UTILIZANDO AS TÉCNICAS DE RANDOM FOREST E REDES NEURAIS PARA PREVISÃO DE NÍVEIS NA CIDADE DE PIAÇABUÇU

Trabalho de Conclusão de Curso

Engenharia de Computação

Gustavo Henrique Nunes da Silva
Orientador: Mêuser Jorge Silva Valença

Universidade de Pernambuco
Escola Politécnica de Pernambuco
Graduação em Engenharia de Computação

GUSTAVO HENRIQUE NUNES DA SILVA

UTILIZANDO AS TÉCNICAS DE RANDOM FOREST E REDES NEURAIS PARA PREVISÃO DE NÍVEIS NA CIDADE DE PIAÇABUÇU

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Recife, Outubro de 2018

Escola Politécnica de Pernambuco
Graduação em Engenharia de Computação
Projeto de Final de Curso

ESCOLA POLITÉCNICA DE PERNAMBUCO

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 26 de dezembro de 2018, às 9:00 horas, reuniu-se para deliberar a defesa da monografia de conclusão de curso do discente **GUSTAVO HENRIQUE NUNES DA SILVA**, orientado pelo professor **Mêuser Jorge Silva Valença**, sob título **UTILIZANDO AS TÉCNICAS DE RANDOM FOREST E REDES NEURAIS PARA PREVISÃO DE NÍVEIS NA CIDADE DE PIAÇABUÇU**, a banca composta pelos professores:

Sérgio Mário Lins Galdino
Mêuser Jorge Silva Valença

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

☒ Aprovada ☐ Aprovada com Restrições* ☐ Reprovada

e foi-lhe atribuída nota: 10,0 (dez)

*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O discente terá 7 dias para entrega da versão final da monografia a contar da data deste documento.

Sérgio Galdino
SÉRGIO MÁRIO LINS GALDINO

MÊUSER JORGE SILVA VALENÇA

* Este documento deverá ser encadernado juntamente com a monografia em versão final.

Aos meus pais Maria Euda e Arnaldo José, pelos ensinamentos que me tornaram a pessoa que sou hoje, e por sempre me incentivarem a ir mais longe.

Agradecimentos

Agradeço primeiramente aos meus pais, Euda e Arnaldo, que me ensinaram a sempre tentar buscar ser uma pessoa melhor, e que estiveram sempre presentes em toda a minha jornada até aqui. Sou muito grato pelos seus exemplos de resiliência, onde mesmo passando por alguns momentos difíceis, me mostraram como ter forças e seguir em frente.

Agradeço também às amizades que construí na universidade, e que tornaram a convivência ao longo do curso algo fundamental para a minha formação. Agradeço em especial aos amigos que contribuíram direta e indiretamente para a conclusão deste trabalho, pelas dúvidas tiradas, pelos conselhos nos momentos mais difíceis e por todo o incentivo dado.

Também agradeço aos professores do Instituto Santa Tereza, que contribuíram com a formação do meu pensamento crítico, e cujo os conhecimentos passados não só permitiram o meu ingresso na universidade, como também me ensinaram valores que levarei para toda a vida.

Por fim, agredeço aos professores da Universidade de Pernambuco, que contribuíram com a minha formação profissional e acadêmica, e que também fizeram parte da minha evolução pessoal, me permitindo alcançar novos horizontes. Agradeço especialmente ao meu orientador Mêuser Valença, por ter permitido a realização deste trabalho, pela sua paciência, compreensão e apoio durante todo o semestre.

Resumo

A qualidade dos recursos hídricos disponíveis em uma grande parcela das cidades brasileiras, vem apresentando índices não satisfatórios para o abastecimento urbano de água. Um dos fatores que contribuem para esses índices, é aumento do nível do mar, podendo contaminar a água das cidades abastecidas por mananciais próximos à maré. A cidade de Piaçabuçu, no estado de Alogoas, está localizada na foz do rio São Francisco, sofrendo influência da vazão e da maré na sua principal fonte de abastecimento de água. Este trabalho tem como objetivo identificar a importância da maré e da vazão nos níveis do rio em Piaçabuçu e obter um modelo de previsão de níveis, visando auxiliar na qualidade da captação de água. Para realizar esse estudo, será utilizado o algoritmo Random Forest para a seleção das variáveis mais importantes, e será gerado um modelo de previsão utilizando a rede neural Multi-Layer Perceptron. As variáveis da maré e vazão também serão analisadas para as cidades de Propriá e Penedo, localizadas a montante de Piaçabuçu, para se obter um melhor resultado. No fim do trabalho, será gerado um modelo de previsão de níveis utilizando a rede neural, e também foi comprovado que a maré possui mais influência que a vazão para os níveis do rio na estação fluviométrica de Piaçabuçu.

Abstract

The quality of the water resources available in a large part of Brazilian cities, has been showing unsatisfactory indices for urban water supply. One of the contributing factors to these indices is sea level rise, which can contaminate the water of towns supplied by water sources near the tide. The city of Piaçabuçu, in the state of Alogoas, is located at the mouth of the São Francisco river, suffering influence of the watershed flow and the tide in its main source of water supply. This work aims to identify the importance of the tide and flow in the river levels in Piaçabuçu and to obtain a level prediction model, aiming to assist in the quality of water catchment. To perform this study, we will use the Random Forests algorithm to select the most important variables, and a prediction model will be generated using the Multi-Layer Perceptron neural network. The tide and watershed flow variables will also be analyzed for the towns of Propriá and Penedo, located upstream of Piaçabuçu, in order to obtain a better result. At the end of the work, a level prediction model will be generated using the neural network, and it was also proved that the tide has more influence than the flow in the fluviometric station of Piaçabuçu.

Sumário

Índice de Figuras

Índice de Tabelas

Tabela de Símbolos e Siglas

ANA – Agência Nacional de Águas

Bagging – Bootstrap Aggregating

CART – Classification and Regression Trees

EMQ – Erro Médio Quadrático

EPMA – Erro Percentual Médio Absoluto

MLP – Multilayer Perceptron

Capítulo 1

Introdução

Este capítulo está divido em três seções. Na primeira seção encontra-se a caracterização do problema e as motivações para o desenvolvimento deste trabalho. A segunda seção apresenta os objetivos e metas do estudo. Na terceira seção é apresentada a estrutura da monografia.

1.1 Caracterização do Problema

O Brasil é um dos países mais ricos em recursos hídricos superficiais do planeta, apresentando uma média de vazões de quase 180 mil m³/s em todo o território nacional. Entretanto, com a grande variabilidade climática característica do Brasil, a distribuição desses recursos para cada região se apresenta de forma bastante desigual.

Segundo a Agência Nacional de Águas (ANA), as regiões Norte e uma grande parcela do Centro-Oeste, representam 81% da disponibilidade hídrica brasileira, enquanto a maior parte da população está concentrada nas regiões onde a oferta de água é mais desfavorável [1].

Para suprir essa necessidade, as regiões que possuem pouca oferta de água utilizam os recursos provenientes de bacias hidrográficas, onde contam com a ajuda de reservatórios que acumulam água e controlam a sua passagem para os períodos mais secos. A bacia do rio São Francisco representa cerca de 2/3 da disponibilidade de água doce do Nordeste brasileiro, sendo de grande importância para o abastecimento de água nesta região [16].

Contudo, a falta de chuvas e a ação humana ao longo do rio e de seus principais afluentes, veio causando um período de seca desde 2012. Por conta desse cenário, o rio São Francisco sofreu uma redução da sua vazão que era de aproximadamente 2.000m³/s para 600 m³/s [13], afetando o abastecimento de água

da população. Por outro lado, a elevação do nível do mar também pode prejudicar as captações de água doce, gerando uma preocupação em relação à qualidade da água nos corpos hídricos [20]. Segundo um estudo realizado pela ANA, o estado de Alagoas é um dos que sofrem com esses problemas, onde algumas de suas regiões apresentaram um avanço da cunha salina do mar [2], o que representa uma ameaça à saúde de quem faz uso dessa água.

Um dos municípios afetados pela presença da salinização nos centros de abastecimento de água, é o de Piaçabuçu, localizado na foz do rio São Francisco, no estado de Alagoas. Por estar localizado em uma região muito próxima ao mar, espera-se que a maré tenha mais influência que a vazão nos níveis do rio na estação fluviométrica da cidade. O algoritmo Random Forest proposto por (Breiman, 2001) é um previsor baseado em árvores de decisões que pode ser utilizado para medir a importância de variáveis em um modelo de previsão [7] .Conforme o estudo realizado por (Bastos, et al.), os critérios de seleção de variáveis utilizados pelo Random Forest apresentou resultados bastante satisfatórios [4]. Por esse motivo, um dos objetivos desse trabalho é utilizar a vazão do rio São Francisco e os níveis da maré como variáveis de entrada para o Random Forest, visando identificar quem possui mais influência nos níveis do rio em Piaçabuçu.

Avaliando esse cenário, também se faz necessário utilizar alguma alternativa que auxilie na captação de água para o abastecimento da cidade. Dependendo do nível do rio, maior pode ser a incidência da salinidade na água, podendo causar doenças a quem utilizá-la. Os estudos de (Pedrollo, 2017), (Valerio, 2016) e (Dornelles, 2007) mostraram-se satisfatórios ao utilizarem Redes Neurais Artificiais (RNA) para auxiliar na previsão de níveis de uma determinada região. [18][23][11] Desse modo, outro objetivo desse trabalho será utilizar redes neurais para a previsão de níveis em Piaçabuçu, buscando gerar um modelo que auxilie na identificação do melhor horário para a captção da água.

1.2 Objetivos e Metas

1.2.1 Objetivo Geral

O principal objetivo deste trabalho é identificar a importância da maré e vazão nos níveis do rio em Piaçabuçu utilizando Random Forests, e também obter a previsão de níveis utilizando a rede neural Multi-Layer Perceptron (MLP).

1.2.2 Objetivos Específicos

1. Utilizar o Random Forest para identificar quais variáveis da vazão e da maré mais influenciam nos níveis em Piaçabuçu.

2. Analisar os dados estatisticamente usando os critérios de seleção de variáveis do Random Forest para encontrar quais delas são mais importantes no processo de previsão de níveis na cidade.

3. Obter a previsão dos níveis em Piaçabuçu para um dia à frente utilizando a rede MLP.

1.3 Estrutura da Monografia

Esta monografia foi desenvolvida seguindo a seguinte estrutura:

- Capítulo 1 – Introdução: neste capítulo é feita uma introdução ao problema, apresentando as motivações e os principais objetivos a serem alcançados através desse trabalho.

- Capítulo 2 – Fundamentação teórica: neste capítulo é feita a descrição de todos os conteúdos necessários para o entendimento deste trabalho, abordando os conceitos de árvores de decisão, Random Forest e Redes Neurais, focando na rede MLP.

- Capítulo 3 – Metodologia: este capítulo aborda a metodologia utilizada para a realização deste trabalho. Será descrito como foram obtidas as bases utilizadas, o pré-processamento desses dados e as

configurações utilizadas no Random Forest e na MLP para que fossem obtidos os resultados de previsão. Também serão descritos os testes estatísticos que serão utilizados para a escolha da configuração mais eficiente da MLP.

- Capítulo 4 – Resultados: neste capítulo serão apresentados os resultados obtidos após a utilização do Random Forest nas bases obtidas no capítulo 3, além dos resultados do modelo de previsão gerado pela MLP.

- Capítulo 5 – Considerações Finais: este capítulo apresenta as conclusões obtidas através dos resultados, além de descrever os possíveis trabalhos futuros para este estudo.

Capítulo 2
Fundamentação Teórica

Neste capítulo serão abordados os principais conceitos relacionados ao problema proposto e à sua solução. Seu objetivo é garantir um melhor entendimento dos assuntos abordados nesta monografia.

Na seção 2.1 serão abordados conceitos de Árvores de Decisão, como uma maneira introdutória para o entendimento do algoritmo Random Forest, que será utilizado nesse trabalho. Já a seção 2.2 apresenta os conceitos de Redes Neurais, descrevendo também o funcionamento de uma rede MLP, que será utilizada para realizar o modelo de previsão.

2.1 Árvores de Decisão

2.1.1 Conceitos

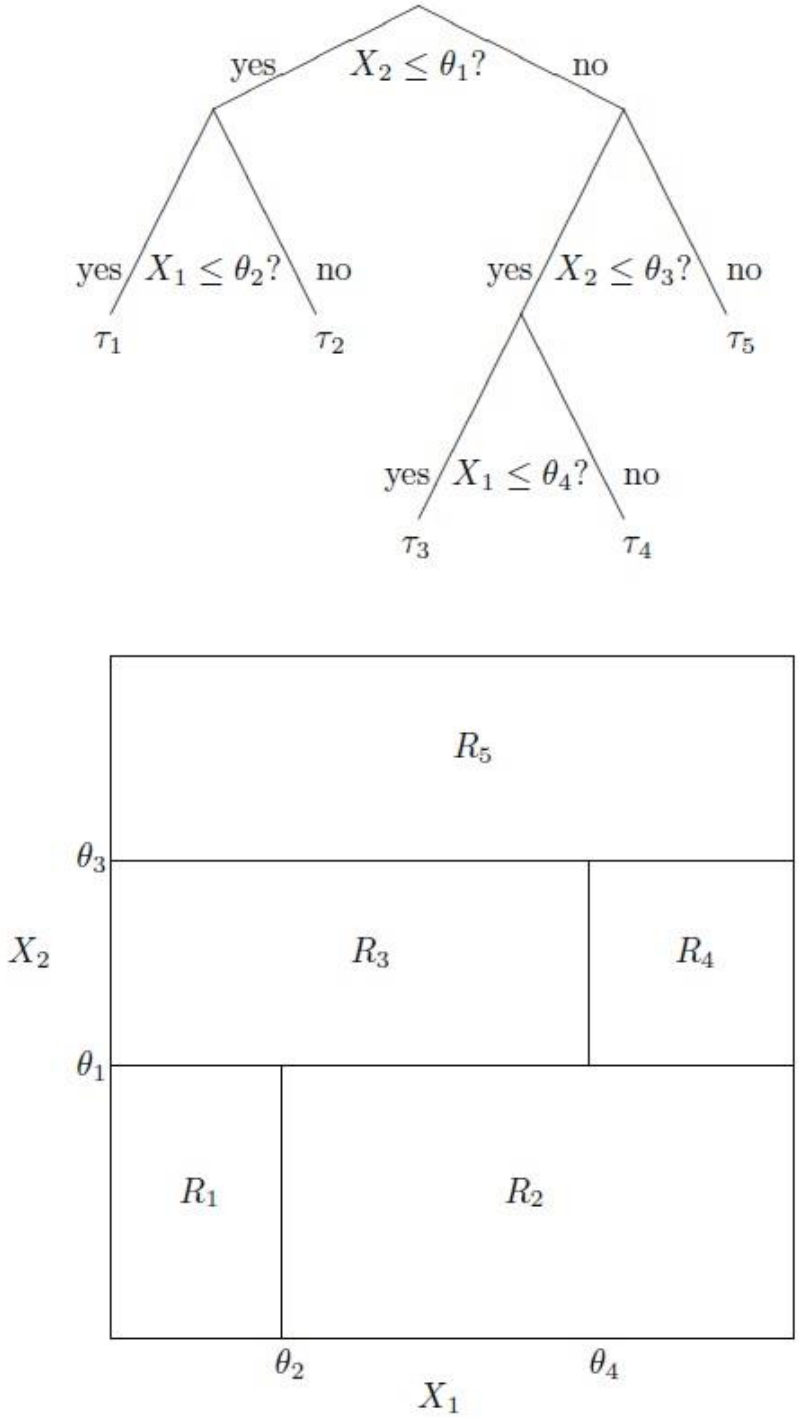
Árvores de decisão são modelos de aprendizado de máquina supervisionado muito utilizados em problemas de tomadas de decisões. Sua premissa consiste em dividir um problema complexo em problemas menores, a fim de agrupar os resultados que pertencem a um grupo em comum. Uma árvore de decisão recebe como entrada um vetor de atributos e retorna uma decisão, ou seja, um conjunto de valores previstos com base nos atributos de entrada [19]. O modelo retornado consiste em um gráfico em estrutura de árvore que facilita a visualização dos caminhos escolhidos para cada variável de entrada.

A metodologia CART (Classification and Regression Trees) proposta por (Breiman et al., 1984), apresenta uma forma simplificada para a construção de árvores de decisão, que podem ser representadas da seguinte forma [6]:

- **Árvores de Classificação:** Consiste em uma árvore de decisão onde as variáveis dependentes são discretas, sendo utilizada para identificar

qual a classe os valores passados como entrada pertencem. A Figura 1 ilustra a estrutura de uma árvore de classificação e a divisão das regiões de agrupamento.

Figura 1. Estrutura de uma árvore de classificação e regiões de agrupamento



Fonte: IZENMAN , A.J. Modern Multivariate Statistical Techniques, 2008 [14]

Uma árvore de classificação é o resultado de uma sequencia de perguntas ordenadas, e o tipo da pergunta feita em cada passo na sequencia depende das respostas passadas. A sequencia de perguntas chega ao fim quando se encontra a previsão da classe em que o atributo de entrada faz parte. O início da árvore de decisão é chamado de nó raiz, e ele consiste no conjunto completo de treinamento, localizado no topo da árvore. Um nó é um subconjunto do conjunto de variáveis, podendo ser um nó não terminal ou terminal. Um nó não terminal, também chamado de nó pai, é um nó que se divide em dois nós filhos, através de uma condição binária aplicada a um único atributo. Esse nó é responsável pela divisão do conjunto em outros nós. O nó "$X_2 \leq \theta_3$?", exibido na Figura 1, é um exemplo de nó não terminal. Após uma sucessão de condições e divisões em subconjuntos, a árvore chega em um nó que não se divide, chamado de nó terminal, o qual é atribuído uma classe. Os nós τ_1 e τ_5 na Figura 1, são exemplos de nós terminais. No final da classificação, o conjunto de treinamento é subdividido em em pequenos conjuntos (as regiões de agrupamento $R_1 - R_5$ na Figura 1) onde serão agrupados cada atributo de entrada [14].

- **Árvores de Regressão:** Assim como as árvores de classificação, uma árvore de regressão também é construída através de condições booleanas aplicadas aos valores de entrada, porém, sua principal diferença é que ela é utilizada para prever um valor de resposta contínuo, ao invés de classicar as entradas em classes.

Através das divisões dos dados em subconjuntos, é possível identificar o nível de pureza de cada nó da árvore de decisão. Com isso, é possível identificar o melhor ponto de corte no conjunto de treinamento. No algoritmo CART, esse nível é medido através do índice de Gini, que corresponde à heterogeneidade dos dados. O índice de Gini para um determinado nó, é dado pela seguinte fórmula:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{2.1}$$

Onde *c* representa o número de classes existentes e p_i é a probabilidade de ocorrência da classe *i* em cada nó. Quanto mais próximo de zero for o índice de Gini, mais puro será o nó, e consequentemente, mais homogêneas serão as classes a partir da divisão do conjunto nesse determinado ponto [3].

Uma outra medida muito utilizada para se escolher a melhor divisão em uma árvore de decisão é a entropia. Quanto menos classes um subconjunto possuir, mais puro ele será. O objetivo da entropia é encontrar o grau de incerteza para que determinado evento ocorra, variando seus valores entre 0 e 1. Quanto mais próximo de 0 for a entropia, menor será a incerteza nos subconjuntos gerados, gerando dados mais homogêneos.

A fórmula da entropia (medida de informação) para um determinado conjunto S pode ser definida por:

$$Entropia(S) = -\sum_{i=1}^{c} p_i \log_2 p_i \tag{2.2}$$

Onde *c* é a quantidade de classes no conjunto S e p_i é a proporção de atributos em S que pertencem à classe *i* [17].

Tendo a entropia como uma medida de impureza de um conjunto de treinamento, também é possível definir uma medida que define o quão importante é determinado atributo para a classificação dos dados. Essa medida é chamada de ganho de informação, e quanto maior o seu valor, menor será a entropia de um novo subconjunto gerado. Para calcular o valor do ganho para um determinado atributo A, primeiro calculamos a entropia obtida ao dividir o conjunto S em função desse atributo, seguindo a seguinte fórmula:

$$Entropia(A) = \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia(S_v) \tag{2.3}$$

Onde *Valores(A)* representa o conjunto de todos os valores possíveis do atributo A, e S_v é o subconjunto de S onde cada atributo A possui o valor *v*. Por fim, o ganho de informação é calculado através da fórmula:

$$Ganho(S, A) = Entropia(S) - Entropia(A) \quad (2.4)$$

Onde a *Entropia(S)* é a entropia calculada para o conjunto de treinamento S e a *Entropia(A)* representa uma nova entropia para o conjunto S caso seja gerado um novo subconjunto a partir do atributo A. Com essa medida espera-se identificar o menor valor de entropia para cada partição do conjunto de treinamento, afim de se obter ramos de árvores mais homogêneos [17].

2.1.2 Métodos Ensemble

Quando um conjunto de treinamento S é submetido a um algoritmo de aprendizado de máquina, os valores retornados da variável dependente *y* são obtidos através de um conjunto de classes discretas, no caso de problemas de classificação, ou através de valores contínuos, para os problemas de regressão. Esses valores de saída representam um conjunto de previsão baseado nos valores passados como entrada no modelo, porém, os valores previstos nem sempre são precisos.

Os métodos ensemble consistem na utilização de vários modelos de previsão que são agrupados de alguma maneira (geralmente através de votação da maioria ou uma média dos resultados) para se obter resultados mais precisos. Um modelo de árvore de classificação, por exemplo, precisa ter um bom índice de generalização para diminuir a taxa de erro ao se fazer uma classificação. Tomando como exemplo um conjunto de classificadores idênticos {c1, c2, c3} obtidos por meio de um método ensemble, caso o classificador c1 realize uma previsão errada, os outros classificadores também errarão [9].

Por outro lado, se os classificadores forem independentes e com uma baixa correlação entre eles, caso o classificador c1 erre a previsão, ainda é possível que os outros dois classificadores estejam corretos, que seriam escolhidos através da escolha de voto da maioria. Por isso, a abordagem dos métodos ensemble oferece uma melhor robustez aos modelos de aprendizado de máquina comuns, pois ela não se baseia apenas em um conjunto específico de resultados.

2.1.3 Bootstrap Aggregating

Utilizando a metodologia dos métodos ensemble, (Breiman, 1996) propôs a técnica Bagging (Bootstrap Aggregating), que oferece uma melhoria nos resultados previstos ao reorganizar o conjunto de treinamento. [5] Essa abordagem é muito útil, pois se uma árvore de decisão for submetida a um treinamento utilizando um subconjunto do conjunto de dados inicial, por exemplo, ela pode apresentar resultados destoantes dos que foram obtidos anteriormente.

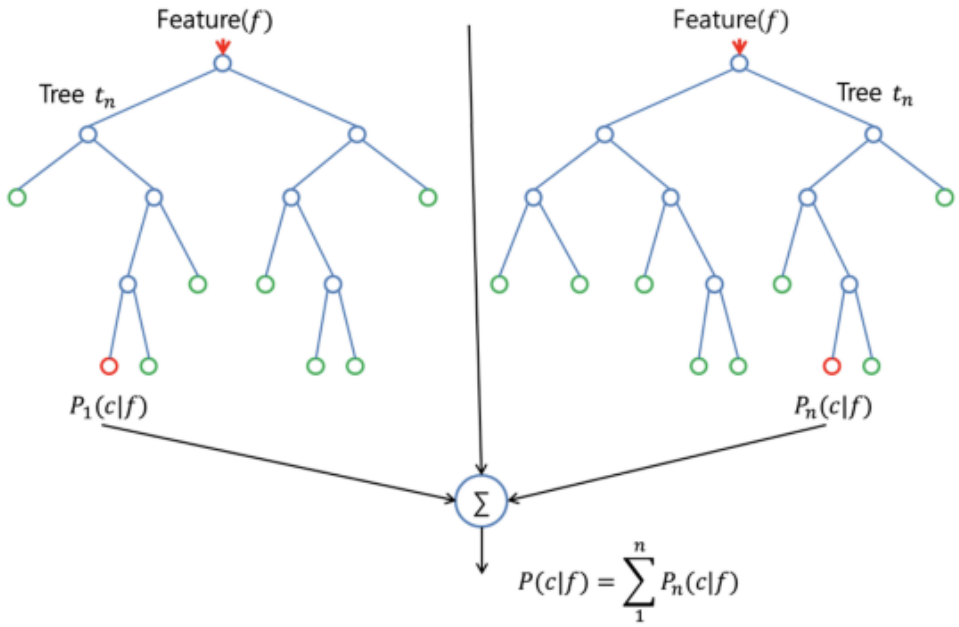
Utilizando a notação empregada por (Breiman, 1996), consideremos um conjunto de dados $S = \{(y_n, x_n), n = 1..., N\}$ onde *y* corresponde ao resultado da classificação ou um valor numérico para os casos de regressão. A previsão de *y* para um dado *x* é feita através do previsor $\varphi = (x, S)$. O conjunto de dados *S* é então subdividido em pequenos subconjuntos $\{S^{(B)}\}$, onde cada elemento (y_n, x_n) é escolhido aleatoriamente do conjunto original *S*. Essa escolha é feita com repetições, o que significa que um mesmo elemento pode estar contido mais de uma vez em um único subconjunto.

Ao rodar o modelo para cada subconjunto em $\{S^{(B)}\}$, teremos então um conjunto de previsores $\{\varphi = (x, S^{(B)})\}$. Por fim, caso o modelo de previsão seja de regressão, a técnica Bagging utiliza uma média dos resultados obtidos por cada um desses previsores. Para os modelos de classificação, o Bagging leva em consideração o voto da maioria dos previsores. O modelo final retorna então o previsor $\varphi_B(x)$, que representa a previsão obtida para o conjunto de treinamento utilizando a estratégia Bagging.

2.1.4 Random Forest

O algoritmo Random Forest foi proposto por (Breiman, 2001), como uma nova abordagem para os métodos Ensemble. O Random Forest utiliza a mesma metodologia do algoritmo Bagging descrito na seção 2.1.3, construindo múltiplas árvores de decisões e agrupando seus resultados para construir uma previsão mais robusta [7]. Na Figura 2 temos uma representação simplificada de um modelo Random Forest:

Figura 2. Previsão em um algoritmo Random Forest



[Fonte: https://machinelearning-blog.com/2018/02/06/the-random-forest-algorithm/]

Através do método Bagging são obtidos os previsores $P_n(c \mid f)$ para cada árvore gerada, e no fim é escolhida uma média desses valores para gerar o previsor do modelo geral. Além da utilização do método Bagging, o Random Forest também utiliza um critério adicional. Ao invés de identificar a melhor variável para a divisão de um nó, o algoritmo utiliza um conjunto aleatório de variáveis e procura a melhor opção dentre elas. Essa estratégia resulta numa maior capacidade de generalização do modelo [10].

O algoritmo utiliza o parâmentro *ntree* para definir a quantidade de árvores que serão geradas no modelo, e a sua documentação aconselha que não seja utilizado um valor muito baixo, para garantir que que cada valor de entrada seja previsto pelo menos algumas vezes. Neste trabalho estaremos utilizando o valor padrão de 500 para o parâmentro *ntree.* Um outro parâmetro importante para o Random Forest é o *mtry*, que representa a quantidade de variáveis que serão escolhidas aleatoriamente durante a divisão em cada nó. Para os problemas de classificação, o valor padrão para o *mtry* deve ser $\sqrt{p}$, onde p é a quantidade de

variáveis no modelo. Já para os problemas de regressão, o *mtry* deve ser definido como $\frac{p}{3}$.

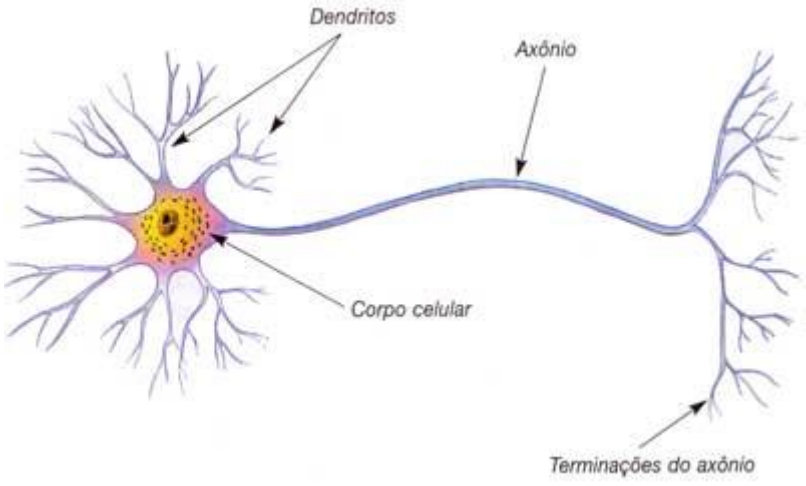
Uma outra característica muito útil no algoritmo Random Forest é a possibilidade de medir a importância das variáveis no modelo. Com isso é possível identificar quais variáveis mais contribuem nos resultados obtidos, e que podem aumentar a taxa de erro caso sejam removidas. A informação da importância dessas variáveis ajuda a evitar o *overfitting* do modelo, ou seja, que ele perca a sua capacidade de generalização e consequentemente obtenha resultados imprecisos. Isso pode ser evitado ao remover do conjunto de entradas, as variáveis que menos contribuem na previsão.

2.2 Redes Neurais Artificiais

Redes Neurais artificiais são modelos computacionais que baseiam-se no comportamento do cérebro humano com objetivo tornar os sistemas mais inteligentes através da auto aprendizagem [22].

Este modelo é proposto a partir da análise do neurônio biológico, que é um conjunto de células que comunicam-se entre si e formam uma rede pela qual circulam impulsos nervosos e informações. No neurônio cerebral o sinal é recebido pelos dendritos, passa pelo axônio e é então transmitido pelas terminações do axônio. E é nessa região de transmissão, conhecida como região de sinapse, que ocorre a comunicação entre os neurônios, formando uma rede neural.

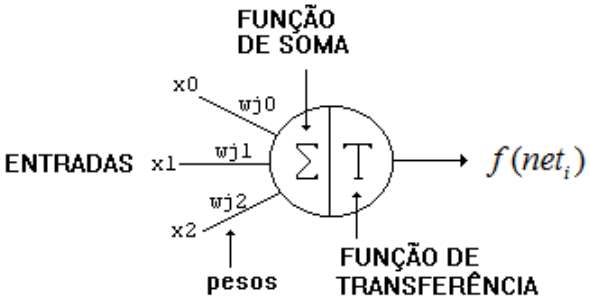
Figura 3. Estrutura do neurônio biológico



[Fonte: https://metodosupera.com.br/neuronios-glossario-do-cerebro/]

A partir desses conceitos biológicos foi proposto o primeiro modelo de Neurônio artificial por McCulloch e Pitts, em 1943, onde o neurônio obedece à Lei do Tudo ou Nada, e sempre estará em dois estados: ativado ou desativado (1 ou 0). Para que a transmissão se propague entre os neurônios, o impulso deve ser maior que um valor mínimo, denominado limiar excitatório. A Figura 4 ilustra o modelo proposto pela dupla:

Figura 4. Modelo do neurônio de McColloch e Pitts



[fonte:http://redesneuraisartificiais.blogspot.com/2010/10/o-primeiro-modelo-de-um-neuronio-criado.html]

O modelo recebe como entrada uma quantidade n de variáveis, representadas por $x_0, x_1, x_2, \ldots, x_n$. Associados à esses pesos está um conjunto de pesos responsáveis por simular as sinapses, conhecidos como pesos sinápticos, representados por $w_{j0}, w_{j1}, w_{j2} \ldots, w_{jn}$. A entrada do neurônio é definida pela regra de propagação:

$$net_i = \sum_{j=1}^{n} w_{ij} x_j - \theta \qquad (2.5)$$

Onde θ é o valor do limiar. A função de ativação é definida pela função degrau:

$$f(net_i) = \begin{cases} 1, \forall net_i \geq 0 \\ 0, \forall net_i < 0 \end{cases} \quad (2.6)$$

Entre as décadas de 1950 e 1960 surgiram outras duas arquiteturas muito importantes: Perceptron e a Adaline. O Perceptron surgiu em 1957, e introduziu o conceito de aprendizado supervisionado (ou com professor) para ajustar os pesos. Já a Adaline, em 1960, permitiu o uso de valores contínuos através de novas funções de ativação. Essas novas funções são de grande importância no estudo das redes neurais e, segundo (Valença, 2010), as principais funções de ativação são: a função linear (2.7), a função sigmóide logística (2.8) e a tangente hiperbólica (2.9) [22].

$$f(net_i) = net_i \quad (2.7)$$

$$f(net_i) = \frac{1}{1+e^{-net_i}} \quad (2.8)$$

$$f(net_i) = \frac{e^{net_i} - e^{-net_i}}{e^{net_i} + e^{-net_i}} \quad (2.9)$$

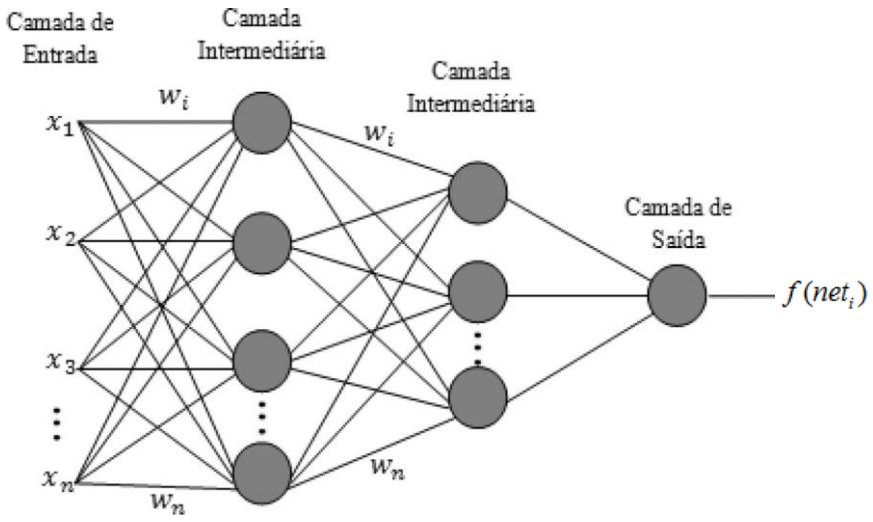
A partir dessas duas arquiteturas, foi possível evoluir conhecimentos e técnicas de aprendizado de máquina e entender o funcionamento de duas das principais redes utilizadas atualmente: a MLP e a RC, que serão abordadas posteriormente.

2.2.1 Multilayer Perceptron

As redes Multilayer Perceptron (MLP) são redes de treinamento supervisionado que nasceram a partir da rede Perceptron. Segundo (Valença, 2010), trata-se de uma generalização dessa rede, onde a principal diferença das redes MLP em relação a sua precursora, é a possibilidade de resolver problemas não linearmente separáveis [22].

Uma rede MLP possui pelo menos três camadas: uma camada de entrada (onde cada neurônio equivale a uma variável de entrada), uma ou mais camadas intermediárias ou escondidas (que permitem as rede resolver problemas reais), e uma camada de saída. Essa estrutura está demonstrada na Figura 5.

Figura 5. Modelo de uma rede MLP



[Fonte: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-77862016000100024]

Cada camada possui um ou mais neurônios e cada neurônio possui uma função de ativação. As principais funções de ativação geralmente são a sigmóide logística ou a tangente hiperbólica, e neste trabalho estaremos utilizando a sigmóide logística. O algoritmo mais utilizado para treinamento de redes Multilayer é o *backpropagation*, que foi desenvolvido em 1974 por Paul Werbos, e consiste no ajuste dos pesos das camadas intermediárias seguindo a ordem inversa da propagação. É um algoritmo simples que é dividido em duas etapas: a fase *forward* e a fase *backward.*

Na fase *forward,* a camada de entrada recebe um dado, que é propagado para cada uma das camadas intermediárias até que chegue na camada de saída, onde será possível calcular o erro. Este erro é o Erro Médio Quadrático (EMQ), e é calculado a partir da resposta obtida e a resposta esperada. Sua fórmula é dada por:

$$EMQ = \frac{1}{N}\sum_{j=1}^{N} e_j^2 \qquad (2.10)$$

Onde N é a quantidade de ciclos, e e_j é o erro calculado ao fim de cada ciclo. O erro deverá ser calculado para cada neurônio da camada de saída e para cada conjunto de entradas apresentado à rede.

Após a propagação da informação da entrada até a camada de saída, é então iniciada a fase *backward*. Nessa fase, o erro calculado na camada de saída é retropropagado para as camadas intermediárias, até que o sinal chegue à camada de entrada. O objetivo dessa etapa é realizar o ajuste dos pesos para que o erro na saída seja minimizado. O reajuste dos pesos é feito através da seguinte fórmula:

$$w_{ij}^m(novo) = w_{ij}^m(antigo) + \alpha\delta_i^m f^{m-1}(net_j^{m-1}) \qquad (2.11)$$

Onde w_{ij}^m são os pesos a serem reajustados, α é a taxa de aprendizagem, $f^{m-1}(net_j^{m-1})$ são os sinais emitidos pelos neurônios da camada anterior e δ_i^m é um termo conhecido como sensibilidade, que na camada de saída é definida pela fórmula:

$$\delta_i^m = f^{m'}(net_i^m).e_i(n) \qquad (2.12)$$

Onde δ_i^m é a sensibilidade a ser calculada para os neurônios da camada de saída m, $f^{m'}(net_i^m)$ representa a derivada da função de ativação utilizada e $e_i(n)$ é o erro calculado para o neurônio i da camada de saída. Para os neurônios das camadas intermidiárias a sensibilidade é calculada da seguinte forma:

$$\delta_j^{m-1} = f^{m-1'}(net_j^{m-1})\sum_{i=1}^{n} w_{ij}^m.\delta_i^m \qquad (2.13)$$

Onde $f^{m-1'}(net_i^{m-1})$ é a derivada da função de ativação da camada $m-1$, w_{ij}^m são os pesos da camada anterior (no sentido da camada de saída para a de

entrada) e δ_i^m é a sensibilidade já calculada na camada anterior. Dessa forma, o algoritmo *backpropagation* utiliza a sensibilidade de forma recursiva desde a camada de saída até a camada de entrada, reajustando os pesos durante vários ciclos, até que o EMQ seja reduzido significativamente.

Ao treinar uma rede neural por uma grande quantidade de ciclos consecutivos, corremos o risco de causar o *overfitting* do modelo. Como no descrito no algoritmo Random Forest, o *overfitting* ocorre quando a rede perde a sua capacidade de generalização e passa a decorar certos valores. Se a rede não for treinada por um número mínimo necessário de repetições, ela também perderá sua capacidade de generalização, pois não irá aprender o suficiente para gerar resultados satisfatórios.

Um dos critérios mais utilizados para parar o treinamento de uma rede neural é o da validação cruzada [22], usado neste trabalho. Isso pode ser obtido ao dividir a base de dados em três conjuntos independentes. O primeiro deles é o conjunto de treinamento, que será utilizado para o ajuste dos pesos da rede neural. Esse conjunto é obtido com aproximadamente 50% dos dados. O próximo conjunto seria o de validação cruzada, com 25% dos dados. Ele será utilizado para calcular a função de erro, e assim servir como critério de parada do treinamento quando o erro tiver o seu valor mínimo atingido. Por fim, utiliza-se os 25% restantes como um conjunto de verificação para avaliar o desempenho da rede após seu treinamento.

Capítulo 3

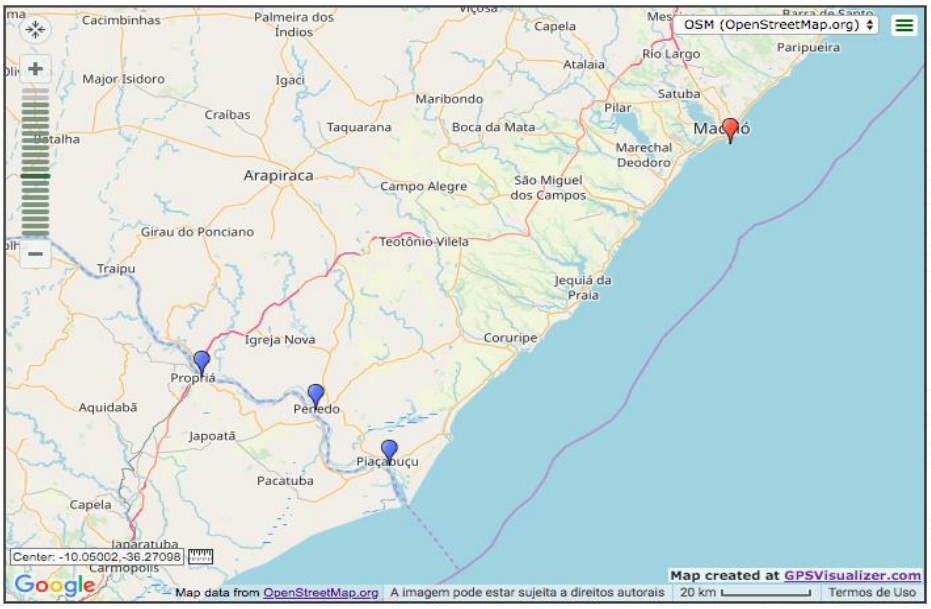
Metodologia

Neste capítulo é abordada a metodologia utilizada para uma solução do problema proposto. Na seção 3.1 serão descritas as bases de dados utilizadas e na seção 3.2 será explicado como foi feito o pré-processamento desses dados. As seções 3.3 e 3.4 apresentam respectivamente as configurações utilizadas para o Random Forest e a rede MLP. Na seção 3.5 serão detalhados os testes estatísticos que são utilizados para escolher a configuração do modelo de previsão que obteve o melhor desempenho.

3.1 Base de Dados

Foram utilizadas três bases de dados, uma delas referente aos níveis medidos nas estações fluviométricas de três municípios próximos à foz do Rio São Francisco: Propriá (SE), Penedo (AL) e Piaçabuçu (AL), este último sendo o foco deste trabalho. As outras bases de dados são referentes à vazão (medida na usina de Xingó) e à maré (medida no posto de Maceió). A localização dos municípios e da estação da maré pode ser vista na Figura 6 a seguir:

Figura 6. Localização dos municípios e da estação da maré utilizados como bases de dados



[Fonte: GPSVisualizer.com]

O objetivo da escolha desses pontos é ter uma melhor visão do comportamento dos dados ao longo do Rio São Francisco até a sua foz, onde se encontra a cidade de Piaçabuçu. As três bases de dados foram obtidas para os anos de 2011, 2013, 2015 e 2017, num período entre Maio e Agosto. A primeira base de dados é referente ao nível da maré medido no posto de Alagoas (Latidude: 9° 40' 60" S, Longitude: 35° 43' 5" W) pela Marinha do Brasil [15]. A sua localização pode ser vista na Figura 7 a seguir:

Figura 7. Localização do posto de Alogoas referente aos dados da maré

[Fonte: Google Maps]

Os níveis foram aferidos durante 3 ou 4 períodos para cada dia (madrugada, manhã, tarde e noite), com as horas sendo variadas. Os períodos representam as mudanças periódicas dos níveis da maré durante um dia.

A segunda base de dados foi disponibilizada pela Chesf (Companhia Hidroelétrica do São Francisco) e possui dados da vazão da usina de Xingó, localizada entre os estados de Alagoas e Sergipe [8]. Entretanto, os dados obtidos são de cunho confidencial, sendo permitido o uso apenas para fins deste estudo. Os valores da vazão foram aferidos em cada um dos três municípios em um intervalo de 5 em 5 minutos.

Os últimos dados a serem coletados foram os das cotas (níveis) do rio São Francisco para as estações fluviométricas de Propriá (Latitude: 10° 12' 50.04" S, Longitude: 36° 49' 26.04" W), Penedo (Latitude: 10° 17' 24.00" S, Longitude: 36° 35' 9.96" W) e Piaçabuçu (Latitude: 10° 25' 0.12" S, Longitude: 36° 26' 0.00" W). A localização de cada estação pode ser vista nas figuras a seguir:

Figura 8. Localização da estação fluviométrica de Propriá

[Fonte: Google Maps]

Figura 9. Localização da estação fluviométrica de Penedo

[Fonte: Google Maps]

Figura 10. Localização da estação fluviométrica de Piaçabuçu

[Fonte: Google Maps]

A base de dados foi obtida pela ANA através do Portal HidroWeb [12], uma ferramenta que faz parte do Sistema Nacional de Informações sobre Recursos Hídricos (SNIRH) e que ofecere acesso ao banco de dados que contém informações coletadas pela Rede Hidrometeorológica Nacional (RHN). Os dados das cotas estão dispostos em um intervalo de 1 em 1 hora.

3.2 Pré-Processamento dos dados

Com as três bases de dados obtidas, o próximo passo é unifica-las em uma uma única base, de tal forma que cada cidade tenha três colunas com os valores do nível da maré, da vazão de Xingó e da cota do rio São Francisco. No fim serão obtidas 12 bases de dados, uma para cada uma das 3 cidades dentre os 4 anos aferidos. O objetivo inicial será identificar se é a vazão ou a maré que mais influencia nos níveis das cotas das cidades, tendo como foco principal deste trabalho a cidade de Piaçabuçu. Após isso, os dados servirão para alimentar um modelo que será utilizado para prever os níveis das cotas em um dia à frente, para a cidade de Piaçabuçu.

Como os dados estão dispostos em intervalos de tempo diferentes, foi escolhido o intervalo de 1 em 1 hora para que todas as três colunas fiquem com os dados dispostos de uma maneira igual. A Tabela 1 ilustra como os dados estarão

dispostos após esse tratamento. Os valores da vazão foram omitidos por conta da confidencialidade dos dados cedidos pela Chesf.

Tabela 1. Estrutura dos dados organizados de hora em hora

Pontos/Horas	Maré	Vazão	Cota
01/05/2013 00:00	3.983	xx	0.6
01/05/2013 01:00	3.757	xx	0.83
01/05/2013 02:00	3.586	xx	1.07
01/05/2013 03:00	3.507	xx	1.30
01/05/2013 04:00	3.537	xx	1.53

A coluna das cotas não precisou sofrer alterações pois os dados já estavam dispostos de hora em hora. Já a coluna da maré não possuía dados para todas as horas do dia, pois como foi descrito na seção 3.1, os dados foram colhidos durante quatro pontos no dia. Para preencher os dados das horas restantes, foi utilizada uma interpolação linear entre os pontos que possuíam horas. A fórmula da interpolação linear utilizada foi a seguinte:

$$y = y_1 + (x - x_1)\frac{y_2 - y_1}{x_2 - x_1} \tag{3.1}$$

Onde:

- x é a posição da linha na tabela onde será atribuído o novo valor da hora
- x_1 é a posição da linha na tabela que contém o valor do primeiro ponto a ser interpolado

- x_2 é a posição da linha na tabela que contém o valor do segundo ponto a ser interpolado
- y é o valor da hora a ser interpolada entre os dois valores já aferidos
- y_1 é o valor da hora contida no primeiro ponto a ser interpolado
- y_2 é o valor da hora contida no segundo ponto a ser interpolado

Aplicando essa fórmula para todos os campos no intervalo entre duas horas medidas, foi construída a coluna da maré com todos os valores dispostos de hora em hora.

Conforme descrito na seção 3.1, a coluna com os dados da vazão possui os dados dispostos em intervalos de 5 em 5 minutos. Por isso foram escolhidos apenas os dados que correspondiam às horas cheias (1:00h, 2:00h, etc), para que ficassem de acordo com as colunas das cotas e maré. Alguns dias não possuíam dados para as horas cheias, por isso foram escolhidos os dados mais próximos desses valores, com 5 minutos de antecedência.

Com as bases de dados já organizadas e com os valores dispostos no mesmo intervalo de horas, o próximo passo foi fazer a normalização dos dados. Segundo (Valença, 2010) , a normalização é importante para que as variáveis em intervalos diferentes tenham o mesmo peso durante o treinamento [22]. Ou seja, é necessário reescrever os dados que estão em intervalos de variação muito grandes para um intervalo menor, pois assim o algoritmo de previsão não perderá a sua capacidade de generalização. A normalização deverá ser feita de uma maneira em que os valores dos dados sejam proporcionais aos limites da função de ativação. Neste trabalho será utilizada a função sigmóide logística, que possui seus valores variando no intervalo entre 0 e 1. No entanto, se os dados forem normalizados nesse intervalo, o treinamento da rede pode ser prejudicado, pois a derivada da função sigmóide se aproxima de zero em valores extremos. Por isso os dados são normalizados geralmente entre 0.1 e 0.9 ou 0.15 e 0.85. Para este trabalho a

normalização dos dados é entre 0.1 e 0.9 através da fórmula da transformação linear descrita a seguir:

$$y = (b - a)\frac{x_i - x_{\min}}{x_{\max} - x_{\min}} + a \tag{3.2}$$

Onde:

- y corresponde ao valor normalizado
- b é o limite máximo utilizado para a normalização da base (0.9)
- a é o limite mínimo utilizado para a normalização da base (0.1)
- x_i corresponde ao valor a ser normalizado, localizado na posição *i*
- $x_{\max}$ é o maior valor encontrado na base
- $x_{\min}$ é o menor valor encontrado na base

Após a normalização da base de dados, o próximo passo é realizar um processo conhecido como defasagem de dados. A defasagem nada mais é do que uma técnica que é aplicada a uma coleção de dados existentes para criar valores passados. A tabela a seguir ilustra a aplicação da defasagem para um conjunto de dados:

Tabela 2. Base de dados defasada em três horas para trás

Pontos/Horas	Maré - 3	Maré - 2	Maré - 1	Maré
14/05/2013 01:00				0.78
14/05/2013 02:00			0.78	0.67
14/05/2013 03:00		0.78	0.67	0.57
14/05/2013 04:00	0.78	0.67	0.57	0.46
14/05/2013 05:00	0.67	0.57	0.46	0.36
14/05/2013 06:00	0.57	0.46	0.36	0.25
14/05/2013 07:00	0.46	0.36	0.25	0.15
14/05/2013 08:00	0.36	0.25	0.15	
14/05/2013 09:00	0.25	0.15		
14/05/2013 10:00	0.15			

Na Tabela 2 temos os dados normalizados na coluna Maré, e a sua defasagem em três horas para trás nas colunas (Maré - 1), (Maré - 2) e (Maré - 3). Essas colunas defasadas são utilizadas como entradas para o Random Forest e a MLP junto com a coluna original, e servem para gerar o conhecimento necessário para que ela possa prever valores futuros. Neste trabalho iremos utilizar os valores de maré e vazão de três dias atrás para prever o valor da cota, então as colunas da maré e vazão foram defasadas em 72 horas cada uma. Após a defasagem, as linhas que possuíam algum dado em branco foram removidas. Após isso, iremos utilizar as colunas defasadas e as originais como entradas do Random Forest e a coluna da cota como saída esperada. Com isso, o algoritmo identifica quais variáveis de entrada têm mais influência para os níveis do rio nas cidades. A mesma base defasada também será utilizada na rede neural para realizar a previsão dos níveis em um dia à frente.

A complexidade da rede neural aumenta quando é utilizada uma grande quantidade de variáveis de entrada. E a introdução de entradas irrelevantes pode acabar prejudicanto o treinamento, pois nem todas as variáveis irão contribuir para a previsão [22]. Por isso, é necessário utilizar uma técnica para selecionar as variáveis mais importantes para o modelo de previsão. Conforme descrito na seção 2.1.4, o Random Forest é uma técnica muito útil para medir a importância das variáveis em um modelo de aprendizado, e por isso estaremos utilizando essa técnica para essa etapa.

3.3 Configuração do Random Forest

O algoritmo do Random Forest foi executado utilizando o software RStudio [21], e inicialmente foram utilizadas as bases separadas de cada cidade para os anos de 2011, 2013, 2015 e 2017. As configurações utilizadas para o Random Forest estão descritas na Tabela 3.

Tabela 3. Configurações utilizadas no Random Forest

Parâmetros	Valores
Quantidade de variáveis de entrada	146
Quantidade de variáveis de saída	1
Variáveis escolhidas aleatoriamente para a divisão de cada nó (mtry)	49
Quantidade de árvores geradas (ntree)	500
Calcular a importância das variáveies (Importance)	True

Cada base possui 146 variáveis de entrada (correspondentes às colunas da maré e vazão e suas defasagens em 72 horas para trás) e uma variável de saída, correspondente à cota da cidade. Como o problema proposto é de regressão, foi utilizado o $mtry = \frac{p}{3}$, onde *p* são as 147 variáveis da base de dados. Então seu valor final foi definido como $mtry = 49$. Como foi descrito na seção 2.1.4, foi

utilizado $ntree = 500$ para a quantidade de árvores geradas no Random Forest. Para que fosse plotado o gráfico com os índices de importância das variáveis, o parâmetro *Importance* foi setado como *true*. As variáveis de entrada *mare73* e *vazao73* são respectivamente as colunas de maré e vazão sem a defasagem, e as demais variáveis correspondem a essas duas colunas defasadas. Após a execução do Random Forest para as bases em cada ano, foi realizada uma nova execução, dessa vez com uma única base contendo os valores de todos os anos para cada cidade. Os resultados obtidos com a execução do Random Forest para todas as bases serão apresentados no capítulo 4.

3.4 Configuração da Rede MLP

Após a execução do Random Forest para as bases contendo os valores de todos os anos, foi utilizada a base de Piaçabuçu para realizar o treinamento da rede MLP. A escolha da quantidade de neurônios da camada de entrada foi baseada nos resultados obtidos pelo Random Forest no capítulo 4, onde foram escolhidas as 6 variáveis mais importantes como entradas da rede. Como o objetivo deste trabalho é gerar um modelo de previsão de níveis para um dia à frente, a rede irá possuir um total de 24 neurônios na camada de saída. Esse valor foi escolhido pelo fato dos dados serem horários, e dessa forma, cada neurônio da saída será equivalente à 1 hora prevista.

O próximo passo será a determinação da quantidade de neurônios na camada escondida, que é um fator muito importante para aumentar a capacidade de mapeamento da rede. Segunda (Valença, 2010), uma maneira prática de definir a quantidade desses neurônios é através do treinamento de algumas redes utilizando uma quantidade diferente de neurônios na camada escondida [22]. Dessa forma, foi realizado o treinamento utilizando três cenários diferentes: o primeiro utilizando a mesma quantidade dos neurônios da entrada para a camada escondida, o segundo utilizando o dobro, e o terceiro utilizando o triplo desse valor. Após a execução dos três cenários, foi utilizado o que teve o menor erro de validação cruzada (utilizou-se testes estatísticos). Essas e as as demais configurações utilizadas na rede MLP estão contidas na Tabela 4.

Tabela 4. Configurações utilizadas na rede MLP

Parâmetros	Valores
Quantidade de neurônios na camada de entrada	6
Quantidade de neurônios na camada escondida	6, 12 e 18
Quantidade de neurônios na camada de saída	24
Algoritmo de treinamento	Backpropagation
Taxa de aprendizado	0.8
Momentum	0.2
Quantidade de ciclos	600
Quantidade de experimentos	30
Função de ativação na camada escondida	Sigmóide Logística
Função de ativação na camada de saída	Linear

A rede foi treinada utilizando o algoritmo Backpropagation, conforme descrito na subseção 2.2.1, e foi utilizada a validação cruzada como critério de parada. Foram executados 600 ciclos ou menos para cada experimento, dependendo do EMQ calculado na validação cruzada. Ao final de cada experimento foi calculado o Erro Percentual Médio Absoluto, definido pela fórmula (3.3) a seguir:

$$EPMA = \sum_{i=1}^{n} \frac{\left| \frac{d_i - y_i}{d_i} \right| * 100}{n} \tag{3.3}$$

Onde d_i é a saída desejada para a previsão i, y_i é a saída calculada para a previsão i, e n representa a quantidade de previsões realizadas. Ao final dos 30 experimentos de cada cenário de configuração, foram executados testes estatísticos para a verificação da configuração com o melhor desempenho.

3.5 Testes estatísticos

Após a execução de cada cenário utilizando a rede MLP, foram obtidos 30 valores de EPMA para cada um, pois com esse valor já se pode considerar que os dados possuem uma distribuição normal. Esses valores servirão como entrada para testes estatísticos, utilizando novamente o RStudio, onde encontra-se o modelo mais eficiente. Os testes podem ser paramétricos ou não paramétricos, e a escolha do teste a ser utilizado depende de alguns fatores que deverão ser analisados em cada conjunto de dados.

Para que um teste paramétrico seja utilizado, é necessário que as amostras sejam normalmente distribuídas e suas variâncias sejam estatisticamente iguais. Caso essas condições não sejam atendidas, deve-se utilizar um teste não paramétrico. As seções a seguir irão detalhar cada teste, explicando os seus objetivos e os critérios para a sua utilização.

3.5.1 Teste Shapiro-Wilk

O teste Shapiro-Wilk é utilizado para verificar se os dados estão normalmente distribuídos, satisfazendo uma das condições para a aplicação de um teste paramétrico. Essa verificação é feitas através das seguintes hipóteses:

- Hipótese Nula (H_0): Os dados estão normalmente distribuídos
- Hipótese Alternativa (H_1): Os dados não estão normalmente distribuídos

As hipóteses são verificadas através da análise do nível de significância (p-value), definido como α = 0.05. Após a execução do teste, caso o p-value possua um valor menor que α, a Hipótese Alternativa é considerada, caso contrário, é considerada a Hipótese Nula.

3.5.2 Teste F

O objetivo do Teste F é verificar se as variâncias de dois conjuntos de dados são estatisticamente iguais. Através do resultado desse teste, pode-se concluir a

segunda condição para a utilização de um teste paramétrico. Seu resultado é avaliado através das seguintes hipóteses:

- Hipótese Nula (H_0): Os dados possuem variâncias estatísticamente iguais
- Hipótese Alternativa (H_1): Os dados não possuem variâncias estatísticamente iguais

Após a execução desse teste, o p-value é verificado, e caso ele possua um valor inferior a α, a Hipótese Alternativa será considerada, caso contrário, será escolhida a Hipótese Nula.

3.5.3 Teste t de Student

O teste t de Student é um teste paramétrico, e para que ele seja executado é necessário que os dados estejam normalmente distribuídos e que suas variâncias sejam estatisticamente iguais. Por isso, é fundamental que o teste Shapiro-Wilk e o teste F sejam executados para confirmar se o teste t de Student poderá ser utilizado. O objetivo desse teste é verificar se a média entre duas amostras são iguais. Da mesma forma que nos testes anteriores, o resultado do t de Student é verificado através das hipóteses a seguir:

- Hipótese Nula (H_0): As médias dos dados são estatisticamente iguais
- Hipótese Alternativa (H_1): As médias dos dados são estatisticamente diferentes

Caso o p-value assuma um valor menor que α, significa que os dados satisfazem a Hipótese Alternativa, e suas médias são estatisticamente diferentes. Caso contrário, a Hipótese Nula é satisfeita.

3.5.4 Teste de Wilcoxon

O teste de Wilcoxon é um teste não paramétrico, ou seja, que não depende de suposições relacionadas à distribuição de probabilidade dos dados. Ele pode ser executado depois que o teste de Shapiro-Wilk e teste F falharem. O objetivo desse

teste é verificar se a mediana de duas amostras independentes são iguais. O teste é verificado através da análise das hipóteses:

- Hipótese Nula (H_0): As medianas dos dados são estatisticamente iguais
- Hipótese Alternativa (H_1): As medianas dos dados são estatisticamente diferentes

Após a execução desse teste, se o valor do p-value for menor que α, a Hipótese Alternativa é satisfeita, significando que as amostras de dados possuem medianas estatisticamente diferentes. Com isso pode-se afirmar que as duas amostras são diferentes, e a que possuir a menor mediana, terá o melhor desempenho.

Capítulo 4

Resultados

O objetivo deste capítulo é detalhar os resultados obtidos ao aplicar o Random Forest nas bases de dados de Propriá, Penedo e Piaçabuçu, visando identificar a importância da maré e vazão para os níveis da cota em cada estação fluviométrica. A técnica será aplicada para cada base durante um período de quatro anos, conforme descrito no Capítulo 3, e por fim, o Random Forest também será utilizado nas bases com os anos unificados. Este capítulo também irá detalhar os resultados obtidos pelo modelo de previsão de níveis em Piaçabuçu através da rede MLP.

Na seção 4.1 serão apresentados os resultados de cada cidade para os anos de 2011, 2013, 2015 e 2017. A seção 4.2 irá detalhar os resultados das cidades ao utilizar uma base com os dados de todos os anos. Por fim, na seção 4.3 serão apresentados os resultados dos três cenários de previsão de níveis em Piaçabuçu utilizando a rede MLP, onde será escolhido o modelo mais eficiente através da utilização de testes estatísticos.

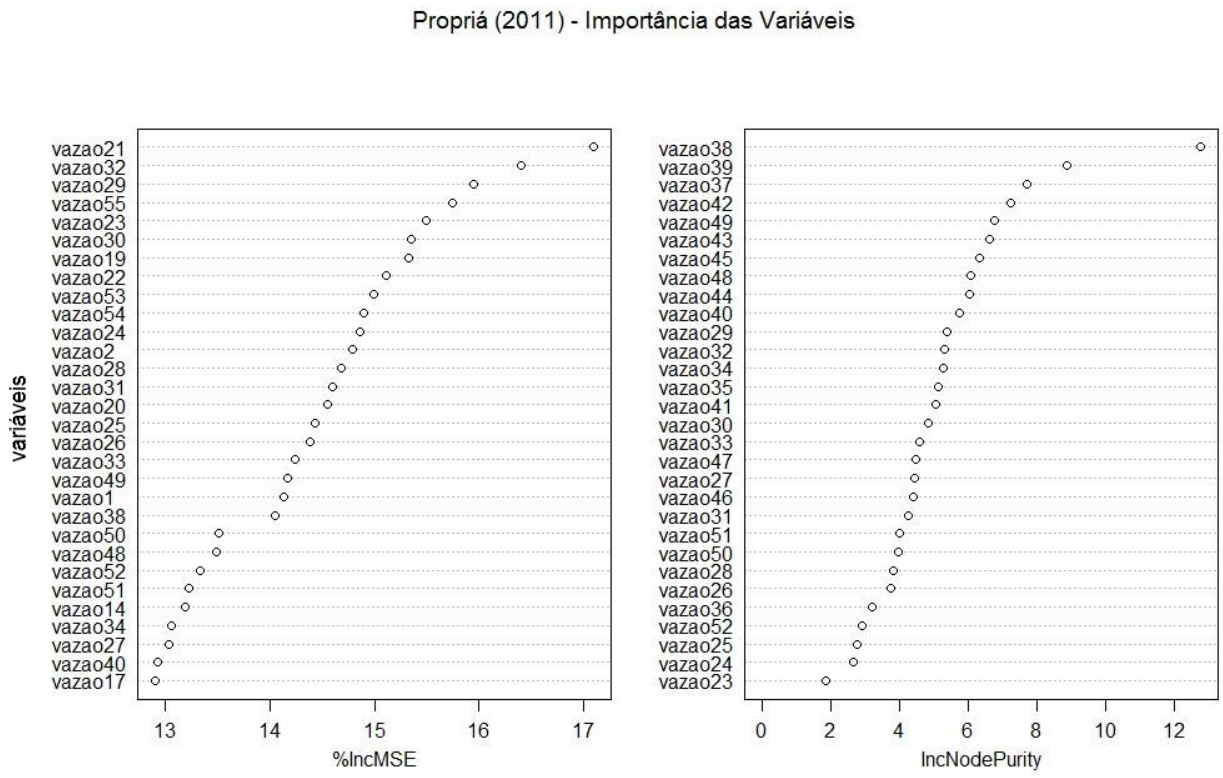
4.1 Random Forest aplicado para cada ano

A importância das variáveis de entrada (maré e vazão) para a variável de saída (cota) em cada ano será descrita nas subseções seguintes.

4.1.1 Importância das variáveis (2011)

Propriá

Figura 11. Importância das variáveis – Propriá (2011)



Após a execução do Random Forest, foram obtidos dois gráficos (Figura 11) contendo as 30 variáveis mais importantes dentre as 146 passadas como entrada. As variáveis foram nomeadas como *vazaox e marex*, onde *x* varia entre 1 e 73. O primeiro gráfico corresponde ao percentual do aumento do EMQ da previsão da cota, caso cada variável tenha seus valores permutados. Como pode ser observado, a variável *vazao21* obteve o maior índice, significando que caso ela seja removida da base de dados, o erro para a previsão da cota será alto. Quanto maior o índice do aumento do erro, mais importante será a variável. O segundo gráfico indica o valor do índice de Gini, descrito na equação (2.1), para cada variável, ou seja, qual delas gera nós mais puros em cada divisão. Como o gráfico do índice de Gini é mais indicado para problemas de classificação, neste trabalho será utilizado apenas o gráfico do EMQ. Para cada índice de importância medido, serão listados os valores das 6 variáveis mais importantes para obetermos uma melhor visualização desse percentual. Para a cidade de Propriá em 2011, foram obtidos os seguintes valores:

Tabela 5. Variáveis mais importantes - Propriá (2011)

Variável	EMQ (%)
vazao21	17.09
vazao32	16.40
vazao29	15.94
vazao55	15.74
vazao23	15.49
vazao30	15.35

Dentre as 30 variáveis mais importantes, exibidas no gráfico do EMQ na Figura 11, todas foram da vazão. As 6 variáveis com maiores percentuais (Tabela 5) obtiveram valores de erro entre 15.35% e 17.09%. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Propriá em 2011 está descrito na Tabela 6.

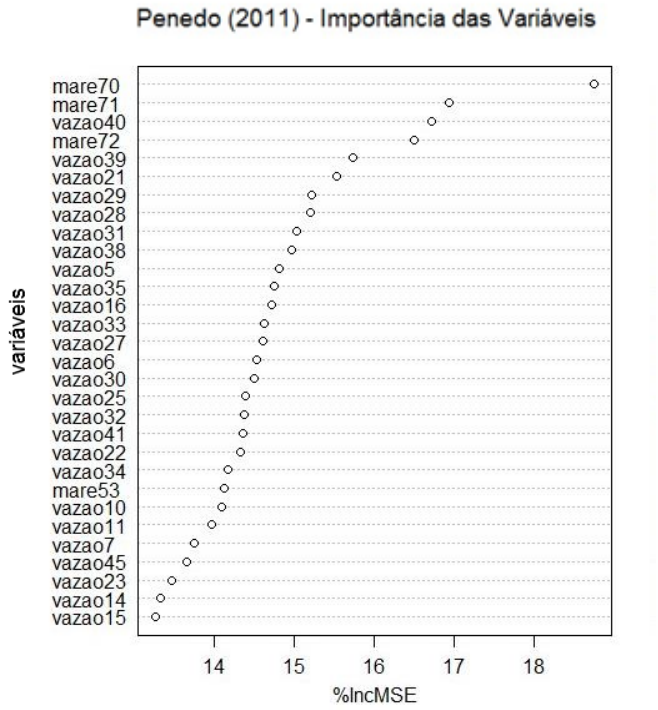
Tabela 6. Percentual das variáveis - Propriá (2011)

Variável	Percentual (%)	Grau de Importância
Maré	12.86	Menos importante
Vazão	87.14	Mais importante

Como a vazão obteve um valor de 87.14% e a maré 12.86%, podemos concluir que a vazão foi mais importante que a maré para os níveis de cota de Propriá em 2011. Esse percentual mais elevado da vazão pode ser justificado ao levar em consideração a localização da cidade de Propriá na Figura 6, já que esta cidade é a mais distante da maré dentre as três analisadas.

Penedo

Figura 12. Importância das variáveis – Penedo (2011)



As variáveis mais importantes para a cidade de Penedo em 2011 estão exibidas na Figura 12, e o percentual das 6 mais importantes pode ser visto na Tabela 7.

Tabela 7. Variáveis mais importantes - Penedo (2011)

Variável	EMQ(%)
mare70	18.76
mare71	16.93
vazao40	16.71
mare72	16.50
vazao39	15.72
vazao21	15.52

Diferente dos resultados encontrados em Propriá para esse mesmo ano, a cidade de Penedo obteve tanto variáveis da maré quanto da vazão dentre as mais importantes, onde a maré de 3 horas passadas obteve o maior valor com 18.76%. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Penedo em 2011 está descrito na Tabela 8.

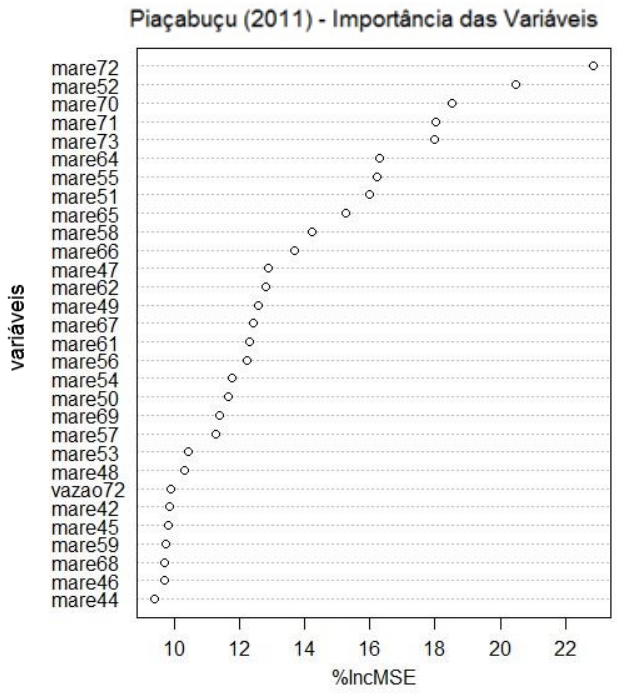
Tabela 8. Percentual das variáveis - Penedo (2011)

Variável	Percentual (%)	Grau de Importância
Maré	41.28	Menos importante
Vazão	58.72	Mais importante

Como a vazão obteve um valor de 58.72% e a maré 41.28%, podemos concluir que a vazão foi mais importante que a maré para os níveis de cota de Penedo em 2011. Como foi apresentado na Figura 6, a cidade de Penedo está localizada entre Propriá e Piaçabuçu, fazendo com que a influência da maré e da vazão sejam mais próximos em importância.

Piaçabuçu

Figura 13. Importância das variáveis – Piaçabuçu (2011)



A cidade de Piaçabuçu apresentou em sua grande maioria variáveis da maré como as mais importantes, como pode ser visto no o gráfico da Figura 13. A variável da maré defasada em 1 hora atrás foi a mais importante, apresentando um percentual de 22.86% conforme a Tabela 9.

Tabela 9. Variáveis mais importantes - Piaçabuçu (2011)

Variável	EMQ (%)
mare72	22.86
mare52	20.49
mare70	18.50
mare71	18.00
mare73	17.96
mare64	16.27

O percentual da importância de todas as variáveis da maré e vazão para a cidade de Piaçabuçu em 2011 está descrito na Tabela 10.

Tabela 10. Percentual das variáveis - Piaçabuçu (2011)

Variável	Percentual (%)	Grau de Importância
Maré	61.65	Mais importante
Vazão	38.35	Menos importante

Como a maré obteve um valor de 61.65% e a vazão 38.35%, podemos concluir que a maré foi mais importante que a vazão para os níveis de cota de Piaçabuçu em 2011. Esse resultado é justificado ao observar a localização de Piaçabuçu na Figura 6, já que a cidade é a que se encontra mais próxima da maré.

4.1.2 Importância das variáveis (2013)

A fim de simplificar a descrição dos resultados obtidos, os gráficos das variáveis mais importantes para cada base a seguir foram dispostos no Apêndice desta monografia.

Propriá

O gráfico encontrado para a cidade de Propriá no ano de 2013 encontra-se no APÊNDICE A, e as 6 variáveis mais importantes estão dispostas na Tabela 11.

Tabela 11. Variáveis mais importantes - Propriá (2013)

Variável	EMQ (%)
vazao1	21.12
vazao13	17.41
vazao14	16.90
vazao20	16.65
vazao15	16.51
vazao73	16.42

Analisando o gráfico do APÊNDICE A, vemos que todas a variáveis mais importantes foram da vazão, assim como nos resultados obtidos em 2011. O valor da importância das 6 variáveis mais importantes variaram entre 16.42% e 21.12%. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Propriá em 2013 está descrito na Tabela 12.

Tabela 12. Percentual das variáveis - Propriá (2013)

Variável	Percentual (%)	Grau de Importância
Maré	26.04	Menos importante
Vazão	73.96	Mais importante

A importância da vazão obteve um valor de 73.96% e a maré 26.04%, dessa forma, podemos concluir que a vazão foi mais importante que a maré para os níveis de cota de Propriá em 2013.

Penedo

O gráfico encontrado para a cidade de Penedo no ano de 2013 encontra-se no APÊNDICE B, e as 6 variáveis mais importantes estão dispostas na Tabela 13.

Tabela 13. Variáveis mais importantes - Penedo (2013)

Variável	EMQ (%)
mare69	13.87
vazao33	11.86
mare64	11.64
mare62	11.38
vazao35	11.08
mare71	10.67

Nos resultados demonstrados no gráfico do APÊNDICE B, podemos observar que a maré obteve mais variáveis importantes do que a vazão, diferente dos resultados obtidos em 2011. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Penedo em 2013 está descrito na Tabela 14.

Tabela 14. Percentual das variáveis - Penedo (2013)

Variável	Percentual (%)	Grau de Importância
Maré	55.01	Mais importante
Vazão	44.91	Menos importante

Avaliando os percentuais obtidos, a maré apresentou um valor de 55.01%, enquanto a vazão obteve 44.91% de importância. Sendo assim, podemos concluir que a maré foi mais importante que a vazão para os níveis de cota de Penedo em 2013.

Piaçabuçu

O gráfico encontrado para a cidade de Piaçabuçu no ano de 2013 encontra-se no APÊNDICE C, e as 6 variáveis mais importantes estão dispostas na Tabela 15.

Tabela 15. Variáveis mais importantes - Piaçabuçu (2013)

Variável	EMQ (%)
mare72	23.31
mare47	17.56
mare60	13.90
mare66	13.50
mare71	10.92
mare22	10.27

Os resultados demonstrados no gráfico do APÊNDICE C foram bem próximos dos escontrados para o ano de 2011, onde a maré obteve mais variáveis importantes do que a vazão, com valores entre 10.27% e 23.31%. O percentual da

importância de todas as variáveis da maré e vazão para a cidade de Piaçabuçu em 2013 está descrito na Tabela 16.

Tabela 16. Percentual das variáveis - Piaçabuçu (2013)

Variável	Percentual (%)	Importância
Maré	59.88	Mais importante
Vazão	40.12	Menos importante

Avaliando os percentuais obtidos, a maré apresentou um valor de 59.88%, enquanto a vazão obteve 40.12% de importância. Sendo assim, podemos concluir que a maré foi mais importante que a vazão para os níveis de cota de Piaçabuçu em 2013.

4.1.3 Importância das variáveis (2015)

Propriá

O gráfico encontrado para a cidade de Propriá no ano de 2015 encontra-se no APÊNDICE D, e as 6 variáveis mais importantes estão dispostas na Tabela 17.

Tabela 17. Variáveis mais importantes - Propriá (2015)

Variável	EMQ (%)
vazao1	14.86
vazao5	13.69
vazao3	12.26
vazao2	11.92
vazao4	11.56
vazao7	9.48

Assim como nos anos seguintes, Propriá continuou apresentanto valores de vazão com uma importância maior que nas outras cidades. Nesse ano, os valores de importância da vazão variaram entre 9.48% e 14.86%, um pouco menor que os resultados obtidos nos anos anteriores para a mesma cidade. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Propriá em 2015 está descrito na Tabela 18.

Tabela 18. Percentual das variáveis - Propriá (2015)

Variável	Percentual (%)	Grau de Importância
Maré	35.75	Menos importante
Vazão	64.25	Mais importante

A vazão obteve um percentual de 64.25% de importância, enquanto a maré obteve 35.75%. Comparando com os resultados obtidos nos anos anteriores, notamos que o percentual da vazão veio diminuindo ao longo dos anos, refletindo a diminuição da vazão devido ao período de seca do São Francisco. Apesar disso, a vazão continuou sendo mais importante que a maré para os níveis de cota de Propriá em 2015.

Penedo

O gráfico encontrado para a cidade de Penedo no ano de 2015 encontra-se no APÊNDICE E, e as 6 variáveis mais importantes estão dispostas na Tabela 19.

Tabela 19. Variáveis mais importantes - Penedo (2015)

Variável	EMQ (%)
mare67	16.81
mare65	14.61
mare71	13.36
vazao33	13.04
mare64	12.67
mare68	12.44

A cidade de Penedo continuou apresentando resultados mais aproximados entre a maré e a vazão, condizente com a sua localização geográfica. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Penedo em 2015 está descrito na Tabela 20.

Tabela 20. Percentual das variáveis - Penedo (2015)

Variável	Percentual (%)	Grau de Importância
Maré	48.52	Menos importante
Vazão	51.48	Mais importante

A vazão apresentou um percentual total de 51.48%, enquanto a maré apresentou 48.52% de importância. Esse ano obteve os dois percentuais mais próximos que nos anos passados, com uma diferença de apenas 2.96% entre a

importância da maré e vazão. Com esses resultados podemos concluir que a vazão foi mais importante que a maré para os níveis de cota de Penedo em 2015.

Piaçabuçu

O gráfico encontrado para a cidade de Piaçabuçu no ano de 2015 encontra-se no APÊNDICE F, e as 6 variáveis mais importantes estão dispostas na Tabela 21.

Tabela 21. Variáveis mais importantes - Piaçabuçu (2015)

Variável	EMQ (%)
mare72	21.93
mare73	16.77
mare59	16.23
mare66	15.89
mare55	14.74
mare70	13.65

Conforme os resultados obtidos nos anos anteriores, a cidade de Piaçabuçu continuou apresentanto a maré como sendo mais influente nos seus níveis de cota, variando seus valores entre 13.65% e 21.93%. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Piaçabuçu em 2015 está descrito na Tabela 22.

Tabela 22. Percentual das variáveis - Piaçabuçu (2015)

Variável	Percentual (%)	Grau de Importância
Maré	60.63	Mais importante
Vazão	39.37	Menos importante

A maré apresentou um percentual total de 60.63%, enquanto a vazão apresentou 39.37% de importância. Com esses resultados podemos concluir que a maré foi mais importante que a vazão para os níveis de cota de Piaçabuçu em 2015.

4.1.4 Importância das variáveis (2017)

Os valores de importância para o ano de 2017 foram referentes aos dados obtidos durante o mês de maio, pois a base da vazão não possuía dados para os meses de maio à outubro.

Propriá

O gráfico encontrado para a cidade de Propriá no ano de 2017 encontra-se no APÊNDICE G, e as 6 variáveis mais importantes estão dispostas na Tabela 23.

Tabela 23. Variáveis mais importantes - Propriá (2017)

Variável	EMQ (%)
vazao2	10.18
vazao1	9.85
vazao3	9.26
vazao40	6.89
vazao5	6.39
vazao36	6.25

Conforme os resultados obtidos nos anos anteriores, a cidade de Propriá continuou apresentando a vazão como sendo mais influente nos seus níveis de cota, embora os valores percentuais das variáveis mais importantes foram menores, variando entre 6.25% e 10.18%. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Propriá em 2017 está descrito na Tabela 24.

Tabela 24. Percentual das variáveis - Propriá (2017)

Variável	Percentual (%)	Importância
Maré	31.52	Menos importante
Vazão	68.48	Mais importante

O percentual da vazão mais uma vez foi mais elevado para a cidade de Propriá, apresentando um valor total de 68.48%, enquanto a maré apresentou 31.52% de importância. Com esses resultados podemos concluir que a vazão foi mais importante que a maré para os níveis de cota de Propriá em 2017.

Penedo

O gráfico encontrado para a cidade de Penedo no ano de 2017 encontra-se no APÊNDICE H, e as 6 variáveis mais importantes estão dispostas na Tabela 25.

Tabela 25. Variáveis mais importantes - Penedo (2017)

Variável	EMQ (%)
mare58	16.89
mare71	12.95
mare70	12.63
mare64	10.58
mare52	10.44
mare69	9.70

No gráfico do APÊNDICE H podemos observar que para esse ano as variáveis da maré foram mais presentes dentre as mais importantes, onde as 6 com maiores percentuais variaram entre 9.70% e 16.89%. O percentual da importância

de todas as variáveis da maré e vazão para a cidade de Penedo em 2017 está descrito na Tabela 26.

Tabela 26. Percentual das variáveis - Penedo (2017)

Variável	Percentual (%)	Grau de Importância
Maré	59.34	Mais importante
Vazão	40.66	Menos importante

Diferente dos resultados obtidos no ano anterior, a maré em Penedo no mês de maio de 2017 foi mais expressiva que a vazão, apresentando um percentual de 59.34%, enquanto a vazão apresentou 40.66% de importância. Dessa forma, concluímos que a maré foi mais importante que a vazão para os níveis de cota de Penedo em 2017.

Piaçabuçu

O gráfico encontrado para a cidade de Piaçabuçu no ano de 2017 encontra-se no APÊNDICE I, e as 6 variáveis mais importantes estão dispostas na Tabela 27.

Tabela 27. Variáveis mais importantes - Piaçabuçu (2017)

Variável	EMQ (%)
mare66	17.30
mare72	15.45
mare60	15.27
mare71	10.36
mare41	9.80
mare59	9.67

Os resultados encontrados para o mês de maio de 2017 foram condizentes com os resultados dos anos passados, onde a maré foi mais dominante entre as variáveis mais importantes. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Piaçabuçu em 2017 está descrito na Tabela 28.

Tabela 28. Percentual das variáveis - Piaçabuçu (2017)

Variável	Percentual (%)	Importância
Maré	61.02	Mais importante
Vazão	38.98	Menos importante

Os percentuais obtidos foram bem próximos aos do ano anterior, com a maré apresentando 61.02% de importância e a vazão 38.98%. Dessa forma, concluímos que a maré foi mais importante que a vazão para os níveis de cota de Piaçabuçu em 2017.

4.2 Random Forest aplicado para todos os anos

Após a execução do Random Forest nas bases das três cidades para os anos de 2011, 2013, 2015 e 2017, o algoritmo foi aplicado para as bases unificadas com os valores de todos os anos. Essa etapa foi realizada tanto para consolidar os resultados encontrados ao longo dos anos, quanto como forma de tentar encontrar resultados finais mais precisos sobre a influência da maré e vazão na cidade de Piaçabuçu.

Os resultados obtidos para as cidades de Propriá e Penedo servirão para entender o comportamento da maré e vazão ao longo do São Francisco, e também como isso influencia no resultado em Piaçabuçu. A importância das variáveis de entrada (maré e vazão) para a variável de saída (cota) em cada cidade será descrita nas subseção seguinte.

4.2.1 Importância das variáveis (todos os anos)

Propriá

O gráfico de importância encontrado para a cidade de Propriá encontra-se no APÊNDICE J, e as 6 variáveis mais importantes estão dispostas na Tabela 29.

Tabela 29. Variáveis mais importantes - Propriá (todos os anos)

Variável	EMQ (%)
vazao48	14.63
vazao9	14.50
vazao50	12.95
vazao20	12.32
vazao29	11.46
vazao40	11.14

Analisando o gráfico do APÊNDICE J, observamos que os resultados encontrados foram condizentes com os já obtidos em Propriá durante os 4 anos, onde a vazão foi a variável presente em todas as 30 mais importantes. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Propriá para a base com todos os anos está descrito na Tabela 30.

Tabela 30. Percentual das variáveis - Propriá (todos os anos)

Variável	Percentual (%)	Importância
Maré	28.57	Menos importante
Vazão	71.43	Mais importante

Conforme esperado, o percentual de importância da vazão foi bem superior ao da maré, obtendo um valor de 71.43%. Já a maré obteve um percentual de 28.57%, visto que a cidade é a que está mais distante da maré dentre as três analisadas. Comparando esse resultado com os obtidos durante os anos individualmente, concluímos que a vazão possui mais influência que a maré nos níveis da cota de Propriá.

Penedo

O gráfico de importância encontrado para a cidade de Penedo encontra-se no APÊNDICE K, e as 6 variáveis mais importantes estão dispostas na Tabela 31.

Tabela 31. Variáveis mais importantes - Penedo (todos os anos)

Variável	EMQ (%)
mare73	16.12
mare67	16.89
mare70	15.38
vazao3	14.65
mare64	13.93
mare62	13.92

Ao avaliar os resultados de Penedo ao longo dos anos, percebemos que existiu uma variação entre as variáveis mais influentes nos níveis da cota. Nos anos de 2011 e 2015 a vazão obteve uma maior importância, enquanto os anos de 2013 e 2017 possuíram a vazão como sendo mais importante. Como foi levantado na seção 4.1.1, essa variação pode ser justificada ao observar a localização da cidade, que está situada em um ponto mediano sob as influências da maré e do rio.

O gráfico de importância encontrado para a base com os dados de todos os anos foi bem próximo do encontrado para 2017, contendo mais variáveis da maré do que da vazão, onde a variável da maré mais importante obteve um percentual de

16.12%. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Penedo para a base com todos os anos está descrito na Tabela 32.

Tabela 32. Percentual das variáveis - Penedo (todos os anos)

Variável	Percentual (%)	Grau de Importância
Maré	56.90	Mais importante
Vazão	43.10	Menos importante

O percentual obtido indicou que a maré obteve um valor de 56.90%, enquanto a vazão obteve 43.10%. Dessa forma, concluímos que a maré possui mais influência que a vazão nos níveis da cota de Penedo.

Piaçabuçu

O gráfico de importância encontrado para a cidade de Piaçabuçu encontra-se no APÊNDICE L, e as 6 variáveis mais importantes estão dispostas na Tabela 33.

Tabela 33. Variáveis mais importantes - Piaçabuçu (todos os anos)

Variável	EMQ (%)
mare72	22.56
mare47	15.67
mare73	15.55
mare66	15.39
mare71	13.53
mare59	12.67

Ao avaliar os resultados encontrados para cada ano individualmente, observamos que a maré veio sendo mais importante para os níveis da cota em

Piaçabuçu, apresentando um valor percentual médio de 60.79%. Conforme os resultados obtidos em Propriá e Penedo, concluímos que a influência da maré nas cotas das cidades é maior quando elas se situam mais próximas da foz do São Francisco. Isso justifica os resultados encontrados em Piaçabuçu até agora.

O gráfico de importância para a base com os dados de todos os anos de Piaçabuçu, foi coerente com os dos anos individuais, contendo mais variáveis da maré do que da vazão. A variável mais importante corresponde à maré defasada em uma hora atrás, com um valor de 22.56%. O percentual da importância de todas as variáveis da maré e vazão para a cidade de Piaçabuçu para a base com todos os anos está descrito na Tabela 34.

Tabela 34. Percentual das variáveis - Piaçabuçu (todos os anos)

Variável	Percentual (%)	Grau de Importância
Maré	57.76	Mais importante
Vazão	42.24	Menos importante

O percentual obtido indicou que a maré obteve um valor de 57.76% de importância, bem próximo da média de 60.79% dos anos separados. Por outro lado, a vazão apresentou um percentual de 42.24%. Após a análise desses resultados em conjunto, concluímos que a maré possui mais influência que a vazão nos níveis de cota de Piaçabuçu.

4.3 Modelo de previsão de níveis utilizando a rede MLP

Com a última execução do Random Forest para a base completa de Piaçabuçu, foram identificadas as variáveis mais importantes para a previsão dos níveis de cota na cidade. Dessa forma, foram utilizadas as 6 variáveis com o maior percentual como entradas da rede MLP, com o intuito de se prever 24 saídas correspondentes aos níveis da cota para cada hora do dia. Porém, conforme descrito

na seção 3.4, uma boa prática para garantir a eficiência do modelo, é variar a quantidade de neurônios da camada escondida e utilizar o modelo que possuir o menor erro de validação cruzada.

Para garantir a eficiência do modelo, a rede foi executada para três cenários distintos, conforme as configurações listadas na Tabela 4. Ao fim de cada execução foram obtidos 30 EPMAs para cada cenário, onde foram aplicados os testes estatísticos para que pudéssemos identificar a configuração mais eficiente. As médias das taxas de erro encontradas para cada cenário estão contidas na tabela a seguir:

Tabela 35. Comparação dos valores do EPMA para cada cenário da rede MLP

Cenários	Quantidade de neurônios na camada escondida	Média dos EPMA (%)
Cenário 1	6	14.31
Cenário 2	12	14.16
Cenário 3	18	14.26

4.3.1 Comparação do Cenário 1 com o Cenário 2

Primeiramente foi utilizado o teste de Shapiro-Wilk para as amostras de erro do Cenário 1 e do Cenário 2. Ao final dos testes, o valor do p-value para ambos os cenários foi inferior à 0.05, indicando que as amostras não são normalmente distribuídas. Dessa forma, não poderemos utilizar um teste paramétrico, pois a primeira condição para a sua utilização não foi satisfeita. Foi então utilizado o teste de Wilcoxon para as duas amostras, onde o p-value obteve um valor superior à 0.05, indicando que os dois modelos são estatisticamente iguais. Como critério de desempate, foi considerado o cenário com o melhor desempenho, aquele que possuísse a menor taxa de erro. Para o Cenário 1, a média da taxa de erro foi de 14.31%, e para o Cenário 2 foi de 14.16%. Logo, o Cenário 2 foi o que possuiu o melhor desempenho. Os valores obtidos em cada teste podem ser observados na tabela a seguir:

Tabela 36. Resultados dos testes estatísticos para a comparação entre o Cenário 1 e o Cenário 2

Cenários	Shapiro-Wilk (p-value)	Wilcoxon (p-value)
Cenário 1	0.0003395	0.398
Cenário 2	2.084e-06	

4.3.2 Comparação do Cenário 1 com o Cenário 3

O teste de Shapiro-Wilk foi aplicado para as amostras de erro do Cenário 1 e do Cenário 3, obtendo um p-value inferior a 0.05 para ambos os resultados, indicando que as amostras não são normalmente distribuídas. Como não podemos utilizar um teste paramétrico, estaremos utilizando o teste de Wilcoxon mais uma vez. Conforme o resultado obtido na comparação anterior, o p-value para o teste de Wilcoxon foi superior à 0.05, o que significa que as medianas dos dois cenários são estatisticamente iguais. Dessa forma, o modelo que possui o melhor desempenho é o que apresenta a menor taxa de erro, ou seja, o Cenário 3. Os valores obtidos em cada teste podem ser observados na tabela a seguir:

Tabela 37. Resultados dos testes estatísticos para a comparação entre o Cenário 1 e o Cenário 3

Cenários	Shapiro-Wilk (p-value)	Wilcoxon (p-value)
Cenário 1	0.0003395	0.9474
Cenário 3	0.001371	

4.3.3 Comparação do Cenário 2 com o Cenário 3

Na última comparação de cenários, obtivemos um valor de p-value inferior a 0.05 para o teste de Shapiro-Wilk, indicando que as amostras não são normalmente distribuídas. Como não pudemos utilizar um teste paramétrico, foi executado o teste de Wilcoxon, obtendo mais uma vez um valor superior à 0.05 para o p-value. Isso indica que as medianas dos dois cenários são estatisticamente iguais. Como o

Cenário 2 é o que possui a menor taxa de erro, então ele é o que possui o melhor desempenho. Os valores obtidos em cada teste podem serobservados na tabela a seguir:

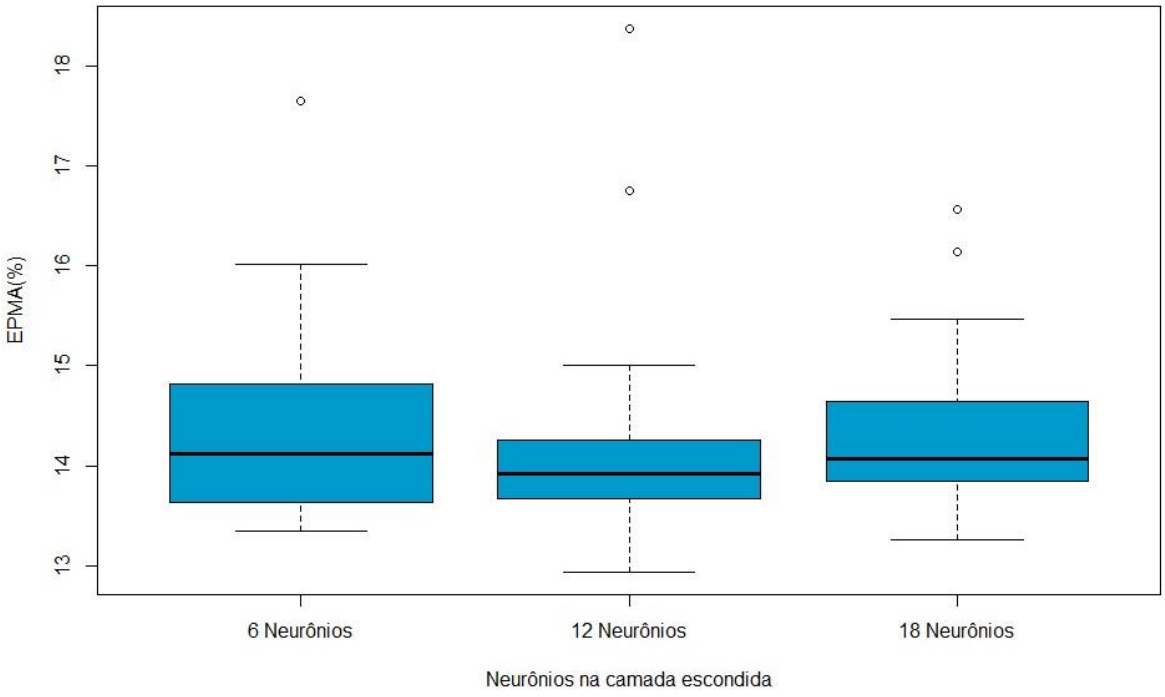
Tabela 38. Resultados dos testes estatísticos para a comparação entre o Cenário 2 e o Cenário 3

Cenários	Shapiro-Wilk (p-value)	Wilcoxon (p-value)
Cenário 2	2.084e-06	0.1973
Cenário 3	0.001371	

4.3.4 Boxplot

Ao avaliar o percentual dos erros obtidos para cada cenário, o que apresentou o menor erro foi o Cenário 2, com uma média de 14.16%. Para se obter uma melhor visualização da distribuição dos valores de EPMA de cada cenário, foi gerado o gráfico de boxplot a seguir:

Figura 14. Gráfico de boxplot para os EPMA de cada cenário da rede MLP



O gráfico nos mostra que o modelo que utilizou 12 neurônios na camada escondida, embora ainda possua dois valores mais distantes da média, foi o que

apresentou os menores erros. Em comparação com os outros cenários, o modelo também foi o que apresentou a menor mediana e limites mais simétricos.

Dessa forma, esse foi o modelo escolhido para treinar a rede MLP e realizar a previsão de níveis em Piaçabuçu para um dia à frente. Os resultados previstos poderão auxiliar na escolha do melhor horário para a captção de água, caso se utilize uma análise mais profunda sobre os índices de salinidade no rio.

Capítulo 5
Considerações Finais

Este capítulo tem como objetivo apresentar as conclusões obtidas neste trabalho. Na seção 5.1 será descrita as conclusões obtidas ao aplicar o Random Forest para medir a importância das variáveis da maré e vazão para a cota em Propriá, Penedo e Piaçabuçu, além da utilização da rede MLP para a previsão de níveis do rio. A seção 5.2 irá detalhar os trabalhos futuros para esse estudo.

5.1 Conclusões

Conforme os resultados obtidos pelo estudo realizado pela ANA, o estado de Alagoas apresentou apenas 18 municípios em condições satisfatórias para o abastecimento de água, onde algumas regiões apresentaram um avanço da salinidade [2]. A cidade de Piaçabuçu, locacalizada na foz do rio São Francisco, é um dos municípios cujo a qualidade do sistema de abastecimento precisa ser aprimorada.

Por esse motivo, esse trabalho teve como objetivo realizar um estudo para avaliar a influência da vazão e da maré nos níveis das cotas na estação fluviométrica de Piaçabuçu, visando auxiliar o sistema de captação de água com os resultados obtidos. O algoritmo Random Forest foi utilizado nas cidades de Propriá, Penedo e por fim, Piaçabuçu, a fim de se obter um panorama dos percentuais de influência nos níveis do rio São Francisco.

Os resultados obtidos para a estação fluviométrica de Propriá em todos os anos, indicaram que a vazão possui uma maior influência nos níveis da cota, obtendo um percentual de 71.43% de importância no total. Já a estação fluviométrica de Penedo apresentou uma certa oscilação nos resultados ao longo dos anos, onde a vazão foi mais importante nos anos de 2011 e 2015, e a maré foi mais importante em 2013 e 2017. Ao utilizar a base com os dados de todos os anos, a cidade de Penedo obteve 56.90% de importância. Esse comportamento serviu

para comprovar que a a influência da maré aumenta conforme o rio se aproxima da sua foz, o que ficou mais claro ao se obter os resultados de Piaçabuçu. Avaliando os percentuais obtidos para cada ano em Piaçabuçu, concluiu-se que a maré obteve maior importância para os níveis do rio do que a vazão, obtendo um percentual médio de 60.79%.

Com os valores de importância obtidos pelo Random Forest, foram utilizadas as 6 variáveis mais importantes como entrada da rede MLP, visando a criação de um modelo de previsão de níveis em Piaçabuçu. Ao rodar a MLP para três cenários diferentes, concluiu-se que utilizando 12 neurônios na camada escondida é possível obter resultados com um menor índice de erro.

5.2 Trabalhos Futuros

Após os resultados obtidos para este presente estudo, podemos listar os seguintes tópicos como trabalhos futuros:

- Adicionar os dados atualizados até 2018, já que neste trabalho foram utilizados dados até maio de 2017.
- Analisar separadamente o período seco (de maio à outubro) e o período úmido (de novembro à abril).
- Analisar os resultados para os meses mais secos (julho, agosto e setembro)
- Utilizar outras técnicas além do Random Forest para identificar a correlação das variáveis, como a Principal Component Analysis (PCA) e Stepwise Regression
- Utilizar o modelo de previsão para identificar o melhor horário para a captação de água, após uma análise dos índices de salinidade
- Obter outro modelo de previsão utilizando a rede Reservoir Computing (RC) e comparar os resultados com os obtidos pela rede MLP

Referências

[1] AGÊNCIA NACIONAL DE ÁGUAS (Brasil). Atlas Brasil: Abastecimento urbano de água. Panorama Nacional - Volume 1. Brasília, 2010.

[2] AGÊNCIA NACIONAL DE ÁGUAS (Brasil). Atlas Brasil: Abastecimento urbano de água. Resultados por Estado - Volume 2. Brasília, 2010.

[3] BARBOSA, Juliana Moreira; CARNEIRO, Tiago Garcia de Senna; TAVARES, Andrea Iabrudi, Métodos de Classificação por Árvores de Decisão, UFOP-Universidade Federal de Ouro Preto, Minas Gerais, Brasil

[4] BASTOS, Denise G. D., NASCIMENTO, Patricia S., LAURETTO, Marcelo S., Análise Empírica de Desempenho de Quatro Métodos de Seleção de Características para Random Forests, Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (EACH - USP), São Paulo, Brasil

[5] BREIMAN, L., Bagging predictors. Machine learning, 1996

[6] BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. and STONE, C. Peter. Classification and regression trees, 1984

[7] BREIMAN, L. (2001). Random forests. Machine Learning

[8] “Companhia Hidrelética do São Francisco (Chesf) - Sistema Chesf Xingó”. Disponível em: <https://www.chesf.gov.br/SistemaChesf/Pages/SistemaGeracao/Xingo.aspx> Acesso em: 20 de novembro de 2018

[9] DIETTERICH,Thomas G., Ensemble Methods in Machine Learning, Oregon State University, Corvallis, Oregon, USA

[10] DONGES, Niklas., The Random Forest Algorithm. Disponível em: <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd> Acesso em: 20 de novembro de 2018

[11] DORNELLES, Fernando. Previsão contínua de níveis fluviais com redes neurais utilizando previsão de precipitação : investigação metodológica da técnica. Dissertação. Universidade Federal do Rio Grande do Sul. Instituto de Pesquisas Hidráulicas. Programa de Pós-Graduação em Recursos Hídricos e Saneamento Ambiental. 2007.

[12] “HIDROWEB – Sistema de Informações Hidrológicas”. Disponível em: <http://www.snirh.gov.br/hidroweb/publico/apresentacao.jsf> Acesso em: 20 de novembro de 2018

[13] IANDOLI, Rafael. De onde vem a crise hídrica que seca a bacia do rio São Francisco. Nexo Jornal. 22 de out. 2017. Disponível em: <https://www.nexojornal.com.br/expresso/2017/10/22/De-onde-vem-a-crise-h%C3%ADdrica-que-seca-a-bacia-do-rio-S%C3%A3o-Francisco>. Acesso em: 24 de setembro de 2018

[14] IZENMAN , A.J. Modern Multivariate Statistical Techniques, 2008

[15] “Marinha do Brasil”. Disponível em: <https://www.marinha.mil.br/> Acesso em: 20 de novembro de 2018

[16] MINISTÉRIO DO MEIO AMBIENTE, Secretaria de Recursos Hídricos. Caderno da Região Hidrográfica do São Francisco – Brasília: MMA, 2006. 148p

[17] MITCHELL, Tom. Machine Learning. McGraw Hill, 1997.

[18] PEDROLLO, Olavo Correa. Previsão de curto prazo de níveis com Redes Neurais artificiais para a cidade de Estrela (RS): Resultados preliminares. In: Simpósio Brasileiro de Recursos Hídricos (22. : Florianópolis, 2017). Anais [recurso eletrônico]. [Porto Alegre : ABRH, 2017]

[19] RUSSEL, Stuart.; NORVIG, Peter. Inteligência Artificial, 3.ed. Rio de Janeiro: Elsevier, 2013

[20] SANTOS, R.T. F. Mudanças climáticas e a zona costeira: uma análise do impacto da subida do nível do mar nos recursos hídricos – o caso do canal de

São Francisco – baía de Sepetiba – RJ. 127 f. Dissertação (Mestrado) – Programa de Pós-graduação em Planejamento Energético, COPPE, da Universidade Federal do Rio de Janeiro. Rio de Janeiro, 2012.

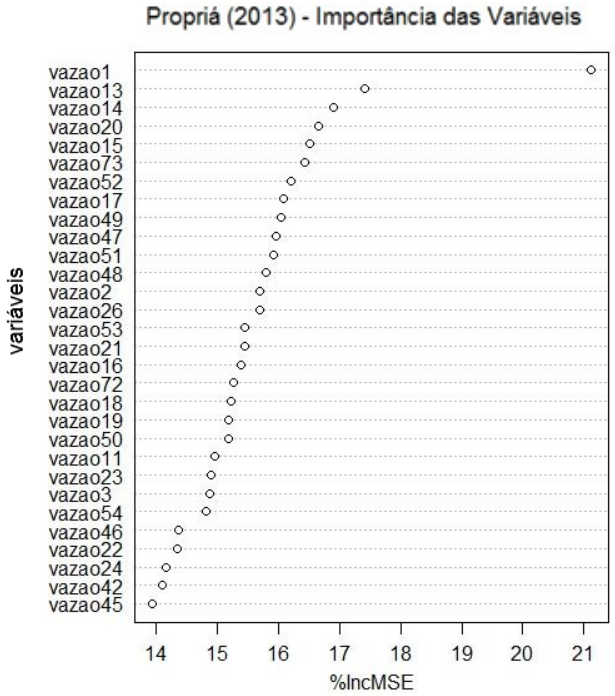
[21] “The R Project for Statistical computing” [Online]. Disponivel em: <https://www.r-project.org/> Acesso em: 15 de novembro de 2018.

[22] VALENÇA, Mêuser. Fundamentos De Redes Neurais. 2.ed. Olinda, PE: Livro Rápido, 2010.

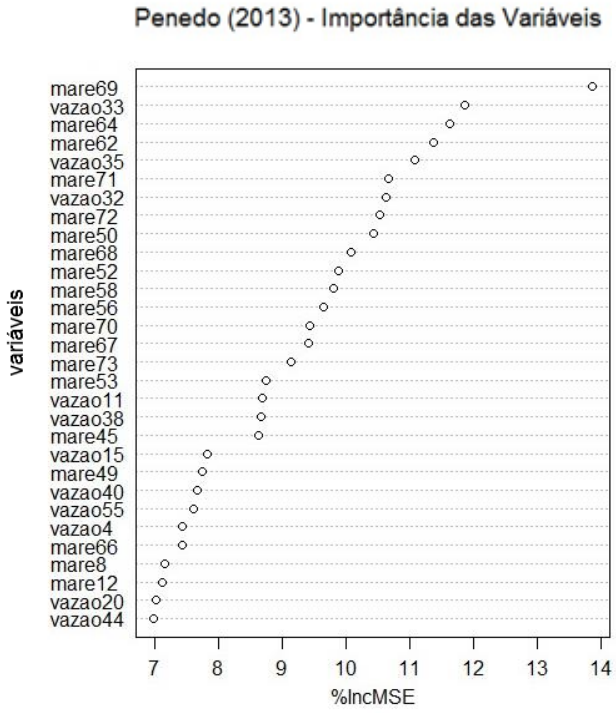
[23] VALERIO, Liziara de Mello. Previsão do Nível do Mar por Redes Neurais Artificiais: Um Estudo de Caso para o Litoral Norte de São Paulo. 2016. 117 f. Dissertação (Mestrado em Meio Ambiente e Recursos Hídricos) – Universidade Federal de Itajubá, Itajubá, 2016.

Apêndice

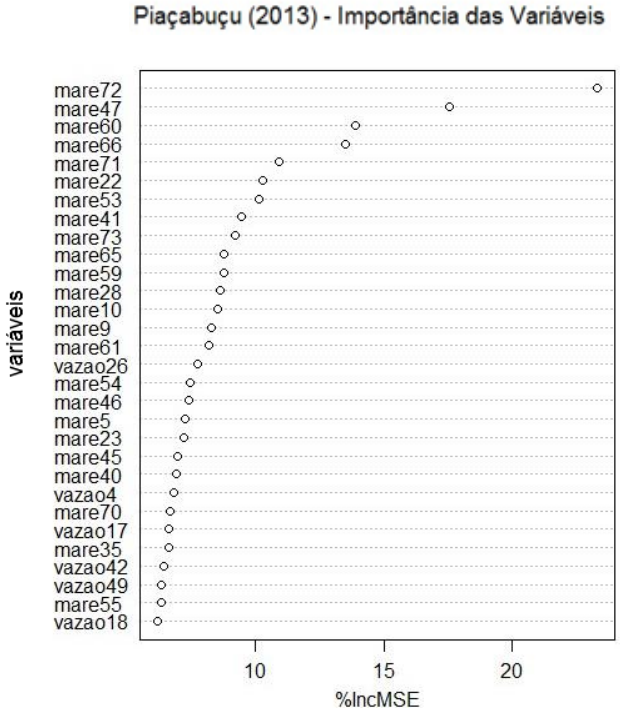
APÊNDICE A – IMPORTÂNCIA DAS VARIÁVEIS – PROPRIÁ (2013)



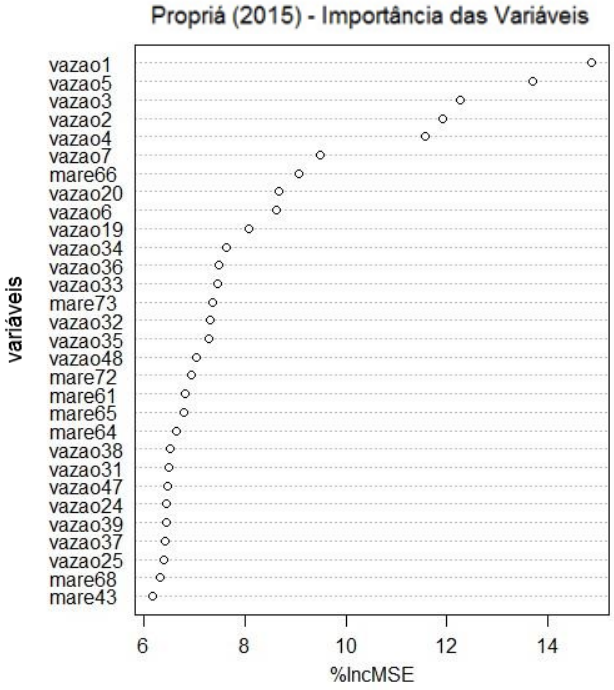
APÊNDICE B – IMPORTÂNCIA DAS VARIÁVEIS – PENEDO (2013)



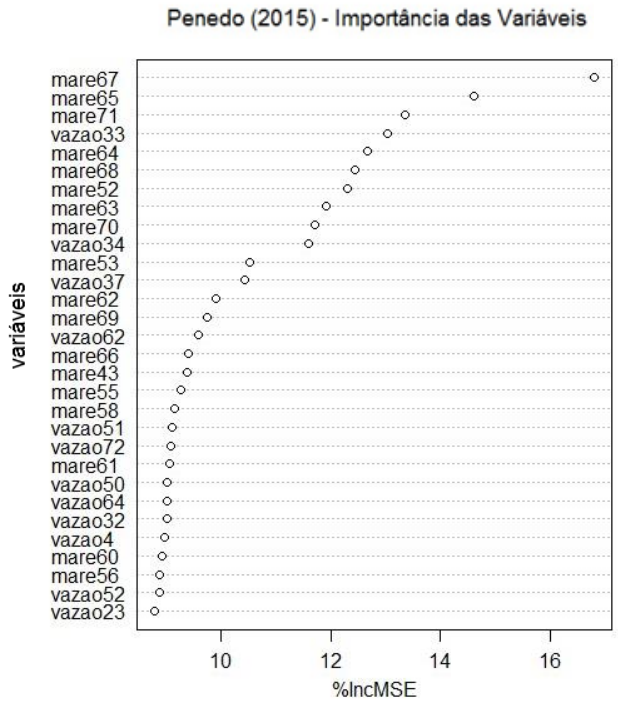
APÊNDICE C – IMPORTÂNCIA DAS VARIÁVEIS – PIAÇABUÇU (2013)



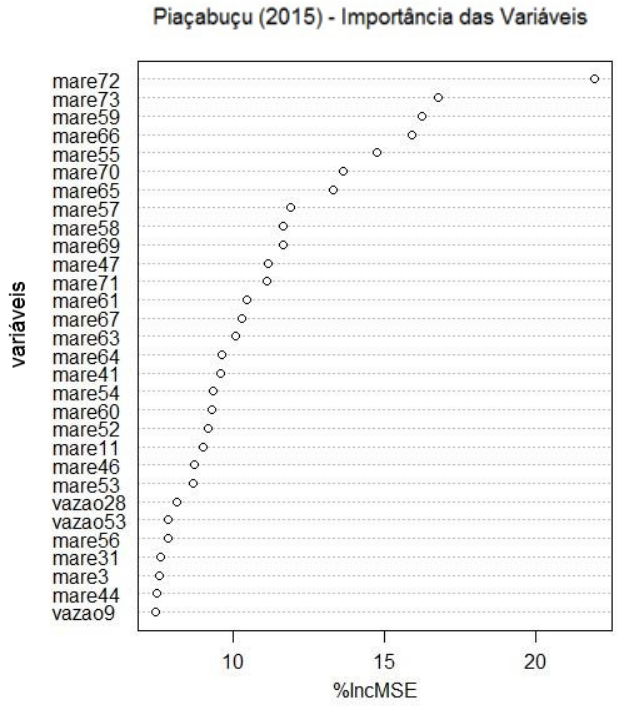
APÊNDICE D – IMPORTÂNCIA DAS VARIÁVEIS – PROPRIÁ (2015)



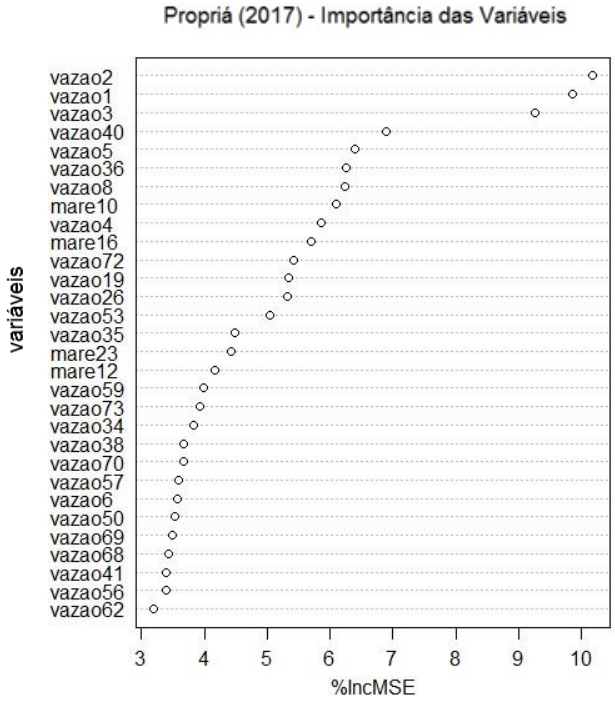
APÊNDICE E – IMPORTÂNCIA DAS VARIÁVEIS – PENEDO (2015)



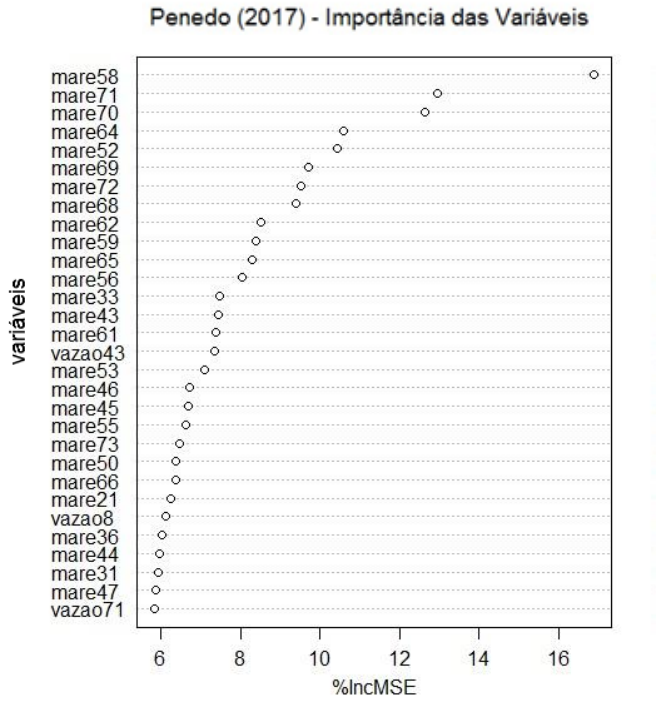
APÊNDICE F – IMPORTÂNCIA DAS VARIÁVEIS – PIAÇABUÇU (2015)



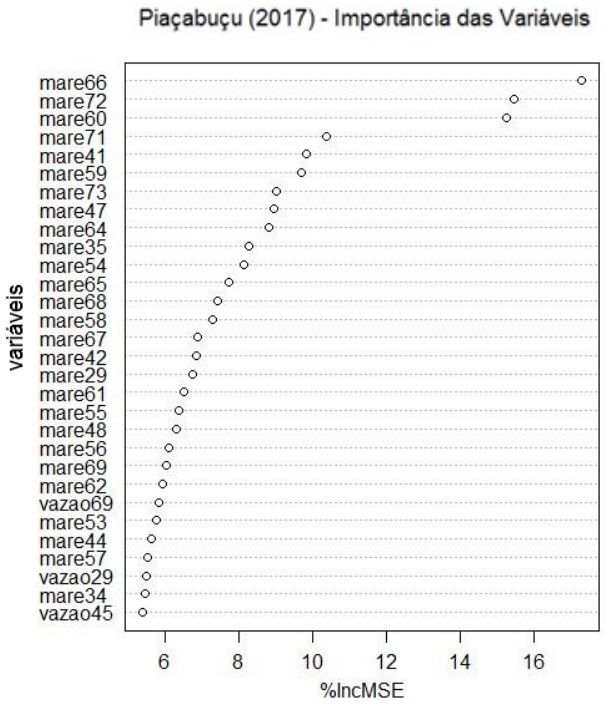
APÊNDICE G – IMPORTÂNCIA DAS VARIÁVEIS – PROPRIÁ (2017)



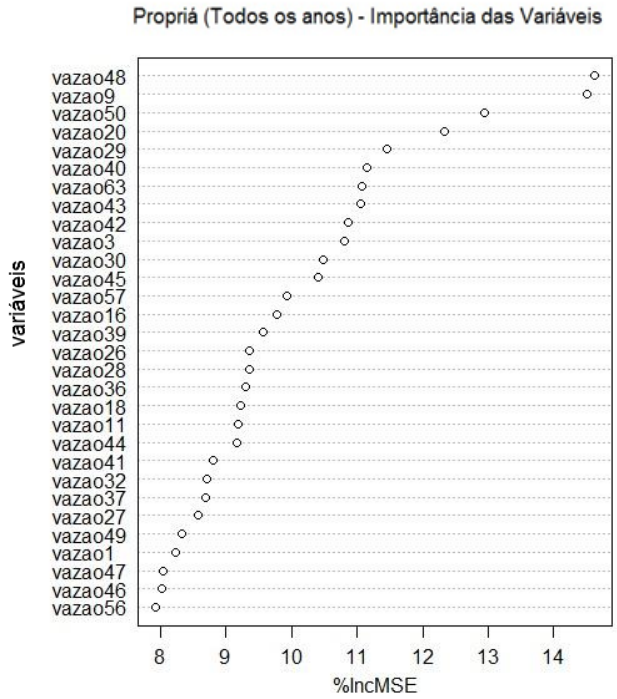
APÊNDICE H – IMPORTÂNCIA DAS VARIÁVEIS – PENEDO (2017)



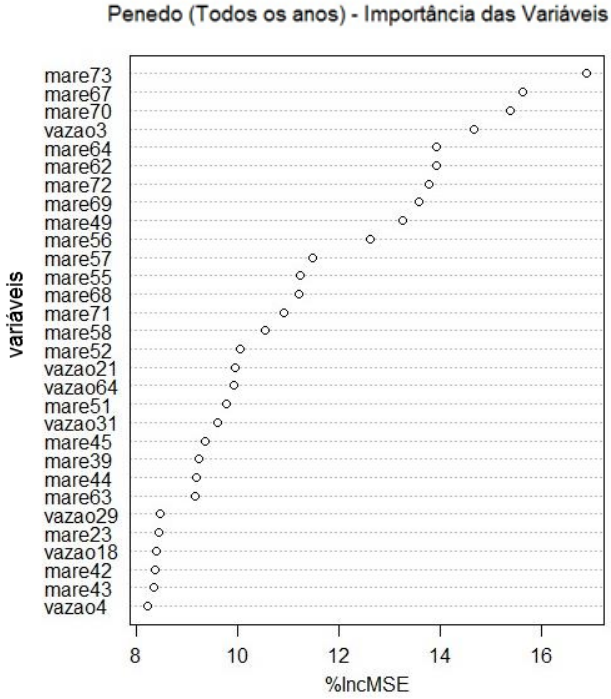
APÊNDICE I – IMPORTÂNCIA DAS VARIÁVEIS – PIAÇABUÇU (2017)



APÊNDICE J – IMPORTÂNCIA DAS VARIÁVEIS – PROPRIÁ (TODOS OS ANOS)



APÊNDICE K – IMPORTÂNCIA DAS VARIÁVEIS – PENEDO (TODOS OS ANOS)



APÊNDICE L – IMPORTÂNCIA DAS VARIÁVEIS – PIAÇABUÇU (TODOS OS ANOS)