Aplicação e avaliação de técnicas de mineração de dados no processo de predição do índice pluviométrico no Norte do estado do Piauí

Ruymar Araújo Mendes Neto[1], **Leinylson Fontinele Pereira**[1]

[1]Universidade Estadual do Piauí (UESPI)
Campus Alexandre Alves de Oliveira – Parnaíba, PI – Brazil

ruymararaujo@gmail.com

Abstract. *The amount of rain is very valuable information when carrying out various activities of society, such as agriculture, tourism and aeronautics. Being important also for day-to-day decisions. One of the most important meteorological variables is the rainfall index in a particular are. In this paper, two data mining techniques were performed, Artificial Neural Networks (ANN) model called Multi-layer Perceptron (MLP), is used in this work to predict a monthly rainfall amount in the coastal region of Piaui, Brazil. and Decision Tree (J48) to perform the categorization of the parameter that has the greatest influence on the prediction of the rainfall index.*

Resumo. *A quantidade de chuva é uma informação muito valiosa na realização de diversas atividades da sociedade, como por exemplo a agricultura, turismo e aeronáutica. Na meteorologia, o índice pluviométrico é uma das variáveis mais importantes para elaboração de métodos preditivos de uma região. Neste trabalho foram realizadas duas técnicas de mineração de dados, Redes Neurais Artificiais (RNA) Multi Layer Perceptron (MLP) para prever a precipitação mensal no litoral do estado do Piauí, e Árvore de Decisão (J48) para realizar a categorização do parâmetro que possui maior influência na predição do índice pluviométrico.*

1. Introdução

Há muito tempo o ser humano busca prever as condições climáticas do planeta terra. Desde o homem primitivo já se via a importância em realizar previsões meteorológicas. Estudos realizados pelo filósofo Aristóteles (384 a.C - 322 a.C) na Grécia antiga já buscavam compreender de que forma acontecia os fenômenos atmosféricos, como as formações de nuvens, precipitações pluviais, vento, raios e trovões [Cunha 1997].

A princípio, o êxito das previsões climáticas dependia da experiência e da desenvoltura do previsor. Com o avanço tecnológico, essa dependência passou a ser responsabilidade de modelos matemáticos e sistemas computadorizados, atrelados a um grande volume de dados [de Carvalho et al. 2013]

Em meados da década de 50, os primeiros testes computadorizados de previsão meteorológica eram realizados nos EUA, pelo computador ENIAC. Técnica que ficou conhecida como previsão numérica de tempo. Os teste consistiam em um modelo matemático baseado em leis físicas que buscava prever o estado final na atmosfera a partir de

um outro inicial conhecido. Algumas décadas depois, esse modelo foi melhorado dando origem a técnica de previsão de tempo por conjuntos [Moura 1996].

Por mais que as técnicas de previsão meteorológica tenham evoluído com o passar do tempo, ainda são incertas e não ajudam de maneira significante aqueles que dependem do comportamento climático. Ressaltando que varia muito de acordo com a antecedência adotada, podendo sofrer alterações na exatidão e nos resultados previstos. Em outras palavras, a previsão para hoje é mais concisa do que para amanhã [Oliveira Filho 2007].

A previsão do tempo é muito importante para a tomada de decisão por órgãos públicos, por exemplo no combate a catástrofes climáticas e nos cuidados com a população. Também tem papel fundamental para a aviação, esporte, construção civil e para o turismo e agricultura, atividades essas que dependem diretamente das condições climáticas para serem bem-sucedidas[Oliveira Filho 2007].

O estado do Piauí é formado por duas estações bem pré-definidas. São elas a estação chuvosa, que geralmente vai do mês de janeiro até maio, e a estação seca que predomina no segundo semestre no ano. A microrregião norte piauiense é formada por 14 municípios, que abrangem toda faixa litorânea do estado e tem como uma das principais atividades econômicas o turismo e agricultura. [Grande et al. 2013]

A mineração de dados é uma etapa do processo conhecido por *Knowledge Discovery in DataBases* (KDD), i.e., descoberta de conhecimento em base de dados. Ele busca identificar nos dados, padrões escondidos e potencialmente uteis, que contenham informações, antes desperdiçadas e que passam a ser usadas a favor do modelo, por meio de ferramentas de visualização e técnicas de inteligencia artificial [Fayyad et al. 1996].

O KDD é formado de algumas etapas, sendo a mineração de dados a mais importante delas. Consiste em extrair padrões, irregularidades e regras, com o intuito de transformar dados, que antes estavam ocultos, em informações relevantes que gerem conhecimento para tomada de decisão e/ou avaliação de resultados [Dantas et al. 2008].

A mineração de dados pode ser classificada pela capacidade de realizar tarefas. Entre elas as mais comuns são: predição, associação, agrupamento e classificação. a tarefa de predição consiste em definir variáveis futuras a partir de variáveis passadas já conhecidas [Larose and Larose 2014]. Para realização da tarefa de predição, uma técnica bastante empregada é a que faz uso de Redes Neurais Artificiais (RNA), que se baseiam no comportamento do neurônio biológico, composto por um conjunto de unidades de entrada e saída interligadas por camadas intermediárias.[Da Silva et al. 2010]

Este trabalho utiliza de Redes Neurais Artificiais com o algoritmo *Backpropagation*, que tem como objetivo prever o índice pluviométrico no norte no estado do Piauí. Como auxílio para o modelo, é usada a técnica de Árvore de Decisão, com o algoritmo J48. Semelhante a estrutura de uma árvore, é uma importante aliada na tomada de decisão, definindo qual parâmetro tem maior importância na realização de tarefas.

Além dessa Seção introdutória o presente artigo é organizado de modo que na Seção 2 são apresentados as pesquisas que inspiraram este trabalho, bem como lacunas que ainda não haviam sido elucidadas. A fundamentação teórica está descrita nas Seções 3 e 4. Na Seção 5 são apresentados os métodos utilizados para a construção do *dataset*, bem como os algoritmos utilizados na tarefa de classificação. Os resultados obtidos são

apresentados na Seção 6. Por fim, na Seção 7 são apresentadas as considerações finais sobre a pesquisa, além de possibilidades de trabalhos futuros.

2. Trabalhos Relacionados

A Mineração de Dados é uma área multidisciplinar que envolve outras grandes áreas. Vem ganhando cada dia mais espaço graças ao rápido avanço tecnológico atrelado a um grande volume de dados.[Feelders et al. 2000]

Souza e Souza (2010) modelaram a relação entre chuva e vazão em base mensal na bacia hidrografia do Rio Piancó, utilizando MLP. A melhor rede teve uma eficiência de 77% [Sousa and de Sousa 2010], mas sem a análise de qual variável poderia ter maior influência neste resultado.

Santhanam e Subhajini (2011) comparam uma MLP com uma Rede de Base Radial na previsão meteorológica. Os resultados mostraram a Rede de Base Radial com uma acurácia de 88,2% e a MLP com 82% [Santhanam and Subhajini 2011]. O trabalho demonstra a aplicação das RNAs, mas sem a percepção dos fatores determinantes para os resultados.

Ferraudo e Ferraudo (2010) apresentam um estudo de série de precipitação mensal acumulada. Constatou-se que a MLP teve um bom desempenho e capacidade de generalização na previsão de seres de precipitação acumulada mensal [Ferraudo and Ferraudo 2010].

3. Previsão do Tempo

Em meados de 1950, nos Estados Unidos, o professor norueguês Ragnar Fjortoft juntamente com Von Neumann realizaram os primeiros testes bem-sucedidos de previsão numérica de tempo no computador ENIAC, dando passo inicial e importante na evolução da meteorologia moderna [Moura 1996].

3.1. Previsão Numérica de Tempo

Um modelo de previsão numérica de tempo (PNT) é responsável por resolver um conjunto de equações matemáticas baseadas em leis físicas, aplicadas para a atmosfera, de modo a prever seu estado futuro a partir de uma situação inicial já conhecida. Em suma, um modelo PNT é desenvolvido em computador para simular o comportamento da atmosfera [de Carvalho et al. 2013].

Os modelos de PNT podem ser classificados em global e regional, variando de acordo com sua escala espacial. No modelo regional, conhecido por *Mesoscale Meteorological Model*(MM5) a escala atinge a marca dos 50 Km, analisando fenômenos mais específicos de uma região. A sua vantagem é a melhor representação dos fenômenos espaciais e temporais em uma escala menor. Modelo usado por: Instituto de Aeronáutica e Espaço (IEA), Instituto de Controle de Espaço Aéreo (ICEA) e Centro Nacional de Meteorologia Aeronáutica (CNMA) [Iriart et al. 2011].

No modelo global, conhecido por *Weather Research and Forecasting Model*(WRF) a escala atinge a ordem de 200 Km, visando identificar o comportamento geral da atmosfera. O WRF é a última geração de modelo numérico de previsão do tempo, responsável por identificar os fenômenos meteorológicos de larga escala, chamados de

sinóticos. Seu diferencial está nos múltiplos núcleos dinâmicos, no sistema variável de assimilação de dados tridimensional e na sua estrutura de software, que permite o paralelismo computacional [Iriart et al. 2011].

O modelo WRF foi desenvolvido por um esforço conjunto de diversos órgãos americanos, entre eles o *National Oceanic and Atmospheric Administration*(NOAA) e o *National Center for Atmospheric Research*(NCAR), que é operado pela *University Corporation for Atmospheric Research*(UCAR). Foi criado para fins de aplicação tanto em pesquisas como também operacionalmente em previsão numérica de tempo, para suceder o modelo MM5 [Silva and Fisch 2014].

3.2. Previsão Numérica por Conjuntos

Um dos problemas encontrados no modelo de PNT é a determinação exata do estado inicial. A atmosfera é um sistema não linear, que depende extremamente das condições inicias. Essas sensibilidade também é vista na PNT, de tal forma que se iniciada com valores ligeiramente diferentes, em pouco tempo o êxito nos resultados da previsão estará comprometido [Meller 2012, Iriart et al. 2011].

A previsão numérica por conjuntos é a técnica que visa melhorar e ampliar os resultados das previsões numéricas por tempo. Leva em consideração a natureza desordenada da atmosfera, inserindo pertubações nas suas condições iniciais, obtendo N cenários distintos e equiprováveis. Podendo determinar as incertezas ou a probabilidade de ocorrência do resultado [Meller 2012, Silva et al. 2008].

As primeiras pesquisas sobre previsão por conjunto aconteceram na década de 80, no NCEP e no *European Centre for Medium-Range Weather Forecasting*(ECMWF), mas somente na década seguinte as pesquisas experimentais começaram a funcionar em modo operacional, ganhando espaço em outros centros meteorológicos como CPTEC/INPE [Meller 2012]. O objetivo principal do modelo é diminuir o impacto da incerteza em relação ao estado inicial da atmosfera na previsão final, que graças as pertubações sofridas, proporciona mais informação para os resultados [Silva et al. 2008].

4. Descoberta de Conhecimento em Base de Dados

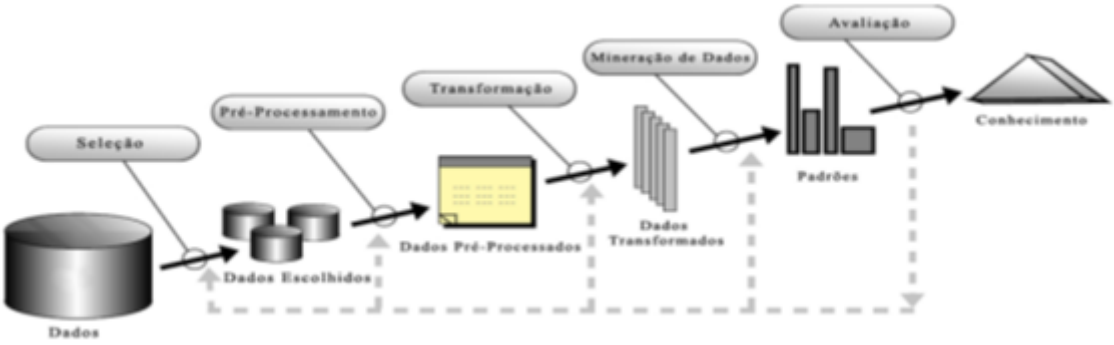
A mineração de dados é uma etapa de um processo mais amplo conhecido como *Knowledge Discovery in Databases*(KDD), ou Descoberta de Conhecimento em Base de Dados. É um processo não trivial de identificação de padrões válidos e potencialmente úteis em uma base de dados. Responsável por descobrir tendências após analise de um grande conjunto de dados, tem como principal etapa a mineração de dados, que consiste de análise e algoritmos específicos que produzem uma relação particular de padrões a partir da base de dados [Fayyad et al. 1996]. O processo de KDD é formado por algumas etapas, como pode ser observado na Figura 1.

4.1. Pré-processamento

A etapa de pré-processamento é composta por várias técnicas para realizar a preparação dos dados. Esse processo pode levar até 50% de todo o trabalho [Olson and Delen 2008].

1. **Limpeza:** Os dados na maioria dos casos vêm inconsistentes, incompletos ou com valores errados. Isso torna os dados “sujos”, o que compromete as outras etapas

Figura 1. Representação do processo de KDD [Camilo and Silva 2009].



do processo de descoberta de conhecimento. A limpeza de dados é uma etapa na qual se busca tratar esse problema por meio de algumas técnicas como: remoção de valores problemáticos, aplicação de técnicas de agrupamento ou de atribuição de valores. Esta etapa visa auxiliar na descoberta dos melhores valores a serem usados na etapas seguintes [Camilo and Silva 2009].

2. **Integração:** Nesta etapa, busca-se resolver dados inconsistentes ou redundantes, dependência entre valores e valores conflitantes. Deve ser realizada de forma cuidadosa pela sua importância. Na maioria dos casos é necessário para realização da mineração [Feelders et al. 2000].
3. **Redução:** O processo de mineração geralmente trabalha com um grande volume de dados, que em alguns casos, compromete o processo de análise e torna a mineração impossível [Han et al. 2011]. Para isso são realizadas estratégias como: seleção de um subconjunto de dados, discretização, criação de estruturas otimizadas e redução de dissimilaridade, que tem como finalidade reduzir o volume de dados, sem comprometer as informações relevantes [Han et al. 2011].
4. **Transformação:** Nesta etapa, a finalidade é tornar os processos de mineração e análise de informações mais eficientes e melhor compreensíveis [Feelders et al. 2000]. As estratégias mais usadas são: agrupamento (agrupa valores por faixas sumarizadas), normalização (coloca os atributos em mesma escala), suavização (faz a remoção de valores errados), generalização (converte valores específicos para genéricos) e a criação de outros atributos, gerados a partir de variáveis já existentes [Han et al. 2011].

4.2. Mineração de Dados (*Data Mining*)

É a principal etapa do processo de descoberta de conhecimento em base de dados. Consiste em extrair padrões nos dados, para então serem analisados, gerando o conhecimento inferido [Simões 2008].

A técnica de mineração é considerada uma área multidisciplinar, que esta localizada entre outras grandes áreas como: inteligência artificial (aprendizado de máquina), banco de dados (modelagem de grande massa de dados) e estatística (avaliação e validação de grandes resultados) (Figura 2) [Feelders et al. 2000].

De acordo com Han e colaboradores, a mineração de dados varia de acordo com o método usado, podendo ser um processo de aprendizado supervisionado (preditivo) ou

Figura 2. Mineração de dados entre outras áreas [Han et al. 2011].

um método de aprendizado não-supervisionado (descritivo) [Han et al. 2011].

As atividades supervisionadas buscam identificar a classe de uma nova amostra de dados (tendência futura) a partir do conhecimento adquirido pela análise de amostras já conhecidas. Já as atividades não-supervisionadas trabalham com um conjunto de dados que não possuem classe determinada, buscando identificar padrões de comportamento comuns nesses dados [Simões 2008]. Após a determinação do método a ser usado, a mineração de dados é definida de acordo com a tarefa a qual ela possui objetivo, as tarefas mais comuns são:

- **Descrição(*Description*):** Tarefa que busca prever a influência de variáveis dentro da base de dados. Descreve os padrões e tendências reveladas e oferece uma possível interpretação para os resultados obtidos, comprovando a influência de certas variáveis nos resultados [Larose and Larose 2014].
- **Classificação(*Classification*):** Uma das tarefas mais comuns. Do tipo supervisionada, consistem em analisar um conjunto de dados e dividi-lo em classes, aprendendo como classificar devidamente novos registros em sua classe correspondente de acordo com os dados conhecidos. Pode ser usado por exemplo, para definição de qual turma de uma escola é mais apropriada para determinado aluno [Camilo and Silva 2009].
- **Estimação(*Estimation*) ou Regressão (*Regresion*):** Similar a classificação, no entanto é usada quando os registros são identificados por valores numéricos, e não categóricos. Como por exemplo, quando se quer saber os batimentos de um paciente de acordo com a sua idade e sexo [Camilo and Silva 2009].
- **Agrupamento(*Clustering*):** Busca identificar e aproximar valores similares. Do tipo não-supervisionado, difere da classificação por não necessitar de registros previamente categorizados responsável apenas por identificar os grupos de dados semelhantes [Larose and Larose 2014].
- **Associação(*Association*):** Consiste em identificar atributos relacionados. Técnica muito conhecida devido aos bons resultados obtidos após analise de cestas de compras, que identifica quais produtos possuem mais tendência de serem comprados

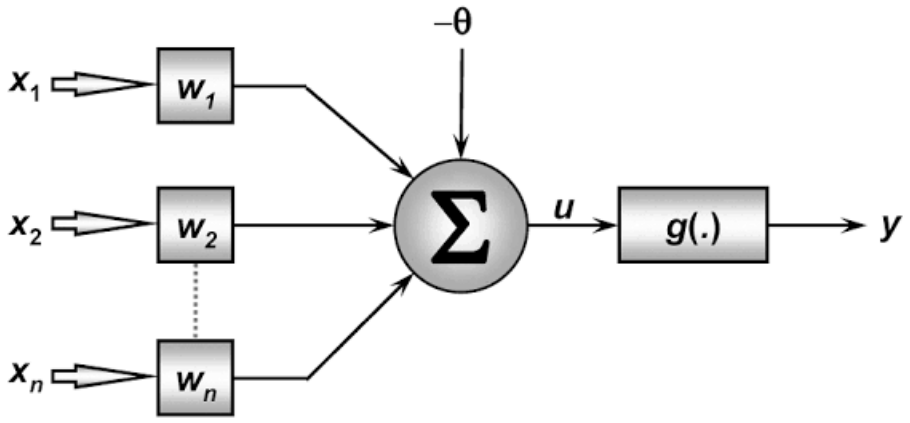
simultaneamente. O aprendizado é não-supervisionado e pode ser representado na forma de SE atributo X, então atributo Y [Larose and Larose 2014].

- **Predição(*Predition*):** Similar as tarefas de classificação e estimação, visa descobrir o valor futuro de um determinado atributo. Muito usado para predizer a quantidade de chuva de um determinado período ou o percentual de crescimento de uma empresa. Consiste em um aprendizado supervisionado e tem como principal técnica o uso de Redes Neurais Artificiais (RNAs) [Marcoulides 2005].

4.3. Rede Neural Artificial (*Artificial Neural Network*)

A neuro-computação é uma área responsável pela simulação computacional do cérebro humano. As Redes Neurais Artificiais (RNA) são sistemas computacionais que tem como unidade de processamento o neurônio artificial (Figura 3) [Muralikrishna 2009], o qual é inspirado no neurônio biológico. O neurônio biológico é composto pelo corpo celular, no qual ocorre o processamento de informação, os dendritos e os axônios responsáveis pela comunicação (sinapse) entre os neurônios e os terminais sinápticos [Moraes and Arraes 2015].

Figura 3. Modelo de um Neurônio Artificial [Da Silva et al. 2010]



Um neurônio artificial é formado pelas suas entradas ($x_1, x_2, ...x_n$), que são multiplicados pelos seus respectivos pesos ($w_1, w_2, ...w_n$). Seus resultados são então somados (E) junto ao limiar de ativação (-0). O resultado (u) passa pela função de ativação (g(.)) produzindo a saída do neurônio (y). A associação de múltiplos neurônios gera uma Rede Neural Artificial [Da Silva et al. 2010].

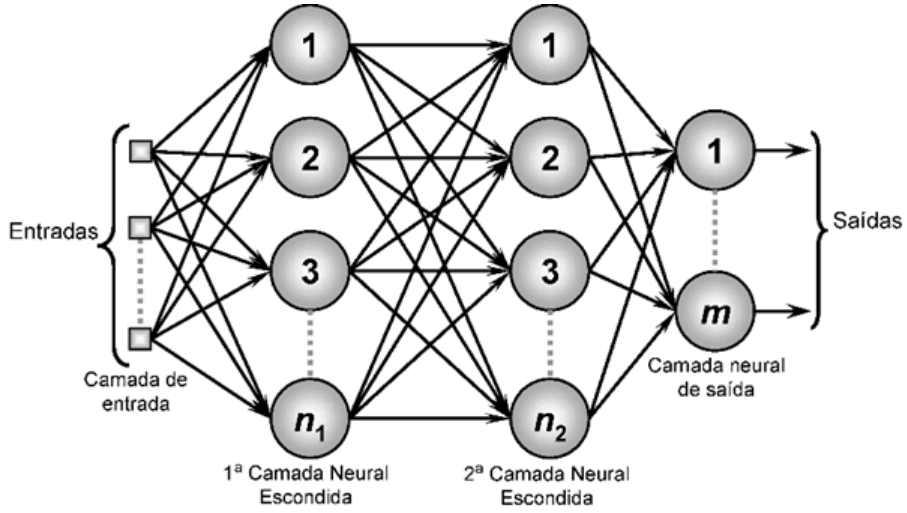
As RNAs têm diferentes tipos de configuração. As mais usadas são: de Camada Única (*Single Layer Net*), Recorrentes (*Recurrent Networks*) e Multicamadas (*Multi Layer Net*), da qual pertence o modelo *Perceptron* de Multicamadas ou MLP(*Multi Layer Perceptron*) [Muralikrishna 2009]. Os diferentes tipos de configurações de uma RNA são definidos de acordo com o arranjo em camadas, processamento inteiro dos neurônios, número de camadas e o tipo de conexão entre elas [Muralikrishna 2009].

4.3.1. Perceptron de Multicamadas(*Multi Layer Perceptron*)

Uma RNA do tipo *Multi Layer Perceptron* é formada por duas ou mais camadas, sendo uma de saída e uma ou mais intermediarias(ocultas). Existe a camada de entrada, res-

ponsável por receber os valores oriundos do *dataset*, mas não é contabilizada por não realizar processamento. Representa-se na Figura 4, uma MLP de três camadas referentes as intermediárias e a terceira referente a camada de saída.

Figura 4. Rede Neural Artificial [Da Silva et al. 2010]



Basicamente, o processo de uma MLP consiste em uma soma ponderada de entradas X_n, pelos seus pesos W_n com o acréscimo do bias B_k, gerando V_k, que é ativado pela função de ativação F, gerando a saída Y_k. O bias tem a função de reforçar ou inibir as entradas dos neurônios, baseando-se no seu sinal positivo ou negativo. Se o neurônio for ativado, a função de ativação é usada gerando sua respectiva saída [Muralikrishna 2009].

4.3.2. Árvore de Decisão (*Decision Tree*)

Redes neurais podem trabalhar em conjunto com outras técnicas de processamento de dados, permitindo que se use o conhecimento acumulado em determinada área de aplicação, devido aos dados que, já pré-processados, evitam que a RNA trabalhe com dados brutos para trabalhar com informações qualificadas [Santos et al. 2005].

Amplamente utilizadas em algoritmos de classificação supervisionada, árvores de decisão são representações simples de conhecimento. Consiste em nodos que representam os atributos, de arcos que recebem os valores e de nodos folhas que representam as diferentes classes do conjunto de treinamento [Shiba et al. 2005].

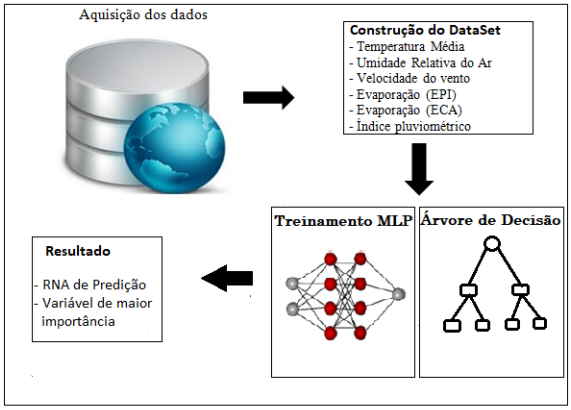
Tem como resultado um conjunto de regras encadeadas do tipo SE - ENTÃO, que forma uma estrutura hierárquica semelhante a de uma árvore. Permitindo mapear um conjunto de dados já rotulados, descobrindo o melhor conjunto de regras para aquela classificação, para em seguida aplicar as regras descobertas em valores desconhecidos e classificá-los [Muralikrishna 2009].

Neste trabalho foram utilizadas técnicas de Redes Neurais Artificiais para previsão de dados meteorológicos, bem como o uso do algoritmo de Árvore de Decisão a fim de verificar quais das variáveis possui maior influência no processo de previsão. A informação do atributo mais relevante pode levar a resultados de predição ainda melhores.

5. Materiais e Métodos

Para a realização deste trabalho, seguiu-se uma sequencia de etapas, todo o processo de KDD foi executado (Figura 5), desde a aquisição e manipulação dos dados para construção do *dataset*, até a análise dos resultados para extração do conhecimento.

Figura 5. Etapas da Metodologia



Fonte:Elaborada pelo autor

5.1. Construção do *dataset*

Na pesquisa foram usados dados meteorológicos diários correspondentes a região norte do Piauí, cedidos pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA) e colhidos pelo Instituto Nacional de Meteorologia (INMET). Os dados abrangem um período de 25 anos (1990 a 2015).

Para a realização das técnicas de mineração, os dados passaram por um pré-processamento envolvendo as etapas de: seleção, limpeza e transformação. Foram dispostos na forma de tabelas, em que cada tabela corresponde a um mês do ano.

As tabelas continham 14 variáveis, nas quais 7 foram selecionadas para os testes, fazendo o total de 282 observações meteorológicas. A primeira variável corresponde ao mês referente e as outras 6 correspondem a variáveis climáticas (Tabela 1).

Tabela 1. Variáveis meteorológicas

VARIÁVEL	TIPO
Temperatura média	Real
Umidade relativa do ar	Inteiro
Velocidade do vento	Real
Evaporação (EPI)	Real
Evaporação (ECA)	Real
Índice pluviométrico	Real

Fonte: Elaborada pelo autor

Após selecionadas as variáveis, os dados passaram pela etapa de limpeza, na qual foi possível constatar dados faltosos (Tabela 2).

Tabela 2. Relação de dados inexistentes no *dataset* original

DADOS FALTANTES
Abril a Dezembro de 2009
2010
Dezembro de 2013
Junho a Dezembro de 2015

Fonte:Elaborada pelo autor

Para realização da limpeza dos dados, inseriu-se o valor nulo nas tabelas incompletas, para total preenchimento de todas as instâncias a serem usadas. O valor final usado, corresponde a média mensal de cada variável meteorológica, com exceção do índice pluviométrico, que já corresponde ao total mensal.

Após o pré-processamento, os dados preparados foram utilizados na construção do *dataset*. A base de dados foi implementada para uso no treinamento das Redes Neurais Artificiais, bem como para no algoritmo de Árvore de Decisão.

5.2. Rede Neural Artificial

A técnica desenvolvida foi a de Rede Neural Artificial, do tipo *MultiLayer Perceptron* (MLP), com o algoritmo de aprendizado *Backpropagation*. Contendo 7 parâmetros de entrada (mês, temperatura média, umidade do ar, velocidade do vento, evaporação EPI, evaporação ECA, índice pluviométrico do mês atual) e 1 parâmetro de saída (índice pluviométrico do mês seguinte).

Para garantir e avaliar a generalização da rede, os dados foram selecionados aleatoriamente e divididos em conjuntos de treinamento, validação e teste (Tabela 3). O número de neurônios das camadas escondidas foi gerado tendo como base o teorema de Kolmogorov [Kolmogorov 1957], no qual observa-se que o número de neurônios das camadas escondidas é igual $2 * n + 1$, com n representando o número de entradas. Sendo assim, definiu-se que a primeira camada escondida deveria varias de 2 a 2n + 1, e o da segunda camada escondida varia de 0 (zero) a (2N + 1)/2. Para cada configuração da RNA foram efetuados 10 treinamentos. Sendo assim, o valor de neurônios da primeira camada escondida variou entre 2 e 15 neurônios, e o da segunda camada variou entre 0 e 7 neurônios.

Tabela 3. Distribuição dos dados

Etapa	Dados (%)
Treinamento	60
Validação	20
Teste	20

Fonte:Elaborada pelo autor

Para assegurar a generalização e minimizar a especialização da rede foi usada a técnica de validação cruzada (*cross-validation*) que corresponde a uma técnica estatística na qual usa o erro quadrático médio para comparar os resultados obtidos com os resultados esperados, e então pode avaliar das topologias candidatas, qual teve melhor resultado. Os

testes foram realizados em um conjunto de dados diferentes daqueles usados nos ajustes de seus parâmetros anteriores.

O número total de épocas de treinamento foi estipulado em 200, os testes realizados em cada topologia foram 10 e o número de validações cruzadas foram 15. Toda a implementação da rede foi realizada no software MATLAB® na versão R2013a(1.0.604). O software usa linguagem de alto nível e seu ambiente de desenvolvimento é muito usado na computação, matemática, visualização e programação. Nas fases de treinamento e validação das RNAs, foi usado uma máquina de processador core i7, 8GB de memória RAM com um sistema de 64 bits. O tempo de processamento foi de aproximadamente 120 horas.

5.3. Árvore de Decisão

Para implementação da técnica de Árvore de Decisão, foi usado a ferramenta WEKA (*Waikato Environment for Knowledge Analysis*) com o algoritmo J48 (Árvore de Decisão C4.5). Para aplicação da técnica foi necessário criar um arquivo com extensão *.arff* para a execução do algoritmo.

Após a criação do arquivo, foi realizado o procedimento de mineração de dados por meio do algoritmo J48. Com isso foi possível determinar qual parâmetro é o mais importante na realização da tarefa de predição de índices pluviométricos. O parâmetro é identificado por meio da visualização dos resultados da técnica, que bem semelhante a uma árvore, se destaca no nó raiz da mesma.

6. Resultados e Discussão

Os resultados do treinamento das Redes Neurais Artificiais e da Árvore de Decisão foram analisados de forma independente por se tratarem de objetivos diferentes.

6.1. Rede Neural Artificial

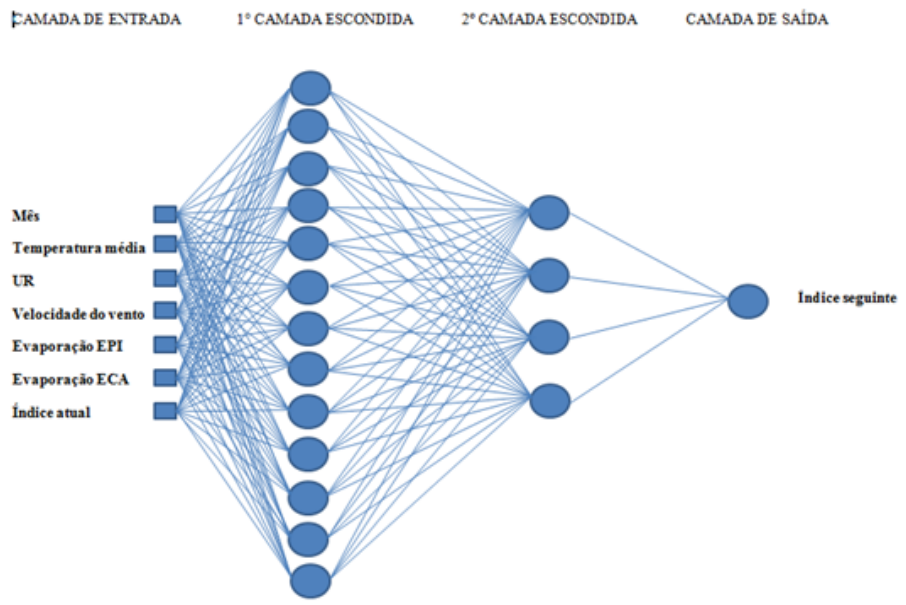
Com a implementação da RNA permitiu-se a obtenção de um sistema inteligente capaz de analisar todas as suas entradas e calcular a saída relacionada em comparação com a desejada, permitindo o aprendizado da rede. Por meio da técnica de validação cruzada que foi aplicada, assegurou-se a generalização da rede, sendo possível identificar a melhor topologia encontrada. As RNAs geradas foram criadas variando a sua quantidade de neurônios na camada escondida, indo de 2 a 15 neurônios na primeira camada escondida e de 0 a 7 neurônios na segunda camada escondida.

Ao final dos testes detectou-se que a melhor rede teve um índice de acerto de 83,8% na validação e sua melhor topologia foi a 13:4:1, treze neurônios na primeira camada escondida, quatro neurônios na segunda camada escondida e um neurônio na camada de saída (Figura 6).

Para avaliação dos resultados foram utilizadas duas métricas diferentes: a correlação coeficiente (CORR2) para indicar a média de todos os erros absolutos percentuais, fornecendo uma indicação do tamanho médio do erro, e o erro médio absoluto percentual (MAPE) para indicar o quão perto os resultados preditos são dos reais.

O treinamento consiste em um processo repetitivo de ajustes de peso que usa todos os valores de entrada da rede ate que o erro quadrático médio das saídas esteja entre um

Figura 6. Representação esquemática da melhor topologia da RNA encontrada pelo processo de validação cruzada.



Fonte:Elaborada pelo autor

valor aceitável. O treinamento da rede durou 120 horas. Ao fim do treinamento, após análise dos gráficos, constatou-se as taxas de acerto correspondente a validação de 83,8% (Figura 7), 77,5% para o treinamento e 72,9% para os testes. Sendo uma série temporal, e com uma saída linear (índice pluviométrico), o resultado obtido foi bastante promissor, uma vez que foi possível prever com um alto índice de acerto o valor exato do volume de chuva em um determinado período do ano, trazendo inovações para essa aplicação, uma vez que trabalhos anteriores apenas indicavam a possibilidade de chuva ou não.

6.2. Árvore de Decisão

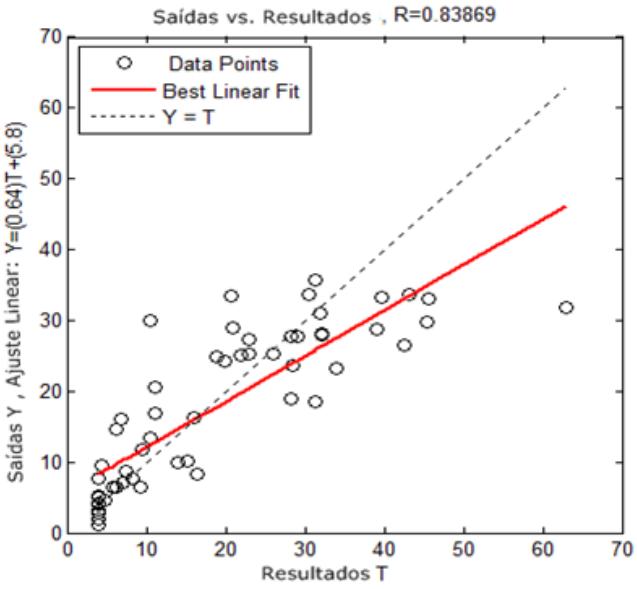
A aplicação da técnica de mineração de dados por meio da árvore de decisão possibilitou a identificação do parâmetro de maior probabilidade de importância. Sendo o mesmo o Índice Pluviométrico Atual (Figura 8). Esta constatação foi possível, pois o parâmetro Índice Pluviométrico Atual veio no nó raiz da árvore, demonstrando a importância desse elemento para a tarefa de predição do índice pluviométrico no norte do estado do Piauí.

7. Conclusões

A Rede neural produzida nesse trabalho teve como objetivo criar um modelo otimizado para predição do índice pluviométrico de um mês baseado nas variáveis já conhecidas. Uma vantagem do uso de redes neurais para este fim é o custo reduzido para ser empregado.

A técnica de mineração de dados árvore de decisão se mostrou eficaz na determinação do parâmetro mais importante na realização da tarefa de predição. O resultado das duas técnicas geram um acumulo de informação que poderá ser usado em novos testes em RNA possibilitando uma melhoria nos resultados encontrados.

Figura 7. Validação dos Dados



Fonte:Elaborada pelo autor

Para trabalhos futuros pode-se aplicar os mesmos procedimentos com outras bases de dados, bem como utilizar outras técnicas de mineração de dados como Algoritmos Genéticos, por exemplo. O conhecimento obtido pode ser utilizado como parâmetro de estudo também por agrônomos e outros cientistas que necessitam trabalhar com dados meteorológicos.

Agradecimentos

A Deus por toda luz, proteção e saúde que deposita sobre minha caminhada.

A meus avós por toda compreensão, apoio e carinho durante minhas horas de estudos. Assim como meus pais, que sempre se fazem presentes.

A minha namorada por sempre passar determinação nas horas em que mais precisava. Agente fundamental nas minhas escolhas e planos.

Ao professor Dario Calçada por todo apoio, participação e por me apresentar a computação com outros olhos.

Aos amigos e membros do Grupo de Estudo e Desenvolvimento de Aplicações Inteligentes (GEDAI) por toda troca de experiências e conhecimentos trocados durante a elaboração da pesquisa.

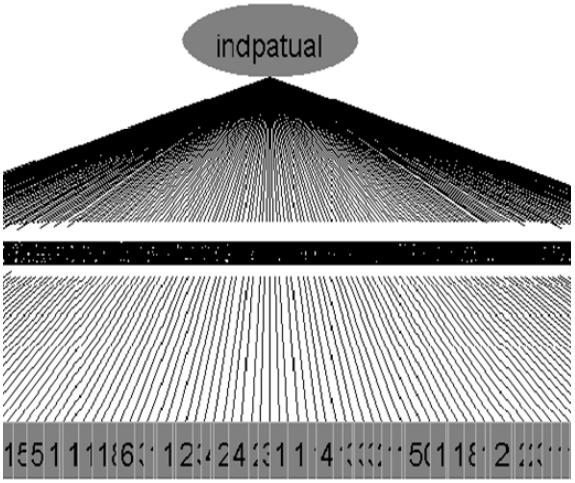
Referências

Camilo, C. O. and Silva, J. C. d. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, pages 1–29.

Cunha, G. R. (1997). Meteorologia. *Fatos e mitos. Passo Fundo. EMBRAPA*.

Da Silva, I. N., Spatti, D. H., and Flauzino, R. A. (2010). Redes neurais artificiais para engenharia e ciências aplicadas curso prático. *São Paulo: Artliber*.

Figura 8. Árvore de Decisão



Fonte:Elaborada pelo autor

Dantas, E. R. G., Almeida, J. C., Júnior, P., de Lima, D. S., and Pessoa-UNIPÊ, J. (2008). O uso da descoberta de conhecimento em base de dados para apoiar a tomada de decisões. *V Simpósio de Excelência em Gestão e Tecnologia*, pages 1–10.

de Carvalho, M. Â. V., Gisler, C. A. F., Silva, A. J. O., and Neto, A. L. C. (2013). Sistema de previsão numérica do tempo do icea.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.

Feelders, A., Daniels, H., and Holsheimer, M. (2000). Methodological and practical aspects of data mining. *Information & Management*, 37(5):271–281.

Ferraudo, A. S. and Ferraudo, G. M. (2010). Ajuste de modelos de redes neurais artificiais na precipitação pluviométrica mensal.

Grande, U., Grande-PB, C., and Grande-PB, C. (2013). Variabilidade pluviométrica entre regimes diferenciados de precipitação no estado do piauí. *Revista Brasileira de Geografia Física*, 6(05):1463–1475.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Iriart, P., Carvalho, M., and Pereira Neto, A. (2011). Manual de instalação, compilação e execução do sistema de modelagem numérica wrf no icea. *São José dos Campos, SP: Instituto de Controle do Espaço Aéreo, Subdivisão de Climatologia e Arquivo Meteorológico (PBCA)*.

Kolmogorov, A. N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. In *Doklady Akademii Nauk*, volume 114, pages 953–956. Russian Academy of Sciences.

Larose, D. T. and Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*, volume 4. John Wiley & Sons.

Marcoulides, G. A. (2005). Discovering knowledge in data: an introduction to data mining.

Meller, A. (2012). Previsão de cheias por conjunto em curto prazo.

Moraes, R. A. and Arraes, C. L. (2015). Análise de uma metodologia para preenchimento de valores faltantes em dados de precipitação, para o estado do paraná. *UNOPAR Científica Ciências Exatas e Tecnológicas*, 11(1).

Moura, A. D. (1996). Von neumann e a previsão numérica de tempo e clima. *Estudos Avançados*, 10(26):227–236.

Muralikrishna, A. (2009). Previsao doíndice geomagnético dst utilizando redes neurais artificiais e arvore de decisao. *Master's thesis, INPE, São José dos Campos*.

Oliveira Filho, C. L. d. (2007). *Prognóstico das variáveis meteorológicas e da evapotranspiração de referência com o modelo de previsão do tempo GFS/NCEP*. PhD thesis, Universidade de São Paulo.

Olson, D. L. and Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

Santhanam, T. and Subhajini, A. (2011). An efficient weather forecasting system using radial basis function neural network. *Journal of Computer Science*, 7(7):962.

Santos, A. M. d., Seixas, J. M. d., Pereira, B. d. B., and Medronho, R. d. A. (2005). Usando redes neurais artificiais e regressão logística na predição da hepatite a. *Revista Brasileira de Epidemiologia*, 8(2):117–126.

Shiba, M. H., Santos, R. L., Quintanilha, J. A., and Kim, H. Y. (2005). Classificação de imagens de sensoriamento remoto pela aprendizagem por árvore de decisão: uma avaliação de desempenho. *SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO*, 12:4319–4326.

Silva, A. F. and Fisch, G. (2014). Avaliação do modelo wrf para a previsão do perfil do vento no centro de lançamento de alcântara. *Revista Brasileira de Meteorologia*, 29(2):259–270.

Silva, M. d., Mendonça, A., and Bonatti, J. (2008). Determinação das previsões de temperaturas mínimas e máximas a partir do histórico das previsões de tempo por conjunto do cptec. *Revista Brasileira de Meteorologia*, 23(4):431–449.

Simões, A. C. A. (2008). Mineração de dados baseada em árvores de decisão para análise do perfil de contribuintes. Master's thesis, Universidade Federal de Pernambuco.

Sousa, W. d. S. and de Sousa, F. d. A. (2010). Rede neural artificial aplicada à previsão de vazão da bacia hidrográfica do rio piancó. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 14(2):173–180.