UNIVERSIDADE FEDERAL DO TOCANTINS

Árvore de Decisão

Discentes: Luana Lorena de Matos Tavares, Pablo Henrique Silva Ribeiro, Arthur Fernandes Ferreira reis, Neudison Nonato Maia Filho e Plácido de Aquino Angelim Neto

Docente: Dra. Glenda Botelho

Universidade Federal do Tocantins

1. Introdução

A Inteligência Artificial (IA) revolucionou inúmeros setores da sociedade, desde a saúde até a educação e as finanças. Dentre as várias técnicas utilizadas na IA, as Árvores de Decisão destacam-se como uma ferramenta poderosa para a tomada de decisões e resolução de problemas.

As Árvores de Decisão são um tipo de algoritmo de aprendizado de máquina supervisionado que é amplamente usado em classificação e regressão. Esses modelos são baseados na estrutura de uma árvore, em que cada nó interno representa um teste em uma determinada característica (ou atributo), cada ramo representa o resultado desse teste e cada folha representa uma classe ou um valor numérico predito. No entanto, a sua maior vantagem é a sua transparência e facilidade de interpretação. Ao contrário de muitos outros algoritmos de aprendizado de máquina, as Árvores de Decisão oferecem resultados que podem ser facilmente interpretados, mesmo sem um conhecimento profundo da matemática subjacente. Isso as torna uma escolha popular em muitos campos onde a interpretabilidade é importante, como a medicina.

No nosso dia a dia, as Árvores de Decisão são usadas em muitas aplicações. Por exemplo, elas podem ser usadas por um serviço de streaming para decidir que tipo de conteúdo recomendar a um usuário, com base em seu histórico de visualizações e preferências. Em um contexto médico, as Árvores de Decisão podem ser usadas para ajudar a prever a probabilidade de um paciente ter uma determinada doença, com base em sintomas, histórico médico e outros fatores. No setor financeiro, elas podem ser usadas para avaliar a probabilidade de um cliente potencial ser capaz de reembolsar um empréstimo.

Em suma, as Árvores de Decisão são uma ferramenta essencial na caixa de ferramentas da IA, com inúmeras aplicações práticas. Uma ferramenta popular para construir e analisar árvores de decisão é o WEKA (Waikato Environment for Knowledge Analysis). O WEKA é uma plataforma de software de código aberto que oferece uma ampla gama de algoritmos de aprendizado de máquina, incluindo árvores de decisão. Neste relatório, abordaremos a utilização de árvores de decisão nos conjuntos de dados Iris e Câncer. Os conjuntos de dados são comumente utilizados para fins de demonstração e experimentação na área de aprendizado de

máquina, sendo úteis para ilustrar conceitos importantes e técnicas de análise de dados. Será feito um comparativo de resultados com a ferramenta WEKA.

2. Metodologia

2.1. Conceitos Importantes:

O conjunto de dados Câncer é amplamente utilizado para classificação de tumores como malignos ou benignos. Ele contém informações sobre características de células tumorais, como raio, textura, suavidade, compactação, simetria, dimensão fractal, entre outros. O objetivo é classificar corretamente os tumores malignos ou benignos com base nessas características.

O conjunto de dados Íris é composto por 150 amostras de flores Íris de três espécies diferentes (Setosa, Versicolor e Virginica). Cada amostra é descrita por quatro características: comprimento da sépala, largura da sépala, comprimento da pétala e largura da pétala. O objetivo é classificar corretamente cada flor em uma das três espécies com base nessas características.

Nós também faremos uso do WEKA, que se trata de uma aplicação onde iremos inserir o dataset Câncer e o Íris para podermos realizar um comparativo. O WEKA (Waikato Environment for Knowledge Analysis) é uma plataforma de software de aprendizado de máquina de código aberto desenvolvida pela Universidade de Waikato, na Nova Zelândia. Ele fornece uma ampla gama de algoritmos de aprendizado de máquina e recursos para análise de dados, incluindo a construção e análise de árvores de decisão. O WEKA possui uma interface amigável e é amplamente utilizado na comunidade acadêmica e na indústria.

2.. O tratamento do dataset Câncer:

Inicialmente, as bibliotecas necessárias são importadas, como matplotlib, pandas, scikit-learn (sklearn) e outras, que fornecem funcionalidades essenciais para o processamento e análise dos dados. Os dados são lidos a partir de um arquivo CSV usando a função apropriada, como a função read_csv do pandas. Em seguida, são realizadas etapas de pré-processamento, como tratamento de valores ausentes, remoção de colunas desnecessárias e transformação de variáveis categóricas em representações numéricas adequadas.

Após o pré-processamento, os dados são divididos em conjuntos de treinamento e teste. O conjunto de treinamento é usado para treinar o modelo de classificação, enquanto o conjunto de teste é utilizado para avaliar seu desempenho em dados não vistos anteriormente.

Na etapa de seleção do modelo, é escolhido um modelo de classificação adequado para o problema em questão, como árvores de decisão, regressão logística ou SVM. O modelo selecionado é treinado com os dados de treinamento, ajustando seus parâmetros internos para otimizar sua capacidade de fazer previsões precisas. Após o treinamento, o modelo é avaliado usando o conjunto de testes. São calculadas métricas de desempenho, como acurácia, precisão, recall e F1-score, além da matriz de confusão, para analisar seu desempenho em diferentes categorias de classificação. Se necessário, ajustes podem ser feitos no modelo, refinando seus parâmetros ou utilizando técnicas de validação cruzada. O objetivo é melhorar seu desempenho e capacidade de generalização.

Por fim, o modelo treinado, avaliado e ajustado está pronto para ser aplicado a novos conjuntos de dados não vistos anteriormente, permitindo fazer previsões em situações reais.

2.2. O tratamento do dataset Íris:

Sabendo que o conjunto de dados Iris é dividido em pares de características e cada par é utilizado para treinar uma árvore de decisão e realizar previsões, o código começa importando as bibliotecas necessárias, como numpy, matplotlib, sklearn, entre outras. Em seguida, é feita a importação do conjunto de dados Íris usando a função load_iris(). O conjunto de dados é dividido em diferentes pares de características, e para cada par, é realizado o treinamento de uma árvore de decisão. Os dados são divididos em conjuntos de treinamento e teste usando a função train_test_split(). A árvore de decisão é então criada e treinada usando o conjunto de treinamento.

Após o treinamento, são feitas previsões usando o conjunto de teste e são calculadas as métricas de desempenho, como acurácia e classification_report. O resultado das previsões é então plotado em um gráfico, mostrando as regiões de decisão da árvore de decisão para cada par de características. No final do código, é exibido o desempenho da última árvore de decisão treinada, incluindo a acurácia e o classification_report. Também é mostrado o tempo de processamento total.

2.3. Nomenclaturas:

É importante ressaltarmos algumas nomenclaturas que utilizaremos para obter nossos resultados:

- TP (True Positive): representa os casos positivos corretamente classificados pelo modelo. Ou seja, são os casos em que o modelo previu corretamente a classe positiva.
- TN (True Negative): representa os casos negativos corretamente classificados pelo modelo. Ou seja, são os casos em que o modelo previu corretamente a classe negativa.
- FP (False Positive): representa os casos negativos incorretamente classificados pelo modelo. Ou seja, são os casos em que o modelo previu erroneamente a classe positiva, quando na verdade eles pertencem à classe negativa. Também é conhecido como "falso alarme" ou erro tipo I.
- FN (False Negative): representa os casos positivos incorretamente classificados pelo modelo. Ou seja, são os casos em que o modelo previu erroneamente a classe negativa, quando na verdade eles pertencem à classe positiva. Também é conhecido como "falso negativo" ou erro tipo II.
- Acurácia (Accuracy): é a proporção de exemplos corretamente classificados em relação ao total de exemplos. Essa métrica fornece uma medida geral do desempenho do modelo, indicando a taxa de acertos global.
- Precisão (Precision): é a proporção de exemplos positivos corretamente classificados em relação ao total de exemplos classificados como positivos (verdadeiros positivos + falsos positivos). Essa métrica avalia a capacidade do modelo em classificar corretamente os exemplos positivos, sendo útil quando o foco é minimizar os falsos positivos.
- Recall (Recall) ou Sensibilidade: é a proporção de exemplos positivos corretamente classificados em relação ao total de exemplos que realmente são positivos (verdadeiros positivos + falsos negativos). Essa métrica mede a capacidade do modelo em identificar corretamente os exemplos positivos, sendo útil quando o foco é minimizar os falsos negativos.
- F1-Score (F1-Score): é a média harmônica entre a precisão e o recall. Essa métrica fornece um equilíbrio entre a precisão e o recall, sendo especialmente útil quando as classes estão desequilibradas.

- Support (Suporte): é o número de exemplos de cada classe presente no conjunto de teste. Essa métrica indica quantos exemplos de cada classe foram considerados na avaliação do modelo.

3. Resultados

3.1. dataset Cancer

Ao aplicarmos o dataset Cancer em nossa Árvore de Decisão, como se pode notar pela Figura 01, ela obteve 92,5% de acurácia que se trata da porcentagem do conteúdo que a árvore conseguiu processar. Dentro dessa porcentagem tivemos 188 dados computados, onde a precisão foi de 94% para os dados de câncer benigno e 90% para os dados de câncer maligno.

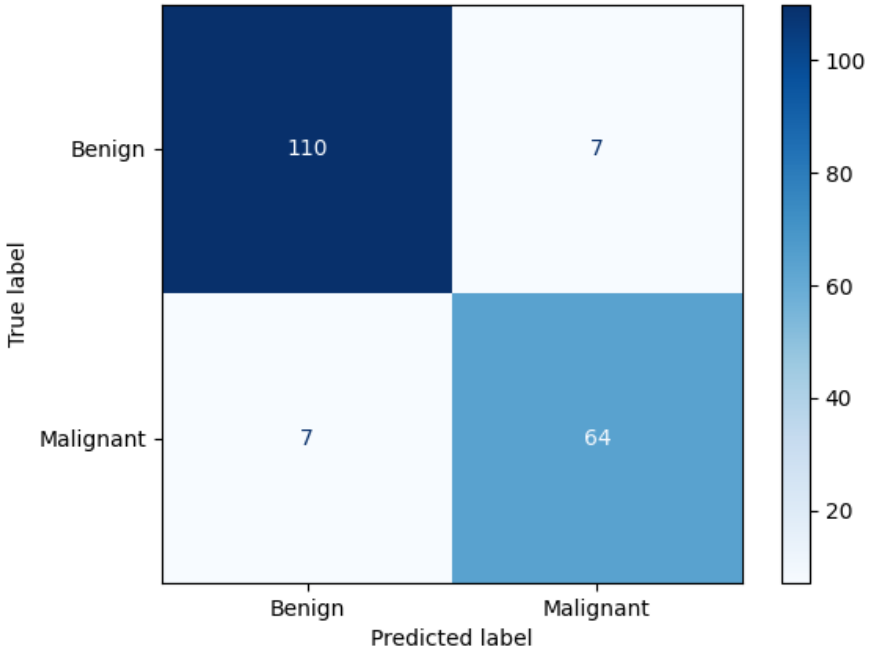
O recall(sensibilidade) que se trata da previsão do que o código esperava consiga ler de forma precisa foi exatamente o valor indicado pela precisão. O f1-score é a média entre os valores de precisão e recall, a tratativa referente a suporte quer dizer a quantidade de dados foram processados pelo algoritmo, tanto benigno quanto maligno.

Figura 01

```
Model Decision Tree:

tn 110, fp 7, fn 7, tp 64

Accuracy: 0.925531914893617

Classification Report:
              precision    recall  f1-score   support

      Benign       0.94      0.94      0.94       117
   Malignant       0.90      0.90      0.90        71

    accuracy                           0.93       188
   macro avg       0.92      0.92      0.92       188
weighted avg       0.93      0.93      0.93       188


O tempo de processamento foi de 0.7957401275634766 segundos.
```

Na Figura 02, podemos analisar melhor os dados mencionados anteriormente em um gráfico aferido com os dados previamente fornecidos:

Figura 02



3.2. dataset Íris

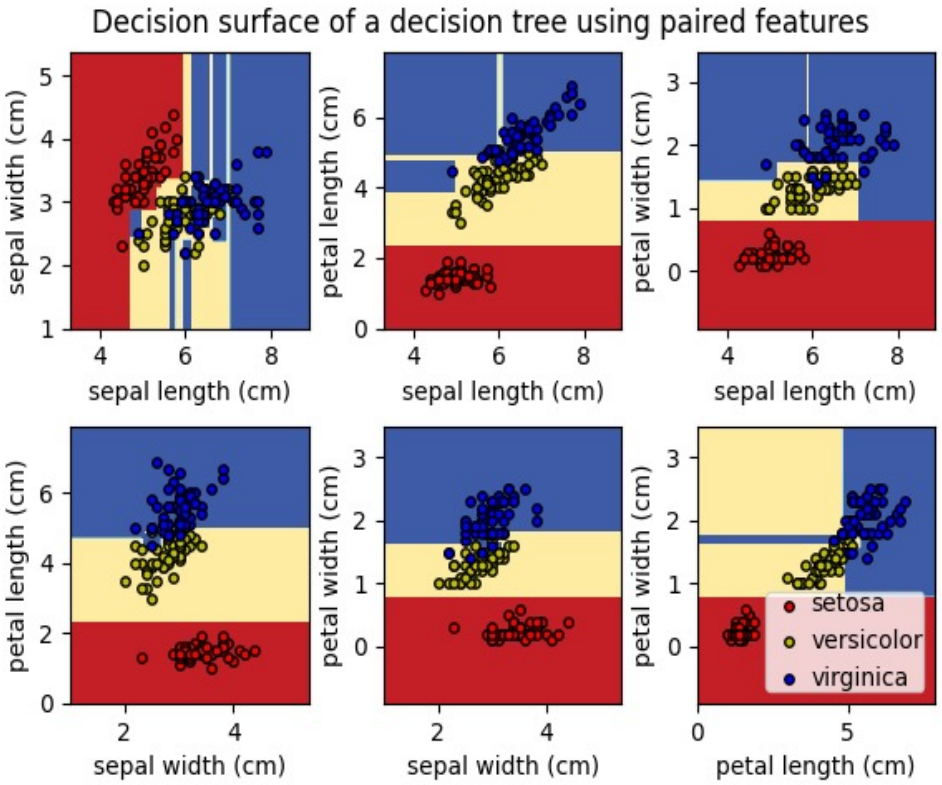
Ao aplicarmos o dataset Íris obtivemos uma acurácia de 97% onde o algoritmo obteve 100% de precisão na espécie setosa e seu recall também foi 100%. Ou seja, o algoritmo já esperava acertar completamente. Foram analisadas 11 flores dessa espécie, logo após obteve 93% de precisão na espécie versicolor, e houveram algumas falhas pois o algoritmo esperava acertar 100% de seus processos. Foram analisadas 13 flores dessa espécie. Por último, o algoritmo obteve 100% de precisão em flores da espécie virginica onde o recall foi de 83 %, ou seja, o algoritmo esperava uma menor assertividade para o conjunto de 6 flores. Como mostrado na Figura 03.

Figura 03

```
Model 6:
Accuracy: 0.9666666666666667
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        11
           1       0.93      1.00      0.96        13
           2       1.00      0.83      0.91         6

    accuracy                           0.97        30
   macro avg       0.98      0.94      0.96        30
weighted avg       0.97      0.97      0.97        30


O tempo de processamento foi de 4.991386651992798 segundos.
```

Na Figura 04, temos um gráfico demonstrando cada parte obtida no relatório para um melhor entendimento:

Figura 04



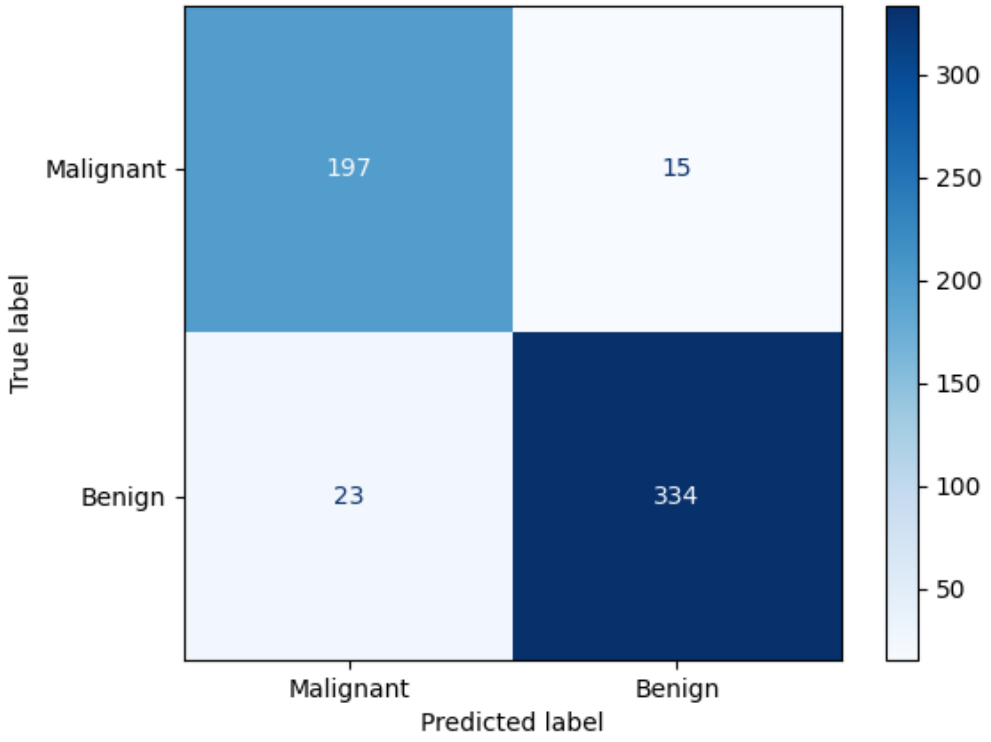
3.3. WEKA

Na Figura 05, realizamos uma iteração do conjunto de dados "Câncer" usando o Weka e examinamos os resultados obtidos. Ao aplicarmos o conjunto de dados no aplicativo, alcançamos uma precisão de aproximadamente 99% em 569 registros analisados, o que indica uma classificação e acurácia altas. Em relação ao câncer maligno, observamos uma precisão de acertos de 99%, com 98% dessa precisão sendo identificados corretamente. No caso do câncer benigno, obtivemos uma precisão de 99% e uma taxa de acertos de 99%.

Figura 05

```
Time taken to build model: 1.11 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.31 seconds

=== Summary ===

Correctly Classified Instances         564               99.1213 %
Incorrectly Classified Instances         5                0.8787 %
Kappa statistic                          0.9812
Mean absolute error                      0.0165
Root mean squared error                  0.0908
Relative absolute error                  3.5281 %
Root relative squared error             18.7856 %
Total Number of Instances              569

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,981    0,003    0,995      0,981    0,988      0,981    0,993     0,991     M
                 0,997    0,019    0,989      0,997    0,993      0,981    0,993     0,992     B
Weighted Avg.    0,991    0,013    0,991      0,991    0,991      0,981    0,993     0,992

=== Confusion Matrix ===

   a   b   <-- classified as
 208   4 |   a = M
   1 356 |   b = B
```

Figura 06

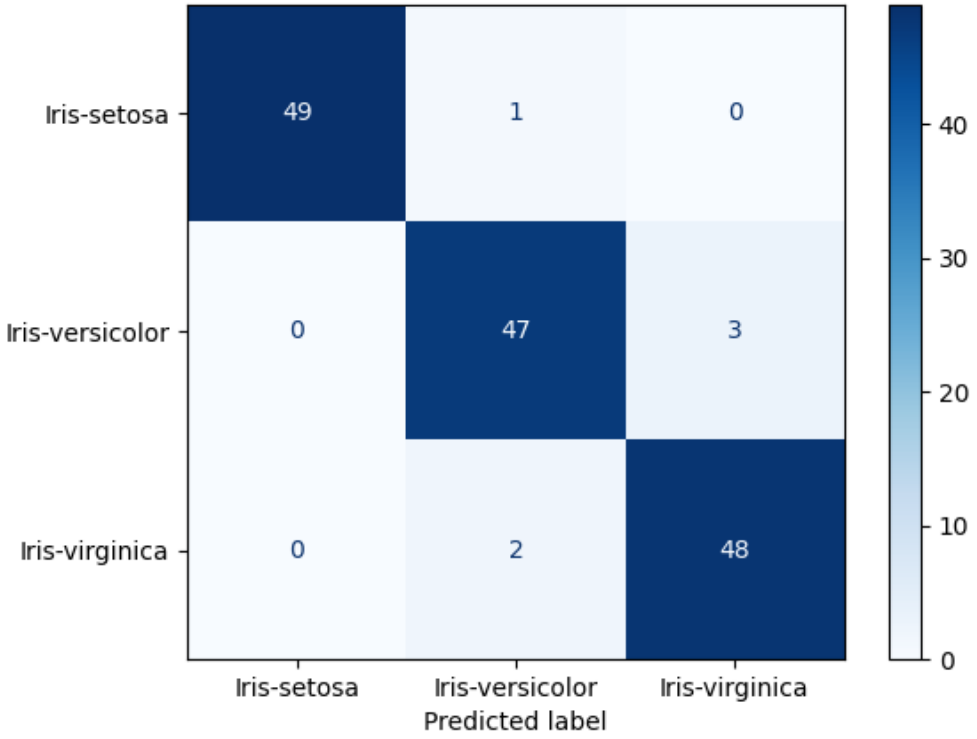


Ao utilizar o Weka com o conjunto de dados íris, alcançamos uma acurácia de 96% no modelo de árvore de decisão. Ao examinar a espécie "setosa", notamos que a precisão foi de 100%, mas o modelo identificou corretamente apenas 98% dos casos, de acordo com o recall. Para a espécie "versicolor", a precisão foi de 94%, e o modelo identificou corretamente 94% dos casos encontrados. Por fim, na espécie "virgínica", a precisão foi de 94%, com um recall de 96% (ver Figura 07).

Figura 7

```
Time taken to build model: 0.09 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         144               96      %
Incorrectly Classified Instances         6                4      %
Kappa statistic                          0.94
Mean absolute error                      0.035
Root mean squared error                  0.1586
Relative absolute error                  7.8705 %
Root relative squared error             33.6353 %
Total Number of Instances              150

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,980    0,000    1,000      0,980    0,990      0,985    0,990     0,987     Iris-setosa
                 0,940    0,030    0,940      0,940    0,940      0,910    0,952     0,880     Iris-versicolor
                 0,960    0,030    0,941      0,960    0,950      0,925    0,961     0,905     Iris-virginica
Weighted Avg.    0,960    0,020    0,960      0,960    0,960      0,940    0,968     0,924

=== Confusion Matrix ===

  a  b  c   <-- classified as
 49  1  0 |  a = Iris-setosa
  0 47  3 |  b = Iris-versicolor
  0  2 48 |  c = Iris-virginica
```

Figura 08



4. Considerações Finais

A Inteligência Artificial, especialmente os algoritmos de aprendizado de máquina, como as Árvores de Decisão, está desempenhando um papel cada vez mais importante na nossa sociedade. O poder desses algoritmos em lidar com grandes volumes de dados e tomar decisões baseadas em padrões complexos os torna uma ferramenta valiosa em uma ampla gama de campos.

Neste trabalho, exploramos o uso de Árvores de Decisão em dois conjuntos de dados diferentes, o conjunto de dados Iris e o conjunto de dados Cancer. Através da implementação e análise desses modelos, conseguimos obter uma compreensão mais profunda de como as Árvores de Decisão funcionam e quais são as suas potenciais aplicações.

Os resultados obtidos evidenciam a eficácia desses algoritmos. Com uma taxa de acurácia de 92,5% no conjunto de dados "Câncer" e 97% no conjunto de dados "Iris", as Árvores de Decisão demonstraram ser uma ferramenta poderosa para classificação e predição em diversas aplicações.

Ao compararmos o desempenho do Weka com outros algoritmos, verificamos que ele alcançou uma acurácia de 99% para todas as colunas do conjunto de dados. Além disso, sua precisão em relação aos acertos foi significativamente superior à do algoritmo com precisão e recall iguais, porém com menor acurácia. Esses resultados mostram que o Weka obteve maior precisão e acurácia, superando o desempenho do algoritmo, com um tempo de processamento ligeiramente maior de 0,31 segundos.

No caso do conjunto de dados "Iris", o desempenho superior foi alcançado pelo algoritmo. Embora ambos tenham apresentado a mesma acurácia, o Weka mostrou diferenças

significativas entre precisão e recall em comparação com o algoritmo. No entanto, o Weka obteve uma vantagem em termos de tempo de processamento, sendo aproximadamente 0,9 segundos mais rápido que o algoritmo, que levou cerca de 4 segundos para processar os dados.

No entanto, é importante lembrar que o sucesso de qualquer algoritmo de aprendizado de máquina depende em grande parte da qualidade dos dados de entrada, bem como da escolha dos parâmetros corretos. Portanto, é essencial passar tempo suficiente na fase de pré-processamento dos dados e na seleção do modelo.

Em conclusão, as Árvores de Decisão representam uma abordagem eficaz e interpretável para a solução de problemas complexos de classificação e regressão, tornando-as uma ferramenta indispensável no campo da Inteligência Artificial.