UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Imputação de Dados Faltantes via Algoritmo EM e Rede Neural MLP com o Método de Estimativa de Máxima Verossimilhança para Aumentar a Acurácia das Estimativas

Elisalvo Alves Ribeiro

SÃO CRISTÓVÃO/SE
2015

UNIVERSIDADE FEDERAL DE SERGIPE
PRÓ-REITORIA DE PÓS-GRADUAÇÃO E PESQUISA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Elisalvo Alves Ribeiro

Imputação de Dados Faltantes via Algoritmo EM e Rede Neural MLP com o Método de Estimativa de Máxima Verossimilhança para Aumentar a Acurácia das Estimativas

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação (PROCC) da Universidade Federal do Sergipe (UFS) como parte de requisito para obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Computação Inteligente.

Orientador: Prof. Dr. Carlos Alberto Estombelo Montesco

SÃO CRISTÓVÃO/SE
2015

COMISSÃO JULGADORA – DISSERTAÇÃO DE MESTRADO

Candidato: Elisalvo Alves Ribeiro

Data da Defesa: 14 de agosto de 2015

Título da Dissertação: "Imputação de Dados Faltantes via Algoritmo EM e Rede Neural MLP com o método de Estimativa de Máxima Verossimilhança para aumentar a acurácia das estimativas"

Prof. Dr. Carlos Alberto Estombelo Montesco (Presidente)

_____

Prof. Dr. Paulo Salgado Gomes de Mattos Neto (Membro Externo)

_____

Prof. Dr. Leonardo Nogueira Matos (Membro Interno)

_____

RESUMO

Base de dados com valores faltantes é uma ocorrência frequentemente encontrada no mundo real, sendo as causas deste problema são originadas por motivos diversos (falha no equipamento que transmite e armazena os dados, falha do manipulador, falha de quem fornece a informação, etc.). Tal situação pode tornar os dados inconsistentes e inaptos de serem analisados, conduzindo às conclusões muito enviesadas. Esta dissertação tem como objetivo explorar o emprego de Redes Neurais Artificiais *Multilayer Perceptron* (RNA MLP), com novas funções de ativação, considerando duas abordagens (imputação única e imputação múltipla). Primeiramente, é proposto o uso do Método de Estimativa de Máxima Verossimilhança (EMV) na função de ativação de cada neurônio da rede, em contrapartida à abordagem utilizada atualmente, que é sem o uso de tal método, ou quando o utiliza é apenas na função de custo (na saída da rede). Em seguida, são analisados os resultados destas abordagens em comparação com o algoritmo *Expectation Maximization* (EM) que é o estado da arte para tratar dados faltantes. Os resultados obtidos indicam que ao utilizar a Rede Neural Artificial MLP com o Método de Estimativa de Máxima Verossimilhança, tanto em todos os neurônios como apenas na função de saída, conduzem a uma imputação com menor erro. Os resultados experimentais foram avaliados via algumas métricas, sendo as principais o MAE (*Mean Absolute Error*) e RMSE (*Root Mean Square Error*), as quais apresentaram melhores resultados na maioria dos experimentos quando se utiliza a RNA MLP abordada neste trabalho para fazer imputação única e múltipla.

Palavra-Chave: Redes Neurais Artificiais MLP, Método de Estimativa de Máxima Verossimilhança, Algoritmo EM, imputação de dados, dados faltantes, novas funções de ativação.

ABSTRACT

Database with missing values it is an occurrence often found in the real world, beiging of this problem caused by several reasons (equipment failure that transmits and stores the data, handler failure, failure who provides information, etc.). This may make the data inconsistent and unable to be analyzed, leading to very skewed conclusions. This dissertation aims to explore the use of Multilayer Perceptron Artificial Neural Network (ANN MLP), with new activation functions, considering two approaches (single imputation and multiple imputation). First, we propose the use of Maximum Likelihood Estimation Method (MLE) in each network neuron activation function, against the approach currently used, which is without the use of such a method or when is used only in the cost function (network output). It is then analyzed the results of these approaches compared with the Expectation Maximization algorithm (EM) is that the state of the art to treat missing data. The results indicate that when using the Artificial Neural Network MLP with Maximum Likelihood Estimation Method, both in all neurons and only in the output function, lead the an imputation with lower error. These experimental results, evaluated by metrics such as MAE (Mean Absolute Error) and RMSE (Root Mean Square Error), showed that the better results in most experiments occured when using the MLP RNA addressed in this dissertation to single imputation and multiple.

Keywords: Artificial Neural Networks MLP, Maximum Likelihood Estimation Method, EM Algorithm, data imputation, missing data, new function activation.

Agradecimentos

Ao Prof. Dr. Carlos Alberto Estombelo Montesco, pelos conselhos, ensinamentos, orientação e acima de tudo paciência.

Aos Dr. Leonardo Nogueira Matos, Dr. Jugurta Montalvão e Drª. Adicinéia, pelo grande ensinamento em suas aulas.

Aos membros da banca, pelas contribuições e comentários junto ao texto da dissertação.

À Universidade Federal de Sergipe e ao Departamento de Ciência da Computação, pelo fornecimento de instalações e condições de trabalho apropriadas.

À CAPES, pelo apoio financeiro ao meu projeto.

Aos meus colegas do Mestrado, pelas críticas, sugestões e companheirismo.

À minha mãe, Lúcia, pelo amor incondicional.

Ao meu pai, Walter, pelo exemplo de honestidade e determinação.

Aos meus irmãos, Alisson, Humberto e Cristiane, meus amigos em todos os momentos, por todos os aprendizados que tivemos e temos juntos.

Aos meus avós, pelo carinho, amor e conselho.

Aos meus sobrinhos, Filipe e Clara, pelas brincadeiras e alegrias que me proporcionaram.

À minha esposa Maise, por todo amor, paciência e companheirismo.

À minha sogra Maria José e ao meu sogro Hermilton, pelo carinho.

Ao meu tio Aloisio, pela motivação e apoio.

Aos meus amigos, Fillipe, Silvaneide, Idilvan, Cleberson, Hallan, Lucas, Nayara, José Rodrigo, Esdras, Armoni, Lúcio pela grande amizade, e todos os outros que de alguma forma contribuíram para este desafio.

Agradeço, enfim, a Deus, pela saúde, harmonia e paz em minha vida.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE SIGLAS

EM - Expectation Maximization
TRI - Teoria de Resposta ao Item
KDD - Knowledge Discovery from Databases
MCAR - Missing Completely At Random
MAR - Missing At Random
MNAR - Missing Not At Random
RNA-MLP - Redes Neurais Artificiais Multilayer Perceptron
EMV - Estimativa de Máxima Verossimilhança
MQO - Mínimos Quadrados Ordinários
MI - Multiple Imputation
df - degrees of freedom
ML - Maximum Likelihood
SIG - Sigmoide
AO - Aranda Ordaz
TH - Tangente Hiperbólica
CLL - Complemento Log-Log
LL - Log-Log
SIGEMV - Sigmoide com Estimativa de Máxima Verossimilhança
AOEMV – Aranda-Ordaz com Estimativa de Máxima Verossimilhança
THEMV - Tangente Hiperbólica com Estimativa de Máxima Verossimilhança
CLLEMV - Complemento Log-Log com Estimativa de Máxima Verossimilhança
LLEMV - Log-Log com Estimativa de Máxima Verossimilhança
MAE - Mean Absolute Error
RMSE - Root Mean Square Error
1º Qu - primeiro quartil
3º Qu - terceiro quartil

CAPÍTULO 1

INTRODUÇÃO

Profissionais e pesquisadores das mais diversas áreas do conhecimento reconhecem que vivemos na "era do *big data*", há até os que falam em *zettabytes* (KURASOVA *et al*., 2014; MINGKUI *et al*., 2014; WANG *et al*., 2014). Esse fenômeno está ocorrendo em virtude do avanço computacional, que facilitou a aquisição e armazenamento de dados das mais variadas fontes, gerando assim novos desafios, que é adquirir e armazenar dados com qualidade e livres de ruídos.

Esta abundância de dados tem conduzido muitas empresas a repensarem seus negócios, pois as organizações que conseguem coletar e analisar seus dados de forma consistente conseguem ter vantagens competitivas no mercado. Prass (2004) afirma que atualmente a informação conseguiu adquirir um valor que há poucos anos atrás era inimaginável. É tanto que, em muitos casos, o maior bem que a organização tem é aquilo que ela sabe sobre seus clientes.

Porém, apesar deste grande volume de dados que está presente em nosso dia a dia, há um problema que é corriqueiro nestes *dataset*, que é a presença de dados faltantes (*missing data*) ocorrendo parcialmente em alguma ou em todas as variáveis de interesse.

De acordo com Liublinska (2013) quando Ronald A. Fisher e Jerzy Neyman iniciaram suas pesquisas seminais, que vieram a tornar-se a base fundamental da Estatística Moderna no início do século 20, o problema de dados faltantes emergiu naturalmente, pois vários trabalhos de campo que foram aplicados por pesquisadores de diversas áreas, se depararam com tal problema. Desde então, diversos estudos foram feitos, principalmente a partir dos trabalhos de Hartley (1958), que propôs simplificar e unificar cálculos de estimadores de máxima verossimilhança para dados incompletos, a partir de amostras completas, Rubin (1976) que apresentou um novo viés para análise de dados faltantes e por fim Dempster, Laird & Rubin (1977) apresentaram formalmente o algoritmo EM (*Expectation Maximization*) o qual facilitou o cálculo iterativo da estimativa de máxima verossimilhança quando as observações podem ser vistas como dados incompletos.

O problema de dados faltantes pode ser ocasionado por diversos fatores, que podem ser desde defeitos em equipamentos às falhas humanas na manipulação dos equipamentos de coleta dos dados, sendo na maioria das vezes inviável, criar mecanismos que evitem tal

problema. Como exemplo, de acordo com Lakshminarayan *et al*. (1996) e Marlin (2008) podem-se listar:

a) Um sensor em uma rede de sensores remotos pode ser danificado e deixar de transmitir dados.

b) Os participantes de um estudo clínico podem sair durante o curso do estudo o que conduz à falta de observações em momentos posteriores.

c) Em ma pesquisa amostral o entrevistado poderá não responder uma determinada pergunta.

Segundo a literatura acadêmica, um sério problema na mineração de bases de dados industriais, é que estas frequentemente contêm dados incompletos, ou erroneamente registrados (BARNARD; MENG, 1999), e que apesar de existir diversas maneiras de lidar com *dataset* em tal situação, a literatura não determina qual o melhor método para todos os tipos de dados (HRUSCHKA JR; EBECKEN, 2002).

Na área de aprendizagem de máquina e análise de dados estatísticos, a fase de aprendizagem, inferência e previsão na presença de dados faltantes é um problema muito comum (MARLIN, 2008), assim como na engenharia de software, que também é comum encontrar nas bases de dados, que são utilizadas para a construção de modelos de previsão de esforço de software, dados omissos (MYRTVEIT; STENSRUD; OLSSON, 2001).

Pereira (2014) cita que, ao analisar bases públicas com indicadores educacionais, através da Teoria de Resposta ao Item (TRI), o mesmo se deparou com este problema, o que dificulta ou às vezes impossibilita a utilização desta técnica (TRI), pois ignorar tais dados faltantes pode criar problemas na estimação de parâmetros. Sendo assim, de acordo com Alisson (2001) mais cedo ou mais tarde, quem faz a análise estatística tem problemas com dados faltantes.

Além disso, os dados que são armazenados e analisados em tempo real, como séries de qualidade do ar, de previsão de demanda de energia, previsão de vazão de água, séries financeiras, entre outras, estão constantemente comprometidas com dados faltantes, tornando-se assim necessário criar mecanismos que possam tornar válidas as análises sob estas premissas, já que é impossível analisá-las diante de dados faltantes (SANTANA; FILIZOLA-JUNIOR; FREITAS, 2010), uma vez que as séries tem que estar ordenadas cronologicamente, para que estejam aptas a serem analisadas (LOPES, 2007).

A qualidade dos dados é tão importante que Silva (2001) enfatizou que no setor elétrico brasileiro, a aquisição e o armazenamento de dados oriundos de fontes confiáveis tornaram-se incontestavelmente parte integrante do patrimônio destas empresas.

Nos últimos anos esta questão tem sido cada vez mais estudada e metodologias foram desenvolvidas para tentar solucioná-las (ASSUNÇÃO, 2012; ENDERS, 2010), destacando-se a utilização de métodos de mineração de dados para imputar dados faltantes, já que estes procedimentos podem ser mais robustos diante de valores extremos e aparentam ser mais fáceis de automatizar (CASTILLO, 2014). Tais técnicas buscam meios de imputar dados, onde há dados faltantes, pois conforme cita Batista (2003) a imputação consiste de um procedimento, no qual trocam-se os valores desconhecidos de um determinado *dataset* por valores admissíveis que, de acordo De Waal *et al.* (2011), é muitas vezes aplicada para simplificar o processo de estimativa.

Veroneze (2011) também afirma que na fase de KDD (*Knowledge Discovery from Databases*), os algoritmos de mineração de dados poderão não ser capazes de fazer inferências da amostra se esta contiver valores faltantes, pois conforme citado em Hruschka Jr *et al*. (2007), estes algoritmos geralmente não são capazes de lidar com dados omissos de uma forma automática, ou seja, sem preparação deles (pré-processamento). Além disso, Honghai *et al*. (2005) alega que na fase de KDD, grande parte do tempo é gasto analisando e preenchendo dados omissos, e que entre 80% a 90% de um projeto de análise de dados é gasto na busca de tornar a base dados confiável o suficiente, que seja capaz de gerar resultados plausíveis, já que problemas de qualidade dos dados podem ser muito caro e causar desperdícios de bilhões de dólares em equipamentos, recursos mal alocado, devido a previsões falhas, e assim por diante . Assim, torna-se imprescindível que haja um pré-processamento nos dados dentro da fase de KDD. Nos problemas da vida real, observa-se que dados completos é uma exceção e não a regra (RAMACHANDRAN; TSOKOS, 2009).

Uma justificativa para a utilização de técnicas estatísticas de imputação de dados é que quando há perda de dados, consequentemente também há uma perda do poder estatístico, uma vez que se diminui o tamanho da amostra em análise (NUNES *et al*., 2009), situação esta que segundo De Waal *et al*. (2011) pode ocasionar um significativo aumento do erro padrão das estimativas dos parâmetros e consequentemente resultar em estimativas enviesadas, além de que, conforme enfatiza Nelwamondo *et al.* (2007) afetará a qualidade das decisões tomadas com base nesses dados. Seguindo esta linha de

pensamento, os métodos de imputação de dados faltantes consistem em preencher tais valores e em seguida analisar o *dataset* resultante, tido como dados completos, usando as técnicas estatísticas matematicamente já bem estabelecidas. Entretanto deve-se ter a consciência de que, com ou sem dados faltantes, o propósito de um procedimento estatístico deve ser o de tornar válida e eficiente as inferências sobre a população de interesse, não para estimar, prever ou recuperar observações faltantes e nem a de obter os mesmos resultados que teria tido com dados completos (SCHAFER; GRAHAM, 2002).

Alguns procedimentos de imputação são simples e implementados na maioria dos aplicativos estatísticos (PINTO, 2013), sendo que, de acordo com Assunção (2012), os métodos mais comuns são os que englobam a remoção ou a troca deste tipo de dado por alguma medida resumo (média ou mediana). Outra conduta que é comum, e que segue na mesma linha de pensamento é a não inclusão, no modelo, das variáveis que possuem dados faltantes.

O conceito de dados faltantes é bastante amplo. Ele inclui, por exemplo, falta de dados em um *layout* desequilibrado, mas se estende às observações de distribuições truncadas, dados censurados, e as variáveis latentes (SORENSE; GIANOLA, 2002).

Neste trabalho, segue-se a abordagem adotada por Liublinska (2013), que é a padrão utilizada na literatura da área, ou seja, o valor é considerado faltante se ele é potencialmente observável e significativo para a análise, embora não esteja disponível no momento.

A literatura estatística cita que os mecanismos de dados faltantes recaem geralmente em três categorias: *Missing Completely At Random* (MCAR), *Missing At Random* (MAR), e *Missing Not At Random* (MNAR) (GRAHAM, 2012). O MCAR ocorre quando o valor faltante não está relacionado com seus valores anteriores ou posteriores, e nem com qualquer outra variável da amostra, o MAR ocorre quando o valor faltante não está relacionado com a variável que o contém, e sim com outra variável da amostra, já o MNAR ocorre quando o valor faltante está relacionado com outros valores de sua própria variável.

Nos últimos dez anos têm surgido alguns trabalhos, que utilizam técnicas de aprendizado de máquina para tratar dados faltantes. Como exemplo disso, Nelwamondo *et al*. (2007) utilizaram Redes Neurais Artificiais *Multilayer Perceptron* (RNA-MLP) com algoritmos genéticos em comparação com o algoritmo EM, sendo que na maioria dos casos a RNA MLP com o algoritmo genético apresentou maior acurácia quando comparada ao

algoritmo EM. No trabalho Arslan (2012) também apresenta bons resultados com o uso de RNA MLP para tratar dados ausentes.

Algumas melhorias para acelerar o processo de aprendizado e aumentar a assertividade de RNA também têm sido estudadas, tais como no trabalho de Gomes (2010) que propôs novas funções de ativação para RNA-MLP, as quais apresentaram em geral melhor resultado do que as funções popularmente utilizadas (sigmoide e tangente hiperbólica).

Outra abordagem que pode, também, ser utilizada para melhorar o processo de aprendizagem de RNA-MLP é através da inferência estatística, que fornece uma maneira objetiva de obter algoritmos tanto para a formação quanto para a avaliação do desempenho de aprendizagem, de uma forma mais sistemática. Neste contexto, o treinamento de uma RNA MLP pelo aprendizado supervisionado é equivalente à regressão não linear; o que cria um elo entre estas duas técnicas, possibilitando que muitos métodos de inferência estatística possam ser aplicados às RNAs. Dentre os métodos de inferência estatística, tem-se o eminente método de estimativa de máxima verossimilhança, que quando utilizado para treinar RNA MLP conduz a resultados estatisticamente eficientes e assintoticamente imparciais (YANG; MURATA; ARMARI, 1998).

Diante do que foi ponderado até o momento, o presente trabalho, tem como objetivo abordar o padrão monotônico nos *dataset* analisados, os quais são oriundos de bases públicas, onde apenas uma variável terá dados faltantes, e para tanto é proposto um *framework* baseado Redes Neurais Artificiais MLP com as funções clássicas (Sigmoide e Tangente Hiperbólica), e com as novas funções proposta por Gomes (2010) (Aranda-Ordaz, Complementar Log-Log e Log-Log), para imputar dados faltantes, pelo viés de imputação única e múltipla, porém com o diferencial que neste trabalho aplicar-se-á o método de Estimativa de Máxima Verossimilhança (EMV) em todas as funções de ativação da RNA-MLP, a fim de verificar se tal abordagem proposta apresenta melhor acurácia, quando comparada com a RNA-MLP clássica (sem tal abordagem) e com a RNA-MLP que usa o EMV apenas na função de custo, ou seja, no último neurônio. Também tais resultados foram comparados com o algoritmo EM, que é o estado da arte para tratar dados faltantes.

1.1 Problemática e Hipótese

1.1.1 Problemática

O pré-processamento dos dados é de fundamental importância e extremamente necessário para melhorar a eficiência dos algoritmos de aprendizado de máquina (SRIDEVI *et al*., 2011). Assim, já que cada registro é único, não tê-lo, dificulta ou inutiliza o uso da base de dados. Sendo assim, diante de um mercado globalizado, altamente competitivo, e onde cada empresa busca obter uma maior acurácia em seus modelos preditivos, para auferir mais lucro, é imprescindível ter dados consistentes, de alta qualidade, antes de se iniciar o processo de modelagem.

De acordo com Pereira (2014), uma boa parte das técnicas estatísticas foram projetadas para analisar dados completos. Devido a isso, procura-se sempre tratar tais dados para que estes tornem-se plausíveis de serem analisados por técnicas já consolidadas, tornando a inferência sobre os dados mais precisas. Já Veroneze (2011) cita que existem vários métodos para o tratamento dos dados faltantes, entretanto para que seja factível encontrar o melhor método, é necessário que se identifique algumas particularidades nos dados, como: mecanismos geradores do dado faltante, padrão e quantidade.

Nesse sentido, Sorjamaa (2010) cita que a acurácia de previsão de valores futuros é fortemente dependente não só de um bom modelo, que é bem treinado e validado, mas também do pré-processamento, sendo os valores faltantes não só um incômodo, mas também um fator proibitivo na utilização de certas metodologias e degrada o desempenho de outras. Assim, imputação de valores faltantes é uma parte imprescindível no pré-processamento de um banco de dados. Esta imputação tem de ser feita com cuidado, a fim de manter a integridade da base de dados, e não para inserir quaisquer valores indesejados, pois se assim o fizer, haverá um agravamento de perda de precisão na análise final dos dados.

Dados reais normalmente contêm valores omissos nos atributos, o que causa perda de informação no processo de mineração de dados. Sendo que vários esquemas têm sido estudados e propostos para sanar tal problema, porém não existe uma solução universal para todos os problemas de dados omissos, sendo que para cada problema terá uma técnica que apresentará melhor desempenho. Dentre as técnicas que tem se destacado, pode-se citar as de aprendizado de máquina (*MLP, Weighted Imputation with K-Nearest Neighbor-*

WKNNI, K-means Clustering Imputation-KMI, Imputation with Fuzzy K-means Clustering-FKMI, Support Vector Machines Imputation-SVMI, Event Covering-EC, Regularized Expectation-Maximization-EM, Singular Value Decomposition Imputation-SVDI, Bayesian Principal Component Analysis-BPCA, Local Least Squares Imputation-LLSI, CART, RBFN methods, Naïve-Bayes, Linear Discriminant Analysis classifiers, C4.5, K2, Data Augmentation (DA), BN-K2Iχ^2, 1BN-K2Iχ^2, algoritmo de biclusterização *SwarmBcluster*) (LUENGO; GARCÍA; HERRERA, 2012; HRUSCHKA JR; EBECKEN, 2002; HRUSCHKA JR *et al*. 2007; VERONEZE, 2011). Outra técnica que também merece destaque é o *Autoclass* (LAKSHMINARAYAN; HARP; SAMADI, 1999). Nos trabalhos de Jerez *et al.* (2010) e Duma (2012) há outras abordagens na mesma linha de pesquisa.

1.1.2 Hipótese

A hipótese admitida neste trabalho é que o método para imputar dados faltantes, em dados multivariados com padrão monotônico, via RNA-MLP combinado com o método de EMV aplicado em todas as funções de ativação (em todos os neurônios da rede), bem como na função de custo (neurônio de saída da rede), poderá apresentar maior acurácia, quando comparado à RNA-MLP padrão e também, quando comparado ao algoritmo EM.

1.2 Objetivo

O presente trabalho tem como objetivo principal, criar um *framework*, através da modificação das cinco funções de ativação, abordada neste trabalho, via o método de estimativa de máxima verossimilhança em todos os neurônios, e no neurônio de saída da RNA MLP, para imputar dados em padrões monótonos, pelo viés de imputação única e múltipla. Sendo assim, torna-se possível que se avalie a acurácia do modelo de RNA-MLP, com estas funções de ativação (Aranda-Ordaz, Complementar Log-Log, Log-Log, Sigmoide e Tangente Hiperbólica). Frise-se que este trabalho restringe-se apenas a fase de pré-processamento dos dados.

1.3 Objetivos Específicos

Para tornar viável esta pesquisa, é necessário que alguns objetivos específicos sejam atingidos:

1. Realizar simulações em ambiente artificial, onde sejam inseridos dados ausentes nos *dataset*, nas proporções de 5%, 10%, 20%, 30%, 40%, 50%, 60% e 70%;
2. Implementar e avaliar o desempenho do algoritmo EM;
3. Implementar e avaliar o desempenho da Rede Neural MLP, com todas as cinco funções de ativação aqui analisadas;
4. Implementar e avaliar o desempenho da Rede Neural MLP, com a função de custo modificada pelo método de EMV.
5. Implementar e avaliar o desempenho da Rede Neural MLP, com todas as funções de ativação modificadas pelo método de EMV proposto.
6. O desempenho teve como métricas o MAE (Mean Absolute Error) e RMSE (Root Mean Square Error).

1.4 Metodologia

O caminho metodológico abordado neste trabalho enquadra-se em Pesquisa Experimental, pois haverá a necessidade de manipulações sistemáticas nos dados a serem analisados, a fim de verificar se cada intervenção produz os resultados esperados (WAZLAWICK, 2009). Para tanto, inicialmente fez-se uma vasta Pesquisa Bibliográfica, em registros disponíveis, decorrentes de pesquisas anteriores, em documentos impressos e *on-line*, como livros, artigos, teses, etc.

A pesquisa foi organizada em fases. No inicio do semestre letivo de 2013/2 iniciou-se o levantamento bibliográfico com livros e artigos que abordavam o conteúdo de dados faltantes ou valores omissos, além de outros assuntos relacionados ao tema que não eram o alvo principal da pesquisa, entretanto, imprescindíveis para uma boa compreensão das diversas abordagens. Questionamentos como: Qual a melhor técnica que deve-se abordar para tratar o problema proposto? Quais os avanços mais recentes na literatura? Alguém já propôs ou fez o que está sendo proposto? Foram estas indagações que embasaram a procura por literaturas que auxiliassem a respondê-las.

Em uma segunda fase da pesquisa, houve a necessidade de uma revisão de álgebra linear e matricial, integrais e probabilidade, que embasam as técnicas do algoritmo EM e de RNA MLP. Abordou-se também, nesta fase, o estudo de testes de hipóteses paramétricos e não paramétricos, que poderiam auxiliar na tomada de decisão, e validação dos modelos.

Na terceira fase deste projeto, iniciou-se o estudo da Rede Neural Artificial *Multilayer Perceptron*, aprofundando-se em seu entendimento e suas vantagens para ser utilizada para imputar dados faltantes em padrão monotônico, buscando sempre apoio na literatura científica. Dentro dos conceitos de RNA-MLP foram feitos estudos sobre quais métricas poderiam ser utilizadas para avaliar os resultados. As principais medidas estudadas foram MAE e RMSE.

Na quarta fase, foram feitos novos e afunilados levantamentos bibliográficos, dada a *expertise* adquirida em fases anteriores, no estado da arte em dados faltantes, que trouxe o entendimento de como foi a evolução científica no processo de imputação de dados, e o quanto tal assunto tem despertado interesse nos últimos anos nas mais diversas áreas científicas, sendo que nesse estudo procurou-se também entender como outras técnicas tem sido utilizadas com tal finalidade. Quais os benefícios e desvantagens que essas técnicas proporcionam, que serão discutidos no Capítulo 2 (seção 2.2).

A quinta fase da pesquisa ocorreu após o exame de qualificação do mestrado, onde foram analisadas as bases de dados, que foram encontradas no repositório público (UCI *Machine Learning*), que possibilitaram testar a acurácia de cada método aqui estudado. A vantagem de se trabalhar com os dados reais, é que estes podem conter erros ou ruídos, fato este que auxiliará para a escolha do melhor método diante de tais circunstâncias. As análises dos dados foram processadas com o auxilio do software científico R desenvolvido pela *Foundation for Statistical Computing* e disponibilizado em (http://www.R-project.org). Para a análise do algoritmo EM, utilizou-se o pacote do R denominado "norm", o qual é específico para análise de dados multivariados sob suposição de normalidade.

Obedeceu-se aos seguintes passos:

1. Fez-se o *download* dos dados da fonte pública (UCI *Machine Learning*);
2. Analisaram-se os dados a fim de verificar se havia presença de dados ausentes, situação esta que não foi encontrada, pois todos os dados adquiridos foram de dados completos;
3. Fez-se uma análise preliminar dos dados, através de análises gráficas, estatísticas descritivas e testes estatísticos.
4. Retiraram-se aleatoriamente amostras de cada conjunto de dados, para que estes passassem a conter dados faltantes;

5. As amostras retiradas foram de respectivamente 5%, 10%, 20%, 30%, 40%, 50%, 60% e 70%;
6. Implementou-se e mediu-se o desempenho do algoritmo EM;
7. Implementou-se e mediu-se o desempenho da Rede Neural MLP.

Depois da execução de todos os passos citados acima, iniciou-se a etapa de interpretação dos resultados, com o objetivo de avaliar o desempenho de cada algoritmo e detectar qual apresentou o melhor desempenho para a tarefa de imputar dados.

Por fim, na dissertação, foram solidificados os conhecimentos adquiridos durante toda a pesquisa, assim como os resultados obtidos da análise dos experimentos. Pretende-se também, que este trabalho seja capaz de gerar publicações que possam disseminar conhecimento para o avanço da ciência.

1.5 Contribuições Esperadas

Dentre as contribuições *a priori* que são esperadas deste trabalho, destacam-se:

- Propor uma metodologia de trabalho para imputar dados em *dataset* com padrão monotônico;
- Avaliar o desempenho das técnicas frente a diversos dados reais com seus ruídos inerentes;
- Verificar se as funções de ativação modificadas pelo método de Estimativa de Máxima Verossimilhança traz algum ganho a RNA-MLP;
- Analisar se a RNA-MLP é uma boa alternativa para tratar dados ausentes frente ao algoritmo EM.

1.6 Organização da Dissertação

Esta proposta de Dissertação apresenta-se organizada em sete capítulos, conforme verifica-se nos tópicos a seguir:

- O capítulo 1 aborda, inicialmente, a apresentação do trabalho, com uma introdução ao problema a ser pesquisado seguido de sua problemática, hipótese, objetivo e metodologia.
- O capítulo 2 apresenta uma revisão geral de dados faltantes;
- O capítulo 3 apresenta toda a fundamentação teórica para tratar dados faltantes, entre elas o algoritmo EM;
- O capítulo 4 aborda toda a fundamentação teórica da Rede Neural MLP e traz alguns dos trabalhos da comunidade científica, que estão atualmente relacionados ao trabalho aqui proposto;
- O capítulo 5 apresenta os resultados experimentais
- O capítulo 6 é destinado às conclusões e trabalhos futuros.

CAPITULO 2

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações Iniciais

Neste capítulo são apresentados os principais conceitos atinentes ao problema de dados faltantes, bem como os principais mecanismos geradores deste. Conhecer tais conceitos com suas nuances, é o primeiro e mais decisivo passo para a correta análise dos dados, pois será nestes conceitos que todas as inferências serão alicerçadas.

2.1.1 Definição e Visão Geral

A imputação é um termo genérico para o preenchimento de dados faltantes por valores plausíveis (CHENG, 1998), o que significa dizer que há algum tipo de omissão de informação sobre os fenômenos em que estamos interessados (MCKNIGHT *et al*., 2007). Sendo que a imputação não é apenas uma ferramenta computacional, mas sim um modo de inferência, que permite a avaliação e obtenção de informação dos dados (MENG, 1994). Além disso, a imputação não deve mudar importantes características do conjunto de dados, sendo necessário defini-las previamente para que possam ser mantidas (HRUSCHKA JR; HRUSCHKA; EBECKEN, 2007).

Geralmente não é possível obter inferências válidas se as imputações foram geradas arbitrariamente. Sendo assim, a imputação deve dar previsões razoáveis para os dados faltantes, bem como a variabilidade entre eles deve refletir um adequado grau de incerteza (SCHAFER, 1999).

Viés, variância e erro quadrado médio descrevem o comportamento de uma estimativa, mas também deseja-se honestidade nas medidas de incerteza dos dados analisados, principalmente quando os valores faltantes ocorrem por motivos alheios a nosso controle, fato este que conduz a necessidade de se fazer suposições sobre os processos que o criaram, porém esses pressupostos são geralmente não testáveis, logo as tentativas para recuperar os valores faltantes podem prejudicar a inferência (SCHAFER; GRAHAM, 2002), já que um método de imputação ingênuo ou sem princípios pode gerar mais problemas que resolvê-lo (SCHAFER, 1999). Em geral, dados faltantes dificultam a capacidade de explicar e compreender os fenômenos estudados (MCKNIGHT *et al*., 2007).

Apesar de dados faltantes surgirem em qualquer tipo de dados, os procedimentos de imputação de dados foram inicialmente introduzidos e desenvolvidos no contexto de não resposta em *surveys*, sendo assim, a maior parte da literatura disponível refere-se às aplicações neste domínio (LAKSHMINARAYAN; HARP; SAMADI, 1999). Rubin (1976) deu um exemplo muito comum de dados faltantes, qual seja: em uma pesquisa realizada com várias variáveis socioeconômicas em 1967 e depois em 1970, certamente muitas destas pessoas não serão encontradas, o que consequentemente gerará dados faltantes no banco de dados a ser analisado. Mcknight *et al.* (2007) também fizeram uma pesquisa por três anos em periódicos na área de psicologia e constataram que aproximadamente 90% dos trabalhos continham em suas amostras dados omissos. Foi Rubin que a partir de 1976 desenvolveu um *framework* denominado Imputação Múltipla, para tratar dados ausentes, que é usado até hoje.

De acordo com Mcknight *et al.* (2007) os problemas de dados faltantes surgem de três fontes principalmente: casos omissos, variáveis omissas, e ocasiões faltantes.

Casos omissos ocorrem quando os participantes do estudo falham ao fornecerem dados para um estudo, por exemplo, por motivo de doença.

Variáveis omissas ocorrem quando os participantes falham ao fornecer dados para alguma, mas não todas as variáveis. Por exemplo, ao responder um questionário, no quesito renda o participante não responde.

Ocasiões faltantes ocorrem quando os participantes estão disponíveis para alguns, mas não todos os períodos de coleta de dados em um estudo longitudinal. Por exemplo, pesquisas de obesidade, onde há a necessidade de coletas dos pesos dos pacientes em períodos pré-determinados, é comum o paciente faltar em um dos períodos da coleta, por motivos alheio a pesquisa.

Basicamente, os procedimentos utilizados para o preenchimento de dados faltantes podem ser amplamente divididos entre os que são baseados em modelo, e aqueles baseados em inferência *quasi-randomization*. Procedimentos baseados em modelo destacam-se os dados faltantes oriundos dos mecanismos MAR, MNAR e MCAR que são: imputação por regressão e regressão estocástica, abordagem baseada em verossimilhança (algoritmo EM), e abordagem baseada em aprendizado de máquina (CART – Classificação e árvore de regressão), *Autoclass*, C4.5, entre outros. Quanto aos procedimentos *quasi-randomization* (ou procedimentos orientado à dados) encontram-se: imputação pela média, imputação *Hot Deck*, imputação *Cold Deck* e Substituição (LAKSHMINARAYAN; HARP; SAMADI, 1999).

A escolha entre as diferentes abordagens listadas acima, em grande parte depende da natureza e quantidade dos dados disponíveis, do uso pretendido, da expertise do usuário dos dados, e do mecanismo causador da omissão (LAKSHMINARAYAN; HARP; SAMADI, 1999), além disso, segundo Hruschka *et al*. (2009) a imputação não pode ser devidamente analisada a parte da tarefa de modelagem.

A quantidade de dados faltantes, que é a porcentagem de dados omissos para todas as variáveis pertencentes à análise, tem grande impacto no poder estatístico, visto que quando o tamanho da amostra diminuiu consequentemente reduzirá o erro padrão, e a precisão das estimativas dos parâmetros. Em consequência disso, as conclusões estatísticas serão menos rígidas ou menos assertivas (MCKNIGHT *et al*., 2007).

2.2 Uso da Imputação *versus* Não Uso da Imputação

2.2.1 Uso da Imputação

De acordo com (LITTLE; RUBIN, 1987; RUBIN, 1987) as principais vantagens do uso de imputação podem ser elencadas como:

a) A partir do momento que os dados faltantes foram preenchidos, os métodos padrões de análise de dados completos podem ser usados.

b) É fácil interpretar os resultados da análise e calcular resumos estatísticos necessários.

c) Em muitos casos a imputação pode ser gerada apenas uma vez pelo coletor de dados, o qual, geralmente, detém melhor conhecimento e compreensão sobre o mecanismo que gerou o caso omisso do que um usuário comum.

d) É fácil especificar a estrutura dos dados usando a terminologia de um modelo experimental.

2.2.2 Não Uso da Imputação

Os principais problemas oriundos dos dados faltantes de acordo com (ROTH, 1994; BARNARD; MENG, 1999; LAKSHMINARAYAN; HARP; SAMADI, 1999; MCKNIGHT *et al*. 2007; BRAND *et al*., 1994), podem ser elencados em:

a) A perda de dados diminui o poder estatístico (que é a habilidade do teste estatístico em descobrir uma relação no conjunto de dados), sendo que um alto nível de poder estatístico geralmente requer uma grande amostra.

b) A perda de informações e eficiência.

c) Complicação no tratamento dos dados, cálculos e análise devido às irregularidades nos padrões dos dados e não aplicabilidade de software padrão.

d) Dados faltantes podem estimar parâmetros enviesados, devido às diferenças sistemáticas entre os dados observados e os dados não observados, e pode subestimar o viés dos coeficientes da correlação. Além de que os vieses são difíceis de eliminar uma vez que as razões precisas para não resposta são geralmente não conhecidas

e) Algumas estatísticas podem ser afetadas, tais como medidas de tendência central que podem ser enviesadas para cima ou para baixo, dependendo onde a distribuição dos dados faltantes aparece, se eles estão dispersos ou concentrados.

f) Medidas de dispersão podem também ser afetadas, dependendo de qual parte da distribuição fornece os dados faltantes.

g) Redução da sensibilidade, a qual ocorre quando um modelo incorpora tanto dados disponíveis quanto o conhecimento do analista sobre o mecanismo de dados faltantes, mas ao mesmo tempo mantém um ajuste viável.

h) Há também perda de qualidade, confiabilidade e validade dos dados.

i) Casos com dados faltantes podem diferir sistematicamente de casos completos, de modo que a amostra já não seja representativa.

2.3 Mecanismos Causadores de dados Faltantes

O primeiro passo a ser adotado por um pesquisador ou analista é conhecer o mecanismo que levou o conjunto de dados a ter valores faltantes, pois a partir daí é que se iniciará o processo de escolha da técnica apropriada, para realizar a correta análise dos resultados. Os mecanismos de dados faltantes são classificados como, MCAR, MAR e MNAR.

Para representar matematicamente estes mecanismos de dados faltantes, parte-se do pressuposto que se tem uma matriz de dados coletada Z, com *i* linhas, que correspondem às amostras, e *j* colunas que correspondem as variáveis, sendo assim pode-se dividir Z em dois conjuntos, ou seja, o conjunto com as amostras que contém todas as variáveis observadas, e o conjunto com as amostras que contém variáveis não observadas. Assim, pode-se representar Z por:

$$Z=\{Z_{obs}, Z_{omis}\} \quad (1)$$

Onde o conjunto Z_{obs} refere-se aos dados presentes e Z_{omis} corresponde aos dados omissos (faltantes). Desta forma tem-se $z_{ij}=(z_{i1}, z_{i2}, ..., z_{in})$, onde cada z_{ij} refere-se ao valor da amostra *i* na variável *j*. Frise-se que, a cada conjunto Z existe um identificador de dado faltante associado, denotado por S, o qual deve ter as mesmas dimensões de Z. Sendo assim, tem-se que $s_{ij}=0$, se o dado é faltante, e $s_{ij}=1$ quando o dado está presente. Desta forma, a distribuição condicional do mecanismo de dados faltante pode ser representada por P(S|Z).

2.3.1 MAR

Dados faltantes são considerados MAR quando a probabilidade de um registro com um valor em falta para um atributo pode depender dos dados observados, mas não do valor dos dados faltantes em si (LAKSHMINARAYAN; HARP; SAMADI, 1999). Em outras palavras, o MAR permite as probabilidades do mecanismo de dados faltantes dependerem de dados observados, mas não de dados faltantes (SCHAFER; GRAHAM, 2002).

Para entender a suposição MAR, considere um conjunto de dados bivariado simples, com uma variável X, que é sempre observado e uma segunda variável Y que às vezes não é observada ou registrada. Assim, a probabilidade de que Y esteja ausente para um indivíduo da amostra pode estar relacionado com o valor do indivíduo da variável X, mas não com o seu próprio valor de Y. Em uma relação estatística (no sentido de regressão) entre Y e X, então pode-se regredir Y em X para os indivíduos entrevistados e, em seguida, usar a relação estimada para obter previsões não enviesadas de Y para os dados faltantes (SCHAFER; OLSEN, 1998). Este mecanismo, também é conhecido como não-resposta ignorável (MOHAMED; MARWALA, 2005).

De acordo com o que foi apresentado na seção 2.3, e exposto aqui, pode-se representar matematicamente o MAR por:

$$P(S|Z) = P(S|Z_{obs}) \qquad (2)$$

2.3.2 MNAR

O mecanismo gerador de dados faltantes MNAR ocorre quando a probabilidade de um registo com um valor faltante em um atributo pode depender do valor do atributo. Exemplos, um sensor pode não detectar temperaturas abaixo de um determinado limite, pessoas não preenchem a renda anual em pesquisas se a renda exceder um determinado valor (LAKSHMINARAYAN; HARP; SAMADI, 1999). Muitas publicações recentes focam

MNAR como uma preocupação séria em ensaios clínicos, em que os participantes podem sair por razões diretamente relacionadas com a resposta a ser medida (SCHAFER; GRAHAM, 2002, LIUBLINSKA, 2013). É também conhecido como o caso não ignorável (MOHAMED; MARWALA, 2005).

Seguindo o que foi exposto na seção 2.3 e 2.3.2, temos o seguinte modelo matemático representando o MNAR:

$$P(S|Z) \neq P(S|Z_{obs}) \quad (3)$$

2.3.3 MCAR

A suposição MCAR ocorre quando, a probabilidade de um registro que tem um valor em falta para um atributo não depende nem do valor observado dos dados e nem do valor faltante. Esta suposição é muito forte, o que conduz a não ser satisfeita na prática, logo na vida real esta suposição não é utilizada. De acordo com Pereira (2014) quando isso ocorre, os dados não observados constituem uma sub amostra aleatória.

Isso significa que nenhuma das variáveis, dependente (Y) ou independente (X), tem *scores* faltando relacionados com os valores da própria variável (ALISSON, 2001).

Suponha que há dados faltantes sobre uma particular variável Y. Dados em Y são ditos ser MCAR se a probabilidade de dados faltantes em Y não está relacionada ao valor do próprio Y ou ao valor de quaisquer outra variável no conjunto de dados. Quando este pressuposto está satisfeito para todas as variáveis, o conjunto de indivíduos com dados completos pode ser considerado como uma sub-amostra aleatória simples do conjunto original de observações. Note-se que MCAR não permite a possibilidade de que o mecanismo de "dados faltantes" em Y esteja relacionado com o mecanismo de "dados faltantes" em algumas outras variáveis X (ALISSON, 2001).

Conforme a descrição apresentada na seção 2.3, o MCAR é representado matematicamente por:

$$P(S/Z) = P(S) \quad (4)$$

CAPITULO 3

3 MÉTODOS PARA TRATAR OS CASOS DE DADOS FALTANTES

3.1 Imputação única

A imputação única ou também conhecida como imputação simples, preenche por um único valor cada dado faltante na amostra.

Esta técnica dá estimativas razoáveis com cálculos padrões, mas não indicam a sensibilidade de inferências para o esquema de imputação (LITTLE; RUBIN, 1987).

Tem a vantagem de poder usar os métodos padrões de dados completos, para o conjunto de dados preenchidos. Em bases de dados de uso público, há geralmente a necessidade de gerar imputação sensível que precisem ser realizadas apenas uma vez, pelo analista, fato este que pode incorporar o conhecimento do mesmo (RUBIN, 1987). Se a proporção de valores em falta é pequena, preferencialmente menos de 5%, então a imputação única pode ser bastante razoável, pois sem medidas corretivas especiais, as inferências de imputação única para um escalar estimado podem ser bastante precisas (SCHAFER, 1999).

Traz consigo a desvantagem de que a imputação de um único valor não captura a variabilidade da amostra do valor imputado, e nem a incerteza associada ao modelo utilizado para a imputação (LAKSHMINARAYAN; HARP; SAMADI, 1999; RUBIN, 1987), podendo causar subestimativas da variância para as variáveis com dados faltantes, e às vezes, covariâncias também (ALISSON, 2001).

3.2 Imputação Múltipla

Imputação múltipla (IM) é uma técnica estatística desenvolvida para tirar vantagem da flexibilidade em cálculos para tratar dados faltantes. Com isso, cada valor faltante é substituído por dois ou mais valores imputados, ao invés de apenas um valor, a fim de representar a incerteza sobre qual valor imputar (RUBIN, 1987), permitindo que as estimativas das variâncias estimadas sejam calculadas usando procedimentos de dados completos (LITTLE; RUBIN, 1987).

As m imputações atribuídas a cada valor faltante gera n conjuntos de dados completados, sendo que cada um destes conjuntos de dados completados é analisado através dos procedimentos padrões para dados completo, como se estes fossem os dados realmente

obtidos caso tivessem sido coletados ou registrados. Esta técnica é muito utilizada no contexto de pesquisas amostrais, já que os dados coletados serão analisados por vários usuários, o que cria a necessidade de se tratar as não respostas ou dados faltantes antes deste chegar ao usuário final. Neste contexto, procura-se organizar tais dados, tornando-os completos, sem lacunas, e aptos a serem usados.

De acordo com Rubin (1987), os *n* conjuntos são ordenados, no sentido de que o primeiro conjunto de valores imputados para os valores faltantes sejam usados para formar o primeiro conjunto de dados completos, e assim por diante.

Dentre as vantagens apontadas pelo uso da IM, pode-se elencar como sendo as principais:

-Incorporar o conhecimento do coletor de dados, permitindo que este use seus conhecimentos para refletir incerteza sobre os valores imputados.

- Quando imputações são sorteadas aleatoriamente tentando representar a distribuição dos dados, imputação múltipla aumenta a eficiência da estimação, refletindo uma variabilidade adicional, simplesmente obtida pela combinação de inferências de dados completo de uma maneira direta.

- Facilita o estudo direto da sensibilidade de inferências, de vários modelos para não resposta simplesmente usando métodos de dados completo repetidamente.

- Em muitas aplicações, apenas 3 ou 5 imputações são suficiente para obter excelente resultados.

- um conjunto de *n* imputações pode ser usado para uma variedade de análises; muitas vezes não há necessidade de reimputar quando uma nova análise é realizada.

Além das vantagens elencadas acima, tem-se também que as inferências de erro padrão, p-valores, etc., obtidas a partir de IM são geralmente válidas porque incorporam incerteza devido à falta de dados, tornando IM atraente porque pode ser altamente eficientes mesmo para pequenos valores de *n* (SCHAFER; OLSEN, 1998). Conforme cita MENG (1994), os estimadores baseados em imputação múltipla são mais eficientes que aqueles baseados em imputação simples, além de que conduzir inferências requer apenas repetir o mesmo padrão de análise de dados completos várias vezes. Outra vantagem deste método é evitar subestimação da verdadeira variância (CHENG, 1998).

Dado que IM é um método estatístico, então este se baseia em certos pressupostos, que são fundamentais conhecê-los antes de se iniciar qualquer análise. Tais pressupostos são:

conhecer a distribuição *a priori* para os parâmetros do modelo, e o mecanismo causador dos dados faltantes.

Analistas experientes sabem que os dados reais raramente estão em conformidade com os modelos convenientes, tais como a normal multivariada. Na maioria das aplicações de IM, o modelo utilizado para gerar as imputações será na melhor das hipóteses apenas uma aproximação da realidade. A experiência tem repetidamente mostrado que IM tende a ser bastante flexível a partir do modelo de imputação. Por exemplo, quando se trabalha com variáveis categóricas binárias ou ordenadas, muitas vezes é aceitável para imputar, admitir um pressuposto de normalidade (SCHAFER; OLSEN, 1998).

Algumas desvantagens, apontadas por Rubin (1987), para o uso da IM são:

- Necessita-se mais trabalho para produzir imputação múltipla que imputação simples.

- Necessita-se mais espaço para armazenar um conjunto de dados múltiplo-imputado.

- Necessita-se mais trabalho para analisar um conjunto de dados múltiplo imputado do que um conjunto de dados simples imputado.

Outra desvantagem é que pode surgir discrepância na variância quando se admite pressupostos equivocados, logo o modelo escolhido é inconsistente para imputar os dados, ou seja, o procedimento de análise não corresponde ao modelo imputado.

De acordo com MENG (1994), a incompatibilidade surge quando o analista ou imputador têm acesso a diferentes níveis e fontes de informação, e têm diferentes avaliações (por exemplo, modelo explícito, opiniões implícitas) sobre ambas as respostas e não respostas.

Estas desvantagens não são graves quando *m* é pequeno. Poucos *m* são adequados quando frações de informações faltantes são pequenas. Quando frações de informações faltantes são grandes, poucos *m* de imputação múltipla não são totalmente satisfatórios (RUBIN, 1987).

A incerteza gerada pelo método de IM é simplesmente um reflexo da variação mútua que ocorre entre os conjuntos de dados imputados, sendo que quando ocorre pouca variação mútua entre os *dataset* imputados, infere-se que existe pouca incerteza acerca dos dados omissos, e quando ocorre muita variação, conclui-se que existe muita incerteza quanto aos dados faltantes, o que conduzirá a decisões imprecisas.

De acordo com Brand *et al*. (1994), as três principais fontes de incerteza podem ser classificadas em:

a) A variação da amostra,

b) O mecanismo causador dos dados faltantes

c) O número finito de imputações usadas

O número finito de imputações é também uma fonte de incerteza, porque a partir de repetidas aplicação do algoritmo de imputação múltipla, diferentes resultados finais são obtidos (BRAND *et al*., 1994).

A imputação múltipla é um método constituído por três passos para manipular dados faltantes, que são (BARNARD; MENG, 1999):

a) No primeiro passo, $n > 1$ conjunto de dados completados são gerados.
b) No segundo passo, m análises de dados completos são realizadas por procedimentos padrões.
c) No terceiro passo, os resultados das m análises dos dados completos são combinados de maneira simples e conveniente para obter as inferências necessárias.

De acordo com Schafer (1999), a quantidade de imputações necessárias, para que uma estimativa de conjunto de dados tenha relativa eficiência, pode ser determinada pelo modelo matemático:

$$\sqrt{1 + \frac{\lambda}{m}} \tag{5}$$

Onde λ (*lambda*) é a taxa de informação faltante e m é quantidade de conjunto de dados completados. Como exemplo, suponha que a quantidade de dados faltantes seja de 50%, e a quantidade de conjuntos completados seja de 5, ou seja, m=5 imputações tem um desvio padrão que é apenas 5% mais amplo do que uma baseada em $m \to \infty$ porque $\sqrt{1 + \frac{0.5}{5}} = 1.049$. A não ser que as taxas de informações omissas sejam não usualmente altas, há uma tendência de ter pouco ou não prático benefício usar mais que 5 a 10 imputações.

Uma das características mais importantes no método de IM é que os valores faltantes para cada participante é predito a partir de seus próprios valores observados, com o ruído aleatório adicionado para preservar uma correta quantidade de variabilidade nos dados imputados (SCHAFER; GRAHAM, 2002).

Com o discorrer da leitura deste texto, observa-se que a validade do IM depende fortemente dos valores imputados, logo é indispensável compreender a metodologia para obter valores imputados de forma que suas estimativas sejam imparciais com um correto intervalo de confiança, pois ao desconsiderar os aspectos relevantes para a criação de modelos

de imputação, estes podem impactar na validade das inferências (JOLANI; VAN BUUREN; FRANK, 2013).

Uma estimativa Q é uma estatística que se tem interesse em medir, desde que se tenha observado toda a população. Como exemplo, podemos citar que estamos interessados em saber a renda média dos alunos universitários no Brasil, então deve-se coletar tal informação de todos os alunos. Caso o estudo tenha interesse em mais de uma estatística, Q será um vetor. Em Van Buuren (2012), encontram-se alguns exemplos do que é e do que não é uma estatística de interesse, que podem ser elencadas como:

Exemplos de medidas que são estatísticas.

- Média Populacional

-Covariância ou Correlação Populacional

- Coeficientes de Regressão

Exemplos de medidas que não são estatísticas.

- Médias Amostrais

- Erro Padrão

- Testes Estatísticos

Dado que o principal objetivo da IM é encontrar uma estimativa $\widehat{\mathrm{Q}}$ da estatística para cada parâmetro estimado, que não seja enviesado.

O primeiro passo a ser dado é combinar os resultados das repetidas imputações, que seria a estimativa global, representada por $\overline{\mathrm{Q}}$.

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m} \hat{Q}_l \tag{6}$$

Esta equação é apenas a média de todas as estimativas, utilizadas para imputar os dados no *dataset*. Onde $\widehat{\mathrm{Q}}_{\mathrm{l}}$ contém k parâmetros, sendo ela representada por um vetor coluna de k x 1. Para uma dada quantidade de número de m imputações deve-se ter a seguinte equação:

$$\bar{U} = \frac{1}{m}\sum_{l=1}^{m} \bar{U}_l \tag{7}$$

Onde o $\overline{U}_l$ representa a matriz de variância-covariância de $\widehat{Q}_l$, obtida a partir da *i*ésima imputação. Quanto à estimativa não enviesada da variância, entre a estimativa dos dados completos *m*, deve ser dada por:

$$B = \frac{1}{m-1}\sum_{l=1}^{m}(\hat{Q}_l - \bar{Q})(\hat{Q}_l - \bar{Q})' \quad (8)$$

Deve-se prestar a atenção quanto à variância total, pois não se pode concluir que T (variância total), seja simplesmente dada por T = $\overline{U}$ + B, pois deve-se levar em conta, que o próprio $\overline{Q}$ é estimado usando quantidades finitas de *m*, logo ele só se aproxima de uma $\overline{Q} \rightarrow \infty$. Sendo assim, desde que B se aproxime de B$\rightarrow\infty$, pode-se reescrever a variância total como:

$$T = \bar{U} + B + \frac{B}{m} \quad (9)$$

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B \quad (10)$$

Em virtude dos pesquisadores, normalmente, preferirem relatar seus resultados na métrica de erro padrão, do que na métrica de variância, então reescreve-se a variância total como sendo (ENDERS, 2010):

$$S = \sqrt{T} \quad (11)$$

Onde S representa o desvio padrão.

Uma importante característica da técnica de IM, é que pode-se estimar a eficiência dos parâmetros com o uso de uma simples equação, qual seja:

$$\frac{1}{1 + \frac{\gamma}{m}} \quad (12)$$

Onde o γ representa a taxa de informação faltante para a quantidade estimada e *m* é igual ao número de imputações. Frise-se que os valores de γ podem variar entre 0 e 1, sendo que quando ele recebe o valor de 1, existem na variável analisada 100% de dados faltantes. Esta equação apresenta uma eficiência relativa das inferências da IM, que está relacionada à

taxa de informação faltante γ em combinação com o número de imputações *m*. A taxa de informação faltante é relacionada ao incremento na variância devido aos dados faltantes.

Uma grande vantagem da IM é permitir que se estime o intervalo de confiança, através da equação:

$$\bar{Q} \pm t_{df}(a/2)\sqrt{T} \tag{13}$$

Onde α representa o nível de significância, e df (do inglês: *degrees of freedom*) que são os graus de liberdade. Normalmente, utiliza-se um nível de significância de 5%, sendo assim tem-se um nível de confiança de 95%, o que possibilita reescrever a equação como:

$$\bar{Q} \pm 1.96\sqrt{T} \tag{14}$$

Cabe ressaltar que o valor de *t* é similar ao teste *t* de Student, o qual pode ser calculado por:

$$t(df) = \frac{\bar{Q}}{\sqrt{T}} \tag{15}$$

Para se saber os graus de liberdade para a os dados analisados, há a necessidade de conhecer os valores de *m*, $\bar{U}$, e B. Para tanto recorre-se a fórmula:

$$df = (m-1)\left(1 + \frac{m\bar{U}}{(m+1)B}\right)^2 \tag{16}$$

Uma forma de se calcular a taxa de informação faltante, através dos graus de liberdade e do incremento relativo da variância, é através da equação:

$$\gamma = \frac{r + 2/(df+3)}{r+1} \tag{17}$$

Sendo o valor de *r* determinado através da seguinte fórmula:

$$r = \frac{(1+m^{-1})B}{\bar{U}} \tag{18}$$

3.3 Métodos Baseados em Deleção e Imputação

Existem duas formas de se tratar uma base dados com valores faltantes, sendo que a primeira consiste em excluir todos os casos com dados incompletos, seja em todas as variáveis ou apenas na variável que será analisada. Já a segunda opção é baseada em métodos de imputação, os quais executam imputações no *dataset* através de medidas de centralidade, modelos estatísticos e de aprendizado de máquina.

3.3.1 Métodos Baseados em Deleção

3.3.1.2 Deleção dos Casos Incompletos

A simplicidade é a principal vantagem do método de deleção dos casos incompletos. No entanto, uma importante desvantagem deste método é a potencial perda de dados coletados com alto custo (BRAND *et al*., 1994). Esta abordagem é viável apenas em situações em que esses registros constituem uma percentagem ignorável do total de dados, e nenhum viés significativo é introduzido por sua eliminação, ou seja, quando o número de registros incompletos é muito pequeno em comparação com o número total de registos. Entretanto ignorar registros incompletos geralmente não é uma boa opção para bases de dados industriais (LAKSHMINARAYAN; HARP; SAMADI, 1999), além de que informações valiosas podem estar sendo descartadas via este método, o que pode ser inadequado (ENNETT; FRIZE, 2003).

3.3.1.3 Deleção *Listwise*

Esta técnica elimina todos os casos com qualquer quantidade de dados faltantes nas variáveis, a partir do cálculo ou séries de cálculos tal como a matriz de correlação, e em seguida, aplica métodos convencionais de análise de conjuntos de dados completos. Conforme Roth (1994) esta técnica sacrifica uma grande quantidade de dados, e pode resultar em perdas ainda maiores de dados porque os sujeitos são frequentemente observados múltiplas vezes. Também é conhecida como análise de casos completos (FICHMAN; CUMMINGS, 2003).

Alisson (2001) cita duas grandes vantagens óbvias para eliminação *listwise*:

a) ela pode ser usada para qualquer tipo de análise estatística, a partir de modelagem de equações estruturais à análise de log-lineares;

b) métodos computacionais especiais não são necessários.

Apenas os casos que não têm dado faltando sobre todas as variáveis independentes e dependentes são considerados para análise. Situações quando há falta de dados, mesmo modestas pode levar a uma grande percentagem de redução nos casos completos, mesmo quando há um pequeno número de variáveis em uma análise (FICHMAN; CUMMINGS, 2003), isso pode ocorrer pelo fato de que cada amostra pode ter um valor faltante para apenas uma variável e não necessariamente para todas as variáveis.

Dependendo do mecanismo de dados faltantes, deleção *listwise* também pode ter algumas propriedades estatísticas atrativas. Especificamente, se os dados são MCAR, então a amostra reduzida será uma sub-amostra aleatória da amostra original. Isto implica que, para qualquer parâmetro de interesse, se as estimativas forem não enviesadas para o conjunto de dados completo, eles também serão não enviesados para o conjunto de dados excluídos por *listwise*. Além disso, os erros padrões e as estatísticas dos testes obtidos com o conjunto de dados excluídos *listwise* será tão apropriado como eles teriam sido no conjunto de dados completo. É claro, o erro padrão geralmente será maior no conjunto de dados excluídos por *listwise* porque menos informação é utilizada (ALISSON, 2001).

Eliminação *Listwise* conduz a inferências válidas quando os dados são MCAR, desde que MCAR implique nos casos completos como sendo uma amostra aleatória de todos os casos. O caso mais geral é que *listwise* gera inferências válidas se os dados faltantes sobre as variáveis de previsão não tenha dependência da variável resposta (FICHMAN; CUMMINGS, 2003).

3.3.1.4 Deleção *Pairwise*

Esta técnica deleta apenas as amostras com dados omissos nas variáveis que serão necessárias para a análise, e também é conhecida como análise de casos disponíveis. Esta abordagem causa perda clara de informação que está disponível nos dados eliminados.

A eliminação *pairwise* é uma alternativa simples que pode ser usada por muitos modelos lineares, incluindo regressão linear, análise fatorial e modelos mais complexos de equações estruturais (ALISSON, 2001), sendo sua ideia principal calcular cada um destes resumos estatísticos usando todos os casos que estão disponíveis. Ela poderá também conduzir a correlações inconsistentes matematicamente, e se houver multicolinearidade, existe o risco de que a matriz de correlação não seja positiva definida (ROTH, 1994), tornando impossível o uso de algumas técnicas, já que um dos requisitos é que a matriz de correlação seja definida positiva, incluindo análise de regressão múltipla. Alisson (2001) frisa que nesta

técnica, para calcular a covariância entre duas variáveis X e Z, todos os casos que têm dados presentes para ambos X e Z são utilizados. Uma vez que as medidas resumos tenham sido calculadas, eles podem ser usados para calcular os parâmetros de interesse, por exemplo, os coeficientes de um modelo de regressão.

Esta técnica é muitas vezes oferecida em pacotes de análise estatística que é aplicado para o cálculo da estatística descritiva (GRAHAM; HOFER; PICCININ, 1994), porém esta técnica é problemática porque a amostra para cada correlação é diferente (FICHMAN; CUMMINGS, 2003).

3.3.2 Métodos Baseados em Inferência *Quasi-Randomization*

Na Inferência *quasi-randomization* imputa-se um valor a partir de medidas de tendência central ou através de amostragem aleatória, da mesma base ou de base semelhante usada em períodos anteriores.

3.3.2.1 Imputação pela média

Esta técnica permite que se substitua um valor omisso pela média dos valores presentes na variável de interesse. Ela também é conhecida como *Unconditional mean imputation* (FICHMAN; CUMMINGS, 2003). Às vezes, esta abordagem pode conduzir os valores imputados a resultados razoáveis, entretanto não leva em consideração a relação entre os atributos, que é útil no processo de tratamento dos valores faltantes, visto que vários autores também argumentam que é mais importante preservar as relações entre os atributos do que obter previsões mais precisas (HRUSCHKA JR; HRUSCHKA; EBECKEN, 2007). Apesar de serem fáceis de usar, outros aspectos da sua distribuição são alteradas com um potencial de sérias consequências, que podem ser elencadas como desvantagens.

- Conduz a uma estimativa de variância atenuada (ROTH, 1994), principalmente quando há uma grande quantidade de dados omissos.

- A variância da variável imputada e a sua covariância com as outras variáveis são sistematicamente subestimada (LAKSHMINARAYAN; HARP; SAMADI, 1999).

- A média da variável é preservada, mas outros aspectos da sua distribuição, como quantis são alteradas (SCHAFER; GRAHAM, 2002).

-Estimativas de quantidade que não são lineares nos dados, tal como a variância ou a correlação entre um par de variáveis, não pode ser estimado consistentemente usando o método padrão de dados completo nos dados completados.

-Este método altera a distribuição empírica dos valores Y amostrados, que é importante quando se estuda a forma da distribuição de Y usando histogramas ou outros plotes dos dados.

3.3.2.2 Imputação pela Mediana

A mediana é uma das medidas de tendência central, que para ser usada necessita primeiramente ordenar os dados, e em seguida escolher a amostra que divide este conjunto de dados no meio, ou seja, em partes iguais.

A mediana proporciona um melhor resumo da distribuição, e assim uma melhor estimativa para valores faltantes, visto que a mediana, frequentemente tem um bom desempenho como uma medida de tendência central, quando a distribuição desvia muito da distribuição normal padrão (MCKNIGHT *et al*., 2007).

3.3.2.3 Imputação por Zero

Imputação por zero é muito utilizada em pesquisas amostrais, onde a resposta pode ser uma variável binária, como por exemplo, sim ou não, concordo ou discordo, aceito ou não aceito. Esta abordagem é muito arriscada, visto que é muito dependente do conhecimento do analista, pois será ele quem decidirá com sua expertise qual melhor resposta para a situação do valor omisso, ou outra abordagem é que pode-se automatizar tal resposta por algum modelo, tal como o modelo de regressão logística binária, que é muito utilizada por operadoras de cartão de crédito, para conceder ou não um cartão a um cliente, ou para aumentar ou não o limite de crédito do cliente.

Tal técnica é bastante comum também nas ciências sociais, principalmente na psicométrica, onde substitui-se os dados faltantes por 0, onde 0 pode indicar falha na medida de interesse. Naturalmente, este método só é apropriado em casos em que 0 é um valor plausível (MCKNIGHT *et al*., 2007).

3.3.2.4 Imputação por substituição

Este método lida com unidades não respondidas na fase de trabalho de campo de pesquisas amostrais, onde uma unidade (caso ou registro) não responde ao questionário. Nesta situação, este caso, é substituído por outro que foi originalmente excluído da amostra. Este método não é aplicável no caso de bancos de dados industriais (LAKSHMINARAYAN; HARP; SAMADI, 1999; LITTLE; RUBIN, 1987).

3.3.2.5 *Hot Deck*

No *Hot Deck* pode-se substituir um valor faltante com o score atual a partir de um caso similar no conjunto de dados atual (ROTH, 1994), ou seja, para cada caso faltante, este irá ser preenchido por outro valor semelhante presente na própria variável (LAKSHMINARAYAN; HARP; SAMADI, 1999), apresentando a vantagem de que todos os valores imputados são valores realmente observados, e consequentemente, não há valores fora do intervalo amostral, além de que a forma da distribuição tende a ser preservada e tende a ter uma maior acurácia (ALISSON, 2001; ROTH, 1994).

Sua desvantagem é que há pouca teoria ou trabalhos empíricos que determinem sua acurácia, e o número de variáveis classificadas pode tornar-se intratável em grandes pesquisas (ROTH, 1994), além de que todas as variáveis preditoras devem ser categóricas (ou tratadas como tal), o que impõe sérias limitações no número de possíveis variáveis preditoras, sacrificando informação (ALISSON, 2001).

Esta abordagem é frequentemente utilizada pelo *U.S. Census Bureau* para produzir valores imputados para conjunto de dados de uso público (ALISSON, 2001).

3.3.2.6 *Cold Deck*

Preenche um valor faltante por um valor de outro conjunto de dados não atualmente em uso (ROTH, 1994; LITTLE; RUBIN, 1987). Ela se assemelha muito ao *Hot Deck*. Por exemplo, valores baseados em amostras de dados anteriores podem ser utilizados num procedimento *cold deck* (LAKSHMINARAYAN; HARP; SAMADI, 1999). Esta técnica tem desvantagens, quando comparada à *Hot Deck*, principalmente, devido ao fato de que os dados que são imputados são oriundos de fonte externa, a qual pode variar sistematicamente do conjunto de dados primários, o que pode conduzir a um nível adicional de viés para o parâmetro estimado. Esta técnica não tem sido amplamente adotada e também não é geralmente recomendada (HAUKOOS; NEWGARD, 2007).

3.3.3 Métodos Baseados em Modelos Estatísticos

Ao utilizar os modelos estatísticos, têm-se os que são baseados em regressão e os que são baseados em modelos probabilísticos, sendo que este último tem como referência o algoritmo EM.

3.3.3.1 Imputação por Regressão

Este método substitui dados faltantes por valores preditos a partir de um modelo de regressão (LITTLE; RUBIN, 1987), ou seja, imputa os dados omissos baseado em outras variáveis no conjunto de dados. Ele é a melhor forma de captar as características da distribuição da variável X, mas ainda subestima o erro padrão e a variabilidade devido à imputação (FICHMAN; CUMMINGS, 2003).

Os valores faltantes para variáveis contínuas podem ser previstos utilizando modelos de regressão linear ou polinomial, enquanto que os valores das variáveis binárias podem ser previstos por meio de regressão logística (LAKSHMINARAYAN; HARP; SAMADI, 1999). Frise-se que existem vários modelos de regressão, desde os modelos de regressão linear simples ou múltiplo, aos modelos robustos de regressão. Sendo uma segunda variante da imputação por regressão a *stepwise regression* ou abordagem de regressão iterativa. *Stepwise regression* procura isolar apenas poucas variáveis chaves que contribuam para imputar (ROTH, 1994). Um exemplo de aplicação dos modelos de regressão pode ser verificado em muitos casos práticos, principalmente, na previsão de esforço de software (MYRTVEIT; STENSRUD; OLSSON, 2001).

Em ciências empíricas, os parâmetros de interesse científico muitas vezes são os coeficientes de regressão de uma modelo linear, que são predominantemente estimados usando estimador de Mínimos Quadrados Ordinários (MQO). Quando há valores em falta nas variáveis de previsão, e a probabilidade de valores em falta depende da variável resposta dada as covariáveis, a análise de caso completo (CCA), conduzirá geralmente para inferências inválidas (DE JONG; VAN BUUREN; SPIESS, 2014).

Na maioria dos pacotes estatísticos, observa-se que quando se utiliza modelos de regressão, nos casos em que no conjunto de dados há valores faltantes, eliminam-se da análise as amostras com dados omissos.

3.3.3.2 Imputação por Regressão Estocástica

A regressão estocástica substitui valores faltantes por um valor predito por uma regressão mais a adição de um termo de ruído, moldando uma incerteza no valor predito, já que com modelos de regressão linear normal, o resíduo irá ser naturalmente normal, com média zero e variância igual a 1 (LITTLE; RUBIN, 1987; LAKSHMINARAYAN; HARP; SAMADI, 1999). Esta abordagem incorpora incertezas, portanto, melhora as estimativas,

principalmente quando os coeficientes da regressão variam não sistematicamente na família de modelos lineares (ROSENBERG, 1973).

3.3.3.3 Método de Máxima Verossimilhança

O método de máxima verossimilhança foi proposto por Fisher (1912), o qual é um método paramétrico, de modo que os pressupostos da distribuição feitos são centrais (SORENSEN; GIANOLA, 2002). Sendo que este método parte do princípio de especificar como a função de verossimilhança deveria ser utilizada como um dispositivo de redução de dados (CASELLA; BERGER, 2010). Fisher (1922) definiu a verossimilhança como: "*The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed*", enfatizando que a verossimilhança não é sinônimo de probabilidade, pois ela não obedece às leis matemáticas probabilísticas, e propõe usar o termo verossimilhança para designar o estado da informação *a priori* no que diz respeito aos parâmetros de uma população hipotética.

De acordo com Fisher (1922) o método de máxima verossimilhança consiste simplesmente na escolha do conjunto de valores para os parâmetros que torne um máximo a função de verossimilhança, o que equivale segundo Batista (2009) a encontrar o valor para o parâmetro que torne mínima a função de log-verossimilhança negativa. Desta forma as inferências auferidas deste método devem ser baseadas na função de verossimilhança (COX, 2006). Cordeiro (1999) ressalta que a teoria da verossimilhança representa um dos métodos mais comuns de inferência estatística, e que seu uso tornou-se crescente a partir de 1930 devido à sua grande contribuição aos problemas de experimentação agrícola.

A função de verossimilhança pode ter vários máximos locais, e pontos de sela na superfície da verossimilhança (MCCULLAGH, 2009; REID; COX, 2013), entretanto, a situação mais comum é que o máximo global seja dominante, principalmente na família exponencial, onde os argumentos de convexidade podem ser utilizados para demonstrar que o log verossimilhança tem um único máximo, possibilitando assim a correta interpretação, e garantindo que a estimativa de máxima verossimilhança esteja perto do valor verdadeiro (REID; COX, 2013).

A função de verossimilhança $L(\theta)$ é definida como sendo igual à função do modelo, embora seja interpretada diferentemente como função de θ para um valor x conhecido. Assim, $L(\theta) = f(x; \theta)$. A inferência de verossimilhança pode ser considerada como um processo de

obtenção de informação sobre um vetor de parâmetros θ, a partir do ponto x do espaço amostral, através da função de verossimilhança L(θ). Vários vetores x's podem produzir a mesma verossimilhança ou, equivalentemente, uma dada verossimilhança pode corresponder a um contorno R(x) de vetores amostrais. Este processo produz uma redução de informação sobre θ, disponível em x (CORDEIRO, 1999).

Definição 1: Suponha que temos $x=(x_1,...,x_n)$, sendo os x_{is} os valores observados de uma amostra aleatória i.i.d. de tamanho *n* da variável aleatória x, com função massa de probabilidade f(x;θ), associada ao parâmetro θ com $\theta \in \Theta$, onde Θ é o espaço paramétrico. Então, a função de verossimilhança θ pode ser definida por (BOLFARINE; SANDOVAL, 2010):

$$L(\theta; x_1, \dots, x_n) = f(x_1; \theta) * \dots * f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta). \tag{19}$$

Geralmente utiliza-se o log-verossimilhança l(θ;x) = log L(θ;x), que é o logaritmo natural da função de verossimilhança de θ, que pode também ser chamado de função suporte. Assim, como a função logarítmica é monótona, constata-se que o valor de θ no espaço paramétrico Θ que maximiza a função de verossimilhança L(θ;x), também irá maximizar l(θ;x). Como a função de verossimilhança é uma função de θ, pode-se reescrevê-la assim:

$$l(\theta) = l(\theta; x) \tag{20}$$

$$l(\theta; x) = \log L(\theta; x) \tag{21}$$

Dado que o objetivo é encontrar uma estimativa de θ, ou seja, um $\hat{\theta}$ que maximize a verossimilhança, pode-se afirmar que a estimativa de máxima verossimilhança é um valor $\hat{\theta}$ tal que $L(\hat{\theta}) \geq L(\theta)$ para todo $\theta \in \Theta$. Assim, o estimador de máxima verossimilhança que se quer encontrar, pode ser obtido pela diferenciação de (21).

$$l'(\theta; x) = \frac{\partial l(\theta; x)}{\partial \theta} = 0 \tag{22}$$

Para inferir se a equação (22) é um ponto de máximo, deve-se aplicar a segunda derivada e verificar se o resultado é menor do que zero (BOLFARINE; SANDOVAL, 2010).

$$l''(\hat{\theta};x) = \frac{\partial^2 \log L(\theta;x)}{\partial \theta^2} < 0 \qquad (23)$$

Uma representação gráfica da verossimilhança é apresentada na Figura 1, a qual tem dois gráficos representando o mesmo conjunto de dados, sendo que no primeiro gráfico (a) foi usada apenas a função de verossimilhança e no segundo gráfico (b) foi usada a função log-verossimilhança.

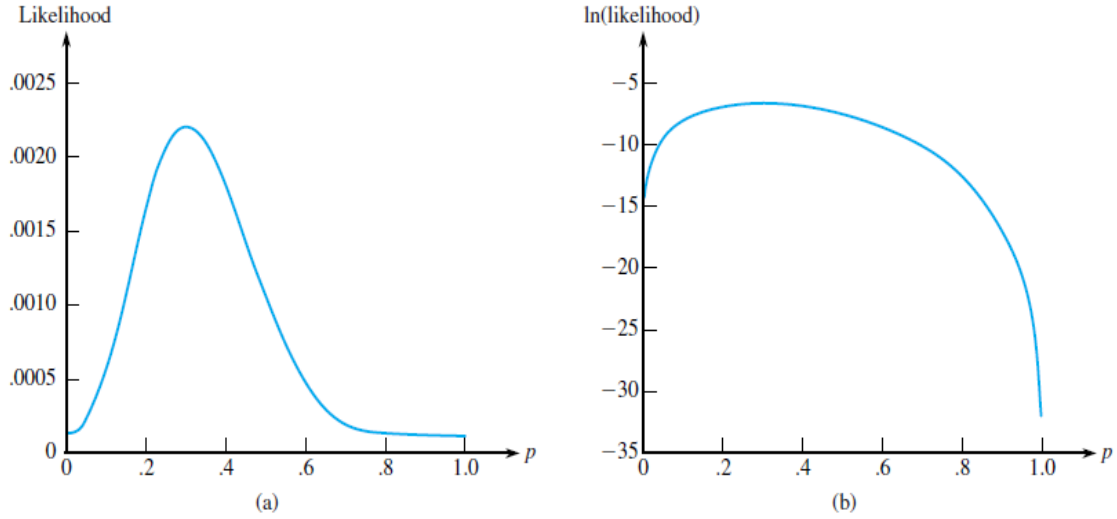

Figura 1: Gráfico da verossimilhança e log-verossimilhança contra p, adaptado de (DEVORE; BERK, 2007).

De acordo com o os gráficos da Figura 1, verifica-se que seus objetivos são encontrar o valor de *p* que maximiza a função de verossimilhança (20) e (21) respectivamente. Neste exemplo, quando *p*=3, ambas as funções conseguem obter o seu máximo.

Se X é uma vetor aleatório, então $L(\theta|x) = P_\theta(X=x)$. Se compararmos a função de verossimilhança em dois pontos do parâmetro e descobrirmos que $P_{\theta 1}(X=x) = L(\theta 1|x) > L(\theta 2|x) = P_{\theta 2}(X=x)$, então é mais provável que a amostra que realmente observamos tenha ocorrido se $\theta = \theta_1$, do que se $\theta = \theta_2$, o que pode ser interpretado como dizendo que θ_1 é um valor mais plausível para o valor verdadeiro de θ do que θ_2 (CASELLA; BERGER, 2010). Generalizando, o vetor de parâmetro mais plausível é aquele de maior verossimilhança (CORDEIRO, 1999).

De acordo com Cordeiro (1999), o foco da função de verossimilhança é retirar dos dados às informações necessárias para fazer inferências sobre um vetor de parâmetro de interesse. Para um melhor esclarecimento deste método, o Exemplo 1, aplica-o em uma distribuição multinomial.

Exemplo 1: Suponha que têm-se Y= (48, 24,115) e P= ($\frac{1-\theta}{2}$ ·, $\frac{\theta}{4}$ ·, $\frac{2+\theta}{4}$), e que deseja-se estimar o parâmetro θ a partir de uma distribuição multinomial, então:

$$P(y) = p(Y_1=y_1, Y_2=y_2, Y_3=y_3) = \frac{n!}{y_1!y_2!y_3!} * P_1^{y_1} P_2^{y_2} P_3^{y_3}$$

Onde:

$n = y_1+y_2+y_3$

$P_1 = (\frac{1-\theta}{2})$; $P_2 = (\frac{\theta}{4})$; $P_3 = (\frac{2+\theta}{4})$

Aplica-se o princípio da Estimativa de Máxima Verossimilhança.

$$\text{Log } P(y) = \log\left(\frac{n!}{y_1!y_2!y_3!}\right) + \log\left(P_1^{y_1} P_2^{y_2} P_3^{y_3}\right)$$

$$\text{Log } P(y) = \log n! - \log y_1! - \log y_2! - \log y_3! + y_1 * \log P_1 + y_2 * \log P_2 + y_3 * \log P_3$$

Agora basta aplicar a derivada parcial em relação ao parâmetro θ, onde há a necessidade apenas da segunda parte do log.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \text{Log } P(y) = y_1 * \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\frac{1-\theta}{2}\right) + y_2 * \frac{\partial}{\partial \boldsymbol{\theta}} \log \frac{\theta}{4} + y_3 * \frac{\partial}{\partial \boldsymbol{\theta}} \log \frac{2+\theta}{4} = 0$$

De acordo com as propriedades da derivada do logaritmo, pode-se reescrever a equação como:

$$y_1 * \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\frac{1-\theta}{2}\right) = y_1 * \left(\frac{\frac{1}{2}}{\frac{1-\theta}{2}}\right) * \frac{d}{d\theta}(1-\theta) = \frac{2*y_1*(-1)}{2(1-\theta)} = \mathbf{-\frac{y_1}{(1-\theta)}}$$

$$y_2 * \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\frac{\theta}{4}\right) = y_2 * \left(\frac{\frac{1}{4}}{\frac{\theta}{4}}\right) * \frac{d}{d\theta}(\theta) = \frac{4*y_2*(1)}{4*\theta} = \mathbf{\frac{y_2}{\theta}}$$

$$y_3 * \frac{\partial}{\partial \boldsymbol{\theta}} \log\left(\frac{2+\theta}{4}\right) = y_3 * \left(\frac{\frac{1}{4}}{\frac{2+\theta}{4}}\right) * \frac{d}{d\theta}(2+\theta) = \frac{4*y_3*(1)}{4*(2+\theta)} = \mathbf{\frac{y_3}{2+\theta}}$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} \text{Log } P(y) = \mathbf{-\frac{y_1}{(1-\theta)} + \frac{y_2}{\theta} + \frac{y_3}{2+\theta} = 0}$$

$$\mathbf{-\frac{y_1}{(1-\theta)} + \frac{y_2}{\theta} + \frac{y_3}{2+\theta} = 0}$$

Tira-se o mínimo múltiplo comum.

$$-y_1\ (2\theta + \theta^2) + y_2\ (2\ - 2\theta) + y_3\ (\theta - \theta^2) = 0$$
$$(-y_1 - y_3)\theta^2 + \ (-2y_1 - 2y_2 + y_3)\theta + 2y_2 = 0$$

$$\Delta = b^2 - 4ac$$

$$\Delta = (-2y_1 - 2y_2 + y_3)^2 - 4(-y_1 - y_3) * 2y_2$$
$$\Delta = (-2*48 - 2*24 + 115)^2 - 4(-48 - 115)*2*24$$

$$\Delta = 32137$$

$$\hat{\theta}_{1,2} = \frac{-b \pm \sqrt{\Delta}}{2a} = \frac{-(-2y_1 - 2y_2 + y_3) \pm \sqrt{\Delta}}{2(-y_1 - y_3)}$$

$$\hat{\theta}_{1,2} = \frac{-b \pm \sqrt{\Delta}}{2a} = \frac{-(-2*48 - 2*24 + 115) \pm \sqrt{32137}}{2(-48 - 115)}$$

$$\hat{\theta}_1 \Theta_1 = -0{,}63885876$$
$$\hat{\theta}_2 \Theta_2 = 0{,}460944649$$

Descarta-se o parâmetro negativo e teremos a estimativa para a ML $\hat{\theta}$ =0,460944649.

Quanto ao uso do método de máxima verossimilhança para facilitar o tratamento de dados faltantes, geralmente assume-se que os dados observados são modelos amostrais a partir da distribuição normal multivariada (LAKSHMINARAYAN; HARP; SAMADI, 1999), sendo assim, a partir deste pressuposto, o método de máxima verossimilhança concentrar-se na estimativa dos parâmetros dos dados observados, ou seja, o vetor de média e matriz de variância-covariância (PIGOTT, 2001), o que torna este método um concorrente próximo à imputação múltipla, já que diante de pressupostos idênticos, ambos os métodos são capazes de produzir estimativas que são consistente, assintoticamente eficiente e assintoticamente normal (ALISSON, 2012).

Todos estes passos, estudados nesta seção, serão necessários para o entendimento e aplicação de uma das técnicas mais importante para encontrar estimativa de máxima verossimilhança, quando há dados faltantes na amostra, que é o algoritmo EM (REID; COX, 2013), o qual é apresentado a seguir.

3.3.3.3.1 Algoritmo *Expectation Maximization*

Antes de se iniciar qualquer processo de análise de dados, deve-se conhecer primeiro sua estrutura, se há ou não ruído, principalmente dados ausentes. Já que, esta informação ausente poderá conter informações relevantes para as inferências, e caso não as trate de forma adequada, aumenta-se as chances de se ter inferências enviesadas.

O algoritmo *Expectation Maximization* (EM) aparece em todos os contextos estatísticos, sendo aplicados em uma variedade de situações, tais como aquelas onde há dados faltantes, dados latentes, dados censurados ou agrupados, e também onde a incompletude dos dados não é natural ou evidente (MCLACHLAN; KRISHNAN, 2008). Este algoritmo também usa técnicas estatísticas para maximizar verossimilhanças complexas, cujo objetivo é calcular a Máxima Verossimilhança a partir de dados incompletos (MLADENOVIC; PORRAT; LUTOVAC, 2011). O termo "dados incompletos" em sua forma geral implica na existência de dois espaços amostrais Y e X e um mapeamento de muitos para um de X à Y (DEMPSTER; LAIRD; RUBIN, 1977).

Ele é aplicado principalmente em duas situações, quais sejam, quando há valores faltantes e quando a função de verossimilhança é difícil de ser obtida por outros métodos analíticos, sendo esta última muito utilizada na área de reconhecimento de padrões (BILMES, 1998), sendo que neste caso, quando é aplicado na área de reconhecimento de padrões, ele é muito útil para o cálculo iterativo da estimativa de máxima verossimilhança (EMV) (MCLACHLAN; KRISHNAN, 2008).

Este algoritmo é utilizado iterativamente para maximizar os parâmetros de um modelo quando há a presença de dados ausentes. Por exemplo, em processo de modelagem no âmbito de classificação poderá surgir dados omissos, tanto na fase de treinamento como na fase de classificação, tornando-o necessário nestes casos, pois caso o profissional não utilize de tal método, geralmente recorre a uma abordagem ingênua que é excluir do *dataset* as amostras que contém dados ausentes (DUDA; HART; STORK, 2001).

Quando as amostras faltantes são oriundas de uma família exponencial, as estimativas de máxima verossimilhança são mais fáceis de serem calculadas pelo algoritmo EM, pois ele é mais robusto, devido ao seu processo iterativo ser baseado no método dos mínimos quadrados reponderados (DEMPSTER; LAIRD; RUBIN, 1977), e também porque o log da verossimilhança será linear nos dados que faltam (CASELLA; BERGER, 2010).

O EM tem a vantagem de assegurar a obtenção de uma convergência, de exigir pouco espaço de memória, baixo custo por iteração e facilidade de ser programado, porém apresenta

a desvantagem de ser lento para convergir em algumas situações práticas (REDNER; WALKER, 1984; WU, 1983), sendo que o problema mais grave atrelado a este algoritmo é o problema de máximos locais, pois tal problema torna o desempenho altamente dependente de um valor inicial do parâmetro (UEDA; NAKANO, 1998). Este algoritmo pode ficar preso em máximos locais ou em um ponto de sela (PAULA, 2013), sendo que ele converge lentamente especialmente em dados incompletos de alta dimensão (OSOBA, 2013), e também é sensível diante da presença de dados discrepantes (*outliers*), além disso, ele pode não responder bem nos casos onde há muitos dados omissos.

O algoritmo EM é uma generalização do de máxima verossimilhança, pois ele herda muitas propriedades do método de EMV. O passo E descreve o melhor modelo de dados completo possível para os dados incompletos, dada todas as informações atuais. O passo M usa esse novo modelo completo para escolher as mais altas estimativas de verossimilhança dos parâmetros da distribuição para dados incompletos. A melhoria das estimativas dos parâmetros a partir do passo M conduz a um melhor modelo completo no passo E (OSOBA, 2013). Isso ocorre de forma iterativa até que haja a convergência. Assim, o objetivo do algoritmo EM é associar um problema onde há dados incompletos a um problema de dados completos, a fim de facilitar as estimativas de máxima verossimilhança (PAULA, 2013).

Na Figura 2, tem-se uma ilustração do mapeamento de muitos para um. Onde o espaço amostral X corresponde aos dados completos e o espaço amostral Y corresponde aos dados incompletos.

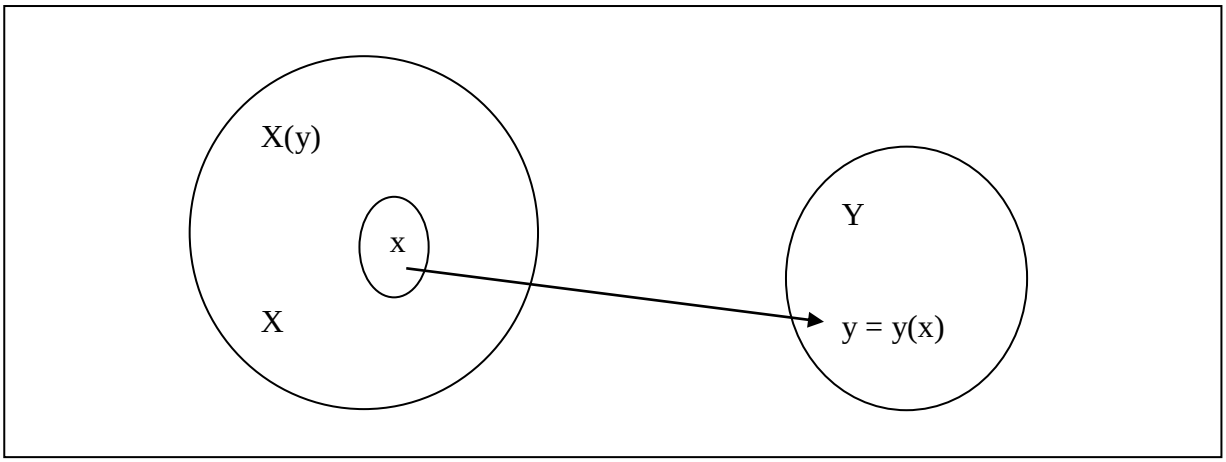

Figura 2: Ilustração do mapeamento de muitos para um de X à Y. O ponto y é a imagem de x e o conjunto X(y) é o mapeamento inverso de y. Adaptado de Moon (1996).

Seguindo a abordagem de (HOGG; MCKEAN; CRAIG, 2012; RAMACHANDRAN; TSOKOS, 2009, CASELLA; BERGER, 2010), para demonstrar o passo a passo deste algoritmo, tem-se:

Supondo-se que em uma amostra de *n* itens, na qual n_1 representa os itens observados e $n_2 = n - n_1$ representa os itens não observados, pode-se demonstrar matematicamente o algoritmo EM, assumindo-se que os dados observados são representados por $X=(x_1, x_2, ..., x_n)$, e os dados não observados são representados por $Y=(y_1, y_2, ..., y_n)$. Assume-se também que os x_i são i.i.d. com fdp conjunta dada por $f(x/\theta)$, onde θ é o vetor de parâmetros com valores em Θ c R^p espaço euclidiano p-dimensional, e que os y_i e os x_i são mutuamente independentes.

$g(x, y/\theta)\rightarrow$ representa a fdp conjunta dos valores observados e dos não observados.

$h(y/\theta, x)\rightarrow$ representando a fdp condicional dos valores ausentes y dado θ e os valores observados x.

$L(\theta/x) = f(x/\theta)\rightarrow$ representa a função de verossimilhança dos dados observados x.

$Lc\ (\theta/x, y) = g(x, y/\theta)\rightarrow$ representa a função de verossimilhança dos dados completos (x, y).

Pela definição de uma fdp condicional, chega-se a seguinte identidade.

$$g(x, y/\theta) = f(x/\theta) * h(y/\theta, x) \tag{24}$$

Ou

$$f(x/\theta) = g(x, y/\theta) \ / \ h(y/\theta, x) \tag{25}$$

O objetivo é maximizar a função de verossimilhança $L(\theta; x)$ pelo uso da verossimilhança completa $Lc\ (\theta/x, z)$. Como $h(y/\theta_0,x)$ é uma fdp, então por definição temos:

$$\int h(y/\theta_0,x)dy = 1 \tag{26}$$

Aplicando agora a EMV (Estimativa de Máxima Verossimilhança) na equação (24) chega-se a seguinte identidade básica para um arbitrário, mas fixo $\theta_0 \in \Theta$.

$$\log g(x,y/\theta) = \log [f(x/\theta) * h(y/\theta,x)] \tag{27}$$

$$\log g(x, y/\theta) = \log f(x/\theta) + \log h(y/\theta, x). \tag{28}$$

$$\log L_c(\theta/x, y) = \log L(\theta/x) + \log h(y/\theta, x). \quad (29)$$

$$\log L(\theta/x) = \log L_c(\theta/x,y) - \log h(y/\theta,x) \quad (30)$$

Tomando a equação (30), têm-se as condições suficientes para aplicar o passo E do algoritmo EM. Para tanto pode-se tomar apenas o primeiro membro (lado esquerdo desta equação), pois conforme explanado em (LITTLE; RUBIN, 1987), pelos pressupostos de ignorabilidade, este não depende de y, então:

$$\log L(\theta/x) = \int \log L(\theta/x) * h(y/\theta,x)\, dy \quad (31)$$

$$\log L(\theta/x) = \int \log f(x/\theta) * h(y/\theta,x)\, dy \quad (32)$$

$$\log L(\theta/x) = \int \log [g(x,y/\theta)/ h(y/\theta,x)] * h(y/\theta,x)\, dy \quad (33)$$

$$\log L(\theta/x) = \int \log [g(x,y/\theta) - h(y/\theta,x)] * h(y/\theta,x)\, dy \quad (34)$$

$$\log L(\theta/x) = \int \log [g(x,y/\theta)] * h(y/\theta,x)\, dy - \int \log [h(y/\theta,x)] * h(y/\theta,x)\, dy \quad (35)$$

$$\log L(\theta/x) = E_{\theta 0}[\log L_c(\theta/x,y)/\theta_0,x] - E_{\theta 0}[\log h(y/\theta,x)/\theta_0,x] \quad (36)$$

O primeiro e o segundo termo do lado direito da equação (36), podem ser reescritos como:

$$Q(\theta/\theta_0,x) = E_{\theta 0}[\log L_c(\theta/x,y)/\theta_0,x] \quad (37)$$

$$H(\theta/\theta_0,x) = E_{\theta 0}[\log h(y/\theta,x)/\theta_0,x] \quad (38)$$

A única equação que interessa para a implementação do algoritmo é a (37), pois a equação (38), de acordo com a desigualdade de Jensen resultará em zero.

Aqui a esperança é obtida com relação à distribuição condicional de y dado θ_0 e x. Vamos agora considerar a maximização deste passo E com relação ao parâmetro θ. Este é o passo M da maximização no algoritmo EM. Temos agora que θ_0 é uma estimativa inicial do θ. Não há regra para se escolher uma valor inicial para o θ_0, o qual pode ser escolhido de

maneira aleatória, ou pode-se, caso se tenha algum conhecimento prévio dos dados, iniciar com a média ou variância dos dados observados. Tem-se:

$$Q(\theta\ /\theta_0,x) = E\ \theta_0[L_c(\theta;x,y)] \quad (39)$$

$$Q(\theta\ /\theta_0,x) = E\ \theta_0[\ln g(x,y/\theta)] \quad (40)$$

Neste passo, θ_0 é usado apenas para calcular a esperança, sendo assim, não se deve substituir θ no log-verossimilhança dos dados completo.

A partir de um valor inicial θ_0 gera-se uma sequência θ(r) conforme o seguinte passo:

Θ(r + 1) = é o valor que maximiza $E\theta_0$ [ln g(x,y/θ)]

Sintetizando o passo E calcula a esperança do log de verossimilhança, e o passo M calcula o máximo do passo E. Sendo assim, o passo E e o passo M, devem ser repetidos até que haja uma convergência para um ponto estacionário, que segundo Wu (1983) pode ser um máximo local ou um ponto de sela. Como critério de parada pode-se adotar $\| \theta(r + 1) – \theta r)\| < \alpha$, sendo α um valor que tem que ser determinado antes de se iniciar a iteração, e ele tem que ser maior que zero. Para uma melhor compreensão destes passos, na Figura 3, tem-se um esquema demonstrativo da iteração deste algoritmo.

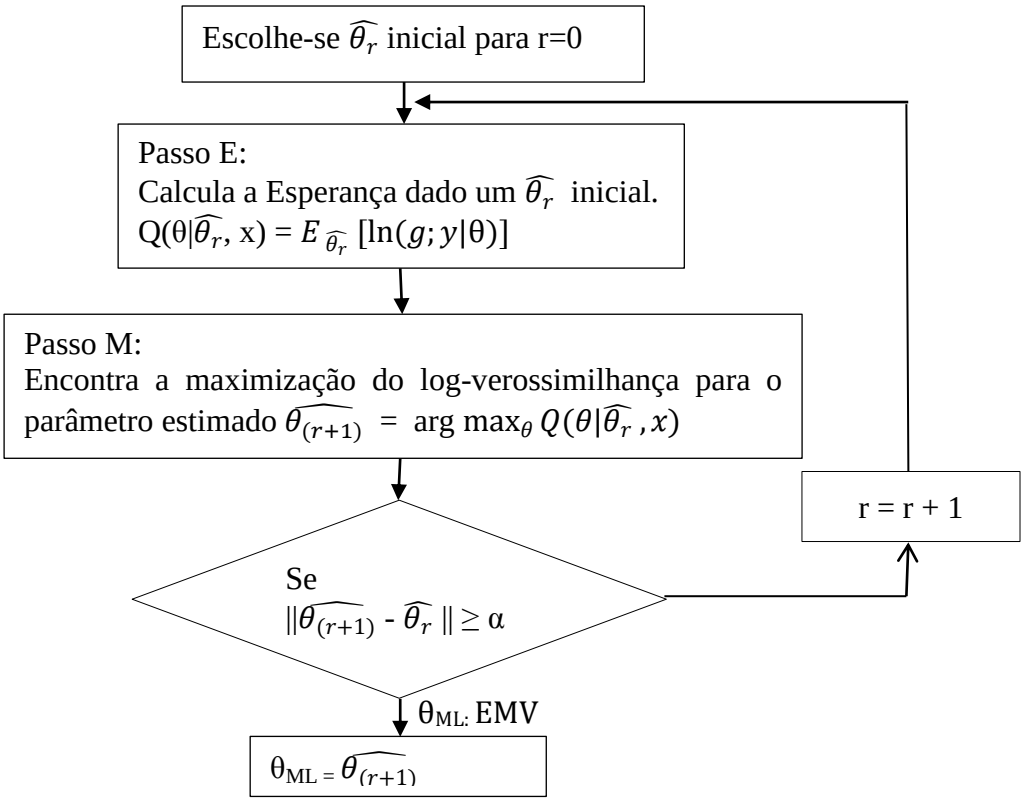

Figura 3: Fluxograma do algoritmo EM.

Semelhante à estimativa de máxima verossimilhança, o algoritmo EM, procura estimar os parâmetros da distribuição conjunta dos dados, tais como o vetor de média e matriz de covariância, resultando em estimativas pontuais destes vetores (PIGOTT, 2001). Ambos EM e MI contam com a suposição de normalidade, e, portanto, tendem a ser sensíveis a *outliers*, em consequência disso, eles podem preencher valores anormais nas variáveis ausentes, além de tenderem a mover potenciais observações atípicas para o centro dos dados (NG-CHI, 1998), principalmente quando há grande quantidade de dados faltantes.

Para uma melhor compreensão do funcionamento deste algoritmo, no Exemplo 2, aplica-o a uma distribuição multinomial.

Exemplo 2: Suponha que tenhamos 197 animais, que são distribuídos multinomialmente em 4 categorias (RUBIN, 1987).

$Y=(125, 18, 20, 34) = (Y_1, Y_2, Y_3, Y_4)$

Um modelo genético para uma população específica tem as seguintes probabilidades, $(1/2+p/4, 1/4-p/4, 1/4-p/4, p/4)$

Representando Y os dados incompletos, $Y_1= x_1 + x_2$, onde $(x_1 + x_2)$ desconhecidos, $Y_2= x_3$, $Y_3= x_4$, $Y_4= x_5$.
Assim, temos:

$P(Y_1) = 1/2+p/4 \rightarrow P(x_1)= 1/2$ e $P(x_2)= p/4$
$P(Y_2) = 1/4-p/4 \rightarrow P(x_3)= 1/4-p/4$
$P(Y_3) = 1/4-p/4 \rightarrow P(x_4)= 1/4-p/4$
$P(Y_4) = p/4 \rightarrow P(x_5)= p/4$

$$P(y) = p(Y_1=y_1, Y_2=y_2, Y_3=y_3, Y_4=y_4)= \frac{n!}{y_1!y_2!y_3!y_{4!}} * P_1^{y_1} P_2^{y_2} P_3^{y_3} P_4^{y_4}$$

OBS: Sabe-se que:

$$E(x_1) = n * p(x_1)$$

Agora, basta normalizar os dados em termos probabilísticos.

$$E(x_1) = n * \frac{p(x_1)}{p(x_1) + p(x_2)} = 125 * \frac{\frac{1}{2}}{\frac{1}{2} + \frac{p}{4}} \equiv x_1^{(k)}$$

$$E(x_2) = n * p(x_2)$$

$$E(x_2) = n * \frac{p(x_2)}{p(x_1) + p(x_2)} = 125 * \frac{\frac{p}{4}}{\frac{1}{2} + \frac{p}{4}} \equiv x_2^{(k)}$$

A função de densidade pode ser expandida para:

$$f(x;p) = \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!\,x_5!} * {p_1}^{x_1} * {p_2}^{x_2} * {p_3}^{x_3} * {p_4}^{x_4} * {p_5}^{x_5}$$

$$f(x;p) = \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!\,x_5!} * \left(\frac{1}{2}\right)^{x_1} * \left(\frac{p}{4}\right)^{x_2} * \left(\frac{1}{4}-\frac{p}{4}\right)^{x_3} * \left(\frac{1}{4}-\frac{p}{4}\right)^{x_4} * \left(\frac{p}{4}\right)^{x_5}$$

$$f(x;p) = \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!\,x_5!} * \left(\frac{1}{2}\right)^{x_1} * \left(\frac{p}{4}\right)^{x_2+x_5} * \left(\frac{1}{4}-\frac{p}{4}\right)^{x_3+x_4}$$

$$c(x) = \frac{n!}{x_1!\,x_2!\,x_3!\,x_4!\,x_5!}$$

$$ln\, f(x;p) = ln\, c(x)\ + x_1 \ln\left(\frac{1}{2}\right) + \ (x_2 + x_5)\, ln\left(\frac{p}{4}\right) + \ \ (x_3 + x_4) ln\left(\frac{1}{4}-\frac{p}{4}\right)$$

Passo E

$$E[ln\, f(x;p) = E[ln\, c(x)\,] + E[x_1 \ln\left(\frac{1}{2}\right)] + \ E[(x_2 + x_5)\, ln\left(\frac{p}{4}\right)] + \ \ E[(x_3 + x_4) ln\left(\frac{1}{4}-\frac{p}{4}\right)]$$

$$E[ln\, f(x;p)] = ln\, c(x) + x_1^{(k)} \ln\left(\frac{1}{2}\right) + \ \left(x_2^{(k)} + x_5\right)\, E[ln\left(\frac{p}{4}\right)] + \ \ (x_3 + x_4) E[ln\left(\frac{1}{4}-\frac{p}{4}\right)]$$

Passo M

$$\frac{\partial}{\partial p} E[ln\, f(x;p)] = ln\, c(x) + x_1^{(k)} \ln\left(\frac{1}{2}\right) + \ \left(x_2^{(k)} + x_5\right)\, E[ln\left(\frac{p}{4}\right)] + \ \ (x_3 + x_4) E[ln\left(\frac{1}{4}-\frac{p}{4}\right)]$$

Antes de aplicar a derivada, para facilitar os cálculos, multiplica-se tudo por 4.

$$\frac{\partial}{\partial p} E[ln\, f(x;p)] = \ \left(x_2^{(k)} + x_5\right) \left(\frac{1}{p}\right) + \ \ (x_3 + x_4) \left(-\frac{1}{1-p}\right) = \ 0$$

$$\left(x_2^{(k)} + x_5\right)(1-p) = (x_3 + x_4)(p)$$

$$x_2^{(k)} + x_5 - x_2^{(k)} * p + x_5 * p = x_3 * p + x_4 * p$$

$$\boldsymbol{p^{(k+1)}} = \frac{x_2^{(k)} + x_5}{x_2^{(k)} + x_5 + x_3 + x_4}$$

Agora para se encontrar os valores de $x_2^{(k)}$, deve-se inicializar o valor de $p^{(k)}$, por uma valor aleatório, e em seguida ir atualizando os valores $x_1^{(k)}$, $x_2^{(k)}$ e $p^{(k+1)}$, até que haja a convergência.

3.3.4 Métodos de Aprendizado de Máquina

Alguns métodos de aprendizado de máquina também têm sido propostos na literatura para tratar os casos de omissão. Dentre eles destacam-se o *Autoclass* e *C4.5*.

Autoclass é um método de agrupamento usado para revelar a estrutura intrínseca nos dados, enquanto *C4.5* é um algoritmo para classificação de aprendizagem de árvore de decisão e baseia-se na teoria de classificação Bayesiana, que poderia ser utilizada para prever diferentes atributos após uma simples sessão de aprendizagem. Isso faz com que seu uso seja econômico em termos de tempo. Uma característica interessante do *Autoclass* é que ele procura por classes automaticamente, e tem limites que impedem dados de *over-fitting* (que é a memorização dos padrões, que tem como consequência um erro quadrático baixo na fase de treinamento, porém um erro quadrático alto na fase de teste) (LAKSHMINARAYAN; HARP; SAMADI, 1999).

Outras técnicas também têm sido abordadas, que são as Redes Neurais MLP, *Weighted Imputation with K-Nearest Neighbor-WKNNI, K-means Clustering Imputation-KMI, Support Vector Machines Imputation-SVMI, Singular Value Decomposition Imputation-SVDI, K2, Data Augmentation (DA), BN-K2Iχ², 1BN-K2Iχ²,* algoritmo de *biclusterização SwarmBcluster*, para melhor detalhe consultar Luengo, García & Herrera (2012), Hruschka Jr. & Ebecken (2002), Hruschka Jr. (2007) e Veroneze (2011). A próxima seção faz uma breve comparação entre EMV e MI.

3.4 Diferenças entre Métodos Baseados em EMV e MI

Alisson (2012) fez uma comparação entre as técnicas de imputação de dados baseada em máxima verossimilhança e as baseada em imputação múltipla, onde o mesmo afirma que a abordagem de máxima verossimilhança é mais promissora por ser mais assertiva. Abaixo está uma pequena explanação desta diferença.

a) Estimativas de máxima verossimilhança são mais eficientes que imputação múltipla, pois, apesar de ambas serem assintoticamente eficiente, o que implica que elas têm variância mínima amostral, a imputação múltipla é quase eficiente, pois para a MI atingir a eficiência, haveria a necessidade de analisar um número infinito de conjuntos de dados.

b) Para um conjunto de dados, EMV sempre produzirá o mesmo resultado, porém a MI dará resultados diferentes toda vez que for usado, visto que MI envolve modelos aleatórios, logo existe uma indeterminação inerente nos resultados, o que conduz a diferentes

investigadores, aplicando os mesmos métodos para os mesmos dados, chegarem a conclusões diferentes.

c) A implementação de MI necessita de muitas decisões diferentes, cada qual envolvendo incertezas. EMV necessita de menos decisões.

d) Com MI, há sempre um potencial conflito entre o modelo de imputação e o modelo de análise. No entanto, não há conflito potencial na ML porque tudo é feito num mesmo modelo.

CAPITULO 4

4 MODELO BASEDO NO FUNCIONAMENTO DO CÉREBRO

Neste capítulo será apresentada a fundamentação teórica do modelo de Redes Neurais Artificias *Multilayer Perceptron*, mais conhecido com Redes Neurais MLP, o qual tem seu modelo de processamento inspirado no cérebro humano. Desta forma, se faz necessário explanar como funciona o cérebro humano.

4.1 Como o Cérebro Funciona

O cérebro humano é a estrutura mais complexa que se conhece, e entender suas operações representa um dos desafios mais difíceis e importantes enfrentados pela ciência (BISHOP, 1994). Ele pode ser considerado como um computador notável, pois interpreta informação imprecisa a partir dos sentidos em uma taxa incrivelmente rápida (HINTON, 1992). Ele é composto de vários neurônios interconectados, e é capaz de regular continuamente a sensibilidade destes. De acordo com Lente (2011), *os neurônios são células excitáveis, que produzem sinais elétricos que codificam informações provenientes de outros neurônios ou provenientes do ambiente, para isso cada neurônio tem sua excitabilidade regulada continuamente, já que nos circuitos cerebrais cada neurônio recebe milhares de sinapses de outros neurônios.*

Todo o processamento de informação realizado no cérebro humano ocorre de forma paralela, entretanto não são todas as operações cerebrais que são processadas de forma lógica, devido à capacidade humana de criar representações físicas que podem de forma simples obter respostas para problemas muito difíceis e abstratos, destacando-se para tanto três habilidades que nos permitem realizar essas conquistas que são exclusivamente humanas de raciocínio formal, quais sejam: habilidade em associar padrões, facilidade em modelar o nosso mundo e sermos bons em manipular nosso ambiente (MCCLELLAND *et al*., 1986). Além disso, o processamento das informações ocorre de forma não linear.

A busca por criar um modelo matemático que possa assemelhar-se ao cérebro humano, impulsionou e ainda impulsiona o interesse de pesquisados em várias partes do mundo. Sendo que os trabalhos seminais foram o de McCulloch & Pitts (1943) que observaram que o caráter "tudo ou nada" da atividade do sistema nervoso, nos eventos neurais e as relações entre eles podiam ser tratada por meio de lógica proposicional, e o trabalho de Widrow & Hoff (1960),

no qual constataram que dentro do neurônio, existe a formação de uma combinação linear dos sinais de entrada. A partir destes trabalhos surgiram vários outros, que possibilitaram a popularização da técnica de Redes Neurais Artificiais. Entretanto como elencado por Minsky (1982) diversas dificuldades e questionamentos ainda permanecem em aberto até o hoje, quanto ao poder de um computador assemelhar-se ao cérebro humano, tais como: poderão os computadores serem criativos, escolherem seus próprios problemas, ter emoção, poderia um computador saber o que significa algo, poderia um computador ser consciente, ter sensibilidade, ter a capacidade de interpretar metáforas (que é a capacidade de relacionar conceitos aparentemente não relacionados, buscando o que tem em comum, isso é uma característica humana).

4.2 Modelo Fisiológico de um Neurônio

O neurônio clássico tem vários dendritos, que são as entradas das informações, geralmente ramificados, que recebem informações de outros neurônios e um único axônio que transfere a informação processada normalmente pela propagação de um "pico" ou um "potencial de ativação". O axônio se ramifica em vários outros ramos que fazem sinapses para os dendritos e corpos celulares de outros neurônios (MCCLELLAND *et al.*, 1986). Na Figura 4, tem-se um modelo de um neurônio.

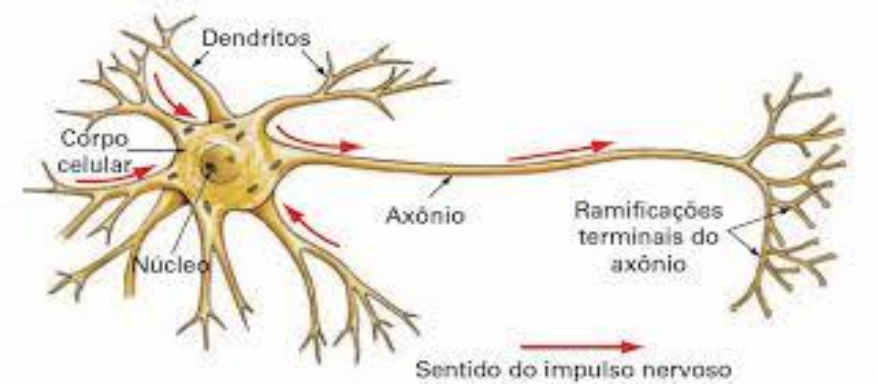

Figura 4: Estrutura fisiológica de um neurônio.

O corpo celular, que também é conhecido como soma, é onde ocorre o processamento da informação, mais precisamente no núcleo, a qual é posteriormente transmitida através do axônio para novas ramificações. As próximas seções irão tratar como modelar matematicamente as RNA MLP.

Os modelos de Redes Neurais tentam imitar a maneira como cérebro humano processa as informações, sendo assim, Haykin (1999) define redes neurais e sua semelhança com o cérebro humano como sendo:

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 uso. Ela se assemelha ao cérebro humano em dois aspectos:

1. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem.

2. Forças de conexões entre neurônios, conhecidas como pesos sinápticos, são utilizadas para armazenar o conhecimento adquirido.

Em um nível prático o cérebro humano tem muitas características que são desejáveis em um computador eletrônico, tais como a capacidade de generalização a partir de ideias abstratas, reconhecer padrões na presença de ruído, recordar rapidamente memórias, e resistir a danos localizados (WARNER; MISRA, 1996). Seguindo esta linha de raciocínio, pode-se citar como umas das principais características das redes neurais artificiais: adaptação por experiência, capacidade de aprendizado, não linearidade, mapeamento entra-saída, habilidade de generalização, organização de dados, resposta a evidências, tolerância a falhas, armazenamento distribuído, facilidade de prototipagem, informação contextual, uniformidade de análise e projeto, analogia neurobiológica.

Apesar dos computadores serem extremamente rápidos em fazer cálculos numéricos, ultrapassando inclusive a capacidade humana, eles ainda não superam muitas das capacidades humanas, que seria desejável em um computador, fato este que é a motivação fundamental para tentar compreender e modelar o cérebro humano de forma matemática (WARNER; MISRA, 1996).

Os métodos de aprendizagem de Redes Neurais Artificiais podem ser divididas em três amplas classes: procedimento supervisionado (classificação, regressão), procedimento por reforço e não supervisionado (HINTO; FREY, 1995). Além disso, há três aspectos muito importantes que caracterizam uma RNA MLP, que são (GOMES; LUDEMIR; LIMA, 2011):

1) O padrão de ligações entre os neurônios (arquitetura),

2) O método de atualização dos pesos das conexões (algoritmo de aprendizado),

3) A função de ativação.

Cada neurônio dentro de uma RNA é representado por uma função de ativação, que recebe várias entradas e apresenta como resposta uma única saída, conforme a equação abaixo (SILVA; SPATTI; FLAUZINO, 2010):

$$Y_i = f_i\left(\sum_{j=1}^{n} w_{ij} x_j + \theta_i\right) \quad (41)$$

Onde*:*

f_i = Função de Ativação

w_{ij} = Peso Sináptico

x_j = Amostra de Treinamento

θ_i= Viés

4.3 Algoritmo *Backpropagation*

Em 1971 Werbos desenvolveu o *backpropagation*, que foi publicado em sua tese de doutorado em 1974, porém ele permaneceu desconhecido da comunidade científica, e que algum tempo depois foi novamente estudado, principalmente por Rumelhart, Hinton & Williams que conseguiram aplicá-lo para solucionar uma variedade de problemas, tornando-o mais conhecido na comunidade científica (WIDROW; LEHR, 1990).

O algoritmo *Backpropagation* é simplesmente um método eficiente e exato para calcular as derivadas da função de erro com relação aos pesos wij (BISHOP, 1991). Seu principal objetivo é encontrar um conjunto de pesos que permita o sistema minimizar o erro a 0 (zero) para cada unidade em cada padrão, através da exposição repetida de todas as amostras do conjunto de padrões. No entanto, é importante observar que a existência de um conjunto de pesos que permitirá o erro ser reduzido a zero não é garantido (MCCLELLAND *et al*., 1986).

No *backpropagation* escolhe-se os pesos wij, de modo a minimizar o erro quadrado sobre o conjunto de treinamento: isto é simplesmente um caso especial do método dos mínimos quadrados, usado muitas vezes em estatística, econometria, e engenharia (WERBOS, 1990). Note-se que no *backpropagation* os erros são propagados para trás através da rede (BISHOP, 1991), e que os pesos são extremamente importantes, pois são eles que determinam o comportamento de cada neurônio, já que eles determinam a intensidade da conexão entre os neurônios. De acordo com Warner & Misra (1996) o cérebro aprende adaptando a força das conexões sinápticas, do mesmo modo ocorre com os pesos sinápticos em redes neurais, que são ajustados para solucionar o problema apresentado para a rede.

No *backpropagation* os pesos são gerados inicialmente de forma aleatória, principalmente no intervalo entre -0.5 e 0.5, por alguma distribuição de probabilidade, que normalmente é a distribuição uniforme. Porém Werbos (1990) afirma que o ideal é que os pesos sejam gerados a partir de alguma informação *a priori*, principalmente nos casos onde esta informação prévia está disponível. Quanto a sua topologia, esta fica a critério do usuário, determinar a quantidade de camadas e neurônio em cada uma respectivamente.

Neste algoritmo as conexões entre os neurônios são unidirecionais (WARNER; MISRA, 1996). Frise-se que o comportamento de uma Rede Neural Artificial depende fortemente dos pesos e da função *input-output* que é especificado para cada neurônio. Estas funções normalmente caem em uma das três categorias: linear, limiar ou sigmoide (HINTON, 1992). Escolhas adequadas dos valores dos pesos, em camadas escondidas, pode conduzir a redução de problemas, tal como o algoritmo ficar preso a um ponto de mínimo local, e, além disso, pesos adequados aumentam significativamente a velocidade do treinamento da rede (WIDROW; LEHR, 1990).

O *backpropagation* pode ser visto como uma generalização da regressão logística à Redes Neurais Artificiais *feedforward* que tem camadas de unidades escondidas entre as unidades de *input* e *output* (HINTO; FREY, 1995). Ele requer mais tempo para aprender à medida que a rede se torna maior, bem como a amostra, entretanto quando este é usado como modelo de aprendizado real ele necessita de um "professor" para fornecer a saída desejada para cada exemplo de treinamento, em contraste, as pessoas aprendem a maioria das coisas sem a ajuda de um professor (HINTON, 1992).

Em relação ao processo de aprendizado da rede, e sua respectiva atualização dos pesos, tem-se o seguinte modelo matemático:

$$w_{ij}(t+1) = w_{ij}(t) - n * \frac{\partial E(t)}{\partial w_{ij}(t)} \tag{42}$$

Onde:

n= Taxa de Aprendizagem

E(t)= Valor do Erro

w_{ij} = Peso Sináptico

É na fase de atualização dos pesos que o algoritmo *backpropagation* se destaca, apresentando uma abordagem simples, pois para atualizar os pesos (w_{ij}) basta aplicar a regra da cadeia, e ir derivando o erro E(t) em relação a cada peso, conforme o modelo abaixo:

$$\partial E = \frac{\partial E}{\partial w_{ij}^{(n)}} = \frac{\partial E}{\partial Y_j^{(n)}} \cdot \frac{\partial Y_j^{(n)}}{\partial I_j^{(n)}} \cdot \frac{\partial I_j^{(n)}}{\partial w_{ij}^{(n)}} \quad (43)$$

Onde:

$$\frac{\partial I_j^{(n)}}{\partial w_{ij}^{(n)}} = Y_j^{(n-1)} \quad (44)$$

$$\frac{\partial Y_j^{(n)}}{\partial I_j^{(n)}} = g'(I_j^{(n)}) \quad (45)$$

$$\frac{\partial E}{\partial Y_j^{(n)}} = (Y_j^{(n)} - d_j) \quad (46)$$

n→ é número da camada

$I_j^{(n)}$→ são as camadas escondidas

g'→ é a derivada

d_j→ é o valor alvo (desejado)

4.4 Redes Neurais Artificiais MLP

As Redes Neurais Multicamadas são compostas pela camada de entrada dos dados, que recebe o nome *inputs* ou de variáveis independentes, a qual é geralmente representada pela variável x, em seguida vem a camada intermediária ou escondida, que pode ter mais que uma, sendo esta representada normalmente por h, e por fim pela camada de saída, que também recebe o nome de *output*, variável resposta ou variável dependente, normalmente representada por y.

Duas camadas de Rede Neural Artificial com um número suficiente de unidades escondidas pode aproximar qualquer função contínua, isto faz a RNA MLP uma poderosa ferramenta de modelagem. As RNA podem ser valiosas quando não sabemos a relação funcional entre as variáveis independentes e dependentes. Ela usa os dados para determinar a relação funcional entre as variáveis dependentes e independentes. Uma vez que ela é dependente de dados, seu desempenho melhora com o tamanho da amostra. É um processo iterativo que utiliza o método do gradiente descendente. Essas redes não impõem uma relação funcional entre as variáveis independentes e dependentes. Em vez disso, a relação funcional é determinada pelos dados no processo de encontrar os valores para os pesos (WARNER;

MISRA, 1996). A desvantagem é que é difícil de interpretar a rede. Outra desvantagem da rede neural segundo Specht (1991) é que a convergência de uma solução pode ser lenta, em virtude da necessidade de se ter um grande número de iterações para convergir para uma solução desejada, além de que conforme enfatiza Hinton (1992) ela depende das condições iniciais da rede. As redes neurais podem, também, ser vistas como um método de regressão não paramétrica (WARNER; MISRA, 1996), sendo que as redes MLPs são geralmente formadas por um algoritmo chamado regra delta generalizada, que calcula derivadas por uma simples aplicação da regra da cadeia, que por fim passa-se a se chamar *backpropagation* (SARLE, 1994).

Redes Neurais MLP oferecem um conjunto poderoso de ferramentas para resolver problemas em reconhecimento de padrões, processamento de dados, e controle não linear, além de ter uma importante capacidade de aprender uma solução geral para um problema, a partir de conjunto específico de exemplos (BISHOP, 1994). Elas podem ser usadas quando você tem pouco conhecimento sobre a forma da relação entre as variáveis independentes e dependentes, e pode-se variar a complexidade do modelo MLP variando o número de camadas escondidas e o número de neurônios escondidos em cada camada escondida (SARLE, 1994). Estas redes neurais podem também ser vistas como membros da classe de modelos estatísticos conhecidos como não paramétrico, logo a teoria geral da estatística não paramétrica está disponível para analisar o comportamento dela (JORDAN; BISHOP, 1996).

Processo de treinamento de uma rede neural (BISHOP, 1994).

1) Selecione o valor do nº de camadas escondidas na rede, inicialize os pesos usando números aleatórios.

2) Defina o critério de minimização do erro com relação ao conjunto de dados usando um dos algoritmos padrão, tal como *backpropagation*.

3) Repita o processo de treinamento um número de vezes usando diferentes inicializações aleatórias para os pesos da rede. Isto representa uma tentativa de encontrar bons mínimos na função de erro. A rede que tiver o menor valor de erro residual é selecionada.

4) Testar a rede treinada para avaliar a função de erro usando o conjunto de dado de teste.

5) Repita o treinamento e teste procedimentos para a rede tendo diferentes números de camadas escondidas e selecione a rede que tem menor erro de teste.

O processo de treinamento de uma rede neural, mencionado acima, na maioria das vezes, a escolha dos parâmetros é realizada manualmente através do método de tentativa e

erro, que é tedioso, menos produtivo, e propenso a erros (ZANCHETTIN; LUDEMIR; ALMEIDA, 2011).

Os modelos de Redes Neurais Artificiais são uma alternativa ao uso dos modelos de regressão, principalmente na presença de ruídos ou dados incompletos (GOMES; LUDEMIR, 2008). A principal vantagem de redes neurais em relação às técnicas estatísticas é que o modelo de RNA não tem de ser explicitamente definida antes do início do experimento, além disso, modelos estatísticos são difíceis de integrar dados de diferentes formatos (ou seja, trabalhando simultaneamente com variáveis contínua, binária, ordinal e nominal), mas isto pode facilmente ser conseguido usando RNAs (ENNETT; FRIZEL, 2003). Graficamente, as RNAS MLP podem ser apresentadas como:

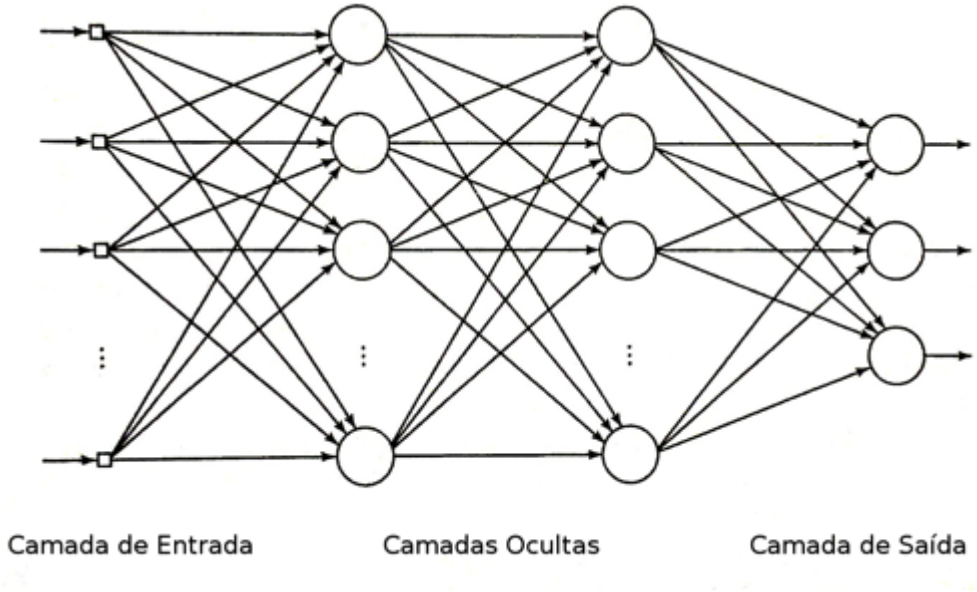

Figura 5: Rede Neural MLP

4.5 Importância do uso de Novas Funções de Ativação

Apesar do foco dos especialistas em desenvolver novos algoritmos de aprendizado, e novas arquiteturas para RNA MLP, alguns estudos têm apresentado a importância das funções de ativação para o aprendizado da RNA MLP, sendo que muitos especialistas a consideram, tão importante quanto à arquitetura e o algoritmo de aprendizado dela (GOMES; LUDEMIR, 2008).

Corriqueiramente, observa-se que as funções de ativação mais utilizadas tanto por pesquisadores quanto por profissionais são a sigmoide, tangente hiperbólica e linear. Devido a isso, os softwares que possuem ferramentas para o uso das RNA, restringem-se também a tais funções.

Com o intuito de reduzir a complexidade e melhorar o desempenho de uma RNA, mesmo diante da presença de *outliers*, Gomes & Ludemir (2008) propuseram novas funções de ativação: função logit e complemento log-log, as quais apresentaram melhor desempenho, quando comparadas com as funções tipicamente utilizadas.

A melhoria adquirida com estas funções exerce um papel muito importante no processo de convergência, além de que novas funções de ativação em algoritmos de aprendizagem é uma tarefa simples, visto que basta substituir as funções padrões por estas novas, com suas respectivas derivadas (GOMES; LUDEMIR; LIMA, 2011).

Na tese de Gomes (2010) foram propostas três funções de parâmetros fixo (complemento log-log, probit e log-log), e uma função de ativação com parâmetro livre, denominada Aranda-Ordaz, pertencente à família de funções de ativação assimétricas. Todas elas são funções monotonicamente crescente, limitadas ao intervalo [0,1], ou seja, o limite tende a 0 quando $x \rightarrow -\infty$, e o limite tende a 1 quando $x \rightarrow +\infty$. Além disso, todas estas funções são contínuas diferenciáveis, o que significa dizer que todas elas são funções não constantes, devido ao fato de que suas respectivas derivadas são diferentes de zero.

4.5.1 Funções de Ativação

4.5.1.1 Sigmoide

A função sigmoide, cujo gráfico tem a forma de s, é a função de ativação mais amplamente utilizada em redes neurais. Ela é definida como uma função estritamente crescente que exibe um equilíbrio adequado entre comportamento linear e não linear. A função sigmoide assume uma faixa contínua de valores entre 0 e 1 e é diferenciável (HAYKIN, 2001).

$$\emptyset(v) = \frac{1}{1 + e^{-a.v}} \tag{47}$$

Onde α é o parâmetro de inclinação da função sigmoide.

4.5.1.2 Aranda-Ordaz

É uma função de ativação baseada na transformação da família assimétrica de Aranda-Ordaz. Esta função de ativação difere da função sigmoide e tangente hiperbólica, por

apresentar um parâmetro livre em sua implementação. Esta função é modelada matematicamente por (GOMES, 2010):

$$\emptyset_i(u_i(t)) = f_\lambda(u_i(t)) = 1 - (1 + \lambda e^{u_i(t)})^{-1/\lambda} \quad (48)$$

Onde λ é considerado um parâmetro livre.

4.5.1.3 Tangente Hiperbólica

A função de ativação tangente hiperbólica se estende no intervalo entre -1 e 1, assumindo, neste caso, uma forma antissimétrica em relação à origem (HAYKIN, 2001). De acordo com Kovács (2006), outra vantagem desta função é que ele possui todas as derivadas contínuas, e é representado por:

$$g(v) = tanh(v) = \frac{e^v - e^{-v}}{e^v + e^{-v}} \quad (49)$$

4.5.1.4 Complemento Log-Log

Esta função é semelhante à sigmoide para valores de π próximos de 0,5, mas difere para π próximo de 0 ou 1. A função Complemento log-log apresenta melhor desempenho quando comparada com outras, principalmente quando a distribuição é muito enviesada (DOBSON, 2002).

$$\pi(x) = 1 - exp(-exp(p * x)) \quad (50)$$

Onde *p* é o parâmetro livre da função Complementar Log-Log.

4.5.1.5 Log-Log

A função Log-Log é semelhante à sigmoide para valores altos de f (x), sendo esta função monotônica crescente (MI) e tem um comportamento positivo assimétrico. A expressão para esta função de ativação é dado por (GOMES; LUDEMIR; LIMA, 2011; GOMES, 2010):

$$f(x) = exp(-exp(-p * x)) \quad (51)$$

A seguir é demonstrada a aplicação do método de estimativa de máxima verossimilhança nas funções de ativação utilizadas neste trabalho, que foram descritas acima.

4.5.2 Funções de Ativação modificadas pela EMV

As redes neurais treinadas pelo método de máxima verossimilhança são geralmente estatisticamente eficiente e assintoticamente imparciais (YANG; MURATA; ARMARI, 1998), desta forma exercerá um papel importante no processo de convergência, melhorando sua capacidade de generalização. A generalização de uma RNA MLP é definida por Daliri & Fatan (2011) como a capacidade de uma rede em estender com acurácia sua resposta aos novos dados ou dados com ruído, bem como a um comportamento da rede em novas situações. De acordo com La Rocca & Perna (2014), o poder de generalização de uma rede aumenta, principalmente quando a amostra de treinamento é grande e também, quando se aumenta o número de nós escondidos. Bishop (2006) também cita que quando o método de máxima verossimilhança é aplicado a distribuição gaussiana, esta consegue atingir a suficiência estatística, dado que toda informação contida na população pode ser obtida pelo parâmetro estimado θ. Seguindo o framework proposto por (Yang; Murata; Armari, 1998; Bishop, 1995), o qual combina o método de máxima verossimilhança com RNA, para melhorar o processo de aprendizagem, tem-se como resultado as funções que serão apresentadas no restante desta seção.

4.5.2.1 Sigmoide com EMV

Em relação à função de ativação sigmoide, temos o seguinte modelo:

$$log\, P(\theta/v) = log[\, 1/(1 + e^{-v})] \tag{52}$$

$$log\, P(\theta/v) = \, -log(1 + e^{-v}) \tag{53}$$

$$\frac{\partial \ log\, P(\theta/v)}{\partial \theta} = \frac{e^{-v}}{1 + e^{-v}} \tag{54}$$

Assume-se que o valor de α é 1. Nesta função, quando os valores de entrada tendem a - ∞, os valores de saída tende a 1, já quando os valores de entrada tendem a + ∞, os valores de saída tende a 0.

4.5.2.2 Aranda-Ordaz com EMV

A aplicação do método de estimativa de máxima verossimilhança na função de ativação Aranda-Ordaz, gera a seguinte função para ser utilizada neste estudo:

$$log\ \emptyset_i(u_i(t)) = log\ \ [1 - (1 + \lambda e^{u_i(t)})^{-1/\lambda}] \tag{55}$$

$$log\ \emptyset_i(u_i(t)) = log((\lambda e^{u_i(t)} + 1)^{-\frac{1}{\lambda}}((\lambda e^{u_i(t)} + 1)^{\frac{1}{\lambda}} - 1)) \tag{56}$$

$$log\ \emptyset_i(u_i(t)) = \ log\left((\lambda e^{u_i(t)} + 1)^{\frac{1}{\lambda}} - 1\right) - log\left((\lambda e^{u_i(t)} + 1)^{\frac{1}{\lambda}}\right) \tag{57}$$

$$\frac{\partial\ log\ \emptyset_i(u_i(t))}{\partial t} = \frac{e^{u_i(t)}}{(\lambda e^{u_i(t)} + 1)\left((\lambda e^{u_i(t)} + 1)^{\frac{1}{\lambda}} - 1\right)} \tag{58}$$

Uma característica importante desta função é que, quando os valores de entrada tendem a - ∞ ou + ∞, os valores de saída tendem a ∞, sendo λ o fator de ajuste livre, responsável por essa mudança.

4.5.2.3 Tangente Hiperbólica com EMV

Nesta função, quando os valores de entrada tendem à - ∞, os valores de saída tende a 1, já quando os valores de entrada tendem à + ∞, os valores de saída tende a -1 .Para a função tangente hiperbólica, temos:

$$log\ g(v) = log\ \ (tanh(v)) \tag{59}$$

$$\frac{\partial\ log\ g(v)}{\partial v} = \frac{\partial\ log(tanh(v))}{\partial v} \tag{60}$$

$$\frac{\partial\ log\ g(v)}{\partial v} = csch(v) * sech(v) \tag{61}$$

4.5.2.4 Complemento Log-Log com EMV

Em relação à função complemento Log-Log, temos os seguintes passos:

$$log\,\pi(x) = log\,[1 - exp(-exp(p * x))] \tag{62}$$

$$log\,\pi(x) = log[exp(-exp(p * x))(exp(exp(p * x)) - 1)] \tag{63}$$

$$\frac{\partial\ log\,\pi(x)}{\partial x} = \frac{p(exp(px))}{exp(exp(px)) - 1} \tag{64}$$

Esta função apresenta um comportamento diferente das outras, bem como tem um fator de configuração livre, p, que deve ser escolhido de forma a melhorar o desempenho da função. Aqui quando os valores de entrada tendem a - ∞ ou + ∞, os valores de saída tendem a ∞.

4.5.2.5 Log-Log com EMV

Por fim, tem-se o método de estimativa de máxima verossimilhança aplicado à função log-log.

$$log\,f(x) = log\,[exp(-exp(-p * x))] \tag{65}$$

$$\frac{\partial\ log\,f(x)}{\partial x} = \frac{\partial log\,[exp(-exp(-p * x))]}{\partial x} \tag{66}$$

$$\frac{\partial\ log\,f(x)}{\partial x} = \frac{exp(-exp(-px)) * (exp(-px) * p)}{exp(-exp(-px))} \tag{67}$$

$$\frac{\partial\ log\,f(x)}{\partial x} = p * exp(-px) \tag{68}$$

A diferença desta função para a função complemento Log-Log é um sinal negativo no parâmetro livre p, o que leva a ter valores de saída de função que tende para + ∞, quando se tem valores de entrada que tendem a - ∞. Assim, quando tem valores de entrada que tendem a + ∞, esta função tem com valores de saída convergindo para zero.

Dois Quadros, com todas as funções de ativação e suas respectivas derivadas, são apresentados respectivamente abaixo.

Quadro 1: Funções de Ativação com suas derivadas.

Rótulo	Função	Derivada
SIG	$\emptyset(v) = \frac{1}{1+e^{-a.v}}$	$\emptyset'(v) = \emptyset(v)(1-\emptyset(v))$
AO	$\emptyset_i(v) = 1-(1+\lambda e^{v})^{-1/\lambda}$	$\emptyset'(v) = (1+\lambda e^{v})^{-(1+\lambda)/\lambda} e^{v}$
TH	$\emptyset(v) = tanh(v)$	$\emptyset'(v) = \frac{1}{cosh^2(v)}$
CLL	$\emptyset(v) = 1-e^{-e^{pv}}$	$\emptyset'(v) = pe^{pv}e^{-e^{pv}}$
LL	$\emptyset(v) = e^{-e^{-pv}}$	$\emptyset'(v) = pe^{-pv}e^{-e^{-pv}}$

Nota: Os significados dos rótulos são: SIG (Sigmoide), AO (Aranda-Ordaz), TH (Tangente Hiperbólica), CLL (Complementar Log-Log) e LL (Log-Log).

Quadro 2: Funções de Ativação com EMV e suas derivadas.

Rótulo	Função com EMV	Derivada com EMV
SIGEMV	$\emptyset(v) = \frac{1}{1+e^{v}}$	$\emptyset'(v) = -\frac{e^{x}}{(1+e^{x})^{2}}$
AOEMV	$\emptyset(v) = \frac{e^{v}}{(1+\lambda e^{v})((1+\lambda e^{v})^{1/\lambda}-1)}$	$\emptyset'(v) = \frac{e^{v}(-e^{v}(e^{v}\lambda+1)^{\frac{1}{\lambda}} + (e^{v}\lambda+1)^{\frac{1}{\lambda}}-1)}{(e^{v}\lambda+1)^{2}((e^{v}\lambda+1)^{\frac{1}{\lambda}}-1)^{2}}$
THEMV	$\emptyset(v) = csch(v) * sech(v)$	$\emptyset'(v) = -4coth(2v) * csch(2v)$
CLLEMV	$\emptyset(v) = \frac{pe^{pv}}{e^{e^{pv}}-1}$	$\emptyset'(v) = \frac{-p^{2}e^{pv}(-e^{e^{pv}} + e^{pv+e^{pv}}+1)}{(e^{e^{pv}}-1)^{2}}$
LLEMV	$\emptyset(v) = pe^{-pv}$	$\emptyset'(v) = -p^{2}e^{-pv})$

Nota: Os significados dos rótulos são: SIGEMV (Sigmoide modificada com o EMV), AOEMV (Aranda-Ordaz modificada com o EMV), THEMV (Tangente Hiperbólica modificada com o EMV), CLLEMV (Complementar Log-Log modificada com o EMV) e LLEMV (Log-Log modificada com o EMV).

Para uma melhor compreensão, as funções foram rotuladas com suas letras iniciais, como aparecem no Quadro 1 e 2. A seguir são apresentados os pseudocódigos que foram utilizados neste trabalho. .

4.6 Pseudocódigos dos Algoritmos Propostos

Inicialmente, para cada porcentagem de dados faltantes foram testadas todas as funções de ativação, que foram apresentadas nas seções 4.5.1 e 4.5.2. Para testar todas as funções propostas, recorreu-se a três frameworks, que estão postos a seguir nos Quadros 3, 4 e 5, através dos pseudocódigos.

Quadro 3: Pseudocódigo que usa a mesma função de ativação em todas as camadas

```
Algoritmo – Pseudocódigo da RNA-MLP
w   ← Peso Inicial
x   ← Dados de Treinamento
t   ← Target
fun ← Função usada em todas as camadas
viés ← 1
n    ← Número de Camadas
iteração  ← 1000
Para i=1 até iteração
  net  ← w*x
  y(i)  ← fun(net) + viés
  Para j=2 até n
    net  ← w*y(i-1)
    y(i) ← fun(net) + viés
  Fim-Para
Fim-Para
```

O Quadro 3, tem o primeiro pseudocódigo utilizado na análise, o qual utiliza em todas as camadas da rede uma única função de ativação. O próximo pseudocódigo tem uma sutil diferença em relação ao primeiro, que é a utilização de método de Estimativa de Máxima Verossimilhança na função de custo (neurônio de saída da rede da última camada), conforme foi apresentado na seção 4.5.2.

Quadro 4: Pseudocódigo que utiliza na camada de saída a função com o EMV

```
   Algoritmo – Pseudocódigo da RNA com EMV
w   ← Peso Inicial
x   ← Dados de Treinamento
t   ← Target
fun ← Função usada na camada inicial e
intermediária
nova_fun  ← Função de ativação usando EMV (ver
seção 4.7)
viés  ← 1
n   ← Número de camadas
iteração  ← 1000
Para i=1 à iteração
  net    ← w*x
  y(i)   ← fun(net) + viés
  Para j=1 a (2:n)
    Se j≠n
      net  ← w*y(i-1)
      y(i) ← fun(net) + viés
    Fim-Se
    Senão
      net   ← w*y(i-1)
      y(i)  ← nova_fun (net) + viés
    Fim-Senão
  Fim-Para
Fim-Para
```

E por fim, o último pseudocódigo, o qual tem em todas as camadas a função de ativação utilizando o método EMV.

Quadro 5: Pseudocódigo que utiliza na camada de saída as funções com EMV

```
Algoritmo – Pseudocódigo da RNA com EMV
w   ← Peso Inicial
x   ← Dados de Treinamento
t   ← Target
nova_fun  ← Função de ativação usando EMV (ver
seção 4.7)
viés  ←  1
n   ←  Número de Camadas
iteração   ←  1000
Para i=1 até iteração
  net   ← w*x
  y(i)  ← nova_fun  (net) + viés
  Para j=2 a n
    net  ← w*y(i-1)
    y(i) ← nova_fun  (net) + viés
  Fim-Para
Fim-Para
```

Para este pseudocódigo (Quadro 5), em todas as camadas as funções de ativação utilizadas seguiram o que foi proposto na seção 4.5.2, ou seja, as cinco funções analisadas neste trabalho foram modificadas pelo método de estimativa de máxima verossimilhança. Cabe salientar também, que todos estes pseudocódigos foram utilizados tanto com o método de imputação única de dados, como com o método de imputação múltipla. A seguir é apresentada a Tabela 19 com os dados analisados pelo viés de RNA-MLP, a qual contém as medidas de sensibilidade obtidas através dos três pseudocódigos elencados acima.

4.7 Trabalhos Relacionados às Redes Neurais MLP para Tratar Dados Faltantes

Ennett & Frizel (2003) propuseram a avaliação da habilidade de um sistema híbrido RNA-RBC (Redes Neurais Artificiais - Raciocínio Baseado em Casos) para imputar os valores faltantes em um banco de dados, que continha 5102 casos completos de paciente de uma Unidade de Terapia Intensiva Neonatal (UTIN) no Canadá. Para realizar o estudo foi necessário remover dados do *dataset* na quantidade de 16%, 40%, 53% e 64% respectivamente, para se criar um banco artificial, possibilitando assim analisar o desempenho do sistema proposto. As saídas do sistema híbrido foram comparadas com duas abordagens de referência: a substituição dos valores faltantes pela média e por valores aleatórios. O

sistema híbrido de imputação apresentou um desempenho ligeiramente melhor do que a imputação pela média e imputação aleatória.

No trabalho de Abdella & Marwala (2005) foi apresentado um método que visava aproximar dados faltantes em um banco de dados de uma cervejaria da África do Sul, usando uma combinação de algoritmos genéticos com Rede Neural MLP e Redes *Radial-Basis Functions* (RBF). Sendo estas redes constituídas de 14 entradas, 10 neurônios na camada intermediária e 14 saídas, com um total de 198 dados de treinamento. Foram retiradas do total de dados, a quantidade de 1, 2, 3, 4 e 5 amostras respectivamente, e para avaliar a precisão dos valores utilizou-se como critério o coeficiente de correlação e o erro padrão. Como resultado observou-se que a rede RBF apresentou melhores resultados, com acurácia de 96% quando comparada a Rede MLP que foi de 93%.

Wen & Lee (2005) apresentaram um estudo de caso com dados coletados de detectores de tráfego de autoestrada. Os dados de campo foram analisados e coletados a partir de uma rodovia em Taiwan, com dez detectores localizados a cerca de 500 m de intervalo ao longo de 6,03 km de comprimento, cujo principal objetivo foi estudar os dados no período da manhã em horário de pico (08h00-09h00), durante uma semana de setembro em 2002. Após inspeções iniciais nos dados coletados, observou-se muitos valores discrepantes e faltantes no *dataset*, sendo que alguns detectores não registraram nenhum dado. O estudo centrou-se no tratamento de dados faltantes e fusão de dados para detectores de tráfego de dados, que tenta integrar modelagem *grey*, que é uma técnica de inteligência artificial, na imputação de dados e modelos de fusão de dados em redes neurais. Este estudo propõe uma inovadora rede neural recorrente *grey-based*, que integrou modelos *grey* na rede neural recorrente, para a estimativa do tempo de viagem dinâmica.

Já no trabalho de Mohamed & Marwala (2005) utilizou-se um conjunto de dados com 5.776 registros, do Departamento Sul Africano de Saúde, com variáveis que continham informações sobre HIV, idade, faixa etária e gravidez (número das gestações), usando também redes neurais para tratar os dados faltantes (MOHAMED; MARWALA, 2005).

Mohamed, Nelwamondo & Marwala (2007) propuseram uma rede híbrida auto associativa, onde seu desempenho em conjunto com o Algoritmo Genético é comparado com de uma RNA MLP. Um sistema de uma PCA e rede neural também foi desenvolvido em comparação com os outros dois sistemas. Os dados utilizados neste experimento são de HIV do Departamento de Saúde da África do Sul. Como resultado o sistema híbrido auto-associativo produz o menor erro padrão médio e tem o coeficiente de correlação global maior.

A rede híbrida apresentou melhor desempenho do que uma RNA padrão, principalmente, quando esta rede híbrida foi aplicada para imputação única, enquanto o modelo PCA e rede neural padrão fornece mais consistência para múltiplas imputações.

Utilizando conjuntos de dados de uma usina de energia industrial e de HIV, para tratar dados faltantes, utilizou-se o algoritmo EM, em comparação a um sistema baseado em redes neurais auto associativa com o algoritmo genético. Os resultados mostram que o algoritmo EM apresentou melhor desempenho nos casos em que há pouca ou nenhuma dependência entre as variáveis de entrada, já a rede neural auto associativa combinada com o algoritmo genético apresentou melhor desempenho quando há alguns relacionamentos não-lineares inerentes entre algumas das variáveis dadas (NELWAMONDO; MOHAMED; MARWALA, 2007).

No trabalho de Randolph-Gips (2008) é apresentado a Rede Neural Cosseno (COSNN), e mostrado como ela pode ser utilizada para processar *dataset* com dados faltantes, sem imputação. Nele é usada uma função baseada em cosseno com uma norma ponderada, que pode ser treinada para combinar os dados de entrada, sem qualquer supressão ou imputação de dados incompletos. Seu desempenho foi comparado com Redes Neurais *Feedforward* usando exclusão e imputação, sendo que a COSNN apresentou melhor resultado que a RNA MLP.

Ssali & Marwala (2007) introduziram um novo paradigma para imputar dados faltantes, que combina um modelo baseado em árvore de decisão com uma rede neural auto associativa (AANN), e um modelo baseado em análise de componentes principais com rede neural (PCA-NN). Os resultados indicam que houve um aumento médio na precisão de 13% com o AANN, cuja exatidão média do modelo passou de 75,8% para 86,3%, enquanto a do modelo de PCA-NN aumentou de 66,1% para 81,6%.

No trabalho proposto por Aydilek & Arslan (2012), foi apresentado um inovador método de rede neural híbrida e K-vizinhos mais próximos ponderados para estimar os valores faltantes. Os resultados mostraram que este método adiciona vantagens significativas ao modelo básico de rede neural. Estimativas NN-KNN são mais sensíveis e exatas, produzindo melhor precisão de imputação, principalmente quando a amostra tiver mais de um valor faltante. Constata-se também que o método apresenta melhor desempenho em casos onde existe dependência entre as variáveis. Finalmente, no livro de Marwala (2009) há uma compilação de vários artigos, que utilizam técnicas de Inteligência Computacional para tratar problemas de dados faltantes.

4.8 Considerações Finais do Capítulo

Neste capítulo foi apresentada a teoria de Redes Neurais MLP, bem como as diversas abordagens atreladas à ela, tais como novas funções de ativação e a combinação da EMV com a RNA-MLP, com foco no tratamento do problema de dados faltantes. Fez-se também uma análise do estado da arte, sendo que na maioria dos casos as RNA foram combinadas com outros algoritmos. Tais abordagens apesar de terem demostrado melhor acurácia, quando comparado com outros métodos clássicos da área de aprendizado de máquina ou da estatística, trazem consigo a desvantagem do alto nível de complexidade e custo computacional.

CAPITULO 5

5 RESULTADOS EXPERIMENTAIS

Para realizar esta fase do trabalho, com o objetivo de utilizar as abordagens propostas nas seções 2.3.1 (viés do MAR, que ocorre quando a probabilidade de um registro com um valor em falta para um atributo pode depender dos dados observados, mas não do valor dos dados faltantes em si), 3.1 (que é a imputação única ou também conhecida como imputação simples, a qual preenche por um único valor cada dado faltante na amostra), 3.3.3.3.1 (que aborda o algoritmo EM, o qual, também usa técnicas estatísticas para maximizar verossimilhanças complexas, cujo objetivo é calcular a Máxima Verossimilhança a partir de dados incompletos), 4.5.1 (que apresentado a importância de novas funções de ativação para o aprendizado da RNA MLP) e 4.5.2 (o qual aborda RNA MLP treinadas pelo EMV para melhorar a capacidade de generalização e acurácia); escolheu-se 4 bases de dados, no contexto de aprendizado supervisionado, via o paradigma de regressão, possibilitando assim a avaliação da acurácia destas abordagens. Frise-se que os *dataset* não continha nenhum dado faltante, sendo necessário gerar artificialmente a porcentagem de faltantes, conforme o mecanismo MAR. Na Tabela 1, a seguir, tem-se uma descrição quantitativa dos dados.

Tabela 1: Bases de Dados utilizadas no experimento

dataset	Emulsão	Breast Tissue	Concrete	Parkinsons
locDow	Amani *et al.* (2008)	UCI - *Machine Learning*		
qtdVarInd	5	9	7	20
qtdVarDep	1	1	3	2
qtdSample	60	106	103	5875
% missing	5%, 10%, 20%, 30%, 40%, 50%, 60% e 70%			
semente	123, 43112, 1234567 e 1802			

Nota: Os rótulos tem o seguinte significado: locDow (local onde os dados foram encontrados), qtdVarInd (quantidade de variáveis independentes), qtdVarDep (quantidade de variáveis dependentes), qtdSample (quantidades de amostras), % Missing (porcentagem de dados faltantes inseridos em cada base de dados), semente (semente utilizada nos experimentos).

A base Emulsão foi inicialmente utilizada no trabalho de Amani *et al*. (2008), que teve como foco determinar os fatores que influenciavam o tamanho da partícula de nano emulsão através de uma RNA-MLP pelo paradigma de regressão.

A segunda base de dados, que é a *Breast Tissue*, foi inicialmente utilizado por Jossinet (1996). Esse *dataset* contém medidas de impedância elétrica em amostras de tecidos mamários recém-extraídos.

Quanto à terceira base de dados, que é a *Concrete*, esta base foi inicialmente utilizada por Yeh (2007), sendo que este *dataset* refere-se à informações que foram coletadas a partir da análise da estrutura do concreto, o qual é um material altamente complexo, que deve ter um bom fluxo, quando está sendo utilizado, porém sua fluidez não é determinada apenas pelo teor de água, mas há também outros componentes que o influenciam, o que gera a necessidade de se estudar e modelar estes outros fatores que determinam sua fluidez.

Em relação à quarta base de dados, que é a *Parkinsons Telemonitoring*, esta também foi inicialmente utilizada por Little *et al*. (2009), sendo que este *dataset* refere-se à informações que foram coletadas a partir de uma série de medições de voz biomédicas, de 42 pessoas com a doença de Parkinson em estágio inicial. Os pacientes, que participaram do experimento, foram recrutados a partir de um estudo clínico, e em seguida utilizaram por seis meses consecutivos um dispositivo que telemonitorava a progressão dos sintomas remotamente e automaticamente na casa do paciente.

Cabe ressaltar que, em consequência das bases de dados serem matrizes, tem-se que as linhas representam as amostras ou instâncias, e as colunas representam as variáveis ou atributos. Dado que estamos trabalhando com padrão monotônico; para se gerar as bases de dados com valores faltantes, via o mecanismo MAR, escolheu-se aleatoriamente uma variável para ser a causadora dos dados omissos e uma variável para possuir os dados faltantes, sendo que a variável que foi escolhida para possuir os dados faltantes é a dependente.

Em virtude dos algoritmos serem estocásticos, ou seja, para cada vez que eles são executados, sempre terá como resultado um conjunto de dados imputados diferentes. Optou-se por escolher uma semente no início do experimento, a fim de assegurar que o experimento possa ser executado várias vezes ou por outra pessoa, e se obtenha os mesmos resultados. Para todas as bases de dados que serão analisadas a partir do método de imputação única, utilizou-se uma única semente (123), a qual garante que sempre ter-se-á os mesmos valores aleatórios, que são gerados no início do experimento. Já quando o método de imputação múltipla foi

usado, escolheu-se estas quatro sementes (123, 43112, 1234567 e 1802). Na próxima seção são apresentadas as medidas de sensibilidade, que foram utilizadas neste trabalho.

5.1 Medidas de Sensibilidade

5.1.1 MAE – *Mean Absolute Error*

A primeira medida de sensibilidade utilizada foi o MAE (do inglês: *Mean Absolute Error*). O Cálculo do MAE é relativamente simples e envolve a soma das grandezas (valores absolutos) dos erros para obter o "erro total", e em seguida, divide-se o erro total por *n*, que é o tamanho da amostra (WILLMOTT; MATSUURA, 2005). O MAE é representado matematicamente por:

$$MAE = \frac{\sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|}{n} \tag{69}$$

O MAE apresenta a informação sobre o desempenho a longo prazo dos modelos; sendo assim, quanto menor for o MAE melhor é a previsão do modelo a longo prazo (DORESWAMY; VASTRAD, 2013). Apesar de o MAE considerar grandes erros em seu cálculo, ele não consegue ponderá-los mais fortemente (TWOMEY; SMITH, 1997), mas segundo Willmott & Matsuura (2005) a medida mais natural do erro médio deve ser o MAE. De acordo com Chai & Draxler (2014) o uso da medida MAE, como critério de desempenho de uma rede neural é mais adequado quando os erros seguem uma distribuição uniforme.

5.1.2 RMSE – *Root Mean Square Error*

A avaliação e validação de modelos de rede neurais artificiais são baseadas na seleção de uma ou mais métricas de erro. Geralmente, estes modelos que realizam uma tarefa de aproximar funções usam uma métrica de erro contínuo, tal como, o erro absoluto médio (MAE), já citado anteriormente, e o erro quadrático médio (MSE do inglês: *Mean Square Error*) ou a raiz do erro quadrático médio (RMSE) (TWOMEY; SMITH, 1997). O RMSE, que também mede a acurácia dos modelos, é mais apropriado para representar o desempenho de um modelo do que o MAE, principalmente quando se espera que os erros não sejam enviesados e siga uma distribuição *Normal* (CHAI; DRAXLER, 2014). Uma justificativa para o uso do RMSE, é que como ele é elevado ao quadrado, isso retira a influência do sinal no

erro, permanecendo apenas a influência da magnitude dos erros na medida de erro médio (WILLMOTT; MATSUURA, 2005). Portanto, o RMSE (e da mesma forma, o MSE) penaliza erros distantes, ou seja, erros claros, com maior variância, mais severamente e, portanto, favorece uma RNA com pouco ou nenhum erro. Isso pode causar uma RNA com muitas previsões incertas (TWOMEY; SMITH, 1997), já que o RMSE é muito sensível a *outliers* (CHAI; DRAXLER, 2014). Matematicamente o RMSE pode ser representado por:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n}} \tag{69}$$

O cálculo do erro quadrático envolve uma sequência de 3 passos simples. Primeiro soma-se os erros quadráticos individuais, isto é, cada erro influencia no total uma proporção ao quadrado, ao invés de uma simples magnitude. Sendo assim, quando o erro é muito grande, tem-se como consequência uma maior influência sobre o erro quadrático total do que quando os erros são menores. No segundo passo, o erro quadrático total é dividido por *n*, o que produz o erro quadrático médio MSE. O terceiro e último passo persiste apenas em tomar a raiz quadrada do MSE (WILLMOTT; MATSUURA, 2005). Os erros de treino geralmente diferem dos erros de teste (TWOMEY; SMITH, 1997), e o RMSE é, por definição, nunca menor do que o MAE (CHAI; DRAXLER, 2014).

O RMSE possui duas componentes associadas a ele, sendo que a primeira mede a variabilidade do estimador (precisão) e a outra mede o seu viés (acurácia), sendo que a precisão está associada à erros aleatórios, enquanto que a acurácia está associada à erros sistemáticos. Mais detalhes podem ser encontrados em (MORETTIN, 2000).

5.2 Análise Preliminar dos Dados

Nesta seção é apresentada uma análise sintética dos dados, na qual tem como objetivo verificar como os dados estão distribuídos, bem como se há *outliers* ou valores discrepantes neles, quais são seus valores das medidas resumos, e por fim testou-se a normalidade dos dados.

5.2.1 Análise Preliminar da Base de dados Emulsão

Em um primeiro momento, há a necessidade de se analisar todas as variáveis, a fim de verificar como estas estão distribuídas, iniciando-se com uma análise gráfica dos dados, a qual nos possibilita perceber o quanto estes oscilam, conforme se observa na Figura 6.

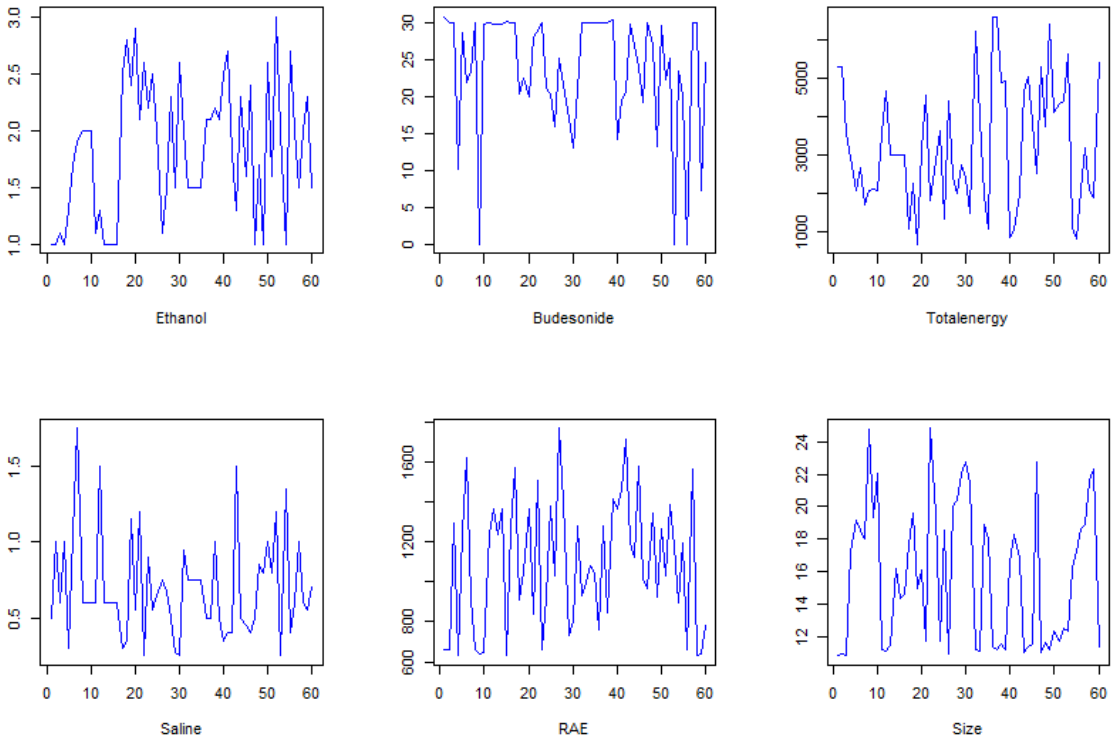

Figura 6: Gráfico da distribuição das variáveis da base emulsão.

Conforme os dados apresentados na Figura 6, onde os cinco primeiros gráficos representam as variáveis independentes e o sexto gráfico (Size) representa a variável dependente, sendo que a variável dependente é a que foi utilizada para conter os dados faltantes. Ao observar os gráficos das variáveis, percebe-se que estes oscilam muito, não apresentando um padrão *a priori*. Tal situação irá influenciar nas medidas de sensibilidade (MAE, RMSE), pois caso o valor faltante esteja próximo de valores extremos, a estimativa do valor a ser imputado será fortemente influenciada por tais valores.

Para uma melhor absorção das informações gráficas, obtidas a partir da Figura 6, tabula-se as medidas descritivas deste conjunto, para cada variável observada, conforme a Tabela 2

Tabela 2: Estatísticas descritivas das variáveis da base emulsão.

Métricas	Ethanol	Budesonide	Totalenergy	Saline	RAE	Size
Mínimo	1	0	650	0,25	629	10,44
1º Qu.	1,3	20,4	2038	0,5	829	11,39
Mediana	1,95	26,9	3000	0,6	1068	16,23
Média	1,823	23,75	3216	0,6984	1094	15,92
3º Qu.	2,3	30	4453	0,8625	1364	19,4
Máximo	3	30,7	6606	1,75	1771	24,81

Nota: As abreviações, 1º Qu → primeiro quartil, 3º Qu → terceiro quartil, As variáveis (Ethanol, Budesonide, Totalenergy, Saline e RAE) → são as variáveis independentes, Size → é a variável dependente, a qual terá os dados faltantes.

Ao fazer uma análise visual na Tabela 2, constata-se que algumas variáveis (*Budesonide* e *Totalenergy*) são fortes candidatas a possuírem *outliers*. Desta forma, como já é sabido que os *outliers* ou dados discrepantes influenciam fortemente o algoritmo EM (NG-CHI, 1998), surge a necessidade, para uma melhor visualização e para se ter maiores evidências de tal suspeitas, de se plotar um histograma, que é uma representação gráfica das frequências do conjunto de dados, que estão ordenados em classes. Tal gráfico torna-se útil, dado que através dele pode-se observar a distribuição dos dados, ou seja, pode-se por exemplo, perceber visualmente se estes dados se aproximam ou não de uma distribuição normal, ou se eles se aproximam de alguma distribuição conhecida, bem como se é unimodal, simétrico ou não. Além disso, possibilita analisar a dispersão dos dados, facilitando a identificação de *outliers* ou valores discrepantes. Tal representação gráfica é apresentada na Figura 7.

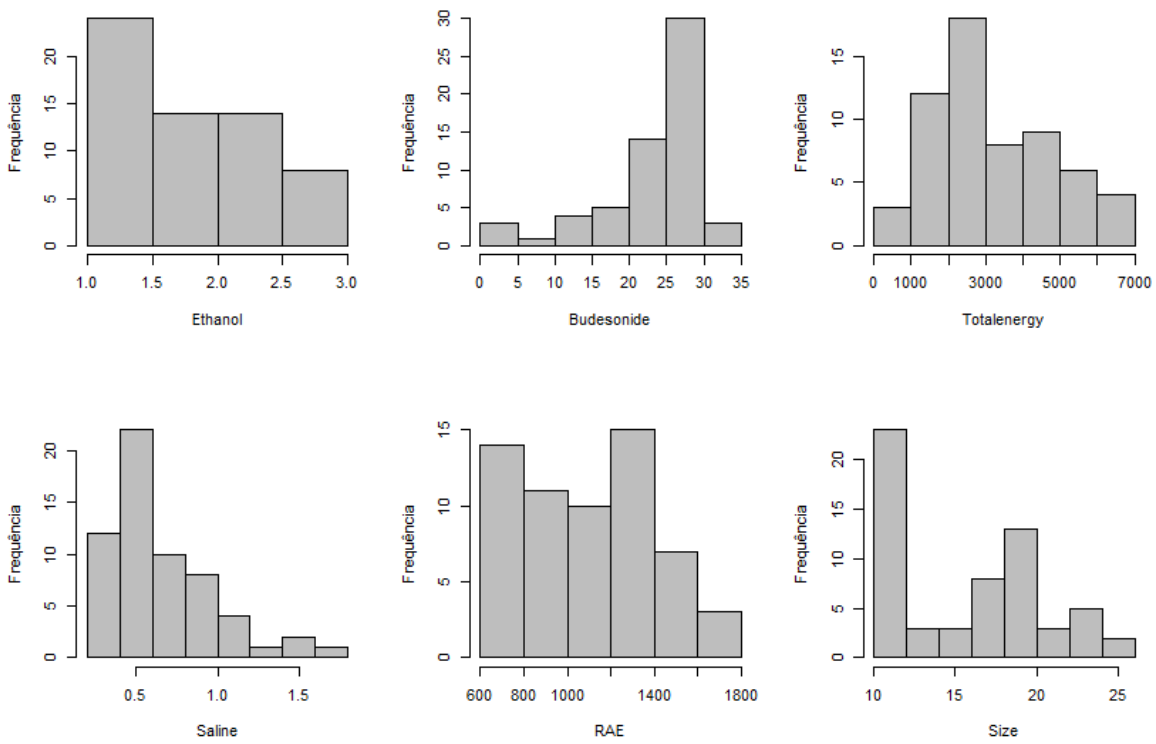

Figura 7: Histograma de todas as variáveis da base emulsão.

Conforme a Figura 7, a maioria das variáveis independentes, bem como a variável dependente possuem valores candidatos a serem *outliers*, porém para confirmar se estes valores são ou não *outliers*, deve-se recorrer ao teste estatístico qui-quadrado para *outliers*, que foi proposto por Dixon (1950).

Inicialmente, verifica-se quais são os possíveis valores candidatos a serem *outliers*, em cada variável de interesse, conforme Tabela 3.

Tabela 3: Valores que são plausíveis de serem *outliers* da base emulsão.

Ethanol	Budesonide	Totalenergy	Saline	RAE	Size
1	30,7	650	0,25	629	10,439
3	0	6606	1,75	1771	24,81

Na Tabela 3, observam-se os valores extremos, para cada variável em estudo, ou seja, o menor e o maior valor respectivamente, que podem ou não ser um *outliers*. Para confirmar as suspeitas, passa-se a fazer o teste de qui-quadrado, conforme apresentado na Tabela 4.

Tabela 4: p-valores para o teste de qui-quadrado da base emulsão.

	Ethanol	Budesonide	Totalenergy	Saline	RAE	Size
p-valor	0,04616	0,00306	0,03554	0,0015	0,03551	0,2123

Conforme os valores da Tabela 4, a única variável que tem *outliers* é a Size, que é a variável dependente. Entretanto, conforme o mecanismo MAR, apenas as outras variáveis é que influenciam nos valores imputados, logo, o fato de ter sido confirmado que a variável **Size** tem *outliers*, esta não exerce nenhuma influência na análise. O teste de qui-quadrado foi realizado a um nível de significância de 5%, que é o padrão.

O próximo passo é analisar a normalidade dos dados, visto que partiu-se do pressuposto que o algoritmo EM, foi modelado por uma distribuição *normal*, sendo que esta análise pode ser inicialmente feita através do gráfico de probabilidade normal, o qual está na Figura 8.

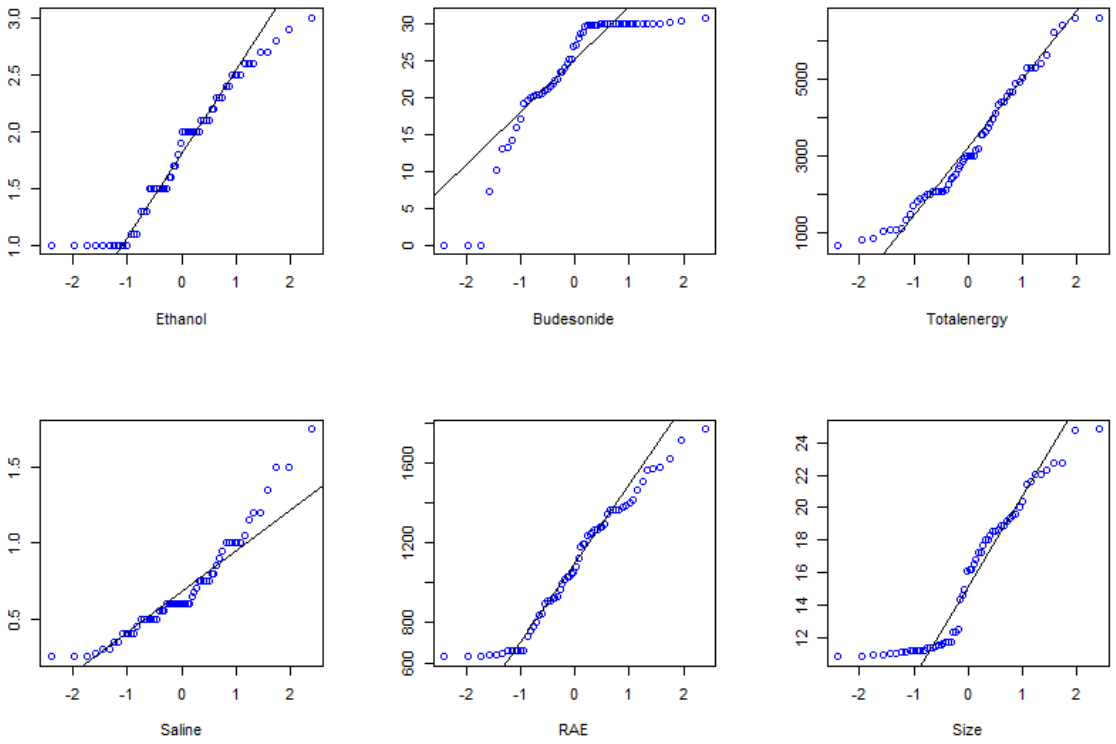

Figura 8: Gráfico de probabilidade normal para a base emulsão.

Nos gráficos da Figura 8, os quais são referentes à distribuição de percentis acumulados, que podem ser interpretados como: caso os pontos plotados sigam o padrão de uma reta, ou se aproxime muito de uma, demonstra-se visualmente que há evidências da variável aleatória em estudo ter uma distribuição que se aproxime da normal. Entretanto ao analisar a referida

Figura 8, verifica-se que os dados em determinados momentos se afastam muito da reta, logo pode-se inferir que estes não seguem uma distribuição normal. Outa alternativa à análise gráfica é através do teste para normalidade de Shapiro-Wilk, que é específico para testar normalidade de dados, o qual tem como resultados os valores apresentados na Tabela 5.

Tabela 5: Teste de normalidade de Shapiro-Wilk para a base emulsão.

	Ethanol	Budesonide	Totalenergy	Saline	RAE	Size
p-valor	0,004989	0,00006005	0,03278	0,000452	0,01474	0,04105

Os valores do p-valor apresentados na Tabela 5, foram comparados ao nível de significância de 5% (0,05), a fim de não rejeitar ou rejeitar a hipótese de nulidade do referido teste (H_0: Os dados seguem uma distribuição Normal). Constata-se que, para todas as variáveis analisadas, nenhuma delas segue uma distribuição normal. Tal situação evidencia que partir do pressuposto de modelar o algoritmo EM via uma distribuição normal não é uma suposição forte para a análise destes dados. Além disso, apesar de não ter havido *outliers* nas variáveis independentes, estas apresentam uma enorme oscilação, fato este que torna ainda mais difícil a modelagem via este algoritmo. A próxima seção tratará da análise da base de dados *Breast Tissue.*

5.2.2 Análise Preliminar da Base de dados Breast Tissue

Passando-se a analisar o *dataset Breast Tissue*, em um primeiro momento, plota-se os dados brutos nos gráficos da Figura 9, os quais servem para analisar as oscilações no *dataset* ao longo do tempo, e possibilitando também, às vezes, perceber tendências ou valores abruptos, tidos como anomalias.

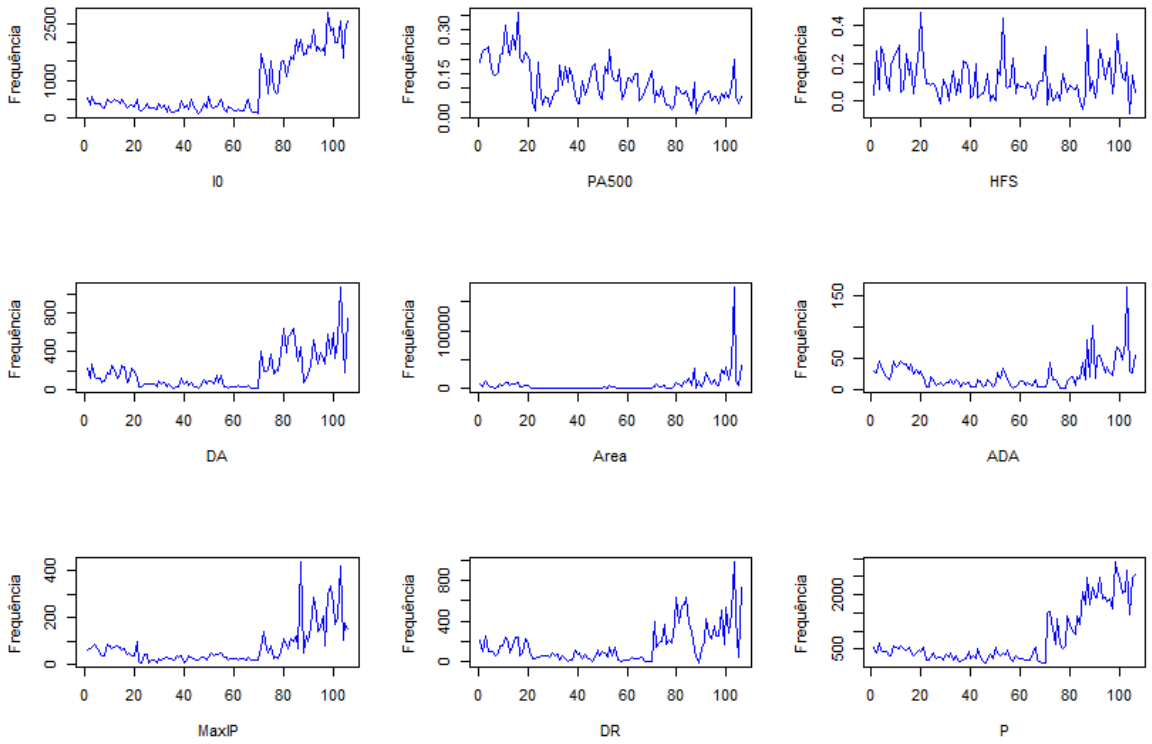

Figura 9: Gráfico dos dados brutos da base *Breast Tissue*.

Os dados apresentados na Figura 9, onde os oito primeiros gráficos representam as variáveis independentes e o nono gráfico (P) representa a variável dependente, sendo que a variável dependente é a que foi utilizada para conter os dados faltantes. Ao analisar os gráficos da Figura 9, percebe-se a presença de valores abruptos em todas as variáveis. Quando ocorre tal circunstância, as medidas de sensibilidade (MAE, RMSE) são muito influenciadas por tais valores, pois caso o valor faltante esteja próximo deles, a estimativa do valor a ser imputado será fortemente determinado por tais valores, principalmente quando se utiliza o algoritmo EM.

Dada a nitidez da presença de valores abruptos, na Figura 9, segue-se a análise tabulando tais dados, com o intuito de se obter as medidas descritivas destes, para cada variável observada, conforme a Tabela 6.

Tabela 6: Estatísticas descritivas das variáveis da base *Breast Tissue*.

Métricas	I0	PA500	HFS	DA	Area	ADA	MaxIP	DR	P
Mínimo	103	0,01239	-0,06632	19,65	70,43	1,596	7,969	-9,258	91,57
1º Qu.	250	0,06741	0,04398	53,85	409,6	8,18	26,89	41,78	277,8
Mediana	384,9	0,1054	0,08657	120,8	2220	16,13	44,22	97,83	439,4
Média	784,3	0,1201	0,1147	190,6	7335	23,47	75,38	166,7	807,5
3º Qu.	1488	0,1696	0,1665	255,3	7615	30,95	83,67	233	1336
Máximo	2800	0,3583	0,4677	1063	174500	164,1	436,1	977,6	2851

Nota: As abreviações, para as variáveis, têm o seguinte significado: I0 → Impedância (ohm) a frequência zero, PA500 → ângulo da fase à 500 KHZ, HFS → inclinação do ângulo da fase de alta frequência , DA → distância da impedância entre as extremidades do espectro, Area → área sobre o espectro, /DA → área normalizada pela DA, MaxIP → máximo do espectro, DR → distância entre I0 e a parte real do ponto de frequência máxima, P → tamanho da curva do espectro (esta é a variável dependente, a qual conterá os dados faltantes). A variável Class, não foi posta nesta tabela, em virtude desta se referir à classe que cada amostra pertence, logo não tem sentido fazer cálculos descritivos dela.

Passando a analisar os valores apresentados na Tabela 6, tem-se a impressão que todas as variáveis apresentam fortes evidências de possuírem *outliers*, já que há uma disparidade enorme entre os valores mínimos e máximos para cada uma das variáveis. Sendo assim, para uma melhor visualização, plota-se um histograma de todas as variáveis, para se ter maiores evidências, conforme é apresentado na Figura 10.

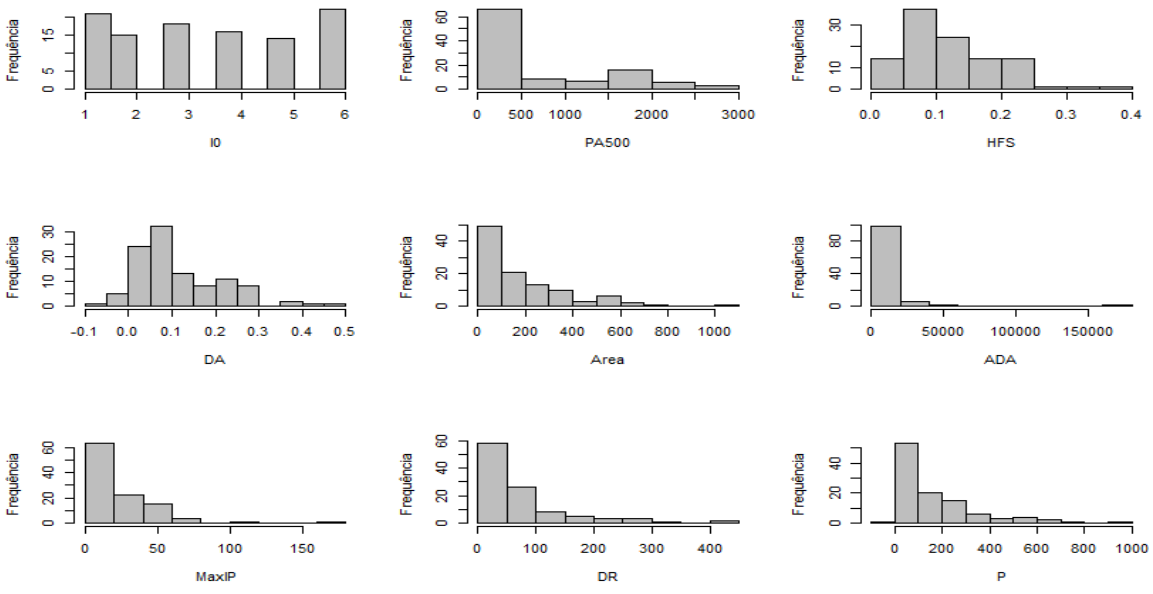

Figura 10: Histograma das variáveis da base *Breast Tissue*.

De acordo com a distribuição dos dados, verificados na Figura 10, a maioria das variáveis independentes, bem como a variável dependente possuem valores concentrados nos extremos, que dão indícios de serem *outliers*, porém para confirmar se estes valores são ou não *outliers*, deve-se recorrer ao teste estatístico de qui-quadrado, como anteriormente citado na seção 5.2.1. Primeiramente, verificam-se quais são os possíveis valores candidatos a serem *outliers*, em cada variável de interesse, conforme é pontuado na Tabela 7.

Tabela 7: Valores que são plausíveis de serem *outliers* para a base *Breast Tissue*.

I0	PA500	HFS	DA	Area	ADA	MaxIP	DR	P
103	0,012392	-0,06632	19,648	70,4262	1,5957	7,96878	-9,2577	91,571
2800	0,358316	0,467748	1063,4	174481	164,07	436,1	977,55	2851,1

Na Tabela 7, têm-se os valores extremos, para cada variável em análise, ou seja, o menor e o maior valor respectivamente, que podem ou não ser um *outlier*. Para confirmar tais indícios, verifica-se via o teste de qui-quadrado, conforme elencado na Tabela 8.

Tabela 8: p-valores para o teste de qui-quadrado para a base *Breast Tissue*.

	I0	PA500	HFS	DA	Area	ADA	MaxIP	DR	P
p-valor	0,008	0,000516	0,000495	5E-06	0,6958	2E-09	9,2E-06	8E-06	0,0074

Conforme os valores da Tabela 8, a única variável que tem *outliers* é a Area, que é uma das variáveis independente. Em consequência disso, conforme o mecanismo MAR, esta variável influenciará fortemente nos valores imputados. O teste de qui-quadrado foi realizado a um nível de significância de 5%, que é o padrão.

Outra forma de verificar a presença de *outliers* é através do gráfico box-plot, o qual apresenta os pontos extremos que estão além dos limites gráficos, conforme apresentado na Figura 11.

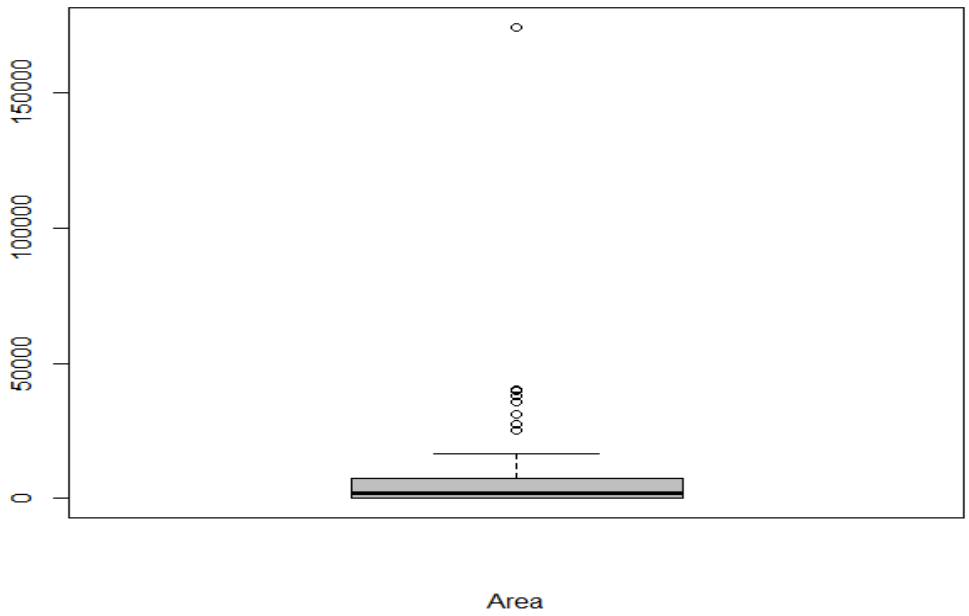

Figura 11: Gráficos de box-plot para a variável da base *Breast Tissue*.

Todos os pontos, que estão visíveis no box-plot da Figura 11, são considerados *outliers*, os quais poderão influenciar nos valores estimados, principalmente ao utilizar o algoritmo EM, visto que a natureza deste algoritmo no passo E utiliza-se da esperança matemática, consequentemente poderá conduzir a valores muito enviesados diante da presença de *outliers*.

Já que partiu-se da premissa que o algoritmo EM, foi modelado por uma distribuição normal, deve-se analisar a normalidade dos dados, análise esta que pode ser inicialmente feita através do gráfico de probabilidade normal, o qual está na Figura 12.

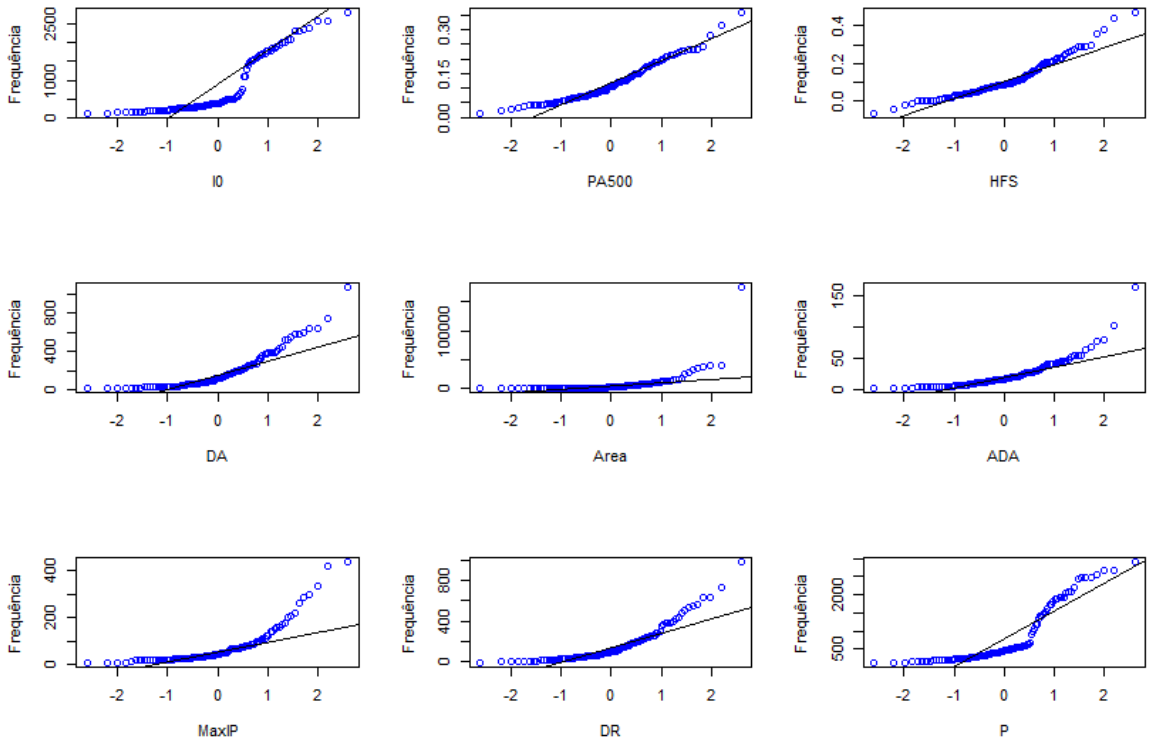

Figura 12: Gráfico de probabilidade normal para a base *Breast Tissue*.

Para os gráficos da Figura 12, os quais se referem à distribuição de percentis acumulados, que caso os valores plotados sigam o padrão de uma reta, ou se aproxime de uma, constata-se visualmente que há evidências da variável aleatória em análise ter uma distribuição que se aproxime da normal. Porém, para a atual situação, verifica-se que os dados em algum momento se afastam muito da reta, logo podemos inferir que estes não seguem uma distribuição normal. Como alternativa à análise gráfica, pode-se recorrer ao teste para normalidade de Shapiro-Wilk, que é específico para tal finalidade, o qual tem como resultados os valores apresentados na Tabela 8.

Tabela 9: Teste de normalidade de Shapiro-Wilk para a base *Breast Tissue*.

	I0	PA500	HFS	DA	Area	ADA	MaxIP	DR	P
p-valor	2E-11	5,77E-05	8,7E-06	1E-10	2,2E-16	4E-12	2E-13	9E-11	2E-11

Os valores do p-valor apresentados na Tabela 9, foram comparados ao nível de significância de 5% (0,05), a fim de não rejeitar ou rejeitar a hipótese de nulidade do referido teste (H0: Os dados seguem uma distribuição Normal). Constata-se que, para todas as variáveis analisadas, nenhuma delas segue uma distribuição normal.

Apesar de ter havido *outliers* em apenas uma variável independente, as demais apresentam valores abruptos, fato este que torna ainda mais difícil a modelagem via o algoritmo EM. Na próxima seção será analisada a base de dados *Concrete*.

5.2.3 Análise Preliminar da Base de dados Concrete

Para a base *Concrete* que foram utilizadas as sete variáveis independentes e uma dependente, como dantes já explicado, também, para iniciar as análises preliminares, plota-se os dados brutos nos gráficos da Figura 13, os quais servem para analisar as oscilações no *dataset* ao longo das instâncias, e possibilitando também, às vezes, perceber tendências ou valores abruptos, tidos como anomalias.

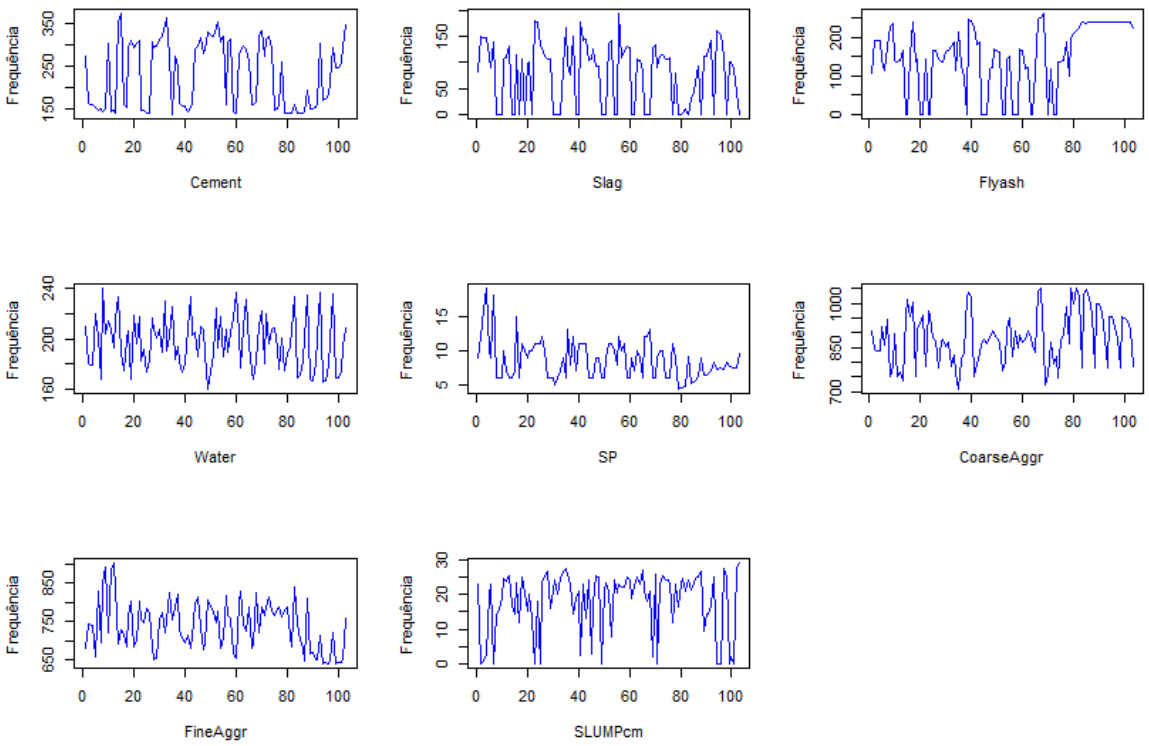

Figura 13: Gráfico da distribuição dos dados brutos da base *Concrete*.

Ao analisar os gráficos da Figura 13, percebe-se a presença de valores abruptos em todas as variáveis. Quando ocorre tal circunstância, as medidas de sensibilidade (MAE, RMSE) são muito influenciadas por tais valores, pois caso o valor faltante esteja próximo deles, a estimativa do valor a ser imputado será fortemente determinado por tais valores. A seguir, na Tabela 10, é apresentada as estatísticas descritivas ou medidas resumos deste *dataset*.

Tabela 10: Estatísticas descritivas para a base *Concrete*.

Métricas	Cement	Slag	Flyash	Water	SP	CoarseAggr	FineAggr	SLUMPcm
Mínimo	137	0	0	160	4,4	708	640,6	0
1º Qu.	152	0,05	115,5	180	6	819,5	684,5	14,5
Mediana	248	100	164	196	8	879	742,7	21
Média	229,9	77,97	149	197,2	8,54	884	739,6	18,04
3º Qu.	303,9	125	236	209,5	10	952,8	788	24
Máximo	374	193	260	240	19	1050	902	40,68

Nota: As abreviações, 1º Qu → primeiro quartil, 3º Qu → terceiro quartil, As variáveis (Cement, Slag Flyash, Water, SP, CoarseAggr e FineAggr) → são as variáveis independentes, SLUMPcm → é a variável dependente, a qual contém os dados faltantes.

Passando a analisar a Tabela 10, constata-se que as variáveis *Slag* e *Flyash* são fortes candidatas a possuírem *outliers*. Sendo assim, para se ter maiores evidências, plota-se um simples gráfico de histograma para verificar a distribuição dos dados, conforme é apresentado na Figura 14.

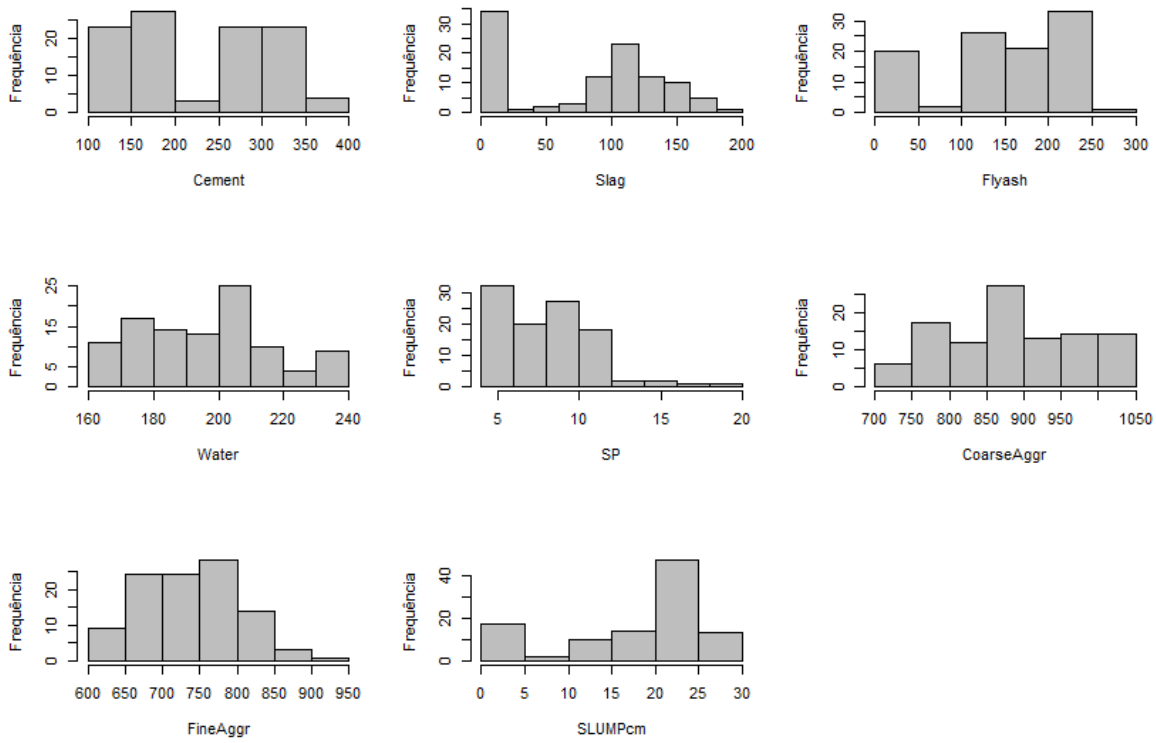

Figura 14: Histograma das variáveis da base *Concrete*.

Conforme visto na Figura 14, a distribuição dos dados para algumas das variáveis independentes, bem como a variável dependente possuem valores concentrados nos extremos, que dão indícios de serem *outliers*, entretanto com o fito de confirmar se estes valores são ou não *outliers*, recorre-se ao teste estatístico de qui-quadrado. Primeiramente, verifica-se quais são os possíveis valores candidatos a serem *outliers*, em cada variável de interesse, conforme os valores postos na Tabela 11.

Tabela 11: Valores que são plausíveis de serem *outliers* para a base *Concrete*.

Cement	Slag	Flyash	Water	SP	CoarseAggr	FineAggr	SLUMPcm
137	0	260	160	4,4	1049,9	640,6	0
374	193	0	240	19	708	902	40,681

Ao fazer uma breve análise na Tabela 11, observam-se os valores extremos, para cada variável em análise, ou seja, o menor e o maior valor respectivamente, que podem ou não ser um *outlier*. Para confirmar tais suspeitas, passa-se a aplicar o teste de qui-quadrado, conforme apresentado na Tabela 12.

Tabela 12: Teste de qui-quadrado para a base *Concrete*.

	Cement	Slag	Flyash	Water	SP	CoarseAggr	FineAggr	SLUMPcm
p-valor	0,06771	0,057	0,0811	0,034	0	0,0605	0,01035	0,01246

Ao analisar os valores da Tabela 12, contata-se que as variáveis (Cement, Slag, Flyash e CoarseAggr) tem p-valores maiores que o nível de significância estabelecido, que é 0,05, porém para esta análise considera-se que tais valores não são convincentes para a não rejeição da hipótese nula, visto que os valores do p-valor estão bem próximo do valor do nível de significância, logo, aceitamos a hipótese alternativa, de que os dados não possuem *outliers*.

Dado que partiu-se da suposição que o algoritmo EM, foi modelado por uma distribuição *normal*, passa-se a analisar a normalidade dos dados; análise esta que pode ser inicialmente feita através do gráfico de probabilidade normal, o qual está na Figura 15.

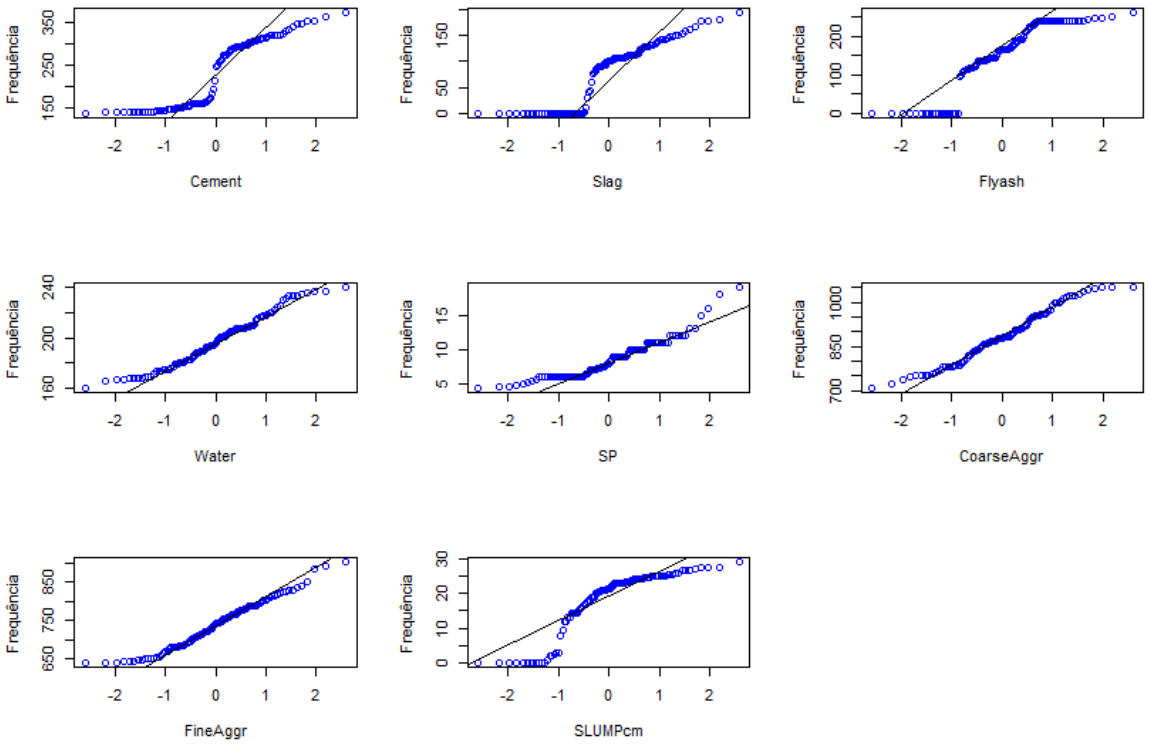

Figura 15: Gráfico de probabilidade normal para a base *Concrete*.

Para os gráficos da Figura 15, os quais se referem à distribuição de percentis acumulados, que caso os valores plotados sigam o padrão de uma reta, ou se aproxime de uma, contata-se visualmente que há evidências da variável aleatória em análise ter uma distribuição que se aproxime da normal. Porém, para a atual situação, verifica-se que os dados em algum momento se afastam muito da reta, logo podemos inferir que estes não seguem uma distribuição normal.

Outra maneira de se testar a normalidade dos dados é através do teste para normalidade de Shapiro-Wilk, que é específico para tal finalidade, o qual tem como resultados os valores apresentados na Tabela 13.

Tabela 13: Teste de normalidade de Shapiro-Wilk para os dados *Concrete*.

	Cement	Slag	Flyash	Water	SP	CoarseAggr	FineAggr	SLUMPcm
p-valor	2,9E-09	2E-08	2E-08	0,0119	0	0,02935	0,01519	8,15E-09

Os valores do p-valor apresentados na Tabela 13, foram comparados ao nível de significância de 5% (0,05), a fim de não rejeitar ou rejeitar a hipótese de nulidade do referido

teste (H_0: Os dados seguem uma distribuição Normal). Constata-se que, para todas as variáveis analisadas, nenhuma delas segue uma distribuição normal. Tal situação evidencia que não partir do pressuposto de modelar o algoritmo EM via uma distribuição normal não é uma premissa forte para a análise destes dados. Além disso, apesar de ter havido *outliers* em apenas uma variável independente, as demais apresentam valores abruptos, fato este que torna ainda mais difícil a modelagem via este algoritmo, e que influenciam fortemente os valores estimados. Na próxima seção será analisada a base de dados *Parkinson.*

5.2.4 Análise Preliminar da Base de dados Parkinson

Seguindo os mesmos passos das análises anteriores, para a base *Parkinson* inicia-se as análises plotando-se os dados brutos nos gráficos da Figura 16, os quais servem para analisar as oscilações no *dataset* ao longo das instâncias, e possibilitando também, às vezes, perceber tendências ou valores abruptos, tidos como anomalias.

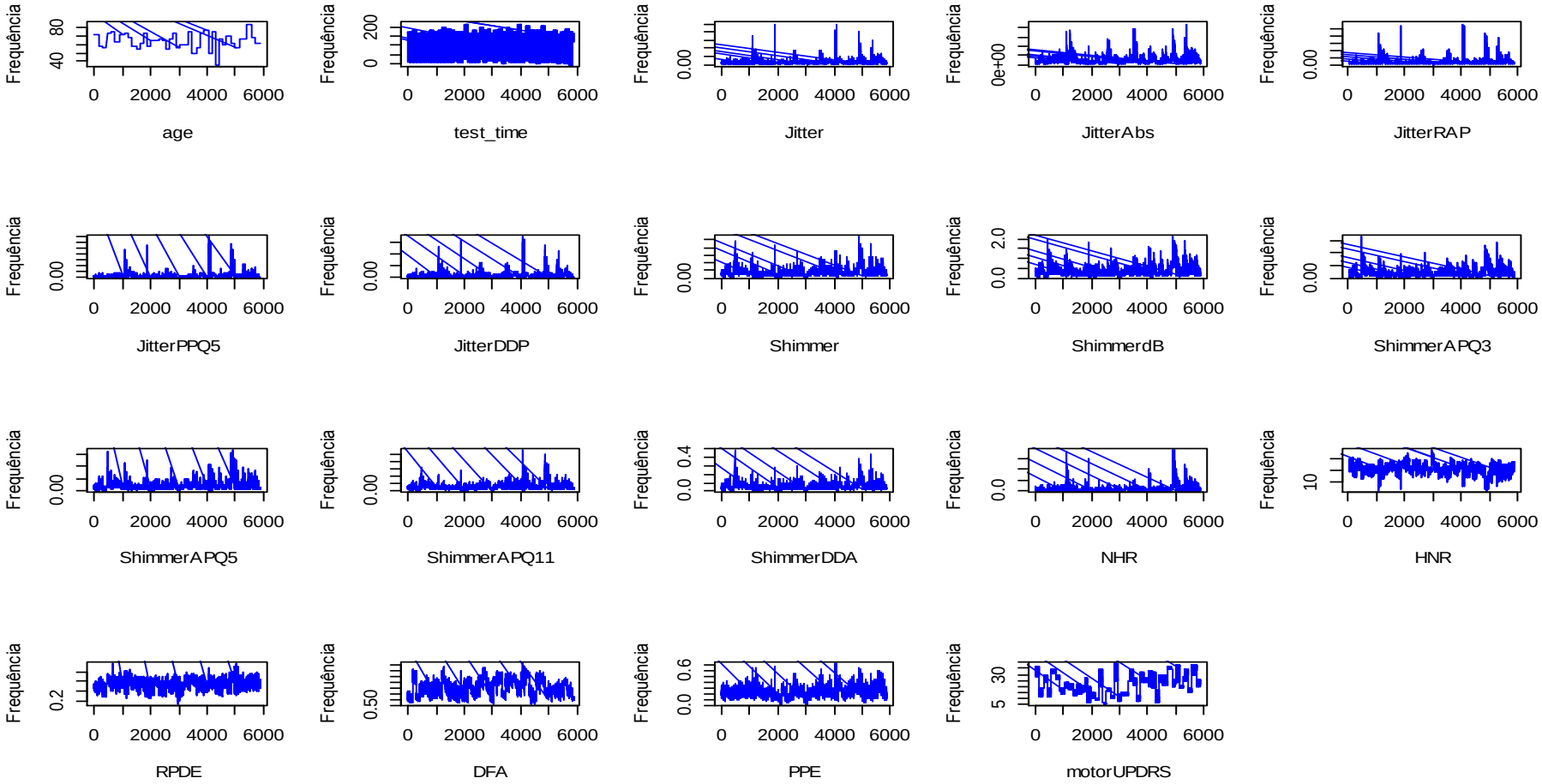

Figura 16: Gráfico da distribuição dos dados brutos da base *Parkinson*.

Ao analisar os gráficos da Figura 16, percebe-se a presença de valores abruptos em todas as variáveis. Quando ocorre tal circunstância, as medidas de sensibilidade (MAE, RMSE) são muito influenciadas por tais valores, pois caso o valor faltante esteja próximo deles, a estimativa do valor a ser imputado será fortemente determinado por tais valores. A seguir, na Tabela 14, são apresentadas as estatísticas descritivas ou medidas resumos da base Parkinson.

Tabela 14: Estatísticas descritivas para a base *Parkinson*.

Métricas	**age**	**testtime**	**Jitter %**	**JitterAbs**	**JitterRAP**	**JitterPPQ5**	**JitterDDP**	**Shimmer**	**ShimmerdB**	**ShimmerAPQ3**
Mínimo	36	-4,262	0,00083	2,25E-06	0,00033	0,00043	0,00098	0,00306	0,026	0,00161
1º Qu.	58	46,85	0,00358	2,24E-05	0,00158	0,00182	0,00473	0,01912	0,175	0,00928
Mediana	65	91,52	0,0049	3,45E-05	0,00225	0,00249	0,00675	0,02751	0,253	0,0137
Média	64,8	92,86	0,006154	4,40E-05	0,002987	0,003277	0,008962	0,03404	0,311	0,01716
3º Qu.	72	138,4	0,0068	5,33E-05	0,00329	0,00346	0,00987	0,03975	0,365	0,02058
Máximo	85	215,5	0,09999	4,46E-04	0,05754	0,06956	0,1726	0,2686	2,107	0,1627

Métricas	**ShimmerAPQ5**	**ShimmerAPQ11**	**ShimmerDDA**	**NHR**	**HNR**	**RPDE**	**DFA**	**PPE**	**motorUPDRS**
Mínimo	0,00194	0,00249	0,00484	0,000286	1,659	0,151	0,514	0,02198	-5,28
1º Qu.	0,01079	0,01566	0,02783	0,01096	19,41	0,4698	0,5962	0,1563	15
Mediana	0,01594	0,02271	0,04111	0,01845	21,92	0,5422	0,6436	0,2055	20,9
Média	0,02014	0,02748	0,05147	0,03212	21,68	0,5415	0,6532	0,2196	21,28
3º Qu.	0,02376	0,03272	0,06174	0,03146	24,44	0,614	0,7113	0,2645	27,56
Máximo	0,167	0,2755	0,488	0,7483	37,88	0,9661	0,8656	0,7317	45,09

Nota: As abreviações, 1º Qu → primeiro quartil, 3º Qu → terceiro quartil, A variável motorUPDRS → é a variável dependente, a qual conterá os dados faltantes. As demais variáveis são todas as independentes.

Passando a analisar a Tabela 14, constata-se que a maioria delas são fortes candidatas a possuírem *outliers*. Sendo assim, para se ter maiores evidências, plota-se um simples gráfico de histograma para verificar a distribuição dos dados, conforme é apresentado na Figura 17.

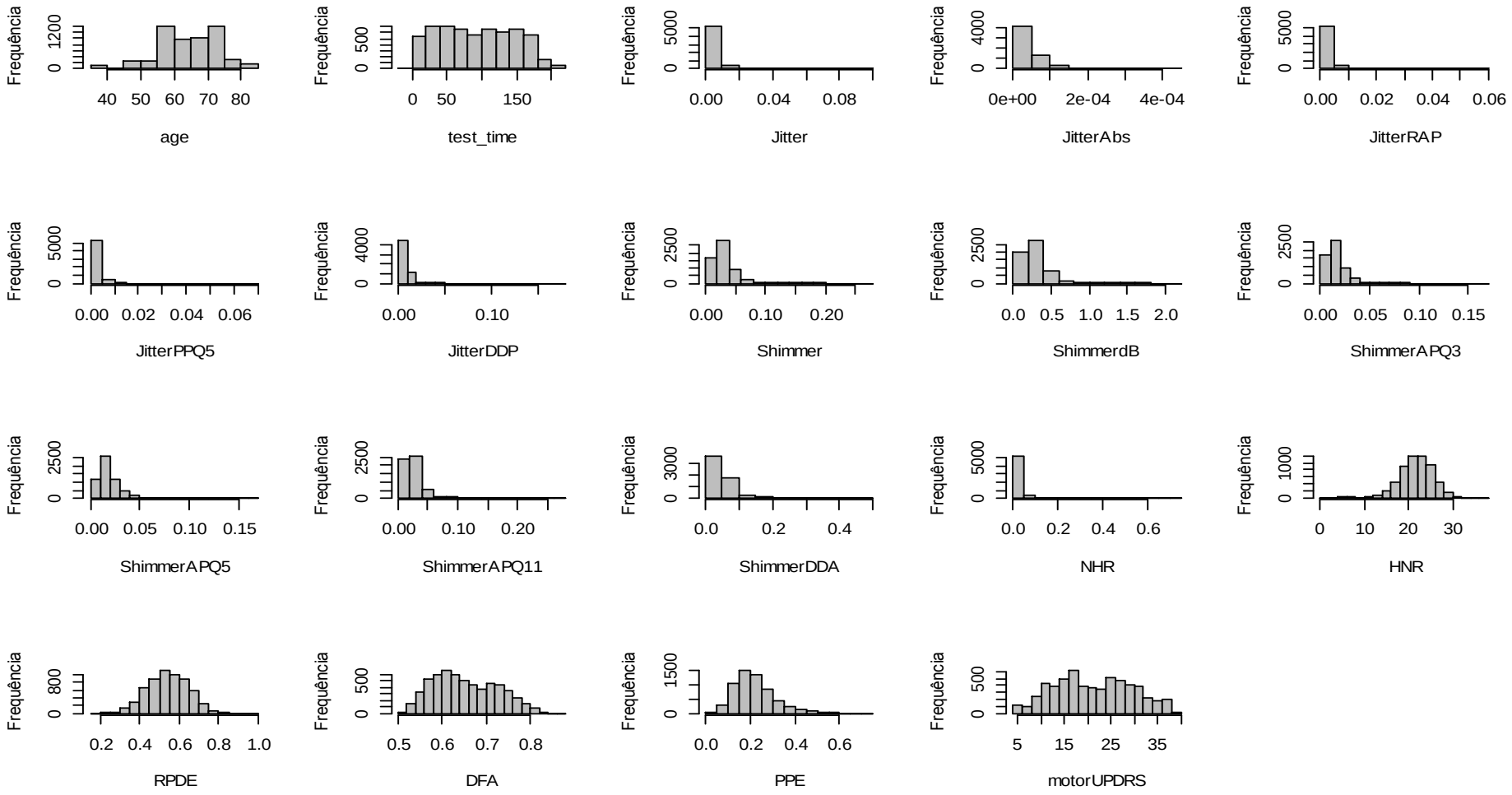

Figura 17: Histograma das variáveis da base *Parkinson*.

Conforme visto na Figura 17, a distribuição dos dados para algumas das variáveis independentes possui valores concentrados nos extremos, que dão indícios de serem *outliers*, entretanto com o objetivo de confirmar se estes valores são ou não *outliers*, recorre-se ao teste estatístico de qui-quadrado. Frise-se também que, a variável *testtime* apresenta o comportamento de uma distribuição uniforme, já as variáveis HNR, RPDE, DFA e PPE apresentam comportamento semelhante a uma distribuição normal. O próximo passo é verificar quais são os possíveis valores candidatos a serem *outliers*, em cada variável de interesse, conforme os valores postos na Tabela 15.

Tabela 15: Valores plausíveis de serem *outliers* da base *Parkinson*.

age	testtime	Jitter %	JitterAbs	JitterRAP	JitterPPQ5	JitterDDP	Shimmer	ShimmerdB	ShimmerAPQ3
36	-4,2625	0,00083	0,00000225	0,00033	0,00043	0,00098	0,00306	0,026	0,00161
85	215,49	0,09999	0,00044559	0,05754	0,06956	0,17263	0,26863	2,107	0,16267
ShimmerAPQ5	**ShimmerAPQ11**	**ShimmerDDA**	**NHR**	**HNR**	**RPDE**	**DFA**	**PPE**	**motorUPDRS**	
0,00194	0,00249	0,00484	0,000286	37,875	0,15102	0,51404	0,021983	-5,279637	
0,16702	0,27546	0,48802	0,74826	1,659	0,96608	0,8656	0,73173	45,0873504	

Ao fazer uma análise na Tabela 15, observam-se os valores extremos, para cada variável em análise, ou seja, o menor e o maior valor respectivamente, que podem ou não ser um *outlier*. Para confirmar tais suspeitas, passa-se a aplicar o teste de qui-quadrado, conforme apresentado na Tabela 16.

Tabela 16: Teste de qui-quadrado para a base *Parkinson*.

	age	testtime	Jitter %	JitterAbs	JitterRAP	JitterPPQ5	JitterDDP	Shimmer	ShimmerdB	ShimmerAPQ3
p-valor	0,001093	0,02177	2,2E-16	0,2456	2,2E-16	2,2E-16	2,2E-16	2,2E-16	6,217E-15	2,2E-16
	ShimmerAPQ5	**ShimmerAPQ11**	**ShimmerDDA**	**NHR**	**HNR**	**RPDE**	**DFA**	**PPE**	**motorUPDRS**	
p-valor	2,2E-16	2,2E-16	2,2E-16	2,2E-16	3,077E-06	2,615E-05	0,002743	2,178E-08	0,001099	

Ao analisar os valores da Tabela 16, verifica-se que apenas a variável *JiterAbs* tem p-valor maior que o nível de significância estabelecido, que é 0,05, o que nos conduz a não rejeitar a hipótese nula, logo, não rejeita-se a hipótese de que esta variável possui *outliers*.

Dado que partiu-se da suposição que o algoritmo EM, foi modelado por uma distribuição *normal*, passa-se a analisar a normalidade dos dados; análise esta que pode ser inicialmente feita através do gráfico de probabilidade normal, o qual está na Figura 18.

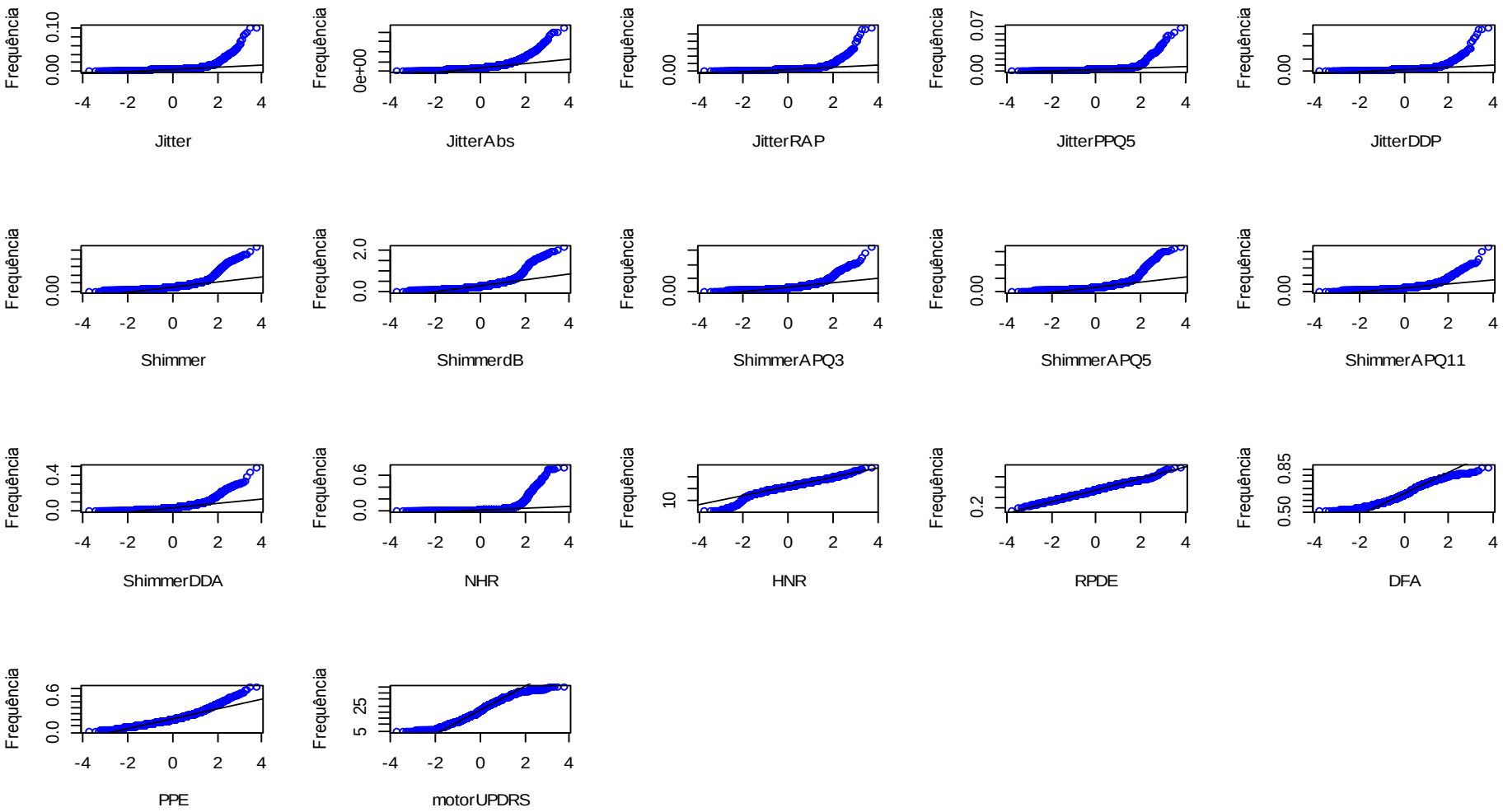

Figura 18: Gráficos de probabilidade normal para a base *Parkinson*.

Para os gráficos da Figura 18, os quais se referem à distribuição de percentis acumulados, que caso os valores apresentados sigam o padrão de uma reta, ou se aproxime de uma, contata-se visualmente que há evidências da variável aleatória em análise ter uma distribuição que se aproxime da normal. Porém, para a atual situação, verifica-se que os dados em algum momento se afastam muito da reta, logo podemos inferir que estes não seguem uma distribuição normal.

O próximo passo é analisar a normalidade dos dados, através do teste de Shapiro-Wilk, visto que partiu-se da premissa que o algoritmo EM, foi modelado por uma distribuição *normal*. Os valores do citado teste estão na Tabela 17.

Tabela 17: Teste de normalidade de Shapiro-Wilk para os dados *Parkinson*.

	age	testtime	Jitter%	JitterAbs	JitterRAP	JitterPPQ5	JitterDDP	Shimmer	ShimmerdB	ShimmerAPQ3
p-valor	2,20E-16	2,20E-16	2,20E-16	2,20E-16	2,20E-16	2,20E-16	2,20E-16	2,20E-16	2,20E-16	2,20E-16
	ShimmerAPQ5	ShimmerAPQ11	ShimmerDDA	NHR	HNR	RPDE	DFA	PPE	motorUPDRS	
p-valor	2,20E-16	2,20E-16	2,20E-16	2,20E-16	2,20E-16	1,98E-07	2,20E-16	2,20E-16	2,20E-16	

Os valores do p-valor apresentados na Tabela 17, foram comparados ao nível de significância de 5% (0,05), a fim de não rejeitar ou rejeitar a hipótese de nulidade do referido teste (H_0: Os dados seguem uma distribuição Normal). Constata-se que, para todas as variáveis analisadas, nenhuma delas segue uma distribuição normal. Diante destas circunstâncias, tem-se que partir do pressuposto de modelar o algoritmo EM via uma distribuição normal não é o melhor meio para a análise destes dados. Além disso, apesar de ter havido *outliers* em apenas uma variável independente, as demais apresentam valores abruptos, fato este que torna ainda mais difícil a modelagem via este algoritmo, e que influenciam fortemente os valores estimados.

5.3 Análise dos dados com imputação única

Nesta seção serão analisados todos os quatro *dataset*, pelo viés da imputação única, onde cada estimativa será gerada apenas uma única vez, para substituir o valor faltante, possibilitando assim, a coleta e análise das medidas de sensibilidade (MAE e RMSE). Após alguns experimentos realizados anteriormente, escolheu-se um valor de "123" para servir como semente para o experimento final, garantindo assim, que caso este experimento seja realizado novamente se obtenha os mesmos resultados. A análise das bases de dados seguirá a mesma sequência da seção 5.2, ou seja, a primeira base a ser analisada é a Emulsão, seguida pela *Breast Tissue*, a *Concrete* e por fim a *Parkinson*. Frise-se que nesta fase iniciam-se as análises através do algoritmo EM e da Rede Neural MLP, conforme apresentado nas seções 3.3.3.3.1, 4.5.1 e 4.5.2.

5.3.1 Base de dados Emulsão

A primeira análise foi realizada através do algoritmo EM, e a segunda pela RNA-MLP com todas as funções de ativação propostas, para auferir as medidas de sensibilidade, para todas as taxas de faltantes já citadas anteriormente. Sendo assim, tem-se todas as medidas auferidas pelo viés do algoritmo EM na Tabela 18.

Tabela 18: Medidas de sensibilidade pelo viés do algoritmo EM para a base emulsão.

%Faltantes	Erros		Dados Imputados					Dados Reais				
	MAE	RMSE	D. Pad	Max	Med	Min	Média	D. Pad	Max	Med	Min	Média
5	4,5096	5,2769	4,3597	24,8100	16,2300	10,4389	15,9219					
10	4,0748	4,7134	4,3061	24,8100	16,1500	8,2467	15,6405					
20	2,5772	3,0092	4,1518	24,8100	15,0000	10,8400	15,6961					
30	3,1044	3,8201	4,1832	24,8100	15,7289	8,4014	15,7473					
40	2,8167	3,6704	4,6728	24,7300	16,1500	6,3683	15,5555	4,3038	24,8100	16,1500	10,8400	15,7598
50	2,8601	3,6188	3,9601	23,9287	16,2300	8,4985	15,8019					
60	3,0725	3,8328	3,7108	22,3500	14,4100	6,9698	14,9605					
70	3,2062	4,1198	3,2580	22,3500	14,6948	7,5994	14,6166					

Nota: As abreviações, para os dados imputados, têm os seguintes significados: D.Pad → desvio-padrão dos dados completados, Max → valor máximo dos dados completados, Med → mediana dos dados completados, Min → valor mínimo dos dados completados, Média → média dos dados completados. Quanto às abreviações dos dados reais, estas têm os mesmo significados da dos dados imputados, porém aplicado aos dados reais.

Ao analisar a Tabela 18, onde os valores destacados em vermelho e negrito representam respectivamente, de cima para baixo, o maior valor dos erros MAE e RMSE, e o menor valor destes erros. Para tais medidas de sensibilidade, era de se esperar que os dados imputados que apresentassem menor erro fossem os que contivessem apenas 5% das observações faltantes, porém não foi isso que ocorreu. Desta forma, como o mecanismo causador dos dados omissos é o MAR, sabe-se que as outras variáveis da amostra são extremamente importantes na influência dos dados que foram imputados. Sendo assim, conforme apresentado na seção 5.2.1, na Figura 6, onde foram plotada todas as variáveis, contata-se que apesar desta base não conter *outliers*, ela contém valores discrepantes, que influenciam fortemente no desempenho do algoritmo, conduzindo às estimativas muito enviesadas. Outro agravante, conforme apresentado também na seção 5.2.1, na Figura 8 e Tabela 5 é que o algoritmo EM foi modelado via uma *normal*, porém os dados não seguem esta distribuição, conforme resultado apresentado na Tabela 5 consequentemente, a precisão das estimativas é comprometida. Além do que foi ponderado até agora, há também de se considerar que esta base de dados contém apenas 60 instâncias, logo, 5% desta base corresponde a apenas 3 instâncias, situação esta, a qual, caso haja um valor muito enviesado, consequentemente as medidas de erro não conseguirão suavizar tal discrepância, devido a baixa quantidade de instâncias. Haja visto tal situação, considerada atípica, houve a necessidade de se fazer uma nova análise, a fim de confirmar tal suspeita (se os dados discrepantes foram o fator determinante para a taxas de 5% ter apresentado o maior erro). Desta forma, recalcularam-se as taxas de erros para esta base. Porém, primeiramente, para esta base, escolheu-se manualmente quem seriam os valores faltantes nela, tendo como critério para determinar em que local da distribuição dos dados estes teriam os valores omissos, o seguinte: verificou-se se as outras variáveis não possuíam valores discrepantes em seus 3 vizinhos mais próximos, tanto para cima, como para baixo. Com o referido critério, procurou-se inserir os dados faltantes em um local da distribuição deste, onde os dados fossem o mais homogêneo possível. Como resultado deste experimento tem-se os dados na Tabela 19.

Tabela 19: Comparação (Antes x Depois) para a base emulsão.

Comparação	%Faltantes	Erros				Dados Imputados				
		MAE	%Melhoria	RMSE	%Melhoria	D. Pad	Max	Med	Min	Média
Antes	5	4,5096	37,0331	5,2769	35,5769	4,3597	24,8100	16,2300	10,4389	15,9219
Depois		**2,8395**		**3,3995**		4,3038	24,8100	16,1500	10,8400	15,7598

Ao passo que se analisa a Tabela 19, fica evidente que o critério adotado para inserir os dados faltantes serviu para corroborar que, se os dados omissos estão longe de valores discrepantes, ter-se-á valores previstos mais assertivos, consequentemente, tem-se como resultado valores de medidas de sensibilidade bem menores. Os valores da Tabela 19, que estão em preto, correspondem ao primeiro experimento (Tabela 18), já os que estão destacados em azul e negrito, correspondem aos valores dos erros medidos, após a inserção de dados faltantes manualmente, e os valores da coluna %Melhoria correspondem à melhoria obtida com tal forma de inserção, que para a medida MAE o ganho foi de 37,03% em termos de redução de erro, e para a medida RMSE o ganho foi de 35,57% em termos de redução de erro. Em geral, o valor médio de melhoria foi em torno de 36%, o que pode ser considerado como um ganho significativo na redução dos erros. Outra situação percebida foi que, para esta base de dados analisada, as porcentagens de dados faltantes influenciam menos na determinação do erro final, do que a presença de dados discrepantes. A seguir será analisada esta mesma base de dados pelo viés de Redes Neurais Artificiais MLP.

Tabela 20: Medidas de sensibilidade para RNA-MLP da base emulsão.

% Faltantes	Função	Dados Imputados								Dados Reais				
		MAE	RMSE	%Melhoria	D. Pad	Max	Med	Min	Média	D. Pad	Max	Med	Min	Média
5	LLEMV	**0,4318**	**0,5469**	62,6121	4,3150	24,8100	16,1248	10,8400	15,7390					
10	CLLEMV	1,2348	2,2061	24,8791	4,2470	24,8100	16,1500	10,8800	15,8194					
20	SIGEMV	1,2149	1,4572	85,5713	4,2778	24,8100	15,3008	10,8400	15,6871					
30	CLL	0,8077	1,1895	24,1411	4,3672	24,8100	16,1500	10,8800	15,8576					
40	LLEMV	1,4060	1,9922	11,4452	4,0331	24,7300	15,5871	10,8400	15,8302	4,6618	24,8100	16,1500	10,8400	15,7598
50	CLL	1,2177	1,7535	4,8030	3,9260	22,7400	16,2300	10,9500	15,7886					
60	EMVCLL	**1,8486**	**2,6926**	17,4380	3,7495	22,3500	14,4100	10,8400	15,2301					
70	CLL	1,6903	2,4616	57,1068	3,2814	22,3500	16,1500	10,9500	15,1344					

Nota: LLEMV ← É a função Log-Log na camada inicial e intermediária com a função Log-Log com o método de estimativa de máxima verossimilhança no neurônio de saída da rede (Conforme Quadro 4). CLLEMV ← É a função Complementar Log-Log na camada inicial e intermediária com a função Complementar Log-Log com o método de estimativa de máxima verossimilhança no neurônio de saída da rede (Conforme Quadro 4). SIGEMV ← É a função Sigmoide na camada inicial e intermediária com a função Sigmoide com o método de estimativa de máxima verossimilhança no neurônio de saída da rede (Conforme Quadro 4).CLL ← É a função Complementar Log-Log em todas as camadas (Conforme Quadro 3). EMVCLL ← É a função de ativação com o método de estimativa de máxima verossimilhança em todas as camadas da rede (Conforme Quadro 5).

Ao analisar a Tabela 20, onde as medidas de sensibilidade esperadas (MAE e RMSE), ou seja, com taxa de faltantes de 5% (destacado em vermelho e negrito) foi a que apresentou o menor erro, e com as taxas de 60% (destacado em vermelho e negrito), foi a que apresentou o maior erro, como era de se esperar. Quanto à coluna referente à "%Melhoria", esta medida foi aferida da seguinte forma: como já citado anteriormente, para cada função de ativação e para cada porcentagem de dados faltantes, foram utilizados os três *frameworks*, apresentados nos Quadro 3, 4 e 5. Em seguida, tomou-se como função padrão a TH (Tangente Hiperbólica) e calculou-se a porcentagem de melhoria que cada uma das outras funções conseguiram obter em relação a ela. Sendo assim, é notório que estas novas funções, principalmente quando se utiliza o EMV, apresentam melhores resultados quando comparadas à função clássica TH.

Para uma melhor visualização e assimilação do ganho conseguido, em relação à redução do erro, através das Redes Neurais MLP, a Tabela 21, tem um comparativo entre as medidas para ambas as técnicas.

Tabela 21: Comparação dos erros do algoritmo EM x RNA-MLP para a base emulsão.

% Faltantes	MAE			RMSE		
	EM	RNA	%Melhoria	EM	RNA	%Melhoria
5	4,5096	0,4318	**90,4259**	5,2769	0,5469	**89,6369**
10	4,0748	1,2348	69,6969	4,7134	2,2061	53,1958
20	2,5772	1,2149	52,8596	3,0092	1,4572	51,5759
30	3,1044	0,8077	73,9808	3,8201	1,1895	68,8607
40	2,8167	1,4060	50,0826	3,6704	1,9922	45,7215
50	2,8601	1,2177	57,4231	3,6188	1,7535	51,5444
60	3,0725	1,8486	39,8348	3,8328	2,6926	29,7490
70	3,2062	1,6903	47,2791	4,1198	2,4616	40,2500

Ao observar as medidas de erros, na Tabela 21, constata-se que ao utilizar as Redes Neurais MLP, para imputar dados, esta apresenta consideravelmente melhor desempenho e assertividade do que o algoritmo EM para todas as porcentagens de faltantes, sendo que a maior melhoria obtida foi para a taxa de faltantes de 5%, com 90,42% de melhoria em relação à medida MAE, e 89,63% em relação a medida RMSE. A próxima seção tratará da base de dados *Breast Tissue*.

5.3.2 Base de dados *Breast Tissue*

Nesta seção, a qual irá analisar a base de dados *Breast Tissue*, também utilizou semente com valor "123", sendo que o primeiro experimento foi realizado com o algoritmo EM, e o segundo com a RNA-MLP proposta.

Seguindo a mesma sequência de análise da seção 5.3.1, as medidas de sensibilidade foram auferidas primeiramente via algoritmo EM, as quais estão expostas na Tabela 22.

Tabela 22: Medidas de Sensibilidade para a base *Breast Tissue*.

%Faltantes	Erros		Dados Imputados					Dados Reais				
	MAE	RMSE	D. Pad	Max	Med	Min	Média	D. Pad	Max	Med	Min	Média
5	**26,2686**	**36,9833**	759,7037	2896,5825	454,1082	124,9786	810,7777					
10	30,5774	50,2260	759,3084	2896,5825	439,3578	124,9786	808,9864					
20	36,8056	44,1354	760,5683	2896,5825	454,1082	134,8927	809,2742					
30	33,0774	43,0782	757,2498	2896,5825	439,3578	124,9786	810,2008					
40	33,0474	43,0012	758,8083	2896,5825	459,4063	124,9786	813,1482	763,0191	2896,5825	445,5133	124,9786	810,6381
50	**27,4065**	**35,4075**	759,8303	2851,0638	439,3578	91,5705	807,5062					
60	**45,2962**	**93,0472**	789,3324	3299,7733	424,1444	127,5166	820,5524					
70	39,1329	64,4070	770,0494	3057,0286	436,4734	127,0944	812,8384					

Nota: As abreviações, para os dados imputados e para os dados reais, têm os mesmos significados dos que foram descritos na Tabela 18.

Para a Tabela 22, onde os valores destacados em vermelho e em negrito representam respectivamente, de cima para baixo, representam o menor valor do erro MAE, e o maior valor deste erro. Nesta medida de sensibilidade, constata-se que o menor valor foi conseguido com taxas de dados faltantes de 5%, como era de se esperar, entretanto para a medida de sensibilidade RMSE, houve uma situação inusitada, a qual tem como menor valor de erro medido à uma taxa de 50% de faltantes que foi o erro de 35,4075. Em suma, a taxa de erro para a porcentagem de faltantes de 50% foi similar à taxa de erro para a porcentagem de faltantes de 5%. Esta situação pode ter ocorrido em virtude dos dados faltantes terem sido gerados de forma aleatória, e certamente estes valores omissos, para a taxa de faltantes de 50%, ficaram bem distribuídos ao longo dos dados e longe dos valores discrepantes.

Outra situação atípica observada é quanto os erros medidos à taxa de faltante de 10%, sendo verificado que o valor do MAE é menor em relação às outras porcentagens de dados falantes, com exceção das de omissos de 5% e 50%, porém quando se analisa o RMSE para esta mesma taxa de 10% de omissão, constata-se que tal valor é bem alto, principalmente

quando comparado às outras porcentagens de faltantes. Tal situação é motivada pela própria estrutura desta medida de sensibilidade, que dá mais ênfase (peso) aos valores de erro que destoam muito do real, já que estes são elevados ao quadrado. Sendo assim, certamente ocorreu a situação onde a maior parte dos valores estimados estava próxima de dados muito discrepantes, conforme foi observado nos gráficos da Figura 9, os quais acabaram influenciando nesta medida. A seguir passa-se a analisar esta base pelo viés da RNA-MLP.

Analisando a base Breast Tissue, via a RNA-MLP, seguindo os mesmos passos dos pseudocódigos dos Quadros 3, 4 e 5, tem-se como resultados para as medidas de sensibilidade os seguintes valores, que estão na Tabela 23.

Tabela 23: Medidas de sensibilidade para imputação única via Redes Neurais MLP para a base *Breast Tissue*.

% Faltantes	Função	Dados Imputados								Dados Reais				
		MAE	RMSE	%Melhoria	D. Pad	Max	Med	Min	Média	D. Pad	Max	Med	Min	Média
5	AO	**88,1880**	**97,0989**	94,4970	758,5277	2896,5825	445,5133	124,9786	815,1757	759,8586	2896,5825	445,5133	124,9786	810,6381
10	EMVSIG	119,8770	143,4113	89,5610	758,9170	2896,5825	445,5133	124,9786	817,0976					
20	CLL	145,5419	160,4857	89,1540	745,3972	2896,5825	493,7018	134,8927	833,8251					
30	CLLEMV	135,0605	148,0840	89,6695	756,5147	2896,5825	448,3908	124,9786	838,0795					
40	CLLEMV	146,8996	178,4546	88,2689	772,5461	2896,5825	508,5404	124,9786	870,1876					
50	LLEMV	**89,5367**	**106,4106**	93,0203	738,6768	2701,9771	445,5133	124,9786	822,2071					
60	EMVSIG	96,2515	113,1779	5,2945	739,4672	2896,5825	384,6662	162,5109	799,1688					
70	EMVSIG	114,3430	139,7314	91,2484	792,9726	2896,5825	374,1761	180,6096	821,6712					

Nota: AO ← É a função Aranda Ordaz em todas as camadas da rede (conforme Quadro 3). EMVSIG ← É a função Sigmoide com o método de estimativa de máxima verossimilhança em todas as camadas da rede (Conforme Quadro 5). CLL ← É a função Complemento Log-Log em todas as camadas da rede (conforme Quadro 3). CLLEMV ← É a função Complementar Log-Log na camada inicial e intermediária, porém com o método de estimativa de máxima verossimilhança na camada de saída da rede (conforme Quadro 5). LLEMV ← É a função de ativação Log-Log na camada inicial e intermediária, porém com o método de estimativa de máxima verossimilhança na camada de saída da rede (conforme Quadro 5).

Passando a analisar a Tabela 23, onde as medidas de sensibilidade esperadas (MAE e RMSE), ou seja, com taxa de faltantes de 5% foi a que apresentou o menor erro; porém uma situação inusitada ocorreu para este *dataset*, que foi na medida quanto à taxa de 50% de faltantes, a qual apresentou valores de erro bem próximos da taxa de faltantes de 5%, situação semelhante quando se utilizou o algoritmo EM. Quanto à coluna referente à "%Melhoria", esta medida foi aferida da seguinte forma: Como já citado anteriormente, para cada função de ativação e para cada porcentagem de dados faltantes, foram utilizados os três frameworks, apresentados nos Quadro 3, 4 e 5. Em seguida, tomou-se como função padrão a TH (Tangente Hiperbólica) e calculou-se a porcentagem de melhoria que cada uma das outras funções conseguiram obter em relação a ela. Sendo assim, é notório que estas novas funções, principalmente quando se utiliza o EMV, apresentam melhores resultados quando comparadas à função clássica TH.

Para uma melhor visualização e assimilação dos erros auferidos com o algoritmo EM e com a Rede Neural MLP, tem-se na Tabela 24, um comparativo entre as medidas para ambas as técnicas.

Tabela 24: Comparação dos erros do algoritmo EM x RNA-MLP para a base *Breast Tissue*.

% Faltantes	MAE			RMSE		
	EM	RNA	%Melhoria	EM	RNA	%Melhoria
5	26,2686	88,188	-235,7164	36,9833	97,0989	-162,5480
10	30,5774	119,877	-292,0445	50,226	143,41113	-185,5317
20	36,8056	145,5419	-295,4341	44,1354	160,4857	-263,6213
30	33,0774	135,0605	-308,3166	43,0782	148,084	-243,7562
40	33,0474	146,8996	**-344,5118**	43,0012	178,4546	**-314,9991**
50	27,4065	89,5367	-226,6988	35,4075	106,4106	-200,5312
60	45,2962	96,2515	-112,4935	93,0472	113,1779	-21,6349
70	39,1329	114,343	-192,1915	64,407	139,7314	-116,9506

Ao observar as medidas de erros, na Tabela 24, constata-se que ao utilizar o algoritmo RNA, para imputar os dados nesta base, este apresenta consideravelmente um desempenho inferior ao algoritmo EM para todas as porcentagens de faltantes, sendo que o pior desempenho ocorreu para a taxa de faltantes de 40%, com uma perda de 344,51% de melhoria em relação à medida MAE, e 314,999% em relação à medida RMSE. Nesta situação infere-se que a tentar moldar esta base via RNA-MLP não é uma boa opção. A próxima seção tratará da base de dados *Concrete*.

5.3.3 Base de dados *Concrete*

Para a base *Concrete*, a qual seguirá os mesmos procedimentos anteriores, também foram feitas as primeiras análises via o algoritmo EM, para o qual foram verificadas as medidas de sensibilidade, obtendo como resultado final para as medidas auferidas os valores pontuados na Tabela 25.

Tabela 25: Medidas de Sensibilidade pelo viés do algoritmo EM para a base *Concrete*.

%Faltantes	Erros		Dados Imputados					Dados Reais				
	MAE	RMSE	D. Pad	Max	Med	Min	Média	D. Pad	Max	Med	Min	Média
5	**11,6605**	**15,0403**	9,0176	40,6809	21,0000	0,0000	18,0375	8,7508	29,0000	21,3750	0,0000	18,0485
10	**4,9205**	**5,5658**	8,4154	29,0000	21,1250	0,0000	18,1712					
20	8,1321	10,4670	8,2424	29,0000	20,2500	0,0000	17,5532					
30	8,4119	9,5906	8,1827	32,2197	21,4699	0,0000	18,6109					
40	8,0540	10,0943	9,2636	37,9808	20,7500	-2,7345	17,7275					
50	8,5227	10,3783	8,7085	39,3686	18,7095	-0,3842	17,5567					
60	**6,9739**	**8,7493**	9,3118	40,2726	20,7500	-8,3431	18,8982					
70	8,1512	10,6574	6,8247	34,1713	20,4979	0,0000	20,0235					

Nota: As abreviações, para os dados imputados e para os dados reais, têm os mesmos significados do que foi descrito na Tabela 17.

Na Tabela 25, onde os valores destacados em vermelho e negrito representam respectivamente, de cima para baixo, o maior valor dos erros MAE e RMSE, e o menor valor destes erros respectivamente. Para a referida medida de sensibilidade, verifica-se que os menores valores foram conseguidos com taxas de dados faltantes de 10%, entretanto para a taxa de valores omissos de 5%, que era de se esperar que apresentasse a menor taxa de erro, o que não ocorreu, verifica-se que o valor do MAE e do RMSE é bem maior em relação às outras porcentagens de faltantes. Tal situação é motivada pela própria estrutura destas medidas de sensibilidade, que dá mais ênfase (peso) aos valores de erro que destoam muito do real. Além disso, como o tamanho da amostra também é pequena, esta influencia em tais valores, ao passo que, para este conjunto de dados tem-se apenas 103 instâncias, logo 5% é equivalente a apenas cinco amostras, desta forma, caso estas estejam próximas de valores abruptos, certamente terá como resultado valores de erros altíssimos, pois ao encontrar o erro médio, terá como fator divisor apenas o tamanho de 5 amostras, impossibilitando reduzir o impacto de valores discrepantes no cálculo do erro. Tal situação é similar ao que ocorreu com a base de dados Emulsão. Quanto aos valores com taxa de omissos de 60% (que está destacada na cor azul e em negrito), esta também apresentou uma situação atípica, neste

conjunto de dados, a qual tem como resultados valores dos erros próximos da taxa de omissão de 10%, que foi a menor entre todas elas, similar ao que ocorreu na base de dados *Breast Tissue*. Tal situação origina-se do fato dos dados faltantes terem sido gerados de forma aleatória, e certamente estes valores omissos ficaram bem distribuídos ao longo dos dados, e longe dos *outliers* ou dados discrepantes. Além disso, mesmo que tenham tido valores próximos de dados muito discrepantes, as medidas dos erros foram suavizadas, dada a maior quantidade de amostras, que para o presente caso seriam 62 amostras que entrariam no denominador como fator de divisão, para se encontrar o erro médio das medidas de sensibilidade.

Em virtude desta situação, considerada atípica, onde as medidas de sensibilidade para a taxa de faltantes de 5% foram as que apresentaram o maior valor, houve a necessidade de se fazer uma nova análise, a fim de confirmar tal suspeita (se os dados discrepantes foram o fator determinante para a uma taxa de 5% ter apresentado o maior erro). Desta forma, recalcularam-se as taxas de erros para esta base. Porém, primeiramente, para esta base, escolheu-se manualmente quem seriam os valores faltantes nela, tendo como critério para determinar em que local da distribuição dos dados estes teriam os valores omissos, o seguinte: verificou-se se as outras variáveis não possuíam valores discrepantes em seus 3 vizinhos mais próximos, tanto para cima, como para baixo. Com o referido critério, procurou-se inserir os dados faltantes em um local da distribuição deste, onde os dados fossem o mais homogêneo possível. Como resultado deste experimento tem-se os dados na Tabela 26.

Tabela 26: Comparação (Antes x Depois) para a base *Concrete*.

Comparação	%Faltantes	Erros				Dados Imputados				
		MAE	%Melhoria	RMSE	%Melhoria	D. Pad	Max	Med	Min	Média
Antes	5	11,6605	38,0124	15,0403	41,6800	9,0176	40,6809	21,0000	0,0000	18,0375
Depois		**7,2281**		**8,7715**		8,7546	29,0000	21,0000	0,0000	17,7333

Ao passo que se analisa a Tabela 26, fica evidente que o critério adotado para inserir os dados faltantes serviu para corroborar que, se os dados omissos estão longe de valores discrepantes, ter-se-á valores previstos mais assertivos, consequentemente, tem-se como resultado valores de medidas de sensibilidade bem menores. Os valores da Tabela 26, que estão em preto, correspondem ao primeiro experimento (Tabela 25), já os que estão destacados em azul e negrito, correspondem aos valores dos erros medidos, após a inserção de dados faltantes manualmente, e os valores da coluna %Melhoria correspondem à melhoria

obtida com tal forma de inserção, que para a medida MAE o ganho foi de 38,02% em termos de redução de erro, e para a medida RMSE o ganho foi de 41,68% em termos de redução de erro. Em geral, o valor médio de melhoria foi em torno de 39,84%, o que pode ser considerado como um ganho significativo na redução dos erros. Outra situação percebida foi que, para esta base de dados analisada, as porcentagens de dados faltantes influenciam menos na determinação do erro final, do que a presença de valores discrepantes. A seguir será analisada esta mesma base de dados pelo viés de Redes Neurais Artificiais MLP.

Ao analisar esta base de dados, *Concrete*, adotando as abordagens propostas nas seções 4.5.1 e 4.5.2, obteve-se como resultados para as medidas de sensibilidade, os valores que estão na Tabela 27.

Tabela 27: Medidas de sensibilidade para imputação única via Redes Neurais MLP para a base *Concrete*.

%Faltantes	Função	Dados Imputados								Dados Reais				
		MAE	RMSE	%Melhoria	D.Pad	Max	Med	Min	Média	D.Pad	Max	Med	Min	Média
5	TH	3,7201	4,0939	0,0000	8,7599	29,0000	21,2999	0,0000	18,1098	8,8245	29,0000	21,3750	0,0000	18,0485
10	LL	5,7723	6,7058	24,9029	8,6776	29,0000	21,0000	0,0000	18,0546					
20	LL	5,2356	6,8074	18,3079	8,7005	29,0000	20,8939	0,0000	17,6639					
30	SIG	4,0122	5,6414	34,4680	8,7545	29,0000	20,5440	0,0000	17,8522					
40	EMVSIG	5,0549	7,4418	6,6057	8,6001	29,0000	20,5174	0,0000	17,4109					
50	CLLEMV	4,1690	6,3463	32,2534	7,6203	27,5000	22,3310	0,0000	19,3685					
60	CLL	4,8750	6,9800	22,4568	7,4435	29,0000	22,0470	0,0000	18,5725					
70	EMVCLL	4,8230	7,0785	28,9458	6,9804	29,0000	21,7866	0,0000	19,4651					

Nota: TH ← É a função Tangente Hiperbólica em todas as camadas da rede (conforme Quadro 3). LL ← É a função Log-Log em todas as camadas da rede (conforme Quadro 3). SIG ← É a função Sigmoide em todas as camadas da rede (conforme Quadro 3). EMVSIG ← É a função Sigmoide com o método de estimativa de máxima verossimilhança em todas as camadas da rede (conforme Quadro 5). CLLEMV ← É a função de ativação Complemento Log-Log na camada inicial e intermediária, porém com o método de estimativa de máxima verossimilhança na camada de saída da rede (conforme Quadro 4). CLL ← É a função de ativação Complemento Log-Log em todas as camadas (conforme Quadro 1). EMVCLL ← É a função de ativação Complemento Log-Log com o método de estimativa de máxima verossimilhança em todas as camadas da rede (conforme Quadro 5).

Passando a analisar a Tabela 27, onde as medidas de sensibilidade esperadas (MAE e RMSE), ou seja, com taxa de faltantes de 5% foi a que apresentou o menor erro. Quanto à coluna referente à "%Melhoria", esta medida foi aferida da seguinte forma: como já citado anteriormente, para cada função de ativação e para cada porcentagem de dados faltantes, foram utilizados os três frameworks, apresentados nos Quadro 3, 4 e 5. Em seguida, tomou-se como função padrão a TH (Tangente Hiperbólica) e calculou-se a porcentagem de melhoria que cada uma das outras funções conseguiram obter em relação a ela. Sendo assim, é notório que estas novas funções, principalmente quando se utiliza o EMV, apresentam melhores resultados quando comparadas à função clássica TH.

Para uma melhor visualização e assimilação dos erros auferidos com o algoritmo EM e com a Rede Neural MLP, na Tabela 28, tem um comparativo entre as medidas para ambas as técnicas.

Tabela 28: Comparação entre as medidas de sensibilidade via as duas técnicas para a base *Concrete*.

%Faltantes	MAE			RMSE		
	EM	RNA	%Melhoria	EM	RNA	%Melhoria
5	11,6605	3,7201	**68,0969**	15,0403	4,0939	**72,7801**
10	4,9205	5,7723	**-17,3127**	5,5658	6,7058	**-20,4831**
20	8,1321	5,2356	35,6183	10,4670	6,8074	34,9633
30	8,4119	4,0122	52,3032	9,5906	5,6414	41,1778
40	8,0540	5,0549	37,2371	10,0943	7,4418	26,2772
50	8,5227	4,1690	51,0831	10,3783	6,3463	38,8508
60	6,9739	4,8750	30,0961	8,7493	6,9800	20,2222
70	8,1512	4,8230	40,8302	10,6574	7,0785	33,5818

Ao observar as medidas de erros, na Tabela 28, constata-se que ao utilizar a Rede Neural MLP, para imputar dados, esta apresenta consideravelmente melhor desempenho e assertividade do que o algoritmo EM, exceto para a taxa de faltantes de 10%. A maior melhoria obtida foi para a taxa de faltantes de 5%, com 68,09% de melhoria em relação à medida MAE, e 72,78% em relação a medida RMSE. A próxima seção tratará da base de dados *Parkinson*.

5.3.4 Base de dados Parkinson

Para a base Parkinson, que teve suas características detalhadas no início do capítulo 5, e suas análises preliminares, na seção 5.2.4, que foram as medidas resumo, verificação de normalidade e presença de *outliers*. Tem nesta seção o objetivo de analisar as medidas de sensibilidade; sendo que, primeiramente utilizou-se o algoritmo EM para tal objetivo, e em seguida a RNA-MLP.

Como anteriormente citado, primeiramente analisou-se a base pelo viés do algoritmo EM, tendo como resultados os valores que estão expostos na Tabela 29.

Tabela 29: Medidas de Sensibilidade para a base *Parkinson*.

%Faltantes	Erros		Dados Imputados					Dados Reais				
	MAE	RMSE	D. Pad	Max	Med	Min	Média	D. Pad	Max	Med	Min	Média
5	8,2231	10,2023	8,1362	45,0874	20,8960	-5,2796	21,2779					
10	7,9235	9,9345	8,1027	44,1359	20,8710	-3,7440	21,2781					
20	8,0999	10,2194	8,1984	51,2855	21,1070	-5,1825	21,3400					
30	7,9836	9,9961	8,1301	49,8597	21,1920	-5,0981	21,3502	8,1293	39,5110	20,8710	5,0377	21,2962
40	8,2218	10,2374	8,1713	56,8068	21,2320	-3,5689	21,3384					
50	8,2544	10,3481	8,1511	51,2974	21,0463	-7,0259	21,2534					
60	8,1555	10,2419	8,2338	48,8976	20,9815	-7,3610	21,2454					
70	7,9759	10,1116	8,3425	83,0182	21,0080	-5,4814	21,2534					

Nota: As abreviações, para os dados imputados e para os dados reais, têm os mesmos significados do que foi descrito na Tabela 17.

Na Tabela 29, onde os valores destacados em vermelho representam respectivamente, de cima para baixo, o menor valor dos erros MAE e RMSE, e o maior valor destes erros respectivamente. Para a referida medida de sensibilidade, verifica-se que os menores valores foram conseguidos com taxas de dados faltantes de 10%, entretanto para a taxa de valores omissos de 5%, que era de se esperar que apresentasse a menor taxa de erro, situação a qual não ocorreu, verifica-se que o valor do MAE e do RMSE é bem próximo das porcentagens de faltantes de 50%. Tal situação é motivada pela própria estrutura destas medidas de sensibilidade, que dá mais ênfase (peso) aos valores de erro que destoam muito do real, já que estes são elevados ao quadrado. Além disso, como o tamanho da amostra para os faltantes de 5%, também é menor, esta influencia em tais valores, desta forma, caso estas estejam próximas de valores abruptos, certamente terá como resultado valores de erros maiores, pois ao calcular o erro médio, terá como fator divisor um tamanho menor de amostras,

impossibilitando reduzir o impacto de valores discrepantes no cálculo do erro. Uma situação que pode ser percebida nesta base, que se diferencia das outras, é que não houve oscilações entre as medidas de sensibilidade, independente da porcentagem de dados faltantes, ou seja, tanto faz esta base ter 5% de dados faltantes, como ter 70%, pois os resultados dos erros serão praticamente os mesmos. A seguir será analisada esta mesma base de dados pelo viés de Redes Neurais Artificiais MLP.

Passando-se a analisar a base de dados *Parkinson*, adotando as abordagens propostas nas seções 3.3.3.3.1 e 4.5, obteve-se como resultados para as medidas de sensibilidade, os valores que estão na Tabela 30.

Tabela 30: Medidas de sensibilidade para imputação única via Redes Neurais MLP para a base *Parkinson*.

% Faltantes	Função	Dados Imputados								Dados Reais				
		MAE	RMSE	%Melhoria	D. Pad	Max	Med	Min	Média	D. Pad	Max	Med	Min	Média
5	CLLEMV	**8,2946**	**9,3964**	26,8099	7,7725	39,5110	21,4670	5,0377	21,4705					
10	TH	7,4667	8,6489	0,0000	7,5679	39,5110	20,7758	5,0377	21,4154					
20	CLLEMV	7,4133	8,5952	32,5794	7,1675	39,5110	22,2741	5,0377	21,6307					
30	CLLEMV	7,3338	8,5081	26,5825	6,6521	39,5110	22,2740	5,0377	21,6775					
40	TH	7,0514	8,2479	0,0000	6,1619	39,5110	21,3700	5,0377	21,5375	8,1293	39,5110	20,8710	5,0377	21,2962
50	CLLEMV	6,6566	7,9254	37,9475	5,7330	39,5110	21,4964	5,0377	21,2384					
60	CLLEMV	**6,6483**	**7,8636**	34,9487	5,2434	39,5110	22,2732	5,0377	21,5584					
70	CLLEMV	6,7753	7,9846	3,5630	4,4994	39,5110	22,2726	5,0377	21,8335					

Nota: CLLEMV ← É a função de ativação Complemento Log-Log na camada inicial e intermediária, porém com o método de estimativa de máxima verossimilhança na camada de saída da rede (conforme Quadro 4). TH ← É a função Tangente Hiperbólica em todas as camadas da rede (conforme Quadro 3).

Passando a analisar a Tabela 30, onde as medidas de sensibilidade (MAE e RMSE) para taxa de faltantes de 5% foi a que apresentou o maior erro, situação não esperada, porém como já enfatizado nas análises das outras bases de dados, provavelmente os valores faltantes ficaram próximos de *outliers*, dificultando assim que o modelo ajustasse bem aos dados. Quanto à coluna referente à "%Melhoria", esta medida foi aferida da seguinte forma: Como já citado anteriormente, para cada função de ativação e para cada porcentagem de dados faltantes, foram utilizados os três frameworks, apresentados nos Quadro 3, 4 e 5. Em seguida, tomou-se como função padrão a TH (Tangente Hiperbólica) e calculou-se a porcentagem de melhoria que cada uma das outras funções conseguiram obter em relação a ela. Sendo assim, é notório que estas novas funções, principalmente quando se utiliza o EMV, apresentam melhores resultados quando comparadas à função clássica TH.

Para uma melhor visualização e assimilação dos erros auferidos com o algoritmo EM e com a Rede Neural MLP, na Tabela 31, tem um comparativo entre as medidas para ambas as técnicas.

Tabela 31: Comparação entre as medidas de sensibilidade via as duas técnicas para a base *Parkinson*.

% Faltantes	MAE			RMSE		
	EM	RNA	%Melhoria	EM	RNA	%Melhoria
5	8,2231	8,2946	-0,8694	10,2023	9,3964	7,8993
10	7,9235	7,4667	5,7651	9,9345	8,6489	12,9409
20	8,0999	7,4133	8,4760	10,2194	8,5952	15,8935
30	7,9836	7,3338	8,1389	9,9961	8,5081	14,8864
40	8,2218	7,0514	14,2358	10,2374	8,2479	19,4339
50	8,2544	6,6566	**19,3560**	10,3481	7,9254	**23,4119**
60	8,1555	6,6483	18,4812	10,2419	7,8636	23,2216
70	7,9759	6,7753	15,0528	10,1116	7,9846	21,0351

Ao observar as medidas de erros, na Tabela 31, constata-se que ao utilizar a Rede Neural MLP, para imputar dados, esta apresenta consideravelmente melhor desempenho e assertividade do que o algoritmo EM, exceto para a taxa de faltantes de 5%, para a medida do MAE, a qual a RNA apresentou um desempenho ligeiramente inferior. A principal melhoria obtida foi para as taxas de faltantes acima de 40%, principalmente para a taxa de faltante de 50%, a qual teve um ganho de 19,35% de melhoria em relação à medida MAE, e 23,41% em relação à medida RMSE. A próxima seção tratará de analisar todas estas bases, aqui analisada na seção 5.3, porém, agora pelo viés da imputação múltipla.

5.4 Análise dos dados com imputação múltipla

Nesta seção, que também analisará as quatro bases de dados, dantes já citadas, utilizou o mesmo mecanismo gerador de dados incompletos, o MAR, sendo assim, a partir desta premissa gerou-se 8 bases de dados incompletos, com as respectivas quantidades omissas de 5%, 10%, 20%, 30%, 40%, 50%, 60% e 70%.

Para todas as bases de dados que serão analisadas a seguir, utilizaram-se quatro sementes (123, 43112, 1234567 e 1802), as quais garantem que sempre ter-se-á os mesmos valores aleatórios, que são gerados no início do experimento.

Além das medidas propostas nas subseções 5.1.1 e 5.1.2, nesta seção também faz-se necessário medir as métricas propostas por Rubin (1976), conforme já apresentadas na subseção 3.2, que são: a média dos valores estimados ($\bar{Q}$), a média da variância ($\bar{U}$), a estimativa não enviesada da variância (B), a variância total (T), os graus de liberdade (df), o intervalo de confiança (que será 95%), a taxa de informação faltante (γ) e o erro padrão (S).

5.4.1 Base de dados Emulsão

Neste experimento, partindo do principio de imputação múltipla, gerou-se quatro bases de dados completadas para cada taxa de dados faltantes, para que fosse possível realizar as análises conforme o paradigma da imputação múltipla. A primeira análise abordou o algoritmo EM, que foi executado para a base quatro vezes, sendo que cada vez utilizou-se de cada uma das sementes, respectivamente, citada na seção anterior, 5.4, e em seguida fez-se a análise via RNA-MLP. A seguir na Tabela 32 têm os resultados referentes ao algoritmo EM.

Tabela 32: Medidas de Sensibilidade para a base Emulsão via imputação múltipla para o algoritmo EM.

%Faltantes	Erros		Dados Imputados													
	MAE	RMSE	Q	Max	Med	Min	$\overline{U}$	B	T	df	95% Conf.		r	y	Efic.	S
5	2,6697	2,8860	16,1230	20,5856	16,2300	10,4389	8,7317	14,9761	28,6998	19631,8380	5,6228	26,6232	2,2868	0,6958	0,9877	5,3572
10	2,3068	2,7219	14,4914	20,3680	16,1500	8,2467	5,2387	21,0339	29,7783	45552,2807	3,7958	25,1870	4,6843	0,8241	0,9756	5,4569
20	2,3009	2,7949	14,0764	18,6264	15,0000	10,8400	5,6780	5,3316	11,4538	9215,1536	7,4431	20,7097	1,0172	0,5044	0,9524	3,3843
30	2,7257	3,0672	15,3887	20,8217	15,7289	8,4014	5,2130	11,0239	16,8493	52256,6676	7,3433	23,4341	2,2322	0,6906	0,9302	4,1048
40	1,9542	2,5212	13,9341	21,9361	16,1500	6,3683	5,1387	17,1135	22,9652	167832,3903	4,5414	23,3268	3,4691	0,7762	0,9091	4,7922
50	2,1530	2,8554	14,9975	20,5617	16,2300	8,4985	4,6297	7,9550	12,8499	38935,5822	7,9715	22,0234	1,7755	0,6397	0,8889	3,5847
60	2,5021	3,4010	14,3492	20,1555	14,4100	6,9698	2,6342	13,5100	16,5195	44424,3530	6,3829	22,3155	5,2711	0,8405	0,8696	4,0644
70	2,3264	3,2208	13,9860	19,2329	14,6948	7,5994	3,9254	9,4329	13,5829	56636,3530	6,7624	21,2096	2,4603	0,7110	0,8511	3,6855

Ao analisar a Tabela 32, observa-se que os valores dos erros, para a taxa de faltantes de 5% (que está destacada em azul e negrito), são altos quando comparados às outras taxas de falantes, porém não é o maior, como ocorreu na imputação única (seção 5.3.1, Tabela 18). Quando à maior taxa de erro MAE, está se deu à taxa de 30%, já o erro RMSE este teve seu maior valor quando se teve uma taxa de faltantes de 60%. Uma situação atípica foi que tanto o MAE quanto o RMSE, que apresentaram a menor taxa de erro ocorreram quando se teve uma taxa de faltante de 40%. Em geral verifica-se que não houve uma enorme discrepância entra os erros (MAE e RMSE), quando analisados em respeito à taxa de faltantes no *dataset*. Quanto à média da variância ($\overline{U}$) dos dados imputados, vê-se que para a taxa de faltantes de 5% foi a que teve maior valor, fato este que é decorrido da pequena quantidade de amostras, já a estimativa não enviesada da variância (B), está oscilou muito entre as respectivas taxas de faltantes. A variância total (T) teve seu melhor desempenho a uma taxa de 20%. O intervalo de confiança, que foi ao nível de confiança de 95%, também apresentou seu melhor resultado à taxa de 20%, já que apresentou a menor distância entre o valor mínimo e o máximo, tornando assim a estimativa mais precisa, mais confiável e, por fim, o erro padrão (S) que, também, apresentou seu melhor desempenho à taxa de 20% de faltantes. A seguir será analisada esta base, via RNA-MLP.

A análise da imputação múltipla para a base Emulsão via Redes Neurais Artificiais MLP, também seguiu os mesmos passos da seção 5.3.1 para a análise da RNA-MLP (ver Quadro 3, 4 e 5), porém com a diferença que cada algoritmo foi executado quatro vezes, sendo que cada vez com uma semente diferente, como anteriormente mencionado. A Tabela 33 contém os resultados para esta abordagem.

Tabela 33: Análise de sensibilidade via imputação múltipla para a base Emulsão via RNA-MLP.

%Faltantes	Erros		Dados Imputados													
	MAE	RMSE	Q	Max	Med	Min	$\overline{U}$	B	T	df	95% Conf.		r	y	Efic.	S
5	0,3382	0,4835	13,1508	16,4134	16,1248	10,8400	0,7488	7,9837	11,3936	60,1349	6,5349	19,7667	14,2167	0,9364	0,8103	3,3754
10	1,3546	2,2761	16,2473	22,4426	16,1500	10,8800	0,0424	20,0927	23,4838	14,9557	6,7491	25,7455	553,4238	0,9984	0,9756	4,8460
20	1,0766	1,5193	14,0782	22,1211	15,3008	10,8400	1,0236	14,6932	16,9413	2436,5924	6,0109	22,1455	15,5505	0,9396	0,9524	4,1160
30	0,6535	0,9445	15,2071	22,2874	16,1500	10,8800	0,1062	19,9974	21,2145	154,1500	6,1795	24,2347	198,8286	0,9951	0,9302	4,6059
40	1,3014	2,2272	14,5297	23,1143	16,2113	10,8400	0,0243	15,4179	16,0847	42,5289	6,6690	22,3905	660,6625	0,9986	0,9091	4,0106
50	1,2345	1,7528	15,1853	21,1876	16,2300	10,9500	0,1900	14,7784	15,4609	400,6045	7,4785	22,8921	80,3914	0,9878	0,8889	3,9320
60	3,5530	5,0468	14,8249	19,2306	14,4100	10,8400	0,2492	13,6783	14,3074	651,9329	7,4111	22,2386	56,4248	0,9826	0,8696	3,7825
70	3,3280	4,4862	14,4093	21,2708	16,1500	10,9500	0,0633	11,0722	11,3991	116,3204	7,7918	21,0267	179,1349	0,9945	0,8511	3,3763

Na Tabela 33, observa-se que os valores que estão em vermelho, de cima para baixo, correspondem respectivamente aos menores e maiores valores dos erros medidos respectivamente, sendo que a uma taxa de 5% o erro foi muito baixo, considerado muito bom para esta situação. Para as outras variáveis medidas, merecem destaque é o intervalo de confiança, que conseguiu manter o intervalo entre o valor mínimo e o valor máximo não muito distante, dando indício de uma maior precisão, principalmente para a taxa de faltantes de 5%,. No geral, para as demais medidas, verifica-se que o melhor resultado foi obtido com a taxa de faltante de 5%. Cabe ressaltar que, as funções de ativação utilizadas para esta análise foram as mesmas que foram utilizadas na Tabela 20, na mesma sequência, visto terem sido aquelas que apresentaram melhor resultado. A próxima análise faz uma comparação entre os resultados dos erros utilizando a imputação múltipla por ambas as técnicas, algoritmo EM e RNA-MLP.

Tabela 34: Medidas de erro via imputação múltipla para comparar o desempenho do algoritmo EM versus RNA-MLP para a base Emulsão.

%Faltantes	Imputação Múltipla			
	MAE		RMSE	
	RNA	EM	RNA	EM
5	**0,3382**	**2,6697**	**0,4835**	**2,8860**
10	1,3546	2,3068	2,2761	2,7219
20	1,0766	2,3009	1,5193	2,7949
30	0,6535	2,7257	0,9445	3,0672
40	1,3014	1,9542	2,2272	2,5212
50	1,2345	2,1530	1,7528	2,8554
60	**3,5530**	**2,5021**	**5,0468**	**3,4010**
70	**3,3280**	**2,3264**	**4,4862**	**3,2208**

Os valores destacados em vermelho na Tabela 34 correspondem respectivamente aos menores e maiores valores de erros, de cima para baixo, para cada técnica. Verifica-se que uma taxa de 5%, a RNA-MLP apresentou um excelente resultado quando comparado ao algoritmo EM, porém quando a taxa de faltantes é igual ou superior a 60%, o algoritmo EM apresentou melhor desempenho. A próxima tabela traz uma comparação para estas medidas, tanto para a imputação única como para a imputação múltipla, a fim de fazer uma análise comparativa geral.

Tabela 35: Medidas de sensibilidade via Imputação Única e Imputação Múltipla para a base Emulsão.

%Faltantes	Imputação Única				Imputação Múltipla			
	MAE		RMSE		MAE		RMSE	
	RNA	EM	RNA	EM	RNA	EM	RNA	EM
5	0,4318	4,5096	0,5469	5,2769	0,3382	2,6697	0,4835	2,8860
10	1,2348	4,0748	2,2061	4,7134	1,3546	2,3068	2,2761	2,7219
20	1,2149	2,5772	1,4572	3,0092	1,0766	2,3009	1,5193	2,7949
30	0,8077	3,1044	1,1895	3,8201	0,6535	2,7257	0,9445	3,0672
40	1,4060	2,8167	1,9922	3,6704	1,3014	1,9542	2,2272	2,5212
50	1,2177	2,8601	1,7535	3,6188	1,2345	2,1530	1,7528	2,8554
60	1,8486	3,0725	2,6926	3,8328	3,5530	2,5021	5,0468	3,4010
70	1,6903	3,2062	2,4616	4,1198	3,3280	2,3264	4,4862	3,2208

Ao observar a Tabela 35, verifica-se que a RNA-MLP apresentou melhor desempenho para todas as taxas de faltantes quando comparada ao algoritmo EM, na imputação única; já para a imputação múltipla a RNA-MLP apresentou um desempenho inferior ao algoritmo EM apenas para as taxas de 60% e 70%. Uma situação que merece ser levantada na presente análise é que, o algoritmo EM quando utilizado via imputação múltipla apresentou melhores resultados do que quando utilizou via imputação única, situação esta, a qual justifica a viabilidade do uso do método de imputação múltipla, pois a melhoria auferida é significativa. Quanto à RNA, quando utilizou-se imputação múltipla trouxe uma ganho em alguns casos, porém este ganho é muito pequeno, o qual desmotiva a utilização da imputação múltipla, pois é muito trabalhosa de se obter. A próxima seção analisará a base *Breast Tissue*.

5.4.2 Base de dados Breast Tissue

Para a base *Breast Tissue*, neste contexto de imputação múltipla, gerou-se também quatro bases de dados completadas para cada taxa de dados faltantes, para que fosse possível realizar as análises. O algoritmo EM foi executado para a base quatro vezes, sendo que cada vez utilizou-se de cada uma das sementes, respectivamente, citada no início da seção 5.4. Seguindo a mesma sequência anterior de análise, da subseção 5.41, a primeira tabela refere-se aos dados auferidos com o algoritmo EM.

Tabela 36: Medidas de Sensibilidade para a base *Breast Tissue* via imputação múltipla para o algoritmo EM.

%Faltantes	Erros		Dados Imputados													
	MAE	RMSE	Q	Max	Med	Min	$\overline{U}$	B	T	df	95% Conf.		r	y	Efic.	S
5	**8,0101**	**12,7864**	679,0818	2090,8919	445,5133	124,9786	962,2263	648354,2252	778987,2965	1,08113E+18	-1050,8192	2408,9829	808,5677	0,9988	0,9877	882,6026
10	23,5336	38,8383	1192,4101	2469,2250	433,2023	124,9786	908,6994	849413,8702	935263,9565	4,43135E+18	-703,0872	3087,9074	1028,2336	0,9990	0,9756	967,0905
20	20,0039	24,9984	503,4936	2502,4028	445,5133	134,8927	861,5414	336269,8005	353144,1895	1,52951E+18	**-661,2548**	**1668,2419**	408,8981	0,9976	0,9524	594,2594
30	32,2271	45,8211	750,6229	2406,4047	433,2023	124,9786	522,3195	467929,6933	483074,8158	1,74128E+18	-611,6474	2112,8931	923,8645	0,9989	0,9302	695,0358
40	25,2984	34,5391	696,2969	2626,9912	456,1096	124,9786	719,1889	469314,7678	481208,1178	4,45615E+18	-663,3387	2055,9326	668,0984	0,9985	0,9091	693,6917
50	23,0227	33,7054	774,9624	2860,2167	433,2023	91,5705	599,9751	556634,2121	567736,7195	5,58694E+18	-701,8644	2251,7892	945,2671	0,9989	0,8889	753,4831
60	40,6601	96,5703	836,8572	3363,3776	423,4923	127,5166	439,5429	690176,6911	701400,2448	5,62078E+18	-804,6358	2478,3502	1594,7492	0,9994	0,8696	837,4964
70	36,1529	59,1865	733,0983	3030,4749	435,9415	127,0944	568,3712	480040,5850	487095,9912	5,29034E+18	-634,8300	2101,0267	856,0033	0,9988	0,8511	697,9226

Ao analisar a Tabela 36, observa-se que os valores dos erros, para a taxa de faltantes de 5% foram os menores, como era esperado. Em geral os demais valores correspondentes às outras taxas de faltantes não oscilaram muito, com exceção da taxa de faltante de 60%, que teve para o MAE e o RMSE, respectivamente os valores de 40,66 e 96,57. Quanto à média da variância ($\overline{U}$) dos dados imputados, vê-se que para a taxa de faltantes de 5% foi a que teve maior valor, fato este que é decorrido da pequena quantidade de amostras, já a estimativa não enviesada da variância (B) não oscilou muito entre as respectivas taxas de faltantes. A variância total (T) teve seu melhor desempenho a uma taxa de 20%. O intervalo de confiança, que foi ao nível de confiança de 95%, também apresentou seu melhor resultado à taxa de 20%, já que apresentou a menor distância entre o valor mínimo e o máximo, o que significa que sua confiabilidade é maior, e por fim o erro padrão (S) que, também, apresentou seu melhor desempenho à taxa de 20% de faltantes. A seguir este *dataset* é analisado pelo viés da RNA-MLP.

A análise da imputação múltipla para a base *Breast Tissue* via Redes Neurais Artificiais MLP, também seguiu os mesmos passos da seção 5.3.1 (Quadro 3, 4 e 5), porém com a diferença que cada algoritmo foi executado quatro vezes, sendo que cada vez com uma semente diferente. A Tabela 37 contém os resultados para esta abordagem.

Tabela 37: Análise de sensibilidade para a base *Breast Tissue* via imputação múltipla para a RNA-MLP.

%Faltantes	Erros		Dados Imputados														
	MAE	RMSE	Q	D. Pad	Max	Med	Min	$\overline{U}$	B	T	df	95% Conf.		r	y	Efic.	S
5	**384,3085**	**427,9031**	55,3158	259,8488	2080,7117	0,0000	0	312550,9825	1370257,2815	1911184,4775	6,73787E+23	-2654,2979	2764,9294	5,1148	0,8365	0,9877	1382,4560
10	102,9684	113,6860	1159,5380	443,9177	2343,7659	0,0000	0	1072,2937	763364,6269	833833,7048	5,63008E+18	-630,2261	2949,3021	776,6169	0,9987	0,9756	913,1450
20	203,3945	232,3542	695,6019	375,6394	2265,9258	0,0000	0	6692,2484	340453,6230	363357,9488	9,45983E+19	**-485,8700**	**1877,0738**	53,2953	0,9816	0,9524	602,7918
30	184,3333	207,7389	912,6825	576,0924	2510,7748	0,0000	0	7311,0621	507771,1450	530469,2115	4,15448E+20	-514,8505	2340,2155	81,0279	0,9878	0,9302	728,3332
40	165,1522	191,6951	842,9665	623,0089	2744,0720	0,0000	0	6173,4443	547202,7625	566101,8524	4,57754E+20	-631,7324	2317,6654	90,6995	0,9891	0,9091	752,3974
50	**88,6193**	**102,6780**	790,4123	627,9670	2700,8256	283,4364	0	720,1333	476420,4339	485963,1678	6,01374E+18	-575,9245	2156,7490	673,8239	0,9985	0,8889	697,1106
60	96,1257	116,3368	782,9324	689,9814	2717,1767	288,6390	0	696,6011	547696,1987	556818,8951	9,0358E+18	-679,6255	2245,4902	798,3368	0,9987	0,8696	746,2030
70	110,0795	133,4647	736,9572	695,8533	2729,6365	307,5665	0	1098,3409	532621,4202	540821,3800	2,46626E+19	-704,4377	2178,3522	491,3985	0,9980	0,8511	735,4056

Na Tabela 37, observa-se que os valores que estão em vermelho, de cima para baixo, correspondem respectivamente aos maiores e menores valores dos erros medidos respectivamente, sendo que a uma taxa de 50% de faltantes foi a que apresentou o menor erro, situação atípica. Para as outras variáveis medidas, merece destaque o intervalo de confiança para a taxa de 20%, que apresentou o melhor resultado. A próxima análise faz uma comparação entre os resultados dos erros utilizando a imputação múltipla por ambas as técnicas, algoritmo EM e RNA-MLP.

A próxima tabela traz uma análise comparativa para as medidas de erro, tanto para a imputação única como para a imputação múltipla.

Tabela 38: Medidas de sensibilidade via Imputação única e Imputação Múltipla para a base *Breast Tissue*.

%Faltantes	Imputação Única				Imputação Múltipla			
	MAE		RMSE		MAE		RMSE	
	RNA	EM	RNA	EM	RNA	EM	RNA	EM
5	88,1880	26,2686	97,0989	36,9833	384,3085	8,0101	427,9031	12,7864
10	119,8770	30,5774	143,4113	50,2260	102,9684	23,5336	113,6860	38,8383
20	145,5419	36,8056	160,4857	44,1354	203,3945	20,0039	232,3542	24,9984
30	135,0605	33,0774	148,0840	43,0782	184,3333	32,2271	207,7389	45,8211
40	146,8996	33,0474	178,4546	43,0012	165,1522	25,2984	191,6951	34,5391
50	89,5367	27,4065	106,4106	35,4075	88,6193	23,0227	102,6780	33,7054
60	96,2515	45,2962	113,1779	93,0472	96,1257	40,6601	116,3368	96,5703
70	114,3430	39,1329	139,7314	64,4070	110,0795	36,1529	133,4647	59,1865

Ao observar a Tabela 38, verifica-se que a RNA apresentou pior desempenho para todas as taxas de faltantes, tanto para a imputação única como para a imputação. Uma situação que merece ser levantada na presente análise é que, o algoritmo EM quando utilizado via imputação múltipla, na maioria dos casos, apresentou melhores resultados do que quando utilizou imputação única, situação esta, a qual é plausível de ser utilizada, pois a melhoria auferida é significativa. Quanto à RNA, quando utilizou-se imputação múltipla trouxe uma ganho em alguns casos, porém este ganho é pequeno, o qual desmotiva a utilização da imputação múltipla, pois é muito trabalhosa de se obter. A próxima subseção irá analisar a base de dados *Concrete*.

5.4.3 Base de dados *Concrete*

Para a base *Concrete*, neste contexto de imputação múltipla, gerou-se também quatro bases de dados completadas para cada taxa de dados faltantes, para que fosse possível realizar tais análises. A primeira análise refere-se ao algoritmo EM, conforme valores da Tabela 39.

Tabela 39: Medidas de Sensibilidade para a base *Concrete* via imputação múltipla para o algoritmo EM.

%Faltantes	Erros		Dados Imputados													
	MAE	RMSE	Q	Max	Med	Min	$\overline{U}$	B	T	df	95% Conf.		r	y	Efic.	S
5	8,2794	10,2196	20,3336	30,2826	21,0000	0,0000	28,6351	54,3271	93,8276	6732807	1,3481	39,3191	2,2767	0,6948	0,9877	9,6865
10	6,1596	7,1263	17,8832	27,9429	21,1250	0,0000	45,3801	38,3384	87,5523	22542617	-0,4564	36,2228	0,9293	0,4817	0,9756	9,3569
20	7,0410	8,1982	16,9951	29,5252	20,2500	0,0000	40,8191	36,8539	79,5157	39054755	-0,4825	34,4727	0,9480	0,4867	0,9524	8,9172
30	6,5940	8,3238	17,3065	26,2010	21,4699	0,0000	36,6792	22,8729	60,3145	19163146	2,0847	32,5284	0,6444	0,3919	0,9302	7,7662
40	6,7258	8,2294	17,6690	30,7094	20,7500	-2,7345	31,5938	42,0478	74,6672	4747443385	0,7326	34,6054	1,3633	0,5769	0,9091	8,6410
50	7,3115	9,0091	17,3903	35,3063	18,7095	-0,3842	33,5591	41,2133	75,5805	9952386665	0,3506	34,4299	1,2522	0,5560	0,8889	8,6937
60	6,6525	8,2233	19,5056	41,5916	20,7500	-8,3431	31,1440	72,1271	104,4344	47331417086	-0,5243	39,5355	2,3533	0,7018	0,8696	10,2193
70	7,3920	9,6659	20,2696	33,3450	20,4979	0,0000	31,3116	15,9429	47,4760	3669873173	6,7647	33,7746	0,5162	0,3405	0,8511	6,8903

Ao analisar a Tabela 39, observa-se que os valores dos erros, para a taxa de faltantes de 5% (que está destacada em azule negrito), são os maiores quando comparados às outras taxas de falantes, inclusive a taxa de faltante de 70%, como ocorreu na imputação única (Tabela 24), já os menores valores de erros ocorreram à taxa de faltantes de 10%. Em geral verifica-se que não houve uma enorme discrepância entra os erros (MAE e RMSE), quando analisados em respeito à taxa de faltantes no *dataset*. Quanto à média da variância ($\overline{U}$) dos dados imputados, vê-se que para a taxa de faltantes de 10% foi a que teve maior valor, fato este que é decorrido da pequena quantidade de amostras, já a estimativa não enviesada da variância (B), esta oscilou muito entre as respectivas taxas de faltantes, sendo que a uma taxa de faltante de 70%, foi onde houve a menor oscilação, isso é devido à natureza do algoritmo EM, o qual diante de muitos dados faltantes, este conduz os dados para a centralidade. O intervalo de confiança, que foi ao nível de confiança de 95%, também apresentou seu melhor resultado à taxa de 70%, já que apresentou a menor distância entre o valor mínimo e o máximo, e por fim o erro padrão (S) que, também, apresentou seu melhor desempenho à taxa de 70% de faltantes. A seguir passa-se a analisar esta base via RNA-MLP.

A análise da imputação múltipla para a base Concrete via Redes Neurais Artificiais MLP, também seguiu os mesmos passos da seção 5.3.1 (Quadro 3, 4 e 5), porém com a diferença que cada algoritmo foi executado quatro vezes, sendo que cada vez com uma semente diferente, conforma já citado no inicia da seção 5.3. A Tabela 40 contém os resultados para esta abordagem.

Tabela 40: Análise de sensibilidade para a base *Concrete* via imputação múltipla para a RNA-MLP.

%Faltantes	Erros		Dados Imputados													
	MAE	RMSE	Q	Max	Med	Min	$\overline{U}$	B	T	df	95% Conf.		r	y	Efic.	S
5	5,5763	6,1788	20,0280	22,7609	0,0000	0	6,4963	7,2056	15,1430	6403	12,4009	27,6552	1,3310	0,5711	0,9877	3,8914
10	5,9389	7,0124	14,7748	19,4376	0,0000	0	0,7391	60,5061	67,2958	15617	-1,3039	30,8535	90,0489	0,9890	0,9756	8,2034
20	5,3368	6,9407	16,2481	20,8913	0,0000	0	0,3508	63,4745	66,9991	9371	0,2049	32,2912	189,9709	0,9948	0,9524	8,1853
30	4,6272	6,2916	14,2887	21,3599	0,0000	0	2,1670	72,1468	76,7186	672626	-2,8788	31,4562	34,4039	0,9718	0,9302	8,7589
40	5,6310	7,9162	16,5442	23,3387	0,0000	0	4,2791	56,9437	62,6117	2282332	1,0352	32,0532	13,6318	0,9317	0,9091	7,9128
50	4,3006	6,1552	20,8173	28,6436	0,2634	0	5,0060	48,0019	53,9491	2800822	6,4211	35,2135	9,7768	0,9072	0,8889	7,3450
60	4,9676	6,9500	18,3648	24,3243	12,4554	0	1,6248	39,4356	41,6965	250320	5,7085	31,0210	24,6620	0,9610	0,8696	6,4573
70	7,7508	8,5184	13,5991	21,5100	12,3308	0	80,3992	13,9659	94,5591	87237777	-5,4602	32,6585	0,1761	0,1497	0,8511	9,7241

Na Tabela 40, observa-se que os valores que estão em vermelho, correspondem respectivamente aos menores valores dos erros medidos, que foram a uma taxa de 50% de faltantes, considerado muito bom para esta situação. Porém, observa-se que não há uma grande oscilação entre as medidas de erros para as demais taxas de faltantes. Quanto ao intervalo de confiança, foi com 5% de faltantes que este apresentou o melhor resultado. A próxima tabela traz uma comparação para estas medidas, tanto para a imputação única como para a imputação múltipla, a fim de fazer uma análise comparativa geral.

Tabela 41: Medidas de sensibilidade via Imputação única e Imputação Múltipla para a base *Concrete*.

%Faltantes	Imputação Única				Imputação Múltipla			
	RNA	EM	RNA	EM	RNA	EM	RNA	EM
	MAE	MAE	RMSE	RMSE	MAE	MAE	RMSE	RMSE
5	3,720066	11,66051	4,093947	15,04027	5,576321	8,27944	6,178778	10,21961
10	5,772313	4,920452	6,705807	5,565767	5,938903	6,15962	7,012446	7,12625
20	5,23558	8,132099	6,807395	10,467	5,336803	7,041021	6,940725	8,198244
30	4,012214	8,411914	5,641415	9,590623	4,627224	6,594011	6,291583	8,323768
40	5,054908	8,05397	7,441822	10,09433	5,630979	6,725792	7,916177	8,229398
50	4,169042	8,522706	6,346252	10,37831	4,300564	7,311521	6,155165	9,009084
60	4,875035	6,973913	6,979971	8,74926	4,967635	6,652516	6,949982	8,223272
70	4,823048	8,151204	7,078481	10,65745	7,750769	7,39204	8,51841	9,665866

Ao observar a Tabela 41, verifica-se que a RNA apresentou um melhor desempenho para a maioria das taxas de faltantes, com exceção da taxa de faltantes de 10%, para a imputação única, já para a imputação múltipla a RNA apresentou um melhor desempenho para todas as taxas de faltantes. Uma situação que merece ser levantada na presente análise é que, o algoritmo EM quando utilizado via imputação múltipla, na maioria dos casos, apresentou melhores resultados do que quando utilizou imputação única, situação esta, a qual, motiva o uso deste método para imputar dados. Quanto à RNA, quando utilizou-se imputação múltipla não trouxe ganho na maioria dos casos, e quando trouxe algum ganho este foi pequeno, o qual desmotiva a utilização da imputação múltipla, pois é muito trabalhosa de se obter. A próxima seção analisará a base *Parkinson*.

5.4.4 Base de dados *Parkinson*

Nesta base, Parkinson, também, no contexto de imputação múltipla, gerou-se quatro bases de dados completadas para cada taxa de dados faltantes, viabilizando assim a realização destas análises. Ao executar o algoritmo EM quatro vezes, sendo que cada vez utilizou-se de cada uma das sementes, respectivamente, citada no início da seção 5.4, teve-se como resultados os valores da Tabela 42.

Tabela 42: Medidas de Sensibilidade para a base *Parkinson* via imputação múltipla para o algoritmo EM.

%Faltantes	Erros		Dados Imputados													
	MAE	RMSE	Q	Max	Med	Min	$\bar{U}$	B	T	df	95% Conf.		r	y	Efic.	S
5	6,7440	8,0266	20,8695	33,7320	20,8960	-5,2796	38,3758	27,0590	65,5268	314409176	5,0036	36,7355	0,7075	0,4143	0,9877	8,0949
10	6,4293	7,8101	21,1686	37,8055	20,8710	-3,7440	38,6629	26,9551	65,6638	636599451	5,2861	37,0511	0,6984	0,4112	0,9877	8,1033
20	6,4573	7,9438	21,0540	42,4922	21,1065	-5,1825	38,7718	27,4428	66,2379	1329335102	5,1022	37,0058	0,7084	0,4147	0,9524	8,1387
30	6,5161	7,9889	21,2532	36,0259	21,1920	-5,0981	38,9611	25,2878	64,2633	1710938458	5,5409	36,9654	0,6494	0,3937	0,9302	8,0164
40	6,5198	8,0134	21,0790	53,1970	21,2200	-3,5689	39,2742	25,9805	65,2658	2448359992	5,2447	36,9133	0,6618	0,3982	0,9091	8,0787
50	6,4933	8,0280	20,9716	40,5101	21,0438	-7,0259	38,4910	26,0044	64,5042	2946368720	5,2299	36,7132	0,6758	0,4033	0,8889	8,0315
60	6,5536	8,0863	21,0078	38,6471	20,9805	-7,3610	38,0434	28,1478	66,1993	4046227108	5,0607	36,9549	0,7401	0,4253	0,8696	8,1363
70	6,5278	8,1326	21,0320	84,9630	21,0076	-5,4814	37,5511	31,5484	69,1071	5776542678	4,7384	37,3256	0,8404	0,4566	0,8511	8,3131

Ao analisar a Tabela 42, observa-se que os valores dos erros, para a taxa de faltantes de 5% (que está destacada em azul e negrito) são os maiores quanto ao MAE, porém não é o maior quanto ao RMSE, quando comparado às outras taxas de falantes, como ocorreu na imputação única. Já os menores valores de erros ocorreram à taxa de faltantes de 10%. Em geral verifica-se que não houve uma enorme discrepância entra os erros (MAE e RMSE), quando analisados em respeito a todas as taxa de faltantes no *dataset*. Quanto à média da variância ($\overline{U}$) dos dados imputados, vê-se que os valores são praticamente homogêneos, fato este que é decorrido de uma maior quantidade de amostras, já a estimativa não enviesada da variância (B), está também, praticamente não oscilou entre as respectivas taxas de faltantes, isso é devido à natureza do algoritmo EM, o qual diante de muitos dados faltantes, este conduz os dados para a centralidade. A seguir será analisada esta base, via RNA-MLP.

Ao analisar a base *Parkinson* via imputação múltipla através de Redes Neurais Artificiais MLP, também seguiram-se os mesmos passos da seção 5.3.1 (Quadro 3, 4 e 5), porém com a diferença que cada algoritmo foi executado quatro vezes, sendo que cada vez com uma semente diferente, conforma já citado no inicia da seção 5.3. A Tabela 43 contém os resultados para esta abordagem.

Tabela 43: Análise de sensibilidade para a base *Parkinson* via imputação múltipla para a RNA-MLP.

%Faltantes	Erros		Dados Imputados													
	MAE	RMSE	Q	Max	Med	Min	$\bar{U}$	B	T	df	95% Conf.		r	y	Efic.	S
5	8,3066	9,4084	21,3051	21,4674	0,0000	0	0,0243	0,5357	0,5614	397	19,8366	22,7736	22,1470	0,9570	0,9877	0,7492
10	**8,5158**	**10,1156**	24,3931	27,0867	0,0000	0	24,8809	5,5618	30,4509	13152836	13,5774	35,2088	0,2239	0,1829	0,9756	5,5182
20	7,4083	8,5918	22,1186	22,2740	0,0000	0	0,0189	0,5712	0,5905	1272	20,6124	23,6248	30,2183	0,9680	0,9524	0,7685
30	7,3195	8,5059	22,1312	22,2739	0,0000	0	0,0785	0,2632	0,3419	1915	**20,9852**	**23,2772**	3,3551	0,7706	0,9302	0,5847
40	7,7523	9,6418	25,7519	32,8681	0,0000	0	23,3312	8,2934	31,6281	91205105	14,7291	36,7747	0,3556	0,2623	0,9091	5,6239
50	6,9419	8,1333	21,2688	21,4964	18,8739	0	0,4899	0,3647	0,8547	4153	19,4568	23,0808	0,7447	0,4271	0,8889	0,9245
60	6,8512	8,0240	21,6603	22,2731	21,0957	0	0,5165	2,0166	2,5337	14817	18,5404	24,7801	3,9056	0,7962	0,8696	1,5918
70	**6,6678**	**7,8983**	21,8037	22,2724	22,2603	0	0,2924	2,0183	2,3112	10464	18,8239	24,7834	6,9041	0,8735	0,8511	1,5203

Na Tabela 43, observa-se que os valores que estão em vermelho, correspondem respectivamente aos maiores e menores valores dos erros medidos, que foram a uma taxa de 10% e 70%. Esta base apresentou uma situação totalmente inesperada, que são os valores dos erros referente a taxa de 70% de faltantes, os quais são o menores entre todos os erros. Esta situação além de ser inesperada é também inexplicável, além disso observa-se que os erros foram praticamente homogêneos entre as taxas de faltantes. Quanto ao intervalo de confiança, foi com 30% de faltantes que este apresentou o melhor resultado, porém na maioria dos casos houve um bom desempenho, o que indica que há uma grande confiabilidade para os valores estimados. A próxima tabela traz uma comparação para estas medidas, tanto para a imputação única como para a imputação múltipla, a fim de fazer uma análise comparativa geral.

Tabela 44: Medidas de sensibilidade via Imputação única e Imputação Múltipla para a base *Parkinson*.

%Faltantes	Imputação Única				Imputação Múltipla			
	MAE		RMSE		MAE		RMSE	
	RNA	EM	RNA	EM	RNA	EM	RNA	EM
5	8,2946	8,2231	9,3964	10,2023	8,3066	6,7440	9,4084	8,0266
10	7,4667	7,9235	8,6489	9,9345	8,5158	6,4293	10,1156	7,8101
20	7,4133	8,0999	8,5952	10,2194	7,4083	6,4573	8,5918	7,9438
30	7,3338	7,9836	8,5081	9,9961	7,3195	6,5161	8,5059	7,9889
40	7,0514	8,2218	8,2479	10,2374	7,7523	6,5198	9,6418	8,0134
50	6,6566	8,2544	7,9254	10,3481	6,9419	6,4933	8,1333	8,0280
60	6,6483	8,1555	7,8636	10,2419	6,8512	6,5536	8,0240	8,0863
70	6,7753	7,9759	7,9846	10,1116	6,6678	6,5278	7,8983	8,1326

Ao observar a Tabela 44, verifica-se que a RNA apresentou um desempenho ligeiramente melhor para todas as taxas de faltantes, para a imputação única, já para a imputação múltipla ocorreu o inverso, ou seja, a RNA apresentou um desempenho ligeiramente inferior, para a todas as taxas de faltantes (com exceção para a taxa de faltante de 60%), em relação ao algoritmo EM. Esta situação, que também era inesperada, pode ser justifica pelo fato de que as sementes escolhidos para a imputação múltipla certamente não eram adequadas para gerar os valores dos pesos para a RNA-MLP. A próxima subseção apresenta alguns resultados comparativos, acerca do ganho obtido da RNA-MLP quando combinada com o EMV.

5.5 Ponderações acerca do uso do EMV combinado com a RNA-MLP

Conforme o que foi ponderado na subseção 4.5.2, no qual o método de EMV exerce um papel importante no processo de convergência, melhorando a capacidade de generalização de uma RNA-MLP; aqui têm alguns resultados, medidos na fase de treinamento, para os quatro *dataset* analisados, seguindo os passos implementados nos pseudocódigos (Quadro 3, 4 e 5). Primeiramente são apresentados na Tabela 47 os parâmetros que foram utilizados para treinar a RNA-MLP para todas as bases de dados.

Tabela 45: Parâmetros para o treinamento da RNA-MLP.

Semente	123 - 43112 - 1234567 - 1802		
Qtd. Camadas	3		
Neurônios por camada	Camada de Entrada	Camada Intermediária	Camada de Saída
	5 - 9 - 7 - 18	5	1
Iteração	1000		
Taxa de aprendizado	0.09		
Função de Ativação	SIG	SIGEMV	EMVSIG
	AO	AOEMV	EMVAO
	TH	THEMV	EMVTH
	CLL	CLLEMV	EMVCLL
	LL	LLEMV	EMVLL

As quatro sementes utilizadas, conforme já explanado no início da seção 5, foram escolhidas após vários experimentos, sendo que estas foram as que apresentaram melhores resultados. Quanto à quantidade de camadas, fixou-se em três, pois durante os experimentos observou-se que ao aumentar o número de camadas para além de três, não correspondia em uma melhoria significativa para o aprendizado da rede, e em alguns casos não havia nenhum ganho a mais, e sim apenas um custo computacional maior, visto que a rede demorava um tempo bem maior para aprender. Para a quantidade de neurônios por camada, também fixou-se cinco neurônios na camada intermediária, já que quando se aumentava para além dos cinco nem sempre trazia algum ganho em desempenho, e às vezes trazia uma perda de desempenho quanto ao erro final, e no neurônio de saída fixou-se um, já que este trabalho abordou o paradigma de regressão via RNA-MLP. Quanto a quantidade de neurônio na entrada da rede, esta foi de acordo com a quantidade de variáveis de entrada, ou seja, para a base emulsão haviam cinco variáveis de entrada, então a rede também teve na camada de entrada cinco

neurônios, e assim foi procedido com as outras bases de dados. As iterações foram fixadas em 1000, já que além destas o ganho em termos de redução do erro não foi tão significativo. A taxa de aprendizagem que melhor se adequou a todos os dados foi 0.09, e por fim todas as funções de ativação que foram utilizadas nos experimentos. O gráfico da Figura 19 mostra o desempenho geral de cada abordagem.

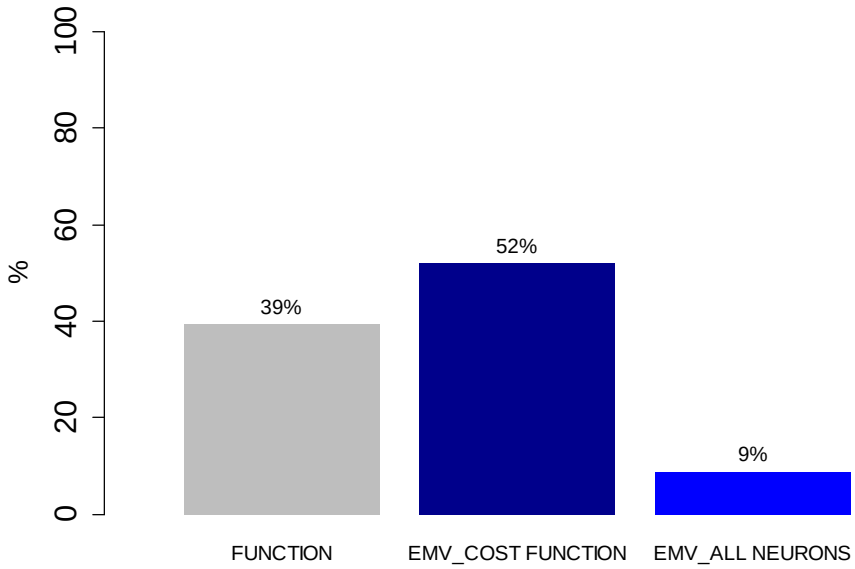

Figura 19: Gráfico com o desempenho da RNA-MLP para todas as bases e todas as abordagens.

Na Figura 19, há o desempenho percentual para as três abordagens utilizadas neste trabalho, ou seja, usar o EMV na função custo (EMV_COST FUNCTION), em todos os neurônios (EMV_ALL NEURONS) e não usar EMV, e sim apenas a função original (FUNCTION). Conforme os valores apresentados no gráfico constata-se que o melhor desempenho foi quando se utilizou EMV na função custo, com 52%, seguido do não uso do EMV, com 39%. Já ao utilizar o EMV em todos os neurônios, houve uma ganho de desempenho em apenas 9% dos casos, porém a vantagem de se utilizar tal abordagem é que ela consegue, geralmente, convergir em poucas iterações, conforme pode-se observar nas Figura 20 e Figura 21.

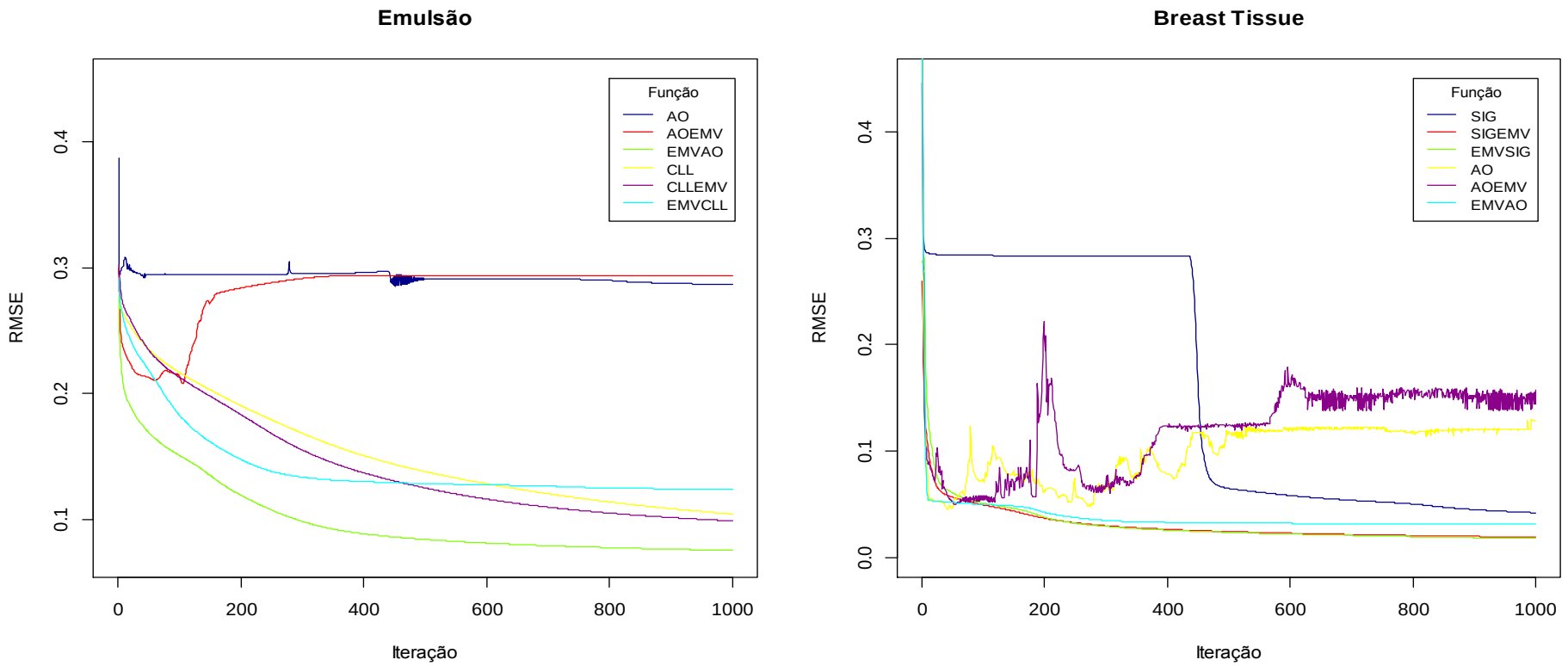

Figura 20: Gráfico de treinamento da RNA-MLP para as bases Emulsão e *Breast Tissue*.

Para a base Emulsão têm-se as funções que apresentaram melhor desempenho, sendo que para a função de ativação AO (Aranda-Ordaz), linha azul, esta apresentou o pior desempenho, a qual praticamente não aprendeu, o que aparenta que a rede ficou presa na fase de treinamento em algum mínimo local. Quando passa-se a analisar esta mesma função com o método de EMV na função erro (AOEMV), verifica-se que inicialmente a rede começou a aprender, porém também deve ter ficada presa em algum mínimo local e ao final do treinamento apresentou erro um pouco maior que a função principal AO. O grande ganho foi quando se utilizou o EMVAO em todas as funções de ativação, linha verde, a qual apresentou o menor erro, bem como seu processo de convergência para um ponto de mínimo foi muito rápido, conforme verifica-se no formato da curva de aprendizado (linha verde). Quanto à função de ativação CLL (Complemento Log-Log), esta apresentou melhor resultado que a AO e AOEMV, porém ao combinar o EMV, na função de erro, este trouxe um ganho pífio, já quando foi utilizado em todas as funções de ativação não houve ganho e sim perda em desempenho.

Analisando agora a base *Breast Tissue*, primeiramente observa-se o desempenho da rede via a função SIG (Sigmoide), a qual apresentou um bom desempenho, porém demorou mais de 400 iterações para começar a aprender, entretanto ao combinar o EMV, tanto na função de custo (SIGEMV), quanto em todas as funções dos neurônios (EMVSIG), houve um ganho significativo no que diz respeito ao processo de convergência, que foi muito rápido, pois conseguiu convergir em aproximadamente 200 iterações para o valor mínimo do RMSE, que foi o menor entre todas as abordagens utilizadas. Outra situação observada, também, é que ao usar o SIGEMV e EMVSIG, praticamente ambas as abordagens apresentaram o mesmo desempenho. Os próximos gráficos (Figura 21) analisam os dados das bases *Concrete* e *Parkinson*.

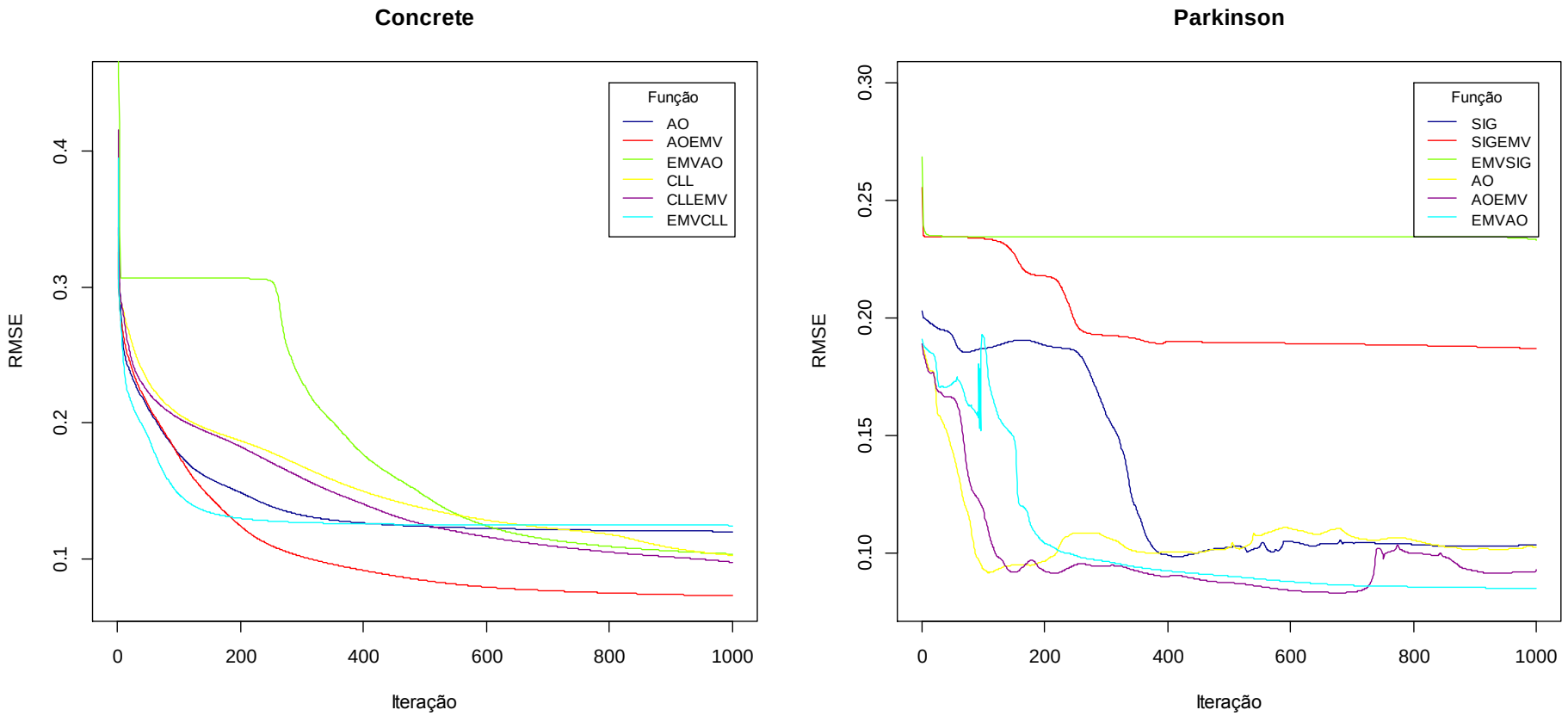

Figura 21: Gráfico de treinamento da RNA-MLP para as bases *Concrete* e *Parkinson*.

Analisando a base *Concrete*, via a função AO (Aranda-Ordaz), esta apresentou um bom desempenho, convergindo em aproximadamente 400 iterações, já ao usar a AOEMV em todas as funções de ativação, obteve-se um bom resultado, tanto em processo da curva de aprendizagem a qual conseguiu aprender muito rápido, bem como no erro RMSE que foi o menor entre todas as outras funções abordadas. Quanto ao desempenho da EMVAO, esta teve um processo de aprendizagem lento, porém após 600 iterações teve seu RMSE reduzindo e chegando ao final do treinamento com este erro bem menor do que quando usou a AO.

Ao analisar a base *Parkinson*, constata-se que ao utilizar a função SIG (Sigmoide), esta apresentou consideravelmente melhor desempenho do que quando se utilizou ela combinada com o EMV (SIGEMV e EMVSIG). Já quando usou-se a função AO, esta apresentou um processo de convergência muito rápido, entretanto ao final do aprendizado o RMSE foi praticamente idêntico quando se utilizou a função SIG. O ganho significativo ocorreu quando combinou o EMV (AOEMV e EMVAO), sendo que estes (AOEMV e EMVAO) apresentaram comportamento bem semelhante à AO, no início da curva de aprendizado, porém a partir da iteração 200ª o RMSE começou a reduzir e chegou ao final do treinamento bem menor do que quando se utilizou a AO. A próxima seção finaliza este trabalho, com as conclusões e trabalhos futuros.

CAPITULO 6

6 CONCLUSÕES

Esta dissertação tem como principal objetivo, apresentar um *framework* para tratar dados faltantes através de Redes Neurais Artificiais *Multilayer Perceptron*, com novas funções de ativação combinadas com o método de estimativa de máxima verossimilhança em todas as funções de ativação e na função custo, facilitando a imputação de dados em *dataset*, que possuam tal problema. Para tanto aborda-se dois vieses para lidar com tal problema, sendo o primeiro a imputação única e o segundo a imputação múltipla, conforme abordado em Rubin (1987).

Vários algoritmos para otimizar RNA-MLP tem sido propostos na literatura, bem como novas topologias, porém pouco esforço tem sido despendido para propor novas funções de ativação, como foi proposta por Gomes (2010). Sendo assim, por se tratar de uma abordagem proposta recentemente na literatura, esta dissertação também analisou o desempenho destas novas funções de ativação, bem como, também, aplicou o método de estimativa de máxima verossimilhança em tais funções a fim de verificar se tal abordagem traz algum ganho em relação à aceleração do processo de aprendizagem, e em relação à redução do erro.

No capítulo 2 foi inicialmente apresentada a fundamentação teórica atinente aos dados faltantes, explanando seus conceitos, com suas principais causas. Introduziu-se as benesses de se usar tal abordagem, com um comparativo entre as consequências da escolha de usar e não usar métodos de imputação. Alicerçaram-se também neste capítulo os principais mecanismos causadores de dados faltantes, os quais são imprescindíveis para os primeiros passos de uma correta análise. Tais mecanismos são classificados em MAR (ocorre quando o valor faltante não está relacionado com a variável que o contém, e sim com outra variável da amostra), MCAR (ocorre quando o valor faltante não está relacionado com seus valores anteriores ou posteriores, e nem com qualquer outra variável da amostra) e MNAR (que ocorre quando o valor faltante está relacionado com outros valores de sua própria variável).

Já o capítulo 3 mencionou as maneiras mais utilizadas para se tratar dados omissos, as quais se dividem nos casos de deleção, imputação única e imputação múltipla. Foram apresentadas as vantagens e desvantagens em utilizar cada uma delas, e por fim foi feita uma comparação entre a escolha de se usar os métodos de imputação baseados em estimativas de máxima verossimilhança e os métodos baseados em imputação múltipla.

O capítulo 4 relatou sobre redes neurais artificiais, sendo que inicialmente, fez-se uma breve descrição de como funciona o cérebro humano, e como é sua estrutura fisiológica e seu funcionamento. Em seguida passou a discorrer acerca de como os pesquisadores tem modelado, de forma matemática, o funcionamento do cérebro, centralizando-se nas RNA-MLP, com o algoritmo de aprendizado *backpropagaion*. Logo após, iniciou-se os comentários sobre a importância do uso de novas funções de ativação, e imediatamente fez-se uma descrição das funções de ativação que são corriqueiramente utilizadas na literatura (sigmoide e tangente hiperbólica), e em seguida expôs as três funções de ativação que foram propostas recentemente na literatura por Gomes (2010). Depois, foram apresentadas todas estas funções, modificadas, com o método de estimativa de máxima verossimilhança, demonstrando o passo a passo matemático de tal tratamento. Por fim, foram expostos somente os artigos que utilizam redes neurais artificiais, para analisar casos de dados faltantes. Sendo que em sua maioria, estes artigos conduziram o uso das RNAs combinadas com outros algoritmos. Todas as abordagens apresentaram algum ganho em termos de redução de erro, quando comparadas com outras abordagens corriqueiramente utilizadas, porém trazem consigo a desvantagem da complexidade e elevado custo computacional.

O capítulo 5 iniciou-se fazendo uma breve apresentação das principais medidas de sensibilidade (MAE e RMSE), que foram utilizadas neste trabalho, e em seguida fez-se uma análise preliminar dos quatro *dataset*, sendo que para tal análise recorreu-se à gráficos, testes estatísticos e estatísticas descritivas. Depois passou a analisar de fato os resultados obtidos, para todas as bases, tanto pelo viés de imputação única como múltipla, por meio do algoritmo EM e a RNA-MLP (com as abordagens propostas).

6.1 Discussão

Os experimentos realizados neste trabalho para imputar dados, pelo paradigma de imputação única e múltipla, mostram que quando o algoritmo EM é aplicado para resolver este problema, obteve-se que na maioria dos casos, as medidas de sensibilidade (MAE e RMSE) não oscilando muito entre as taxas de faltantes, situação que corrobora para inferirmos que estas medidas de sensibilidade são mais influenciadas pela presença de valores discrepantes do que pela quantidade de dados faltantes, situação que dá indício de que as medidas de sensibilidade, utilizadas aqui, tem pouca relação com a quantidade de dados faltantes. Tal situação foi nitidamente percebida ao analisar os *dataset* Emulsão (Tabela 19) e *Concrete* (Tabela 26), os quais apresentaram uma menor taxa de erro, após a retirada manual

de amostras nestes *dataset*, em locais da distribuição distantes de valores discrepantes, situação esta que conduziu a uma redução geral de aproximadamente 37% no erro. Quanto a aplicação da RNA-MLP para resolver este problema, também observou-se que em metade das análises de sensibilidade, estas não oscilaram muito, porém quando se compara o desempenho do algoritmo EM com a RNA-MLP, tem com resultado que a RNA-MLP apresentou para a maioria dos casos um desempenho bem superior ao algoritmo EM. Pelo viés da imputação única, a RNA-MLP apresentou um desempenho superior em 75% dos casos, quando comparado ao algoritmo EM, já pelo viés da imputação múltipla a RNA-MLP apresentou um melhor desempenho em 56,25% dos caso. Ressalte-se que estes desempenhos poderiam ter sido melhores se a base de dados *Breast Tissue* não tivesse apresentado um comportamento atípico (Tabela 23 e Tabela 37), a qual merece ser reanalisada com o auxílio de alguma técnica de seleção de características. Outa situação que também foi notada é que, na maioria dos experimentos, da RNA-MLP, que apresentaram bons desempenhos, ocorreram quando se utilizou as novas funções de ativação, bem como quando se utilizou o EMV, tanto na função de custo quanto em todos os neurônios (Figura 19).

Ao analisar o desempenho do método de imputação única contra o método de imputação múltipla, constata-se que a imputação múltipla não trouxe nenhum ganho em termos de redução de erro, e sim perda de desempenho, situação a qual evidencia que fixar sementes não é uma boa tomada de decisão para utilizar estes métodos, pois se a semente que for utilizada na imputação única for a que apresentar o melhor desempenho, consequentemente a imputação múltipla terá um desempenho inferior. Apesar do presente trabalho ter conseguido responder algumas questões, e também, contribuído com resultados positivos tanto na área de dados faltantes quanto na área de Redes Neurais Artificiais, ainda ficam algumas questões em aberta a serem melhor avaliadas e possivelmente melhoradas.

Inicialmente, frise-se que a presença de dados discrepantes nos dados é um fator determinante na qualidade dos resultados, sendo assim há uma necessidade de um aprofundamento no estudo de meios e técnicas que possam sanar ou suavizar tal problema. Além disso, observou-se uma necessidade de se utilizar alguma técnica de seleção de caraterísticas, a fim de verificar se precisará excluir alguma variável que seja redundante ou que esteja prejudicando o desempenho dos algoritmos.

Também, cabe destacar uma situação que ainda ficou em aberta, que é quanto aos valores dos parâmetros utilizados neste trabalho. Ainda que todos os parâmetros tenham sido auferidos experimentalmente, constata-se ser necessário fazer uma análise com maior

profundidade quanto à influência e sensibilidade que cada parâmetro escolhido exerce na qualidade das análises, o que facilitará uma escolha mais apropriada destes parâmetros em caso de se aplicar os algoritmos, aqui analisados, em outro conjunto novo de dados.

Cabe salientar que um dos parâmetros que deve ser analisado exaustivamente é quanto à escolha da semente, já que esta é fator determinante para que os algoritmos tenham resultados de qualidade. Além disso, caso se consiga chegar a uma semente tida como ótima, certamente não haverá a necessidade de analisar os dados pelo viés da imputação múltipla.

Deve-se, também, analisar o custo computacional dos algoritmos de imputação utilizados neste trabalho, principalmente quando da utilização de RNA-MLP com novas funções de ativação e com o EMV. Além disso, deve-se fazer uma análise do impacto nas medidas de sensibilidade, quanto à relação existente entre as funções de ativação, com suas melhorias via EMV, e a taxa de dados faltantes, já que nem sempre a mesma função apresenta o melhor desempenho para qualquer quantidade de dados faltantes.

E por fim, aplicar em algum modelo preditivo o conjunto dos dados reais e dos dados imputados, o que tornará possível avaliar e comparar o impacto e precisão de cada abordagem de imputação de dados, aqui estudadas, para a modelagem de dados na vida real.

6.2 Perspectivas Futuras

Após a análise dos experimentos realizados nesta dissertação, percebem-se alguns aspectos que podem ser melhores explorados nesta linha de pesquisa, tal como buscar novas medidas de sensibilidade, que possam ter alguma relação com a quantidade de dados faltantes e menos dependência ou influência da presença de dados faltantes. Permanece em aberto a perspectivas de novas pesquisas, que possam extrapolar os conceitos abordados nesta dissertação, tanto para imputação de dados faltantes, quanto para encontrar novas funções de ativação que possam responder com melhor precisão nas análises.

Uma aplicação que poderá ser conveniente da RNA-MLP com novas funções de ativação e com o EMV (na função de custo e em todos os neurônios), reside na área de regressão, classificação, predição e mineração de textos.

Outras técnicas da inferência estatística de estimadores poderão ser utilizadas em RNA-MLP, a fim de verificar se poderão trazer algum ganho em redução do erro. Além disso, dado que houve em algumas situações a convergência dos valores de saída da RNA-MLP para o infinito, deve-se buscar uma alternativa para sanar tal problema, como o uso do algoritmo

TAO que lida bem com problema de modelagem diante de dados discrepantes ou *outliers* (PERNíA-ESPINOZA, *et al*. 2005).

Dado que as bases de dados analisadas neste trabalho não seguiam uma distribuição normal, certamente deve-se procurar novas alternativas que possam modelar de forma robusta as mesmas, principalmente por distribuição assimétrica, tal como a *Birnbaum-Saunders*, que de acordo com Soto (2014), ela tem sido muito utilizada, na área médica, para descrever a resposta de sobrevivência, que leva em conta a informação que não foi medida, que neste caso pode ser considerada como um valor não observado (dado faltante). No artigo de Käärik (2006), também é encontrada uma proposta interessante, onde o mesmo usa cópulas para tratar dados faltantes. No trabalho de Acuna et. al (2014), também é proposto o Filtro de partícula para tratar dados faltantes, apresentando bons resultados. E por fim, pode-se abordar também, o HMC (Hamiltonian Monte Carlo), que na tese de Liublinska (2013) apresentou excelente resultado quando comparado a outras técnicas, já bem consolidadas na literatura de dados faltantes. Sendo assim, seria interessante tentar utilizar estas abordagens que foram propostas recentemente na literatura, de forma comparativa com os algoritmos analisados nesta dissertação.

REFERÊNCIAS

ABDELLA, M.; MARWALA, T. The use of genetic algorithms and neural networks to approximate missing data in database. **3rd International Conference on Computational Cybernetics – ICCC, IEEE**, p. 207-212, 2005.

ACUNA, D. E., ORCHARD, M. E., SILVA, J. F., Pérez, A. Multiple-imputation-particle-filtering scheme for Uncertainty Characterization in Battery State-of-Charge Estimation Problems with Missing Measurement Data. **Annual Conference of the Prognostics and Health Management Society**, 2014.

ALISSON, P. D. Handling Missing Data by Maximum Likelihood. **Statistics and Data Analysis, SAS Global Forum**, 2012.
ALLISON, Paul D. **Missing data**. Sage publications, 2001.

AMANI, A.; YORK, P.; CHRYSTYN, H.; CLARK, B. J.; DO, D. Q. Determination of factors controlling the particle size in nanoemulsions using Artificial Neural Networks. European Journal of Pharmaceutical Sciences, v. 35, n. 1, p. 42-51, 2008.

ASSUNÇÃO, F. **Estratégias para tratamento de variáveis com dados faltantes durante o desenvolvimento de modelos preditivos**. Dissertação de Mestrado da Universidade de São Paulo, São Paulo, 2012.

AYDILEK, I. B.; ARSLAN, A.. A novel hybrid approach to estimating missing values in databases using k-nearest neighbors and neural networks. **International Journal of Innovative Computing, Information and Control**, v. 7, n. 8, p. 4705-4717, 2012.

BARNARD, J.; MENG, Xiao-Li. Applications of multiple imputation in medical studies: from AIDS to NHANES. **Statistical Methods in Medical Research**, v. 8, n. 1, p. 17-36, 1999.

BATISTA, G.E.A.P.A. **Pré-processamento de Dados em Aprendizado de Máquina Supervisionado**. Tese de Doutorado do Instituto de Ciências Matemática e de Computação da Universidade de São Paulo, São Carlos, 2003.

BATISTA, J. L. F. **Verossimilhança e Máxima Verossimilhança**. Notas de aula do Departamento de Ciências Florestais da Escola Superior de Agricultura ”Luiz de Queiroz”. Universidade de São Paulo, Campus Piracicaba. Disponível em: http://cmq.esalq.usp.br/, 2009.

BILMES, J. A. *et al*. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. **International Computer Science Institute**, v. 4, n. 510, p. 126, 1998.

BISHOP, C. A fast procedure for retraining the multilayer perceptron. **International Journal of Neural Systems**, v. 2, n. 03, p. 229-236, 1991.

BISHOP, C. M. Neural networks and their applications. **Review of scientific instruments**, v. 65, n. 6, p. 1803-1832, 1994.

BISHOP, C. M. **Neural networks for pattern recognition**. New York: Oxford University Press, 1995.

BISHOP, C. M. **Pattern recognition and machine learning**. New York: springer, 2006.

BLACKWELL, M.; HONAKER, J.; KING, G. **Multiple overimputation: a unified approach to measurement error and missing data**. 2012.

BOLFARINE, H.; SANDOVAL, M. C. **Introdução à inferência estatística**. 2. Ed. Rio de janeiro, SMB, 2010.

BRAND, J. *et al*. Multiple imputation as a missing data machine. **Proceedings of the Annual Symposium on Computer Application in Medical Care**. American Medical Informatics Association, 1994.

CASELLA, G., BERGER, R. L. **Inferência Estatística**. Cengage Learning, São Paulo, 2010.

CASTILLO, P.R. On the Use of Data Mining for Imputation. **United Nations Economic Commission for Europe, Conference of European Statisticians, Work Session on Statistical Data Editing**. Paris, France, 8-30 April 2014.

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. **Geoscientific Model Development**, v. 7, n. 3, p. 1247-1250, 2014.

CHARNET, R.; FREIRE, C. A. L.; CHARNET, E. M. R.; BONVINO, H. **Análise de modelos de regressão linear com aplicações**. Campinas, São Paulo, Unicamp, 356p, 1999.

CHENG, T.C. **Very robust statistics in the presence of missing data**. 1998. Tese de Doutorado. London School of Economics and Political Science (University of London), 1998.

CORDEIRO, G.M. **Introdução à Teoria Assintótica**. 22º Colóquio Brasileiro de Matemática, IMPA 26-30 julho, 1999.

COX, D.R. **Principles of Statistical Inference**. Cambridge University Press, 2006.

DALIRI, M. R.; FATTAN, M. Improving the Generalization of Neural Networks by Changing the Structure of Artificial Neuron. **Malaysian Journal of Computer Science**, v. 24, n. 4, p. 195, 2011.

DE JONG, R.; VAN BUUREN, S.; SPIESS, M. Multiple imputation of predictor variables using generalized additive models. **Communications in Statistics - Simulation and Computation**, DOI:10.1080/03610918.2014.911894, 2014.

DE WAAL, T.; PANNEKOEK, J.; SCHOLTUS, S. **Handbook of statistical data editing and imputation**. John Wiley & Sons, 2011.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. **Journal of the Royal Statistical Society. Series B (Methodological)**, Vol. 39, pp.1-38, 1977.

DEVORE, J. L.; BERK, K. N. **Modern Mathematical Statistics with Applications**. Thomson Books, 2007.

DIXON, W. J. Analysis of extreme values. **The Annals of Mathematical Statistics**, p. 488-506, 1950.

DOBSON, A. J.; BARNETT, A. **An Introduction to Generalized Linear Models**. CRC, Second Edition, 2002.

DORESWAMY, K. K.; VASTRAD, C. M. Performance Analysis Of Neural Network Models For Oxazolines And Oxazoles Derivatives Descriptor Dataset. **International Journal of Information Sciences and Techniques (IJIST)**, Vol.3, No.6, November 2013.

DUDA, R. O., HART, P. E., STORK, D. G. **Pattern Classification**. John Wiley & Sons, 2nd Edition, 2001.

DUMA, M. S. **Improving Classification Performance In Missing Insurance Data**. Thesis submitted in fulfilment of the requirements for the degree Doctor Philosophie in Electrical and Electronic Engineering in the Faculty of Engineering and the built environment at the University of Jahannesburg, October 2012.

ENDERS, C. K. **Applied missing data analysis**. Guilford Publications, 2010.

ENNETT, C. M.; FRIZE, M. Validation of a hybrid approach for imputing missing data. **Proceedings of the 25th Annual International Conference of the Engineering in Medicine and Biology Society, IEEE**. p. 1268-1271, 2003.

FICHMAN, M.; CUMMINGS, J. M. Multiple Imputation for Missing Data: Making the Most of What you Know. **Organizational Research Methods**, v. 6, n. 3, p. 282- 308. Disponível em: http://repository.cmu.edu/tepper, 2003.

FISHER, R. A. **On an absolute criterion for fitting frequency curves**. Messenger of Mathmatics, v. 41, p.155-160. Encontrado em http://hdl.handle.net/2440/15165, 1912.

FISHER, R. A. On the Mathematical Foundations of Theoretical Statistics. **Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character**, Vol. 222, pp. 309-368. Encontrado em: http://www.jstor.org/stable/91208, 1922.

FRANÇA, F. O. **Biclusterização na análise de dados incertos**. Tese de Doutorado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação. Campinas - SP, 2010.

GOMES, G. S. S. **Novas Funções de Ativação em Redes Neurais Artificiais Multilayer Perceptron**. Tese em Ciência da Computação, UFPE, Recife, 2010.

GOMES, G. S. S.; LUDEMIR, T. B. REDES NEURAIS ARTIFICIAIS COM FUNÇÕES DE ATIVAÇÃO COMPLEMENTO LOGLOG E PROBIT PARA APROXIMAR FUNÇÕES NA PRESENÇA DE OBSERVAÇÕES EXTREMAS. **Learning and Nonlinear Models, Revista da Sociedade Brasileira de Redes Neurais (SBRN)**, Vol. 6, No. 2, pp. 142-153, 2008.

GOMES, G. S. S.; LUDEMIR, T. B; LIMA, L. M. M. R. Comparison of new activation functions in neural network for forecasting financial time series. **Neural Comput and Applications,** v. 20, n. 3, p. 417–439, 2011.

GRAHAM, J. W. **Missing Data: Analysis and Design**. Springer Science & Business Media, 2012.

GRAHAM, J. W.; HOFER, S. M.; PICCININ, A. M. Analysis with missing data in drug prevention research. **NIDA research monograph**, v. 142, p. 13-13, 1994.

HARTLEY, H. O. Maximum Likelihood Estimation from Incomplete Data. **Biometrics**, v. 14, n. 2, p. 174-194, 1958.

HAUKOOS, J. S.; NEWGARD, C. D. Advanced Statistics: Missing Data in Clinical Research—Part 1: An Introduction and Conceptual Framework. **Society for Academic Emergency Medicine,** Vol. 14, No. 7, 2007.

HAYKIN, S. **Redes Neurais: Princípios e Prática**. Trade. Paulo Martins Engel. Segunda edição-Porto Alegre, Bookman, 2001.

HINTON, G. E. How Neural Networks Learn from Experience. **Scientific American**, v. 267, September 1992.

HINTON, G. E.; FREY, B. J. Using Neural Networks to Monitor for Rare Failures. **Proceedings of the 37th Mechanical Working and Steel Processing Conference**. IRON AND STEEL SOCIETY OF AIME, 1996.

HOGG, R. V., MCKEAN, J., CRAIG, A. T. **Introduction to Mathematical Statistics**. Pearson Education, 7ª Ed., 2012.

HONGHAI, F.; GUOSHUN, C.; CHENG, Y.; BINGRU, Y.; YUMEI, C. A SVM Regression Based Approach to Filling in Missing Values. **Knowledge-Based Intelligent Information and Engineering Systems**. Springer-Verlag Berlin Heidelberg, p. 581-587, 2005.

HRUSCHKA JR, E. R.; EBECKEN, N. F. F. Missing values prediction with K2. **Intelligent Data Analysis**, v. 6, n. 6, p. 557-566, 2002.

HRUSCHKA JR, E. R.; HRUSCHKA, E. R.; EBECKEN, Nelson FF. Bayesian networks for imputation in classification problems**. Journal of intelligent information systems**, v. 29, n. 3, p. 231-252, 2007.

HRUSCHKA, E. R.; GARCIA, A. J. T.; HRUSCHKA JR, E. R.; EBECKEN, N. F .F. On the influence of imputation in classification: practical issues. **Journal of Experimental & Theoretical Artificial Intelligence**, v. 21, n. 1, p. 43-58, 2009.

JEREZ, J. M. *et al*. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. **Artificial intelligence in medicine**, v. 50, n. 2, p. 105-115, 2010.

JOLANI, S.; VAN BUUREN, S.; FRANK, L. E. Combining the complete-data and nonresponse models for drawing imputations under MAR. **Journal of Statistical Computation and Simulation**, v. 83, n. 5, p. 868-879, 2013.

JORDAN, M. I.; BISHOP, C. M. Neural Networks. **ACM Computing Surveys**, Vol. 28, No. 1, March 1996.

JOSSINET, J. Variability of impedivity in normal and pathological breast tissue. **Med. & Biol. Eng. & Comput**, 34: 346-350, 1996.

KÄÄRIK, E. Imputation algorithm using copulas. **Advances in Methodology and Statistics**, v. 3, n. 1, p. 109-120, 2006.

KOVACS, Z. L. **Redes Neurais Artificiais: Fundamentos e Aplicações**. São Paulo, Editora Livraria da Física, 2006.

KURASOVA, O.; MARCINKEVICIUS, V.; MEDVEDEV, V.; RAPECKA, A.; STEFANOVIC, P. Strategies for Big Data Clustering. **IEEE 26th International Conference on Tools with Artificial Intelligence**, 1082-3409, 2014.

LA ROCCA, M.; PERNA, C. DESIGNING NEURAL NETWORKS FOR MODELING BIOLOGICAL DATA: A STATISTICAL PERSPECTIVE. **Mathematical biosciences and engineering: MBE**, v. 11, n. 2, p. 331-342, 2014.

LAKSHMINARAYAN, K., HARP, S.A., SAMAD, T. Imputation of Missing Data in Industrial Databases. **Applied Intelligence**, v. 11, n. 3, p. 259–275, 1999.

LAKSHMINARAYAN, K.; HARP, S.A; GOLDMAN, R.; SAMAD, T. Imputation of missing data using machine learning techniques. **KDD Proceedings, AAAI**, 1996.

LENT, R. **Sobre neurônios, cérebros e pessoas**. Ed. Atheneu, São Paulo, 2011.

LI, H.; ZHANG, K.; JIANG, T. The regularized EM algorithm. **Proceedings of the national conference on artificial intelligence**. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 2005.

LITTLE, M. A.; MCSHARRY, P. E.; HUNTER, E. J.; SPIELMAN, J.; RAMING, L. O. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. **IEEE Transactions on Biomedical Engineering**, 1015-1022, 2009.

LITTLE, R. J. J.; RUBIN, D. B. **Statistical Analysis with Missing Data**. John Wiley & Sons, 1987.

LIUBLINSKA, V. **Sensitivity Analyses in Empirical Studies Plagued with Missing Data**. Dissertation for the degree of Doctor of Philosophy in the subject of Statistics, Havard University, Cambridge, Massachuster, 2013.

LOPES, M.M. **Programação Genética para Otimização de Séries Temporais com Dados Faltantes**. Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências, jullho de 2007.

LUENGO, J.; GARCÍA, S.; HERRERA, F. On the choice of the best imputation methods for missing values considering three groups of classification methods. **Knowledge and information systems**, v. 32, n. 1, p. 77-108, 2012.

MARLIN, B. M. **Missing Data Problems in Machine Learning**. A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy Graduate Department of Computer Science University of Toronto, 2008.

MARWALA, T. (Ed.). **Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques: Knowledge Optimization Techniques**. IGI Global, 2009.

MCCLELLAND, J. L. *et al*. Parallel Distributed Processing: **Explorations in the Microstructures of Cognition**, v. 2: Psychological and Biological Models, 1986.

MCCULLAGH, P.. Marginal likelihood for distance matrices. **Statistica Sinica**, v. 19, n. 2, p. 631, 2009.

MCCULLOCH, W. S.; PITTS, W. H. A Logical Calculus of the Ideas Immanent in Nervous Activity. **The Bulletin of Mathematical Biophysics**, v. 5, p. 115-133, 1943.

MCKNIGHT, P. *et al.* **Missing data: A gentle introduction**. Guilford Press, 2007.

MCLACHLAN, G.; KRISHNAN, T. **The EM Algorithm and Extensions**, John Wiley & Sons, 2nd Edition, 2008.

MENG, Xiao-Li. Multiple-imputation inferences with uncongenial sources of input. **Statistical Science**, v. 9, n. 4, p. 538-558, 1994.

MINGKUI, T.; TSANG, I. W.; WANG, L. Towards ultrahigh dimensional feature selection for big data. **The Journal of Machine Learning Research,** 1371-1429, 2014.

MINSKY, M. Why People Think Computers Can't. **AI Magazine**, v. 3, n. 4, Fall 1982.

MLADENOVIC, V. M.; PORRAT, D.; LUTOVAC, M. D. The direct execution of the expectation-maximization algorithm using symbolic processing. **10th International Conference on Telecommunication in Modern Satellite Cable and Broadcasting Services (TELSIKS),** IEEE, p. 265-268, 2011.

MOHAMED, A. K.; NELWAMONDO, F. V.; MARWALA, T. Estimating Missing Data Using Neural Network Techniques, Principal Component Analysis and Genetic Algorithms.

Proceedings of the Eighteenth Annual Symposium of the Pattern Recognition Association of South. 2007.

MOHAMED, S.; MARWALA, T. Neural Network Based Techniques for Estimating Missing Data in Databases. **Proceedings of the Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa , Langebaan**. 2005.

MOON, T. K. The Expectation-Maximization algorithm. **Signal Processing Magazine, IEEE**, v. 13, n. 6, p. 47-60, 1996.

MORETTIN, L. G. **Estatistica Basica Vol 2 - Inferencia**, MAKRON BOOKS, 2000.

MYRTVEIT, I.; STENSRUD, E.; OLSSON, U. H. Analyzing Data Sets with Missing Data: An Empirical Evaluation of Imputation Methods and Likelihood-Based Methods. **IEEE Transactions on Software Engineering**, v. 27, n. 11, p. 999-1013, 2001.

NEAL, R. M.; HINTON, G. E. A View of the EM Algorithm that Justifies Incremental, Sparse, and orher Variants. **Learning in graphical models**. Springer Netherlands, p. 355-368. 1998.

NELWAMONDO, F. V.; MOHAMED, S.; MARWALA, T. Missing data: A comparison of neural network and expectation maximization techniques. **Current Science (00113891)**, v. 93, n. 11, 2007.

NG-CHI, C. **Robust Statistics in the Presence of Missing Data**. A thesis submitted to the University of London in fulfillment of the requirement for the degree of Doctor of Philosophy, 1998.

NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. **Cad. Saúde Pública**, v. 25, n. 2, p. 268-278, 2009.

OSOBA, O. A. **Noise Benefits in Expectation-Maximization Algorithms**. A Dissertation Presented to Faculty of the USC Graduate School University of Southern California Doctor of Philosophy in Electrical Engineering, 2013.

PAULA, A. V. **Determinação de Parâmetros que caracterizam o Fenômeno da Bioestabilidade em Escoamento Turbulento**. Tese em Engenharia. Porto Alegre, 2013.

PEREIRA, E. A. **Algumas Propostas para Imputação de Dados Faltantes em Teoria de Resposta ao Item**. Dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília como requisito parcial a obtenção do título de Mestre em Estatística, Brasília, Julho de 2014.

PERNÍA-ESPINOZA, A. V.; ORDIERES-MERÉ, J. B.; MARTINEZ-DE-PISÓN, F. J.; GONZÁLEZ-MARCOS, A. TAO-robust backpropagation learning algorithm. **Neural Networks**, v. 18, n. 2, p. 191-204, 2005.

PINTO, W. P. **Uso da Metodologia de Dados Faltantes em Séries Temporais com Aplicações a dados de Concentração (PM10) Observados na Região da Grande Vitória**. Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Ambiental do Centro Tecnológico da Universidade Federal do Espírito Santo, 2013.

PRASS, F. S. **Estudo Comparativo entre Algoritmo de Análise de Agrupamentos em Data Mining**. Dissertação de Mestrado submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de Mestre em Ciência da Computação, Florianópolis, Novembro de 2004.

RAMACHANDRAN, K. M., TSOKOS, C. P. **Mathematical Statistics with Applications**. Elsevier, 2009.

RANDOLPH-GIPS, M. A new neural network to process missing data without Imputation. **Seventh International Conference on Machine Learning and Applications - ICMLA**, IEEE, 2008.

REDNER, R. A.; WALKER, H. F. Mixture densities, maximum likelihood and the EM algorithm. **Society for Industrial and Applied Mathematics - SIAM review**, v. 26, n. 2, p. 195-239, 1984.

REID, N.; COX, D. R. Principles of Statistical Inference. **This paper is based on a talk given at the World Statistics Congress of the International Statistical Institute in Hong Kong**, August 2013..

ROSENBERG, B. A survey of stochastic parameter regression. In: **Annals of Economic and Social Measurement**, Volume 2, number 4, p. 380-396, 1973

ROTH, P. L. Missing Data: A conceptual review for applied psychologists. **Personnel psychology**, v. 47, n. 3, p. 537-560, 1994.

RUBIN, D. B. Inference and missing data. **Biometrika**, v. 63, n. 3, p. 581-592, 1976.

RUBIN, D. B. **Multiple Imputation for Nonresponse in Surveys**. John Wiley & Sons, 1987.

RUBIN, D.B. Inference and Missing Data. **Biometrika**, v. 63, n. 3, p. 581-592, 1976. Disponível em: http://www.jstor.org/.

RUMELHART, D.E. *et al*. Parallel Distributed Processing: **Explorations in the Microstructures of Cognition**, Volume 1: Foundations, Chapter 8, 1986.

SANTANA, I. F.; FILIZOLA-JUNIOR, N. P.; FREITAS, C. E. C. Recuperação de Valores Perdidos de Dados de Desembarque: Uma Aplicação com Dados de Desembarque em Santarém, Estado do Pará, Brasil. **Rev. Bras. Eng. Pesca**. v. 5, n. 1, p. 43-55, 2010.

SARLE, W. S. Neural Networks and Statistical Models. **Proceedings of the Nineteenth Annual SAS Users Group International Conference**, April, 1994.

SCHAFER, J. L. Multiple imputation: a primer. **Statistical Methods in Medical Research**, v. 8, n. 1, p. 3-15, 1999.

SCHAFER, J. L.; GRAHAM, J. W. Missing Data: Our View of the State of the Art. **Psychological Methods**, v. 7, n. 2, p. 147–177, 2002.

SCHAFER, J. L.; OLSEN, M. K. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. **Multivariate behavioral research**, v. 33, n. 4, p. 545-571, 1998.

SEVERINO, A.J. **Metodologia do Trabalho Científico**. Editora Cortez, São Paulo, 2007.

SILVA, H. F. **Um Sistema Integrado de Monitoramento e Previsão de Carga Elétrica de Curto Prazo**. Tese apresentada ao Departamento de Engenharia Elétrica da PUC-RIO como requisito parcial para a obtenção do título de Doutor em Ciências em Engenharia Elétrica, na área de concentração Energia Elétrica, 2001.

SILVA, I. N.; SPATTI, D. H.; FLAUZINO, R.A. **Redes Neurais Artificiais: para engenharia e ciências aplicadas**. Ed. Artliber Ltda, SP, 2010.

SORENSEN, D., GIANOLA, D. **Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics**. Springer-Verlag New York, 2002.

SORJAMAA, A. **Methodologies for the Time Series Prediction and Missing Value Imputation**. Dissertation for the degree of Doctor of Science in Technology to be presented with due permission of the Faculty of Information and Natural Sciences at the Aalto University School of Science and Technology (Espoo, Finland) on the 19th of November, 2010.

SOTO, D. C. F. **Modelos Birnbaum-Saunders para sobrevivência com dados longitudinais**. Tese de Doutorado, Universidade de Sao Paulo, 2014.

SPECHT, D. F. A General Regression Neural Network. **IEEE Transaction on Neural Networks**, v. 2, n. 6, p. 568-576, 1991.

SRIDEVI, S. *et al*. Imputation for the analysis of missing values and prediction of time series data. **International Conference on Recent Trends in Information Technology (ICRTIT), IEEE**, p. 1158-1163, 2011.

SRIDEVI, S., RAJARAM, S.,PARTHIBAN C., SIBIARASAN, S., SWADHIKAR. Imputation for the Analysis of Missing Values and Prediction of Time Series Data. IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011 MIT, Anna University, Chennai. June 3-5, 2011.

SSALI, G.; MARWALA, T. Estimation of missing data using computational intelligence and decision trees. **Disponível em: http://arxiv.org/abs/0709.1640**, 2007.

TWOMEY, J. M.; SMITH, A. E. Validation and verification. **Artificial neural networks for civil engineers: fundamentals and applications, ASCE**, New York, p. 44-64, 1997.

UEDA, N.; NAKANO, R. Deterministic annealing EM algorithm. **Neural Networks**, v. 11, n. 2, p. 271-282, 1998.

VAN BUUREN, S. **Flexible Imputation of Missing Data**. CRC Press, 2012.

VERONEZE, R. **Tratamento de dados faltantes empregando biclusterização com imputação múltipla**. Dissertação de Mestrado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos exigidos para a obtenção do título de Mestre em Engenharia Elétrica. Área de Concentração: Engenharia de Computação, Campinas – SP – Brasil Junho de 2011.

WANG, Y., Li, B., LUO, R., CHEN, Y., XU, N., YANG, H. Energy efficient neural networks for big data analytics. **In Design, Automation and Test in Europe Conference and Exhibition , IEEE**, 2014.

WARNER, B.; MISRA, M. **Understanding Neural Networks as Statistical Tools**. American Statistical Association, v. 50, n. 4, November 1996.

WAZLAWICK, R. S. **Metodologia de Pesquisa para Ciência da Computação**. Elsevier, Rio de Janeiro, 6ª reimpressão, 2009.

WEN, Yuh-horng; LEE, Tsu-tian; CHO, Hsun-Jung. Missing data treatment and data fusion toward travel time estimation for ATIS. **Journal of the Eastern Asia Society for Transportation Studies**, v. 6, p. 2546-2560, 2005.

WERBOS, P. J. Backpropagation Through Time: What it Does and How to Do it. **Proceeding of the IEEE**, v. 78, n. 10, October 1990.

WIDROW, B.; HOFF, M. E. Adaptive Switching Circuits. **IRE WESCON Convention Record**, Part 4, New York IRE, pp. 96–104, 1960.

WIDROW, B.; LEHR, M. A. 30 years of adaptive neural networks: perceptron, madaline, and backpropagation. **Proceedings of the IEEE**, v. 78, n. 9, p. 1415-1442, 1990.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. **Climate research**, v. 30, n. 1, p. 79, 2005.

WU, C. F. J.. On the convergence properties of the EM algorithm. **The Annals of statistics**, v. 11, n. 1, p. 95-103, 1983.

YANG, H. H.; MURATA, N.; AMARI, Shun-ichi. Statistical inference: learning in artificial neural networks. **Trends in Cognitive Sciences**, v. 2, n. 1, p. 4-10, 1998.

Yeh, I. C. Modeling slump flow of concrete using second-order regressions and artificial neural networks. **Cement and Concrete Composites**, 474-480, 2007.

ZANCHETTIN, C.; LUDEMIR, T. B.; ALMEIDA, L. M. Hybrid Training Method for MLP: Optimization of Architecture and Training. **IEEE Transaction on Systems, Man, and Cybernetics, Part B: Cybernetics**, v. 41, n. 4, p. 1097-1109, 2011.

Trabalhos Publicados Pelo Autor

RIBEIRO, E. A.; FARIAS, A. F.; COLACO JR., M.; MONTESCO, C. A. E. UTILIZANDO REDES NEURAIS ARTIFICIAIS MLP PARA CLASSIFICAÇÃO DE CÉLULAS CANCERÍGENAS EM AMOSTRAS DE TECIDOS MAMÁRIOS. **Simpósio Nacional de Probabilidade e Estatística**, 2014.

MOTA, F. S.; RIBEIRO, E. A.; MONTESCO, C. A. E. A Constituição da Provisão de Devedores Duvidosos Utilizando Aprendizado de Máquina. **Simpósio Nacional de Probabilidade e Estatística**, 2014.

RIBEIRO, E.A.; NUNES, M.A.S.N. PROSPECTION IN SIMULATOR ELECTRICAL ENERGY. **Revista GEINTEC: gestao, inovacao e tecnologias** , v. 4, p. 453-459, 2014.