INSTITUTO FEDERAL DO SERTÃO PERNAMBUCANO
CAMPUS SALGUEIRO
PÓS-GRADUAÇÃO EM EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

FRANCICLEIDE GEREMIAS DA COSTA SOUZA

PREVISÃO DE EVASÃO E RETENÇÃO ESCOLAR NO ENSINO MÉDIO PROFISSIONAL: UMA ABORDAGEM BASEADA EM REDES NEURAIS ARTIFICIAIS

Dissertação de Mestrado

SALGUEIRO
2021

FRANCICLEIDE GEREMIAS DA COSTA SOUZA

PREVISÃO DE EVASÃO E RETENÇÃO ESCOLAR NO ENSINO MÉDIO PROFISSIONAL: UMA ABORDAGEM BASEADA EM REDES NEURAIS ARTIFICIAIS

Trabalho apresentado ao Programa de Pós-graduação em Educação Profissional e Tecnológica ofertado pelo Campus Salgueiro do Instituto Federal do Sertão Pernambucano como requisito parcial para obtenção do grau de Mestre em Educação Profissional e Tecnológica.

Orientador: *Ricardo de Andrade Araújo*

SALGUEIRO
2021

FICHA

INSTITUTO FEDERAL DO SERTÃO PERNAMBUCANO
Autarquia criada pela Lei n 11.892 de 29 de Dezembro de 2008

PROGRAMA DE PÓS-GRADUAÇÃO EM EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

FRANCICLEIDE GEREMIAS DA COSTA SOUZA

PREVISÃO DE EVASÃO E RETENÇÃO ESCOLAR NO ENSINO MÉDIO PROFISSIONAL: UMA ABORDAGEM BASEADA EM REDES NEURAIS ARTIFICIAIS

Trabalho apresentado ao Programa de Pós-graduação em Educação Profissional e Tecnológica ofertado pelo Campus Salgueiro do Instituto Federal do Sertão Pernambucano como requisito parcial para obtenção do grau de Mestre em Educação Profissional e Tecnológica.

Aprovado em 18 de Fevereiro de 2021.

COMISSÃO EXAMINADORA

Prof. Dr. Ricardo de Andrade Araújo
Instituto Federal do Sertão Pernambucano
Orientador

Prof. Dr. Francisco Kelsen de Oliveira
Instituto Federal do Sertão Pernambucano

Prof. Dr. Adelson Dias de Oliveira
Universidade Federal do Vale do São Francisco

A Minha filha, Maria Eduarda, Meu Mundo singelo de ternura e Paz! Por fazer meus dias mais felizes...

AGRADECIMENTOS

A Deus, pela força para vencer as lutas diárias e por tantos momentos de aprendizado a qual o Mestrado em Educação Profissional e Tecnológica - Profept tem me proporcionado. A coragem e persistência se sobressairão em meio as dificuldades pra seguir em frente na busca por objetivos e sonhos almejados.

Ao meu orientador, Prof. Ricardo Araújo, agradeço pelo acompanhamento, pela oportunidade de novos conhecimentos e crescimento dentro da pesquisa científica, pelo aprendizado, estímulo e parceria. Obrigada!

A minha filha, Maria Eduarda Geremias de Souza, pela presença, alegria, sorriso, pelo amor presente e acolhedor, me trazendo alegria em dias turbulentos. Ao meu esposo Claudiomar Cicero de Souza pela compreensão e companheirismo.

Aos meus pais, Maria Francisca da Costa e Francisco Geremias da Costa, por não me fazer desanimar, pelas mãos estendidas sempre, pelos sim em momentos esperados ou inesperados. Aos meus irmãos, companheiros de caminhada, sempre dispostos a ouvir e acolher.

Ao Instituto Federal de Educação do Ceará (IFCE), Campus Crateús, nas pessoas da gestão de ensino, a qual se despuseram em compreender esse momento na qual em meio ao trabalho e também inserida como mestranda na pesquisa científica. Obrigada pelo apoio!

Aos colegas de aula e professores pelo apoio, em especial a Yane Ferreira Machado parceira de caminhada, seguindo na busca de vencer todo e qualquer desafio.

A todos aqueles que durante esse período se fizeram presente a esse processo construtivo. Muito Obrigada!

Ontem um menino
Que brincava me falou
Hoje é a semente do amanhã
Para não ter medo
Que este tempo vai passar
Não se desespere, nem pare de sonhar
Nunca se entregue
Nasça sempre com as manhãs
Deixe a luz do sol brilhar no céu do seu olhar
Fé na vida, fé no homem, fé no que virá
Nós podemos tudo, nós podemos mais
Vamos lá fazer o que será..

—GONZAGUINHA (1945-1991)

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 artificiais, do tipo multicamadas, para prever este tipo particular de série temporal. Para o processo de aprendizagem, foi utilizado o algoritmo de retropropagação do erro (*back propagation*, BP). Uma análise experimental é conduzida com a abordagem proposta utilizando séries temporais relacionadas aos índices de evasão e retenção do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), no período de 2009 a 2018. Nestes experimentos, são utilizadas as medidas erro médio quadrático (*mean squared error*, MSE) e erro médio absoluto percentual (*mean absolute percentage error*, MAPE) 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 Escolar. Retenção Escolar. Séries Temporais. Redes Neurais Artificiais. Algoritmo de Retropropagação.

ABSTRACT

This work presents a study about time series, related to rates of evasion and retention in high school, aiming to identify peculiar characteristics of these series and, based on such study, to propose an approach based on artificial 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 Education, Science and Technology of *Ceará* (IFCE), from 2009 to 2018. In these experiments, both measures the mean squared error (MSE) and the mean absolute percentage error (MAPE) 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 are able to efficiently predict these series within evaluated period, being feasible options for the prediction of evasion and retention rates in high school institutions.

Keywords: School Evasion. School Retention. Time Series. Artificial Neural Networks. Back Propagation Algorithm.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACF	Função de autocorrelação (*autocorrelation function*)
ANN,	Redes neurais artificiais (*artificial neural networks*)
AR	Autorregressivo (*autoregressive*)
ARIMA	Autoregressivo integrado de médias móveis (*Autoregressive integrated moving average*)
ARMA	Autorregressivo de médias móveis (*autoregressive moving average*)
BP	Retro-propagação do erro (*back-propagation*)
DM	Mineração de dados (*data mining*)
DT	Árvores de decisão (*decision trees*)
EDM	Mineração de dados educacionais (*educational data mining*)
EPT	Educação Profissional e Tecnológica
GP	Programação genética (*genetic programming*)
IEE	Índice de Evasão Escolar
IFCE	Instituto Federal de Educação, Ciência e Tecnologia do Ceará
IRE	Índice de Retenção Escolar
LM	*levenberg-marquardt*
MA	Médias móveis (*moving average*)
MAPE	Erro médio percentual absoluto (*mean absolute percentage error*)
ML	Aprendizagem de máquina (*machine learning*)
MMI	informação mútua média (*mean mutual information*)
MLP	*Perceptron* multicamadas (*multilayer perceptron*)
MSE	Erro médio quadrático (*mean squared error*)
OSSCG	gradiente conjugado de um passo secante (*one step secant conjugate gradient*)
PACF	Função de autocorrelação parcial (*partial autocorrelation function*)
QUICKPROP	*quick-propagation*

RFEPT	Rede Federal de Educação Profissional e Tecnológica
RPROP	*resilient back-propagation*
SCG	gradiente conjugado escalar (*scaled conjugate gradient*)
SVR	Regressor de vetor de suporte (*support vector regressor*)
TL	Retardos temporais (*time lags*)

SUMÁRIO

1

INTRODUÇÃO

Este capítulo apresenta o problema de evasão e retenção de alunos no ensino médio profissional, o contexto da educação profissional no Brasil e as políticas públicas de educação profissional para ensino médio, bem como a motivação, a justificativa e os objetivos desta dissertação. Ao final é apresentada a estrutura de seus capítulos subsequentes.

1.1 Contextualização

A análise de índices relacionados a evasão ou retenção escolar na Educação Profissional e Tecnológica (EPT) tem sido considerada um grande desafio na literatura da área de educação, uma vez que a estimativa destes índices no futuro poderiam fornecer informações indispensáveis para tomada de decisão, com efeitos preventivos, em instituições de ensino, na tentativa de reverter a ocorrência de evasão ou retenção escolar (CUNHA; MOURA; ANALIDE, 2016).

Neste sentido, alguns trabalhos relevantes têm desenvolvido conceitos e teorias na tentativa de explicar o fenômeno da evasão ou retenção escolar. Viadero (VIADERO, 2001) e Dore e Lucher (DORE; LUSCHER, 2008) argumentam que o fenômeno da evasão escolar está relacionado ao nível escolar (fundamental, médio, técnico ou superior). Cunha *et al.* (CUNHA et al., 2013) argumentam que a evasão e retenção escolar possuem características multiformes, dificultando a construção de um conceito aplicável aos diversos tipos de situações, tais como familiar, individual ou grupo social.

Além disso, Dore *et al.* (DORE; ARAUJO; S. MENDES, 2014) identificou que a evasão na EPT está relacionada a heterogeneidade dos alunos, associada a déficits de aprendizagem e carência de subsídios socioeconômicos e familiares. Dore e Lucher (DORE; LUSCHER, 2008) e Tavares Junior *et al.* (JUNIOR; SANTOS; S. MACIEL, 2017) argumentam que o fator condicional para ocorrência da evasão está vinculado a democratização do acesso ao ensino, isto é, o ingresso, as oportunidades e as condições mínimas oferecidas pela instituição de ensino

para permanência do estudante até a conclusão dos seus estudos.

Alternativamente, Silva *et al.* (SILVA; DIAS; SILVA, 2015) realizou um estudo e identificou que há índices preocupantes de evasão em toda Rede Federal de Educação Profissional e Tecnológica (RFEPT), tratando-se de um problema social com necessidades de investigação e novas políticas públicas, ou seja, a evasão está diretamente ligada a renda dos estudantes, uma vez que os que possuem baixa renda (de até dois salários mínimos) são os que mais se evadem. Cocco e Sudbrack (COCCO; SUDBRACK, 2016) relacionaram a evasão escolar a renda familiar insuficiente (levando o jovem ao trabalho com horários incompatíveis com o da instituição de ensino), a baixa escolaridade familiar (levando a falta de estímulo ao jovem), a condição estrutural da instituição e a didática dos docentes.

Mais recentemente, Trombini *et al.* (TROMBONI; OLEGARIO; LAROQUE, 2017) realizou um estudo aprofundado sobre as causas da evasão, chegando a conclusão que o histórico da vida escolar anterior ao ensino médio, as expectativas não satisfatórias ao curso, o complexo funcionamento do sistema educacional, e as experiências pessoais e vivências acontecidas dentro da escola, são fatores cruciais para a ocorrência do fenômeno da evasão. Por fim, Silva *et al.* (SILVA; DIAS; SILVA, 2017) relacionou a evasão ao modo como o estudante reage as dificuldades e mudanças enfrentados no período escolar, tanto no ambiente interno quanto externo a instituição.

Neste sentido, mesmo com a ocorrência de uma expansão expressiva do acesso a educação, Rumberger e Thomas (RUMBERGER; THOMAS, 2000) argumentam que os níveis dos índices evasão são alarmantes, principalmente no tocante ao ensino médio. Vale mencionar que dados do Censo Escolar de 2015 revelaram que aproximadamente 1 milhão de estudantes abandonam seus estudos no ensino médio. Além disso, de acordo com Oliveira (OLIVEIRA, 2016), o quantitativo de matrículas no ensino médio teve sua maior queda entre 2014 e 2015 (a quantidade de estudantes caiu de 8.3 para 8.1 milhões, representando uma fatia significativa de 2.7%).

No contexto da expansão da RFEPT no Brasil, podemos confirmar a hipótese apresentada por Rumberger e Thomas (RUMBERGER; THOMAS, 2000), haja visto que ao longo dos últimos anos os seus índices de evasão e de retenção acadêmica têm sido alarmantes, contrariando a perspectiva de universalização do acesso à educação e da garantia da permanência dos discentes.

De maneira geral, Rumberger e Thomas (RUMBERGER; THOMAS, 2000) afirmam que as causas da evasão são complexas, sendo influenciadas por um conjunto de fatores que se relacionam tanto ao estudante quanto à sua família, tanto à escola quanto à comunidade em que esta vive. Além disso, Rumberger e Thomas (RUMBERGER; THOMAS, 2000) identifica-

ram os dois principais contextos de investigação do problema: i) a perspectiva individual, que abrange o estudante e as circunstâncias de seu percurso escolar, e ii) a perspectiva institucional, que leva em conta a família, a escola, a comunidade e os grupos de amigos.

Considerando a complexidade das causas da evasão apresentadas por Rumberger e Thomas (RUMBERGER; THOMAS, 2000), Rebelo (REBELO, 2009) identificou como uma das principais causas da evasão a permanência do aluno no mesmo ano escolar (devido a insucessos no seu desempenho nas disciplinas), isto é, a retenção escolar e, levando em consideração o fato da reprovação ser ineficaz do ponto de vista pedagógico. Neste contexto, verifica-se que a retenção tem efeitos danosos, sobretudo a longo prazo, com consequências futuras para o abandono escolar.

Dessa forma e devido à complexidade destes fatores, como poderíamos desenvolver de uma forma diferenciada, o problema de estimar índices futuros de evasão e retenção escolar? Levando em consideração a relevância dessa problemática, esse trabalho visa desenvolver uma nova abordagem para lidar com o problema, através do uso de séries temporais modeladas e de redes neurais artificiais, na tentativa de prevenção de situações adversas no tocante ao abandono escolar, de forma a minimizar as consequências negativas deste fenômeno.

1.2 Motivação e Justificativa

O problema de prever séries temporais é difícil de ser solucionado devido a complexidade das características encontradas em seu fenômeno gerador (NIU; WANG, 2014). Na literatura, diversos modelos estatísticos, lineares e não-lineares, têm sido propostos para previsão de fenômenos temporais (ENGLE, 1982; RAO; GABR, 1984; OZAKI, 1985; RUMELHART; MCCLELAND, 1987; PRIESTLEY, 1988; BOX; JENKINS; REINSEL, 1994; CLEMENTS; FRANSES; SWANSON, 2004), com destaque para o modelo autoregressivo integrado de médias móveis (*autoregressive integrated moving average*, ARIMA) (BOX; JENKINS; REINSEL, 1994).

No entanto, a grande limitação do modelo ARIMA se dá pelo fato deste ser puramente linear, e não há nenhuma garantia que o fenômeno de uma série temporal é gerado por processos lineares (CLEMENTS; FRANSES; SWANSON, 2004). Clements *et al.* (CLEMENTS; FRANSES; SWANSON, 2004) argumentam que os modelos estatísticos não-lineares não possuem desempenho expressivo e possuem alto custo computacional, implicando na impossibilidade de seu uso na prática (CLEMENTS; FRANSES; SWANSON, 2004).

Devido a esta limitação, modelos de redes neurais artificiais (*artificial neural networks*, ANN) (HAYKIN, 1998) do tipo *perceptron* multicamadas (*multilayer perceptron*, MLP) (GAM-

BOGI; COSTA, 2014), recorrentes (MENEZES; BARRETO, 2008), difusas (*fuzzy*) (VELLA; NIG, 2014) e morfológicas (A. ARAUJO; OLIVEIRA; MEIRA, 2017; A. ARAUJO; OLIVEIRA; L. MEIRA, 2017a,b; A. ARAUJO et al., 2018, 2019), de programação genética (*genetic programming*, GP) (MOUSAVI; ESFAHANIPOUR; ZARANDI, 2014) e do regressor de vetor de suporte (*support vector regressor*, SVR) (ZHIQIANG; HUAIQING; QUAN, 2013; WANG; HUANG; WANG, 2013) têm sido aplicados para modelagem de séries temporais.

No contexto do problema de evasão e retenção escolar, é possível verificar algumas abordagens propostas na literatura para análise do problema. Nascimento *et al.* (NASCIMENTO et al., 2018) desenvolveram um modelo baseado em regressor de vetor de suporte para estimar índices de evasão. Yu *et al.* (YU et al., 2010) apresentaram um abordagem baseada em mineração de dados (*data mining*, DM) (BAKER; ISOTANI; CARVALHO, 2011) para estimar retenção escolar. kabra e Bichkar (KABRA; BICHKAR, 2011) apresentaram um classificador baseado em árvores de decisão (*decision trees*, DT) (KIM, 2008) para estimar o desempenho de estudantes da graduação e pós-graduação da *Can Tho University* e do *Asian Institute of Technology*.

Alternativamente, Martinho *et al.* (MARTINHO; NUNES; MINUSSI, 2013a,b) apresentaram um sistema inteligente, baseado em redes neurais fuzzy-artmap, para estimar o risco de evasão em grupos de estudantes do Instituto Federal do Mato Grosso, obtendo taxas de acerto de aproximadamente 76%. Marquez-Vera *et al.* (MARQUEZ-VERA; ROMERO; VENTURA, 2013) investigaram o uso de técnicas de mineração de dados para classificar grupos de estudantes com perfil evasor. Yasmin (YASMIN, 2013) apresentou um modelo baseado em árvores de classificação para estimar a evasão escolar na Índia, no contexto do ensino à distância. Ahmed e Elaraby (AHMED; ELARABY, 2014) desenvolveram uma abordagem baseada em árvores de decisão para classificar o desempenho de estudantes de graduação. Kawase (KAWASE, 2015) apresentou um sistema inteligente composto por redes neurais de função de base radial para analisar a evasão discente do curso se sistemas de informação da Unviersidade Federal Rural do Rio de Janeiro. Oliveira (OLIVEIRA JUNIOR, 2015) propôs uma abordagem computacional para detecção de padrões a serem utilizados na análise de evasão de estudantes, classificando-os como "haverá evasão" ou "não haverá evasão", da Universidade Tecnológica Federal do Paraná.

Mais recentemente, Meedech *et al.* (MEEDECH; IAM-ON; BOONGOEN, 2016) apresentaram uma abordagem híbrida baseada em árvores de decisão e modelos de regras de indução para descoberta de conhecimento a partir de dados de estudantes da Universidade de *Mae Fah Luang*. Cunha *et al.* (CUNHA; MOURA; ANALIDE, 2016) propuseram um abordagem de aprendizagem de máquina (*machine learning*, ML) para detectar comportamentos de estudantes evasores do Instituto Federal do Rio Grande do Norte. Jaiswal *et al.* (G. JAISWAL; YADAV,

2019) apresentou uma abordagem analítica, 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.

Neste sentido, a literatura argumenta que a evasão e retenção escolar gera anualmente perdas financeiras na ordem de bilhões de reais, bem como não há trabalhos publicados relacionando a abordagem de previsão de séries temporais para índices de evasão e retenção escolar. Desta forma, 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.

1.3 Objetivos

Neste contexto, os objetivos deste trabalho são: i) realizar um estudo sobre séries temporais de índices de evasão e retenção escolar, e ii) desenvolver um modelo com processo de aprendizagem baseada em gradiente descendente capaz de prevê-las; iii) realização de uma análise experimental com o modelo de previsão a partir de gráficos e de duas medidas consolidadas na literatura para determinação do desempenho preditivo.

Para alcançar estes objetivos, foram estabelecidos os seguintes objetivos específicos:

1. Definição, determinação e coleta das séries temporais de interesse:
    - Índice de Evasão Escolar (IEE),
    - Índice de Retenção Escolar (IRE);
2. Realização de um estudo sobre o fenômeno gerador das séries temporais investigadas;
3. Desenvolvimento do modelo de previsão;
4. Desenvolvimento do processo de aprendizagem para projetar o modelo de previsão;
5. Realização de uma análise experimental com o modelo de previsão (utilizando gráficos, duas medidas consolidadas na literatura e testes estatísticos) para determinação do desempenho preditivo;
6. Apresentar o modelo proposto como ferramenta que pode ser utilizada por outras instituições em formato de produto educacional, denominado: Fluxograma didático para modelo de previsão de evasão e retenção escolar.

1.4 Estrutura deste Trabalho

A estrutura deste trabalho é composta por seis capítulos descritos a seguir:

Capítulo 1 - Introdução: este Capítulo apresenta uma introdução ao problema de evasão e retenção de alunos no ensino médio profissional, bem como apresenta os objetivos e motivações deste trabalho;

Capítulo 2 - Séries Temporais: este Capítulo define formalmente as séries temporais, apresenta o problema de previsão de séries temporais abordado neste trabalho e realiza um estudo sobre as séries temporais de índices de evasão e retenção;

Capítulo 3 - Modelos para Previsão de Séries Temporais: este Capítulo apresenta a definição dos modelos para previsão de séries temporais utilizados na literatura;

Capítulo 4 - Simulações e Resultados Experimentais: este Capítulo descreve o procedimento empregado para realização das simulações, define um conjunto de medidas para avaliação da previsão, e apresenta os resultados obtidos, que foram analisados através de medidas de desempenho e gráficos, demonstrando que o modelo de previsão apresentado é uma opção viável para previsão de índices de evasão e retensão;

Capítulo 5 - Conclusões e Trabalhos Futuros: neste capítulo final são apresentadas as considerações finais deste trabalho. Também, é feita uma discussão sobre o modelo de previsão apresentado, suas limitações e sugestões para trabalhos futuros.

2

SÉRIES TEMPORAIS

Neste capítulo, são apresentados a definição formal de uma série temporal e o problema de previsão de séries temporais, bem como descreve as séries temporais de índices de evasão e retenção investigadas neste trabalho.

2.1 Definição

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

$$\mathbf{X} = \{x_t \in \mathbb{R} \mid t = 1, 2, 3 \ldots N\}, \tag{2.1}$$

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

Neste contexto, x_t pode ser modelada em termos de três componentes principais (de acordo com a abordagem aditiva), e definida por (MORETTIN; TOLOI, 2004)

$$x_t = L_t + S_t + R_t, \tag{2.2}$$

em que L_t é a componente de tendência, S_t é a componente sazonal e R_t é a componente aleatória.

Portanto, de acordo com Araújo (A. ARAÚJO, 2016), a análise de uma série temporal consiste em identificar e estimar o grau de contribuição de cada componente visando a construção de um mapeamento capaz de aproximar o fenômeno gerador da série temporal, ou seja, prever o futuro.

2.2 O Problema de Previsão

Como visto na seção anterior, 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 (A. ARAÚJO, 2016).

No entanto, antes da construção de um modelo de previsão, é necessário a definição de alguns elementos (A. ARAÚJO, 2016): *i)* período de previsão: unidade básica de tempo, *ii)* horizonte de previsão: período coberto, no futuro, pela previsão, e *iii)* intervalo de previsão: frequência da série temporal. Note que, dependendo do valor do horizonte, podemos ter modelos de previsão de longo, médio ou curto prazos. Neste trabalho focamos apenas em previsões de curto prazo (um-passo-adiante).

Logo, 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.

A função de autocorrelação (*autocorrelation function*, ACF) (BOX; JENKINS; REINSEL, 1994) pode ser utilizada para definição dos retardos temporais quando há apenas dependência linear entre as observações do fenômeno temporal. No entanto, não há razão para generalizar a existência de dependência linear entre as observações de uma série temporal, uma vez que pode existir dependência não-linear. Logo, a análise e definição dos retardos temporais relevantes via análise ACF é considerada um processo bastante complexo (BOX; JENKINS; REINSEL, 1994).

Diversas abordagens na literatura foram propostas para a definição de retardos temporais, tendo maior destaque: Savit e Green (SAVIT; GREEN, 1991), Pi e Peterson (PI; PETERSON, 1994) e Tanaka *et al.* (TANAKA; OKAMOTO; NAITO, 2001). Entretanto, uma metodologia bastante utilizada na literatura, conhecida como *lagplot* (PERCIVAL; WALDEN, 1998; KANTZ; SCHREIBER, 2003), tem sido aplicada para determinação e análise das relações entre os retardos temporais.

O *lagplot* é um gráfico de dispersão relacionado com os diferentes retardos temporais da série ($x_t \times x_{t-1}$,...,$x_t \times x_{t-n}$), possibilitando a caracterização da relevância de um determinado retardo temporal a partir da existência de alguma estrutura bem definida em seu *lagplot*. No

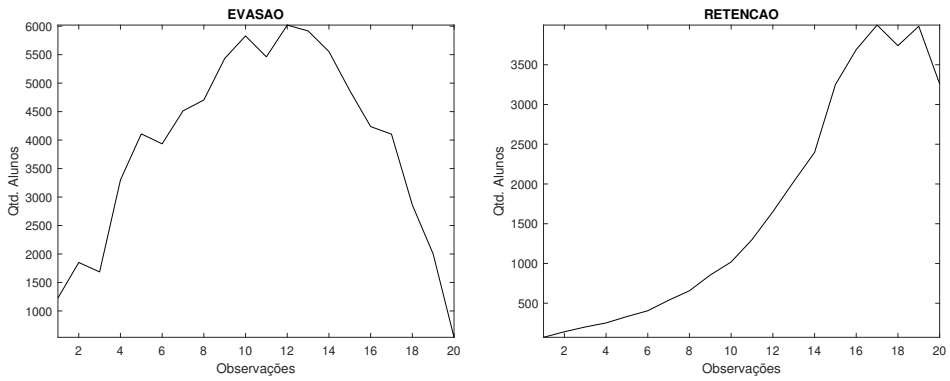
entanto, surge uma limitação, haja visto que esta depende de interpretação humana dos gráficos, e as relações não-lineares nem sempre são humanamente passíveis de identificação. Entretanto, sua simplicidade é um forte argumento para sua utilização (A. ARAÚJO, 2016).

Geralmente, quando há dependência não-linear entre as observações, a informação mútua média (*mean mutual information*, MMI) (FRASER; SWINNEY, 1986; KRASKOV; STG-BAUER; GRASSBERGER, 2004), pode ser utilizada para análise e definição dos retardos temporais (AMJADY; KEYNIA, 2009; STOJANOVIC et al., 2014; TRAN; MUTTIL; PERERA, 2015; CHEN; LEE, 2015), uma vez que a MMI representa uma generalização da ACF para sistemas não-lineares. Note que a inexistência de relacionamento não-linear entre os retardos temporais implica em um valor nulo para a MMI (STOJANOVIC et al., 2014).

2.3 Análise das Séries Temporais

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 (A. ARAÚJO, 2016), este trabalho considerou realizar a análise do fenômeno gerador a partir de seus gráficos, que podem ser ilustrados na Figura 2.1.

Figura 2.1 Gráfico das séries temporais dos índices de evasão e retenção escolar.



(a) IEE (b) IRE

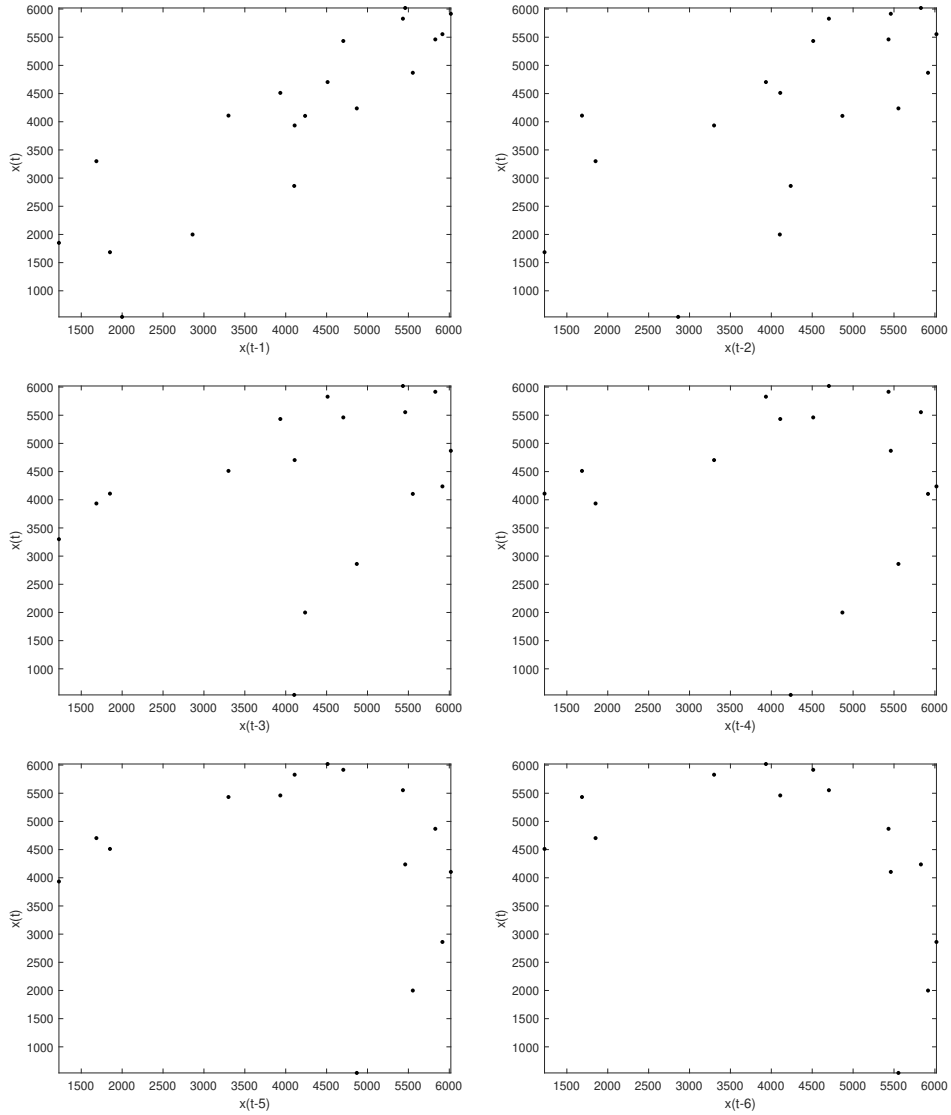
Fonte: Elaborada pelo autor.

De acordo com a Figura 2.1 (a), é possível verificar a existência de componentes de tendência com características crescente e decrescente. Além disso, ao analisar a Figura 2.1

(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 2.2 e 2.3) para determinar e analisar as relações entre os retardos temporais das séries investigadas.

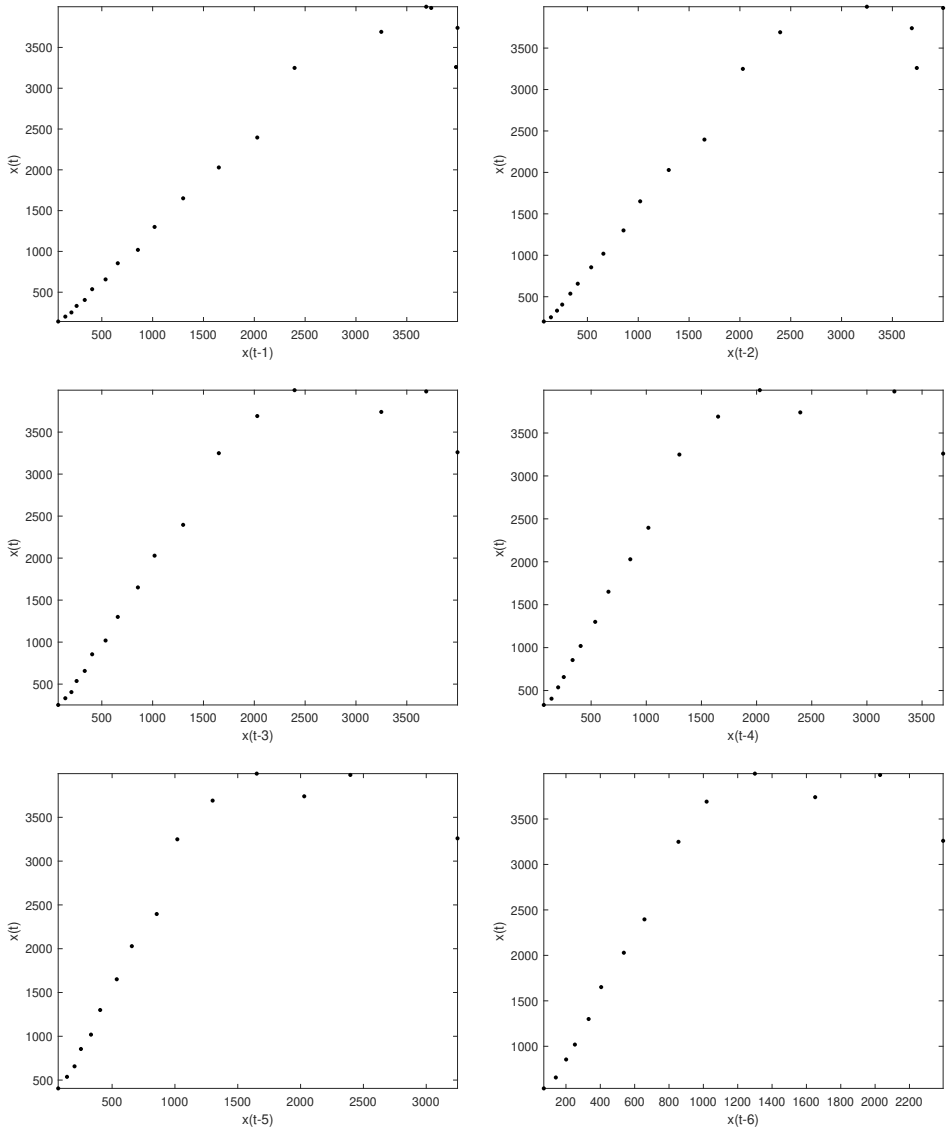
Figura 2.2 *Lagplot* da série IEE.



Fonte: Elaborada pelo autor.

A partir da análise da Figuras 2.2 e 2.3, foi possível identificar estruturas que caracterizam a presença de relacionamento linear e não-linear em todas as séries investigadas. No

Figura 2.3 *Lagplot* da série IRE.

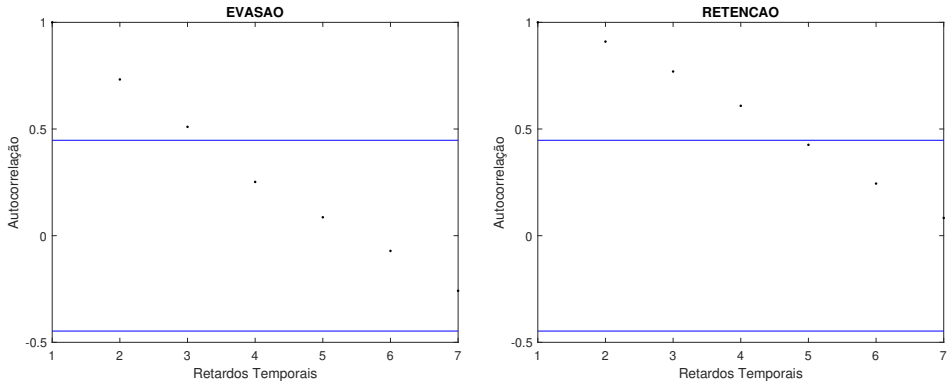


Fonte: Elaborada pelo autor.

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 2.4, é utilizada para analisar o comportamento da componente linear.

Note que, de acordo com a Figura 2.4, 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.

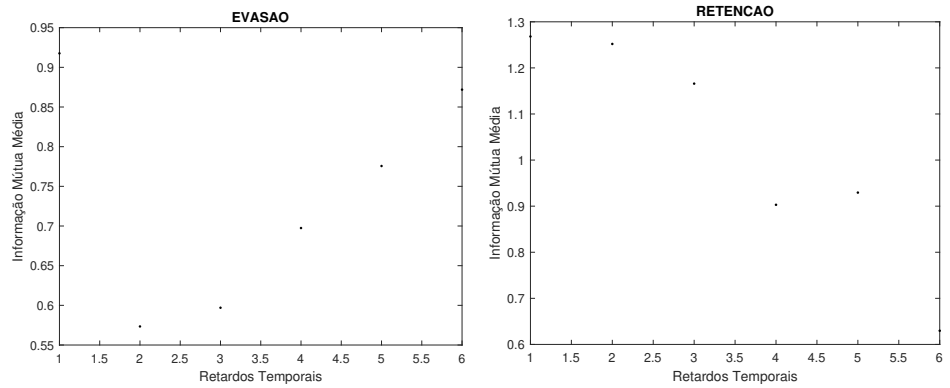
Figura 2.4 ACF das séries temporais dos índices de evasão e retenção escolar.



Fonte: Elaborada pelo autor.

Entretanto, nada se pode observar em relação a natureza da componente não-linear a 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 2.5, é utilizada para analisar a componente não-linear.

Figura 2.5 MMI das séries temporais dos índices de evasão e retenção escolar.



Fonte: Elaborada pelo autor.

De acordo com a Figura 2.5, 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.

2.4 Resumo do Capítulo

Este capítulo apresentou a definição formal de séries temporais e do problema de previsão, bem como apresentou um estudo sobre o fenômeno gerador de dos índices de evasão e retenção escolar. Este apresenta evidências sugerindo que este tipo particular de série temporal é gerada por um processo construído a partir de combinações entre componentes lineares e não-lineares.

3

MODELOS PARA PREVISÃO DE SÉRIES TEMPORAIS

Neste capítulo serão apresentados os modelos de séries temporais que serão investigados no problema de previsão de índices de evasão e retenção escolar.

3.1 Introdução

Na literatura existem três classes de modelos para previsão de séries temporais (A. 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 escolha de um modelo adequado para descrever um dado fenômeno gerador depende de diversos fatores, tais como o comportamento do fenômeno ou o conhecimento a *priori* da sua natureza (A. ARAÚJO, 2016). Desta forma, os modelos clássicos de previsão são baseados em funções temporais, e definidos por (A. ARAÚJO, 2016)

$$x_t = b_1 f_1(t) + b_2 f_2(t) + \cdots + b_i f_i(t) + r_t, \tag{3.1}$$

onde b_i e $f_i(t)$ com $i = 1, 2, \ldots, I$ são, respectivamente, os parâmetros constantes e funções matemáticas de t. O termo r_t é uma componente aleatória.

No entanto, Box *et al.* (BOX; JENKINS; REINSEL, 1994) apresentou uma forma alternativa para modelagem de séries temporais em termos de uma função das componentes aleatórias $(r_t, r_{t-1}, r_{t-2}, \ldots)$. Esta representação conhecida como "operador linear de resposta finita ao impulso", sendo amplamente aplicado para prever séries temporais, e definida por (BOX;

JENKINS; REINSEL, 1994)

$$x_t = m + y_0 r_t + y_1 r_{t-1} + y_2 r_{t-2} + \cdots + y_i r_{t-i}, \tag{3.2}$$

onde m e y_i $(i = 1, 2, \ldots, I)$ são termos constantes.

A seguir, serão apresentados os modelos para previsão de séries temporais escolhidos para análise comparativa apresentada neste trabalho.

3.2 Modelos Estatísticos

Diversos modelos estatísticos têm sido propostos na literatura para solucionar o problema de previsão de séries temporais (CLEMENTS; FRANSES; SWANSON, 2004). No entanto, o modelo de Box *et al.* (BOX; JENKINS; REINSEL, 1994) tem recebido destaque como solução para problemas reais de previsão de séries temporais. Antes de definir formalmente o modelo de proposto por Box *et al.* (BOX; JENKINS; REINSEL, 1994), surge a necessidade de definição dos modelos estatísticos lineares autorregressivos e de médias móveis.

O modelo autorregressivo (*autoregressive*, AR) é definido por (BOX; JENKINS; REINSEL, 1994)

$$\widetilde{x_t} = \phi_1 \widetilde{x}_{t-1} + \phi_2 \widetilde{x}_{t-2} + \cdots + \phi_p \widetilde{x}_{t-p} + r_t, \tag{3.3}$$

onde $\widetilde{x_i} = x_i - \mu$ e o termo μ é o nível médio da série. Os termos ϕ_i $(i = 1, 2, \ldots, p)$ são os coeficientes autorregressivos.

O modelo de médias móveis (*moving average*, MA) é definido por (BOX; JENKINS; REINSEL, 1994)

$$x_t = \mu + r_t - \theta_1 r_{t-1} - \theta_2 r_{t-2} - \cdots - \theta_q r_{t-q}. \tag{3.4}$$

Como $\widetilde{x_i} = x_i - \mu$, então (BOX; JENKINS; REINSEL, 1994)

$$\widetilde{x} = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) r_t = \Theta(B) a_t \tag{3.5}$$

onde $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ é o operador de médias móveis.

Na tentativa de se construir um modelo de previsão parcimonioso, foi apresentado o modelo autorregressivo de médias móveis (*autoregressive moving average*, ARMA), que é composto da combinação de ambos os modelos AR e MA, e definido por (BOX; JENKINS; REINSEL, 1994)

$$\widetilde{x_t} = \phi_1 \widetilde{x}_{t-1} + \cdots + \phi_p \widetilde{x}_{t-p} + r_t - \theta_1 r_{t-1} - \cdots - \theta_q r_{t-q}, \tag{3.6}$$

onde p é o número de termos autorregressivos e q é o número de termos da média móvel.

3.2.1 Autorregressivo Integrado de Médias Móveis

Box *et al.* (BOX; JENKINS; REINSEL, 1994) apresentaram um modelo, referido como modelo autorregressivo integrado de médias móveis (*autoregressive integrated moving average*, ARIMA), que consiste na aplicação de um filtro passa-alta na série temporal para que estes sejam aplicados a um modelo ARMA. Este procedimento é conhecido como diferenças entre dados e é utilizado para tornar a série temporal um processo estacionário. Assim, considerando uma série temporal estacionária, esta pode ser representado por um modelo ARMA(p,q). Como a série é representada por uma diferença, esta é uma integral da série, e é representada pela letra "I" (*integrated*) na notação do modelo ARIMA(p,q,d), em que os termos p e q são as ordens dos operador AR e MA, respectivamente, e o termo d representa a ordem das diferenças.

Box *et al.* (BOX; JENKINS; REINSEL, 1994) apresentaram um procedimento capaz de encontrar o modelo com melhor desempenho para solucionar o problema de previsão de séries temporais: *i)* utiliza-se um mecanismo para medir a correlação entre as observações dentro de um conjunto de dados da série temporal. Este mecanismo é representado pela ACF e a função de autocorrelação parcial (*partial autocorrelation function*, PACF) (BOX; JENKINS; REINSEL, 1994), ambas formalmente definidas em (BOX; JENKINS; REINSEL, 1994), *ii)* estimam-se os coeficientes do modelo escolhido no passo *(i)*, e *iii)* procedimentos de validação são empregados a fim de determinar a adequação do modelo candidato como solução para o problema em questão.

3.2.2 Considerações

Apesar do modelo ARIMA (BOX; JENKINS; REINSEL, 1994) ser uma das escolhas mais comuns dentre as técnicas apresentadas na literatura de previsão de séries temporais, ele é um modelo linear. Entretanto, o fenômeno gerador de séries temporais dos índices de evasão e retenção escolar possuem componentes não-lineares (como foi apresentado na análise das séries temporais no Capítulo 2) e tal fato introduz uma limitação na precisão das previsões geradas, uma vez que estes modelos assumem que tais séries temporais são geradas por processos puramente lineares.

Neste contexto, diversos modelos estatísticos não-lineares foram propostos para superar as limitações do modelo ARIMA. Entretanto, de acordo com Clements *et al.* (CLEMENTS; FRANSES; SWANSON, 2004), não se encontram evidências, em termos de desempenho preditivo, a favor dos modelos estatísticos não-lineares, mesmo com a inclusão de alta complexidade

matemática e computacional, quando comparados ao modelo ARIMA. Por este motivo o modelo ARIMA foi escolhido para representar a classe dos modelos estatísticos.

3.3 Modelos de Redes Neurais Artificiais

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

$$y = f(u), \tag{3.7}$$

com

$$u = \sum_{j=1}^{J} w_j x_j + b, \tag{3.8}$$

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.

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)

$$y(\mathbf{w}, \mathbf{x}, b) = f(\mathbf{w}^T \mathbf{x} + b), \tag{3.9}$$

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

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

3.3.1 *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 (A. ARAÚJO, 2016).

O modelo de rede MLP com melhor desempenho para previsão de séries temporais reportado na literatura utiliza função de ativação sigmóide logística para todas as unidades de processamento escondidas (A. ARAÚJO, 2016). A unidade de processamento de saída utiliza função de ativação linear com seu bias passando por função sigmóide logística (A. ARAÚJO, 2016). Portanto, a saída da rede MLP é dada por:

$$y_k(t) = \sum_{j=1}^{n_h} W_{jk} Sig\left[\sum_{i=1}^{n_{in}} W_{ij} x_i(t) + b_j^1\right] + Sig(b_k^2), \tag{3.10}$$

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:

$$Sig(x) = \frac{1}{1 + exp(-x)}. \tag{3.11}$$

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). No entanto, outros algoritmos de aprendizagem têm sido utilizados com sucesso para treinamento de redes MLP, tais como métodos adaptativos (PACK; EL-SHARKAWI; II, 1991; JACOBS, 1998), *quick-propagation* (QUICKPROP) (FAHLMAN, 1989), o *resilient back-propagation* (RPROP) (RIEDMILLER; BRAUN, 1993), *levenberg-marquardt* (LM) (HAGAN; MENHAJ, 1994), gradiente conjugado escalar (*scaled conjugate gradient*, SCG) (MOLLER, 1993), gradiente conjugado de um passo secante (*one step secant conjugate gradient*, OSSCG) (BATTITI, 1992), etc.

3.3.2 Considerações

A fim de se definir uma solução a um dado problema de previsão de séries temporais, as ANNs requerem a definição de um conjunto de parâmetros que são bastante difíceis de determinar, sendo esta escolha de fundamental importância para a capacidade de ajuste de fase temporal. Assim, a grande questão para otimizar o desempenho de uma rede neural para problemas de previsão é como determinar os valores subótimos destes parâmetros, bem como determinar os retardos temporais relevantes para caracterizar o fenômeno gerador da série temporal.

Apesar da rede neural ser capaz de aproximar fenômenos temporais, estas possuem um custo computacional bastante elevado para determinar um conjunto de parâmetros (como a quantidade de neurônios na camada escondida, os pesos sinápticos, a arquitetura, o algoritmo de treinamento e seus respectivos parâmetros, dentre outros (HAYKIN, 1998)) que solucione o problema de previsão com desempenho aceitável. A definição incorreta destes parâmetros afeta diretamente a eficiência do processo de aprendizagem da rede.

3.4 Resumo do Capítulo

Este capítulo apresentou a definição dos modelos estatísticos e dos modelos de redes neurais artificiais considerados para previsão de séries temporais de índices de evasão e retenção escolar. Também são elencados os motivos para escolha deste modelos, em particular, para previsão das séries temporais estudadas nesta dissertação.

4

SIMULAÇÕES E RESULTADOS EXPERIMENTAIS

Este capítulo apresenta o processo utilizado para realização dos experimentos com os modelos investigados, bem como apresenta as medidas utilizadas para avaliação de desempenho. Ao final, os resultados alcançados pelos modelos serão analisados, discutidos e validados estatisticamente.

4.1 Metodologia

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

$$xn_i = \frac{x_i - min(\mathbf{x})}{max(\mathbf{x}) - min(\mathbf{x})}. \tag{4.1}$$

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.

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 no Capítulo 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.

4.1.1 Medidas para Desempenho de Previsão

A principal e mais utilizada medida para avaliação da previsão é o erro médio quadrático (*mean squared error*, MSE), dada por (CLEMENTS; HENDRY, 1993)

$$\text{MSE} = \frac{1}{N}\sum_{j=1}^{N}(e_j)^2, \tag{4.2}$$

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

$$e_j = x_j - \widehat{x}_j \tag{4.3}$$

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 $\to 0$.

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 deve ser considerada para permitir uma avaliação 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, dada por (CLEMENTS; HENDRY, 1993)

$$\text{MAPE} = \frac{1}{N}\sum_{j=1}^{N}\left|\frac{e_j}{x_j}\right|. \tag{4.4}$$

Vale mencionar que em um modelo de previsão ideal, MAPE $\to 0$.

4.2 Resultados

A seguir será apresentada uma análise comparativa entre o modelo ARIMA e o modelo proposto (MLP) a partir das medidas de desempenho definidas na Seção 4.1.1. Foram calculadas a média e o desvio padrão do desempenho preditivo para cada medida investigada. Além disso, a fim de validar estatísticamente o modelo proposto com o melhor desempenho preditivo, foi aplicado o teste de Friedman com nível de significância de $\alpha = 0.05$, uma vez que este estabele um ranking de desempenho para os modelos investigados. Também foi utilizado um teste *post hoc*, conhecido como teste de Tukey, com $\alpha = 0.05$, na tentativa de analisar o desempenho par a par dos modelos investigados. Vale mencionar que ambos os testes consideraram todas as

séries temporais em conjunto (utilizando a abordagem proposta definida em (A. ARAUJO et al., 2019)).

4.2.1 Análise da Medida MSE

Na Tabela 4.1, 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 4.1 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 $\pm$0.0003	0.0027 $\pm$0.0020	1	1.0		
ARIMA	0.0091 $\pm$0.0000	0.2413 $\pm$0.0000	2	2.0	-1.0	1.57e-01

De acordo com a Tabela 4.1 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 MSE no intervalo [3.E-4,8E-3] indicam que as previsões geradas pelo modelo proposto estão bastante próximas dos valores reais da série temporal. De acordo com os resultados do Teste de Friedman, é possível confirmar, estatisticamente, os resultados apresentados na Tabela 4.1. Além disso, note que o modelo proposto alcançou o menor valor de posto para o Teste de Friedman, 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.

4.2.2 Análise da Medida MAPE

A Tabela 4.2 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.

Note que os resultados apresentados na Tabela 4.1 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

Tabela 4.2 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 $\pm$0.0268	0.0635 $\pm$0.0287	1	1.0		
ARIMA	0.2112 $\pm$0.0000	0.6129 $\pm$0.0000	2	2.0	-1.0	1.57e-01

0.03% a 0.06%. Novamente, o Teste de Friedman pôde confirmar, estatisticamente, os resultados apresentados na Tabela 4.2, em que 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.

4.2.3 Análise do Comportamento da Previsão

A Figura 4.1 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, isto é, os valores estimados estão mais próximos aos valores reais das séries temporais consideradas neste trabalho. No caso particular da série IEE, é possível verificar que 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, ou seja, é capaz de prever o futuro dos índices de evasão e retenção escolar investigados e, portanto, é uma opção viável para prever com eficácia tais fenômenos temporais.

4.2.4 Considerações

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

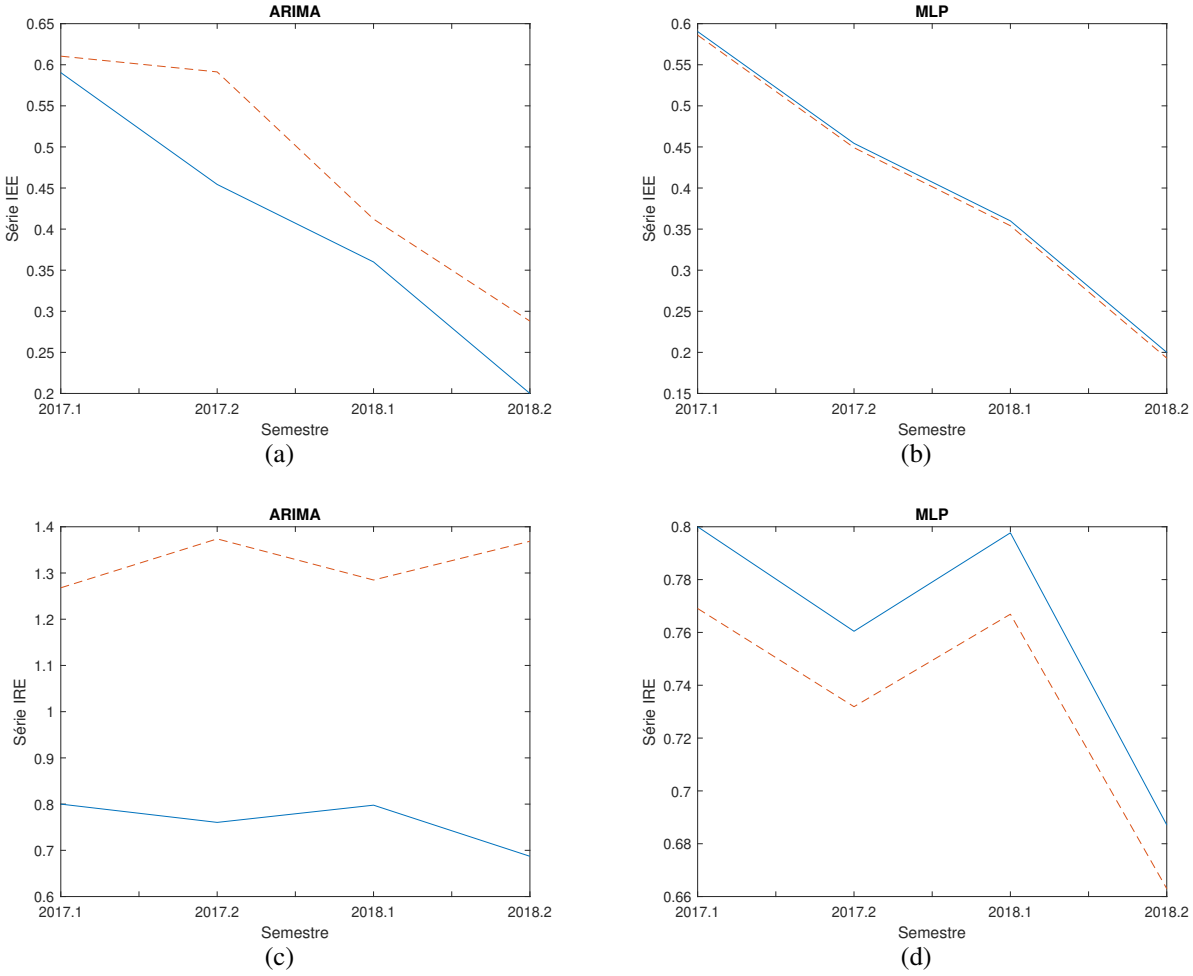


Figura 4.1 Gráfico de previsão (conjunto de teste - semestre 2017.1 ao semestre 2018.2): linha sólida azul (valor real) e a linha vermelha tracejada (valor previsto): a) Série IEE - Modelo ARIMA, b) Série IEE - Modelo MLP, c) Série IRE - Modelo ARIMA, d) Série IRE - Modelo MLP.

modelo MLP.

4.3 Resumo do Capítulo

Neste Capítulo foi apresentada uma análise comparativa entre os resultados obtidos pelos modelos ARIMA e MLP para séries temporais de índices de evasão e retenção escolar. Inicialmente foi apresentada a metodologia empregada para realização dos experimentos, para posterior análise do desempenho preditivo, através das medidas MSE e MAPE e dos testes de Friedman e Tukey. Por fim, foi possível comprovar que o modelo MLP alcançou desempenho de previsão superior ao modelo ARIMA.

5

CONCLUSÕES E TRABALHOS FUTUROS

Este Capítulo apresenta as considerações finais sobre o desenvolvimento deste trabalho. Além disso, são apresentadas as limitações e algumas sugestões para trabalhos futuros.

5.1 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 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).

Neste contexto, 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 apresentado 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 o 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 50 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.

5.2 Trabalhos Futuros

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 outras séries temporais de índices de evasão e retenção escolar, provenientes de outras instituições de ensino.

Também, um estudo particular sobre a complexidade computacional deve ser realizada 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

A. ARAÚJO, R. de. **Mercado de Ações Brasileiro em Alta-Frequência**: evidências de sua previsibilidade com modelagem morfológica-linear. 2016. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal de Pernambuco.

A. ARAUJO, R. de et al. Evolutionary-morphological learning machines for high-frequency financial time series prediction. **Swarm and Evolutionary Computation**, [S.l.], v.42, p.1 – 15, 2018.

A. ARAUJO, R. de et al. A deep increasing-decreasing-linear neural network for financial time series prediction. **Neurocomputing**, [S.l.], v.347, p.59 – 81, 2019.

A. ARAUJO, R. de; OLIVEIRA, A. L. I.; L. MEIRA, S. R. de. A Morphological Neural Network for Binary Classification Problems. **Engineering Applications of Artificial Intelligence**, [S.l.], v.65, p.12 – 28, 2017.

A. ARAUJO, R. de; OLIVEIRA, A. L. I.; L. MEIRA, S. R. de. A Class of Hybrid Multilayer Perceptrons for Software Development Effort Estimation Problems. **Expert Systems with Applications**, [S.l.], v.90, p.1 – 12, 2017.

A. ARAUJO, R. de; OLIVEIRA, A. L. I.; MEIRA, S. On the problem of forecasting air pollutant concentration with morphological models. **Neurocomputing**, [S.l.], v.265, p.91 – 104, 2017.

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**, [S.l.], v.2, n.2, p.43–47, 2014.

AMJADY, N.; KEYNIA, F. Day-ahead price forecasting of electricity markets by mutual information technique and cascaded neuro-evolutionary algorithm. **IEEE Transactions on Power Systems**, [S.l.], v.24, n.1, p.306–318, 2009.

BAKER, R.; ISOTANI, S.; CARVALHO, A. Mineração de Dados Educacionais: oportunidades para o brasil. **Brazilian Journal of Computers in Education**, [S.l.], v.19, n.02, 2011.

BATTITI, R. One Step Secant Conjugate Gradient. **Neural Computation**, [S.l.], v.4, p.141–166, 1992.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. **Time Series Analysis**: forecasting and control. 3.ed. New Jersey: Prentice Hall, 1994.

CHEN, T. T.; LEE, S. J. A weighted LS-SVM based learning system for time series forecasting. **Information Sciences**, [S.l.], v.299, p.99–116, 2015.

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

CLEMENTS, M. P.; HENDRY, D. F. On the Limitations of Comparing Mean Square Forecast Errors. **Journal of Forecasting**, [S.l.], 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**, [S.l.], 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. **Anais...** [S.l.: s.n.], 2016. p.189–198.

DORE, R.; ARAUJO, A. D. de; S. MENDES, J. de. **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. **Anais...** [S.l.: s.n.], 2008. p.197–203.

ENGLE, R. F. Autoregressive conditional heteroskedasticity with estimates of the variance of UK onflation. **Econometrica**, [S.l.], v.50, p.987–1008, 1982.

FAHLMAN, S. E. Faster-learning variations of back-propagation: an empirical study. **Proceedings of the 1998 Connectionist Models Summer School**, [S.l.], 1989.

FRASER, A.; SWINNEY, H. Independent Coordinates for Strange Atractors from Mutual Information. **Physical Review A**, [S.l.], v.33, n.2, p.1134–1140, 1986.

G. JAISWAL, A. S.; YADAV, S. K. Analytical Approach for Predicting Dropouts in Higher Education. **International Journal of Information and Communication Technology Education**, [S.l.], v.3, n.15, p.1–14, 2019.

GAMBOGI, J. A.; COSTA, O. L. V. Stock Market Trading System Implemented by Artificial Neural Networks. **Learning and Nonlinear Models**, [S.l.], v.11, n.2, p.92 – 102, 2014.

HAGAN, M.; MENHAJ, M. Training feedforward networks with the Marquardt algorithm. **IEEE Transactions on Neural Networks**, [S.l.], v.5, n.6, p.989–993, November 1994.

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

JACOBS, R. A. Incresed rates od convergence through learning rate adaptation. **Neural Networks**, [S.l.], v.1, n.4, p.295–308, 1998.

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

KABRA, R. R.; BICHKAR, R. Performance Prediction of Engineering Students using Decision Trees. **International Journal of Computer Application**, [S.l.], v.36, n.11, p.1–12, 2011.

KANTZ, H.; SCHREIBER, T. **Nonlinear Time Series analysis**. 2.ed. 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**. 2015. Dissertação (Mestrado em Ciência da Computação) — Universidade Federal Rural do Rio de Janeiro.

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**, [S.l.], 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**, [S.l.], 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**, [S.l.], v.8, n.1, p.7–14, 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**, [S.l.], p.111–114, 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**, [S.l.], 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. **Anais...** Springer International Publishing, 2016. p.143–155.

MENEZES, J. M. P.; BARRETO, G. A. Long-term time series prediction with the NARX network: an empirical evaluation. **Neurocomputing**, [S.l.], v.71, n.16-18, p.3335–3343, 2008.

MOLLER, M. F. A scaled conjugate gradient algorithm for fast supervised learning. **Neural Networks**, [S.l.], v.6, p.525–533, 1993.

MORETTIN, P. A.; TOLOI, C. M. C. **Analise de Series Temporais**. [S.l.]: ABE-Projeto Fisher e Editora Edgard Blucher, 2004.

MOUSAVI, S.; ESFAHANIPOUR, A.; ZARANDI, M. H. F. A novel approach to dynamic portfolio trading system using multitree genetic programming. **Knowledge-Based Systems**, [S.l.], v.66, n.0, p.68 – 81, 2014.

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. **Anais...** Springer International Publishing, 2018. p.246–257.

NIU, H.; WANG, J. Financial time series prediction by a random data-time effective RBF neural network. **Soft Computing**, [S.l.], v.18, n.3, p.497–508, 2014.

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**, [S.l.], v.10, n.17, p.187–198, 2016.

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

OZAKI, T. **Nonlinear Time Series Models and Dynamical Systems**. Amsterdam: Noth-Holland, 1985. (HandBook of Statistics, v.5).

PACK, D. C.; EL-SHARKAWI, M. A.; II, R. J. M. An adaptively trained neural network. **IEEE Transactions on Neural Networks**, [S.l.], v.2, n.3, p.334–345, 1991.

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

PI, H.; PETERSON, C. Finding the Embedding Dimension and Variable Dependences in Time Series. **Neural Computation**, [S.l.], v.6, p.509–520, 1994.

PRECHELT, L. **Proben1**: a set of neural network benchmark problems and benchmarking rules. [S.l.: s.n.], 1994. (21/94).

PRIESTLEY, M. B. **Non-Linear and Non-Stacionary Time Series Analysis**. [S.l.]: Academic Press, 1988.

RAO, T. S.; GABR, M. M. **Introduction to Bispectral Analysis and Bilinear Time Series Models**. Berlin: Springer, 1984. (Lecture Notes in Statistics, v.24).

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

RIEDMILLER, M.; BRAUN, H. A direct adaptive method for faster backpropagation learning: the rprop algorithm. In: IEEE INT. CONF. ON NEURAL NETWORKS (ICNN), San Francisco. **Proceedings...** [S.l.: s.n.], 1993. p.586–591.

ROMERO, C.; VENTURA, S. Data mining in education. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, [S.l.], 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**, [S.l.], v.73, n.1, p.39–67, 2000.

RUMELHART, D. E.; MCCLELAND, J. L. **Parallel Distributed Processing, Explorations in the Microstructure of Cognition**. [S.l.]: MIT Press, 1987. v.1 & 2.

SAVIT, R.; GREEN, M. Time Series and Dependent Variables. **Physica D**, [S.l.], v.50, p.95–116, 1991.

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**, [S.l.], v.3, n.6, p.279–293, 2015.

SILVA, J.; DIAS, P.; SILVA, M. C. Fatores de influência no processo de evasão escolar em três cursos técnicos do Instituto Federal de Educação, Ciência e Tecnologia de Brasília. **Revista da UIIPS**, [S.l.], v.5, n.3, p.6–21, 2017.

STOJANOVIC, M. B. et al. A methodology for training set instance selection using mutual information in time series prediction. **Neurocomputing**, [S.l.], v.141, p.236–245, 2014.

TANAKA, N.; OKAMOTO, H.; NAITO, M. Estimating the Active Dimension of the Dynamics in a Time Series Based on a Information Criterion. **Physica D**, [S.l.], v.158, p.19–31, 2001.

TRAN, H. D.; MUTTIL, N.; PERERA, B. J. C. Selection of significant input variables for time series forecasting. **Environmental Modelling & Software**, [S.l.], v.64, p.156–163, 2015.

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

VELLA, V.; NIG, W. L. Enhancing risk-adjusted performance of stock market intraday trading with Neuro-Fuzzy systems. **Neurocomputing**, [S.l.], v.141, n.0, p.170 – 187, 2014.

VIADERO, D. The Dropout dilemma: research hindered by lack of uniform way to count students who quit school. **Education Week**, [S.l.], v.20, n.21, p.26–29, 2001.

WANG, B.; HUANG, H.; WANG, X. A support vector machine based MSM model for financial short-term volatility forecasting. **Neural Computing and Applications**, [S.l.], v.22, n.1, p.21–28, 2013.

YASMIN, D. Application of the classification tree model in predicting learner dropout behaviour in open and distance learning. **Distance Education**, [S.l.], 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**, [S.l.], 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**, [S.l.], v.14, p.35–62, 1998.

ZHIQIANG, G.; HUAIQING, W.; QUAN, L. Financial time series forecasting using LPP and SVM optimized by PSO. **Soft Computing**, [S.l.], v.17, n.5, p.805–818, 2013.

Apêndice

A

PRODUTO EDUCACIONAL

O produto Educacional apresentado nesta dissertação constitui-se em um formato de fluxograma, ferramenta utilizada como uma imagem visual daquilo que foi estudado. Representa um método, que se constitui de etapas a quais devem ser percorridas para que se tenha uma padronização na sequência de atividades necessárias para realização do processo.

A.1 Justificativa

Embora a expansão da Rede Federal de Educação Profissional, Científica e Tecnológica, tenha ampliado a oferta de vagas e criação de políticas de ações afirmativas de acesso, os índices de evasão na rede tornaram-se preocupantes. Buscando conter a evasão e retenção escolar nos Institutos Federais, o MEC em conjunto com os Institutos Federais, institui um plano voltado ao tratamento da evasão na Rede Federal de Educação Profissional.

Esse plano deveria contemplar ente outras metas o levantamento de dados de variáveis que permitam identificar alunos com maior propensão de evasão; inserção nos termos de acordos de metas e compromissos de indicadores de evasão, retenção, além de uma política voltada a linhas de assistência estudantil, voltadas ao atendimento de alunos com risco de evasão.

Diante disso o IFCE construiu o Plano Estratégico Institucional para Permanência e Êxito dos Estudantes PPE, no objetivo de fortalecer a qualidade do ensino através de ações de incentivo à permanência e à promoção acadêmica. O PPE está estruturado relatando a trajetória do IFCE, com enfoque na sua identidade, na organização multicampi, ambiente de atuação do IFCE e aspectos socioeducacionais do Ceará. Apresenta a base conceitual de evasão adotada pelo IFCE, diagnóstico quantitativo e qualitativo relativos à evasão e retenção, bem como medidas de intervenção que visam a superação ou minimização dos índices e evasão.

A plataforma a qual foi extraída os dados quantitativos para e trabalho de dissertação, (IFCE em Números) incluída dentro das ações do PPE, possibilita diagnosticar a situação de-

talhada da evasão na instituição, consolidando os dados dos sistemas acadêmicos num local único, de uso simplificado para que os próprios educadores (docentes, técnico-administrativos, entre outros) conseguissem acessar e manusear as informações de acordo com as suas necessidades específicas. Os dados do sistema acadêmico são mantidos pelas coordenações de registro acadêmico presente em cada um dos 33 campi do IFCE. A plataforma IFCE em Números apresenta visualizações desses dados para subsidiar a execução de ações para permanência e êxito dos estudantes da instituição. O fluxograma foi aplicado aos dados extraídos da plataforma a qual prentedeu-se desenvolver uma análise experimental utilizando séries temporais de índices de evasão e retenção escolar do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), no período de 2009 a 2018.

A.2 Produto: Fluxograma Didático para Modelo de Previsão de Evasão e Retenção Escolar

O fluxograma para um modelo de previsão, referido como um caminho, fluxo que se perfaz de várias etapas, no objetivo de abordagem a rede neural artificial perceptron multicamadas (MLP), apresentando-se, 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. O fluxograma, disponível para *download* em *http://educapes.capes.gov.br/handle/capes/599150*, pode ser utilizado como uma ferramenta, para estimar índices de evasão e retenção escolar, em outras plataformas educacionais, por outras instituições de educação. As subseções a seguir descrevem os três blocos deste fluxograma.

A.2.1 Coleta de Séries Temporais de Interesse

Inicialmente, foi acessado a plataforma "IFCE em Números" com o intuito de extrair informações relevantes referentes ao índice de Evasão Escolar (IEE) e ao índice de retenção Escolar (IRE). Note que ambos os índices estão relacionados com a quantidade de alunos evadidos e retidos do IFCE. Posteriormente, deve-se escolher a frequência semestral e o período de 2009 a 2018 de cada índice. A seguir, para definição das janelas temporais, deve-se calcular a função de autocorrelação, a informação mútua média e o lagplot. Além disso, estas técnicas permitirão avaliar a dependência linear e não linear existente em cada série temporal investigada.

A.2.2 Desenvolvimento do Modelo de Previsão

Inicialmente, deve-se considerar os modelos estatísticos clássicos propostos na literatura como solução para o problema, de maneira a se construir um referencial, uma vez que os modelos estatísticos são, naturalmente, a primeira abordagem a ser pensada para solucionar o problema em questão. Neste contexto, deve ser investigado o modelo ARIMA por ser aquele dotado de melhor desempenho para esta classe de modelos.

Vale mencionar que em outros problemas de previsão, redes neurais artificiais tem apresentado resultados de previsão mais eficazes, quando comparados a modelos estatísticos. Dentre os diversos modelos de redes neurais, o perceptron multicamadas tem apresentado os melhores resultados em diversas aplicações de previsão de séries temporais e, por isso, foi considerado como modelo proposto, que possui três camadas de neurônios artificiais (camada de entrada, camada escondida e camada de saída).

A.2.3 Simulações e Medidas utilizadas para Avaliação de Desempenho

Com as séries temporais armazenadas, deve ser aplicado o processo de normalização para o intervalo [0,1], de forma a cobrir o domínio dos modelos investigados. Para realização das simulaçoes, deve-se, para cada modelo, definir uma arquitetura básica, composta dos parâmetros dos modelos. No caso do modelo ARIMA, deve-se definir os parâmetros p, d e q. No caso do modelo MLP, deve-se definir os parâmetros quantidade de neurônios artificiais na camada de entrada, na camada escondida e na camada de saída, bem como as funções de ativação de cada neurônio artificial. Note que como a previsão considerada é de apenas um passo a frente, a camada de saída deve ter apenas um neurônio artificial. Para o processo de treinamento da MLP, deve-se utilizar o algoritmo de retropropagação do erro, onde deve-se definir os parâmetros "taxa de aprendizagem" e "termo de *momentum*". Para realização dos experimentos, deve ser utilizada a ferramenta *Neural Network Toolbox* (para o modelo MLP) e a *Econometrics Toolbox* (para o modelo ARIMA) do software MATLAB.

Com os resultados gerados a partir de cinquenta simulações para cada série, deve-se calcular as medidas de desempenho consideradas (MSE e MAPE) e, posteriormente, calcular a média e o desvio padrão desses experimentos, de forma a permitir utilizar os testes de Friedman e Tukey para avaliação quantitativa dos resultados. Também, deve-se gerar o gráfico entre a previsão e o valor real da série temporal, de maneira a permitir avaliar, qualitativamente, os resultados preditivos dos modelos investigados.

FLUXOGRAMA DIDÁTICO PARA MODELO DE PREVISÃO DE ESTIMATIVAS FUTURAS DE EVASÃO E RETENÇÃO ESCOLAR
coleta de séries tempoais de interesse
Desenvovimento do modelo de previsão
Simulações e medidas utilizadas para avaliação de desempenho
Coleta semestral na base de dados 2009 a 2018
Determinação e coleta das séries temporais. IEE e IRE
Definir uma janela temporal (d) contendo as observações do passado da série temporal. Time lags. TL
Dependência linear/ Dependência não-linear
Gráfico de dispersão:lagplot ACF/MMI
considerar o tipo particular de série temporal encontrada suas combinações entre componentes lineares e não-lineares.
Definição de modelos estatísticos disponíveis na literatura
modelos estatísticos para solucionar o problema de previsão de séries temporais
ARMA; ARIMA; ANN; MLP
ANN superar as limitações dos modelos estatísticos lineares e não lineares
MLP tem melhor desempenho para previsão de séries temporais na literatura
considerar MLP/ neurônios são dispostos em uma ou mais camadas /capacidade de aprendizagem através de ajustes aplicados aos seus pesos sinápticos e bias. Determinar retardos temporais relevantes
Processo de normalização das séries temporais investigadas
Comparar o desempenho preditivo com o modelo proposto utilizar o ARIMA
Definir uma arquitetura básica para todos os experimentos MLP.Definir quantidade de camadas/treinamento da rede (BP).
medida para avaliação do desempenho preditivo MSE/MAPE.
Análise comparativa entre os valores reais e as previsões geradas pelo modelo MLP e ARIMA
calcular a média e o desvio padrão do desempenho preditivo. validar com teste de Friedman e Tukey
considerar o modelo proposto MLP obteve desempenho preditivo estatisticamente superior ao modelo ARIMA, sob as mesmas condições, com capacidade de estimar, no futuro, índices de evasão e retenção escolar.
Índice de Evasão Escolar (IEE)
Índice de Retenção Escolar (IRE)
Função de autocorrelação (ACF)
Informação mútua média (MMI)
Modelo autorregressivo de médias móveis (ARMA)
Modelo autorregressivo integrado de médias móveis (ARIMA)
Redes neurais artificiais ANN / Perceptron multicamadas (MLP)
Algoritmo de retro-propagação do erro (BP)
Erro médio quadrático (MSE)
Erro médio Percentual absoluto (MAPE)