CBIE 2023
CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO
Uma escola para o futuro: Tecnologia e conectividade a serviço da educação

# Mineração de Dados Educacionais: Investigando a Relação entre os Microdados do INEP e o Desempenho do IDEB

**Mariana de Lira Farias [1], Renê Pereira de Gusmão[1], Cleonides Silva Dias Gusmão[2]**

[1]Departamento de Computação– Universidade Federal de Sergipe (UFS)
São Cristóvão – SE – Brazil

[2]Departamento de Fundamentação da Educação – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brazil

{marianalf,rene}@dcomp.ufs.br, cleonides.silva@academico.ufpb.br

***Abstract.*** *The objective of this work was to investigate the Basic Education Development Index in search of the main factors that influence its result, through educational data mining. For this, information obtained from the INEP Microdata referring to the Basic Education Census and the Basic Education Assessment System were analyzed. The data used refer to the socioeconomic status of students, the physical structure of schools and the working conditions of teachers in third-year high school classes in three states, namely: Sergipe, Bahia and Alagoas. It was identified that the student's socioeconomic variables have a great influence on achieving the IDEB's projected goal.*
***Keywords****: EDM. Data Mining. IDEB. Quality of Education.*

***Resumo.*** *O objetivo deste trabalho foi investigar o Índice de Desenvolvimento da Educação Básica em busca dos principais fatores que influenciam o seu resultado, através de mineração de dados educacionais. Para isso, foram analisados informações obtidas dos Microdados do INEP referentes ao Censo da Educação Básica e ao Sistema de Avaliação da Educação Básica. Os dados utilizados são referentes à situação socioeconômica de estudantes, à estrutura física das escolas e às condições de trabalho dos professores de turmas de terceiro ano do ensino médio de três estados, a saber: Sergipe, Bahia e Alagoas. Identificou-se que variáveis socioeconômicas dos estudantes possuem influência na obtenção da meta projetada do IDEB.*
***Palavras-chave****: MDE. Mineração de dados. IDEB. Qualidade Educacional.*

## 1. Introdução

A qualidade educacional é um assunto de extrema importância e no Brasil, o principal indicador de qualidade da educação básica é o Índice de Desenvolvimento da Educação Básica (IDEB) [Fernandes and GREMAUD 2009]. Criado em 2007, pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), torna possível o monitoramento do aprendizado e a criação de metas para a aumentar a qualidade do ensino brasileiro [Inep 2020c].

De acordo com [Fernandes 2007], o IDEB é uma nota que varia entre 0 e 10, que combina informações de rendimento escolar oriundas do Censo Escolar e a média de proeficiência em exames no Sistema de Avaliação da Educação Básica (SAEB) e Prova Brasil.

O Censo Escolar é o maior e principal meio de acompanhamento da situação educacional do país, abrangendo todas as modalidades de ensino básico [Inep 2020a]. Já o SAEB e a Prova Brasil são avaliações aplicadas a cada dois anos para diagnóstico do nível de aprendizagem dos estudantes, através de um conjunto de questionários socioeconômicos e testes padronizados [MEC 2009].

O IDEB é um importante condutor das políticas públicas na área da educação [Inep 2020c], por isso, o conhecimento dos fatores que mais impactam esse indicador pode direcioná-las a alcançar as metas estabelecidas pelo Ministério da Educação (MEC). Contudo, a qualidade da educação não pode ser resumida apenas em variáveis voltadas ao desempenho escolar. De acordo com [Chirinéa and Brandão 2015], o processo educativo é complexo, portanto, é preciso aliar os dados do desempenho estudantil com informações além das fronteiras escolares.

Diante desse cenário, cresce a necessidade de obtenção de informações de qualidade e, nesse sentido, a utilização da Mineração de Dados Educacionais (MDE) pode ganhar destaque. A MDE é uma área de pesquisa que utiliza técnicas de mineração de dados com o objetivo de descobrir informações que impactem a tomada de decisão no âmbito da educação [do Nascimento et al. 2018]. Ela vem se estabelecendo cada vez mais como uma linha de pesquisa, devido ao grande potencial de propiciar uma melhora na qualidade do ensino [Baker et al. 2011].

Levando em conta os aspectos citados, o presente trabalho pretende realizar um estudo para investigar as variáveis que influenciam a nota do IDEB para os estados de Alagoas e Bahia, separadamente, e também em conjunto envolvendo os dados de Sergipe. Para tanto, serão analisados os dados do Censo Escolar e SAEB de três categorias: informações socioeconômicas dos estudantes, estrutura física das escolas e condições de trabalho dos professores.

## 2. Trabalhos Relacionados

Nessa seção serão descritos trabalhos relacionados ao presente estudo. A pesquisa realizada em [da Silva Pinto et al. 2019] buscou identificar os fatores que influenciam o desempenho escolar dos alunos do 9º ano do ensino fundamental de escolas públicas da cidade de Teotônio Vilela- AL. Neste estudo, empregou-se algoritmos preditivos, incluindo NaiveBayes, J48, JRip, LibSVM, RandomForest, IBK, OneR e REPTree. Além disso, foram aplicadas técnicas de seleção de cada grupo de métodos, a saber: filtro, embrulhamento e incorporação.

No estudo realizado por [da Silva et al. 2020], técnicas de mineração de dados foram utilizadas com o objetivo de encontrar quais fatores das atividades dos Diretores de Escola que estão relacionados com o desempenho dos alunos do Ensino Médio. Para tanto, as avaliações do SAEB no ano de 2017 foram consideradas. Os achados do estudo apontaram que o tempo de experiência, nível de pós-graduação, ações para controle de reprovações e nível socioeconômico da escola interferem no desempenho escolar dos estudantes.

[Canedo et al. 2019] analisaram a relação da qualificação acadêmica dos professores com o IDEB das escolas. Esse estudo foi realizado com professores do Estado de Goiás. O algoritmo utilizado nessa pesquisa foi o priori, o qual apontou que o nível de pós-graduação dos docentes tem influência na nota final do IDEB.

Os autores em [Gusmão et al. 2021] investigaram o IDEB no estado de Sergipe utilizando Mineração de Dados. Os dados sobre a estrutura física de 96 escolas em adição a variáveis derivadas sobre os professores dessas escolas foram usados para investigar o IDEB. Quatro modelos de aprendizagem de máquina foram avaliados para prever se as escolas haviam atingido a meta do IDEB para o ano avaliado. Percebeu-se que houve uma diferença estatisticamente significativa entre o IDEB médio das escolas que não atingiram a meta e as escolas que atingiram. O modelo que obteve o melhor desempenho em termos de acurácia foi o Máquina de Vetor de Suporte.

[de Farias et al. 2023] realizaram um estudo envolvendo variáveis obtidas dos Microdados do Censo da Educação Básica e do Sistema de Avaliação da Educação Básica (SAEB) para investigar a previsão do IDEB em escolas de Sergipe. Seis algoritmos de aprendizagem de máquina supervisionada foram usados para prever se as escolas atingiram a meta do IDEB com base em três conjuntos de dados. Além disso, técnicas de seleção de atributos foram utilizadas para investigar a influência da redução das bases na previsão. Os algoritmos conseguiram maior acurácia na previsão para a base de dados com maior quantidade de atributos. O algoritmo de Floresta Aleatória foi o que apresentou maior acurácia.

O diferencial do presente estudo consiste na continuidade do que foi realizado por [de Farias et al. 2023] ao considerar também dados dos estados de Alagoas e Bahia usando variáveis contidas nos Microdados do Censo da Educação Básica e do Sistema de Avaliação da Educação Básica (SAEB). As variáveis analisadas pertencem a três categorias de informações: dados socioeconômicos dos alunos, estrutura física das escolas e atividade e capacitação profissional dos professores. Os dados deste trabalho são referentes ao terceiro ano do ensino médio de 2017.

## 3. Metodologia

Nessa seção, serão apresentadas as etapas de desenvolvimento do estudo. Para o desenvolvimento dos experimentos foi adotada a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*) [Wirth and Hipp 2000], esta é amplamente utilizada e fornece uma estrutura para condução de projetos de mineração de dados.

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 dos estados de Sergipe, Bahia e Alagoas. As suas fases do CRISP-DM serão descritas, a seguir.

### 3.1. Compreensão do Domínio

Nesta etapa, define-se o problema de Mineração de Dados. O objetivo desta pesquisa é investigar variáveis associadas a previsão do IDEB das escolas para turmas do terceiro ano do ensino médio dos estados de Sergipe, Alagoas e Bahia. As questões norteadoras são:

1. Quais as variáveis associadas ao IDEB identificadas nas análises?
2. Como a situação socioeconômica dos alunos influenciam o IDEB das escolas?
3. Como as diferenças entre as estruturas físicas das escolas influenciam o IDEB?
4. Como as condições de trabalho dos professores influenciam o IDEB das escolas?
5. Há diferenças entre as variáveis relevantes para os estados considerados?

### 3.2. Compreensão dos Dados

Para a formação das bases de dados utilizadas neste estudo, foram utilizadas as bases de dados do SAEB e do Censo Escolar, disponíveis no portal de microdados do INEP[1]. Foram utilizadas as informações relativas ao questionário dos alunos do 3º ano do ensino médio do SAEB e as amostras de dados de docentes, escolas e turmas do Censo Escolar.

O questionário do SAEB para o aluno é composto por 60 itens que abordam assuntos como nível socioeconômico, participação da família e atividades pedagógicas [Inep 2020b]. 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.

### 3.3. Preparação dos Dados

A preparação dos dados seguiu o mesmo procedimento realizado por [de Farias et al. 2023]. Inicialmente, os arquivos originais foram filtrados a partir dos estados de Sergipe, Bahia e Alagoas e a etapa seguinte consistiu na criação de três bases de dados. Na Tabela 1 podem ser vistas as quantidades de escolas por estado. As três bases foram criadas com objetivo de facilitar a identificação dos grupos de informações mais importantes para previsão do desempenho escolar através das notas do IDEB.

A primeira base, chamada de base A, é composta por variáveis socioeconômicas dos estudantes derivadas do questionário do SAEB e por outros atributos tamém presentes na base do SAEB relacionadas a localização das escolas; já a segunda, denominada de base B, é formada por atributos derivados de informações das escolas e professores provenientes do Censo Escolar, além de variáveis ligadas a localização, categorias e administração das escolas também presentes no Censo; e a última, intitulada de base C, representa a união das bases A e B. A quantidade de atributos e descrição das bases podem ser vistas na Tabela 2.

**Tabela 1. Quantidade de escolas por estado**

| Estado | Quantidade de Escolas |
|---|---|
| Sergipe | 77 |
| Bahia | 396 |
| Alagoas | 137 |

**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 |

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

Os atributos originais já estavam presentes nas bases e não requeriam manipulação adicional. As variáveis socioeconômicas derivadas do questionário do aluno do SAEB foram obtidas a partir do agrupamento a nível de escola, da quantidade de respostas para cada item das perguntas presentes no questionário. Já os atributos derivados do Censo Escolar foram elaborados a partir do agrupamento a nível de escola de cada coluna das tabelas referentes a professores e estrutura física das escolas. Na Tabela 3 está apresentado um exemplo de variável derivada.

**Tabela 3. Exemplo de criação das variáveis derivadas**

| **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 |

Outra etapa realizada foi a aplicação da técnica de discretização das notas do IDEB das instituições, coletadas manualmente no portal do IDEB[2]. A discretização de dados é o processo de transformar variáveis contínuas em variáveis discretas com a finalidade de simplificar sua análise [Dougherty et al. 1995]. Desse modo, através do cálculo da média do IDEB das instituições, foi adicionada nas três bases a variável "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: retirada de colunas com 60% ou mais dos valores nulos e preenchimento de valores nulos. Por fim, foram aplicadas técnicas de seleção de atributos [Han et al. 2011] nas três bases. Seu objetivo é encontrar as melhores variáveis como possíveis preditoras. Foram aplicadas quatro tipo de técnicas: filtragem, embaralhamento e embutida, além da aplicação do método merge. Segundo [Pinto et al. 2019], esses métodos podem ser explicados da seguinte maneira:

- Filtragem: A seleção é feita a partir de características dos dados. Foi utilizado o método SelecKBest da biblioteca *scikit-learn*[3], que seleciona k atributos com base nos maiores valores de alguma pontuação, nesse caso foi utilizada o valor F da análise de variância (ANOVA);
- Embaralhamento: Reconhece a seleção de atributos como um problema de busca. São criados vários subconjuntos de atributos que são testados em um determinado modelo de aprendizagem e o melhor deles é selecionado. Para tanto foi aplicado o algoritmo RFE também da biblioteca *scikit-learn* aliado SVC;
- Embutida: O processo de escolha dos melhores subconjuntos se dá a partir de algoritmos preditivos. Nesse caso, foi utilizado o método SelectFromModel da biblioteca *scikit-learn* aliado a Regressão Logística..

O método *merge*, desenvolvido por [de Lima 2016], deve ser aplicado após alguma abordagem de seleção de atributos. Ele consiste em identificar os atributos mais frequentes nos melhores subconjuntos de atributos selecionados, através de uma pontuação gerada para cada variável. Além disso, *Python* foi a linguagem de programação utilizada tanto nesta fase, quanto na condução dos experimentos.

[2]Portal IDEB: `http://ideb.inep.gov.br/`

[3]Biblioteca Scikit-learn: `https://scikit-learn.org/stable/`

### 3.4. Modelagem

Nessa fase foram realizados 3 experimentos, um para cada base. Os experimentos foram realizados em duas etapas: a primeira com as bases completas e a segunda com um subconjunto de variáveis filtradas a partir da aplicação da técnica de seleção de atributos. A métrica de avaliação utilizada foi a acurácia, esta 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 literatura da área. Assim, os seis algoritmos selecionados foram: Árvore de Decisão (AD), Florestas Aleatórias (FA), SVM, Naive Bayes (NB), K-vizinhos aleatórios (KNN) e Regressão Logística (RL). Foi utilizado o método Holdout [Han et al. 2011], onde 80% das amostras foram utilizadas para treinamento dos modelos de classificação e 20% para teste dos modelos criados.

Por fim, as variáveis presentes nos melhores modelos foram analisadas quanto ao seu nível de influência para previsão do IDEB. A identificação das variáveis mais influentes foi feita com o uso do algoritmo de floresta aleatória na base de dados com maior acurácia. A fase 5 está apresentada na seção seguinte.

## 4. Resultados

Nessa seção, serão apresentados os resultados obtidos nos experimentos realizados. A seguir, os resultados serão detalhados para os dados dos estados de Alagoas e Bahia de forma separada e, por fim, com os três estados combinados. Ressalta-se que os resultados para o estado de Sergipe isoladamente foram apresentados em [de Farias et al. 2023].

### 4.1. Alagoas

Os resultados podem ser observados na Tabela 4. Os melhores modelos foram desenvolvidos utilizando o subconjunto oriundo da base de dados C com método *merge*, a partir dos algoritmos de Floresta Aleatória, Árvore de Decisão, SVM e Regressão Logística, suas acurácias foram respectivamente: 92.37%, 90.68%, 91.53% e 91.53%.

**Tabela 4. Resultados das Análises na base de dados de Alagoas**

| **Algoritmos** | **Acurácia** | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sem Seleção de Atributos | | | Com Seleção de Atributos | | | | | | | | | |
| | | | | Filtragem | | | Embutida | | | Wrapper | | | Merge |
| | A | B | C | A | B | C | A | B | C | A | B | C | C |
| FA | 54.17 % | **83.33 %** | 50.0 % | 58.33 % | **87.5 %** | **92.37 %** | 54.17 % | 63.29 % | 62.5 % | 58.33 % | 58.33 % | 62.5 % | **92.37 %** |
| NB | 50.0 % | 50.0 % | 62.5 % | 54.17 % | 50.0 % | 61.02 % | 54.17 % | 63.29 % | 70.83 % | 62.5 % | 62.5 % | 66.67 % | 68.64 % |
| DT | 33.33 % | **90.0%** | 54.17 % | 37.5 % | **90.0%** | **89.83 %** | 41.67 % | 64.56 % | 54.17 % | 45.83 % | 45.83 % | 45.83 % | **90.68 %** |
| SVM | 41.67 % | 37.5 % | 41.67 % | 45.83 % | 37.5 % | **82.2 %** | 62.5 % | 69.62 % | 45.83 % | 58.33 % | 54.17 % | **75.0 %** | **91.53 %** |
| LR | 41.67 % | 45.83 % | 54.17 % | 58.33 % | 45.83 % | **80.51 %** | 62.5 % | 64.56 % | 54.17 % | 58.33 % | 50.0 % | 62.5 % | **91.53 %** |
| KNN | 50.0 % | 45.83 % | 50.0 % | 33.33 % | 45.83 % | 69.49 % | 54.17 % | 55.7 % | 62.5 % | 41.67 % | 37.5 % | 66.67 % | 70.34 % |

Na Tabela 5 é possível observar as variáveis mais importantes a partir do melhor modelo gerado. Observa-se a ocorrência de variáveis relacionadas às condições socioeconômicas dos estudantes, participação dos pais em reuniões escolares e grau de escolaridade da mãe.

### 4.2. Bahia

O estado da Bahia seguiu o mesmos padrões de resultados dos anteriores. Novamente, a filtragem e o método *merge*, utilizando a base de dados C foram as abordagens que

**Tabela 5. Principais variáveis associadas ao IDEB do Estado de Alagoas**

| Questão | Descrição |
|---|---|
| TX_RESP_Q026_C | Com qual frequência seus pais, ou responsáveis por você, vão à reunião de pais? R: Nunca ou quase nunca |
| TX_RESP_Q037_E | Em dias de aula, quanto tempo você gasta fazendo trabalhos domésticos (ex.: lavando louça, limpando o quintal etc.)? R: Não realizo |
| TX_RESP_Q008_C | Na sua casa tem geladeira? R: Duas |
| TX_RESP_Q036_E | Em dias de aula, quanto tempo você gasta assistindo à TV, navegando na internet ou jogando jogos eletrônicos? R: Não utiliza |
| TX_RESP_Q017_B | Em sua casa trabalha empregado(a) doméstico(a) pelo menos cinco dias por semana? R: Um |
| TX_RESP_Q046_A | O que você consulta para fazer o dever de casa de Língua Portuguesa? Jornais. R: Sim |
| TX_RESP_Q040_C | A partir da primeira série do Ensino Médio, em que tipo de escola você estudou? R: Em escola pública e em escola particular. |
| TX_RESP_Q019_G | Até que série sua mãe, ou a mulher responsável por você, estudou? R: Completou o Ensino Médio |
| TX_RESP_Q007_C | Na sua casa tem videocassete e/ou DVD? R: Duas |
| TX_RESP_Q036_C | Em dias de aula, quanto tempo você gasta assistindo à TV, navegando na internet ou jogando jogos eletrônicos? R: Entre 2 e 3 horas |

geraram os melhores modelos, conforme pode ser visto na Tabela 6. Os modelos desenvolvidos pelos algoritmos de floresta aleatória e árvore de decisão obtiveram as melhores taxas de acerto, sendo 93.38% e 91.35% na filtragem e 93.89% e 91.09% no *merge*. A partir da Tabela 7, é possível observar variáveis associadas ao IDEB referentes ao grau de formação e letramento dos pai e da mãe, condições socioeconômicas e hábitos de pesquisa na realização de atividades.

**Tabela 6. Resultados das Análises na base de dados da Bahia**

| **Algoritmos** | **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 | C |
| RF | 63.29 % | 67.09 % | 62.03 % | 62.03 % | 64.56 % | **93.38 %** | 60.76 % | 63.29 % | 64.56 % | 64.56 % | 65.82 % | 66.95 % | **93.89 %** |
| NB | 63.29 % | 60.76 % | 64.56 % | 64.56 % | 60.76 % | 63.36 % | 64.56 % | 63.29 % | 64.56 % | 65.82 % | 64.56 % | 59.32 % | 67.43 % |
| DT | 62.03 % | 59.49 % | 63.29 % | 50.63 % | 51.9 % | **91.35 %** | 58.23 % | 64.56 % | 55.7 % | 54.43 % | 53.16 % | 59.32 % | **91.09 %** |
| SVM | 64.56 % | 58.23 % | 63.29 % | 63.29 % | 58.23 % | 74.3 % | 70.89 % | 69.62 % | 68.35 % | 70.89 % | 69.62 % | 69.49 % | **79.9 %** |
| LR | 64.56 % | 63.29 % | 67.09 % | 65.82 % | 63.29 % | 74.55 % | 70.89 % | 64.56 % | 68.35 % | 69.62 % | 70.89 % | 68.64 % | **78.12 %** |
| KNN | 48.1 % | 58.23 % | 59.49 % | 55.7 % | 58.23 % | 70.74 % | 58.23 % | 55.7 % | 56.96 % | 62.03 % | 56.96 % | 58.47 % | 72.77 % |

## 4.3. Sergipe, Bahia e Alagoas

Na análise com os três estados, novamente os melhores modelos foram desenvolvidos com a base C, a partir dos algoritmos de floresta aleatória e árvore de decisão, com acurácias de 93.19 % e 90.63 %, para os modelos que utilizam filtragem, além de 94.38 % e 91.14 % nos modelos que aplicam o método *merge*. As variáveis que mais influenciaram no IDEB estão ligadas a informações socioeconômicas dos estudantes, conforme pode ser observado na Tabela 9.

## 4.4. Discussão

De maneira geral, pode-se perceber uma forte influência das variáveis socioeconômicas, sendo as mais importantes: a estrutura econômica familiar, grau de formação e letra-

**Tabela 7. Principais variáveis associadas ao IDEB do Estado de Bahia**

| Questão | Descrição |
|---|---|
| TX_RESP_Q020_B | Sua mãe, ou a mulher responsável por você, sabe ler e escrever? R: Não |
| TX_RESP_Q024_B | Seu pai, ou o homem responsável por você, sabe ler e escrever? R: Não |
| TX_RESP_Q043_A | Você concluiu o Ensino Fundamental na Educação de Jovens e Adultos(EJA), antigo supletivo? R: Sim |
| TX_RESP_Q055_A | O que você consulta para fazer o dever de casa de Matemática? Livros didáticos. R: Sim |
| TX_RESP_Q012_B | Na sua casa tem carro? R: Uma |
| TX_RESP_Q014_C | Na sua casa tem banheiro? R: Duas |
| TX_RESP_Q015_B | Na sua casa tem quartos para dormir? R: Uma |
| TX_RESP_Q019_A | Até que série sua mãe, ou a mulher responsável por você, estudou? R: Nunca estudou |
| TX_RESP_Q023_G | Até que série seu pai, ou o homem responsável por você, estudou? R: Completou o Ensino Médio |
| TX_RESP_Q018_B | Você mora com sua mãe? R: Não |

**Tabela 8. Resultados das Análises na base de dados com os três estados**

| **Algoritmos** | **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 | C |
| RF | 66.1 % | 63.56 % | 66.95 % | 66.1 % | 62.71 % | **93.19 %** | 64.41 % | 66.1 % | 72.03 % | 66.95 % | 65.25 % | 69.49 % | **94.38 %** |
| NB | 62.71 % | 61.86 % | 51.69 % | 66.95 % | 61.86 % | 63.88 % | 59.32 % | 68.64 % | 64.41 % | 61.86 % | 61.86 % | 64.41 % | 66.95 % |
| DT | 57.63 % | 54.24 % | 55.08 % | 56.78 % | 52.54 % | **90.63 %** | 49.15 % | 58.47 % | 58.47 % | 50.0 % | 44.92 % | 61.02 % | **91.14 %** |
| SVM | 69.49 % | 64.41 % | 65.25 % | 71.19 % | 64.41 % | 71.38 % | 65.25 % | 70.34 % | 72.88 % | 68.64 % | 68.64 % | 77.12 % | **76.66 %** |
| LR | 67.8 % | 66.95 % | 69.49 % | 70.34 % | 66.95 % | 72.74 % | 65.25 % | 72.88 % | 75.42 % | 67.8 % | 67.8 % | 78.81 % | **78.36 %** |
| KNN | 57.63 % | 62.71 % | 60.17 % | 65.25 % | 62.71 % | 71.21 % | 60.17 % | 61.02 % | 64.41 % | 60.17 % | 63.56 % | 66.1 % | 77.0 % |

mento dos pais e responsáveis, o seu acompanhamento e incentivo na vida estudantil e a participação dos estudantes nas atividades pedagógicas. Os resultados relatados são semelhantes ao que foi encontrado por [Menezes-Filho 2007], em que as variáveis que mais 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.

Além da família, a vivência acadêmica dos alunos também causou grande impacto nas escolas com notas abaixo da média, sobretudo o abandono (justamente com a evasão) e a reprovação. Segundo [Filho et al. 2017], o abandono - juntamente com a evasão - é uma das maiores fraquezas da educação no país e a situação socioeconômica dos estudante pode influenciar fortemente essa decisão. Ressalta-se também que foram apresentadas apenas as dez variáveis de maior influência para o melhor modelo dos experimentos. No entanto, outras variáveis relacionadas à estrutura física das escolas e professores também tiveram influência.

No que diz respeito ao grupo de atributos ligados as escolas, foi encontrada relação entre notas de IDEB acima da média e escolas que tem depedências, como biblioteca, sala de leitura, laboratório de informática, quadra de esportes, além de posse de equipamentos eletrônicos, a exemplo de computadores, retroprojetores e impressoras. Também foram

**Tabela 9. Principais variáveis associadas ao IDEB nos três estados**

| Questão | Descrição |
|---|---|
| TX_RESP_Q043_A | Você concluiu o Ensino Fundamental na Educação de Jovens e Adultos(EJA), antigo supletivo? Resposta: Sim |
| TX_RESP_Q042_C | Você já abandonou a escola durante o período de aulas e ficou fora da escola o resto do ano? Resposta: Duas vezes ou mais |
| TX_RESP_Q055_B | O que você consulta para fazer o dever de casa de Matemática? Livros didáticos. R: Não |
| TX_RESP_Q045_B | Você faz o dever de casa de Língua Portuguesa? Resposta: De vez em quando |
| TX_RESP_Q014_C | Na sua casa tem banheiro? Resposta: Duas |
| TX_RESP_Q055_A | O que você consulta para fazer o dever de casa de Matemática? Livros didáticos. R: Sim |
| TX_RESP_Q036_A | Em dias de aula, quanto tempo você gasta assistindo à TV navegando na internet ou jogando jogos eletrônicos? R: Menos de 1 hora. |
| TX_RESP_Q019_G | Até que série sua mãe, ou a mulher responsável por você, estudou? R: Completou o Ensino Médio |
| TX_RESP_Q013_A | Na sua casa tem computador? R: Não tem |
| TX_RESP_Q022_B | Você mora com seu pai? R: Não |

identificadas variáveis voltadas a serviços básicos, como água encanada e banheiros com chuveiro. O resultado corrobora o que foi encontrado por [Sátyro et al. 2007], que ressaltou a importância do acesso a serviços básicos e instalações educacionais adequadas para o alcance de um desempenho escolar satisfatório.

Em relação aos professores, apesar de não terem aparecido entre as 10 mais influentes, a quantidade de professores efetivos, quantidade de professores terceirizados e quantidade de professores com formação continuada de no mínimo 80 horas também estiveram entre associadas as notas do IDEB.

É importante também mencionar a Educação de Jovens e Adultos, visto que, os alunos que iriam concluir o ensino médio no EJA apresentam notas abaixo da média. O EJA é uma modalidade voltada ao público que não completou à educação na idade adequada e por isso, segundo [Ferreira et al. 2016] podem apresentar baixo desempenho em escrita e matemática, fato que pode explicar as notas abaixo da média.

Percebe-se que houve uma predominância das variáveis socioeconômicas, principalmente as voltadas ao ambiente familiar. Esse panorama indica que o desempenho escolar, de fato é uma síntese de múltiplas determinações e analisar o desempenho escolar apenas por testes padronizados, pode não ser capaz de contemplá-las, necessitando assim de estudos envolvendo outros setores da vida dos discentes conforme indica [Chirinéa and Brandão 2015].

## 5. Conclusões

O presente trabalho teve por objetivo realizar um estudo em busca das variáveis associadas ao IDEB. Foram utilizadas três bases de dados derivadas das informações do SAEB e Censo Escolar e a partir das análises dos resultados, pode-se perceber que vários aspectos socioeconômicos podem estar relacionados as notas do IDEB, dentre eles o envolvimento familiar nas atividades escolares e a formação acadêmica dos responsáveis. 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 também são importantes.

Para trabalhos futuros, pode-se realizar experimentos com bases de dados envolvendo outros estados e outros anos, a fim de realizar mais análises. Ademais, a inserção de novas variáveis, como voltadas a gestão escolar, adoção de políticas públicas também pode levar a resultados relevantes. Por fim, a adoção de outras abordagens de seleção de atributos, com o intuito de aumentar a acurácia.

## Referências


Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. Revista Brasileira de Informática na Educação, 19(02):03.

Canedo, E. D., De Carvalho, R. R., Leão, H. A. T., Costa, P. H. T., and Okimoto, M. V. (2019). How the academics qualification influence the students learning development. In 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), volume 1, pages 336–345. IEEE.

Chirinéa, A. M. and Brandão, C. d. F. (2015). O ideb como política de regulação do estado e legitimação da qualidade: em busca de significados. Ensaio: Avaliação e Políticas Públicas em Educação, 23:461–484.

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

da Silva Pinto, G., Júnior, O. d. G. F., and 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.

de Farias, M. L., Gusmão, C. S. D., and de Gusmão, R. P. d. G. (2023). Mineração de dados para investigar o ideb usando o censo da educação básica e saeb: um estudo de caso em sergipe. RENOTE, 21(1):No prelo.

de Lima, R. A. F. (2016). Estratégias de seleção de atributos para detecção de anomalias em transações eletrônicas.

do Nascimento, R. L. S., da Cruz Junior, G. G., and de Araújo Fagundes, R. A. (2018). Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. RENOTE, 16(1).

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In Machine learning proceedings 1995, pages 194–202. Elsevier.

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

Fernandes, R. and GREMAUD, A. P. (2009). Qualidade da educação: avaliação, indicadores e metas. Educação básica no Brasil: construindo o país do futuro. Rio de Janeiro: Elsevier, 1:213–238.

Ferreira, Aparecida, A., and de Cássia Martinelli, S. (2016). Estudantes da educação de jovens e adultos: considerações sobre o perfil e desempenho escolar. Educação: teoria e prática, 26(52):312–331.

Filho, S., Barbosa, R., de Lima Araújo, and Marcos, R. (2017). Evasão e abandono escolar na educação básica no brasil: fatores, causas e possíveis consequências. Educação por escrito, 8(1):35–48.

Gusmão, R., Gusmão, C., and Dias, M. (2021). A qualidade da educação para além do ideb: Um estudo através de técnicas de mineração de dados. In Anais do XXXII Simpósio Brasileiro de Informática na Educação, pages 803–812, Porto Alegre, RS, Brasil. SBC.

Han, J., Pei, J., and Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Inep (2020a). Censo escolar.

Inep (2020b). Testes e questionários.

Inep (2020c). Índice de desenvolvimento da educação básica (ideb) — inep. (Accessed on 11/05/2021).

Matos, P. F., Lombardi, L., Ciferri, R., Pardo, T., Ciferri, C., and Vieira, M. (2009). Relatório técnico "métricas de avaliaçao". Universidade Federal de Sao Carlos.

MEC (2009). Prova brasil - ministério da educação. (Accessed on 11/05/2021).

Menezes-Filho, N. A. (2007). Os determinantes do desempenho escolar do brasil.

Pinto, G. d. S. et al. (2019). 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.

Sátyro, N., Soares, S., et al. (2007). A infra-estrutura das escolas brasileiras de ensino fundamental: um estudo com base nos censos escolares de 1997 a 2005. Technical report.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, volume 1, pages 29–39. Manchester.