Identificação dos fatores de melhorias no IDEB pelo uso de mineração de dados: Um estudo de caso em escolas municipais de Teotônio Vilela-Alagoas

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

[1]Universidade Federal de Alagoas (UFAL) – AL – Brasil
{glevsonsilva@ic.ufal.br, olival@ic.ufal.br, evandro@ic.ufal.br}

Resumo. *A Mineração de Dados Educacionais vem auxiliando educadores e gestores no apoio a tomada de decisões, permitindo extração de informações relevantes de bases de dados. O objetivo deste artigo é identificar os fatores que afetam o desempenho escolar dos alunos (IDEB) das escolas de ensino fundamental do município de Teotônio Vilela através dos resultados obtidos na Prova Brasil. Neste artigo explorou-se técnicas de seleção de atributos e mineração de dados, identificando quais fatores impactam no IDEB das escolas municipais de Alagoas para posteriores estudos e reflexões na área da educação. Para tanto, utilizou-se dados do SAEB das escolas públicas municipais de Teotônio Vilela, conduzindo um estudo experimental, produzindo relevantes resultados na tarefa de identificação de atributos relevantes para apoiar os gestores educacionais.*

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

Identification of IDEB improvement factors through the use of data mining: A case study in Teotônio Vilela-Alagoas municipal schools

Abstract. *Educational data mining has been helping educators and decision makers for the purpose of extracting useful academical information on the students, from large data sources. In this paper we explore the use of attribute selection techniques in data mining, with the aim of identifying the most relevant variables that impact on the Basic Education Development Index (Ideb) of the students of the municipal schools of Teotônio Vilela. We used data from the Saeb test of municipal schools and applied attribute selection and classification methods. The conducted experimental study and the obtained results are discussed at the end, showing the most relevant attributes increasing the predictive performance and providing relevant information to decision makers in educational settings.*

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

1. Introdução

Percebe-se cada vez mais a necessidade de informações de qualidade por parte dos gestores educacionais, visando à efetividade na tomada de decisões no sentido de melhorar o processo de ensino e aprendizagem nas instituições educacionais públicas do Brasil. Assim, por exemplo, 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 daEducação Básica (SAEB), a Prova Brasil, a Avaliação Nacional de Alfabetização entre outras (INEP, 2019; INEP/MEC, 2007).

Os dados educacionais, incluindo-se fatores socioeconômicos, 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. Nos últimos anos diversos trabalhos (Coelho *et al.*, 2015; Pasta, 2011) têm explorados os benefícios que a MDE traz ao ambiente educacional.

O objetivo deste artigo é identificar os fatores que afetam o desempenho escolar dos alunos (IDEB) das escolas de ensino fundamental do município de Teotônio Vilela através dos resultados obtidos na Prova Brasil. Para isso, aplicam-se técnicas de seleção de atributos para descobrir as questões que mais impactam o IDEB. Em seguida, aplicam-se diversos classificadores, visando determinar qual método é aconselhado para realizar a previsão com a melhor acurácia.

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, 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. Esses dois períodos foram selecionados devido aos avanços alcançados pelo município conforme disponível no portal do INEP. Salienta-se que em 2015 houve avanço de 4.6 (meta era 3.7 no geral do município para o 9^0 ano) e em 2017 um avanço ainda mais expressivo de 5.8 (meta era 3.9 no geral do município para o 9^0 ano). Esse aumento no desempenho desperta nosso olhar para MDE com objetivo de descobrir quais atributos estão influenciando na melhoria de desempenho educacional deste referido município.

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. A seção 4 apresenta as conclusões 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.

Nascimento *et al.* (2018) aplicou 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).

Bezerra *et al.* (2016) abordou 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. O objetivo deste trabalho é a extração de conhecimento a partir dos dados do censo escolar disponibilizado pelo INEP, visando identificar o perfil do aluno evasor e estimar a

propensão à evasão através de Árvore de Decisão, Indução de Regras e Regressão Logística. Os resultados mostraram que fatores como idade, turno das aulas e região geográfica das escolas influenciam fortemente a evasão.

Márquez-Vera *et al* (2013) aplicou 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. Utilizou-se técnicas de mineração de dados focadas em regras de indução e árvores de decisão, para prever falhas acadêmicas dos alunos em escolas de ensino médio. No entanto, para alcançar esses resultados, este trabalho teve que considerar variáveis de várias fontes de dados, incluindo dados não acadêmicos. Com isso, algumas ações preventivas poderiam ser tomadas para evitar a evasão escolar desses alunos. Eles conseguem excelentes resultados finais, como uma acurácia de 97,3%.

Romero *et al* (2013) construiu uma ferramenta de MDE dentro do ambiente virtual de aprendizagem Moodle. Para isso, inicialmente, são implementados quatro algoritmos que realizam a previsão do desempenho de alunos em sete cursos de engenharia. Os autores concluem que não há, em geral, nenhum algoritmo que obtenha uma melhor classificação com a utilização de todas as evidências, mesmo com o uso de técnicas de pré-processamento dos dados (filtragem, discretização ou balanceamento), por isso a acurácia de 65% é apresentada como uma "média" entre os algoritmos. Mas acreditam que adição dos dados de mais alunos e mais informações off-line sobre eles poderiam ajudar. As informações off-line seriam sobre atendimentos em sala de aula, pontualidade, participação, atenção, predisposição, etc.

O presente artigo tem como foco identificar os fatores que afetam o desempenho escolar dos alunos das escolas de ensino fundamental de um 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 IDEB dos alunos das escolas municipais de Teotônio Vilela. 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. 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. Dessa forma, o desenvolvimento do trabalho seguiu as etapas do CRISP-DM, apresentadas na **Figura 1** descrita a seguir.

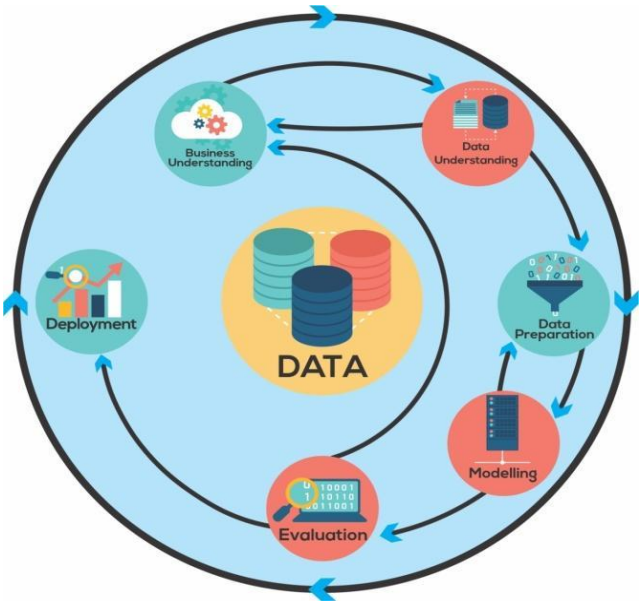

Figura 1 - Etapas da Metodologia CRISP-DM.
Fonte: Adaptado de Chapman *et al.* (2000)

1ª Etapa: Compreensão do domínio. 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 o IDEB. Este estudo fará a análise dos dados educacionais de escolas públicas municipais de Teotônio Vilela relativos aos alunos dos anos finais do ensino fundamental (9º ano). 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. Cada escola deve melhorar seus indicadores, contribuindo para que o Brasil chegue à meta 6,0 em 2022. O IDEB da rede pública de ensino do município de Teotônio Vilela cresceu, porém ainda não atingiu a meta 6,0. No ano de 2017, o 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.

2ª Etapa: Entendimento dos dados. 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 realizada com a ferramenta Anaconda, conforme **Figura 2**, 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.

```
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'

Figura 2 - Preparação dos dados usando ferramenta Anaconda.
Fonte: Elaborada pelos autores

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 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 (**Tabela 1**).

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*). 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 282 instâncias para cada classe de "*Condicao"* na matéria de língua portuguesa e 278 instâncias para cada classe em matemática.

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. A abordagem "Todos" representa a combinação das três abordagens anteriores (**Tabela 2**).

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: Elaborada 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 **Figura 3** 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.

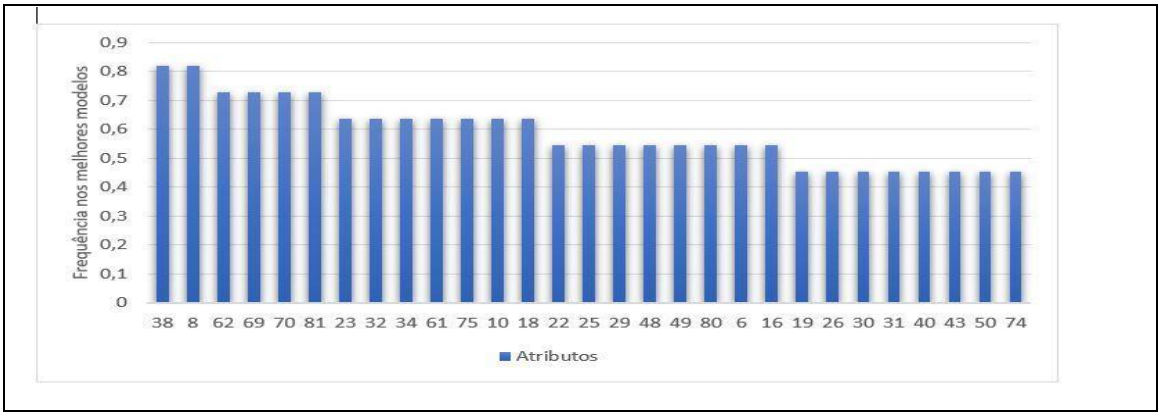

Figura 3 - Ranking gerado pelo método de combinação de atributos Merge.
Fonte: Elaborada pelos autores

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 3** 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 3 - 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: Elaborada pelos autores

De acordo com a **Tabela 3**, 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 visualizam-se os atributos reduzidos e dessa forma já utilizar as informações dessa base para análise dos dados. Os resultados com acurácia acima de 90% foi possível devido à qualidade do conjunto de instâncias, e ao fato de que o município estudado tem excelentes resultados no IDEB.

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 obteve-se o mesmo comportamento da abordagem anterior.

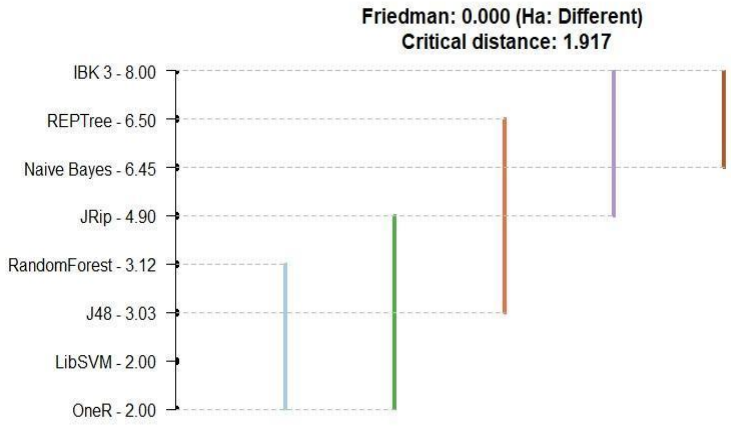

Figura 4 - Aplicação do método de avaliação estatística de Friedman e Nemenyi.
Fonte: Elaborada pelos autores

A **Figura 4** apresenta o resultado do teste estatístico de Friedman e Nemenyi, aplicado à base de dados "Todos". 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 NaiveBayes, 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 4** 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 no qual os conhecimentos encontrados são interpretados e utilizados em processo decisório. É possível observar na **Tabela 3** 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 4** apresenta os atributos que tiveram maior incidência por disciplina.

De acordo com a **Tabela 4**, 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. Dessa forma têm-se 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 3**.

Tabela 4 - Atributos com maior incidência.

Matéria	Atributo
Português	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
Matemática	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

É possível verificar 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), ou seja, uma redução de 22% dos atributos, mais ainda durante essa etapa manteve-se os dados socioeconômicos 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; representando os atributos mais forte do conjunto de dados, 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 com os 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.

4. Conclusão

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.

Neste estudo foram descobertos 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.

Esse 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.

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.

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

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.

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.

INEP/MEC. **Indicadores da Qualidade da Educação**. São Paulo: ação educativa, 2007.

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

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.

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, RENOTE, 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.

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.

ROMERO, Cristobal; ESPEJO, Pedro G.; ZAFRA, Amelia; VENTURA, Sebastian. ***Web usage mining for predicting final marks of students that use Moodle courses***. Computer Applications in Engineering Education, v. 21, n. 1, p. 135-146, 2013.

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.