UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Glevson da Silva Pinto
glevsonsilva@ic.ufal.br

Dissertação de Mestrado

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

Novembro
2019

GLEVSON DA SILVA PINTO

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

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Curso de Mestrado em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Evandro de Barros Costa

Coorientador: Olival de Gusmão Freitas Júnior

Novembro
2019

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 - 1767

P659m Pinto, Glevson da Silva.
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 / Glevson da Silva Pinto. – 2020.
84 f. : il.

Orientador: Evandro de Barros Costa.
Co-orientador: Olival de Gusmão Freitas Júnior.
Dissertação (mestrado em Informática) - Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2019.

Bibliografia: f. 74-77.
Anexos: f. 78-84

1. Índice de Desenvolvimento da Educação Básica (Brasil. Ministério da Educação) - Alagoas. 2. Mineração de dados (Computação) - Educação. 3. Aprendizagem de máquina. I. Título.

CDU: 004.624

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PPGI
Instituto de Computação
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Folha de Aprovação

Glevson da Silva Pinto

"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"

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 18 de novembro de 2019.

Banca Examinadora:

Prof. Dr. Evandro de Barros Costa
UFAL – Programa de Pós-graduação em Informática
Orientador

Prof. Dr. Olival de Gusmão Freitas Júnior
UFAL – Programa de Pós-graduação em Informática
Coorientador

Prof. Dr. Thales Miranda de Almeida Vieira
UFAL – Programa de Pós-graduação em Informática
Examinador Interno

Prof. Dr. Rodrigo Lins Rodrigues
UFRPE – Universidade Federal Rural de Pernambuco
Examinador Externo

Dedicamos este trabalho a todos que acreditaram, e que ainda acreditam em nosso sucesso e na importância da computação na vida das pessoas e da sociedade como um todo.

AGRADECIMENTOS

A Deus por ter nos dado a vida e a força para chegarmos à universidade com coragem e fé.
A minha esposa Vanessa pelo carinho e atenção, por ter suportado minha ausência e meu estresse durante a jornada em busca dessa formação. Também por ter dado à luz ao meu bem mais precioso minha filha LIS, justamente no terceiro semestre do programa de pós-graduação.
As nossas famílias pelas motivações e apoio.
Aos colegiados da Secretaria de Educação de Teotônio Vilela, pois foram minha fonte de inspiração na busca por conhecimentos valiosos para essa dissertação e também a Noêmia Pereira por todo apoio dado para eu alcançar esse título.
Aos colegas do curso de mestrado em informática pelo compartilhamento das alegrias e momentos de estudos.
Ao meu orientador prof. Evandro pelo empenho e atenção.
Ao meu amigo prof. Olival Freitas pelo empenho e atenção durante toda essa pesquisa em que esteve ao meu lado junto ao Grupo de pesquisa Gestão do Conhecimento e da Inovação.
E a todos que direta ou indiretamente fizeram parte de nossa formação.

RESUMO

As transformações ocorridas na gestão e organização dos sistemas educacionais vêm demandando a tomada de decisões no sentido de melhorar o processo de ensino e aprendizagem nas instituições educacionais públicas do Brasil. Nesse sentido, o Ministério da Educação criou o Índice de Desenvolvimento da Educação Básica (IDEB), para avaliar e monitorar o processo educativo das escolas brasileiras. Neste contexto, particularmente, a Mineração de Dados Educacionais tem provido técnicas e, deste modo vem auxiliando educadores e gestores no apoio a tomada de decisões, permitindo extração de informações relevantes de bases de dados. Situado na mencionada área este trabalho objetiva explorar técnicas de seleção de atributos e algoritmos preditivos, visando identificar quais fatores impacta no IDEB das escolas municipais de Alagoas. Assim, visa-se auxiliar no processo decisório dos gestores educacionais, para posteriores estudos e reflexões na área da educação. Para tanto, utilizou-se dados do teste Saeb das escolas públicas de Maceió e de Teotônio Vilela, conduzindo um estudo experimental, produzindo relevantes resultados na tarefa de identificação de atributos relevantes para apoiar os gestores educacionais. Os resultados indicam que diversos fatores influenciam o desempenho dos alunos, tais como: a escolaridade dos pais do aluno, o incentivo aos estudos, o compromisso do docente e o estilo de gestão.

Palavras-chave: Mineração de dados educacionais, Seleção de atributos, Aprendizagem de máquina, IDEB.

ABSTRACT

The transformations that have occurred in the management and organization of educational systems have been demanding decisions to improve the teaching and learning process in public educational institutions in Brazil. In this sense, the Ministry of Education created the Basic Education Development Index (IDEB), to evaluate and monitor the educational process of Brazilian schools. In this context, in particular, Educational Data Mining has provided techniques and, thus, has been helping educators and managers to support decision making, enabling the extraction of relevant information from databases. Situated in the aforementioned area, this work aims to explore attribute selection techniques and predictive algorithms, aiming to identify which factors impact on the IDEB of Alagoas municipal schools. Thus, it aims to assist in the decision-making process of educational managers, for further studies and reflections in the area. of education. Therefore, data from the Saeb test of the public schools of Maceió and Teotônio Vilela were used, conducting an experimental study, producing relevant results in the task of identifying relevant attributes to support educational managers. The results indicate that several factors influence student performance, such as the student's parents' education, the incentive to study, the teacher's commitment and the management style.

Keywords: Educational data mining, Attribute selection, Machine learning, IDEB.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS

MDE - Educational Data Mining

EM - Expectation–Maximization

IBGE - Instituto Brasileiro de Geografia e Estatística

IDEB - Índice de Desenvolvimento da Educação Básica

INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira

KDD - Knowledge Discovery in Databases

MD - Mineração de Dados

MDE - Mineração de Dados Educacionais

SAEB - Sistema de Avaliação da Educação Básica

SEMED - Secretaria Municipal de Educação

SQL - Structured Query Language

UFRGS - Universidade Federal do Rio Grande do Sul

WEKA - Waikato Environment for knowledge Analysis

Sumário

1 INTRODUÇÃO

1.1 Justificativa da escolha do tema

O século 21 se caracteriza por ser a era do conhecimento em que as atividades humanas se organizam em torno da geração, disseminação, compartilhamento, recuperação e uso de ativos de informação e conhecimento. Esse novo contexto, proveniente da revolução tecnológica e do processo de globalização das relações econômicas e culturais, fez com que a educação fosse revisitada, em relação às suas finalidades e aos seus meios, com vistas a garantir a formação de pessoas capazes de enfrentar os múltiplos desafios dessa nova era.

É de interesse de toda sociedade que cada geração de cidadãos seja mais eficiente e qualificada que seus predecessores. O rendimento escolar é um fator determinístico para o sucesso profissional de cada aluno, abrindo oportunidades e expandindo seus horizontes. Diversas pesquisas vêm sendo realizadas para o monitoramento e a melhoria da aprendizagem em sala de aula; desta forma, estão ocorrendo avanços e aplicações de diversas técnicas pedagógicas nas escolas, o que está resultando no aperfeiçoamento das novas gerações.

Em 2007, o Ministério da Educação criou o Índice de Desenvolvimento da Educação Básica (IDEB) para avaliar o processo de ensino e aprendizagem nas escolas brasileiras. Esse índice tem sido influenciado por vários fatores educacionais, presentes nas escolas oriundos de avaliações sobre o aproveitamento escolar dos alunos, por meio do censo escolar e as médias de desempenho nas avaliações do Sistema de Avaliação da Educação Básica (SAEB), a Prova Brasil, a Avaliação Nacional de Alfabetização entre outras (INEP, 2019; INEP/MEC, 2007).

Dentro dessa perspectiva, há entre os países, em diversos âmbitos, a necessidade de efetuar ações que possibilitem a melhoria do ensino em todos os níveis, desde a alfabetização até os níveis mais altos de pós-graduação.

A avaliação em larga escala é um instrumento significativo para as atuais demandas sobre a qualidade do ensino e relevância da educação escolar, oferecendo subsídios para formulação, reformulação e monitoramento de políticas públicas de educação no Brasil, e também para a gestão da educação em nível de sistemas estaduais e municipais em suas respectivas escolas.

Essa premissa subsidia a afirmação de que a criação, o aprimoramento e a evolução das avaliações educacionais externas e as políticas públicas educacionais estão entrelaçadas, uma vez que, os índices resultantes destas avaliações, sejam elas nos parâmetros nacionais, como o IDEB, ou internacionais como o PISA (*Programme for Internacional Student Assessment* – Programa Internacional de Avaliação de Alunos), balizam e orientam as diretrizes governamentais.

Segundo INEP (2019), o Brasil gasta 5,8% do PIB em educação pública por ano. Esse índice é superior à média dos países da OCDE, que é de 5,5%, e acima de países como Chile (4,8%) e Estados Unidos (5,4%). Verifica-se assim, que a baixa qualidade da educação não se deve à insuficiência de recursos. Embora tenha havido aumento da oferta de vagas, a qualidade da educação básica brasileira ainda é muito precária, quando comparado internacionalmente.

Em 2015 no PISA, o Brasil ficou na 63^0 posição em ciências, na 59^0 em leitura e na 66^0 colocação em matemática entre 70 países. Em 2018 no PISA, o Brasil ficou na 66^0 posição em ciências, na 57^0 em leitura e na 70^0 colocação em matemática entre 79 países. Ou seja, o modelo de ensino adotado no Brasil se mostra pouco eficaz e ineficiente, comprometendo o desenvolvimento econômico futuro (INEP, 2019).

Diante desse sistema de avaliação de alta escala, percebe-se cada vez mais a necessidade de informações de qualidade por parte dos atores do ecossistema educacional (gestores, diretores, coordenadores pedagógicos, docentes e discentes), visando à efetividade na tomada de decisões no sentido de melhorar o processo de ensino e aprendizagem nas instituições educacionais (privadas e públicas) do Brasil.

Com o avanço tecnológico e o desenvolvimento da sociedade do conhecimento, as tecnologias da informação e comunicação, vêm produzindo novas formas de tratamento dos dados, tornando as informações produzidas por eles em conhecimento. E um dos processos atuais e importantes na produção do conhecimento é a mineração de dados.

A mineração de dados teve suas origens nos cursos de matemática, estatísticas e computação, mas surgiu como área de pesquisa e aplicação nos meados dos anos 90. Mas ganhou destaque a partir de 2011, quando teve evidências depois de ser cunhado o termo *Big Data* e com a publicação do relatório intitulado *Big Data: The Next Frontier for Innovation, Competition, and Productivity* pelo *Mckinsey Global Institute* (CASTRO E FERRARI, 2016).

Essa área ganhou um avanço muito grande no campo tecnológico, a capacidade de coletar dados e armazená-los têm superado a habilidade de analisar e extrair conhecimentos

destes. Para isto se faz necessário à aplicação de técnicas e ferramentas que transformem, de maneira inteligente e automática, os dados disponíveis em informações úteis, que representem conhecimento para uma tomada de decisão estratégica nos negócios e até no nosso cotidiano. Neste caso, muitos pesquisadores buscam estudar a análise dos dados e especificamente a mineração de dados.

A aplicação de algoritmo na mineração de dados é uma ferramenta capaz de acessar a informação e extrair conhecimento e a possível tomada de decisões estratégicas dentro das organizações. Essa área é também conhecida como análise inteligente de dados, que se foca no uso de técnicas de aprendizagem de máquina.

A mineração de dados emerge como um processo sistemático, interativo e iterativo, de preparação e extração de conhecimentos a partir de grandes bases de dados. De acordo com Castro e Ferrari, conceitua-se:

> O termo mineração de dados (MD) foi cunhado como alusão ao processo de mineração (...) uma vez que se explora uma base de dados (mina) usando algoritmos (ferramentas) adequados para obter conhecimento (minerais preciosos). Os dados são símbolos ou signos não estruturados, sem significados, como valores em uma tabela, e a informação está contida nas descrições, agregando significado e utilidade aos dados, como o valor da temperatura do ar. Por fim, o conhecimento é algo que permite uma tomada de decisão para a agregação de valor, então, por exemplo, saber, que vai chover no fim de semana pode influenciar sua decisão de viajar ou não para a praia (2016, p.1).

Na atualidade vem sendo utilizada e aplicada na área da educação para tratamento dos dados educacionais, visando à melhoria na gestão educacional, na organização do trabalho pedagógico e na melhoria da qualidade do ensino e da aprendizagem.

O Ministério da Educação criou o Índice de Desenvolvimento da Educação Básica (IDEB), para avaliar o processo de ensino e aprendizagem nas escolas brasileiras. Viu-se que este índice tem sido influenciado por vários fatores educacionais, presentes nas escolas oriundos de avaliações sobre o aproveitamento escolar dos alunos, através do censo escolar e as médias de desempenho nas avaliações do Sistema de Avaliação da Educação Básica, a Prova Brasil, a Avaliação Nacional de Alfabetização entre outras.

Pensando neste índice, resolveu-se utilizar a mineração de dados para analisar o IDEB das escolas municipais de Alagoas e trabalhar com diferentes atributos com a utilização de algoritmos de seleção de atributos e algoritmos de mineração de dados que podem fornecer resultados significativos para a melhoria da gestão nas escolas públicas de Alagoas.

1.2 Questão de pesquisa

Este trabalho procura responder à questão: Como a mineração de dados educacionais, a partir de um conjunto de atributos da base de dados do IDEB, pode contribuir na reflexão da melhoria da gestão e do processo de ensino e aprendizagem das escolas públicas de Alagoas?

1.3 Objetivos

Objetivo Geral:

Desenvolver um modelo de análise e predição para identificação dos fatores que possam influenciar positivamente o desempenho escolar na rede de ensino básico, visando à melhoria do IDEB.

Objetivos Específicos:

- Analisar o conjunto de dados do IDEB a partir dos diversos atributos retirados do INEP para a utilização de diversas abordagens de seleção de atributos, para posteriores estudos e reflexões na área da educação;
- Analisar os dados educacionais das escolas públicas de Maceió e de Teotônio Vilela que compõem o IDEB, através da utilização de diversos algoritmos de aprendizagem de máquina;
- Avaliar o desempenho dos algoritmos de classificação na predição do IDEB com base nos atributos do questionário SAEB-2015;
- Desenvolver um preditor para classificar um aluno como IDEB "Satisfatório" ou não, onde seja possível visualizar os atributos que estão influenciando na predição.
- Avaliar o desempenho dos métodos de seleção de atributos em relação ao IDEB;
- Avaliar o desempenho dos algoritmos de classificação na predição do IDEB com base nos atributos do questionário SAEB-2015;
- Analisar os dados educacionais das escolas públicas de Maceió e de Teotônio Vilela que compõem o IDEB, através da utilização de diversos algoritmos de aprendizagem de máquina.

1.4 Principais Contribuições da Pesquisa

Destaca-se como principais contribuições deste trabalho:

1. Investigação da eficácia de métodos tradicionais de seleção de atributos na área da educação básica;
2. Utilização do método Merge para combinar os atributos mais frequentes nos melhores conjuntos de seleção, gerando um score.
3. Utilização do teste estatístico de Friedman e Namenyi para representar a distância critica entre os algoritmos, levando-se em conta o número de interações, particionamento e acurácia de um conjunto de dados;
4. Construção de modelo utilizando técnicas de seleção de atributos adequadas entre as classes e técnicas de classificação;
5. Realização de estudo de casos com dados reais;
6. Publicações de artigos em veículos científicos, que será citado nas referências.

1.5 Metodologia de pesquisa

Trata-se de uma pesquisa de cunho quantitativa e exploratória. Do ponto de vista dos procedimentos técnicos, trata-se de um estudo de caso, analisando-se dados educacionais das escolas públicas municipais de Maceió e de Teotônio Vilela, a partir de uma pesquisa no portal do INEP. Convém ressaltar que este portal apresenta os dados educacionais de diversos anos, mas com foco no Índice de Desenvolvimento da Educação Básica das instituições que realizaram a Prova Brasil.

Neste trabalho, utilizam-se apenas os dados obtidos nos anos de 2015 e 2017 relativos aos alunos dos anos finais do ensino fundamental (9º ano) das escolas municipais da cidade de Maceió e de Teotônio Vilela.

Conforme observa-se na **Figura 1**, na primeira etapa busca-se quais são os atributos que influenciam no IDEB, partir dos questionários aplicados aos alunos e dos resultados da Prova Brasil. Para isso, na segunda etapa da metodologia utilizam-se diversas técnicas de seleção de atributos, visando identificar quais fatores impactam no IDEB das escolas municipais de Maceió e de Teotônio Vilela. A seleção de atributos tem como objetivo excluir atributos redundantes e que não são úteis para a criação do modelo de predição. Para tanto, foram utilizados algoritmos de cada grupo de algoritmos de seleção.

A terceira etapa consiste em utilizar o método Merge que combina os atributos mais frequentes nos melhores conjuntos de seleção. A partir desse ranking, utiliza-se como

estratégia de corte, selecionar um subconjunto de atributos com frequência superior a dois. Na quarta etapa foram utilizados diversos algoritmos de classificação para analisar a precisão dos dados selecionados aplicados ao nono ano nas matérias de português e matemática. Na quinta etapa aplica-se o teste estatístico de Friedman e Namenyi para mostrar que os algoritmos de classificação são diferentes estatisticamente e quais deles apresentam melhores eficácias.

Com objetivo de extrair informações relevantes para esse trabalho no contexto da Mineração de dados Educacionais foi necessário a utilização de softwares para apoio a análise de dados para realização da pesquisa de forma empírica, já que não há necessidade de implementação dos algoritmos e há diversos deles a serem testados de forma rápida e eficiente, sendo necessário apenas conhecer em que categoria da mineração de dados o problema se enquadra. No caso desta dissertação, utilizou-se diversas ferramentas de apoio, entre elas: a linguagem R, Weka e Anaconda.

Figura 1 - Metodologia de pesquisa.

Fonte: Elaborado pelo autor

1.6 Estrutura do trabalho

O presente trabalho está organizado da seguinte forma: o **capítulo 2** abordará a fundamentação teórica bem como os trabalhos relacionados a esta temática, tentando mostrar a originalidade do presente trabalho. O **capítulo 3** tratará da aplicação da metodologia CRISP-DM adaptada a nossa proposta, destacando as suas seis fases, mas enfatizando mais a que endereça mais diretamente a identificação das variáveis mais relevantes. O **capítulo 4** apresenta as considerações finais obtidas com este trabalho.

2 REVISÃO DA LITERATURA

2.1 Mineração de dados educacionais

Com os avanços na tecnologia e nos meios de comunicação, o crescimento exponencial dos dados fez com que os conceitos como *Big Data* e *Data Mining* surgissem e fossem intensamente explorados. Devido à imensa quantidade de dados disponíveis fizeram-se necessárias soluções para estruturar e avaliar todas essas informações. Surgiu assim um processo para descobrir e entender estes dados, o *Knowledge Discovery in Databases* (KDD) ou Descoberta de Conhecimento em Banco de Dados. O KDD representa uma metodologia onde a descoberta de informações pode ser definida por uma sequência de passos, sendo eles: limpeza de dados, integração dos dados, seleção dos dados, transformação dos dados, mineração dos dados, avaliação dos padrões e apresentação e assimilação do conhecimento (HAN E KAMBER, 2001).

A Mineração de Dados também se relaciona com diversas áreas, sendo possível perceber que as aplicações possuem um caráter multidisciplinar, podendo ser utilizada em diversos campos: educação, marketing, antropologia, judiciário, saúde, economia, política, ações contra o terrorismo, etc. Geralmente, as aplicações podem ser divididas entre (TAN *et al.*., 2009):

- Tarefas preditivas: prevê o valor de um atributo específico com base nos valores dos demais atributos. O atributo a ser previsto é normalmente conhecido como alvo, variável preditiva ou dependente, enquanto os demais, como variáveis preditoras, explicativas ou independentes.
- Tarefas descritivas: deriva padrões (associações, correlações, tendências, agrupamentos, trajetórias e anomalias) que resumem as relações subjacentes aos dados. Ou seja, consiste na identificação de características intrínsecas do conjunto de dados, definidas por padrões compreensíveis por humanos. As tarefas descritivas são de natureza exploratória e, frequentemente, necessitam de técnicas de pós-processamento para validar e explicar os resultados.

Um passo necessário para a realização da Mineração de Dados é o pré-processamento dos dados. Após o pré-processamento, os métodos de Mineração de Dados poderão ser

aplicados. Os padrões descobertos pelos métodos são extraídos e pode haver a necessidade da aplicação de um pós-processamento (Han e Kamber, 2001), onde ocorrerá a avaliação dos padrões encontrados e a apresentação do conhecimento no qual se utiliza técnicas de visualização e representação de conhecimento para apresentar o conhecimento extraído. Han e Kamber (2001) considera a Mineração de Dados como parte de um processo iterativo, que ocorre entre as etapas de pré-processamento e pós-processamento, podendo ser realizada através da interação com o usuário ou com uma base de dados.

A Mineração de Dados faz parte do processo de KDD e pode ser definida como a procura por identificar relações entre dados e, a partir desse processo, criar informações que podem gerar novos conhecimentos (PATRICIO E MAGNONI, 2018). De acordo com Weis e Indurkhya (1999):

> Mineração de dados é a busca de informações valiosas em grandes bancos de dados. É um esforço de cooperação entre homens e computadores. Os homens projetam bancos de dados, descrevem problemas e definem seus objetivos. Os computadores verificam dados e procuram padrões que casem com as metas estabelecidas pelos homens.

A área da Mineração de Dados surge para descobrir padrões e gerar conhecimento quando o homem não possui a capacidade de processar a enorme quantidade de dados existente nos dias atuais; e principalmente com a possibilidade de gerar hipóteses de forma automática a partir desses dados.

Segundo Costa *et al.* (2012), Mineração de Dados Educacionais (MDE) procura desenvolver ou adaptar métodos e algoritmos de mineração existentes de tal modo que se prestem a compreender melhor os dados em contextos educacionais, produzidos principalmente por estudantes e professores, considerando os ambientes nos quais eles interagem, tais como: AVAs, Sistemas Tutores Inteligentes (STIs), entre outros. Com tais métodos visa-se, por exemplo, entender melhor o estudante no seu processo de aprendizagem, analisando-se sua interação como ambiente. Assim há a necessidade por exemplo, de adequação dos algoritmos de mineração de dados existentes para lidar com especificidades inerentes aos dados educacionais, tais como a não independência estatística e a hierarquia dos dados.

De acordo Costa *et al.* (2012) apesar de algumas iniciativas primeiras com workshops específicos dentro das conferências sobre *Artificial Intelligence in Education* (AIEd) e sobre *Intelligent Tutoring Systems* (ITS), foi somente em 2005, em Pittsburgh, EUA, que foi organizado o primeiro *Workshop on Educational Data Mining*, como parte do 20th *National*

Conference on Artificial Intelligence (AAAI 2005). Daí em diante, houve mais algumas realizações deste workshop entre 2006 e 2007. Seguindo-se, em 2008 lança-se, em Montreal, Canadá, a primeira conferência em MDE: *First International Conference on Educational Data Mining*, evento este que se estabeleceu e ganhou regularidade de realização anual, estando agora em 2012 na sua décima segunda edição. Em 2009, esta sociedade investiu na criação de um periódico e publicou o seu primeiro volume do JMDE - Journal of Educational Data Mining. Em 2011 constituiu-se a sociedade científica para MDE (*International Educational Data Mining Society*). Quanto ao Brasil, de acordo com Baker et al. (2011), os primeiros trabalhos publicados ocorreram por volta de 2006 e mesmo assim, até hoje, o número de publicações nos principais periódicos é mínimo.

Segundo Costa *et al.* (2012) há diversas tarefas envolvidas em MDE, notadamente as que decorrem diretamente da análise de dados gerados nas interações dos estudantes com os ambientes de aprendizagem. Dessa análise surgem demandas para responder questões relacionadas a como melhorar a aprendizagem do estudante, como desenvolver ambientes educacionais mais eficazes que contribuam efetivamente para os estudantes aprenderem mais e em menos tempo? Em outra perspectiva, pretende-se saber quais métodos de mineração de se adequam às necessidades presentes na área de MDE? Quais ajustes devem ser feitos nas técnicas de forma a suprir a necessidade de MDE? Do ponto de vista computacional, alguns desafios práticos que se apresentam em vários contextos educacionais estão relacionados, por exemplo, a falta de padronização dos dados, o que acaba exigindo grande esforço de pré-processamento. Além disso, há a necessidade de adequação dos algoritmos clássicos de mineração de dados para lidar com especificidades inerentes aos dados educacionais, tais como a não independência estatística e a hierarquia dos dados (BAKER, 2011).

Entre as tarefas e métodos de mineração de dados educacionais a serem discutidos no presente texto, incluem-se: classificação e regressão, agrupamento de dados, mineração de regras de associação. Outras abordagens, entretanto, serão comentadas.

A tarefa de classificação diz respeito ao processo de encontrar um modelo que descreve e distingue classes de dados ou conceitos. Os modelos são derivados com base nas análises de coleções de dados, denominadas conjuntos de treinamentos, os quais correspondem a objetos de dados para os quais os rótulos de classes são conhecidos. O modelo é usado para prediz erro rótulo de classe de objetos para os quais o rótulo de classe é desconhecido. Ele associa um item de dado a uma ou várias classes predefinidas. Os modelos derivam dos podem ser representados sem várias formas, tais como: árvore de decisão, regras

de classificação, funções matemáticas, redes neurais (HAN AND KAMBER, 2000). Enquanto na classificação a predição é feita para um atributo classificador que assume valores discretos, em modelos de regressão a variável alvo é continua, ou seja, associa um item de dado a uma ou mais variáveis de predição de valores reais. Por sua vez, a análise de agrupamento de dados procura associar um item de dado com um ou vários agrupamentos determinados pelos dados, valendo-se principalmente de medidas de similaridades. Já a abordagem de mineração de regras de associação busca encontrar possíveis relações interessantes entre atributos de uma base de dados. Estas abordagens serão discutidas em detalhes na próxima seção deste capítulo, onde se discutem outras abordagens.

A partir dos métodos tradicionais de análise de dados, a mineração de dados abriu oportunidades excitantes para explorar novos tipos de dados e para analisar antigos tipos de dados de novas maneiras. Assim, a mineração de dados educacionais surge com o recente aumento dos cursos à distância e do suporte computacional aos cursos presenciais.

Os pesquisadores da área da educação descobriram que os métodos de mineração de dados precisam ser constantemente modificados para atender às particularidades dos dados educacionais (BAKER, 2010; ROMERO E VENTURA, 2013):

- Há uma variedade de contextos educacionais, onde os dados podem, por exemplo, ser organizados em termos de estrutura do material de aprendizagem (habilidades, problemas, unidades, aulas) e da estrutura de contexto de aprendizagem (alunos, professores, pares de colaboração, classes e escolas) (COSTA et al., 2012).
- A variedade de níveis hierárquicos dos dados sugere que a informação pode estar entre vários níveis.
- Existe uma constante falta de independência estatística entre os dados educacionais, provavelmente pela variedade de níveis hierárquicos. A estatística aponta que dados não independentes devem ser analisados em conjunto (VICINI E SOUZA, 2005). Dizer que um dado é dependente significa que ele pode ser usado para inferir informação sobre outro. A estatística também aponta que a falta de independência pode ocasionar interpretações equivocadas nas análises de inferência estatística.

O principal objetivo da MDE é antigo e estudado há muito tempo, que é conseguir compreender como ocorre o processo da aprendizagem. A diferença é que agora os pesquisadores possuem uma grande escala de dados e conseguem analisar uma aprendizagem prática e real (sem a antiga necessidade de realizar experimentos para obter dados). Sendo possível, por exemplo, observar estudantes em um curso durante oito meses e descobrir quais

atividades geram melhor aprendizado a longo prazo, verificar o desempenho do aluno em sala de aula de acordo com o tempo que ele inicia a tarefa de casa ou se revisões do conteúdo visto em sala trazem algum benefício ou não (ROMERO et al., 2011).

A MDE pode ser aplicada para resolver muitas tarefas (Romero e Ventura, 2010; Romero e Ventura, 2013), dentre as quais destacam-se:

- Criação de alertas. Como forma de comunicação aos interessados no processo de aprendizagem, auxiliando gestores e educadores na tomada de decisão, através da análise das atividades realizadas pelos alunos e informações de uso dos recursos/materiais do curso, por exemplo. Os alertas podem ser sobre comportamentos indesejáveis, como baixa motivação, uso indevido dos recursos, probabilidade de evasão, etc. As técnicas mais frequentes são mineração de processo e análise exploratória dos dados, através de análise estatística e visualizações/relatórios.
- Manutenção e melhoria dos cursos. Envolve tarefas de pesquisa científica, construção de material didático, planejamento e programação, onde se deve analisar como ocorre a aprendizagem do aluno, verificando, por exemplo, o que foi e como foi utilizado. Sendo importante a realização de testes de teorias sobre a aprendizagem baseada em novas tecnologias, para a formulação de novas hipóteses científicas, que poderão apoiar a construção ou adequação dos materiais didáticos, o planejamento de futuros cursos, quais disciplinas um aluno deve cursar em um período, alocação de recursos, etc. As técnicas mais utilizadas são associação, agrupamento e classificação.
- Geração de recomendação. Verifica as necessidades do aluno em um dado momento, gerando uma recomendação, que pode lhe proporcionar aprofundamento em um determinado domínio ou auxiliá-lo em uma dúvida. As recomendações podem ser links para visitas, dicas ou a resolução para um problema, a indicação de algum recurso ou material, um curso que pode ser feito, etc. A maioria das técnicas para esta tarefa envolvem associação, sequenciamento, classificação e agrupamento.
- Previsão de notas e resultados de aprendizagem. Essa tarefa é considerada como a mais antiga e popular da MDE. Consiste em utilizar os dados de atividades do curso para prever as notas finais do aluno ou algum outro tipo de resultado de aprendizagem, como uma possível evasão ou futura capacidade de aprender algo. As técnicas utilizadas podem ser regressão linear, classificação, agrupamento e associação.
- Criação de perfis de alunos. Utilizada para detectar o estado e as características dos estudantes, como a satisfação, motivação, progresso de aprendizagem, estilo de aprendizagem, preferências e assim por diante. Para a criação desses perfis, também é

considerado alguns problemas que possam impactar negativamente nos resultados de aprendizagem, como a apresentação de muitos erros na realização de uma tarefa, a má utilização ou subutilização de tutores inteligentes, manipulação dos sistemas, exploração dos recursos de forma ineficiente, etc. As técnicas mais frequentes além de agrupamento, classificação e análise de associação, são análises estatísticas, redes Bayesianas, modelos psicométricos e aprendizado por reforço.

- Análise da estrutura de domínio. Realizada para determinar a qualidade do conteúdo apresentado e a sequência em que ele foi dado, através da previsão do desempenho dos alunos, descrevendo o domínio de instrução em termos de conceitos, habilidades, itens de aprendizagem e suas inter-relações. As técnicas mais utilizadas são regras de associação, agrupamento e algoritmos *spacesearching*.

Em algumas aplicações, as tarefas apresentadas podem está relacionadas entre si, como: (i) geração de recomendação de acordo com a criação do perfil do aluno, (ii) previsão de notas e a criação de alertas, com as possíveis evasões, (iii) manutenção e melhoria dos cursos, através dos perfis dos alunos, da previsão de notas e da estrutura de domínio; etc. Essas possibilidades de relacionamento, também justificam a segregação da mineração de dados como MDE e caracterizam o quão complexa e desafiante essa área de pesquisa.

2.2 Seleção de Atributos

Ao todo existem 57 questões no questionário da Prova Brasil. O objetivo desta etapa será aplicar técnicas para reduzir ao máximo esse número de questões. O conjunto inicial de questões é chamado modelo completo, enquanto o conjunto após a redução é chamado de modelo reduzido.

A seleção de atributos pode ser entendida como a tentativa de selecionar o menor subconjunto de atributos de forma que a precisão dos classificadores não seja significativamente reduzida (DASH E LIU, 2003). Em muitos bancos de dados existem atributos que são irrelevantes ou redundantes, sendo assim, faz-se necessário o uso de técnicas de seleção de atributos. Dentre os benefícios destas técnicas está à facilitação da visualização de dados, a compreensão dos mesmos, a redução dos requisitos computacionais para a análise dos dados e a redução de tempo de treinamento e resposta dos algoritmos aplicados. Além disso, a redução de atributos também pode aumentar a precisão dos algoritmos de predição (GUYON E ELISSEEFF, 2003).

Técnicas de seleção de atributos, também conhecidas como seleção de *features*, podem ser aplicadas de várias formas. Os principais métodos serão utilizados neste trabalho a fim de reduzir o modelo completo de 57 questões. Tais métodos são descritos a seguir.

2.2.1 Filtragem

Métodos de filtragem analisam características intrínsecas dos dados, tais como: informação, consistência, dependência, entre outros (DASH E LIU, 2003). Porém, esses métodos são suscetíveis a seleção de atributos redundantes. Exemplos de algoritmos de filtragem são:

Teste Qui-quandrado: Esse método avalia os atributos individualmente usando a medida x^2 com relação à classe. Quanto maior o valor de x^2, mais provável é a correlação das variáveis (atributo e classe).

Ganho de Informação: Ranqueia os atributos através do ganho de informação. Avalia por meio do conjunto de testes (ou validação cruzada). Nesse processo ao algoritmo realiza a redução da entropia, porém tem uma desvantagem essa abordagem se torna impraticável quando o número de atributos é muito grande.

Seleção baseada em Correlação (CSF): É um método em que um subconjunto de atributos é selecionado através da alta relação com a classe. Onde um atributo individual é verificado também em relação aos demais atributos.

2.2.2 Embaralhamento

Métodos de embaralhamento (do inglês, *Wrapper methods*) consideram a seleção como um problema de busca, no qual avaliam o modelo preparando vários subconjuntos de atributos e comparam os resultados das combinações. Exemplos desse método são: Wrapper com o NaiveBayes, *Forward* e *Backward Selection*.

2.2.3 Embutida

O método embutida consiste na busca por encontrar o conjunto de atributos ótimo enquanto constrói um modelo preditivo. Seu uso é menos custoso computacionalmente do que os métodos de empacotamento, porém são mais complexos conceitualmente e modificações no algoritmo de classificação podem resultar em uma fraca performance. Um exemplo para esta técnica é o método *Recursive Feature Elimination*.

2.3 Validação cruzada (Cross-validation)

A validação cruzada *(cross-validation)* é uma técnica amplamente usada em algoritmos de seleção de atributos. Esta técnica consiste na divisão dos dados em subgrupos, onde parte destes subgrupos são destinados a amostra de treinamento do algoritmo de seleção ou predição utilizado, enquanto o restante dos subgrupos é usado como amostras de testes, ou também chamados de amostras de validação, e são utilizados para estimar o erro do algoritmo. Então, a validação cruzada seleciona o(s) grupo(s) de menor risco. Esse método evita o sobre ajuste de dados (do inglês, *overfitting*) devido ao fato dos dados de treinamento serem independentes dos dados de validação (WITTEN et al., 2011).

Com o objetivo de reduzir qualquer viés que esteja presente nos algoritmos aplicados foi usado o método *K-Fold Cross-Validation*. Esta técnica divide aleatoriamente os dados em k subconjuntos de tamanho aproximadamente iguais, onde k-1 subconjuntos são usados no treinamento do algoritmo e a estimativa de erro é calculada usando o subconjunto restante que aplicado como amostra de teste. Em seguida os resultados de cada grupo são comparados entre eles, sendo possível a visualização da performance média dos algoritmos (RODRIGUEZ et al., 2010). De acordo com Wong (2015) existem 4 fatores que impactam a precisão do método *K-Fold Cross-Validation*: o número de partições (*folds*); o número de instâncias em cada partição; a distribuição de dados por partição e a quantidade de repetição.

Para este trabalho foi utilizado o método com dez grupos, ou seja, *10-fold Cross-validation*, que para encurtamento será chamado de *10-fold*, executado 30 vezes.

2.4 Técnicas de Classificação de Mineração de Dados Educacionais

Segundo Costa *et al.* (2012), grande parte das técnicas utilizadas na área de MDE são provindas da área de mineração de dados. Entretanto, na maioria das vezes há a necessidade de adaptá-las devido às particularidades existentes em ambientes educacionais e seus dados. As técnicas estão apresentadas conforme sua categorização nas subáreas de MDE, seguindo-se o que consta na taxonomia proposta por (Baker *et. al*., 2011), tal como segue.

2.4.1 Classificadores Baseados em Regras

Esse tipo de técnica deve expor o conhecimento "escondido" nos dados, mostrando possíveis inconsistências, evitando desta forma resultados que não podem ser previstos. Este

tipo de técnica é eficaz em situações que não é necessário um grande conjunto de regras para descrever os dados (DUCH, 2013).

O classificador baseado em regra, também conhecido como método separar e conquistar é um processo interativo no qual são criadas regras do tipo SE-ENTÃO (IF-THEN) que definem um subconjunto de amostras de treinamento. Após essas regras serem definidas, o classificador é limpo e então uma nova interação ocorre preenchendo o classificador com outro subconjunto que também irá gerar novas regras até não haver mais subconjuntos (MAHAJAN E GANPATI, 2014).

No modelo de regras SE-ENTÃO gerados, as condições *SE* são conhecidas como regras antecedentes ou pré-condições e as regras *ENTÃO* são chamadas de regras consequentes.

2.4.2 Árvores de Decisão

A árvore de decisão é um modelo estatístico que utiliza treinamento supervisionado para construir uma estrutura de árvore, onde cada nó interno pode ser entendido como um teste de atributo e as folhas da árvore representam um determinado rótulo. A árvore irá classificar a instância de acordo com o caminho percorrido na árvore que termina em um nó folha, atribuindo a instância o rótulo do mesmo (COSTA et al., 2012).

Classificadores de árvores de decisão quebram problemas com um conjunto de regras complexas em um modelo multicamada (estrutura de árvore) que utiliza a união de regras mais simples (SAFAVIAN E LANDGREBE, 1991).

Uma das vantagens de modelos baseados em árvore é a fácil visualização e compreensão de dados. Muitas ferramentas, como Weka, que é utilizado neste trabalho, contém a funcionalidade de gerar diagramas visuais de árvores geradas pelos algoritmos.

2.4.3 Redes Bayesianas

Redes Bayesianas são classificadores estatísticos que podem predizer a relação entre classes. Essa rede produz um modelo casual das relações, o qual é usado para a aprendizagem e para a classificação (HAN E KAMBER, 2001).

O modelo construído na Rede Bayesiana pode ser descrito como um grafo direcionado acíclico que permite a representação eficiente da distribuição de probabilidade de junção, dado um conjunto de variáveis. Cada vértice representa uma variável e cada aresta representa a correlação direta entre as variáveis (FRIMDEAN *et al.*, 1997).

Essas redes são modelos que trabalham com conhecimento incerto ou incompleto através do Teorema de Bayes, criado pelo matemático Thomas Bayes em 1973, que define a probabilidade de um evento dado em um conhecimento prévio. Tal teorema é conhecido por sua famosa fórmula:

$$P(A \backslash B) = \frac{P(B|A).P(A)}{P(B)}$$

onde $P(A)$ significa a probabilidade de A ser verdade e $P(A|B)$ a probabilidade de A ser verdade caso B seja verdadeiro.

2.4.4 Máquina de Vetores de Suporte

Máquina de vetores de suporte (*Support Vector Machine* ou SVM) é um algoritmo supervisionado para classificação que procura separar instâncias de rótulos diferentes usando um hiperplano, o qual é descoberto usando vetores de suporte (COSTA et al., 2012).

A utilização de SVM tem se tornado popular, pois tem demonstrado bons resultados quando comparado a outros algoritmos de aprendizagem em diversas áreas, tais como, a detecção de faces em imagens e a categorização de textos. Lorena e Carvalho (2002) descrevem as principais vantagens da SVM: boa capacidade de generalização; robustez diante de objetos com grandes dimensões, como imagens; convexidade da função objetivo, pois contém apenas um mínimo local para funções quadráticas e teoria bem definida dentro da matemática e estatística.

Todavia, Zeng et al. (2008) apresentam um desafio para as SVM que é o espaço em memória para problemas quadráticos com um grande conjunto de dados de treinamento, pois este espaço de memória aumenta em $O(n^2)$, onde n é o número de dados de treinamento.

2.5 Teste estatístico de Friedman e Namenyi

O teste de Friedman é um teste não paramétrico equivalente ao teste ANOVA (Fisher, 1959) para medições repetidas na versão não paramétrica, o método consiste em classificar os dados por linhas ou blocos, substituindo-os em sua respectiva ordem. Ao fazer o pedido, devemos considerar a existência de dados idênticos. Os algoritmos k são ordenados separadamente para cada instância, de acordo com o desempenho alcançado; o melhor algoritmo tem classificação 1, o segundo melhor tem classificação 2 e assim por diante. Em caso de empate, intervalos intermediários são atribuídos. Seja i o intervalo do j-ésimo

algoritmo sobre a i-ésima instância. O teste de Friedman compara os intervalos médios dos algoritmos calculados pela equação abaixo (3). Sob a hipótese nula H0, o desempenho de todos os algoritmos é o mesmo e os intervalos Rj devem ser semelhantes. A estatística proposta por Friedman na equação 4 é distribuída de acordo com um XF2 com k-1 graus de liberdade, quando n e k são suficientemente grandes, como regra experimental n> 10 e k> 5. Para um número menor de casos e algoritmos, os valores críticos exatos foram calculados (ZAR, 1998, SHESKIN, 2000).

$$R_j = \frac{1}{n}\sum_{i=1}^{n} r_i^j \quad (3) \qquad X_F^2 = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \quad (4) \qquad F_F = \frac{(n-1)X_F^2}{n(k-1) - X_F^2} \quad (5)$$

Iman e Davenport (1980) mostraram uma estatística melhor para XF2 que é distribuída de acordo com uma distribuição Fisher-Snedecor F com graus de liberdade k - 1 y (k - 1) (n - 1) de FF veja na equação acima (5). Sendo assim o teste de Namenyi é utilizado para verificar a distância estatística existente entre as bases de dados.

2.5 Avaliação da Educação Básica

Com a universalização do acesso, começaram debates em torno da qualidade da educação e a avaliação passou a ser um importante elemento, tendo se expandido em todo o Brasil. Essa mobilização em torno das avaliações motivou o desenvolvimento de indicadores que pudessem dar conta de analisar essa qualidade.

Em 1990, o Ministério da Educação criou o Sistema de Avaliação da Educação Básica (SAEB). O SAEB é uma avaliação em larga escala que acontece bienalmente no Brasil e objetiva principalmente diagnosticar a situação da educação brasileira a fim de buscar alternativas para melhorá-la e sanar deficiências em todas as esferas – federal, estadual e municipal (SAEB, 2019).

Na primeira edição, o SAEB avaliou escolas com 1ª, 3 ª, 5ª e 7ª séries do Ensino Fundamental, atendo-se apenas a escolas públicas e localizadas em áreas urbanas. Ao longo dos anos, o SAEB se aperfeiçoou e em 1995 começou a utilizar a Teoria de Resposta ao Item (TRI), ferramenta que permite a comparação de dados ao longo dos tempos (SAEB, 2019).

Segundo INEP (2007), nos anos de 1997 e 1999, os testes foram aplicados a alunos da 4ª e 8ª séries, com questões de Língua Portuguesa, Matemática e Ciências. Já os alunos da 3ª série do Ensino Médio responderam ainda questões de História e Geografia. Na edição seguinte (2001) as avaliações passaram a ter questões apenas de Língua Portuguesa e Matemática e se mantém assim até hoje. Em 2003, as provas foram aplicadas em apenas algumas escolas, tendo um caráter amostral. Em 21 de março de 2005, a Portaria n° 931 reestruturou o SAEB, que passou a ser composto por duas avaliações: a Avaliação Nacional da Educação Básica (ANEB) e a Avaliação Nacional do Rendimento Escolar (ANRESC), popularmente conhecida como Prova Brasil. Dois anos mais tarde, o MEC criou o IDEB, que objetiva primordialmente aferir o nível das redes de ensino e das escolas em particular.

IDEB reúne em um só indicador, dois importantes conceitos na qualidade da educação: fluxo escolar e as médias de desempenho nas avaliações. Esse indicador é calculado a partir dos dados sobre aprovação escolar, obtido no Censo Escolar, e médias de desempenho nas avaliações do INEP, o SAEB – para as unidades da federação e para o país, e a Prova Brasil – para os municípios (INEP, 2007).

Um sistema de ensino ideal seria aquele em que todas as crianças e/ou adolescentes têm acesso à escola, não desperdiçam tempo com repetências, não abandonam a escola precocemente e, ao final de tudo, aprendam. O IDEB foi desenvolvido para ser um indicador

que sintetiza informações de desempenho em exames padronizados com informações sobre rendimento escolar (taxa média de aprovação dos estudantes na etapa de ensino).

Segundo INEP (2017), o Brasil tem quase 50 milhões de alunos matriculados em pouco mais de 180 mil escolas particulares e públicas, da Educação Infantil ao Ensino Médio, incluindo EJA (Ensino de Jovens e Adultos) e Educação Especial. Desse total de educandos, 80% estão em escolas públicas. Os índices educacionais de hoje são alarmantes e apontam que na rede pública de ensino somente 30% dos alunos aprendem o adequado na competência de leitura e interpretação de textos até o 9º ano. Já em Matemática, apenas 14% deles aprendem o adequado em resolução de problemas até o último ano do Ensino Fundamental.

No Brasil, nas últimas décadas, a Educação Básica vem sendo monitorada por meio de índices que avaliam o desempenho dos alunos em testes padronizados, utilizando também as taxas de aprovação e reprovação da escola.

Em nível nacional, o IDEB, criado pelo Decreto nº 6094 de 24 de abril de 2007, foi implantado para acompanhar e monitorar a qualidade das escolas de Ensino Fundamental, focando 5º e 9º ano. Tal índice combina o desempenho dos alunos nas disciplinas de Português e Matemática e o fluxo escolar.

A criação do IDEB representa uma iniciativa precursora no Brasil, pois a partir dele passou a ser possível acompanhar a qualidade da educação sob dois aspectos igualmente importantes: o rendimento escolar, traduzido pelas taxas de aprovação, reprovação e abandono, e a aprendizagem, captada por meio das médias de desempenho na Prova Brasil e no SAEB (INEP, 2007).

De acordo com INEP (2007), o IDEB procura ampliar as possibilidades de mobilização da sociedade em prol da educação, pois o indicador é comparável nacionalmente por meio dos resultados que reúnem aprendizagem e fluxo. Trata-se de uma política pública que busca a qualidade da educação, ou seja, uma ferramenta para o acompanhamento das metas de qualidade do Plano de Desenvolvimento da Educação que estabelece como meta que, em 2022, o IDEB do Brasil será de 6,0 (Meta média comparável a um sistema educacional de qualidade dos países desenvolvidos).

O IDEB, considerado mais do que simplesmente um indicador estatístico, é um condutor de políticas para a melhoria da qualidade da educação das escolas, em nível nacional, estadual e municipal. Ele permite o estabelecimento de metas individuais intermediárias para as escolas, possibilitando a busca pela qualidade do ensino. Tais metas são os percursos estabelecidos de evolução individual dos índices, para que o Brasil atinja o patamar educacional dos países da Organização para a Cooperação e Desenvolvimento

Econômico (OCDE). A ideia central é que cada escola melhore seus indicadores, contribuindo para que o Brasil chegue à meta 6,0 em 2022. Mesmo para aqueles estabelecimentos de ensino que já atingiram a meta, está prevista a continuidade na melhoria. E para aquelas redes de ensino e escolas que apresentarem problemas está previsto apoio específico a fim de auxiliá-las em seu desempenho.

O IDEB varia de zero até 10 pontos em uma escala de qualidade, e quanto maior a nota melhor o desempenho dos alunos e maior a regularidade no fluxo escolar. Neste caso, o resultado da Prova Brasil, que é apresentado em pontos em uma escala de proficiência (Escala SAEB), é posicionado e distribuído em quatro níveis: Insuficiente, Básico, Proficiente e Avançado.

O IDEB apresenta, segundo Ronca (2013), alguns pontos positivos, como a articulação entre os conceitos de fluxo e de desempenho, favorecendo o entendimento da qualidade das escolas e dos sistemas, auxiliando os gestores e os professores na identificação de competências e dificuldades dos alunos nas disciplinas avaliadas – Português e Matemática. O IDEB auxilia também no diagnóstico, no monitoramento e no agrupamento de informações para a criação de políticas públicas, estabelecendo metas individualizadas para cada escola, sistema de ensino, municípios e estados brasileiros. Tal situação tem o objetivo, entre outros, de controlar e elaborar ações para combater a evasão escolar. No entanto, o IDEB também apresenta várias limitações. De fato, não é possível assumir o IDEB como dispositivo de compreensão da qualidade da Educação Básica. As avaliações em larga escala, utilizando somente dois fatores, o fluxo escolar e o desempenho dos alunos, associados diretamente com a questão da leitura e da matemática, não são capazes de captar a complexidade da escola.

Segundo o MEC (2018), o IDEB no ano de 2017 para os anos iniciais da rede pública de Maceió cresceu e atingiu a meta, alcançando 5,6 (**Tabela 1**).O IDEB da rede pública de ensino municipal de Maceió cresceu, porém ainda não atingiu a meta 6,0. O município de Maceió possui55 instituições de Educação Infantil (sendo 32 creches) e 91 escolas de anos iniciais do Ensino Fundamental e 29 de anos finais do Ensino Fundamental. Essas escolas oferecem o ensino fundamental na modalidade regular e na Educação de Jovens e Adultos, além de Programas Educacionais do tipo: Brasil Alfabetizado, Projovem Urbano, Mais Educação e outros (PME, 2014).

Tabela 1 - Metas para a educação básica de Alagoas.

	IDEB Observado							Metas			
	2009	**2011**	**2013**	**2015**	**2017**	**2007**	**2009**	**2011**	**2013**	**2015**	**2017**
Total	4.6	5.0	5.2	5.5	5.8	3.9	4.2	4.6	4.9	5.2	5.5
Estadual	4.9	5.1	5.4	5.8	6.0	4.0	4.3	4.7	5.0	5.3	5.6
Municipal	4.4	4.7	4.9	5.3	**5.6**	3.5	3.8	4.2	4.5	4.8	**5.1**
Privada	6.4	6.5	6.7	6.8	7.1	6.0	6.3	6.6	6.8	7.0	7.2
Pública	4.4	4.7	4.9	5.3	5.5	3.6	4.0	4.4	4.7	5.0	5.2

Fonte: MEC (2018)

O IDEB da rede pública de ensino municipal de Teotônio Vilela cresceu, porém ainda não atingiu a meta 6,0. No ano de 2017, o referido município obteve o segundo melhor IDEB nas séries finais de ensino (5,8) e o terceiro lugar nas séries iniciais (6,9) no Estado dentre as escolas públicas. O município de Teotônio Vilela possui 43 escolas públicas de educação básica com 9.309 alunos matriculados, sendo 3.157 matriculados nos anos finais. Verifica-se que 51% dos alunos até o 9º ano aprenderam adequadamente a competência de leitura e interpretação de textos, enquanto que apenas 43% aprenderam adequadamente a resolução de problemas de matemática (QEdu, 2019).

Tabela 2 - IDEB por municípios séries finais do Estado de Alagoas

MUNICÍPIO	IDEB
Coruripe	6,3
Teotônio Vilela	5,8
Jequiáda Praia	5,3
Campo Alegre	5,2
Junqueiro	5,2
Feliz Deserto	5,0
Roteiro	4,8
Santana do Mundaú	4,8
Pindoba	4,7
...	...
Maceió	3,8

Fonte: QEdu (2019)

2.6 Trabalhos Relacionados

Neste tópico são apresentados alguns trabalhos relacionados a esta temática, assim como suas respectivas formas de abordagem. Nos últimos anos várias pesquisas relacionadas a tópicos de Mineração de Dados Educacionais (MDE) vêm sendo realizadas.

Freitas Júnior *et al.* (2019) utilizaram a ferramenta de mineração de dados, para analisar o Índice de Desenvolvimento da Educação Básica (IDEB) das escolas públicas do município de Maceió, visando auxiliar no processo decisório dos gestores educacionais pela adoção de medidas de melhoria da gestão escolar. Aplicou duas técnicas de mineração de dados: regressão linear e árvore de decisão, visando identificar fatores que influenciam no desempenho do IDEB. Os resultados indicam que diversos fatores influenciam o desempenho do aluno, tais como: a escolaridade dos pais do aluno, o incentivo aos estudos e o compromisso do docente.

Silva e Nunes (2015) aplicaram técnicas de Mineração de Dados Educacionais com a finalidade de oferecer uma análise dos alunos, permitindo identificar várias situações, tais como desistências ou reprovações. Para isso utilizou-se uma base dados de alunos do Ensino Médio, por Série, a partir da utilização da técnica de Classificação usando o Algoritmo J48. Com a estratégia percebeu-se a importância da técnica de classificação da mineração de dados educacionais, gera benefícios e contribui na observação e classificação dos alunos aprovados e reprovados.

Fonseca e Namen (2016) aplicaram técnicas de mineração de dados educacionais com o intuito de analisar o Ensino Fundamental público brasileiro, utilizando a base de dados do Saeb com o intuito de identificar fatores que relacionam o perfil de professores que lecionam matemática com a proficiência obtida por seus alunos. São apresentados os passos deste processo no contexto desta aplicação, explicitando, principalmente, a Mineração de Dados (Data Mining).

Paiva *et al*. (2013) desenvolveram uma ferramenta para a recomendação pedagógica cujo objetivo é prover aos professores de cursos baseados na web, recomendações pedagógicas personalizadas geradas por especialistas no domínio com base nos resultados da mineração dos dados educacionais dos alunos. A ferramenta segue o Processo de Recomendação Pedagógica e foi utilizada em um estudo de caso com dados reais de um curso de língua Espanhola com 200 alunos. Os resultados permitiram detectar padrões de interação úteis, usados na criação de recomendações, avaliadas (relevância) por especialistas no

domínio educacional/pedagógico, e disponibilizadas para que os professores as oferecessem a seus alunos.

Diniz *et al.* (2012) descrevem a distribuição da educação na população do Rio Grande do Norte (RN), mostrando a sua disparidade entre seus municípios. Logo, foram utilizados alguns indicadores da educação do Instituto Brasileiro de Geografia e Estatística (IBGE), juntamente com o Sistema de Informação Geográfico (SIG) TerraView para a elaboração de mapas, utilizando técnicas do geoprocessamento e da estatística. Além disto, também foram utilizadas técnicas de Mineração de Dados (MD) para identificar padrões que ajudam a caracterizar a qualidade do sistema educacional do RN, mostrando assim, a desigualdade entre seus municípios e evidenciando um desenvolvimento educacional concentrado em algumas regiões. Por fim, os resultados mostram que a maioria dos municípios do RN apresenta qualidade educacional abaixo da meta e que a situação é pior em regiões do estado que possuem pequenas atividades econômicas.

Nascimento *et al.* (2018) aplicaram técnicas de mineração de dados com a finalidade de explicar indicadores como a evasão e reprovação escolar no ensino fundamental. Tentar identificar fatores que colocam o desempenho do aluno em risco ou até sua desistência é um desafio aceito por muitos pesquisadores como Sarra *et al*. (2018).

Patrício e Magnoni (2018) discorrem sobre as vantagens de se armazenar grandes quantidades de dados e interpretá-los em busca de melhor compreender o comportamento dos alunos.

Bezerra *et al.* (2016) abordaram a evasão escolar no último ano do ensino fundamental nas escolas públicas estaduais e municipais do estado de Pernambuco, com base nos dados dos Censos Escolares 2011 e 2012. Utilizou-se de técnicas de mineração de dados para identificar o perfil do aluno evadido e estimar a propensão à evasão.

Manhães (2015) apresentou uma proposta de arquitetura baseada em MDE que oferecia informações úteis sobre o desempenho acadêmico dos graduandos e predizia os que estão em risco de abandonar o sistema de ensino através da predição do seu desempenho acadêmico.

Márquez-Vera *et al* (2013) aplicaram técnicas de mineração de dados, investindo em seleção de atributos, a um *data set* de 670 alunos do ensino médio de Zacatelas (México) para descrever o insucesso escolar através da identificação de quais alunos poderiam evadir, considerando um modelo preditivo aplicado a uma coleção de atributos selecionados. Com isso, algumas ações preventivas poderiam ser tomadas para evitar a evasão escolar desses alunos. Por sua vez, Pasta (2011) aplicou técnicas de *Data Mining* em ambientes de gestão educacional, apontando as vantagens da utilização destas técnicas na base de dados destes

ambientes para a gestão das informações de uma instituição de ensino superior, apresentando à mesma o perfil de seus ingressantes e egressos, contribuindo na gestão e organização de campanhas dirigidas a estes diferentes tipos de perfis de seus futuros e ex-alunos.

No trabalho de Silva e Nunes (2015) foi investigado os fatores que contribuem para evasão e reprovação de alunos do ensino médio, por série em um dado município. Utilizou-se a técnica de classificação J.48. Verificou-se que a quantidade de alunos que são desistentes é pequena. O número de alunos aprovados é maior que reprovados e os alunos de cidades vizinhas são menos reprovados que os de Campina Grande. Percebeu-se também que o número de reprovados na 3ª série é menor que nas demais.

Brito et al (2015) realizaram a previsão das notas dos alunos do primeiro período do curso de Ciência da Computação baseado nas notas dos alunos ingressantes via vestibular. Os resultados obtidos mostram que os potenciais alunos evadidos podem ser identificados com taxas de acerto de até 86,9%. Observa-se que a acurácia da classificação dos alunos se manteve superior a 84% em todos os algoritmos, sendo árvore de decisão o que obteve o melhor resultado.

MayaPérez et al. (2018) desenvolveram um modelo de predição baseado em seleção de atributos e classificadores. Objetivo desse estudo fui identificar padrões relacionados com os aspectos de maior influência da desistência (Evasão) dos estudantes de uma instituição de educação superior no México. Utilizou-se a abordagem Filtro de seleção atributos, aplicando em seguida diversões algoritmos de classificação (JRip. OneR, ZeroR, J48 e RipTree). Verificou-se que o algoritmo J48 apresentou o melhor resultado com 75% de acurracia, nesse estudo identificou-se 15 atributos mais relevantes no universo de 39, casas de desistências dos alunos.

Calixto *et al.* (2017) identificaram as variáveis concernentes à evasão escolar, utilizando os dados do censo educacional de 2014, 2015 e 2016 dos estados de Ceará e Sergipe. Utilizou-se nesse trabalho a metodologia CRISP-DM. Na fase de preparação dos dados foi utilizado a ferramenta SPSS. Utilizou-se a abordagem Filtro (Ripper) bem como os algoritmos de Indução de Regras e Regressão Logística. Os modelos criados apresentaram acurácia em torno de 87%. As variáveis relevantes foram: a idade, etapa de ensino, modalidade de ensino, existência de laboratórios e localização da escola se destacaram como variáveis influentes na evasão escolar.

Tabela 3 - Análise comparativa de trabalhos relacionados.

ARTIGO	SELEÇÃO DE ATRIBUTOS	TÉCNICAS DE MDE	CRISP-DM/KDD
FREITAS JÚNIOR et al. (2019)	--	J48 e Regressão Linear	CRISP-DM
SILVA E NUNES (2015)	--	J48	KDD
FONSECA e NAMEN (2016)	--	Naïve Bayes	KDD
PAIVA *et al.* (2013)	--	Árvore de decisão e Regra de Associação	KDD
DINIZ *et al.* (2012)	--	J48	KDD
MARQUEZ-VERAS (2013)	Filtro (10 algoritmos)	J48, JRip, NNge, OneR, Prism, Ridor, ADTree, RandomTree, REPTree, SimpleCart	KDD
BEZERRA et al. (2016)	CHi-squared	J48, Apriori, Regressão logística	CRISP-DM
LIMA et al. (2016)	Três abordagens (filtro, embutida e embaralhado)	Redes Bayesianas, Regressão Logísticas, J48,	KDD
BRITO et al (2015)	---	NaiveBayes, J48, Ada BoostM1, Simple Logistic, IBK, Decision Table, VFI	KDD
PINTO (2019)	Três abordagens (filtro, embutida e embaralhado) 11 algoritmos	NaveBayes, J48, JRip, OneR, REPTree, RandomForest, IBK, SVM	CRISP-DM
MAYA PÉREZ et al. (2018)	Filtro	OneR, J48, ZeroR, JRip e REPTree	KDD
CALIXTO et al. (2017)	Filtro (Ripper)	Indução de Regras e Regressão Logística	CRISP-DM

Fonte: Elaborado pelo autor

O presente trabalho tem como foco identificar os fatores que afetam o desempenho escolar dos alunos (IDEB) das escolas de ensino fundamental do município através dos resultados obtidos na Prova Brasil. Este estudo de caso se propõe a utilizar técnicas de seleção de atributos, no contexto de mineração de dados, visando identificar quais fatores impactam positivamente no Índice de Desenvolvimento da Educação Básica (IDEB) dos alunos das

escolas municipais de Maceió. Analisando as diversas abordagens dos trabalhos consultados, verifica-se mais proximidade com Márquez-Vera *et al* (2013), no processo de seleção de atributos, mas o presente trabalho tem um processo diferente na identificação dos atributos e foca nas instituições educacionais municipais de ensino básico.

3 ESTUDO DE CASO

3.1 Introdução

Nesta seção busca-se realizar um estudo de caso sobre uma experiência com tratamento de dados educacionais em escolas públicas municipais de Maceió e Teotônio Vilela. Nesta perspectiva realizou-se uma análise de vários atributos que influenciam na determinação do Índice de Desenvolvimento da Educação Básica (IDEB) das escolas *lócus* da pesquisa, utilizando diversas ferramentas.

3.2 Delineando os *lócus* e o cenário da pesquisa

A pesquisa realizou um estudo de caso sobre uma experiência com tratamento de dados educacionais em escolas públicas municipais situadas no município de Maceió e de Teotônio Vilela. Visto as dificuldades encontradas durante as coletas dos dados junto as instituições responsáveis pelos dados educacionais, optou-se por utilizar as informações contidas no Portal do INEP.

Município de Maceió, Alagoas

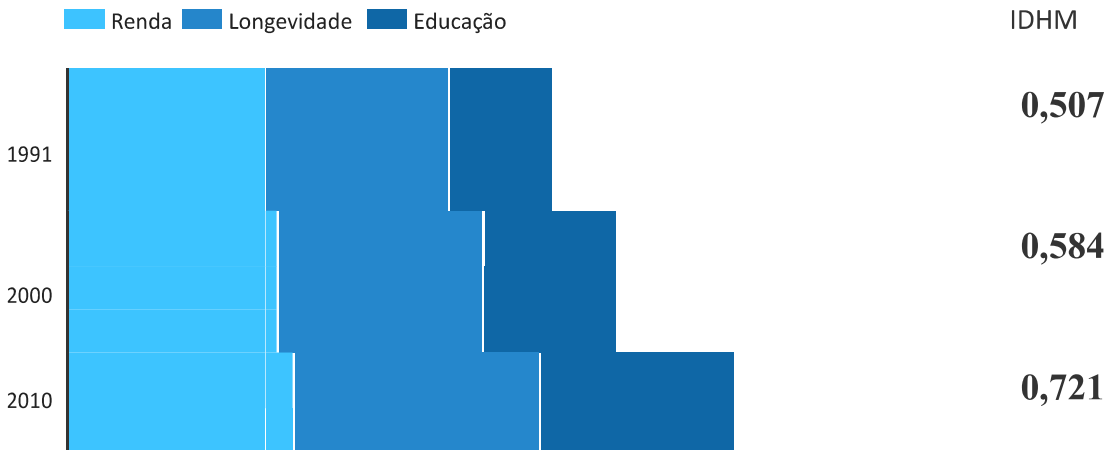
O município de Maceió possui as seguintes características conforme pode-se observar na **Tabela 4,** segundo dados do censo 2010 realizado pelo IBGE**.**

Tabela 4 - Caracterização do município de Maceió.

Área	IDHM 2010	Faixa do IDHM	População (Censo 2010)
516,46 km²	0,721	Alto (IDHM entre 0,700 e 0,799)	932.748 hab.
Densidade demográfica	**Ano de instalação**	**Microrregião**	**Mesorregião**
1805,77 hab/km²	1815	Maceió	Leste Alagoano

Fonte: Adaptada de atlasbrasil.org.br

Figura 2 - IDHM de Maceió.



Fonte: PNUD, Ipea e FJP

O Índice de Desenvolvimento Humano (IDHM) de Maceió é 0,721, em 2010, o que situa esse município na faixa de Desenvolvimento Humano Alto (IDHM entre 0,700 e 0,799). A dimensão que mais contribui para o IDHM do município é Longevidade, com índice de 0,799, seguida de Renda, com índice de 0,739, e de Educação, com índice de 0,635.

Tabela 5 - Índice de Desenvolvimento Humano Municipal e seus componentes - Município - Maceió - Alagoas

IDHM e componentes	1991	2000	2010
IDHM Educação	0,339	0,433	0,635
% de 18 anos ou mais com fundamental completo	41,00	44,61	59,10
% de 5 a 6 anos na escola	52,66	73,73	87,01
% de 11 a 13 anos nos anos finais do fundamental REGULAR SERIADO ou com fundamental completo	36,14	46,51	83,80
% de 15 a 17 anos com fundamental completo	19,17	30,31	49,70
% de 18 a 20 anos com médio completo	15,21	20,34	42,62
IDHM Longevidade	0,594	0,667	0,799
Esperança de vida ao nascer	60,65	65,03	72,94
IDHM Renda	0,649	0,689	0,739
Renda per capita	455,26	583,12	792,54

Fonte: PNUD, Ipea e FJP

Pode-se observa nas características do município de Maceió que ainda apresenta valores altos de IDHM, contudo nessa pesquisa percebe-se que apesar do índices com bons resultados, o município tem resultados inferiores quando se tratando do IDEB com base no conjunto de dados de 2015 e 2017 que foram o alvo desse trabalha em relação a cidades menores do estado de Alagoas. Talvez o fato de que Maceió possuem um conjunto populacional bastante grande apresenta bem mais demandas a serem resolvidas e as decisões tomadas pela gestão possivelmente não estão conseguindo resolver tantas demandas para melhorar os resultados do município em relação ao IDEB.

Ainda sobre o município de Maceió na **Tabela 5,** pode-se observar o aumento do número de alunos cursando o ensino fundamental e médio, mas ainda assim pode ser considerado baixo em relação ao tamanho do município, tendo em vista que ambos os valores estão abaixo de cinquenta por cento do valor que poderia ser alcançado, o que demostra uma evolução no acesso à escola para os dois públicos que poderia ser ainda melhor.

Município de Teotônio Vilela, Alagoas

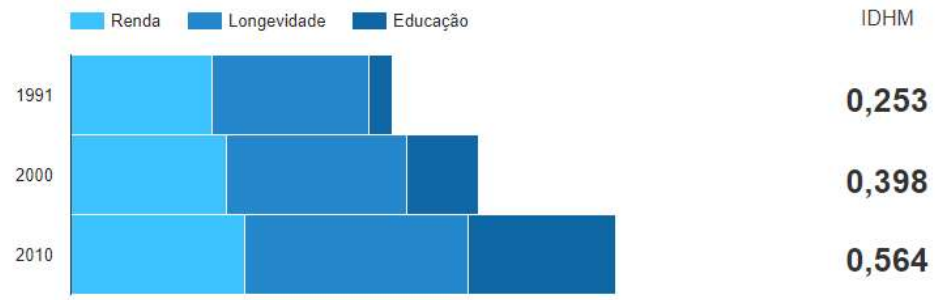
O município possui as seguintes características conforme pode-se observar na **Tabela 6,** segundo dados do censo 2010 realizado pelo IBGE**.**

Tabela 6 - Caracterização do município de Teotônio Vilela.

Área	IDHM 2010	Faixa do IDHM	População (Censo 2010)
298,26 km²	0,564	Baixo (IDHM entre 0,500 e 0,599)	41.152 hab.
Densidade demográfica	**Ano de instalação**	**Microrregião**	**Mesorregião**
137,99 hab/km²	1987	São Miguel dos Campos	Leste Alagoano

Fonte: Adaptada de atlasbrasil.org.br

Figura 3 - IDHM de Teotônio Vilela.



Fonte: Adaptada de atlasbrasil.org.br

O Índice de Desenvolvimento Humano (IDHM) de Teotônio Vilela é 0,564, em 2010, o que situa esse município na faixa de Desenvolvimento Humano Baixo (IDHM entre 0,500 e 0,599). A dimensão que mais contribui para o IDHM do município é Longevidade, com índice de 0,700, seguida de Renda, com índice de 0,549, e de Educação, com índice de 0,466.

Tabela 7 - Índice de Desenvolvimento Humano Municipal e seus componentes - Município de Teotônio Vilela -Alagoas

IDHM e componentes	1991	2000	2010
IDHM Educação	0,074	0,227	0,466
% de 18 anos ou mais com fundamental completo	6,89	13,33	30,33
% de 5 a 6 anos na escola	17,12	65,01	87,18
% de 11 a 13 anos nos anos finais do fundamental REGULAR SERIADO ou com fundamental completo	6,80	36,01	87,91
% de 15 a 17 anos com fundamental completo	4,66	14,79	39,24
% de 18 a 20 anos com médio completo	1,94	2,58	16,37
IDHM Longevidade	0,496	0,569	0,700
Esperança de vida ao nascer	54,77	59,14	67,01
IDHM Renda	0,443	0,488	0,549
Renda per capita	126,33	166,55	244,39

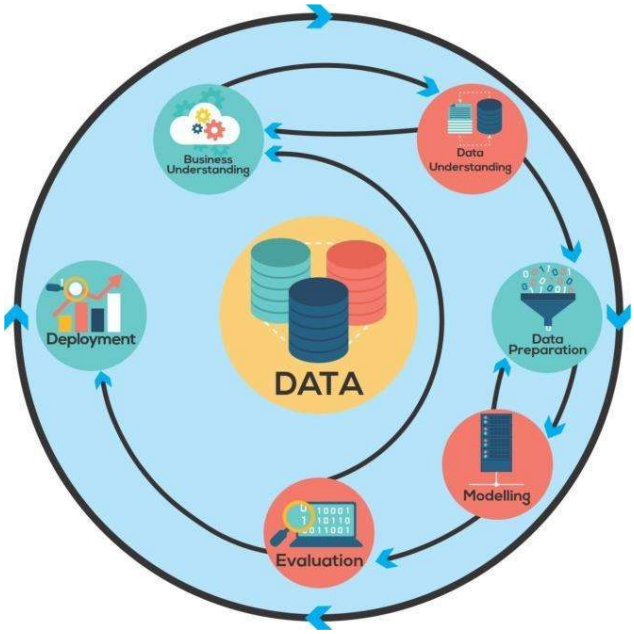
Fonte: PNUD, Ipea e FJP

Pode-se observa nas características do município de Teotônio Vilela que apresenta valores baixos de IDHM, mas através dessa pesquisa podemos investigar que conforme os anos se passaram as decisões tomadas pela gestão possivelmente melhorou o IDHM e por consequência também possibilitou a melhoria da educação do município, conforme pode ser visto no capítulo de resultados. Ainda sobre o município de Teotônio na **Tabela 7,** pode-se observar o aumento do número de alunos cursando o ensino fundamental e médio o que demostra uma evolução a nível de interesse pela educação por parte dos munícipes.

3.3 Metodologia CRISP-DM

Uma das metodologias mais populares para aumentar o sucesso dos processos de mineração de dados é o CRISP-DM (Chapman *et al*., 2000). Essa metodologia define uma sequência não rígida de seis etapas, que permite a construção e implementação de um modelo de mineração para ser usado em um ambiente real, auxiliando as decisões de negócio. Dessa forma, o desenvolvimento do trabalho seguiu as etapas do CRISP-DM, apresentadas na **Figura 4**.

Figura 4 - Etapas da Metodologia CRISP-DM.



Fonte: Adaptado de Chapman *et al.* (2000)

1ª Etapa: Compreensão do domínio. Esta etapa foca no conhecimento dos objetivos do projeto e, então, converte-se esse conhecimento em uma definição de problema de mineração de dados e um plano de projeto preliminar com a intenção de alcançar os objetivos. Nesta etapa, além de definir os objetivos a serem alcançados, determinam-se os critérios de sucesso do projeto. Também serão definidos os recursos necessários à execução de todo o plano de mineração. O objetivo do projeto, nesse caso, é encontrar os fatores responsáveis pela melhoria do desempenho escolar (IDEB), para isso aplicam-se técnicas de seleção de atributos para descobrir as questões que mais impactam no IDEB. Converte-se este problema de pesquisa em uma óptica guiada para mineração de dados, além de definir os objetivos específicos a serem alcançados. Este estudo fará a análise dos dados educacionais de escolas públicas municipais de Maceió e Teotônio Vilela relativos aos alunos dos anos finais do ensino fundamental (9º ano).

2ª Etapa: Entendimento dos dados. Os dados podem vir de diversas fontes e ter diversos formatos. Assim, a partir de uma coleção inicial, já livre de possíveis problemas, é necessário descrevê-los, explorá-los e, por fim, verificar sua qualidade. As bases de dados usadas nesse estudo foram disponibilizadas abertamente pelo INEP em seu portal. Coletou-se os dados educacionais das escolas selecionadas no portal do INEP, não houve muito a ser feito em

relação à pré-processamento e transformação de dados, pois os dados já haviam sido limpos e validados pelo próprio portal.

Foram coletados os dados educacionais da rede municipal de Maceió e de Teotônio Vilela por meio do portal do INEP, mediante a ferramenta Anaconda Distribuição para visualização e conferência dos tipos de dados antes de avançar para próxima etapa. Segundo o INEP (2016), o questionário do aluno dos anos finais do ensino fundamental consiste de 57 itens, distribuídos em 6 (seis) categorias: caracterização sociodemográfica, informações socioeconômicas, capital social, capital cultural, trajetória escolar e atitudes em relação a estudos específicos, conforme o **Tabela 8**.

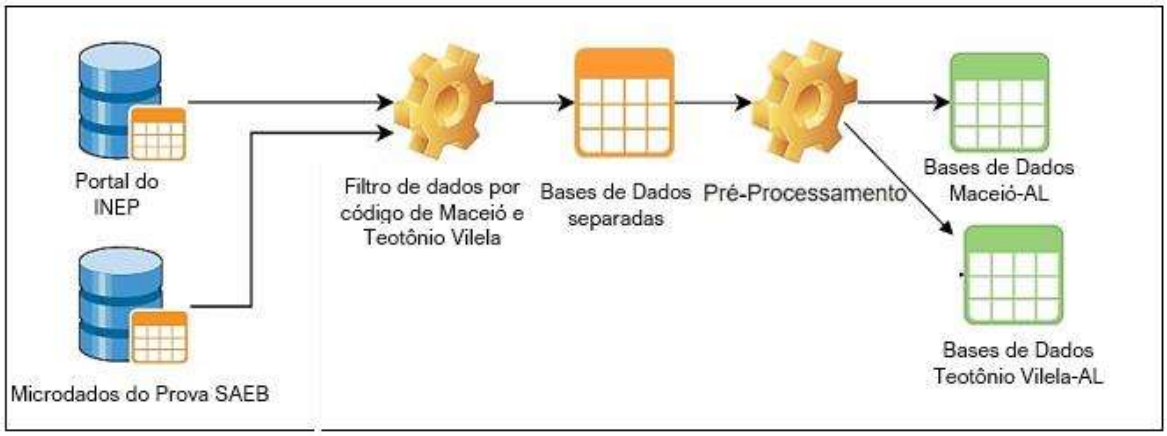
Tabela 8 - Estrutura básica do questionário do aluno.

Categoria	Descrição
Caracterização sociodemográfica	Sexo, cor, raça e idade
Informações socioeconômicas	Capital econômico
Capital social	Convívio, formação e atitude dos pais ou responsáveis na educação do aluno
Capital cultural	Hábitos de leitura e gestão do tempo
Trajetória escolar	Tempo de permanência na escola, reprovação e abandono
Atitudes em relação a estudos específicos	Atitudes do aluno e do professor em relação ao estudo de português e matemática, somados a um item sobre uso da biblioteca ou sala de leitura

Fonte: Adaptado de INEP (2016)

3ª Etapa: Preparação dos dados. Esta etapa tem o objetivo construir o conjunto final de dados que será utilizado nas ferramentas de modelagem. As tarefas de preparação podem ser realizadas muitas vezes, e sem uma ordem pré-determinada o caminho utilizado nessa etapa está discriminado na **Figura 5**.

Figura 5 - Etapa de preparação dos dados.



Fonte: Elaborado pelo autor

Conforme **Figura 5** essa etapa envolve operações como tratar a falta de dados em alguns campos, limpeza de dados como a verificação de inconsistências, redução da quantidade de campos em cada registro, o preenchimento ou a eliminação de valores nulos, remoção de dados duplicados. Inicialmente, devido ao fato do INEP disponibilizar apenas os dados nacionais, foi necessário um filtro para selecionar apenas os alunos das escolas públicas do município de Maceió e Teotônio Vilela, dividindo em duas bases de dados distintas. A **Figura 6** apresenta o processo de limpeza dados realizado com ferramenta Anaconda com Pandas e Numpy.

Figura 6 - Preparação dos dados usando a ferramenta Anaconda.

```
In [11]: import numpy as np
         import pandas as pd

         filepath = os.sep.join(data_path + ['Base814 Teotonio VilelaAL-CONDICAO-SEM-VAZIO-SMOTE-EXResult.csv'])
         print(filepath)
         data = pd.read_csv(filepath)
         data.head()

         data\Base814 Teotonio VilelaAL-CONDICAO-SEM-VAZIO-SMOTE-EXResult.csv
```

Out[11]:

TX_RESP_Q049	TX_RESP_Q050	TX_RESP_Q051	TX_RESP_Q052	TX_RESP_Q053	TX_RESP_Q054	TX_RESP_Q055	TX_RESP_Q056	TX_RESP_Q057	CONDICAO
A	A	B	A	A	B	B	B	C	'ABAIXO DA MEDIA'
A	A	A	A	A	A	A	B	C	'ACIMA DA MEDIA'
A	A	A	A	A	A	A	B	C	'ACIMA DA MEDIA'
A	A	A	A	A	A	A	B	A	'ABAIXO DA MEDIA'
A	A	A	A	A	A	A	B	A	'ACIMA DA MEDIA'

Fonte: Elaborado pelo autor

Os dados baixados são compostos por vários atributos dos alunos, que foram mantidos e também aqueles referentes às respostas dos questionários contextuais e a nota de proficiência de cada aluno nas disciplinas de Língua Portuguesa e Matemática. Essas notas serão usadas na etapa de pré-processamento e mantidas, pois permitem avaliar se o conjunto de alunos possui tendência para tirar nota satisfatória ou não no SAEB, visto que o mesmo é

composto por três avaliações externas em larga escala: a ANEB, a Anresc (Prova Brasil) e a ANA.

Ao invés de usar a nota de proficiência como variável dependente, foi decidido utilizar uma técnica de discretização nas notas para simplificar o problema. Essa técnica consiste na transformação de uma variável numérica para uma variável categórica, que será denominada *CONDICAO*, referente à condição do aluno nas matérias de Português e Matemática. Essa nova variável classifica cada aluno em duas possíveis condições: acima da média e abaixo da média. Foram calculadas a média e a mediana para as notas de proficiência de Português e Matemática do nono ano como segue nas **Tabelas 9 e 10**.

Tabela 9 - Estatística dos alunos de Maceió.

Médias e medianas dos alunos			**Quantidade de alunos por condição**		
	LP	**MT**		**LP**	**MT**
Média	251,64	254,65	**Acima da média**	640	640
Mediana	253,03	255,05	**Abaixo da média**	680	680

Nota: LP-Língua Portuguesa; MT-Matemática. Fonte: Elaborada pelos autores

Tabela 10 - Estatística dos alunos de Teotônio Vilela.

			Quantidade de alunos por condição		
	LP	**MT**		**LP**	**MT**
Média	251,64	254,65	**Acima da média**	282	278
Mediana	253,03	255,05	**Abaixo da média**	267	271

Nota: LP-Língua Portuguesa; MT-Matemática. Fonte: Elaborada pelos autores

A importância da mediana para esses casos é ver a proximidade da média, podendo assim detectar a existências de *outliers* que possam interferir na representação da média, já que a mediana não é suscetível a tal fenômeno. Como se pode observar os valores da média e mediana são próximos, o que valida o uso da média para esse caso. Com isso, cada aluno foi separado em uma das duas possíveis condições.

4ª Etapa: Modelagem. Tipicamente, existem diversas técnicas para o mesmo tipo de problema de mineração. A fim de atingir o balanceamento completo e maximizar a precisão dos algoritmos, foi decidido utilizar técnicas de balanceamento de dados. Essas técnicas consistem em gerar dados sintéticos para equilibrar a base de dados para as variáveis dependentes. Neste estudo foi utilizado o método SMOTE (*Synthetic Minority Oversampling Techniques*).

O conjunto de dados de Teotônio Vilela sem o balanceamento tem 513 instancias, sendo 70% das instancias classificadas como IDEB "abaixo da média" (Alunos com nota final < 5), um total de 360 alunos. O restante do conjunto de dados dos alunos corresponde a

30% nesse caso 153 alunos apresentaram resultado da nota IDEB maior que 5 “acima da média”. A classe majoritária é a de alunos com nota abaixo da média o que prejudicaria a criação do modelo influenciando o classificador a classificar de forma errada. Para ser aceitável, acurácia de predição de dados desbalanceado deveria ser maior que acurácia obtida atribuindo todo novo objeto à classe majoritária. Em nossos casos de teste não houve um bom desempenho da na acuraria preditiva sem o balanceamento, o que indicou a necessidade de balanceamento uma vez que foi possível criar novas instancias artificiais removendo a diferença existente entre os dados majoritário e minoritário a estratégia utilizada foi realizada utilizando o algoritmo SMOTE.

Durante o processo de criação da base de dados balanceada foi tomado os devidos cuidados para evitar os problemas de *overfitting*, situação onde o modelo é supera ajustado aos dados de treinamento. Também foi evitado o *underfitting*, quando os dados são eliminados de classe majoritária e os dados relevantes de indução do modelo correto sejam perdidos provocando a não aprendizagem do modelo, quando preditor não se ajusta ao modelo de treinamento. Os dados de Maceió, sem o balanceamento tem 910 instancias, sendo 69% das instancias classificadas como IDEB “abaixo da média” (Aluno com nota final < 5), um total de 628 alunos. O restante do conjunto de dados dos alunos corresponde a 31% nesse caso 282 alunos apresentaram resultado da nota maior que 5 “acima da média”. A mesma estratégia de balanceamento dos dados foi aplicada ao conjunto de dados de Maceió, aplicando o SMOTE e também avaliando o comportamento da base ajustada e tomando os devidos cuidados para evitar problemas no processo de treinamento do modelo.

Nos dois casos, o método SMOTE foi utilizado para gerar mais dados das classes de minoria através da adição de instâncias em segmentos de linhas que juntam os k membros de uma determinada minoria. A partir disso, essa pesquisa terá 813 instâncias (Base Teotônio Vilela) e 1320 instancias (base Maceió) para cada classe de “*Condicao”* na matéria de língua portuguesa e 813 instâncias (Base Teotônio Vilela) e 1320 instancias (Base Maceió) para cada classe em matemática.

Nesta etapa, várias técnicas de modelagem são selecionadas e aplicadas. Tipicamente, existem diversas técnicas para o mesmo tipo de problema de mineração. No entanto, há algumas que dependem do objetivo desejado. Esta etapa pode ser apoiada por diversas áreas, entre as quais: aprendizagem de máquina, visualização e estatística.

A fim de atingir o balanceamento completo e maximizar a precisão dos algoritmos de aprendizagem de máquina, foi decidido utilizar técnicas de balanceamento de dados. Essas técnicas consistem em gerar dados sintéticos para equilibrar a base de dados para as

variáveis dependentes.

Existem vários algoritmos de balanceamento de dados, nesse estudo foi utilizado o método SMOTE (*Synthetic Minority Oversampling Techniques*). Neste método, são gerados mais dados das classes de minoria através da adição de instâncias em segmentos de linhas que juntam os k membros de uma determinada minoria. A partir disso, essa pesquisa terá para a rede pública de Maceió: 910 instâncias (abaixo da média) para cada classe de *Condição* e 410 instâncias (acima da média) para cada classe. Enquanto que para a rede pública de Teotônio Vilela têm-se 529 instâncias (abaixo da média) para cada classe e 284 instâncias (acima da média).

A etapa de seleção de atributos tem como objetivo excluir atributos redundantes e que não são úteis para a criação do modelo de predição. Ao utilizar a seleção de atributos, busca-se um melhor desempenho e a simplificação do modelo, reduzindo com isso o custo computacional (Márquez-Vera *et al.*, 2013). Na **Tabela 11** tem-se a descrição dos algoritmos de seleções utilizados e sua função de processamento na busca para encontrar os melhores atributos de um conjunto de dados.

Tabela 11 - Descrição dos Algoritmos de Seleção de Atributos adotados na pesquisa e implementados no Weka.

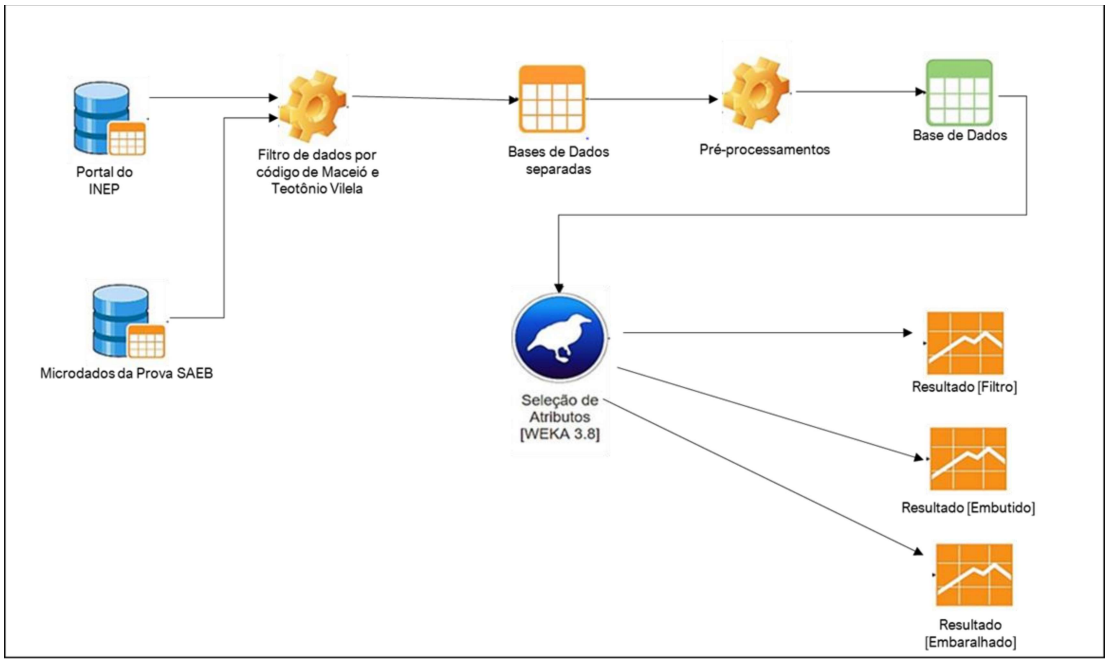
Algoritmo	**Função**
CfsSubsetEval	Considera o valor preditivo de cada atributo individualmente, juntamente com o grau de redundância entre eles
ChiSquared-AttributeEval.	Calcular qui-quadrado estatístico de cada atributo em relação à classe
GainRatio-AttributeEval	Avalia atributos medindo sua razão de ganho em relação à classe.
InfoGain-AttributeEval	Avalia atributos medindo seu ganho de informação com relação à classe.
OneRAttributeEval	Usa a medida de precisão (acurácia) simples adotada pelo classificador OneR

ReliefFAttributeEval	Baseado em instância: faz amostragem de instâncias aleatoriamente e verifica instâncias vizinhas das mesmas e diferentes classes
SymmetricalUncertAttribute Eval	Avalia atributo com base na incerteza Simétrica
Embutido: J48 e REPTree	Avalia atributo com base no ganho de informação.
Wrapper com NaiveBayes	Avalia conjunto de atributos usando um algoritmo Naive Bayes que avalia um conjunto de atributos com base no Teorema de Bayes

Fonte: Elaborado pelo autor

Para selecionar os dados mais significativos para este trabalho foram utilizados algoritmos de cada grupo de método de seleção que são: filtro, embaralhamento e embutida. Entre as técnicas de filtro foram selecionados: *Chi-Square*, Incerteza Simétrica, Correlação, OneR, Taxa de Ganho, Ganho de Informação, *Relief* e o *Correlation-based Feature Selection* (*CFS*). Para o algoritmo de embaralhamento foi utilizado o *Wrapper* com o NaiveBayes; enquanto que os algoritmos REPTree e o J48 irão representar os algoritmos embutida.

Figura 7 - Processo de Seleção de atributos.



Fonte: Elaborado pelo autor

Na **Figura 7** tem-se etapa de seleção de atributos onde foram apresentados os atributos selecionados e as respectivas quantidades, utilizando as seguintes bases de dados: completa, embutida, embaralhada, filtro e "Todos". "Todos" representa o conjunto de atributos após seleção dos mesmos, utilizando as três abordagens (embutida, filtro e embaralhado), com uso do método de validação *cross validation* com *fold* 10 e 30 interações, em cada execução tem-se um conjunto de instâncias diferentes. Para este estudo, as notas dos alunos se mantiveram separadas entre Língua Portuguesa e Matemática. Os dados das escolas públicas de Maceió são mostrados na **Tabela 12**. Ressalta-se que o dicionário de dados referentes aos atributos contidos na **Tabela 12** encontra-se em **Anexo B**.

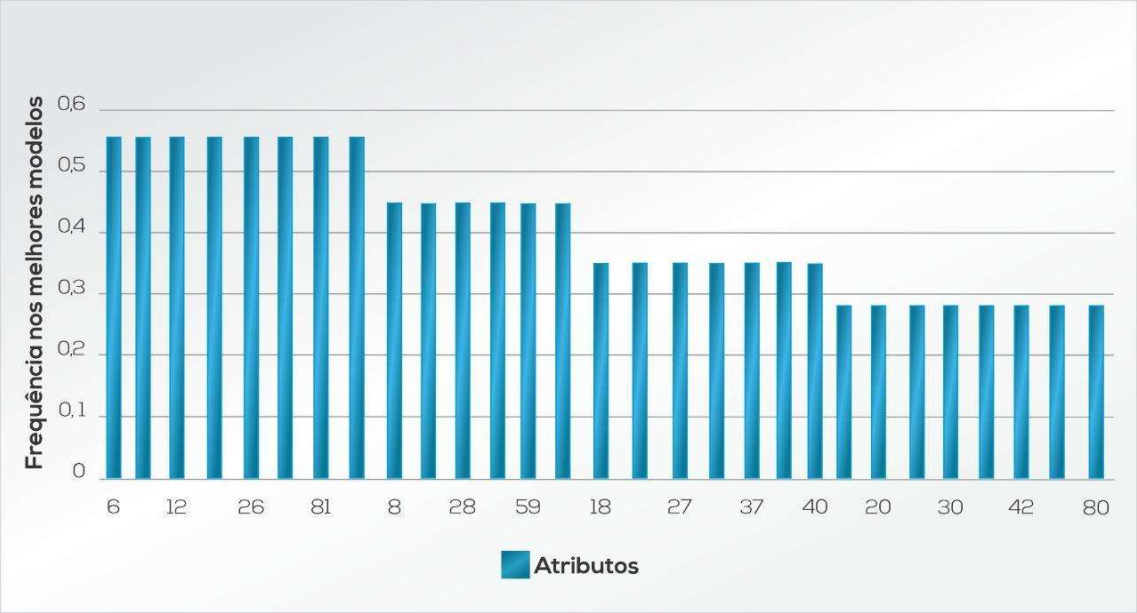
Tabela 12 - Atributos selecionados para alunos do 9º ano das escolas públicas de Maceió.

Abordagem	Algoritmo	Língua Portuguesa		Matemática	
		Atributos	Quantidade	Atributos	Quantidade
Embutida	*REPTree e J48*	5,7,27,58,82	5	5,7,26,30,57,81	6
Filtro	CfsSubsetEval, CorrelationAttribute, *ChiSquaredSubsetEval*, GainRatio-AttributeEval, InfoGain-AttributeEval, OneRAttributeEval, SymmetricalUncertAttributeEval e ReliefFAttributeEval	6,8,9,12,18,19,20,21, 24,25,26,27,28,29,30, 35,37,38,39,40,42,58,59,6 8,74,79,80,81,82	29	6,8,12,18,19,20,21,24,25, 26,27,28,29, 30,35,37,38,39,40,42,58, 59,68,74,79,80,81,82	28
Embaralhamento	WrapperSubsetEval com o NaiveBayes	7,27,58,81,82	5	7,26,30,57,81	5
Todos		5,6,7, 8,9,12,18,19,20,21, 24,25,26,27,28,29,30, 35,37,38,39,40,42,58,59,6 8,74,79,80,81,82	31	5,6,7,8,12,18,19,20,21,24, 25,26,27,28,29,30,35,37, 38,39,40,42,57,58,59,68, 74,79,80,81,82	31

Fonte: Elaborado pelo autor

Além das abordagens de seleção de atributos tradicionais apresentadas, utiliza-se um método, denominado de Merge, que combina os atributos mais frequentes nos melhores conjuntos de seleção (LIMA, 2016). Neste método, para cada atributo será gerado um *score* e, ordenando esses *scores* será possível obter um ranking da mesma forma que qualquer técnica individual de uma das abordagens apresentadas anteriormente. A partir deste ranking, utiliza-se como estratégia de corte, selecionar um subconjunto de atributos com frequência superior a dois. A **Figura 8** apresenta o ranking gerado, onde no eixo y são apresentados os méritos de cada atributo, calculado pela frequência de vezes que esse atributo se encontra entre os melhores conjuntos de atributos gerados.

Figura 8 - Ranking gerado pelo método de combinação de atributos Merge para o município de Maceió.



Fonte: Elaborado pelo autor

À medida que se realizou a etapa de construção do método Merge teve-se os atributos com melhor ranking e também melhoria do tempo de processamento da base de dados, com destaque para os atributos com score entre 0,5 e 0,8; representando os atributos mais forte do conjunto de dados, conforme **Figura 8**, tendo em vista o número de ocorrências entre os algoritmos de seleção de atributos aplicados que são:

- (Questão 16): Desvio Padrão Língua Portuguesa. Refere-se ao valor estatístico do extrato da prova de português que permite visualizar o conjunto de alunos com desempenho médio ou individual de um discente;
- (Questão 18): Desvio Padrão Língua Portuguesa SAEB. Refere-se ao valor estatístico do extrato de português do resultado da prova, permitindo a tomada de decisão a respeito dos alunos com deficiência nesta disciplina, conforme descritores selecionados pelo MEC;
- (Questão 22): Em que ano você nasceu. Informação que permite acompanhar o conjunto de alunos que estão em distorção idade-série, fator que influencia bastante na evasão escolar.

Um fator fundamental para esse método é uma boa avaliação sobre os conjuntos de atributos gerados. Dessa forma, esse método somente pode ser aplicado após a utilização de um método de avaliação dos subconjuntos de atributos gerados pelas técnicas de seleção de atributos. Assim, foram construídos 4 (quatro) modelos reduzidos, os quais serão introduzidos em algoritmos de classificação para validar cada modelo. Para analisar a

precisão dos dados selecionados, um algoritmo de cada categoria descrita neste trabalho foi arbitrariamente escolhido dentre as opções já desenvolvidas na ferramenta Weka, foram eles: NaivesBayes, *J48*, *JRip*, *LibSVM, RandomForest, IBK, OneR* e REPTree. Os quais foram parametrizados conforme **Tabela 13.**

Tabela 13 - Algoritmos de Classificação do Weka abordados e parametrização utilizada.

Nome do algoritmo implementado no Weka	Função	Parametrização utilizada	Categoria	Método de Busca	Tipo
JRip	implementa *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER), incluindo a otimização global heurística do conjunto de regras.	*usePruning: True e optimizations: 2*	Regras	Baseado em Procura	Caixa Branca
OneR	é o classificador 1R com um parâmetro – o tamanho mínimo do *bucket* para discretização.	*minBucketSize: 6 e numDecimalPlace: 2*			
J48	utiliza C4.5 árvore de decisão para aprendizagem. Implementa o C4.5 revisão 8.	*ConfidenceFactor: 0.25 e minNumObj: 2*	Árvore de Decisão		
RandomForest	constrói varias árvores que considera um dado número de características aleatórias utilizado para classificar um novo exemplo.	*breakTiesRandomly: False e outputoutOfBangComplexity statistic: False*			
REPTree	*reduced-error pruning*. constrói uma árvore de decisão ou regressão usando a redução de ganho / variação de informação e remove-a usando a podagem de erro reduzido.	*minVarianceProp: 0.001 e noPruning: False*			
LibSVM	biblioteca que inclui classificadores do tipo *wrapper* que permitem implementações de máquinas de vetores de suporte (SVM) e regressão logística de terceiros para o Weka.	*KernelType: exp(-gamma*|u-v|^2)*	SVM	Baseado em Otimização	Caixa Preta
Naive Bayes	implementa o classificador probabilístico Naïve Bayes padrão	*useKernelEstimador: False e useSupervisedDiscretization: False*	Naive Bayes	Probabilístico	
iBk	*k-nearest-neighbor* (k-NN), é um classificador de k-vizinhos mais próximos.	*KNN: 1 e nearestNeigh:LinearNNSearch*	k-NN	Baseado em Distância	

Fonte: Elaborado pelo autor

Métricas de Desempenho em Aprendizagem de Máquina

Segundo Borges (2019), conforme também aplicadas em Márquez, Morales e Soto (2013); Márquez *et al.*, (2016) e conceituado em Faceli; Gama e Carvalho (2011), as métricas comumente utilizadas para predições em MDE, e que também serão exploradas o decorrer desta pesquisa são:

Taxa de acerto ou acurácia total (Accuracy): calculada pela soma dos valores da diagonal principal da matriz de confusão, dividida pela soma de todos os elementos da matriz. É a métrica principal utilizada nesta pesquisa.

$$Accurácia\ (ACC) = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population}$$

Taxa de Verdadeiros Positivos (True Positive Rate): Calcula a soma dos verdadeiros positivos sobre os verdadeiros.

$$TP\ rate = \frac{\sum True\ Positive}{\sum True\ Positive + \sum False\ Negative}$$

Taxa de Verdadeiros Negativos (True Negative Rate): Calcula a soma dos verdadeiros negativos sobre os negativos.

$$TN\ rate = \frac{\sum True\ Negative}{\sum True\ Negative + \sum False\ Positive}$$

No intuito de facilitar a identificação de padrões, comparação entre algoritmos, visualização gráfica, e análise exploratória do comportamento dos atributos, neste trabalho optou-se por mensurar o desempenho das técnicas utilizando-se predominantemente a acurácia total combinado ao uso do método estatística de Friedman e Namenyi. Pretende-se, em trabalhos futuros, explorar outras métricas importantes que venham a explicitar outras perspectivas de eficiência.

Na **Tabela 14** são mostradas as precisões dos algoritmos de classificação aplicados ao nono ano nas matérias de Língua Portuguesa e Matemática. Também é mostrada a precisão média de cada modelo reduzido gerado para que seja possível avaliar o desempenho médio da redução, usando o método de validação de algoritmos *cross validation* (validação cruzada) com fold de tamanho 10 e executado 30 vezes para gerar um ranking e, por fim, realizado o teste estatístico de Friedman e Nemenyi.

Tabela 14 - Precisão dos classificadores para Língua Portuguesa e Matemática – escolas públicas de Maceió.

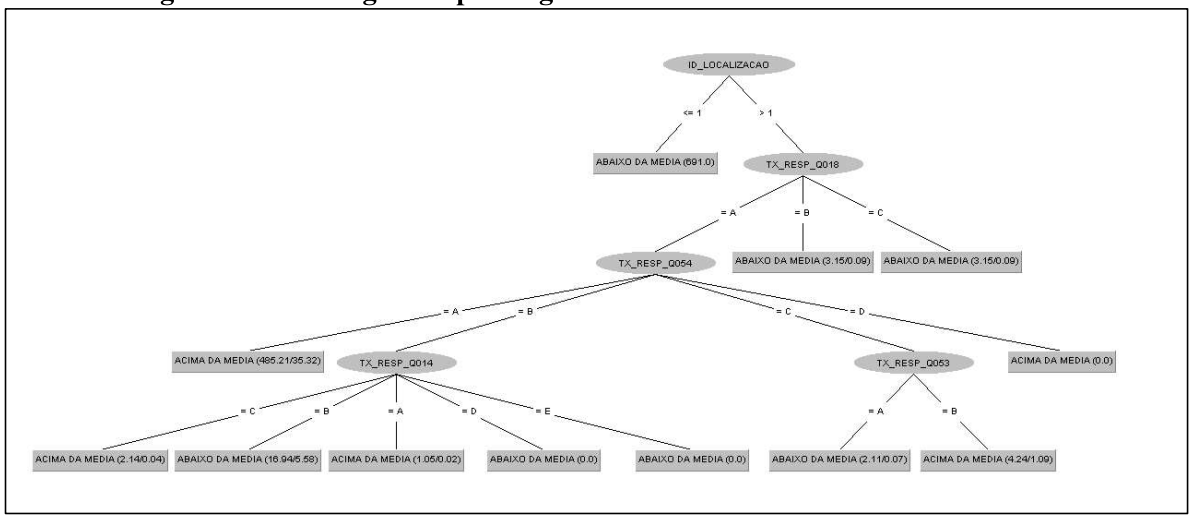
Algoritmo	Completo		Embutida		Filtro		Embaralhado		Todos	
	LP	MT	LP	MT	LP	MT	LP	MT	LP	MT
NaiveBayes	92,86%	86,60%	55,24%	54,93%	93,05%	91,02%	55,24%	54,93%	93,05%	90,88%
J48	100%	100%	51,72%	52,18%	100%	100%	51,72%	52,18%	100%	100%
JRip	100%	100%	54,42%	54,52%	100%	100%	54,42%	54,31%	100%	100%
LibSVM	100%	100%	51,51%	51,51%	100%	100%	51,51%	51,51%	100%	100%
RandomForest	97,26%	91,67%	54,51%	54,32%	93,67%	87,07%	54,51%	54,31%	94,71%	88,54%
IBK	73,15%	67,25%	54,51%	54,91%	87,53%	75,54%	54,91%	54,91%	87,53%	75,82%
OneR	100%	100%	48,62%	48,42%	100%	100%	48,62%	48,42%	100%	100%
REP Tree	51,49%	51,48%	51,51%	51,51%	51,49%	51,48%	51,51%	51,51%	51,49%	51,48%
Precisão Média	**89,34%**	**87,13%**	**52,81%**	**52,79%**	**90,72%**	**88,14%**	**52,81%**	**52,76%**	**90,85%**	**88,34%**
LP Lingua Portuguesa										
MT Matemática										

Fonte: Elaborado pelo autor

A melhor abordagem verificada, conforme a **Tabela 14**, foi "Todos" que representa o conjunto de atributos após seleção dos mesmos, utilizando as três abordagens (embutida, filtro e embaralhado), com uso do método de validação *cross validation* com fold 10 e 30 interações, em cada execução tem-se um conjunto de instâncias diferentes e nesse caso uma precisão média de 90,98% para a base de dados de Português e 88,34% para Matemática. Em relação à abordagem de filtro, os classificadores apresentaram uma precisão média de 90,72% (português) e 88,14% (matemática), comprovando que o método filtro aplicado individualmente apresentou um resultado satisfatório para o processo de seleção dos melhores atributos do conjunto de dados, porém, a junção com as demais abordagens melhora os resultados da base de dados e evidência os atributos mais fortes para a base de dados de Português e Matemática. O conjunto de dados completo sem seleção de atributos possui precisão média de 89,34% (Português) e 87,13% (Matemática) apesar dos valores serem altos, fica evidenciado que os dados nesse primeiro momento já podem ser usados nos classificadores.

Na **Figura 9,** tem-se a árvore gerada pelo algoritmo J48 no qual teve um dos melhores resultados nas predições dos fatores que influenciam no desempenho do IDEB desses municípios. Pode-se destacar os nós das questões 14, 18, 54 e 53 conforme já discutido na seção de seleção de atributos o que significa cada um deles.

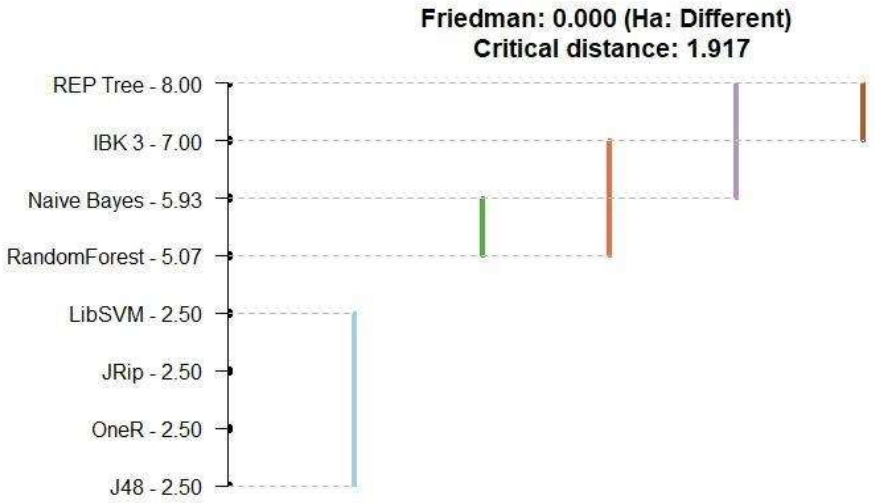
Figura 9 - Árvore gerada pelo algoritmo J48 com base de dados "Todos".



Fonte: Elaborado pelo autor

Todavia, existe a necessidade de realizar o processo de seleção de atributos com o objetivo de entender melhor a importância dos atributos mais relevantes que neste caso fica evidenciado quando utilizando as diferentes abordagens de seleção de atributos. Apesar da precisão média da abordagem embutida ser 52,81% (Português) e 52,79% (Matemática) e da abordagem embaralhado ser 52,81% (Português) e 52,76% (Matemática), é perceptível o fato de que o conjuntos das abordagens formam uma seleção de atributos fortes que permitem uma acurácia de classificação alta e também um conjunto de atributos possíveis de serem discutidos como importantes, dada a sua incidência nessa etapa de processamento dos dados com diferentes categorias de técnicas de seleção de atributos.

Figura 10 - Aplicação do método de avaliação estatística de Friedman e Namenyi no R para comparar as saídas dos classificadores (Maceió).



Fonte: Elaborado pelo autor

A **Figura 10** apresenta o resultado do teste estatístico de Friedman e Namenyi, aplicado à base de dados "Todos" conforme pode-se observar os algoritmos J48, OneR, JRip e LibSVM apresentaram os melhores resultados com 100% de acurácia de classificação para o conjunto de dados de Português. Por outro lado, verifica-se que os algoritmos RandomForest, IBK e Naive Bayes apresentaram resultados satisfatórios, enquanto o REPTree apresentou o pior resultado entre todos.

O valor de distância crítica, conforme mostrado na **Figura 10**, corresponde há diferença estatística entre dois algoritmos quando utilizados em um determinado conjunto de dados. Essa diferença é descoberta realizando a subtração entre os valores da colocação de dois algoritmos no ranking, se o resultado obtido for maior que a distância crítica, isto corresponde que os algoritmos são diferentes estatisticamente e que um deles realmente possui uma melhor eficácia quando aplicado em um cenário exercido. Nesse contexto, o algoritmo REPTree obteve o pior ranking, sendo diferente estatisticamente dos demais algoritmos.

Tabela 15 - Atributos selecionados para alunos do 9º ano das escolas públicas de Teotônio Vilela.

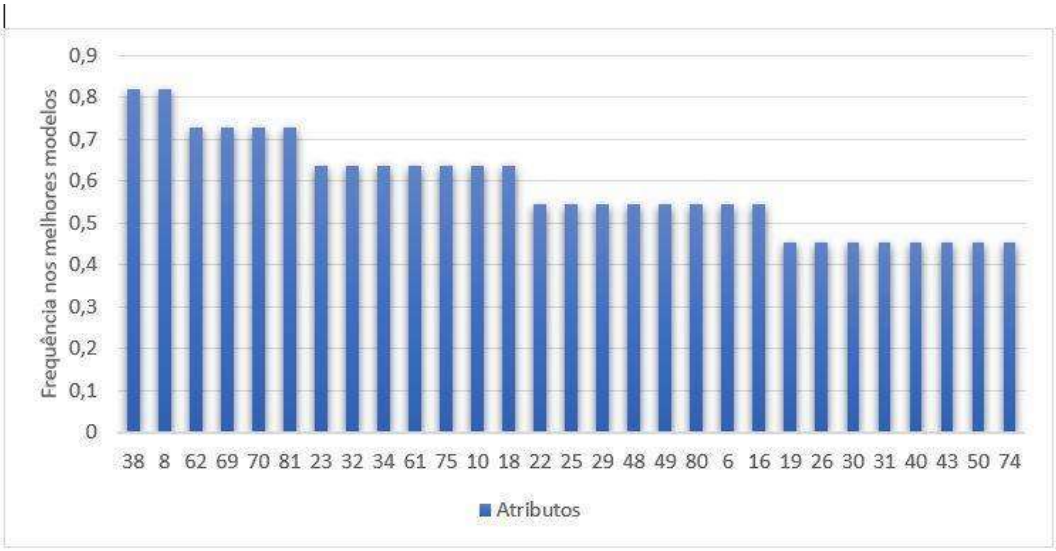
Abordagem	Algoritmo	Língua Portuguesa		Matemática	
		Atributos	*Quantidade*	*Atributos*	*Quantidade*
Embutida	*REPTree e J48*	6,8,16,24,29,32,34,38, 39,48,49,61,63,65,66, 70,72,74,75,77,81	21	6,8,16,29,32,34,38,39,40,49,61 ,63,65,66,70,72,74,75,77,81	20
Filtro	*CfsSubsetEval, CorrelationAttribute, ChiSquaredSubsetEval, GainRatio-AttributeEval, InfoGain-AttributeEval, OneRAttributeEval, SymmetricalUncertAttributeEval e ReliefFAttributeEval*	*1,5,6,7,8,9,10,12,15,1 6,17,18,19,21,22,23,2 5,26,28,29,30,31,32,3 3,34,35,36,37,38,39,4 0,42,43,45,46,47,48,4 9,50,51,52,53,59,60,6 1,62,63,64,65,66,67,6 8,69,70,71,72,73,74,7 5,76,77,78,79,80,81,8 2,83,84*	*68*	*5,6,7,8,9,10,12,15,16,17,18,19, 21,22,23,25,26,28,29,30,31,32, 33,34,35,36,37,38,39,40,42,43, 45,46,47,48,49,50,51,52,53,59, 60,61,62,63,64,65,66,67,68,69, 70,71,73,74,75,76,77,78,79,80, 81,82,83,84*	*66*
Embaralhamento	*WrapperSubsetEval com o NaiveBayes*	*8,9,10,18,25,28,30,34, 41,46,47,53,57,58,63, 73,80,81,83,84*	*20*	*10,18,23,25,30,31,32,34,52,62, 65,67,70,72,75,77,81,84*	*18*
Todos		1,5,6,7,8,10,12,15,16, 17,18,19,21,22,23,24, 25,26,28,29,30,31,32, 33,34,35,36,37,38,39, 40,41,42,43,45,46,47, 48,49,50,51,52,53,57, 58,59,60,61,62,63,64, 65,66,67,68,69,70,71, 72,73,74,75,76,77,78, 79,80,81,82,83,84	71	5,6,7,8,9,10,12,15,16,17,18,19, 21,22,23,25,26,28,29,30,31,32, 33,34,35,36,37,38,39,40,42,43, 45,46,47,48,49,50,51,52,53,59, 60,61,62,63,64,65,66,67,68,69, 70,71,72,73,74,75,76,77,78,79, 80,81,82,83,84	68

Fonte: Elaborado pelo autor

Além das abordagens de seleção de atributos tradicionais apresentadas, utiliza-se um método, denominado de Merge, que combina os atributos mais frequentes nos melhores conjuntos de seleção (LIMA, 2016). Neste método, para cada atributo será gerado um *score* e, ordenando esses *scores* será possível obter um ranking da mesma forma que qualquer

técnica individual de uma das abordagens apresentadas anteriormente. A partir deste ranking, utiliza-se como estratégia de corte, selecionar um subconjunto de atributos com frequência superior a dois. A **Figura 11** apresenta o ranking gerado, onde no eixo y são apresentados os méritos de cada atributo, calculado pela frequência de vezes que esse atributo se encontra entre os melhores conjuntos de atributos gerados.

Figura 11 - Ranking gerado pelo método de combinação de atributos Merge para o município de Teotônio Vilela.



Fonte: Elaborado pelo autor

À medida que se realizou a etapa de construção do método Merge teve-se os atributos com melhor ranking e também melhoria do tempo de processamento da base de dados, com destaque para os atributos com score entre 0,5 e 0,8; representando os atributos mais forte do conjunto de dados, conforme **Figura 11**, tendo em vista o número de ocorrências entre os algoritmos de seleção de atributos aplicados que são:

- (Questão 16): Desvio Padrão Língua Portuguesa. Refere-se ao valor estatístico do extrato da prova de português que permite visualizar o conjunto de alunos com desempenho médio ou individual de um discente;
- (Questão 18): Desvio Padrão Língua Portuguesa SAEB. Refere-se ao valor estatístico do extrato de português do resultado da prova, permitindo a tomada de decisão a respeito dos alunos com deficiência nesta disciplina, conforme descritores selecionados pelo MEC;
- (Questão 22): Em que ano você nasceu. Informação que permite acompanhar o conjunto de alunos que estão em distorção idade-série, fator que influencia bastante na evasão escolar;

- (Questão 69): Professor corrige o dever de matemática. Refere-se ao conjunto de dados de alunos que possuem tarefas de casa corrigidas pelo professor. Esses alunos possuem um melhor desempenho do que os alunos que não fazem parte do conjunto;
- (Questão 70): Você utiliza biblioteca ou sala de leitura da escola. Característica importante para os alunos com bom desempenho é o caso dos alunos que frequentam a biblioteca ou sala de leitura;
- (Questão 38): Você ver sua mãe, ou a mulher responsável por você, lendo. Verifica-se que o hábito de leitura dos pais têm melhor resultado na avaliação do SAEB;
- (Questão 62): Você já foi reprovado. Alunos com reprovação são tendenciosos a evadirem ou tirar notas baixas na avaliação, tendo em vista que possuem deficiências em um conjunto de habilidades que deveriam ser desenvolvidas;
- (Questão 61): A partir da 5ª série que tipo de escola você estudou (Pública ou Privada). Alunos de escolas privadas apresentaram resultados melhores que alunos que estudaram em escolas públicas;
- (Questão 70): Você utiliza biblioteca ou sala de leitura da escola. Alunos com hábito de leitura possuem melhores desempenhos no sistema de avaliação;
- (Questão 48): Qual frequência você lê revistas em geral. Atributos que compõem o conjunto de atributos de alunos com bom desempenho em português;
- (Questão 49): Qual frequência você ler revistas de comportamentos, celebridades, esportes ou TV. Mais uma vez a prática de leitura é evidenciada na melhoria dos resultados;
- Os demais atributos correspondem aos fatos socioeconômicos, tais como: (Questão 32) sua casa tem banheiro, (Questão 23) sua casa tem TV a cores, (Questão 25) sua casa tem vídeo cassete/DVD e (Questão 29) sua casa tem máquina de lavar roupa. Esse conjunto de fatores socioeconômicos influência no desempenho dos alunos. Nesse caso decisões de gestão com objetivo de melhorar o social dos discentes (e seus familiares) impactam de forma positiva na melhoria do resultado do IDEB para esse conjunto de dados.

Um fator fundamental para esse método é uma boa avaliação sobre os conjuntos de atributos gerados. Dessa forma, esse método somente pode ser aplicado após a utilização de um método de avaliação dos subconjuntos de atributos gerados pelas técnicas de seleção de atributos. Assim foram construídos 4 (quatro) modelos reduzidos, os quais serão introduzidos em algoritmos de classificação para validar cada modelo. Para analisar a precisão dos dados selecionados, um algoritmo de cada categoria descrita neste trabalho foi arbitrariamente escolhido dentre as opções já desenvolvidas na ferramenta Weka, foram

eles: NaiveBayes, *J48*, *JRip*, *LibSVM, RandomForest, IBK, OneR* e REPTree. Os quais foram parametrizados conforme **Tabela 13.**

Na **Tabela 16** são mostradas as precisões dos algoritmos de classificação aplicados ao nono ano nas matérias de Língua Portuguesa e Matemática para as escolas do município de Teotônio Vilela. Também é mostrada a precisão média de cada modelo reduzido gerado para que seja possível avaliar o desempenho médio da redução, usando o método de validação cruzada com *fold* de tamanho 10 e executado 30 vezes para gerar um *ranking* e, por fim, realizado o teste estatístico de Friedman e Namenyi.

Tabela 16 - Precisão dos classificadores para Língua Portuguesa e Matemática escolas públicas de Teotônio Vilela.

Algoritmo	Completo		Embutida		Filtro		Embaralhado		Todos	
	LP	MT	LP	MT	LP	MT	LP	MT	LP	MT
NaiveBayes	98,26%	98,22%	96,34%	96,36%	98,26%	98,33%	98,19%	98,35%	98,33%	98,33%
J48	99,56%	99,56%	96,12%	96,12%	99,74%	99,73%	96,50%	100%	99,68%	99,73%
JRip	98,83%	98,92%	98,95%	99,27%	99%	98,71%	99,92%	100%	98,94%	98,88%
LibSVM	100%	100%	100%	100%	100%	100%	92,33%	95,48%	100%	100%
RandomForest	99,82%	99,83%	98,75%	98,92%	99,90%	99,89%	98,69%	99,56%	99,85%	99,85%
IBK	93,86%	94,16%	95,50%	96,15%	94,98%	94,11%	98,07%	94,39%	94,05%	94,27%
OneR	100%	100%	100%	100%	100%	100%	100%	97,73%	100%	100%
REPTree	97,50%	97,73%	95,95%	95,95%	97,50%	97,73%	97,55%	97,73%	97,50%	97,73%
Precisão Média	**98,48%**	**98,55%**	**97,70%**	**97,84%**	**98,67%**	**98,56%**	**97,66%**	**97,91%**	**98,54%**	**98,60%**

LP Língua Portuguesa
MT Matemática

Fonte: Elaborado pelo autor

De acordo com a **Tabela 16**, houve empate entre "Completo", "Todos" e "Filtro". Em relação à abordagem de "Filtro", os classificadores apresentaram uma precisão média de 98,67% (Português) e 98,56% (Matemática). Já a abordagem "Todos", os classificadores apresentaram uma precisão média de 98,54% para a base de dados de Português e 98,60% para Matemática. O conjunto de dados "Completo" sem seleção de atributos possui precisão média também considerável de 98,48% (Português) e 98,55% (Matemática). A precisão média da abordagem embutida também foi muito boa 97,70% (Português) e 97,84% (Matemática) bem próxima da abordagem embaralhado 97,66% (Português) e 97,91% (Matemática). Dessa forma, é perceptível o fato de que o conjunto das abordagens forma uma seleção de atributos fortes que permitem uma acurácia de classificação alta e também um conjunto de atributos possíveis de serem discutidos como importantes.

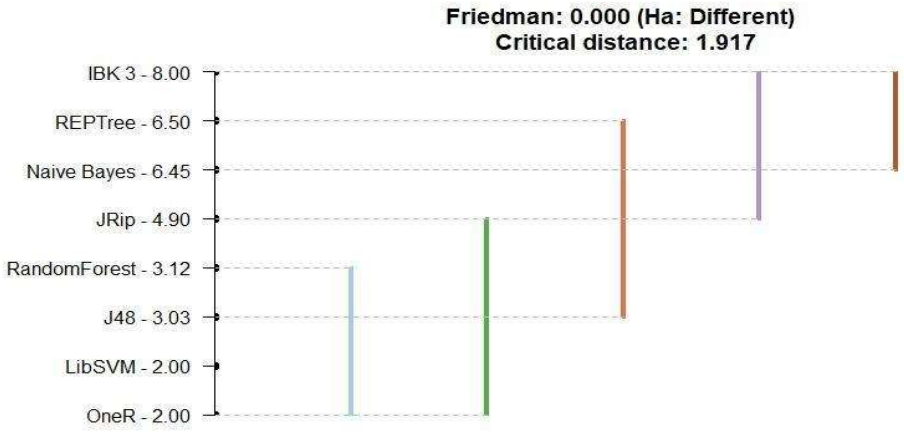
A abordagem utilizada mostra ganho de informação aproximada de 1% entre as categorias de seleção de atributos, esse ganho de informação pode parecer pequeno, mas é muito significativo se levarmos em consideração as bases de dados criadas por meio desse processo, pois através dessas tabelas podemos visualizar os atributos reduzidos e dessa forma já utilizar as informações dessa base para análise dos dados. Os resultados com acurácias

acima de 90% foram possíveis devido à qualidade do conjunto de instâncias, e ao fato de que o município estudado tem excelentes resultados no IDEB.

Dessa forma, os classificadores com resultados altos contribuem para criação do preditor que será mostrado na **Figura 13,** onde será possível inserir uma instância com configuração de um aluno para escola do município e, por meio do modelo treinado pela base, obter uma indicação com precisão acima de 90%, informando se esse discente terá IDEB "Satisfatório" (acima da média) ou "Não Satisfatório" (abaixo da média).

É importante lembrar que esse estudo mostra também que as três abordagens de seleção de atributos são boas para MDE, uma vez que o conjunto de dados reduzidos são muito próximos em seus resultados, evidenciando estudos anteriores como o de (Marquez-Vera *et al*., 2013; Lima, 2015) dependendo apenas da qualidade dos conjuntos de dados e da escolha do intervalo dos atributos mais bem ranqueados, baseado no cálculo do *score* quando na abordagem Filtro (35 atributos) em cada algoritmo dessa categoria. Em relação à abordagem Embutida (21 atributos), realizou-se a poda contemplando o objetivo desejado dessa pesquisa, ou seja, selecionar atributos com características relevantes para o desempenho acadêmico do aluno. Por último, na abordagem Embaralhado (20 atributos) também se obteve o mesmo comportamento da abordagem anterior.

Figura 12 - Aplicação do método de avaliação estatística de Friedman e Namenyi no R para comparar as saídas dos classificadores (Teotônio Vilela).



Fonte: Elaborado pelo autor

A **Figura 12** apresenta o resultado do teste estatístico de Friedman e Namenyi, aplicado à base de dados "Todos" para o município de Teotônio Vilela. Conforme pode-se observar os algoritmos OneR, LibSVM e J48 apresentaram os melhores resultados com mais de 98% de acurácia de classificação para o conjunto de dados de Português e de Matemática. Por outro lado, verifica-se que os algoritmos Naive Bayes, RandomForest e JRip

apresentaram resultados satisfatórios, enquanto o IBK e REPTree apresentou resultados mais baixos entre todos.

O valor de distância crítica de 1.917, conforme mostrado na **Figura 12** corresponde à diferença estatística entre dois algoritmos quando utilizados em um determinado conjunto de dados. Essa diferença é descoberta realizando a subtração entre os valores da colocação de dois algoritmos no ranking, se o resultado obtido for maior que a distância crítica, isto corresponde que os algoritmos são diferentes estatisticamente e que um deles realmente possui uma melhor eficácia quando aplicado em um determinado cenário. Nesse contexto, o algoritmo IBK obteve o pior ranking, sendo diferente estatisticamente dos demais algoritmos.

5ª Etapa: Avaliação. Nesta etapa, tem-se construído um ou mais modelos que aparentam ter alta qualidade. Ao final será tomada uma decisão a partir dos resultados da mineração, sem, entretanto, desconsiderar alguma questão que seja importante. Esta é a etapa na qual os conhecimentos encontrados são interpretados e utilizados em processo decisório.

Em relação à rede municipal de Maceió, é possível observar na **Tabela 14** que as precisões médias aumentaram nos modelos reduzidos, exceto pelo modelo do *REPTree*. Dentre os algoritmos selecionados, os algoritmos J48, OneR, JRip e LibSVM apresentaram os melhores resultados com 100% de acurácia de classificação para o conjunto de dados de Português e Matemática. O desempenho dos classificadores para Língua Portuguesa (90,85%) se manteve ligeiramente maior em relação ao de Matemática (88,34%), indicando uma maior relação com dados socioeconômicos. De acordo com os atributos selecionados que compuseram os modelos reduzidos (sem considerar o modelo *Todos*), os atributos que foram escolhidos mais de uma vez foram considerados como fortemente impactantes. A **Tabela 17** apresenta os atributos que tiveram maior incidência por disciplina.

Tabela 17 - Atributos com maior incidência (Maceió).

Matéria	Atributos
LP	5,6,7,8,9,12,18,19,20,21,24,25,26,27,28,29,30,35,37,38,39,40,42,58,59,68,74,79, 80,81,82
MT	5,6,7,8,12,18,19,20,21,24,25,26,27,28,29,30,35,37,38,39,40,42,57,58,59,68,74,79,80,81,82

Fonte: Elaborado pelo autor

De acordo com a **Tabela 18** são colocados como destaques os atributos referentes às questões: 7, 26, 27, 30, 57, 58, 81 e 82; que se apresentam como fortemente impactantes para as duas disciplinas. Abaixo estão os atributos que foram selecionados para avaliar o desempenho do aluno em uma dada matéria (Língua Portuguesa e Matemática) com base nos algoritmos de seleção.

Tabela 18 - Atributos mais relevantes do município de Maceió.

Atributos mais relevantes para Português	Atributos mais relevantes para Matemática
▪ **Dependência administrativa** (pública - federal, estadual e municipal - e privada localização (urbana e rural); e área (capital e interior)). Este atributo envolve diversos fatores (entre eles, o geográfico, o econômico, o social e o cultural do meio ambiente) que interfere diretamente no desempenho escolar dos alunos.	▪ **Dependência administrativa** (pública - federal, estadual e municipal - e privada localização (urbana e rural); e área (capital e interior)). Foi observado que alunos que estudaram anteriormente em escolas particulares contribuem de maneira expressiva na melhoria do IDEB.
▪ **Proficiência em Língua Portuguesa** (interpretação pedagógica do desempenho nas avaliações, desempenho do aluno em leitura). A avaliação do desempenho do aluno em leitura é fundamental para se ter fluência na Língua Portuguesa, tanto na parte gramatical quanto no vocabulário.	▪ **Proficiência em Matemática**. A avaliação do desempenho do aluno em Matemática é fundamental para se ter domínio e base nesta disciplina.
▪ **Seus pais ou responsáveis incentivam você a estudar**. O incentivo dos pais ao estudo dos filhos é fundamental para que os mesmos comecem a ter um hábito regular de estudo.	▪ **Seus pais ou responsáveis incentivam você a estudar**. Quanto à análise dos incentivos aos estudos, ficou claro que é o fator que mais pesa para obtenção de um melhor IDEB.
▪ **Você gosta de estudar Língua Portuguesa**. O aluno tem que gostar da disciplina de Português para obter um bom aproveitamento. Este ponto envolve o compromisso e a formação docente que despertem e motivem os alunos para a disciplina de Língua Portuguesa.	▪ **Qual é o seu sexo**. A seleção deste atributo reafirma os conceitos de preconceito estrutural. Percebe-se que os discentes do sexo feminino têm um destaque maior na melhoria do IDEB.
▪ **Você faz o dever de casa de Língua Portuguesa.** O aluno deve destinar um tempo diário para o estudo e para realizar o dever de casa de Português, aprimorando o seu vocabulário bem como a parte gramatical da linguagem.	▪ **Você faz o dever de casa de matemática**. Criando esse hábito de estudo o aluno conseguirá um alto desempenho escolar em matemática.

Fonte: Elaborado pelo autor

É possível ver que a estratégia de seleção de atributos usada por categorias como seleção embutida, filtro e embaralhada, combinadas ao modelo de ranking Merge, permitiu evidenciar os melhores atributos do conjunto de 88 para 31, redução de 65% dos atributos

mais ainda durante essa etapa manteve-se os dados socioeconômico e os extratos dos resultados dos alunos nas provas ANEB, Prova Brasil e ANA, permitindo assim fazer uma correlação entre os dois tipos de dados. Esses atributos permitem verificar de imediato a tendência dos alunos para o sucesso ou não do resultado no IDEB. À medida que se realizou a etapa de construção do método Merge teve-se os atributos com melhor ranking e também melhoria do tempo de processamento da base de dados, com destaque para os atributos com *score* entre 0,5 e 0,6 que são: (1) Com qual frequência você costuma ir à biblioteca; (2) Quando você entrou na escola; (3) Localização da escola; (4) Extrato da Avaliação Nacional da Educação Básica (ANEB); (5) Desvio Padrão em Língua Portuguesa; (6) Até que série seu pai, ou o homem responsável por você, estudou; (7) Seu pai, ou o homem responsável por você, sabe ler e escrever e (8) Em dias de aula, quanto tempo você gasta fazendo trabalhos domésticos.

Conforme se verifica o processo de Merge foi uma validação da etapa de seleção de atributos por categoria de técnicas. O que se pode provar a partir desse processo é que ao juntar diferentes estratégias, obtém-se ganho de atributos valiosos para a base de dados, auxiliando na melhor correção entre os atributos que envolvem o problema estudado. A aplicação de diferentes estratégias de seleção resultou em uma base de dados com mais atributos, contribuindo assim para entender melhor os aspectos socioeconômicos e cognitivos que envolvem o conjunto de dados estudado.

Em relação à rede municipal de Teotônio Vilela, é possível observar na **Tabela 16** que entre os algoritmos selecionados, os algoritmos OneR, LibSVM e J48 apresentaram os melhores resultados com quase 99% de acurácia de classificação para o conjunto de dados de Português e Matemática. De acordo com os atributos selecionados que compuseram os modelos reduzidos (sem considerar o modelo *Todos*), os atributos que foram escolhidos mais de uma vez foram considerados como fortemente impactantes. A **Tabela 19** apresenta os atributos que tiveram maior incidência por disciplina.

Tabela 19 - Atributos com maior incidência (Teotônio Vilela).

Matéria	Atributo
LP	1,5,6,7,8,10,12,15,16,17,18,19,21,22,23,24,25,26,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,45,46,47,48,49,50,51,52,53,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84
MT	5,6,7,8,9,10,12,15,16,17,18,19,21,22,23,25,26,28,29,30,31,32,33,34,35,36,37,38,39,40,42,43,45,46,47,48,49,50,51,52,53,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84

Fonte: Elaborada pelo autor

De acordo com a **Tabela 19** observa-se que as questões 1, 24, 41, 57 e 58; são atributos exclusivamente referentes à Língua Portuguesa. Enquanto o atributo 9 é exclusivamente referente a disciplina de Matemática. Na **Tabela 20** estão os 18 atributos (6,8,10,16,18,22,23,25,29,32,34,38,48,49,61,62,69,70) que foram selecionados para avaliar o desempenho do aluno em uma dada matéria (Língua Portuguesa e Matemática) com base nos algoritmos de seleção e *score* acima de 0,5 conforme apresentado na **Figura 12**.

Tabela 20 - Atributos mais relevantes do município de Teotônio Vilela.

Atributos mais relevantes para Português/ Matemática	
▪ ID TURNO	▪ Sua casa tem banheiro?
▪ ID Caderno	▪ Incluindo você, quantas pessoas vivem na sua casa?
▪ ID Bloco 2	▪ Você ver sua mãe, ou a mulher responsável por você, lendo?
▪ Desvio Padrão Língua Portuguesa:	▪ Qual frequência você lê revistas em geral?
▪ Desvio Padrão Língua Portuguesa SAEB:	▪ Qual frequência você ler revistas de comportamentos, celebridades, esportes ou tv?
▪ Em que ano você nasceu?	▪ A partir da 5a série que tipo de escola você estudou (Pública ou Privada)?
▪ Sua casa tem TV a cores?	▪ Você já foi reprovado?
▪ Sua casa tem vídeocassete e/ou DVD?	▪ Professor corrige o dever de Matemática?
▪ Sua casa tem máquina de lavar roupa?	▪ Você utiliza biblioteca ou sala de leitura da escola?

Fonte: Elaborada pelo autor

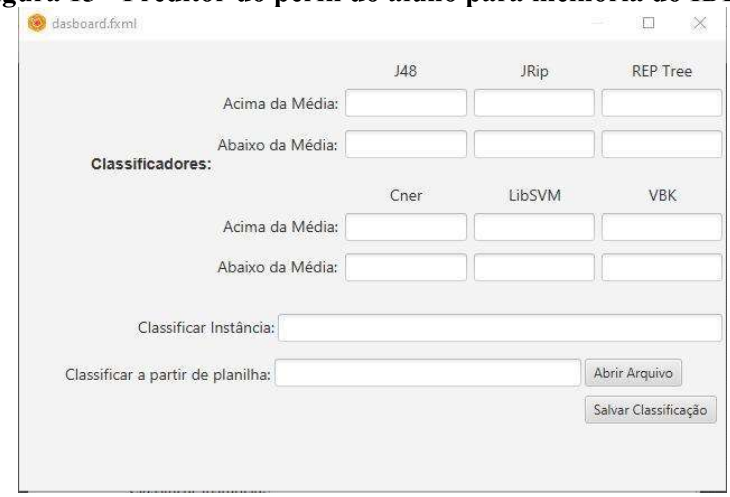
É possível ver que a estratégia de seleção de atributos usada por categorias como embutida, filtro e embaralhada, combinadas ao modelo de ranking Merge, permitiu evidenciar os melhores atributos do conjunto de 91 para 71 (língua portuguesa), redução de 22% dos atributos. Ainda durante essa etapa manteve-se os dados socioeconômico e os extratos dos resultados dos alunos nas provas ANEB, Prova Brasil e ANA, permitindo assim fazer uma correlação entre os dois tipos de dados. Esses atributos permitem verificar de imediato a tendência dos alunos para o sucesso ou não do resultado no IDEB. À medida que se realizou a etapa de construção do método Merge obteve-se os atributos com melhor ranking e também melhoria do tempo de processamento da base de dados, com destaque para os atributos com *score* entre 0,5 e 0,6.

Verifica-se que o processo de Merge foi uma validação da etapa de seleção de atributos por categoria de técnicas. O que se pode provar a partir desse processo é que ao juntar diferentes estratégias, obtém-se ganho de atributos valiosos para a base de dados,

auxiliando na melhor correção entre os atributos que envolvem o problema estudado. A aplicação de diferentes estratégias de seleção resultou em uma base de dados com mais atributos, contribuindo assim para entender melhor os aspectos socioeconômicos e as habilidades que envolvem o conjunto de dados estudado.

Com base nos dados obtidos pela etapa de pré-processamento dos dados e de seleção de atributos, foi gerado um preditor conforme **Figura 13** com objetivo de encontrar os perfis de alunos para obtenção de um resultado "Satisfatório" ou "Não Satisfatório" no IDEB. Dessa forma, utilizou-se a base de dados de 2015 como instâncias de treinamento e um conjunto de instâncias correspondente ao perfil do aluno com foco nos atributos elencados na **Tabela 20** e aplicando essas instâncias de teste retiradas da base de dados de 2017 no preditor obteve-se a classificação e características dos alunos.

Figura 13 - Preditor do perfil do aluno para melhoria do IDEB.



Fonte: Elaborado pelo autor

Com base na **Tabela 21** pode-se visualizar na coluna classificadores, os algoritmos que melhor classificam o conjunto de instâncias na predição visando a melhoria do resultado do IDEB. Nesse processo aplica-se a rejeição de classificadores em que excluímos os classificadores com acurácias baixas (REPTree, IBK e Naive Bayes), estratégia comumente utilizada na mineração de dados que classificaram as instâncias de teste, possibilitando encontrar os perfis de alunos.

Tabela 21 - Classificação do perfil do aluno e Rejeição de classificadores (Teotônio Vilela).

Instância do Aluno(Perfil do aluno)	Classificação no IDEB Satisfatório/ Não satisfatório	Classificadores	Precisão média dos classificadores
1	Satisfatório	J48,OneR, LibSVM e JRip	99,65%

2	Satisfatório	J48,OneR, LibSVM e RandomForest	98,94%
3	Não satisfatório	J48,OneR, LibSVM e JRip	99,65%
4	Não satisfatório	J48,OneR, LibSVM e RandomForest	98,94%

Fonte: Elaborado pelo autor

Após todo processo de treinamento do modelo de predição pode ser testado esse modelo na busca de predizer através da base de treinamento reduzida "Todos" se de fato é possível encontrar o perfil de um determinado aluno, podendo classificá-lo como IDEB satisfatório ou não satisfatório. Este estudo comprovou que a aplicação para predição do resultado do IDEB 2015 (Treino) e 2017 (Teste) funciona e conseguimos identificar alguns perfis de alunos que podem contribuir no melhor resultado do IDEB. A ferramenta está disponível para ser utilizada por usuários de Alagoas através do site do grupo de pesquisa em www.edumach.com.br. Salienta-se que esse preditor é específico para o município de Teotônio Vilela-Alagoas, dispondo apenas sobre as características dos alunos da rede pública deste município.

É possível ver que a estratégia de seleção de atributos usada por categorias como seleção embutida, filtro e embaralhada, combinadas ao modelo de ranking Merge, permitiu evidenciar os melhores atributos do conjunto de 91 para 71 (língua portuguesa), redução de 22% dos atributos mais ainda durante essa etapa manteve-se os dados socioeconômico e os extratos dos resultados dos alunos nas provas ANEB, Prova Brasil e ANA, permitindo assim fazer uma correlação entre os dois tipos de dados. Esses atributos permitem verificar de imediato a tendência dos alunos para o sucesso ou não do resultado no IDEB. À medida que se realizou a etapa de construção do método Merge teve-se os atributos com melhor ranking e também melhoria do tempo de processamento da base de dados, com destaque para os atributos com *score* entre 0,5 e 0,8. que são: (16) Desvio Padrão Língua Portuguesa; (18) Desvio Padrão Língua Portuguesa SAEB; (22) Em que ano você nasceu; (69) Professor corrige o dever de Matemática; (70) Você utiliza biblioteca ou sala de leitura da escola; (38) Você ver sua mãe, ou a mulher responsável por você, lendo; (62) Você já foi reprovado; (61) A partir da 5a série que tipo de escola você estudou (Pública ou Privada); (70) Você utiliza biblioteca ou sala de leitura da escola; (48) Qual frequência você lê revistas em geral; (49) Qual frequência você ler revistas de comportamentos, celebridades, esportes ou tv; os demais atributos correspondem aos fatos socioeconômicos, tais como: (32) Sua casa tem banheiro,

(23) sua casa tem TV a cores, (25) Sua casa tem vídeo cassete/DVD e (29) sua casa tem máquina de lavar roupa.

Conforme se verifica o processo de Merge foi uma validação da etapa de seleção de atributos por categoria de técnicas. O que se pode provar a partir desse processo é que ao juntar diferentes estratégias, obtém-se ganho de atributos valiosos para a base de dados, auxiliando na melhor correção entre os atributos que envolvem o problema estudado. A aplicação de diferentes estratégias de seleção resultou em uma base de dados com mais atributos, contribuindo assim para entender melhor os aspectos socioeconômicos e cognitivos que envolvem o conjunto de dados estudado.

Figura 14 - Perfil do aluno encontrado pelo preditor para melhoria do IDEB com utilizando a base de dados "Todos" e o Método Merge.

B	C	D	E	F
PERFIL	ID_TURNO	ID_CADERNO	ID_BLOCO_2	DESVIO_PADRAO_LP
1	2	5	6	0.35591
2	2	19	6	0.327086
PERFIL	Na sua casa tem banheiro?	Incluindo você, quantas pessoas vivem atualmente em sua casa?	Você vê sua mãe, ou a mulher responsável por você, lendo?	Com qual frequência você lê: Livros de literatura.
1	B: Sim um	D: Quatro	A: Sim	B: De vez em quando
2	B: Sim um	D: Quatro	A: Sim	B: De vez em quando

H	I	J	K	L	M
PERFIL	PROFICIENCIA_LP_SAEB	Em que ano você nasceu?	Na sua casa tem televisão em cores?	Na sua casa tem videocassete e/ou DVD?	Na sua casa tem máquina de lavar roupa?
1	248.943.670.000.000.000	C: 2001	B: Sim uma	B: Sim um	A: Não
2	314.456.841.000.000.000	B: 2002	B: Sim uma	B: Sim um	A: Não
PERFIL	Com qual frequência você lê: Revistas em geral.	A partir da quinta série ou sexto ano, em que tipo de escola você	Você já foi reprovado?	O(A) professor(a) corrige o dever de casa de Matemática?	Você utiliza a biblioteca ou sala de leitura da sua escola?
1	B: De vez em quando	A: Escola pública.	A: Não	A: Sempre.	B: De vez em quando
2	B: De vez em quando	A: Escola pública.	A: Não	A: Sempre.	B: De vez em quando

PERFIL	ID_TURNO	ID_CADERNO	ID_BLOCO_2	DESVIO_PADRAO_LP
3	1	6	7	0.314075
4	2	12	7	0.29506
PERFIL	TX_RESP_Q014	Incluindo você, quantas pessoas vivem atualmente em sua casa?	Você vê sua mãe, ou a mulher responsável por você, lendo?	Com qual frequência você lê: Livros de literatura.
3	C: Sim, dois.	E: Cinco.	B: Não	B:De vez em quando.
4	C: Sim, dois.	F: Seis pessoas ou mais.	B: Não	A: Sempre

PERFIL	PROFICIENCIA_LP_SAEB	Em que ano você nasceu?	Na sua casa tem televisão em cores?	Na sua casa tem videocassete e/ou DVD?	Na sua casa tem máquina de lavar roupa?
3	217.187.144.000.000.000	E: 1999.	A:Não.	A:Não.	A: Não tem
4	216.710.049.000.000.000	B: 2002	B: Sim, uma.	B: Sim, uma.	B: Sim, uma.
PERFIL	Com qual frequência você lê: Revistas em geral.	A partir da quinta série ou sexto ano, em que tipo de escola você estudou?	Você já foi reprovado?	O(A) professor(a) corrige o dever de casa de Matemática?	Você utiliza a biblioteca ou sala de leitura da sua escola?
3	B: De vez em quando	A: Escola pública	A: Sim, uma vez.	A: Sempre	B: De vez em quando
4	C: Nunca	A: Escola pública	A: Não	A: Sempre	B: De vez em quando

Fonte: Elaborado pelo autor

Através da **Figura 14** é possível validar os perfis de alunos obtidos pela classificação por meio do preditor construído através do desenvolvimento dessa pesquisa, nesse processo também se verifica que as informações da **Tabela 21** são evidenciadas por meio dos classificadores e que é possível predizer quais atributos contribuem para a melhoria do resultado do IDEB. Na **Tabela 21** são identificados os classificadores que atribuíram resultado satisfatório para as instâncias de alunos da base de dado de 2017 (Teste) por meio das instâncias 1 e 2, e resultado de IDEB não Satisfatório para a instância 3 e 4.

É importante lembrar que existem outros perfis de alunos que podem ser identificados dentro das bases de dados de Maceió e também de Teotônio Vilela e que para encontrar as demais características teríamos que submeter mais instâncias de alunos ao preditor criado e

ao verificarmos o resultado da classificação analisar os valores dos atributos existentes, sendo que a etapa de seleção de atributos realizada neste trabalho torna esse trabalho bem menos oneroso uma vez que o número de atributos irrelevantes diminuiu consideravelmente com o uso das abordagens por categoria e, por fim, o uso do método Merge.

3.4 Análise comparativa entre os municípios de Maceió e Teotônio Vilela

Durante esse processo de investigação foi possível visualizar a diferença entre os dois municípios e a princípio mesmo utilizando os mesmos métodos para ambos os conjuntos de dados assim como os mesmo algoritmos de seleção de atributos e classificadores, o comportamento das bases foram diferentes entre os municípios, mostrando através dos atributos selecionados que os perfis de alunos que influenciam na melhoria do IDEB são distintos, tendo em vista que na cidade de Maceió o fator infraestrutura evidenciado pelos atributos dependências administrativas e as *features* relacionadas a interesse do discente pelas disciplinas de Português e Matemática, assim como a influência de seus pais ao processo ficaram evidentes no estudo aqui apresentado como também os níveis de proficiência desses alunos nos exames do SAEB são mais baixos que os do município de Teotônio Vilela.

Tabela 22 - Comparação entre Maceió x Teotônio Vilela

Município	Número de Instâncias	Melhor algoritmo de classificação	Pior algoritmo de classificação	Número de atributos relevantes
Maceió	1.320	J48,LibSVM,OneR e JRip	REPTree	31
Teotônio Vilela	572	J48,LibSVM,OneR e RandomForest	IBK	18

Fonte: Elaborado pelo autor

Conforme a **Tabela 22,** tem-se uma síntese comparativa entre os dois municípios onde destaca-se as mudanças de comportamento das as bases de dados, assim como das características dos discentes apresentadas por meio dos atributos identificados em Maceió (31) e Teotônio Vilela (18). Outro importante ponto encontrado é o comportamento dos Algoritmos que tiveram melhor acurácia de classificação em Maceió (JRip) e em Teotônio Vilela (Random Forest) e os algoritmos com pior resultado de classificação foram REPTree (Maceió) e IBK (Teotônio Vilela).

Vale ressaltar também que para o município de Maceió o comportamento dos classificadores foram menores para a base de dados na abordagem Embutido e Embaralhado o que implica em dizer que essa base de dados se encontrava muito mais desorganizada em

termos de instâncias disponíveis para classificação, o fato que foi corrigido com a utilização do algoritmo SMOTE e também com processo de Randomização e exclusão de atributos e instâncias imprestáveis mesmo como todo o tratamento dos dados.

Os resultados mais relevantes para o teste estatístico de Friedman e Namenyi foi a base de dados "Todos" que contêm os atributos que mais influenciam no melhor desempenho do IDEB para o Município de Maceió, também é importante mencionar que esse comportamento que se aplicou a base "Todos" através da inserção do método de ranqueamento Merge provou ser bastante eficiência para ambas as bases de dados, pois a junção dos diferentes métodos de seleção de atributos são de fato positivos para a investigação das melhores *features* para melhoria do IDEB de Maceió em relação a outra cidade.

Ainda se tratando sobre a diferença em ambos os municípios foi possível analisar a respeito da cidade de Teotônio Vilela, a influência dos atributos socioeconômicos muito mais forte que no município de Maceió, situação que já se imaginava devido ao fato de que por se tratar de um município do interior o acesso a recursos é muito menor e o nível social dos estudantes é diferente dos discentes da capital, nesse caso tivemos atributos como: quantidade de banheiro em casa, aparelhos eletrônicos e eletrodomésticos disponíveis em casa que discorrem sobre as característica do aluno e a falta desses equipamentos em casa demonstram o nível social deles que quanto mais baixos menor é o resultado no IDEB.

Já em se tratando do comportamento da base dados para os algoritmos de seleção de atributos devido ao fato a base de dados de Teotônio Vilela está bem mais organizada necessitou de poucos ajustes no que se trata do uso do SMOTE, randomização e da exclusão de atributos insignificantes. Dessa forma, tivemos um processo de seleção com maior número de atributos e também como escolhemos um limite de corte para os algoritmos de ranqueamento do tipo filtro que no processo de Maceió foi de valor 30 (mais bem ranqueados) onde encontramos mais atributos que desrespeitam a influência de decisões para o gestor, enquanto nos dados de seleção para Teotônio Vilela ampliamos a escolha dos atributos para os algoritmos do tipo filtro para o valor 35 (mais bem ranqueados), possibilitando avançar na compreensão dos atributos mais relacionados com perfil pedagógico aluno assim como também o social. Essa mudança durante o processo de MDE foi bastante importante, pois mostra que um mesmo conjunto de dados dá condições de estudarmos para tomada de decisão tanto a nível de gestão, quanto para os professores que estão no dia a dia com os alunos, atacando as habilidades a serem desenvolvidas.

No entanto como a prova SAEB é um sistema de avaliação é necessário atacar bem mais que apenas a aprendizagem como o foi o caso apresentado mais fortemente na de base de dados de Maceió e também visível na de Teotônio Vilela.

Em relação ao comportamento dos classificadores, a base de dados de Teotônio Vilela teve melhores resultados que a de Maceió, provando que as abordagens (Embutido, Embaralhado e Filtro) para a seleção de atributos combinadas ao método Merge, possibilitam encontrar excelentes atributos de melhoria do resultado do IDEB, possibilitando encontrar as características dos alunos a serem atacadas posteriormente por seus professores ou gestores na tentativa de melhoramento dos índices de desenvolvimento da educação do município para ambos os casos.

No mais durante esse processo foram aplicados métodos equivalentes que se mostram diferentes devido aos aspectos distintos dos municípios, mas sendo eficientes para seleção de atributos relevantes na melhoria do IDEB e também na classificação da predição para identificação dos fatores que influenciam o desempenho escolar na rede de ensino básico.

4 CONSIDERAÇÕES FINAIS

Neste trabalho foi mostrado que a combinação de diferentes categorias de seleção de atributos torna possível obter um conjunto de atributos ainda melhor que usar apenas um tipo de categoria de seleção de atributos. Essa estratégia enriquece a base de dados, tornando possível encontrar atributos que são imprescindíveis para análise de dados. A base de dados "Todos" representa esses atributos, pois combina as três abordagens de seleção de atributos.

Para a rede pública do município de Maceió foi descoberto 31 (trinta e um) atributos que influenciam o desempenho escolar do aluno nas duas disciplinas: língua portuguesa e matemática. Foram também analisados os principais algoritmos com o objetivo de identificar a melhor precisão na classificação dos atributos via mineração de dados, além de identificar a diferença estatística entre os algoritmos: J48, OneR, JRip, LibSVM, RandomForest, IBK, Naive Bayes e o REPTree. A melhor abordagem verificada foi "Todos" que representa o conjunto de atributos após seleção dos mesmos, utilizando as três abordagens (embutida, filtro e embaralhado). No processo de avaliação foi empregado testes estatísticos de Friedman e Nemenyi implementado na linguagem R. O resultado do teste estatístico de Friedman e Nemenyi, aplicado à base de dados "Todos" comprova que os algoritmos J48, OneR, JRip e LibSVM apresentaram os melhores resultados com 100% de acurácia de classificação para o conjunto de dados de português. Por outro lado, verifica-se que os algoritmos RandomForest, IBK e Naive Bayes apresentaram resultados satisfatórios, enquanto o REPTree apresentou o pior resultado entre todos.

Para a rede pública do município de Teotônio Vilela ficou evidenciado neste estudo que ao aplicar o método Merge visualiza-se o comportamento dos atributos na base "Todos", buscando junto ao *score* estabelecido através do conjunto de técnicas de seleção quais são os atributos mais relevantes ao estabelecer valores superiores a 0,5 sendo possível encontrar informações sobre as características acadêmicas e socioeconômicas dos discentes. Também foi possível verificar ao aplicar o processo de seleção encontrar atributos que possibilitam aos gestores e/ou professores identificar dificuldades na obtenção de conhecimento dos alunos através de informações presentes na base de dados do IDEB como desvio padrão e médias obtidas em anos anteriores no conjunto de provas que faz parte do sistema de avaliação do SAEB. Neste estudo foi descoberto 18 (dezoito) atributos que influenciam mais

fortemente o desempenho escolar do aluno nas duas disciplinas: língua portuguesa e matemática. Foram também analisados os principais algoritmos com o objetivo de identificar a melhor precisão na classificação dos atributos através da mineração de dados, além de identificar a diferença estatística entre os algoritmos: J48, OneR, JRip, LibSVM, RandomForest, IBK, NaiveBayes e o REPTree. Os algoritmos OneR, LibSVM e J48 apresentaram os melhores resultados com mais de 98% de acurácia de classificação para o conjunto de dados de português e de matemática.

O mais interessante é que para realizar esse processo de identificação não seria necessário conhecimento aprofundado em MDE, pois como mostrado na etapa de preparação de dados do CRISP-DM, pode-se identificar por meio de uma ferramenta como Anaconda todas as bases de dados e com isso os valores de seus respectivos atributos e a partir dessa visualização inicia-se um processo de investigação do que representa cada atributo, utilizando essa etapa para atacar os problemas reais. Sendo que nas etapas seguintes do método aqui aplicado consegue-se demonstrar que cada vez mais o processo de seleção de atributos permite encontrar atributos valiosos para tomada de decisões e que esse procedimento ajuda na melhoria da interpretação dos dados, assim como na diminuição do tempo de processamento e também na melhoria da acurácia de predição.

Neste trabalho foi mostrado que a combinação de diferentes categorias de seleção de atributos torna possível obter um conjunto de atributos ainda melhor que usar apenas um tipo de categoria de seleção de atributos. A base de dados "Todos" representa esses atributos, pois combina as três abordagens de seleção de atributos.

Após todo processo de treinamento do modelo de predição pode ser testado esse modelo na busca de predizer através da base de treinamento reduzida "Todos" se de fato é possível encontrar o perfil de um determinado aluno, podendo classificá-lo como IDEB satisfatório ou não satisfatório. Este estudo comprovou que a aplicação para predição do resultado do IDEB 2015 (Treino) e 2017 (Teste) funciona e conseguimos identificar alguns perfis de alunos que podem contribuir no melhor resultado do IDEB. A ferramenta está disponível para ser utilizada por usuários de Alagoas através do site do grupo de pesquisa em www.edumach.com.br.

Salienta-se que esse preditor é específico para o município de Teotônio Vilela-Alagoas, dispondo apenas sobre as características dos alunos da rede pública deste município. Para trabalhos futuros pretende-se ampliar o preditor criado aqui, colocando disponível no mesmo todas as etapas do CRISP-DM e dessa forma ao escolhermos a base de dados ainda sem tratamento de seleção de atributos poderemos conseguir automatizar esse processo

através da ferramenta Anaconda podendo inclusive disponibilizar para qualquer município do Brasil as informações dos atributos que devem ser atacados para melhoria do IDEB e, com isso, contribuir com a compreensão dos perfis de alunos envolvidos nesse sistema de avaliação através da MDE.

Também como trabalhos futuros podem-se aplicar outros tipos de técnicas de IA para alcançar resultados ainda melhores a respeito desses alunos suas habilidades desenvolvidas ou ainda não desenvolvidas exigidas pelo sistema de avaliação aplicando técnicas como algoritmos genéticos para mutar as bases de dados e encontrar possíveis alunos vencedores com objetivo de melhorar os resultados do exame por parte das instituições de ensino.

Neste trabalho foi mostrado que muitos dos atributos podem ser descartados para avaliar o desempenho do aluno em uma dada matéria, e que tal avaliação, as seleções tiveram um consenso geral para a maioria dos atributos. Que procedimento de seleção criar uma base de dados reduzida passível de compreender melhor os alunos envolvidos no sistema de avaliação para atacar os problemas desses alunos em busca de melhores resultados e também foi possível criar preditor para testar o processo de seleção de atributos criados ao visualizar os perfis de alunos com maiores chances de bons resultados foi possível enxergar a influência do fator socioeconômico o que impacta ainda mais na busca por melhores resultados tendo em vista a necessidade de investimento por parte dos responsáveis também na qualidade de vida dos discentes assim como compreender atributos que estão em ênfase hoje como a escola integral e a influência dos incentivos dos responsáveis para o bom desempenho dos alunos, entre outros atributos aqui obtidos pela MDE que são ricos para a tomada de decisão dos gestores.

Como contribuições do trabalho destacam-se a metodologia empregada nos testes e os resultados que demonstram quais os melhores algoritmos a ser empregados em um sistema real para classificação dos atributos relevantes para melhoria do IDEB. Para estudos futuros é possível acrescentar mais dados socioeconômicos, tais como: categorizar alunos por bairro, cursos e estudos adjacentes aos da escola, entre outros fatores que afetam o desenvolvimento, atenção e dedicação dos mesmos.

O estudo de caso serviu para perceber o quanto cada atributo influência na determinação da classe IDEB, influenciando assim na melhoria dos índices nas escolas públicas do estado de Alagoas. A partir deste trabalho os gestores poderão utilizar estas informações e refleti-las para o melhoramento da gestão educacional, da organização do trabalho pedagógico e no processo de ensino e aprendizagem nas escolas públicas.

REFERÊNCIAS

BAKER, Ryan Shaun; INVENTADO, Paul Salvador. Educational data mining and learning analytics. In: Learning analytics, p. 61-75. Springer New York, 2014.
BAKER, Ryan. Data mining for education. International encyclopedia of education, v. 7, n. 3, p. 112-118, 2010.

BEZERRA, C.; SCHOLZ, R.; ADEODATO, P.; PONTES, T.; SILVA, I. **Evasão Escolar: Aplicando Mineração de Dados para Identificar Variáveis Relevantes.** V Congresso Brasileiro de Informática na Educação (CBIE 2016). Anais do XXVII Simpósio Brasileiro de Informática na Educação (SBIE 2016). 2016.

BORGES, José Estevam Vilar, et al. Um estudo exploratório sobre os aspectos que influenciam na eficiência da predição de sucesso acadêmico de alunos de programação introdutória. Dissertação (Dissertação em Ciência da Computação), Universidade Federal de Alagoas, p.25. 2019.

BRITO, Daniel Miranda; PASCOAL, Túlio Albuquerque; ARAÚJO, Jairo Gustavo G. de O.; LEMOS, Marcílio O. Rêgo; THAÍS Gaudêncio. Identificação de Estudantes do Primeiro Semestre com risco de Evasão através de Técnica de Data Mining. Nuevas Ideas en Informatica Educativa. TISE. 2015.

CALIXTO, Kennet E. A.; SEGUNDO, Caetano V. N.; Renê, P. de Gusmão. Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. VI CBIE. SBIE – 2017.

CASTRO, LEANDRO NUNES; FERRARI, DANIEL GOMES. Introdução a Mineração de Dados. Editora Saraiva, 2016.

COELHO, V. C. G.; COSTA, J. P. C. L. da. **Mineração de dados educacionais no ensino a distância governamental.** In: Conferências Ibero-Americanas WWW/Internet e Computação Aplicada. Brasília, Brasil: [s.n.], 2016. p. 77–84.

COELHO, V. C.; COSTA, J. P. C. L.; SOUSA, D. C. R.; CANEDO, E. D.; SILVA, D. G.; SOUSA JÙNIOR, R. T. **Mineração de dados educacionais para identificação de barreiras na utilização da educação a distância**. ENAP. Ministério do Planejamento, Orçamento e Gestão, Brasília – DF, 2015.

COSTA, Evandro et al. Mineração de Dados Educacionais: Conceitos, Técnicas, Ferramentas e Aplicações. Jornada de Atualização em Informática na Educação, v. 1, n. 1, p. 1-29, 2012.

DASH, M., LIU, H. Consistency-based search in feature selection. Articial Intelligence 151 (2003).

Diniz, F. A., de Moura Silva, F. R., da Costa, R. D., Reis, T., da Silva, Í. B. G. D., & Moura, A. F. C. ANÁLISE ESPACIAL DOS ÍNDICES EDUCACIONAIS DO RIO GRANDE DO NORTE COM O USO DE TÉCNICAS DE MINERAÇÃO DE DADOS.

FONSECA, Stella Oggioni da; NAMEN, Anderson Amendoeira. Mineração em bases de dados do Inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. Educação em Revista, 2016, 32.1: 133-157.

Freitas Júnior, O. G., Rodrigues, W. R. M., Barbirato, J. C. C., & de Barros Costa, E. (2019). Melhoria da gestão escolar através do uso de técnicas de mineração de dados educacionais: um estudo de caso em escolas municipais de Maceió. RENOTE-Revista Novas Tecnologias na Educação, 17(1), 296-305.

FRIMDEAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. Machine learning, Springer, 1997.

GUYON, I., ELISSEEFF, A. An introduction to variable and feature selection. Journal of Machine Learning Research 3 (2003).

HAN, Jiawei; KAMBER, Micheline. Data Maning: Concepts and Techniques. San Francisco: Morgan Kaufmanni, 2011.

INEP. **Ideb**. 2019. Acesso em: 31 Janeiro 2019. Disponível em:http://portal.inep.gov.br/ideb.
INEP. **Prova Brasil**. Sistema de Avaliação da Educação Básica (Saeb). Disponível em: http://provabrasil.inep.gov.br/. Acessado em: 10 de setembro de 2016.

INEP/MEC. **Indicadores da Qualidade da Educação**. São Paulo: ação educativa, 2007.
Laisa, J.; Nunes, I. (2015, October). 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) (Vol. 26, No. 1, p. 1112).

LIMA, R. A. F. et al. **Estratégias de seleção de atributos para detecção de anomalias em transações eletrônicas**. Dissertação (Dissertação em Ciência da Computação), Universidade Federal de Minas Gerais, p. 25. 2016.

LORENA, A. C. & CARVALHO, A. C. P. L. F. (2002). As máquinas de vetores suporte, relatório interno, Laboratório de Inteligência Computacional, ICMC/USP, São Carlos – SP, junho.

MANHÃES, L. M. B. **Predição do desempenho acadêmico de graduandos utilizando mineração de dados educacionais.** Tese (Doutorado em Engenharia de Sistemas e Computação), Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2015.

MÁRQUEZ-VERA, C.; Morales, C. R.; Soto, S. V. ***Predicting School Failure and Dropout by Using Data Mining Techniques.*** IEEE Journal of Latin American Learning Technologies, Vol. 8, no. 1, February, 2013.

MAYAPÉREZ, Petra Norma; AGUILAR, C. Jorge R; ZAMORA, R. Rosa A.; BARRON, A. J. Miguel. Diseño de un Modelo predictivo aplicando Minería de Datos para identificar causas de Deserción Estudiantil Universitaria Predictive Model Design applying Data

Mining to identify causes of Dropout in University Students. Technology & Society, vol. 7 (2018). 11-39. 2018.

NASCIMENTO, R. L. S.; Cruz Junior, G. G; Fagundes, R. A. A. **Mineração de Dados Educacionais: Um estudo sobre indicadores da educação em bases de dados do INEP**. Novas Tecnologias na Educação, CINTED, UFRGS, 2018.

PAIVA, R.; BITTENCOURT, I. I.; PACHECO, H.; DA SILVA, A. P.; JACQUES, P.; ISOTANI, S. **Mineração de dados e a gestão inteligente da aprendizagem: desafios e direcionamentos.** Instituto de Computação – Universidade Federal de Alagoas (UFAL), Alagoas – AL, 2012.

PAIVA, Ranilson; BITTENCOURT, Ig Ibert; DA SILVA, Alan Pedro. Uma ferramenta para recomendação pedagógica baseada em mineração de dados educacionais. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2013.

PASTA, A. **Aplicação da técnica de data mining na base de dados do ambiente de gestão educacional: um estudo de caso de uma instituição de ensino superior de Blumenau-sc.** Dissertação (Mestrado em Computação Aplicada) Universidade do Vale do Itajaí, São José-SC, 2011.

PATRÍCIO, T. S.; MAGNONI, M. da G. M. **Mineração de dados e big data na educação**. In: Revista GEMInIS. São Carlos, Brasil: [s.n.], 2018. p. 57–75.

PINTO, GLEVSON DA SILVA; FREITAS JÚNIOR, OLIVAL DE GUSMÃO; COSTA, EVANDRO DE BARROS; BARBIRATO, JOÃO CARLOS CORDEIRO; RODRIGUES, WANDERSON RUBIAN MARTINS. Identificação dos Fatos de Melhoria do IDEB pelo uso de mineração de dados: Um estudo de caso em Escolas Municipais de Maceió. SBIE, Trilha 4, 2019.

PINTO, GLEVSON DA SILVA; FREITAS JÚNIOR, OLIVAL DE GUSMÃO; COSTA, EVANDRO DE BARROS, Evandro. 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-Revista Novas Tecnologias na Educação, 2019, 17.3: 183-193.

ROMERO, Cristobal et al. Handbook of educational data mining. CRC Press, 2011.
ROMERO, Cristobal; VENTURA, Sebastian. Data mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, v. 03, n. 1, p. 12-27, 2013.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining: a review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), v. 40, n. 6, p. 601-618, 2010.

SAEB. SISTEMA DE AVALIÇÃO DA EDUCAÇÃO BÁSICA. Acesso em: 16/10/2019. Disponível em: http://portal.inep.gov.br/educacao-basica/saeb.

SARRA, A.; FONTANELLA, L.; ZIO, S. D. ***Identifying students at risk of academic failure within the educational data mining framework***. Social Indicators Research, apr. 2018.

SILVA, Jéssica Laísa Dias; NUNES, Isabel Dillmann. Mineração de dados Educacionais Como apoio para a classificação de alunos do ensino médio. CBIE-LACLO 2015. SBIE 2015. 2015.

TAN, P. N.; STEINBACH, M.; KUMAR, V. Introduction to data mining. Pearson, 2009.
VICINI, Lorena; SOUZA, Adriano Mendonça. Análise multivariada da teoria à prática. Santa Maria: UFSM, CCNE, 2005.

WEIS, SHOLOM M.; Indurkhya, Nitim. “Predict Data Mining”. Morgan Kaufmann Publisheis Inc., 1999.

WITTEN, I. H., FRANK, E.; HALL, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition.

ANEXO A – FERRAMENTAS DE APOIO

Foi visto que o processo de descoberta de conhecimento possui as seguintes etapas básicas: pré-processamento, mineração de dados e pós-processamento. Cada uma dessas etapas possui técnicas que podem ser específicas a determinadas aplicações. Além disso, cada técnica pode apresentar um ou mais algoritmos que agregam diferentes estratégias para melhorar a desempenho de cada etapa. Implementar todos esses algoritmos sempre que for necessário, extraindo dados de uma base é extremamente dispendioso. Por isso, softwares, frameworks e bibliotecas foram construídos, para auxiliar esse processo:

- Sistemas de Gerenciamento de Bancos de Dados (SGBDs): como o Oracle, Sql server, MySQL ou o PostgreSQL, são conhecidos por administrar e manter bases de dados além de gerenciar os acessos e as manipulações aos dados. Atualmente, alguns deles também são capazes de realizar tarefas de pré-processamento de dados, análises descritivas e visualização (gráficos e relatórios), permitindo até mesmo, programar algoritmos para tarefas mais complexas de mineração, como agrupamento, classificação, detecção de anomalias, entre outras.

- Weka: é um software gratuito de código aberto, desenvolvido em Java e mantido pela Universidade de Waikato (www.cs.waikato.ac.nz/ml/weka). Com ele, é possível realizar tarefas de pré-processamento, classificação, regressão, agrupamento e visualização dos dados. Também é possível planejar e executar análises/experimentos mais complexos através de fluxogramas que encadeiam as tarefas de mineração de dados. Por se tratar de um software livre, as suas bibliotecas podem ser integradas a ambientes de desenvolvimento Java (Eclipse ou NetBeans) para a realização de alterações ao processo.

- RapidMiner: é um software que possui versões gratuitas e pagas (rapidminer.com). Em sua plataforma, há três produtos principais: RapidMiner Studio, RapidMiner Server e RapidMiner Radoop. O primeiro é um ambiente visual de programação, que constrói projetos de análise de dados por meio de blocos e fluxogramas, possui conexão direta à base de dados, ferramentas específicas para a realização de um pré-processamento dos dados e pode realizar a mineração através da classificação, regressão, agrupamento e associação. O segundo produto, é utilizado para a replicação e compartilhamento dos modelos construídos pelo primeiro, possui recursos para agendamento, controle de versão,

acesso remoto, etc. O terceiro visa análises de Big Data e possui plug-ins com funções de mineração de dados da web, mineração de textos e integração com o Weka, assim como, com as linguagens Python e R, sendo possível o desenvolvimento de algoritmos customizados.

- Python: é uma linguagem de programação, orientada a objetos, que por meio de suas bibliotecas pode realizar coleta de dados, engenharia de dados, análise, web scraping (extração de dados/conteúdo em sites), construção de aplicativos na web, etc. Alguns de seus pacotes úteis para a Mineração de Dados são o SciPy/NumPy (computação científica), Pandas (manutenção/análise de dados), Matplotlib (gráficos) e Sckit-learn (aprendizado de máquina). Essa linguagem é utilizada para a análise de dados principalmente, quando se deseja ter o acompanhamento das análises por meio de aplicativos na web ou quando códigos estatísticos precisam estar integrados com servidores em ambiente de produção.

- R: também é uma linguagem de programação e um ambiente de desenvolvimento integrado para a realização de cálculos estatísticos e gráficos. Ela foi desenvolvida por estatísticos para estatísticos, engenheiros e cientistas sem conhecimento de programação de computadores. Possui um grande número de pacotes para análise de dados, com modelos, fórmulas e testes estatísticos. Com o R é fácil escrever fórmulas complexas e praticamente todos os tipos de testes e modelos estatísticos estão disponíveis para o uso.

- Anaconda: é uma plataforma de código aberto que une a linguagem Python e R, com várias bibliotecas para a análise de dados.

ANEXO B – DICIONÁRIO DE DADOS

Atributo	Tipo	Descrição	Código de Preenchimento
ID_PROVA_BRASIL	Num	Ano da ANEB/Prova Brasil	
ID_REGIAO	Num	Código da Região	1 - Norte
			2 - Nordeste
			3 - Sudeste
			4 - Sul
			5 - Centro-Oeste
ID_UF	Num	Código da Unidade da Federação	11-RO
			12-AC
			13-AM
			14-RR
			15-PA
			16-AP
			17-TO
			21-MA
			22-PI
			23-CE
			24-RN
			25-PB
			26-PE
			27-AL
			28-SE
			29-BA
			31-MG
			32-ES
			33-RJ
			35-SP
			41-PR
			42-SC
			43-RS
			50-MS
			51-MT
			52-GO
			53-DF
ID_MUNICIPIO[6]	Num	Código do Município	
ID_AREA	Num	Área	1 - Capital
			2 - Interior
ID_ESCOLA[6]	Num	Código da Escola	
ID_DEPENDENCIA_ADM	Num	Dependência Administrativa	1 - Federal
			2 - Estadual
			3 - Municipal
			4 - Privada
ID_LOCALIZACAO	Num	Localização	1 - Urbana
			2 - Rural
ID_TURMA	Num	Código da turma na ANEB/Prova Brasil	
ID_TURNO	Num	Turno da Turma	1 - Matutino[3]
			2 - Vespertino[4]
			3 - Noturno[5]

ID_SERIE	Num	Série	9 - 8ª série/9º ano Ensino Fundamental
ID_ALUNO	Num	Código do aluno na ANEB/Prova Brasil	
IN_SITUACAO_CENSO	Num	Indicador de Consistência com o Censo Escolar 2015	0 - Não consistente
			1 - Consistente
IN_PREENCHIMENTO_PROVA	Num	Indicador de preenchimento da prova	0 - Prova não preenchida
			1 - Prova preenchida
ID_CADERNO	Num	Caderno de Provas	Prova Regular (Cadernos 1 a 21)
			Macrotipo (Cadernos 22, 23 e 24)
			Ledor (Caderno 25)
ID_BLOCO_1	Num	Identificador do Bloco 1 (Língua Portuguesa e Matemática)	De 1 a 7
ID_BLOCO_2	Num	Identificador do Bloco 2 (Língua Portuguesa e Matemática)	De 1 a 7
TX_RESP_BLOCO_1_LP[2]	Char	Resposta do Aluno ao Bloco 1 da Prova de Língua Portuguesa	A, B, C, D, . (branco), * (nulo)
TX_RESP_BLOCO_2_LP[2]	Char	Resposta do Aluno ao Bloco 2 da Prova de Língua Portuguesa	A, B, C, D, . (branco), * (nulo)
TX_RESP_BLOCO_1_MT[2]	Char	Resposta do Aluno ao Bloco 1 da Prova de Matemática	A, B, C, D, . (branco), * (nulo)
TX_RESP_BLOCO_2_MT[2]	Char	Resposta do Aluno ao Bloco 2 da Prova de Matemática	A, B, C, D, . (branco), * (nulo)
IN_PROFICIENCIA	Num	Indicador para cálculo da proficiência (no mínimo três itens respondidos na prova)	0 - Não
			1 - Sim
IN_PROVA_BRASIL	Num	Indicador de participação na Prova Brasil	0 - Não
			1 - Sim
ESTRATO_ANEB	Char	Descrição dos estratos da ANEB	Os estratos são compostos por: Série, Região, UF, Área, Dependência Administrativa e Localização. Para mais detalhes consulte o Relatório de Amostragem da ANEB 2015.
PESO_ALUNO_LP	Num	Peso do Aluno em Língua Portuguesa	Valor com 15 casas decimais
PESO_ALUNO_MT	Num	Peso do Aluno em Matemática	Valor com 15 casas decimais
PROFICIENCIA_LP	Num	Proficiência do aluno em Língua Portuguesa calculada na escala única do SAEB, com média = 0 e desvio = 1 na população de referência	Valor com 15 casas decimais

DESVIO_PADRAO_LP	Num	Desvio padrão da proficiência em Língua Portuguesa	Valor com 15 casas decimais
PROFICIENCIA_LP_SAEB	Num	Proficiência em Língua Portuguesa transformada na escala única do SAEB, com média = 250, desvio = 50 (do SAEB/97)	Valor com 15 casas decimais
DESVIO_PADRAO_LP_SAEB	Num	Desvio padrão da proficiência transformada em Língua Portuguesa	Valor com 15 casas decimais
PROFICIENCIA_MT	Num	Proficiência do aluno em Matemática calculada na escala única do SAEB, com média = 0 e desvio = 1 na população de referência	Valor com 15 casas decimais
DESVIO_PADRAO_MT	Num	Desvio padrão da proficiência em Matemática	Valor com 15 casas decimais
PROFICIENCIA_MT_SAEB	Num	Proficiência do aluno em Matemática transformada na escala única do SAEB, com média = 250, desvio = 50 (do SAEB/97)	Valor com 15 casas decimais
DESVIO_PADRAO_MT_SAEB	Num	Desvio padrão da proficiência transformada em Matemática	Valor com 15 casas decimais
IN_PREENCHIMENTO_QUESTIONARIO	Num	Indicador de preenchimento do questionário	0 - Não preenchido 1 - Preenchido parcial ou totalmente

Atributo	Tipo	Questão	Enunciado
TX_RESP_Q001	Char	Questão 1	Qual é o seu sexo?
TX_RESP_Q002	Char	Questão 2	Como você se considera?
TX_RESP_Q003	Char	Questão 3	Você poderia nos dizer qual é o mês de seu aniversário?
TX_RESP_Q004	Char	Questão 4	Em que ano você nasceu?
TX_RESP_Q005	Char	Questão 5	Na sua casa tem televisão em cores?
TX_RESP_Q006	Char	Questão 6	Na sua casa tem rádio?
TX_RESP_Q007	Char	Questão 7	Na sua casa tem videocassete e/ou DVD?
TX_RESP_Q008	Char	Questão 8	Na sua casa tem geladeira?
TX_RESP_Q009	Char	Questão 9	Na sua casa tem freezer (parte da geladeira duplex)?
TX_RESP_Q010	Char	Questão 10	Na sua casa tem freezer separado da geladeira?
TX_RESP_Q011	Char	Questão 11	Na sua casa tem máquina de lavar roupa (O tanquinho NÃO deve ser considerado)?
TX_RESP_Q012	Char	Questão 12	Na sua casa tem carro?
TX_RESP_Q013	Char	Questão 13	Na sua casa tem computador?
TX_RESP_Q014	Char	Questão 14	Na sua casa tem banheiro?
TX_RESP_Q015	Char	Questão 15	Na sua casa tem quartos para dormir?
TX_RESP_Q016	Char	Questão 16	Incluindo você, quantas pessoas vivem atualmente em sua casa?
TX_RESP_Q017	Char	Questão 17	Em sua casa trabalha empregado(a) doméstico(a) pelo menos cinco dias por semana?

TX_RESP_Q018	Char	Questão 18	Você mora com sua mãe?
TX_RESP_Q019	Char	Questão 19	Até que série sua mãe, ou a mulher responsável por você, estudou?
TX_RESP_Q020	Char	Questão 20	Sua mãe, ou a mulher responsável por você, sabe ler e escrever?
TX_RESP_Q021	Char	Questão 21	Você vê sua mãe, ou mulher responsável por você, lendo?
TX_RESP_Q022	Char	Questão 22	Você mora com seu pai?
TX_RESP_Q023	Char	Questão 23	Até que série seu pai, ou o homem responsável por você, estudou?
TX_RESP_Q024	Char	Questão 24	Seu pai, ou homem responsável por você, sabe ler e escrever?
TX_RESP_Q025	Char	Questão 25	Você vê o seu pai, ou homem responsável por você, lendo?
TX_RESP_Q026	Char	Questão 26	Com qual frequência seus pais, ou responsáveis por você, vão à reunião de pais?
TX_RESP_Q027	Char	Questão 27	Seus pais ou responsáveis incentivam você a estudar?
TX_RESP_Q028	Char	Questão 28	Seus pais ou responsáveis incentivam você a fazer o dever de casa e/ou os trabalhos da escola?
TX_RESP_Q029	Char	Questão 29	Seus pais ou responsáveis incentivam você a ler?
TX_RESP_Q030	Char	Questão 30	Seus pais ou responsáveis incentivam você a ir a escola e/ou não faltar às aulas?
TX_RESP_Q031	Char	Questão 31	Seus pais ou responsáveis conversam com você sobre o que acontece na escola?
TX_RESP_Q032	Char	Questão 32	Com qual frequência você lê: Jornais.
TX_RESP_Q033	Char	Questão 33	Com qual frequência você lê: Livros em geral.
TX_RESP_Q034	Char	Questão 34	Com qual frequência você lê: Livros de literatura.
TX_RESP_Q035	Char	Questão 35	Com qual frequência você lê: Revistas em geral.
TX_RESP_Q036	Char	Questão 36	Com qual frequência você lê: Revistas em quadrinhos (gibis).
TX_RESP_Q037	Char	Questão 37	Com qual frequência você lê: Revistas de comportamento, celebridades, esportes ou TV.
TX_RESP_Q038	Char	Questão 38	Com qual frequência você lê: Notícias na internet (ex.: blog, notícia).
TX_RESP_Q039	Char	Questão 39	Com qual frequência você costuma ir à biblioteca?
TX_RESP_Q040	Char	Questão 40	Com qual frequência você costuma ir ao cinema?
TX_RESP_Q041	Char	Questão 41	Com qual frequência você costuma ir a algum tipo de espetáculo ou exposição (teatro, museu, dança, música)?
TX_RESP_Q042	Char	Questão 42	Com qual frequência você participa de festas na sua vizinhança ou comunidade?
TX_RESP_Q043	Char	Questão 43	Em dia de aula, quanto tempo você gasta assistindo à TV, navegando na internet ou jogando jogos eletrônicos?
TX_RESP_Q044	Char	Questão 44	Em dias de aula, quanto tempo você gasta fazendo trabalhos domésticos (ex.: lavando louça, limpando o quintal etc.)?
TX_RESP_Q045	Char	Questão 45	Atualmente você trabalha fora de casa (recebendo ou não um salário)?
TX_RESP_Q046	Char	Questão 46	Quando você entrou na escola?

TX_RESP_Q047	Char	Questão 47	A partir da quinta série ou sexto ano, em que tipo de escola você estudou?
TX_RESP_Q048	Char	Questão 48	Você já foi reprovado?
TX_RESP_Q049	Char	Questão 49	Você já abandonou a escola durante o período de aulas e ficou fora da escola o resto do ano?
TX_RESP_Q050	Char	Questão 50	Você gosta de estudar Língua Portuguesa?
TX_RESP_Q051	Char	Questão 51	Você faz o dever de casa de Língua Portuguesa?
TX_RESP_Q052	Char	Questão 52	O(A) professor(a) corrige o dever de casa de Língua Portuguesa?
TX_RESP_Q053	Char	Questão 53	Você gosta de estudar Matemática?
TX_RESP_Q054	Char	Questão 54	Você faz o dever de casa de Matemática?
TX_RESP_Q055	Char	Questão 55	O(A) professor(a) corrige o dever de casa de Matemática?
TX_RESP_Q056	Char	Questão 56	Você utiliza a biblioteca ou sala de leitura da sua escola?
TX_RESP_Q057	Char	Questão 57	Quando você terminar o 9º ano(8ª série) você pretende:

1 - Consistência entre os dados a aplicação da Prova Brasil 2015 com o Censo da Educação Básica 2015 finalizado.
2 - Para o 5º ano/4ª série há 11 respostas e 9º ano/8ª série há 13 respostas
3 - Turno Matutino: Início entre 5:00h às 10:59h
4 - Turno Vespertino: Início entre 11:00h às 16:59h
5 - Turno Noturno: Início entre 17:00h às 4:59h
6 - Os códigos dos Municípios e das Escolas que começam com o dígito "6" são máscaras, isto é, são códigos fictícios.