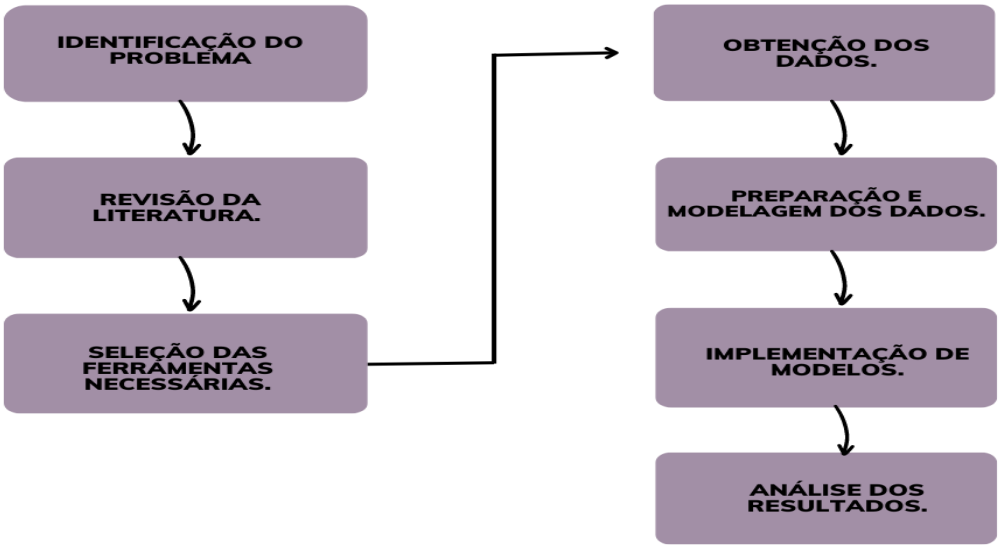
A fórmula realiza a normalização ao ajustar o valor original em relação à amplitude total do conjunto de dados, sendo a diferença entre $Xmax$ e $Xmin$. O resultado Xc é um valor normalizado que varia entre 0 e 1, indicando a posição relativa do valor original dentro do intervalo do conjunto de dados.

Diante da fundamentação teórica apresentada, no capítulo a seguir é apresentada toda a metodologia utilizada neste trabalho.

3 PROCEDIMENTOS METODOLÓGICOS

Neste capítulo serão abordados os procedimentos metodológicos empregados nesta pesquisa, sendo empregadas as seguintes etapas: identificação do problema, revisão da literatura, seleção das ferramentas necessárias, obtenção dos dados, preparação e modelagem dos dados, implementação de modelos, análise dos resultados. Na Figura 4 é apresentado de forma gráfica esse procedimento metodológico.

Figura 4: Pipeline da metodologia.



Fonte: Autor (2024).

Como evidenciado na Figura 4, a etapa inicial foi destinada para identificação do problema que estava relacionado aos fatores socioeconômicos e o desempenho dos alunos na avaliação do SAEB para assim prosseguir para a etapa seguinte na qual foi feita a revisão da literatura a fim de se obter um aprofundamento a respeito da mineração de dados e suas técnicas assim como o entendimento das tecnologias utilizadas. Seguindo a interpretação da Figura 4, temos a etapa de seleção de ferramentas necessárias para este trabalho, dentre as ferramentas estão, a plataforma Google Colab, um ambiente de programação em nuvem, juntamente com a linguagem de programação Python na versão 3.10, além das bibliotecas e algoritmos mencionados no Capítulo 2.

Prosseguindo com a apresentação dos procedimentos metodológicos, aborda-se a quarta etapa que se refere a obtenção dos dados, fase esta em que é empregada a coleta e seleção dos dados para este trabalho. Logo em seguida está a fase de preparação e modelagem de dados, nesta etapa acontecem o pré-processamento e a busca de uma modelagem adequada para o objetivo desta pesquisa.

Após a etapa de preparação e modelagem vem a implementação do modelo, em que ocorre a implementação de modelos de inteligência artificial usando o algoritmo Random Forest para trabalhar com os dados antes preparados. E no que diz respeito à etapa de análise dos resultados, a mesma está presente na seção de resultados e discussões desta pesquisa.

3.1 SELEÇÃO E COLETA DE DADOS

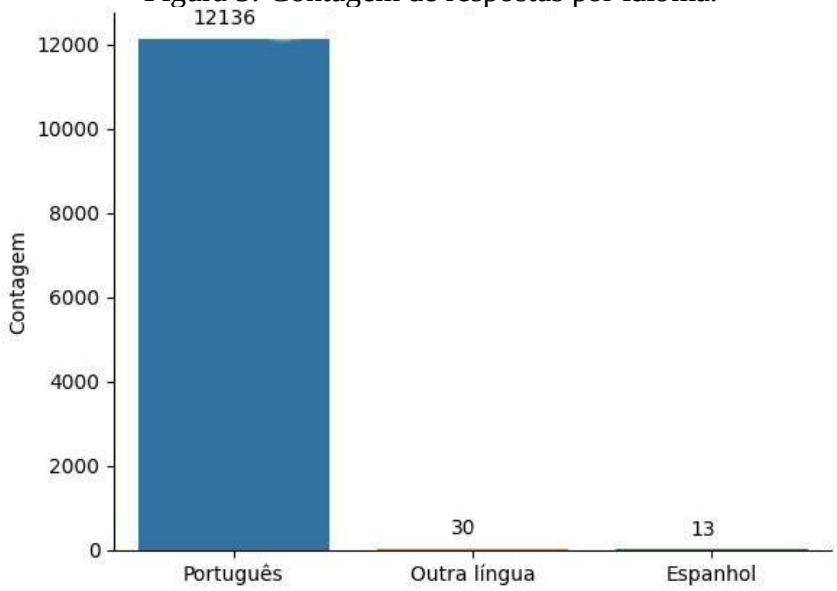
Para esta pesquisa os dados foram oriundos do SAEB 2021 por meio do portal de microdados abertos do INEP. Para o andamento da pesquisa foi utilizado o arquivo denominado TS_ALUNO_34EM.csv que corresponde ao questionário socioeconômico dos alunos do terceiro ano do ensino médio de todos os estados do país nessa edição do SAEB.

No decorrer do processo de coleta de dados, foram efetuadas seleções criteriosas alinhadas aos objetivos da pesquisa. Em um primeiro critério, optou-se por concentrar a análise exclusivamente nos dados dos alunos do estado do Piauí, uma vez que a base de dados inicial continha informações de todos os estados brasileiros. Simultaneamente, uma segunda seleção foi aplicada, restringindo a amostra aos estudantes exclusivos da rede pública de ensino. Essa abordagem resultou na delimitação mais precisa e focalizada, atendendo aos parâmetros específicos de interesse para a condução da pesquisa.

Foi realizada uma análise de distribuição das variáveis para identificar desbalanceamento em algumas variáveis para assim tomar decisões de como trabalhar ou até mesmo eliminá-las. Diante disso foram encontradas algumas variáveis do questionário socioeconômico que possuíam uma distribuição desigual entre as opções de respostas, sendo assim desbalanceada. Um exemplo disso é a variável TX_RESP_Q03, que aborda o idioma usualmente falado pelos pais do aluno em casa. Na Figura 5 é apresentada a disparidade entre as diversas respostas possíveis para essa pergunta do questionário.

Como apresentado na Figura 5, observa-se um desbalanceamento considerável na variável TX_RESP_Q03. Diante dessa disparidade, ao utilizar os dados em um modelo de inteligência artificial, surgiram opções válidas, como realizar tratamentos específicos de reestruturação dos dados ou optar pela exclusão da variável problemática. Devido a complexidade das técnicas de reestruturação de dados a decisão tomada foi remover a variável em questão, juntamente com outras que apresentavam o mesmo tipo de distribuição.

Figura 5: Contagem de respostas por idioma.



Fonte: Autoria própria (2023).

As variáveis utilizadas para a classificação foram PROFICIENCIA_LP_SAEB, que representa os níveis de proficiência em Língua Portuguesa, e PROFICIENCIA_MT_SAEB, que refere-se aos níveis de proficiência em Matemática.

3.2 PREPARAÇÃO E MODELAGEM

Nesta etapa da pesquisa houveram etapas de pré-processamento e modelagem dos dados para assim serem submetidos ao modelo, inicialmente, realizou-se a exclusão de dados nulos, seguida por uma transformação nas respostas do questionário. As respostas foram reorganizadas de acordo com a ordem alfabética, sendo posteriormente convertidas para um formato numérico. Nessa transformação, atribuiu-se o valor 1 às respostas originalmente marcadas como "A", o valor 2 para aquelas marcadas como "B" e assim por diante. Após esse processo de tratamento dos dados, a base resultante contou com 12.179 registros e 46 atributos, incluindo 2 atributos alvo.

Na etapa subsequente do pré-processamento realizou-se o procedimento de identificação e tratamento de anomalias nas respostas preenchidas pelos alunos. Para isso foi utilizado o algoritmo KNN oriundo da biblioteca PyOD. O valor escolhido para o parâmetro *n_neighbors* foi estabelecido como 5 e cálculo de distância escolhido foi o *Minkowski*, conforme a recomendação da documentação oficial da referida biblioteca. As anomalias foram encontradas e excluídas para fins de limpeza da base de dados, após esse procedimento a base resultou em 11.207 registros.

Dando continuidade a fase de pré-processamento, implementou-se a técnica de normalização dos dados, especificamente através da abordagem minmax, como apresentado na Seção 2. Destaca-se que a normalização foi realizada em virtude das características dos algoritmos a serem empregados, os quais demandam a entrada de dados normalizados para otimizar a convergência e o desempenho.

Após os procedimentos anteriores, foi necessário realizar uma transformação nas variáveis de proficiência em matemática e língua portuguesa. Essas variáveis estavam no formato decimal e, para este estudo, foi necessário convertê-las para um padrão binário. É importante ressaltar que o limiar escolhido para essa análise foi estabelecido em 225. Esta

escolha baseou-se no SAEB, cuja escala de proficiência indica que estudantes da 3ª série com pontuações inferiores a 225 demandam atenção especial. Isso se justifica pelo fato de que esses alunos ainda não demonstram habilidades elementares esperadas para essa etapa escolar, tornando crucial a identificação e o suporte apropriado para seu desenvolvimento educacional.

Para realizar a transformação, os atributos PROFICIENCIA_MT_SAEB e PROFICIENCIA_LP_SAEB foram modificados da seguinte forma: se o valor do atributo fosse menor que 225, ele foi convertido para 0; se fosse maior, foi convertido para 1. Após esse procedimento os dados estavam prontos para serem processados pelo modelo.

3.3 IMPLEMENTAÇÃO DO MODELO

Nesta etapa da pesquisa foi utilizado o modelo Random Forest, uma técnica de árvores de decisão, oriundo da biblioteca Scikit-learn. Para o treinamento do modelo foram inseridos os dados pré-processados nas etapas anteriores. Dois modelos foram implementados, um para cada área de classificação, isto é, um para Proficiência em Língua Portuguesa e outro para Proficiência em Matemática. Para fins de treinamento foram utilizados 70 % dos dados e 30 % para testes.

Neste processo é relevante ressaltar que o parâmetro *n_estimators* foi fixado em 100 durante a configuração do modelo Random Forest, conforme a escolha realizada com base em considerações técnicas e experimentações. Essa decisão buscou otimizar a eficiência e desempenho do modelo, alinhando-se com práticas recomendadas na literatura e documentação associada à biblioteca Scikit-learn.

Para os procedimentos de busca de um modelo mais aprimorado, foi decidido não apenas utilizar o mesmo modelo, mas também implementar novas instâncias com reduções de atributos no conjunto de dados. Essa escolha foi baseada na ideia de que manter o mesmo modelo, porém com menos atributos, poderia resultar em benefícios tanto em eficiência quanto em interpretabilidade.

A decisão de reduzir a dimensionalidade do modelo foi orientada pela avaliação da importância de cada variável, obtida por meio do atributo *feature_importances_* do modelo Random Forest. Esse atributo fornece pontuações que indicam a contribuição de cada variável para as previsões do modelo como antes explicado neste trabalho na subseção 2.2.1 no capítulo 2. Optou-se por manter as variáveis mais relevantes e remover as consideradas menos informativas.

Esse processo não apenas contribui para a eficácia do modelo, destacando as variáveis mais impactantes, mas também simplifica a interpretação do modelo. Vale ressaltar que a busca por outros modelos foi evitada para manter a consistência e a organização dos resultados. Essa estratégia teve como objetivo reduzir possíveis complicações na comparação entre diferentes modelos, concentrando-se na otimização do modelo original por meio da seleção cuidadosa de variáveis.

No início do procedimento, 44 atributos foram utilizados para o treinamento do modelo, tanto para avaliar a proficiência em língua portuguesa quanto em matemática. No entanto, ao final do processo de seleção e redução de atributos, restaram apenas 25 atributos em ambos os conjuntos de dados. Essa redução, baseada na importância das variáveis conforme o atributo *feature_importances_* do modelo Random Forest, teve como objetivo simplificar o modelo, concentrando-se nas características mais relevantes para as previsões em ambas as áreas de proficiência.

Agora, procederemos à apresentação dos resultados derivados da aplicação da metodologia delineada anteriormente. Neste capítulo serão analisadas e discutidas as percepções obtidas através da implementação das técnicas presentes na metodologia deste trabalho.

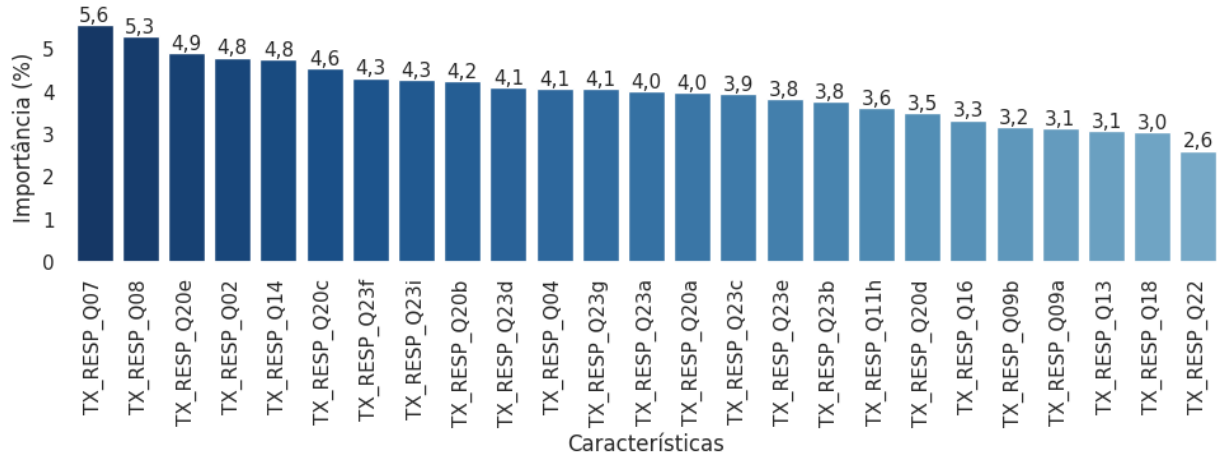
4 RESULTADOS E DISCUSSÕES

Nesta seção, serão abordados os resultados alcançados no decorrer do procedimento metodológico, fornecendo uma análise detalhada tanto da proficiência em Língua Portuguesa quanto da proficiência em Matemática.

4.1 PROFICIÊNCIA EM LÍNGUA PORTUGUESA

Após seguir os passos detalhados no capítulo anterior, com o objetivo de encontrar o modelo mais aprimorado, utilizando o atributo relacionado à nota em Língua Portuguesa como variável-alvo, foi obtido um modelo que apresentou uma acurácia de 82,93%. Em seguida, foram conduzidos novas implementações de modelos, implementando reduções adicionais de variáveis, contudo, observaram-se decréscimos na acurácia. Por meio da aplicação do método *feature_importances_* do Random Forest, foram identificadas as variáveis mais relevantes nos resultados do modelo. Na Figura 6 é fornecida uma representação gráfica desses resultados.

Figura 6: Variação de importância de variáveis.



Fonte: Autor (2024).

Conforme evidenciado na Figura 6, o atributo TX_RESP_Q07, associado ao nível de escolaridade da mãe ou da mulher responsável pelo aluno, destacou-se com o índice mais significativo de relevância, exercendo uma forte influência no desempenho do aluno na avaliação de Língua Portuguesa. Em contrapartida, a variável TX_RESP_Q22, que indica se o aluno concluiu o ensino fundamental por meio de supletivo ou similar, foi identificada como a menos importante.

Para facilitar a compreensão, o Quadro 1 apresenta uma classificação em ordem decrescente das cinco variáveis mais influentes, acompanhadas de suas descrições.

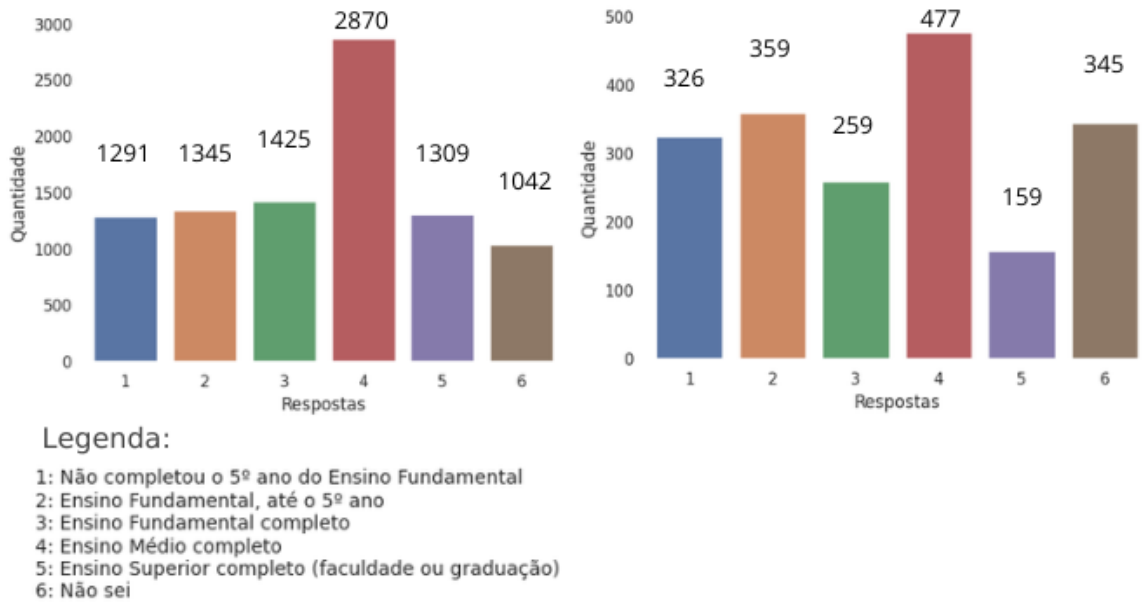
Quadro 1: Ranking das cinco variáveis mais significativas.

Fonte: Autor (2024).

Variável	Descrição
TX_RESP_Q07	Qual é a maior escolaridade da sua mãe (ou mulher responsável por você)?
TX_RESP_Q08	Qual é a maior escolaridade de seu pai (ou homem responsável por você)?
TX_RESP_Q20e	Fora da escola em dias de aula, quanto tempo você usa para: - Lazer (TV, internet, brincar, música etc.).
TX_RESP_Q02	Qual é a sua idade?
TX_RESP_Q14	Considerando a maior distância percorrida, normalmente de que forma você chega à sua escola?

Conforme evidenciado no Quadro 1, a variável TX_RESP_Q07 se destaca como a de maior impacto no desempenho do aluno em Língua Portuguesa, indicando claramente a influência significativa do nível de escolaridade da mãe ou da mulher responsável, sobre o desempenho acadêmico. Na Figura 7 é proporcionada uma representação visual da distribuição das respostas relacionadas a essa variável no questionário socioeconômico.

Figura 7: Distribuição de respostas para variável TX_RESP_Q07.

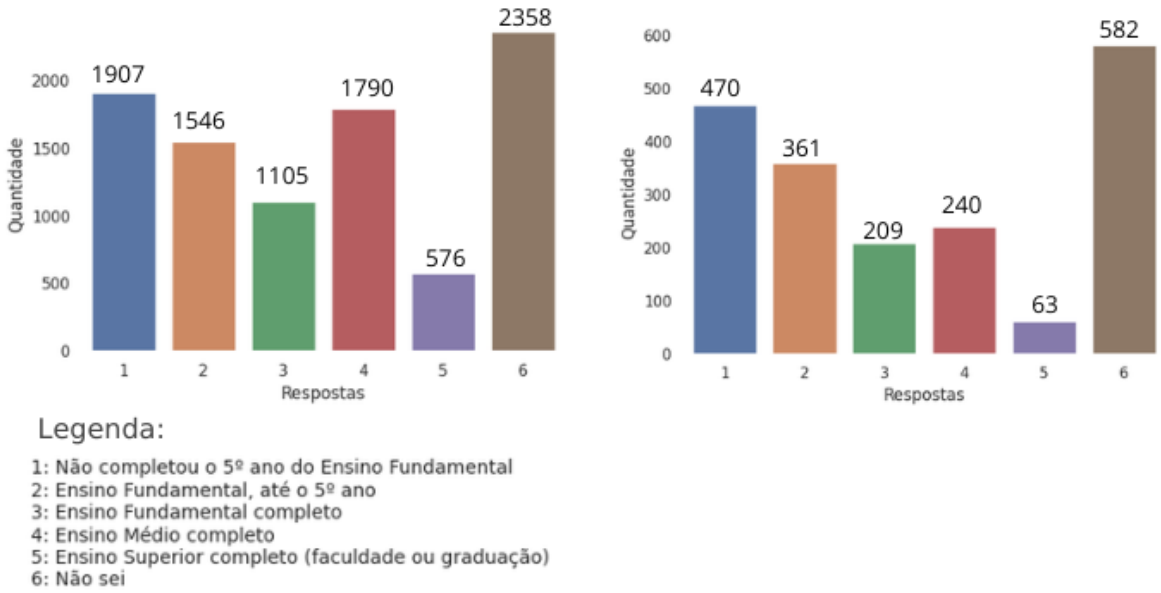


Fonte: Autor (2024).

Com base na Figura 7, destaca-se que a escolha mais proeminente entre as respostas foi a indicação de ensino médio completo, tanto para os alunos classificados com proficiência crítica quanto para os demais. No entanto, uma análise mais detalhada da distribuição revela que, em relação às demais alternativas no gráfico dos alunos com proficiência não crítica, a opção de ensino médio completo é praticamente o dobro das outras. Por outro lado, para os alunos em situação crítica, embora a opção de Ensino Médio completo seja notável, não apresenta uma discrepância tão expressiva em comparação com as demais alternativas.

Agora destacando o segundo atributo que está associado ao nível de escolaridade do pai, ou homem responsável, pelo aluno o mesmo padrão não se repete com a mesma intensidade. Na Figura 8 é apresentada a distribuição para essa variável.

Figura 8: Distribuição de respostas para variável TX_RESP_Q08.

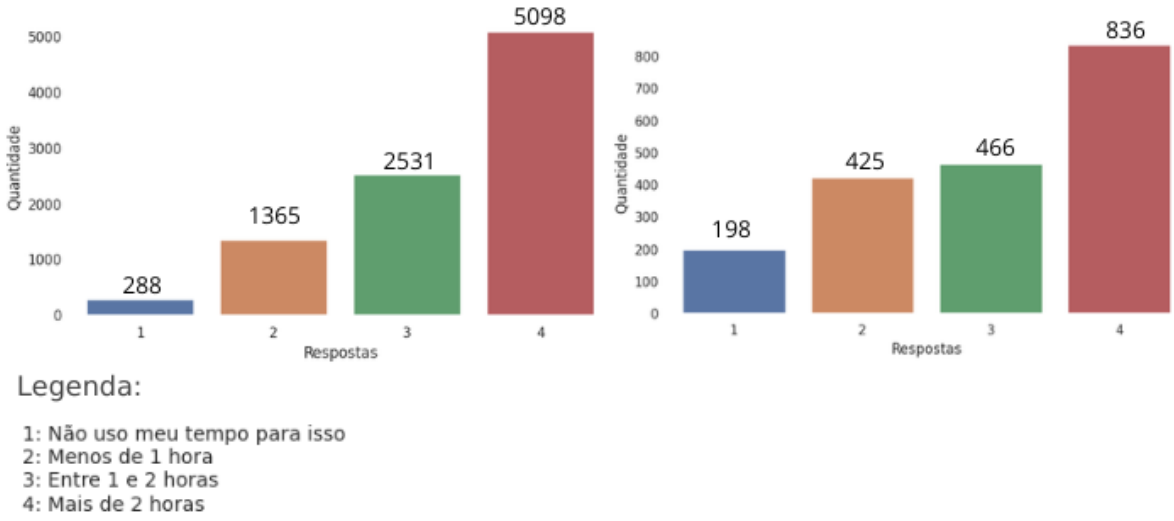


Fonte: Autor (2024).

A partir da Figura 8 é possível compreender que a distribuição de respostas para a variável indicou que a opção mais proeminente foi a que reflete a falta de conhecimento do aluno sobre o nível de escolaridade do pai, ou do homem responsável, tanto para alunos em situação crítica quanto para os demais. Embora essa opção se destaque em relação às demais, é importante notar que não chega a representar quase o dobro das outras, como observado na variável relacionada à escolaridade da mãe.

Dando continuidade ao destaque das variáveis, a terceira variável apresentada no Quadro 1 que se refere ao tempo gasto com lazer fora da escola, na Figura 9 é apresentada de forma gráfica a distribuição relacionada a variável TX_RESP_Q20e.

Figura 9: Distribuição de respostas para variável TX_RESP_Q20e.

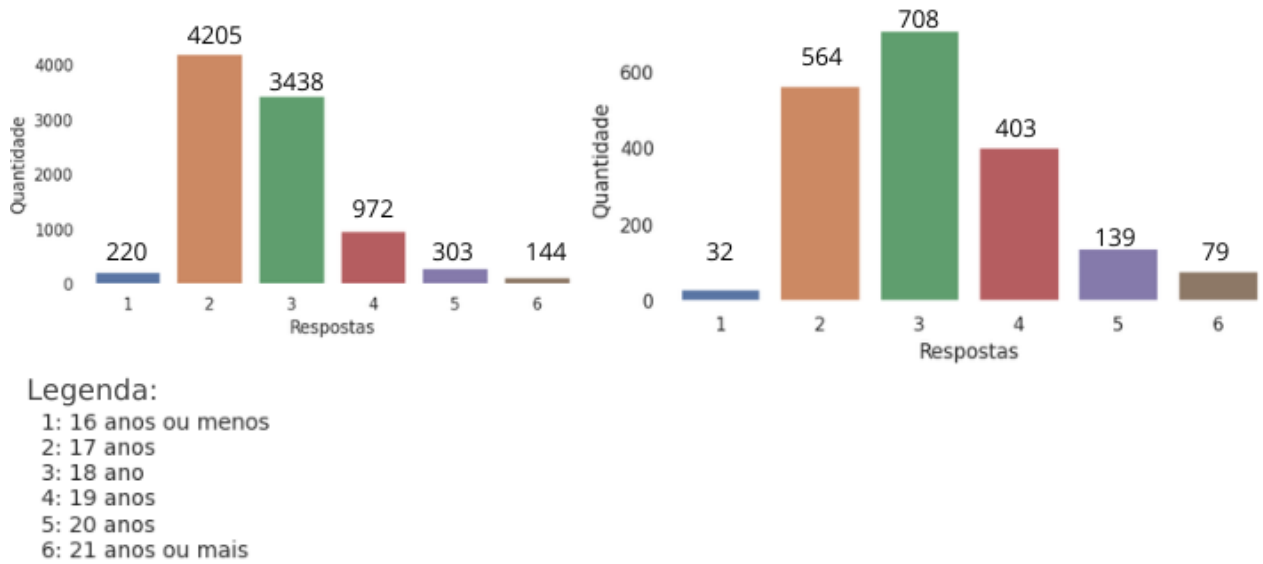


Fonte: Autor (2024).

Na Figura 9, observa-se que tanto os alunos em situação crítica quanto os demais dedicam mais de 2 horas diárias ao lazer fora da escola. Contudo, é relevante ressaltar que, ao analisar a resposta "Não uso meu tempo para isso", os alunos com proficiência crítica exibiram uma representatividade mais significativa em comparação aos demais.

Avançando para a quarta variável, que representa a idade dos alunos avaliados, na Figura 10 é exibida a distribuição das diversas categorias para a variável TX_RESP_Q02.

Figura 10: Distribuição de respostas para variável TX_RESP_Q02.

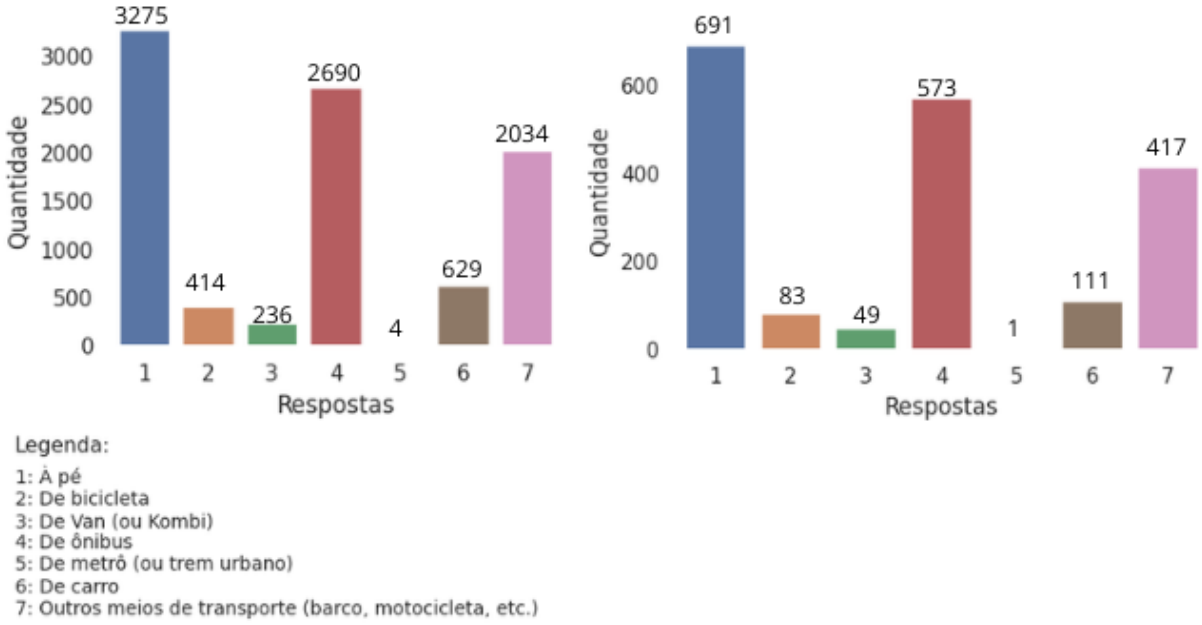


Fonte: Autor (2024).

Conforme evidenciado na Figura 10, observa-se que a distribuição desta variável concentra-se principalmente em duas respostas possíveis, independentemente do desempenho crítico ou não do aluno na avaliação. No entanto, destaca-se um aumento na distribuição de respostas para alunos com proficiência crítica na faixa etária de 18 e 19 anos. Neste contexto, a maior representatividade de respostas para esses alunos ocorre aos 18 anos, contrastando com alunos de proficiência normal, cuja maior representatividade é observada na faixa etária de 17 anos.

Por fim, é ressaltado o quinto atributo destacado no Quadro 1, relacionada à forma como os alunos se deslocam de casa para a escola, considerando a maior distância percorrida. A Figura 11 exibe graficamente a distribuição das respostas para a variável TX_RESP_Q14.

Figura 11: Distribuição de respostas para variável TX_RESP_Q14.



Fonte: Autor (2024).

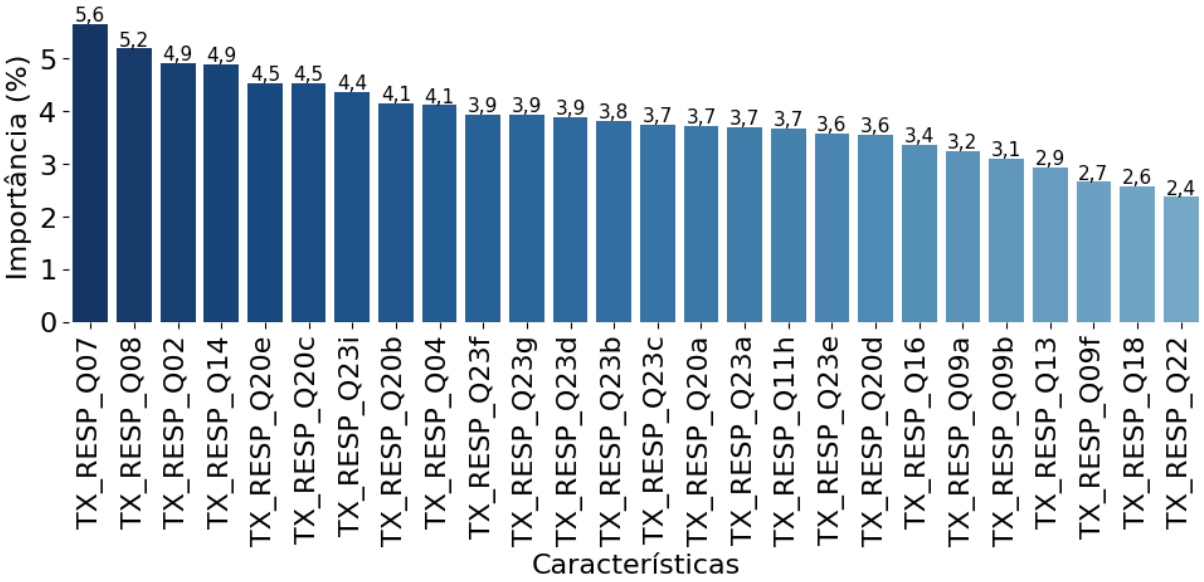
Como evidenciado na Figura 11, observa-se uma similaridade na distribuição de respostas para esta variável tanto entre alunos em situação crítica quanto entre os demais. Destaca-se que a resposta mais predominante é a afirmação de ir a pé para a escola, seguida pela opção de utilizar o ônibus como meio de deslocamento. A terceira opção mais proeminente é a utilização de outros meios, como barco e motocicleta, para se dirigir à escola.

4.2 PROFICIÊNCIA EM MATEMÁTICA

Após fazer os mesmos passos com os modelos usando a proficiência em Língua Portuguesa com variável alvo, nesta etapa o mesmo foi feito com o intuito de encontrar o modelo mais otimizado, porém utilizando o atributo relacionado à proficiência em Matemática como variável alvo, e assim alcançamos um modelo que demonstrou uma acurácia de 82,63 %. Posteriormente, foram realizadas novas implementações de modelos, implementando reduções extras de variáveis, no entanto, observou-se uma diminuição na acurácia. Utilizando o método

*feature_importances_* do Random Forest, conseguimos identificar as variáveis mais relevantes nos resultados do modelo. A representação gráfica dos resultados está apresentada na Figura 12.

Figura 12: Variação de relevância de variáveis.



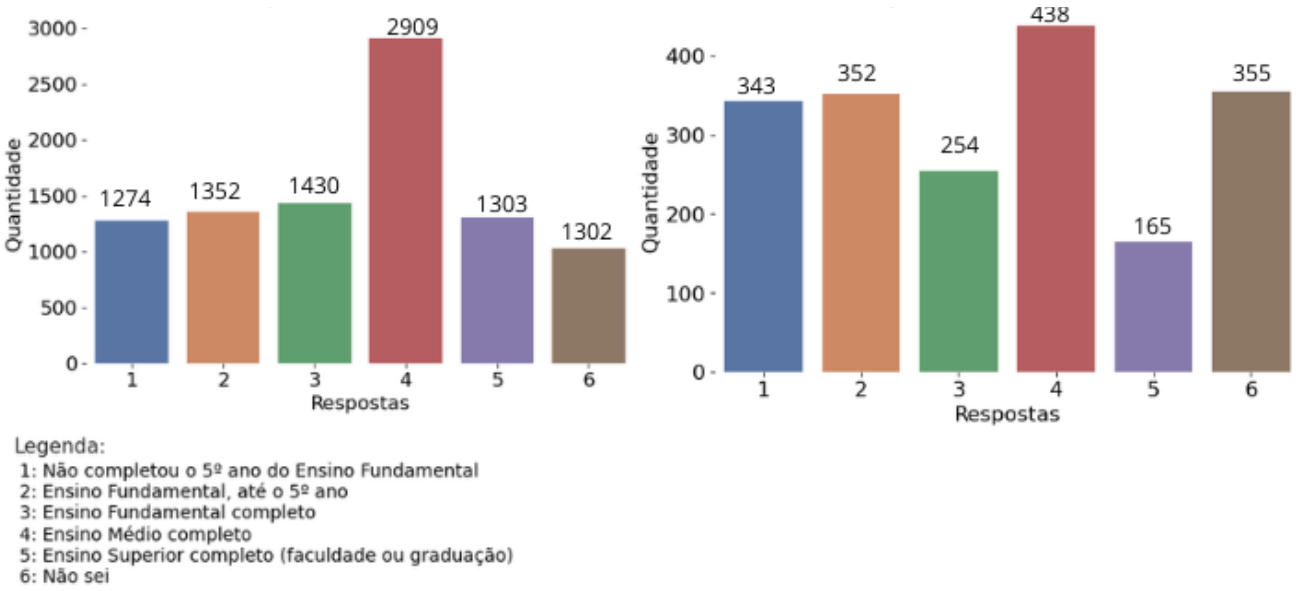
Fonte: Autor (2024).

Assim como nos resultados obtidos no modelo tendo a variável de proficiência em Língua Portuguesa como alvo, aqui também serão analisadas as cinco variáveis de maior relevância tendo a proficiência com variável alvo.

Como ilustrado na Figura 12, após a implementação de novos modelos com reduções de variáveis, utilizando a proficiência em Matemática como variável alvo, foi empregado o método *feature_importances_* do Random Forest para identificar as variáveis mais impactantes no desempenho do modelo. Notavelmente, as mesmas variáveis foram identificadas, com mudança em apenas na ordem de importância entre algumas delas, quando o modelo foi aplicado à proficiência em Língua Portuguesa. Cabe ressaltar que, inicialmente, essa correspondência nem sempre foi evidente.

Em relação à distribuição de registros para as variáveis, observou-se uma notável semelhança, destacando-se, por exemplo, a variável TX_RESP_Q07, associada ao nível de escolaridade da mãe ou mulher responsável, que foi identificada como a mais influente. A Figura 13 apresenta a distribuição de registros para a mencionada variável TX_RESP_Q07.

Figura 13: Distribuição de opções para variável TX_RESP_Q07.

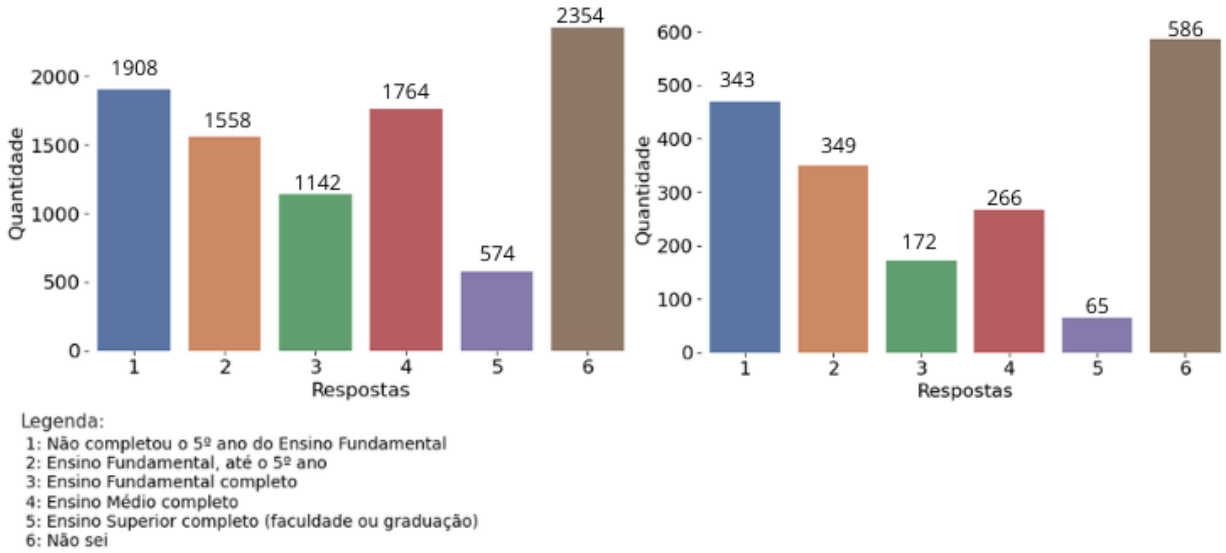


Fonte: Autor (2024).

Com base na Figura 13, observa-se que a distribuição de registros para a variável TX_RESP_Q07 no modelo, quando considerada a proficiência em matemática como variável alvo, é semelhante à distribuição correspondente à proficiência em Língua Portuguesa, mantendo a alternativa de ensino médio completo como a de maior destaque para alunos com proficiência normal, chegando a ser quase o dobro das demais quando comparando-as uma a uma. E para alunos em situação crítica a alternativa de ensino médio também é a de maior destaque, porém não tendo tanta discrepância das demais. Cabe ressaltarmos a quantidade significativa de alunos em ambas as situações em que marcaram a opção de não saberem o grau de escolaridade da mãe ou mulher responsável.

Continuando a análise dos atributos mais significativos, destacamos o atributo TX_RESP_Q08, que está vinculado ao nível de escolaridade do pai ou responsável pelo aluno. Sua alternativa mais proeminente é a opção 6, na qual o aluno declara não estar ciente do nível de escolaridade do pai ou responsável, independentemente do contexto, abrangendo tanto alunos em situação crítica quanto aqueles em situação normal na avaliação. A Figura 14 ilustra visualmente essa distribuição.

Figura 14: Distribuição de opções para variável TX_RESP_Q08.

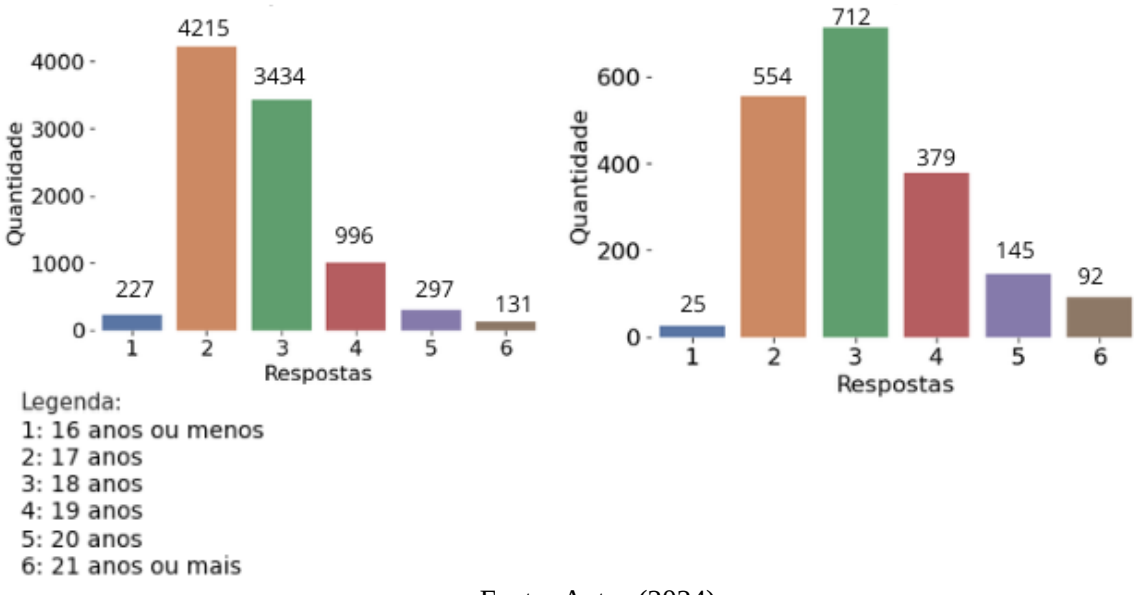


Fonte: Autor (2024).

Como evidenciado na Figura 14, apesar da opção 6 ser a de mais destaque nota-se que a distribuição entre as demais opções ocorre de forma mais diversificada no cenário em que o aluno não está em situação crítica, fator esse que não ocorre na mesma situação na variável referente ao nível de escolaridade da mãe ou mulher responsável.

Prosseguindo a análise da Figura 12, evidenciando a terceira variável mais influente sobre o desempenho do aluno na avaliação, temos a variável TX_RESP_Q02, que está associada a idade do estudante, a Figura 15 está ilustrado a distribuição entre as possíveis respostas para esta variável.

Figura 15: Distribuição de opções para variável TX_RESP_Q02.

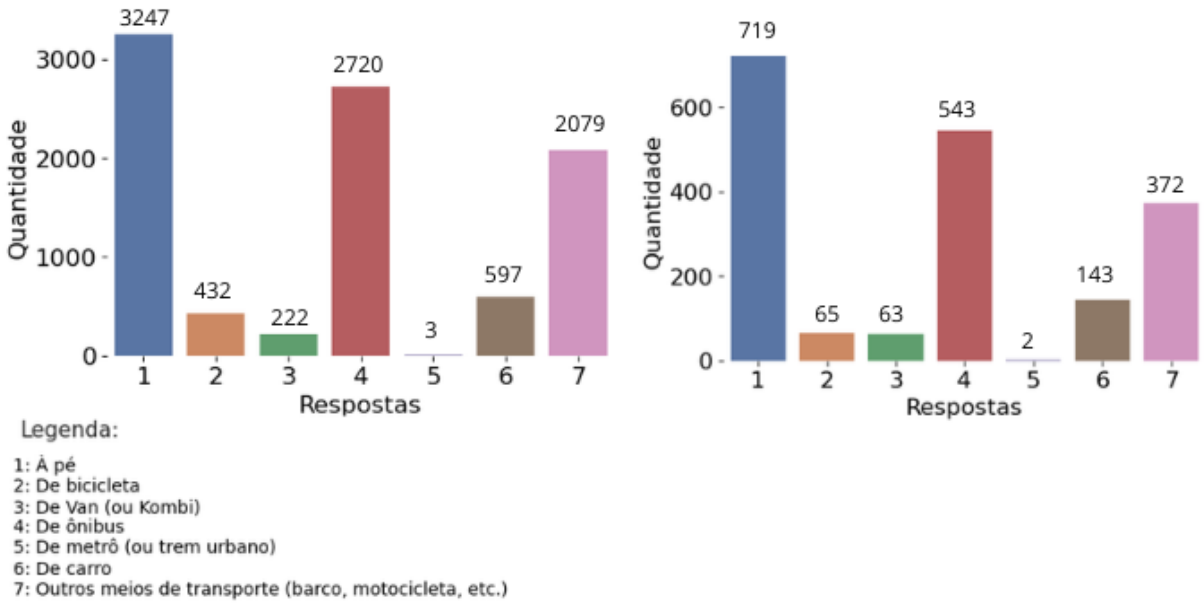


Fonte: Autor (2024).

Como evidenciado na Figura 15, é possível visualizar que a maior concentração de seleção das respostas está em duas das possíveis para o atributo em questão independentemente da situação do aluno na avaliação, opções estas sendo a 2 em que o aluno afirma ter 17 anos de idade, e a opção 3 em que o aluno afirma ter 18 anos de idade. No caso do aluno com proficiência normal a opção 1 está em maior destaque, tendo a opção 2 como a segunda de maior destaque, e para os alunos de proficiência crítica ocorre exatamente o contrário.

Analisando a figura 12, temos como a quarta variável mais influente a TX_RESP_Q14, variável esta que corresponde ao seguinte questionamento: “Considerando a maior distância percorrida, normalmente de que forma você chega à sua escola?”. Na Figura 16 é apresentada a distribuição de respostas para esta variável.

Figura 16: Distribuição de opções para variável TX_RESP_Q14.

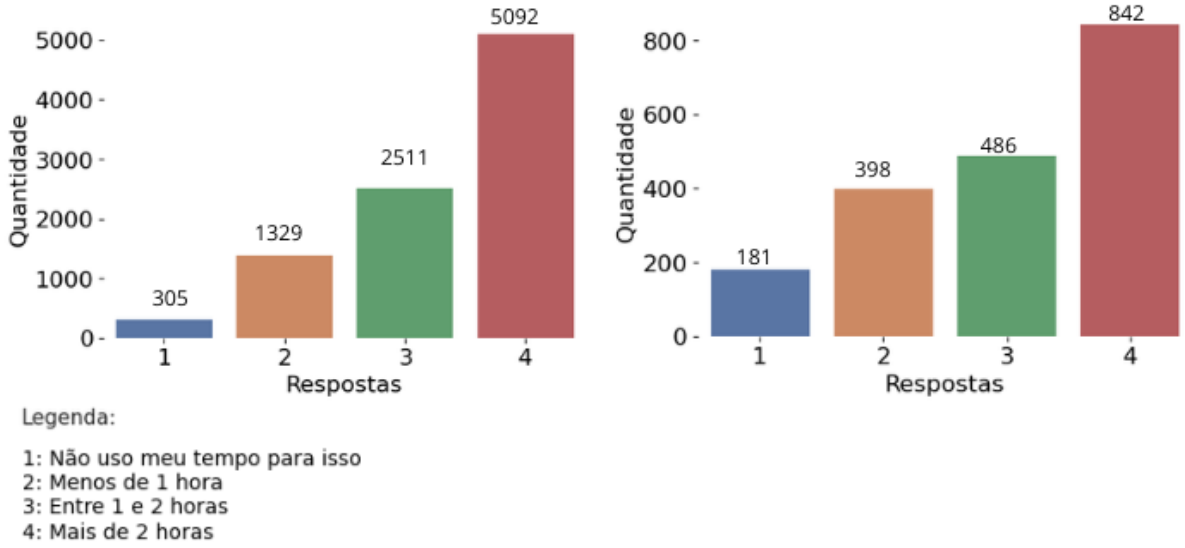


Fonte: Autor (2024).

Conforme a Figura 14 é possível notar que há uma distribuição com maior destaque em 3 opções de respostas para este atributo, e considerando que esta característica é presente tanto para alunos com proficiência normal quanto para alunos com proficiência crítica. Das três opções em destaque neste cenário a de maior predominância é a opção 1 na qual o aluno afirma ir se deslocar a pé de sua residência, considerando a maior distância, até a escola, logo em seguida temos a opção 4 tendo segunda maior relevância em que o aluno afirma ir de ônibus para a escola e pôr fim a terceira seleção para este atributo, sendo a de número 7 em que o aluno afirma ir para a escola por com outros meios de transportes como motocicleta e barcos.

Dando continuidade ao detalhamento da Figura 12, temos a variável TX_RESP_Q20e que está diretamente relacionada a quanto tempo de lazer o aluno tem fora da escola. Na Figura 17 está ilustrada a distribuição das possíveis alternativas para esta variável.

Figura 17: Distribuição de opções para variável TX_RESP_Q20e.



Fonte: Autor (2024).

Como apresentado na Figura 15, nota-se que a alternativa de maior escolha dentre as demais foi a de número 4, na qual o aluno, tanto com proficiência normal quanto crítica, afirma ter mais de 2 horas destinadas a lazer fora da escola. Uma outra opção que merece ser abordada é o fato de ter alunos afirmando no questionário que não usam seu tempo para tal atividade. Ao encerrar este capítulo, avançamos para o próximo estágio. No Capítulo 5, serão apresentadas as conclusões derivadas dos resultados discutidos anteriormente, além de oferecer sugestões para futuras pesquisas.

5 CONCLUSÕES E SUGESTÕES DE TRABALHOS FUTUROS

Após a condução completa da pesquisa, emerge uma série de conclusões fundamentadas nos resultados obtidos. Destaca-se que o fator de maior influência sobre o desempenho dos alunos da 3ª série do ensino médio das escolas públicas no estado do Piauí nos resultados do SAEB está relacionado ao atributo TX_RESP_Q07, relacionada à escolaridade da mãe ou da mulher responsável pelo aluno. Essa influência se mantém consistente em ambas as áreas avaliadas, tanto em proficiência em Língua Portuguesa quanto em Matemática. Além disso, é relevante ressaltar a influência significativa da variável TX_RESP_Q08, referente ao nível de escolaridade do pai ou do homem responsável pelo aluno, embora esta se posicione logo após a variável anteriormente mencionada em ambas as categorias de proficiência.

Um ponto a ser enfatizado é a presença de registros significativos em ambos os casos em que os alunos informaram não saber o nível de escolaridade do respectivo responsável, com uma prevalência ainda maior no caso do pai ou homem responsável, e isso sugere uma lacuna na informação socioeconômica dos alunos, o que pode afetar a compreensão completa do impacto da escolaridade dos pais ou responsáveis no desempenho dos alunos. Assim, considerando que essas duas variáveis se destacaram como as mais influentes em ambas as áreas de aplicação da prova, conclui-se que os fatores socioeconômicos mais determinantes para o desempenho dos alunos são os níveis de escolaridade dos pais ou responsáveis.

Tendo em vista os resultados encontrados e as conclusões criadas em cima destes resultados fica compreensível que para um aumento do desempenho acadêmico dos alunos seria interessante um maior acompanhamento dos pais ou responsáveis na vida escolar dos alunos, assim como alguma forma de capacitação escolar por parte dos pais dos alunos.

5.1 SUGESTÕES DE TRABALHOS FUTUROS

Apesar das conclusões alcançadas com os resultados apresentados, ainda há oportunidades para explorar ideias de pesquisa futuras relacionadas ao tema. Uma dessas possibilidades é a investigação e aplicação de técnicas adicionais de mineração de dados, como redes neurais, com o intuito de comparar os resultados obtidos com o algoritmo Random Forest. Tal comparação permitiria avaliar se existem diferenças significativas nas variáveis identificadas como mais relevantes para o desempenho acadêmico dos alunos.

Outra abordagem promissora seria a análise das disparidades de gênero na relação entre os fatores socioeconômicos e o desempenho escolar. Especial atenção seria dada à variação nas respostas dos alunos em relação à escolaridade dos pais ou responsáveis. Essa investigação poderia fornecer insights valiosos sobre como fatores socioeconômicos específicos afetam meninos e meninas de maneiras distintas no contexto educacional.

REFERÊNCIAS

ARAÚJO, Herlane Martins et al. Mineração de dados educacionais: um estudo sobre a proficiência em matemática no Ceará. Brazilian Journal of Development, 2022. Disponível em: https://doi.org/10.34117/bjdv8n10-214. Acesso em: 14 nov. 2023.

BREIMAN, L. Random Forests. Machine Learning, 45, 5-32, 2001.

BIAU, G. Analysis of a Random Forests Model. Journal of Machine Learning Research, 13, 1063-1095, 2012.

DA SILVA LIMA, M. .; PEREIRA DE ANDRADE, A. .; MARIA DA SILVA, E.; TARCIANA CUNHA SILVA MARTINS, M. . A IMPORTÂNCIA DO SISTEMA DE AVALIAÇÃO DA EDUCAÇÃO BÁSICA. Humanas em Perspectiva, [S. l.], v. 4, 2022. DOI: 10.51249/hp04.2022.868. Disponível em: https://periodicojs.com.br/index.php/hp/article/view/868. Acesso em: 15 fev. 2024.

FAYYAD, U. M. et al. Advances in Knowledge Discovery & Data Mining. 1. ed. Menlo Park, Califórnia: American Association for Artificial Intelligence, 1996. 611 p.

IBGE. Mapa de pobreza e desigualdade, 2023. Disponível em: https://cidades.ibge.gov.br/brasil/pi/pesquisa/36/30252. Acesso em: 15 nov. de 2023.

INÊS, Maria Pestana. Trajetória do Saeb: criação, amadurecimento e desafios. Em Aberto, 2016. Disponível em: https://doi.org/10.24109/2176-6673.emaberto.29i96.%25p. Acesso em: 15 fev. 2024.

J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, USA, 2001.

PYOD. Python Outlier Detection, 2017. Disponível em: https://pyod.readthedocs.io/en/latest/. Acesso em: 17 nov. de 2023.

RAMOS, Jorge L. Cavalcanti; RODRIGUES, Rodrigo Lins; SILVA, João C. Sedraz; OLIVEIRA, Pamella L. Silva de. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. Sociedade Brasileira de Computação, 2020. Disponível em: https://doi.org/10.5753/cbie.sbie.2020.1092. Acesso em: 12 nov. 2023.

SAEB, Sistema de Avaliação da Educação Básica - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2023 , Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb. Acesso em: 17 nov. 2023.

SOARES, Raimundo de Castro; NETO, Nelson Weber; COUTINHO, Luciano Reis; SANTOS, Davi Viana dos; SILVA, Francisco José da Silva e; TELES, Ariel Soares. Minerando Dados para Entender os Fatores de Influência da Qualidade Educacional do Maranhão. REVISTA BRASILEIRA DE INFORMÁTICA NA EDUCAÇÃO, 2023. Disponível em: https://doi.org/10.5753/rbie.2023.2831. Acesso em: 12 nov. 2023.

SOTNIKOV, Dmitry V.; LYLY, Mika; SALMI, Tiina. Prediction of 2G HTS Tape Quench Behavior by Random Forest Model Trained on 2-D FEM Simulations. Institute of Electrical and Electronics Engineers(IEEE), 2023. Disponível em: https://doi-org.ez117.periodicos.capes.gov.br/10.1109/TASC.2023.3262212. Acesso em: 12 nov. 2023.

SCIKIT-LEARN. Machine Learning in Python, 2023. Disponível em: https://scikit-learn.org/stable/. Acesso em: 17 nov. de 2023.

VERDIER, G.; ROSA, A.F.P. Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing. IEEE Transactions on Semiconductor Manufacturing, v.24, n.1, p.59-68, 2011.

WANG, Ying. Lung Cancer Prediction Based on Learning Machine. Institute of Electrical and Electronics EngineerWANG, Ying. Lung Cancer Prediction Based on Learning Machine. Institute of Electrical and Electronics Engineers(IEEE), 2023. Disponível em: https://doi-org.ez117.periodicos.capes.gov.br/10.1109/ICIPCA59209.2023.10257678. Acesso em: 20 jan. 2024.s(IEEE), 2023. Disponível em: https://doi-org.ez117.periodicos.capes.gov.br/10.1109/ICIPCA59209.2023.10257678. Acesso em: 20 jan. 2024.

YU, Jing et al. Research on Feature Importance of Gait Mechanomyography Signal Based on Random Forest. Institute of Electrical and Electronics Engineers(IEEE), 2020. Disponível em: https://doi-org.ez117.periodicos.capes.gov.br/10.1109/CVIDL51233.2020.00045. Acesso em: 11 jan. 2024.