Mineração de Dados Educacionais: Um Estudo Sobre Indicadores da Educação em Bases de Dados do INEP

[1]Rafaella Leandra Souza do Nascimento, [2,3]Geraldo Gomes da Cruz Junior, [1]Roberta Andrade de Araújo Fagundes

[1]Universidade de Pernambuco (UPE) – PE – Brasil
[2]Universidade Federal Rural de Pernambuco (UFRPE) – PE – Brasil
[3]Instituto SENAI de Inovação para Tecnologias da Informação e Comunicação (ISI-TICs) – PE – Brasil

{rlsn@ecomp.poli.br, geraldoj8@gmail.com, roberta.fagundes@upe.br}

Resumo: A Mineração de Dados Educacionais possibilita o conhecimento de fatores que melhorem a proposta educacional, além de prever o desempenho dos alunos e de fatores que influenciam o aprendizado. Através dessas características, esse trabalho utiliza bases de dados educacionais fornecidas pelo INEP e aplica técnicas de mineração de dados com a finalidade de melhor explicar indicadores como a evasão e reprovação escolar no ensino fundamental. Após análises correlacionais das variáveis, modelos de regressões linear e robusta foram desenvolvidos afim de comparar o desempenho e fornecer um modelo que minimize o erro de previsão. Os modelos foram avaliados pelo erro médio absoluto, além de desvio padrão e gráficos. Os resultados indicam que a regressão robusta obteve melhores resultados na estimação das variáveis elencadas nesse estudo.

Palavras-chaves: Indicadores Educacionais; MDE; Regressão; Correlação.

Educational Data Mining: A Study on Education Indicators in INEP Databases

***Abstract*:** *Educational Data Mining makes it possible to know factors that improve the educational proposal, as well as to predict student performance and factors that influence learning. Through these characteristics, this work uses educational databases provided by INEP and applies data mining techniques in order to better explain indicators such as dropout, and school disapproval in primary school. After correlational analyzes of the variables, linear and robust regression models were developed in order to compare the performance and provide a model that minimizes the prediction error. The models were evaluated by absolute mean error, as well as standard deviation and graphs. The results indicate that the robust regression obtained better results in the estimation of the variables listed in this study.*

***Keywords*:** *Educational Indicators; EDM*; *Regression*; *Correlation*.

1. Introdução

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. Segundo o INEP, entre 2014 e 2015, a repetência na 1ª série do ensino médio chegou a 15,3%. O índice também é alto no 6º ano do ensino fundamental, com taxas de 14,4%. Sobre a evasão, em 2007, 14,5% dos matriculados no ensino médio abandonavam

os estudos antes de se formarem, em 2015 diminuiu para 11,2%. No ensino fundamental, 7,5% dos alunos deixavam as escolas antes da formatura nos anos finais, passou a 5,4% em 2015. Já nos anos iniciais, a evasão reduziu de 3,5% para 2,1% (INEP 2017).

Apesar dos avanços em alguns aspectos, combater problemas como a evasão e a reprovação escolar ainda é um dos grandes desafios para a área de educação, tornando-se um tema bastante relevante e em expansão. Diversos trabalhos vêm sendo desenvolvidos a partir de Mineração de Dados Educacionais (MDE) para fins de tomada de decisão (Rodrigues et al. 2013). A aplicação da MDE tem como objetivo a descoberta de informações que ajudem na proposta educacional, no melhoramento das condições de infraestrutura escolar, no processo ensino-aprendizagem, na previsão de desempenho dos alunos, além de outros fatores que influenciam a aprendizagem (Baker et al. 2011), dentre os quais pode-se destacar a reprovação e a evasão escolar.

No contexto da evasão escolar, o objetivo do trabalho de Calixto et al. (2017) consistiu na identificação de variáveis relacionadas a este indicador educacional, utilizando os dados do censo escolar no âmbito do Ceará e Sergipe. As análises se deram por meio de técnicas de indução de regras e regressão logística. A idade, a etapa e a modalidade de ensino, a existência de laboratórios e localização da escola se destacaram como variáveis influentes na evasão escolar. Ainda, Machado et al. (2015) faz um estudo bibliométrico com o objetivo de identificar os trabalhos que abordam o problema da evasão escolar utilizando técnicas de mineração de dados. O estudo permitiu identificar que os principais métodos utilizados são árvores de decisão, redes neurais, regressão logística e algoritmos de agrupamento.

Já no âmbito do desempenho escolar, o trabalho de Laisa e Nunes (2015) analisou a aprovação e a reprovação utilizando base de dados de alunos do ensino médio. Utilizou técnica de classificação usando o Algoritmo J48. Em Detoni et al. (2015), é investigado a reprovação no cenário acadêmico do ensino à distância (EaD). Utiliza-se como atributo as contagens de interações com ambiente virtual e demonstrou que as redes bayesianas se mostraram o modelo mais adequado de predição. Ainda no EaD, Da Silva e Imram (2015) investigam variáveis relacionadas a conclusão de alunos do ensino superior.

Tendo em vista a abordagem destes trabalhos apresentados, a realização desta pesquisa traz contribuição uma vez que aspectos educacionais carecem de tópicos de análises sob diferentes abordagens. Em relação as técnicas aplicadas a problemas MDE, são pouco aplicadas técnicas de predição diferentes daquelas apresentadas por Machado et a. (2015). Desta forma, utilizando modelos lineares de regressão, as análises de previsão da evasão e reprovação escolar são construídas. Em Rodrigues et al. (2013), os resultados obtidos no trabalho demonstraram que é possível utilizar a técnica de regressão linear para obter inferências com boas taxas de precisão.

Seguindo as fases do *Cross-Industry Standard for Data Mining* - CRISP-DM (Chapman et al. 2000), realiza-se uma análise dos dados das bases e investiga-se quais variáveis educacionais, fornecidas abertamente pelo INEP, possuem relação com as taxas de evasão e reprovação escolar, através da análise correlacional. O estudo se concentrou em dados do ensino fundamental do estado de Pernambuco. Após isso, os modelos de regressão linear e de regressão robusta são aplicados e comparados.

A regressão linear é uma boa técnica para ser aplicada quando os erros de predição são normais. Quando os erros não são normais, outros métodos podem ser considerados. Uma abordagem é a remoção de *outliers*, mas isso pode não ser eficaz quando existem vários pontos discrepantes na base de dados. Nesse ponto, a regressão robusta fornece uma alternativa para esse problema (Ortiz 2006). Portanto, este trabalho realiza uma análise de qual das duas técnicas minimizam o erro de previsão da evasão e da reprovação escolar com base nos dados utilizados para esse estudo.

2. Metodologia

Uma das metodologias mais populares para aumentar o sucesso dos processos de mineração de dados é o CRISP-DM (Chapman et al. 2000). A metodologia define uma sequência não rígida de seis fases, que permite a construção e implementação de um modelo de mineração para ser usado em um ambiente real, ajudando as decisões de negócios (Moro et al. 2011). Assim, o desenvolvimento desse trabalho segue as fases do CRISP-DM, mostradas na Figura 1 e descritas nas subseções seguintes.

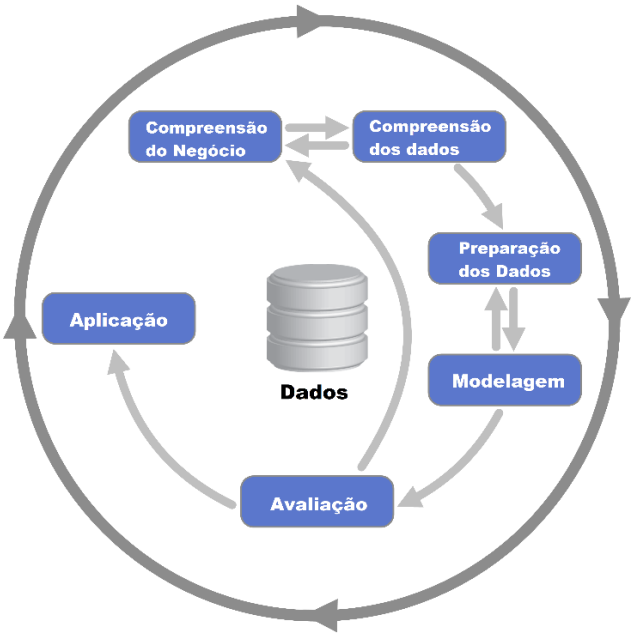

Figura 1 – Fases do CRISP-DM.
Fonte - Adaptado de Chapman et al. (2000).

2.1.Compreensão do negócio

A fase inicial do CRISP-DM concentra-se no entendimento dos objetivos e requisitos do projeto e converte esse conhecimento em uma definição de problema de mineração de dados e um plano preliminar projetado para atingir os objetivos. Para esse trabalho, foi verificado todo material já elaborado na literatura relacionado a cenários educacionais, mineração de dados educacionais e modelos de regressão.

2.2.Compreensão dos dados

Esta fase do CRISP-DM começa com a coleta de dados, e prossegue com atividades que permitem familiarizar-se, identificar problemas de qualidade, descobrir *insights* sobre os dados e detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas.

Descrição das bases de dados. As bases de dados utilizadas nesse estudo são disponibilizadas abertamente pelo INEP em seu portal, desde 1995 (INEP 2016). São utilizados os dados dos indicadores educacionais referentes ao ano de 2016.

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. Eles consideram informações como o acesso, a permanência e a aprendizagem dos alunos. Pode-se obter esses dados em diferentes granularidades como nível nacional, regional ou nível das escolas. Para este estudo é

utilizada a base de dados de nível das escolas. Os indicadores educacionais considerados para esse estudo são:

- Taxa de Rendimentos: Refere-se a dois indicadores desse estudo, sendo a taxa de evasão escolar e a taxa de reprovação.
- Adequação da formação dos docentes: 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 leciona. Os grupos são descritos na Tabela 1.

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

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

Fonte - INEP

- Alunos por turma: Média de alunos por turma da educação básica.
- Complexidade da gestão: Nível de complexidade de gestão da escola. O indicador classifica as escolas em níveis de 1 a 6 de acordo com sua complexidade de gestão, níveis elevados indicam maior complexidade. Com base nos dados disponíveis do Censo da Educação Básica, considerou-se que complexidade de gestão está relacionada às seguintes características: porte da escola, número de turnos de funcionamento, quantidade e complexidade de modalidades/etapas oferecidas.
- Distorção entre idade e série do aluno: Taxa de distorção idade-série da escola.
- Docente com curso superior: Percentual de docentes com curso superior na escola.
- Esforço do docente: 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 dos docentes do ensino fundamental e ensino médio.

Nível	Descrição
1	Docente que, em geral, tem até 25 alunos e atua em um único turno, escola e etapa.
2	Docente que, em geral, tem entre 25 e 150 alunos e atua em um único turno, escola e etapa.
3	Docente que, em geral, tem entre 25 e 300 alunos e atua em um ou dois turnos em uma única escola e etapa.
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.
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.
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.

Fonte - INEP

- Média de horas-aula diária: Número médio de horas-aula diária na escola.
- Regularidade do docente: Média do indicador de regularidade do docente. Para cada docente em cada escola foi atribuída uma pontuação de forma que fosse valorizado: o total de anos em que o docente atuou na escola nos últimos 5 anos, a atuação do docente na escola em anos mais recentes e a atuação em anos

consecutivos. O Indicador de Regularidade do Docente (IRD) varia de 0 a 5, quanto mais próximo de 0, mais irregular é o vínculo do docente com a escola e quanto mais próximo de 5, mais regular é esse vínculo.

Com estas bases de dados descritas, foram derivadas duas bases novas. Uma base, possuindo o indicador de evasão escolar, e a segunda o indicador de reprovação escolar. Nessas bases, o indicador de evasão e de reprovação escolar são as variáveis resposta (y) e as demais são as variáveis explicativas (x). As 18 variáveis geradas para as bases são mostradas na Tabela 3.

Tabela 3 – Variáveis presentes nas bases.

Variáveis		
1	TE / TR	Taxa de Evasão / Taxa de Reprovação
2	IRD	Índice de Regularidade Docente
3	TDI	Taxa de dispersão Idade-Série
4	ICG	Índice de Complexidade da Gestão
5	HAU	Média de Horas-Aula
6	ATU	Alunos por Turma
7	DSU	Docente com Curso Superior
8 a 12	AFD	Percentual da Formação docente por grupo (5 grupos)
13 a 18	IED	Nível de Esforço Docente (6 níveis)

Fonte - Autores.

2.3. Preparação dos dados

Nesta etapa, as tarefas incluem seleção de atributos, além de limpeza de dados e transformação para as ferramentas de modelagem. Nesse trabalho, após a formação da base, foi feita uma análise das variáveis com a finalidade de realizar as prováveis transformações. Inicialmente, foi realizada uma seleção dos dados para o estado de Pernambuco, pois consiste no âmbito estudado. Excluiu as instâncias que possuíam valor 0 (zero) para a evasão e reprovação pois não representam a existência de valores para as variáveis a serem estudadas.

Após isso, foi possível observar poucos valores ausentes para algumas variáveis, e para resolver esse problema, foi realizada a inserção de valores utilizando a mediana dos valores das colunas. Isto foi realizado uma vez que o número de *missing values* não foi grande e o objetivo foi ter o mínimo de perda de instâncias. No entanto, a variável HAU, a qual se refere a média de horas-aula na escola, continha um grande número de valores faltantes. Segundo a análise da representatividade dessa variável em relação a taxa de evasão e reprovação (variáveis TE e TR respectivamente), optou-se pela exclusão de HAU.

Nessa fase, também foi aplicada a normalização dos dados, a qual consiste em ajustar a escala dos valores dos atributos para que os valores fiquem em pequenos intervalos, tais como entre 0 a 1. Tal ajuste se faz necessário para evitar que alguns atributos, por apresentarem uma escala de valores maior que outros, influenciem de forma tendenciosa na mineração de dados. A fórmula utilizada é mostrada na Equação 1.

$$\left(z_i^k\right) = \frac{z_i^k - z_{min}^k}{z_{max}^k - z_{min}^k} \tag{1}$$

onde, z_i^k se refere a cada valor na base, z_{max}^k e z_{min}^k consistem nos valores máximo e mínimo, respectivamente. Após realizar as tarefas presentes nessa fase, formam-se as bases finais para o ensino fundamental com um total de 17 variáveis, conforme mostra a Tabela 4.

Tabela 4 – Bases de dados após preparação dos dados.

Base de Dados	Cenário	Nº de instâncias
I	Evasão do Ensino Fundamental	3.683
II	Reprovação do Ensino Fundamental	6.591

Fonte - Autores.

2.4. Modelagem

Apesar das fases já descritas serem de suma importância para o resultado final do CRISP-DM, a modelagem é considerada a principal etapa do processo. Representa o desenvolvimento dos modelos para o problema, com base nos dados que já estão adequados para serem utilizados. Nesse trabalho, será implementado modelos de regressão de forma comparativa. Para gerar o modelo, é realizada a análise correlacional entre as variáveis.

Análise Correlacional: É necessário avaliar o grau de relacionamento entre as variáveis das bases de dados, ou seja, descobrir o quanto uma variável (x) interfere no resultado de outra (y), quando o relacionamento possui forma linear. Esse grau de associação pode ser medido pelo coeficiente de correlação. A medida de correlação mais 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.

Os índices de correlação apresentados na Tabela 5 mostram o grau de relacionamento entre variáveis da base I, em relação a variável resposta taxa de evasão (TE). Nota-se que as variáveis TDI e IED1 (índices 0.31 e 0.25, respectivamente) possuem uma maior correlação positiva. Ou seja, quanto maior o valor de evasão escolar, maior a dispersão idade-série dos alunos e maior o nível de esforço 1 dos docentes para exercer a profissão (esforço mais baixo). Essas variáveis variam no mesmo sentido.

Já as variáveis ATU, ICG e AFD1 (índices -0.23, -0.20 e -0.17, respectivamente) possuem maior correlação negativa para a variável resposta. Isto significa que o relacionamento entre taxa da evasão escolar com as variáveis explicativas como complexidade da gestão da escola, alunos por turmas e percentual de docentes formados em nível são variados em sentidos opostos.

Tabela 5 – Análise correlacional das variáveis da base de dados I.

Variável resposta	Valor Correlacional com as Variáveis Explicativas							
TE	IRD	**TDI**	**ICG**	DSU	**ATU**	**IED1**	IED2	IED3
	-0.06	**0.31**	**-0.19**	-0.13	**-0.23**	**0.25**	-0.09	-0.15
	IED4	IED5	IED6	**AFD1**	AFD2	AFD3	AFD4	AFD5
	-0.11	-0.05	-0.06	**-0.17**	-0.03	0.05	-0.05	0.12

Fonte - Autores.

Os índices de correlação apresentados na Tabela 6 mostram o grau de relacionamento entre variáveis da base II, em relação a variável resposta taxa de reprovação (TR). Em análise, as variáveis TDI e IED1 (índices 0.47 e 0.17, respectivamente) possuem uma maior correlação positiva. Ou seja, quanto maior o valor de evasão escolar, maior a dispersão idade-série dos alunos e maior o nível de esforço 1 dos docentes para exercer a profissão (esforço mais baixo). Isso mostra que as variáveis variam no mesmo sentido. Já a variável IED3 (índices -0.16) possui maior correlação negativa com a variável resposta. Isto significa que quanto maior a evasão escolar, menor o nível de docentes que possuem esforço 3 para exercer a profissão.

Tabela 6 – Análise correlacional das variáveis da base de dados II.

Variável resposta	**Valor Correlacional com as Variáveis Explicativas**							
TR	IRD	**TDI**	ICG	DSU	ATU	**IED1**	IED2	**IED3**
	-0.09	**0.47**	-0.08	-0.02	-0.04	**0.17**	-0.01	**-0.16**
	IED4	IED5	IED6	AFD1	AFD2	AFD3	AFD4	AFD5
	-0.04	-0.03	-0.03	-0.05	-0.06	0.06	-0.01	0.01

Fonte - Autores.

Modelos de Regressão: A regressão é uma técnica que permite inferir a relação de uma variável de resposta (y) com variáveis explicativas (x). Os modelos utilizados nesse trabalho possuem as seguintes definições: seja $x = (x_0, ..., x_p)$ um vetor de variáveis explicativas, seja $\beta = (\beta_{0, ..., }\beta_{p)}$ um vetor de parâmetros e seja $\varepsilon = (\varepsilon_{0, ..., }\varepsilon_{p)}$ um vetor de erro aleatório, a equação do modelo linear para conjunto de dados $i = (1, ..., n)$ é dada por

$$y_i = \beta x_i + \varepsilon_i \quad (2)$$

Segundo a equação 2, as regressões utilizadas nesse trabalho possuem as seguintes particularidades:

- Regressão Linear: O vetor β para a regressão linear, é estimado pelo método dos mínimos quadrados minimizando uma função baseada na soma dos resíduos quadrados (ε_i) que é dada por

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

  onde, y_i é a variável resposta real e $\hat{y}_i$ a variável resposta predita pelo modelo.
- Regressão Robusta: O vetor β é estimado, minimizando uma função critério. A função critério é dada por

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \beta x_i}{\sigma}\right) \quad (4)$$

  onde, σ é um estimador robusto e ρ uma função particular.

O modelo final das regressões foi obtido passando como o vetor de variáveis explicativas x com maior valor correlacional, conforme análises realizadas. Para a regressão linear e para a regressão robusta foram gerados 2 modelos diferentes: análise da evasão e análise da reprovação escolar.

A configuração experimental se deu por execuções em simulações Monte Carlo com 30 iterações. A estimação dos valores é realizada através 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.

2.5. Avaliação

Nesta etapa da metodologia, os modelos desenvolvidos são avaliados. Um dos índices de desempenho mais utilizados nas técnicas de previsão é o cálculo baseado no erro de previsão. O índice de desempenho utilizado nesse trabalho é o erro médio absoluto (MAE), como denotado em

$$\frac{1}{n} \sum_{j=1}^{n} |y_j - \hat{y}_j|$$

(5)

Onde, n é a quantidade de valores, y_j é a variável resposta real e $\hat{y}_j$ a variável resposta predita pelo modelo. O MAE é uma medida amplamente utilizada em avaliações de modelos (Chai e Drexler 2014). Com a amostra das execuções, pode-se calcular o desvio padrão do erro, realizar testes estatísticos, gráficos *boxplots* para também avaliar o desempenho dos modelos de regressão.

2.6. Aplicação

Todo o conhecimento obtido através do trabalho de mineração tornou-se subsídio para o desenvolvimento de estratégias que resolvam o problema proposto. Nesse trabalho, serão listadas estratégias e considerações para os cenários estudados após todas as etapas do CRISP-DM.

3. Resultados e Discussões

Os resultados dos experimentos são descritos nas próximas subseções. As análises se concentram em tabelas e gráficos *boxplots* sobre o desempenho dos modelos de regressão utilizados.

3.1. Evasão Escolar

Para o cenário da evasão, no âmbito do ensino fundamental das escolas de Pernambuco, a Tabela 8 fornece os resultados da média do erro calculado e o desvio padrão associado. 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.

Outra forma de analisar o resultado para esse cenário é por meio da Figura 2. Ela mostra o *boxplot*, que representa a variação de dados contidos numa amostra, no caso, a variação do valor do erro associado a previsão da evasão escolar. Como pode ser visto, a regressão robusta 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). Dessa forma, pode-se concluir que a regressão robusta é o melhor modelo para previsão da evasão escolar.

Tabela 8 – Média dos valores de erro e desvio padrão para estimação da variável TE.

Modelo	MAE	Desvio Padrão
Regressão Linear	0.0331	0.00078
Regressão Robusta	**0.0306**	0.00087

Fonte - Autores.

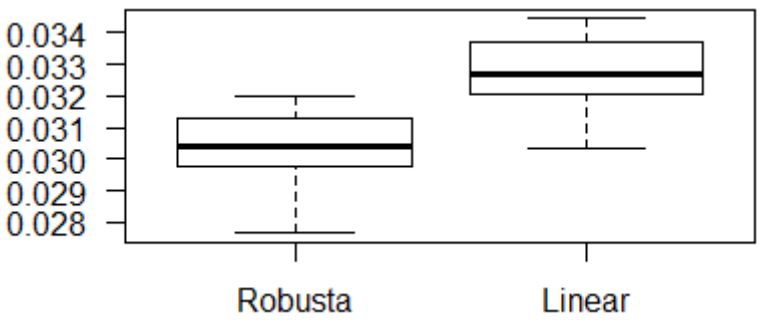

Figura 2 – Gráfico *boxplot* entre os modelos para estimação da variável TE
Fonte - Autores.

3.2. Reprovação Escolar

Para o cenário da reprovação, no âmbito do ensino fundamental das escolas de Pernambuco, a Tabela 9 fornece os resultados da média do erro calculado e o desvio padrão associado. Como pode ser analisado, a 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 em uma diferença significativa.

Outra forma de analisar o resultado para esse cenário é por meio da Figura 3. Ela mostra o *boxplot*. Como pode ser visto, a regressão robusta obteve uma menor mediana para o valor do erro. Também possui menor valores em comparação a regressão linear. Dessa forma, pode-se concluir que a regressão robusta é o melhor modelo para previsão da reprovação escolar.

Tabela 9 – Média dos valores de erro e desvio padrão para estimação da variável TR.

Modelo	MAE	Desvio Padrão
Regressão Linear	0.04944938	0.0008648008
Regressão Robusta	0.04819603	0.0009360381

Fonte - Autores.

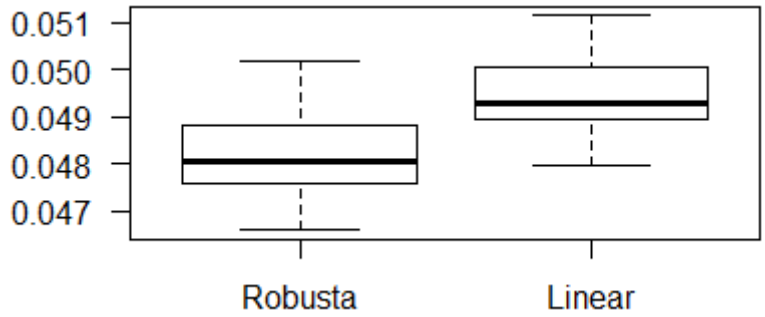

Figura 3 – Gráfico *boxplot* entre os modelos para estimação da variável TR.

Fonte - Autores.

A análise de dois modelos de regressão tratou de encontrar o método que fornecesse um melhor ajuste aos dados, de forma a minimizar o erro e estimar com maior precisão o relacionamento entre as variáveis em estudo. Nesse sentido, o modelo de regressão robusta alcançou este objetivo, pois minimizou o erro de predição para os indicadores educacionais dos cenários definidos. O motivo pela qual a regressão robusta obteve melhores resultados pode ser explicado pelo fato desse método tratar melhor os dados na presença de *outliers*, e essa característica é presente nas bases dessa pesquisa.

4. Conclusões

Esse estudo buscou investigar e explicar variáveis educacionais que podem estar relacionadas com a evasão e reprovação escolar. Os modelos aplicados nos experimentos contam com os resultados de técnicas lineares distintas (regressão linear e robusta). Através de bases de diferentes indicadores educacionais fornecidas pelo INEP, pode-se gerar novas bases de dados para realizar os experimentos.

Todas as fases aplicadas na metodologia CRISP-DM foram importantes para que se obtivesse resultados de forma mais assertiva. O entendimento do relacionamento entre as variáveis da base, através da análise correlacional, tornou possível a implementação de um modelo com variáveis mais correlacionadas com os indicadores de evasão e reprovação. A dispersão idade-série dos alunos, o esforço dos docentes para exercer a profissão, a quantidade de alunos por turma e a formação dos docentes, mostraram-se variáveis mais associadas com estes indicadores.

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, comprovada por tabelas e gráficos *boxplots*.

Portanto, a principal contribuição desse trabalho consiste em fornecer uma aplicação de análise correlacional e de regressores que minimiza o erro de predição e estima com maior precisão o relacionamento entre variáveis e indicadores educacionais, como a evasão e a reprovação escolar. Assim, utilizar a mineração de dados educacionais possibilita a identificação prévia de aspectos que podem precisar de melhorias e investimentos mais adequados, aprimorando aspectos do ensino e mitigando problemas. Além disso, melhor estimar as variáveis relacionadas a esse cenário traz um grande ganho a literatura e aos diversos interessados como estudantes, educadores, pesquisadores, governo, entre outros. São ferramentas que podem ser aplicadas de forma extensiva, gerando conhecimento, servindo como base para soluções de problemas e desenvolvimento de mecanismos em apoio ao ensino

Dessa forma, como trabalhos futuros, pretende-se refinar as técnicas de previsão, assim como explorar outros aspectos educacionais, expandir o escopo de estudo para outros níveis de escolaridade e para outros cenários como o regional ou nacional.

Referências

BAKER, R. S. J., ISOTANI, S., CARVALHO, A. de: Mineração de dados educacionais: Oportunidades para o Brasil. **Revista Brasileira de Informática na Educação**, 2011, v. 12, n. 2, p. 3 – 13.

CALIXTO, K., SEGUNDO, C., and DE GUSMÃO, R. P.: Mineração de dados aplicada a educação: um estudo comparativo acerca das características que influenciam a evasão escolar. **In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**, 2017, volume 28, page 1447.

CHAI, T., DRAXLER, R. R.: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. **Geoscientific Model Development**, 7, 1247-1250, 2014.

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

DA SILVA, J.M.C., IMRAN, H.: Um estudo sobre as variáveis para predição de alunos não concluintes em cursos suportados por Ambientes Virtuais de Ensino e Aprendizagem. **Revista Novas Tecnologias na Educação**, v. 13, n. 2, 2015.

DETONI, D., CECHINEL, C., ARAÚJO, R. M.:Modelagem e Predição de Reprovação de Acadêmicos de Cursos de Educação a Distância a partir da Contagem de Interações. **Revista Brasileira de Informática na Educação**, 2015, v. 23, n. 3.

INEP, Portal: Inep divulga dados inéditos sobre fluxo escolar na educação básica. 20 jun. 2017. **Notícias**. Disponível em: <http://portalinep.gov.br/>. Acesso em 14/02/2018.

LAISA, J., and NUNES, I.: Mineração de Dados Educacionais como apoio para a classificação de alunos do Ensino Médio. **In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)**, 2015, volume 26, page 1112.

MACHADO, R. D., NARA, E. O. B., SCHREIBER, J. N. C., and SCHWINGEL, G. A.: Estudo bibliométrico em mineração de dados e evasão escolar. **XI Congresso Nacional de Excelência em Gestão**, 2015.

MORO, S., LAUREANO, R., CORTEZ, P.: Using data mining for bank direct marketing: An Application of the crisp-dm methodology. EUROSIS. **Proceedings of European Simulation and Modelling Conference-ESM**, 2011, page 117–121.

ORTIZ, M. C., SARABIA, L. A. and, HERRERO, A.: Robust regression techniques: A useful alternative for the detection of outlier data in chemical analysis.**Talanta**, 2006, v. 70, n.3, pages 499-512.

RODRIGUES, R. L., DE MEDEIROS, F. P., and GOMES, A. S.: 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)**, 2013, volume 24, page 607.