Avaliação de classificadores para relacionar características escolares a indicadores educacionais *

Doglas W. Sorgatto[1], Bruno M. Nogueira[1], Edson N. Cáceres[1], Henrique Mongelli[1]

[1]Faculdade de Computação – Universidade Federal do Mato Grosso do Sul (UFMS)
Av. Costa e Silva, s/nº – Bairro Universitário – 79.070-900 – Campo Grande – MS – Brazil

{doglas.sorgatto, bruno.nogueira, edson.caceres, h.mongelli}@ufms.br

Abstract. *In Brazil, there are many educational data on Basic Education, among these data, those that make up the School Census and educational indicators stand out. In this work, we propose an evaluation of nine classification algorithms that tried to relate the characteristics of schools in Mato Grosso do Sul with educational indicators - IDEB, ENEM and pass rate -, using machine learning techniques. The results were achieved in terms of accuracy and indicated that they existed between the characteristics of the schools and the educational indicators provided.*

Resumo. *No Brasil existem muitos dados educacionais sobre a Educação Básica, entre estes dados destacam-se os que compõem o Censo Escolar e os indicadores educacionais. Neste trabalho, propõe-se uma avaliação de nove algoritmos de classificação que procuraram relacionar as características das escolas de Mato Grosso do Sul com indicadores educacionais - IDEB, ENEM e taxa de aprovação -, utilizando técnicas de aprendizado de máquina. Os resultados foram analisados quanto à acurácia e indicaram existir relação entre as características das escolas e os indicadores educacionais analisados.*

1. Introdução

O Brasil possui diversos indicadores educacionais para a Educação Básica provenientes de avaliações de larga escala que permitem ter uma visão geral de todas as escolas do país, organizadas por etapa de ensino, modalidade, localização e dependência administrativa [Horta Neto 2007]. Com esses dados é possível verificar as características comuns das escolas com melhores, ou piores, resultados nestes indicadores.

Os indicadores educacionais são importantes, pois permitem avaliar efetividade dos programas públicos destinados à educação e a qualidade da educação básica, bem como, verificar resultado/retorno de investimentos na educação e permitem comparar instituições e avaliar o resultado de mudanças metodológicas, além de possibilitar o acompanhamento e fiscalização por parte da sociedade. O uso de indicadores, apesar de não ser garantia de sucesso para os programas e medidas públicas destinadas à educação, "potencializam as chances de sucesso, já que permitem, em tese, a avaliação dos resultados tecnicamente bem respaldados, além de diagnósticos sociais abrangentes e empiricamente referidos" [Rezende and Jannuzzi 2008].

*Artigo com financiamento da Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul – FUNDECT pelo termo de outorga 247/2020; e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico [426663/2018-7 e 433082/2018-6].

São vários os trabalhos existentes que analisam a aplicação e importância dos indicadores educacionais [Rezende and Jannuzzi 2008, Horta Neto 2007, Horta Neto 2018]. Estes trabalhos são apresentados e discutidos na fundamentação teórica e se caracterizam por defender o uso dos indicadores como ferramentas para acompanhamento da efetividade dos programas educacionais e apoiar a tomada de decisão sobre mudanças no currículo, metodologias de trabalho, investimentos e estratégias educacionais.

Sabendo que o conhecimento dos indicadores pode fornecer importantes informações para o acompanhamento da efetividade das políticas públicas, a análise das características das escolas que possuem os melhores e/ou piores indicadores permitiria identificar padrões e tendência e apoiar a tomada de decisão (maior investimento, corte de gastos, abertura ou encerramento de salas, contratação de profissionais etc.). Podem se beneficiar do conhecimento e da classificação das escolas por seus indicadores educacionais os gestores, professores e gerentes das mantenedoras, pois podem antecipar e apoiar em dados a tomada de decisão; gerenciar melhor os investimentos na educação e antecipar problemas no processo de ensino-aprendizagem. A sociedade também pode se beneficiar, pois pode acompanhar a evolução dos indicadores e realizar cobranças por melhorias ou mudanças no sistema de ensino [Rezende and Jannuzzi 2008].

Neste contexto, algoritmos de Aprendizado de Máquina (AM) emergem como uma importante ferramenta para predição de indicadores educacionais. Algoritmos de AM podem aprender funções que relacionem as características escolares com o desempenho nos indicadores educacionais, permitindo a generalização para dados de outras escolas. Com isso, pode-se realizar a predição dos indicadores garantindo a isonomia no processo de obtenção, tratamento e processamento dos dados, bem como, na análise dos resultados. Alguns trabalhos já utilizam a predição de indicadores educacionais, mas são focados no Ensino Superior [Motta et al. 2016, Silva Filho 2017], na educação a distância [Rabelo et al. 2017], ou em analisar os indicadores educacionais em si mesmo [Nascimento et al. 2018] não relacionando-os às características da escola.

O modelo de análise proposto utilizou dados do Censo Escolar, de onde foram extraídas as informações sobre as características da escola, e dos seguintes indicadores educacionais: Índice de Desenvolvimento da Educação Básica (IDEB), Exame Nacional do Ensino Médio (ENEM) e Taxa de Aprovação. Estas bases de dados foram organizadas por etapa de ensino e indicador educacional, sendo submetidas a nove algoritmos de classificação, cujos nomes, configurações e metodologia estão descritos na seção 4. Os resultados para o estudo de caso realizado com as escolas de Mato Grosso do Sul (MS) indicaram que o melhor classificador é o *Naive Bayes* e a maior correlação entre características da escola e indicadores ocorre quando se considera os valores do ENEM.

O Texto está organizado da seguinte forma: na próxima seção é apresentado a fundamentação teórica sobre mineração de dados educacionais (MDE); na Seção 3 são descritas as bases de dados dos indicadores educacionais selecionados; na Seção 4 é realizada uma breve descrição da metodologia empregada e os algoritmos selecionados para a avaliação; na Seção 5 são discutidos os resultados de acurácia encontrados e, por fim, as conclusões deste estudo.

2. Fundamentação Teórica

Desde a década de 90, o Brasil acumula dados educacionais oriundos de avaliações de larga escala que visam avaliar a qualidade e o desempenho do o sistema educativo [Horta Neto 2018, Vitelli et al. 2018]. Devido ao volume destes dados, para melhor analisá-los, se faz necessário o uso de técnicas de Mineração de Dados Educacionais (MDE), uma sub-área da Mineração de Dados tradicional com algumas especificidades [Romero and Ventura 2007, Campbell et al. 2007]. Com o uso de MDE, é possível empregar métodos da mineração de dados para realizar descoberta de padrões potencialmente úteis a partir de dados educacionais [Borges 2017, Maschio et al. 2018]. Com isso, é possível aliar grandes volumes de dados com técnicas estatísticas e modelos gerados por algoritmos de aprendizado de máquina para prover apoio à tomada de decisão orientada a dados [Campbell et al. 2007].

As avaliações de larga escala no Brasil são processos importantes, pois os resultados são fontes de dados para o estabelecimento de políticas públicas, para identificação de problemas educacionais, servem de critérios de classificação e medição da qualidade da educação. Este sistema pode gerar, como efeito negativo, o uso dos resultados para culpabilização dos profissionais da educação, especialmente quando os resultados interferem no fornecimento de recursos para a escola, e como pressão por mudanças no currículo, objetivando melhorar os resultados nas avaliações e não o processo de ensino aprendizagem [Bauer et al. 2015a, Bauer et al. 2015b].

Os dados oriundos das avaliações de larga escala, aplicadas no Brasil pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), são armazenados, tratados e divulgados pelo INEP através de seu site[1]. Estes dados fazem parte da Política Nacional de Dados Abertos (PNDA) [Brasil 2020] e estão disponíveis para serem utilizados em estudos, comparações e utilização da comunidade de pesquisadores, jornalistas e demais interessados na Educação Básica.

A MDE é uma atividade não trivial que pode influenciar a tomada de decisão no processo educacional devido ao conhecimento utilizável que produz [Borges et al. 2016]. Essa descoberta de conhecimento utilizável em sistemas educacionais tem sido investigada com o objetivo de auxiliar professores, administradores e pesquisadores a realizarem uma melhor gestão destes sistemas e, consequentemente, promover melhorias no processo educacional [Borges 2017, Baker et al. 2011, Rigo et al. 2014]. A maioria dos trabalhos em MDE são dirigidos a Educação Superior [Baker et al. 2011, Borges 2017], havendo escassez de trabalhos sobre a Educação Básica e isso se deve às necessidades específicas e aos problemas pedagógicos e administrativos que são próprios dessa fase da educação [Maschio et al. 2018].

Uma revisão sistemática sobre MDE no Brasil, cobrindo o período de 2001 a 2017, encontrou 49 trabalhos publicados e destes apenas 9 eram sobre a Educação Básica [Maschio et al. 2018]. Outros achados importantes sobre o uso de MDE no Brasil são:

- As técnicas de MDE mais utilizadas são aprendizado de máquina (29%), agrupamento (10%) e regras de associação (5%);
- Os tópicos de investigação mais abordados são desempenho dos aprendizes (43%) e comportamento dos estudantes em sua interação com Ambientes Virtuais de

[1]Disponível em http://www.gov.br/inep.

Aprendizagem (AVA) correspondendo a 29% dos trabalhos.
- O nível de escolaridade mais abordado é o superior (25 trabalhos entre 49 analisados), seguido de 14 trabalhos que não informam o nível de escolaridade analisado e apenas 9 trabalhos sobre a Educação Básica.
- A maioria dos trabalhos analisados trata de Ambientes Virtuais de Aprendizagem (AVA) utilizados na educação a distância, sendo os dados mais analisados aqueles relacionados ao número de interações (23%), dados pessoais (15%) e textos produzidos (10%).

A Educação Básica é a educação oferecida aos Brasileiros de 4 a 17 anos de idade, com matrícula obrigatória, abrangendo toda a população jovem do país. Ela é organizada em três etapas: a Educação Infantil, o Ensino Fundamental e o Ensino Médio. Sua organização é garantida pela Lei de Diretrizes e Bases da Educação Nacional (LDB lei 9.495/96) [Brasil 1996] e uma série de regulamentos específicos [Brasil 2018, Brasil 2013]. O que é importante sobre a Educação Básica, para o escopo deste trabalho, é que há muitos dados oriundos das avaliações em larga escala, mas que são subutilizados para a tomada de decisão em favor do processo educativo [Bauer et al. 2015b, Bauer et al. 2015a, Horta Neto 2007], e que há carência de estudos que possam auxiliar para melhorar este cenário [Maschio et al. 2018].

3. Bases de Dados Educacionais Utilizadas

O INEP publicou em 2020 o PNDA. Neste documento, é possível encontrar a relação de todas as bases de dados abertos sobre educação que são gerenciados pelo órgão, sendo fornecidas as seguintes informações: sítios de disponibilização; período de cobertura dos dados; situação quanto à atualização em cada sítio de disponibilização e informações sobre auditoria pelo Tribunal de Contas da União (TCU). O documento afirma em suas diretrizes que há o compromisso do INEP com a disponibilidade, integridade, autenticidade, sensibilidade e atualização periódica, de modo a garantir a perenidade dos dados abertos [Brasil 2020].

Para a execução deste trabalho, utilizaram-se bases de dados pertencentes ao PNDA, todas relativas ao ano de 2015. A escolha do ano de 2015 foi necessária para garantir a integridade e conformidade dos dados, pois este é o último ano em que o INEP divulgou a consolidação dos dados do ENEM por escola.

A base de dados do **Censo Escolar** contém informações sobre docentes, gestores, escolas, matrículas e turmas da Educação Básica de todo o Brasil, tanto de escolas públicas quanto privadas. Ele é o "principal instrumento de coleta de informações da Educação Básica e a mais importante pesquisa estatística educacional brasileira" [INEP 2020a]. Sua aplicação é regulamentada por instrumentos normativos que estabelecem obrigatoriedade, prazos, responsáveis e responsabilidades para a coleta de dados, sendo esta coordenada pelo INEP, em regime de cooperação com as secretarias estaduais e municipais de educação.

Da base de dados do Censo Escolar manteve-se como objeto de análise para classificação apenas as informações relacionadas às características das escolas (139 atributos). São informações sobre quantidade de salas, número de funcionários, equipamentos pedagógicos, localização, infraestrutura, entre outras características físicas e pedagógicas [INEP 2020a].

A segunda base de dados contém informações sobre o ENEM [INEP 2020b], empregado para avaliar o desempenho dos alunos da Educação Básica ao final desta etapa de ensino e o resultado deste indicador é utilizado como forma de acesso ao Ensino Superior em diversas instituições e para obtenção de financiamento estudantil. A base de dados possui resultados individuais, com o desempenho nas áreas de conhecimento de ciências humanas, ciências da natureza, linguagens e em produção de texto. O desempenho médio dos alunos por escola está disponível, pré-computado pelo INEP, apenas para o período entre 2005 e 2015 e seu último resultado divulgado serviu como restrição para a seleção das bases de dados empregadas neste estudo de caso.

A terceira base de dados utilizada contém as informações com o resultado do IDEB por escola [INEP 2020c]. Estas planilhas possuem informações detalhadas sobre os índices utilizados para calcular o IDEB de cada escola no período de 2003 a 2019, por etapa de ensino. As informações do IDEB estão disponíveis por etapa de ensino da educação básica: ensino fundamental anos iniciais, ensino fundamental anos finais e ensino médio. Esta última etapa começou a ser computada apenas em 2017, portanto os valores do IDEB do ensino médio não aparecem nesta análise.

O IDEB é um indicador de desempenho composto pela média do resultado da aplicação do Sistema de Avaliação da Educação Básica (SAEB) um conjunto de provas aplicadas aos alunos concluintes do quinto e nono ano do ensino fundamental e, a partir de 2017, aos alunos do ensino médio [INEP 2020c]. O outro valor que compõe o indicador é a taxa de aprovação média da escola. O IDEB é calculado por amostragem e com intervalo de dois anos entre as aplicações. O IDEB para o ensino médio começou a ser calculado pelo INEP apenas em 2017, e como a restrição temporal das bases de dados neste estudo se aplica a 2015, o IDEB para o ensino médio não foi utilizado na aplicação da avaliação de acurácia proposta, mas seu cálculo será considerado em trabalhos futuros.

A quarta base de dados contém informações sobre a taxa de rendimento educacional. Estas taxas computam a aprovação, reprovação e abandono dos alunos, por turma e etapa de ensino da Educação Básica. Os dados disponíveis cobrem o intervalo de 2007 a 2019, sendo utilizado os dados de 2015. A taxa de aprovação indica a quantidade percentual de alunos que progrediram para a série seguinte em seu processo educativo. Considera-se que quanto mais alta a taxa de aprovação, melhor o desempenho da escola em ensinar seus alunos[INEP 2020d].

Na Tabela 1 apresenta-se um resumo sobre as bases de dados formadas. Em cada linha, a primeira parte do nome representa o indicador analisado (ENEM, Aprovação e IDEB) e a segunda parte do nome contém a etapa de ensino (Inicial, Médio e Final). Além do nome das bases é apresentada a quantidade de atributos considerados para as escolas de MS, associadas com o indicador educacional analisado (139 atributos e o indicador), e a quantidade de escolas em cada base.

3.1. Criação das classes para predição dos indicadores

Todos os indicadores educacionais utilizados são valores contínuos, alguns variando de zero a mil, como o ENEM; outros variando de zero a dez, caso do IDEB; e, por fim, a taxa de aprovação, que é apresentada como um valor percentual. Estes valores contínuos precisaram ser transformados em categóricos, formando classes, para então serem processados pelos algoritmos de classificação.

Tabela 1. Composição das Bases de Dados

Base	Atributos	Escolas
ENEM	140	278
APROVACAO_FIN	140	841
APROVACAO_INI	140	1000
APROVACAO_MED	140	419
IDEB_FIN	140	440
IDEB_INI	140	592

Para a formação das classes, os indicadores foram submetidos a divisão estatística, pela média aritmética do indicador, de modo que foi possível classificar as escolas em duas classes: abaixo da média e aquelas com valor igual ou acima da média. Na Tabela 2 são apresentados as bases de dados, a média do indicador e a quantidade de escolas em cada classe criada.

Tabela 2. Formação das classes

Base	Média do indicador	Total de Escolas	Escolas Abaixo da Média	Escolas Acima da Média
ENEM	52.65	278	185	93
APROVACAO_FIN	84.34	841	381	460
APROVACAO_INI	90.62	1000	443	557
APROVACAO_MED	81.39	419	200	219
IDEB_FIN	4.36	440	214	226
IDEB_INI	5.24	592	305	287

4. Metodologia

Para a construção de modelos preditivos de indicadores educacionais da Educação Básica a partir de características das escolas, neste trabalho, foram utilizados nove algoritmos de classificação: *Decision tree*, *Extra tree*, *K Neighbors*, *Random Forest*, *Label Propagation*, *Gaussian Process*, *SGD Classifier*, *Bernoulli Naive Bayes* e *MLP Classifier*. Para implementação, foi utilizada a biblioteca *Scikit-learn*[Pedregosa et al. 2011], da linguagem de programação Python. A metodologia experimental utilizada está sumarizada na Figura 1.

Em um pré-processamento das bases, foi realizado o preenchimento dos valores ausentes com -1. Em seguida, a base foi embaralhada, para que a distribuição dos exemplos fosse aleatória e não influenciasse o resultado do processamento. Após isso, a fim de selecionar os atributos que mais influenciam na classificação, foi medida a informação mútua (MIC) entre os atributos e as classes. Foram testadas diferentes porcentagens de atributos com maior informação mútua para a classe, variando de 10% a 100% do total inicial.

Após os dados passarem pelo procedimento de pré-processamento, eles encontram-se aptos a serem apresentados para os algoritmos de classificação. A fim de obter o conjunto de parâmetros mais adequados para cada algoritmos em cada uma das bases de dados, um processo de ajuste de parâmetros foi realizado. Para evitar o enviesamento dos dados, já que a base usada no treino também seria utilizada no teste devido à pouca

quantidade de exemplos em algumas bases, optou-se por usar a melhor configuração retornada pelo *GridSearchCV* para cada algoritmo e aplicar o algoritmo puro, com a mesma proporção da base, mas com uma validação cruzada de *10-folds*. Esta etapa do teste produziu as informações de avaliação sobre a acurácia utilizadas para analisar os resultados dos algoritmos na Seção 5.

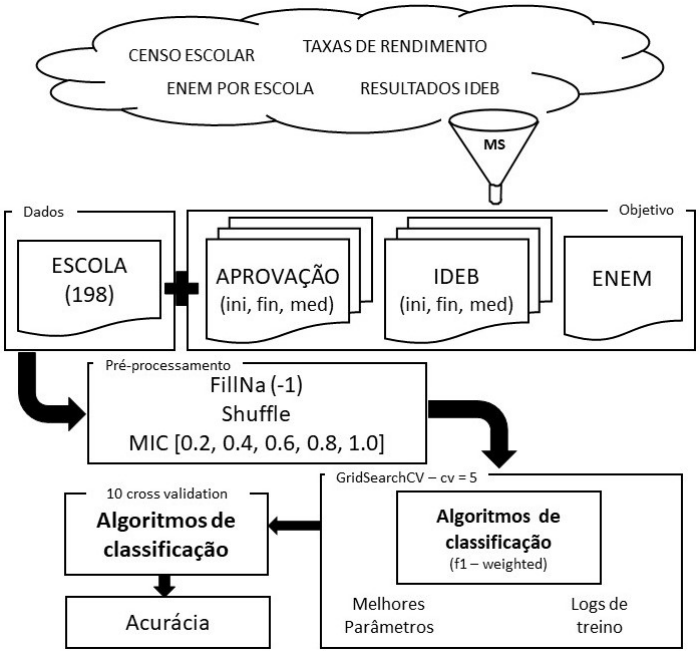

Figura 1. Esquema do modelo de análise utilizado

Os nove algoritmos de classificação utilizados e os parâmetros de entrada do *GridSearchCV* podem ser visualizados na Tabela 3. Estes algoritmos foram testados com as bases contendo as classes descritas na Tabela 2.

5. Análise dos resultados

Os nove algoritmos utilizados para realizar a classificação com as seis bases descritas na Tabela 1 da Seção 3, apresentaram resultados promissores.

Os valores de acurácia média e do desvio padrão para cada um dos nove algoritmos utilizados são apresentados na Tabela 4. Por questão de espaço, optou-se por não apresentar o valor de acurácia obtido para cada proporção retornada pela função MIC. Para compreensão das informações nas colunas considere: AP, bases sobre o indicador *Taxa de aprovação*; ID, bases sobre o indicador *IDEB*; FI, etapa ensino fundamental anos finais; IN, etapa ensino fundamental anos iniciais; ME, etapa ensino médio. Os valores acompanhados de asterisco indicam os melhores resultados para a base em análise.

Analisando as informações apresentadas na Tabela 4, para as bases que tratam do indicador **Aprovação**, observa-se que o melhor resultado para a base APROVACAO_FIN foi obtida com os algoritmos *Naive Bayes*, *MLP* e *Random Forest* cuja acurácia média foi de 71%. Analisando a acurácia média para todos os algoritmos, esta base apresentou 68% como resultado. Já a base APROVACAO_INI teve a melhor acurácia com o algoritmo *Naive Bayes*, com média de 67% e acurácia média, para todos os algoritmos testados, em

Tabela 3. Algoritmos e parâmetros utilizados

Algoritmo	Parâmetros para o teste
Decision Tree	criterion: gini, entropy splitter: best, random max_depth: 6, 10, 12, None
Extra Tree	criterion: gini, entropy splitter: best, random max_depth: 6, 10, 14, None
K Neighbors	n_neighbors: 3, 5, 7 weights: uniform, distance algorithm: auto, ball_tree, kd_tree, brute
Random Forest	criterion: gini, entropy n_estimators: 5, 10, 15, 20 max_depth: 4, 8, 10, None
Label Propagation	n_neighbors: 5, 7, 10 kernel: knn, rbf gamma: 5, 10, 20
Gaussian Process	kernel: RBF(1.0), 1.5*RBF(1.0), Matern(1.0)
SGD Classifier	loss: hinge alpha: 0.01, 0.001, 0.0001
Bernoulli Naive Bayes	alpha: 0.5, 1, 5, 10
MLP Classifier	hidden_layer_sizes: 50, 100, 200 activation: identity, logistic, tanh, relu solver: lbfgs, sgd, adam

Tabela 4. Acurácia média, por base e algoritmo

Algoritmo	AP-FI	AP-IN	AP-ME	ID-FI	ID-IN	ENEM
Decision Tree	0.70 (±0.01)	0.62 (±0.02)	0.68 (±0.01)	0.57 (±0.02)	0.54 (±0.01)	0.85 (±0.03)
Extra Tree	0.69 (±0.02)	0.63 (±0.01)	0.68 (±0.02)	0.59 (±0.03)	0.53 (±0.02)	0.85 (±0.02)
K Neighbors	0.63 (±0.04)	0.59 (±0.02)	0.66 (±0.02)	0.56 (±0.02)	0.53 (±0.01)	0.78 (±0.00)
Naive Bayes	0.71* (±0.02)	0.67* (±0.01)	0.70 (±0.01)	0.65* (±0.02)	0.54 (±0.01)	0.88* (±0.01)
Gaussian	0.66 (±0.06)	0.61 (±0.03)	0.69 (±0.03)	0.55 (±0.02)	0.57 (±0.00)	0.80 (±0.04)
Label Propagation	0.63 (±0.03)	0.58 (±0.02)	0.66 (±0.02)	0.55 (±0.01)	0.56 (±0.01)	0.78 (±0.01)
MLP	0.71* (±0.01)	0.65 (±0.01)	0.72* (±0.02)	0.64 (±0.01)	0.58* (±0.02)	0.86 (±0.01)
Random Forest	0.71* (±0.01)	0.65 (±0.01)	0.70 (±0.01)	0.62 (±0.02)	0.57 (±0.01)	0.88* (±0.01)
SGD Classifier	0.66 (±0.03)	0.59 (±0.02)	0.63 (±0.04)	0.54 (±0.01)	0.53 (±0.01)	0.74 (±0.07)
Média	**0.68**	**0.62**	**0.68**	**0.58**	**0.55**	**0.82**

62%. Para a base APROVACAO_MED o melhor algoritmo foi o *MLP*, com acurácia média de 72% e resultado final, para todos os algoritmos em 68%.

Ao analisar as informações da Tabela 4 para os valores referentes ao indicador **IDEB**, tem-se que, para a base IDEB_FIN o melhor algoritmo foi *Naive Bayes*, com

acurácia média de 65%, e com acurácia média de todos os algoritmos de 58%. Já para a base IDEB_INI, o melhor algoritmo foi o *MLP*, com acurácia média de 58%, e o resultado médio de todos os algoritmos ficou em 55%.

O último indicador educacional analisado é o **ENEM**. A base ENEM tem como melhor resultado aqueles retornados pelos algoritmos *Naive Bayes* e *Random Forest*, com acurácia média de 88%. O valor médio, para todos os algoritmos ficou em 82%.

Em resumo, é possível concluir que o melhor algoritmo de classificação, pelo ranking médio das performances de classificação, foi o *Naive Bayes*, seguido pelo *MLP*. Este resultado é especialmente interessante, uma vez que o algoritmo *Naive Bayes* é simples e de baixo custo computacional, o que permite a sua utilização escalável em bases de dados de maior dimensionalidade. Assim, este seria um classificador adequado a ser utilizado nas diferentes bases abordadas neste trabalho para a predição de indicadores utilizando bases de dados educacionais.

6. Conclusão

A partir do resultado dos algoritmos de classificação é possível concluir que há relação entre as características das escolas e os indicadores educacionais selecionados, para as escolas de Mato Grosso do Sul, indicando que o modelo de análise de indicadores educacionais a partir das características das escolas funciona para os indicadores educacionais analisados.

Quais características das escolas são determinantes para estes resultados é uma das questões que não foram abordadas neste trabalho, mas que podem ser analisadas em trabalhos futuros, com esse objetivo específico. Saber que existe uma relação entre as características das escolas e os indicadores educacionais é um passo importante para uma posterior análise dos elementos que caracterizam esta relação.

O uso de algoritmos de classificação para analisar a característica da escolas com os valores dos indicadores pode contribuir para desmitificar algumas opiniões sobre as escolas e os investimentos em educação, pois permite uma análise estatística, lógica e reprodutível dos dados disponíveis. Este trabalho mostrou que é possível encontrar relação entre os indicadores educacionais e as características das escolas.

O estudo poderá, em trabalhos futuros, ser reproduzido para escolas e indicadores de outros estados da federação, ou do Brasil todo, para confirmar os resultados ou ser o ponto de partida para a busca dos fundamentos das relações apresentadas pela análise de dados aqui iniciada. Também pretende-se testar o modelo com bases de dados e indicadores mais recentes, o que exigiria o tratamento dos dados do ENEM para obter seu valor agregado por escola para anos posteriores a 2015, e tratamento para estimativa do IDEB nos anos pares, utilizando predição por série temporal, por exemplo.

Referências

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, 19(02):03–13.

Bauer, A., Alavarse, O. M., and Oliveira, R. P. d. (2015a). Avaliações em larga escala: uma sistematização do debate. *Educação e Pesquisa*, 41(SPE):1367–1384.

Bauer, A., Pimenta, C. O., Horta Neto, J. L., and Sousa, S. Z. L. (2015b). Avaliação em larga escala em municípios brasileiros:o que dizem os números? *Estudos em Avaliação Educacional*, 26(62):326–352.

Borges, V. A. (2017). *Definição de um modelo de referência de dados educacionais para a descoberta de conhecimento*. PhD thesis, USP - ICMC, São Carlos - SP.

Borges, V. A., Nogueira, B. M., and Barbosa, E. F. (2016). A multidimensional data model for the analysis of learning management systems under different perspectives. In *Frontiers in Education Conference (FIE)*, pages 1–8, Erie, PA, USA. IEEE.

Brasil (1996). Lei nº 9.394, de 20 de dezembro de 1996. estabelece as diretrizes e bases da educação nacional. *Diário Oficial [da] República Federativa do Brasil*.

Brasil (2013). *Diretrizes Curriculares Nacionais da Educação Básica*. MEC, SEB, DICEI, Brasília, DF.

Brasil (2018). *Base Nacional Comum Curricular*. MEC, SEB, CNE, Brasília, DF.

Brasil (2020). *Política e plano de dados abertos do INEP (Biênio 2020-2021)*. MEC, INEP, Brasília, DF.

Campbell, J. P., DeBlois, P. B., and Oblinger, D. G. (2007). Academic analytics: A new tool for a new era. *EDUCAUSE review*, 42(4):40.

Horta Neto, J. L. (2007). Um olhar retrospectivo sobre a avaliação externa no brasil: das primeiras medições em educação até o saeb de 2005. *Revista Iberoamericana de Educación*, 42(5):3.

Horta Neto, J. L. (2018). Avaliação educacional no brasil para além dos testes cognitivos. *Revista de Educação PUC-Campinas*, 23(1):37–53.

INEP (2020a). Censo escolar. Diponível em: http://portal.inep.gov.br/censo-escolar. Acessado em mar 2020.

INEP (2020b). Enem - exame nacional do ensino médio. Disponível em: http://portal.inep.gov.br/web/guest/enem. Acessado em abr 2020.

INEP (2020c). Ideb - Índice de desenvolvimento da educação básica. Disponível em: http://portal.inep.gov.br/web/guest/educacao-basica/ideb. Acessado em mar 2020.

INEP (2020d). Indicadores educacionais. Disponível em: http://portal.inep.gov.br/indicadores-educacionais. Acessado em mar 2020.

Maschio, P., Vieira, M., Costa, N., Melo, S. M., and Júnior, C. (2018). Um panorama acerca da mineração de dados educacionais no brasil. *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE)*, 29(1):1936.

Motta, P. R. d. A. et al. (2016). Estudo exploratório do uso de classificadores para a predição de desempenho e abandono em universidade. Master's thesis, Universidade Federal de Goiás.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rabelo, H., Burlamaqui, A., Valentim, R., de Souza Rabelo, D. S., and Medeiros, S. (2017). Utilização de técnicas de mineração de dados educacionais para predição de desempenho de alunos de ead em ambientes virtuais de aprendizagem. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, page 1527.

Rezende, L. M. d. and Jannuzzi, P. d. M. (2008). Monitoramento do plano de desenvolvimento da educação: proposta de aprimoramento do ideb e de painel de indicadores. *Revista do Serviço Público (RSP)*.

Rigo, S. J., Cambruzzi, W., Barbosa, J. L., and Cazella, S. C. (2014). Aplicações de mineração de dados educacionais e learning analytics com foco na evasão escolar: oportunidades e desafios. *Revista Brasileira de Informática na Educação*, 22(01):132.

Romero, C. and Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, 33(1):135–146.

Silva Filho, R. L. C. (2017). Modelo de análise e predição do desempenho dos alunos dos institutos federais de educação usando o enem como indicador de qualidade escolar. Master's thesis, Universidade Federal de Pernambuco.

Vitelli, R. F., Fritsch, R., and Corsetti, B. (2018). Indicadores educacionais na avaliação da educação básica e possíveis impactos em escolas de ensino médio no município de porto alegre, rio grande do sul. *Revista Brasileira de Educação*, 23.