PONTIFÍCIA UNIVERSIDADE CATÓLICA DE GOIÁS

ESCOLA POLITÉCNICA E DE ARTES

ENGENHARIA DE CONTROLE E AUTOMAÇÃO

Trabalho Final de Curso II

Roberto Gerais Freitas

Erick Pereira de Souza

ANÁLISE DINÂMICA E DECISÃO DA EVAPOTRANSPIRAÇÃO DAS PLANTAS COM ÁRVORE DE DECISÕES

Trabalho Final de Curso II como parte dos requisitos para obtenção do título de bacharel em Engenharia de Controle e Automação apresentados à Pontifícia Universidade Católica de Goiás.

BANCA EXAMINADORA:

Orientador – Prof. Dr. Cássio Hideki Fujisawa – POLI-PUC Goiás.

Coorientador – Prof. Dr. Bruno Quirino de Oliveira – IFG-Trindade.

Prof. Dr. Antônio Marcos de Melo Medeiros – POLI-PUC Goiás.

Goiânia, 13 de dezembro de 2024.

Análise Dinâmica e Decisão da Evapotranspiração das Plantas com Árvore de Decisões

Roberto Gerais Freitas
Pontifícia Universidade Católica de Goiás
Goiânia, GO, Brasil

Erick Pereira de Souza
Pontifícia Universidade Católica de Goiás
Goiânia, GO, Brasil

Prof. Dr. Cássio Hideki Fujisawa
Pontifícia Universidade Católica de Goiás
Goiânia, GO, Brasil

Prof. Dr. Bruno Quirino de Oliveira
Pontifícia Universidade Católica de Goiás
Goiânia, GO, Brasil

Prof. Dr. Antônio Marcos de Melo Medeiros
Pontifícia Universidade Católica de Goiás
Goiânia, GO, Brasil

***Abstract*— This final paper presents the development of Artificial Intelligence (AI) with the objective of doing a dynamic calculation of the evapotranspiration of the plants. And for that the AI will consider the average temperature obtained from the database. The system works based on input analytics, next to a database, and calculated by a decision tree resulting in the evapotranspiration, allowing a more efficient decision of water or not the plants. The dynamic part comes from the fact that depending on the meteorological conditions, its ETO changes and the AI calculates this new one. The performance measurement was obtained based on a database of 2010 to 2022 and compared with the 2023 data to prove its accuracy. The results were presented and discussed.**

***Keywords* – Dynamic evapotranspiration, artificial intelligence, AI, ETO, prediction, analysis, decision tree, irrigation.**

Resumo – Este trabalho de conclusão de curso apresenta o desenvolvimento de uma Inteligência Artificial (IA) com o objetivo de calcular de forma dinâmica e eficiente a quantidade a evapotranspiração das plantas (ETO). E para isso a IA irá considerar a temperatura média obtida do banco de dados. O Sistema funciona a partir da análise das variáveis inseridas, junto a um banco de dados, e calculado a partir de uma árvore de decisões resultando na predição da evapotranspiração, permitindo que haja uma escolha mais eficiente de quando regar ou não as plantas. A parte dinâmica do sistema vem do fato que dependendo das condições meteorológicas, a ETO muda e a IA já calcula o novo valor. A apuração do desempenho foi obtida a partir de um banco de dados de 2010 até 2022 e comparados aos dados de 2023 para comprovar a acuracidade.

Palavras-chave – Evapotranspiração dinâmica, inteligência artificial, IA, ETO, predição, análise, árvore de decisão, irrigação.

I. Introdução

A água é um dos recursos mais importantes para o desenvolvimento das plantas, e uma boa gestão desse meio é essencial tanto financeiramente quanto sustentavelmente. E devido a todo avanço tecnológico é cada vez mais cotidiano a utilização de inteligência artificial (IA), uma maneira rápida, prática e compreensível de conhecimento, se adaptando e aprendendo muito bem quando educada.

Utilizando da evapotranspiração calculada a partir do Método de Hargreaves [1] , junto a dados obtido do site do INMET (Instituto Nacional de Meteorologia) para prever a quantidade a Evapotranspiração das plantas, onde tal informação é calculada de forma dinâmica por meio de uma árvore de decisões treinada por um banco de dados. Ou seja, caso haja alguma mudança nos dados obtidos, todo cálculo de decisão da evapotranspiração é feito novamente.

A evapotranspiração nada mais é que o processo combinado de evaporação do solo e transpiração das plantas, possuindo uma extrema relevância para a agricultura. [1] Segundo Pereira et al. (1997), a evapotranspiração é controlada pela disponibilidade de energia, pela demanda atmosférica e pelo suprimento de água do solo às plantas. Dessa maneira, é perceptível que a evapotranspiração é extremamente necessária para o ciclo da água, para o clima e para o crescimento das plantas.

O objetivo desse trabalho de conclusão de curso é desenvolver uma inteligência artificial capaz de prever a evapotranspiração das plantas de forma dinâmica, otimizando sistemas de agricultura, economizando água e maximizando rendimento das colheitas. Alimentando a inteligência artificial com dados diários de informações de cultivos do estado de Goiás, educando e treinando-a com dados de 2010 a 2023, para que possa prever os próximos resultados da região.

II. Fundamentação Teórica

O método conhecido como árvore de decisões é um dos mais simples conceitualmente e mais efetivo para construção de um algoritmo de aprendizagem. Este método começa com uma entrada de dados e uma situação a qual ela deve achar a melhor resposta, podendo ser denominada como classificação e regressão. Onde classificação é quando deseja-se prever a classe a qual o objeto pertence, como se um indivíduo está ou não doente, e regressão é quando deseja-se prever um valor numérico, como o preço de algum produto [2].

Uma árvore de decisões desenvolve uma sequência de cálculos com possíveis soluções para a situação apresentada, chamados de nós ou ramos, a partir de uma raiz, nó primário com os dados iniciais (Fig. 1). Onde cada nó novo tem seu peso, ou seja, quão útil é a informação gerada para chegar à solução final, chamada de folha, caso

tenha alguma relevância o cálculo é feito novamente, gerando mais nós, e caso não tenha relevância o nó é descartado. Portanto a estrutura dessa IA acaba se tornando com uma árvore, onde cada galho se ramifica em vários outros até chegar nas folhas.

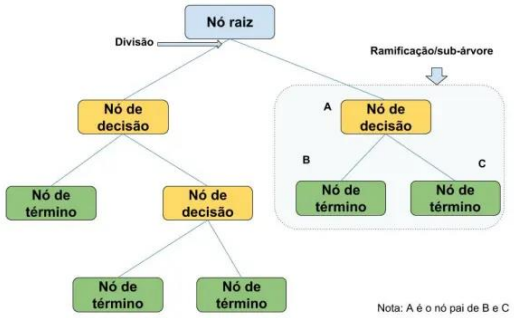


Figura 1 – Estrutura de funcionamento da árvore de decisões.

Toda árvore começa com a raiz, e para definir quais e como serão usados os dados nesse nó inicial (raiz) existem alguns métodos como o Índice Gini e a Entropia e Ganho. Esses métodos trabalham com o nível de impureza dos dados para definir qual o melhor caminho a ser seguido. E essa impureza nada mais é do que quão heterogêneos vão ser os dados utilizados.

O método Índice Gini mede a impureza baseando-se em uma fórmula e uma média ponderada dos resultados obtidos dela, e quanto mais perto de zero, melhor o resultado. Pode ser calculado pela equação (1) e (2):

$$Gini = 1 - \sum_{i=1}^{n} Pi^2 \quad (1)$$

$\boldsymbol{P_i}$ = probabilidade de i

$$Gt = \frac{\sum_{i=1}^{n}(Gi*Di)}{\sum_{i=1}^{n} Di} \quad (2)$$

Gt = Gini total

Gi = Gini individual

Di = quantidade de dados

E uma alternativa ao Índice Gini, com o mesmo objetivo, é a Entropia e Ganho. Essa alternativa tem sua origem na estatística, mas foi adaptada para poder calcular a pureza de dados. Essencialmente ela é similar ao Índice Gini, porém faz uso de log, como mostra na equação (3):

$$Entropia\ (S) = \sum_{i=1}^{c} - P_i \log_2 P_i \quad (3)$$

$\boldsymbol{P_i}$ = probabilidade de i

O ganho de informação aponta a medida de diminuição da entropia, quanto maior o ganho, melhor para fazer a divisão do conjunto de dados. A equação (4) mostra como é calculado [3]:

$$Ganho\ (S, A) = Entropia\ (S) - \sum_{v\ \in\ valores\ (A)} \frac{|S_v|}{|S|}\ Entopia\ (S_v) \quad (4)$$

S = conjunto total

Sv = subconjunto

A = atributo escolhido

Quando uma árvore fica com muitos ramos e folhas, tornando-se muito complexa, ou é alimentada com dados incorretos e imprecisos, pode ocorrer um fenômeno chamado *Overfitting* ou Superdatação. E ao ocorrer isso a árvore irá se ajustar excessivamente aos dados apresentados, por exemplo, uma árvore que prevê se uma sorveteria qual o sabor de sorvete mais vendido, se ocorrer o *Overfitting*, ela irá criar regras muito específicas como "se for um dia ensolarado com 20% de humildade no ar e estiver 30° Celsius e 10% de chance de chuva e o cliente for um pai com dois filhos pequenos, então o sorvete mais vendido será o sabor chocolate". Mesmo que essa regra seja algo esperado, ela se aplica a um grupo muito único de clientes, assim levando a previsões incorretas [2].

Um método para evitar esse fenômeno é a poda da árvore, que consiste em remover ramos insignificantes para a acuracidade do modelo, e é feita depois do processo de criação da árvore. A poda usa a validação cruzada para decidir se o ramo é descartável ou não, e ela funciona separando os dados em um grupo de teste e outro de treinamento como mostra a Fig. 2, ou seja, os resultados do treinamento serão comparados ao grupo teste, e esse processo se repetirá até que a árvore tenha um parâmetro para usar como métrica da poda. Exemplificando com a mesma situação da sorveteria, ao realizar a poda a mesma regra criada antes pode ser resumida a "se for um dia ensolarado com 30° Celsius e o cliente for um pai com dois filhos pequenos o sorvete mais vendido será o sabor chocolate"[2].

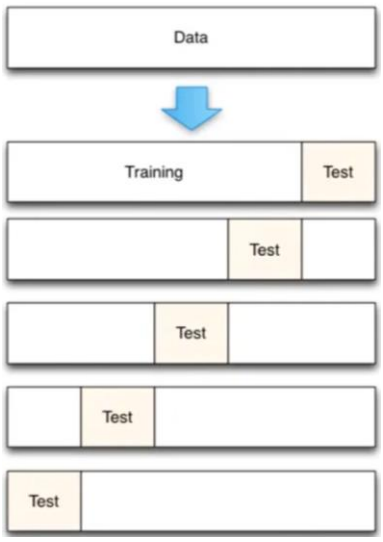


Figura 2 – Representação do grupo teste e grupo de treinamento

III. EXEMPLOS

Exemplificando o Índice Gini, mostre qual o melhor processo para saber se o empréstimo para a pessoa solteira que tem 6000 de renda será aprovado ou não.

Tabela I

Estado civil	Renda mensal	Aprovação empréstimo
Solteiro	1000	0
Casado	2000	1
Solteiro	3500	0
Casado	4000	1

Solteiro	5000	1
Solteiro	6000	?

Obs: D = dados totais, 0 = Não e 1 = Sim.

D	5
0	2
1	3

SIM NÃO

A decisão de qual ramo será o ideal para chegar no resultado é dada por:

Ramo 1:

É solteiro?

D	5
0	2
1	3

SIM NÃO

D	3		D	2
0	2		0	0
1	1		1	2

$$G1 = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] \quad G2 = 1 - \left[0 + \left(\frac{2}{2}\right)^2\right]$$

G1 = 0.44 G2 = 0

$$Gt = \frac{(0.44 * 3) + (0 * 2)}{3 + 2} = 0.264$$

Ramo 2:

Tem renda menor que 2000?

D	5
0	4
1	1

SIM NÃO

D	1		D	4
0	1		0	1
1	0		1	3

$$G1 = 1 - \left[\left(\frac{1}{1}\right)^2 + \left(\frac{0}{1}\right)^2\right] \quad G2 = 1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right]$$

G1 = 0 G2 = 0.375

$$Gt = \frac{(0 * 1) + (0.375 * 4)}{1 + 4} = 0.3$$

Ou seja, o melhor parâmetro para usar como raiz é o Ramo 1, pois ele foi mais próximo de zero (pureza maior).

Continuação do Ramo 1:

Como o grupo para os "Não solteiros" mostrou apenas resultados com valor de "1" não há mais necessidade de calcular um novo Índice Gini, logo esse ramo se tornou uma folha da árvore. Porém para os que caíram como "Solteiros" Ramo continua se desenvolvendo:

Ramo 3:

Tem renda menor que 4000?

D	3
0	2
1	1

SIM NÃO

D	2		D	1
0	2		0	0
1	0		1	1

$$G1 = 1 - \left[\left(\frac{2}{2}\right)^2 + \left(\frac{0}{2}\right)^2\right] \quad G2 = 1 - \left[\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right]$$

G1 = 0 G2 = 0

$$Gt = \frac{(0 * 2) + (0 * 1)}{2 + 1} = 0$$

Como o Índice Gini deu "0" os dados estão puros, ou seja, todos os ramos se tornaram folhas então, respondendo o questionamento feito, se a pessoa for solteira e tiver uma renda maior que 4000 ela vai cair no grupo onde há apenas dados "1", tendo seu empréstimo aprovado.

Exemplificando a Entropia e Ganho, mostre qual o melhor processo para saber se em um dia ensolarado, frio e vento fraco, irá ocorrer o jogo de tênis.

Tabela II

Tempo	Temperatura	Vento	Jogar tênis
Ensolarado	Quente	Fraco	Não
Ensolarado	Quente	Forte	Não
Nublado	Quente	Fraco	Sim
Chuvoso	Confortável	Fraco	Sim
Chuvoso	Frio	Fraco	Sim
Chuvoso	Frio	Forte	Não
Nublado	Frio	Forte	Sim
Ensolarado	Confortável	Fraco	Não
Ensolarado	Frio	Fraco	Sim
Chuvoso	Confortável	Fraco	Sim
Ensolarado	Frio	Fraco	?

Entropia = E

Cálculo do Ganho para os valores do "Vento" (Forte e Fraco):

$$E = -(\frac{6}{10}\log_2\left(\frac{6}{10}\right) + \frac{4}{10}\log_2(\frac{4}{10})) = 0.97$$

$$E(Forte) = -(\frac{1}{3}\log_2\left(\frac{1}{3}\right)$$

$$+\frac{2}{3}\log_2(\frac{2}{3})) = 0.92$$

$$E(Fraco) = -(\frac{5}{7}\log_2\left(\frac{5}{7}\right)$$

$$+\frac{2}{7}\log_2(\frac{2}{7})) = 0.86$$

$$Ganho(E, Vento) = 0.97 - \left(\frac{7}{10}\right) * 0.86 - \left(\frac{3}{10}\right) * 0.92 = 0.092$$

Cálculo do Ganho para os valores da "Temperatura" (Quente, Confortável e Frio):

$$E = -(\frac{6}{10}\log_2\left(\frac{6}{10}\right) + \frac{4}{10}\log_2(\frac{4}{10})) = 0.97$$

$$E(Quente) = -(\frac{1}{3}\log_2\left(\frac{1}{3}\right)$$

$$+\frac{2}{3}\log_2(\frac{2}{3})) = 0.92$$

$$E(Confortável) = -(\frac{2}{3}\log_2\left(\frac{2}{3}\right)$$

$$+\frac{1}{3}\log_2(\frac{1}{3})) = 0.92$$

$$E(Frio) = -(\frac{3}{4}\log_2\left(\frac{3}{4}\right)$$

$$+\frac{1}{4}\log_2(\frac{1}{4})) = 0.81$$

$$Ganho(E, Temperatura) =$$

$$0.97 - \left(\frac{3}{10}\right) * 0.92 - \left(\frac{3}{10}\right) * 0.92 - \left(\frac{4}{10}\right) * 0.81 = 0.094$$

Cálculo do Ganho para os valores do "Tempo" (Ensolarado, Nublado e Chuvoso):

$$E = -(\frac{6}{10}\log_2\left(\frac{6}{10}\right) + \frac{4}{10}\log_2(\frac{4}{10})) = 0.97$$

$$E(Ensolarado) = -(\frac{1}{4}\log_2\left(\frac{1}{4}\right)$$

$$+\frac{3}{4}\log_2(\frac{3}{4})) = 0.81$$

$$E(Nublado) = -(\frac{2}{2}\log_2\left(\frac{2}{2}\right)$$

$$+0 = 0$$

$$E(Chuvoso) = -(\frac{3}{4}\log_2\left(\frac{3}{4}\right)$$

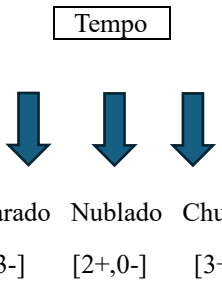
$$+\frac{1}{4}\log_2(\frac{1}{4})) = 0.81$$

$$Ganho(E, Tempo) =$$

$$0.97 - \left(\frac{4}{10}\right) * 0.81 - \left(\frac{2}{10}\right) * 0 - \left(\frac{4}{10}\right) * 0.81 = 0.97$$

Ou seja, o melhor parâmetro para usar como raiz é o "Tempo", pois ele foi o que obteve o maior ganho (pureza maior). Montando o diagrama da árvore fica:

Obs: + indica os "sim" da tabela e – indica os "não" da tabela



Como o grupo para o Tempo "Nublado" mostrou apenas resultados com valor de "+" não há mais necessidade de calcular uma nova Entropia e Ganho, logo esse ramo se tornou uma folha da árvore. Porém para o grupo com o Tempo "Ensolarado" e "Chuvoso" é preciso outro cálculo para achar a resposta. Como o problema está perguntando para o Tempo "Ensolarado" será feito apenas o ramo desse parâmetro:

Cálculo do Ganho para os valores da "Temperatura" (Quente, Confortável e Frio):

$$E2 = -(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2(\frac{3}{4})) = 0.81$$

$$E(Quente) = -(0$$

$$+\frac{2}{2}\log_2(\frac{2}{2})) = 0$$

$$E(Confortável) = -(0$$

$$+\frac{1}{1}\log_2(\frac{1}{1})) = 0$$

$$E(Frio) = -(\frac{1}{1}\log_2\left(\frac{1}{1}\right)$$

$$+0 = 0$$

$$Ganho(E2, Temperatura) =$$

$$0.81 - \left(\frac{2}{4}\right) * 0 - \left(\frac{1}{4}\right) * 0 \ - \left(\frac{1}{4}\right) * 0 = 0.81$$

Cálculo do Ganho para os valores da "Vento" (Forte e Fraco):

$$E2 = -(\frac{1}{4}\log_2\left(\frac{1}{4}\right) + \frac{3}{4}\log_2(\frac{3}{4})) = 0.81$$

$$E(Fraco) = -(\frac{1}{3}\log_2(\frac{1}{3})) + \frac{2}{3}\log_2(\frac{2}{3})) = 0.92$$

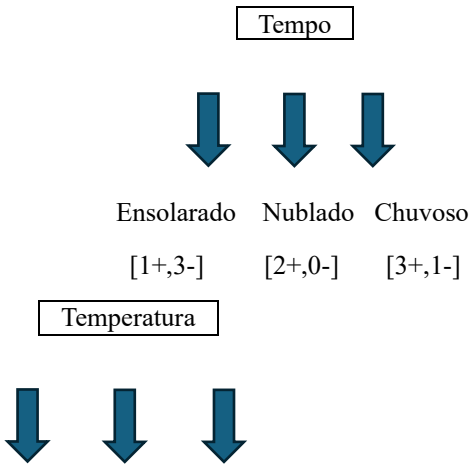
$$E(Forte) = -(0$$

$$+\frac{1}{1}\log_2(\frac{1}{1})) = 0$$

$$Ganho(E2, Vento) =$$

$$0.81 - \left(\frac{3}{4}\right) * 0.92 - \left(\frac{1}{4}\right) * 0 = 0.12$$

Ou seja, o melhor parâmetro para usar como continuação do ramo é a "Temperatura", pois ele foi o que obteve o maior ganho (pureza maior). Montando o diagrama da árvore fica:



Como todos os grupos mostraram valores com apenas "+" ou "-" eles chegaram à pureza completa, logo esses ramos se tornaram folhas da árvore. Portando, observando os resultados a resposta do problema é, se o dia estiver "Ensolarado", "Frio" e Vento "Fraco" o jogo de tênis irá acontecer. Após os dois exemplos é possível ver que ambos os métodos são efetivos mudando basicamente os cálculos feitos.

IV. MATERIAIS E MÉTODOS

Os dados utilizados para o desenvolvimento do modelo de regressão por arvore de decisão foram obtidos através do Instituto nacional de meteorologia (IMET), onde foi recolhidos dados no estado de Goiás na região de Goiânia da estação de meteorologia automática, código da estação A002, latitude -16,64277777, longitude -49,2199999 e altitude 770, foram recolhidos dados de 2010 a 2023 e destes dados foi usado para realização do estudo os três últimos meses de cada ano, pelo fato da criação de um senário hipotético onde usaremos o período da soja no estado de Goiás onde se inicia no mês outubro e termina no mês de dezembro.

Como material de desenvolvimento do projeto foi utilizado o Visual Studio code como ambiente de desenvolvimento, em conjunto com o software foi utilizado a linguagem Python para o desenvolvimento do algoritmo, A linguagem de programação Python foi escolhida para o desenvolvimento deste modelo de machine learning devido à sua facilidade de aprendizado e ampla adoção na área de ciência de dados e produtividade, tendo que o Python permite escrever programas mais simples, e possui um grande acervo de bibliotecas permitindo agilizar alguns processos, outro bom ponto desta linguagem é sua compatibilidade com diversos sistemas operacionais [2]. Para efetuar a manipulação dos dados e criar os *dataframes*, que seriam usados no treinamento do modelo. Foi utilizada a biblioteca pandas que permite que seja possível o tratamento e a análise dos dados, que irão ser usados para o treino e validação do modelo de *machine lerning* [6]. Em conjunto foi utilizado o método de Hargreaves para a obtenção da evapotranspiração de referência (ETo), o método de Hargreaves aplica a análise de regressão em dados diários de evapotranspiração de referência [1], [4],[17], obtendo a equação (5):

$$ETO = 0{,}0023\ (T_{med} + 17{,}18)(T_{max} - T_{min})^{1/2} R_a\ 0{,}480 \quad (5)$$

ETO : evapotranspiração de referência, em $mm\ /dia$

Tmed: Temperatura média diária, em °C;

Tmax: Temperatura máxima diária, em °C;

Tmin: Temperatura mínima diária, em °C;

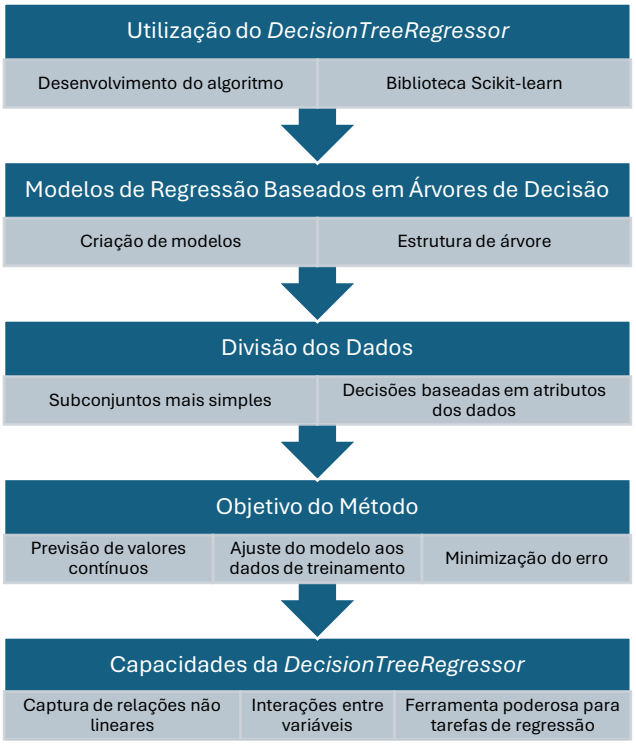
Ra: Radiação do topo da atmosfera, MJ $m^{-2}\ dia^{-1}$

Os dados utilizados para o desenvolvimento do modelo de regressão por arvore de decisão foram obtidos através do Instituto nacional de meteorologia (IMET), onde foi recolhidos dados no estado de Goiás na região de Goiânia da estação de meteorologia automática, código da estação A002, latitude -16,64277777, longitude -49,2199999 e altitude 770, foram recolhidos dados de 2010 a 2023 e destes dados foi usado para realização do estudo os três últimos meses de cada ano, pelo fato da criação de um senário hipotético onde usaremos o período da soja no estado de Goiás onde se inicia no mens outubro e termina no mês de dezembro.

Outra biblioteca amplamente usada na programação foi a Numpy é o pacote fundamental para a computação cientifica em Python. É uma biblioteca que oferece um objeto *array* multidimensional, é um pacote muito utilizado por outras bibliotecas para efetuar cálculos de maior complexidade como operações matemáticas, logica, manipulação de formas, classificação, seleção,

transformadas discretas de Fourier, álgebra linear básica, operações estatísticas simples. Com este acervo de funcionalidade torna a Numpy uma das bibliotecas mais importantes da linguagem de programação Python [8].

A biblioteca Scikit-learn é uma ferramenta essencial para aprendizado de máquina em Python. Desenvolvida como um projeto de código aberto, ela oferece uma ampla gama de algoritmos eficientes para tarefas como classificação, regressão, clusterização e redução de dimensionalidade. Construída sobre as bibliotecas NumPy, SciPy e Matplotlib, a Scikit-learn é conhecida por sua simplicidade e facilidade de uso, tornando-se uma escolha popular tanto para iniciantes quanto para profissionais experientes. Além disso, ela inclui diversas ferramentas de pré-processamento de dados, o que facilita a preparação dos dados para modelagem. A Scikit-learn é amplamente utilizada em ambientes acadêmicos e industriais devido à sua robustez e versatilidade [7], [9], [10].



O *DecisionTreeRegressor* é um algoritmo de aprendizado de máquina desenvolvido com a biblioteca *Scikit-learn*. Este modelo de regressão é baseado em árvores de decisão, uma abordagem poderosa para prever valores contínuos. O desenvolvimento do algoritmo envolve a criação de modelos que utilizam uma estrutura de árvore, construída dividindo os dados em subconjuntos mais simples, com decisões baseadas em atributos específicos dos dados. O principal objetivo do *DecisionTreeRegressor* é prever valores contínuos, ajustando o modelo aos dados de treinamento e buscando minimizar o erro de previsão. Uma das grandes vantagens do *DecisionTreeRegressor* é sua capacidade de capturar relações não lineares e interações complexas entre variáveis, tornando-o uma ferramenta poderosa para diversas tarefas de regressão. [13].

V. RESULTADOS

O treinamento do algoritmo de regressão por árvore de decisão se iniciou com a separação de 90% dos dados totais para treinamento e 10% dos dados totais para a validação e utilizando uma profundidade máxima de ramificações da árvore na função *max_depth=10* este comando obriga o modelo a ter sempre a profundidade de ramificação igual a 10 ou seja ele poderá criar subi grupos até uma profundidade de 10 evitando assim o *Overfitting*, com esta configuração obtivemos os resultados expressados na figura (3).

```
Soma dos erros ao quadrado : 7947.47
Erros quadrático médio : 0.05
Erro Medio Abisoluto : 0.17
Raiz do erro quedratico medio : 0.22
R2-score : 0.57
```

Figura 3- Resultados estatísticos da regressão

É possível observar na figura (3), o R2-score O coeficiente de determinação, conhecido como R^2, é amplamente utilizado em análises de regressão para avaliar a qualidade de um modelo. Ele mede o quão bem a linha de regressão linear se ajusta aos dados e o poder explicativo das variáveis independentes. Embora o R^2 não comprove uma relação de causalidade entre as variáveis X e Y, ele é fundamental para entender a força e a direção da relação entre elas, assim obtivemos um $R^2 = 0.57$ podemos considerar que o modelo está se ajustando bem aos dados. O R2 varia de 0 a 1 e indica o quanto a linha de regressão se ajusta aos dados, quanto mais próximo de 1 estiver o R2, melhor é o ajuste da regressão linear aos dados, [12].

$$R^2 = \frac{SQexp}{SQtot} = 1 - \frac{SQres}{SQtot} \qquad (6)$$

R^2: Coeficiente de determinação

SQtot: Soma total dos quadrados

SQres: Soma dos quadrados dos resíduos

SQexp: soma dos quadrados explicada

A raiz do erro quadrático médio (REQM) é um dos principais indicadores de desempenho para modelos de previsão de regressão. Ela quantifica a diferença média entre os valores previstos pelo modelo e os valores reais observados. Esse indicador fornece uma estimativa da precisão das previsões feitas pelo modelo. Quanto menor for o valor da REQM, melhor será o desempenho do modelo de previsão. Em um cenário ideal, onde o modelo sempre prevê o valor exato, a REQM seria igual a 0. Com REQM do modelo igual a 0.22 podemos afirmar que o modelo está entregando uma boa precisão das estimações de valores do modelo, [13].

$$REQM = \sqrt{\frac{SSEw}{W}} = \sqrt{\frac{1}{W}\sum_{i=1}^{N} wi\ ui^2} \qquad (7)$$

REQM: Raiz do erro quadrático médio

SSEw: Soma ponderada dos quadrados

W: Peso total da população

N: Número de observações

Wi: Peso da observação número i

Ui: Erro associados á observação número i

O erro médio absoluto (MAE) é uma métrica usada para avaliar a performance de modelos de regressão em machine learning. Ele calcula a média das diferenças absolutas entre os valores previstos pelo modelo e os valores reais observados. Foi obtido um MAE igual a 0.17 que pode ser considerado um erro médio absoluto aceitável para o modelo, [14].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|yi - pi| \qquad (8)$$

MAE: Erro médio absoluto

n: Número de amostras

y: Valor absoluto para cada amostra

p: Valor previsto pelo modelo para cada amostra

O Erro Quadrático Médio (MSE) representa o desvio padrão dos erros de previsão, também conhecidos como valores residuais. Esses valores residuais indicam a distância entre os pontos de dados e a linha de regressão. O RMSE mede a dispersão desses erros, fornecendo uma indicação de quão próximos os dados estão da linha de melhor ajuste. Em outras palavras, ele mostra o grau de concentração dos dados em torno da linha de regressão. O modelo apresentou com resultado uma RMSE de 0.05 podemos considerar que nosso RMSE é tolerável, [15].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Yi - \widehat{Yi})^2 \qquad (9)$$

MSE: Erro quadrático médio

n: Números de pontos de dados

Yi: Valores observados

$\widehat{Yi}$: Valores previstos

A soma dos erros quadráticos (SSE) é calculada somando as diferenças ao quadrado entre cada valor observado da variável dependente e o valor previsto por um modelo de regressão para os mesmos pontos da variável independente. Em outras palavras, essa métrica quantifica a discrepância entre os valores reais e os valores previstos da variável dependente. O SSE do modelo foi de 7947.47, [16].

$$SSE = \sum(Yi\ \widehat{Yi}) \qquad (10)$$

SSE: Somo dos erros quadráticos

Yi: Valores observados

$\widehat{Yi}$: Valores previstos

Podemos ver o comportamento da regressão na figura (4), (5),para este treinamento foi escolhido o senário onde os dados são do período de produção da soja no estado de Goiás englobando os três meses do período do plantio da soja por ano, é notável que a reta de regressão do modelo de regressão por arvore de regressão é atópico de um modelo de regressão convencional, é possível observar que a reta de regressão não tomo um caminho linear, mas um caminho que se ajuste o mis próximo possível ao pontos isso e possível porque o nosso modelo efetua um agrupamento dos valore assim permitindo subi dividir os valore tornando mais fácil a regressão.

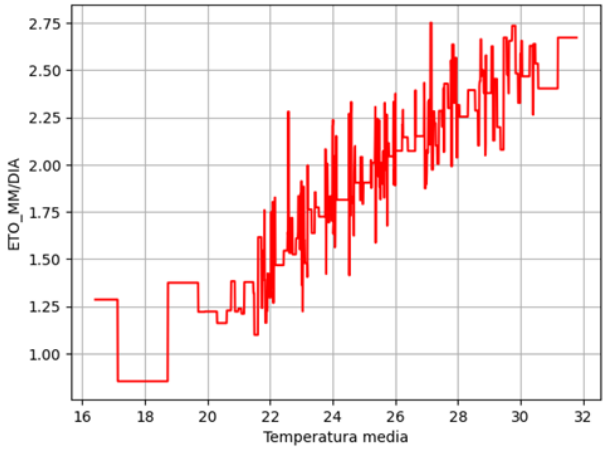


Figura (4) - Gráfico da ETO estimada

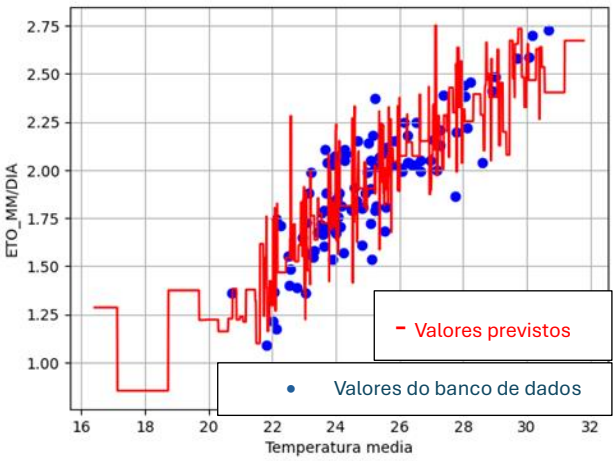


Figura (5) – Gráfico de relação dos valores de ETO prevista x ETO amostra

Com os dados de saída do modelo de arvore de regressão foi possível fazer um comparativo das medidas de ETO previstas pelo modelo e os valores calculados pelo métodos convencional que pode ser conferido no apêndice 1.

O gráfico de relação dos resultados da regressão e os valore calculados convencionalmente nos mostra uma boa assertividade do nosso modelo este gráfico foi elaborado com cerca de 6 meses de dados, onde para efetuar o treinamento do modelo foi usado como exemplo o período da soja onde seu ciclo e de 3 meses do ano sendo assim o gráfico compreende 2 anos dos dados usados para validação, expressado figura (6).

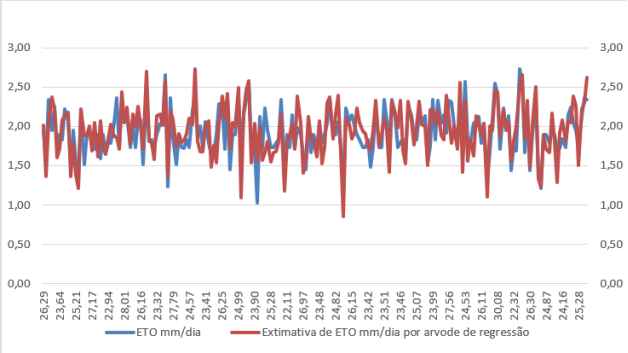


Figura (6) - gráfico de relação dos resultados

As configurações da máquina que foi utilizado para o desenvolvimento deste modelo e um notebook modelo Del G7 com as seguintes configurações processador core i7 8 geração, 21 G bit de memória Ram, placa de vídeo Nvidia Geforce GTX 1050, foi utilizado um computador com configurações básicas para o desenvolvimento do projeto, ou fazer a execução do treinamento e teste de validação a máquina leva apenas 9 segundos para efetuar todos os cálculos e nos dar uma resposta, demostrando que o modelo não necessita de um equipamento de grande potencia para funcionar sendo um bom ponto já que o valor para investimento de aplicação fica mais barato podendo chegar a mais pessoas.

VI. CONCLUSÃO

Os resultados obtidos demonstram a eficiência da técnica de regressão por árvore de decisão na previsão da ETO a partir de dados de temperatura diária. Com um R^2 de 0,57, REQM de 0,22 e RMSE de 0,5, o modelo mostrou-se funcional e alcançou resultados satisfatórios, validando sua aplicabilidade para este tipo de previsão.

O modelo desenvolvido poderá ser utilizado para criar modelos preditivos da ETO, auxiliando na tomada de decisão do melhor cenário de irrigação.

O objetivo é que este modelo possa ser aplicado tanto por grandes produtores quanto por pequenos. Pois não é necessário ter bases climáticas de grande complexidade para obter os dados de entrada, já que o modelo de arvore de regressão se baseia apenas na temperatura média diária.

VII. REFERÊNCIAS BIBLIOGRÁFICAS

[1] BERNARDO, Salassier. Manual de irrigação. Viçosa: Universidade Federal de Viçosa, 2009.

[2] **RUSSELL**, Stuart J.; NORVIG, Peter. Inteligencia Artificial: un enfoque moderno. Rio de Janeiro: Campus 2004.

[3] SILVA, L. M. **Uma aplicaçao de árvores de decisao, redes neurais e KNN para a identificaçao de modelos ARMA nao-sazonais e sazonais**. Rio de Janeiro. Tese de Doutorado-Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro, 2005.

[4] ALLEN, Richard G. et al. Crop evapotranspiration-Guidelines for computing crop water requirements-FAO Irrigation and drainage paper 56. Fao, Rome, v. 300, n. 9, p. D05109, 1998.

[5] PYTHON SOFTWARE FOUNDATION. Python 3.9.1 documentation. Python.org, 2020. Disponível em: https://docs.python.org/3.9/. Acesso em: 04 dez. 2024.

[6] MCKINNEY, Wes. Pandas: powerful Python data analysis toolkit. Versão 1.3.3. [S.l.]: The Pandas Development Team, 2021. Disponível em: <https://pandas.pydata.org/>. Acesso em: 4 dez. 2024.

[7] SCIKIT-LEARN. Scikit-learn: Machine Learning in Python. Versão 0.24.2. [S.l.]: The Scikit-learn Development Team, 2021. Disponível em: <https://scikit-learn.org/stable/documentation.html>. Acesso em: 4 dez. 2024.

[8] NUMPY DEVELOPERS. NumPy: Scientific Computing Tools for Python. Versão 1.21.0. [S.l.]: The NumPy Development Team, 2021. Disponível em: <https://numpy.org/doc/stable/>. Acesso em: 4 dez. 2024.

[9] SCIPY DEVELOPERS. SciPy: Scientific Computing Tools for Python. Versão 1.7.1. [S.l.]: The SciPy Development Team, 2021. Disponível em: <https://scipy.org/docs.html>. Acesso em: 4 dez. 2024.

[10] HUNTER, John D. et al. Matplotlib: Visualization with Python. Versão 3.4.3. [S.l.]: The Matplotlib Development Team, 2021. Disponível em: <https://matplotlib.org/stable/contents.html>. Acesso em: 5 dez. 2024.

[11] INSTITUTO NACIONAL DE METEOROLOGIA. Portal INMET. Disponível em: <https://portal.inmet.gov.br/>. Acesso em: 3 nov. 2024.

[12] ESCOLA DNC. Análise de Regressão e Coeficiente de Determinação: Guia Completo. Disponível em: <https://www.escoladnc.com.br/blog/analise-de-regressao-e-coeficiente-de-determinacao-guia-completo/#:~:text=O%20R2%20%C3%A9%20um%20coeficiente%20de%20determina%C3%A7%C3%A3o%20usado%20para%20avaliar,como%20as%20vari%C3%A1veis%20se%20relacionam>. Acesso em: 25 nov. 2024.

[13] SAP. SAP Analytics Cloud. Disponível em: <https://help.sap.com/docs/SAP_ANALYTICS_CLOUD/00f68c2e08b941f081002fd3691d86a7/12cda9ce48b049a8adda7a6f3c240fa4.html>. Acesso em: 25 nov. 2024.

[14] FILHO, Mario. MAE: Erro Médio Absoluto em Machine Learning. Disponível em: <https://mariofilho.com/mae-erro-medio-absoluto-

em-machine-learning/#:~:text=O%20erro%20m%C3%A9dio%20absoluto%20(MAE,modelo%20e%20os%20valores%20observados>. Acesso em: 25 nov. 2024.

[15] ORACLE. Insights Metrics: RMSE. Disponível em: <https://docs.oracle.com/cloud/help/pt_BR/pbcs_common/PFUSU/insights_metrics_RMSE.htm#PFUSU-GUID-FD9381A1-81E1-4F6D-8EC4-82A6CE2A6E74>. Acesso em: 1 dez. 2024.

[16] DATACAMP. Regression Sum of Squares. Disponível em: <https://www.datacamp.com/pt/tutorial/regression-sum-of-squares>. Acesso em: 3 dez. 2024

[17] VIANA, Patrick T.; MEDEIROS, Antônio M. M. Programa para cálculo da evapotranspiração por métodos

APENDICE 1

TEMPERATURA MÉDIA. Dia	**ETO mm/dia**	**Estimativa de ETO mm/dia por árvore de regressão**	**ERRO %**
26,29	2,14	2,04	5,19
25,24	1,95	2,02	3,25
23,03	1,54	1,36	12,94
28,59	2,35	2,04	15,12
25,24	1,95	2,37	17,78
28,13	2,26	2,22	1,70
23,64	1,74	1,61	8,22
23,81	1,89	1,71	10,92
24,30	1,83	2,08	11,92
27,26	2,22	2,13	4,13
25,17	2,07	2,18	5,07
22,06	1,52	1,37	10,99
25,21	1,95	1,79	9,19
22,77	1,56	1,39	12,75
22,02	1,30	1,21	6,72
27,26	2,22	2,21	0,36
23,15	1,51	1,88	19,68
24,91	1,90	1,88	0,96
27,17	1,97	2,00	1,27
23,63	1,74	1,69	2,81
24,26	1,71	2,05	16,51
24,81	1,90	1,61	18,13
25,38	1,59	2,07	23,23
25,11	1,90	1,72	10,42
22,94	1,65	1,65	0,38
23,90	1,78	1,84	2,90
23,91	1,78	2,03	12,12
24,10	2,05	1,88	9,08
27,74	2,36	1,87	26,60
23,51	1,74	1,72	1,31
	ETO MEDIO MÊS	**ETO MEDIO EXTIMADO POR MÊS**	**ERRO MEDIO %**
	1,809289109	1,793589398	0,875323577