UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS QUIXADÁ

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

HINESSA DANTAS CAMINHA

ESTIMATIVA DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA UTILIZANDO MODELOS PREDITIVOS E SELEÇÃO DE ATRIBUTOS

QUIXADÁ

2017

HINESSA DANTAS CAMINHA

ESTIMATIVA DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA UTILIZANDO MODELOS PREDITIVOS E SELEÇÃO DE ATRIBUTOS

Monografia apresentada no curso de Sistemas de Informação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Sistemas de Informação. Área de concentração: Computação.

Orientadora: Profª Dra. Ticiana Linhares Coelho da Silva

Coorientadora: Profª Dra. Atslands Rego da Rocha

QUIXADÁ

2017

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C191e Caminha, Hinessa Dantas.
Estimativa da evapotranspiração de referência utilizando modelos preditivos e seleção de atributos / Hinessa Dantas Caminha. – 2017.
81 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Sistemas de Informação, Quixadá, 2017.
Orientação: Profa. Dra. Ticiana Linhares Coelho da Silva.
Coorientação: Profa. Dra. Atslands Rego da Rocha.

1. Agricultura-Irrigação. 2. Mineração de dados. 3. Evapotranspiração. 4. Regressão linear múltipla. I. Título.
CDD 005

HINESSA DANTAS CAMINHA

ESTIMATIVA DA EVAPOTRANSPIRAÇÃO DE REFERÊNCIA UTILIZANDO MODELOS PREDITIVOS E SELEÇÃO DE ATRIBUTOS

Monografia apresentada no curso de Sistemas de Informação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Sistemas de Informação. Área de concentração: Computação.

Aprovada em: __/__/__

BANCA EXAMINADORA

Profª Dra. Ticiana Linhares Coelho da Silva (Orientadora)
Universidade Federal do Ceará – UFC

Profª Dra. Atslands Rego da Rocha (Coorientadora)
Universidade Federal do Ceará - UFC

Profª Ma. Lívia Almada Cruz Rafael
Universidade Federal do Ceará - UFC

Prof. Me. Carlos Diego Andrade de Almeida
Universidade Federal do Ceará - UFC

AGRADECIMENTOS

À minha família, em especial aos meus pais, por compartilharem comigo esse sonho e por todo o suporte oferecido durante essa jornada. Nada disso seria possível sem vocês. Amo vocês com minha vida.

À minha orientadora, Profª Ticiana Linhares, que durante os últimos três anos me ofereceu não apenas orientação, mas confiança, companheirismo, inspiração e aprendizados pelos quais palavras serão poucas para agradecer.

À Profª Atslands Rocha, que se fez presente mesmo estando fisicamente distante, e acrescentou tanto à este trabalho.

Às minhas queridas Natália Lionel e Talhita Rabelo, pela amizade sincera e por todos os momentos únicos que vivemos juntas. Vocês foram minha segunda família em Quixadá e tornaram toda essa jornada mais prazerosa.

Aos amigos Rafael Costa, Rogerio Carvalho, Erick Bhrener, Sérgio Filho, Raquel Vitoriano, Juliana Sousa, Marcelo Gonçalves e Saiane Lins pelo companheirismo e apoio prestados durante a produção deste trabalho.

À toda comunidade acadêmica da UFC Quixadá que me proporcionou um ambiente de aprendizagem único, no qual pude obter lições que carregarei comigo durante toda a minha vida.

"Não sou nada.
Nunca serei nada.
Não posso querer ser nada.
À parte isso, tenho em mim todos os sonhos do mundo."

(Fernando Pessoa)

RESUMO

No Brasil, a agricultura irrigada é o setor responsável pela maior demanda de água consumida. Desse modo, é necessário o desenvolvimento de técnicas que permitam a utilização dessa água de forma sustentável. A evapotranspiração é a ocorrência simultânea dos processos de evaporação e transpiração em uma superfície vegetada, e designa a quantidade de água perdida por uma cultura. Através da estimativa dessa perda, pesquisadores e agricultores podem gerenciar de forma mais eficiente o consumo de água de seus cultivos. Este trabalho propôs a criação de modelos de predição para a evapotranspiração de referência, a partir de dados meteorológicos. A multidimensionalidade dos dados pode gerar modelos que necessitem de muitas variáveis, logo, também foi proposta uma solução que emprega a execução de técnicas de seleção de atributos antes da construção dos modelos. Ao final do estudo, foi possível concluir que, modelos de alta acurácia podem ser criados a partir do algoritmo *M5'* em conjunto com técnicas de seleção de atributos.

Palavras-chave: Agricultura-Irrigação. Mineração de Dados. Evapotranspiração. Regressão Linear Múltipla.

ABSTRACT

In Brazil, irrigated agriculture is the sector responsible for the greater water consumption demand. Therefore, it is necessary to develop techniques that allow the use of this water in a sustainable way. Evapotranspiration is the simultaneous occurrence of evaporation and transpiration processes on a vegetated surface, and designates the amount of water lost by a crop. By estimating this loss, researchers and farmers can efficiently manage the water consumption of their crops. This work proposes the creation of prediction models for reference evapotranspiration, based on climatic data. The multidimensionality of the data can generate models that need many climatic variables, so a solution that employs the execution of feature selection techniques before the construction of the models was also proposed. All in all, it was possible to conclude that models with high accuracy can be generated by using M5' trees and feature selection techniques.

Keywords: Agriculture-Irrigation. Data Mining. Evapotranspiration. Multiple Linear Regression.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

O esforço para reduzir de forma significativa a insegurança alimentar e a pobreza se mantém no topo das prioridades humanas. A agricultura irrigada exerce um papel importante para que esse objetivo seja alcançado, assegurando abordagens inovadoras que levam a uma maior produtividade (FAO, 2015). Segundo Agência Nacional de Águas (2015), a irrigação é responsável por 72% da demanda de água consumida no Brasil. À medida que outros setores como abastecimento de água, indústria, manufatura e o próprio meio ambiente se expandem, a concorrência por essa demanda aumenta. Assim, cabe ao setor de agricultura revisar e ajustar seus métodos de acordo com a quantidade de água disponível para utilização (GARCES-RESTREPO; VERMILLION; MUOZ, 2007).

O manejo da irrigação tem por finalidade estabelecer técnicas que possibilitem aumentar a conservação de água e energia sem reduzir a produção econômica da cultura, a partir do conhecimento das necessidades de água do cultivo, além das características do solo e da sua capacidade de armazenar água. Dentre as diversas técnicas existentes, pode-se citar o monitoramento via clima, que consiste na utilização de dados climáticos para estimativa do consumo de água de uma cultura. A estimativa é determinada pela evapotranspiração, conceito este que designa a ocorrência simultânea dos processos de evaporação e transpiração de uma superfície vegetada. O cálculo das necessidades de água das culturas é feito a partir da evapotranspiração de referência e do coeficiente de cultura (FRIZZONE; SOUZA; LIMA, 2013).

Métodos indiretos de monitoramento podem ser utilizados para a realização da estimativa da evapotranspiração de referência. Enquadram-se, nessa categoria, as estações meteorológicas automáticas, compostas por uma série de sensores de parâmetros meteorológicos (FRIZZONE; SOUZA; LIMA, 2013). A partir desses sensores, dados climáticos são coletados com uma determinada frequência e são usados para a realização das estimativas.

Esses dados também podem ser analisados por meio da mineração de dados, que consiste na aplicação de técnicas e algoritmos que reconhecem padrões e modelos sobre os dados, de modo que esses gerem algum conhecimento. Dentre as várias técnicas existentes, está a regressão linear, que possui a finalidade de mapear os dados em função de uma variável de predição real (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). O pré-processamento é a etapa que precede a mineração e possui grande importância para que os resultados obtidos pelas análises sejam realmente úteis e precisos. A seleção de atributos é uma das técnicas utilizadas mais importantes, que visa a redução do número de atributos, de modo que sejam retirados dados

irrelevantes, redundantes ou que não possuam valor significativo para os resultados (LIU; YU, 2005).

Em Xavier, Tanaka e Revoredo (2015b) foi realizado um estudo que possuía por objetivo responder ao questionamento: é possível estimar a evapotranspiração independentemente da disponibilidade de todas as variáveis? Para solucionar esse problema, os autores utilizaram conjuntos de dados com séries históricas, gerados por estações meteorológicas no estado do Rio de Janeiro. Foram criados modelos de predição a partir do algoritmo $M5'$, que utiliza árvores de decisão para criar equações lineares. Para cada conjunto de dados, foram obtidas algumas equações para calcular a evapotranspiração. Ao final do estudo, os autores notaram que nenhum dos modelos utilizava todos os atributos dos conjuntos para gerar as equações.

Os resultados gerados pela metodologia de Xavier, Tanaka e Revoredo (2015b) estimaram valores referentes especificamente às culturas presentes nos ambientes onde os dados climáticos foram coletados. Partindo desse cenário, este trabalho propôs a criação de modelos preditivos para o cálculo da evapotranspiração de referência, utilizando os dados produzidos pela estação meteorológica instalada na UFC Quixadá. Foram criados modelos com e sem seleção de atributos, através da aplicação da regressão linear e do $M5'$. Ao final, foi realizada uma comparação entre as duas abordagens para constatar qual delas resultou em modelos de melhor acurácia. A principal contribuição deste trabalho, deu origem ao artigo de Caminha et al. (2017), aceito na *19th International Conference on Enterprise Information Systems* (ICEIS 2017).

No Capítulo 2 serão abordados os principais conceitos envolvidos neste trabalho: Evapotranspiração (seção 2.1) e Evapotranspiração de Referência (subseção 2.1.1), Métodos de Seleção de Atributos (seção 2.2), Regressão Linear (seção 2.3) e $M5'$ (seção 2.4). Os trabalhos relacionados serão apresentados no Capítulo 3 e o Capítulo 4 mostrará os objetivos deste trabalho. No Capítulo 5, serão abordados os procedimentos metodológicos e no Capítulo 6 serão descritos os experimentos e seus resultados. Por fim, a conclusão e os trabalhos futuros serão apresentados no Capítulo 7.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, serão apresentados os principais conceitos presentes neste trabalho. O capítulo está organizado dessa maneira: na Seção 2.1 será mostrada a definição de evapotranspiração, seguida pela Subseção 2.1.1, onde será abordada a evapotranspiração de referência. As Seções 2.2, 2.3 e 2.4 irão descrever os métodos de seleção de atributos, a regressão linear e o $M5'$, respectivamente.

2.1 Evapotranspiração

O ciclo hidrológico é composto por diversos componentes que são fortemente impulsionados pelas condições climáticas. Dentre os mais importantes processos desse ciclo, pode-se destacar a evaporação e a transpiração (TUNDISI, 2003). Em ambos os processos, a água sofre uma transformação física e passa do estado líquido para o vapor. A evaporação ocorre em diversas superfícies como lagos, rios, solos e vegetações úmidas, enquanto a transpiração ocorre especificamente nas superfícies das plantas (ALLEN et al., 1998).

De acordo com Frizzone, Souza e Lima (2013), o termo evapotranspiração (ET) designa a ocorrência simultânea da evaporação e transpiração em uma superfície vegetada. Sua taxa é normalmente descrita em milímetros (mm) por uma determinada unidade de tempo e expressa a quantidade de água perdida da cultura (ALLEN et al., 1998).

A evapotranspiração pode ser estimada por medidas diretas ou indiretas. No primeiro grupo se enquadram os diferentes tipos de lisímetros e o método do balanço de água no solo. No segundo grupo, encontram-se os modelos micrometeorológicos teóricos e empíricos, baseados na utilização de dados climáticos. Os métodos diretos oferecem melhores estimativas, porém sua realização é dispendiosa e demorada, difíceis de serem realizadas no campo. Logo, esse métodos são geralmente utilizados por pesquisadores para calibrar os métodos indiretos (FRIZZONE; SOUZA; LIMA, 2013). Um dos mecanismos utilizados nos métodos indiretos são as estações meteorológicas automáticas, compostas por diversos sensores que coletam dados climáticos. A partir desses dados, a estimativa da evapotranspiração é realizada.

2.1.1 Evapotranspiração de Referência

A taxa de evapotranspiração de uma superfície de referência é denotada pelo termo evapotranspiração de referência (ET_0), a qual utiliza a superfície de cultura da grama como padrão.

Os únicos fatores que influenciam na ET_0 são de parâmetros climáticos e, consequentemente, seus valores podem ser computados a partir de dados meteorológicos (ALLEN et al., 1998).

A determinação da ET_0 é uma das principais atividades para entender o consumo de água de uma certa cultura. Esse consumo é designado pelo coeficiente ET_c, obtido a partir da Equação 2.1:

$$ET_c = ET_0 \times K_c \tag{2.1}$$

Onde ET_c corresponde à quantidade de água usada pela cultura, ET_0 a evapotranspiração de referência e K_c o coeficiente da cultura.

Diversos métodos para calcular a ET_0 são apresentados na literatura e cada um possui suas particularidades como, por exemplo, a utilização somente de variáveis climáticas que tenham maior influência em locais com um determinado clima. Embora haja essa diversidade, o método recomendado pela *Food and Agriculture Organization of the United Nations* (FAO), é a equação de *Penman-Monteith*. Entretanto, sua utilização é complexa, dispendiosa, e intolerante à indisponibilidade de alguma das variáveis climáticas requeridas.

Neste trabalho, foram criados modelos de predição da ET_0 alternativos à equação de *Penman-Monteith*, que utilizassem as variáveis disponibilizadas pela estação meteorológica do campus.

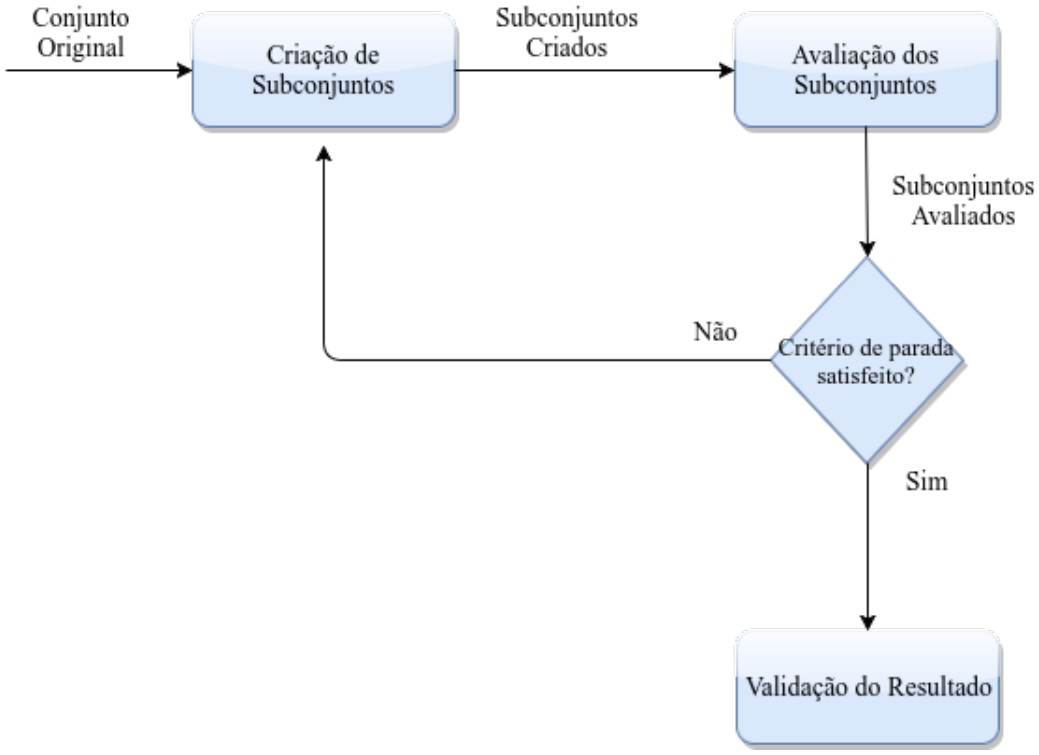
2.2 Métodos de Seleção de Atributos

O pré-processamento dos dados é uma etapa de grande importância no processo de extração de conhecimento. Suas técnicas incluem limpeza, integração, transformação e redução dos dados. Quando aplicadas antes da mineração de dados, essas técnicas podem aumentar substancialmente a qualidade dos resultados ou até mesmo diminuir o tempo gasto durante o processo de mineração (KAREGOWDA; MANJUNATH; JAYARAM, 2010). Dentre as técnicas de redução de dados, encontram-se os métodos de seleção de atributos.

A seleção de atributos é o processo pelo qual um subconjunto de atributos é selecionado a partir do seu conjunto de atributos original. Durante esse processo, dados irrelevantes e redundantes são removidos, de modo que, ao final, resultem apenas os atributos que melhor representem a base de dados (LIU; YU, 2005).

Esse processo possui várias etapas, conforme são apresentadas na Figura 1. Primeiramente, um conjunto de dados é recebido como entrada. A partir dele, subconjuntos com atributos candidatos são criados utilizando um algoritmo de busca. Após essa fase, os subconjuntos gerados são avaliados mediante um critério de parada. Caso a avaliação selecione o melhor subconjunto, o processo é encerrado e o resultado é validado; caso contrário, os passos anteriores são executados novamente até que o melhor subconjunto seja gerado e avaliado.

Figura 1 – Passos da Seleção de Atributos



Fonte – Adaptado de Liu e Yu (2005).

Existem duas abordagens para a seleção de atributos: a *wrapper* e a *filter* (KAREGOWDA; MANJUNATH; JAYARAM, 2010). A abordagem *wrapper* utiliza o próprio algoritmo de mineração de dados para avaliar a importância do conjunto de atributos. Sua principal vantagem é a capacidade de gerar um conjunto ótimo para o algoritmo utilizado. Entretanto, sua utilização é de grande custo computacional devido à sua complexidade. Diferente da anterior, a abordagem *filter* é executada de forma independente e ocorre antes da aplicação do algoritmo de mineração de dados. Isso permite que o conjunto de atributos selecionados por ela seja dado como entrada no algoritmo de mineração.

Em Hall (2000) foi proposto o algoritmo *Correlation-based Feature Selection* (CFS), que enquadra-se na abordagem *filter* e emprega uma heurística para avaliar a importância dos

atributos. Essa heurística considera a capacidade de predição do valor de classe que os atributos possuem baseando-se na correlação interna entre eles. A hipótese básica da heurística diz que um bom conjunto de atributos possui seus membros altamente correlacionados com o atributo de classe, mas pouco correlacionados entre si.

O CFS é um avaliador, portanto necessita ser executado juntamente com um algoritmo de busca, responsável por criar os subconjuntos de atributos a serem avaliados. Dentre os vários algoritmos de busca, encontram-se o *BestFirst*, o *ExhaustiveSearch* o *GeneticSearch* e o *RandomSearch*.

O *BestFirst* emprega a heurística de sempre explorar primeiro o subconjunto com melhor escore. Esse escore é definido a partir de uma função de avaliação, utilizada em todos os subconjuntos. O algoritmo mantém duas listas nas quais, uma mantém os subconjuntos que ainda não foram explorados (*open*) e a outra, mantém os subconjuntos já explorados (*closed*). Ambas as listas são ordenadas de acordo com os escores. No Algoritmo 1[1] é apresentado o passo-a-passo da busca em pseudocódigo.

Algoritmo 1: *BestFirst Search*

```
Data: List open, closed
closed ← {};
open ← {start};
score ← {};
score[start] ← evaluationFunction(start) ;
while open ≠ empty do
    current ← minimum(open) ;
    if current = goal then
        return path(current);
    end
    closed ← add(current);
    open ← remove(current);
    neighbors ← neighborhood(current);
    for neighbor to neighbors do
        if neighbor ⊄ closed then
            score[newNode] ← evaluationFunction(neighbor);
            open ← add(neighbor);
        end
    end
end
```

O algoritmo *ExhaustiveSearch* testa sequencialmente cada possibilidade a fim de determinar a solução esperada. A busca pode iniciar com um conjunto vazio ou não vazio. Caso seja iniciado com um conjunto vazio, o melhor subconjunto é retornado como solução. Se

[1] Adaptado de: <http://wiki.roblox.com/index.php?title=Best-first_search>.

iniciado com um conjunto não vazio, o subconjunto com avaliação melhor ou igual ao conjunto inicial é retornado. Um pseudocógido generalizado para o *ExhaustiveSearch* é apresentado no trecho de código[2] abaixo:

```
backtrack(int solution, int depth)
{
   if (isSolution(solution))
      return solution
   else{
      newSolution = generatesolution()
      backtrack(newSolution, depth+1)
   }
}
```

Goldberg e Holland (1988) definiram o *GeneticSearch* como um método probabilístico projetado para grandes espaços de busca envolvendo estados que podem ser representados por *strings*. Esse método é paralelo e utiliza amostras do espaço (conjuntos de *strings*) para gerar um novo conjunto de amostras. Também são analisados conjuntos de *substrings* com a finalidade de produzir regras de criação para as populações futuras.

Após a criação do espaço de busca, o algoritmo desenvolve-se fazendo uso de três operações: a seleção, o *crossover* e a mutação. A seleção é responsável por manter a sobrevivência da amostra que melhor se encaixa ao objetivo da busca. O *crossover* permite que os membros das amostras (*strings*) sejam recombinados e, a mutação, introduz modificações de modo a aleatorizar as sequências de *bits* repetidas, evitando uma convergência prematura. O Algoritmo 2 [3] contém o pseudocódigo em alto nível do *GeneticSearch*.

Algoritmo 2: *GeneticSearch Algorithm*

produce an initial population of individuals;
evaluate the fitness of all individuals;
while *termination condition not met* **do**
    select filter individuals for reproduction;
    recombine between individuals;
    mutate individuals;
    evaluate the fitness of the modified individuals;
    generate a new population;
end

O termo *RandomSearch* faz referência a um algoritmo que utiliza algum tipo de aleatoriedade ou probabilidade. De modo geral, o *RandomSearch* prioriza encontrar uma boa

[2] Adaptado de: <http://www.algorithmist.com/index.php/Exhaustive_Search>.
[3] Adaptado de: <https://www.slideshare.net/kancho/genetic-algorithm-by-example>.

solução de forma rápida, cujo resultado converge em probabilidade, a encontrar uma solução ótima. Tomando X como um vetor de n variáveis de decisão e S um espaço de solução, Zabinsky (2009) definiu genericamente o *RandomSearch* de acordo com os passos abaixo:

- **Passo 0:** Inicialize os parâmetros do algoritmo Θ_0 , os pontos iniciais $X_0 \subset S$ e o índice de iteração $K = 0$.
- **Passo 1:** Gere uma coleção de pontos candidatos $V_{k+1} \subset S$ de acordo com um gerador específico e seu respectivo distribuidor de amostras.
- **Passo 2:** Atualize X_{k+1} baseando-se nos pontos candidatos V_{k+1}, o índice de iteração e os parâmetros do algoritmo anteriores. Também atualize os parâmetros Θ_{k+1}.
- **Passo 3:** Se o critério de parada for satisfeito, encerre o algoritmo. Caso contrário, incremente K e retorne ao Passo 1.

Este trabalho utilizou o CFS juntamente com os algoritmos de busca descritos acima, para selecionar o conjunto de atributos que gerou modelos de predição da ET_0 com alta acurácia.

2.3 Regressão Linear

A análise de regressão é uma técnica estatística que possui a finalidade de investigar e modelar a relação entre variáveis por meio de uma equação matemática (MONTGOMERY; PECK; VINING, 2015). Essa técnica é amplamente utilizada para extração de conhecimento em bases de dados com variáveis contínuas.

Segundo Montgomery, Peck e Vining (2015) um modelo de regressão linear múltipla pode ser representado pela Equação (2.2):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \varepsilon \tag{2.2}$$

Onde y corresponde à variável a ser prevista e $\beta_j, \forall j, j \in \{0, 1, ..., k\}$ os coeficientes de regressão, que representam a variação do valor-resposta de y por cada unidade de x_j. O ε corresponde ao erro estatístico para mapear os dados de acordo com o modelo.

A regressão linear múltipla é implementada no algoritmo *LinearRegression*, na ferramenta *Waikato Enviroment for Knowledge Analysis* (WEKA) (HALL et al., 2009), e foi utilizada neste trabalho para criar o modelo preditivo da ET_0, o qual utilizou como entrada os dados de todos os atributos do *dataset* e o conjunto de atributos selecionados pelos métodos abordados na Seção anterior.

2.4 M5'

Uma das abordagens para criação de modelos preditivos consiste na aplicação de árvores de decisão. As árvores de decisão tradicionais, e suas regras de aprendizagem, foram desenvolvidas para utilizarem atributos cujos valores são discretos (WANG; WITTEN, 1996). Contudo, em dados do mundo real, valores contínuos são muito comuns e fizeram-se necessárias melhorias nessa abordagem.

Partindo desta necessidade, Quinlan (1992) propôs o algoritmo $M5$, que possui como principal característica a utilização de árvore de decisão na qual, ao contrário da tradicional, oferece a possibilidade de uso de modelos de regressão linear multivariados em suas folhas.

Posteriormente, Wang e Witten (1996) perceberam que o $M5$ não possuía tratamento de atributos enumerados ou com valores ausentes e propuseram o algoritmo $M5'$ que implementava essas melhorias. No $M5'$, a solução para os atributos com valores ausentes é feita a partir do processo de particionamento dos mesmos, que é um método utilizado por diversos algoritmos de seleção de atributos. Devido essa característica, é possível que os modelos lineares gerados nas folhas da árvore não precisem utilizar todos os atributos dados como entrada.

O $M5P$ é a implementação do $M5'$ na ferramenta WEKA (HALL et al., 2009) e foi utilizado neste trabalho para gerar modelos de predição da ET_0.

3 TRABALHOS RELACIONADOS

O uso de modelos preditivos para auxiliar na estimativa da quantidade de água necessitada por uma cultura já foi objeto de estudo de diversos autores e abordados de diferentes formas.

No trabalho de Xavier, Tanaka e Revoredo (2015a) os autores possuíam o objetivo de encontrar padrões regionais na estimativa da evapotranspiração potencial. Inicialmente, foram obtidas séries históricas de sete cidades a partir das bases de dados geradas por estações meteorológicas e disponibilizadas pelo Instituto Nacional de Meteorologia (INMET). As cidades que possuíam latitudes próximas eram consideradas similares quanto aos seus aspectos climáticos. Baseando-se nisso, foi gerado um modelo de predição para estimar a evapotranspiração sobre os dados de Resende - RJ utilizando regressão linear. Esse mesmo modelo foi reutilizado sobre as bases de dados das outras cidades (Cordeiro, Itaperuna, São Carlos, Cruzeiro do Sul, Maceió e Manaus) para descobrir qual a precisão dele ao gerar as estimativas de locais diferentes. Os resultados gerados foram comparados com as séries históricas das respectivas cidades. Ao final, o modelo obteve boa acurácia com os dados históricos nas três cidades que foram consideradas similares, porém eram falhos nas cidades não similares. Este trabalho assemelha-se com Xavier, Tanaka e Revoredo (2015a) na utilização de dados meteorológicos para criar um modelo preditivo a partir da técnica de regressão linear. Entretanto, diferencia-se na base de dados que será utilizada para criar os modelos e no emprego de algoritmos de seleção de atributos. Além disso, este trabalho buscou estimar o valor da evapotranspiração de referência, enquanto Xavier, Tanaka e Revoredo (2015a) estimou a evapotranspiração potencial.

Hendrawan e Murase (2011) pretendiam desenvolver um sistema de irrigação precisa baseado em *machine vision*. Nesse sistema, era necessário que houvesse um modelo preditivo capaz de estimar a quantidade de água necessária para a planta a partir de sua respectiva textura. Para tal, foi realizada a observação do estresse causado pela água em uma plantação de *sunagoke moss*.[1] Tal observação foi executada a partir de fotos, onde essas passavam por um processo de conversão em diversas escalas de cores (*Grey*, HSV, HSL e L*a*b*). A partir dessas conversões uma *Colour Co-occurrence Matrix* (CCM) foi gerada para cada escala de cor, com a finalidade de descobrir as texturas das plantas. Tais texturas são representadas por escalas de cores e indicam o estado da planta, causado pelo excesso ou pela falta de água. Logo após esse processo, foram aplicados quatro algoritmos de seleção de atributos para a detecção das

[1] Um tipo de planta sem flores que cresce em aglomerados e ambientes úmidos.

melhores texturas a serem dadas de entrada para a criação do modelo preditivo. Os algoritmos escolhidos foram *Neural-Intelligent Water Drops* (N-IWD), *Neural-Simulated Annealing* (N-SA), *Neural-Genetic Algorithm* (NGA) e *Neural-Discrete Particle Swarm Optimization* (N-PDSO). As texturas selecionadas na etapa anterior foram passadas como entrada para o classificador *Back-Propagation Neural Network* (BPNN) e um modelo, para cada conjunto de texturas selecionados, foi gerado. Após os experimentos, foi concluído que as texturas escolhidas pelos algoritmos de seleção de atributos geraram modelos preditivos com maior acurácia.

Assim como Hendrawan e Murase (2011), este trabalho se propôs a aplicar algoritmos de seleção de atributos e gerar modelos de predição a partir do conjunto retornado por esses algoritmos. Entretanto, este trabalho utilizou essa abordagem para estimar a evapotranspiração de referência a partir de dados com valores numéricos.

Em um outro estudo, Xavier, Tanaka e Revoredo (2015b) obtiveram as bases de dados meteorológicos do INMET, as quais correspondiam às seis estações meteorológicas presentes em cidades do Rio de Janeiro. O objetivo dos autores era criar um modelo de predição da evapotranspiração potencial para cada base, o qual conseguisse gerar bons resultados mesmo que houvesse a indisponibilidade de alguns atributos. Dentre os diversos algoritmos de classificação para dados numéricos presentes na ferramenta WEKA, o que obteve melhor acurácia nos testes foi o $M5'$, escolhido para realização dos experimentos. Todas as bases possuíam pelo menos um atributo com valores ausentes. Os modelos geraram equações para calcular a evapotranspiração potencial e, em algumas bases, o atributo "temperatura compensada média"era utilizado como decisor para qual equação ser empregada. Os resultados gerados pelo modelo obtiveram alto grau de correlação com as séries históricas disponibilizadas pelo INMET. É importante ressaltar que nenhuma das equações criadas pelo $M5'$ utilizaram todos os atributos das bases de dados.

Este trabalho possui semelhança com o de Xavier, Tanaka e Revoredo (2015b) no fato de que foi utilizado o $M5'$ para a criação de um modelo preditivo a partir de dados meteorológicos. Porém, distingue-se na evapotranspiração estimada e na abordagem, na qual o $M5'$ não foi o único algoritmo aplicado para gerar os modelos.

Sawalkar e Dixit (2015), obtiveram bases de dados climáticos coletados em diferentes regiões da Florida, disponibilizadas pela *United States Geological Survey* (USGS). As bases correspondiam ao monitoramento de três estações meteorológicas instaladas em lugares distintos, durante os anos de 2000 a 2004. Os autores estimaram, para cada ano, as taxas de evapotranspiração de referência diárias e mensais utilizando o método de *Penman-Monteith* e o

algoritmo *M5*. Ao final, foram realizadas comparações com a finalidade de descobrir qual das duas abordagens resultava em estimativas com maior acurácia. Ao final do estudo concluiu-se que, o *M5* criou modelos de predição da ET_0 com melhor acurácia do que o método de *Penman-Monteith*.

O trabalho de Sawalkar e Dixit (2015) assemelha-se a este na criação de modelos de predição da ET_0, na utilização de mais de uma abordagem para realizar as estimativas e na comparação de acurácia das abordagens adotadas. Entretanto, diferencia-se deste trabalho por utilizar o método de *Penman-Monteith* como uma das técnicas para calcular a ET_0.

No estudo de Rahimikhoob (2014), foram coletadas bases de dados climáticos da província Sistão-Baluchistão, no Irã. As bases continham dados mensais do período de 1998 a 2007, provenientes 4 estações meteorológicas diferentes em regiões de clima árido. O objetivo do autor consistiu na criação de modelos de predição da ET_0 utilizando *Artificial Neural Network* (ANN) e árvores *M*5, a partir dos dados citados acima. Os modelos foram comparados e avaliados utilizando as métricas coeficiente de correlação, erro médio e erro médio dos quadrados. Concluiu-se que, ambos os modelos obtiveram alta acurácia, com coeficientes de correlação muito próximos. Entretanto, o modelo criado pela ANN, obteve erros menores do que o modelo criado pelo *M*5.

O trabalho proposto possui semelhanças com o de Rahimikhoob (2014) na utilização de duas abordagens diferentes para a criação de modelos preditivos da ET_0. Entretanto, diferencia-se no emprego de técnicas de seleção de atributos.

A Figura 2 apresenta as similaridades e as dissimilaridades dos trabalhos citados neste capítulo com o trabalho proposto, em relação às suas principais características.

Figura 2 – Comparação entre os trabalhos relacionados e o trabalho proposto

	Xavier, Tanaka e Revoredo (2015a)	**Hendrawan e Murase (2011)**	**Xavier, Tanaka e Revoredo (2015b)**	**Sawalkar e Dixit (2015)**	**Rahimikhoob (2014)**	**Trabalho Proposto**
Dados Meteorológicos	Sim	Não	Sim	Sim	Sim	Sim
Seleção de Atributos	Não	Sim	Não	Não	Não	Sim
Algoritmo de Mineração de Dados	Regressão Linear	BPNN	$M5'$	$M5$	ANN e $M5$	Regressão Linear e $M5'$
Comparação de Modelos Preditivos	Não	Sim	Não	Não	Sim	Sim
Estimativa de ET_0	Não	Não	Não	Sim	Sim	Sim

Fonte – Elaborada pela autora.

4 OBJETIVOS

Neste Capítulo, serão apresentados o objetivo geral e os objetivos específicos deste trabalho.

4.1 Objetivo Geral

Criar modelos de predição da ET_0 que possuam tolerância à indisponibilidade de variáveis e que sejam de simples utilização.

4.2 Objetivos específicos

- Aplicar os métodos de seleção de atributos sobre os dados;
- Gerar modelos preditivos da ET_0 utilizando regressão linear a partir dos conjuntos de atributos selecionados;
- Gerar um modelo preditivo da ET_0 a partir de todos atributos da base de dados, utilizando o algoritmo *Linear Regression*;
- Gerar modelos preditivos da ET_0 utilizando o *M5'* a partir dos conjuntos de atributos selecionados;
- Gerar um modelo preditivo da ET_0 utilizando o *M5'* e todos os atributos da base de dados;
- Comparar os modelos gerados em relação à métrica: coeficiente de correlação.

5 PROCEDIMENTOS METODOLÓGICOS

Neste Capítulo, serão descritas as etapas executadas para atingir os objetivos deste trabalho.

5.1 Coleta dos dados

A primeira etapa consistiu na coleta de dados climáticos gerados pela estação meteorológica instalada no campus da UFC Quixadá. As informações e maiores detalhes sobre a estação são apresentados no Anexo A.

5.2 Preparação dos dados

Nesta etapa, os dados foram preparados para serem recebidos pelos algoritmos de seleção de atributos. Inicialmente, foi necessário converter o arquivo de dados para o formato *.csv*, visto que o arquivo original fornecido pela estação possuía extensão *.dat*. Em seguida, foram removidas as tuplas consideradas *outliers*, para garantir resultados mais precisos e de melhor acurácia.

5.3 Aplicação da seleção de atributos

Após a preparação dos dados, foi realizada a aplicação da seleção dos atributos para a criação dos modelos preditivos. Foi executado o algoritmo CFS juntamente com cada um dos métodos de busca apresentados no Capítulo 2: *BestFirst*, *GeneticSearch*, *ExhaustiveSearch* e *RandomSearch*.

5.4 Criação dos modelos preditivos

Esta etapa, consistiu na criação dos modelos de predição utilizando a regressão linear e o *M5'*. Primeiramente, para cada algoritmo, foram obtidos os modelos utilizando todos os atributos da base de dados. Após isso, de modo similar, foram criados os modelos utilizando cada conjunto de atributos selecionados na etapa abordada na Seção 6.3.

5.5 Validação e análise dos resultados

Após a criação dos modelos na etapa descrita anteriormente, os resultados de cada um foram comparados e avaliados baseando-se na métrica coeficiente de correlação. Também foram executados testes nos modelos, de modo a descobrir o quão próximos estavam os valores de ET_0 estimados pela estação meteorológica, com os valores estimados pelos modelos de predição. Para melhor visualização dos resultados dos testes, foram elaborados gráficos, apresentados no capítulo seguinte.

6 EXPERIMENTOS E RESULTADOS

Neste Capítulo, serão abordados com maiores detalhes os experimentos realizados e seus respectivos resultados.

6.1 Coleta dos dados

Como já mencionado anteriormente na Seção 5.1, esta etapa foi constituída pela coleta dos dados que foram utilizados nos experimentos. Os dados obtidos correspondem às condições climáticas monitoradas no período de 16 de junho à 19 de outubro de 2016. A coleta foi realizada por meio de uma conexão serial com o *datalogger* da estação, provida pelo *software* PCW200. O arquivo de dados recebido possuía 3191 tuplas e os atributos presentes na Figura 3.

Figura 3 – Atributos presentes nos arquivos de dados

ATRIBUTOS
ET_0
Record
Timestamp
Precipitação
Velocidade do vento
Radiação solar (total e média)
Temperatura (máxima e mínima)
Umidade relativa (mínima, máxima e média)
Temperatura do ar (mínima, máxima e média)
Pressão atmosférica (mínima, máxima e média)

Fonte – Elaborada pela autora.

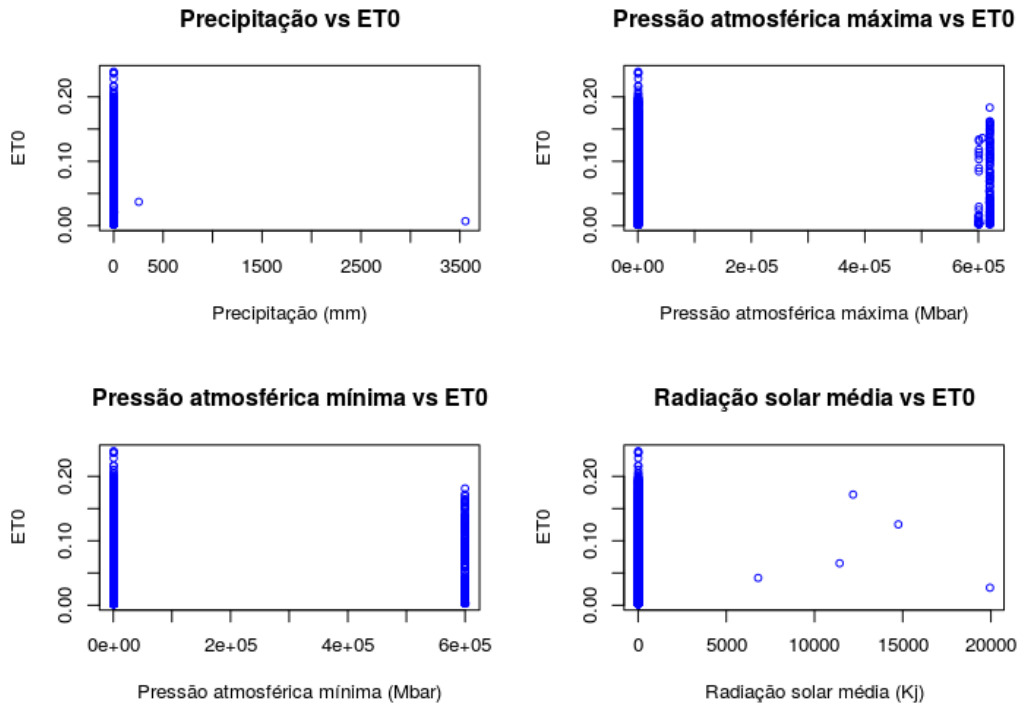
6.2 Preparação dos dados

A etapa de preparação dos dados é uma das mais importantes no processo de descoberta de conhecimento, pois é nela que são retiradas as instâncias vazias, inconsistentes e fora do padrão. Instâncias fora do padrão são chamadas de *outliers* e quando presentes na base de dados, podem reduzir de forma significativa a acurácia e a confiança dos resultados obtidos após a mineração.

A fim de visualizar o padrão e detectar os *outliers* nos dados coletados, foram elaborados gráficos nos quais, cada atributo presente na Figura 3, foi plotado em função da ET_0. Para a criação dos gráficos, foi utilizada a linguagem estatística R (R Core Team, 2015), que

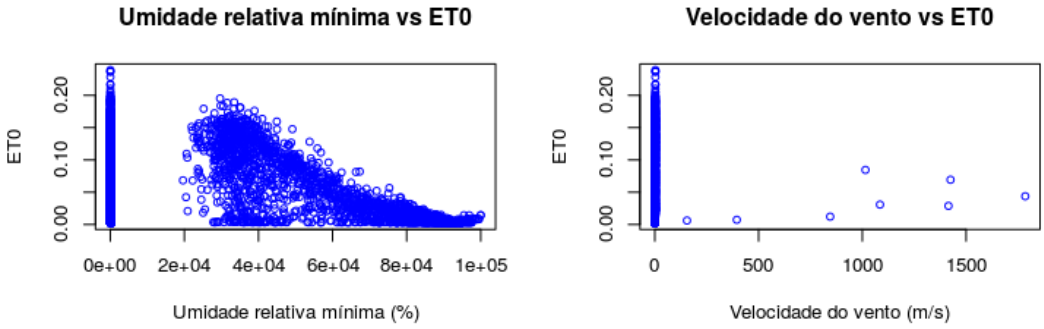
oferece diversas bibliotecas e funções para auxiliar na produção de representações gráficas de conjuntos de dados. As Figuras 4 à 7 mostram os gráficos criados.

Figura 4 – Atributos em função de ET_0



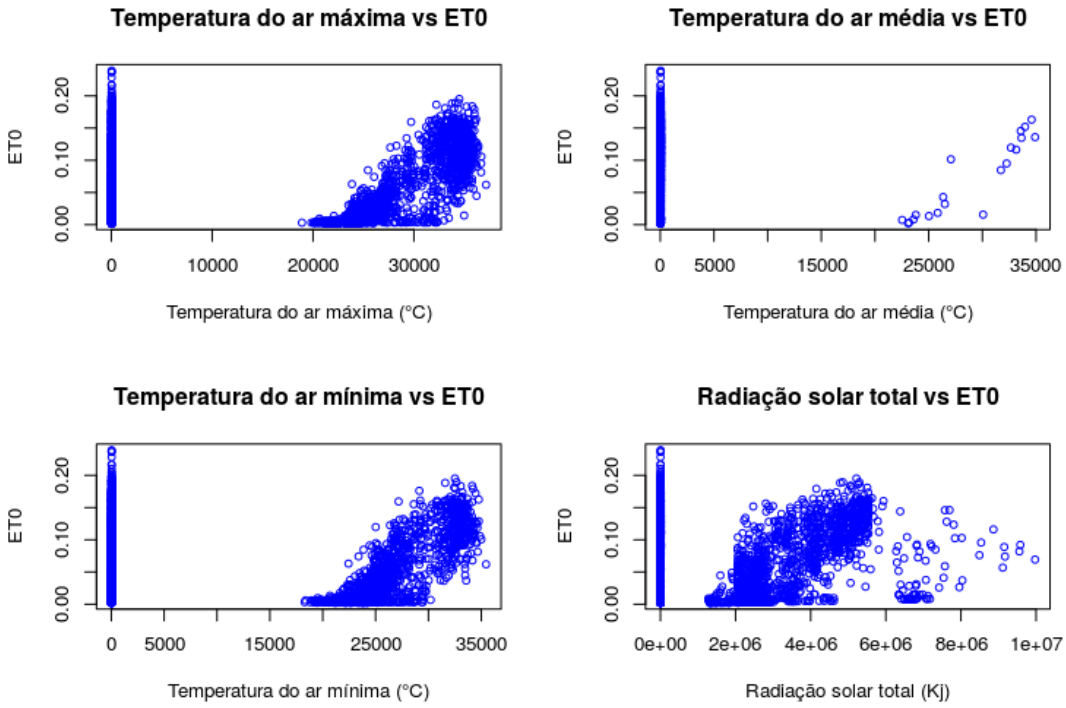
Fonte – Elaborada pela autora.

Figura 5 – Atributos em função de ET_0



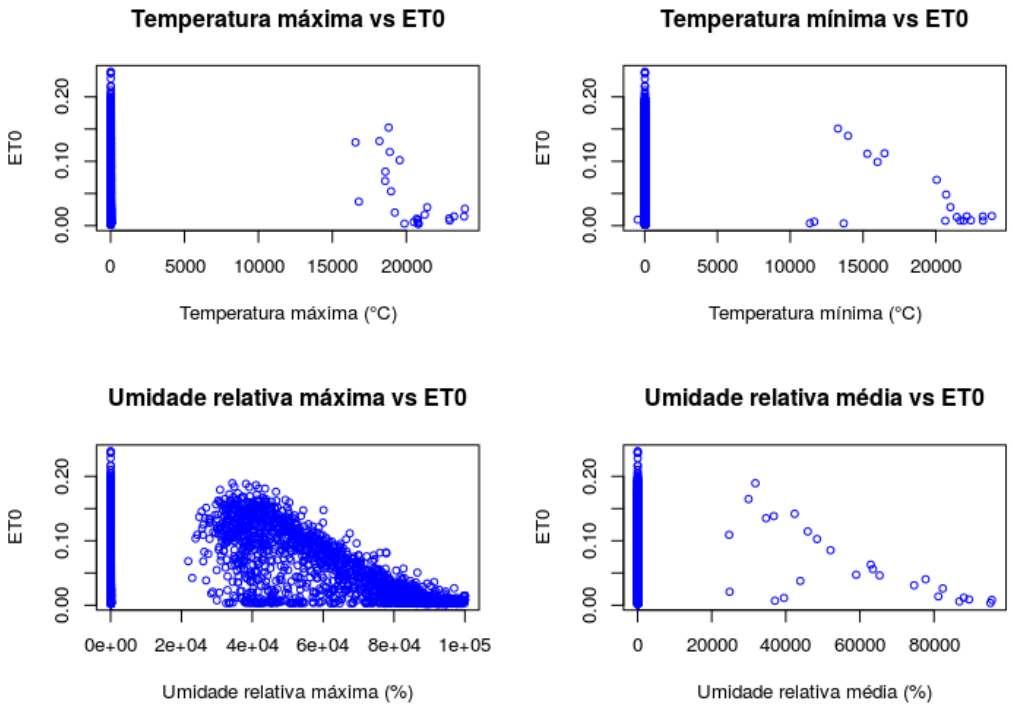
Fonte – Elaborada pela autora.

Figura 6 – Atributos em função de ET_0



Fonte – Elaborada pela autora.

Figura 7 – Atributos em função de ET_0

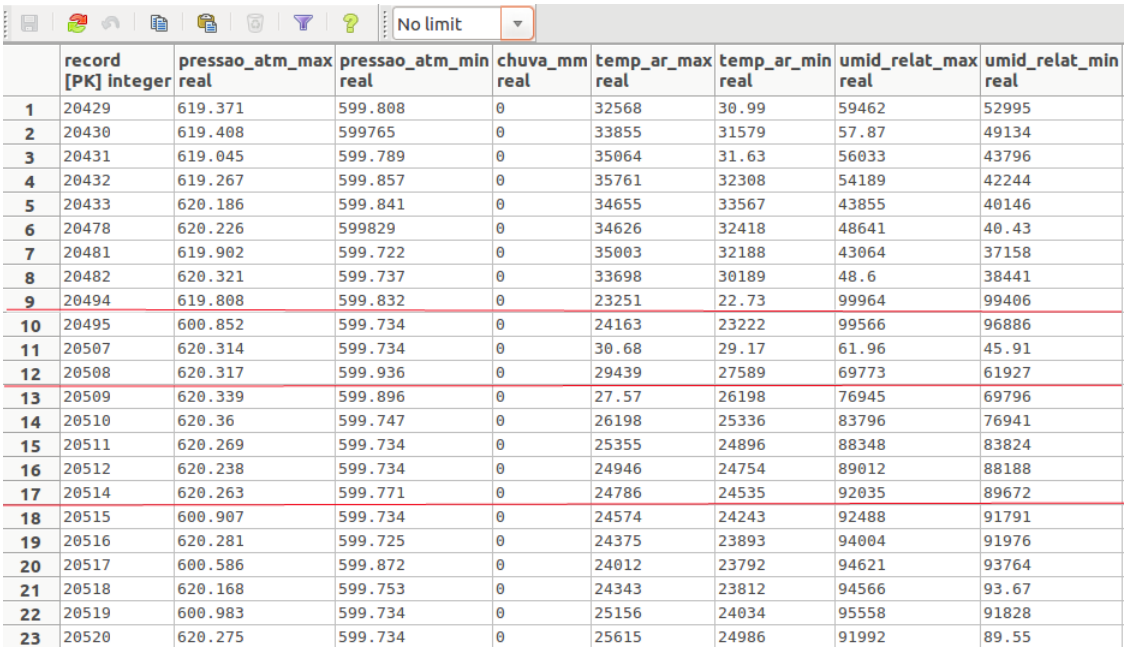


Fonte – Elaborada pela autora.

Analisando as Figuras acima, é possível notar que cada atributo possui mais de um padrão quando seus valores são colocados em função da ET_0, o que indica fortemente a existência de *outliers* na base de dados. Tomando como exemplo o gráfico da temperatura do ar mínima presente na Figura 6, pode-se perceber que existe uma grande concentração de valores iguais a 0 quando a ET_0 varia. Desse modo, todas as instâncias que possuem esse valor de temperatura devem ser removidas, pois devido a localização e as condições climáticas de Quixadá, é bastante improvável que esse dado tenha sido monitorado corretamente pela estação.

Após a observação dos gráficos, os dados foram inseridos em uma tabela no PostgreSQL para que a remoção dos *outliers* pudesse ser efetuada. Na Figura 8, estão destacadas algumas tuplas, e nelas é possível perceber a ausência da casa decimal em valores de atributos como temperatura do ar máxima, umidade relativa máxima e umidade relativa mínima. Esse padrão ocorreu em diversas tuplas de diversos atributos em toda a base de dados. Essa falha pode ter sido gerada por comportamentos inesperados de leitura dos sensores da estação, possivelmente causados por falta de manutenção.

Figura 8 – Dados presentes no banco de dados



No limit

	record [PK] integer	pressao_atm_max real	pressao_atm_min real	chuva_mm real	temp_ar_max real	temp_ar_min real	umid_relat_max real	umid_relat_min real
1	20429	619.371	599.808	0	32568	30.99	59462	52995
2	20430	619.408	599765	0	33855	31579	57.87	49134
3	20431	619.045	599.789	0	35064	31.63	56033	43796
4	20432	619.267	599.857	0	35761	32308	54189	42244
5	20433	620.186	599.841	0	34655	33567	43855	40146
6	20478	620.226	599829	0	34626	32418	48641	40.43
7	20481	619.902	599.722	0	35003	32188	43064	37158
8	20482	620.321	599.737	0	33698	30189	48.6	38441
9	20494	619.808	599.832	0	23251	22.73	99964	99406
10	20495	600.852	599.734	0	24163	23222	99566	96886
11	20507	620.314	599.734	0	30.68	29.17	61.96	45.91
12	20508	620.317	599.936	0	29439	27589	69773	61927
13	20509	620.339	599.896	0	27.57	26198	76945	69796
14	20510	620.36	599.747	0	26198	25336	83796	76941
15	20511	620.269	599.734	0	25355	24896	88348	83824
16	20512	620.238	599.734	0	24946	24754	89012	88188
17	20514	620.263	599.771	0	24786	24535	92035	89672
18	20515	600.907	599.734	0	24574	24243	92488	91791
19	20516	620.281	599.725	0	24375	23893	94004	91976
20	20517	600.586	599.872	0	24012	23792	94621	93764
21	20518	620.168	599.753	0	24343	23812	94566	93.67
22	20519	600.983	599.734	0	25156	24034	95558	91828
23	20520	620.275	599.734	0	25615	24986	91992	89.55

Fonte – Elaborada pela autora.

Devido a inconsistência presente nessas tuplas, elas foram removidas para garantir a acurácia dos modelos a serem criados. Através da realização de consultas SQL, foi possível detectar, para cada atributo, qual era o último valor que possuía casa decimal. Desse modo, todas as tuplas que possuíam valores maiores do que esse limite, foram removidas. A Figura 9

apresenta tais valores.

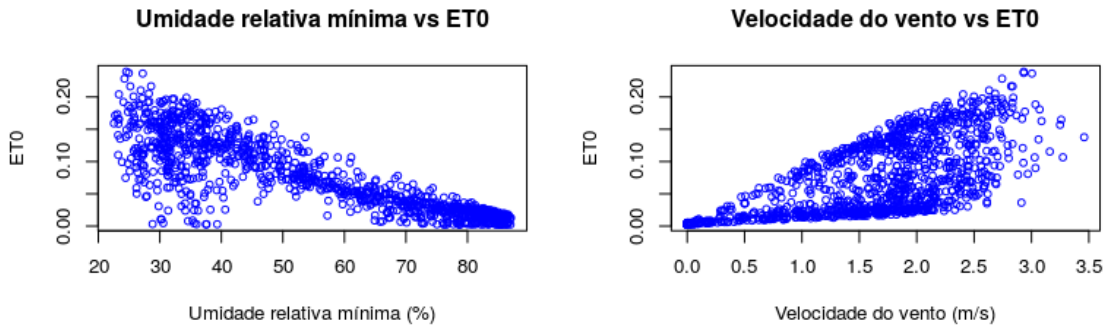
Figura 9 – Instâncias removidas

VALORES REMOVIDOS
Precipitação $>$ 1
Velocidade do vento $>$ 4
Pressão atmosférica máxima $>$ 625.000
Radiação solar média $>$ 25.4809
Radiação solar total $>$ 6000000
Temperatura mínima $>$ 23.4105
Umidade relativa média $>$ 99.8006
Temperatura máxima $>$ 24.8820
Temperatura do ar mínima $<$ 1
Temperatura do ar mínima $>$ 36.692
Temperatura do ar máxima $<$ 1
Temperatura do ar máxima $>$ 37.965
Umidade relativa mínima $>$ 87.062
Umidade relativa máxima $>$ 95.5

Fonte – Elaborada pela autora.

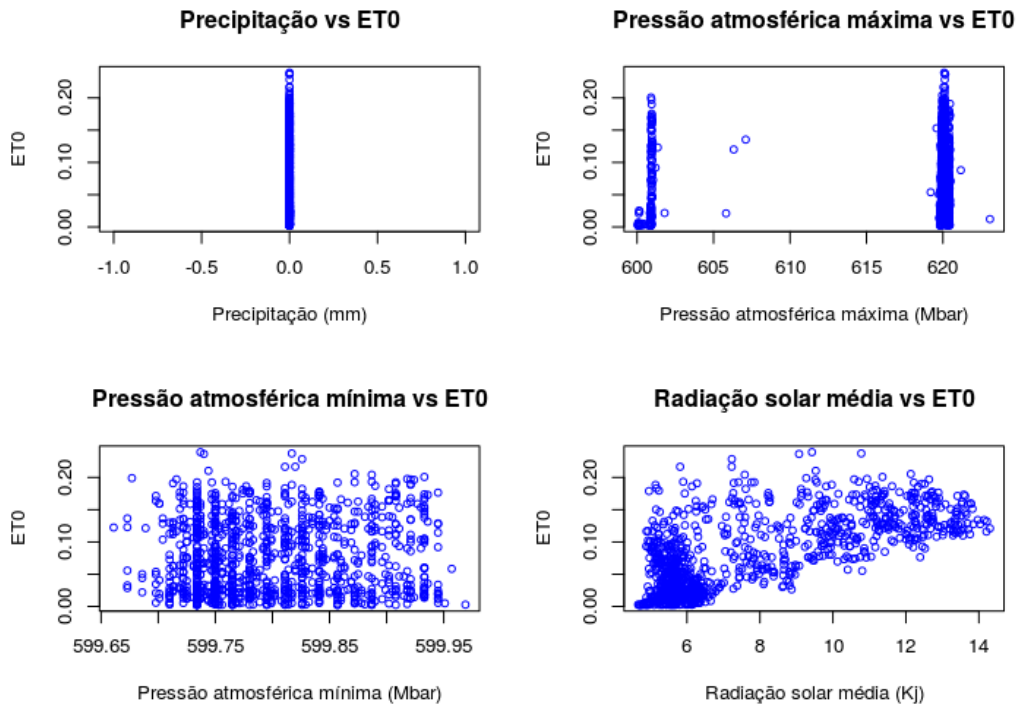
Após a retirada dos *outliers*, a quantidade de tuplas da base de dados reduziu para 1120 e foram produzidos novos gráficos para visualizar a mudança no padrão dos dados. Nas Figuras 10 à 13 é possível constatar padrões mais consistentes, diferentes dos gráficos mostrados anteriormente.

Figura 10 – Atributos em função de ET_0 após a remoção de *outliers*



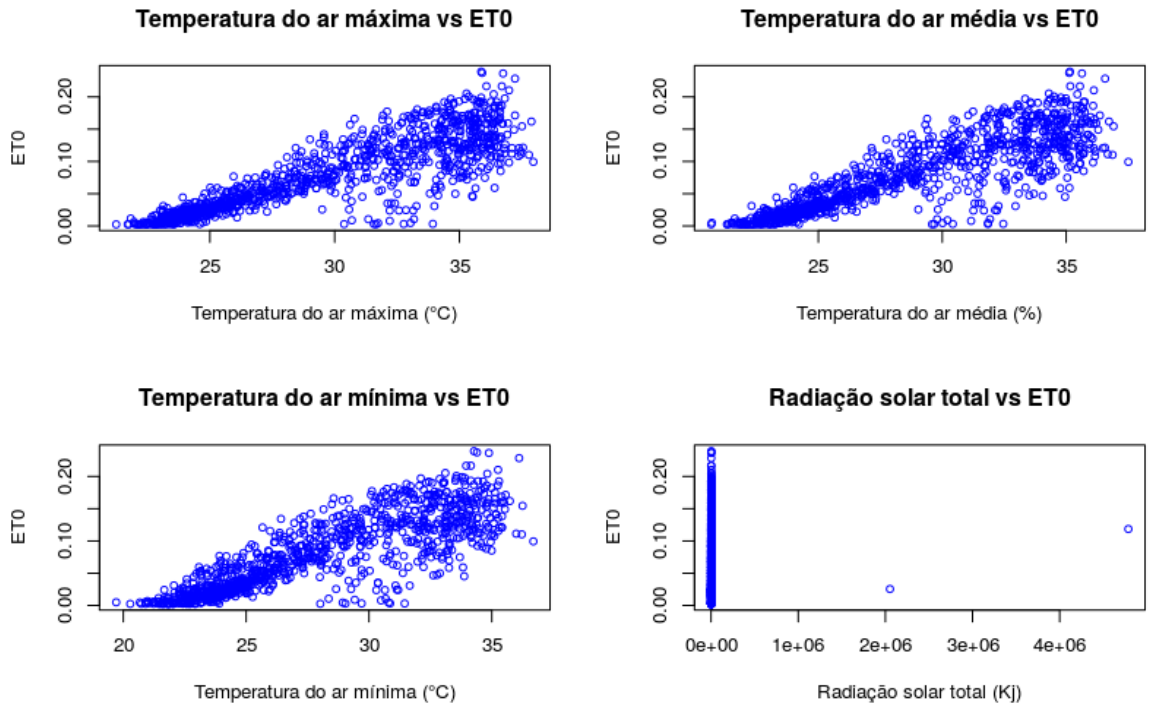
Fonte – Elaborada pela autora.

Figura 11 – Atributos em função de ET_0 após a remoção de *outliers*



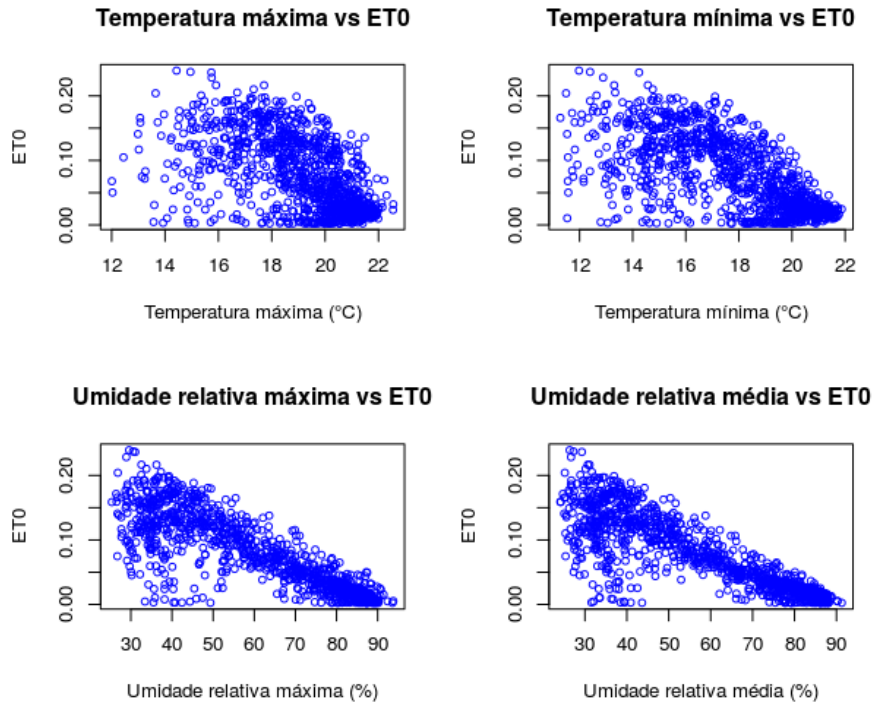
Fonte – Elaborada pela autora.

Figura 12 – Atributos em função de ET_0 após a remoção de *outliers*



Fonte – Elaborada pela autora.

Figura 13 – Atributos em função de ET_0 após a remoção de *outliers*



Fonte – Elaborada pela autora.

Analisando o novo gráfico de temperatura do ar mínima (Figura 12) é possível notar que, diferente da Figura 6, não existem mais instâncias com valor de temperatura do ar mínima iguais a 0 e todos os outros valores estão concentrados no intervalo de 20 à 35 graus *celsius*. Logo, analisando também os outros gráficos, é seguro afirmar que a remoção dos *outliers* contribuiu para o aumento da qualidade dos dados.

6.3 Aplicação da seleção de atributos

Após a etapa de preparação dos dados, foi executado o algoritmo de seleção de atributos CFS, juntamente com cada um dos algoritmos de busca explorados na seção 2.2: *BestFirst*, *ExhaustiveSearch*, *GeneticSearch* e *RandomSearch*.

O WEKA é uma ferramenta livre que implementa uma coleção de algoritmos de *machine learning* para mineração de dados, desenvolvida na *University of Waikato* (HALL et al., 2009) e foi utilizada para a realização de todos os experimentos dessa etapa.

A execução de todos os algoritmos de busca utilizaram como modo de teste o método de validação cruzada com 10 conjuntos. Na Figura 14, são apresentados os atributos selecionados por cada algoritmo e no Apêndice A são expostas as telas de saída geradas pelo WEKA durante

as execuções.

Figura 14 – Atributos selecionados por cada algoritmo

ALGORITMO	ATRIBUTOS SELECIONADOS
CFS + *BestFirst*	Pressão atmosférica mínima, temperatura do ar máxima, velocidade do vento, radiação solar média
CFS + *ExhaustiveSearch*	Pressão atmosférica mínima, temperatura do ar mínima, temperatura do ar média, velocidade do vento, radiação solar média
CFS + *GeneticSearch*	Pressão atmosférica mínima, temperatura do ar mínima, temperatura do ar máxima, temperatura do ar média, velocidade do vento, radiação solar média
CFS + *RandomSearch*	Pressão atmosférica mínima, temperatura máxima, temperatura do ar média, velocidade do vento, radiação solar média

Fonte – Elaborada pela autora.

6.4 Criação dos modelos preditivos

A criação dos modelos preditivos deu-se logo após a etapa de aplicação dos algoritmos de seleção de atributos. Os modelos foram produzidos a partir dos algoritmos *M5'* e a regressão linear, implementados na ferramenta WEKA. Para criar os modelos, a base de dados foi separada em conjuntos de treino e teste. Para treino foram utilizadas 70% das instâncias e, para teste, os 30% restantes.

Ao final da execução dos algoritmos, são informadas as taxas de erros dos modelos e seus coeficientes de correlação. O coeficiente de correlação é uma medida de avaliação e indica a relação entre os valores estimados pelo modelo e os valores reais presentes no conjunto de teste. O coeficiente varia de 0 a 1, logo, quanto mais próximo de 1 ele for, mais os valores estimados pelo modelo estarão próximos de seus valores originais do conjunto de teste.

Na Figura 15 é possível observar a correlação obtida em cada um dos modelos criados durante os experimentos. De um modo geral, todos os algoritmos produziram modelos de alta acurácia, visto que seus valores estão todos acima de 0.9. Entretanto, o modelo gerado a partir do conjunto de atributos selecionados pelo CFS + *RandomSearch* apresentou o coeficiente de correlação mais próximo ao do modelo criado sem a seleção de atributos.

Figura 15 – Coeficiente de correlação dos modelos criados

ALGORITMO	***M5'***	**REGRESSÃO LINEAR**
CFS + *BestFirst*	0.988	0.9511
CFS + *ExhaustiveSearch*	0.9897	0.9536
CFS + *GeneticSearch*	0.9899	0.9536
CFS + *RandomSearch*	0.9963	0.9653
Sem seleção de atributos	0.9969	0.9659

Fonte – Elaborada pela autora.

Para melhor organização do texto, apenas os melhores modelos serão expostos nessa Seção e os demais podem ser encontrados no Apêndice B. A seguir, na Figura 16, é apresentado o modelo criado pela regressão linear. Nas Figuras 17, 18 e 19 são mostradas, respectivamente, a árvore e suas equações correspondentes, produzidas pelo *M5'*. Ambos os algoritmos utilizaram como entrada o conjunto de atributos selecionado pelo CFS + *RandomSearch* para criar seus modelos.

Figura 16 – Modelo gerado pela Regressão Linear utilizando os atributos selecionados pelo CFS + *RandomSearch*

REGRESSÃO LINEAR
et0 = -0.006 * tempMax + 0.0057 * tempArMedia + 0.0289 * ventoVeloc + 0.0055 * radSolarMedia + -0.0601

Fonte – Elaborada pela autora.

Figura 17 – Árvore do *M5'* utilizando os atributos selecionados pelo CFS + *RandomSearch*

```
tempArMedia <= 26.4 :
|   ventoVeloc <= 0.784 : EQ1
|   ventoVeloc >  0.784 :
|   |   tempArMedia <= 24.718 :
|   |   |   radSolarMedia <= 6.249 : EQ2
|   |   |   radSolarMedia >  6.249 : EQ3
|   |   tempArMedia >  24.718 :
|   |   |   radSolarMedia <= 6.963 : EQ4
|   |   |   radSolarMedia >  6.963 : EQ5
tempArMedia >  26.4 :
|   radSolarMedia <= 6.122 :
|   |   tempArMedia <= 29.416 : EQ6
|   |   tempArMedia >  29.416 :
|   |   |   ventoVeloc <= 0.822 : EQ7
|   |   |   ventoVeloc >  0.822 :
|   |   |   |   ventoVeloc <= 1.433 :
|   |   |   |   |   tempArMedia <= 32.715 : EQ8
|   |   |   |   |   tempArMedia >  32.715 : EQ9
|   |   |   |   ventoVeloc >  1.433 :
|   |   |   |   |   tempArMedia <= 31.746 : EQ10
|   |   |   |   |   tempArMedia >  31.746 : EQ11
|   radSolarMedia >  6.122 :
|   |   tempArMedia <= 30.43 :
|   |   |   ventoVeloc <= 1.851 : EQ12
|   |   |   ventoVeloc >  1.851 :
|   |   |   |   tempMax <= 20.49 : EQ13
|   |   |   |   tempMax >  20.49 : EQ14
|   |   tempArMedia >  30.43 :
|   |   |   ventoVeloc <= 1.92 :
|   |   |   |   ventoVeloc <= 1.356 : EQ15
|   |   |   |   ventoVeloc >  1.356 : EQ16
|   |   |   ventoVeloc >  1.92 :
|   |   |   |   tempArMedia <= 32.648 :
|   |   |   |   |   ventoVeloc <= 2.264 : EQ17
|   |   |   |   |   ventoVeloc >  2.264 : EQ18
|   |   |   |   tempArMedia >  32.648 :
|   |   |   |   |   ventoVeloc <= 2.376 : EQ19
|   |   |   |   |   ventoVeloc >  2.376 : EQ20
```

Fonte – Elaborada pela autora.

Figura 18 – Equações geradas pelo *M5'* utilizando os atributos selecionados pelo CFS + *Random Search*

NOME	**EQUAÇÃO**
EQ01	et0 = -0.0022 * tempMax + 0.002 * tempArMedia + 0.0145 * ventoVeloc + 0.0024 * radSolarMedia - 0.0115
EQ02	et0 = -0.0057 * tempMax + 0.0054 * tempArMedia + 0.0097 * ventoVeloc + 0.001 * radSolarMedia - 0.0088
EQ03	et0 = -0.0103 * pressaoAtmMin - 0.0064 * tempMax + 0.0035 * tempArMedia + 0.0156 * ventoVeloc + 0.0052 * radSolarMedia + 6.1857
EQ04	et0 = -0.0069 * tempMax + 0.0065 * tempArMedia + 0.0138 * ventoVeloc + 0.0031 * radSolarMedia - 0.0293
EQ05	et0 = -0.0068 * tempMax + 0.0086 * tempArMedia + 0.0187 * ventoVeloc + 0.0023 * radSolarMedia - 0.0787
EQ06	et0 = -0.0043 * tempMax + 0.0071 * tempArMedia + 0.0275 * ventoVeloc - 0.0024 * radSolarMedia - 0.0932
EQ07	et0 = -0.0021 * tempMax + 0.0044 * tempArMedia + 0.0617 * ventoVeloc + 0.0004 * radSolarMedia - 0.1006
EQ08	et0 = -0.002 * tempMax + 0.009 * tempArMedia + 0.0521 * ventoVeloc + 0.0004 * radSolarMedia - 0.2454
EQ09	et0 = -0.0023 * tempMax + 0.0083 * tempArMedia + 0.0602 * ventoVeloc + 0.0004 * radSolarMedia - 0.2245
EQ10	et0 = -0.0034 * tempMax + 0.0108 * tempArMedia + 0.0393 * ventoVeloc + 0.0004 * radSolarMedia - 0.2582
EQ11	et0 = -0.0041 * tempMax + 0.012 * tempArMedia + 0.0551 * ventoVeloc - 0.0031 * radSolarMedia - 0.294
EQ12	et0 = -0.0062 * tempMax + 0.0074 * tempArMedia + 0.0407 * ventoVeloc + 0.002 * radSolarMedia - 0.0994
EQ13	et0 = -0.0071 * tempMax + 0.0091 * tempArMedia + 0.0403 * ventoVeloc + 0.0016 * radSolarMedia - 0.1248
EQ14	et0 = -0.0067 * tempMax + 0.0089 * tempArMedia + 0.0357 * ventoVeloc + 0.0041 * radSolarMedia - 0.1412
EQ15	et0 = 0.003 * pressaoAtmMin - 0.0029 * tempMax + 0.0061 * tempArMedia + 0.0679 * ventoVeloc + 0.0008 * radSolarMedia - 1.9219

Fonte – Elaborada pela autora.

Figura 19 – Equações geradas pelo *M5'* utilizando os atributos selecionados pelo CFS + *Random Search*

NOME	EQUAÇÃO
EQ16	et0 = 0.0078 * pressaoAtmMin - 0.0041 * tempMax + 0.0068 * tempArMedia + 0.0643 * ventoVeloc + 0.0014 * radSolarMedia - 4.8287
EQ17	et0 = 0.0031 * pressaoAtmMin - 0.0066 * tempMax + 0.009 * tempArMedia + 0.0497 * ventoVeloc + 0.001 * radSolarMedia - 1.9765
EQ18	et0 = 0.0025 * pressaoAtmMin - 0.0065 * tempMax + 0.0103 * tempArMedia + 0.0499 * ventoVeloc + 0.0009 * radSolarMedia - 1.659
EQ19	et0 = 0.0015 * pressaoAtmMin - 0.005 * tempMax + 0.0077 * tempArMedia + 0.0643 * ventoVeloc + 0.0015 * radSolarMedia - 1.0391
EQ20	et0 = 0.0018 * pressaoAtmMin - 0.0058 * tempMax + 0.0086 * tempArMedia + 0.0635 * ventoVeloc + 0.0018 * radSolarMedia - 1.2602

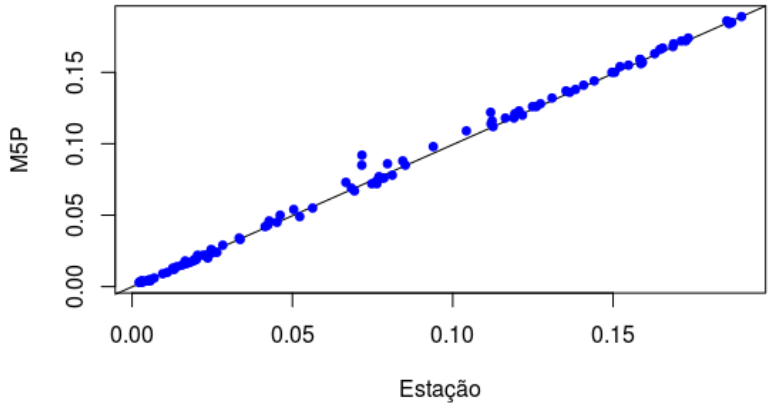
Fonte – Elaborada pela autora.

6.5 Validação e análise dos resultados

A fim de validar os modelos obtidos, foram selecionadas 100 instâncias do conjunto de teste usado na etapa anterior, para compor um outro conjunto de testes. Nesse novo conjunto, para cada instância, foi removido o seu respectivo valor de ET_0. Após essa remoção, essas instâncias foram submetidas aos modelos criados, de modo que novos valores de ET_0 fossem estimados. Esse procedimento foi executado com o auxílio da ferramenta WEKA (HALL et al., 2009).

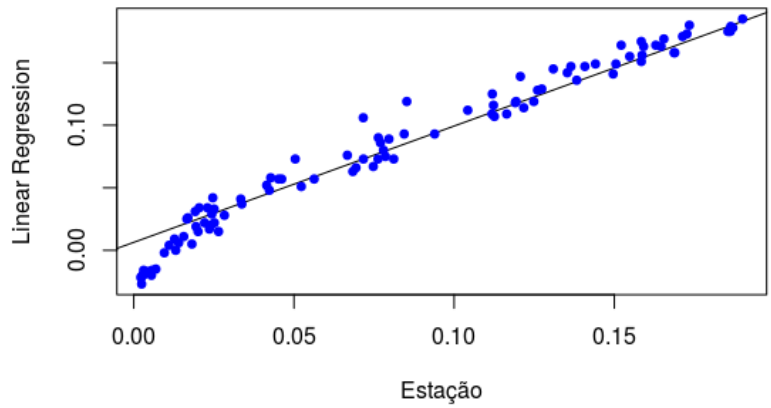
Utilizando os novos valores de ET_0 fornecidos pelos modelos e os valores de ET_0 originais fornecidos pela estação meteorológica, foram criados gráficos para ilustrar a proximidade desses valores. Para cada gráfico, foi executada a regressão linear, que gerou uma função de ajuste dos dados e seu respectivo coeficiente de determinação. O coeficiente de determinação (R^2) é uma medida de ajustamento em relação aos valores observados. Sua variação é de 0 a 1, e indica quanto o modelo pode explicar esses valores. Quanto mais próximo de 1 é o R^2, mais o modelo se ajusta ao conjunto de dados. Os gráficos foram produzidos com a linguagem R (R Core Team, 2015), e são apresentados a seguir nas Figuras 20 à 29. A Figura 30 mostra os valores de R^2 de cada gráfico.

Figura 20 – Correlação entre os valores de ET_0 estimados pelo *M5'* e pela estação



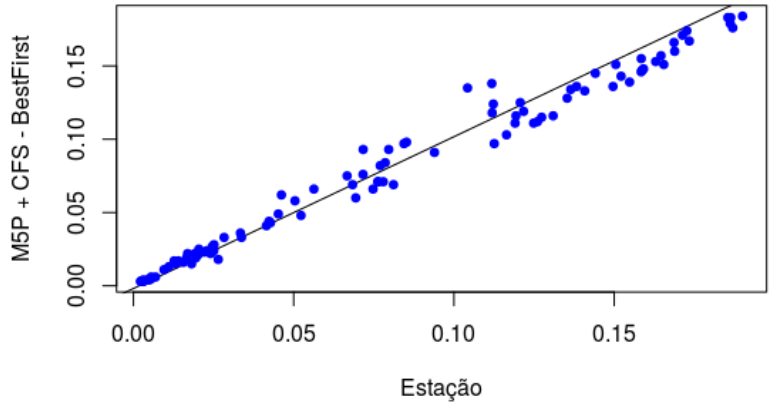
Fonte – Elaborada pela autora.

Figura 21 – Correlação entre os valores de ET_0 estimados pela Regressão Linear e pela estação



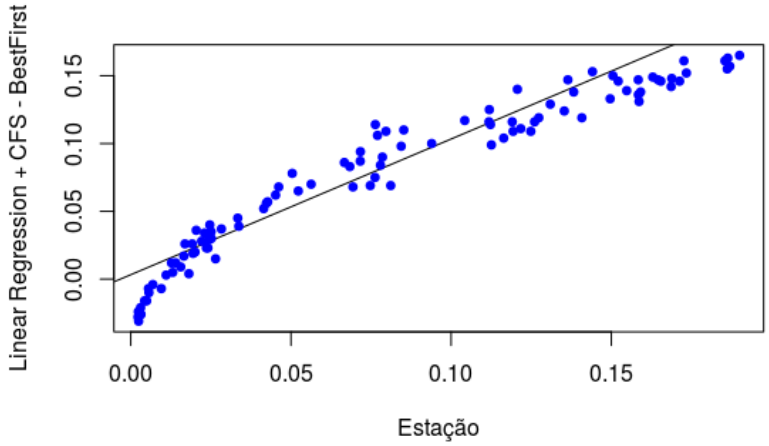
Fonte – Elaborada pela autora.

Figura 22 – Correlação entre os valores de ET_0 estimados pelo *M5'* + *BestFirst* e pela estação



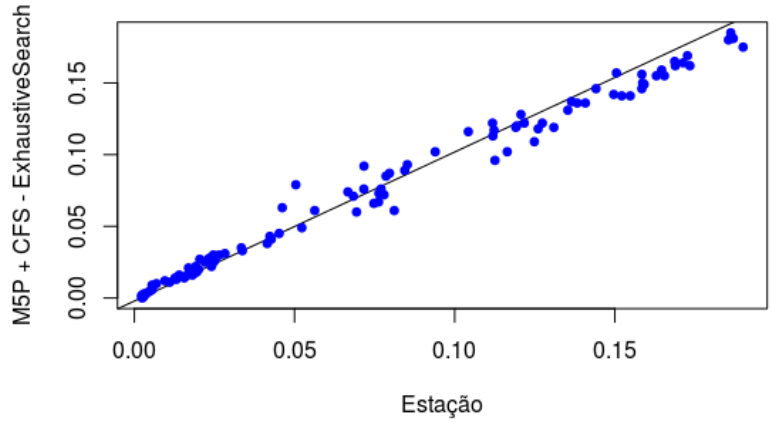
Fonte – Elaborada pela autora.

Figura 23 – Correlação entre os valores de ET_0 estimados pela Regressão Linear + *BestFirst* e pela estação



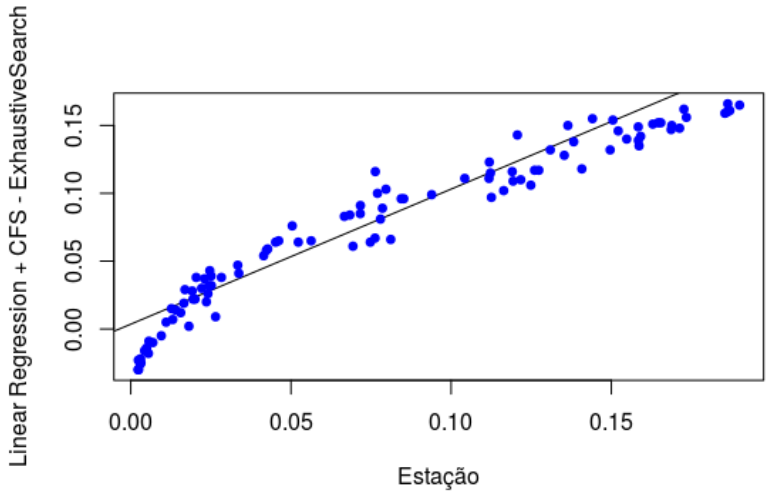
Fonte – Elaborada pela autora.

Figura 24 – Correlação entre os valores de ET_0 estimados pelo *M5'* + *ExhaustiveSearch* e pela estação



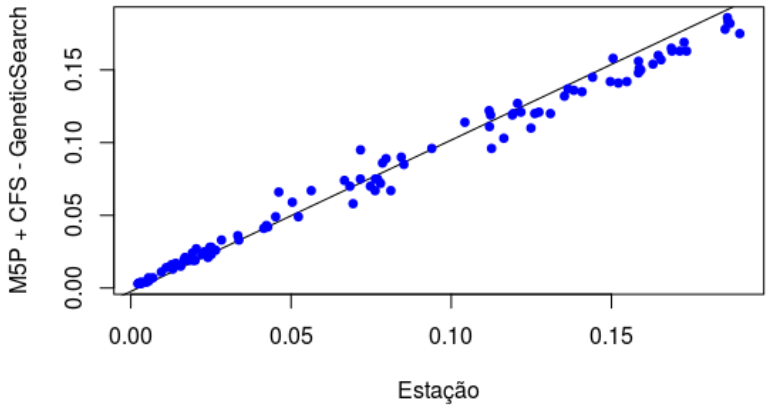
Fonte – Elaborada pela autora.

Figura 25 – Correlação entre os valores de ET_0 estimados pela Regressão Linear + *ExhaustiveSearch* e pela estação



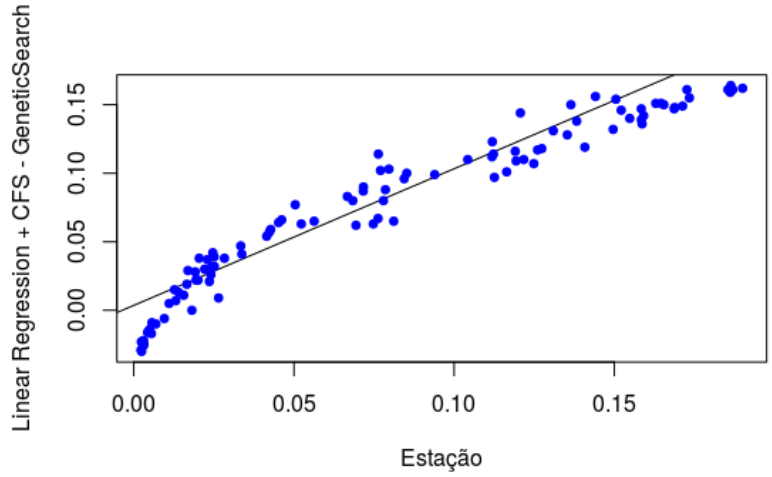
Fonte – Elaborada pela autora.

Figura 26 – Correlação entre os valores de ET_0 estimados pelo *M5'* + *GeneticSearch* e pela estação



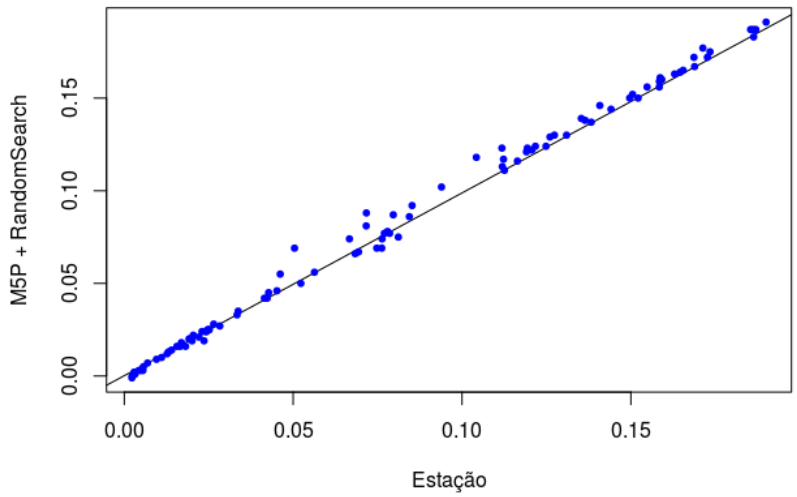
Fonte – Elaborada pela autora.

Figura 27 – Correlação entre os valores de ET_0 estimados pela Regressão Linear + *GeneticSearch* e pela estação



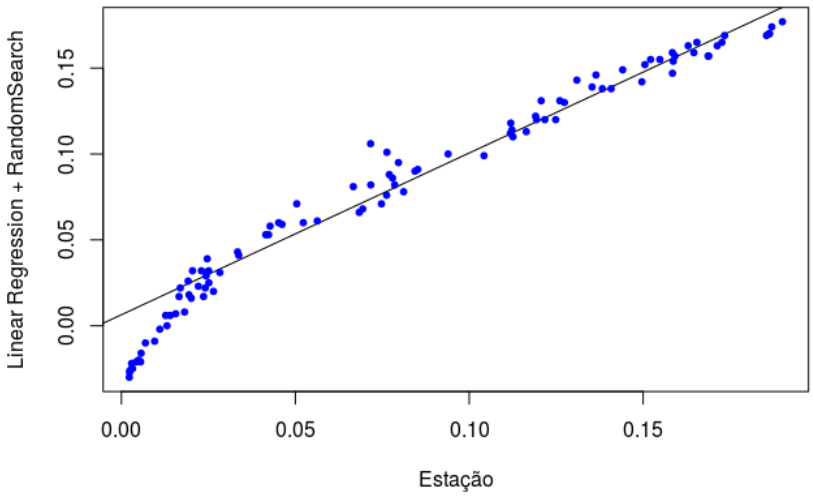
Fonte – Elaborada pela autora.

Figura 28 – Correlação entre os valores de ET_0 estimados pelo *M5'* + *RandomSearch* e pela estação



Fonte – Elaborada pela autora.

Figura 29 – Correlação entre os valores de ET_0 estimados pela Regressão Linear + *RandomSearch* e pela estação



Fonte – Elaborada pela autora.

Figura 30 – Coeficientes de determinação (R^2)

ALGORITMO	***M5'***	**REGRESSÃO LINEAR**
CFS + *GeneticSearch*	0.989	0.9302
CFS + *ExhaustiveSearch*	0.989	0.9308
CFS + *BestFirst*	0.9821	0.9262
CFS + *RandomSearch*	0.9957	0.9618
Sem seleção de atributos	0.9972	0.9666

Fonte – Elaborada pela autora.

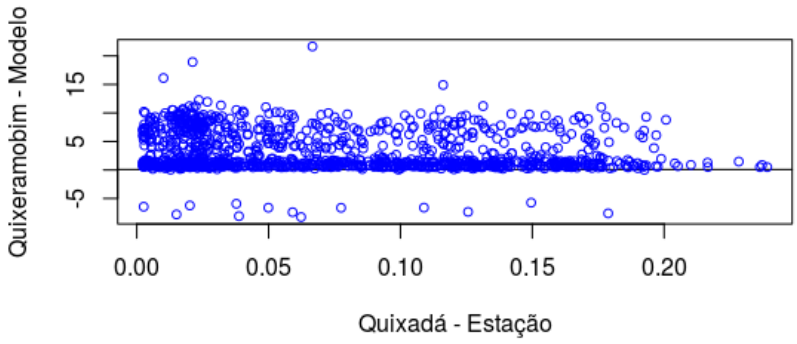
Tomando como base a Figura 15 na Seção 6.4, e a Figura 30, é possível afirmar que o *M5'* criou modelos de maior correlação que a regressão linear. Na abordagem que utilizou a seleção de atributos, o algoritmo de busca que obteve maior acurácia quando executado junto ao *CFS* foi o *RandomSearch*, que apresentou para o *M5'*, coeficiente de correlação e R^2 iguais à 0.9963 e 0.9957, e para a regressão linear iguais à 0.9653 e 0.9618, respectivamente. Note que esses valores são muito próximos aos coeficientes obtidos pelos modelos que não utilizaram seleção de atributos, o que indica que o conjunto de atributos selecionados pelo *CFS + RandomSearch* é o que melhor representa a base dados original usada para os experimentos.

Xavier, Tanaka e Revoredo (2015a) concluíram em seu trabalho, que era possível utilizar o mesmo modelo de predição da evapotranspiração potencial para cidades com condições climáticas semelhantes. De modo análogo, foram realizados testes com dados climáticos fornecidos pelo INMET, da cidade de Quixeramobim, a fim de investigar a possibilidade da

reutilização dos modelos criados para Quixadá, ou da criação de um modelo unificado que pudesse ser empregado nas duas cidades.

Para tal, foi criado a partir dos dados da estação de Quixadá um novo modelo com os atributos correspondentes à base do INMET. Entretando, os dados de Quixeramobim não possuíam o atributo radiação solar média, logo, ele foi retirado dos dados de Quixadá antes da criação do modelo. A criação do modelo deu-se pela execução do *M5'* sem seleção de atributos. A Figura 31 apresenta o gráfico que ilustra a correlação entre os valores de ET_0 e a Figura 32 apresenta o coeficiente de determinação R^2.

Figura 31 – Correlação entre os valores de ET_0 de Quixeramobim estimados pelo modelo e os valores de ET_0 de Quixadá estimados pela estação



Fonte – Elaborada pela autora.

Figura 32 – Coeficiente de determinação entre os valores de ET_0 de Quixeramobim estimados pelo modelo e os valores de ET_0 de Quixadá estimados pela estação

```
Call:
lm(formula = et0EstacaoReal$et0 ~ QxbEt0$predicted)

Residuals:
     Min       1Q   Median       3Q      Max
-0.10103 -0.04783 -0.01179  0.04532  0.15574

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.0848253  0.0022926  36.999  < 2e-16 ***
QxbEt0$predicted -0.0029208  0.0004898  -5.963 3.31e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05584 on 1118 degrees of freedom
Multiple R-squared:  0.03083,   Adjusted R-squared:  0.02996
F-statistic: 35.56 on 1 and 1118 DF,  p-value: 3.313e-09

> |
```

Fonte – Elaborada pela autora.

Como pode ser observado nas figuras acima, o R^2 é igual a 0.03083, um valor muito baixo de correlação. Isso pode indicar que, embora as cidades de Quixadá e Quixeramobim sejam próximas e possuam condições climáticas semelhantes, o mesmo modelo de predição da ET_0 não pode ser utilizado. Um outro possível motivo para a baixa correlação do modelo, foi a retirada do atributo de radiação solar média. É possível notar que em todos os modelos apresentados Seção 6.4, a radiação solar média fazia-se presente nas equações, fato este que caracteriza sua importância para criação dos modelos. Entretanto, ainda faz-se necessário um estudo mais detalhado para confirmar os motivos da baixa correlação, visto que Xavier, Tanaka e Revoredo (2015a) mostraram que é possível a reutilização do mesmo modelo em locais de condições climáticas parecidas.

7 CONCLUSÃO E TRABALHOS FUTUROS

A agricultura irrigada é o principal meio de garantir a segurança alimentar, bem como a qualidade dos alimentos. O manejo da irrigação consiste em técnicas que visam utilizar a água disponível para irrigação de modo eficiente a partir do conhecimento das características do solo, das condições climáticas e do tipo de cultura. A estimativa da evapotranspiração de referência (ET_0) é uma das principais atividades para entender a necessidade de água de uma determinada cultura. Entretanto, os métodos existentes para realizar tal estimativa são complexos e dispendiosos.

Este trabalho apresentou modelos preditivos da ET_0 compostos de equações lineares simples e de fácil aplicação, criados a partir de dados meteorológicos fornecidos pela estação meteorológica instalada na UFC Quixadá. Foram apresentadas alternativas com e sem seleção de atributos, e ambas as abordagens resultaram em modelos de alta acurácia. Dentre os algoritmos escolhidos para a criação dos modelos, o *M5'* apresentou taxas de acerto maiores do que a regressão linear. Além disso, dentre os quatro algoritmos de busca executados (*BestFirst*, *ExhaustiveSearch*, *GeneticSearch* e *RandomSearch*) juntamente com o avaliador *CFS* na etapa de seleção de atributos, o *RandomSearch* selecionou um conjunto de atributos que resultou em modelos com acurácia muito próximas dos modelos que empregaram todos os atributos. Logo, ao fim deste estudo, foi possível concluir que estimar a ET_0 usando modelos preditivos pode ser uma alternativa viável e menos complexa para agrônomos, agricultores e pesquisadores da área.

Como trabalhos futuros, pode-se destacar a replicação da metodologia apresentada neste trabalho para analisar dados climáticos de outras cidades, e tentar identificar a possibilidade de desenvolver um modelo unificado capaz de estimar a ET_0 de uma região sem grandes percas de acurácia. Uma outra proposta também é a construção de uma ferramenta *online* que crie e disponibilize esses modelos, de modo que possa auxiliar pessoas interessadas em realizar essa estimativa a partir do seu próprio conjunto de dados.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ÁGUAS. **Relatório de Conjuntura dos Recursos Hídricos no Brasil 2014**. [S.l.], 2015.

ALLEN, R. G.; PEREIRA, L. S.; RAES, D.; SMITH, M. **Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56**. FAO, Rome, v. 300, n. 9, p. D05109, 1998.

CAMINHA, H.; SILVA, T.; ROCHA, A.; LIMA, S. **Estimating Reference Evapotranspiration using Data Mining Prediction Models and Feature Selection**. Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017), v. 1, p. 272–279, 2017.

FAO. **Irrigation Sector Reform**. [S.l.], 2015. Portal Corporativo. Disponível em: <http://www.fao.org/nr/water/topics__irrig_reform.html>. Acesso em: 24 set. 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. AI magazine, v. 17, n. 3, p. 37, 1996.

FRIZZONE, J. A.; SOUZA, F. de; LIMA, S. C. R. V. **Manejo da irrigação: Quando, Quanto e Como Irrigar**. [S.l.]: INOVAGRI, 2013.

GARCES-RESTREPO, C.; VERMILLION, D.; MUOZ, G. **Irrigation management transfer: Worldwide efforts and results**. Roma, FAO, 2007.

GOLDBERG, D. E.; HOLLAND, J. H. **Genetic algorithms and machine learning**. Machine learning, Springer, v. 3, n. 2, p. 95–99, 1988.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. **The WEKA data mining software: an update**. ACM SIGKDD explorations newsletter, ACM, v. 11, n. 1, p. 10–18, 2009.

HALL, M. A. **Correlation-based feature selection of discrete and numeric class machine learning**. University of Waikato, Department of Computer Science, 2000.

HENDRAWAN, Y.; MURASE, H. **Neural-intelligent water drops algorithm to select relevant textural features for developing precision irrigation system using machine vision**. Computers and Electronics in Agriculture, Elsevier, v. 77, n. 2, p. 214–228, 2011.

KAREGOWDA, A. G.; MANJUNATH, A.; JAYARAM, M. **Comparative study of attribute selection using gain ratio and correlation based feature selection**. International Journal of Information Technology and Knowledge Management, v. 2, n. 2, p. 271–277, 2010.

LIU, H.; YU, L. **Toward integrating feature selection algorithms for classification and clustering**. IEEE Transactions on knowledge and data engineering, IEEE, v. 17, n. 4, p. 491–502, 2005.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. [S.l.]: John Wiley & Sons, 2015.

QUINLAN, J. R. **Learning with continuous classes**. 5th Australian joint conference on artificial intelligence, v. 92, p. 343–348, 1992.

R Core Team. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2015. Disponível em: <https://www.R-project.org/>.

RAHIMIKHOOB, A. **Comparison between M5 model tree and neural networks for estimating reference evapotranspiration in an arid environment**. Water resources management, Springer, v. 28, n. 3, p. 657–669, 2014.

SAWALKAR, N.; DIXIT, P. **Evapotranspiration Modeling Using M5 Model Tree**. International journal of earth sciences and engineering, v. 8, n. 2, p. 491–496, 2015.

TUNDISI, J. G. **Ciclo hidrológico e gerenciamento integrado**. Ciência e Cultura, scielocec, v. 55, p. 31 – 33, 12 2003. Disponível em: <http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252003000400018&nrm=iso>. Acesso em: 11 out. 2016.

WANG, Y.; WITTEN, I. H. **Induction of model trees for predicting continuous classes**. Working paper series, University of Waikato, Department of Computer Science, 1996.

XAVIER, F.; TANAKA, A. K.; REVOREDO, K. C. **Aplicação de KDD em dados meteorológicos para identificação de padrões regionais na estimativa da evapotranspiração**. Proceedings of satellite events of the 30th braziliam symposium on databases, 2015a.

XAVIER, F.; TANAKA, A. K.; REVOREDO, K. C. **Application of knowledge discovery in databases in evapotranspiration estimation: an experiment in the State of Rio de Janeiro**. SBSI 2015 Proceedings, 2015b.

ZABINSKY, Z. B. **Random search algorithms**. Wiley Encyclopedia of Operations Research and Management Science, Wiley Online Library, 2009.

APÊNDICE A – SELEÇÃO DE ATRIBUTOS

Neste apêndice serão expostas as telas de saídas com os resultados, obtidas pela execução dos algoritmos de seleção de atributos utilizando a ferramenta WEKA.

Resultados da execução do *CFS* com os algoritmos de busca

A seguir, é apresentada a saída gerada pela execução do *CFS* com o algoritmo de busca *BestFirst*. Note que o campo *number of folds* indica a quantidade de conjuntos em que um determinado atributo estava presente. Desse modo, é possível inferir que os atributos que obtiveram essa quantidade acima de 0, foram selecionados pelo algoritmo como candidatos para representar a base de dados.

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval
Search:weka.attributeSelection.BestFirst -D 1 -N 5
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    1120
Attributes:   15
              pressao_atm_max
              pressao_atm_min
              chuva_mm
              temp_ar_max
              temp_ar_min
              umid_relat_max
              umid_relat_min
              rad_solar_total
              temp_max
              temp_min
              temp_ar_media
              umid_relat_media
              vento_veloc
              rad_solar_media
              et0
Evaluation mode:10-fold cross-validation

=== Attribute selection 10 fold cross-validation seed: 1 ===

number of folds (%)  attribute
           0(  0 %)    1 pressao_atm_max
           2( 20 %)    2 pressao_atm_min
           0(  0 %)    3 chuva_mm
```

```
          10(100 %)    4 temp_ar_max
           0(  0 %)    5 temp_ar_min
           0(  0 %)    6 umid_relat_max
           0(  0 %)    7 umid_relat_min
           0(  0 %)    8 rad_solar_total
           0(  0 %)    9 temp_max
           0(  0 %)   10 temp_min
           0(  0 %)   11 temp_ar_media
           0(  0 %)   12 umid_relat_media
          10(100 %)   13 vento_veloc
          10(100 %)   14 rad_solar_media
```

Saída gerada pela execução do *CFS* com o algoritmo de busca *ExhaustiveSearch*:

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval
Search:weka.attributeSelection.ExhaustiveSearch
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    1120
Attributes:   15
              pressao_atm_max
              pressao_atm_min
              chuva_mm
              temp_ar_max
              temp_ar_min
              umid_relat_max
              umid_relat_min
              rad_solar_total
              temp_max
              temp_min
              temp_ar_media
              umid_relat_media
              vento_veloc
              rad_solar_media
              et0
Evaluation mode:10-fold cross-validation



=== Attribute selection 10 fold cross-validation seed: 1 ===

number of folds (%)  attribute
           0(  0 %)    1 pressao_atm_max
           2( 20 %)    2 pressao_atm_min
           0(  0 %)    3 chuva_mm
           0(  0 %)    4 temp_ar_max
```

```
           1( 10 %)    5 temp_ar_min
           0(  0 %)    6 umid_relat_max
           0(  0 %)    7 umid_relat_min
           0(  0 %)    8 rad_solar_total
           0(  0 %)    9 temp_max
           0(  0 %)   10 temp_min
           9( 90 %)   11 temp_ar_media
           0(  0 %)   12 umid_relat_media
          10(100 %)   13 vento_veloc
          10(100 %)   14 rad_solar_media
```

Saída gerada pela execução do *CFS* com o algoritmo de busca *GeneticSearch*:

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval
Search:weka.attributeSelection.GeneticSearch -Z 20 -G 20 -C 0.6 -M 0.033 -R 20 -S 1
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    1120
Attributes:   15
              pressao_atm_max
              pressao_atm_min
              chuva_mm
              temp_ar_max
              temp_ar_min
              umid_relat_max
              umid_relat_min
              rad_solar_total
              temp_max
              temp_min
              temp_ar_media
              umid_relat_media
              vento_veloc
              rad_solar_media
              et0
Evaluation mode:10-fold cross-validation



=== Attribute selection 10 fold cross-validation seed: 1 ===

number of folds (%)  attribute
           0(  0 %)    1 pressao_atm_max
           3( 30 %)    2 pressao_atm_min
           0(  0 %)    3 chuva_mm
           2( 20 %)    4 temp_ar_max
```

```
 1( 10 %)    5 temp_ar_min
 0(  0 %)    6 umid_relat_max
 0(  0 %)    7 umid_relat_min
 0(  0 %)    8 rad_solar_total
 0(  0 %)    9 temp_max
 0(  0 %)   10 temp_min
 7( 70 %)   11 temp_ar_media
 0(  0 %)   12 umid_relat_media
10(100 %)   13 vento_veloc
10(100 %)   14 rad_solar_media
```

Saída gerada pela execução do *CFS* com o algoritmo de busca *RandomSearch*:

```
=== Run information ===

Evaluator:    weka.attributeSelection.CfsSubsetEval
Search:weka.attributeSelection.RandomSearch -F 25.0 -seed 1
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    1120
Attributes:   15
              pressao_atm_max
              pressao_atm_min
              chuva_mm
              temp_ar_max
              temp_ar_min
              umid_relat_max
              umid_relat_min
              rad_solar_total
              temp_max
              temp_min
              temp_ar_media
              umid_relat_media
              vento_veloc
              rad_solar_media
              et0
Evaluation mode:10-fold cross-validation



=== Attribute selection 10 fold cross-validation seed: 1 ===

number of folds (%)  attribute
           0(  0 %)    1 pressao_atm_max
          10(100 %)    2 pressao_atm_min
           0(  0 %)    3 chuva_mm
           0(  0 %)    4 temp_ar_max
```

```
 0(  0 %)    5 temp_ar_min
 0(  0 %)    6 umid_relat_max
 0(  0 %)    7 umid_relat_min
 0(  0 %)    8 rad_solar_total
10(100 %)    9 temp_max
 0(  0 %)   10 temp_min
10(100 %)   11 temp_ar_media
 0(  0 %)   12 umid_relat_media
10(100 %)   13 vento_veloc
10(100 %)   14 rad_solar_media
```

APÊNDICE B – MODELOS PREDITIVOS

Neste apêndice, serão apresentados os modelos preditivos criados a partir da execução dos algoritmos *M5'* e regressão linear, implementados pela ferramenta WEKA.

Modelos criados pelo *M5'*

O *M5'* é um algoritmo que utiliza árvores de decisão com modelos lineares em suas folhas para criar seus modelos. A seguir, é mostrado o modelo de predição da ET_0 criado por esse algoritmo, sem a utilização de seleção de atributos.

```
=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    1120
Attributes:   15
              pressao_atm_max
              pressao_atm_min
              chuva_mm
              temp_ar_max
              temp_ar_min
              umid_relat_max
              umid_relat_min
              rad_solar_total
              temp_max
              temp_min
              temp_ar_media
              umid_relat_media
              vento_veloc
              rad_solar_media
              et0
Test mode:split 70.0% train, remainder test

=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

umid_relat_min <= 62.089 :
|   rad_solar_media <= 6.122 :
|   |   vento_veloc <= 0.928 :
|   |   |   vento_veloc <= 0.196 : LM1 (11/0.294%)
|   |   |   vento_veloc >  0.196 : LM2 (30/4.717%)
```

```
|   |   vento_veloc >  0.928 :
|   |   |   temp_ar_min <= 28.789 :
|   |   |   |   temp_ar_max <= 28.815 : LM3 (34/2.654%)
|   |   |   |   temp_ar_max >  28.815 :
|   |   |   |   |   vento_veloc <= 2.395 : LM4 (30/2.811%)
|   |   |   |   |   vento_veloc >  2.395 : LM5 (16/2.468%)
|   |   |   temp_ar_min >  28.789 :
|   |   |   |   vento_veloc <= 1.433 :
|   |   |   |   |   temp_ar_min <= 31.36 : LM6 (10/5.641%)
|   |   |   |   |   temp_ar_min >  31.36 : LM7 (28/4%)
|   |   |   |   vento_veloc >  1.433 :
|   |   |   |   |   temp_ar_min <= 30.455 : LM8 (19/3.321%)
|   |   |   |   |   temp_ar_min >  30.455 : LM9 (25/7.658%)
|   rad_solar_media >  6.122 :
|   |   umid_relat_max <= 55.536 :
|   |   |   vento_veloc <= 1.844 :
|   |   |   |   vento_veloc <= 1.356 : LM10 (29/1.586%)
|   |   |   |   vento_veloc >  1.356 : LM11 (91/1.862%)
|   |   |   vento_veloc >  1.844 :
|   |   |   |   umid_relat_min <= 37.489 :
|   |   |   |   |   vento_veloc <= 2.369 : LM12 (84/1.363%)
|   |   |   |   |   vento_veloc >  2.369 : LM13 (50/2.168%)
|   |   |   |   umid_relat_min >  37.489 : LM14 (80/3.192%)
|   |   umid_relat_max >  55.536 :
|   |   |   vento_veloc <= 1.875 : LM15 (38/3.014%)
|   |   |   vento_veloc >  1.875 : LM16 (63/6.014%)
umid_relat_min >  62.089 :
|   temp_ar_max <= 25.109 :
|   |   vento_veloc <= 0.44 : LM17 (76/3.804%)
|   |   vento_veloc >  0.44 :
|   |   |   vento_veloc <= 1.571 : LM18 (149/3.46%)
|   |   |   vento_veloc >  1.571 : LM19 (109/4.107%)
|   temp_ar_max >  25.109 :
|   |   vento_veloc <= 1.754 : LM20 (60/5.452%)
|   |   vento_veloc >  1.754 :
|   |   |   rad_solar_total <= 2501.705 :
|   |   |   |   umid_relat_max <= 76.802 :
|   |   |   |   |   temp_ar_max <= 26.983 : LM21 (22/4.908%)
|   |   |   |   |   temp_ar_max >  26.983 : LM22 (13/3.862%)
|   |   |   |   umid_relat_max >  76.802 : LM23 (25/3.566%)
|   |   |   rad_solar_total >  2501.705 : LM24 (28/1.509%)

LM num: 1
et0 =
0.0001 * pressao_atm_max
+ 0.0003 * temp_ar_max
+ 0.0029 * temp_ar_min
```

```
+ 0.0001 * umid_relat_max
+ 0.0002 * umid_relat_min
- 0.0005 * temp_min
- 0.0003 * temp_ar_media
- 0.0008 * umid_relat_media
+ 0.0621 * vento_veloc
+ 0.0004 * rad_solar_media
- 0.1013

LM num: 2
et0 =
0.0001 * pressao_atm_max
+ 0.0011 * temp_ar_max
+ 0.0029 * temp_ar_min
+ 0 * umid_relat_max
+ 0.0002 * umid_relat_min
- 0.0005 * temp_min
- 0.0003 * temp_ar_media
- 0.0008 * umid_relat_media
+ 0.0599 * vento_veloc
+ 0.0004 * rad_solar_media
- 0.1242

LM num: 3
et0 =
0 * pressao_atm_max
- 0.0002 * temp_ar_max
+ 0.0024 * temp_ar_min
+ 0 * umid_relat_max
+ 0.0002 * umid_relat_min
- 0.0005 * temp_min
+ 0.0004 * temp_ar_media
- 0.0016 * umid_relat_media
+ 0.0281 * vento_veloc
+ 0.0004 * rad_solar_media
+ 0.0033

LM num: 4
et0 =
0 * pressao_atm_max
- 0.0004 * temp_ar_max
+ 0.002 * temp_ar_min
+ 0 * umid_relat_max
+ 0.0002 * umid_relat_min
- 0.0005 * temp_min
+ 0.0006 * temp_ar_media
- 0.0017 * umid_relat_media
```

```
+ 0.0313 * vento_veloc
+ 0.0004 * rad_solar_media
+ 0.0171

LM num: 5
et0 =
0 * pressao_atm_max
- 0.0004 * temp_ar_max
+ 0.0021 * temp_ar_min
+ 0 * umid_relat_max
+ 0.0002 * umid_relat_min
- 0.0005 * temp_min
+ 0.0006 * temp_ar_media
- 0.0018 * umid_relat_media
+ 0.0285 * vento_veloc
+ 0.0004 * rad_solar_media
+ 0.024

LM num: 6
et0 =
0 * pressao_atm_max
+ 0.0006 * temp_ar_max
+ 0.0022 * temp_ar_min
- 0.0002 * umid_relat_max
+ 0 * umid_relat_min
- 0.0005 * temp_min
+ 0.0015 * temp_ar_media
- 0.0011 * umid_relat_media
+ 0.0528 * vento_veloc
+ 0.0004 * rad_solar_media
- 0.0841

LM num: 7
et0 =
0 * pressao_atm_max
+ 0.0007 * temp_ar_max
+ 0.0022 * temp_ar_min
- 0.0002 * umid_relat_max
- 0 * umid_relat_min
- 0.0005 * temp_min
+ 0.0015 * temp_ar_media
- 0.0011 * umid_relat_media
+ 0.0615 * vento_veloc
+ 0.0004 * rad_solar_media
- 0.0932

LM num: 8
```

```
et0 =
0.0001 * pressao_atm_max
- 0.0008 * temp_ar_max
+ 0.004 * temp_ar_min
+ 0 * umid_relat_max
+ 0.0002 * umid_relat_min
- 0.0005 * temp_min
+ 0.0015 * temp_ar_media
- 0.0023 * umid_relat_media
+ 0.0462 * vento_veloc
+ 0.0004 * rad_solar_media
- 0.087

LM num: 9
et0 =
0 * pressao_atm_max
- 0.0035 * temp_ar_max
+ 0.0037 * temp_ar_min
+ 0 * umid_relat_max
+ 0.0002 * umid_relat_min
- 0.0005 * temp_min
+ 0.0072 * temp_ar_media
- 0.0016 * umid_relat_media
+ 0.0555 * vento_veloc
+ 0.0004 * rad_solar_media
- 0.1793

LM num: 10
et0 =
0 * pressao_atm_max
+ 0.0004 * temp_ar_max
+ 0.0013 * temp_ar_min
+ 0.0001 * umid_relat_max
- 0.0008 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0002 * temp_max
- 0.0001 * temp_min
- 0.0002 * temp_ar_media
- 0.0011 * umid_relat_media
+ 0.067 * vento_veloc
+ 0.0002 * rad_solar_media
+ 0.0152

LM num: 11
et0 =
0 * pressao_atm_max
+ 0.0008 * temp_ar_max
```

```
+ 0.001 * temp_ar_min
+ 0.0001 * umid_relat_max
- 0 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0002 * temp_max
- 0.0001 * temp_min
- 0.0002 * temp_ar_media
- 0.0021 * umid_relat_media
+ 0.0649 * vento_veloc
+ 0.0002 * rad_solar_media
+ 0.0243

LM num: 12
et0 =
0 * pressao_atm_max
+ 0.0009 * temp_ar_max
+ 0.0001 * temp_ar_min
- 0.0001 * umid_relat_max
- 0 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0009 * temp_max
+ 0.0004 * temp_min
+ 0.0001 * temp_ar_media
- 0.0028 * umid_relat_media
+ 0.0634 * vento_veloc
+ 0.0002 * rad_solar_media
+ 0.0445

LM num: 13
et0 =
0 * pressao_atm_max
+ 0.0004 * temp_ar_max
+ 0.0001 * temp_ar_min
- 0.0004 * umid_relat_max
- 0 * umid_relat_min
+ 0 * rad_solar_total
- 0.0007 * temp_max
- 0.0016 * temp_min
+ 0.0041 * temp_ar_media
- 0.0011 * umid_relat_media
+ 0.0629 * vento_veloc
+ 0.0002 * rad_solar_media
- 0.0562

LM num: 14
et0 =
0 * pressao_atm_max
```

```
+ 0.0006 * temp_ar_max
+ 0.0013 * temp_ar_min
+ 0.0006 * umid_relat_max
- 0.0001 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0003 * temp_max
- 0.0001 * temp_min
- 0.0001 * temp_ar_media
- 0.0032 * umid_relat_media
+ 0.0501 * vento_veloc
+ 0.0002 * rad_solar_media
+ 0.0765

LM num: 15
et0 =
0 * pressao_atm_max
- 0.0026 * pressao_atm_min
+ 0.0027 * temp_ar_max
+ 0.0001 * temp_ar_min
- 0.0008 * umid_relat_max
+ 0.0006 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0002 * temp_max
- 0.0016 * temp_min
- 0.0012 * temp_ar_media
- 0.0011 * umid_relat_media
+ 0.0385 * vento_veloc
+ 0.0002 * rad_solar_media
+ 1.6061

LM num: 16
et0 =
0 * pressao_atm_max
- 0.0162 * pressao_atm_min
+ 0.0037 * temp_ar_max
+ 0.0001 * temp_ar_min
+ 0.0004 * umid_relat_max
+ 0.0011 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0002 * temp_max
- 0.0012 * temp_min
- 0.0021 * temp_ar_media
- 0.0034 * umid_relat_media
+ 0.0372 * vento_veloc
+ 0.0002 * rad_solar_media
+ 9.8092
```

```
LM num: 17
et0 =
0 * pressao_atm_max
+ 0.0007 * temp_ar_max
+ 0 * temp_ar_min
- 0 * umid_relat_max
- 0 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0001 * temp_max
- 0.0003 * temp_min
- 0.0005 * temp_ar_media
- 0.0001 * umid_relat_media
+ 0.0032 * vento_veloc
+ 0.0002 * rad_solar_media
+ 0.0116

LM num: 18
et0 =
0 * pressao_atm_max
+ 0.0007 * temp_ar_max
+ 0 * temp_ar_min
- 0 * umid_relat_max
- 0.0005 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0001 * temp_max
- 0.0003 * temp_min
- 0.0005 * temp_ar_media
- 0.0001 * umid_relat_media
+ 0.012 * vento_veloc
+ 0.0002 * rad_solar_media
+ 0.0586

LM num: 19
et0 =
0 * pressao_atm_max
+ 0.0007 * temp_ar_max
+ 0 * temp_ar_min
- 0 * umid_relat_max
- 0.0006 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0001 * temp_max
- 0.0003 * temp_min
- 0.0005 * temp_ar_media
- 0.0005 * umid_relat_media
+ 0.0111 * vento_veloc
+ 0.0002 * rad_solar_media
+ 0.0775
```

```
LM num: 20
et0 =
0 * pressao_atm_max
+ 0.0019 * temp_ar_max
- 0.0001 * temp_ar_min
+ 0 * umid_relat_max
- 0.0007 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0014 * temp_max
- 0.0003 * temp_min
- 0.001 * temp_ar_media
- 0.0006 * umid_relat_media
+ 0.016 * vento_veloc
+ 0.0002 * rad_solar_media
+ 0.0683

LM num: 21
et0 =
0.0028 * pressao_atm_max
+ 0.0028 * temp_ar_max
- 0.0022 * temp_ar_min
- 0.0006 * umid_relat_max
+ 0.0001 * umid_relat_min
+ 0 * rad_solar_total
- 0.0008 * temp_max
- 0.0003 * temp_min
- 0.0023 * temp_ar_media
- 0.0007 * umid_relat_media
+ 0.0149 * vento_veloc
+ 0.0002 * rad_solar_media
- 1.5778

LM num: 22
et0 =
-0.0001 * pressao_atm_max
+ 0.003 * temp_ar_max
- 0.0022 * temp_ar_min
- 0.0006 * umid_relat_max
+ 0.0001 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0003 * temp_max
- 0.0003 * temp_min
- 0.0023 * temp_ar_media
- 0.0007 * umid_relat_media
+ 0.0117 * vento_veloc
+ 0.0002 * rad_solar_media
```

```
+ 0.1625

LM num: 23
et0 =
-0.0001 * pressao_atm_max
+ 0.0023 * temp_ar_max
- 0.0025 * temp_ar_min
- 0.0006 * umid_relat_max
- 0.0003 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0021 * temp_max
- 0.0003 * temp_min
- 0.0023 * temp_ar_media
- 0.0007 * umid_relat_media
+ 0.0095 * vento_veloc
+ 0.0002 * rad_solar_media
+ 0.1918

LM num: 24
et0 =
-0.0001 * pressao_atm_max
- 0.0029 * pressao_atm_min
+ 0.0017 * temp_ar_max
- 0.0013 * temp_ar_min
- 0.0004 * umid_relat_max
+ 0.0002 * umid_relat_min
+ 0 * rad_solar_total
+ 0.0028 * temp_max
- 0.0003 * temp_min
- 0.0023 * temp_ar_media
- 0.0021 * umid_relat_media
+ 0.0192 * vento_veloc
+ 0.0002 * rad_solar_media
+ 1.9542

Number of Rules : 24

Time taken to build model: 1.01 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.9969
Mean absolute error                      0.0025
Root mean squared error                  0.0044
Relative absolute error                  5.1208 %
Root relative squared error              7.8692 %
```

```
Total Number of Instances              336
```

A seguir, o modelo criado utilizando o *M5'*, a partir do conjunto de atributos selecionados pelo *CFS* + *BestFirst*:

```
=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1-2,4,7-11,13-weka.
filters.unsupervised.attribute.Remove-R3-4
Instances:    1120
Attributes:   5
              pressao_atm_min
              temp_ar_max
              vento_veloc
              rad_solar_media
              et0
Test mode:split 70.0% train, remainder test

=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

temp_ar_max <= 27.664 :
|   temp_ar_max <= 25.109 :
|   |   vento_veloc <= 0.44 : LM1 (76/3.804%)
|   |   vento_veloc >  0.44 :
|   |   |   vento_veloc <= 1.571 :
|   |   |   |   vento_veloc <= 0.886 :
|   |   |   |   |   temp_ar_max <= 24.198 : LM2 (34/4.485%)
|   |   |   |   |   temp_ar_max >  24.198 : LM3 (15/5.652%)
|   |   |   |   vento_veloc >  0.886 :
|   |   |   |   |   temp_ar_max <= 24.368 : LM4 (68/6.299%)
|   |   |   |   |   temp_ar_max >  24.368 :
|   |   |   |   |   |   temp_ar_max <= 24.579 :
|   |   |   |   |   |   |   rad_solar_media <= 5.747 : LM5 (9/3.66%)
|   |   |   |   |   |   |   rad_solar_media >  5.747 : LM6 (5/2.061%)
|   |   |   |   |   |   temp_ar_max >  24.579 : LM7 (19/4.397%)
|   |   |   vento_veloc >  1.571 :
|   |   |   |   rad_solar_media <= 6.183 : LM8 (83/5.2%)
|   |   |   |   rad_solar_media >  6.183 :
|   |   |   |   |   rad_solar_media <= 6.43 : LM9 (21/7.42%)
|   |   |   |   |   rad_solar_media >  6.43 : LM10 (6/22.215%)
|   temp_ar_max >  25.109 :
```

```
|   |   vento_veloc <= 1.582 : LM11 (51/9.494%)
|   |   vento_veloc >  1.582 :
|   |   |   rad_solar_media <= 6.949 : LM12 (79/10.308%)
|   |   |   rad_solar_media >  6.949 : LM13 (34/13.466%)
temp_ar_max >  27.664 :
|   rad_solar_media <= 6.122 :
|   |   vento_veloc <= 0.928 :
|   |   |   vento_veloc <= 0.196 : LM14 (11/0.294%)
|   |   |   vento_veloc >  0.196 :
|   |   |   |   vento_veloc <= 0.506 : LM15 (12/2.769%)
|   |   |   |   vento_veloc >  0.506 : LM16 (17/4.696%)
|   |   vento_veloc >  0.928 :
|   |   |   temp_ar_max <= 32.963 : LM17 (93/10.983%)
|   |   |   temp_ar_max >  32.963 : LM18 (61/24.676%)
|   rad_solar_media >  6.122 :
|   |   temp_ar_max <= 31.693 : LM19 (115/18.441%)
|   |   temp_ar_max >  31.693 :
|   |   |   vento_veloc <= 1.92 : LM20 (146/10.984%)
|   |   |   vento_veloc >  1.92 : LM21 (165/16.194%)

LM num: 1
et0 =
0.0007 * pressao_atm_min
+ 0.0008 * temp_ar_max
+ 0.0025 * vento_veloc
+ 0.0006 * rad_solar_media
- 0.4661

LM num: 2
et0 =
0.0007 * pressao_atm_min
+ 0.0018 * temp_ar_max
+ 0.0035 * vento_veloc
+ 0.0009 * rad_solar_media
- 0.4866

LM num: 3
et0 =
0.0007 * pressao_atm_min
+ 0.0021 * temp_ar_max
+ 0.0087 * vento_veloc
+ 0.0009 * rad_solar_media
- 0.4953

LM num: 4
et0 =
0.0007 * pressao_atm_min
```

```
+ 0.0015 * temp_ar_max
+ 0.0071 * vento_veloc
+ 0.001 * rad_solar_media
- 0.4821

LM num: 5
et0 =
0.0007 * pressao_atm_min
+ 0.0018 * temp_ar_max
+ 0.0039 * vento_veloc
+ 0.003 * rad_solar_media
- 0.4953

LM num: 6
et0 =
0.0062 * pressao_atm_min
+ 0.0018 * temp_ar_max
+ 0.0039 * vento_veloc
+ 0.0033 * rad_solar_media
- 3.7829

LM num: 7
et0 =
0.0007 * pressao_atm_min
+ 0.0018 * temp_ar_max
+ 0.0039 * vento_veloc
+ 0.0018 * rad_solar_media
- 0.4874

LM num: 8
et0 =
0.0022 * pressao_atm_min
+ 0.002 * temp_ar_max
+ 0.0101 * vento_veloc
+ 0.0015 * rad_solar_media
- 1.3445

LM num: 9
et0 =
0.0112 * pressao_atm_min
+ 0.0009 * temp_ar_max
+ 0.017 * vento_veloc
- 0.0015 * rad_solar_media
- 6.7621

LM num: 10
et0 =
```

```
0.0166 * pressao_atm_min
+ 0.0009 * temp_ar_max
+ 0.0126 * vento_veloc
+ 0.0083 * rad_solar_media
- 9.9978

LM num: 11
et0 =
0.0061 * temp_ar_max
+ 0.0158 * vento_veloc
+ 0.0033 * rad_solar_media
- 0.1679

LM num: 12
et0 =
0.0026 * pressao_atm_min
+ 0.0079 * temp_ar_max
+ 0.0128 * vento_veloc
- 0.0022 * rad_solar_media
- 1.7171

LM num: 13
et0 =
0.0363 * pressao_atm_min
+ 0.0098 * temp_ar_max
+ 0.0052 * vento_veloc
+ 0.0022 * rad_solar_media
- 21.984

LM num: 14
et0 =
-0.0091 * pressao_atm_min
+ 0.005 * temp_ar_max
+ 0.059 * vento_veloc
+ 0.0016 * rad_solar_media
+ 5.2986

LM num: 15
et0 =
-0.0091 * pressao_atm_min
+ 0.0059 * temp_ar_max
+ 0.0561 * vento_veloc
+ 0.0021 * rad_solar_media
+ 5.2694

LM num: 16
et0 =
```

```
-0.0091 * pressao_atm_min
+ 0.0065 * temp_ar_max
+ 0.0592 * vento_veloc
+ 0.0019 * rad_solar_media
+ 5.2473

LM num: 17
et0 =
-0.003 * pressao_atm_min
+ 0.0079 * temp_ar_max
+ 0.0246 * vento_veloc
- 0.0092 * rad_solar_media
+ 1.6124

LM num: 18
et0 =
-0.003 * pressao_atm_min
+ 0.0124 * temp_ar_max
+ 0.0482 * vento_veloc
+ 0.0003 * rad_solar_media
+ 1.3819

LM num: 19
et0 =
0.0021 * pressao_atm_min
+ 0.0074 * temp_ar_max
+ 0.0393 * vento_veloc
+ 0.0048 * rad_solar_media
- 1.4922

LM num: 20
et0 =
0.0123 * pressao_atm_min
+ 0.0085 * temp_ar_max
+ 0.0716 * vento_veloc
- 0.0003 * rad_solar_media
- 7.6557

LM num: 21
et0 =
0.0021 * pressao_atm_min
+ 0.0122 * temp_ar_max
+ 0.0568 * vento_veloc
+ 0.0008 * rad_solar_media
- 1.6481

Number of Rules : 21
```

```
Time taken to build model: 0.39 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.988
Mean absolute error                      0.0061
Root mean squared error                  0.0086
Relative absolute error                 12.5002 %
Root relative squared error             15.4738 %
Total Number of Instances              336
```

A seguir, o modelo criado utilizando o *M5'*, a partir do conjunto de atributos selecionados pelo *CFS* + *ExhaustiveSearch*:

```
=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1-weka.
filters.unsupervised.attribute.Remove-R1,3-4,6-10,12
Instances:    1120
Attributes:   6
              pressao_atm_min
              temp_ar_min
              temp_ar_media
              vento_veloc
              rad_solar_media
              et0
Test mode:split 70.0% train, remainder test

=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

temp_ar_media <= 26.4 :
|   vento_veloc <= 0.784 : LM1 (121/3.449%)
|   vento_veloc >  0.784 : LM2 (366/10.519%)
temp_ar_media >  26.4 :
|   rad_solar_media <= 6.122 :
|   |   temp_ar_media <= 29.416 : LM3 (79/8.961%)
|   |   temp_ar_media >  29.416 :
|   |   |   vento_veloc <= 0.822 :
|   |   |   |   vento_veloc <= 0.196 : LM4 (11/0.294%)
```

```
|   |   |   |   vento_veloc >  0.196 : LM5 (24/3.695%)
|   |   |   vento_veloc >  0.822 :
|   |   |   |   vento_veloc <= 1.433 :
|   |   |   |   |   temp_ar_media <= 32.715 : LM6 (14/5.9%)
|   |   |   |   |   temp_ar_media >  32.715 : LM7 (29/7.142%)
|   |   |   |   vento_veloc >  1.433 :
|   |   |   |   |   temp_ar_min <= 30.455 : LM8 (25/8.312%)
|   |   |   |   |   temp_ar_min >  30.455 : LM9 (25/7.216%)
|   rad_solar_media >  6.122 :
|   |   temp_ar_min <= 28.484 : LM10 (90/18.758%)
|   |   temp_ar_min >  28.484 :
|   |   |   vento_veloc <= 2.132 : LM11 (204/10.248%)
|   |   |   vento_veloc >  2.132 : LM12 (132/13.679%)

LM num: 1
et0 =
-0.0053 * temp_ar_min
+ 0.0067 * temp_ar_media
+ 0.0151 * vento_veloc
+ 0.001 * rad_solar_media
- 0.0364

LM num: 2
et0 =
-0.0178 * temp_ar_min
+ 0.0227 * temp_ar_media
+ 0.013 * vento_veloc
+ 0.002 * rad_solar_media
- 0.13

LM num: 3
et0 =
0.0017 * temp_ar_min
+ 0.0065 * temp_ar_media
+ 0.0248 * vento_veloc
- 0.0107 * rad_solar_media
- 0.1543

LM num: 4
et0 =
0.0028 * temp_ar_min
+ 0.0033 * temp_ar_media
+ 0.0578 * vento_veloc
+ 0.0007 * rad_solar_media
- 0.1831

LM num: 5
```

```
et0 =
0.0034 * temp_ar_min
+ 0.0038 * temp_ar_media
+ 0.0556 * vento_veloc
+ 0.0014 * rad_solar_media
- 0.2209

LM num: 6
et0 =
0.0018 * temp_ar_min
+ 0.0085 * temp_ar_media
+ 0.0502 * vento_veloc
- 0.0002 * rad_solar_media
- 0.3117

LM num: 7
et0 =
0.0029 * temp_ar_min
+ 0.0058 * temp_ar_media
+ 0.0571 * vento_veloc
- 0.0002 * rad_solar_media
- 0.2629

LM num: 8
et0 =
0.0018 * temp_ar_min
+ 0.0102 * temp_ar_media
+ 0.0384 * vento_veloc
- 0.0022 * rad_solar_media
- 0.3404

LM num: 9
et0 =
0.0027 * temp_ar_min
+ 0.0094 * temp_ar_media
+ 0.0499 * vento_veloc
+ 0.0028 * rad_solar_media
- 0.3825

LM num: 10
et0 =
0.0029 * pressao_atm_min
+ 0.0008 * temp_ar_min
+ 0.0074 * temp_ar_media
+ 0.0359 * vento_veloc
+ 0.0049 * rad_solar_media
- 2.027
```

```
LM num: 11
et0 =
0.0135 * pressao_atm_min
+ 0.0041 * temp_ar_min
+ 0.0044 * temp_ar_media
+ 0.0686 * vento_veloc
+ 0.0004 * rad_solar_media
- 8.3884

LM num: 12
et0 =
0.0023 * pressao_atm_min
+ 0.0073 * temp_ar_min
+ 0.0044 * temp_ar_media
+ 0.0565 * vento_veloc
+ 0.0016 * rad_solar_media
- 1.756

Number of Rules : 12

Time taken to build model: 0.53 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.9897
Mean absolute error                      0.0054
Root mean squared error                  0.008
Relative absolute error                 11.1767 %
Root relative squared error             14.4098 %
Total Number of Instances              336
```

A seguir, o modelo criado utilizando o *M5'*, a partir do conjunto de atributos selecionados pelo *CFS* + *GeneticSearch*:

```
=== Run information ===

Scheme:weka.classifiers.trees.M5P -M 4.0
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1-2,4,7-11,13
Instances:    1120
Attributes:   7
              pressao_atm_min
              temp_ar_max
              temp_ar_min
              temp_ar_media
```

```
              vento_veloc
              rad_solar_media
              et0
Test mode:split 70.0% train, remainder test

=== Classifier model (full training set) ===

M5 pruned model tree:
(using smoothed linear models)

temp_ar_max <= 27.664 :
|   temp_ar_max <= 25.109 :
|   |   vento_veloc <= 0.44 : LM1 (76/3.804%)
|   |   vento_veloc >  0.44 :
|   |   |   vento_veloc <= 1.571 : LM2 (150/5.692%)
|   |   |   vento_veloc >  1.571 : LM3 (110/7.053%)
|   temp_ar_max >  25.109 :
|   |   vento_veloc <= 1.582 : LM4 (51/9.494%)
|   |   vento_veloc >  1.582 :
|   |   |   rad_solar_media <= 6.949 : LM5 (79/10.207%)
|   |   |   rad_solar_media >  6.949 : LM6 (34/13.143%)
temp_ar_max >  27.664 :
|   rad_solar_media <= 6.122 :
|   |   vento_veloc <= 0.928 :
|   |   |   vento_veloc <= 0.196 : LM7 (11/0.294%)
|   |   |   vento_veloc >  0.196 :
|   |   |   |   vento_veloc <= 0.506 : LM8 (12/2.769%)
|   |   |   |   vento_veloc >  0.506 : LM9 (17/3.077%)
|   |   vento_veloc >  0.928 :
|   |   |   temp_ar_min <= 28.789 : LM10 (72/8.907%)
|   |   |   temp_ar_min >  28.789 :
|   |   |   |   vento_veloc <= 1.433 :
|   |   |   |   |   temp_ar_min <= 31.36 : LM11 (10/5.998%)
|   |   |   |   |   temp_ar_min >  31.36 : LM12 (28/6.857%)
|   |   |   |   vento_veloc >  1.433 :
|   |   |   |   |   temp_ar_min <= 30.455 : LM13 (19/9.68%)
|   |   |   |   |   temp_ar_min >  30.455 : LM14 (25/7.658%)
|   rad_solar_media >  6.122 :
|   |   temp_ar_min <= 28.484 : LM15 (90/17.745%)
|   |   temp_ar_min >  28.484 :
|   |   |   vento_veloc <= 2.132 : LM16 (204/9.949%)
|   |   |   vento_veloc >  2.132 : LM17 (132/13.574%)

LM num: 1
et0 =
0.0012 * temp_ar_max
- 0.001 * temp_ar_min
```

```
+ 0.0004 * temp_ar_media
+ 0.0029 * vento_veloc
+ 0.0003 * rad_solar_media
- 0.0145

LM num: 2
et0 =
0.0006 * pressao_atm_min
+ 0.0073 * temp_ar_max
- 0.0049 * temp_ar_min
+ 0 * temp_ar_media
+ 0.0105 * vento_veloc
+ 0.0002 * rad_solar_media
- 0.4154

LM num: 3
et0 =
0.0075 * pressao_atm_min
+ 0.0122 * temp_ar_max
- 0.0142 * temp_ar_min
+ 0.0047 * temp_ar_media
+ 0.0112 * vento_veloc
+ 0.0002 * rad_solar_media
- 4.587

LM num: 4
et0 =
0.0043 * temp_ar_max
- 0.0035 * temp_ar_min
+ 0.0051 * temp_ar_media
+ 0.0162 * vento_veloc
+ 0.0028 * rad_solar_media
- 0.1603

LM num: 5
et0 =
0.0111 * temp_ar_max
+ 0.0015 * temp_ar_min
- 0.0051 * temp_ar_media
+ 0.0129 * vento_veloc
- 0.0029 * rad_solar_media
- 0.1646

LM num: 6
et0 =
0.0006 * temp_ar_max
- 0.006 * temp_ar_min
```

```
+ 0.0153 * temp_ar_media
+ 0.0063 * vento_veloc
+ 0.0015 * rad_solar_media
- 0.2234

LM num: 7
et0 =
-0.0015 * temp_ar_max
+ 0.0018 * temp_ar_min
+ 0.0051 * temp_ar_media
+ 0.0597 * vento_veloc
- 0.0012 * rad_solar_media
- 0.1508

LM num: 8
et0 =
-0.0011 * temp_ar_max
+ 0.0023 * temp_ar_min
+ 0.0051 * temp_ar_media
+ 0.0577 * vento_veloc
- 0.0012 * rad_solar_media
- 0.1798

LM num: 9
et0 =
-0.0009 * temp_ar_max
+ 0.0028 * temp_ar_min
+ 0.0051 * temp_ar_media
+ 0.0605 * vento_veloc
- 0.0012 * rad_solar_media
- 0.2036

LM num: 10
et0 =
-0.0007 * temp_ar_max
+ 0.0047 * temp_ar_min
+ 0.0062 * temp_ar_media
+ 0.0265 * vento_veloc
- 0.0075 * rad_solar_media
- 0.2283

LM num: 11
et0 =
-0.0008 * temp_ar_max
+ 0.0031 * temp_ar_min
+ 0.0069 * temp_ar_media
+ 0.0498 * vento_veloc
```

```
- 0.001 * rad_solar_media
- 0.2719

LM num: 12
et0 =
-0.0003 * temp_ar_max
+ 0.0027 * temp_ar_min
+ 0.0069 * temp_ar_media
+ 0.0578 * vento_veloc
- 0.001 * rad_solar_media
- 0.2772

LM num: 13
et0 =
-0.0024 * temp_ar_max
+ 0.0063 * temp_ar_min
+ 0.0088 * temp_ar_media
+ 0.0398 * vento_veloc
- 0.001 * rad_solar_media
- 0.364

LM num: 14
et0 =
-0.005 * temp_ar_max
+ 0.0036 * temp_ar_min
+ 0.0141 * temp_ar_media
+ 0.0511 * vento_veloc
- 0.001 * rad_solar_media
- 0.3767

LM num: 15
et0 =
0.0029 * pressao_atm_min
- 0.0006 * temp_ar_max
+ 0.0008 * temp_ar_min
+ 0.0074 * temp_ar_media
+ 0.0369 * vento_veloc
+ 0.0044 * rad_solar_media
- 1.9882

LM num: 16
et0 =
0.0151 * pressao_atm_min
- 0.0047 * temp_ar_max
+ 0.0018 * temp_ar_min
+ 0.0111 * temp_ar_media
+ 0.0684 * vento_veloc
```

```
+ 0.0003 * rad_solar_media
- 9.3162

LM num: 17
et0 =
0.0025 * pressao_atm_min
- 0.004 * temp_ar_max
+ 0.006 * temp_ar_min
+ 0.0094 * temp_ar_media
+ 0.0563 * vento_veloc
+ 0.0015 * rad_solar_media
- 1.8292

Number of Rules : 17

Time taken to build model: 0.4 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.9899
Mean absolute error                      0.0054
Root mean squared error                  0.0079
Relative absolute error                 11.1612 %
Root relative squared error             14.2651 %
Total Number of Instances              336
```

Modelos criados pela Regressão Linear

A regressão linear é uma técnica que visa mapear os dados em função de uma variável de predição real. A seguir, encontra-se o modelo gerado a partir da sua execução, sem a utilização de seleção de atributos.

```
=== Run information ===

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1
Instances:    1120
Attributes:   15
              pressao_atm_max
              pressao_atm_min
              chuva_mm
              temp_ar_max
              temp_ar_min
              umid_relat_max
```

```
              umid_relat_min
              rad_solar_total
              temp_max
              temp_min
              temp_ar_media
              umid_relat_media
              vento_veloc
              rad_solar_media
              et0
Test mode:split 70.0% train, remainder test

=== Classifier model (full training set) ===


Linear Regression Model

et0 =

      0.0098 * temp_ar_max +
     -0.0022 * temp_ar_min +
     -0.0016 * umid_relat_max +
      0.0019 * umid_relat_min +
     -0.0096 * temp_min +
      0.0009 * umid_relat_media +
      0.0314 * vento_veloc +
      0.0054 * rad_solar_media +
     -0.1273

Time taken to build model: 0.01 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.9659
Mean absolute error                      0.0109
Root mean squared error                  0.0144
Relative absolute error                 22.34   %
Root relative squared error             25.9766 %
Total Number of Instances              336
```

A seguir, o modelo criado utilizando a regressão linear, a partir do conjunto de atributos selecionados pelo *CFS + BestFirst*:

```
=== Run information ===
```

```
Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1-2,4,7-11,13-weka.
filters.unsupervised.attribute.Remove-R3-4
Instances:    1120
Attributes:   5
              pressao_atm_min
              temp_ar_max
              vento_veloc
              rad_solar_media
              et0
Test mode:split 70.0% train, remainder test

=== Classifier model (full training set) ===


Linear Regression Model

et0 =

      0.0077 * temp_ar_max +
      0.0246 * vento_veloc +
      0.0044 * rad_solar_media +
     -0.2242

Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.9511
Mean absolute error                      0.0137
Root mean squared error                  0.0171
Relative absolute error                 28.1222 %
Root relative squared error             30.9578 %
Total Number of Instances              336
```

A seguir, o modelo criado utilizando a regressão linear, a partir do conjunto de atributos selecionados pelo *CFS + ExhaustiveSearch*:

```
=== Run information ===

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1-weka.
filters.unsupervised.attribute.Remove-R1,3-4,6-10,12
Instances:    1120
```

```
Attributes:   6
              pressao_atm_min
              temp_ar_min
              temp_ar_media
              vento_veloc
              rad_solar_media
              et0
Test mode:split 70.0% train, remainder test

=== Classifier model (full training set) ===


Linear Regression Model

et0 =

      0.0081 * temp_ar_media +
      0.0249 * vento_veloc +
      0.0047 * rad_solar_media +
     -0.2314

Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.9536
Mean absolute error                      0.0134
Root mean squared error                  0.0167
Relative absolute error                 27.4272 %
Root relative squared error             30.1868 %
Total Number of Instances              336
```

A seguir, o modelo criado utilizando a regressão linear, a partir do conjunto de atributos selecionados pelo *CFS + GeneticSearch*:

```
=== Run information ===

Scheme:weka.classifiers.functions.LinearRegression -S 0 -R 1.0E-8
Relation:     2016total-semoutlier2-weka.filters.unsupervised.attribute.Remove-R1-2,4,7-11,13
Instances:    1120
Attributes:   7
              pressao_atm_min
              temp_ar_max
              temp_ar_min
              temp_ar_media
```

```
              vento_veloc
              rad_solar_media
              et0
Test mode:split 70.0% train, remainder test

=== Classifier model (full training set) ===


Linear Regression Model

et0 =

      0.0039 * temp_ar_max +
      0.0042 * temp_ar_min +
      0.0247 * vento_veloc +
      0.0048 * rad_solar_media +
     -0.2329

Time taken to build model: 0 seconds

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient                  0.9536
Mean absolute error                      0.0134
Root mean squared error                  0.0167
Relative absolute error                 27.4272 %
Root relative squared error             30.1868 %
Total Number of Instances              336
```

ANEXO A – ESTAÇÃO METEOROLÓGICA

Neste anexo serão apresentadas as características da estação meteorológica automática responsável pela medição das condições climáticas e pelo fornecimento dos dados utilizados nos experimentos deste trabalho.

Componentes da estação meteorológica

A estação meteorológica encontra-se instalada no campus da Universidade Federal do Ceará, na cidade de Quixadá e não recebe manutenções frequentes de especialistas. Abaixo, é possível observar os componentes da estação e seus respectivos modelos.

ITEM	MODELO	QUANTIDADE
Coletor de dados 900MHz-5 S.E.	CR206	01
Bateria 12VDC 7AH.	BAT 12V.7	01
Painel/Gerador/Módulo Solar 10W	KS10	01
Caixa plástica selada IP67 com suportes	CSB2916	01
Sensor de direção e velocidade do vento	03002-L	01
Sensor de temperatura e umidade relativa	SDI12 CSL	01
Cabo 4M	CS215-L12	01
Abrigo termométrico 6 pratos R.M. Young	41303-5A	01
Sensor de radiação solar global	Apogee	01
Cabo 4M	CS300-L12	01
Base de nivelamento CS300	CSB18356	01
Pluviômetro R.M. Young 0.2MM/Tip	52203	01
Tripe de metal alumínio 3M com braço superior para sensores	CSB-CM10	01
Suporte para sensor de radiação solar	CM225	01
Braço superior de alumínio com adaptador CM210	-	01
Suporte para sensor de vento em ângulo reto	CM220	01
Transmissor de dados *spread spectrum* 910 a 918 MHz	RF401	01
Antena 900MHz para RF401	ANTRF401	01