UTILIZANDO A TÉCNICA RANDOM FOREST PARA DETERMINAR A IMPORTÂNCIA DE VARIÁVEIS PARA PREVISÃO DE NÍVEIS NO POSTO DA CIDADE DE PIAÇABUÇU

Trabalho de Conclusão de Curso

Engenharia de Computação

Marcela Maria Coelho Campos
Orientador: Prof. Mêuser Jorge Silva Valença

UNIVERSIDADE DE PERNAMBUCO

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

MARCELA MARIA COELHO CAMPOS

UTILIZANDO A TÉCNICA RANDOM FOREST PARA DETERMINAR A IMPORTÂNCIA DE VARIÁVEIS PARA PREVISÃO DE NÍVEIS NO POSTO DA 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, Junho de 2019

ESCOLA POLITÉCNICA
DE PERNAMBUCO

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 02/07/2019, às 10h, reuniu-se para deliberar sobre a defesa da monografia de conclusão de curso do(a) discente **MARCELA MARIA COELHO CAMPOS**, orientado(a) pelo(a) professor(a) **MÊUSER JORGE SILVA VALENÇA**, sob título UTILIZANDO A TÉCNICA RANDOM FOREST PARA DETERMINAR A IMPORTÂNCIA DE VARIÁVEIS PARA PREVISÃO DE NÍVEIS NO POSTO DA CIDADE DE PIAÇABUÇU, a banca composta pelos professores:

SERGIO MARIO LINS GALDINO (PRESIDENTE)
MÊUSER JORGE SILVA VALENÇA (ORIENTADOR)

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

Sérgio Galdino
AVALIADOR 1: Prof (a) **SERGIO MARIO LINS GALDINO**

Mêuser Jorge Silva Valença
AVALIADOR 2: Prof (a) **MÊUSER JORGE SILVA VALENÇA**

AVALIADOR 3: Prof (a)

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

ESCOLA POLITÉCNICA
DE PERNAMBUCO

Autorização de publicação de PFC

Eu, **Marcela Maria Coelho Campos** autor(a) do projeto de final de curso intitulado: **UTILIZANDO A TÉCNICA RANDOM FOREST PARA DETERMINAR A IMPORTÂNCIA DE VARIÁVEIS PARA PREVISÃO DE NÍVEIS NO POSTO DA CIDADE DE PIAÇABUÇU;** autorizo a publicação de seu conteúdo na internet nos portais da Escola Politécnica de Pernambuco e Universidade de Pernambuco.

O conteúdo do projeto de final de curso é de responsabilidade do autor.

Marcela Campos

Marcela Maria Coelho Campos

Orientador(a): **Mêuser Jorge Silva Valença**

Coorientador(a):

Prof, de TCC: **Daniel Augusto Ribeiro Chaves**

Data: 02/07/2019

Agradecimentos

Primeiramente gostaria de agradecer ao professor Mêuser pela oportunidade de realizar este trabalho sob sua supervisão. Obrigada professor, pela paciência, orientação, e apoio ao longo deste semestre. Agradeço aos outros professores do curso de Engenharia da Computação da UPE, que fizeram parte desta caminhada, colaborando para que eu pudesse evoluir cada vez mais com a minha formação profissional e acadêmica.

Gostaria de agradecer imensamente aos amigos que fiz na graduação e que fizeram parte da minha jornada, fazendo com que esse período fosse repleto de boas lembranças e de saudades. Agradeço em especial a Everalda por me ajudar na execução deste projeto e ao meu grande amigo Gustavo, companheiro desde o primeiro semestre da graduação, que esteve sempre ao meu lado, me ajudando em diversos momentos de dificuldades.

Agradeço a empresa CESAR, que me acolheu desde 2014 e que, além de permitir que eu conciliasse meus estudos com o trabalho, sempre incentivou para que eu jamais deixasse de me dedicar a minha graduação. Obrigada Ticiana, Paloma, Tutuca e Gep, pela compreensão em todos os momentos que precisei me ausentar, ou trabalhar em horários alternativos.

Agradeço muito a Felipe, meu namorado, não somente pela paciência e apoio, mas pela determinação em me ajudar. Pela dedicação em parar e aprender um conteúdo para poder me ensinar, pela paciência e preocupação comigo e com meus resultados na universidade, e em especial pela ajuda neste trabalho. Você foi meu companheiro em todos os momentos.

Agradeço a minha família, por estar sempre ao meu lado, me dando suporte em todas as situações de dificuldades e celebrando minhas vitórias como se fossem de cada um de deles. Agradeço a meu pai, Fernando, por me ensinar a lutar por pelos meus objetivos, apesar das dificuldades. Agradeço as minhas irmãs Carolina e Fernanda pelo incentivo e pela revisão deste trabalho. Agradeço as minhas tias e tios pela torcida e incentivo ao longo dos anos. Agradeço a meu primo Philipe, grande

amigo e incentivador incansável, que me ajudou a enxergar várias vezes a grandeza do que eu estava construindo.

Por fim agradeço a minha mãe, Angela, base e fortaleza incessante, suporte para todas os momentos de dificuldade e maior incentivadora deste sonho, meu e dela chamado Graduação. Obrigada por cada palavra de apoio, por todos os abraços de celebração e também pelos de consolo, pela torcida e pela compreensão nas dificuldades ao longo deste caminho.

Resumo

Em termos globais de distribuição de água, o Brasil é um dos países que mais se destaca, pois, estima-se que o país possua cerca de 12% da disponibilidade de água doce do planeta. Entretanto, com a grande variabilidade climática característica do Brasil, a distribuição desses recursos para cada região se apresenta de forma desigual. Diante deste cenário, algumas regiões do Brasil dependem de bacias hidrográficas para abastecerem seus postos, como a cidade Piaçabuçu, no estado de Alagoas, que se encontra na foz do rio São Francisco. Devido a sua localização, Piaçabuçu sofre influência direta da vazão do rio e maré na sua principal fonte de abastecimento de água. Este trabalho tem como objetivo propor uma metodologia para avaliar qual variável (vazão ou maré) mais influência os níveis de água desta cidade. Para realizar este objetivo será utilizado o algoritmo de aprendizagem de máquina Random Forest, para regressão de valores, a partir de uma base de dados com valores da maré e da vazão entre os anos de 2011 a 2017. Para efeito de comparação e avaliação da metodologia proposta se fará igual análise nos postos hidrométricos localizados nas cidades de Propriá (SE) e Penedo (AL), sendo Piaçabuçu o foco deste trabalho.

Palavras-chave: árvores de decisão, métodos ensemble, random forest, variable importance

Abstract

In terms of global water distribution, Brazil is one of the countries that stands out the most, because it is estimated that the country has about 12% of the availability of fresh water on the planet. However, with the great climatic variability characteristic of Brazil, the distribution of these resources to each region is uneven. Given this scenario, some regions of Brazil rely on hydrographic basins to supply their stations, such as the city of Piaçabuçu, in the state of Alagoas, which lies at the mouth of the São Francisco River. Due to its location, Piaçabuçu suffers direct influence of the flow of the river and tide in its main source of water supply. This work aims to propose a methodology to evaluate which variable (flow or tide) more influence the water levels of this city. To achieve this objective, the Random Forest machine learning algorithm will be used for value regression from a database with tidal and flow values between 2011 and 2017. For the purpose of comparison and evaluation of the proposed methodology, the same analysis will be done at the hydrometric stations located in the cities of Propriá (SE) and Penedo (AL), Piaçabuçu being the focus of this work.

Keywords: random forest, decision trees, ensemble methods, bagging, variable importance

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
OOB	Out of Bagging

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 principais objetivos e metas do estudo. Na terceira seção é apresentada a estrutura da monografia.

1.1 Caracterização do Problema

Em termos globais de distribuição de água, o Brasil é um dos países que mais se destaca, pois, estima-se que o país possua cerca de 12% da disponibilidade de água doce do planeta [1]. Entretanto, com a grande variabilidade climática característica do Brasil, a distribuição desses recursos para cada região se apresenta de forma desigual, podendo-se verificar extremos: Região Hidrográfica Atlântico Nordeste Oriental, que abrange 5 estados do Nordeste, apresenta disponibilidade hídrica inferior a 100m³/s. Por outro lado, na região Hidrográfica do Amazônia, a disponibilidade hídrica é extremamente elevada, alcançando vazões da ordem de 74 mil m³/s [2].

Grandes centros urbanos, onde há elevada densidade populacional, como cidades próximas ao oceano atlântico, que concentram cerca de 45% da população brasileira, tem dificuldade de disponibilidade de água. Estes centros detêm menos de 3% dos recursos hídricos do país e dependem de bacias hidrográficas para utilização de água doce [2].

A região Hidrográfica São Francisco ocupa 7,5% do território brasileiro, abrangendo sete estados: Bahia, Minas Gerais, Pernambuco, Alagoas, Sergipe, Goiás e Distrito Federal. Esta região hidrográfica, que abastece as cidades Piaçabuçu, Penedo e Propriá vem sofrendo períodos de baixa vazão, apresentando situações de escassez de água, o que afeta diretamente o abastecimento de água da população [19]. Devido ao fato da cidade Piaçabuçu estar localizada em uma região limítrofe

entre o mar e o rio, torna-se mais complicado determinar se é a vazão do rio ou a maré que mais influência nos níveis de água desta cidade.

Diante do cenário de períodos de escassez de água nestas cidades, e dos resultados satisfatórios observados em (Bastos, et al.) para seleção de variáveis utilizando Random Forest [5], este trabalho propõe utilizar a técnica de aprendizagem de máquina Random Forest, proposta por Breiman (2001), com o objetivo de investigar qual a relação das variáveis de vazão e maré com os níveis de água destas cidades. Ou seja, determinar qual destas variáveis (maré ou vazão) tem maior importância para previsão dos níveis nos referidos postos hidrométricos.

Para realizar esta análise iremos utilizar dados cedidos pela Chesf, ANA e dados da maré obtidos do site da Marinha do Brasil [17] entre os anos de 2011 a 2017, que correspondem a valores médios da vazão liberada pela usina Hidrelétrica de Xingó, níveis dos postos hidrométricos localizados nas cidades de Propriá, Penedo e Piaçabuçu e da maré no Terminal Marítimo Inácio Barbosa localizado em Sergipe. Estes dados serão aplicados a ferramentas utilizando o algoritmo de aprendizado de máquina Random Forest, que é uma técnica que utiliza árvores de decisão, e pode ser utilizado para classificação ou regressão, dependendo do objetivo do projeto [13].

Neste projeto, será utilizado para regressão, uma vez que nosso objetivo é determinar a importância de variáveis dentro de um problema de previsão.

1.2 Objetivos

1.2.1 Objetivos Gerais

Objetivo geral deste projeto é utilizar a técnica de aprendizagem de máquina Random Forest para avaliar a influência da maré e da vazão nos postos de Própria, Penedo e Piaçabuçu localizados no trecho baixo do rio São Francisco.

1.2.2 Objetivos Específicos

- Determinar para cada posto a importância das cotas de maré e da vazão do rio São Francisco de forma a estabelecer quem mais controla cada posto.
- Avaliar qual a métrica de seleção de variáveis a ser utilizada na técnica de Random Forest para representar melhor o problema.

- Comparar os resultados de Random Forest com duas ferramentas: RStudio e XLSTAT

1.3 Estrutura da Monografia

No capítulo 1 encontra-se a introdução ao problema, apresentando as motivações e os principais objetivos a serem alcançados através desse trabalho. No capítulo 2 será apresentado a fundamentação teórica, abordando o conteúdo necessário para a compreensão deste trabalho, explorando os temas: Árvores de Decisão, Métodos Ensemble, Bagging e Random Forest. O Capítulo 3 aborda a metodologia utilizada para a realização deste trabalho. Será descrito como foram obtidas as bases utilizadas, e as configurações utilizadas no Random Forest para que fossem obtidos os resultados esperados. No capítulo 4 serão apresentados os resultados obtidos após a utilização do Random Forest nas bases de dados das cidades em estudo. O capítulo 5 apresenta as conclusões obtidas através dos resultados, além de descrever os possíveis trabalhos futuros para este estudo.

Com o objetivo de se utilizar o algoritmo Random Forest citado anteriormente, com o intuito de determinar quais das variáveis vazão ou maré mais influenciam os postos das cidades em estudo, será abordado em seguida o conteúdo necessário para melhor compreensão do algoritmo Random Forest.

Capítulo 2
Fundamentação Teórica

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

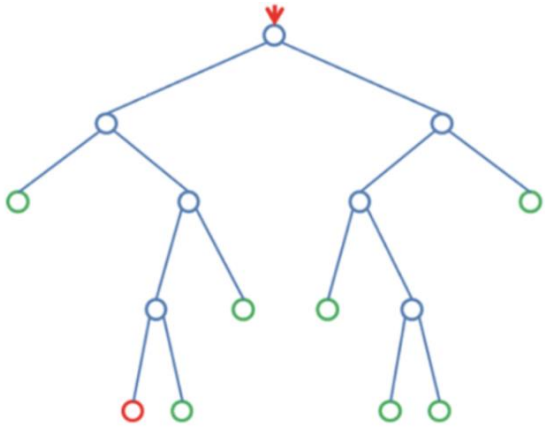
Na seção 2.1 serão abordados conceitos de Árvores de Decisão, como uma maneira introdutória para o entendimento do algoritmo Random Forest, que será utilizado nesse trabalho. A seção 2.2 aborda o conteúdo Método Ensemble, de modo a introduzir o conceito que será abordado em seguida, na seção 2.3 Método Bagging. Por fim, será abordado o conteúdo Random Forest, foco principal deste trabalho, na seção 2.4

2.1 Árvores de Decisão

As árvores de decisão são modelos estatísticos representadas por um conjunto de condições que divide a amostra de aprendizagem em partes cada vez menores [16]. Esta abordagem utiliza o conceito “dividir para conquistar”, onde cada subgrupo gerado está cada vez mais refinado e próximo do resultado desejado. Trata-se de uma das metodologias mais utilizadas na área de mineração de dados para problemas de classificação ou regressão de objetos.

A figura 1 mostra um modelo para uma árvore de decisão: No topo, encontra-se o nó raiz, seguido de nós internos (nós subsequentes, abaixo do nó raiz), constituídos a partir de uma sequência de decisões. A cada decisão, um conjunto de nós subsequentes será definido para cada resposta da decisão e esses conjuntos de nós formarão os ramos da árvore. São nas folhas, os últimos nós de cada ramo (também chamados de nós terminais) que as árvores de decisão determinam seus resultados, pois, estes nós contém os valores de predição definidos a partir de todas as decisões tomadas previamente.

Figura 1. Exemplo de uma árvore de decisão

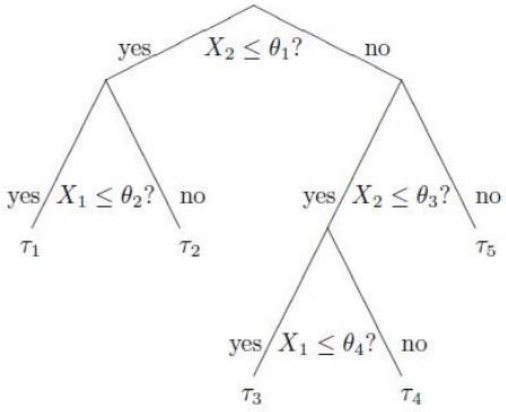


[Fonte:https://machinelearning-blog.com/2018/02/06/the-random-forest-algorithm/]

As características das decisões de uma árvore de decisão as dividem em árvores de classificação ou árvores de regressão. Breiman (1998), utiliza a metodologia CART (Classification and Regression Trees) para construir árvores de decisão a partir das seguintes maneiras:

- **Árvores de Classificação**: Consistem em árvores determinísticas. Ou seja, a partir de um conjunto de decisões discretas, determinadas a partir de uma condição, tal qual $\times \geq 0$, por exemplo, o conjunto de dados será separado em subgrupos que contém as mesmas características. Daí então se dará uma sequência de decisões similares aos exemplos citados acima. Essa ação irá se repetir até que haja um conjunto de nós terminais ideais, de onde seja possível determinar um resultado satisfatório, tal qual, como mostra a figura 2 a seguir:

Figura 2. Exemplo de árvore de classificação



Fonte: IZENMAN, A.J. Modern Multivariate Statistical Techniques, 2008 [15]

Neste caso, o nó raiz é $X_2 \leq \theta_1$, a partir do qual o conjunto de dados irá ser subdivido nos ramos subsequentes que possuem a mesma característica determinada a partir do primeiro nó de decisão. $X_1 \leq \theta_2$ e $X_2 \leq \theta_3$ são nós subsequentes dos quais serão construídos os próximos ramos desta árvore. As folhas, ou nós terminamos para este exemplo serão os nós $\tau_1 - \tau_5$, de onde será obtido os resultados. É importante observar que a sequência de ações em uma árvore de classificação é dependente. Ou seja, o tipo de pergunta feita para dividir cada subgrupo é determinado a partir da pergunta de decisão feita anteriormente.

- **Árvores de Regressão**: A determinação de árvores de regressão segue o mesmo caminho das árvores de classificação no que diz respeito ao processo de decisões ir dividindo o problema em subgrupos menores. Porém, para BREIMAN et al. (1996), este tipo de árvore é mais simples, pois, não há antecedentes para levar em consideração e também porque cada caso possui o mesmo peso.

  Ainda segundo BREIMAN et al. (1996), a análise de regressão pode ter 2 objetivos principais: (1) predizer a variável de resposta correspondente a futuros vetores de medida com a maior precisão possível; (2) compreender as relações estruturais entre a resposta e as variáveis medidas.

Um conceito fundamental sobre árvores de decisão, é a entropia. Este termo refere-se ao grau de pureza, ou, grau de desordem de uma amostra [10]. Pode-se entender que se uma amostra possui um alto grau de entropia, teremos muitos dados, porém estes dados estarão desordenados e, portanto, pouca informação pode ser inferida desta amostra.

Por outro lado, quanto menor o grau de entropia em uma amostra, mais informação poderá ser coletada sobre esta, uma vez que os dados estão mais homogêneos. A Entropia em uma amostra pode ser determinada através do seguinte cálculo:

$$Entropia\ (S) = -\sum_{i=1}^{c} p_i \log_2 p_i \qquad (2.1)$$

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* [10].

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, resultando em dados mais homogêneos e, por consequência, mais ganho [19].

Para determinar o ganho de um atributo A, é necessário calcular primeiramente a entropia resultante ao particionar o conjunto S, em função deste atributo. [10] Este cálculo é obtido através da seguinte fórmula:

$$Entropia\ (A) = \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropia\ (S_v) \qquad (2.2)$$

Onde

- S_v é um subconjunto de S onde cada atributo possui o valor v
- $Valores\ (A)$ representa o conjunto de todos os valores possíveis do atributo A

Por fim, o cálculo do ganho do atributo A pode ser determinado da seguinte maneira:

$$Ganho\ (S, A) = Entropia\ (S) - Entropia\ (A) \tag{2.3}$$

Onde $Emtropia\ (S)$ corresponde a entropia calculada para o conjunto de treinamento S e $Entropia\ (A)$ corresponde a uma nova entropia para o conjunto S, caso seja gerado um novo conjunto a partir do atributo A [19].

Um conceito abordado na metodologia CART, a respeito do grau de pureza dos nós, conforme estes são subdivididos é o índice de Gini [7]. Este índice corresponde a heterogeneidade dos dados e pode ser calculado através da seguinte expressão:

$$Gini = 1 - \sum_{i=1}^{c} p_i^2 \tag{2.4}$$

Onde *c* corresponde ao número de classes e p_i representa a frequência relativa de cada classe em cada nó. Diz-se que um nó é puro, conforme o índice de Gini deste nó obtém valores cada vez mais próximos de zero. Por outro lado, conforme os valores deste índice se aproximam de um, este nó é considerado impuro e o número de classes uniformemente distribuídas por este nó aumenta [4].

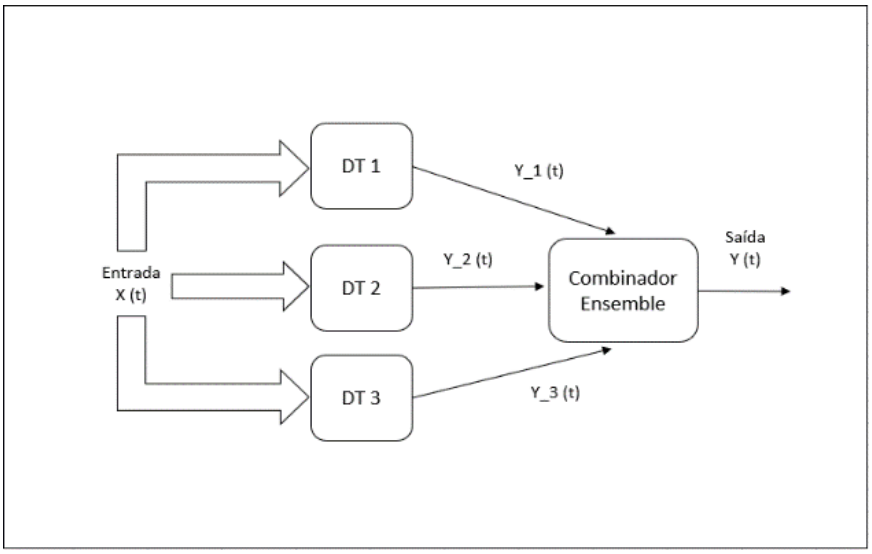
2.2 Métodos Ensemble

O método Ensemble consiste na combinação de um conjunto de modelos, onde cada um resolve a mesma atividade, com o intuito de se encontrar um modelo global composto que tenha resultados melhores, com estimativas e decisões mais precisas e confiáveis do que a que se pode obter utilizando apenas um destes modelos individualmente [20].

Dietterich (2000) define o método Ensemble como algoritmos de aprendizagem que constroem um conjunto de classificadores e então classificam novas bases utilizando pesos para votarem em suas predições [12]. Este conjunto de classificadores pode ser de diversos tipos, como redes neurais ou outros algoritmos de aprendizado de máquina, como árvores de decisão [18].

A figura 4 demonstra o comportamento de um método Ensemble em que 3 algoritmos de Árvores de Decisão DT1, DT2 e DT3 são utilizados para classificar um objeto. Seus resultados Y_1(t), Y_2(t), Y_3(t) são então combinados em um método Ensemble para gerar uma nova saída Y(t):

Figura 3. Exemplo do método Ensemble utilizando Árvores de Decisão como classificadores



Adaptado de: DE CASTRO; BRAGA; ANDRADE (2019) [11]

Uma questão fundamental sobre o método Ensemble é a precisão dos classificadores utilizados na combinação. Se tomarmos como exemplo a figura 4, e considerarmos DT1, DT2, DT3 como modelos de árvore de decisão que utilizam algoritmos iguais, teremos a mesma generalização para este conjunto, e por esta razão essa combinação não terá um bom desempenho. [11] Sendo assim, se DT1 errar na sua predição, existem grandes chances de DT2 e DT3 também errarem. Por esta razão, se faz necessário a utilização de classificadores independentes, com baixa correlação entre eles, para que caso um dos classificadores erre, os outros ainda possam acertar [12].

2.3 Bootstrap Aggregating

A técnica Bootstrap Aggregating, ou, Bagging, criada por Breiman (1996), foi definida como um método onde múltiplas versões de um preditor são geradas para então se obter um preditor agregado, gerando um único resultado final [6]. Trata-se de um método do tipo Ensemble bastante utilizado na área de aprendizagem de máquina.

Para construir o método Bagging é necessário utilizar um número y de conjuntos de treinamento, tal que estes não sejam métodos iguais. Em seguida deve-se replicar este conjunto de treinamento, de forma aleatória, a fim de construir um conjunto independente, por remostarem com reposição [3]. Breiman (1996) define esta metodologia da seguinte maneia:

Inicialmente considere o seguinte conjunto de dados $S = \{(y_n, x_n), n = 1 \dots, N\}$ onde y representa o resultado da metodologia de classificação aplicada. Esta metodologia poderá ser classificação ou regressão. O previsor utilizado nesta abordagem é representado por $\varphi = (x, S)$, onde x representa o dado passado como parâmetro de entrada. O subconjunto de preditores deste conjunto é demonstrado por $\{S^{(B)}\}$, onde cada par (y_n, x_n) será escolhido aleatoriamente a partir do conjunto inicial S [6].

A principal vantagem do método Bagging é o fato de não se precisar de um conjunto de dados grande para realizar um treinamento, uma vez que este algoritmo utiliza um conjunto *Bootstrap* formado por repetições da amostra original. Ou seja, o algoritmo seleciona aleatoriamente um conjunto das variáveis da amostra original e recombina estas variáveis em sub-árvores. Para teste, o algoritmo utiliza variáveis que não foram selecionadas para o conjunto Bootstrap (variáveis Out-Of-Bag ou OOB) para validar os resultados [6].

2.4 Random Forest

É um algoritmo de aprendizagem de máquina, supervisionado que baseia-se em árvores de decisão. É um dos algoritmos mais utilizados por sua simplicidade e pelo fato de poder tratar tanto de problemas de classificação quanto de problemas de regressão [13].

Breiman (2001) define Random Forest da seguinte maneira: Uma coleção de classificadores estruturados em árvores $\{h\,(X, \theta K), k = 1, \ldots\}$, onde $\{\theta K\}$ são vetores independentemente e identicamente distribuídos, e cada árvore vota pela classe mais popular em X [8]. Ou seja, é uma combinação de árvores preditoras, de tal forma que cada árvore depende dos valores de um vetor aleatório, amostrado de forma independente e com a mesma distribuição para todas as árvores na floresta [8].

O algoritmo Bagging, mencionado anteriormente, é utilizado em Random Forest para definir as variáveis de forma aleatória em um conjunto de treinamento. Porém, a abordagem utilizada é a partir da construção de uma grande coleção de árvores correlacionadas, no qual se considera a média dos resultados. [16]

No algoritmo de Random Forest, alguns atributos são fundamentais para a execução do código de maneira eficiente, tais como o número de árvores que se deseja possuir na floresta, que corresponde ao parâmetro *"ntree"*, o número de variáveis escolhidas para a divisão de cada nó *"mtry"* e a importância de uma variável em relação ao aumento da taxa de erro *"importance".*

De modo geral, o parâmetro *"ntree"* deverá ser definido de acordo com o problema a ser resolvido. Uma abordagem para selecionar o número de árvores correto é através do Out-Of-Bag (OOB) *error*. Este erro corresponde a média de erros das predições em um conjunto de treinamento. Trata-se de uma técnica de tentativa e erro, onde relacionamos o OOB *error* com o número de árvores utilizado, selecionando o número de árvores em que a floresta obteve o menor OOB *error*.

O parâmetro "*mtry*", corresponde ao o número de variáveis escolhidas para a divisão de cada nó e pode ser determinado da seguinte maneira:

- Para problemas de classificação *"mtry*", é determinado por $\sqrt{p}$, onde p é a quantidade de variáveis no modelo
- Para problemas de Regressão *"mtry"*, é determinado por $\frac{p}{3}$, onde p é a quantidade de variáveis no modelo

O parâmetro relacionado a importância das variáveis *"importance"* permite definir quais variáveis mais contribuem nos resultados obtidos, e o quanto a taxa de erro pode aumentar caso estas sejam removidas do grupo de dados. Esta informação é relevante, não somente para análise dos dados, mas, para problemas em que se deseja determinar a importância de uma variável em um modelo de previsão, por exemplo.

Com os conceitos aprendidos neste capítulo, tornou-se possível colocar em prática os objetivos deste estudo. Em seguida, serão apresentados os métodos utilizados nesse trabalho.

Capítulo 3

Metodologia

Neste capítulo será abordada a metodologia utilizada para a 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 a configuração do algoritmo Random Forest nas ferramentas XLSTAT e RStudio.

Figura 4. Localização geográfica das cidades em estudo

Fonte: Google Maps

3.1 Configuração da Base de Dados

Para a realização deste trabalho será utilizado o histórico de dados coletados dos níveis dos postos das cidades Propriá (SE), Penedo (AL) e Piaçabuçu (AL), disponíveis na ANA (Agência Nacional de Águas), Sistema Nacional de Gerenciamento de Recursos Hídricos (SINGREH), na Companhia Hidrelétrica do São Francisco (CHESF) e da maré para o terminal marítimo Inácio Barbosa no site da Marinha do Brasil [17], para que sirva como base na utilização da técnica Random Forest.

As informações coletadas correspondem a dados horários dos anos de 2011, 2013, 2015 e 2017, referentes a vazão liberada pela usina de Xingó, aos dados horários dos níveis registrados nos postos localizados nas cidades de Propriá, Penedo e Piaçabuçu, bem como os dados de maré do terminal Inácio Barbosa localizado em Sergipe. Mais detalhes sobre a base de dados são descritos a seguir:

- Base de dados 1: Esta base corresponde a dados obtidos no site da Marinha do Brasil [17] e correspondem ao nível da maré. Estes dados foram coletados para o Terminal Inácio Barbosa para ser representativo das variações na foz do Rio São Francisco. Estes dados são fornecidos para quatro valores diários. Desta forma se fez necessário a interpolação linear hora a hora. Neste trabalho se fez uma interpolação linear.
- Base de dados 2: Base cedida pela Chesf (Companhia Hidroelétrica do São Francisco) [9], contendo dados da vazão da usina de Xingó, localizada entre os estados de Alagoas e Sergipe. Os valores foram de defluência horária da referida usina.
- Base de dados 3: Base cedida pela Chesf e ANA através do Portal HidroWeb. [14] Corresponde a dados do rio São Francisco para cada posto utilizado coletados a cada hora.

Esses dados encontravam-se, inicialmente, separados por cidade e por cada ano. Ou seja, quatro planilhas para cada cidade (para os anos 2011, 2013, 2015 e 2017). Nunes (2018) realizou em seu trabalho a interpolação linear e defasagem dos dados de cada cidade e, em seguida, agrupou todas as informações de vazão, maré e cota de uma cidade em uma única planilha. Para este trabalho iremos utilizar estas planilhas resultantes. Sendo assim, os dados utilizados estão organizados da seguinte maneira:

1. Uma planilha para a cidade Piaçabuçu – dados horários de nível, maré e vazão para os anos de 2011, 2013, 2015 e 2017.
2. Uma planilha para a cidade Propriá – dados horários de maré e vazão para os anos de 2011, 2013, 2015 e 2017.
3. Uma planilha para a cidade Penedo - dados horários de maré e vazão para os anos de 2011, 2013, 2015 e 2017.

Cada planilha encontra-se organizada da seguinte maneira:

- 73 colunas correspondem a dados horários da Maré nos anos de 2011 a 2017
- 73 colunas correspondem a dados horários da Vazão entre os anos de 2011 a 2017
- 1 coluna corresponde ao nível no posto (cota arbitrária)

3.2 Configuração do Random Forest

Para executar o algoritmo Random Forest, iremos utilizar duas ferramentas: XLSTAT e RStudio. XLSTAT é uma ferramenta de análise e estatística de dados, que é executada a partir da ferramenta Excel [22]. XLSTAT possui uma série de funções de aprendizado de máquina, entre elas Random Forest. A segunda ferramenta utilizada foi o RStudio, que é uma ferramenta para desenvolvimento de software a partir da linguagem R [21].

Ao final, deseja-se confrontar os resultados, a fim de garantir mais confiança nos resultados obtidos. As duas ferramentas irão utilizar as seguintes configurações base, demostradas na tabela abaixo:

Tabela 1. Configurações utilizadas no Random Forest

Parâmetros	Valores
Número de variáveis de entrada	146
Número de variáveis de saída	1
Variáveis escolhidas aleatoriamente para a divisão de cada nó (mtry)	48
Training data	70%
Número de árvores geradas (ntree)	300
Calcular a importância das variáveis (Importance)	True

Desta forma, para cada cidade, vamos executar os seguintes passos:

1. Executar primeiramente o algoritmo Random Forest na ferramenta XLSTAT 10 vezes para obter o parâmetro *seed* que deverá ser utilizado nas execuções da ferramenta RStudio. Ou seja, para cada execução do XLSTAT, deveremos salvar o *seed* obtido para utilizá-los nas rodadas do RStudio;
2. Calcular o valor absoluto das importâncias por execução e a média das variáveis;
3. Calcular a média final das execuções por cidade obtidas a partir do passo 2;
4. Executar o algoritmo Random Forest na ferramenta RStudio utilizando o parâmetro *seed* obtido no passo 1;
   Calcular o valor absoluto das importâncias por execução e a média das variáveis a partir das execuções no RStudio;
5. Calcular a média final das execuções por cidade obtidas a partir do passo 5;
6. Gerar a tabela dos resultados obtidos em 3 e 6;
7. Gerar um gráfico da importância das variáveis obtidas no XLSTAT

A partir desta metodologia de execução apresentada, bem como das configurações realizadas nas ferramentas XLSTAT e RStudio para o algoritmo Random Forest, serão abordados a seguir os resultados obtidos e as considerações finais deste trabalho.

Capítulo 4

Resultados

Este capítulo aborda os resultados encontrados após aplicar o algoritmo Random Forest às bases de dados das cidades Propriá (SE), Penedo (AL) e Piaçabuçu (AL), utilizando as duas ferramentas XLSTAT e RStudio e as configurações citadas no capítulo 3. É objetivo deste capítulo detalhar os resultados obtidos visando identificar a importância da maré e vazão para os níveis e cota destas cidades.

Na seção 4.1 serão detalhados os resultados para a base de dados da cidade Propriá. A seção 4.2 irá detalhar os resultados obtidos para a base de dados da cidade Penedo e, por fim, a seção 4.3 irá detalhar os resultados obtidos para a base de dados da cidade Piaçabuçu.

4.1 Random Forest aplicado a base de dados da cidade Propriá

A importância das variáveis e a média predominante destas variáveis ao longo das execuções de Random Forest para a base de dados da cidade Propriá será descrita nas subseções a seguir.

4.1.1 XLSTAT

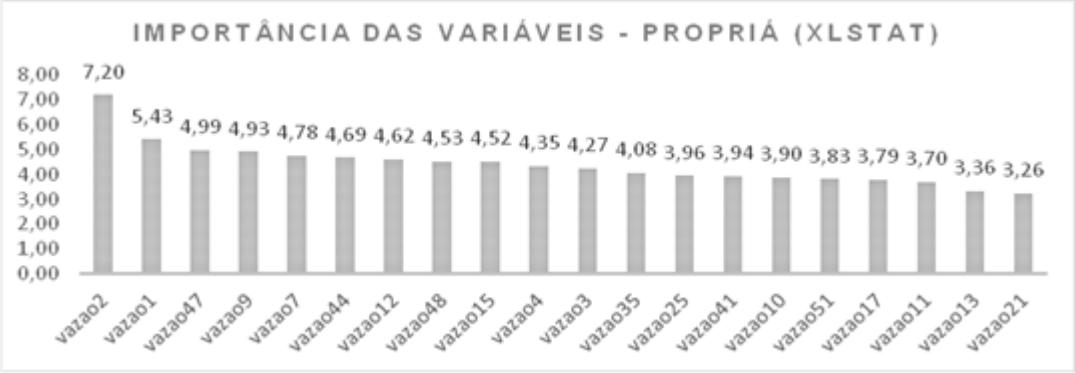
A tabela 2 agrupa os resultados dos percentuais de importância por execução da ferramenta XLSTAT aplicado a base de dados da cidade Propriá. A partir desta tabela é possível inferir que o percentual de importância da vazão foi bem superior ao da maré, correspondendo a 79,64%, enquanto a maré obteve 20,35%. Este resultado é coerente uma vez que esta é a cidade que se encontra mais distante do mar. Sendo assim, pode-se concluir que a vazão representa a variável mais importante para o posto desta cidade.

Tabela 2. Média do percentual de importância – Propriá (XLSTAT)

Propriá XLSTAT			
Rodadas	**Maré Imp. %**	**Vazão Imp. %**	**Total %**
1	18,38	81,61	100
2	21,41	78,58	100
3	20,96	79,03	100
4	18,29	81,70	100
5	20,92	79,07	100
6	20,59	79,40	100
7	20,69	79,30	100
8	19,52	80,47	100
9	20,99	79,00	100
10	21,74	78,25	100
Total	**203,54**	**796,45**	**1000**
Média	**20,35**	**79,64**	**100**

A figura 5 representa as 20 variáveis mais importantes levantadas pela ferramenta XLSTAT, onde todas as variáveis correspondem a vazão, sendo vazão 2 é variável que obteve o maior índice para o posto desta cidade. Isto significa dizer que a remoção desta variável da base de dados ocasionará um aumento no erro na previsão da cota.

Figura 5. Importância das variáveis – Propriá (XLSTAT)



4.1.2 RStudio

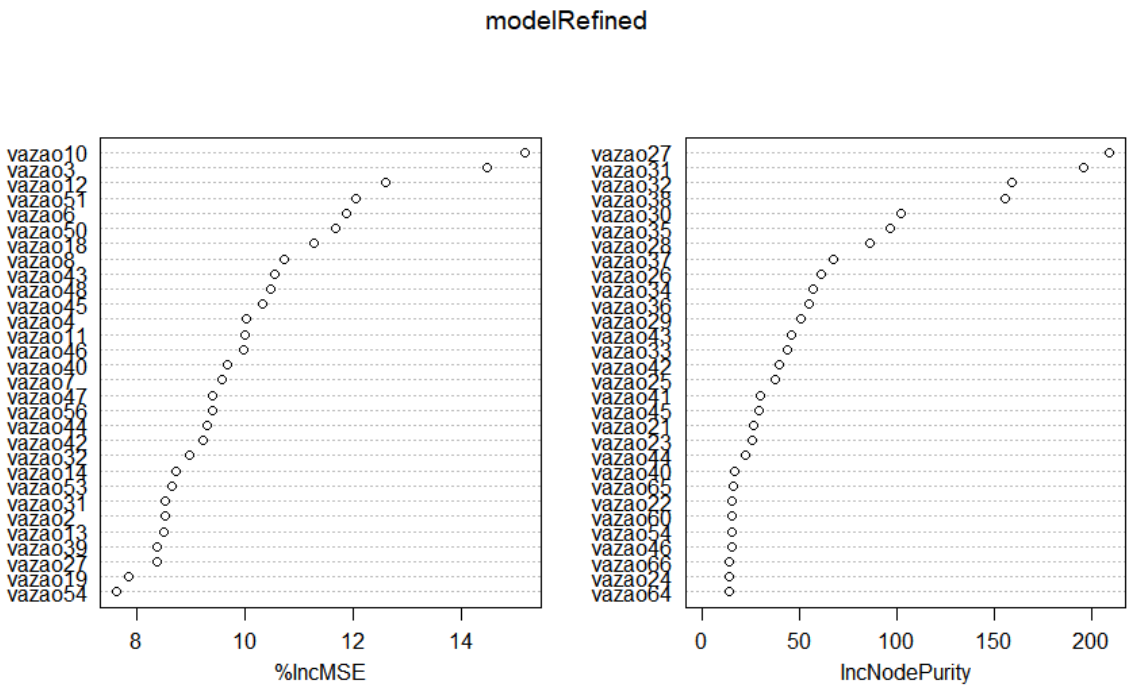
O resultado da ferramenta RStudio foi semelhante aos da ferramenta XLSTAT, apontando a variável vazão com o maior índice de importância ao longo das execuções. A tabela 3 agrupa os resultados dos percentuais de importância por execução da ferramenta RStudio para o posto da cidade Propriá. A partir desta tabela é possível inferir que o percentual de importância da vazão foi bem superior ao da maré, correspondendo a 72,06%, enquanto a maré obteve 27,94%. Este resultado é coerente uma vez que esta é a cidade que se encontra mais distante do mar.

Tabela 3. Média do percentual de importância – Propriá (RStudio)

Propriá RStudio			
Rodadas	**Maré Imp. %**	**Vazão Imp. %**	**Total %**
1	25,69	74,31	100
2	28,97	71,03	100
3	26,97	73,03	100
4	26,01	73,99	100
5	26,76	73,24	100
6	32,83	67,17	100
7	28,16	71,84	100
8	31,43	68,57	100
9	28,75	71,25	100
10	23,86	76,14	100
Total	**279,41**	**720,59**	**1000**
Média	**27,94**	**72,06**	**100**

A figura 6 representa os resultados obtidos para importância de variável na ferramenta RStudio passando como entrada a base de dados da cidade Propriá. O gráfico a esquerda corresponde ao percentual do aumento do EMQ da previsão da cota, caso uma determinada variável seja removida. Neste caso a variável vazão 10 obteve o maior índice. Isto significa dizer que a remoção desta variável da base de dados ocasionará um aumento no erro na previsão da cota. O gráfico a direita na figura 6 representa o valor do índice de Gini, descrito na equação 2.4, representando quais das variáveis geram nós mais puros em cada divisão.

Figura 6. Importância das variáveis – Propría (RStudio)



4.2 Random Forest aplicado a base de dados da cidade Penedo

A importância das variáveis e a média predominante destas variáveis ao longo das execuções na base de dados da cidade Penedo será descrita nas subseções a seguir.

4.2.1 XLSTAT

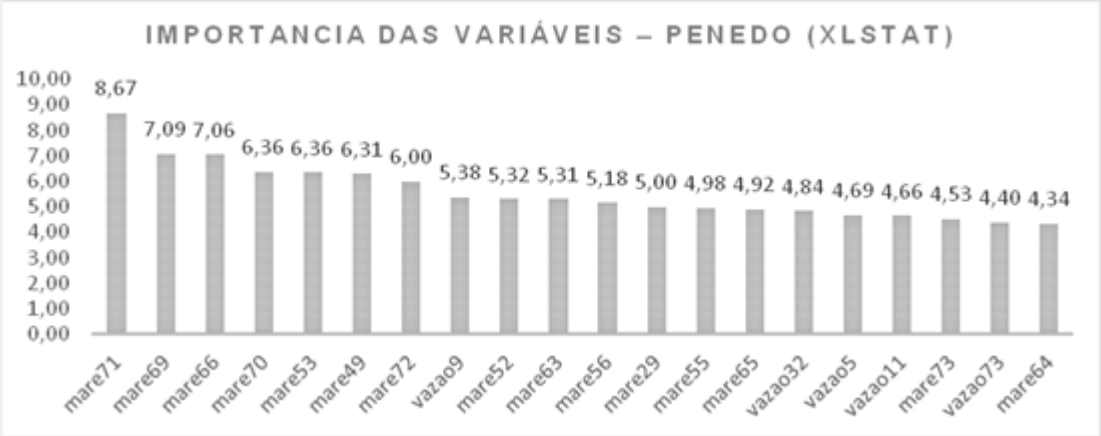
A tabela 4 agrupa os resultados dos percentuais de importância por execução da ferramenta XLSTAT para a base de dados da cidade Penedo. A partir desta tabela é possível inferir que o percentual de importância da maré foi 51,69%, enquanto a vazão obteve 48,30%. Devido a localização da cidade Penedo, que fica entre as cidades Propriá e Piaçabuçu, pode-se perceber que as variáveis maré e vazão influenciam o posto desta cidade, onde a variável vazão obteve um percentual ligeiramente maior.

Tabela 4. Média do percentual de importância – Penedo (XLSTAT)

	Penedo XLSTAT		
Rodadas	**Maré Imp. %**	**Vazão Imp. %**	**Total %**
1	51,43	48,56	100
2	53,04	46,95	100
3	53,49	46,50	100
4	55,73	44,26	100
5	47,66	52,33	100
6	51,26	48,73	100
7	54,33	45,66	100
8	52,23	47,76	100
9	47,30	52,69	100
10	50,41	49,58	100
Total	**516,93**	**483,06**	**1000**
Média	**51,69**	**48,30**	**100**

A figura 7 representa as 20 variáveis mais importantes levantadas pela ferramenta XLSTAT para a base de dados da cidade Penedo (percentual do aumento MEQ). Devido a sua localização, Penedo apresentou as variáveis maré e vazão em seu gráfico de importância. Porém, a variável maré se apresentou com mais predominância, sendo maré 71 a variável mais importante segundo o gráfico. Isto significa dizer que a remoção desta variável da base de dados ocasionará um aumento no erro na previsão da cota.

Figura 7. Importancia das variáveis – Penedo (XLSTAT)



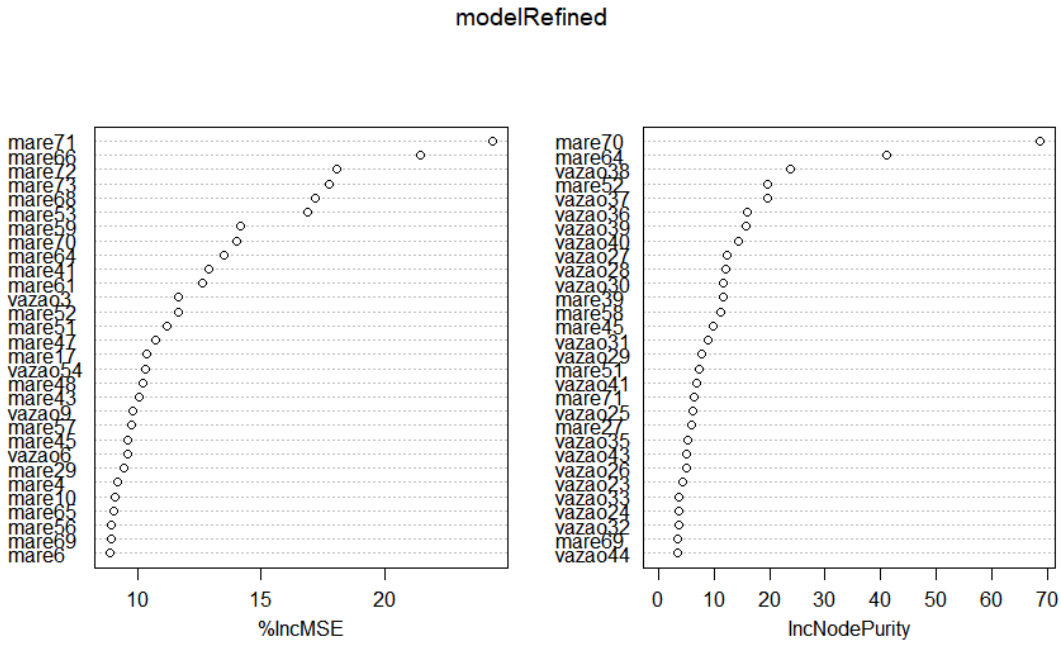
4.2.2 RStudio

O resultado da ferramenta RStudio foi semelhante ao da ferramenta XLSTAT. A varável maré obteve 53,29% enquanto a variável vazão obteve 46,71%. A tabela 5 agrupa os resultados dos percentuais de importância por execução da ferramenta RStudio para a base de dados da cidade Penedo. Devido a localização da cidade Penedo, que fica entre as cidades Propriá e Piaçabuçu, pode-se perceber que as variáveis maré e vazão influenciam o posto desta cidade, onde a variável maré obteve um percentual ligeiramente maior.

Tabela 5. Média do percentual de importância – Penedo (RStudio)

Penedo RStudio			
Rodadas	**Maré Imp. %**	**Vazão Imp. %**	**Total %**
1	47,99	52,01	100
2	52,58	47,42	100
3	51,27	48,73	100
4	54,67	45,33	100
5	50,19	49,81	100
6	49,75	50,25	100
7	57,35	42,65	100
8	62,52	37,48	100
9	56,05	43,95	100
10	50,48	49,52	100
Total	**532,87**	**467,13**	**1000**
Média	**53,29**	**46,71**	**100**

Figura 8. Importância das variáveis – Penedo (RStudio)



A figura 8 representa os resultados obtidos para importância de variável na ferramenta RStudio passando como entrada a base de dados da cidade Penedo. O gráfico a esquerda corresponde ao percentual do aumento do EMQ da previsão da cota, caso uma determinada variável seja removida. Neste caso a variável maré 71 obteve o maior índice. Isto significa dizer que a remoção desta variável da base de dados ocasionará um aumento no erro na previsão da cota. O gráfico a direita na figura 6 representa o valor do índice de Gini, descrito na equação 2.4, representando quais das variáveis gera nós mais puros em cada divisão.

4.3 Random Forest aplicado a base de dados da cidade Piaçabuçu

A importância das variáveis e a média predominante destas variáveis ao longo das execuções para a base de dados da cidade Piaçabuçu será descrita nas subseções a seguir.

4.3.1 XLSTAT

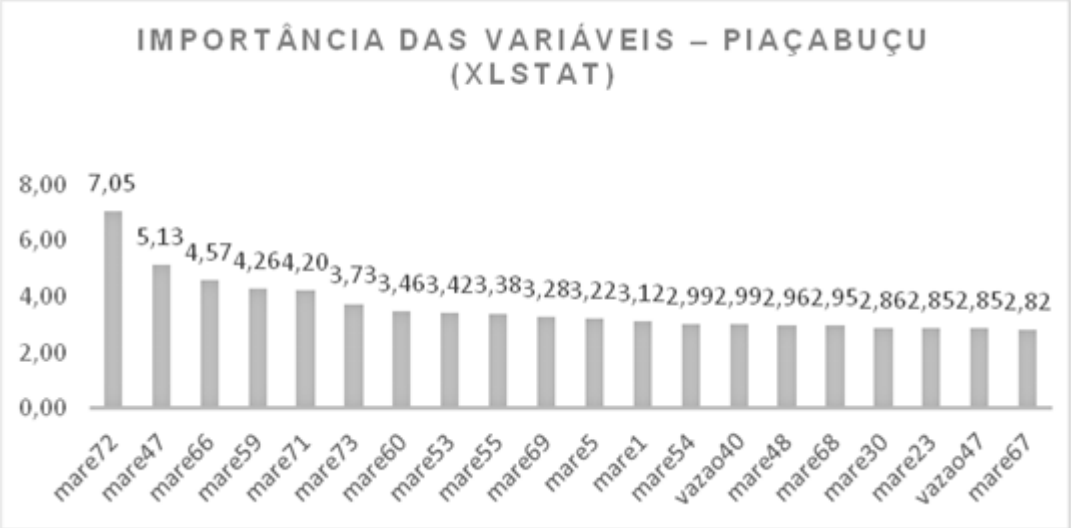
A tabela 6 agrupa os resultados dos percentuais de importância por execução da ferramenta XLSTAT para a base de dados da cidade Piaçabuçu. A partir desta tabela é possível inferir que o percentual de importância da maré foi superior ao da vazão, correspondendo a 56,84%, enquanto a vazão obteve 43,15%. Este resultado é coerente com a localização geográfica da cidade Piaçabuçu, que fica mais próxima do mar.

Tabela 6. Média do percentual de importância – Piaçabuçu (XLSTAT)

	Piaçabuçu XLSTAT		
Rodadas	**Maré Imp. %**	**Vazão Imp. %**	**Total %**
1	55,29	44,70	100
2	57,82	42,17	100
3	57,31	42,68	100
4	56,62	43,37	100
5	57,92	42,07	100
6	56,42	43,57	100
7	57,91	42,08	100
8	56,15	43,84	100
9	55,24	44,75	100
10	57,68	42,31	100
Total	**568,41**	**431,58**	**1000**
Média	**56,84**	**43,15**	**100**

A figura 9 representa as 20 variáveis mais importantes levantadas pela ferramenta XLSTAT para a base de dados da cidade Piaçabuçu, através do percentual do aumento MEQ. A variável maré 72 obteve o índice mais alto de importância. Isto significa dizer que a remoção desta variável da base de dados ocasionará um aumento no erro na previsão da cota.

Figura 9. Importância das variáveis – Piaçabuçu (XLSTAT)



4.3.2 RStudio

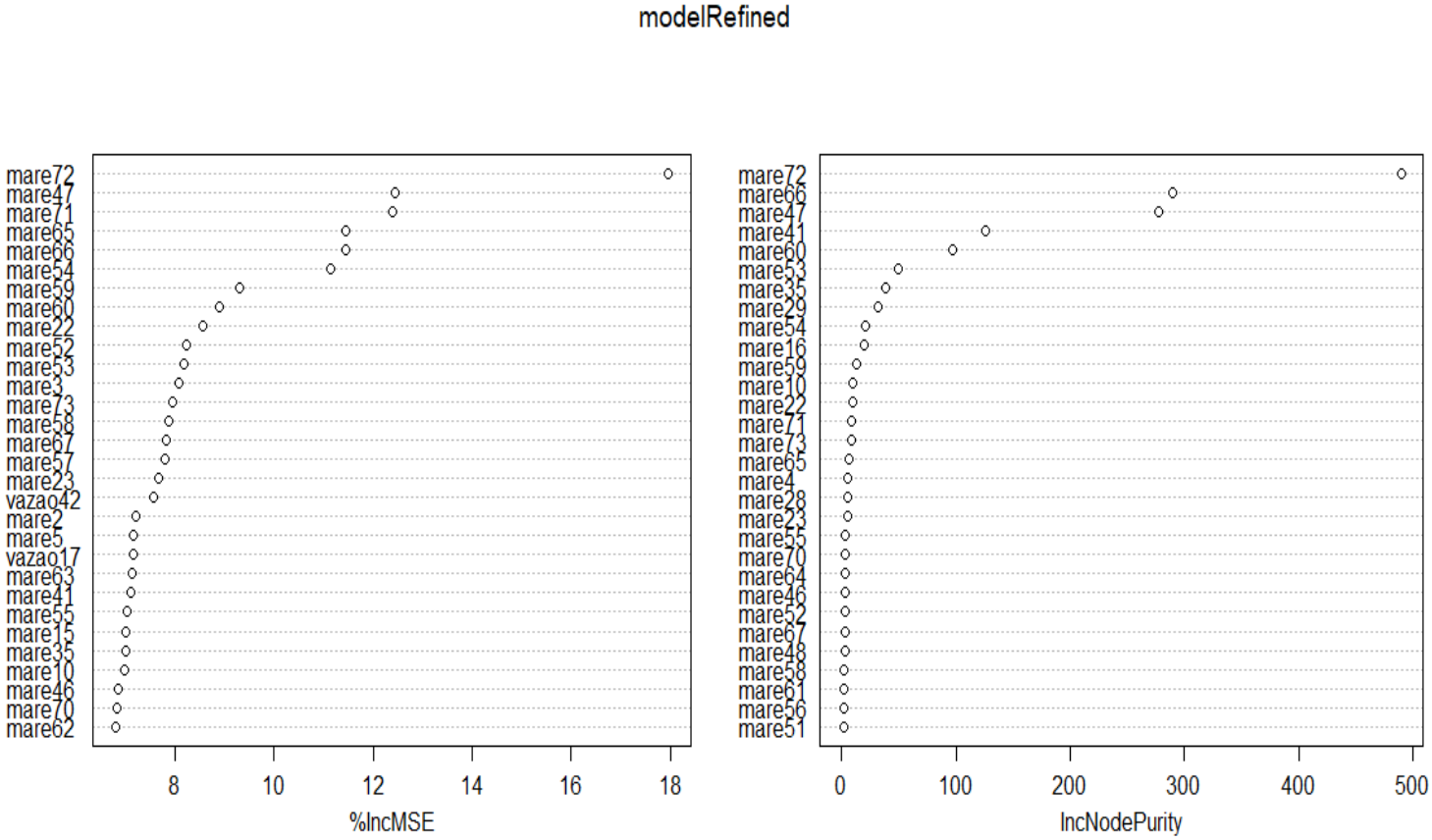
O resultado da ferramenta RStudio foi semelhante ao da ferramenta XLSTAT, apontando a variável maré com um percentual maior de importância ao longo das execuções. A tabela 7 agrupa os resultados dos percentuais de importância por execução da ferramenta RStudio para a cidade Piaçabuçu. A partir desta tabela é possível inferir que o percentual de importância da maré foi superior ao da vazão, correspondendo a 57,10%, enquanto a vazão obteve 42,90%.

Tabela 7. Média do percentual de importância – Piaçabuçu (RStudio)

Piaçabuçu RStudio			
Rodadas	**Maré Imp. %**	**Vazão Imp. %**	**Total %**
1	56,46	43,54	100
2	56,29	43,71	100
3	58,35	41,65	100
4	55,85	44,15	100
5	58,33	41,67	100
6	57,30	42,70	100
7	57,23	42,77	100
8	57,04	42,96	100
9	56,54	43,46	100
10	57,58	42,42	100
Total	**570,98**	**429,02**	**1000**
Média	**57,10**	**42,90**	**100**

A figura 10 representa os resultados obtidos para importância de variável na ferramenta RStudio passando como entrada a base de dados da cidade Piaçabuçu. O gráfico a esquerda corresponde ao percentual do aumento do EMQ da previsão da cota, caso uma determinada variável seja removida. Neste caso a variável maré 72 obteve o maior índice. Isto significa dizer que a remoção desta variável da base de dados ocasionará um aumento no erro na previsão da cota. O gráfico a direita na figura 10 representa o valor do índice de Gini, descrito na equação 2.4, representando quais das variáveis geram nós mais puros em cada divisão.

Figura 10. Importância das variáveis – Piaçabuçu (RStudio)



A seguir serão discutidas as conclusões e trabalhos futuros a partir dos resultados obtidos por este estudo.

Capítulo 5
Conclusões

Este capítulo tem como objetivo apresentar as conclusões obtidas neste trabalho. Será descrita as conclusões obtidas ao aplicar o Random Forest para medir a importância das variáveis da maré e vazão para a cota nos postos das cidades Piaçabuçu, Penedo e Propriá. A seção 5.1 irá detalhar os trabalhos futuros para esse estudo.

Este trabalho teve como objetivo investigar de que maneira as variáveis vazão e maré influenciavam os níveis no posto hidrométrico localizado na cidade Piaçabuçu. Para realizar este objetivo, o algoritmo de aprendizagem de máquina Random Forest foi utilizado, visto que este é um algoritmo bastante difundido e de fácil execução tanto para problemas de classificação quanto para regressão de variáveis [4]. O algoritmo Random Forest também foi utilizado nos postos de outras duas cidades próximas ao rio São Francisco Propriá (SE) e Penedo (AL), para fins de comparação.

Os resultados obtidos na cidade Propriá demonstraram que a vazão obteve um maior índice de influência nos níveis da cota do posto desta cidade ao longo do período em estudo, alcançando um percentual de 79,64% a partir da ferramenta XLSTAT e 72,05% na fermenta RStudio. Por outro lado, a maré alcançou 20,35% para a ferramenta XLSTAT enquanto na ferramenta RStudio alcançou o valor 27,94%. Este resultado se mostrou coerente com a localização geográfica da cidade, mais próxima do rio.

Após utilizar o algoritmo Random Forest passando a base de dados da cidade Penedo, obteve-se um percentual de 51,69% a partir da ferramenta XLSTAT e 53,28% na ferramenta RStudio para a variável maré. Por outro lado, a variável vazão alcançou 48,30% para a ferramenta XLSTAT enquanto na ferramenta RStudio alcançou o valor 46,71%. Ou seja, em ambas as ferramentas, a variável maré obteve um percentual ligeiramente maior.

Após utilizar o algoritmo Random Forest passando a base de dados da cidade Piaçabuçu, obteve-se um percentual de 56,84% a partir da ferramenta XLSTAT e 57,09% na ferramenta RStudio para a variável maré. Por outro lado, a vazão alcançou 43,15% para a ferramenta XLSTAT enquanto na ferramenta RStudio alcançou o valor 42,90%. Ou seja, em ambas as ferramentas, a variável maré obteve um percentual ligeiramente maior.

5.1 Trabalhos Futuros

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

- Adicionar os dados atualizados até 2019, 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 as variáveis mais significativas tendo em vista a forte correlação entre algumas das variáveis independentes.

- Utilizar redes neurais para previsão dos níveis na cidade de Piaçabuçu de forma a 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). Disponível em: <http://www3.ana.gov.br/portal/ANA/aguas-no-brasil/usos-da-agua/abasteciment> Acesso em: 21 de Abril de 2019

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

[3] ANICETO, Maísa. **Estudo Comparativo entre Técnicas de Aprendizado de Máquina para Estimação de Risco de Crédito**, Universidade de Brasília, Programa de Pós-Graduação em Administração, 2016.

[4] BARBOSA, Juliana Moreira; CARNEIRO, Tiago Garcia de Senna; TAVARES, Andrea Iabrudi, **Métodos de Classificação por Árvores de Decisão**, Universidade Federal de Ouro Preto.

[5] 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.

[6] BREIMAN, L., **Bagging predictors. Machine learning**, 1996.

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

[8] BREIMAN, L. **Random forests. Machine Learning**, 2001.

[9] “Companhia Hidrelétrica do São Francisco (Chesf) - Sistema Chesf Xingó”. Disponível em: <https://www.chesf.gov.br/SistemaChesf/Pages/SistemaGeracao/Xingo.aspx> Acesso em: 20 de Abril de 2019.

[10] DA SILVA, Luiza Maria Oliveira. **Uma aplicação de Árvores de Decisão, Redes Neurais e KNN para identificação de modelos Arma Não-Sazonais e Sazonais**. Tese (Doutorado). Pontífica Universidade Católica do Rio de Janeiro, 2005.

[11] DE CASTRO, Cristiano; BRAGA, Antonio; ANDRADE, Alessandro. **Aplicação de um Modelo Ensemble de Redes Neurais Artificiais para Previsão de Séries Temporais não Estacionárias**. Centro Universitário de Belo Horizonte-UNI-BH, 2019.

[12] Dietterich T.G. **Ensemble Methods in Machine Learning**. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg, 2000.

[13] DONGES, Niklas.The Random Forest Algorithm. Disponível em: <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd> Acesso em: 25 de Março de 2019

[14] “HIDROWEB – Sistema de Informações Hidrológicas”. Disponível em: <http://www.snirh.gov.br/hidroweb/publico/apresentacao.jsf > Acesso em: 25 de Abril de 2019

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

[16] LENTO, Gabriel. **Random forest em dados desbalanceados: uma aplicação na modelagem de churn em seguro saúde** -Fundação Getúlio Vargas, 2019.

[17] “Marinha do Brasil”. Disponível em: <https://www.marinha.mil.br/chm/tabuas-de-mare>, Acesso em: 15 de Abril de 2019

[18] NEVES, Filipe. **Predicting customer response to cross- market discounts using ensemble methods**. Faculdade de Engenharia da Universidade do Porto, 2015.

[19] NUNES, Gustavo. **Utilizando as Técnicas de Random Forest e Redes Neurais para Previsão de Níveis na Cidade de Piaçabuçu**. Universidade de Pernambuco, 2018

[20] ROKACK, L. **Ensemble Methods for Classifiers**. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA, 2005

[21] “The R Project for Statistical computing” [Online]. Disponível em: <https://www.r-project.org/ > Acesso em: 27 de Março de 2019

[22] “XLSTAT by Addinsoft” Online. Disponível em: <https://www.xlstat.com/en/> Acesso em 27 de Março de 2019