Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
http://br-ie.org/pub/index.php/rbie

Submission: 15/Oct/2022; 1st round notif.: 25/Nov/2022; New version: 16/Dec/2022; 2nd round notif.: 10/Apr/2023;
Camera ready: 20/Apr/2023; Edition review: 28/Jun/2023; Available online: 03/Jul/2023; Published: 03/Jul/2023;

Minerando Dados para Entender os Fatores de Influência da Qualidade Educacional do Maranhão

Title: Mining Data to Understand Influence Factors of Educational Quality in Maranhão

Raimundo de Castro Soares
Universidade Federal do Maranhão (UFMA)
ORCID: 0000-0002-0904-0370
soares.raimundo@ufma.br

Nelson Weber Neto
Universidade Federal do Maranhão (UFMA)
ORCID: 0000-0002-1136-768X
nelsonweberneto@gmail.com

Luciano Reis Coutinho
Universidade Federal do Maranhão (UFMA)
ORCID: 0000-0001-7996-7334
luciano.rc@ufma.br

Davi Viana dos Santos
Universidade Federal do Maranhão (UFMA)
ORCID: 0000-0003-0470-549X
davi.viana@lsdi.ufma.br

Francisco José da Silva e Silva
Universidade Federal do Maranhão (UFMA)
ORCID: 0000-0001-8339-3679
fssilva@lsdi.ufma.br

Ariel Soares Teles
Instituto Federal do Maranhão (IFMA)
ORCID: 0000-0002-0840-3870
ariel.teles@ifma.edu.br

Resumo

O estado do Maranhão apresenta índices baixos na qualidade da educação básica, conforme pode ser verificado nas avaliações de desempenho nacionais ao longo dos anos. O problema da baixa qualidade educacional pode ser abordado e investigado do ponto de vista de diversas áreas. Uma delas é a utilização de Mineração de Dados Educacionais, que está cada vez mais presente em estudos científicos, e também é utilizada para dar suporte à tomada de decisão por elaboradores de políticas públicas. No entanto, as pesquisas com os dados de uma região geográfica, ou determinada localidade, ainda podem ser escassas, sendo o caso do estado do Maranhão. Essa pesquisa tem o objetivo de entender quais são os fatores que influenciam na qualidade da educação pública do estado do Maranhão. Para isso, foram utilizadas técnicas de mineração de dados, tais como análise exploratória de dados, análise de correlação, análise de fatores, regressão e árvore de decisão. Os dados utilizados são das escolas públicas estaduais de ensino médio do Maranhão. Os resultados desse estudo mostram um diagnóstico da situação educacional do estado, com a identificação do que influencia significativamente no desempenho da educação.

Palavras-chave: *Mineração de Dados Educacionais, Maranhão, Educação, Análise de Correlação, Análise de Fatores, Regressão, Árvore de Decisão.*

Abstract

The state of Maranhão has low levels of quality in basic education, as can be seen in the national performance assessments over the years. The problem of low educational quality can be approached and investigated from the view point of several areas. One of them is the use of Educational Data Mining, which is increasingly present in scientific studies, and is also used to support decision-making by public policy makers. However, research with data from a geographic region, or a particular location, may still be scarce, as is the case in the state of Maranhão. This research aims to understand what are the factors that influence the quality of public education in the state of Maranhão. For this purpose, data mining techniques were used, such as exploratory data analysis, correlation analysis, factor analysis, regression and decision tree. The data used are from state public high schools in Maranhão. Results of this study show a diagnosis of the state's educational situation, with the identification of what significantly influences the performance of education.

Keywords: *Educational Data Mining, Maranhão, Education, Correlation Analysis, Factor Analysis, Regression, Decision Tree.*

Cite as: Soares, R. de C., Weber Neto, N., Coutinho, L. R.; Santos, D. V.; Silva, F. J. S. & Teles, A. S. Minerando Dados para Entender os Fatores de Influência da Qualidade Educacional do Maranhão. Revista Brasileira de Informática na Educação, 31, 378-406. DOI: 10.5753/rbie.2023.2831.

1 Introdução

A mineração de dados, através da junção da estatística e a inteligência computacional (Namen, Borges, & Sadala, 2013), e usando metodologias próprias, trata e processa bases de dados, visando extrair informações relevantes (Tan, Steinbach, & Kumar, 2009) e encontrar padrões relacionados a eles (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Ela está cada vez mais sendo utilizada nas diversas áreas do conhecimento, incluindo a educação. Considerando a mineração de dados como uma área que realiza o processamento de bases de dados com técnicas apropriadas, pode-se afirmar que todas as áreas que dispõem de dados suficientes podem fazer uso dela.

A mineração de dados dispõe de ferramentas que podem analisar os mais diversos conjuntos de dados em busca de padrões e evidências, e proporcionar a descoberta de conhecimento não disponíveis em ferramentas comuns de banco de dados (Agarwal, 2013). Quando essa análise é feita com dados da educação, buscando padrões em estudantes, predizendo resultados, identificando fatores, formas de aprendizagem e de interação, é chamada de Mineração de Dados Educacionais (MDE) (Peña-Ayala, 2014).

O Ministério da Educação (MEC) do Brasil, através de avaliações e questionários direcionados a professores, alunos e gestão escolar, produz diversos dados relevantes (Fonseca & Namen, 2016), tais como a quantidade de alunos matriculados, dados socioeconômicos, dependência administrativa das escolas, formação dos professores, localização das escolas, dentre outros. Esses dados são fontes de informações que, quando analisados detalhadamente, podem evidenciar o diagnóstico educacional em diversos aspectos. O Instituto Nacional de Estudos Pedagógicos Anísio Teixeira (INEP) é uma agência do MEC que disponibiliza publicamente bases de dados sobre a educação nacional. Os dados disponibilizados pelo INEP, quando analisados, podem mostrar a realidade boa ou ruim da educação brasileira. Essa análise pode ser feita usando técnicas e algoritmos de mineração de dados.

O diagnóstico da educação pode ser feito a partir de uma série de indicadores, como o desempenho escolar , a distorção entre idade e série, bem como a evasão e progressão dos alunos ao longo dos anos (i.e., a eficiência do sistema de ensino por meio do fluxo escolar) (Rigo, Cambruzzi, Barbosa, & Cazella, 2014), as condições oferecidas pela rede de ensino, os recursos disponíveis na escola, e a qualificação dos professores (Rigotti & Cerqueira, 2015). No Brasil, o Índice de Desenvolvimento da Educação Básica (IDEB) é o indicador utilizado para medir a qualidade da educação básica.

A educação pode ser influenciada por diversos fatores, que podem causar impactos positivos ou negativos. Por exemplo, evasão e reprovação de alunos são itens que impactam negativamente na educação. Além disso, esses fatores podem contribuir para a diminuição dos índices de qualidade educacional. Já a participação dos pais e da família, dando incentivo aos alunos, são itens que podem impactar positivamente na educação. Uma vez identificados os fatores que podem influenciar na qualidade educacional, pode-se criar políticas públicas para alavancar esses fatores no sentido de fazer com que eles sejam melhorados ou aperfeiçoados para causar um impacto ainda melhor e mais evidente da educação na sociedade. Ou, caso o fator seja negativo (e.g., evasão, reprovação), pode ser estudado sua origem, para ser combatido.

O estado do Maranhão apresenta baixos índices educacionais. A pontuação média do IDEB no Maranhão nos últimos anos, referente às escolas públicas de ensino médio, melhorou, mas não

significativamente. O último índice (3,7), referente ao ano de 2019, ainda está abaixo da nota dos estados com melhor pontuação (>=4,0), ou mesmo da nota média nacional (3,9) (INEP, 2022b). Esse baixo indicador pode ter várias causas, que precisam ser investigadas e estudadas, para a identificação de eventuais fatores que podem proporcionar sua melhoria. Dessa forma, o problema abordado nesta pesquisa está relacionado ao fato de não haver um conhecimento profundo, sob o aspecto da análise e mineração de dados, dos fatores que causam impactos e influência na qualidade educacional do estado do Maranhão.

O motivo pelo qual o IDEB no Maranhão não evoluiu significativamente é algo que deve ser estudado, buscando identificar os possíveis fatores que mais influenciam o desempenho das escolas do estado. Uma análise dos dados relacionados ao IDEB pode gerar evidências que expliquem o problema do baixo rendimento escolar, ao menos em parte, e contribuir para a tomada de decisão dos gestores educacionais.

O problema dos baixos índices educacionais maranhenses pode ser causado por diversos fatores, os quais essa pesquisa procura identificar e analisar. Portanto, o objetivo desta pesquisa é identificar os fatores de influência na qualidade educacional do Maranhão através do uso de técnicas de mineração das bases de dados fornecidas pelo INEP: IDEB e Sistema de Avaliação da Educação Básica (SAEB). Especificamente, este estudo objetiva realizar uma análise exploratória dos dados do ensino médio da educação pública do Maranhão, buscando identificar as variáveis de maior influência na nota do IDEB das escolas públicas estaduais. Além disso, este trabalho também cria modelos para estudar o ensino médio da educação pública estadual maranhense.

Este artigo está organizado como segue. A Seção 2 discute os trabalhos relacionados que se assemelham ou estudam tópicos similares a este estudo. A Seção 3 apresenta a metodologia utilizada. A Seção 4 expõe o entendimento do negócio e entendimento dos dados. A Seção 5 explica a preparação e modelagem dos dados, enquanto a Seção 6 discute os resultados. A Seção 7 finaliza o artigo com as considerações finais.

2 Trabalhos Relacionados

Inicialmente, os trabalhos relacionados a esta pesquisa foram identificados e analisados através de uma Revisão Sistemática da Literatura (RSL) (Castro Soares et al., 2021), onde foi realizado o levantamento do estado da arte referente à mineração de dados com as bases disponibilizadas pelo INEP. Embora a RSL tenha identificado 19 (Castro Soares et al., 2021) artigos, a seguir são descritos somente aqueles que mais se relacionam com esta pesquisa, devido realizarem estudos de MDE com dados de uma região específica do Brasil. Adicionalmente, considerou-se nesta análise de trabalhos relacionados outros estudos que também focaram na mineração de dados da educação básica brasileira, não necessariamente utilizando as bases do INEP.

Júnior et al. (Júnior, Nascimento, Alves, & Gouveia, 2017) utilizaram a base de dados do Censo Escolar e do Exame Nacional do Ensino Médio (ENEM), referentes aos anos 2014 e 2015, para realizar a análise de correlação e identificação de *outliers*, com os dados do estado de Pernambuco. Essa pesquisa visou analisar a correlação entre as características das escolas (água filtrada, água da rede pública, sistema de esgoto e coleta de lixo) e o desempenho dos alunos no ENEM (notas de Ciências Humanas, Linguagens, Ciências Naturais e Ciências Exatas). O

índice de correlação foi superior a 0,7 para todas as áreas do conhecimento. Os pesquisadores concluíram que, seja qual for a área do conhecimento, se a nota de um aluno aumentar em uma determinada área, também tende a aumentar em outra. Ao tentar identificar *outliers*, não foram encontradas discrepâncias entre as notas do ENEM e as características de infraestrutura das escolas. Os índices atribuídos às infraestruturas das escolas e notas analisadas apresentaram características semelhantes.

Carvalho et al. (Carvalho, Cruz, & Gouveia, 2017) utilizaram a base de dados dos Censos da Educação Básica e Superior, referentes aos anos de 2014 e 2015, para estudar o problema da evasão de alunos no âmbito do estado de Pernambuco. A mineração de dados foi realizada pela metodologia *Knowledge Discovery in Databases* (KDD) com o software WEKA, que dispõe de algoritmos de mineração de dados. Os algoritmos usados pelos pesquisadores foram: Árvore de Decisão (Algoritmo J48) e Classificação Bayesiana (*Naive Bayes*). Segundo os autores, a escolha do Algoritmo J48 foi devido à técnica de classificação por árvores de decisão gerar regras objetivas que facilitam a interpretação dos resultados (Witten, Frank, Hall, Pal, & DATA, 2005), e o J48 alcançar resultados considerados aceitáveis na métrica de acurácia. Já o algoritmo *Naive Bayes* foi escolhido por elaborar classificação probabilística simples. Como resultado da pesquisa, eles conseguiram elaborar perfis das escolas quanto à infraestrutura, perfis de alunos da educação básica consoante a região da escola em que estudam, perfis de alunos da educação superior dos cursos da área de Tecnologia da Informação e Comunicação, e também fizeram a comparação entre os perfis desses alunos que residiam na região metropolitana da capital (Recife) e no interior do estado.

Os pesquisadores Bem et al. (Nascimento Bem, da Silva Pereira, & Souza, 2017) utilizaram os dados do IDEB para a criação de um *data mart* (Barbieri, 2001) com o objetivo de realizar a análise comparativa das cidades da microrregião do Pajeú, no estado de Pernambuco. A solução desenvolvida foi capaz de mostrar os dados graficamente através de tabelas, com números, gráficos e cores, possibilitando a análise, comparação e tomada de decisão. A conclusão dos pesquisadores foi que os dados precisam ser disponibilizados de maneira facilitada e entendível, o que pode ser feito através da ferramenta utilizada.

Os pesquisadores Pinto et al. (Silva Pinto, Júnior, & de Barros Costa, 2019) utilizaram a base de dados do SAEB para identificar os fatores que afetam o desempenho escolar dos alunos (notas do IDEB) dos anos finais (9º ano) do ensino fundamental das escolas públicas municipais da cidade de Teotônio Vilela, estado de Alagoas, referente aos anos de 2015 e 2017. A metodologia CRISP-DM foi utilizada pelos autores para a realização do estudo. Visando gerar dados sintéticos para equilibrar a base de dados para as variáveis dependentes, foi utilizada técnicas de balanceamento de dados através do *Synthetic Minority Oversampling Techniques* (SMOTE). O estudo identificou os atributos que mais influenciam nas notas do IDEB dos alunos da rede pública municipal de Alagoas e, dessa forma, os autores concluíram que, uma vez conhecendo esses atributos, pode-se ter uma ideia de como melhorar os índices educacionais.

Os pesquisadores Pinto et al. (Silva Pinto, Júnior, Costa, Barbirato, & Rodrigues, 2019), através de mineração de dados, analisaram os resultados de avaliações oficiais realizadas pelo INEP para analisar a influência no desempenho do IDEB. O estudo recorreu à metodologia CRISP-DM e trabalhou com dados de 13 escolas da rede pública municipal do ensino médio da capital do estado de Alagoas, Maceió. A pesquisa identificou 10 atributos que mais influenciam o desempenho dos alunos nas disciplinas Português e Matemática. Os algoritmos utilizados J48, OneR,

JRip e LibSVM tiveram resultados com 100% de acurácia de classificação dos atributos com mais influência nas notas dos alunos para os dados das provas de português e matemática.

A pesquisadora Pacini (Pacini, 2020) realizou uma pesquisa de mineração de dados para analisar os indicadores educacionais de esforço do corpo docente (esforço realizado pelos docentes da educação básica brasileira no exercício da sua profissão), regularidade do corpo docente (permanência dos professores nas escolas nos últimos cinco anos) e adequação da formação do docente (se a formação do professor é adequada ou não à disciplina que leciona). A pesquisa considerou essas variáveis em relação à média de proficiência da Prova Brasil de Língua Portuguesa e Matemática, na edição do ano de 2015, do 5º e 9º ano do Ensino Fundamental, da rede estadual de ensino do estado de Tocantins, com os dados do INEP. Na exploração dos resultados da pesquisa, foram utilizadas soluções de software e técnicas de estatística em mineração de dados, juntamente com planilhas eletrônicas. A análise identificou atributos dos indicadores com significância estatística para as escolas que tiveram melhor desempenho na Prova Brasil.

Os autores Silva et al. (Santos & de Medeiros, 2020) utilizaram a base do IDEB e dados do Portal da Transparência de financiamento federal da educação básica do estado da Paraíba, visando estudar se havia uma relação do investimento federal com a qualidade da educação nos anos de 2016 e 2017. Eles utilizaram a metodologia KDD, e os experimentos foram realizados no software R, recorrendo à Regressão Linear Simples e Correlação de Pearson. As notas do IDEB foram divididas em duas categorias: anos iniciais e anos finais. A regressão realizada não mostrou resultado significativo, portanto, não foi encontrada linearidade na relação entre o investimento Federal e o IDEB. Os resultados da correlação de Pearson se mostraram também pouco significativos, uma vez que o coeficiente de correlação resultante para o IDEB nos anos iniciais e finais foram, respectivamente, de 0,01 e 0,05. Os pesquisadores concluíram não haver relação direta entre o investimento federal e o IDEB. Portanto, somente os dados de financiamento federal na educação municipal são pouco representativos ou não são suficientes para obter uma relação entre o investimento federal bianual e a média do IDEB para os municípios.

Fernandes et al. (Fernandes et al., 2019) realizaram a predição do desempenho de alunos de escolas públicas localizadas no Distrito Federal referente aos anos letivos de 2015 e 2016. Eles desenvolveram modelos de classificação baseados no *Gradient Boosting Machine* (GBM) para realizar a predição da aprovação dos alunos no final do ano letivo. O estudo utilizou dois conjuntos de dados: um obtido antes do início do ano letivo, e outro com variáveis obtidas dois meses após o início do ano letivo. Os pesquisadores concluíram que os atributos relacionados a notas e ausências foram os mais relevantes para predição do resultado acadêmico. Adicionalmente, outras variáveis (e.g., bairro e escola) também contribuíram no desempenho dos alunos.

Lima et al. (Lima, Ferreira, & Silva, 2021), considerando o fato de poucas mulheres no mundo se interessarem por estudarem ou atuarem nas áreas de ciências, tecnologia, engenharia e matemática (do inglês, *Science, Technology, Engineering, and Mathematics* - STEM), realizaram um estudo de caso sobre o projeto "Elas na Robotica" em uma cidade de médio porte do interior da região sudeste do Brasil. O estudo foi feito por meio da análise de sentimentos, opiniões e análises estatísticas nos dados de respostas de questionários aplicados a professoras e alunas mulheres, e demais respondentes externos ao projeto. Por fim, ao aplicar técnicas de aprendizado de máquina não supervisionado (agrupamento), a análise focou no grupo de interesse do projeto: mulheres que se interessam por STEM, mas sem conhecimento em robótica.

Fonseca e Namen (Fonseca & Namen, 2016) realizaram um estudo de mineração de dados utilizando as bases do SAEB e a metodologia KDD visando identificar variáveis relacionadas ao perfil de professores de Matemática que podem influenciar, tanto de forma positiva quanto negativa, o desempenho dos alunos. Os dados utilizados foram do ano de 2011, referentes ao 9º ano do ensino fundamental do estado do Rio de Janeiro. Os pesquisadores puderam identificar características do perfil dos professores que contribuem para o bom desempenho do aluno, tais como a estabilidade do vínculo profissional, professores que não faltam e professores com visão positiva do aluno.

A Tabela 1 apresenta uma síntese dos trabalhos relacionados. A primeira coluna da tabela traz a referência e o ano das publicações. A segunda coluna informa as bases de dados, enquanto a terceira apresenta o problema estudado. A quarta coluna informa a metodologia utilizada, e a última mostra a(s) técnica(s) de mineração de dados utilizada(s) em cada trabalho.

Tabela 1: Comparação dos trabalhos selecionados.

Referência	**Dados**	**Problema**	**Metodologia**	**Técnica**
(Fonseca & Namen, 2016)	SAEB	KDD	Desempenho dos alunos	Algoritmo de classificação
(Júnior et al., 2017)	Censo	Baixo desempenho escolar	KDD	Análise de correlação e identificação de *outliers*
(Carvalho et al., 2017)	Censo	Evasão de alunos	KDD	Árvore de decisão e Classificação Bayesiana
(Nascimento Bem et al., 2017)	IDEB	Falta de padronização entre bases educacionais	Proposta de solução	Ferramentas OLAP
(Silva Pinto, Júnior, & de Barros Costa, 2019)	SAEB	Baixo Desempenho escolar	CRISP-DM	Algoritmos de classificação
(Silva Pinto, Júnior, Costa, et al., 2019)	SAEB	Baixo desempenho escolar	CRISP-DM	Algoritmos de classificação
(Fernandes et al., 2019)	iEducar software	Reprovação	CRISP-DM	Gradient Boosting Machine
(Pacini, 2020)	SAEB	Baixo desempenho escolar	Análise exploratória	Análise de correlação
(Santos & de Medeiros, 2020)	IDEB	Baixo investimento na educação	KDD	Análise de correlação e Modelo de regressão
(Lima et al., 2021)	Questionário próprio	Interesse das mulheres por STEM	Análise estatística	Clusterização
Este estudo	SAEB e IDEB	Baixo desempenho escolar	CRISP-DM	Análise exploratória de dados, análise de correlação, análise de fatores, regressão e árvore de decisão

Embora existam diversos trabalhos que abordam a qualidade educacional do ensino básico brasileiro, buscando identificar as causas para os índices baixos, esta pesquisa se faz necessária,

pois, para o melhor do nosso conhecimento, não foram identificadas estudos semelhantes, utilizando mineração de dados sobre a educação do estado do Maranhão. Portando, há uma carência, em particular na literatura científica, em realizar o processo de mineração em bases de dados educacionais do estado do Maranhão. Nesse sentido, a ausência de estudos voltados à mineração de dados educacionais no Maranhão demonstra uma subutilização dos dados disponíveis, evidenciando a necessidade deste estudo.

3 Metodologia

As metodologias empregadas em mineração de dados são utilizadas para uma melhor otimização do processo, seguindo uma sequência de etapas e procedimentos e, dessa forma, conseguir atingir os objetivos pretendidos. A metodologia CRISP-DM foi adotada para realizar esta pesquisa. Ela pode ser adaptada a qualquer categoria de negócio e trata-se de uma metodologia interativa, ou seja, executam-se suas etapas, podendo voltar para uma etapa anterior caso esta precise ser refeita. Ela está dividida em 6 (seis) etapas, ilustradas na Figura 1 (adaptada de (Shearer, 2000)) e explicadas em seguida (Azevedo & Santos, 2008; Shearer, 2000).

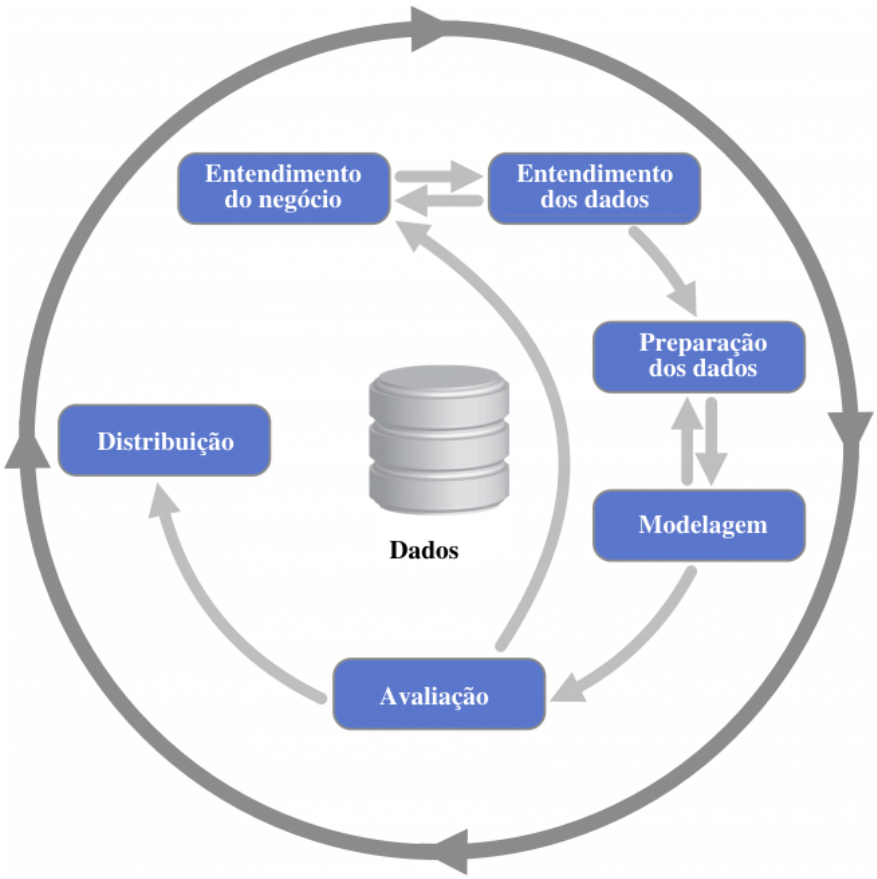

Figura 1: Ciclo da Metodologia CRISP-DM.

- **Entendimento do negócio:** nesta primeira etapa é feito o entendimento dos objetivos do

negócio, bem como a elaboração de uma ou mais questões de pesquisa visando encontrar respostas através da mineração de dados;

- **Entendimento dos dados:** nessa etapa é feita a coleta dos dados e já se inicia uma análise prévia para seu entendimento. Eventualmente, poderão ser identificados problemas e os primeiros *insights*;
- **Preparação dos dados:** esta etapa envolve a preparação da base de dados que pode ter que passar por diversos processos. A preparação pode incluir: padronização e normalização dos dados, exclusão de variáveis, junção com outras bases, alteração de nomes de variáveis para um melhor entendimento, e eventuais outras modificações necessárias;
- **Modelagem:** nesta fase ocorre a elaboração de um ou mais modelos para a análise dos dados (e.g., árvores de decisão, regressão). Diversos modelos podem ser criados, analisados e testados, e posteriormente escolher um deles (ou mais de um) para resolver o problema e/ou responder à questão de pesquisa;
- **Avaliação:** nesta etapa é feita uma verificação se o modelo (ou os modelos, em caso de mais de um) adotado respondeu à questão elaborada na primeira fase. Caso a avaliação não seja satisfatória, nessa etapa é possível voltar para fase inicial para refazer o processo, ou parte dele. Isso explica a iteração presente na metodologia CRISP-DM;
- **Distribuição:** entrega dos resultados da análise dos dados, feita através de relatórios, imagens, gráficos e ilustrações. Essa etapa é concebida através deste artigo.

Nas próximas seções são apresentadas detalhadamente todas as etapas realizadas da metodologia CRISP-DM, juntamente com os resultados alcançados.

4 Entendimento do Negócio e dos Dados

Essa seção apresenta, primeiramente, o entendimento da educação do Maranhão, e posteriormente o entendimento dos dados e as ferramentas computacionais utilizadas, equivalendo às duas primeiras etapas da metodologia CRISP-DM usada neste trabalho. A etapa de entendimento do negócio consiste na proposição de uma situação-problema relacionada à mineração de dados e um plano preliminar para solucionar esse problema. Nesta fase, são definidos os objetivos do projeto e os recursos a serem utilizados na mineração. Já a etapa de entendimento dos dados consiste em buscar a base de dados e fazer a análise preliminar deles para verificar se é possível resolver o problema abordado, ou pelo menos parte dele. Nessa etapa também já é feita uma análise preliminar dos dados.

4.1 Entendendo a Educação Básica Maranhense

O estado do Maranhão está localizado na região Nordeste do Brasil. De acordo com o Instituto Brasileiro de Geografia e Estatística (IBGE), a população estimada é de 7 milhões de habitantes, distribuídos em 216 cidades, com uma densidade demográfica de 19 habitantes por quilômetro

quadrado. A renda domiciliar *per capita* é de R$ 636,00 (IBGE, 2021). Ainda segundo dados do IBGE de 2010, o Índice de Desenvolvimento Humano Municipal (IDHM) é 0,639 (IBGE, 2021). Esse número coloca o estado do Maranhão na 26ª posição entre os estados brasileiros e o Distrito Federal. Esse dado também demonstra o estado do Maranhão como um dos mais pobres e menos desenvolvidos dentre os estados brasileiros.

Segundo dados do Censo Escolar de 2019, ano abordado nesse estudo, o estado do Maranhão possui 1.031 escolas de ensino médio, sendo: 28 escolas federais, 14 escolas municipais, 193 escolas privadas e 796 escolas estaduais. No que diz respeito ao IDEB, os números do Maranhão, referente às escolas públicas de ensino médio, não tiveram uma grande variação nas últimas avaliações (SEDUC-MA, 2020). As médias das notas das escolas nas últimas avaliações foram: 2,8 em 2013, 3,1 em 2015, 3,4 em 2017, e 3,7 em 2019. Considerando o IDEB em uma escala de 0 (zero) a 10 (dez), as médias referentes às escolas públicas de ensino médio do Maranhão estão muito aquém de uma nota considerada elevada para os padrões nacionais. Isso evidencia a necessidade de verificação dos fatores que influenciam essas notas do IDEB.

No estado do Maranhão, além das tradicionais escolas públicas de ensino médio, o estado recentemente criou escolas de ensino técnico e em tempo integral. As escolas técnicas criadas pelo Governo do Maranhão são chamadas de Institutos Estaduais de Educação, Ciência e Tecnologia do Maranhão (IEMA) e têm o objetivo de ofertar cursos técnicos integrados ao ensino médio. Já as escolas de tempo integral são chamadas Centros Educa Mais. Elas oferecem educação em tempo integral, em que o aluno passa o dia (de 7h às 17h) na escola, fazendo refeição, e tendo horas de estudo e lazer na própria escola. Os IEMAs e os Centros Educa Mais já estão presentes em 33 cidades.

Em 2022, o piso salarial nacional dos professores da educação básica é de R$ 3.845,63, referente a jornada semanal de 40 horas. No Maranhão, o salário do professor de ensino médio em início de carreira é de R$ 6.867,68 (Maranhão, 2022), pela jornada semanal de 40 horas. Esse valor do salário de professores no Maranhão é composto de salário-base e mais Gratificação de Atividade do Magistério (GAM), pago a professores efetivos concursados. Há também os professores concursados com jornada semanal de 20 horas, cujo valor do salário é proporcional ao valor de 40 horas. O Maranhão também conta com professores contratados temporariamente para suprir as necessidades de carência de professores. Atualmente (2022), o professor contratado no Maranhão recebe salário bruto mensal de R$ 1.876,06 pela jornada semanal de 20 horas.

A seguir, como parte da etapa de entendimento do negócio, definimos o objetivo do processo de mineração de dados, a questão de pesquisa e os critérios de sucesso.

O objetivo deste estudo, ou seja, do processo de mineração de dados realizado, é entender os fatores que influenciam na qualidade educacional do estado do Maranhão. Para isso, o estudo tira vantagem do uso de técnicas de mineração das bases de dados fornecidas pelo INEP: as bases do SAEB e do IDEB.

A pesquisa busca responder à seguinte questão de pesquisa primária: O que influencia o desempenho acadêmico das escolas estaduais de ensino médio do Maranhão?

Como critérios de sucesso (CSs) dessa pesquisa, podemos listar:

- (CS1) A coleta, análise preliminar e entendimento dos dados apresentados nas bases do SAEB e IDEB;

- (CS2) A identificação do que influencia o desempenho acadêmico das escolas estaduais do Maranhão a partir da mineração dos dados;
- (CS3) A criação de modelos (preditivo e de análise exploratória) relacionados à educação básica do Maranhão;
- (CS4) Criação de relatório com os resultados para eventual publicação em forma de artigo científico, para a realização da etapa de distribuição da metodologia CRISP-DM.

4.2 Entendimento dos Dados do IDEB e SAEB do Maranhão

Existem diversas bases de dados educacionais brasileiras disponibilizadas pelo INEP e, dentre elas, estão o Censo Escolar, o SAEB, o ENEM, o Exame Nacional de Certificação de Competências da Educação de Jovens e Adultos (ENCCEJA), o Exame Nacional de Desempenho dos Estudantes (ENADE), o Censo da Educação Superior e Indicadores Educacionais. As bases SAEB e IDEB[1] são utilizadas nesta pesquisa e, por esse motivo, são detalhadas a seguir. Adicionalmente, os valores de IDHM das cidades do Maranhão também são utilizados nesse estudo.

As avaliações aplicadas pelo SAEB acontecem a cada dois anos. As provas de duas disciplinas são aplicadas, Língua Portuguesa e Matemática, para estudantes do 5º ano do Ensino Fundamental I, 9º ano do Ensino Fundamental II, e 3º ano do Ensino Médio.

O IDEB é um indicador que mede a qualidade da educação básica brasileira, calculado a partir da média da proficiência das provas de Língua Portuguesa e Matemática aplicadas pelo SAEB, padronizada entre 0 (zero) e 10 (dez), multiplicada pelo indicador de rendimento baseado na taxa percentual de aprovação dos alunos, a qual é padronizada entre 0 (zero) e 1 (um). Além das avaliações, os questionários do SAEB também são aplicados aos alunos, professores e gestores da educação. Eles visam obter dados sobre a infraestrutura das escolas, aspectos administrativos e socioeconômicos.

Todos os dados produzidos por meio de avaliações e questionários são disponibilizados pelo INEP através de uma página para acesso a dados abertos (INEP, 2022a). O volume de dados disponível é grande e tem sido analisado para gerar conhecimento para gestores e demais envolvidos na educação. Isso permite a realização de tomadas de decisão baseadas em evidências. Considerando essa disponibilização de bases, pesquisas em mineração de dados educacionais podem ser conduzidas . Para esta pesquisa, foram analisados os dados de escolas públicas estaduais de ensino médio do Maranhão, obtidos através das bases de dados do SAEB e IDEB referentes ao ano de 2019, por serem os mais recentes disponíveis no momento de realização do estudo.

Quando se realiza uma análise de dados, as informações podem vir de uma variedade de fontes e em vários formatos. Nesta etapa, os dados foram compreendidos, e a qualidade e adequação desses dados foram verificadas. Nessa etapa, também foram feitas as primeiras constatações e questionamentos baseados nos dados.

O primeiro conjunto de dados utilizado foi o que continha as respostas do questionário socioeconômico aplicado aos alunos do SAEB. Ele é um conjunto com dados apenas de alunos do 3º e 4º ano do ensino médio de todo o Brasil. São mais de 2 milhões de linhas, que correspondem aos dados de cada aluno que respondeu ao questionário, e mais de 90 variáveis, entre elas:

[1] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados

- Dados de identificação: código da escola, região, estado, cidade;
- Dados das respostas da Prova Brasil: respostas do questionário socioeconômico com informações como raça, meio de transporte utilizado para ir à escola, escolaridade dos pais, questões sobre incentivo dos pais para estudar, acesso à tecnologia.

O segundo conjunto de dados analisado possui o resultado do IDEB do ano de 2019. Ele contém mais de 20 mil registros, que correspondem a cada uma das escolas participantes do SAEB, e 28 variáveis, tais como os dados de identificação da escola, como o código INEP, estado, região, cidade. O INEP também disponibiliza um dicionário em formato XLS, com descrição detalhada do conjunto de dados.

As duas bases de dados utilizadas possuem qualidade e podem ser consideradas idôneas. A justificativa para isso é que os dados são obtidos a partir de fontes oficiais. No entanto, devido a forma como os dados são coletados, via preenchimento de questionários por estudantes, as duas bases contêm dados faltantes e nulos. Dados nulos são decorrentes de preenchimento com caracteres fora da resposta padrão esperada. Na base de dados do SAEB, os dados faltantes são dos alunos que não responderam ao questionário socioeconômico. Essa identificação é possível, pois há uma variável com resposta binária em que informa se o aluno respondeu ou não ao questionário. Quanto aos dados do IDEB, os dados ausentes se referem às escolas que, por algum motivo, não realizaram a prova para obtenção da nota ou não tiveram quantidade de alunos suficiente para que a nota fosse divulgada.

A Tabela 2 apresenta a lista de variáveis, com respectivas questões do SAEB (16 variáveis) e suas possíveis respostas, IDHM do município da escola (1 variável) e nota do IDEB (1 variável na última linha). Algumas variáveis da base de dados do SAEB foram desconsideras, por serem referentes a dados de localização da escola, identificação da escola e do estudante, e também dados de controle, os quais não têm relação com o objetivo deste estudo.

Na etapa de entendimento dos dados, é necessário realizar o levantamento das ferramentas a serem utilizadas. A realização desta pesquisa exigiu a utilização de diversas soluções tecnológicas. O manuseio das bases de dados para o processo de mineração foi realizado através das bibliotecas, na linguagem *Python*: *Pandas*, *NumPy*, *Matplotlib* e *Scikit-Learn*. Todas elas foram utilizadas no ambiente de desenvolvimento *Google Colaboratory*. Adicionalmente, para realizar tarefas específicas, também foram utilizados o *Orange Data Mining* e o *IBM SPSS*.

5 Preparação e Modelagem

Esta seção apresenta a preparação dos dados do SAEB e do IDEB, as análises e modelagens realizadas para a geração dos resultados para a questão de pesquisa. Primeiramente, é apresentado como os dados foram preparados, em seguida os modelos e resultados obtidos após os dados estarem prontos para uso: análise de correlação, análise de fatores, modelos de regressão e árvore de decisão.

Tabela 2: Descrição das variáveis utilizadas.

Variável	***Descrição***	***Resposta***
TEM COMPUTADOR	O aluno tem computador em casa?	A. Nenhum; B. um; C. dois; D. três ou mais.
TEM WIFI	O aluno tem acesso à rede de internet sem fio em casa?	a. Sim; B. Não.
ESCOL MAE	Nível de escolaridade da mãe	A. Não completou o 5º ano do Ensino Fundamental; B. Ensino Fundamental, até o 5º ano; C. Ensino Fundamental completo; D. Ensino Médio completo. E. Ensino Superior completo (faculdade ou graduação); F. Não sei.
ESCOL PAI	Nível de escolaridade do pai	A. Não completou o 5º ano do Ensino Fundamental; B. Ensino Fundamental, até o 5º ano; C. Ensino Fundamental completo; D. Ensino Médio completo; E. Ensino Superior completo (faculdade ou graduação); F. Não sei.
PAIS CONVERSAM ESCOLA	Os pais do aluno conversam com ele sobre a escola?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre.
PAIS INCENTIVAM	Os pais incentivam o aluno a estudar?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre.
PAIS TAREFA CASA	Os pais incentivam o aluno a fazer as tarefas de casa?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre.
INCENTIVAR IR ESCOLA	Os pais incentivam o aluno a ir para a escola?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre.
COMPARECER REUNIOES	Os pais comparecem às reuniões com os professores e gestão da escola?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre.
LER NOTICIAS	Os alunos lêm noticias?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre.
LER LIVROS	Os alunos lêm livros	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre.
LER QUADRINHOS	Os alunos lêm revistas em quadrinhos?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre.
DEVER CASA	Os alunos fazem o dever de casa?	A. Nunca ou quase nunca; B. De vez em quando; C. Sempre ou quase sempre.
IDADE INIC ESTUDAR	Com que idade os alunos começaram a estudar?	A. 3 anos ou menos; B. 4 ou 5 anos; C. 6 ou 7 anos; D. 8 anos ou mais.
REPROVOU	O aluno já reprovou algum ano durante sua trajetória escolar?	A. Nunca; B. Sim, uma vez; C. Sim, duas vezes ou mais.
EVADIU	O aluno já evadiu da escola durante algum ano da sua trajetória escolar?	A. Nunca; B. Sim, uma vez; C. Sim, duas vezes ou mais.
IDHM	IDHM do município em que a escola é situada.	Valores entre 0 e 1
IDEB 2019	Nota do IDEB referente ao ano de 2019	Notas de 0 a 10

5.1 Pré-processamento

Todo o processo realizado na etapa de preparação dos dados é apresentado na Figura 2, descrito a seguir.

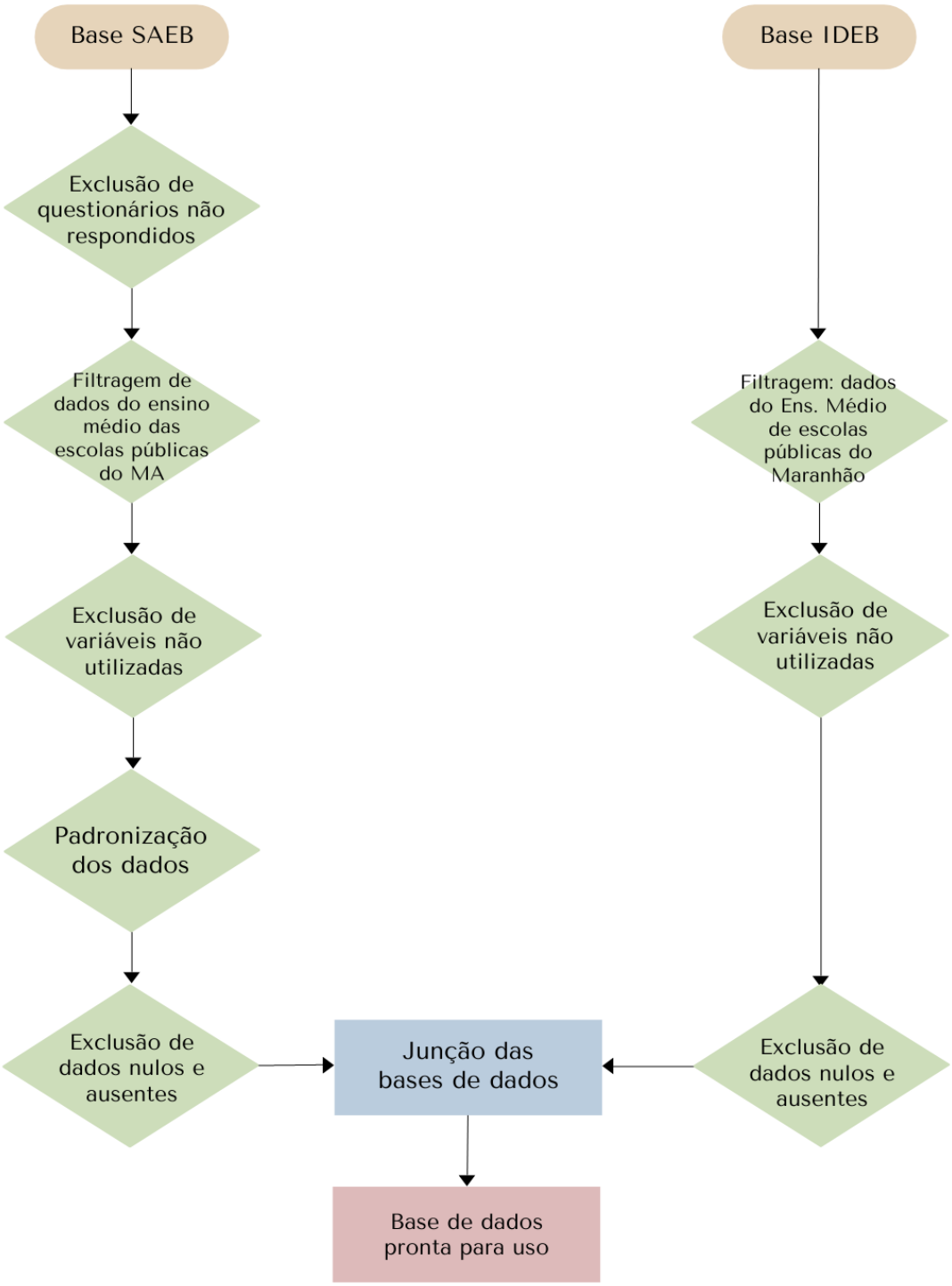

Figura 2: Organização da etapa de preparação dos dados.

As bases do INEP fornecem dados educacionais de todos os estados do Brasil, sendo possível filtrar e separá-los por regiões (estados ou municípios), por dependência administrativa (municipal, estadual, federal e privada), por nível de ensino (fundamental e médio), conforme os dados disponíveis. Nesta pesquisa, foram utilizados dados de escolas públicas de ensino médio da rede estadual do Maranhão.

A etapa de preparação pode envolver, caso necessário, a padronização e normalização dos dados, para produzir o conjunto final de dados, pronto para ser analisado. Os principais dados da base do SAEB são as respostas do questionário socioeconômico e da prova. Como o que interessa nessa base de dados são as respostas do questionário socioeconômico, a primeira filtragem feita foi para deixar apenas as linhas com os alunos que haviam respondido o questionário socioeconômico. Outra filtragem foi deixando apenas os dados do estado do Maranhão.

Com o conjunto de dados apresentando dados de todas as escolas do Maranhão, linhas foram excluídas para manter apenas as escolas públicas estaduais. Em seguida, o conjunto de dados foi alterado para exibir apenas as variáveis a serem trabalhadas. Embora o arquivo possua mais de 90 variáveis, para fins da pesquisa, neste primeiro conjunto de dados, 16 variáveis foram utilizadas, como visto na Tabela 2. Como a base contém dados categóricos e numéricos, os dados tiveram que ser padronizados e normalizados. Os dados categóricos ordinais foram convertidos para numéricos, por exemplo: para respostas “Sim” ou “Não”, foram atribuídos os valores 1 (um) e 0 (zero), respectivamente; para valores como “Nunca” ou “Quase nunca”, “De vez em quando”, “Sempre” ou “Quase sempre”, foram atribuídos os valores 0 (zero), 1 (um) e 2 (dois), respectivamente.

O conjunto de dados com os resultados do IDEB também precisou ser tratado, de forma que apenas se manteve os resultados (notas) e a variável com o código das escolas, para realizar a junção dos conjuntos de dados. Todas as linhas que não possuíam dados do IDEB foram excluídas. A Tabela 3 sumariza (funções *describe* e *info* do *Pandas*) o conjunto de dados utilizado no estudo após a realização da etapa de preparação dos dados. O conjunto de dados possui um total de 27.595 amostras não nulas. A Tabela 3 apresenta o valor médio, o desvio padrão, o valor mínimo, os percentis de 25%, 50% e 75%, o valor máximo, e o tipo de dado. Todos os dados são numéricos: a maioria é do tipo inteiro, e o último, que corresponde ao IDEB 2019, é do tipo *float*.

5.2 Análise Descritiva

Com os dados prontos, foi possível realizar uma análise descritiva para efeito de estudo e observação dos dados e variáveis disponíveis. A Figura 3 mostra o grau de escolaridade das mães e dos pais. Ambos possuem como escolaridade predominante o ensino médio completo. Adicionalmente, é possível observar que a quantidade de mães nos dados é maior que a de pais. Isso levanta uma hipótese de que há muitos alunos que convivem exclusivamente com a mãe, sem a presença do pai. No entanto, essa hipótese até o momento não pôde ser comprovada através desta análise.

A Figura 4 ilustra a quantidade de computadores em que os alunos possuem (variáveis com respostas definidas de acordo como descrito na Tabela 2; ou seja: nenhum computador = 0; um computador = 1; dois computadores = 2; e três ou mais computadores = $>=3$) e o acesso à Internet através de rede sem fio (Sim = Tem acesse a internet Wi-Fi; Não = Não tem acesso à internet Wi-Fi). A maioria dos alunos não possui computador. No entanto, mais da metade dos alunos possuem acesso à Internet via rede sem fio (Wi-Fi ou conexão via telefonia móvel).

Tabela 3: Visão geral (funções *describe* e *info* do *Pandas*) do conjunto de dados resultante utilizado no estudo. Nota: DP - Desvio Padrão; MÍN - Mínimo; MÁX - Máximo.

Variável	***Média***	***DP***	***MÍN***	***25%***	***50%***	***75%***	***MÁX***	***Tipo de dado***
TEM COMPUTADOR	0,45	0,69	0,0	0,0	0,0	1,0	3,0	int64
TEM WIFI	0,51	0,49	0,0	0,0	1,0	1,0	1,0	int64
ESCOL MAE	3,13	1,33	1,0	2,0	4,0	4,0	5,0	int64
ESCOL PAI	2,74	1,34	1,0	1,0	3,0	4,0	5,0	int64
PAIS CONVERSAM ESCOLA	2,24	0,62	1,0	2,0	2,0	3,0	3,0	int64
PAIS INCENTIVAM	2,82	0,436	1,0	3,0	3,0	3,0	3,0	int64
PAIS TAREFA CASA	2,50	0,66	1,0	2,0	3,0	3,0	3,0	int64
INCENTIVAR IR ESCOLA	2,87	0,41	1,0	3,0	3,0	3,0	3,0	int64
COMPARECER REUNIOES	2,47	0,66	1,0	2,0	3,0	3,0	3,0	int64
LERNOTICIAS	2,20	0,61	1,0	2,0	2,0	3,0	3,0	int64
LERLIVROS	2,07	0,64	1,0	2,0	2,0	2,0	3,0	int64
LERQUADRINHOS	1,76	0,70	1,0	1,0	2,0	2,0	3,0	int64
DEVER CASA	2,04	0,91	0,0	1,0	2,0	3,0	3,0	int64
IDADE INIC ESTUDAR	2,40	0,71	0,0	2,0	3,0	3,0	3,0	int64
REPROVOU	2,61	0,61	1,0	2,0	3,0	3,0	3,0	int64
EVADIU	2,90	0,35	1,0	3,0	3,0	3,0	3,0	int64
IDEB 2019	3,82	0,67	1,7	3,4	3,7	4,2	6,20	float64

O IDHM mede o desenvolvimento das cidades considerando três dimensões: longevidade, educação e renda. A Figura 5 ilustra, através de um *scatter plot*, a distribuição do IDEB em função do IDHM das cidades maranhenses. A figura mostra haver um número maior de cidades no primeiro quadrante (cor púrpura). O segundo quadrante (cor verde) ilustra uma menor quantidade de cidades com IDEB acima da média em função do IDHM baixo. O terceiro quadrante (cor azul), embora tenha um número considerável de cidades com IDEB acima da média em função do alto IDHM, não chega a ser maior que o primeiro quadrante. Por fim, o quadrante de cor laranja mostra poucas cidades com bom desempenho no IDEB e com IDHM abaixo da média.

Foi possível tirar algumas conclusões a partir da análise descritiva preliminar dos dados realizada neste estudo, as quais seguem as principais. Primeiramente, a análise mostrou que a região metropolitana da capital São Luís se destaca, com os melhores indicadores educacionais. As notas do IDEB das escolas da capital são em geral acima da média, mas há também escolas nas cidades do interior com IDEB acima da média. A região metropolitana da capital maranhense é melhor desenvolvida em vários aspectos, com disponibilização de uma infraestrutura que não está presente em outras regiões do estado. Embora não seja a melhor e nem a mais perfeita infraestrutura, isso acaba influenciando nos resultados dos indicadores educacionais. As escolas das regiões do interior do estado possuem, em sua grande maioria, um IDEB com índice abaixo da média.

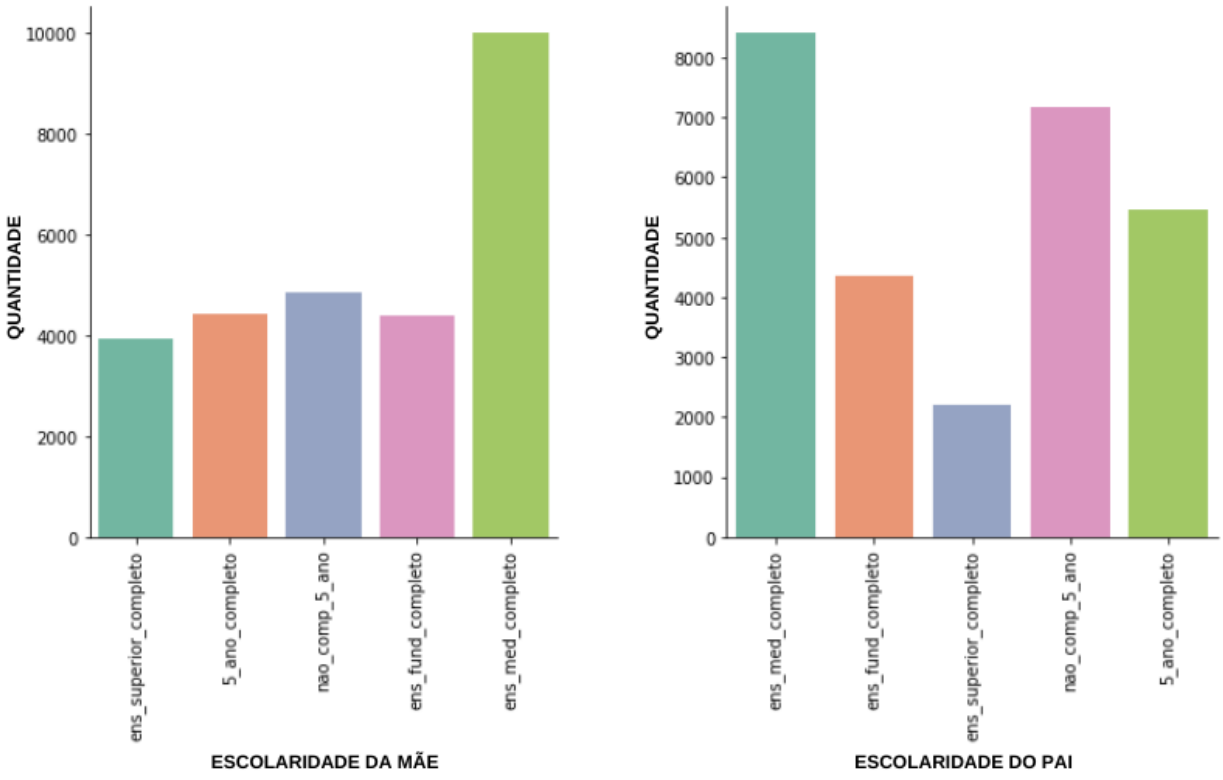

Figura 3: Escolaridade das mães e dos pais.

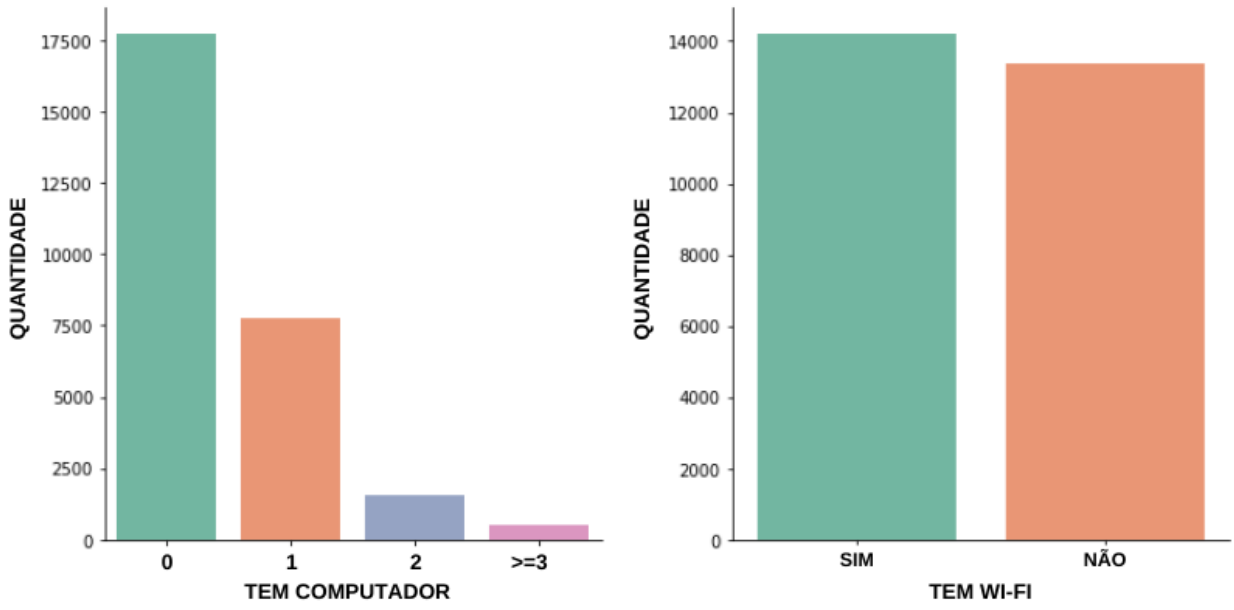

Figura 4: Recursos tecnológicos: quantidade de computadores que cada aluno possui e acesso à Internet via rede sem fio.

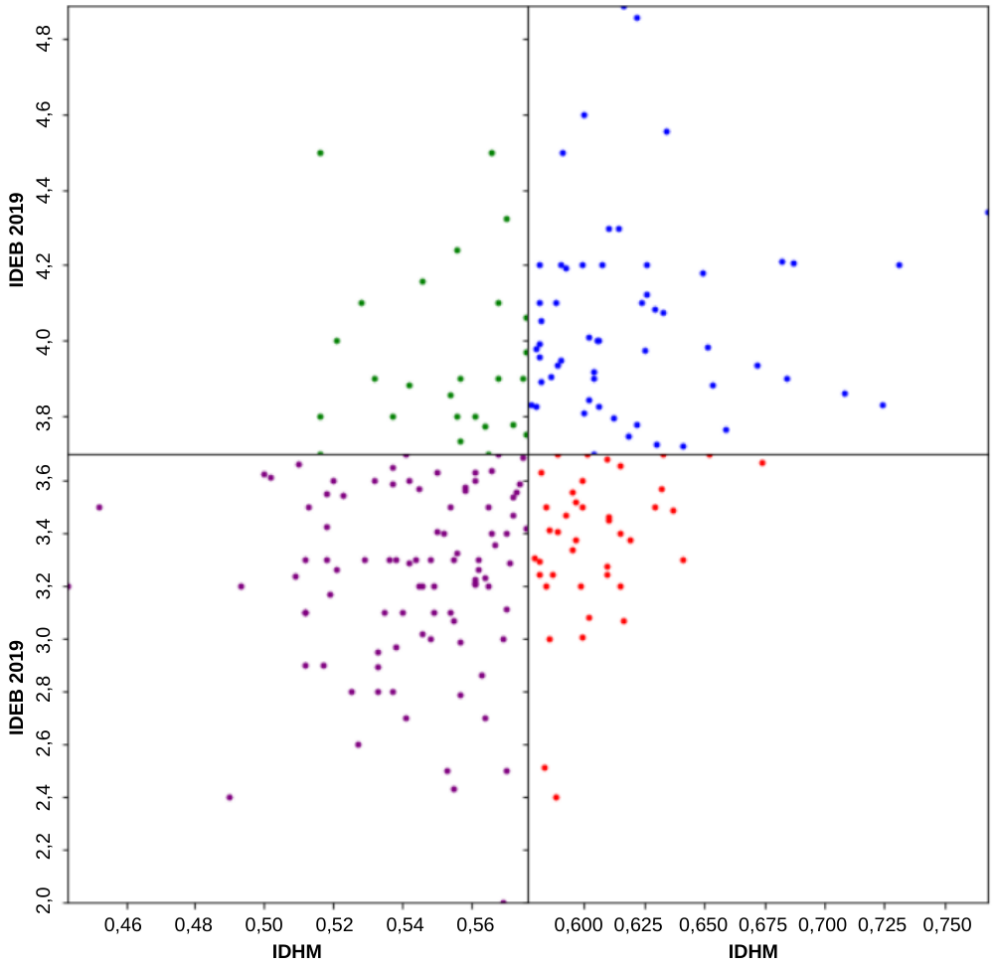

Figura 5: *Scatter plot* das variáveis IDHM e IDEB de 2019, em que cada ponto representa um município do estado.

5.3 Análise de Correlação

Correlação é uma medida estatística que indica o grau de relacionamento (i.e., a força de associação) entre duas ou mais variáveis. A correlação pode ser positiva ou negativa. Se a correlação for positiva, ela se refere ao aumento ou diminuição das variáveis conjuntamente. No entanto, se a correlação for negativa, ela indica que uma variável aumenta à medida que a outra diminui, e vice-versa. A correlação também pode ser chamada de medida de associação, medida de interdependência, medida de intercorrelação ou medida de relação entre as variáveis (Lira, 2004).

Existem três principais medidas de correlação: de Pearson, de Kendall, e de Spearman. A correlação de Pearson é utilizada para verificar a relação linear entre variáveis quantitativas. A correlação de Kendal é um teste não paramétrico para medir a dependência entre duas variáveis. A correlação de Spearman é um teste também não paramétrico para medir o grau de associação entre duas variáveis, sem exigir que a relação entre as variáveis seja linear. A correlação também pode ser classificada como simples, múltipla ou parcial. A simples analisa a dependência de duas variáveis, sejam elas X e Y, sendo uma dependente e a outra independente. A correlação múltipla estuda a relação entre uma variável dependente e outras duas ou mais variáveis independentes. A correlação parcial ocorre quando é realizada a correlação múltipla, após eliminar uma

das variáveis independentes.

Nesta pesquisa, foi utilizada a análise de correlação simples de Spearman, uma vez que não foi possível garantir a existência de relacionamento linear entre as variáveis analisadas e devido algumas delas serem categóricas ordinais. Através da correlação de Spearman, obtém-se o coeficiente de correlação. O coeficiente fica em um intervalo de -1 a 1 e indica o grau de relação entre a variável que se obtém a correlação e a variável alvo (Akoglu, 2018). Se o valor da correlação for 0 (zero), ele indica que não existe associação entre as variáveis, mas isso não significa que não existe uma relação entre elas. É importante destacar que, embora a correlação seja um método estatístico importante para verificar o relacionamento entre duas ou mais variáveis, ela não indica causalidade. Ou seja, se uma determinada variável tem uma correlação forte, conforme especificado na Tabela 4 (Akoglu, 2018), não significa que tal variável seja a causa da variável com a qual se relaciona. As variáveis podem ser influenciadas por um fator desconhecido.

Tabela 4: Interpretações dos resultados de uma análise de correlação.

Correlação (+ ou -)	***Interpretação***
0 a 0,19	Correlação bem fraca
0,20 a 0,39	Correlação fraca
0,40 a 0,69	Correlação moderada
0,70 a 0,89	Correlação forte
0,90 a 1	Correlação muito forte

A correlação foi realizada a partir de cada uma das variáveis com a variável dependente (i.e., a nota do IDEB de 2019). Os dados foram divididos por microrregião e, em cada uma delas, buscou-se saber qual variável mais se correlacionava e a que menos se correlacionava com o IDEB, a fim de comparar os resultados. As microrregiões são agrupamentos de municípios limítrofes, para integrar a organização, o planejamento e a execução de funções públicas de interesse comum (Brasil, 1988). Essa divisão foi necessária por percebermos que os dados do estado do Maranhão eram muito heterogêneos. Com um grande território, o estado possui peculiaridades diversas entre suas regiões, tais como uma elevada desigualdade social, incluindo o IDHM visto na Figura 5, e também diversidade cultural entre microrregiões.

A Tabela 5 apresenta as variáveis de menor (valores mais próximos a 0) e maior (valores mais próximos a -1 ou 1) correlação para cada uma das 21 microrregiões do estado, em que é possível observar, entre as maiores (coluna da direita), majoritariamente correlações moderadas e fortes. Ao analisar a correlação das variáveis com as notas do IDEB, há uma grande variação entre microrregiões. Não há como concluir que uma variável específica tenha menor ou maior correlação para todo o estado. Quando comparada uma microrregião com a outra, existem diversidades tanto nas notas do IDEB como nos dados das variáveis, acarretando variações nas correlações. Isso evidencia as disparidades regionais do estado, com cada microrregião tendo suas particularidades. Entretanto, é possível observar uma recorrência das variáveis relacionadas a escolaridade dos pais (Mãe e Pai), com 8 ocorrências, entre as maiores correlações.

Tabela 5: Variáveis de maior e menor correlação de Spearman para cada microrregião do Maranhão.

Microrregião	***Menor Correlação (Valor)***	***Maior Correlação (Valor)***
Litoral Ocidental Maranhense	Idade que começou a estudar (0,022101)	Tem WIFI (0,463914)
Aglomeração Urbana de São Luís	Ler Livros (0,046474)	Escolaridade da Mãe (0,629817)
Rosário	Tem computador (-0,022512)	Ler Notícias (0,721049)
Lençóis Maranhenses	Tem WI-FI (0,001040)	Faz Dever de Casa (0,498704)
Gurupi	Pais comparecem a reuniões (0,014773)	Faz dever de casa (0,519469)
Codó	Pais conversam sobre a escola (0,004531)	Ler notícias (0,644174)
Coelho Neto	Pais conversam sobre escola (-0,071429)	Tem WI-FI (0,892857)
Caxias	Pais incentivam estudar (-0,018281)	Escolaridade da mãe (0,748134)
Porto Franco	Pais conversam sobre a escola (-0,035714)	Pais comparecem a reuniões (0,792825)
Baixada Maranhense	Pais incentivam tarefa de casa (-0,043136)	Escolaridade do Pai (0,560112)
Itapecuru-mirim	Pais conversam sobre escola (0,036738)	Ler Notícias (0,737693)
Pindaré	Ler Livros (0,030832)	Pais comparecem a reuniões (0,466886)
Imperatriz	Idade que iniciou a estudar (-0,009840)	Escolaridade da Mãe (0,530312)
Médio Mearim	Pais conversam sobre a escola (-0,010420)	Tem computador (0,530472)
Alto Mearim e Grajaú	Pais incentivam ir a escola (-0,031749)	Pais comparecem a reuniões (0,307918)
Presidente Dutra	Pais incentivam ir a escola (-0,011122)	Escolaridade da Mãe (0,550663)
Baixo Parnaíba Maranhense	Ler Quadrinhos (0,027160)	Escolaridade da Mãe (0,524466)
Chapadinha	Pais comparecem a reuniões (-0,017837)	Escolaridade do Pai (0,631164)
Chapadas do Alto Itapecuru	Pais conversam sobre escola (-0,142857)	Ler Notícias (0,774806)
Gerais de Balsas	Ler Livros (0,099376)	Escolaridade da Mãe (0,695183)
Chapada das Mangabeiras	Pais conversam sobre escola (-0,024714)	Pais incentivam tarefa de casa (0,678571)

5.4 Regressão Linear da Relação do IDEB com o IDHM

A regressão linear é uma tentativa de modelar uma equação que descreva o relacionamento entre duas variáveis (Curral, 1994). Um dos objetivos da regressão é realizar identificação e avaliação da relação entre uma variável dependente e uma ou mais variáveis independentes, também chamadas de preditoras ou explicativas. A regressão também pode ser aplicada para predizer valores futuros de uma variável (Rodrigues, De Medeiros, & Gomes, 2013). A regressão linear pode ser de dois tipos: simples ou múltipla. Simples é quando há uma única variável dependente e uma única

variável independente. A regressão múltipla ocorre quando há mais de uma variável independente, e uma única variável dependente. Nesta pesquisa, foi usada a regressão linear simples.

A Figura 6, a partir do gráfico de dispersão *scatter plot*, ilustra a regressão linear simples realizada com os dados do IDHM (variável independente X) e os dados do IDEB 2019 (variável dependente Y). A reta traçada em vermelho representa a função linear de regressão com os possíveis valores preditos de Y em função do valor já conhecido de X.

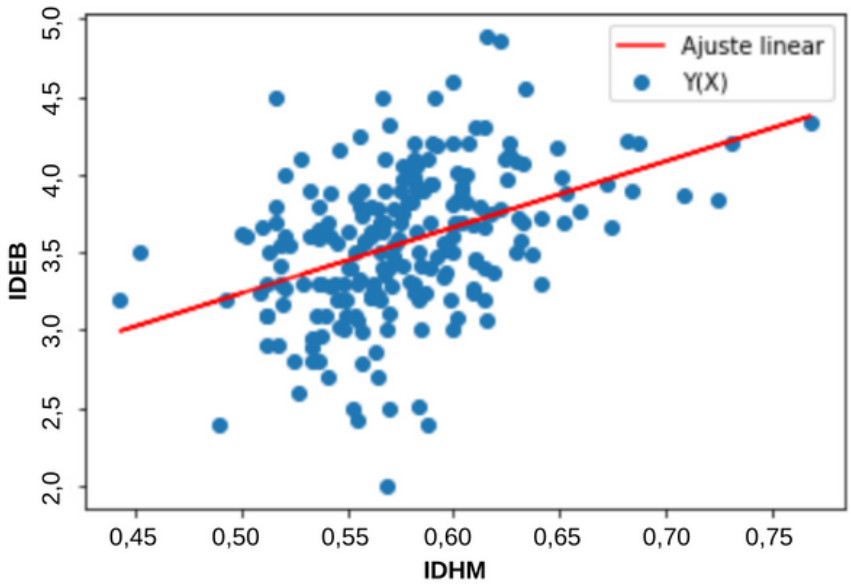

Figura 6: Regressão linear considerando as variáveis IDHM e o IDEB de 2019.

Para medir a qualidade da regressão, foi utilizada a Raiz Quadrada do Erro Médio (RMSE) que, ao ser calculada, utiliza a mesma escala da variável dependente (de 0 a 10), como visto na Equação 1. Nessa regressão, o RMSE foi de 0,43. Portanto, é possível predizer, com um erro aceitável, o valor do IDEB, em função do IDHM de uma determinada cidade do estado do Maranhão.

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

5.5 Análise de Fatores

A Análise de Fatores é uma técnica estatística utilizada para reduzir a dimensão de um conjunto de dados com fatores em comum (Gorsuch, 2014). As variáveis com características em comum podem ser representadas por um único fator. A análise fatorial é uma técnica utilizada para, quando houver muitas variáveis observadas, gerar fatores subjacentes não observados. Através da técnica de análise de fatores, as variáveis podem ser condensadas em fatores comuns, tornando o conjunto de dados com menos variáveis, possibilitando sua análise e entendimento de forma mais fácil (DiStefano, Zhu, & Mindrila, 2009).

A análise fatorial pode ser de dois tipos: exploratória e confirmatória (Figueiredo Filho

& Silva Júnior, 2010). A Análise Fatorial Exploratória (AFE) é geralmente utilizada nos estágios iniciais da pesquisa para explorar os dados. Nessa fase, procura-se explorar a relação entre um conjunto de variáveis, identificando padrões de correlação. A Análise Fatorial Confirmatória (AFC) é utilizada para testar hipóteses. O pesquisador testa em que medida determinadas variáveis são representativas de uma dimensão. Nesta pesquisa, foi utilizada a análise fatorial confirmatória.

Para ocorrer a análise fatorial, o processo passa por três fases (Figueiredo Filho & Silva Júnior, 2010), como segue:

- Passo 1: Verificação de adequação da base de dados, tais como o tamanho da amostra, variáveis contínuas ou discretas, exclusão de variáveis (e.g., sexo, cor), limite *Kaiser-Meyer-Olkin* (KMO) mínimo de 0,6, e realização do Teste de Esfericidade de Bartlett (do inglês, *Bartelett Test of Spherecity* - BTS) ($p<0,05$).

O teste KMO verifica se é apropriado usar as variáveis de manifesto para a análise fatorial. O teste realiza o cálculo da proporção de variância entre as variáveis de manifesto. Os valores de KMO variam entre 0-1, uma proporção abaixo de 0,6 sugere que o conjunto de dados é inapropriado para a análise fatorial. Já o Teste de Esfericidade de Bartlett é uma verificação de intercorrelação entre variáveis manifestas, ou seja, a comparação da matriz de correlação observada e a matriz de identidade. Se a análise fatorial for um método apropriado a ser usado, a matriz de correlação e a matriz de identidade não serão as mesmas e o teste será significativo ($p<0,05$).

- Passo 2: Determinar o número de fatores a serem extraídos através do *Scree test* (analisar graficamente a dispersão dos fatores) e do *Eigenvalue* igual ou acima de 1 (variância em todas as variáveis devida ao fator, a qual é uma variância explicada por cada fator da variância total).

Embora não exista uma regra absoluta de quantos fatores devem se extrair, a regra do *Eigenvalue* acima de 1 deve ser seguida, pois, se for abaixo desse valor, o fator pouco contribui para a explicação das variáveis.

- Passo 3. Decidir o tipo de rotação dos fatores: rotação Ortogonal do tipo *Varimax* ou rotação oblíqua.

A rotação é um método matemático que rotaciona os eixos no espaço geométrico com o propósito de facilitar a determinação de quais variáveis são carregadas em quais componentes. A rotação contribui para que o resultado empírico encontrado seja mais facilmente interpretável, conservando as suas propriedades estatísticas.

A modelagem de análise de fatores foi realizada, pois, dentre as variáveis disponíveis, este estudo quer saber quais delas podem ser condensadas em fatores comuns que influenciam a nota do IDEB. Para fazer a análise de fatores, foi utilizado o software *IBM SPSS*. Após toda a preparação dos dados, como descrito na Seção 5.1, o arquivo do tipo CSV foi importado para o software e executado os comandos de análise de fator. O software realiza todo o processo e também acrescenta ao conjunto de dados uma tabela com as pontuações dos fatores.

Através do gráfico de escarpa, também chamado *Scree Plot*, ou popularmente conhecido como gráfico de análise de cotovelo, pudemos avaliar a quantidade de fatores a serem resumidos a partir dos autovalores iguais ou superiores a 1. Essa mesma análise também pode ser feita através da matriz de variância total explicada, a qual mostra a porcentagem dos dados que podem ser explicados ou resumidos pelas principais variáveis mais relevantes.

Para efeito de elaboração dos fatores, foram executadas e observadas as três fases do processo da análise de fatores, conforme descrito a seguir:

- A base de dados estava adequada, pois ela foi importada totalmente pronta para o software com os dados devidamente tratados. Ao realizar o teste de KMO, o valor obtido foi 0,734. Portanto, sendo um valor acima de 0,6, o que possibilitou fazer a análise de fatores;
- Ao realizar a análise do *Scree test*, foi evidenciado o total de 4 (quatro) fatores (i.e., *eigenvalues* superiores a um), conforme mostra o *scree plot* na Figura 7;
- A rotação escolhida foi a Ortogonal *Varimax*, por ser comumente utilizada (Pallant, 2020).

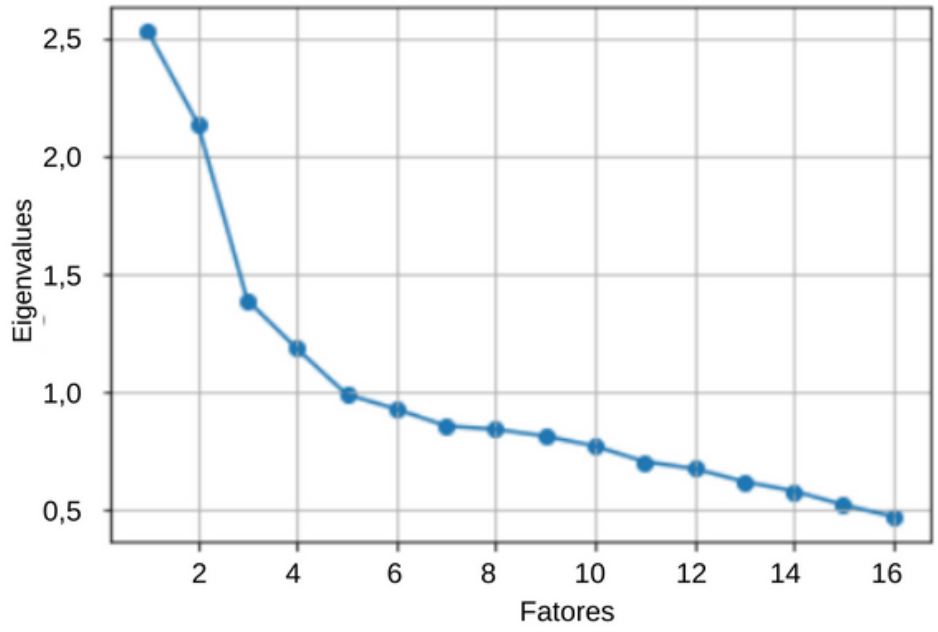

Figura 7: *Scree plot* da análise de fatores.

A Tabela 6 apresenta os 4 (quatro) fatores obtidos a partir das variáveis disponíveis, por ordem de significância da representatividade dos fatores.

5.6 Árvore de Decisão

A Árvore de Decisão é um dos algoritmos supervisionados de aprendizado de máquina que considera a divisão dos dados para realizar uma classificação. O objetivo é encontrar atributos que geram a melhor divisão dos dados em subconjuntos com maior pureza, pertencente à classe alvo

Tabela 6: Fatores identificados e variáveis correspondentes.

Fatores	Variáveis
Escolaridade dos Pais e Tecnologia	TEM COMPUTADOR
	TEM WI-FI
	ESCOL MAE
	ESCOL PAI
Incentivo dos Pais	PAIS CONVERSAM ESCOLA
	PAIS INCENTIVAM ESTUDAR
	PAIS TAREFA CASA
	INCENTIVAR IR ESCOLA
	PAIS COMPARECEM REUNIOES
Cultura	LER NOTICIAS
	LER LIVROS
	LER QUADRINHOS
Vida escolar	DEVER CASA
	IDADE INIC ESTUDAR
	REPROVOU
	EVADIU

(Garcia, 2003). Através da árvore de decisão, ilustra-se um mapeamento dos possíveis resultados de uma série de escolhas, com a possibilidade de fazer comparações entre as opções propostas da árvore.

A árvore de decisão prioritariamente se origina a partir de um nó, que se subdivide em galhos com os possíveis resultados formando outros nós, e assim sucessivamente. Ela pode ser formada por três categorias de nós:

- Nó de decisão: mostra a decisão a ser tomada;
- Nó de probabilidade: mostra resultados incertos;
- Nó de desfecho: indica o resultado final.

O algoritmo de árvore de decisão (Orange Data Mining, n.d.) foi utilizado após ser feita a análise de fatores. O objetivo em construir e apresentar visualmente a árvore de decisão foi para exibir o relacionamento dos fatores identificados na análise de fatores com a variável alvo. As variáveis foram condensadas em fatores, gerando um conjunto de dados com 4 (quatro) novas variáveis correspondentes à pontuação de cada fator, mais a variável alvo (i.e., notas do IDEB de 2019). Para criar o modelo de árvore de decisão, foi utilizado o software *Orange Data Mining*. Após a importação do conjunto de dados para o software, todas as variáveis foram categorizadas. Cada uma das variáveis correspondentes aos fatores foram categorizadas em quatro classes conforme sua influência baseada na pontuação de sua respectiva carga de fator: muito baixo, baixo, alto e muito alto. A variável IDEB 2019 foi transformada em duas classes: abaixo da média e acima da média.

A Figura 8 apresenta o modelo de árvore de decisão binária resultante. Para uma melhor visualização, a árvore foi podada do nível 3 para o 4, deixando-a com 4 níveis (do 0 ao 3), uma

vez que o próximo nível (nível 4) possuiria uma grande quantidade de nós. Uma árvore de decisão pode ser interpretada da seguinte maneira: os nós da árvore representam atributos (i.e., os fatores obtidos a partir da análise de fatores); os ramos a partir de cada nó representam os valores possíveis de cada atributo, os quais refletem as cargas dos fatores categorizados; cada nova instância é analisada a partir do nó raiz da árvore, comparando o valor do atributo da instância contra o valor do atributo representado no nó, seguindo o ramo correspondente ao resultado da comparação; essas comparações continuam até se atingir um nó folha, os quais representam a classe a que pertence uma instância. Como pode ser visto na Figura 8, o modelo teve como nó raiz a escolaridade dos pais e tecnologia. O fator "Escolaridade dos Pais e Tecnologia", quando apresenta valores das classes alto e muito alto, tende à nota do IDEB do aluno ser acima da média; enquanto se esse fator for baixo ou muito baixo, a nota do IDEB será abaixo da média. Isso evidencia a escolaridade dos pais e o uso de tecnologias como um dos fatores determinantes e influentes na qualidade da educação do Maranhão.

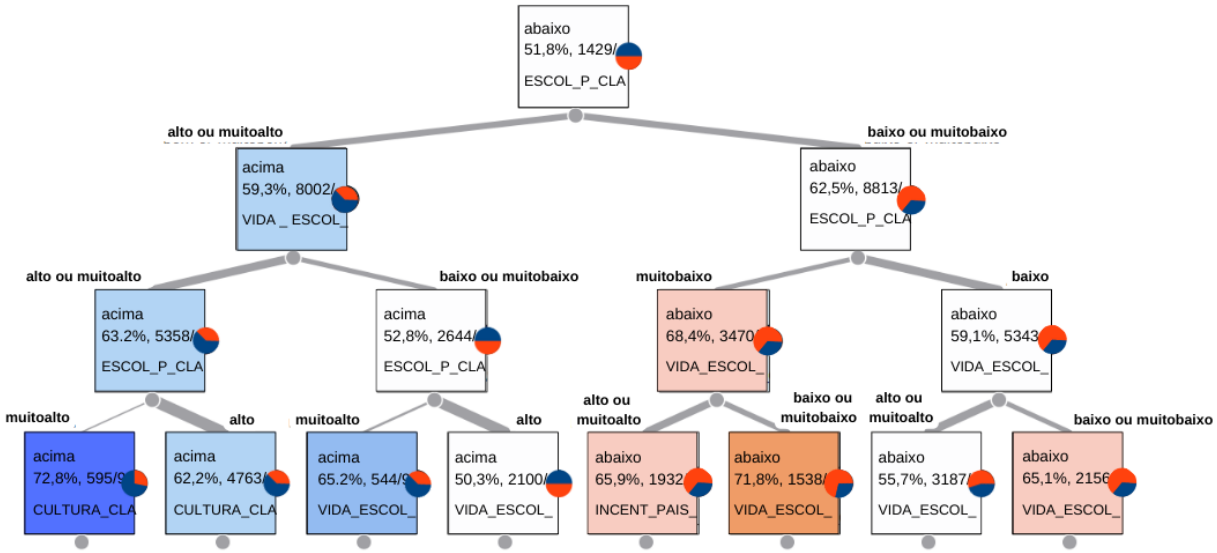

Figura 8: Modelo de árvore de decisão gerado a partir dos fatores.

6 Discussão

Esta seção responde à questão de pesquisa baseado nos resultados encontrados e, em seguida, são discutidas as limitações do estudo.

6.1 Respondendo à Questão de Pesquisa

A questão de pesquisa deste estudo foi a seguinte: O que influencia o desempenho acadêmico das escolas estaduais de ensino médio do Maranhão?

Este estudo realizou a aplicação de diferentes técnicas de MDE, seguindo a metodologia iterativa CRISP-DM, objetivando responder à questão de pesquisa. Após às etapas iniciais da

metodologia envolvendo o entendimento do negócio e dos dados e também a preparação destes últimos, a análise de correlação de Spearman foi utilizada para identificar as variáveis do SAEB com menor e maior correlação com a nota do IDEB, a cada microrregião do estado (ver Tabela 5). Isso permitiu identificar as variáveis com maior recorrência na análise de correlação entre as diferentes regiões do estado. Observou-se uma grande variação entre microrregiões, não sendo possível concluir que uma variável específica tenha menor ou maior correlação para todo o estado, demonstrando a disparidade entre as diferentes regiões do estado.

Posteriormente, o modelo de regressão linear criado com os dados do IDHM mostrou conseguir predizer as notas do IDEB do ano de 2019. Entretanto, deve-se considerar que há uma variação muito grande entre as notas do IDEB de uma cidade para outra. Esse modelo de regressão linear pode ser usado na predição de notas do IDEB do estado do Maranhão, com um erro aceitável.

Considerando os resultados da análise de fatores, o principal fator (mais significativo) que influencia a nota do IDEB é o fator da escolaridade dos pais, que considera o grau de estudo da mãe, pai ou responsável pelo aluno, e também a tecnologia, que considera se o aluno tem computador e acesso à Internet via rede sem fio. O fator de incentivo dos pais à educação dos filhos também é bastante relevante para determinar os indicadores educacionais. Esse fator considera o nível de incentivo dos pais ou responsáveis no que diz respeito a participar de reuniões escolares, conversar com os filhos sobre a escola, incentivar a estudar, incentivar a fazer as tarefas de casa, e incentivar a ir à escola. Os outros dois fatores (i.e., Cultura e Vida Escolar) compõem o resultado obtido na análise, os quais foram considerados menos significativos pela análise.

Por fim, a árvore de decisão criada com as variáveis correspondentes à pontuação dos fatores (obtidos na análise de fatores), em relação à variável alvo (i.e., notas do IDEB de 2019), foi explicativa. O fator relacionado a escolaridade dos pais e tecnologia foi considerado o nó raiz da árvore e, portanto, mais determinante na qualidade da educação, corroborando com a significância da análise de fatores. Embora tenha sido realizada uma poda na árvore, o modelo conseguiu discriminar se a nota seria abaixo ou acima da média estadual, baseado em um conjunto reduzido de fatores relacionados ao aluno.

6.2 Limitações

Um estudo de mineração de dados é um processo iterativo e, consequentemente, ele vem com limitações. Primeiramente, essa pesquisa se limitou apenas aos dados educacionais da rede pública estadual. Essa restrição aconteceu devido os dados do IDEB de escolas federais e particulares não estarem totalmente presentes, portanto pouco representativos. O motivo desses dados ausentes se deve ao fato de que poucos alunos nas turmas dessas escolas terem realizado a prova do SAEB e, dessa forma, os dados não foram suficientes para o cálculo da nota do IDEB. Em alguns casos, as escolas também solicitaram a não publicação dos seus dados do IDEB.

Uma vez que a qualidade educacional pode variar (i.e., evoluir ou regredir) de um ano para outro (no caso do IDEB, com avaliação a cada dois anos), outra limitação foi a não realização de uma análise histórica dos dados. Essa análise de série temporal permitiria verificar a evolução das diversas variáveis estudadas ao longo dos últimos anos, permitindo uma verificação das mudanças que ocorreram historicamente.

7 Considerações Finais

Este trabalho se propôs a identificar os fatores de influência da educação pública do Maranhão através da mineração de dados. Para isso, foram realizadas análises de dados e desenvolvimento de modelos para estudar e entender a educação maranhense. O objetivo do estudo foi alcançado e todos os critérios de sucesso foram atingidos, incluindo a identificação de respostas para a questão de pesquisa. O estudo conseguiu identificar os fatores que influenciam no ensino médio da rede pública estadual do Maranhão. Além disso, para uma melhor compreensão dos dados do IDEB, foram desenvolvidos modelos que contribuíram para o entendimento dos fatores de influência da educação do Maranhão. Os resultados alcançados são relevantes por contribuírem para um entendimento da educação do Maranhão, considerando as notas do IDEB e também dos dados adquiridos através do questionário socioeconômico do SAEB. Adicionalmente, a pesquisa evidenciou que a nota do IDEB tem relação direta com o IDHM do município em que o aluno cursa o ensino médio.

Como trabalhos futuros, pretende-se realizar uma comparação da qualidade educacional do Maranhão com a de outros estados, em particular, os do nordeste, por possuírem características socioeconômicas similares, mas desempenho no IDEB de 2019 diversificado (e.g., Pernambuco tem nota média 4,4 e Rio Grande do Norte tem 3,2). Também pretendemos analisar historicamente os dados do Maranhão ao longo dos anos, considerando anos anteriores a 2019 e incluindo também anos posteriores, quando os dados forem disponibilizados pelo INEP. Planos para trabalhos futuros incluem também a mineração de dados para analisar o impacto da pandemia do Coronavírus (COVID-19) na educação maranhense, por meio de uma análise comparativa de anos e uso de mais bases de dados. Por fim, uma vez que este estudo restringiu-se a análises computacionais, consideramos importante a realização de experimentos complementares para validar os resultados encontrados.

Agradecimentos

Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001; do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) - processo 308059/2022-0; e da Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) - processos UNIVERSAL-01026/18, UNIVERSAL-00745/19, BEPP-01608/21, BEPP-01768/21, UNIVERSAL-06123/22, e APP-09405/22.

Referências

Agarwal, S. (2013). Data mining: Data mining concepts and techniques. In *2013 international conference on machine intelligence and research advancement* (pp. 203–207). doi: 10.1109/ICMIRA.2013.45 [GS Search]

Akoglu, H. (2018). User's guide to correlation coefficients. *Turkish Journal of Emergency Medicine*, *18*(3), 91-93. doi: 10.1016/j.tjem.2018.08.001 [GS Search]

Azevedo, A., & Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview. In A. Abraham (Ed.), *Iadis european conf. data mining* (p. 182-185).

Barbieri, C. (2001). *Bi–business intelligence–modelagem & tecnologia*.

Brasil (1988). Constituição da república federativa do brasil de 1988. *Diário Oficial da República Federativa do Brasil*.

Carvalho, J., Cruz, L., & Gouveia, R. (2017). Descoberta de conhecimento com aprendizado de máquina supervisionado em dados abertos dos censos da educaçao básica e superior. In *Anais dos workshops do congresso brasileiro de informática na educação* (Vol. 6, pp. 674–683). doi: 10.5753/cbie.wcbie.2017.674 [GS Search]

Castro Soares, R., Neto, N. W., Coutinho, L. R., da Silva, F. J., dos Santos, D. V., & Teles, A. S. (2021). Mineração de dados da educaçao básica brasileira usando as bases do inep: Uma revisão sistemática da literatura. *RENOTE*, *19*(1), 361–370. doi: 10.22456/1679-1916.118526 [GS Search]

Curral, J. (1994). Statistics packages: A general overview.

DiStefano, C., Zhu, M., & Mindrila, D. (2009). Understanding and using factor scores: Considerations for the applied researcher. *Practical Assessment, Research, and Evaluation*, *14*(1), 20. doi: 10.7275/da8t-4g52 [GS Search]

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37–37. doi: 10.1609/aimag.v17i3.1230 [GS Search]

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Van Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of brazil. *Journal of Business Research*, *94*, 335–343. doi: 10.1016/j.jbusres.2018.02.012 [GS Search]

Figueiredo Filho, D. B., & Silva Júnior, J. A. d. (2010). Visão além do alcance: uma introdução à análise fatorial. *Opinião pública*, *16*(1), 160–185. doi: 10.1590/S0104-62762010000100007 [GS Search]

Fonseca, S. O. d., & Namen, A. A. (2016). Data mining on INEP databases: An initial analysis aiming to improve brazilian educational system. *Educação em Revista*, *32*, 133–157. doi: 10.1590/0102-4698140742 [GS Search]

Garcia, S. C. (2003). *O uso de árvores de decisão na descoberta de conhecimento na área da saúde*. Unpublished master's thesis. [GS Search]

Gorsuch, R. L. (2014). *Factor analysis: Classic edition*.

IBGE (2021). *Maranhão - IBGE Cidades. Disponível em: https://cidades.ibge.gov.br/brasil/ma/panorama.*

INEP (2022a). *Portal de Dados Abertos do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos.*

INEP (2022b). *Índice de Desenvolvimento da Educação Básica (IDEB) - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb.*

Júnior, G. C., Nascimento, R., Alves, G., & Gouveia, R. (2017). Identificando correlaçoes e outliers entre bases de dados educacionais. In *Anais dos workshops do congresso brasileiro de informática na educação* (Vol. 6, pp. 694–703). doi: 10.5753/cbie.wcbie.2017.694 [GS Search]

Lima, D. A., Ferreira, M. E. A., & Silva, A. F. F. (2021). Machine learning and data visualization

to evaluate a robotics and programming project targeted for women. *Journal of Intelligent & Robotic Systems*, *103*(1), 1–20. doi: 10.1007/s10846-021-01443-w [GS Search]

Lira, S. A. (2004). *Análise de correlação: Abordagem teórica e de construção dos coeficientes com aplicações*.

Maranhão, G. (2022). *Piso salarial do professor com jornada de 40 horas no maranhão é r$ 3 mil a mais que o nacional.*

Namen, A. A., Borges, S. X. d. A., & Sadala, M. d. G. S. (2013). Indicadores de qualidade do ensino fundamental: o uso das tecnologias de mineração de dados e de visões multidimensionais para apoio à análise e definição de políticas públicas. *Revista Brasileira de Estudos Pedagógicos*, *94*(238), 677–700. doi: 10.1590/S2176-66812013000300003 [GS Search]

Nascimento Bem, L., da Silva Pereira, V., & Souza, E. (2017). Data mart para análise comparativa de dados do ideb em municípios da microrregião do pajeú em pernambuco. In *Anais dos workshops do congresso brasileiro de informática na educação* (Vol. 6, pp. 704–713). doi: 10.5753/cbie.wcbie.2017.704 [GS Search]

Orange Data Mining (n.d.). *Orange documentation. tree viewer.*

Pacini, I. B. d. A. (2020). Indicadores educacionais: Um estudo dos limites e potencialidades da prova brasil da rede estadual de ensino do tocantins. *Humanidades & Inovação*, *7*(18), 242–257. [GS Search]

Pallant, J. (2020). *Spss survival manual: A step by step guide to data analysis using ibm spss*. Routledge. [GS Search]

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4, Part 1), 1432-1462. doi: 10.1016/j.eswa.2013.08.042 [GS Search]

Rigo, S., Cambruzzi, W., Barbosa, J., & Cazella, S. (2014). Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação*, *22*(01), 132. doi: 10.5753/rbie.2014.22.01.132 [GS Search]

Rigotti, J. I. R., & Cerqueira, C. A. (2015). As bases de dados do inep e os indicadores educacionais: conceitos e aplicações. *Livros*, 71–88. [GS Search]

Rodrigues, R. L., De Medeiros, F. P., & Gomes, A. S. (2013). Modelo de regressão linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. In *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)* (Vol. 24, p. 607-616). doi: 10.5753/cbie.sbie.2013.607 [GS Search]

Santos, A., & de Medeiros, F. P. A. (2020). Relationship of federal funding to ideb results in a state in brazil: an approach based on educational data mining. In *15th iberian conference on information systems and technologies (cisti)* (pp. 1–4). doi: 10.23919/CISTI49556.2020.9140924 [GS Search]

SEDUC-MA, S. E. E. M. (2020). *Maranhão mantém trajetória de crescimento e atinge 3,7 no ideb, maior nota da história.*

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining.

Silva Pinto, G., Júnior, O. d. G. F., & de Barros Costa, E. (2019). Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de teotônio vilela-alagoas. *RENOTE*, *17*(3), 183–193. doi: 10.22456/1679-1916.99469 [GS Search]

Silva Pinto, G., Júnior, O. F., Costa, E., Barbirato, J. C. C., & Rodrigues, W. R. M. (2019). Iden-

tificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de maceió. In *Brazilian symposium on computers in education* (Vol. 30, pp. 1828–1837). doi: 10.5753/cbie.sbie.2019.1828 [GS Search]

Tan, P.-N., Steinbach, M., & Kumar, V. (2009). *Introdução ao datamining: mineração de dados*.

Witten, I. H., Frank, E., Hall, M. A., Pal, C. J., & DATA, M. (2005). Practical machine learning tools and techniques. In *Data mining* (Vol. 2, p. 4). doi: 10.1016/C2009-0-19715-5 [GS Search]