Estudo sobre Docentes do Ensino Básico através de Indicadores Educacionais e Modelos de Regressão

Rafaella Leandra Souza do Nascimento[1], **Geraldo Gomes da Cruz Júnior**[2,3]

[1]Escola Politécnica de Pernambuco – Universidade de Pernambuco (UPE) – Recife, PE – Brasil

[2]Departamento de Estatística e Informática – Universidade Federal Rural de Pernambuco (UFRPE) – Recife, PE – Brasil

[3]Instituto SENAI de Inovação para Tecnologias da Informação e Comunicação (ISI-TICs) – Recife, PE – Brasil

{rafaellalsn, geraldoj8}@gmail.com

Abstract. *This work uses educational databases provided by INEP and applies regression techniques with the purpose of estimating the indicator of teachers with higher education in the scenario of basic education. This application of Educational Data Mining is based on the phases of the CRISP-DM methodology. After correlational analyzes of the variables, linear and robust regression models were applied in order to compare the performance and provide a model that minimizes the prediction error. The models were evaluated by absolute mean error in addition to graphs and statistical tests. The results indicate that the robust regression obtained better results in the estimation of the variable listed in this study.*

Resumo. *Este trabalho utiliza bases de dados educacionais fornecidas pelo INEP e aplica técnicas de regressão com o objetivo de estimar o indicador de docentes com curso superior no cenário da educação básica. Essa aplicação de Mineração de Dados Educacionais segue como base as fases da metodologia CRISP-DM. Após análises correlacionais das variáveis, modelos de regressões linear e robusta foram aplicados com a finalidade de comparar o desempenho e fornecer um modelo que minimize o erro de previsão. Os modelos foram avaliados pelo erro médio absoluto além de gráficos e testes estatísticos. Os resultados indicam que a regressão robusta obteve melhores resultados na estimação da variável elencada nesse estudo.*

1. Introdução

Os indicadores educacionais atribuem valor estatístico à qualidade do ensino, atendo-se não somente ao desempenho dos alunos mas também ao contexto econômico e social em que as escolas estão inseridas [INEP 2018]. Eles consideram informações como o acesso, permanência, a aprendizagem dos alunos. No Brasil, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) é responsável pelo levantamento e divulgação de informações sobre a educação no país, em todas etapas de ensino, por meio de avaliações, exames e indicadores.

Por meio do levantamento e estudo desses dados coletados em ambientes educacionais pode-se desenvolver métodos para explorar tais dados utilizando Mineração de

DOI: 10.5753/cbie.wcbie.2018.379

Dados Educacionais (MDE). De forma geral, MDE é uma abordagem que tem como objetivo a descoberta de informações que ajudem na proposta educacional, no melhoramento das condições de infraestrutura escolar, no processo ensino, na previsão de desempenho dos alunos, além de outros fatores que influenciam a aprendizagem. Nesse sentido, diversos trabalhos vem sendo desenvolvidos utilizando MDE para fins de tomada de decisão [Rodrigues et al. 2013].

A MDE consiste na aplicação de técnicas da Mineração de Dados (MD) nos conjuntos de dados obtidos nos mais variados contextos educacionais, os quais diversificados, demandam adaptações e novas técnicas. Ao mesmo tempo, essa diversidade representa potencial de implementar resoluções de problemas nos mais diferentes setores da educação [Rigo et al. 2012]. Existem várias linhas de pesquisa na área de educação e muitas delas derivadas diretamente da área de MD, como tarefas preditivas, de agrupamento ou associação [Baker et al. 2011].

No campo da predição, técnicas podem ser aplicadas para encontrar estruturas ou associações em conjunto de dados, realizar previsões, entre outros. Dentre elas, destacam-se modelos de regressão, que são modelos matemáticos e tem como um dos objetivos prever o valor da variável dependente (Y) a partir das informações provenientes de uma variável ou um conjunto de variáveis independentes (X) [Montgomery et al. 2012]. Esse tipo de técnica pode estimar o relacionamento entre benefícios e problemas educacionais por meio das variáveis consideradas no estudo.

Portanto, o objetivo desse trabalho consiste em investigar bases de dados educacionais fornecidas pelo INEP e estimar o indicador de docentes com curso superior da educação básica do ensino fundamental, mediante a aplicação de regressão linear simples e robusta. Dessa forma, este trabalho realiza uma análise de qual das duas técnicas minimiza o erro de previsão da variável alvo. Utiliza-se como metodologia o CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Chapman et al. 2000].

2. Trabalhos Relacionados

Em Rodrigues et al. (2013), o objetivo consistiu em utilizar o modelo de regressão linear para a obtenção de inferências em etapas iniciais da realização de cursos *online*, obtendo como resultado boas taxas de precisão. Em relação a estimar variáveis a partir de indicadores educacionais utilizando de modelos de regressão, em do Nascimento et al. (2018a) aplica-se técnicas de regressão linear e robusta com a finalidade de melhor explicar indicadores como a evasão e reprovação escolar no ensino fundamental. Ainda, em do Nascimento et al. (2018b), estima-se a evasão escolar aplicando técnicas como a regressão quantílica não-paramétrica e a *Support Vector Regression* (SVR).

Outras técnicas de MD são abordadas em trabalhos a partir de indicadores e exames educacionais fornecidos pelo INEP. Como em da Fonseca e Namen (2016) que utilizam base do Saeb com o intuito de identificar fatores que relacionam o perfil de professores de Matemática com a proficiência obtida por seus alunos. Em Simon e Cazella (2017) o objetivo é gerar um modelo preditivo do indicador de desempenho médio em ciências da natureza e suas tecnologias dos alunos de escolas do ensino médio a partir do ENEM. Utiliza-se árvores de decisão nos experimentos. Por fim, em de Souza et al. (2017) buscou-se analisar o desempenho de um curso no ENADE levando em consideração o desempenho dos alunos obtido na prova do ENEM e suas características

sociais e econômicas, a partir da técnica de regressão logística e árvores de decisão.

Tendo em vista a abordagem destes trabalhos apresentados, a realização dessa pesquisa traz contribuição uma vez que as atividades preditivas no contexto da MDE, em especial a regressão, é um tema importante. Apesar de muitos trabalhos levarem em consideração aplicações a partir de indicadores e exames educacionais, esse tópico carece de outros pontos de análises. Dessa forma, utilizando a abordagem de regressão, as análises de previsão de indicadores educacionais voltados aos docentes da educação básica são realizadas.

3. Metodologia

O CRISP-DM define uma sequência de seis fases, que permite a construção e implementação de um modelo de MD para ser usado em um ambiente real. Essa metodologia define um projeto como um processo iterativo, no qual etapas podem ser usadas para permitir o resultado final baseado nos objetivos de MD. As fases consistem em:

1. Compreensão do Problema: Essa fase inicial se concentra na compreensão dos objetivos e requisitos do problema abordado, convertendo esse conhecimento em uma definição de problema de MD.
2. Compreensão dos Dados: A fase de compreensão dos dados permite familiarizar-se com os dados, identificar problemas de qualidade, descobrir primeiros *insights* sobre os dados e detectar subconjuntos interessantes para formar hipóteses sobre informações relevantes.
3. Preparação dos Dados: Nessa etapa, as tarefas incluem seleção de atributos, além de transformação e limpeza de dados para melhor aplicação das técnicas.
4. Modelagem: Nessa fase, técnicas de MD são selecionadas e aplicadas. Representa o desenvolvimento dos modelos para o problema, com base nos dados que já foram adequados para serem utilizados.
5. Avaliação: Nessa fase do projeto o modelo desenvolvido é avaliado para ratificar a sua adequação ao problema em estudo.
6. Aplicação: Todo conhecimento obtido por meio do trabalho de MD tornam-se subsídios para o desenvolvimento de estratégias para tomada de decisão no contexto do problema estudado.

Tendo em vista esse processo de desenvolvimento de soluções em MD, a aplicação desenvolvida nesse trabalho possui como base as fases do CRISP-DM. Os próximos subcapítulos descrevem as atividades desenvolvidas em cada fase.

3.1. Compreensão do Problema

Nessa primeira etapa, foi realizada um estudo relacionado a cenários educacionais, MDE e modelos preditivos. Posteriormente, foram listadas as principais variáveis educacionais que envolvem essa pesquisa, obtidas a partir de bases de dados educacionais fornecidas abertamente pelo INEP [INEP 2018].

3.2. Compreensão dos Dados

Os dados utilizados nesse trabalho são fornecidos abertamente pelo INEP em seu portal, e referem-se a indicadores educacionais da educação básica referentes aos docentes. Pode-se obter esses dados em diferentes granularidades como nível nacional, regional ou nível

das escolas. Para esse estudo é considerado o nível das escolas, ou seja, a menor granularidade significa maior detalhamento e volume dos dados. Os indicadores educacionais considerados são:

- Adequação da formação dos docentes: A base fornece o percentual de docentes por grupo de adequação da formação à disciplina que leciona. As categorias de adequação da formação dos docentes em relação à disciplina que lecionam são mostradas na Tabela 1.

Tabela 1. Categorias da adequação da formação dos docentes em relação a disciplina lecionada.

	Descrição
Grupo 1	Docentes com formação superior de licenciatura (ou bacharelado com complementação pedagógica) na mesma área da disciplina que leciona.
Grupo 2	Docentes com formação superior de bacharelado (sem complementação pedagógica) na mesma área da disciplina que leciona.
Grupo 3	Docentes com formação superior de licenciatura (ou bacharelado com complementação pedagógica) em área diferente daquela que leciona.
Grupo 4	Docentes com formação superior não considerada nas categorias anteriores.
Grupo 5	Docentes sem formação superior.

- Docente com curso superior: A base possui o percentual de docentes com curso superior. Essa é a variável explicativa deste trabalho.
- Esforço do docente: Apresenta o percentual de docentes que atuam no ensino fundamental e ensino médio por nível de esforço necessário para o exercício da profissão. Os níveis do indicador são descritos na Tabela 2 de acordo com as características usuais dos docentes pertencentes a cada um deles.

Tabela 2. Níveis do esforço docente da educação básica.

	Descrição
Nível 1	Docente que, em geral, tem até 25 alunos e atua em um único turno, escola e etapa.
Nível 2	Docente que, em geral, tem entre 25 e 150 alunos e atua em um único turno, escola e etapa.
Nível 3	Docente que, em geral, tem entre 25 e 300 alunos e atua em um ou dois turnos em uma única escola e etapa.
Nível 4	Docente que, em geral, tem entre 50 e 400 alunos e atua em dois turnos, em uma ou duas escolas e em duas etapas.
Nível 5	Docente que, em geral, tem mais de 300 alunos e atua nos três turnos, em duas ou três escolas e em duas etapas ou três etapas.
Nível 6	Docente que, em geral, tem mais de 400 alunos e atua nos três turnos, em duas ou três escolas e em duas etapas ou três etapas.

Então, formou-se uma base de dados contendo todas as variáveis citadas considerando o âmbito do ensino fundamental. Cada base de dados possui uma coluna que

representa o código identificador de cada escola, e utilizando esse código relacionou-se as bases. As 12 variáveis são mostradas na Tabela 3.

Tabela 3. Variavéis presentes nas bases.

Variáveis da base		
1	DSU	Docente com curso superior
2 a 6	AFD	Percentual de formação docente por grupo (1 a 5)
7 a 12	IED	Nível de esforço docente necessário (1 a 6)

3.3. Processamento dos Dados

Nessa fase foi realizada uma análise das variáveis e transformações destas, de acordo com as necessidades identificadas. Inicialmente, foi realizada uma seleção dos dados para um âmbito de estudo, pois o número de possibilidades de cenários a ser estudado é alto. Então, os indicadores relacionados ao cenário do estado de Pernambuco foram utilizados para essa aplicação. Após isso, foi possível observar valores ausentes para algumas variáveis, e para resolver esse problema, foi realizada a inserção de valores utilizando a mediana dos valores das colunas. Isso foi realizado uma vez que o número de *missing values* não foi grande e o objetivo foi ter o mínimo de perda de instâncias.

Nessa fase, também foi aplicada a normalização dos dados, que consiste em ajustar a escala dos valores dos atributos para que fiquem em pequenos intervalos. Para esse trabalho os dados foram normalizados entre 0 a 1. Após realizar as tarefas presentes nessa fase, forma-se as bases finais para o ensino fundamental com um total de 12 variáveis, conforme mostra a Tabela 4.

Tabela 4. Bases de dados após preparação dos dados.

Base de dados	Cenário	Nº de Instâncias
I	Ensino Fundamental	8.133

3.4. Modelagem

Nesse trabalho, será implementado modelos de regressão. Para gerar os modelos, é realizada a análise correlacional entre as variáveis.

3.4.1. Análise correlacional entre as variáveis

O grau de associação entre y e x pode ser medido pelo coeficiente de correlação. Uma medida de correlação comum que reflete o grau de relacionamento linear entre duas variáveis é o coeficiente de correlação de Pearson (r). O coeficiente r pode assumir valores entre -1 e 1, o que significa $r = 1$ uma correlação perfeita positiva e $r = -1$ correlação perfeita negativa. Quanto mais próximo de 0 o r, torna-se mais fraca a correlação. A Figura 1 mostra os índices de correlação entre as variáveis elencadas para esse estudo, sendo a correlação positiva representada pela cor azul, e a negativa pela cor vermelha.

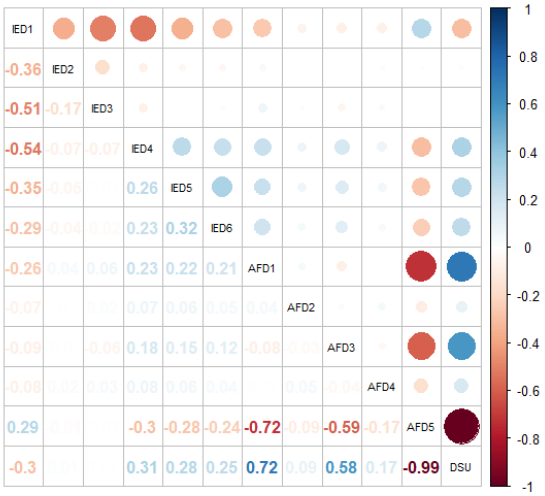


Figura 1. Matriz de Correlação entre as variáveis

Como pode-se observar na matriz de correlação, as variáveis com maior correlação positiva a variável resposta (DSU) consistem na AFD1 (0,72) e AFD3 (0,58). Isso significa que a medida que a taxa de professores nas escolas que possuem ensino superior cresce, também cresce o número de professores com formação docente do grupo 1 e 3. Ou seja, são professores com formação docente em licenciatura. Em relação as variáveis IED4 (0,31), IED5 (0,28) e IED6 (0,25), apesar de representar correlações positivas inferiores com a DSU, mostra que há uma tendência (baixa) a docentes que possuem um esforço de exercer a profissão maior, com mais alunos e atuando em mais turnos.

Em análise as variáveis de maior correlação negativa, destaca-se a AFD5 (-0,99). Esse valor é esperado, já que essa variável representa a taxa de professores que não possuem formação superior, ou seja, há uma relação inversa a DSU. Ainda pode-se verificar na matriz de correlação que a variável AFD5 possui, em relação as demais variáveis, um relacionamento oposto ao comparar com a DSU. Outra variável que possui uma correlação negativa com a variável resposta é IED1 (-0,3), o que representa uma relação inversa entre a taxa de formação docente e a taxa de professores que possuem um esforço mais baixo para exercer a profissão (menos alunos e turmas).

Portanto, pode-se elencar a variável com maior correlação a variável DSU, sendo a AFD1. Apesar da variável AFD5 possuir um maior valor correlacional (negativo), não foi considerado pois essa variável representa o oposto da DSU.

3.5. Modelos de Regressão

A regressão é uma técnica que permite inferir a relação de uma variável resposta (y) com uma ou várias variáveis explicativas (x). Diferentes modelos podem ser considerados para essa estimação. O modelo de Regressão Linear (referenciada como RL) [Montgomery et al. 2012] definida para essa situação consiste em

$$y_i = \alpha + \beta x_i + \epsilon_i \ (i = 1, 2, ..., n) \quad (1)$$

onde, α representa o intercepto da reta com o eixo dos y, β o parâmetro que representa a variação de y em função da variação de x, x a variável explicativa, e $\epsilon = (\epsilon_1, ...$

, ϵ_n) é um vetor de erro aleatório da i-ésima observação. Os parâmetros β são estimados minimizando uma função baseada no Método dos Mínimos Quadrados (MMQ), que é dada por

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{2}$$

onde, n é a quantidade de observações, y_i é a variável resposta real e $\widehat{y}_i$ a variável resposta estimada. No entanto, a reta ajustada pelo MMQ é fortemente influenciada pelos *outliers*, pois o método minimiza a soma dos resíduos ao quadrado. Então, para ajustar um modelo que produza boas estimativas dos parâmetros β na presença de *outliers*, ou quando y não segue uma distribuição normal, outros métodos podem ser utilizados para obter a relação linear entre as variáveis. Como exemplo, tem-se a Regressão Linear Robusta (RLR), a qual utiliza M-estimadores, no qual o vetor β é estimado minimizando uma função critério baseada na função ρ. Essa função critério é dada por

$$\sum_{i=1}^{n} \rho(\frac{y_i - \beta x_i}{\sigma}) \tag{3}$$

onde, n é o tamanho da amostra, y_i é o variável resposta real, x_i o vetor da variável explicativa, β o vetor de parâmetros, σ é um estimador robusto e ρ uma função particular.

O modelo final das regressões foi obtido utilizando a variável explicativa x com maior valor correlacional a variável resposta y, conforme análises realizadas. Foram gerados 2 modelos diferentes, um para a regressão linear e para a regressão robusta. A configuração experimental se deu por execuções em 30 iterações. A estimação dos valores é realizada por meio do método *holdout*, o qual particiona os dados em 25% para a base de teste e 75% para a base de treino do modelo. Todos experimentos e análises foram implementados no ambiente *open source* R.

3.6. Avaliação

O índice de desempenho utilizado nesse trabalho é o *Mean Absolute Error* (MAE), ou erro médio absoluto, no qual os erros são tratados igualmente de acordo com a sua magnitude. O MAE é definido na Equação 4,

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{4}$$

onde $\hat{y} = \{\hat{y}_1, ..., \hat{y}_n\}$ são os valores preditos e $y = \{y_1, ..., y_n\}$ são os valores reais da variável resposta. Com as amostras geradas após as execuções dos experimentos, pode-se calcular a média e desvio padrão do erro, realizar testes estatísticos, gráficos *boxplots* para assim também avaliar o desempenho dos modelos de regressão.

3.7. Aplicação

Todo o conhecimento obtido por meio da MDE fornece subsídio para conhecer as variáveis educacionais abordadas nesse trabalho. Portanto, serão listadas considerações para o cenário estudado após todas as etapas da metodologia proposta.

4. Resultados

Para o cenário estudado, a Tabela 5 apresenta os modelos construídos para a estimação do percentual de docentes com curso superior. A Tabela 6 fornece os resultados da média do erro calculado e o desvio padrão associado entre o modelo RL e RLR com as amostras geradas após as execuções. Como pode ser analisado, o erro da regressão robusta possui menor valor em comparação a regressão linear. Apesar do desvio padrão do erro da regressão linear ser um pouco inferior, não consiste numa diferença significativa.

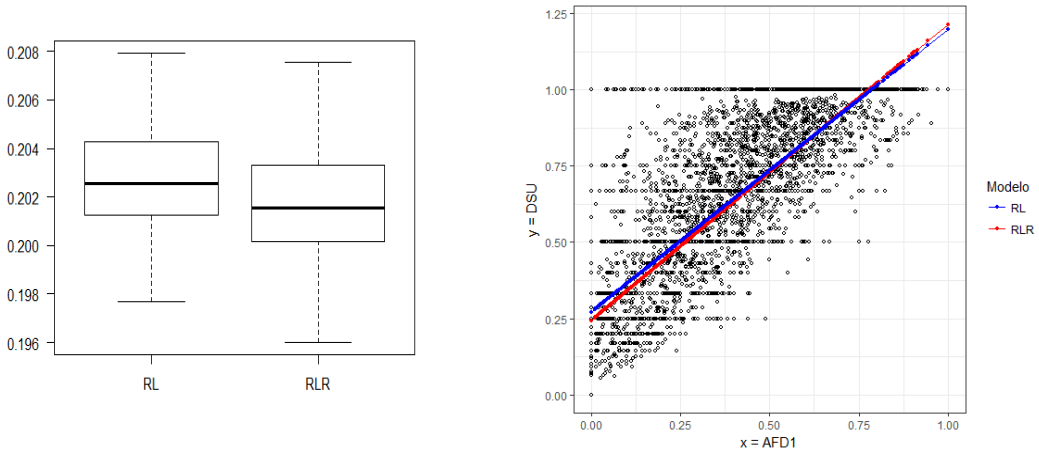
Tabela 5. Modelos para a estimação do percentual de docentes com curso superior.

Fórmula = DSU ∼AFD1			
RL		**RLR**	
Coeficients:		Coeficients:	
(Intercept)	AFD1	(Intercept)	AFD1
0.2713	0.9240	0.24150	0.97149

Tabela 6. Resultados dos experimentos para a média e desvio padrão das amostras do MAE.

Modelo	**Média MAE**	**Desvio Padrão MAE**
RL	0,2028045	0.002423
RLR	**0,2017858**	0,002637

Outra forma de analisar o resultado para esse cenário é por meio da Figura 2 (a). Ela mostra o *boxplot*, que representa a variação de dados contidos numa amostra, no caso, a variação do valor do erro associado a partir das 30 execuções de cada modelo de regressão. Como pode ser visto, RLR obteve uma menor mediana para o valor do erro (representada pela linha central na caixa). Também possui menor valores em comparação a regressão linear (indicadas pelas linhas horizontais acima e abaixo das caixas).



(a) *Boxplots* (b) Retas de Regressão

Figura 2. Comparação dos resultados experimentais entre o modelo RL e RLR.

Já a Figura 2 (b) mostra as curvas de regressão formadas a partir de uma execução de cada modelo de regressão. Como pode ser visto, em relação a dispersão dos pontos no gráfico, os modelos buscam ajuste aos dados, no entanto a variação entre retas entre si é pequena, com uma baixa inclinação do modelo RL nas extremidades.

Por fim, o teste de normalidade Shaphiro-Wilk [Shaphiro and Wilk 1965] foi aplicado nas amostras geradas a partir das simulações. Devido as amostras apresentarem uma distribuição normal, foi estabelecido a aplicação do teste estatístico T-Student [Student 1908] para amostras emparelhadas. O teste aconteceu sobre a média do MAE das duas amostras formadas a partir das execuções dos modelos, a fim de verificar se estatisticamente RLR é menor do que RL. Ou seja, se utilizar o modelo RLR é mais vantajoso para a estimação da variável resposta.

Considerando-se um nível de significância de 5%, o *p-value* é mostrado na Tabela 7. O resultado do teste mostra que, estatisticamente, há evidências para aceitar diferença entre os modelos e que a amostra do modelo RLR é menor do que a do modelo RL.

Tabela 7. Resultado do teste de hipótese T-Student.

Hipótese	**p-value** (α = 5%)
RLR $<$ RL	2,2e-16

5. Conclusões e Trabalhos Futuros

Esse trabalho buscou investigar e variáveis educacionais que podem estar relacionadas com o índice de professores que possuem curso superior (DSU). Os modelos aplicados nos experimentos contam com os resultados de técnicas lineares de regressão (simples e robusta). Através de bases de diferentes indicadores educacionais fornecidas pelo INEP, pode-se estudar variáveis para realizar os experimentos, seguindo as fases do CRISP-DM.

O entendimento do relacionamento entre as variáveis da base, por meio da análise correlacional, tornou possível a implementação de um modelo com a variável mais correlacionada com a variável resposta (DSU). Pode ser observado que variáveis relacionadas a formação docente em licenciatura e variáveis que representam maior esforço do docente para lecionar, mostraram-se mais associadas a DSU. Como resultados, as análises dos experimentos realizados mostraram melhores resultados da regressão robusta em relação a regressão linear, para o âmbito do ensino fundamental do estado de Pernambuco. A regressão robusta obteve menor erro de predição, representado na tabela e gráfico *boxplots*.

Portanto, uma contribuição desse trabalho consiste em fornecer uma aplicação de análise correlacional e de regressores que minimiza o erro de predição entre variáveis e indicadores educacionais relacionados a docentes do ensino básico. Assim, utilizar a mineração de dados educacionais possibilita a identificação prévia de aspectos voltados a área educacional. São ferramentas que podem ser aplicadas de forma ampla, gerando conhecimento, servindo como base para soluções de problemas e desenvolvimento de mecanismos em apoio ao ensino. Dessa forma, como trabalhos futuros, pretende-se abordar outras técnicas de previsão, assim como explorar outros aspectos educacionais e expandir o escopo de estudo para outros níveis de escolaridade.

Referências

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineraçao de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, 19(02):03.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., and Wirth, R. (2000). Crisp-dm 1.0 step-by-step data mining guide. *CRISP-DM Consortium*.

da Fonseca, S. O. and Namen, A. A. (2016). Mineração em bases de dados do inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. *Educação em Revista*, 32(1).

de Souza, H. V. L., Neiva, D., Cavalcanti, R. P., Rodrigues, R., Gomes, A. S., and Adeodato, P. (2017). Uma análise preditiva de desempenho dos alunos dos cursos no enade com base no perfil socioeconômico e de desempenho no enem. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 6, page 684.

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

do Nascimento, R. L. S., das Neves Junior, R. B., de Almeida Neto, M. A., and de Araújo Fagundes, R. A. (2018b). Educational data mining: An application of regressors in predicting school dropout. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 246–257. Springer.

INEP (2018). Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. Disponível em: http://portalinep.gov.br/. Acesso em 14/03/2018.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). *Introduction to linear regression analysis*, volume 821. John Wiley & Sons.

Rigo, S. J., Cazella, S. C., and Cambruzzi, W. (2012). Minerando dados educacionais com foco na evasão escolar: oportunidades, desafios e necessidades. In *Anais do Workshop de Desafios da Computação Aplicada à Educação*, pages 168–177.

Rodrigues, R. L., de Medeiros, F. P., and Gomes, A. S. (2013). Modelo de regressão linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 24, page 607.

Shaphiro, S. and Wilk, M. (1965). An analysis of variance test for normality. *Biometrika*, 52(3):591–611.

Simon, A. and Cazella, S. (2017). Mineração de dados educacionais nos resultados do enem de 2015. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 6, page 754.

Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.