Estimação de Índices de Aprovação e Reprovação Escolar do Ensino Médio

Ricardo B. das Neves Junior[1], Rafaella L. S. do Nascimento[2], Roberta A. A. Fagundes[1], Paulo S. G. de Mattos Neto[2]

[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco (UPE)
R. Benfica, 455 - Madalena, Recife PE, 50720-001

[2]Centro de Informática, Universidade Federal de Pernambuco (UFPE)
Av. Jorn. Aníbal Fernandes, s/n - Cidade Universitária, Recife -

rbnj@ecomp.poli.br, rlsn@cin.ufpe.br,roberta.fagundes@upe.br, psgmn@cin.ufpe.br

Abstract. *Educational data mining seeks to study and contribute results that explain variables and find possible solutions to problems in the area of education. Considering this motivation, this article describes a study based on educational data provided by INEP and the construction of prediction models using nonparametric quantile regression with and without parameter optimizer. We present a descriptive study of the explanatory variables of the model, which is used to predict school approval and disapproval. The results obtained show that the quantile regression with optimization received a smaller prediction error. The study shows the relevance in the application of nonparametric regression techniques.*

Resumo. *A mineração de dados educacionais busca estudar e contribuir com resultados que expliquem variáveis e encontrem possíveis soluções para problemas na área da educação. Tendo em vista essa motivação, este artigo descreve um estudo com base em dados educacionais fornecidos pelo INEP e a construção de modelos de predição utilizando a regressão quantílica não paramétrica com e sem otimizador de parâmetros. Apresenta-se um estudo descritivo das variáveis explicativas do modelo o qual é utilizado para predizer a aprovação e a reprovação escolar. Os resultados obtidos mostram que a regressão quantílica com otimização obteve menor erro de predição. O estudo mostram a relevância na aplicação de técnicas não paramétricas de regressão.*

1. Introdução

O estudo de variáveis educacionais é uma corrente e objeto de várias pesquisas, com o objetivo de compreender o contexto do fenômeno nos mais diferentes aspectos do ensino como ambientes educacionais, modalidades (à distância e presencial) e estágios da educação. Um dos fatores que facilitam o desenvolvimento destes trabalhos é a disponibilidade de dados relacionados à educação, aumentando a aplicabilidade da Mineração de Dados Educacionais (MDE) [Romero and Ventura 2010].

A MDE é uma abordagem que objetiva 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

DOI: 10.5753/cbie.sbie.2019.339

influenciam a aprendizagem e que representam a qualidade do ensino, como os indicadores educacionais. 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 2019].

De forma geral, a MDE explora técnicas estatísticas, de aprendizado de máquina e de mineração de dados (MD) sobre os diferentes tipos de dados educacionais. Existem várias linhas de pesquisa na área de educação e muitas delas derivadas da área de MD, como tarefas preditivas, de agrupamento ou de associação [Baker et al. 2011]. Entre as preditivas, podem ser utilizadas técnicas de regressão, que têm como um dos objetivos predizer o valor da variável resposta a partir da informação proveniente de uma variável ou de um conjunto de variáveis explicativas [Montgomery et al. 2012].

Estas técnicas são abordadas neste trabalho na estimação dos indicadores educacionais da reprovação e aprovação escolar de alunos do ensino médio no âmbito do estado de Pernambuco. As bases de dados utilizadas na pesquisa são fornecidas abertamente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP e são referentes ao ano de 2016. Dentre as possíveis técnicas de regressão, investiga-se a aplicação da abordagem não paramétrica a partir da regressão quantílica.

A regressão quantílica difere dos demais tipos de regressão, pois permite o uso de vários quantis para obter uma visão mais completa da relação entre as variáveis estudadas [Koenker and Bassett Jr 1978]. Na regressão quantílica não paramétrica (RQNP), um kernel pode ser aplicado ajustando a largura de banda (parâmetro que controla o grau de suavidade da função estimada). Investiga-se, portanto, o ganho que essas técnicas podem oferecer para a área educacional em direção a uma melhor resposta de predição.

Este artigo está dividido da seguinte forma: o item 2 apresenta trabalhos relacionados ao tema deste artigo; em 3 apresenta a base de dados utilizada bem como o processamento realizado; em 4, os modelos propostos; 5 mostra a avaliação experimental e os resultados obtidos; e a seção 6 compõe as conclusões, desenvolvidas após a análise dos resultados obtidos ao final dos experimentos.

2. Trabalhos Relacionados

No trabalho de [Laisa and Nunes 2015] o objetivo foi analisar uma base dados de alunos do ensino médio dos anos de 2011 a 2014, por série, a partir da utilização da técnica de classificação usando o Algoritmo J48. As bases obtidas foram de uma escola particular. Busca entender as variáveis relacionadas aos alunos aprovados e reprovados. Os modelos gerados utilizaram-se dos seguintes atributos de cada aluno: nível (1ª, 2ª ou 3ª série), cidade de origem (Local ou Vizinha), bolsista (sim ou não), desistentes (sim ou não), sexo do aluno, idade e ano (2011 a 2014).

O trabalho de [do Nascimento et al. 2018a] utiliza bases de dados educacionais fornecidas pelo INEP e aplica técnicas de mineração de dados com a finalidade de melhor explicar indicadores educacionais como a evasão e reprovação escolar no ensino fundamental. Realiza análise correlacional entre variáveis e aplica modelos de regressões linear e robusta para predição das variáveis resposta. Estas bases do INEP são utilizadas em [Colpani 2018], com dados do ano de 2017. Aplicou-se regressão linear a fim de analisar as variáveis que se relacionam com a evasão escolar. Como resultados, foi possível verificar que a taxa de distorção idade-série apresentou a maior correlação positiva.

A pesquisa desenvolvida por [do Nascimento et al. 2018b] utiliza duas técnicas não paramétricas de regressão, Regressão Quantílica e SVR, para prever a evasão no estado de Pernambuco, para alunos do ensino fundamental. O desenvolvimento do trabalho seguiu as fases do CRISP-DM (*Cross-Industry Standard for Data Mining*) e utilizou os dados do Censo Escolar de 2015. Algumas variáveis explicativas utilizadas são: número de computadores, de salas, de funcionários e localização da escola.

No trabalho de [Calixto et al. 2017] objetivo foi identificar as variáveis relacionadas à evasão escolar, utilizando os dados do Censo Escolar dos anos de 2014, 2015 e 2016 dos estados de Ceará e Sergipe. Aplicou-se técnicas de Indução de Regras e Regressão Logística. Como resultados obtidos, a idade, etapa de ensino, modalidade de ensino, existência de laboratórios e localização da escola se destacaram como variáveis influentes na evasão escolar no cenário de estudo.

Para realizar a estimação dos índices de aprovação e reprovação em escolas, este trabalho utiliza uma abordagem de RQNP utilizada por [Li et al. 2013]. O modelo proposto pelos autores apresenta um método automático para estimação do parâmetro Largura de Banda presente na RQNP baseado em validação cruzada. O método está disponível no R (*R Core Team*) através do pacote "np"(função "npcdistbw") versão 0.40-14 ou superior [Hayfield and Racine 2008]. Como segunda alternativa de estimação do parâmetro Largura de Banda presente na RQNP, este trabalho utiliza os Algoritmos Genéticos (AG), fundamentados por [Holland 1992], os AGs são algoritmos inteligentes de otimização da família da computação evolucionária e são inspirados na teoria das origens das espécies proposta por Darwin (1909) [Darwin 1909].

Tendo em vista estes trabalhos relacionados, esta pesquisa traz como diferencial a aplicação de técnicas de regressão não paramétricas, enquanto a maioria dos estudos com indicadores educacionais consistem em aplicar modelos de regressão paramétricos ou de classificação. Assim, utilizando a técnica de RQNP, serão construídas as análises preditivas da aprovação e reprovação escolar. As técnicas não paramétricas trazem uma modelagem mais flexível e busca uma resposta que melhor represente os dados avaliados.

3. Bases de Dados

As bases de dados utilizadas neste estudo são disponibilizadas abertamente pelo INEP [INEP 2019]. Os dados são referentes a indicadores educacionais do ano de 2016. Os indicadores educacionais considerados são:

- **Taxa de eficiência escolar (TR/TA)**: Taxa de Reprovação/Aprovação da escola.
- **Adequação da formação de professores (AFD)**: Porcentagem de professores por grupo da adequação da formação à disciplina que ensinam (cinco grupos, AFD11 a AFD5).
- **Alunos por turma (ATU)**: O número médio de alunos por turma.
- **Complexidade da gestão (ICG)**: Nível de complexidade de gestão escolar. O indicador classifica as escolas em níveis de 1 a 6 de acordo com sua complexidade de gestão, níveis elevados indicam maior complexidade.
- **Distorção idade-série (TDI)**: Taxa de distorção idade-série dos alunos por escola.
- **Professor com ensino superior (DSU)**: O percentual de professores com ensino superior na escola.

- **Esforço do professor (IED)**: O percentual de professores que trabalham no ensino fundamental e médio pelo nível de esforço exigido para o exercício da profissão (seis níveis, IED1 a IED6).
- **Média de horas de aula (HAU)**: O número médio de horas/aulas diário da escola.
- **A regularidade do professor (IRD)**: A média do indicador de regularidade do professor.

Todas essas variáveis foram integradas em um base de dados únicas, levando em consideração o identificador único de cada escola. Nessas bases, os indicadores de eficiência escolar como reprovação e aprovação são as variáveis resposta (y), e os demais são as variáveis explicativas (x).

3.1. Pré-processamento e Tratamento dos Dados

Como a base inclui informações sobre todos os estados do Brasil, foi realizada a extração de dados apenas para o estado de Pernambuco, o qual é o alvo desta pesquisa. A base também informa sobre várias etapas do ensino, como fundamental e médio. Considerou-se a divisão em duas bases distintas, uma para o ensino médio e outra para o ensino fundamental. No entanto, nesta aplicação utiliza-se os dados relacionados ao ensino médio. Isto significa que busca-se estimar os valores de de Reprovação e Aprovação escolar no ensino médio das escolas do estado de Pernambuco a partir dos indicadores educacionais.

Após isso, foi possível observar poucos valores ausentes para algumas variáveis explicativas, 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 objetivo foi ter o mínimo de perda de instâncias. Diferente disso, a variável HAU, a qual se refere a média de horas-aula na escola, continha um grande número de valores faltantes. Optou-se pela exclusão da HAU da base de dados pois utilizar estratégia de resolução de *missing value* alteraria a variabilidade deste indicador. Nessa fase, também foi aplicada a padronização 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. Portanto, no final deste período a base é composta por 16 variáveis explicativas, em que a base de reprovação possui 1.012 instâncias, enquanto a base de aprovação possui 1.135 instâncias.

3.2. Seleção de Variáveis

Para selecionar quais variáveis explicativas seriam utilizadas na execução das previsões, foi escolhido o algoritmo *Random Forest* (RF) proposto por Breiman em 2001 [Breiman 2001]. Embora o objetivo inicial do algoritmo, fosse desempenhar tarefas de classificação, a literatura indica que o RF é capaz de executar atividades de seleção de variáveis, gerando um *rank* das variáveis de acordo com a importância em relação à variável dependente, ou seja, as variáveis mais importantes deverão ser utilizadas no modelos de previsão [Coutinho 2015]. As 5 variáveis mais importantes selecionadas pelo RF foram utilizadas no modelo de previsão, estas variáveis são: TDI, AFD1, ATU, AFD3, AFD5 (conjunto de dados Aprovação do Ensino Médio) e TDI, AFD3, ATU, AFD1, IRD (conjunto de dados Reprovação do Ensino Médio).

Como pode ser visto, nos dois cenários há a seleção das mesmas variáveis, em maioria. A distorção idade-série é um dos grandes problemas na área da educação, e possui como uma das principais causas a reprovação e o abandono escolar, fazendo com que

o aluno fique atrás duas ou mais séries indicadas para sua idade [INEP 2019]. Portanto, pode-se notar uma forte relação entre TDI e as variáveis resposta (TA com uma relação inversa). Características dos docentes também mostraram-se influentes, como o índice de professores formados em licenciatura (AFD1) e bacharelado (AFD3) na mesma área que ensinam e a regularidade de ensino que esses professores desempenham nas escolas (IRD). Percebe-se também em ambos os cenários a relação das variáveis resposta com a quantidade de alunos por turma das escolas (ATU). No cenário da aprovação, por fim, é identificada a presença da variável AFD5. Esta variável está relacionada a taxa de professores sem formação superior, indicando que a taxa de professores sem formação apresenta uma relação negativa com a taxa de aprovação das escolas.

4. Modelos propostos para estimação de indicadores educacionais

4.1. Regressão Quantílica Não Paramétrica

O modelo de RQNP foi executado através do software R (*R Core Team*) utilizando o pacote "np" [Hayfield and Racine 2008]. Como função de estimação de parâmetros utilizou-se a função "npcdistbw" proposta por Li et al. (2013) [Li et al. 2013]. Como função de ajuste de modelo de RQNP empregou-se o a função "npqreg" que segundo Koenker [Koenker 2006] o modelo de RQNP é ajustado de acordo com a equação:

$$\sum_{i=1}^{N} w_i(x)\rho_\tau(y_i - \beta_0 - \beta_1(x_i - x)) \tag{1}$$

em que ρ_τ é o quantil escolhido, y_i é a variável dependente que se deseja estimar, x_i e x são as variáveis independentes, β_0 e β_1 são os coeficientes de correlação e $w_i(x)$ corresponde ao *kernel* gaussiano, que corresponde à distribuição normal padrão que segundo [Fagundes and Cysneiros 2013], pode ser representado por:

$$K(d(x, x_j)) = \frac{1}{(\sqrt{2\pi})^{1/p}} \frac{1}{h^p} e^{\frac{-d(x,c_i)}{2h^2}} \tag{2}$$

em que $d(x, x_j)$ é a raiz quadrada da distância euclideana entre x e a localização de interesse x_i.

4.2. Otimização por Algoritmos Genéticos

Os Algoritmos Genéticos (AG) atuaram na estimação do parâmetro Largura de Banda na RQNP com o objetivo de melhorar o desempenho do modelo. A função matemática na qual o parâmetro de suavização está presente é representado pela Equação 2, em que h é o parâmetro à ser otimizado, aqui representado por h_D, em que D é a dimensão do problema e h é um indivíduo de D dimensões (i.e. uma matriz $1xD$).

A população inicial P é um conjunto de n indivíduos representados por $P = (h_1, h_2, \cdots, h_D)$. Sabendo que h_D é um conjunto de valores entre os limites inferiores e superiores do espaço de busca. Após gerar a população inicial, é necessário avaliar os indivíduos de acordo com a função objetivo, que a RQNP avalia pelo cálculo do Erro Médio Absoluto (i.e Equação 5), após a estimação realizada pela RQNP (Equação 1).

Para seleção de pais foi utilizado o método roleta, que tem como principal característica fornecer maior e menor probabilidade de seleção de indivíduos bons e ruins,

respectivamente. A principal vantagem é manter dentro da população indivíduos bons e ruins, pois, os bons têm maior probabilidade de gerar filhos com alta aptidão e os ruins mantém a diversidade na população, desde modo, evitando mínimos locais. A função matemática o método roleta é dado por:

$$p_i = \frac{f_i}{\sum_{j=1}^{N} f_i} \tag{3}$$

em que N é o número total de indivíduos na população, f é o resultado retornado pela função objetivo e p_i é a probabilidade de seleção do indivíduo.

Na fase de cruzamento, o método *crossover point* foi empregado utilizando dois pais para gerar um filho (novo indivíduo da população), escolhendo um ponto aleatório dos cromossomos (pais) e efetuando o cruzamento de informação entre eles para gerar um novo indivíduo. Dado que temos um problema de D dimensões, deverá ser escolhido um ponto p aleatório ente o intervalo [1, D].

Na etapa de mutação, para procurar melhores soluções, uma busca local b foi inserida nos genes dos cromossomos. Em um problema de D dimensões, deverá se escolher um valor inteiro aleatório y entre o intervalo [1, D], este valor selecionado aleatoriamente é a quantidade de genes do cromossomo que sofrerão mutação. Logo após, é necessário selecionar aleatoriamente y genes do cromossomo e aplicar nestes genes uma busca local. A mutação é realizada conforme a Equação 4

$$h_D = h_D + b \tag{4}$$

em que b é um valor aleatório entre o intervalo [-1,1]. Posteriormente, o indivíduo modificado fará parte da próxima geração.

Uma vez realizado os procedimentos de *crossover* e mutação, os indivíduos da próxima geração h_D serão avaliados pela função objetivo elucidada na Equação 5 e um novo ciclo inicia conforme fluxograma da Figura 1.

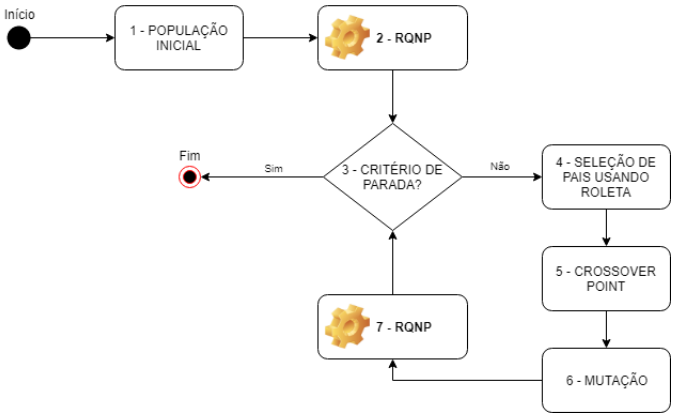


Figura 1. Diagrama Algoritmos Genéticos utilizado para otimizar a RQNP

A Figura 1 mostra um fluxograma do AG presente na literatura. O detalhamento do algoritmo é mostrado de acordo com os passos abaixo.

1. Inicializar população aleatoriamente.
2. Calcular *fitness* da População, utilizando a RQNP apresentada na equação 1.
3. Verifica se o critério de parada foi alcançado. Se sim, finaliza a execução do algoritmo. Se não, executa o próximo passo.
4. Seleção de pais utilizando Roleta, conforme elucidado pela equação 3.
5. Executa *crossover point* usando os pais selecionados no passo anterior.
6. Executa mutação.
7. Calcular *fitness* dos indivíduos sobreviventes.
8. Se o critério de parada estiver satisfeito, interrompe a execução do algoritmo. Se não, voltar para o passo 4.

5. Avaliação Experimental

Para todos os testes realizados, foram utilizados os seguintes parâmetros: Tamanho da População = 30; Quantidade de iterações = 100; Chance de ocorrer um *crossover* = 80%; Chance de ocorrer uma mutação = 5%; Limites inferior e superior do espaço de busca = -1 e 3 (considerados somente na inicialização aleatória da população); Dimensão do problema = 6; Limites inferior e superior da busca local utilizada nos processos de *crossover* e mutação = -1 e 1, Condições de parada: (*i*) quando executar o número pré-estabelecido de iterações e (*ii*) se o algoritmo não evoluir durante 30% do total de iterações.

A estimação da variável dependente foi realizada utilizando o método de validação cruzada *holdout* [Kohavi et al. 1995]. Este método separa os dados em dois conjuntos, chamados de treino e teste. Para cada experimento, os dados foram divididos em 75% como treino e 25% como conjunto de teste.

A avaliação de desempenho dos modelos baseiam-se na Erro Médio Absoluto (EMA), que é dado pela equação 5:

$$EMA = \frac{\sum_{i=1}^{n} |y_i - \hat{y_i}|}{n} \tag{5}$$

em que y_i é o valor real, $\hat{y_i}$ é o valor predito e n é o número de instâncias da base.

O teste estatístico de Wilcoxon para amostras não pareadas com um nível de significância de 5% foi aplicado, a fim de verificar estatisticamente a superioridade de um modelo sobre outro [Montgomery et al. 2000]. Considerou-se μ_1 e μ_2 como média EMA dos modelos comparativos. A hipótese nula é $H_0 : \mu_1 = \mu_2$ e hipótese alternativa é $H_1 : \mu_1 < \mu_2$. Em que μ_1 é a média do resultado da RQNP otimizada por AG e μ_2 é a média do resultado da RQNP padrão.

5.1. Discussões e Resultados

A Tabela 1 mostra o resultado (quanto menor, melhor) comparativo entre as estimações realizadas pelo modelo RQNP padrão e RQNP otimizado por Algoritmos Genéticos. De acordo com os resultados de média e desvio padrão (entre parêntesis), percebe-se que o modelo otimizado por AG obteve uma média de erro menor com maior desvio padrão. Para avaliar a confiabilidade estatística dos modelos, executamos o teste estatístico de Wilcoxon não pareado com 5% de significância. O resultado alcançado pelo teste estatístico de Wilcoxon indica que os resultados alcançados pelos modelos são de fato diferentes, logo, rejeitamos H_0.

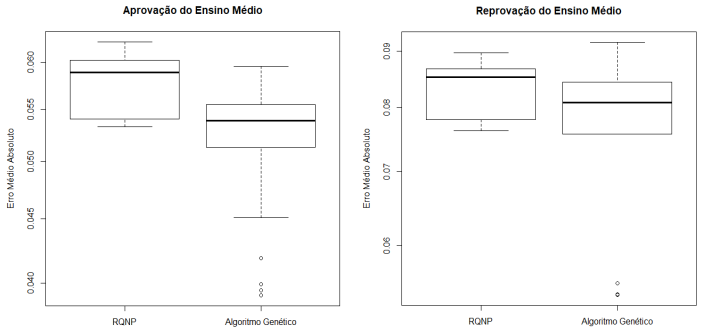
Tabela 1. Resultado comparativo entre a RQNP e os Algoritmos Genéticos, com a média e desvio padrão (entre parêntesis) para o EMA.

Conjunto de Dados	RQNP	Algoritmo Genético
Aprovação do Ensino Médio	0.0576 (0.0030)	**0.0521** (0.0058)
Reprovação do Ensino Médio	0.0834 (0.0042)	**0.0764** (0.0129)

As Figuras 2(a) e 2(b) mostram o resultado comparativo entre a estimação realizada pelo RQNP padrão e a estimação realizada pela RQNP otimizada por AG, nos conjuntos de dados de aprovação e reprovação. De acordo com os gráficos em *boxplot* apresentados, nota-se que os modelos propostos possuem variações semelhantes, no entanto o modelo otimizado por AG, possui menor mediana amostral.

O modelo otimizado por AG alcança melhor resultado porque enquanto o modelo de RQNP utiliza uma abordagem estatística de estimação baseada em validação cruzada, o AG utiliza uma abordagem de computação inteligente baseada em população, em que para cada geração, 10 indivíduos candidatos são testados, e os melhores resultados são considerados para a próxima geração. Portanto, o modelo com otimizador obteve melhor predição da taxa de reprovação e aprovação escolar no cenário aplicado. O modelo de RQNP padrão, possui menor custo computacional, logo é recomendado em situações que se deseja uma estimação mais rápida. No entanto, se o objetivo é maior precisão de estimação, indica-se a utilização do modelo de RQNP otimizado por AG.

Os modelos propostos neste trabalho podem ser utilizados como sistemas de apoio a decisão, já que indicam em que casos pode haver uma reprovação. Logo, diante da indicação de que pode haver uma alta taxa de reprovação, a escola será capaz de agir preventivamente.



(a) Resultados Aprovação (b) Resultados Reprovação

Figura 2. *Boxplot* comparativo entre a RQNP e os Algoritmos Genéticos

6. Conclusões e Trabalhos Futuros

Este trabalho apresentou duas abordagens para previsão de aprovação e reprovação escolar do ensino médio. O conjunto de dados utilizado neste estudo estão disponibilizadas abertamente pelo INEP, referente ao ano de 2016, para o âmbito do estado de Pernambuco. Este estudo buscou investigar e explicar indicadores educacionais que podem estar relacionadas com a taxa de aprovação e reprovação das escolas. Na seleção de variáveis através da técnica Random Forest, pode ser observado que variáveis como a dispersão

idade-série das escolas, nível de formação dos docentes e quantidade de alunos por turma tem maior importância na construção do modelo de estimação (variáveis explicativas).

Outra contribuição deste trabalho são os modelos propostos de Regressão Quantílica Não Paramétrica (RQNP) e RQNP Otimizada por Algoritmos Genéticos (AG) na estimação da aprovação e reprovação escolar. Embora as abordagens apresentem médias parecidas, o teste estatístico de Wilcoxon indica que a abordagem otimizada por Algoritmos Genéticos é capaz de fornecer uma previsão mais precisa, em relação ao modelo sem otimização, uma vez que minimizou o erro de predição.

A Mineração de Dados Educacionais possibilita o conhecimento acerca de variáveis educacionais, resultando em ferramentas que ofereçam melhorias neste cenário. Desta forma, a aplicação de modelos não paramétricos foi satisfatório, oferecendo contribuições através de modelos mais robustos na estimação de índices educacionais. Como trabalhos futuros, pretende-se investigar outros fatores educacionais, assim como aplicar outras formas de otimização de modelos para diminuir o erro da predição.

Através de técnicas de predição (na aplicação da Mineração de Dados Educacionais), sistemas e plataformas podem ser desenvolvidos e adotados no ambiente educacional para fins de identificação de fatores relacionadas as taxas de eficiência escolar, ou quaisquer outros indicadores educacionais. Este trabalho revela o primeiro passo de um estudo que pode ser expandido para auxilio no ensino/aprendizagem, pois identificar fatores influentes na educação pode revelar *insights* que ajudem a melhorar este cenário, e assim criar programas, intervenções ou investimentos.

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.

Breiman, L. (2001). Random forests machine learning. 45: 5–32. *View Article PubMed/NCBI Google Scholar*.

Calixto, K., Segundo, C., and de Gusmão, R. P. (2017). 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)*, volume 28, page 1447.

Colpani, R. (2018). Mineração de dados educacionais: um estudo da evasão no ensino médio com base nos indicadores do censo escolar. *Informática na educação: teoria & prática*, 21(3).

Coutinho, L. D. (2015). Utilizando redes neurais artificiais e algoritmos de seleção de variaveis para realizar diagnóstico precoce da doença de alzheimer e do déficit cognitivo leve. Master's thesis, Universidade de Pernambuco.

Darwin, C. (1909). *The origin of species*. Dent.

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. *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 regres-

sors in predicting school dropout. In *International Conference on Machine Learning and Data Mining in Pattern Recognition*, pages 246–257. Springer.

Fagundes, R. A. d. A. and Cysneiros, F. J. d. A. (2013). Métodos de regressão robusta e kernel para dados intervalares. *Universidade Federal de Pernambuco*.

Hayfield, T. and Racine, J. S. (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software*, 27(5).

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

INEP (2019). Instituto nacional de estudos e pesquisas educacionais anísio teixeira. url: http://portalinep.gov.br/. Accesso em 14 de Janeiro de 2019.

Koenker, R. (2006). Quantile regresssion. *Encyclopedia of environmetrics*.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Laisa, J. and Nunes, I. (2015). 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)*, volume 26, page 1112.

Li, Q., Lin, J., and Racine, J. S. (2013). Optimal bandwidth selection for nonparametric conditional distribution and quantile functions. *Journal of Business & Economic Statistics*, 31(1):57–65.

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

Montgomery, D. C., Runger, G. C., and Calado, V. (2000). *Estatística Aplicada E Probabilidade Para Engenheiros* . Grupo Gen-LTC.

Romero, C. and Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 40(6):601–618.