Mineração de Dados para investigar o IDEB usando o Censo da Educação Básica e SAEB: um estudo de caso em Sergipe

Mariana Lira de Farias, UFS, marilira1998@gmail.com
ORCID ID: 0009-0007-3113-2849
Renê Pereira de Gusmão, UFS, renepgusmao@gmail.com
ORCID ID: 0000-0002-4806-6506
Cleonides Silva Dias Gusmão, UFPB, cleonides.silva@academico.ufpb.br
ORCID ID: 0000-0002-6094-5642

Resumo: O Índice de Desenvolvimento da Educação Básica (IDEB) é a métrica que diagnostica a qualidade do ensino a partir do desempenho médio e do fluxo escolar. Este estudo utilizou mineração de dados e métodos de aprendizagem de máquina para investigar variáveis associadas ao IDEB. Foram utilizados dados do estado de Sergipe de 2017 do Sistema de Avaliação da Educação Básica (SAEB) e Censo Escolar. As categorias de variáveis usadas foram: contexto familiar dos estudantes, estrutura das escolas e condições de trabalho dos professores. Verificou-se que as variáveis socioeconômicas dos estudantes tem uma forte relação com as notas do IDEB das escolas, sobretudo as variáveis relacionadas aos pais ou responsáveis pelos alunos e situação financeira familiar.

Palavras-chave: Mineração de dados, IDEB, Educação Básica, SAEB.

Data Mining to investigate IDEB using School Census of Basic Education and SAEB: a case study in Sergipe

Abstract: The Basic Education Development Index (IDEB) is a metric that diagnoses the quality of education based on average performance and school flow. This study uses data mining and machine learning methods to investigate variables related to IDEB. Data from the state of Sergipe from 2017 from the Basic Education Assessment System (SAEB) and School Census were used. The categories of variables used were: students' family context, school structure and teachers' working conditions. It was found that the socioeconomic variables of students have a strong relationship with the IDEB scores of schools, especially as students related to parents or guardians and family financial situation.

Keywords: Data Mining, IDEB, Basic Education, SAEB.

1. Introdução

A qualidade educacional é um tema de grande importância. Atualmente, o principal mecanismo nacional utilizado para sua avaliação é o Índice de Desenvolvimento da Educação Básica (IDEB). O indicador relaciona informações do rendimento escolar, obtidos pelo Censo Escolar, e de desempenho oriundos do Sistema Educacional de Avaliação Básica (SAEB) e Prova Brasil (Fernandes, 2007).

Através do IDEB, o governo realiza o diagnóstico do contexto educacional do país e orienta as políticas desse cunho, com o objetivo de melhorar a qualidade do ensino básico (Inep, 2020c). Contudo, autores como Chirinéa e Brandão (2015) e Gusmão, Gusmão e Dias (2021) alertam para o simplismo de se avaliar o desempenho escolar apenas pela ótica de exames padronizados. Para Fernandes e Gremaud (2009) e Freitas (2007), os resultados de exames padronizados são insuficientes para medir o desempenho escolar, visto que este não é consequência apenas do ambiente escolar.

Entender o IDEB como o índice que revela a qualidade da educação é tratá-la a partir de uma visão reducionista e psicologizante por meio do direcionamento apenas para variáveis individuais, excluindo-se uma estrutura mais ampla de variáveis sociais e econômicas que afetam este fenômeno (Chirinéa e Brandão, 2015). Nesse sentido, as condições materiais em que se dá essa educação devem ser consideradas. É importante destacar, ainda, que a ideia da existência de igualdade de oportunidades e a atribuição do fracasso escolar a variáveis individuais esteve presente no cerne da história da Psicologia da Educação e vem influenciando certos discursos no campo da educação até hoje (Patto, 1999).

Portanto, na atualidade, o IDEB é tratado como sendo um índice ”neutro”para avaliação do desempenho e da qualidade educacional baseado na ideologia meritocrática, ignorando a grande importância da desigualdade social e o seu impacto em relação ao fenômeno educacional (Freitas, 2007). Adicionalmente, como pontuam Souza (2020) e Viégas (2020), não é possível analisar o desempenho dos alunos com foco direcionado apenas a variáveis individuais e familiares, devendo-se considerar um conjunto de variáveis políticas, econômicas, históricas, sociais, pedagógicas e institucionais.

Dentro desse contexto, segundo Baker, Isotani e Carvalho (2011), a Mineração de Dados Educacionais (MDE) surge como alternativa para investigar dados obtidos em ambientes educacionais. A MDE vem sendo utilizada para investigar a previsão de problemas como: evasão (Calixto; Segundo e Gusmão, 2017; Colpani, 2018), desempenho (Laisa e Nunes, 2015) e IDEB (Gusmão; Gusmão e Dias, 2021). Assim sendo, o presente trabalho realizou um estudo em busca dos principais fatores que influenciam a nota do IDEB, utilizando informações de três categorias: informações socioeconômicas dos estudantes, estrutura física das escolas e condição de trabalho dos professores. Para tanto, utilizou-se a MDE em um conjunto de dados de escolas do 3º ano do ensino médio do estado de Sergipe.

Este trabalho está organizado da seguinte forma: a seção 2 apresenta os trabalhos relacionados, a seção 3 apresenta a metodologia aplicada, os resultados obtidos e discussão são mostrados na seção 4. Por fim, a última seção contém as conclusões da pesquisa.

2. Trabalhos Relacionados

Nesta seção serão descritos trabalhos relacionados, bem como o diferencial do presente estudo. Em Pinto, Júnior e Costa (2019) foi realizada uma pesquisa a fim de identificar os fatores que afetam o desempenho escolar dos alunos do 9º ano do ensino fundamental de escolas públicas da cidade de Teotônio Vilela- AL. Para isso, foram exploradas técnicas de seleção de atributos e algoritmos preditivos (NaiveBayes, J48, JRip, LibSVM, RandomForest, IBK, OneR e REPTree). Desses algoritmos, o OneR, LibSVM e J48 apresentaram maior acurácia. Alguns atributos que se destacaram nessa pesquisa por influenciar o desempenho escolar foram: notas em português, a atenção do professor em relação a correção de atividades, a utilização da biblioteca/sala de leitura da escola, a prática de leitura da mãe, a prática de leitura do aluno e fatores socioeconômicos (a casa possuir TV, banheiro, máquina de lavar).

Já em Medeiros e Santiago (2019), o objetivo foi compreender a relação dos perfis escolares dos estudantes e o impacto das políticas públicas educacionais no Estado de Pernambuco. Foram utilizadas amostras do ano de 2017 de turmas do 3º ano do ensino médio e algoritmo de florestas aleatórias. Os resultados apontaram a necessidade de atuação de políticas públicas no combate à distorção idade-ano e a reprovação.

Silva, Silva e Lima (2020) aplicaram técnicas de mineração de dados a fim

de identificar as atividades desenvolvidas pelos Diretores da Escola que impactaram no desempenho dos alunos do Ensino Médio nas avaliações do SAEB de 2017. Foi identificado que o tempo de experiência, a titulação do diretor, as ações para reduzir as reprovações e nível socioeconômico da escola contribuíram para o desempenho escolar dos alunos.

Góes e Steiner (2016) desenvolveram uma metodologia para criação de etiquetas de classificação das escolas públicas em Araucária- PR a partir das notas da Prova Brasil. Para isso, foram explorados três algoritmos que obtiveram resultados satisfatórios, a saber: SVM, Redes Neurais Artificiais e Algoritmos Genéticos.

Em Canedo *et al.* (2019) foi analisado como a qualificação acadêmica dos professores impacta nas notas do IDEB das escolas. O estudo de caso compreendeu docentes que trabalhavam em escolas do Estado de Goiás e o algoritmo utilizado é o priori. Foi constatado que o nível de pós-graduação dos professores tem impacto na nota final do IDEB.

Calixto, Segundo e Gusmão (2017) buscaram identificar as variáveis relacionadas a evasão escolar. Para tal, utilizou-se dados do censo educacional dos anos de 2014, 2015 e 2016 dos estados de Ceará e Sergipe e os algoritmos utilizados foram de regressão logística e indução da regra. A idade, etapa de ensino, modalidade de ensino, existência de laboratórios e localização da escola foram variáveis associadas à evasão escolar.

O presente estudo tem seu diferencial em combinar amostras de dados com três categorias de informações: dados socioeconômicos dos alunos, estrutura física da escola e condições de trabalho dos professores, devido a compreensão de que o desempenho educacional envolve diversos aspectos além das fronteiras escolares. Além de utilizar dados de um estado pouco explorado em pesquisas desse cunho.

3. Metodologia

Nesta seção serão apresentadas as etapas de desenvolvimento do estudo. O objetivo das análises é encontrar os fatores educacionais que mais influenciam o IDEB. Para tanto, para compor as amostras de dados, foram escolhidos dados educacionais do ano de 2017 de turmas do 3º ano do ensino médio do estado de Sergipe.

É importante salientar que, para nortear a seleção das técnicas de mineração de dados utilizadas nos experimentos, foi desenvolvida uma busca, junto à literatura. A partir dessa etapa, foi possível perceber a vasta aplicação de algoritmos de aprendizagem supervisionada, sobretudo, voltados a tarefa de classificação e a utilização das técnicas de seleção de atributos como método de preparação dos dados. Além disso, a CRISP-DM (*Cross Industry Standard Process for Data Mining*) foi a metodologia de mineração de dados mais encontrada nos estudos encontrados.

Portanto, a metodologia aplicada no processo de mineração de dados foi a CRISP-DM, uma abordagem composta por 6 fases bem definidas, a saber: Compreensão do Domínio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação. Cada fase do CRISP-DM desenvolvida neste estudo será descrita a seguir, exceto a fase de implantação.

3.1. Compreensão do Domínio

Para alcançar o objetivo dos experimentos, serão aplicadas técnicas de seleção de atributos e algoritmos de classificação em uma amostra de dados derivadas do SAEB e Censo escolar. As questões norteadoras das análises são:

1. Quais os modelos de classificação possuem o melhor desempenho?
2. Quais foram as variáveis associadas ao IDEB identificadas nas análises?

3.2. Compreensão dos Dados

As bases de dados originais foram obtidas no portal de microdados do INEP*. A base do SAEB é composta por respostas de uma série de questionários aplicados em professores, diretores e estudantes, enquanto o Censo Escolar é composto por características dos estabelecimentos de ensino, gestores, turmas, alunos e profissionais escolares em sala de aula, coletadas em duas fases: no período de matrícula e no final do ano letivo (Inep, 2020a).

Neste estudo, foram utilizadas as tabelas referentes ao questionário dos alunos do 3º ano do ensino médio (Estado de Sergipe) do SAEB e as tabelas de docentes, escolas e turmas do Censo Escolar, ambas estão disponíveis em formato CSV. O questionário do aluno é composto por 60 itens que abordam assuntos como nível socioeconômico, participação da família e atividades pedagógicas (Inep, 2020b). Já as tabelas do Censo Escolar, são compostas por informações voltadas a estrutura física das escolas, composição das turmas e corpo docente, como por exemplo, nível de especialização, o tipo de atividade profissional dos professores, tipo de contratação. (Inep, 2020a).

3.3. Preparação dos Dados

Inicialmente, as bases de dados originais foram filtradas a partir do estado de Sergipe. Em seguida, foram criadas três bases de dados. A primeira com informações dos estudantes (chamada de Base A), a segunda com informações das escolas e professores (chamada de Base B) e a última com a união dessas duas bases (chamada de Base C). A primeira contém variáveis derivadas do questionário do aluno do SAEB e a segunda atributos derivados de professores e da estrutura física das escolas, oriundas do censo escolar, mas também da nota do IDEB das instituições, coletada manualmente no portal do IDEB†. As variáveis derivadas do questionário do aluno do SAEB foram criadas a partir do agrupamento a nível de escola, da quantidade de respostas para cada item das perguntas presentes no questionário, conforme o exemplo apresentado na Tabela 1.

Tabela 1. Exemplo de criação das variáveis deriavadas

Questão 20		
Nome Original da Questão na base do SAEB	**Variáveis Derivadas**	
TX_RESP_Q020	TX_RESP_Q020_A	TX_RESP_Q020_B
Descrição		
Sua mãe, ou a mulher responsável por você, sabe ler e escrever?	Variável referente a quantidade de resposta ”Sim”para uma escola	Variável referente a quantidade de resposta ”Não”para uma escola

A integração das bases A e B, para formação da base C, foi feita utilizando o código das escolas presente nas bases do SAEB e do Censo Escolar. Portanto, este estudo teve como diferencial a criação de uma base de dados (base C) contendo informações de bases distintas, agregando dados sobre a estrutura das escolas, informações sobre professores, ambiente familiar dos alunos e IDEB das escolas.

Tabela 2. Composição das bases de dados desenvolvidas

Base	Quantidade Total de Atributos	Quantidade de Atributos Derivados	Quantidade de Atributos das Bases originais	Descrição
A	235	227	8	Base com informações do questionário do SAEB
B	183	18	165	Base com informações das escolas e Professores vindas do Censo Escolar
C	418	245	173	União das bases A e B

*Portal do INEP: ⟨https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados⟩

†Portal IDEB: ⟨http://ideb.inep.gov.br/⟩

A aplicação da técnica de discretização das notas do IDEB também foi realizada nesta etapa. Esse processo é aplicado em variáveis contínuas, que consiste em mapear os valores contínuos para valores discretos. Dessa forma, criou-ou, através do cálculo da média do IDEB para cada base, a variável dependente "Classe-Ideb", composta por três valores: abaixo da média, média e acima da média.

Após a criação, as bases ainda passaram por outras etapas de pré-processamento, a saber: retirada de colunas com 60% dos valores nulos e preenchimento de valores faltantes por zero. Esse procedimento foi realizado de modo a facilitar a condução da pesquisa. Por fim, foram aplicadas as técnicas de seleção de atributos com o objetivo de encontrar o subconjunto de variáveis que realizam a melhor previsão das notas do IDEB. Dessa maneira, foram utilizadas as seguintes abordagens de seleção de atributos: filtragem, embaralhamento e embutida, além da aplicação do método merge (Lima, 2016). A escolha dessas abordagens foi feita baseada em um estudo semelhante realizado em (Pinto *et al.*, 2019)

3.4. Modelagem

Nessa fase foram realizados dois conjuntos de experimentos. O primeiro conjunto sem a seleção de atributos das bases de dados e o segundo conjunto com a seleção de atributos das bases de dados. A métrica de avaliação utilizada foi a acurácia, que indica a porcentagem de acerto de um classificador, ou seja, quanto maior a acurácia, melhor é o modelo gerado (Matos *et al.*, 2009).

Para criação dos modelos de aprendizagem de máquina, foram escolhidos algoritmos de classificação devido as características das bases de dados e a ampla utilização desses algoritmos na literatura em problemas desse cunho, conforme encontrado na revisão sistemática. Portanto, os seis algoritmos selecionados, os mais encontrados na revisão, foram: Árvore de Decisão (AD), Florestas Aleatórias (FA), Máquina de Vetor de Suporte (SVM), Naive Bayes (NB), K-vizinhos aleatórios (KNN) e Regressão Logística (RL). É importante destacar que foi utilizado o método Holdout (Han; Pei e Kamber, 2011), onde 80% das amostras foram utilizadas para treinamento e 20% para teste.

Além de avaliar os modelos quanto a acurácia, as variáveis presentes no melhor modelo foram analisadas quanto ao seu nível de importância e impacto na classificação. O processo de identificação das variáveis mais influentes se deu a partir da aplicação e anaálise do algoritmo de floresta aleatória na base de dados com maior acurácia. No que diz respeito as tecnologias utilizadas, *Python* é a linguagem de programação utilizada tanto na fase de pré-processamento quanto na criação dos modelos de aprendizagem de máquina e os ambientes utilizados são a IDE *Spyder* e a ferramenta Jupyter Notebook. A fase 5 será apresentada na seção seguinte.

4. Resultados e Discussões

Nesta seção serão apresentados os resultados encontrados a partir da utilização das três bases de dados que podem ser observados detalhadamente na Tabela 3.

Os resultados revelaram que a aplicação da técnica de seleção de atributos contribuiu significamente com o aumento da acurácia dos modelos para a maioria dos casos. Além disso, pode-se perceber que o experimento com a base de dados C obteve os melhores resultados. Ainda, os modelos desenvolvidos a partir da base C com a utilização da filtragem e o método merge obtiveram os melhores resultados. Dentre eles, os algoritmos de Floresta Aleatória, Árvore de Decisão, SVM e Regressão Logística alcançaram as maiores acurácias sendo, 96.05%, 90.79 %, 93.42 % e 94.74 %

Tabela 3. Resultados das Análises na base de dados de Sergipe

Método	**Acurácia**												
	Sem Seleção de Atributos			Com Seleção de Atributos									
				Filtragem			Embutida			Embaralhamento			Merge
	A	B	C	**A**	B	C	A	B	C	A	B	C	
FA	75.0 %	43.75 %	56.25 %	81.25 %	50.0 %	**96.05 %**	62.5 %	68.75 %	87.5 %	75.0 %	56.25 %	87.5 %	**97.37 %**
NB	68.75 %	56.25 %	75.0 %	75.0 %	56.25 %	78.95 %	62.5 %	62.5 %	81.25 %	62.5 %	56.25 %	68.75 %	82.89 %
AD	62.5 %	25.0 %	56.25 %	62.5 %	37.5 %	**90.79 %**	56.25 %	37.5 %	68.75 %	68.75 %	68.75 %	68.75 %	**93.42 %**
SVM	75.0 %	37.5 %	50.0 %	68.75 %	37.5 %	**93.42 %**	50.0 %	56.25 %	81.25 %	62.5 %	68.75 %	**93.75 %**	**93.42 %**
RL	81.25 %	37.5 %	56.25 %	87.5 %	37.5 %	**94.74 %**	56.25 %	62.5 %	75.0 %	62.5 %	62.5 %	75.0 %	**93.42 %**
KNN	50.0 %	81.25 %	68.75 %	68.75 %	81.25 %	69.74 %	62.5 %	56.25 %	43.75 %	75.0 %	56.25 %	62.5 %	76.32 %

respectivamanete na filtragem e 97.37 %, 93.42 %, 93.42 % e 93.42 % no *merge*.

O destaque da eficiência dos algoritmos com a utilização da base de dados C, a qual considera aspectos relacionados ao contexto familiar dos alunos, condições de trabalho dos professores e estrutura das instituições escolares, reforça a importância da consideração de múltiplas dimensões quando trata-se de qualidade educacional e desempenho escolar dos estudantes, como bem apontam Viégas (2020), Souza (2020), Gusmão, Gusmão e Dias (2021), Chirinéa e Brandão (2015), Freitas (2007).

Especificamente, as variáveis que mais influenciaram a nota do IDEB podem ser vistas na Tabela 4. Como foi dito anteriormente, pode-se perceber que com a utilização da base de dados C os melhores resultados foram obtidos. Portanto, a combinação das informações socioeconômicas dos estudantes, atividades profissionais dos docentes e a estrutura física das escolas aumentou a acurácia das previsões das classes de notas do Ideb. Esse panorama indica que o desempenho escolar, de fato, é uma síntese de múltiplas determinações.

Tabela 4. Principais variáveis associadas ao IDEB de Sergipe

Questão	**Resposta**
Você já abandonou a escola durante o período de aulas e ficou fora da escola o resto do ano?	Duas vezes ou mais
Você concluiu o Ensino Fundamental na Educação de Jovens e Adultos(EJA), antigo supletivo?	Sim
Na sua casa tem computador?	Dois
Total de Funcionários da Escola	Maior número de funcionários
Na sua casa tem carro?	Um

Tal resultado se assemelhou com os achados de Silva *et al.* (2020), estes encontraram dentre as causas associadas ao desempenho escolar fatores ligados a três grupos: familía do aluno, comportamento do aluno e a escola. As variáveis que contém um maior grau de importância na classificação podem ser vistas na Tabela 4, sendo elas: os estudantes que abandonaram a escola mais de uma vez, estudantes que concluiram o ensino na modalidade EJA, a quantidade total de funcionários na escola e os alunos que possuem dois computadores em casa. Segundo Filho e Araújo (2017), o abandono - juntamente com a evasão - é uma das maiores fraquezas da educação no país e a situação socioeconômica dos estudantes pode influenciar fortemente essa decisão.

Em relação a isso, destacaram-se as variáveis relacionadas a estrutura econômica familiar e o grau de formação dos pais. A presença de atributos ligados a estrutura econômica familiar pode estar relacionada ao IDEB da escola e, consequentemente, ao desempenho escolar, pois uma maior renda pode propiciar um melhor acesso a educação (Simões, 2016). De acordo com Freitas (2007), o nível socioeconômico destaca-se como importante aspecto quando trata-se de desempenho escolar, ressaltando, com isso, a necessidade do combate à desigualdade social.

É importante destacar que foi identificado que alunos de escolas com nota do IDEB acima da média apresentaram mais posses em suas casas, como carro, televisão, dvd, computador, geladeira, além da quantidade de quartos e banheiros.

Em relação ao grau de formação dos pais, foi percebido que quanto maior o grau de instrução, maior a nota do IDEB. Além disso, os alunos tendem a ter notas abaixo da média quando os pais ou responsáveis não frequentam reuniões escolares ou não estimulam e acompanham a vida escolar dos estudantes. De forma concordante, segundo Silva, Silva e Martins (2019), o acompanhamento dos responsáveis nas atividades escolares contribui diretamente no desempenho dos alunos.

Segundo Silva *et al.* (2020), o nível de escolaridade dos responsáveis impacta no desempenho escolar, na medida em que aumenta a probabilidade do acompanhamento no processo educativo do estudante. Desse modo, foi constatado que, para amostra de dados utilizadas, estudantes de escolas com nota do IDEB acima da média, tem pais ou responsáveis com o ensino médio completo.

Ainda a respeito do núcleo familiar, foi encontrado outro dado relevante. Os alunos que fazem as atividades domésticas tiveram um desempenho inferior se comparado aos que não realizam. Segundo Alberto *et al.* (2009), as horas empregadas em tarefas desse cunho deixam de ser dispendidas em outras atividades que teriam impacto positivo na vida do indivíduo, como, por exemplo, o estudo fora da escola.

Outro grupo de variáveis que merecem destaque são as voltadas a vida estudantil. As questões mostram que o início da vida letiva, a modalidade de conclusão do ensino fundamental e o abandono escolar se apresentaram como importantes em relação a nota do IDEB. Além disso, o tempo gasto em entretenimento também se destacou.

No conjunto de variáveis relacionadas as instituições de ensino, foi identificado que quanto melhor a estrutura física, melhor é o desempenho escolar, sobretudo a presença de biblioteca, sala de leitura, quadra de esportes, existência de saneamento básico e laboratório de informática. Apesar da escola não influenciar diretamente na situação socioeconômica dos discentes, pode criar mecanismos para garantir acesso a educação (Silva; Silva e Lima, 2020). Dessa maneira, o acesso a serviços básicos, como água e esgoto sanitário; dependências escolares, como biblioteca ou sala de leitura; infraestrutura de comunicação e informação; são fundamentais para o desempenho escolar (Sátyro; Soares *et al.*, 2007).

No grupo de variáveis dos professores, temos que a pós-graduação se mostrou como um item de grande importância, mais especificamente quantidade de professores com mestrado e doutorado. Esse resultado se assemelha a Canedo *et al.* (2019) que constatou que as regiões do país que contém docentes com maior nível de formação também apresentam maiores notas do IDEB.

De maneira geral, pode-se perceber uma forte influência das variáveis socioeconômicas, sendo as mais importantes: a estrutura econômica familiar, grau de formação e letramento dos pais ou responsáveis, o seu acompanhamento e incentivo na vida estudantil e a participação dos estudantes nas atividades pedagógicas. Além das variáveis desse grupo, também foi identificado a presença de atributos ligados a estrutura física da escola e capacitação profissional do professor.

Os resultados relatados se assemelham ao que foi encontrado por Filho (2012), em que as variáveis que melhor explicaram o desempenho escolar estão relacionadas a família e comportamento do aluno, como a educação da mãe, atraso escolar, início da vida acadêmica, reprovação prévia, presença de computador em casa e trabalho fora de casa.

Como pode-se perceber houve uma predominância das variáveis socioeconômicas, principalmente as voltadas ao ambiente familiar, mas também uma parcela significativa daquelas voltadas a escola. O que comprova que o desempenho escolar não depende apenas de variáveis individuais (Vasconcelos *et al.*, 2020). Portanto,

analisar o desempenho escolar apenas por testes padronizados, pode não trazer um panorama verídico, necessitando assim de estudos envolvendo múltiplos contextos e aspectos que estão relacionados ao complexo contexto educacional (Chirinéa e Brandão, 2015; Gusmão; Gusmão e Dias, 2021).

5. Conclusões

Este estudo teve como objetivo a investigação do IDEB a partir de bases de dados disponibilizadas no portal de Microdados do INEP. Destaca-se a integração das bases do Censo Escolar da Educação Básica e do SAEB, além da coleta manual da nota de IDEB das escolas. A integração das bases foi feita a partir do código das escolas e foi criada uma variável dependente de acordo com a média do IDEB das escolas.

Os resultados deste estudo demonstraram como a computação e, mais especificamente, a mineração de dados, podem contribuir para a análise de importantes questões relacionadas a educação e para a melhoria do sistema educacional, apontando relevantes aspectos que devem ser considerados ao se pensar políticas públicas, especialmente as educacionais. Aspectos que merecem destaque a partir dos resultados encontrados são: a importância do investimento em educação para a melhoria da estrutura material da escola; importância da qualificação profissional do docente; e a necessidade do combate à desigualdade social.

Nesse sentido, os objetivos estabelecidos foram alcançados, tendo sido possível identificar as técnicas de mineração de dados mais eficientes para o estudo em questão e as variáveis associadas ao IDEB.A partir das análises realizadas, pôde-se perceber que vários aspectos podem estar relacionados as notas do IDEB, dentre eles a renda familiar, a formação acadêmica dos responsáveis, as condições da estrutura física das instituições de ensino, bem como o nível de especialização dos professores. Esse panorama mostra a importância de avaliar o desempenho escolar por uma ótica mais abrangente, não apenas da perspectiva de testes em sala de aula, pois outros aspectos envolvendo aluno, professor, instituição de ensino também são importantes.

Para trabalhos futuros, pode-se realizar experimentos com bases de dados mais robustas. Ademais, a inserção de novas variáveis, tais como variáveis relacionadas a gestão escolar, políticas públicas educacionais, remuneração docente, investimento governamental em educação, também pode levar a resultados relevantes. Por fim, a adoção de outras abordagens de seleção de atributos com o intuito de aumentar a acurácia, pode ser uma abordagem a ser adotada em pesquisas futuras.

Referências

Alberto, M. D. *et al.* Trabalho infantil doméstico: Perfil bio-sócio-econômico e configuração da atividade no município de joão pessoa, pb. **Cadernos de Psicologia Social do Trabalho**, v. 12, n. 1, p. 57, 2009.

Baker, R.; Isotani, S.; Carvalho, A. Mineração de Dados Educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, SBC, v. 19, n. 02, p. xx, xx 2011. ISSN 1414-5685. GS Search.

Calixto, K.; Segundo, C.; Gusmão, R. P. de. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2017. v. 28, n. 1, p. 1447.

Canedo, E. D.; Carvalho, R. R. D.; Leão, H. A. T.; Costa, P. H. T.; Okimoto, M. V. How the academics qualification influence the students learning development. In: IEEE. **2019**

IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC). [S.l.], 2019. v. 1, p. 336–345.

Chirinéa, A. M.; Brandão, C. d. F. O ideb como política de regulação do estado e legitimação da qualidade: em busca de significados. **Ensaio: Avaliação e Políticas Públicas em Educação**, SciELO Brasil, v. 23, p. 461–484, 2015.

Colpani, R. Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do censo escolar. **Informática na educação: teoria & prática**, v. 21, n. 3, dez. 2018. Disponível em: ⟨https://seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/87880⟩.

Fernandes, R. Índice de desenvolvimento da educação básica (ideb): metas intermediárias para a sua trajetória no brasil, estados, municípios e escolas. **Brasil: INEP/MEC**, 2007.

Fernandes, R.; Gremaud, A. P. Qualidade da educação: avaliação, indicadores e metas. **Educação básica no Brasil: construindo o país do futuro. Rio de Janeiro: Elsevier**, v. 1, p. 213–238, 2009.

Filho, N. A. M. Os determinantes do desempenho escolar do brasil. In: _____. [S.l.]: Saraiva, 2012.

Filho, R. B. S.; Araújo, R. M. de L. Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. **Educação por escrito**, v. 8, n. 1, p. 35–48, 2017.

Freitas, L. C. d. Eliminação adiada: o ocaso das classes populares no interior da escola e a ocultação da (má) qualidade do ensino. **Educação & Sociedade**, SciELO Brasil, v. 28, n. 100, p. 965–987, 2007.

Góes, A. R. T.; Steiner, M. T. A. Proposta de metodologia para a criação de etiqueta de classificação–estudo de caso: desempenho escolar. **Gestão & Produção**, SciELO Brasil, v. 23, p. 177–191, 2016.

Gusmão, R. P.; Gusmão, C. S.; Dias, M. S. A qualidade da educação para além do ideb: Um estudo através de técnicas de mineração de dados. In: SBC. **Anais do XXXII Simpósio Brasileiro de Informática na Educação**. [S.l.], 2021. p. 803–812.

Han, J.; Pei, J.; Kamber, M. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011.

Inep. **Censo Escolar**. 2020. (Accessed on 11/05/2021). Disponível em: ⟨https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar⟩.

Inep. **Testes E questionários**. 2020. (Accessed on 11/05/2021). Disponível em: ⟨https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/saeb/testes-e-questionarios⟩.

Inep. **Índice de Desenvolvimento da Educação Básica (Ideb) — Inep**. 2020. ⟨https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb⟩. (Accessed on 11/05/2021).

Laisa, J.; Nunes, I. D. Mineração de dados educacionais como apoio para a classificação de alunos do ensino médio. In: **Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**. [S.l.: s.n.], 2015. v. 16.

Lima, R. A. F. de. Estratégias de seleção de atributos para detecção de anomalias em transações eletrônicas. Universidade Federal de Minas Gerais, 2016.

Matos, P. F. *et al.* Relatório técnico "métricas de avaliaçao". **Universidade Federal de Sao Carlos**, 2009.

Medeiros, H.; Santiago, K. Políticas públicas educacionais baseadas em evidências: tomada de decisão apoiada em algoritmos de mineração de dados a partir dos questionários da avaliação nacional da educação básica (aneb). **Tecnologias da Educação**, 2019.

Patto, M. H. S. **A produção do fracasso escolar: histórias de submissão e rebeldia**. [S.l.]: Casa do Psicólogo, 1999.

Pinto, G. d. S. *et al.* Modelo de análise e predição para identificação dos fatores que influenciam o desempenho escolar na rede de ensino básico: estudo de caso em escolas municipais de alagoas. Universidade Federal de Alagoas, 2019.

Pinto, G. da S.; Júnior, O. d. G. F.; Costa, E. de B. 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**, v. 17, n. 3, p. 183–193, 2019.

Sátyro, N.; Soares, S. *et al.* **A infra-estrutura das escolas brasileiras de ensino fundamental: um estudo com base nos censos escolares de 1997 a 2005**. [S.l.], 2007.

Silva, I. V.; Silva, M. T.; Martins, S. A. S. Predictive success factors in school performance: An analysis of the large-scale asessment in brazil. In: **Conference: International Conference ICT, Society, and Human Beings 2019**. [S.l.: s.n.], 2019.

Silva, I. V. da; Silva, M. T. da; Lima, N. D. da S. Fatores preditivos de desempenho escolar em avaliações do saeb: influência da gestão escolar. **Research, Society and Development**, v. 9, n. 10, p. e9509109423–e9509109423, 2020.

Silva, P. M.; Lima, M. N.; Fagundes, R. A.; Souza, F. da F de. Dep–dm: Uma abordagem baseada em ensemble regression para o diagnóstico de problemas educacionais. **RENOTE**, v. 18, n. 1, 2020.

Simões, A. A. As metas de universalização da educação básica no plano nacional de educação o desafio do acesso e a evasão dos jovens de famílias de baixa renda no brasil. **Série PNE em Movimento**, n. 4, p. 52–52, 2016.

Souza, B. P. Orientação à queixa escolar. In: ____. 1. ed. [S.l.]: Casa do Psicólogo, 2020. cap. 1, p. 241–278.

Vasconcelos, J. C.; Lima, P. V. P. S.; Rocha, L. A.; Khan, A. S. Infraestrutura escolar e investimentos públicos em educação no brasil: a importância para o desempenho educacional. **Ensaio: Avaliação e Políticas Públicas em Educação**, SciELO Brasil, v. 29, p. 874–898, 2020.

Viégas, L. S. Psicologia escolar e educacional no brasil: a importância da autocrítica. **Psicologia escolar e educacional [recurso eletrônico]: processos educacionais e debates contemporâneos**, Edições do Bosque UFSC/CFH, p. 14–32, 2020.