Identificação dos fatores de melhorias no IDEB pelo uso de mineração de dados: Um estudo de caso em escolas municipais de MACEIÓ

[1]Glevson da Silva Pinto, [1]Olival de Gusmão Freitas Júnior, [1]Evandro de Barros Costa, [1]João Carlos Cordeiro Barbirato, [1]Wanderson Rubian Martins Rodrigues

[1]Universidade Federal de Alagoas (UFAL) – AL – Brasil
{glevsonsilva@ic.ufal.br, olival@ic.ufal.br, evandro@ic.ufal.br, jccb@ctec.ufal.br, rubian64@gmail.com}

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 Maceió. We used data from the Saeb test of 13 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.*

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. Neste artigo explorou-se técnicas de seleção de atributos em mineração de dados, visando identificar quais fatores impactam no IDEB das escolas municipais de Maceió. Para tanto, utilizou-se dados do teste Saeb de 13 escolas municipais de Maceió, conduzindo um estudo experimental, produzindo relevantes resultados na tarefa de identificação de atributos relevantes para apoiar os gestores educacionais.*

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 da Educação Básica (SAEB), a Prova Brasil, a Avaliação Nacional de Alfabetização entre outras (INEP, 2019; INEP/MEC, 2007; INEP, 2016).

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; Manhães, 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 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. Trata-se de uma pesquisa de

DOI: 10.5753/cbie.sbie.2019.1828

cunho quantitativa e exploratória. Do ponto de vista dos procedimentos técnicos, trata-se de um estudo de caso, analisando-se dados educacionais de 13 escolas públicas municipais de Maceió, 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.

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 fases mais diretamente relacionada à identificação das variáveis mais relevantes para a melhoria do IDEB. 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). 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) 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.

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) 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. 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.

O presente artigo 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. Descrição da Base de Dados e do Pré-Processamento

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 13 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. No entanto utilizou-se a ferramenta Anaconda Distribuição para visualizar os dados e conferir os tipos de dados.

Esta etapa tem o objetivo de construir o conjunto final de dados que serão utilizados nas ferramentas de modelagem. As tarefas de preparação podem ser realizadas muitas vezes, e sem uma ordem pré-determinada. Esta 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 de Maceió.

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 matérias 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. O questionário da Prova Brasil é composto por 57 questões para alunos do nono ano do ensino fundamental. “Esse questionário serve como instrumento de coleta de informações sobre aspectos da vida escolar, do nível socioeconômico, capital social e cultural dos alunos”, de acordo com o próprio portal do INEP (2019b).

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 na **Tabela 1**.

Tabela 1. Estatística dos alunos.

Média e Mediana			**Condição dos alunos**		
	Língua Portuguesa	**Matemática**		**Língua Portuguesa**	**Matemática**
Média	251,64	254,65	**Acima da média**	282	278
Mediana	253,03	255,05	**Abaixo da média**	267	271

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**).

4. Mineração de Dados

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. 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.

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 282 instâncias para cada classe de CONDICAO na matéria de língua portuguesa e 278 instâncias para cada classe em matemática.

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, embrulhamento e incorporação. Entre as técnicas de filtro foram selecionados: ChiSquaredSubsetEval, SymmetricalUncertAttributeEval, CorrelationAttribute, OneRAttributeEval, GainRatio-AttributeEval, InfoGain-AttributeEval, ReliefFAttributeEval e o CfsSubsetEval. Para o algoritmo de embrulhamento foi utilizado o WrapperSubsetEval com o NaiveBayes; e os algoritmos REPTree e o J48 irão representar os algoritmos de incorporação.

Nesta etapa foram apresentados os atributos selecionados e as respectivas quantidades, utilizando a base de dados completa. Para este estudo, as notas dos alunos se mantiveram separadas entre Língua Portuguesa e Matemática. Os dados são mostrados na **Tabela 2**.

Tabela 2. Atributos selecionados para alunos do 9º ano.

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: 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 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.

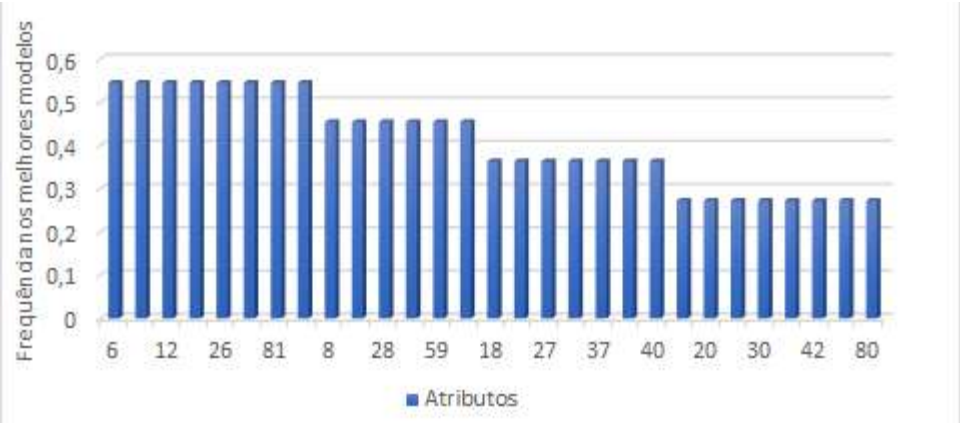

Figura 1. Ranking gerado pelo método de combinação de atributos Merge.
Fonte: Elaborada pelos autores

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 validação. 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.

A **Tabela 3** apresenta 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	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: Elaborada pelos autores

A melhor abordagem verificada, conforme a **Tabela 3**, 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 90,85% 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. 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.

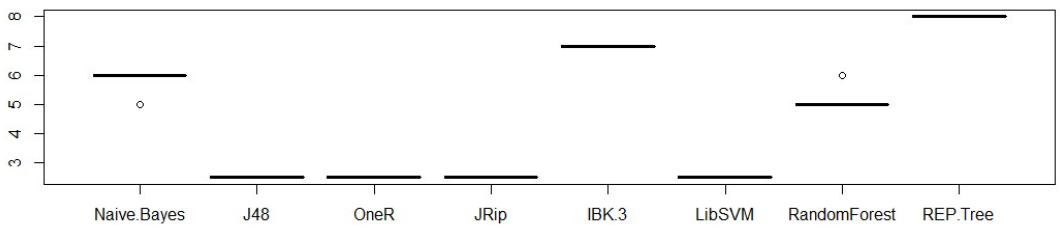

Figura 2. Aplicação do método de avaliação estatística de Friedman e Nemenyi no R para comparar as saídas dos classificadores.

Fonte: Elaborada pelos autores

A **Figura 2** apresenta o resultado do teste estatístico de Friedman e Nemenyi, 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 NaiveBayes apresentaram resultados satisfatórios, enquanto o REPTree apresentou o pior resultado entre todos.

O valor de distância crítica, conforme mostrado na **Figura 2**, 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.

5. Análise dos Resultados

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 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 4** apresenta os atributos que tiveram maior incidência por disciplina.

Tabela 4. Atributos com maior incidência.

Matéria	Atributo
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: Elaborada pelos autores

De acordo com a **Tabela 5** são colocados como destaques os atributos referentes às questões: 7, 26, 27, 30, 57, 58, 81 e 82; que se apresentam como fortemente impactante 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 5. Atributos mais relevantes.

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) que interfere diretamente no desempenho	▪ **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.

escolar dos alunos.	
▪ **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 como 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: Elaborada pelos autores

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.

6. Conclusões

Neste trabalho foi mostrado que a combinação de diferentes categorias de

seleção de atributos torna possível obter um conjunto de atributos ainda melhor em relação apenas um tipo de categoria. Sendo assim, uma vez que a seleção de atributos do tipo filtro é satisfatória para a etapa de processamento dos dados, está combinada com outras categorias enriquece a base de dados, tornando possível encontrar atributos que são imprescindíveis para análise de dados. Foi descoberto neste estudo 31 (trinta e um) atributos que influenciam o desempenho escolar do aluno nas duas disciplinas: língua portuguesa e matemática.

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 NaiveBayes apresentaram resultados satisfatórios, enquanto o REPTree apresentou o pior resultado entre todos.

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. Vale considerar que o impacto de cada atributo socioeconômico pode variar de acordo com a matéria, pois como foi visto a matéria de língua portuguesa apresentou uma maior correlação com os atributos socioeconômicos usados, e outro fator que deve ser considerado é a etapa da vida do aluno que será avaliada. Por exemplo, pode-se esperar resultados diferentes se aplicados às turmas de terceiro ano do ensino médio.

É importante frisar que o modelo em questão não é genérico, envolvendo um assunto particular sobre o desempenho acadêmico dos estudantes da Educação Básica.

Como contribuições do trabalho pode-se destacar 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.

7. Referências

BEZERRA, C.; SCHOLZ, R.; ADEODATO, P.; PONTES, T.; SILVA, I. (2016) **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).

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

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

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

INEP. (2016) **Prova Brasil**. Sistema de Avaliação da Educação Básica (Saeb). Disponível em: http://provabrasil.inep.gov.br/. Acessado em: 10 de setembro.

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

LIMA, R. A. F. et al. (2016) **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.

MANHÃES, L. M. B. (2015) **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.

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

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

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

PASTA, A. (2011) **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.

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

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