Mineração de Dados Educacionais: Um Modelo de Predição do Perfil do Aluno para Melhoria do IDEB

Glevson da Silva Pinto[1], Olival de Gusmão Freitas Júnior[1], Evandro de Barros Costa[1]

[1]Universidade Federal de Alagoas (UFAL) – AL – Brasil

{glevsonsilva@ic.ufal.br, olival@ic.ufal.br, evandro@ic.ufal.br}

Abstract. *This work aims to explore techniques of attribute selection and predictive algorithms, aiming at developing a model of a predictor to identify which factors positively impact the Basic Education Development Index of Brazilian public schools. This is a quantitative and exploratory research. From the point of view of technical procedures, it is a case study, analyzing educational data from the municipal public schools of Teotônio Vilela (Alagoas), conducting an experimental study, producing relevant results in the task of identifying relevant attributes to support educational managers. This study proved that the application for predicting the outcome of the Basic Education Development Index works and we were able to identify some student profiles that can contribute to the best outcome of the Basic Education Development Index*.

Resumo. *Este trabalho tem como objetivo explorar técnicas de seleção de atributos e algoritmos preditivos, visando desenvolver um modelo de um preditor para identificar quais fatores impactam positivamente no IDEB (Índice de Desenvolvimento da Educação Básica) das escolas públicas brasileiras. 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 Teotônio Vilela (Alagoas), conduzindo um estudo experimental, produzindo relevantes resultados na tarefa de identificação de atributos relevantes para apoiar os gestores educacionais. Este estudo comprovou que a aplicação para predição do resultado do IDEB funciona e conseguimos identificar alguns perfis de alunos que podem contribuir no melhor resultado do IDEB.*

1. Introdução

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 (CALIXTO *et al.*, 2017).

Segundo INEP (2019), a criação, o aprimoramento e a evolução das avaliações de larga escala e as políticas públicas educacionais estão entrelaçadas, uma vez que, os índices resultantes destas avaliações, sejam elas nos parâmetros nacionais, como o IDEB (Índice de Desenvolvimento da Educação Básica), ou internacionais como o PISA (*Programme for Internacional Student Assessment* – Programa Internacional de Avaliação de Alunos), balizam e orientam as diretrizes governamentais.

Os índices gerados pelas avaliações de larga escala, além de diagnosticar o desempenho dos alunos, expõem o trabalho docente. O IDEB é concebido como um indicador que permitirá o monitoramento da evolução da situação educacional, compreendendo metas intermediárias (a cada dois anos) e finais em 2022, estimado com base nas avaliações que os professores fazem nas escolas, das quais resultam as taxas de

DOI: 10.5753/cbie.sbie.2020.1172

promoção, e as avaliações de desempenho dos alunos em modalidades do Sistema de Avaliação da Educação Básica (SAEB).

Os questionários aplicados aos alunos servem como instrumentos de coleta de informações sobre aspectos da vida escolar, do nível socioeconômico, capital social e cultural dos alunos. Os questionários aplicados aos professores de português e matemática das séries avaliadas e aos diretores das escolas, por sua vez, possibilitam conhecer a formação profissional, práticas pedagógicas, nível socioeconômico e cultural, estilos de liderança e formas de gestão. Na mesma ocasião, os aplicadores dos testes preenchem um formulário sobre a escola indicando as condições de infraestrutura, segurança e os recursos pedagógicos disponíveis (INEP, 2019).

Esses dados educacionais constituem fontes de informação que podem ser analisadas por meio de técnicas de mineração de dados, visando à melhoria na gestão educacional, na organização do trabalho pedagógico e na melhoria da qualidade do ensino e da aprendizagem (PAIVA *et al.*, 2012).

A mineração de dados educacionais (MDE) é um campo de pesquisa que busca descobrir padrões ou evidências sobre alunos e formas de aprendizagem. Na MDE, os problemas empíricos geralmente podem ser decompostos em três tipos (Junker, 2011): realização de inferências sobre as características de atividades; realização de inferências sobre as características dos alunos e previsões sobre o desempenho dos alunos em tarefas futuras.

Atualmente, no Brasil, falta um sistema para analisar e monitorar o desempenho dos alunos no ensino básico. Há duas razões para isso. Primeiro, o estudo dos métodos de predição ainda é insuficiente para identificar os métodos mais eficazes para a visualização do desempenho dos alunos nas instituições educacionais. O segundo é devido à falta de investigação mais aprofundada sobre os fatores que interferem no desempenho dos alunos. 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 (MayaPérez *et al.*, 2018).

O objetivo deste artigo é explorar técnicas de seleção de atributos e algoritmos preditivos, visando desenvolver um modelo de um preditor para identificar quais fatores impactam positivamente no IDEB das escolas públicas brasileiras. 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 Teotônio Vilela (Alagoas), 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 Teotônio Vilela.

O artigo está organizado da seguinte forma: a seção 2 abordará alguns trabalhos relacionados a esta temática, tentando mostrar a originalidade do presente trabalho. A seção 3 tratará da aplicação da metodologia CRISP-DM adaptada a nossa proposta, destacando as suas seis etapas, mas enfatizando mais a que interessa diretamente a identificação dos fatores mais relevantes. A seção 4 apresenta as considerações finais obtidas com este trabalho.

2. 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.

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 etapa de mineração de dados.

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. Neste trabalho se utilizou técnicas de mineração de dados para identificar o perfil do aluno evadido e estimar a propensão à evasão.

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 utilizada 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 influentes na evasão escolar foram: idade, etapa de ensino, modalidade de ensino, existência de laboratórios e localização da escola.

MayaPérez *et al*. (2018) desenvolveram um modelo de predição baseado em seleção de atributos e classificadores. O objetivo desse estudo é 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 de 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 acurácia. Nesse estudo, identificou-se 15 atributos mais relevantes no universo de 39.

Freitas Júnior *et al.* (2019) utilizaram uma 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. Aplicaram 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.

O presente artigo tem como foco explorar técnicas de seleção de atributos e algoritmos preditivos, visando desenvolver um modelo de um preditor para identificar quais fatores impactam positivamente no IDEB das escolas públicas brasileiras. Utilizou-se três abordagens de seleção de atributos (filtro, embaralhamento e embutida), aplicando em seguida diversões algoritmos de classificação (NaiveBayes, *J48*, *JRip*, *LibSVM*, *RandomForest, IBK, OneR* e REPTree). Analisando as diversas abordagens dos trabalhos consultados, verifica-se uma maior proximidade com MayaPérez *et al.* (2018), porém o presente trabalho tem um processo diferente na identificação dos atributos e foca nas instituições educacionais municipais de ensino fundamental.

3. Metodologia

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.

1ª Etapa: Compreensão do domínio. O objetivo deste projeto é explorar técnicas de seleção de atributos e algoritmos preditivos, visando desenvolver um modelo de um preditor para identificar quais fatores impactam positivamente no IDEB das escolas públicas brasileiras.

2ª Etapa: Entendimento dos dados. As bases de dados usadas nesse estudo foram disponibilizadas abertamente pelo INEP em seu portal. Coletou-se os dados educacionais das escolas municipais de Teotônio Vilela no portal do INEP, utilizando a ferramenta Anaconda Distribuição para visualizar os dados e conferir os 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.

3ª Etapa: Preparação dos dados. Esta etapa foi utilizada a ferramenta Pandas com a linguagem em Python, envolvendo operações para 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.

Ao invés de usar a nota de proficiência como variável dependente, utiliza-se 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 "*Condição"*, 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 9º ano como segue na **Tabela 1**.

Tabela 1. Estatística dos alunos

Médias e medianas dos alunos			**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 observa na **Tabela 1**, 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.

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. Para a rede pública de Teotônio Vilela têm-se 312 instâncias (abaixo da média) para cada classe e 260 instâncias (acima da média).

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.

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). 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.

Na **Tabela 2** são apresentados os atributos selecionados e as respectivas quantidades, utilizando as seguintes bases de dados: "Completa", embutida, filtro, embaralhamento e "Todos". O conjunto de dados "Completa" representa o questionário sem seleção de atributos. Já "Todos" representa o conjunto de atributos após seleção dos mesmos, utilizando as três abordagens (embutida, filtro e embaralhamento), 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.

Tabela 2. Atributos selecionados para alunos do 9º ano

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

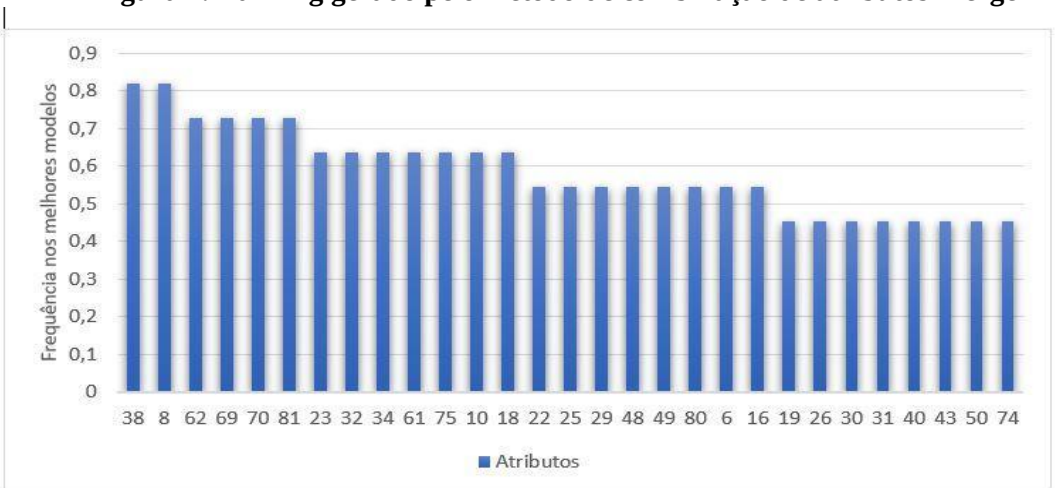
Fonte: Elaborado pelos autores

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 1** 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. À medida que se realizou a etapa de construção do método Merge se teve 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.

Figura 1. Ranking gerado pelo método de combinação de atributos Merge



Fonte: Elaborado pelos autores

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.

Na **Tabela 2** 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 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 2. Precisão dos classificadores para Língua Portuguesa e Matemática

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 pelos autores

De acordo com a **Tabela 2**, 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 Embaralhamento 97,66% (Português) e 97,91% (Matemática). Dessa forma, é perceptível o fato de que o conjunto dessas diversas 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 e que contribuem para uma boa predição do modelo proposto.

Os classificadores com resultados altos contribuem para criação do preditor que será mostrado na **Figura 2,** no qual 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.

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. É possível observar na **Tabela 2** 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 3** apresenta os atributos que tiveram maior incidência por disciplina.

Tabela 3. 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 pelos autores

De acordo com a **Tabela 3**, 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. Já 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 são: ID TURNO, ID Caderno, ID Bloco 2, Desvio Padrão Língua Portuguesa, Desvio Padrão Língua Portuguesa SAEB, Em que ano você nasceu?, Sua casa tem TV a cores?, Sua casa tem vídeocassete e/ou DVD?, Sua casa tem máquina de lavar roupa?, Sua casa tem banheiro?, Incluindo você, quantas pessoas vivem na sua casa?, Você ver sua mãe, ou a mulher responsável por você, lendo?, Qual frequência você lê revistas em geral?, Qual frequência você ler revistas de comportamentos, celebridades, esportes ou TV?, A partir da 5ª série que tipo de escola você estudou (Pública ou Privada)?, Você já foi reprovado?, Professor corrige o dever de Matemática? e Você utiliza biblioteca ou sala de leitura da escola?.

É possível ver que a estratégia de seleção de atributos usada por categorias como embutida, filtro e embaralhamento, combinadas ao modelo de ranking Merge, permitiu evidenciar os melhores atributos do conjunto de 91 para 71 (língua portuguesa), redução de 22%. 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 se obteve 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.

Com base nos dados obtidos pela etapa de pré-processamento dos dados e de seleção de atributos, foi gerado um preditor, conforme **Figura 2**, com objetivo de encontrar os perfis dos 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 4** e aplicando essas instâncias de teste retiradas da base de dados de 2017 no preditor obteve-se a classificação e características dos alunos.

Figura 2. Preditor do perfil do aluno para melhoria do IDEB

Fonte: Elaborado pelos autores

Com base na **Tabela 4,** 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 dos alunos.

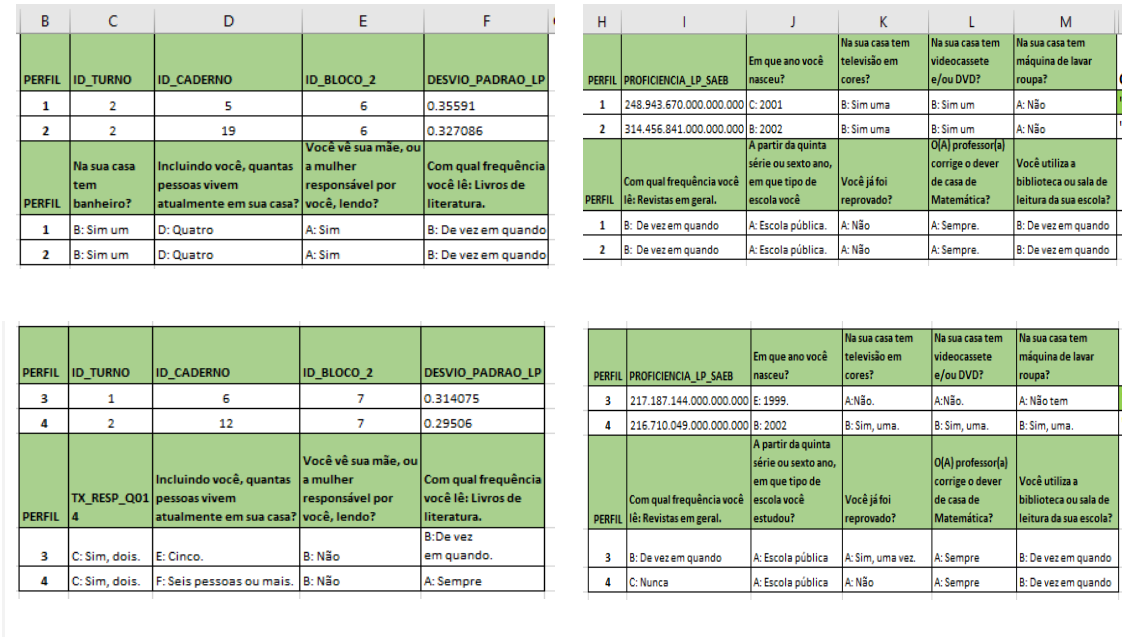
Tabela 4. Classificação do perfil do aluno e rejeição de classificadores (Teotônio Vilela)

Perfil do aluno	Classificação no IDEB	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 pelos autores

De acordo com a **Figura 3** é possível validar os perfis dos alunos obtidos pela classificação por meio do preditor. Nesse processo também se verifica que as informações da **Tabela 4** são evidenciadas por meio dos classificadores e que é possível predizer quais atributos contribuem para a melhoria do resultado do IDEB. 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.

Figura 3. 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 pelos autores

Conforme se observa, o processo de Merge valida a 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 avaliação entre os atributos que envolvem o problema estudado.

Com esse modelo de preditor será possível identificar o perfil de um determinado aluno bem como identificar os perfis de alunos com maiores chances de obterem bons resultados no IDEB. Além disso, 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.

4. Considerações Finais

Neste trabalho foi mostrado que o procedimento de seleção de atributos cria 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 no IDEB.

Com isso, foi possível desenvolver um modelo de preditor para testar o processo de seleção de atributos criados ao visualizar os perfis de alunos com maiores chances de bons resultados no IDEB, sendo 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, tais como: a escola integral e a influência dos incentivos dos pais ou responsáveis para o bom desempenho dos alunos, entre outros atributos aqui obtidos pela MDE que são ricos para a tomada de decisão dos gestores escolares.

Neste trabalho foi mostrado também 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.

Após todo processo de treinamento do modelo de predição pode ser testado esse modelo em 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. Como contribuições desse 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.

5. Referências

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.

CALIXTO, K. E. A.; SEGUNDO, C. V. N.; RENÊ, P. G.. **Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar**. VI CBIE, SBIE, 2017.

CHAPMAN, P. *et al*. **CRISP-DM 1.0** *step-by-step data mine guide*. CRISP-DM Consortium. 2000.

FONSECA, S. O.; NAMEN, A. A. **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.; COSTA, E. B. **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, 2019.

INEP. **Ideb**. 2019. Acesso em: 31 Janeiro 2019. Disponível em: http://portal.inep.gov.br/ideb.

JUNKER, Brian W. ***Modeling hierarchy and dependence among task responses in educational data mining***. Handbook of Educational Data Mining, p. 143- 155, 2011.

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.

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, P. N.; AGUILAR, C. J. R; ZAMORA, R. R. A.; BARRON, A. J. M. **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.

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.