Revista Semiárido De Visu

ISSN 2237-1966

DOI: 00.0000/0000-0000.0000000000

Previsão de Evasão e Retenção no Ensino Médio Profissional: Uma Abordagem Baseada em Redes Neurais

Evasion and Retention Prediction in Vocational High School: A Neural Network based Approach

SOUZA, Francicleide Geremias da Costa. M.e. em Educação Profissional
Instituto Federal do Ceará. Av. Geraldo Barbosa Marques, 567 - Crateús - CE - Brasil. CEP: 63708-260. Telefone: +558821512943 / E-mail: francicleide.geremias@ifce.edu.br

ARAÚJO, Ricardo de Andrade. Ph.D. em Ciência da Computação
Instituto Federal do Sertão Pernambucano. Estr. do Tamboril, s/n - Ouricuri- PE - Brasil. CEP: 56.200-000. Telefone: +558721012350 / E-mail: ricardo.araujo@ifsertao-pe.edu.br

RESUMO

Este trabalho apresenta um estudo sobre séries temporais, relacionadas a índices de evasão e retenção escolar no ensino médio profissional, visando a identificação das características peculiares a estas séries e, baseado neste estudo, propor uma abordagem baseada em redes neurais, do tipo multicamadas, para prever este tipo particular de série temporal. Para o processo de aprendizagem, é utilizado o algoritmo de retropropagação do erro (*back propagation*, BP). Uma análise experimental é conduzida com a abordagem proposta utilizando séries temporais, com frequência semestral, relacionadas aos índices de evasão e retenção do Instituto Federal do Ceará. Nestes experimentos, são utilizadas as medidas relevantes para avaliar o desempenho preditivo, e os testes de Friedman e Tukey para validá-lo estatisticamente. Os resultados alcançados indicam que a abordagem proposta é capaz de prever eficientemente estas séries no período avaliado, sendo opções viáveis para previsão de índices de evasão e retenção escolar em instituições de ensino médio profissional.

Palavras-chave: Evasão, Retenção, Séries Temporais, Redes Neurais, Previsão.

ABSTRACT

This work presents a study about time series, related to rates of evasion and retention in vocational high school, aiming to identify peculiar characteristics of these series and, based on such study, to propose an approach based on neural networks, multilayer-like, to predict this particular kind of time series. For the learning process, it is used the back propagation (BP) algorithm. An experimental analysis is conducted with the proposed approach using time series related to the evasion and retention rates of the Federal Institute of Ceará. In these experiments, relevant measures are used to assess the prediction performance, and the Friedman and Tukey tests to validate it statistically. The achieved results indicate that the proposed approach in this work is able to efficiently predict these series within evaluated period, being feasible options for the prediction of evasion and retention rates in vocational high school institutions.

Keywords: Evasion, Retention, Time Series, Neural Network, Prediction.

Periódico científico de acesso livre

ISSN 2237-1966

1 Introdução

O acesso à Educação Profissional e Tecnológica (EPT) tem sido elevado significativamente na última década através da expansão acelerada da rede federal (CUNHA; MOURA; ANALIDE, 2016). No entanto, com a expansão surgiram diversos problemas, dentre os quais a evasão e retensão têm despertado interesse de diversos pesquisadores na área da educação (RUMBERGER; THOMAS, 2000; VIADERO, 2001; DORE; LUSCHER, 2008; CUNHA et al., 2013; DORE; ARAUJO; MENDES, 2014; SILVA; DIAS; SILVA, 2015; COCCO; SUDBRACK, 2016; CUNHA; MOURA; ANALIDE, 2016; JUNIOR; SANTOS; MACIEL, 2017; TROMBONI; OLEGARIO; LAROQUE, 2017).

Neste contexto, vale mencionar que aproximadamente 1 milhão de estudantes abandonam seus estudos na modalidade de ensino EPT (OLIVEIRA, 2016), o que representa um nível alarmante, uma vez que o quantitativo de matrículas nesta modalidade de ensino está por volta de oito milhões (CUNHA; MOURA; ANALIDE, 2016), o que na prática gera um alto índice percentual de aproximadamente doze porcento, contrariando a perspectiva de universalização do acesso à educação e da garantia da permanência dos discentes, bem como gera anualmente perdas financeiras na ordem de bilhões de reais.

De maneira geral, as causas da evasão são complexas, sendo influenciadas por um conjunto de fatores de difícil identificação (RUMBERGER; THOMAS, 2000). No entanto, é possível relacioná-la com a retenção escolar (permanência do aluno no mesmo ano escolar devido ao desempenho insuficiente nas disciplinas cursadas) (REBELO, 2009), que tem gerado efeitos danosos, sobretudo a longo prazo, no que tange o abandono escolar (TROMBONI; OLEGARIO; LAROQUE, 2017).

Na literatura, é possível encontrar algumas abordagens alternativas para análise do problema de evasão e retensão escolar, tais como modelos baseados em regressor de vetor de suporte para estimar índices de evasão (NASCIMENTO et al., 2018), abordagens baseada em mineração de dados (*data mining*, DM) (BAKER; ISOTANI; CARVALHO, 2011) para estimar retenção escolar (YU et al., 2010), classificadores baseados em árvores de decisão (*decision trees*, DT) (KIM, 2008) para estimar o desempenho de estudantes (KABRA; BICHKAR, 2011), sistemas inteligentes baseados em redes neurais *fuzzy-artmap* para estimar o risco de evasão em grupos de estudantes (MARTINHO; NUNES; MINUSSI, 2013b, 2013a), técnicas de mineração de dados para classificar grupos de estudantes com perfil evasor (MARQUEZ-VERA; ROMERO; VENTURA, 2013), modelos baseados em árvores de classificação para estimar a evasão escolar, no contexto do ensino à distância (YASMIN, 2013), dentre outros.

Também, vale destacar abordagens baseadas em árvores de decisão para classificar o desempenho de estudantes (AHMED; ELARABY, 2014), sistemas inteligentes compostos por redes neurais de função de base radial para analisar a evasão discente (KAWASE, 2015), abordagens computacionais para detecção de padrões a serem utilizados na análise de evasão de estudantes, classificando-os como "haverá evasão" ou "não haverá evasão" (JUNIOR, 2015), abordagens híbridas baseada em árvores de decisão e modelos de regras de indução para descoberta de conhecimento a partir de dados de estudantes (MEEDECH; IAM-ON; BOONGOEN, 2016), modelos de aprendizagem de máquina (*machine learning*, ML) para detectar comportamentos de estudantes evasores (CUNHA; MOURA; ANALIDE, 2016), abordagens analíticas, usando técnicas de mineração de dados educacionais (*educational data mining*, EDM) (ROMERO; VENTURA, 2013), para estimar o risco de um estudante ser classificado como possível evasor (JAISWAL; YADAV, 2019).

Mesmo com uma grande diversidade de modelos, métodos, técnicas e abordagens para lidar com o problema de previsão, esforços ainda devem ser realizados para uma análise mais aprofundada do fenômeno gerador de séries temporais de índices de evasão e retenção e, consequentemente para o desenvolvimento de modelos para previsão deste tipo particular de série temporal. Portanto, este trabalho visa desenvolver um estudo sobre séries temporais, relacionadas a índices de evasão e retenção escolar na modalidade de ensino EPT, para identificar as características peculiares a estas séries e, baseado neste estudo, propor uma abordagem baseada em redes neurais, do tipo multicamadas, capaz de prevê-las.

Para o processo de aprendizagem, é utilizado o algoritmo de retropropagação do erro (*back propagation*, BP) (HAYKIN, 2007). Uma análise experimental é conduzida com a abordagem proposta utilizando séries temporais, com frequência semestral, relacionadas aos índices de evasão e retenção do Instituto Federal do Ceará, e os resultados alcançados indicam que a abordagem proposta é capaz de prever eficientemente estas séries no período avaliado, permitindo o desenvolvimento de políticas institucionais de prevenção de situações adversas no tocante ao abandono escolar e minimizando as consequências negativas deste fenômeno.

Este trabalho foi organizado da seguinte forma. Na Seção 2, é apresentada uma análise das séries temporais relacionadas a índices de evasão e retenção escolar, na modalidade de ensino EPT, investigadas. Em seguida, na Seção 3, é descrito o modelo baseado em redes neurais artificiais proposto, bem como seu processo de aprendizagem. Posteriormente, na Seção 4, são mostradas e analisadas as simulações e os resultados experimentais realizados com o modelo proposto. Ao final, na Seção 5, são apresentadas as considerações finais e os trabalhos futuros promissores em relação a esta temática.

ISSN 2237-1966

2 Séries Temporais

De acordo com Box *et al.* (BOX; JENKINS; REINSEL, 1994), uma série temporal (**X**) pode ser representada como uma sequência de observações de um determinado fenômeno que evolui com o tempo, sendo definida por

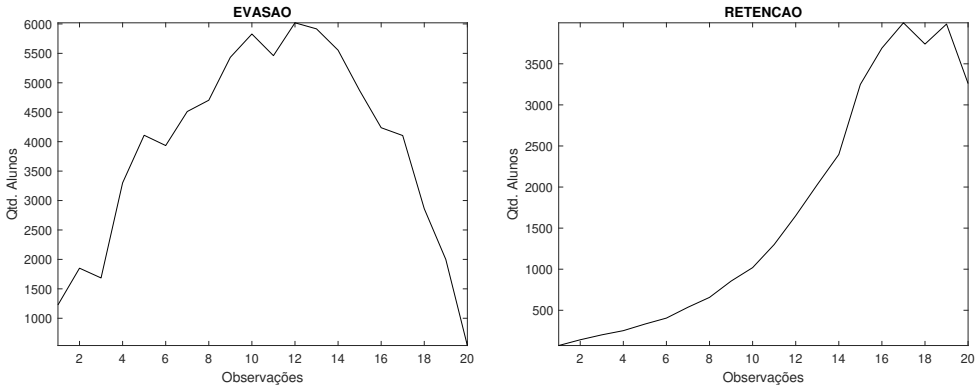
$$\mathbf{X} = \{x_t \in \mathbb{R} \mid t = 1, 2, 3 \ldots N\}, \tag{1}$$

em que x_t representa uma observação no tempo t, e N representa o total de observações.

O principal objetivo de se construir um modelo de previsão é gerar um mapeamento capaz de estimar, com certa precisão, as observações futuras de uma série temporal, dados por x_{t+h}, em que h é o horizonte de previsão de h passos a frente (ARAÚJO, 2016). A ideia básica do problema de previsão é definir uma janela temporal (d) contendo as observações do passado da série temporal. Esta janela deve conter as informações e características necessárias para melhor aproximação possível do fenômeno gerador da série temporal de interesse. O conjunto de observações nesta janela temporal é conhecido como retardos temporais (*time lags*, TL) (BOX; JENKINS; REINSEL, 1994). Note que o principal elemento para um desempenho acurado em estimar um fenômeno temporal é a escolha correta dos retardos temporais de maneira a caracterizar as leis que governam tal fenômeno.

Um estudo de caso sobre o fenômeno gerador de séries temporais provenientes do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE) é apresentado neste trabalho. Para tal, são investigadas duas séries temporais (com frequência semestral) referentes ao Índice de Evasão Escolar (IEE) e ao Índice de Retenção Escolar (IRE) do IFCE no período de 2009 a 2018. Vale mencionar que ambos os índices estão relacionados com a quantidade de alunos evadidos e retidos, respectivamente. Inicialmente, seguindo a metodologia proposta por Araújo (ARAÚJO, 2016), este trabalho considerou realizar a análise do fenômeno gerador a partir de seus gráficos, que podem ser ilustrados na Figura 1.

Figura 1: Gráfico das séries temporais dos índices de evasão e retenção escolar.



(a) IEE (b) IRE

Fonte: Elaborada pelo autor.

De acordo com a Figura 1 (a), é possível verificar a existência de componentes de tendência com características crescente e decrescente. Além disso, ao analisar a Figura 1 (b), foi possível identificar a existência de componentes com características exponenciais em relação aos valores observados.

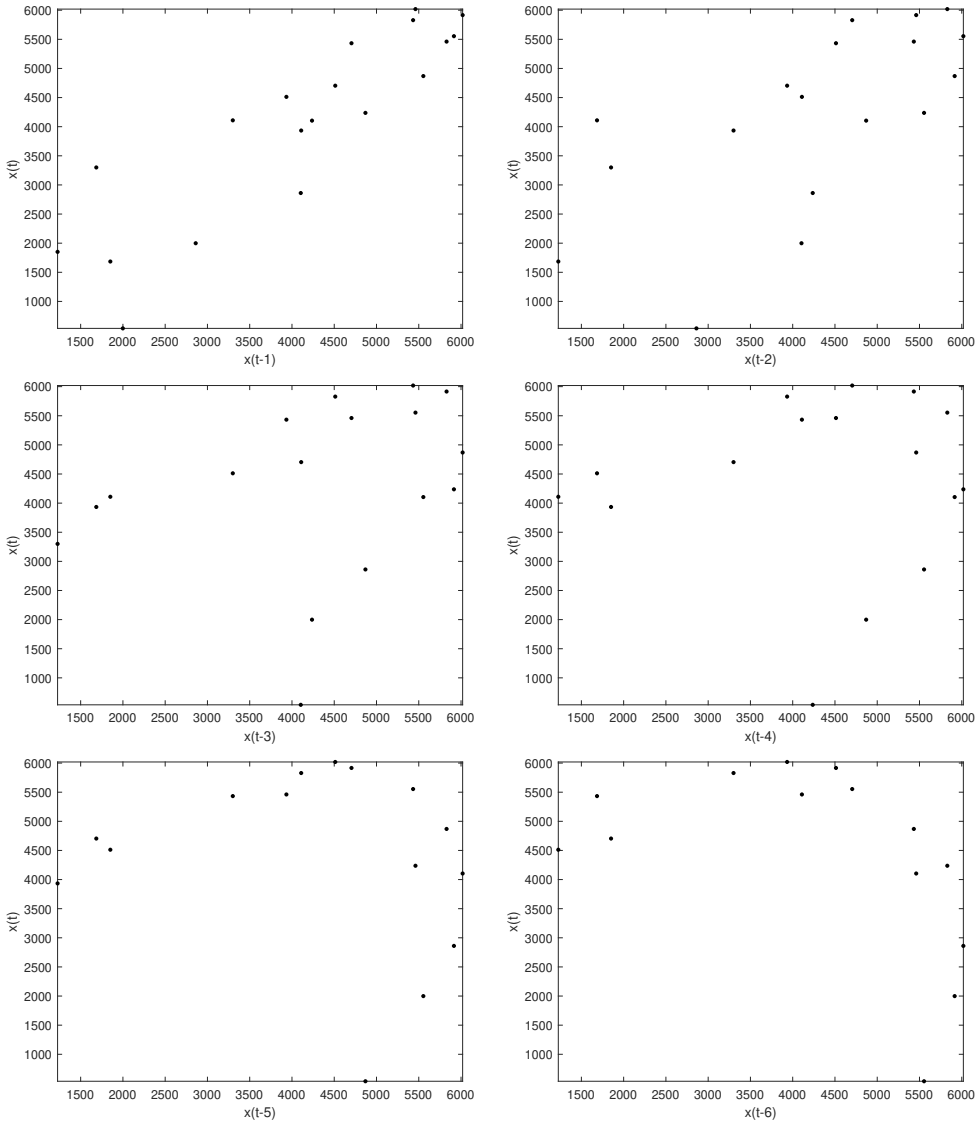
Devido ao fato do principal problema na caracterização do fenômeno gerador de uma série temporal ser, naturalmente, a escolha dos retardos temporais relevantes (dimensionalidade d), utiliza-se o gráfico *lagplot* (PERCIVAL; WALDEN, 1998; KANTZ; SCHREIBER, 2003) (apresentado na Figuras 2 e 3), que visa determinar e analisar as relações entre os retardos temporais das séries investigadas.

A partir da análise da Figuras 2 e 3, foi possível identificar estruturas que caracterizam a presença de relacionamento linear e não-linear em todas as séries investigadas. No entanto, como o *lagplot* é fortemente dependente da interpretação humana dos gráficos e, em alguns casos, as relações contidas nestes gráficos podem não refletir claramente as características do fenômeno gerador da série (a medida que a dimensionalidade n aumenta), outras técnicas devem ser consideradas. Neste contexto, a ACF, ilustrada na Figura 4, é utilizada para analisar o comportamento da componente linear.

Note que, de acordo com a Figura 4, a ACF das séries apresentam um característico decaimento hiperbólico, o que confirma a suposição da presença de dependência linear no fenômeno gerador destas séries, uma vez que é possível verificar altas correlações em retardos temporais de baixa ordem, bem como baixas correlações em retardos temporais de alta ordem. Entretanto, nada se pode observar em relação a natureza da componente não-linear a

ISSN 2237-1966

Figura 2: *Lagplot* da série IEE.



Fonte: Elaborada pelo autor.

partir da análise da ACF, uma vez que de acordo com (BOX; JENKINS; REINSEL, 1994) estas funções só podem ser utilizadas para análise da dependência linear presente no fenômeno gerador da série temporal. Assim, a MMI, (KRASKOV; STGBAUER; GRASSBERGER, 2004), ilustrada na Figura 5, é utilizada para analisar a componente não-linear.

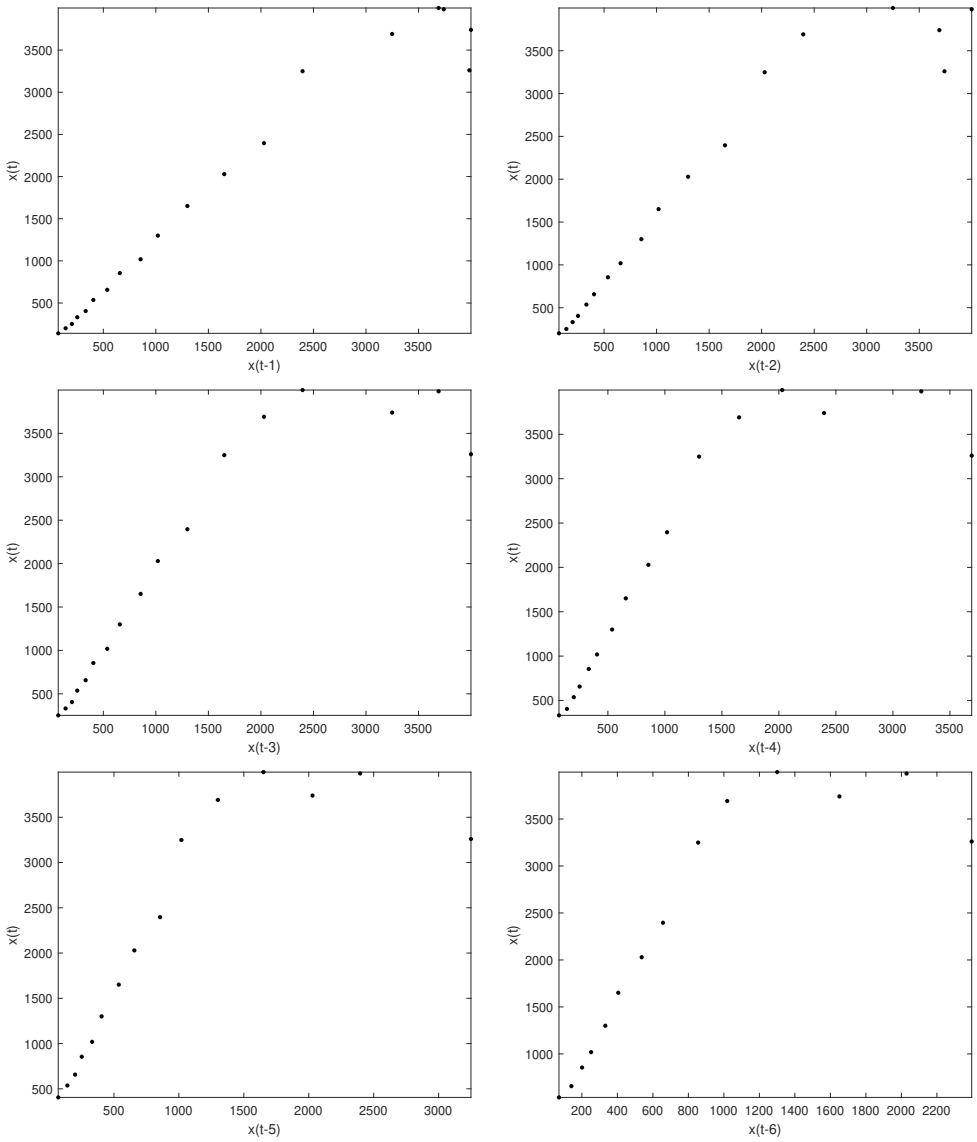
De acordo com a Figura 5, é possível verificar a existência de dependência não-linear em todas as séries investigadas (MMI > 0). Note-se que a inexistência de dependência não-linear implicaria em um valor nulo para a MMI.

3 O Modelo Proposto

Na literatura existem três classes de modelos para previsão de séries temporais (ARAÚJO, 2016): *i)* Univariados: utilização das observações de um único fenômeno gerador de interesse para realizar previsões, *ii)* Função de Transferência: utilização das observações de mais de um fenômeno gerador (que devem ser, obrigatoriamente, não-correlacionados) de interesse para realizar previsões, e *iii)* Multivariados: utilizam mais de um fenômeno gerador (não havendo nenhuma imposição no tocante a causalidade entre si) de interesse para realizar previsões. Apesar da diversidade de modelos encontrados na literatura, a sua escolha ainda representa um processo complexo

ISSN 2237-1966

Figura 3: *Lagplot* da série IRE.



Fonte: Elaborada pelo autor.

no problema de previsão de séries temporais, uma vez que descrever um dado fenômeno gerador depende de diversos fatores, como o conhecimento a *priori* das leis que governam a dinâmica do fenômeno (ARAÚJO, 2016).

Neste contexto, as redes neurais artificiais (*artificial neural networks*, ANNs) (HAYKIN, 1998) são consideradas uma alternativa para superar as limitações dos modelos de previsão clássicos, uma vez que ANNs são modelos não-lineares com baixo grau de complexidade matemática e computacional, quando comparadas aos modelos clássicos. Em uma ANN, a unidade fundamental de processamento da informação é conhecida como neurônio artificial, definida por (HAYKIN, 1998)

$$y = f(u), \tag{2}$$

com

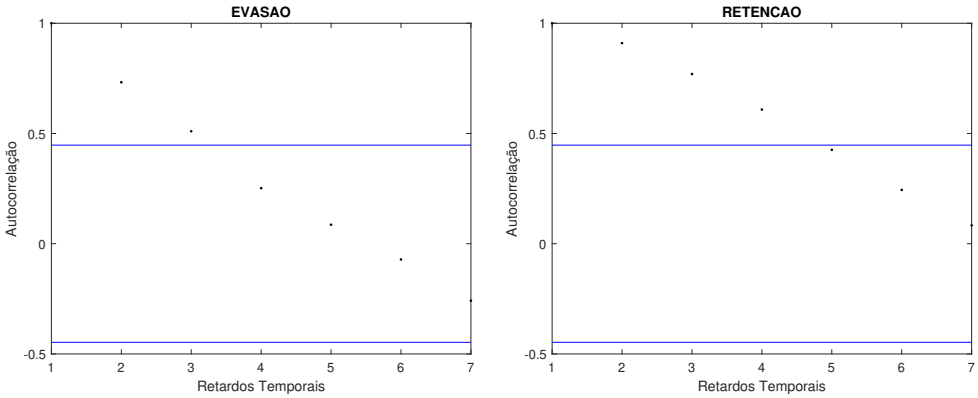
$$u = \sum_{j=1}^{J} w_j x_j + b, \tag{3}$$

em que J é a dimensionalidade do sinal de entrada, x_j com $j = 1, 2, \ldots, J$ é o sinal de entrada, w_j são os pesos sinápticos, u é o nível de ativação interna, $f(\cdot)$ é a função de ativação, b é o termo de *bias*, e y representa a ativação da saída do neurônio.

Periódico científico de acesso livre

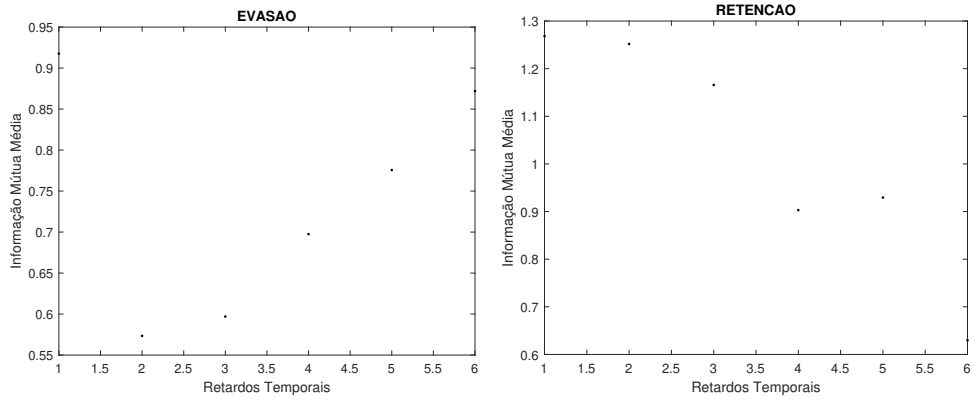
ISSN 2237-1966

Figura 4: ACF das séries temporais dos índices de evasão e retenção escolar.



Fonte: Elaborada pelo autor.

Figura 5: MMI das séries temporais dos índices de evasão e retenção escolar.



Fonte: Elaborada pelo autor.

Alternativamente, o neurônio pode ser definido em função da notação vetorial. Seja $\mathbf{x} = (x_1, x_2, \ldots, x_J)$ o vetor que representa o sinal de entrada, $\mathbf{w} = (w_1, w_2, \ldots, w_J)$ o vetor que representa os pesos sinápticos do neurônio e b um escalar (*bias*). Portanto, a saída do neurônio é definida por (HAYKIN, 1998)

$$y(\mathbf{w}, \mathbf{x}, b) = f(\mathbf{w}^T \mathbf{x} + b), \tag{4}$$

em que $\cdot^T$ é uma operação de transposição.

Neste sentido, ANNs têm apresentado desempenho expressivo na tarefa de aproximar o fenômeno gerador de séries temporais (ARAÚJO, 2016). Neste contexto, é possível encontrar uma série de ANNs propostas na literatura para solucionar o problema de previsão de séries temporais (ARAÚJO, 2016). Dentre elas, vale destacar que o modelo mais difundido é o *perceptron* multicamadas e, por esta razão, este será escolhido para representar o modelo proposto.

3.1 *Perceptron* Multicamadas

O *perceptron* multicamadas (*multilayer perceptron*, MLP) (HAYKIN, 1998), é uma rede neural com arquitetura em camadas, onde os neurônios são dispostos em uma ou mais camadas de processamento, sendo a ANN mais frequentemente encontrada na literatura de previsão de séries temporais (ARAÚJO, 2016).

O modelo de ANN do tipo MLP com melhor desempenho para previsão de séries temporais reportado na literatura utiliza função de ativação sigmóide logística (Equação 6) para todas as unidades de processamento escondidas (ARAÚJO, 2016). A unidade de processamento de saída utiliza função de ativação linear com seu bias

ISSN 2237-1966

passando por função sigmóide logística (ARAÚJO, 2016). Portanto, a saída da rede MLP é dada por:

$$y_k(t) = \sum_{j=1}^{n_h} W_{jk} Sig\left[\sum_{i=1}^{n_{in}} W_{ij} x_i(t) + b_j^1\right] + Sig(b_k^2), \tag{5}$$

onde $x_i(t)$ $(i = 1, 2, \ldots, n_{in})$ são os valores de entrada da rede MLP (retardos temporais), n_{in} e n_h são a quantidade de entradas da rede MLP e a quantidade de unidades de processamento na camada escondida, respectivamente. Como a previsão pretendida é de um-passo-adiante, utiliza-se apenas uma unidade de processamento na camada de saída ($k = 1$). O termo $Sig(\cdot)$ é uma função sigmóide logística definida por:

$$Sig(x) = \frac{1}{1 + exp(-x)}. \tag{6}$$

De acordo com Haykin (HAYKIN, 1998), a propriedade mais relevante de uma rede MLP é sua capacidade de aprendizagem através de um processo iterativo de ajustes aplicados aos seus pesos sinápticos e *bias*. O processo de aprendizagem de uma rede MLP é do tipo supervisionado. Este tipo de aprendizado é caracterizado pela presença de um agente externo que induz a rede MLP a uma resposta desejada a um determinado estímulo apresentado pelo ambiente, de forma a conseguir realizar o mapeamento entre a entrada e saída desejada, através da minimização de uma função de custo f, de modo que a resposta observada se aproxime da resposta desejada a cada iteração, definida como época, no processo de aprendizagem (HAYKIN, 1998).

A função de custo f define uma superfície de erro sobre o espaço de pesos (HAYKIN, 1998). Se P representa a dimensionalidade dos vetor de pesos ajustáveis na rede neural e N representa a dimensionalidade do padrão de saída do problema, então $f : \mathbb{R}^P \rightarrow \mathbb{R}^N$. Nesta superfície, tipicamente tem-se a presença de mínimos locais e globais (HAYKIN, 1998). Os métodos de otimização tipicamente utilizados para minimizar a função f utilizam informações do gradiente descendente do erro para ajustar os parâmetros da rede. Teoricamente tais métodos sempre encontram pontos de mínimo (local ou global) na superfície de erro a partir de uma condição inicial arbitrária (HAYKIN, 1998). O método clássico utilizado no processo de aprendizagem de redes neurais MLP, que utiliza informações do gradiente descendente do erro, é o algoritmo de retro-propagação do erro (*back-propagation*, BP) (HAYKIN, 1998).

4 Simulações e Resultados Experimentais

Ambas as séries temporais investigadas devem passar por um processo de normalização (etapa de pré-processamento (ZHANG; PATUWO; HU, 1998)). O principal objetivo da etapa é prover conformidade, em termos de domínio, entre os valores da série temporal e os valores gerados pelo modelo de previsão. Zhang (ZHANG; PATUWO; HU, 1998) discute diversas maneiras para realizar a normalização dos dados. Neste trabalho, utilizou-se a normalização linear para o intervalo $[0, 1]$, uma vez que torna possível a utilização de todo o domínio de atuação do modelo de redes neurais investigado, sendo definida por

$$xn_i = \frac{x_i - min(\mathbf{x})}{max(\mathbf{x}) - min(\mathbf{x})}. \tag{7}$$

em que x_i e xn_i com $i = 1, 2, \ldots, I$ são os valores reais e normalizados, respectivamente, da série temporal, e $min(\cdot)$ e $max(\cdot)$ são as operações de mínimo e máximo, respectivamente, de um arranjo de elementos.

Após a etapa de normalização, cada uma das séries temporais foi dividida em dois conjuntos, de acordo com Prechelt (PRECHELT, 1994) (padronização da divisão do conjunto de dados em problemas de classificação e previsão): i) conjunto de treinamento (utilizado no processo de aprendizagem do modelo de previsão) e ii) conjunto de teste (utilizado para confirmar o desempenho prático do modelo de previsão). Para definição da cardinalidade de cada um destes conjuntos, também foi utilizado o conjunto de regras apresentado em (PRECHELT, 1994), onde foi definido 90% dos dados para o conjunto de treinamento e 10% dos dados para o conjunto de teste.

Tendo em vista comparar o desempenho preditivo do modelo proposto, foi utilizado o modelo estatístico de Box *et al.* (BOX; JENKINS; REINSEL, 1994) (ARIMA), uma vez que este é uma das escolhas mais comuns dentre as técnicas apresentadas na literatura de previsão de séries temporais. Para realização dos experimentos com o modelo ARIMA$(p; q; d)$ foi utilizado o termo de diferenciação $d = 1$ como sugerido por Box (BOX; JENKINS; REINSEL, 1994).

Para realização dos experimentos com o modelo de rede neurais proposto, foi necessário definir uma arquitetura básica para todos os experimentos, que consiste em uma MLP de três camadas (uma camada de entrada, uma camada intermediária e uma camada de saída), formalmente descrita utilizando a notação MLP(I;H;O), onde I representa a camada de entrada, H representa a quantidade de unidades de processamento na camada intermediária, e O representa a quantidade de unidades de processamento na camada de saída.

ISSN 2237-1966

A camada de entrada é definida pela quantidade de retardos temporais utilizados para a descrição da série temporal. Para a definição dos retardos temporais foi utilizada uma metodologia empírica de acordo com a análise apresentada na Seção 2, a partir dos quais foram escolhidos os valores 1-3 (IEE) e 1-5 (IRE). A quantidade de unidades de processamento na camada escondida foi determinada empiricamente através de uma série de experimentos, a partir dos quais foram escolhidos os valores 1, 5, 10, 25 e 50. A quantidade de unidades de processamento na camada de saída foi fixada em 1, uma vez que este trabalho foca apenas em previsões de um-passo-adiante, isto é, com horizonte de previsão unitário ($h = 1$). Em termos de arquitetura do modelo MLP, foi utilizada função de ativação sigmóide logística para todas as unidades de processamento escondidas, e para unidade de processamento de saída foi utilizada a função de ativação linear, uma vez que esta arquitetura possui o melhor desempenho para previsão das séries temporais investigadas.

Para treinamento da rede, foi utilizado o algoritmo de retro-propagação do erro (*back-propagation*, BP) (HAYKIN, 1998), utilizando os seguintes critérios de parada (PRECHELT, 1994): i) A quantidade máxima de épocas de treinamento (10^4), ii) O aumento no erro de validação ou *generalization loss* ($Gl > 5\%$), e iii) A queda no erro de treinamento ou *process training* ($Pt \leq 10^{-6}$). Foram realizadas cinquenta execuções distintas para cada configuração investigada, tendo em vista se obter um comportamento médio do modelo MLP. O experimento que obtiver o melhor desempenho no conjunto de treinamento será eleito como representante do modelo MLP.

Para avaliação dos resultados obtidos são utilizadas duas medidas de desempenho relevantes na literatura (CLEMENTS; FRANSES; SWANSON, 2004). A principal e mais utilizada medida para avaliação da previsão é o erro médio quadrático (*mean squared error*, MSE), dada por (CLEMENTS; HENDRY, 1993)

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N}(e_j)^2, \tag{8}$$

onde N é a quantidade de padrões, e e_j é o erro instantâneo para o padrão j, que é definido por

$$e_j = x_j - \widehat{x}_j \tag{9}$$

em que x_j e $\widehat{x}_j$ representam, respectivamente, o valor real e previsto da série temporal no tempo j. Note que, em um modelo de previsão ideal, MSE $\rightarrow 0$.

Vale mencionar que a medida MSE é frequentemente utilizada no processo de aprendizagem de modelos de previsão. Entretanto, esta não pode ser considerada como uma medida conclusiva em uma análise comparativa entre diversos modelos de previsão (CLEMENTS; HENDRY, 1993). Por esta razão, outra medidas devem ser consideradas para permitir uma avaliação mais apurada do desempenho de previsão. Nesse contexto, o erro médio percentual absoluto (*mean absolute percentage error*, MAPE) é uma medida que permite identificar precisamente os desvios percentuais do modelo de previsão (note que em um modelo de previsão ideal, MAPE $\rightarrow 0$), dada por (CLEMENTS; HENDRY, 1993)

$$\text{MAPE} = \frac{1}{N}\sum_{j=1}^{N}\left|\frac{e_j}{x_j}\right|. \tag{10}$$

A seguir será apresentada uma análise comparativa entre o modelo ARIMA e o modelo proposto (MLP) a partir das medidas de desempenho previamente definidas. Foram calculadas a média e o desvio padrão para cada medida investigada. Além disso, a fim de validar estatísticamente o modelo proposto, foi aplicado o teste de Friedman com nível de significância de $\alpha = 0.05$ e o teste de Tukey com $\alpha = 0.05$.

4.1 Análise da Medida MSE

Na Tabela 1, são apresentados os resultados obtidos para as séries temporais IEE e IRE, considerando as estatísticas média e desvio padrão, bem como os resultados do teste de Friedman e de Tukey para a medida MSE.

Tabela 1: Desempenho de teste para a medida MSE.

	Medida MSE		Teste de Friedman		Teste de Tukey	
Modelo	Série IEE	Série IRE	Posição	Posto	Estatística	p-valor
MLP	0.0002 ±0.0003	0.0027 ±0.0020	1	1.0		
ARIMA	0.0091 ±0.0000	0.2413 ±0.0000	2	2.0	-1.0	1.57e-01

De acordo com a Tabela 1 é possível verificar que o modelo proposto obteve melhor desempenho preditivo, considerando a medida MSE, para ambas as séries temporais investigadas neste trabalho. Os valores para a medida

Periódico científico de acesso livre

ISSN 2237-1966

MSE no intervalo [3.E-4,8E-3] indicam que as previsões geradas pelo modelo proposto estão bastante próximas aos valores reais da série temporal. De acordo com os resultados do Teste de Friedman, é possível confirmar, estatisticamente, os resultados apresentados na Tabela 1. Além disso, note que o modelo proposto alcançou o menor valor de posto para o teste, sugerindo que este pode ser considerado o melhor modelo de previsão para as séries temporais IEE e IRE, considerando a medida MSE. Por fim, é possível notar que o maior valor para o Teste de Tukey para o par MLP-ARIMA é -1.00, sugerindo que o modelo proposto tem um desempenho de previsão estatisticamente superior ao modelo ARIMA.

4.2 Análise da Medida MAPE

A Tabela 2 apresenta os resultados alcançados, levando em consideração as estatísticas média e desvio padrão, para as séries temporais IEE e IRE, bem como os resultados do teste de Friedman e de Tukey para a medida MAPE.

Tabela 2: Desempenho de teste para a medida MAPE.

	Medida MAPE		Teste de Friedman		Teste de Tukey	
Modelo	Série IEE	Série IRE	Posição	Posto	Estatística	*p*-valor
MLP	0.0344 ±0.0268	0.0635 ±0.0287	1	1.0		
ARIMA	0.2112 ±0.0000	0.6129 ±0.0000	2	2.0	-1.0	1.57e-01

Note que os resultados apresentados na Tabela 2 sugerem que o modelo proposto obteve melhor desempenho preditivo, considerando a medida MAPE, para ambas as séries temporais investigadas neste trabalho. Os valores para a medida MAPE no intervalo [3.5E-2,6.4E-1] indicam que as previsões geradas têm um desvio percentual relativamente baixo, variando entre 0.03% a 0.06%. Novamente, o Teste de Friedman pôde confirmar, estatisticamente, os resultados apresentados na Tabela 2, onde o modelo proposto obteve o menor valor de posto, levantando a hipótese deste ser o melhor modelo de previsão para as séries temporais IEE e IRE, considerando a medida MAPE. Note que o maior valor para o Teste de Tukey para o par MLP-ARIMA é -1.00, sugerindo que o modelo proposto tem desempenho, considerando a medida MAPE, estatisticamente superior ao modelo ARIMA.

4.3 Análise do Comportamento da Previsão

A Figura 6 apresenta uma análise comparativa entre os valores reais e as previsões geradas pelo modelo MLP e ARIMA para as séries temporais IEE e IRE. Note que em ambas as séries, a previsão gerada pelo modelo MLP está mais precisa quando comparado à previsão gerada pelo modelo ARIMA. No caso particular da série IEE, a previsão está quase sobreposta ao valor real da série. Tal fato sugere que o modelo MLP é capaz de reproduzir o fenômeno gerador das séries temporais investigadas neste trabalho, sendo uma opção viável para prever com eficácia tais fenômenos temporais.

Os resultados apresentados nas seções anteriores deram suporte a hipótese do modelo MLP ter alto desempenho preditivo, quando comparado ao modelo ARIMA, e poder ser utilizado, na prática, para prever séries temporais de índices de evasão e retenção escolar. Também foi possível confirmar o alto poder de generalização do mapeamento gerado pelo modelo MLP para prever esse tipo particular de série temporal, considerando as medidas de desempenho MSE e MAPE. Note que o processo de aprendizagem utilizado no modelo MLP foi capaz de convergir para pontos de ótimo na superfície do erro, uma vez que foram alcançados valores próximos de 0 para ambas as medidas MSE e MAPE. Neste sentido, o teste de Friedman e o teste de Tukey forneceram a base estatística para confirmar o desempenho preditivo superior do modelo MLP.

5 Conclusões

Este trabalho apresentou um estudo sobre o fenômeno gerador de séries temporais de Índices de Evasão e Retenção escolar. Estas séries são compostas por observações semestrais relacionadas ao quantitativo de alunos evadidos e retidos do IFCE no período de 2009 a 2018. A análise do *lagplot* destas séries permitiu a identificação de estruturas que caracterizaram a presença de relacionamento linear e não-linear em seus retardos temporais. No entanto, como o *lagplot* é fortemente dependente da interpretação humana, uma vez que as relações contidas nestes gráficos podem não refletir claramente as características do fenômeno gerador da série, foi empregada a função de autocorrelação, que confirmou a existência de dependência linear (devido ao característico decaimento encontrado nos gráficos, isto é, altos índices de correlação em retardos temporais de baixa ordem e baixos índices de correlação em retardos

ISSN 2237-1966

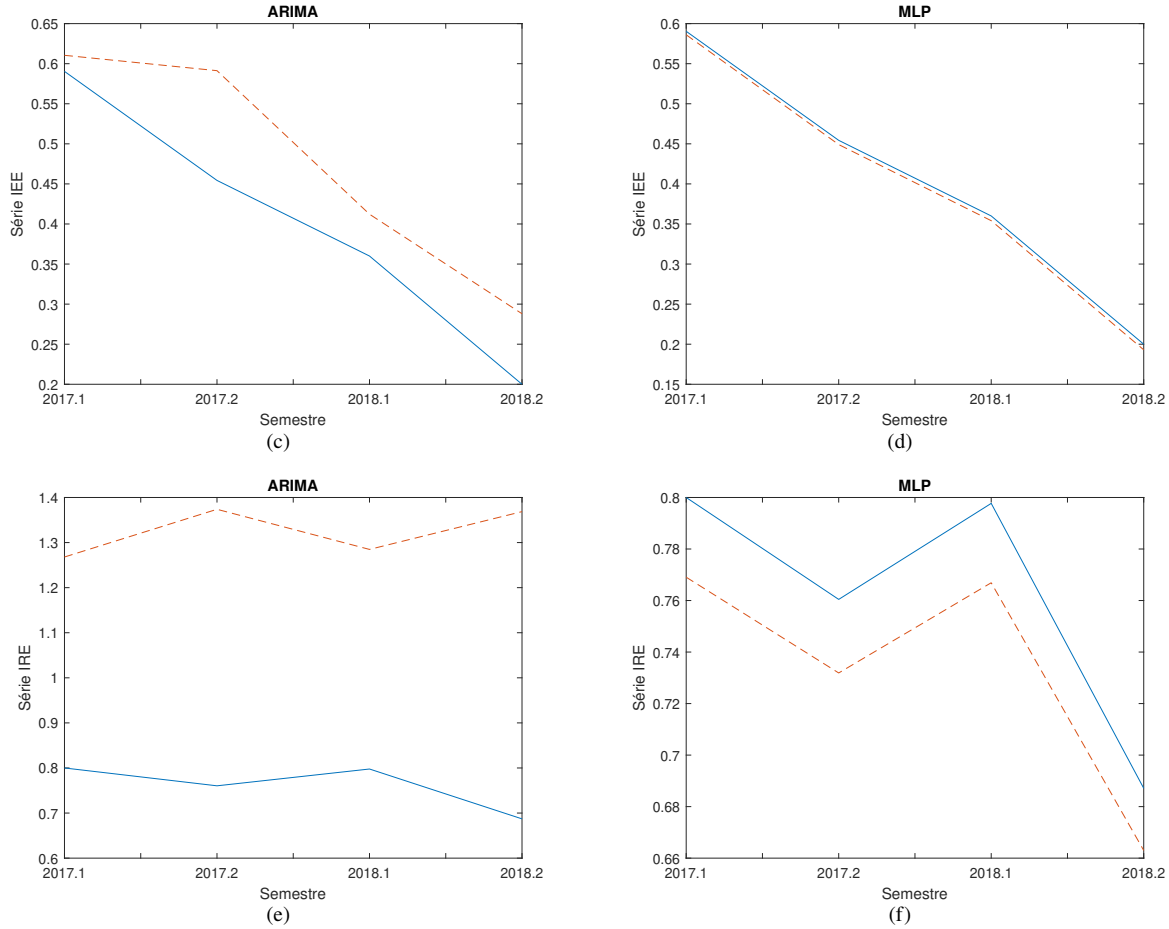


Figura 6: Gráfico de previsão (conjunto de teste - semestre 2017.1 ao semestre 2018.2): linha sólida azul (valor real) e a linha vermelha tracejada (valor previsto): a) Série IEE - Modelo ARIMA, b) Série IEE - Modelo MLP, c) Série IRE - Modelo ARIMA, d) Série IRE - Modelo MLP.

temporais de alta ordem). De maneira análoga, a informação mútua média também pôde confirmar a presença de dependência não-linear (devido a curva do gráfico ter valores superiores a 0).

Baseado nas evidências encontradas na análise do fenômeno gerador destas séries temporais, este trabalho apresentou um modelo de previsão, referido como rede neural artificial *perceptron* multicamadas (MLP), capaz de estimar, no futuro, índices de evasão e retenção escolar. A sua escolha foi baseada em estudos que demonstraram a capacidade acurada deste tipo de modelo aproximar as características encontradas na análise das séries temporais realizada neste trabalho. Para o projeto do modelo proposto, foi investigado um método baseado em gradiente descendente utilizando o algoritmo de retropropagação do erro.

Para se estabelecer um nível de referência para o desempenho preditivo, foram realizados experimentos com um modelo estatístico comumente empregado para previsão de séries temporais (ARIMA). Além disso, para avaliação do desempenho preditivo, foram investigadas duas de medidas de desempenho com características distintas (MSE e MAPE). Para cada configuração estudada com os modelos investigados neste trabalho foram realizados cinquenta experimentos e, para cada medida de desempenho, foi calculada a média e o desvio padrão dos resultados para se ter noção do comportamento médio do modelo.

A análise dos resultados obtidos revelou que o modelo proposto obteve desempenho preditivo estatísticamente superior ao modelo ARIMA, sob as mesmas condições de experimentação, para ambas as medidas de desempenho analisadas. Além do desempenho preditivo mais acurado, uma vantagem do modelo proposto é a sua capacidade de reproduzir o fenômeno gerador de índices de evasão e retenção escolar, o que possíbilitará seu uso na prática em outras instituições de ensino. Portanto, pode-se concluir que o modelo proposto é viável, em termos de desempenho preditivo, para previsão de índices de evasão e retenção escolar.

Embora o modelo proposto tenha alcançado desempenho preditivo expressivo, existem algumas questões que ainda necessitam ser investigadas como trabalhos futuros. A formalização e uma investigação mais detalhada sobre as propriedades do modelo proposto deve ser realizada visando determinar as limitações práticas e teóricas em

ISSN 2237-1966

outras séries temporais de índices de evasão e retenção escolar, provenientes de outras instituições de ensino, bem como a realização de um estudo particular sobre a complexidade computacional do modelo e de seu processo de aprendizagem para se estabelecer uma avaliação completa em termos de custo-benefício. Além disso, a investigação de sistemas híbridos deve ser considerada, uma vez que um ponto crucial para o desempenho preditivo é a otimização dos retardos temporais e dos parâmetros do modelo de previsão.

Referências

AHMED, A. B. E. D.; ELARABY, I. S. Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology*, v. 2, n. 2, p. 43–47, 2014.

ARAÚJO, R. de A. *Mercado de Ações Brasileiro em Alta-Frequência: Evidências de sua Previsibilidade com Modelagem Morfológica-Linear*. Tese (Doutorado) — Universidade Federal de Pernambuco, 2016.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de dados educacionais: Oportunidades para o brasil. *Brazilian Journal of Computers in Education*, v. 19, n. 02, 2011.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Analysis: Forecasting and Control*. Third. New Jersey: Prentice Hall, 1994.

CLEMENTS, M. P.; FRANSES, P. H.; SWANSON, N. R. Forecasting economic and financial time-series with non-linear models. *International Journal of Forecasting*, v. 20, p. 169–183, 2004.

CLEMENTS, M. P.; HENDRY, D. F. On the limitations of comparing mean square forecast errors. *Journal of Forecasting*, v. 12, n. 8, p. 617–637, Dec. 1993.

COCCO, E. M.; SUDBRACK, E. M. Ensino médio no contexto atual e os desafios de acesso e permanência. *Impulso*, v. 26, n. 67, p. 7–22, 2016.

CUNHA, D. M. et al. *Formação-profissionalização de professores e formação profissional e tecnológica: fundamentos e reflexões contemporâneas*. [S.l.]: PUC Minas Gerais, 2013.

CUNHA, J. A. da; MOURA, E.; ANALIDE, C. Data mining in academic databases to detect behaviors of students related to school dropout and disapproval. In: *New Advances in Information Systems and Technologies*. [S.l.: s.n.], 2016. p. 189–198.

DORE, R.; ARAUJO, A. D. de; MENDES, J. de S. *Evasão na educação: estudos, políticas e propostas de enfrentamento*. [S.l.]: Editora do IFB/RIMEPES, 2014.

DORE, R.; LUSCHER, A. Educação profissional e evasão escolar. In: *Encontro Internacional de Pesquisadores de Políticas Educativas*. [S.l.: s.n.], 2008. p. 197–203.

HAYKIN, S. *Neural networks: A comprehensive foundation*. New Jersey: Prentice Hall, 1998.

HAYKIN, S. *Neural Networks and Learning Machines*. Canada: McMaster University, 2007.

JAISWAL, A. S. G.; YADAV, S. K. Analytical approach for predicting dropouts in higher education. *International Journal of Information and Communication Technology Education*, v. 3, n. 15, p. 1–14, 2019.

JUNIOR, F. T.; SANTOS, J. R. dos; MACIEL, M. de S. Análise da evasão no sistema educacional brasileiro. *Revista Pesquisa e Debate em Educação*, v. 6, n. 1, p. 73–92, 2017.

JUNIOR, J. G. de O. *Identificação de Padrões para Análise da Evasão em Cursos de Graduação Usando Mineração de Dados Educacionais*. Dissertação (Mestrado) — Universidade Tecnológica Federal do Paraná, 2015.

KABRA, R. R.; BICHKAR, R. Performance prediction of engineering students using decision trees. *International Journal of Computer Application*, v. 36, n. 11, p. 1–12, 2011.

KANTZ, H.; SCHREIBER, T. *Nonlinear Time Series analysis*. Second. New York, NY, USA: Cambridge University Press, 2003.

KAWASE, K. H. F. *Aplicação de Redes Neurais RBF e MLP na Análise de Evasão Discente do Curso de Sistemas de Informação da UFRRJ*. Dissertação (Mestrado) — Universidade Federal Rural do Rio de Janeiro, 2015.

KIM, Y. S. Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. *Expert Systems with Applications*, v. 34, n. 2, p. 1227 – 1234, 2008.

KRASKOV, A.; STGBAUER, H.; GRASSBERGER, P. A new auto-associative memory based on lattice algebra. *Phys. Rev. E*, v. 69, n. 6, 2004.

MARQUEZ-VERA, C.; ROMERO, C.; VENTURA, S. Predicting school failure and dropout by using data mining techniques. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, v. 8, n. 1, p. 7–14, 2013.

MARTINHO, V. R. C.; NUNES, C.; MINUSSI, C. R. An intelligent system for prediction of school dropout risk group in higher education classroom based on artificial neural networks. *IEEE International Conference on Tools with Artificial Intelligence*, 2013.

MARTINHO, V. R. C.; NUNES, C.; MINUSSI, C. R. Prediction of school dropout risk group using neural network. *2013 Federated Conference on Computer Science and Information Systems*, p. 111–114, 2013.

MEEDECH, P.; IAM-ON, N.; BOONGOEN, T. Prediction of student dropout using personal profile and data mining approach. In: *Intelligent and Evolutionary Systems*. Cham: Springer International Publishing, 2016. p. 143–155.

NASCIMENTO, R. L. S. do et al. Educational data mining: An application of regressors in predicting school dropout. In: *Machine Learning and Data Mining in Pattern Recognition*. Cham: Springer International Publishing, 2018. p. 246–257.

OLIVEIRA, D. A. As políticas para o ensino médio na realidade brasileira: uma agenda em disputa. *Poiésis-Revista do Programa de Pós-Graduação em Educação*, v. 10, n. 17, p. 187–198, 2016.

PERCIVAL, D. B.; WALDEN, A. T. *Spectral Analysis for Physical Applications – Multitaper and Conventional Univariate Techniques*. New York: Cambridge University Press, 1998. ISBN 0-521-43541-2.

PRECHELT, L. *Proben1: A set of Neural Network Benchmark Problems and Benchmarking Rules*. [S.l.], 1994.

REBELO, J. A. S. Efeitos da retenção escolar, segundo os estudos científicos, e orientações para uma intervenção eficaz: Uma revisão. *Revista portuguesa de pedagogia*, v. 43, n. 1, p. 27–52, 2009.

ROMERO, C.; VENTURA, S. Data mining in education. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, v. 3, n. 1, p. 12–27, 2013.

RUMBERGER, R.; THOMAS, S. The distribution of dropout and turnover rates among urban and suburban high schools. *Sociology of Education*, v. 73, n. 1, p. 39–67, 2000.

SILVA, J.; DIAS, P. C.; SILVA, M. C. Evasão escolar em cursos técnicos do instituto federal de educação, ciência e tecnologia de brasília: perfil socioeconômico de estudantes de cursos técnicos subsequentes do campus brasilia. *Revista da UIIPS*, v. 3, n. 6, p. 279–293, 2015.

TROMBONI, J.; OLEGARIO, F.; LAROQUE, L. F. S. As políticas para o ensino médio na realidade brasileira: uma agenda em disputa. *Revista Intersabes*, v. 12, n. 25, p. 144–151, 2017.

VIADERO, D. The dropout dilemma: Research hindered by lack of uniform way to count students who quit school. *Education Week*, v. 20, n. 21, p. 26–29, 2001.

YASMIN, D. Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. *Distance Education*, v. 34, n. 2, p. 218–231, 2013.

YU, C. H. et al. A data mining approach for identifying predictors of student retention from sophomore to junior year. *Journal of Data Science*, v. 2, n. 8, p. 307–325, 2010.

ZHANG, G.; PATUWO, B. E.; HU, M. Y. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, v. 14, p. 35–62, 1998.

ISSN 2237-1966

Periódico científico de acesso livre