Modelos Regressão Aplicados a Predição do Desempenho Escolar de Estudantes do Ensino Fundamental

Paulo M. Silva[1], Rafaella L. S. Nascimento[2], Marilia N. C. A Lima[3], Roberta A. A. Fagundes[4], Fernando F. de Souza[5]

[1,2,5] Centro de Informática – Universidade Federal de Pernambuco (UFPE)

Avenida Jornalista Aníbal Fernandes – Cidade Universitária – Recife – PE - Brazil

[3,4]Departamento de Engenharia da Computação – Universidade de Pernambuco (UPE)

Rua Benfica, Madalena – Recife – PE – Brazil

{pms3, rafaellalsn, fdfd }@cin.ufpe.br, roberta.fagundes@upe.br, mncal@ecomp.poli.br

Abstract. *This paper investigates the applicability of regression models for predicting student performance in public schools in the state of Pernambuco. The study used information from 2015 and 2017 from the National Basic Education Assessment System (SAEB). Knowledge extracted from the data through automatic selection using the Stepwise method, and it is possible to identify the associated factors that most influence school performance. Parametric and nonparametric Regression models were applied to predict this performance. The results showed that factors such as the number of people living in the residence, the parents' encouragement of schoolwork and the area where the student lives. These factors had a strong influence on school performance.*

Resumo. *O presente artigo investiga a aplicabilidade de modelos de Regressão para a previsão desempenho de alunos pertencentes as escolas públicas do estado de Pernambuco. O estudo utilizou informações dos anos de 2015 e 2017 do Sistema Nacional de Avaliação da Educação Básica (SAEB). O conhecimento extraído dos dados através da seleção automática usando o método Stepwise, sendo possível identificar os fatores associados que mais influenciam o desempenho escolar. Foram aplicados modelos de Regressão paramétricos e não paramétricos, para a previsão desse desempenho. Os resultados mostraram que os fatores, como: a quantidade de pessoas que moram na residência, o incentivo dos pais as tarefas escolares e a área onde o estudante reside. Esse fatores obtiveram forte influência sobre o desempenho escolar.*

1. Introdução

O direito à Educação de qualidade ainda está longe de ser assegurado e configura um desafio nos dias atuais. De acordo com o Anuário da Educação Brasileira, em 2018, menos da metade dos alunos das escolas públicas brasileiras atingiram níveis de proficiência considerados adequados ao fim do 3º ano do Ensino Fundamental em Língua Portuguesa (LP) e Matemática (MT). Principalmente em LP, os níveis de proficiência estão distantes do razoável: 33,8% dos alunos encontram-se em níveis insuficientes [ANUÁRIO 2019]. Ainda de acordo com a publicação, a cada 100 estudantes que ingressam no ensino fundamental, 76 concluem aos 16 anos, sendo que 39,5% e 21,5%

DOI: 10.5753/cbie.sbie.2019.1621

desses estudantes tem aprendizagem adequada em LP e MT, respectivamente, ao final dessa etapa de ensino.

Os fatores associados a aprendizagem que afetam o desempenho escolar, são provenientes de três grupos: a família, o aluno e a escola. O primeiro influencia com sua própria estrutura e seu envolvimento no processo de aprendizagem. O segundo, com suas características pessoais e atitudes em relação a escola e o terceiro com equipes de profissionais competentes, metodologia de ensino, recursos físicos e pedagógicos [SOARES 2004].

A compreensão dos desafios da Educação em busca de um ensino de qualidade perpassam por estudos que sejam baseados nas evidências, a partir de estatísticas e análise dos dados oriundos de bases de dados educacionais. No Brasil, o Instituto Nacional de Pesquisas Educacionais Anisio Teixeira (INEP), detém a mais completa base de dados sobre educação do país. Os dados educacionais são de origem censitária ou de avaliações em larga escala, como por exemplo: SAEB [INEP 2018].

Os dados contidos nas bases do SAEB, caracterizam-se pelas proficiências nas avaliações de LP e MT, e das respostas dos questionários socioeconômicos respondidos pelos estudantes. As evidências da realidade da educação básica brasileira contidas nesses dados, nos permite extrair informações relevantes que serão importantes para a identificação dos fatores associados a aprendizagem, que afetam desempenho escolar dos estudantes.

Na busca de soluções computacionais que fossem capazes de facilitar e potencializar o processo de ensino aprendizagem, utilizou-se a Mineração de Dados Educacionais (MDE), como uma alternativa para investigar e encontrar padrões entre o desempenho escolar e os fatores associados a aprendizagem. Nesse contexto, [Baker, Isotani e Carvalho 2011], conceituam a MDE, como a área de pesquisa que tem como principal foco o desenvolvimento de métodos para explorar um conjunto de dados coletados em cenários educacionais.

Em [Calderón et al. 2015] os autores utilizaram a MDE aplicando o algoritmo de Classificação *Artificial Neural Network*, para detectar os fatores com maior influência na nota do semestre de alunos do ensino superior. Com acurácia de 84,86%, detectou-se que os fatores mais relevantes são: idade, gênero, nota do exame de admissão nas respostas objetivas e discursivas, práticas para a aprendizagem, aspectos mais significativos a vida, calma em relação a situações difíceis e grau de dificuldade em lidar com fatos desagradáveis.

Em [Rabelo *et al*. 2017] os autores propuseram um modelo baseado em classificação para prever o desempenho dos alunos de um curso de graduação por meio da participação e interação de um ambiente virtual de aprendizagem (AVA). Como resultado, o modelo obteve uma acurácia de 96,5%, utilizando o classificador de árvore de decisão.

No trabalho de [Rodrigues *et al*. 2013] os autores investigam a viabilidade da utilização do modelo de regressão linear para obtenção de inferências nas etapas iniciais de cursos online. A regressão linear buscou estimar o desempenho de alunos baseados em suas interações dentro da plataforma virtual de aprendizagem. Os resultados obtidos mostraram que é possível utilizar a técnica de regressão linear para obter inferências com taxas de confiança de 95%.

Este trabalho propõe modelos de regressão para previsão do desempenho escolar. Para isso, investiga-se os fatores associados a aprendizagem de LP e MT dos alunos do ensino fundamental das escolas públicas do estado de Pernambuco. O objetivo é trazer informações que contribuam para o desenvolvimento de políticas públicas visando a melhoria da aprendizagem e consequentemente do desempenho nessa fase escolar.

Nesta perspectiva, este trabalho se diferencia dos demais por propor a aplicação de modelos de Regressão utilizando métodos paramétricos e não-paramétricos para identificação dos fatores associados a aprendizagem que afetam o desempenho dos estudantes em LP e MT dos anos de 2015 e 2017 do SAEB. De acordo com os seguintes aspectos: i) realizar análise do problema, investigando os fatores mais preponderantes que afetam o desempenho escolar dos estudantes do 5º ano ensino fundamental das escolas públicas do Estado de Pernambuco; ii) análise das bases de dados do SAEB, para determinar quais fatores exercem influência significativa no desempenho dos estudantes em LP e MT; iii) a utilização de modelos de Regressão proporciona a previsão de condições futuras, dando suporte quantitativo a decisão, apontando falhas, e fornecendo *insights* que podem ajudar os gestores educacionais e professores a tomar decisões estratégicas em relação a aprendizagem de seus alunos; iv) contribuir com o fortalecimento do campo da MDE, na resolução de problemas envolvendo a previsão do desempenho no contexto educacional.

O trabalho está organizado da seguinte forma: Na seção 2 apresentamos o processo de MDE aplicado ao desempenho escolar. Na seção 3 são descritos os experimentos realizados. A seção 4 apresenta os resultados e suas respectivas análises. Conclusões e possíveis trabalhos futuros são apresentados na seção 5.

2. Processo MDE aplicado ao Desempenho Escolar

Nesta fase será descrita a metodologia utilizada neste trabalho. O CRISP-DM (*Cross Industry Standard Process for Data Mining*) [Chapman et al. 2000]. É descrito hierarquicamente e consiste em seis fases. No entanto, nesse trabalho as fases do processo foram alteradas para o contexto da aplicação. Dessa forma será possível a construção e implementação de modelos de MDE para serem aplicados a um contexto educacional.

2.1. Entendimento do Contexto Educacional

A fase inicial concentra-se na compreensão dos objetivos e requisitos do problema abordado. O conhecimento adquirido será convertido em uma definição de problema de MDE. Realizamos análise da literatura especializada e coleta de informações sobre desempenho escolar e fatores associados a aprendizagem, em particular nos dados referentes a educação fundamental brasileira, fornecidos pelo INEP/SAEB, caracterizando como o problema a ser abordado.

2.2. Entendimento dos Dados Educacionais

O SAEB é composto por um conjunto de avaliações externas em larga escala, que permite realizar um diagnóstico da educação básica brasileira dos fatores que possam interferir no desempenho do estudante, fornecendo um indicativo sobre a qualidade do ensino ofertado [INEP 2019]. O INEP disponibiliza publicamente desde 1995, os dados do SAEB. Este estudo utilizou a base de dados de 2015 e 2017, para a extração de conhecimento e caracterização dos fatores relacionados a aprendizagem. Foi levado em consideração os argumentos propostos na literatura como fatores associados a aprendizagem como por exemplo, atributos demográficos, socioeconômicos e acadêmicos.

Com o objetivo de identificar, os fatores associados a aprendizagem, que afetam de forma expressiva o desempenho dos alunos pernambucanos. As bases de dados do SAEB, foram divididas por ano/disciplina, os estudantes que realizaram a avaliação nos anos de 2015 e 2017, foram classificados pelos níveis de proficiência em LP e MT. Sendo considerados nesse estudo, os alunos cujas proficiências em LP e MT foram inferiores à média regional e nacional definidas pelo INEP, para o 5º ano do ensino fundamental nos respectivos anos e disciplinas. Assim foram identificadas as diversas variáveis que compõem os grupos de fatores associados a aprendizagem conforme descrito no trabalho de [Soares 2004].

2.3. Preparação dos Dados para MDE

Nessa fase o escopo da pesquisa foi definido pela filtragem para remoção dos registros que não estavam dentro do grupo de interesse, como por exemplo as escolas particulares. Nessa etapa foram mantidos os alunos do 5º ano do ensino fundamental das escolas públicas estaduais e municipais do estado de Pernambuco.

Foram excluídos da base os alunos que não participaram da avaliação. A caracterização do desempenho (variável dependente) foi definida pelas proficiências dos alunos nas disciplinas de LP e MT. Já os fatores associados a aprendizagem (variáveis independentes), foram definidas pelas respostas do questionário socioeconômico preenchido pelo estudante que realizou a avaliação. As variáveis redundantes ou irrelevantes foram excluídas e analisou-se a base de dados em busca de registros com valores não preenchidos (*missing values*). Para as variáveis nessa situação, os registros foram preenchidos utilizando a mediana entre os atributos. Os dados do questionário socioeconômico para efeitos da aplicação das técnicas de regressão, foram transformados em dados numéricos e em alguns casos dicotômizados. Também foi utilizado o método Stepwise para a seleção automática das variáveis independentes. Assim, a Tabela 1 apresenta a quantidade de instâncias antes e depois do pré-processamento realizado.

Tabela 1 - Dimensões do Conjunto de Dados

Base SAEB	Antes do pré-processamento		Depois do pré-processamento	
	Nº variáveis	Nº instâncias	Nº de variáveis	Nº de instâncias
2015	86	105.456	52	85.036
2017	86	113.667	52	94.554

O método Stepwise foi executado no software Rstudio [R 2019]. O modelo de regressão, inicialmente, incorpora todas as variáveis independentes e depois, por etapas, retira ou não cada variável desse modelo. A Tabela 2 ilustra as variáveis independentes com maior influência, seja ela: positiva ou negativa, na aprendizagem de LP e MT, além do respectivo coeficiente Correlação de Pearson. Além disso, utiliza-se como referência, para a correlação com o desempenho escolar, o trabalho de Soares (2014), o qual propõe a classificação dos fatores associados nos grupos: aluno, família e escola.

Tabela 2 – Variáveis Associadas a Aprendizagem

Variável	Coef.
Qt. de pessoas que moram na residência	-0,97
Área de residência do estudante	0,95
Pais incentivam a fazer o dever de MT	0,93
Qt. de quartos na residência	0,90
Qt. de trabalhadores domestico (a)	-0,90
Pais incentivam a Ler	-0,86
O aluno faz o dever de MT	0,80
Qt. de geladeiras na residência	0,81
Os pais vão as reuniões na escola	0,82
Qt. de banheiros na residência	0,76
Nivel de escolaridade da mãe	0,77
Mãe alfabetizada	0,78

Dos fatores associados a aprendizagem identificados, os que exercem influência fortemente são: a quantidade de pessoas que moram na residência, a área de residência do estudante (urbana ou rural) e o incentivo dos pais a realização das tarefas de matemática. Com relação a classificação os fatores identificados pertencem ao grupo de fatores relacionados ao aluno e a família [Soares 2004].

Como fator relacionado as atitudes do aluno em relação a escola, a realização das tarefas de MT mostrou-se um item importante. Ainda relacionado ao aluno os itens funcionais e estruturais da sua residência tais como, quantidade de trabalhadores domésticos, quantidade de quartos, quantidade de geladeiras, quantidade de banheiros, também possui participação significativa no seu desempenho. Por fim, como fatores relacionados a família temos: o incentivo dos pais a leitura, os pais vão as reuniões na escola, o nível de escolaridade da mãe e se a mãe sabe ler e escrever, apresentam significante influencia no desempenho escolar.

2.4. Modelagem

Nesta fase, define-se as técnicas de modelagem de dados, especificamente um conjunto de algoritmos de previsão, de acordo com as variáveis selecionadas pelo método Stepwise presentes na Tabela 2. Os métodos paramétricos de Regressão Linear (RL), Regressão Linear Robusta (RLR), Regressão Quantílica (RQ) e não paramétricos, Regressão Vetorial de Suporte (SVR), foram utilizadas para verificar o desempenho dos alunos em relação aos fatores associados a aprendizagem.

As técnicas de regressão utilizadas neste trabalho possuem as seguintes definições: seja $x = (x_0, ..., x_n)$ um vetor de variáveis explicativas, seja $\beta = (\beta_0, ..., \beta_n)$ um vetor de parâmetros e seja $\varepsilon = (\varepsilon_0, ..., \varepsilon_n)$ um vetor de erro aleatório, a equação do modelo linear para conjunto dos dados $i = (1, ..., n)$ é dada por

$$yi = \beta xi + \varepsilon i \tag{1}$$

Na RL o vetor β é estimado pelo método dos mínimos quadrados minimizando uma função baseada na soma dos resíduos quadrados (ε_i) que é dada por

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

onde, y_i é a variável resposta real e $\hat{y}_I$ a variável resposta predita pelo modelo. Na RLR vetor β é estimado, minimizando uma função critério. A função critério é dada por

$$\sum_{i=1}^{n} \rho\left(\frac{y_i - \boldsymbol{\beta} \boldsymbol{x}_i}{\sigma}\right) \tag{3}$$

onde, σ é um estimador robusto e ρ uma função particular. A RQ possui o vetor de parâmetros e os resíduos associados ao Θ-ésimo quantil, $\Theta \in (0, 1)$. O estimador β_0 é encontrado a partir da minimização da seguinte função objetivo

$$\sum_{i=1}^{n} \rho_\theta \left(y_i - \boldsymbol{\beta}_\theta \boldsymbol{x}_i\right) \tag{4}$$

onde, ρ_Θ é a função de *check* definida por

$$\rho_\theta(\epsilon) = \begin{cases} \theta\epsilon & , \epsilon \geq 0 \\ (\theta - 1)\epsilon & , \epsilon < 0 \end{cases} \tag{5}$$

Para o caso da aplicação de técnica não-paramétrica, a SVR como regressão não-linear a equação consiste em

$$\hat{y}_q = \langle \phi(x_i), \phi(x_q) \rangle + b \text{ com } b \in \Re \quad (6)$$

onde, *Φ* consiste numa função não-linear que mapeia o espaço de entrada para um espaço de característica dimensional superior. Nesse caso, o algoritmo SVR envolve o uso de multiplicadores Lagrangeanos, que dependem exclusivamente de produtos de pontos de *Φ*. As variáveis independentes selecionadas em cada cenário proposto foram utilizadas para explicar as relações com o desempenho escolar das proficiências de LP e MT. Todos os experimentos foram desenvolvidos no ambiente *open source* do R Studio, seguindo os passos do Algoritmo 1.

Algoritmo 1 Composição Experimental

1: **Definir** $n = 30$
2: **Para** todo i igual $1 \leq i \leq n$ **faça**:
3: **Particionar** aleatoriamente a base dados em treino e teste (75-25%).
4: **Construir** os modelos de regressão a partir da base de treino (Equações 1, 2, 3, 4, 5 e 6.)
5: **Estimar** a variável resposta de novos exemplos a partir da base de teste.
6: **Calcular** o erro de predição (MAE) para cada modelo construído.
7: **Salvar** os valores de erro de cada modelo.
8: **Fim Para**
9: **Gerar** as análises com base na amostra do MAE para cada modelo (boxplot, média, desvio padrão e teste estatístico.

2.5.Avaliação

Nesta fase os modelos desenvolvidos serão avaliados de acordo com os critérios de MDE definidos. Um dos índices de desempenho mais utilizados para o cálculo da previsão, baseia-se no erro de previsão. Analisa-se os resultados através do erro absoluto médio (MAE). Assim, foi possível desenvolver testes estatísticos (teste de hipótese) para avaliar os modelos de regressão propostos.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \quad (7)$$

onde, *n* é o tamanho do conjunto de dados, y_i é o valor real da variável e $\hat{y}_i$ é a variável e estimada pelo modelo.

2.6. Implementação

Após avaliar os modelos desenvolvidos, as considerações para os cenários de dados do SAEB e todo o conhecimento adquirido com a aplicação do CRISP-DM neste estudo será organizado e apresentado na seção de resultados.

3. Análise dos Resultados

Nas subseções seguintes, serão apresentados os resultados das técnicas de previsão para os quatro cenários propostos. Os modelos foram construídos para cada Cenário de acordo com as variáveis independentes selecionadas pelo método estatístico *Stepwise* (apresentada na Tabela 2). A primeira subseção mostra os resultados para a predição da

proficiência em LP, (Cenário 1 e 2) e a segunda, apresenta os resultados para a predição da proficiência em MT (Cenário 3 e 4).

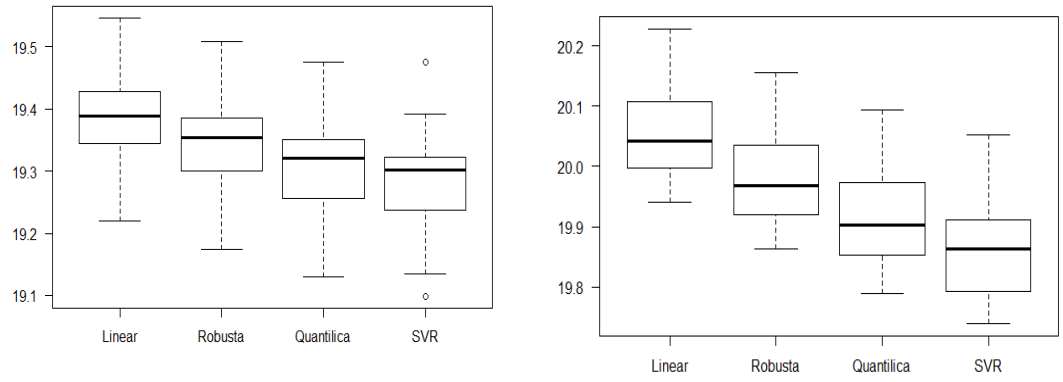
3.1. SAEB 2015 e 2017 para proficiência em Língua Portuguesa

A Tabela 3 mostra os valores de média e desvio padrão (entre parenteses) obtidos das amostras de cada modelo de regressão construídos. O menor valor médio de erro obtido é do SVR tanto para o Cenário 1 quanto para o Cenário 2. A Figura 1 mostra o *boxplot* após as execuções das técnicas de regressão utilizando as variáveis explicativas definidas para este grupo. Em relação à variância amostral, observa-se que entre os métodos há alta variância, mas sem muita diferença entre um modelo e outro (também observado nos valores de desvio padrão da Tabela 3).

Tabela 3 – Valor do erro médio das execuções (desvio padrão) – Cenários 1 e 2.

Cenário/Modelo	RL	RLR	RQ	SVR
1 – LP 2015	19,3823 (**0,0739**)	19,3404 (0,0746)	19,3033 (0,0769)	**19,2826** (0,0791)
2 – LP 2017	20,0586 (**0,0716**)	19,9834 (0,0747)	19,9204 (0,0860)	**19.8688** (0,0888)

Os modelos paramétricos (RL, RLR e RQ) obtiveram uma mediana amostral maior nos dois cenários estudados. O modelo SVR apresenta a menor mediana amostral para o valor do erro médio absoluto. Já o modelo RL possui maior mediana tanto em (a) quanto em (b). A presença de maior dispersão nos dados pode contribuir para esse pior desempenho, enquanto pode ser visto na análise do modelo RLR que a baixa sensibilidade a *outliers* fez com que obtivesse um desempenho um pouco melhor. Ainda, a análise dos resíduos do modelo RL mostra uma não normalidade, não satisfazendo uma das suposições de sua aplicação. Ou seja, não há uma garantia da explicação do modelo, portanto a regressão linear múltipla não seria a mais indicada.



(a) Cenário 1 **(b) Cenário 2**

Figura 1. Boxplot para amostra dos erros gerados para Cenários 1 e 2.

O teste de hipótese de *wilcoxon* foi realizado para ratificar o melhor desempenho do modelo SVR, apresentando menor erro que os demais, ou seja, teste unilateral para amostras pareadas. Os resultados foram analisados através dos valores de *p-value* ($9{,}313 \times 10^{-10}$) para estes dois Cenários, a um nível de significância de 5%. Indicando, estatisticamente, que há evidências de que o modelo SVR tem uma média amostral menor que outros modelos.

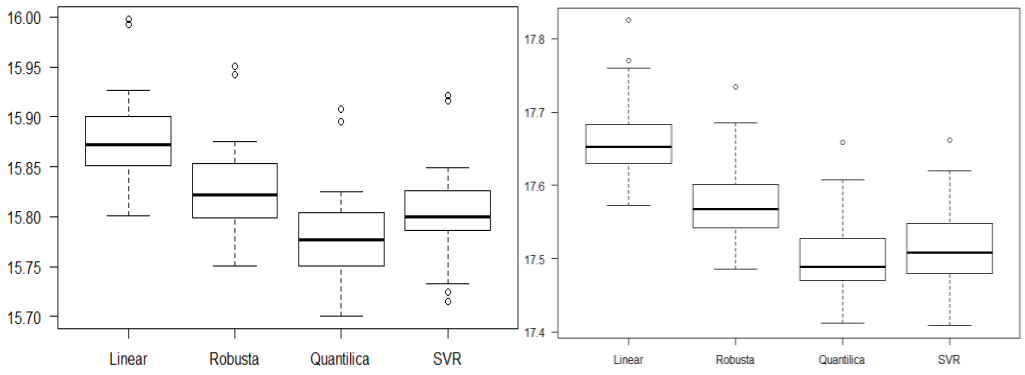
3.2. Cenários 3 e 4: SAEB 2015 e 2017 para proficiência em Matemática

A Tabela 4 apresenta os valores de média e desvio padrão (entre parênteses) obtidos das amostras referente aos Cenários 3 e 4. O menor valor médio de erro obtido foi do modelo RQ em ambos os Cenários. A Figura 2 mostra o *boxplot* após as execuções dos modelos de regressão utilizando as variáveis independentes definidas para estes grupos. Pode ser observado *outliers* nos *boxplots* dos modelos. Indicando a presenta de valores discrepantes nos dados. Dessa forma, verifica-se que o desempenho do modelo RL, a qual possui sensibilidade a *outliers*, e apresenta a maior mediana amostral em (a) e (b); o modelo RLR apresentou um erro menor comparada a RL, pois o desempenho o modelo RLR utiliza o método dos mínimos quadrados reponderados, possuindo uma estimativa melhor quando existe aos *outliers*.

No entanto, a menor mediana do MAE é apresentada pelo modelo RQ, pois além de ser robusta em resposta aos *outliers*, utiliza em sua estimativa diferentes quantis dos dados. Neste estudo, utiliza-se a mediana dos dados (quantil 0.5), sendo mais representativa faixa para estes cenários. Apesar do modelo SVR ser de natureza não-paramétrica e os resultados anteriores apresentar melhor performance, não conseguiu superar o modelo RQ para esse grupo de experimentos.

Tabela 4 – Valor do erro médio das execuções (desvio padrão) – Cenários 3 e 4.

Cenário/Modelo	RL	RLR	RQ	SVR
3 –MT 2015	15,8781 (**0,0458**)	15,8285 (0,0464)	15,7795 (0,0478)	**15,8038** (0,0461)
4 – MT 2017	17,6607 (**0,0558**)	17,5758 (0,0562)	**17,5026** (0,0560)	17,5175 (0,0563)



(b) Cenário 3 **(b) Cenário 4**

Figura 2. Boxplot para amostra dos erros gerados para Cenários 3 e 4.

Analisa-se os resultados do teste de hipóteses de *wilcoxon* unilateral pareado, através dos valores de *p-value* ($9{,}313 \times 10^{-10}$), para este cenário. Com um nível de significância de 5%, os testes indicaram que, estatisticamente, há evidências de que o modelo RQ possui menor média amostral do que outros modelos de regressão nestes cenários. Portanto, podemos ver que a aplicação do modelo RQ proporcionou um melhor resultado na previsão do desempenho dos alunos nos Cenários 3 e 4.

4. Discussões

Após a análise desses resultados, percebe-se que a aplicação do modelo SVR proporcionou um melhor resultado na predição do desempenho dos alunos nos Cenários que buscam a estimação do valor proficiência em Língua Portuguesa. A vantagem de

utilizar uma técnica não paramétrica é que não há suposição da relação entre a variável resposta e a variável explicativa. Assim, formando curvas nos ajustes dos dados (e não uma suposição de parâmetros, como nas regressões paramétricas utilizadas), esse tipo de técnica proporciona melhor desempenho. Além disso, SVR pode obter melhores resultados porque procura o valor ideal no espaço de pesquisa.

Já para os Cenários que consideram a proficiência em Matemática como variável resposta, identifica-se o modelo RQ como o que minimoza o erro de predição. O modelo SVR nestes grupos de experimentos não conseguiu ser superior, mesmo apresentando resultados mais significativos que as demais técnicas paramétricas. Pode-se perceber nos *boxplots* dos modelos a presença de *outliers*. Ao considerar a mediana dos dados, o modelo RQ estima sobre a faixa de dados mais significativas, sendo mais robustas a *outliers*. Estas características da RQ minimizaram o erro de predição para este grupo.

Com esses resultados, pode-se observar que técnicas de regressão não paramétrica podem ser aplicadas em cenários educacionais. O poder de ajustes e flexibilidade aos dados justifica a aplicabilidade desse tipo de modelo, quando os modelos paramétricos são insuficientes. A técnica não-paramétrica utilizada traz uma modelagem de regressão mais flexível, buscando uma resposta que melhor corresponda aos dados. Essas técnicas trazem ganhos para a área educacional em direção a predição de fatores educacionais com mais precisão e menor erro. Contudo, percebe-se que regressão paramétrica pode apresentar resultados superiores em relação as não-paramétricas, tendo em vista as características dos dados utilizados e do modelo utilizado. A regressão quantílica tem a vantagem de estimar a mediana (em vez da média em RL), seu resultado vai ser mais robusto, em resposta a existência de *outliers* nos dados.

Nesse estudo utiliza-se um conjunto de variáveis independentes, as quais foram selecionadas por cenários de aplicação. Estes cenários correspondem a predizer o valor da proficiência dos alunos em Língua Portuguesa e Matemática utilizando dados do SAEB dos anos de 2015 e 2017. Enquanto a maioria dos trabalhos encontrados aplicam tarefas de classificação para os dados educacionais, foi proposta a estimação do valor da proficiência investigando os resultados de diferentes modelos de regressão (paramétrico e não paramétrico) para encontrar o que minimize o erro de predição.

5. Conclusões e Trabalhos Futuros

Neste estudo, foi utilizado o método *Stepwise* para selecionar as variáveis independentes, que servirão como base para a construção dos modelos preditivos. A vantagem da utilização de modelos paramétricos e não paramétricos para previsão proporciona um maior ajuste e flexibilidade na busca de melhores respostas para os dados. Partindo de uma visão mais holística em um todo único e integrado permite resultados mais abrangentes e precisos. O modelo proposto que utiliza SVR, obteve melhores resultados nos cenários 1 e 2, em relação aos modelos RL, RLR, RQ, obtendo métricas de predição do desempenho de 79% e 88% respectivamente. Nos cenários 3 e 4 o modelo proposto que utiliza RQ, obteve melhores resultados em relação aos demais modelos, obtendo métricas de predição de 47% e 56% respectivamente.

Verificou-se a existência dos fatores associados a aprendizagem e sua relação aos grupos de fatores propostos por [Soares 2004], principalmente os relacionados ao aluno e a família. Constata-se uma forte relação positiva com o desempenho escolar, a quantidade de banheiros e quartos na residência, o fato de possuir geladeira em casa, a

área de domicilio do estudante, o incentivo dos pais a leitura, o incentivo dos pais a fazer o dever de MT, a realização das tarefas escolares, a frequência com que os pais vão as reuniões na escola, o nível de escolaridade da mãe e se a mãe sabe ler e escrever. Em contrapartida, os fatores associados a aprendizagem que possuem uma maior relação negativa com o desempenho escolar, a quantidade de reprovações, o fato de trabalhar fora de casa, a idade do aluno, a quantidade de domésticas e a quantidade de pessoas que moram na mesma residência. A partir do conhecimento desses fatores, os gestores educacionais e professores podem criar mecanismos para melhorar o desempenho de seus alunos, bem como avaliar a eficácia de seus sistemas de ensino. Através da adoção de estratégias que minimizem as desigualdades de aprendizagem existentes nas escolas públicas do estado de Pernambuco.

Como trabalhos futuros, a partir dos modelos de Regressão utilizados nesta pesquisa, o desenvolvimento de uma abordagem que combine esses modelos com técnicas de *Ensemble* no contexto de *EDM,* para o diagnóstico de problemas educacionais. Além disso, a utilização dos modelos de regressão propostos (RL, RLR, RQ e SVR) no contexto da *Academic Analytics* para predição de desempenho escolar utilizando dados e informações oriundas de sistemas educacionais de apoio ao ensino aprendizagem e de gestão acadêmica em universidade públicas, como por exemplo: SIGA, ATRIO.

6. Referencias

Anuário. (2019) Anuário Brasileiro da Educação Básica. Todos pela Educação, Editora Moderna.

Baker, R., S., J., Carvalho, M., J., B., D., Isotani, (2011). Mineração de Dados Educacionais Oportunidades para o Brasil. Revista Brasileira de Informática na Educação.

Calderon, O., A., B., Aranibar, D. (2015). Optimal selection of factors using generic algorithms and neural networks for the Prediction of student´s academic performance. Latin America Congress on Computacional Intelligence (*LA-CCI*), p 341-346.

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

Inep, (2018). Instituto Nacional de Pesquisas Educacionais Anisio Teixeira. (Acessado em 20 de maio de 2018). www.inep.gov.br/SAEB.

Rabelo, H., Medeiros, S., R., S., Burlamaqui, A., M., F., Valentim, R., A., M., Rabelo, D., S., S., (2017). Utilização de técnicas de Mineração de Dados Educacionais para a predição de desempenho de alunos de EAD, em Ambientes Virtuais de Aprendizagem. VI Congresso Brasileiro de Informática na Educação (CBIE). P. 1527-1536.

Rodrigues, R., L., Medeiros, F., P., A., Gomes, A., S., (2013). Modelo de Regressão Linear aplicado à Previsão de Desempenho de Estudantes em Ambientes de Aprendizagem. II Congresso Brasileiro de Informática na Educação (CBIE). P. 607-616.

Soares, J., F., (2004). O efeito da escola no desempenho cognitivo de seus alunos. Revista Eletronica Iberoamericana sobre Calidad, Eficacia y Cambio em Educacion, Madrid, v.2, n.2, p 83-104.