



UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

UNIDADE ACADÊMICA DE SERRA TALHADA

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Redes Neurais Artificiais Aplicadas à Previsão de Surtos de Leptospirose.

Por

Elias Ferraz Dantas

Serra Talhada,
Agosto/2018



UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

Elias Ferraz Dantas

Redes Neurais Artificiais Aplicadas à Previsão de Surtos de Leptospirose

Trabalho de Conclusão de Curso apresentada ao
Curso de Bacharelado em Sistemas de Informação
da Unidade Acadêmica de Serra Talhada da
Universidade Federal Rural de Pernambuco como
requisito parcial à obtenção do grau de Bacharel.

Orientador: Prof. M.e Glauber Magalhães Pires

Serra Talhada,
Agosto/2018

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema Integrado de Bibliotecas da UFRPE
Biblioteca da UAST, Serra Talhada - PE, Brasil.

D192r Dantas, Elias Ferraz

Redes neurais artificiais aplicadas à previsão de surtos de leptospirose / Elias Ferraz Dantas. – Serra Talhada, 2018.
50 f.: il.

Orientador: Glauber Magalhães Pires

Trabalho de Conclusão de Curso (Graduação em Bacharel em Sistema de Informação) – Universidade Federal Rural de Pernambuco. Unidade Acadêmica de Serra Talhada, 2018.

Inclui referências.

1. Leptospirose. 2. Séries temporais. 3. Redes Neurais Artificiais.
I. Pires, Glauber Magalhães, coorient. II. Título.

CDD 004

**UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE SERRA TALHADA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO**

ELIAS FERRAZ DANTAS

Redes Neurais Artificiais Aplicadas à Previsão de Surto de Leptospirose

Trabalho de Conclusão de Curso julgado adequado para obtenção do título de Bacharel em Sistemas de Informação, defendida e aprovada por unanimidade em 30/08/2018 pela banca examinadora.

Banca Examinadora:

Prof. M.e Glauber Magalhães Pires
Orientador
Universidade Federal Rural de Pernambuco

Prof. M.e Hidelberg Oliveira Albuquerque
Universidade Federal Rural de Pernambuco

Prof. Dr. Richarlyson Alves D'Emery
Universidade Federal Rural de Pernambuco

AGRADECIMENTOS

A conclusão deste trabalho só foi possível graças à Deus que sempre me deu forças para continuar quando as dificuldades vinham, à minha mãe Núbia que sempre me incentivou a estudar desde sempre, à minha filha Elisa de Souza Ferraz, com seu amor e gentileza de ceder um pouco do tempo que deveria ser dedicado à ela para que eu pudesse continuar os estudos, agradeço também ao meu orientador Glauber Magalhães Pires que sempre teve a paciência e a disponibilidade para dar o suporte necessário, agradeço de todo meu coração à Marleide Gomes da Silva que me deu total apoio para que fosse possível concluir meu trabalho e sem ela eu jamais teria conseguido em tempo, e por fim agradeço aos demais amigos, em especial Katty, Fabiana, Ana Paula, e todos aqueles que acreditaram em mim e que direta ou indiretamente contribuíram para realização deste trabalho.

RESUMO

A Leptospirose é uma doença de ocorrência global, principalmente em regiões tropicais com altas incidências de chuvas e em locais com precárias condições sanitárias. Ela é uma doença considerada negligenciada, subestimada e de difícil diagnóstico por apresentar sintomas comuns à outras doenças, além de possuir uma alta taxa de letalidade ao atingir níveis graves. No Brasil, seu tratamento possui elevado custo hospitalar e a doença possui alta incidência em grandes centros urbanos. No estado de Pernambuco, em especial a cidade do Recife, destaca-se das demais localidades por possuir números significativamente maiores de registros de infecções por Leptospirose, estes dados preocupantes enaltecem a necessidade de iniciativas para apoiar a vigilância epidemiológica no combate à doença. O objetivo deste trabalho é analisar a viabilidade de utilização de Redes Neurais Artificiais aplicadas à previsão de Séries Temporais para realizar identificação antecipada do número de casos de Leptospirose a fim de demonstrar a possibilidade de utilização da técnica para previsão de surtos como um meio adicional de informação para contribuir com o planejamento de ações de prevenção em saúde, minimizando o impacto social e econômico causados pela doença.

Palavras-chave: Leptospirose, Séries Temporais, Previsão, Redes Neurais Artificiais

ABSTRACT

Leptospirosis is a disease of global occurrence, especially in tropical regions with high rainfall incidence and in places with poor sanitary conditions. It is a disease considered neglected, underestimated and difficult to diagnose because it presents symptoms common to other diseases, besides having a high lethality rate when reaching serious levels. In Brazil, its treatment has a high hospital cost and the disease has a high incidence in large urban centers. In the state of Pernambuco, Recife, it stands out from the other cities for having significantly larger numbers of records of Leptospirosis infections, this worrying data extols the need to support epidemiological surveillance in the fight against the disease. The objective of this work is to analyze the feasibility of using Artificial Neural Networks applied to the forecast of Time Series to perform early identification of the number of cases of Leptospirosis in the city of Recife / Pernambuco in order to demonstrate the possibility of using the technique for forecasting outbreaks as an additional means of information to contribute to the planning of health prevention actions, minimizing the social and economic impact caused by the disease.

Keywords: Leptospirosis, Time Series, Forecasting, Artificial Neural Networks

LISTA DE FIGURAS

Figura 1.1 – Casos confirmados de infecção por Leptospirose na cidade de Recife – PE em 2011, distribuídos por mês.	12
Figura 2.1 – Grafo Arquitetural de um <i>Perceptron</i> de múltiplas camadas com duas camadas ocultas.	23
Figura 2.2 – Exemplo de validação cruzada <i>K-Fold</i> .	24
Figura 3.1 – Representação do conjunto de dados aplicando o conceito de janelas deslizantes.	28
Figura 4.1 – Variação do RMSE para 10 repetições da configuração inicial.	32
Figura 4.2 – Erro da configuração inicial por iteração.	32
Figura 4.3 – Valores médios do RMSE por número de neurônios da camada oculta.	33
Figura 4.4 – Valores do desvio padrão entre 10 repetições por número de neurônios na camada oculta.	34
Figura 4.5 – Valores médios do RMSE por tamanho do mini-lote do gradiente descendente estocástico.	35
Figura 4.6 – Valores do desvio padrão por tamanho do mini-lote do gradiente descendente estocástico.	36
Figura 4.7 – Região não convexa possuindo dois mínimos locais sendo um deles global.	37
Figura 4.8 – RMSE médio por taxa de aprendizagem constante.	38
Figura 4.9 – RMSE Médio por taxa de aprendizagem adaptativa.	39
Figura 4.10 – Desvio padrão do RMSE por taxa de aprendizagem adaptativa.	39
Figura 4.11 – RMSE médio por valor da constante de <i>momentum</i> .	40
Figura 4.12 – Desvio padrão por valor da constante de <i>momentum</i> .	41
Figura 4.13 – Teste de 100 repetições para configuração final da rede.	42
Figura 4.14 – Resultado da previsão para o ano de 2017 (valores normalizados).	43
Figura 4.15 – Comparativo entre os casos da doença, entre 2015 e 2017 com os casos previstos para 2017.	44

LISTA DE TABELAS

Tabela 2.1 – Regra prática para interpretar o tamanho de um coeficiente de correlação	19
Tabela 4.1 – Coeficiente de Correlação de Pearson por variável meteorológica	30

LISTA DE ABREVIATURAS E SIGLAS

ARIMA *Autoregressive Integrated Moving Average*

CSV *Comma-separated values*

FII *Ficha Individual de Investigação*

FIN *Ficha Individual de Notificação*

INMET *Instituto Nacional de Meteorologia*

MAPE *Mean Absolute Percentage Error*

MLP *Multilayer Perceptron*

NARX *Nonlinear Auto-regressive with eXogenous inputs*

OMS *Organização Mundial de Saúde*

RMSE *Root Mean Square Error*

RNA *Rede Neural Artificial*

SINAN *Sistema de Informação de Agravos de Notificação*

SUMÁRIO

1	INTRODUÇÃO'	11
1.1	Contextualização	11
1.1	Objetivos	13
1.2	Motivação e justificativa	13
2	REFERENCIAL TEÓRICO	15
2.1	A Leptospirose	15
2.2	O Sistema de Informação de Agravos de Notificações (SINAN)	17
2.3	Séries Temporais	18
2.4	Coefficiente de Correlação de Pearson	19
2.5	Redes Neurais Artificiais	20
2.5.1	<i>Multilayer Perceptron</i>	22
2.5.2	Validação Cruzada <i>K-Fold</i>	24
2.6	Trabalhos relacionados	25
2.6.1	Previsão da incidência de dengue no estado do Rio de Janeiro através de Redes Neurais Artificiais	25
2.6.2	Previsão de casos de dengue em Itajaí - SC por meio de Redes Neurais Artificiais Multicamadas e Recorrentes	26
3	METODOLOGIA	27
3.1	Seleção e Pré-processamento dos Dados	27
3.2	Treinamento e Testes da Rede Neural Artificial	28
3.3	Métricas dos resultados	29
4	RESULTADOS	30
4.1	Análise da base de dados	30
4.2	Testes Iniciais da Rede Neural Artificial	31
4.3	Otimização da Rede Neural Artificial	33
4.4	Testes Finais da Rede Neural Artificial	43
5	CONCLUSÃO	45
5.1	Considerações Finais	45
5.2	Proposta de Trabalhos Futuros	46
	REFERÊNCIAS	48

1 Introdução

Neste capítulo é apresentada uma introdução sobre este trabalho. A Seção 1.1 traz uma contextualização sobre o problema abordado. Na Seção 1.2 são descritos os objetivos gerais e específicos. E na Seção 1.3 são apresentadas as motivações e justificativas para a escolha do tema.

1.1 Contextualização

A Leptospirose é uma doença infecciosa febril de início abrupto, cujo espectro pode variar de quadros assintomáticos à letais (BRASIL, 2009). Ela possui grande importância social e econômica por possuir alto custo hospitalar e elevados índices de mortalidade, podendo chegar a 50%, nos casos mais graves (BRASIL, 2014). É uma zoonose de importância mundial e apresenta alta incidência em países de clima tropical e em desenvolvimento (LEVETT, 2001), como no Brasil.

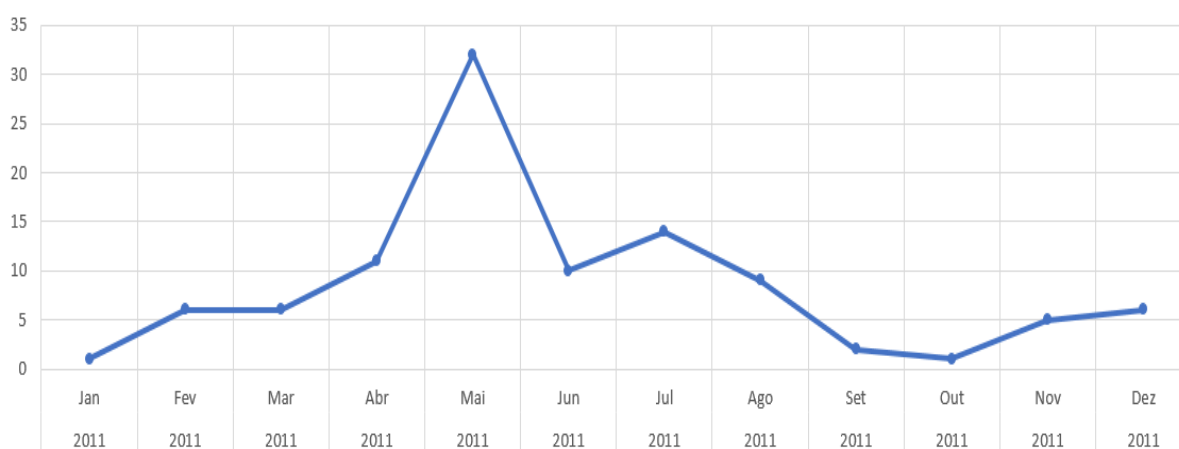
A forma de contágio se dá através do contato direto ou indireto com água ou alimentos contaminados pela urina de animais infectados, os principais reservatórios de contágio no meio urbano são os roedores, especialmente os ratos (ANDRADE et al., 2002). Estudos apontam que, dentre outros fatores, a doença está associada a baixos níveis socioeconômicos, má qualidade das infraestruturas sanitárias e que elevações das precipitações pluviométricas precedem surtos da enfermidade, uma vez que em grandes centros urbanos estes fatores combinados provocam enchentes ou alagamentos facilitando o contato do homem com a urina de animais infectados (VASCONCELOS et al., 2012; PELISSARI et al., 2011).

Segundo dados do Sistema de Informação de Agravos de Notificação (SINAN) no Brasil durante o período de 2008 a 2017 foram notificados um total de 33.804 casos confirmados de infecção por Leptospirose (desconsiderando casos ignorado, em branco, inválido ou exterior). O estado de Pernambuco, no mesmo período, foi responsável por 1170 casos da doença, 3,46% do total do país, onde Recife, dentre todas as cidades do estado, apresentou o maior número de incidências, 560 casos correspondendo à 47,86%, seguidos por Olinda com 98 ocorrências (8,37%), indicando que Recife possui um número de registros de

infecção por Leptospirose significativamente maior em relação aos demais municípios de Pernambuco.

Observando os números de registros de infecção em Recife no ano de 2011, ano em que ocorreu o maior número de casos durante o período de 2008 a 2017, percebe-se que a distribuição dos casos não é uniforme, indicando a presença de surtos, especificamente no mês de maio para o ano selecionado, conforme Figura 1.1.

Figura 1.1 – Casos confirmados de infecção por Leptospirose na cidade de Recife – PE em 2011, distribuídos por mês.



Fonte: Adaptada de SINAN (2017)

A existência de surtos da doença em determinados períodos enaltece a necessidade do apoio à vigilância epidemiológica em combate a Leptospirose, uma vez que a epidemiologia “ [...] usa métodos quantitativos para estudar a ocorrência de doenças nas populações humanas e para definir estratégias de prevenção e controle” (BONITA et al., 2010, p. 1). Uma das formas de apoio à epidemiologia é através de previsão de surtos como mostra uma matéria publicada no *site* da revista Times, em que um programa de computador administrado por epidemiologistas em Boston conseguiu alertar as principais agências do governo dos Estados Unidos sobre o surto de Ebola da África Ocidental, quatro dias antes da Organização Mundial de Saúde (OMS) anunciá-lo pela primeira vez (SCHERER, 2014).

Através da análise dos fatores que contribuem para o surgimento de surtos de uma doença e com o apoio de sistemas de informação aplicados à previsão de ocorrências é possível fornecer subsídios para que possam ser elaboradas políticas de combate a doenças, e prevenção em saúde em determinadas regiões, minimizando o impacto social e econômico.

1.1 Objetivos

Este trabalho tem como objetivo analisar a viabilidade de utilização de Redes Neurais Artificiais (RNAs) para previsão de casos de Leptospirose.

Para atingir o objetivo principal destacam-se os seguintes objetivos específicos:

- Selecionar e coletar dados de variáveis que contribuem para o surgimento de surtos de Leptospirose;
- Correlacionar as variáveis selecionadas com os casos de infecção;
- Selecionar e treinar um modelo de RNA para previsão da série temporal;
- Avaliar e otimizar o modelo de RNA;
- Analisar a qualidade da previsão.

1.2 Motivação e justificativa

Recife, dentre todas as cidades de Pernambuco, possui o maior número de casos de infecção por Leptospirose, este número é preocupante e por ser uma doença com ocorrências geralmente concentrada nos períodos chuvosos, faz-se necessário tomar medidas preventivas que antecipem a gravidade do aumento do número de casos nestes períodos de maior risco para a população de forma a preparar os serviços de saúde para demanda futura, uma das alternativas para prever estes cenários é a utilização de RNAs aplicadas à previsão de séries temporais.

A escolha das RNAs para previsão de casos de Leptospirose foi motivada devido à escassez de publicações utilizando esta abordagem, especificamente para a doença, muito embora seja frequente encontrar trabalhos com bons resultados que utilizem RNAs para prever outras doenças, como a Dengue, por exemplo.

Dessa forma, é possível contribuir com a epidemiologia através da utilização de RNAs, alimentadas com dados ambientais e epidemiológicos, como ferramenta de previsão de surtos de Leptospirose e auxiliar no planejamento da gestão pública que poderá antecipadamente realizar ações de combate à doença antes que ocorram os surtos, como campanhas de conscientização junto à população, aquisição em maiores quantidades de

medicamentos específicos para o tratamento da doença, preparação dos laboratórios para demandas maiores de realização de exames, entre outras ações.

2 Referencial Teórico

Neste capítulo é apresentada uma breve explanação sobre os temas abordados neste trabalho. Na Seção 2.1 é apresentado um resumo sobre a zoonose Leptospirose. Na Seção 2.2 é descrito o Sistema de Informação de Agravos e Notificações, principal fonte de informações sobre o registro de infecções por Leptospirose. Na Seção 2.3 é apresentado o conceito de Séries Temporais e algumas de suas características. A Seção 2.4 traz uma breve definição do coeficiente de correlação de Pearson, como calcular e suas interpretações. Na Seção 2.5 é explanado o conceito de Redes Neurais Artificiais e suas aplicações, bem como uma breve descrição do modelo selecionado, além da técnica da validação cruzada, utilizada para melhorar a qualidade da previsão dos casos de Leptospirose, objeto de estudo deste trabalho.

2.1 A Leptospirose

A Leptospirose, descrita por Adolf Weil em 1886, é uma doença de ocorrência global, negligenciada e subestimada, principalmente por apresentar sintomas comuns a outras doenças e seu diagnóstico ser dado através de testes laboratoriais que muitas vezes não estão disponíveis com fácil acesso, principalmente em países em desenvolvimento (WHO, 2003).

É uma doença febril e infecciosa de início abrupto, com período de incubação de 1 a 30 dias, possui manifestações clínicas variáveis desde quadro clínico assintomático à formas graves que podem atingir letalidade de 50%. É causada por bactérias patogênicas *Leptospiras*, e sua principal forma de contágio se dá através do contato direto ou indireto do homem com a urina de animais infectados, sendo as ratazanas, no meio urbano, o principal reservatório da doença, embora suínos e outros animais domésticos ou selvagens possam também ser reservatórios (BRASIL, 2014). A penetração da bactéria pode ocorrer através da pele lesionada, mucosas, ou através da pele íntegra submersa por longos períodos em água contaminada (BRASIL, 2009).

Conforme Brasil (2014), a Leptospirose possui alto custo hospitalar e significativo percentual de internações, ocorrendo em todos os meses do ano, principalmente nos períodos chuvosos, e há registros de infecções em todo território nacional, em áreas urbanas e rurais, com destaque nos grandes centros urbanos, devido a aglomerações de pessoas vulneráveis economicamente morando em locais com precárias condições sanitárias.

Os sintomas da doença podem ter grandes variações e são classificados pela WHO (2003) em quatro manifestações clínicas:

- I. Doença leve, como a gripe;
- II. Síndrome de Weil caracterizada por icterícia, insuficiência renal, hemorragia e miocardite com arritmias;
- III. Meningite / meningoencefalite; e
- IV. Hemorragia pulmonar com insuficiência respiratória

Diante da grande variação de sintomas, a Leptospirose pode ser comumente confundida com outras doenças, o que muitas vezes dificulta o diagnóstico, podendo ser realizado somente através de métodos específicos e dentre eles destacam-se:

- Elisa-IgM: “[...] método sensível e específico, que permite a detecção de anticorpos, a partir da 1ª semana da doença até 1 ou 2 meses [...] o resultado “reagente” confirma o caso e o resultado “não reagente” descarta o caso.” (BRASIL, 2009, p. 509)
- Microaglutinação: “[...] realizada a partir de antígenos vivos, é considerada como o exame laboratorial “padrão ouro” [...]. Além de detectar anticorpos específicos, é usada na identificação e classificação dos sorovares isolados. ” (BRASIL, 2009, p. 509).

Para o tratamento, a OMS (2003) recomenda o uso de antibióticos assim que identificado o caso suspeito, uma vez que os exames laboratoriais podem levar semanas para apresentar um resultado e em casos graves é realizada a internação do paciente sob intensos cuidados, podendo até ser submetido a hemodiálises.

A Leptospirose, de acordo com a Portaria nº 204 de 2016 do Ministério da Saúde é uma das doenças de notificação compulsória no Brasil e casos suspeitos ou confirmados devem ser notificados o mais breve possível e posteriormente inseridos no SINAN.

2.2 O Sistema de Informação de Agravos de Notificações (SINAN)

O SINAN foi concebido no início da década de 90, com o objetivo de coletar, transmitir e disseminar dados gerados rotineiramente pelo Sistema de Vigilância Epidemiológica das três esferas do governo, através de uma rede informatizada e dar suporte a vigilância epidemiológica para investigação e análise das doenças de notificação compulsórias, sendo permitido aos estados e municípios incluir outras doenças conforme a necessidade de cada região (BRASIL, 2006).

Os dados que alimentam o sistemas são oriundos das notificações e investigações de casos e doenças presentes na lista nacional de doenças de notificação compulsória através dos informações preenchidas nas Fichas Individuais de Notificação (FIN) e nas Ficha Individuais de Investigação (FII) (BRASIL, 2006). A FIN é preenchida nas unidades de saúde quando ocorrem casos suspeitos ou confirmados de doenças de notificação compulsória e a FII que é um roteiro de investigação, permite identificar a fonte de infecção e os mecanismos de transmissão da doença (IBGE, 2017).

Diversas informações são coletadas e inseridas no SINAN, elas podem variar conforme o agravo registrado, mas os principais dados armazenados são o mês, ano, estado e município de notificação; sexo, escolaridade, raça, área residencial (urbano ou rural), estado e município de residência do paciente; e data dos primeiros sintomas, município, estado e país da infecção do agravo (IBGE, 2017). A divulgação dos resultados normalmente é realizada anualmente, podendo ocorrer atrasos. A divulgação dos dados de Leptospirose mais recentes são de 2017.

A visualização dos dados é dada através de gráficos e tabelas, também sendo possível exportá-los no formato *Comma-separated values* (CSV). No entanto, o sistema é limitado a apenas exibir dados, não sendo possível nativamente realizar análises ou até mesmo projeções de caso de determinado agravo. O que acaba por ser um infortúnio considerando que o sistema possui dados temporais de doenças que poderiam ser usados juntamente com técnicas de previsão de séries temporais para realizar projeções de surtos.

2.3 Séries Temporais

Série Temporal é qualquer conjunto de observações realizadas em sequência ao longo do tempo, podendo a variável tempo ser substituída por outra, como: espaço, profundidade, entre outras. A ordem das observações é fundamental para análise da série, bem como a dependência com observações vizinhas (EHLERS, 2005). As aplicações de séries temporais são inúmeras e podem ser utilizadas em diversas áreas de pesquisa. São alguns exemplos de Séries Temporais: temperatura média mensal de uma cidade, batimentos cardíacos de um paciente, registro anual de notificações de casos de Tuberculose, etc.

Séries Temporais podem ser discretas, quando as observações são feitas em períodos de tempos iguais ou contínuas quando as observações estão distribuídas continuamente no tempo (MORETIN; TOLOI, 2006). Para Moretin e Toloi (1981), elas são compostas por três componentes:

- Tendência: quando há um deslocamento do sentido da série, crescente ou decrescente, por exemplo;
- Sazonalidade ou Ciclo: Quando há periodicidade de movimentos na série, para períodos até um ano, denomina-se sazonalidade, acima disso serão Ciclos; e
- Componente aleatório ou residual: Processo puramente aleatório, que não pode ser medido.

Os principais objetivos para a análise de Séries Temporais, segundo Ehlers (2005), podem ser representados como:

- Descrever as propriedades das séries, como padrões, variações sazonais, etc;
- Explicar a variação de uma série com base na variação de outra série;
- Prever dados futuros com base em dados passados, minimizando os erros de previsão, uma vez que há incertezas nas previsões; e
- Controlar processos, como, por exemplo: controle estatístico de qualidade.

Para o objetivo de predição de observações em um dado intervalo de tempo, existem diversas abordagens no campo de séries temporais, com grande destaque ao modelo de Box e Jenkins (1976), conhecido como o Modelo Autoregressivo Integrado de Médias Móveis (ARIMA) e suas variações. Neste trabalho, o método para previsão de dados futuros, no caso para prever a incidência de Leptospirose, será a utilização de RNAs, que é um ramo da Inteligência Artificial.

A adição de outras variáveis à série de estudo, contribui para uma melhor capacidade de previsão. Entretanto, para que esta adição possa melhorar a previsão realizada, é imprescindível, existir um grau de relação entre as variáveis selecionadas como informações de entrada. Uma das formas de identificar dependências entre estas variáveis, ou seja, analisar o relacionamento entre elas, é o método conhecido como coeficiente de correlação de Pearson.

2.4 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson, também conhecido como coeficiente de correlação produto-momento de Pearson, tem este nome em homenagem ao estatístico inglês Karl Pearson (1857-1936) e é uma medida da força e direção de uma relação linear entre duas variáveis (LARSON, 2010). Representado pela letra r sua fórmula é descrita na Equação 2.1.

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n \sum x^2 - (\sum x)^2} \sqrt{n \sum y^2 - (\sum y)^2}} \quad (2.1)$$

Onde n representa o número de pares de dados, e x e y as variáveis a serem correlacionadas.

Segundo Figueiredo Filho (2009), o coeficiente de correlação de Pearson varia entre -1 e 1, onde o sinal indica a direção (positiva ou negativa) de uma relação e o valor indica a força desse relacionamento. Valores próximos dos extremos indicam maiores níveis de correlação, e valores próximos de zero indicam a inexistência de correlação entre as variáveis selecionadas.

O coeficiente de correlação de Pearson pode ser classificado, segundo sua interpretação, conforme a Tabela 2.1.

Tabela 2.1 – Regra prática para interpretar o tamanho de um coeficiente de correlação

Valor da correlação	Interpretação
0.9 a 1.0 (-0.9 a -1.0)	Correlação muito forte positiva (negativa)
0.7 a 0.9 (-0.7 a -0.9)	Correlação forte positiva (negativa)
0.5 a 0.7 (-0.5 a -0.7)	Correlação moderada positiva (negativa)
0.3 a 0.5 (-0.3 a -0.5)	Correlação fraca positiva (negativa)
0.0 a 0.3 (0.0 a -0.3)	Correlação insignificante

Fonte: Adaptada de Mukaka (2012)

Variáveis que possuam maiores níveis de correlação com a série de estudo, no caso deste trabalho o número de registro de infecção por Leptospirose, servirão melhor como entrada para uma RNA, e impactarão diretamente na qualidade das previsões realizadas.

2.5 Redes Neurais Artificiais

Para Braga et al. (2011), as RNAs são sistemas paralelos distribuídos, inspirados no cérebro humano, compostos por unidades de processamento simples capazes de realizar cálculos de determinadas funções matemáticas, estas unidades são dispostas em camadas e interligadas por conexões que, na maioria dos modelos, estão associadas à pesos que armazenam o conhecimento representado no modelo.

Haykin (2001) descreve uma RNA como uma máquina projetada para modelar a maneira como o cérebro realiza um determina função e pode ser implementado em componentes eletrônicos ou simulado em computador. Ainda segundo o autor:

Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para o uso. Ela se assemelha ao cérebro em dois aspectos:

1. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.
2. Forças de conexão entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

Assim como no cérebro humano, uma RNA é composta de elementos de processamento, denominados neurônios que estão conectados entre si paralelamente através de sinapses, essas conexões são ponderadas por pesos sinápticos que na maioria dos casos retém o conhecimento adquirido pela rede. Haykin (2001) ainda afirma que a forma de aprendizagem de um modelo de RNA está diretamente ligada à maneira como os elementos de processamento (ou neurônios) estão estruturados na rede e classifica estas estruturas em basicamente três classes de arquitetura de rede:

- Rede Alimentada Adiante de Camada Única: A forma mais simples de uma rede em camadas, onde uma camada de entrada de nós de fonte repassa valores diretamente à uma camada de saída composta por neurônios de saída, porém não o inverso, o que garante a alimentação adiante. A referência à camada única diz respeito a apenas a camada de saída, composta por neurônios;

- Rede Alimentada Adiante com Múltiplas Camadas: De forma semelhante a rede de camada única o fluxo dos dados é também no sentido da camada de entrada para camada de saída, entretanto nesta arquitetura existem uma ou mais camadas ocultas entre elas, composta por neurônios ocultos, a adição de camadas extras permite a rede resolver problemas mais complexos; e
- Redes Recorrentes: Podem possuir uma ou mais camadas e diferem das redes de múltiplas camadas no que diz respeito ao fluxo dos dados na rede, por existir laços de realimentação, onde é possível que saída de um neurônio alimente a entrada de outros neurônios, e laços de autorrealimentação onde a saída de um neurônio é a alimentada à sua própria entrada.

Quanto ao aprendizado, no contexto de RNA, Haykin (2001, p. 75) apresenta uma definição adaptada de Mendel e McLaren (1970):

Aprendizagem é um processo pelo qual os parâmetros livres de uma rede neural são adaptados através de um processo de estimulação pelo ambiente no qual a rede está inserida. O tipo de aprendizagem é determinado pela maneira pela qual a modificação dos parâmetros ocorre.

Ainda segundo o autor, não existe um algoritmo único de aprendizagem e estes diferem na forma como é realizado o ajuste dos parâmetros da rede e a forma como a rede interage com o ambiente. Braga et al. (2011) classificam o processo de aprendizagem em três grandes paradigmas:

- Aprendizado Supervisionado: Nesta forma de aprendizado, são apresentados à rede entradas e saídas desejadas fornecidas por um supervisor, e a rede, por sua vez, tenta se ajustar de modo que possa minimizar a diferença entre o valor de saída encontrada por ela e o valor esperado;
- Aprendizado Não Supervisionado: Não há existência de um supervisor e somente são apresentados os padrões a rede sem qualquer saída esperada. A rede se ajusta de modo a encontrar características semelhantes nos padrões apresentados para agrupá-los ou classificá-los; e
- Aprendizado Por Reforço: Assemelha-se ao aprendizado supervisionado, entretanto não é apresentado o valor correto para a saída de um padrão apresentado, apenas se esta saída foi correta ou não. A rede por sua vez tenta se aproximar da saída esperada após consecutivas ações de acertos.

Os diferentes tipos de arquitetura de rede e aprendizado permitem que as RNAs possam resolver inúmeros problemas, Braga et al. (2011) exemplificam algumas das tarefas que podem ser realizadas por uma RNA como:

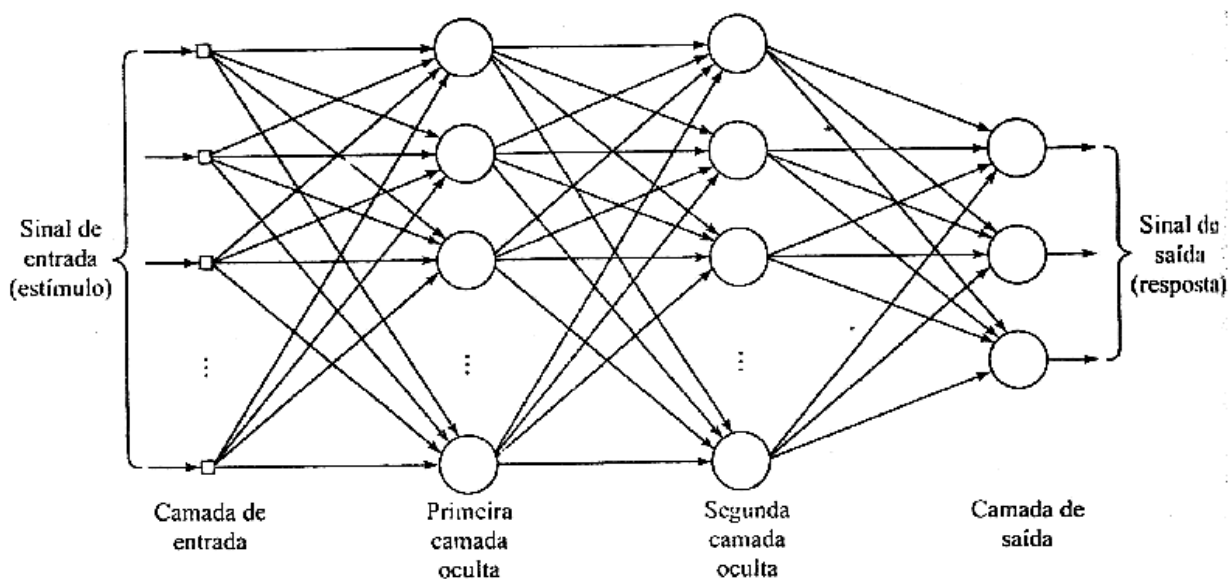
- Classificação: Atribui um padrão desconhecido a uma classe entre um grupo de classes conhecidas;
- Categorização (ou Reconhecimento de padrão): Criar agrupamentos de padrões apresentados a rede;
- Aproximação: Entradas e saídas são apresentadas a rede, e esta tem por objetivo mapear funções contínuas a partir das variáveis de entrada;
- Previsão: Estimativa de situações futuras baseadas em dados atuais ou dados anteriores; e
- Otimização: Minimização ou maximização de funções de custo.

Existem uma variedade de modelos de RNA, com diferentes tipos de arquitetura, que se adaptam de formas diferentes para cada tipo de tarefa, alguns exemplos de modelos de RNA são: *Self Organized Maps* (SOM), *Radial Basis Function* (RBF), Rede de Hopfield, entre outros. Este trabalho irá utilizar um modelo de múltiplas camadas conhecido como *Multilayer Perceptron* (MLP), também utilizado nos trabalhos relacionados, descritos na Seção 2.5, e que apresentaram resultados satisfatórios na predição da incidência de doenças.

2.5.1 *Multilayer Perceptron*

Silva et al. (2010) definem o modelo MLP como sendo um modelo de múltiplas camadas, possuindo uma ou mais camadas ocultas entre a camada de entrada e a camada de saída, possuindo alimentação *feedforward* – sem realimentação – e aprendizado supervisionado. A Figura 2.1 ilustra, através de um grafo orientado, a arquitetura de uma rede MLP, onde cada nó do grafo representa neurônios.

Figura 2.1 – Grafo Arquitetural de um *Perceptron* de múltiplas camadas com duas camadas ocultas.



Fonte: Haykin (2001, p. 186)

O fluxo do sinal é propagado na rede da esquerda para direita, partindo da camada de entrada passando por todas as camadas seguintes em direção a camada de saída. A primeira camada oculta é constituída de neurônios e é alimentada a partir da camada de entrada, composta de unidades sensoriais, a saída da primeira camada oculta alimenta a segunda camada oculta e assim sucessivamente até que se chegue a camada de saída da rede, composta por neurônios de saída, que apresentam a resposta da rede ao padrão fornecido (HAYKIN, 2011).

Redes MLP possuem diversos algoritmos de treinamento, em sua maioria supervisionados, o mais conhecido é o algoritmo *Backpropagation* (RUMELHART et al., 1986) que utiliza pares (entrada, saída desejada) para ajustar o peso da rede por meio de um mecanismo de correção de erros. O *Backpropagation* possui duas fases: A primeira fase define a saída da rede a partir de uma padrão apresentado e é chamada de fase *forward*, na segunda fase, fase *backward*, é utilizada a saída da rede e a saída desejada para ajustar os pesos das conexões. Este processo é repetido durante a fase de treinamento, de modo a minimizar o erro produzido na saída dos padrões apresentados, até um certo limite ou até atingir um número pré definido de iterações (BRAGA et al., 2011).

O algoritmo *Backpropagation*, é o mais comumente utilizado em redes MLP, porém se houver treinamentos em excesso, pode fazer com que a rede memorize os dados de treinamento e acabe comprometendo a capacidade de generalização, não sendo capaz de realizar projeções corretas. Haykin (2001) sugere como técnica para contornar este tipo de situação a utilização de Validação Cruzada.

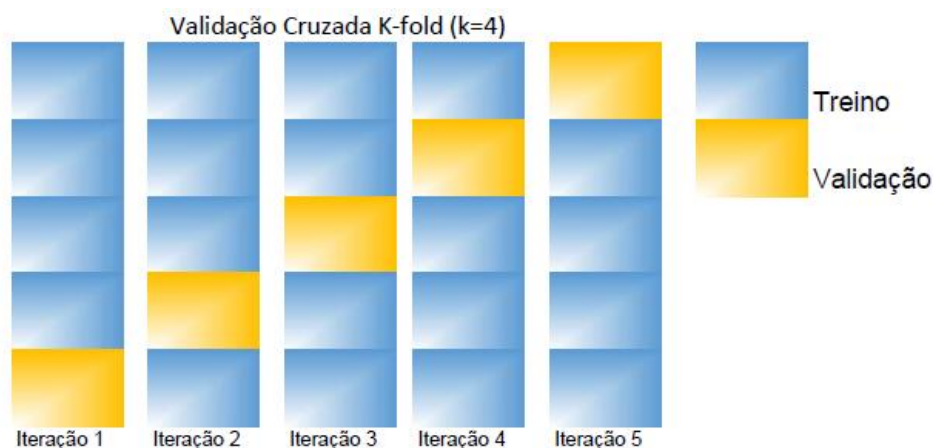
2.5.2 Validação Cruzada *K-Fold*

Ao treinar uma RNA, espera-se obter um modelo treinado com a melhor capacidade de generalização, capaz de mapear corretamente as entradas-saídas e produzir resultados satisfatórios para um conjunto de dados não apresentados na fase de teste (HAYKIN 2001). Para isto, Haykin (2001) ainda descreve a validação cruzada como uma ferramenta padrão da estatística que divide o conjunto de dados em treinamento e teste, adicionalmente divide o conjunto de treinamento em mais dois subconjuntos: treino e validação, tendo como principal atrativo garantir uma melhor generalização da RNA.

Na validação cruzada, a RNA é treinada de maneira usual, inicialmente com subconjunto de treino, com o diferencial que após um determinado número de iterações (épocas) o treinamento é interrompido e a rede é validada com o subconjunto de validação. O processo é repetido enquanto a curva de aprendizagem continuar a decrescer (HAYKIN, 2001). Este processo de treinamento contribui para evitar o *overfitting*, situação em que a rede memoriza os dados de treino ao invés de extrair características dos padrões apresentados, e então, não sendo capaz de produzir uma saída satisfatória para um conjunto novo (BRAGA et al., 2011).

Uma variação da validação cruzada, conhecida como *K-Fold*, separa o conjunto de treino em k “*folds*” e em seguida retirar um destes “*folds*” para validação. Na próxima iteração o conjunto de dados de validação é devolvido e outro é retirado para validação, o processo se repete até que todos os conjuntos tenham sido utilizados, e o desempenho do modelo é medido como a média do erro quadrado obtido em todas as iterações (HAYKIN, 2001). A Figura 2.2 ilustra a divisão de um conjunto de dados utilizando validação cruzada *K-Fold*.

Figura 2.2 – Exemplo de validação cruzada *K-Fold*.



Fonte: Adaptada de Bisgin et al. (2011).

Dessa maneira evita-se que a rede se adapte a um conjunto de dados na fase de treinamento e não seja capaz de produzir a generalização adequada e, conseqüentemente comprometa a capacidade de previsão de surtos de Leptospirose.

2.6 Trabalhos relacionados

É possível encontrar diversos trabalhos que utilizem RNAs como técnica de predição de incidências de doenças, sendo a Dengue uma doença com diversos exemplos, como nos trabalhos relacionados descritos nesta seção. O primeiro trabalho é uma Monografia que tem como objetivo auxiliar a tomada de decisão da gestão pública em saúde e demonstra resultados satisfatórios na previsão de casos de Dengue no Rio de Janeiro utilizando MLP. O segundo trabalho faz uma comparação entre uma rede recorrente e uma MLP para prever casos de Dengue em Itajaí-SC, concluindo que a MLP demonstrou um desempenho melhor.

2.6.1 Previsão da incidência de dengue no estado do Rio de Janeiro através de Redes Neurais Artificiais

A monografia do curso de Especialização em *Business Intelligence* do Departamento de Engenharia Elétrica da PUC/RIO, apresentada por Dos Santos (2013), teve como objetivo auxiliar na previsão de casos de dengue em alguns municípios do estado do Rio de Janeiro, predizendo 5 semanas epidemiológicas futuras.

Para alcançar seu objetivo, o autor coletou dados de casos de dengue de janeiro de 2008 à dezembro de 2012 dos municípios de Rio de Janeiro, São João de Meriti e Belford Roxo no estado do Rio de Janeiro. Após a coleta, foram efetuados pré-processamento nos dados para alimentar a RNA. O modelo selecionado foi o *MLP*, diversos testes foram realizados utilizando o software *MathLab* R2011b para encontrar a configuração ideal para as cidades estudadas, o algoritmo de aprendizado usado foi o *Backpropagation* sem *momentum* e como critério de parada o número de épocas de treinamento entre 20 a 80. Após análise dos resultados o autor conseguiu chegar a números satisfatórios de previsão de incidência de dengue, atingindo uma taxa de Erro Médio Percentual Absoluto (MAPE, do inglês *Mean Absolute Percentage Error*) de 6,46% no melhor dos resultados.

O autor concluiu que RNAs possuem capacidade de previsão de incidência de Dengue e demonstram resultados satisfatórios, mesmo com as adversidades encontradas como poucos padrões apresentados à rede, poucos dados para teste e validação e o pouco tempo de implementação que ele dispunha.

2.6.2 Previsão de casos de dengue em Itajaí - SC por meio de Redes Neurais Artificiais Multicamadas e Recorrentes

O artigo publicado por Mittelman e Soares (2017) na *Computer on the Beach* 2017, teve como objetivo implementar um modelo para prever a incidência de Dengue na cidade de Itajaí – SC, utilizando RNAs. Como informações para alimentar a rede, os autores utilizaram dados de casos de dengue no período de julho de 2010 a junho de 2015 e dados meteorológicos do mesmo período, como: precipitação total, temperatura máxima e mínima, temperatura média do ar e umidade relativa do ar. Após a coleta, foi realizada a normalização dos dados para servir de entrada à rede e utilizada a validação cruzada para selecionar o conjunto de treino, teste e validação.

Foram selecionados dois modelos de RNAs para tentar realizar a previsão, um modelo recorrente e um de múltiplas camadas, o *Nonlinear Auto-regressive with eXogenous inputs* (NARX) e MLP, respectivamente. Diversas configurações foram testadas a fim de encontrar a que apresentasse melhores resultados e o índice para avaliar a qualidade da previsão foi a Raiz do Erro Quadrático Médio (RMSE, do inglês *Root Mean Squared Error*) e o MAPE. Após as comparações, a rede que obteve o melhor desempenho foi a MLP treinada com 8 neurônios na camada oculta, função de ativação tangente hiperbólica e o algoritmo de Levenberg-Marquardt, com erros RMSE e MAPE no conjunto de testes de $2,68 \cdot 10^{-3}$ e 0,0369%, respectivamente.

Após as análises dos resultados, os autores concluíram que é possível realizar a previsão de casos de dengue utilizando RNAs, especificamente o modelo MLP, porém salientam que o trabalho realizado apenas é capaz de prever casos da doença na área específica de estudo, no caso em Itajaí-SC e, como toda modelagem matemática, é necessário levar em consideração as características únicas de cada região.

3 Metodologia

Neste capítulo é descrito a metodologia utilizada para realização deste trabalho, que será dividido em três partes: Seleção e pré-processamento dos dados (Seção 3.1), treinamento e testes da RNA (Seção 3.2) e, por fim, as métricas escolhidas para avaliar os resultados (Seção 3.3).

3.1 Seleção e Pré-processamento dos Dados

Para realizar a previsão de casos de Leptospirose, utilizaram-se dados de Recife-PE, coletados no período de 01/01/2008 à 31/12/2017 e distribuídos em intervalos mensais, formando séries temporais com 120 amostras. Além dos números passados de casos de infecção por Leptospirose, extraídos do SINAN, foram selecionados também dados pluviométricos da cidade obtidos através do Instituto Nacional de Meteorologia (INMET), uma vez que diversos trabalhos apontam as chuvas como um dos principais contribuintes no aumento de casos da doença (VASCONCELOS et al., 2012; MASI, 2014). Para este trabalho foram escolhidos duas variáveis climatológicas: o índice de precipitação pluviométrica, que indica a quantidade em milímetros de chuva em um determinado local e a umidade do ar por contribuir na dificuldade de evaporação da água. A escolha destas variáveis se deu através da aplicação de um método de correlação entre as séries temporais, o método empregado para avaliar o grau de correlação foi o Coeficiente de Correlação de Pearson.

O próximo passo, após a seleção das variáveis, foi a normalização dos dados de entrada da RNA, convertendo-os nos intervalos das funções de ativação dos neurônios para garantir uma melhor generalização do modelo e evitar que o dados de entrada gerem efeitos indesejados à saída da rede. Todos os valores de entradas foram convertidos para o intervalo [0,1], em que o novo valor de uma amostra x , é definida como sendo a diferença entre o valor de x e o valor mínimo da série dividido pela diferença entre o maior e menor valor da série, conforme demonstrado na Equação 3.1.

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (3.1)$$

3.2 Treinamento e Testes da Rede Neural Artificial

As RNAs recebem um conjunto de dados de entrada e geram uma saída. No caso do modelo supervisionado MLP, utilizado neste trabalho, é necessário que haja uma saída desejada para um determinado conjunto de dados para que seja possível efetuar o cálculo do erro e atualizar os pesos da rede.

Para montar os pares de dados de entrada e saídas desejadas foi aplicado o conceito de Janela Deslizante, dividindo a série em um conjunto menor como sendo os dados de entrada e o valor seguinte a este conjunto como sendo a saída esperada pela rede. Para o próximo par de entrada e saída, é movido o ponto inicial deste conjunto para o valor seguinte ao ponto inicial do conjunto anterior, repetindo este processo sucessivamente como mostra a Figura 3.1

Figura 3.1 – Representação do conjunto de dados aplicando o conceito de janelas deslizantes.



Fonte: Elaborada pelo autor

Em seguida o conjunto de dados foi dividido em duas partes separadas, sendo uma dedicada para fase de treinamento e outra para fase de teste nas proporções 90% e 10%, respectivamente. O conjunto de treinamento foi subdividido ainda, em mais duas partes, na proporção 90% e 10%, correspondendo ao conjunto de treino e validação respectivamente.

Com a utilização da validação cruzada é possível treinar a rede com um conjunto de dados e validar o resultado a cada iteração. Após o fim do processo de treinamento é possível executar o teste com o conjunto de teste previamente separado e avaliar o desempenho final da rede.

Devido a sua popularidade na área de aprendizado de máquina, com inúmeras bibliotecas e a vasta quantidade de material de estudo disponível em livros e na *Internet*, todo o trabalho foi realizado utilizando a linguagem de programação *python* (PEDREGOSA, 2011) na versão 3.6.5. A fim de agilizar o processo de desenvolvimento e concentrar-se na avaliação dos resultados foram utilizadas as seguintes bibliotecas para *python*: *Pandas* v0.22.0 para leitura e manipulação dos dados, *scikit-learn* v0.19.1 para implementação da rede neural e coleta de métricas dos resultados (SCIKIT-LEARN, 2011), e *matplotlib* v2.2.2. Para os gráficos também foi utilizado o *Microsoft Office Excel* 2016.

A biblioteca *scikit-learn* possui uma classe chamada *MLPRegressor* que implementa o modelo MLP, treinando com o algoritmo *Backpropagation*, com função identidade na camada de saída, apropriada para problemas de regressão.

A classe *MLPRegressor* utiliza uma série de parâmetros para otimizar a capacidade generalização da rede para um modelo preditivo, e diversas combinações foram testadas durante o treinamento para encontrar a melhor configuração da rede que fosse capaz de realizar previsões satisfatórias de quando ocorreriam surtos de Leptospirose em Recife-PE.

3.3 Métricas dos resultados

A métrica utilizada para avaliar o desempenho da rede em relação às previsões realizadas foi o RMSE, selecionada por ser uma das medidas padrões de avaliar o desempenho de predição de séries temporais (BROWNLEE, 2017). O RMSE calcula a raiz dos erros médios quadráticos entre a diferença dos valores reais e valores obtidos pela previsão, como descrito na Equação 3.2.

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (x_i - \hat{x}_i)^2}{N}} \quad (3.2)$$

Onde, o termo $(x_i - \hat{x}_i)^2$ é a diferença entre o valor real e o valor da predição, respectivamente, elevado ao quadrado e N representa a quantidade de valores da saída. O RMSE possui a mesma unidade do valor das predições, e um valor igual a zero indica a ausência de erros..

4 Resultados

Neste capítulo é descrito os resultados obtidos para cada teste da rede neural e as escolhas de otimização. A Seção 4.1 apresenta uma análise da base de dados escolhida. A Seção 4.2 mostra o desempenho da rede para uma dada configuração inicial. A Seção 4.3 descreve as etapas realizadas para otimizar o desempenho da rede e, por fim, a Seção 4.4 exibe os resultados da rede otimizada aplicada ao conjunto de testes e avalia a qualidade da previsão.

4.1 Análise da base de dados

Inicialmente foram selecionadas diversas variáveis meteorológicas, a fim de encontrar as que possuísem um maior grau de correlação com os casos de infecção da doença no mesmo período, e descartar aquelas que possuísem grau insignificante para evitar que impactassem na capacidade de generalização da rede. O coeficiente de correlação r para cada variável em relação aos casos de infecção por Leptospirose pode ser observado na Tabela 4.1.

Tabela 4.1 – Coeficiente de Correlação de Pearson por variável meteorológica

Variável	Coeficiente de correlação
Precipitação total	0,64403767329
Umidade relativa média	0,62406728622
Nº de dias de com precipitação	0,47825340306
Nebulosidade média	0,43690197387
Evapotranspiração real BH	0,23463841000
Direção do vento predominante	0,03797993656
Pressão atmosférica média	0,03520069523
Velocidade do vento máxima média	-0,01908416361
Pressão atmosférica do nível do mar média	-0,02923596885
Temperatura mínima média	-0,33683577071
Velocidade do vento media	-0,39821422950
Insolação total	-0,45733530367
Temperatura máxima média	-0,46883886009
Temperatura compensada média	-0,47627988950
Evapotranspiração potencial BH	-0,49927713834
Evaporação do piche	-0,51796663517

Fonte: Elaborada pelo autor

Como mostrado na Tabela 4.1, é possível perceber que nenhuma das variáveis possui um coeficiente de correlação considerado positivo forte ($r > 0,7$). Entretanto, foi selecionado as duas que mais se aproximaram deste valor, neste caso as variáveis “Precipitação total” e “Umidade relativa média”, sendo consideradas de correlação positiva moderada (r entre 0,5 e 0,7).

Após a escolha das variáveis a serem utilizadas, foi necessário adaptar o conjunto de dados para alimentar a rede neural, para isto foi aplicado o conceito de Janela Deslizante. Como dados de entrada foram selecionados os números de casos, a precipitação total, e a umidade relativa média, e como saída apenas a série contendo o número de casos.

4.2 Testes Iniciais da Rede Neural Artificial

Para encontrar a configuração ideal da rede, diversos parâmetros foram testados. A maioria dos parâmetros da configuração inicial foram escolhidos intuitivamente, uma vez que seriam ajustados posteriormente para encontrar uma configuração que produzisse uma saída satisfatória.

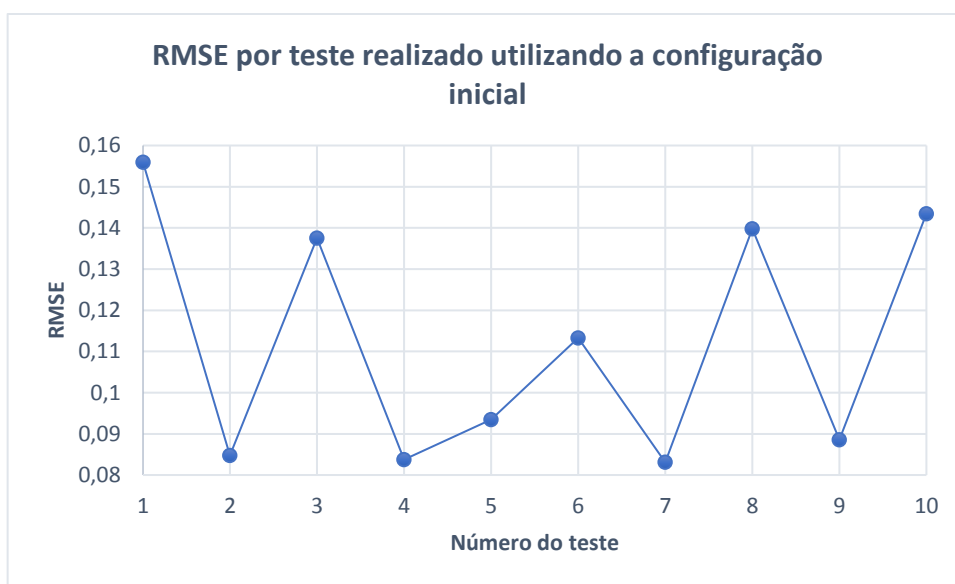
Inicialmente, devido a pequena quantidade de dados disponível para treinamento, os dados e a rede foram configurados da seguinte forma: a base de dados foi dividida em 10% para testes e 90% para treino, destes retirou-se 10% dos dados para validação. Após utilização do conceito de janela deslizante, foi produzido, como dados a serem alimentados na camada de entrada da RNA, um vetor contendo 9 valores, correspondendo ao tamanho de janela 3 aplicado em cada uma das 3 séries (casos, precipitação e umidade).

Para a RNA, foram usados as seguintes características: 9 neurônios na camada de entrada; 10 neurônios na única camada oculta; uma função logística para ativação das camadas intermediárias, uma vez que os dados foram normalizados para o intervalo $[0,1]$ que correspondem a saída da função logística; uma função identidade para ativação da camada de saída; um número máximo de iterações igual a 1000 e um gradiente descendente estocástico como algoritmo de aprendizado porém, no primeiro momento, foi utilizado todo o conjunto de treino para atualização dos pesos, com taxa de aprendizagem constante e igual a 0,1 e sem *momentum*.

Com esta configuração inicial, a rede demonstrou resultados insatisfatórios, com respostas distantes do esperado e variando bastante o resultado entre repetições dos testes com

as mesmas configurações. Para observar esta oscilação, foram realizadas 10 execuções com a mesma configuração, tendo a rede chegado a respostas diferentes em cada uma delas. O resultado da rede para cada uma das 10 execuções variava conforme o vetor inicial de pesos que, por inicializarem de maneira aleatória, levavam a rede a obter saídas muito diferentes sem que nenhum parâmetro fosse alterado. A situação é apresentada na Figura 4.1, onde é possível perceber diferentes valores para o RMSE a cada teste realizado.

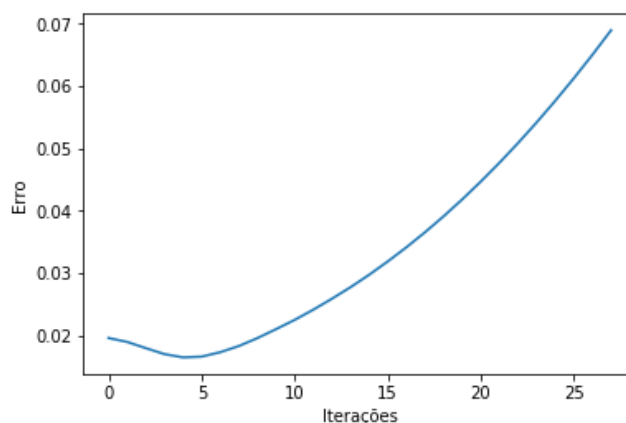
Figura 4.1 – Variação do RMSE para 10 repetições da configuração inicial.



Fonte: Elaborada pelo autor

Ainda no mesmo teste, observando a curva do erro por iteração para uma das 10 repetições realizadas, escolhida ao acaso, é possível perceber que o algoritmo não convergia de maneira adequada, e o erro ao invés de diminuir estava crescendo com o passar das épocas fazendo com que o algoritmo encerrasse o treinamento sem convergir para um valor, conforme demonstra a Figura 4.2.

Figura 4.2 – Erro da configuração inicial por iteração.



Fonte: Elaborada pelo autor

4.3 Otimização da Rede Neural Artificial

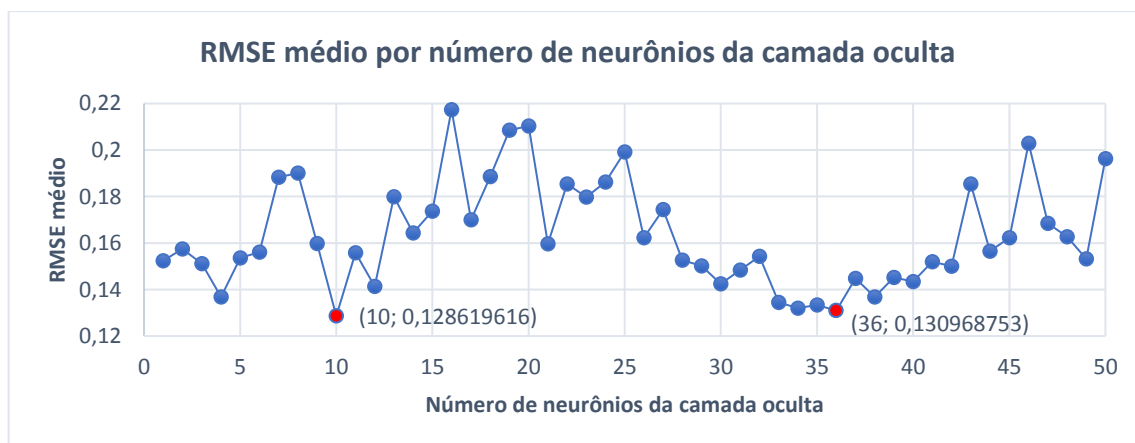
Observando os resultados do teste inicial, o modelo estava encontrando dificuldades de convergência, e como demonstrado na Figura 4.1, variava seu resultado conforme o vetor de pesos inicial. Para melhorar o desempenho do modelo foi necessário realizar ajustes nos dados e nos parâmetros da rede.

Para o ajuste dos dados, uma possibilidade seria alterar o tamanho da janela, de modo a aumentar o número de dados que justificam a previsão de um determinado mês. Porém devido a limitação da quantidade de amostras das séries, o tamanho da janela foi mantido em 3, ou seja, três meses anteriores justificam o mês atual, valores maiores que estes reduziram a quantidade de amostras ao deslocar o ponto de origem da série, consequentemente comprometendo a capacidade da rede de extrair as características dos dados apresentados.

Dessa forma, as tentativas de melhorar o desempenho do modelo foram direcionadas para os ajustes dos parâmetros da rede. A primeira tentativa de aprimorar a saída da rede neural foi encontrar um número ideal de neurônios da camada oculta que gerasse melhores respostas, e a partir daí começar a ajustar os demais parâmetros.

Nesta etapa foram testados valores para o número de neurônios na camada oculta no intervalo entre [1,50]. Devido as oscilações nas saídas, para configurações idênticas, terem sido significativamente grandes, cada quantidade de número de neurônios foi testada por 10 vezes e obtido o valor médio do RMSE, de modo a reduzir o impacto da aleatoriedade inicial dos pesos na saída da rede. O resultado do teste do número de neurônios pode ser observado na Figura 4.3.

Figura 4.3 – Valores médios do RMSE por número de neurônios da camada oculta.



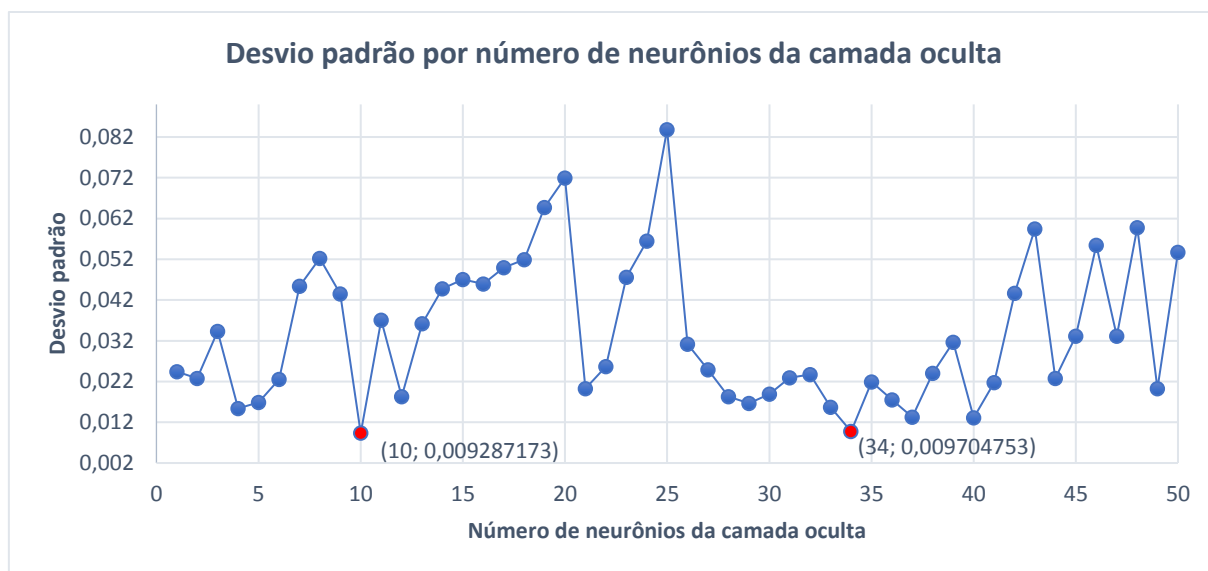
Fonte: Elaborada pelo autor

A Figura 4.3 apresenta o RMSE médio das 10 repetições de cada teste do número de neurônios. O melhor resultado obtido foi para o teste com 10 neurônios na camada oculta com RMSE médio de 0,1286. Contudo, ainda na Figura 4.3, é possível perceber que outros pontos apresentaram resultados próximos do teste com 10 neurônios, como é o caso do ponto com 36 neurônios que obteve um RMSE médio de 0,1309.

Considerando que, como visto na Figura 4.1, há uma grande oscilação nas saídas para uma mesma configuração, e que no teste anterior houve valores de saída próximos, o menor valor do RMSE médio não poderia ser utilizado como única base para escolha do número de neurônios da camada oculta. Logo, um critério adicional foi inserido como base para escolha do número de neurônios e demais parâmetros: a uniformidade das saídas obtidas nas 10 repetições realizadas.

A medida escolhida para avaliar esta uniformidade foi o desvio padrão, que é uma medida de dispersão que mede o quanto um conjunto de dados é uniforme (GOUVEIA, 2018). Quanto mais próximo de zero, mais uniforme será o conjunto de dados e, para o caso dos testes realizados, menor será a oscilação presente nas saídas do teste.

Figura 4.4 – Valores do desvio padrão entre 10 repetições por número de neurônios na camada oculta.



Fonte: Elaborada pelo autor

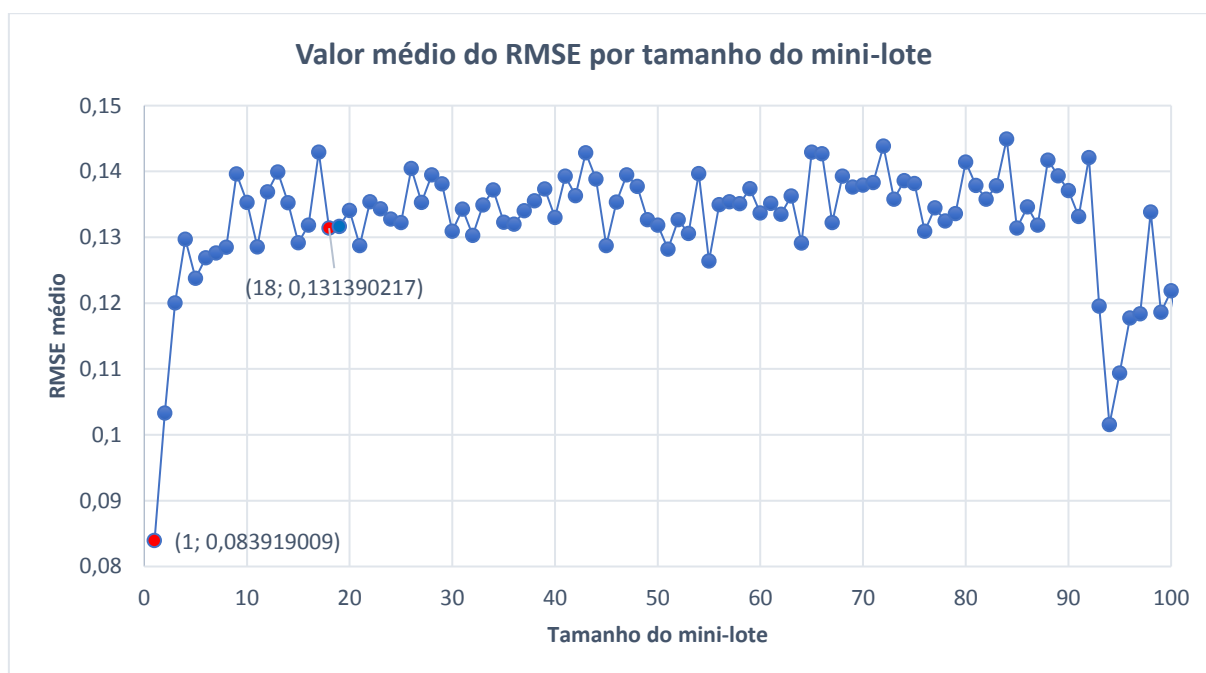
Observando a Figura 4.4, o menor desvio padrão encontrado foi no ponto com 10 neurônios na camada oculta com desvio padrão de 0,0092, seguido pelo ponto com 34 neurônios com desvio padrão de 0,0097. Como os testes realizados com 10 neurônios apresentaram o menor RMSE e também o menor desvio padrão, foi mantido este, como o

valor para o número de neurônios da camada oculta e prosseguido com os demais ajustes de otimização.

O próximo passo, após selecionado o número de neurônios, foi verificar o impacto do tamanho do lote do gradiente descendente estocástico. Diferentemente do gradiente descendente tradicional, que utiliza todo o conjunto de treino para executar um passo na otimização, o gradiente descendente estocástico utiliza apenas uma certa quantidade dos dados de treino para realizar a otimização por cada iteração. Este processo garante uma maior agilidade na atualização dos pesos, uma vez que não é necessário treinar com todos os dados da amostra para executar um único passo do gradiente, além de permitir que a rede explore outras regiões na curva do erro (FERREIRA, 2015).

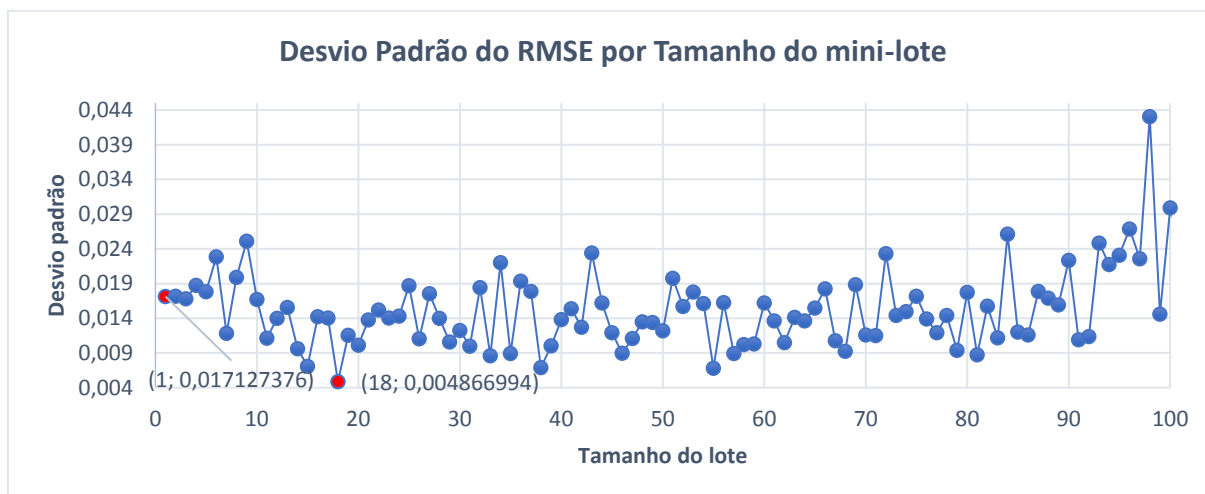
Ao utilizar o gradiente descendente estocástico, um parâmetro adicional precisa ser configurado, que é a porção de dados de treinamento que serão utilizados para computar o gradiente, denominado mini-lote. Encontrar o valor ideal para o tamanho do mini-lote é um processo empírico e foram testados valores entre 1 e o máximo de amostras, repetidos 10 vezes cada, mantendo-se as demais configurações, a Figura 4.5 mostra o RMSE médio e a Figura 4.6 mostra o desvio padrão para cada valor de mini-lote testado.

Figura 4.5 – Valores médios do RMSE por tamanho do mini-lote do gradiente descendente estocástico.



Fonte: Elaborada pelo autor

Figura 4.6 – Valores do desvio padrão por tamanho do mini-lote do gradiente descendente estocástico.



Fonte: Elaborada pelo autor

Após os testes, conforme pode ser visualizado na Figura 4.5, a rede produzia resultados significativamente melhores com tamanho do mini-lote igual a 1, piorando as respostas conforme se aumentavam o seu tamanho, e novamente apresentando melhoras para valores muito próximos do máximo, porém ainda longe do resultado obtido para o tamanho igual a 1.

A Figura 4.6, mostra o desvio padrão das 10 repetições para cada um dos mini-lotes testados, tendo o ponto com tamanho igual a 18 como sendo a saída mais estável, apresentando um desvio padrão igual a 0,0048.

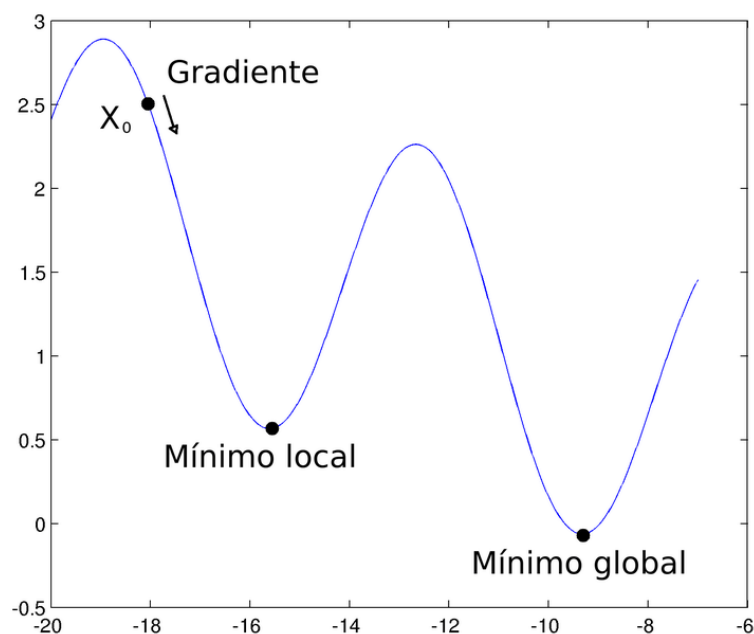
Destacados nas Figuras 4.5 e 4.6, os pontos com tamanho de mini-lote igual a 1 e 18 apresentaram o menor RMSE médio e o menor desvio padrão respectivamente. A partir desta observação, foi decidido por manter valores menores para o tamanho do mini-lote uma vez que ambos, 1 e 18 possuem uma certa proximidade. Sendo o ponto de tamanho 1, apesar da piora na uniformidade da saída, o ponto com RMSE médio significativamente melhor que os demais, este foi escolhido como sendo o tamanho do mini-lote definitivo para os testes seguintes.

Embora a rede tenha apresentado melhoras no estágio atual, possuindo agora um RMSE médio de 0,0839 com desvio padrão de 0,0171, este valor ainda está insatisfatório e bastante instável, o que demonstra claras dificuldades no treinamento.

A instabilidade atual da rede, remete a um comportamento indesejado descrito na literatura como o problema dos mínimos locais onde a rede encontra uma solução estável, mas que não é a saída ótima. Este problema ocorre devido ao algoritmo *Backpropagation* tentar

minimizar o erro ajustando os pesos na direção do valor mínimo, utilizando o método do gradiente descendente (BRAGA et al., 2011). A Figura 4.7 ilustra um exemplo de uma função onde há mínimos locais.

Figura 4.7 – Região não convexa possuindo dois mínimos locais sendo um deles global.



Fonte: Carlon (2015)

Como ilustrado na Figura 4.7, o gradiente descendente leva a rede a caminhar na direção dos valores mínimos para o erro obtido. Porém, em superfícies complexas este comportamento pode fazer a rede encontrar o mínimo local e convergir erroneamente para este resultado, ignorando o mínimo global, que seria o valor desejado.

Braga et al. (2011) propõe como possíveis soluções, tanto para evitar os mínimos locais quanto para acelerar o treinamento, as seguintes técnicas:

- Adição de nós intermediários;
- Taxa de aprendizagem decrescente; e
- Adição do termo *momentum*;

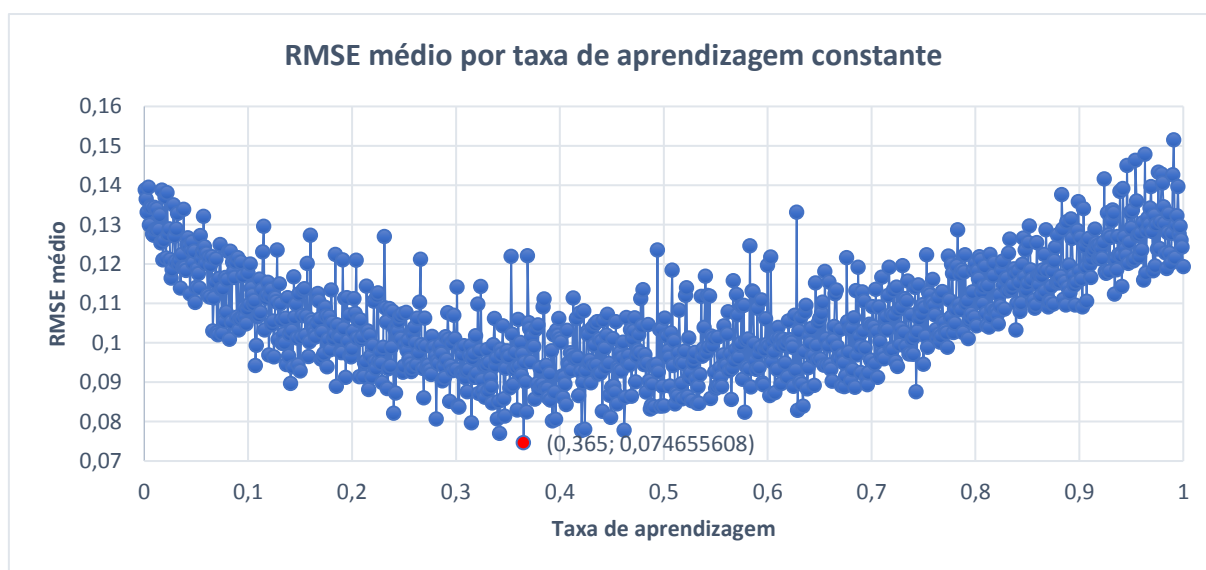
Como a adição de nós intermediários já havia sido realizada, e não houve melhoras significativas, a próxima técnica utilizada foi a alteração na taxa de aprendizado para que fosse possível a rede encontrar uma solução.

Com relação a taxa de aprendizagem duas abordagens foram testadas: observar o impacto de diferentes taxas de aprendizagem constantes e diferentes taxas de aprendizagem adaptativas ou decrescentes. Quanto a segunda abordagem, a classe `MLPRegressor` do *scikit-*

learn, implementa uma taxa de aprendizagem adaptativa, reduzindo-a cada vez que o modelo falhar em decrementar a taxa de erro por duas épocas consecutivas ou não melhorar o resultado da validação.

Na primeira abordagem, para taxa de treinamento constante, foram testados valores entre 0,001 e 1,000 e observado o RMSE médio. Todas as demais configurações foram mantidas.

Figura 4.8 – RMSE médio por taxa de aprendizagem constante.



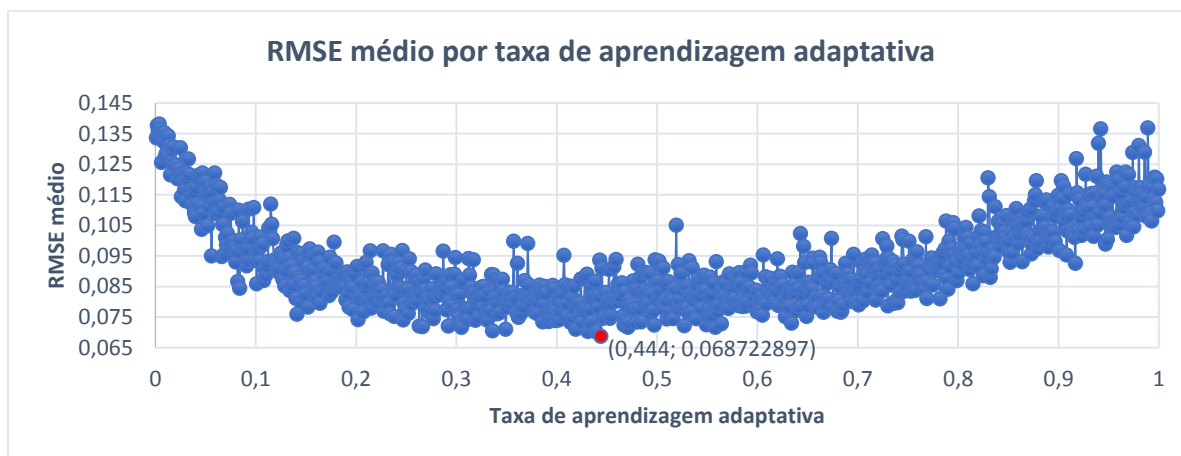
Fonte: Elaborada pelo autor

Observando a Figura 4.8, nota-se que valores muito pequenos ou muito grandes de taxa de aprendizagem resultam em piores resultados. A melhor saída (RMSE 0,07465) foi possível com taxa de aprendizagem igual a 0,365.

Na segunda abordagem, tornando a taxa de aprendizagem adaptativa foram testados valores no intervalo entre 0,001 e 1,000 e observado o valor médio do RMSE, conforme pode ser visto na Figura 4.9.

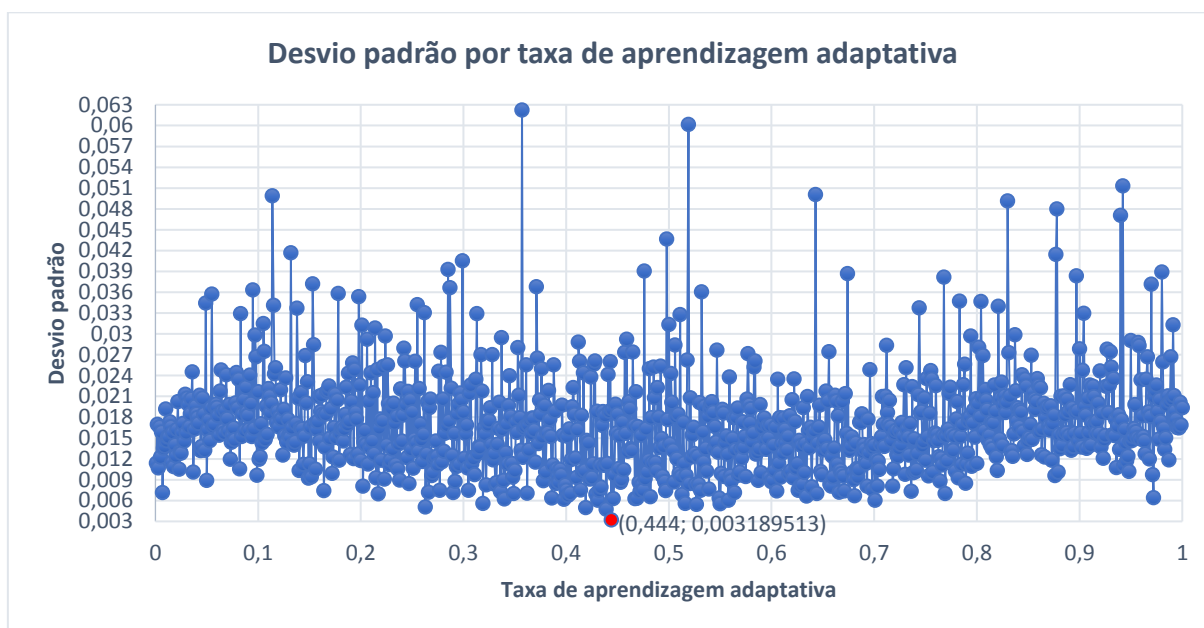
Com a taxa de aprendizagem adaptativa Figura 4.9, a resposta da rede, seguiu a mesma tendência que foi observado na taxa constante, mostrada na Figura 4.8, onde valores pequenos ou grandes demais resultavam em um RMSE médio maior. O melhor RMSE médio, valor de 0,0687 foi obtido para uma taxa de aprendizagem igual a 0,444. Como de uma maneira geral os resultados na taxa de aprendizagem adaptativa foram melhores, foi observado também a uniformidade dos resultados para cada valor e foi verificado o desvio padrão conforme mostra a Figura 4.10.

Figura 4.9 – RMSE Médio por taxa de aprendizagem adaptativa.



Fonte: Elaborada pelo autor

Figura 4.10 – Desvio padrão do RMSE por taxa de aprendizagem adaptativa.



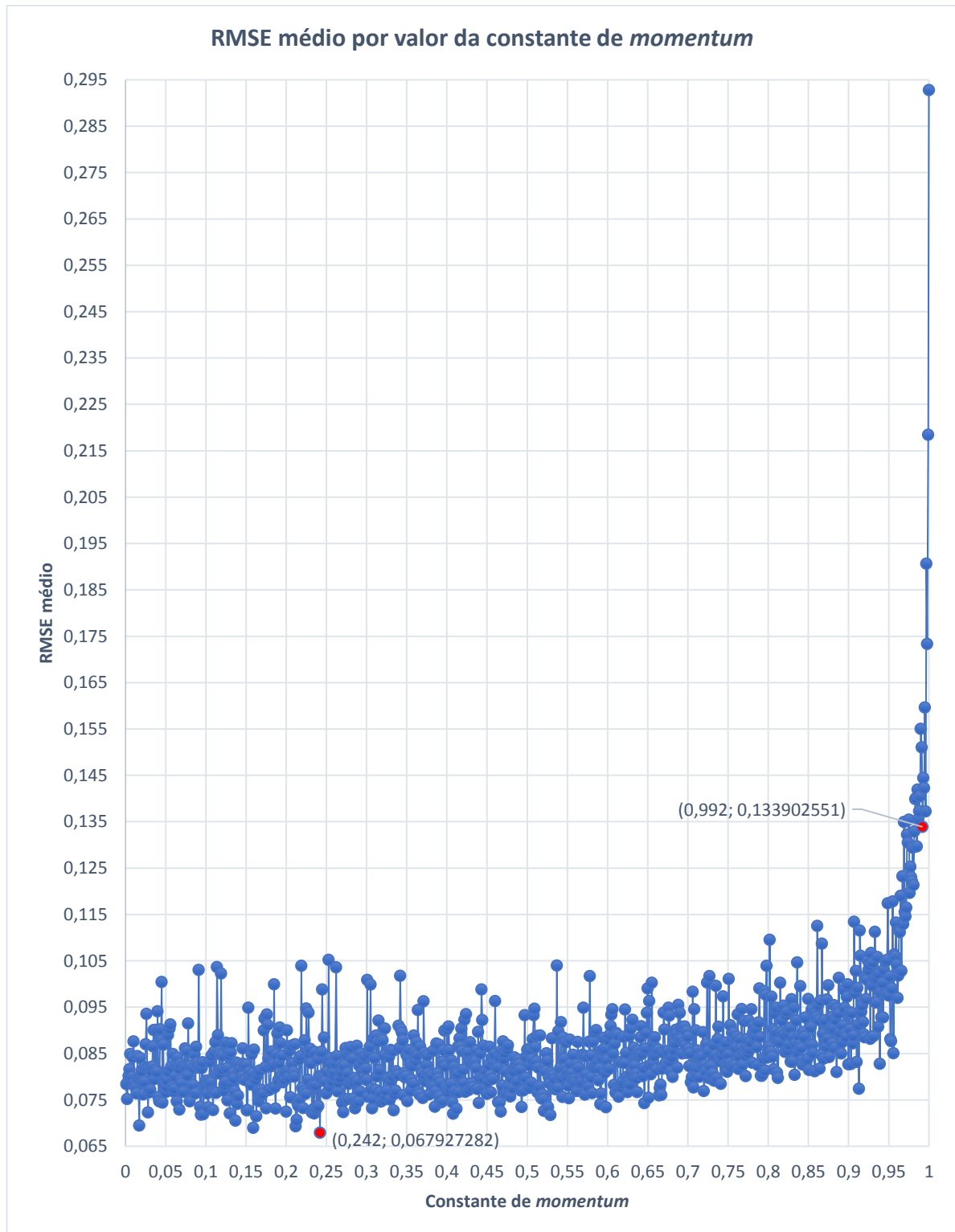
Fonte: Elaborada pelo autor

Destacado na Figura 4.10, o ponto com taxa de aprendizagem 0,444 apresentou o menor desvio padrão, com valor igual a 0,0031. Logo uma taxa de aprendizagem adaptativa igual 0,444 mostrou, na configuração atual, ser o melhor valor de taxa de aprendizagem para rede, por possuir tanto o menor RMSE médio quanto o menor valor de desvio padrão, gerando uma saída melhor e mais estável que nos testes anteriores.

A próxima etapa na otimização foi a utilização do termo *momentum*, que é uma constante que leva em consideração a alteração dos pesos anteriores no sentido atual do

movimento do gradiente, e é capaz de acelerar o processo de treinamento, além de contribuir para evitar mínimos locais. (BRAGA, et al. 2011).

Figura 4.11 – RMSE médio por valor da constante de *momentum*.

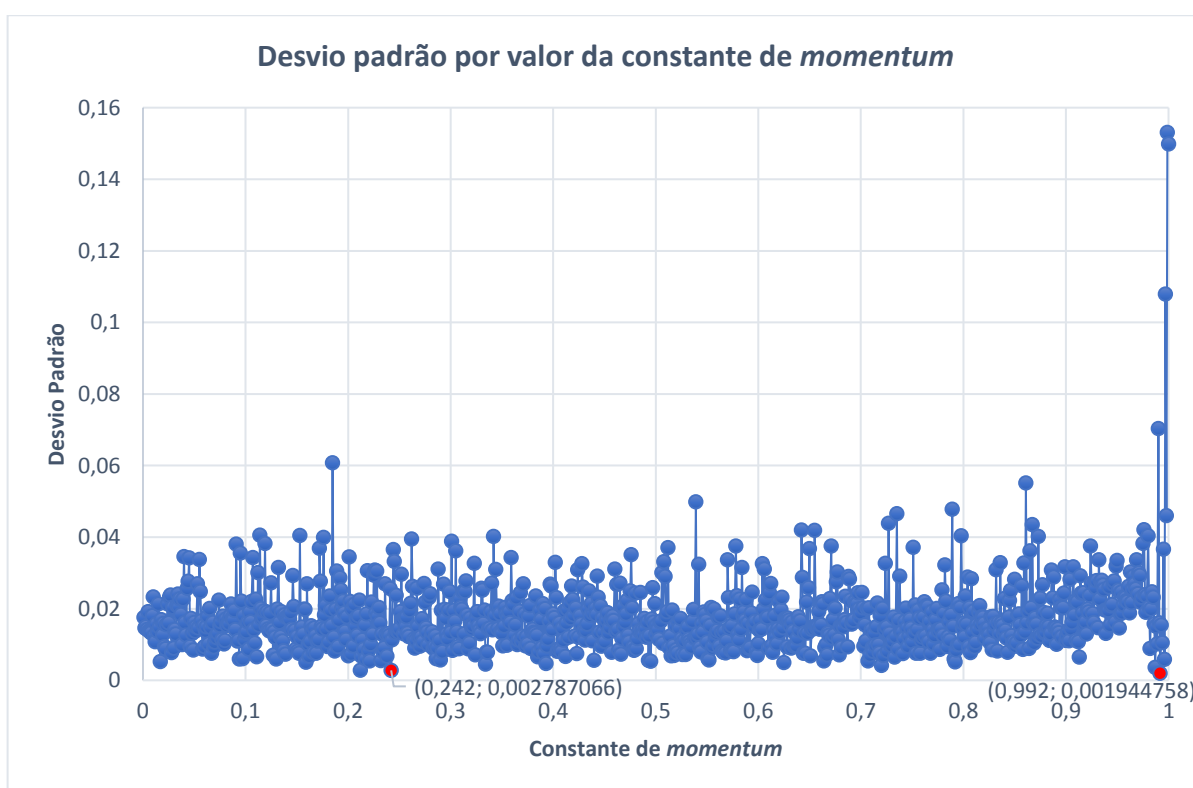


Fonte: Elaborada pelo autor

Mantendo-se as demais configurações da rede, com a nova taxa de aprendizagem adaptativa (0,444), foram testados valores para o termo *momentum* no intervalo de 0,001 e 0,999 conforme mostrado na Figura 4.11.

O menor valor para o RMSE médio foi igual 0,0679 e foi encontrado no ponto em que a constante de *momentum* era igual a 0,242. Adicionalmente também foi verificado a uniformidade das saídas, avaliando o desvio padrão, como pode ser visualizado na Figura 4.12, logo abaixo.

Figura 4.12 – Desvio padrão por valor da constante de *momentum*.



Fonte: Elaborada pelo autor

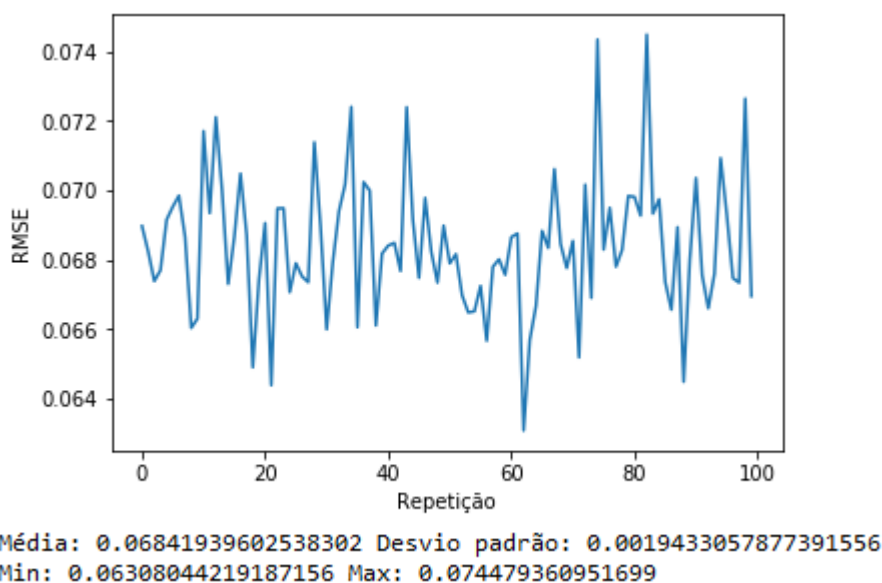
Na Figura 4.12, a saída mais uniforme foi encontrada para o valor de constante de *momentum* igual a 0,992 com desvio padrão igual a 0,0019. Entretanto, o RMSE médio para este mesmo valor, como destacado na Figura 4.11, é demasiadamente elevado por ser igual a 0,1339, e não poderia ser utilizado. Em contrapartida, o menor RMSE médio que pode ser encontrado na Figura 4.11 no ponto 0,242 possuía um desvio padrão igual 0,0027. Dessa forma foi mantido como 0,242 o valor para constante de *momentum*.

A rede no estado atual, após as aplicações das técnicas de otimização sugeridas por Braga et al. (2011), possui como saídas um RMSE médio igual a 0,0679 e desvio padrão igual

0,0027. Estes valores indicam melhores e mais estáveis previsões do que nos estados iniciais da rede.

Para certificar-se de que este resultado é um valor realmente estável, ou seja, que não está sofrendo influência do efeito aleatório inicial dos pesos, e nem recaindo no problema dos mínimos locais, a configuração atual foi submetida a um teste de 100 repetições e calculado o valor médio do RMSE e o desvio padrão.

Figura 4.13 – Teste de 100 repetições para configuração final da rede.



Fonte: Elaborada pelo autor

Analisando-se o teste representado na Figura 4.13, percebe-se que a oscilação na resposta ainda está presente, porém estas oscilações ficaram muito menores do que no estado inicial da rede, convergindo sempre em respostas muito próximas.

A média dos valores (0,0684) ficou bem próxima do que foi encontrado no teste anterior e o desvio padrão foi ainda menor 0,0019, com a diferença entre o valores máximos e mínimos do RMSE médio na ordem de 0,01.

A melhor configuração da rede encontrada foi a seguinte:

Para os dados: 10% para testes e 90% para treino, sendo 10% dos dados de treino para validação; e janela de tamanho 3.

Para a rede neural: 9 neurônios na camada de entrada; 10 neurônios na única camada oculta; uma função logística para ativação das camadas intermediárias; uma função identidade para ativação da camada de saída; um número máximo de iterações igual a 1000 e um gradiente descendente estocástico como algoritmo de aprendizado com tamanho de mini-lote

igual a 1, com taxa de aprendizagem adaptativa e igual a 0,444 e com o termo momentum igual a 0,242.

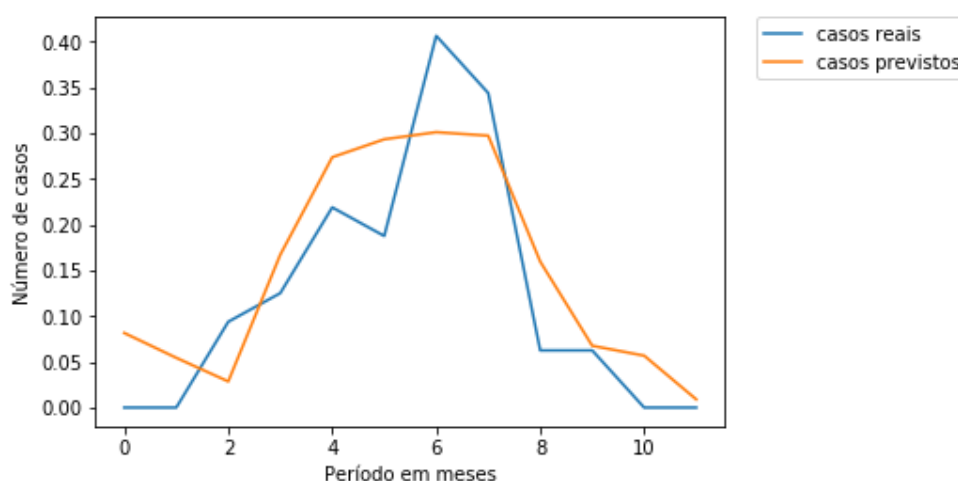
4.4 Testes Finais da Rede Neural Artificial

Após realizada a etapa de otimização das saídas da RNA, o modelo foi testado com um conjunto de 12 meses normalizados, correspondente aos 10% reservados para o teste, que não haviam sido utilizados durante a fase de treino e validação.

Figura 4.14 – Resultado da previsão para o ano de 2017 (valores normalizados).

Precisão FINAL:

R2: 0.74 MSE: 4.65e-03 RMSE: 6.82e-02



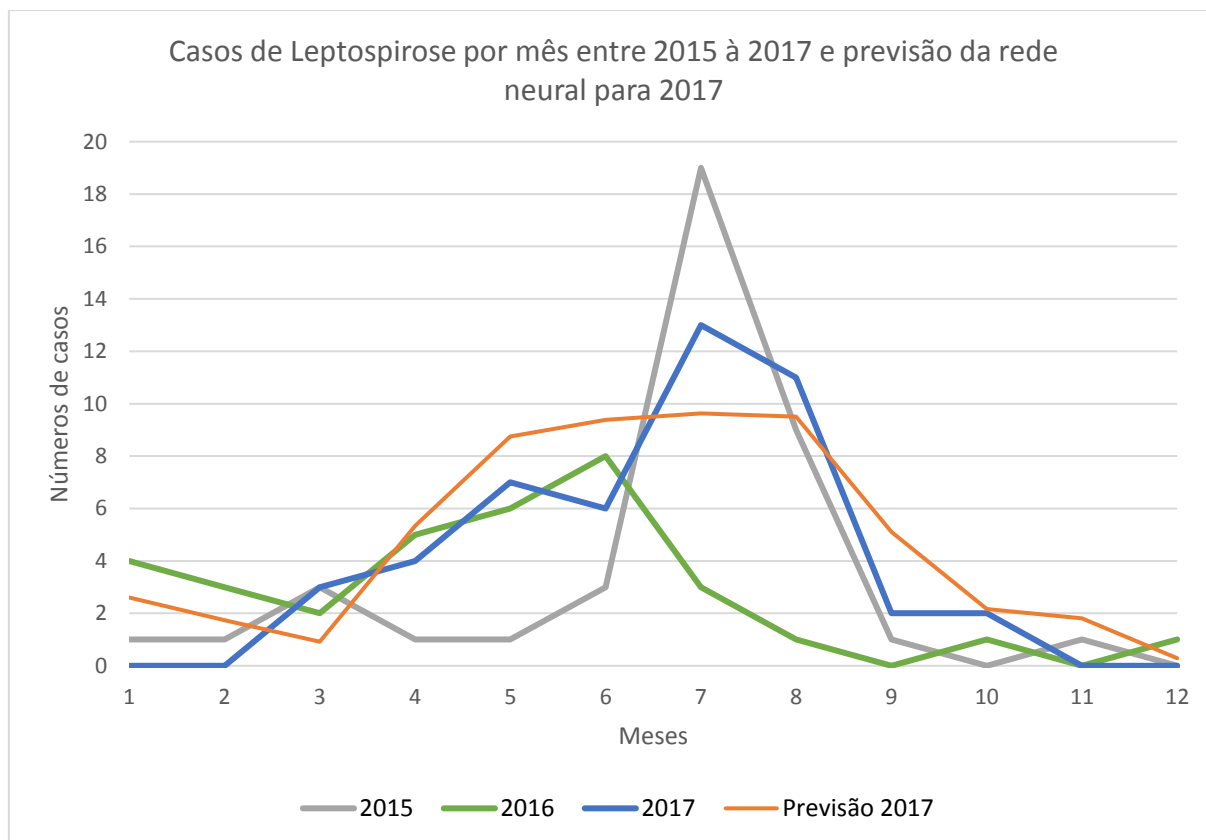
Fonte: Elaborada pelo autor

Analizando a Figura 4.14, percebe-se que a rede foi capaz de prever com certa aproximação quando o número de casos da doença tende a subir e diminuir. Entretanto não foi possível prever com precisão os valores iniciais e dos picos, comportamento possivelmente causado por fatores externos, como, por exemplo, subnotificações dos casos da doença no SINAN, uma vez que os dados de 2017 ainda estão sujeitos a revisão, e como verificado pelo autor, após a revisão, o número de casos de Leptospirose, em anos anteriores, sofreu alterações.

É conveniente ressaltar, que as possíveis revisões do SINAN sobre os dados de 2017 de Leptospirose podem comprometer o teste final da RNA. Por esta razão, adicionalmente foi avaliado a tendência de surgimento de surtos da doença, comparando os resultados da

previsão do ano de 2017, com os casos reais para o mesmo ano, juntamente com dois anos anteriores. O resultado é mostrado na Figura 4.15 com saídas convertidas para valores reais.

Figura 4.15 – Comparativo entre os casos da doença, entre 2015 e 2017 com os casos previstos para 2017.



Fonte: Elaborada pelo autor

A Figura 4.15 permite observar que os valores iniciais nos últimos anos anteriores, não foram zero, como ocorreu em 2017, e os casos da doença tendem a crescer entre abril e agosto, sendo o período de junho a agosto o período de maior incidência, tendência também acompanhada pela rede, bem como a redução dos casos nos meses seguintes, fator que contribui para viabilidade de utilização da previsão da RNA, uma vez que a rede foi capaz de prever em que período ocorreriam os surtos da doença.

5 Conclusão

Esta seção descreve as conclusões alcançadas pelo autor após a realização deste trabalho. Na Seção 5.1 são apresentadas as considerações finais e na Seção 5.2 são propostos temas para futuros trabalhos relacionados a este.

5.1 Considerações Finais

Este trabalho, abordou a viabilidade do uso de RNAs como uma ferramenta adicional na previsão de surtos de Leptospirose na cidade de Recife-PE, utilizando o modelo supervisionado *Multilayer Perceptron*.

A partir dos números de casos de infecção da doença, obtidos no SINAN, e de dados sobre a precipitação pluviométrica e umidade do ar, extraídos do INMET por serem variáveis meteorológicas correlacionadas com os casos de Leptospirose, foi montada uma base de dados que serviu de alimentação para o modelo de RNA.

Durante o treinamento foram observadas algumas dificuldades na convergência da rede neural, e, por isso, diversos parâmetros foram testados e ajustados com base na análise dos resultados e recomendações da literatura, de forma a otimizar a saída da rede para realizar uma previsão satisfatória de surtos da doença.

Outra dificuldade encontrada foi à disponibilização dos dados do SINAN, podendo comprometer o resultado da previsão final da RNA. Muitas vezes esses dados chegam a levar um ou mais anos para estarem acessíveis e, após a divulgação, ainda ficam sujeitos a revisões futuras que podem alterar seus valores.

Após a análise dos testes finais, realizando previsões para um conjunto de dados não apresentado ao modelo durante a fase de treino, foi possível concluir que, apesar das dificuldades, a RNA do tipo *Multilayer Perceptron*, da maneira como foi implementada foi capaz de acompanhar a tendência de surtos de Leptospirose na cidade de Recife-PE.

Entretanto, por se tratar de uma previsão, não é possível garantir que o número de casos da doença previsto seja igual ao real, uma vez que estes valores podem sofrer alterações por fatores externos.

Dessa forma, as redes neurais artificiais mostraram-se uma aliada da gestão pública, provendo informações de quando ocorrerão surtos de Leptospirose para que ações de prevenção, conscientização, diagnóstico e tratamento possam ser preparadas antecipadamente, evitando a sobrecarga dos serviços de saúde e consequentemente melhorando a qualidade de vida das pessoas ao minimizar o impacto social e econômico causado pelos surtos.

5.2 Proposta de Trabalhos Futuros

Como proposta para futuros trabalhos, sugere-se que sejam realizados aperfeiçoamentos tanto na preparação dos dados quanto no treinamento da RNA, de modo a garantir uma maior qualidade e agilidade na previsão realizada.

Devido aos atrasos na disponibilização dos dados do SINAN, como já mencionado anteriormente, é aconselhado que a coleta dos registros de infecção por Leptospirose possa ser realizada prontamente na cidade de estudo (na Secretaria Municipal de Saúde, por exemplo) de preferência em intervalos mensais, que foi o intervalo de tempo utilizado pela RNA da maneira como foi implementada.

Propõe-se ainda, algumas modificações na preparação dos dados de alimentação, e testes mais amplos na RNA, em busca de uma configuração que melhor se ajuste às entradas apresentadas, produzindo previsões com qualidade superior. Para isto recomenda-se algumas alterações:

- Explorar diferentes tamanhos de janelas para montar os pares de entrada e saída;
- Testar outros intervalos de tempo ao montar os pares de entrada e saída: semanais, bimestrais, trimestrais, etc.
- Avaliar o impacto de diferentes proporções nos conjuntos de dados de treino e validação; e
- Realizar testes mais exaustivos em busca de melhores combinações de configurações da RNA.

Adicionalmente é sugerido medir o custo computacional de cada teste realizado a fim de encontrar a melhores configurações com menor custo, uma vez que a alteração de certos parâmetros na rede fará com que o modelo leve consideravelmente mais tempo para

convergir, o que pode vir a ser desnecessário caso uma configuração diferente com menor custo computacional possa alcançar resultados próximos.

Por fim, embora, como descrito na literatura, as variáveis meteorológicas possuam maiores graus de correlação com os casos da doença, as características únicas de cada local não podem ser desprezadas, devendo ser observadas outras variáveis que possam contribuir para a qualidade da previsão da rede.

REFERÊNCIAS

- ANDRADE, A.; PINTO, S. C.; OLIVEIRA, R. S. (Orgs). **Animais de Laboratório: criação e experimentação** [online]. Rio de Janeiro: Editora FIOCRUZ, 2002. 388 p. Disponível em <<http://books.scielo.org/id/sfwjtj/pdf/andrade-9788575413869-26.pdf>>. Acesso em: 12 jul. 2018.
- BISGIN, H.; USAME KILINC, O.; UGUR, A.; XU, X.; TUZCU, V. **Diagnosis of long QT syndrome via support vector machines classification** - Scientific Figure on ResearchGate. Disponível em: <https://www.researchgate.net/k-fold-cross-validation-scheme-example_fig2_228403467> Acesso em: 01 ago. 2018
- BONITA, R.; BEAGLEHOLE, R.; KJELLSTRÖM, T. Epidemiologia básica. 2. ed. São Paulo, Santos. 2010. p1.
- BOX, G. E. P.; JENKINS, G. M. **Time series analysis: forecasting and control., revised edition**. San Francisco: Holden-day. 1976.
- BRAGA, A. P.; LUDERMIR, T. B.; CARVALHO, A. **Redes Neurais Artificiais: Teoria e Aplicações**. LTC Editora, 2011.
- BRASIL. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica. **Sistema de Informação de Agravos de Notificação–Sinan: normas e rotinas / Ministério da Saúde**, Secretaria de Vigilância em Saúde, Departamento de Vigilância Epidemiológica. – Brasília: Editora do Ministério da Saúde, 2006.
- _____. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância Epidemiológica. **Guia de vigilância epidemiológica**. 7a ed. Brasília, DF; 2009; 816 p.
- _____. Ministério da Saúde. Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis. **Leptospirose: diagnóstico e manejo clínico/Ministério da Saúde**, Secretaria de Vigilância em Saúde. Departamento de Vigilância das Doenças Transmissíveis. – Brasília: Ministério da Saúde, 2014.
- _____. Ministério da Saúde. **Portaria nº 204 de 17 de fevereiro de 2016**. 2016. Disponível em <http://bvsms.saude.gov.br/bvs/saudelegis/gm/2016/prt0204_17_02_2016.html>. Acesso em: 10 ago. 2018
- _____. MINISTÉRIO DA SAÚDE. Secretaria de Vigilância em Saúde, **Sistema de Informação de Agravos de Notificação (SINAN)**. Disponível em <<http://portalsinan.saude.gov.br/dados-epidemiologicos-sinan>>. Acesso em: 16 jul. 2018
- BROWNLEE, J. Machine Learning Mastery. **Time Series Forecasting Performance Measures With Python**, 2017. Disponível em: <<https://machinelearningmastery.com/time-series-forecasting-performance-measures-with-python/>>. Acesso em: 20 fev. 2018

CARLON, A. **Desenvolvimento de um novo operador para algoritmos metaheurísticos baseado na maximização da rigidez aplicado à otimização de estruturas treliçadas** - Scientific Figure on ResearchGate. Disponível em: <https://www.researchgate.net/Figura-3-Regiao-nao-convexa-possuindo-dois-minimos-locais-sendo-um-deles-global_fig2_324976745>. Acesso em: 01 jan. 2018

DOS SANTOS, C. S. S. **Previsão da Incidência de Dengue no Estado do Rio de Janeiro através de Redes Neurais Artificiais**. 2013. Tese de Doutorado. PUC/Rio, 2013.

EHLERS, R. S. **Análise de Séries Temporais**. Departamento de Estatística, UFPR. 2005. Disponível em: <<http://www.each.usp.br/rvicente/AnaliseDeSeriesTemporais.pdf>>. Acesso em: 11 ago. 2018.

FERREIRA, E. V. **Gradiente descendente (batch, stochastic e boosting)**. Laboratório de Estatística e Geoinformação – LEG, UFPR, 2015. Disponível em: <<http://cursos.leg.ufpr.br/ML4all/gradiente/2015/04/18/aula24>>. Acesso em: 12 ago. 2017

FIGUEIREDO FILHO, D. B.; SILVA JÚNIOR, J. A. **Desvendando os Mistérios do Coeficiente de Correlação de Pearson (r)**. 2009. Disponível em: <<https://periodicos.ufpe.br/revistas/politica hoje/article/viewFile/3852/3156>>. Acesso em: 03 set. 2018.

GOUVEIA, R. Toda Matéria. **Desvio Padrão**, 2018. Disponível em: <<https://www.todamateria.com.br/desvio-padrao/>>. Acesso em 16 de jul. 2018

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2001.

IBGE. Instituto Brasileiro de Geografia e Estatística. Comitê de Estatísticas Sociais Disponível. Base de Dados. Metadados. Ministério da Saúde. **Sistema de Informações de Agravos de Notificação – SINAN**. Disponível em <<http://ces.ibge.gov.br/base-de-dados/metadados/ministerio-da-saude/sistema-de-informacoes-de-agravos-de-notificacao-sinan>>. Acesso em: 30 ago. 2018

LARSON, R; FARBER, B. **Estatística aplicada**. 4. ed. São Paulo: Pearson Prentice Hall, 2010.

LEVETT, P.N. **Leptospirosis**. Clinincal Microbiology vol14. v14 n.2. 2001; p. 296–326. Disponível em: <http://leptospirosis.org.nz/Portals/0/Images/Leptospirosis/Levett_2001_review.pdf> Acesso em : 30 ago. 2018.

MASI, E. **Análise de intervenção em séries temporais de dengue e leptospirose da cidade de São Paulo: influência de fatores políticos, administrativos, técnicos e ambientais**. 2014. Tese de Doutorado. Universidade de São Paulo 2014.

MITTELMANN, M.; SOARES, D. G. **Previsão de Casos de Dengue em Itajaí-SC por meio de Redes Neurais Artificiais Multicamadas e Recorrentes**. Anais do Computer on the Beach, 2017. p. 130-139.

MORETTIN, P. A.; TOLOI, C. M. C. **Previsão de séries temporais**. 2. ed., São Paulo: Editora Edgard Blücher. 2006.

_____; _____. **Modelos para previsão de séries temporais**. Instituto de matemática pura e aplicada, 1981.

MUKAKA M. **A guide to appropriate use of Correlation coefficient in medical research**. Malawi Medical Journal : The Journal of Medical Association of Malawi. v.24, n. 3, 2012. p. 69-71. Disponível em: <<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/#R4>>. Acesso em: 18 mai. 2018

PEDREGOSA, F. et al. **Scikit-learn: Machine learning in Python**. Journal of machine learning research, v. 12, n. Oct, p. 2825-2830, 2011.

PELISSARI, D. M.; MAIA-ELKHOURY A. N. S; SIMÕES, M. L. N.; NUNES, A. M. L. **Revisão sistemática dos fatores associados à leptospirose no Brasil, 2000-2009**. Epidemiol Serv Saúde [periódicos na Internet]. v20, n. 42011; p. 565-74 Disponível em: <<http://scielo.iec.pa.gov.br/pdf/ess/v20n4/v20n4a16.pdf>> Acesso em 18 jul. 2018

RUMELHART, D. E., HINTON, G. E. e WILLIAMS, R. J. **Learning Representations by Backpropagation Errors**. In: Nature, v. 323, 1986. p. 533-536.

SCHERER, M. Time Health. **Meet the Bots That Knew Ebola Was Coming**, 2014. Disponível em: <<http://time.com/3086550/ebola-outbreak-africa-world-health-organization/>>. Acesso em: 20 jul. 2018

SCIKIT-LEARN: **Machine Learning in Python**, PEDREGOSA *et al.*, JMLR 12, 2011. p. 2825-2830.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais para engenharia e ciências aplicadas**. São Paulo: Artliber, 2010. p. 33-111.

VASCONCELOS C.H., FONSECA FR, LISE M.L.Z., ARSKY M.L.N.S. **Fatores ambientais e socioeconômicos relacionados à distribuição de casos de leptospirose no Estado de Pernambuco, Brasil, 2001-2009**. Cad Saúde Colet. 2012. Disponível em: <http://www.iesc.ufrj.br/cadernos/images/csc/2012_1/artigos/CSC_v20n1_49-56.pdf>. Acesso em: 08 jul. 2018

WORLD HEALTH ORGANIZATION. **Human Leptospirosis: Guidance for Diagnosis, Surveillance and Control**. Geneva, Switzerland: WHO, 2003. Disponível em <http://www.who.int/csr/don/en/WHO_CDS_CSR_EPH_2002.23.pdf>. Acesso em: 10 ago. 2018