

UNIVERSIDADE DE SÃO PAULO
FACULDADE DE FILOSOFIA, CIÊNCIAS E LETRAS DE RIBEIRÃO PRETO
DEPARTAMENTO DE COMPUTAÇÃO E MATEMÁTICA

CAÍQUE AUGUSTO FERREIRA

**Indução de Árvore de Decisão utilizando
Meta-Aprendizado**

Ribeirão Preto-SP

2022

CAÍQUE AUGUSTO FERREIRA

Indução de Árvore de Decisão utilizando Meta-Aprendizado

Versão Original

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) da Universidade de São Paulo (USP), como parte das exigências para a obtenção do título de Mestre em Ciências.

Área de Concentração: Computação Aplicada.

Orientador: Prof. Dr. José Augusto Baranauskas

Ribeirão Preto-SP

2022

CAÍQUE AUGUSTO FERREIRA

Decision Tree Induction Using Meta-Learning

Original Version

Dissertation presented to Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto (FFCLRP) from the Universidade de São Paulo (USP), as part of the requirements to hold the Master of Science degree.

Field of Study: Applied Computing.

Supervisor: Prof. Dr. José Augusto Baranauskas

Ribeirão Preto-SP

2022

Caíque Augusto Ferreira

Indução de Árvore de Decisão utilizando Meta-Aprendizado. Ribeirão Preto–SP,
2022.

100p. : il.; 30 cm.

Dissertação apresentada à Faculdade de Filosofia, Ciências e Letras
de Ribeirão Preto da USP, como parte das exigências para
a obtenção do título de Mestre em Ciências,
Área: Computação Aplicada.

Orientador: Prof. Dr. José Augusto Baranauskas

1. Meta-Aprendizado. 2. Árvore de Decisão. 3. Combinação de Modelos.

Caíque Augusto Ferreira

Indução de Árvore de Decisão utilizando Meta-Aprendizado

Modelo canônico de trabalho monográfico
acadêmico em conformidade com as normas
ABNT.

Trabalho aprovado. Ribeirão Preto–SP, 20 de Setembro de 2022

Orientador:
Prof. Dr. José Augusto Baranauskas

Professor
Convidado 1

Professor
Convidado 2

Ribeirão Preto–SP
2022

Dedico este trabalho ao Criador de todas as coisas, em especial a Jesus Cristo, o autor e consumador da minha fé.

Agradecimentos

Agradeço, primeiramente, a Deus por ter me dado saúde e vida para conseguir chegar até aqui e realizar este sonho que tenho desde a minha adolescência. Agradeço a toda minha família e amigos, em especial minha querida esposa Danielle, meus pais Andréia e Eduardo e meu irmão Diego, que me suportaram, em todos os sentidos, por toda esta caminhada, me apoiando e me motivando a perseverar na busca dos meus sonhos. Agradeço também, imensamente, ao meu orientador Prof. Dr. José Augusto Baranauskas, que me concedeu o privilégio de poder trabalhar com ele todo esse tempo, e com muita paciência e dedicação conduziu, de uma forma excelente, o nosso trabalho, no qual conseguimos contribuir um pouquinho com a ciência. Por fim, agradeço o meu colega Adriano Henrique Cantão, colaborou de forma incrível com a publicação da nossa pesquisa.

*“Não vos amoldeis às estruturas deste mundo,
mas transformai-vos pela renovação da mente,
a fim de distinguir qual é a vontade de Deus:
o que é bom, o que Lhe é agradável, o que é perfeito.
(Bíblia Sagrada, Romanos 12, 2)*

Resumo

Os modelos de aprendizado simbólico destacam-se dentro da área do Aprendizado de Máquina devido às suas representações serem interpretáveis pelo homem. Uma característica deste modelo é ser excessivamente responsivo ao conjunto de exemplos utilizados, o que pode resultar em uma piora significativa no desempenho caso haja pequenas variações no conjunto de treinamento. A estratégia de combinação de modelos (*ensembles*) apresenta-se como uma alternativa para melhorar a precisão e a estabilidade dos modelos. A estratégia consiste em gerar diferentes modelos por meio do mesmo conjunto de treinamento e combiná-los em um único modelo final, geralmente, por meio de um processo de votação. Uma característica indesejável da estratégia *ensemble* é a complexidade do modelo final, já que este é formado por um conjunto de modelos. Nesta pesquisa é proposta uma abordagem para induzir uma meta-árvore de decisão com base na combinação das árvores de decisão de uma floresta (*Random Forest*). Experimentos foram realizados em 150 *datasets* de diferentes domínios. A abordagem proposta aplicada em 43 *datasets* categóricos dos 150 analisados, obteve um desempenho tão bom quanto uma floresta com 128 árvores sem diferenças estatisticamente significativas. Trata-se de um resultado interessante, levando em consideração a interpretabilidade fornecida por uma única árvore de decisão como modelo resultante.

Palavras-chave: Meta-Aprendizado. Árvore de Decisão. Combinação de Modelos. Inteligência Artificial Explicável.

Abstract

Symbolic learning models stand out within the Machine Learning area due to their representations being human-interpretable. A characteristic of this model is that it is excessively responsive to the set of examples used, which can result in a significant decrease in performance if there are small variations in the training set. The strategy of combining models (ensembles) is presented as an alternative to improve the accuracy and stability of the models. The strategy is to generate different models using the same training set and combine them into a single final model, usually through a voting process. An undesirable characteristic of the ensemble strategy is the complexity of the final model, since it is formed by a set of models. In this research, an approach is proposed to induce a meta-decision tree based on the combination of decision trees of a forest (Random Forest). Experiments were performed on 150 datasets from different domains. The proposed approach applied to 43 categorical datasets of the 150 analyzed, performed as well as a forest with 128 trees without statistically significant differences. This is an interesting result, considering the interpretability provided by a single decision tree as the resulting model.

Keywords: Meta-Learning. Decision tree. Model Combination. Explainable Artificial Intelligence.

Lista de figuras

Figura 1 – Exemplo de transformação de Árvore de Decisão (a) para Tabela de Decisão (b).	37
Figura 2 – CMM adaptado de Domingos (1997).	45
Figura 3 – MDT adaptado de Strecht, Mendes-Moreira e Soares (2014).	47
Figura 4 – Fusão do MDT adaptado de Strecht, Mendes-Moreira e Soares (2014).	48
Figura 5 – Processo de Indução da Meta Árvore de Decisão.	52
Figura 6 – (a) Exemplo de árvore de decisão contendo atributos numéricos induzida a partir do conjunto de exemplos da Tabela 3. (b) Tabela de decisão correspondente à árvore de decisão em (a).	54
Figura 7 – Exemplo de tabela de decisão para a árvore da Figura 6 utilizando a média para representar o intervalo de valores.	55
Figura 8 – Exemplo de tabela de decisão para a árvore da Figura 6 utilizando os limites inferior e superior para representar o intervalo de valores.	55
Figura 9 – Ilustração do algoritmo de meta aprendizado proposto e implementado.	58
Figura 10 – Resultados de ROC AUC (Tabelas 8-11 do Apêndice A). A cor azul destaca o resultado obtido pela DT, a cor vermelha destaca o resultado obtido pela RF. A cor preta representa as variações do algoritmo proposto.	71
Figura 11 – Resultados sobre número de folhas. A cor azul destaca o resultado obtido pela DT. A cor preta representa as variações do algoritmo proposto.	72
Figura 12 – Diagrama de diferença crítica considerando os resultados obtidos de ROC AUC (Tabelas 8-11 do Apêndice A).	73
Figura 13 – Diagrama de diferença crítica considerando o desvio padrão da ROC AUC.	73
Figura 14 – Diagrama de diferença crítica considerando o número de folhas (escala logarítmica) da DT e das MTs.	74
Figura 15 – Diagrama de diferença crítica considerando o desvio padrão do número de folhas (escala logarítmica) da DT e das MTs.	74

Lista de tabelas

Tabela 1 – Conjunto de dados T no formato atributo-valor, com n exemplos e m atributos. A linha i refere-se ao i -ésimo exemplo ($i = 1, 2, \dots, n$) e a entrada x_{ij} refere-se ao valor do j -ésimo ($j = 1, 2, \dots, m$) atributo X_j do exemplo i	33
Tabela 2 – Conjunto exemplos hipotético sobre robôs amigos e inimigos contendo quatro atributos (Cabeça, Corpo, Segura e Sorri) e duas classes (amigo e inimigo).	36
Tabela 3 – Conjunto exemplos hipotético sobre jogar ou não jogar com base nas características climáticas contendo quatro atributos (Clima, Temperatura, Umidade e Ventando) e duas classes (sim e não).	53
Tabela 4 – Matriz de contingência para a folha (regra) $L \rightarrow R$	55
Tabela 5 – Conjunto com 43 datasets somente com atributos categóricos. Atr: número de atributos, Ex: número de exemplos, Cl: número de classes..	63
Tabela 6 – Conjunto com 51 datasets somente com atributos numéricos. Atr: número de atributos, Ex: número de exemplos, Cl: número de classes..	64
Tabela 7 – Conjunto com 56 datasets com atributos categóricos e numéricos. Cat: número de atributos categóricos, Num: número de atributos numéricos, Atr: número de atributos, Ex: número de exemplos, Cl: número de classes.	65
Tabela 8 – Valores AUC: árvore de decisão, meta-árvore e floresta - 43 Datasets categóricos	88
Tabela 9 – Valores AUC: árvore de decisão, meta-árvore e floresta - 51 Datasets numéricos	89
Tabela 10 – Valores AUC: árvore de decisão, meta-árvore e floresta - 56 Datasets categóricos e numéricos	91
Tabela 11 – Média dos valores de AUC e o ranking médio: árvore de decisão, meta-árvore e floresta - Todos os 150 Datasets	93
Tabela 12 – Número de folhas: árvore de decisão e meta-árvore - 43 Datasets categóricos	94
Tabela 13 – Número de folhas: árvore de decisão e meta-árvore - 51 Datasets numéricos	96
Tabela 14 – Número de folhas: árvore de decisão e meta-árvore - 56 Datasets categóricos e numéricos	98
Tabela 15 – Média do número de folhas e o ranking médio: árvore de decisão e meta-árvore - Todos os 150 Datasets	100

Sumário

Introdução	27
1 FUNDAMENTAÇÃO TEÓRICA	31
1.1 Aprendizado de Máquina	31
1.2 Árvore de Decisão	33
1.2.1 Métricas de Impureza	34
1.3 Tabela de Decisão	36
1.4 <i>Random Trees</i> e <i>Random Forests</i>	37
1.5 Meta-Aprendizado	38
1.6 <i>Explainable Artificial Intelligence</i>	40
1.7 Considerações Finais	41
2 TRABALHOS RELACIONADOS	43
2.1 <i>Combined Multiple Models</i>	44
2.2 <i>Merging Decision Trees</i>	46
2.3 Considerações Finais	48
3 INDUÇÃO DE META ÁRVORE DE DECISÃO	49
3.1 Visão Geral	49
3.2 Tratamento de Nós de Decisão de Atributos Numéricos	52
3.3 Métricas para o Peso de uma Folha	53
3.4 Normalização dos Pesos de cada Árvore	57
3.5 Considerações Finais	59
4 CONFIGURAÇÃO EXPERIMENTAL	61
4.1 Datasets	61
4.2 Algoritmos e Parametrização	62
4.3 Validação	66
4.4 Considerações Finais	66
5 RESULTADOS & DISCUSSÃO	67
5.1 Resultados	67
5.2 Discussão	68
5.2.1 Datasets Categóricos	68
5.2.2 Datasets Numéricos	68
5.2.3 Datasets Categóricos e Numéricos	69
5.2.4 Todos os Datasets	69

5.2.5	Tratamento de Atributos Numéricos	70
5.2.6	Ponderação das Folhas	70
5.3	Considerações Finais	70
6	CONCLUSÃO	75
6.1	Considerações Iniciais	75
6.2	Principais Contribuições	76
6.3	Trabalhos Futuros	77
	Referências	79
	APÊNDICES	85
	APÊNDICE A – RESULTADOS DOS EXPERIMENTOS . . .	87

Introdução

A estratégia de combinação de modelos (*ensemble*) proposta por Dietterich (1997), apresenta-se como uma alternativa para melhorar o desempenho dos modelos simbólicos. Esta estratégia apoia-se na ideia de que o desempenho de um conjunto de modelos considerados fracos é geralmente melhor do que um único modelo considerado forte (SIRIKULVIRIYA; SINTHUPINYO, 2011). A estratégia consiste em gerar diferentes modelos com base no mesmo conjunto de treinamento e combiná-los em um modelo final. Geralmente o processo de combinação ocorre por meio de algum mecanismo de votação, no qual o conjunto de modelos define, por votação majoritária, o rótulo dos novos exemplos. Além da estabilidade, a utilização da estratégia *ensemble* pode melhorar a precisão do modelo, de forma que o modelo resultante da combinação é geralmente mais preciso e mais estável do que um modelo individual.

Apesar dos algoritmos baseados em *ensemble* aperfeiçoarem o desempenho do modelo, para a categoria simbólica, há uma dificuldade com relação a interpretabilidade do modelo final, já que o mesmo é formado por diferentes modelos combinados. Mesmo considerando que os modelos sejam individualmente interpretáveis, quando o conjunto é muito grande, torna-se humanamente complexo o processo de interpretação do modelo resultante, até mesmo para os especialistas do domínio em questão.

Com base nisso, combinar o conjunto de modelos resultantes da estratégia *ensemble* em apenas um único modelo, poderia resgatar a interpretabilidade e também preservar, de certa forma, a precisão e a estabilidade. Neste projeto de pesquisa está sendo proposto um novo algoritmo para combinar as árvores de decisão geradas pelo algoritmo *Random Forest* em uma única árvore de decisão usando meta-aprendizado.

A partir de uma análise preliminar da literatura estão sendo propostos os Algoritmos 3 e 4, os quais realizam o processo de indução de uma meta-árvore de decisão com base em um conjunto de árvores de decisão geradas pelo algoritmo *Random Forest*.

Motivação

Por meio dos estudos realizados na área, caracterizam-se alguns critérios para avaliar os resultados produzidos pelo Aprendizado de Máquina. Um desses critérios, que segundo Domingos (1997) pode ser considerado um dos mais importantes, é a **compreensibilidade do resultado**. Para certas aplicações, devido aos riscos envolvidos, tem-se a necessidade de compreender o modelo antes de tomar alguma decisão. Para tais aplicações, fazer o modelo produzir um resultado preciso não é suficiente, o modelo também precisa ser humanamente comprehensível para considerá-lo útil, confiável e aceitável. Além disso, existem situações nas quais o objetivo principal é obter informações sobre o domínio e não apenas um resultado preciso. Até mesmo quando a precisão preditiva é o objetivo mais importante, a **compreensibilidade é um recurso crucial, pois facilita o processo de renovação interativa do modelo** (DOMINGOS, 1997).

Para aplicações não críticas como recomendações de filmes, produtos e conteúdos digitais, a não compreensão de como o modelo chegou ao resultado não oferece riscos ou impactos significativos caso o resultado obtido não seja tão bom. Porém, existem áreas, como por exemplo a medicina, onde o impacto de uma predição errada pode causar grandes prejuízos. Ainda que o modelo resultante seja utilizado somente como apoio à tomada de decisão, o fato de não conseguir compreender como se chegou àquele resultado é um fator que inviabiliza sua utilização (RIBEIRO; SINGH; GUESTRIN, 2016a). Além disso, com a popularização da Inteligência Artificial, novas leis surgiram para controlar o seu uso, como fez a União Européia ao criar no Regulamento Geral sobre a Proteção de Dados (GDPR, do inglês *General Data Protection Regulation*); o artigo 22, que define o direito de explicação, garante que qualquer pessoa afetada pela decisão de um algoritmo tenha o direito de saber o porquê aquela decisão foi tomada (UNION, 2016).

Objetivo Geral

O objetivo geral deste projeto de pesquisa consiste em desenvolver um novo algoritmo para combinar as árvores de decisão geradas pelo algoritmo *Random Forest* em uma única árvore de decisão utilizando meta-aprendizado; o modelo produzido pelo algoritmo proposto será referenciado como *Meta-Tree* (MT). Os objetivos específicos desta proposta consistem em:

- Avaliar se existe alguma relação de desempenho de uma meta-árvore de decisão, uma única árvore de decisão e uma floreta de árvores de decisão e
- Avaliar se existe alguma relação de tamanho de uma meta-árvore de decisão e uma única árvore de decisão.

Organização

O restante deste documento está organizado da seguinte forma: no Capítulo 1 é apresentada a fundamentação teórica seguida por trabalhos relacionados ao aqui apresentado no Capítulo 2. O projeto de pesquisa é apresentado no Capítulo 3. No Capítulo 4 são apresentadas a metodologia de avaliação experimental. No Capítulo 5 são apresentados os resultados obtidos e a discussão e, finalmente, no Capítulo 6 são apresentadas as conclusões e propostas atividades de pesquisa futuras em continuidade ao trabalho aqui desenvolvido.

Fundamentação Teórica

Nas seções seguintes são apresentados conceitos básicos sobre: Aprendizado de Máquina, Árvores de Decisão, Métricas de Impureza, Tabelas de Decisão, Combinação de Classificadores (*ensembles*) e Meta-Aprendizado. Tais conceitos, juntamente com os trabalhos relacionados apresentados no Capítulo 2, formam a base da proposta de pesquisa que é apresentada no Capítulo 3.

1.1 Aprendizado de Máquina

Segundo Weiss e Kulikowski (1991), o Aprendizado de Máquina (AM) é uma das subáreas que compõe a Inteligência Artificial. Seu objetivo é propor técnicas computacionais para construção de sistemas capazes de adquirir conhecimento de forma automática, tomando como premissa um conjunto de exemplos (experiência acumulada) que caracterizam as situações a serem aprendidas. O processo de aquisição de conhecimento baseia-se em um algoritmo de aprendizado, também chamado de indutor, que, por meio da inferência indutiva sobre os exemplos apresentados, cria uma hipótese (modelo) com a finalidade de classificar corretamente novos exemplos pertencentes àquele domínio. A inferência indutiva é uma das principais metodologias utilizadas para extrair conhecimento e predizer eventos futuros. Esse recurso é o mais utilizado pelo cérebro humano para gerar hipóteses ou induzir novos conhecimentos com base nas suas experiências.

O aprendizado pela inferência indutiva pode ocorrer de várias maneiras, uma delas é quando os exemplos do conjunto de treinamento são rotulados, ou seja, o exemplo, que é representado por um vetor de atributos, possui um rótulo definido. Esse tipo de aprendizado é denominado Aprendizado Supervisionado. Dentro deste, é importante ressaltar que quando os exemplos do conjunto de treinamento possuem rótulos discretos, denomina-se o processo como classificação, caso os exemplos possuam rótulos contínuos, denomina-se o processo como regressão. Uma outra maneira de realizar o aprendizado

indutivo é quando os exemplos do conjunto de treinamento não possuem rótulo. Para este caso, o algoritmo de aprendizado tenta identificar a relação que há entre os exemplos com o objetivo de agrupá-los. Normalmente após a criação dos grupos é realizado uma análise, com o auxílio de um especialista, para determinar o que cada grupo significa dentro do contexto analisado. Esse tipo de aprendizado é denominado Aprendizado Não-Supervisionado (BARANAUSKAS, 2001). Existem outros tipos de aprendizado, tais como: Aprendizado Semi-Supervisionado, Aprendizado por Reforço, Aprendizado Ativo, entre outros.

Segundo Michalski (1983), os algoritmos de AM podem ser categorizados como simbólicos e não-simbólicos. A categoria simbólica é caracterizada por representações do conhecimento que podem ser facilmente interpretadas pelo homem em uma escala de entendimento que abrange desde o nível do senso comum até o nível especialista. A categoria não-simbólica, também conhecida como caixa-preta, caracteriza-se por representações que não são facilmente interpretadas pelo homem. Para esta categoria o algoritmo desenvolve sua própria representação do conhecimento o que geralmente não fornece nenhum esclarecimento, dificultando assim a sua compreensão. Os algoritmos de aprendizado simbólico colaboram muito para a compreensão do conhecimento induzido, porém, geralmente, são menos precisos se comparados com os algoritmos de aprendizado não-simbólicos, os quais, tendem a ter uma precisão maior, porém não possuem representações de fácil interpretação. A escolha de qual algoritmo utilizar está relacionada com os objetivos da análise que será realizada.

Mitchell (1997) define Aprendizado de Máquina como um programa de computador que aprende a partir de uma experiência D com respeito a alguma classe de tarefas \mathcal{T} e medida de performance A , se sua performance nas tarefas em \mathcal{T} , medida por A , melhora com a experiência D .

Formalmente, um problema de aprendizado \mathcal{P} pode ser denotado pelo par $\mathcal{P} = (D, H)$, com $D \in \mathcal{D}$, onde \mathcal{D} representa o espaço de descrição de exemplos ou espaço de dados e H representa o espaço de descrição de hipóteses ou modelos (GRABCZEWSKI, 2014). Dessa forma, um problema de aprendizado pode ser visto como um problema de seleção de hipóteses. Geralmente o espaço de descrição de exemplos é representado por sequências de pares atributo-valor e, como descrito anteriormente, o objetivo do aprendizado é encontrar um mapeamento do espaço de atributos para o espaço de valores, refletindo melhor a relação entre eles.

Assim, os dados brutos coletados são transformados no formato atributo-valor definido na Tabela 1, também conhecido como formato padrão (REZENDE, 2003). Embora outros formatos possam ser definidos, o formato adotado como padrão representa os dados brutos de uma forma simples e uniforme, que é utilizado universalmente por várias técnicas de Mineração de Dados que utilizam algoritmos de Aprendizado de Máquina.

Tabela 1 – Conjunto de dados T no formato atributo-valor, com n exemplos e m atributos.

A linha i refere-se ao i -ésimo exemplo ($i = 1, 2, \dots, n$) e a entrada x_{ij} refere-se ao valor do j -ésimo ($j = 1, 2, \dots, m$) atributo X_j do exemplo i .

Exemplo	X_1	X_2	...	X_m	Y
z_1	x_{11}	x_{12}	...	x_{1m}	y_1
z_2	x_{21}	x_{22}	...	x_{2m}	y_2
\vdots	\vdots	\vdots	..,	\vdots	\vdots
z_n	x_{n1}	x_{n2}	...	x_{nm}	y_n

Em aprendizado de máquina exemplos são tuplas $z_i = (x_{i1}, x_{i2}, \dots, x_{im}, y_i) = (\vec{x}_i, y_i)$ também denotados por (x_i, y_i) , onde fica subentendido o fato que x_i é um vetor. A última coluna, $y_i = f(x_i)$, é a função que tenta-se predizer a partir dos atributos. Observa-se que cada x_i é um elemento do conjunto $X_1 \times X_2 \times \dots \times X_m$ e quando se trata de classificação y_i pertence a uma das k classes ($y_i \in \{C_1, C_2, \dots, C_k\}$), no caso de regressão y_i assume valores reais.

Dado um conjunto de exemplos de treinamento, um indutor (algoritmo de aprendizado ou máquina de aprendizado) gera como saída um classificador (também denominado hipótese ou modelo ou descrição de conceito) de forma que, dado um novo exemplo, ele possa predizer com a maior precisão possível a sua classe. Do ponto de vista formal, um algoritmo ou máquina de aprendizado M é um processo $M : K^{(M)} \times \mathcal{D} \rightarrow H$, onde $K^{(M)}$ é o espaço de parâmetros de configuração de M .

Formalmente, em classificação, um exemplo é um par $(x_i, f(x_i))$ onde x_i é a entrada e $f(x_i)$ é a saída. A tarefa de um indutor L é, dado um conjunto de exemplos, induzir uma função $h(\cdot) \in H$ que aproxima $f(\cdot)$, normalmente desconhecida. Neste caso, $h(\cdot)$ é chamada uma hipótese sobre a função objetivo $f(\cdot)$, ou seja, $h(x_i) \approx f(x_i)$.

1.2 Árvore de Decisão

Segundo Breiman et al. (1984), Quinlan (1986), a construção de uma árvore de decisão realiza-se da seguinte forma, ilustrada pelo Algoritmo 1: Se a condição de terminação (linha 2) é atingida, o conjunto de treinamento é homogêneo, ou seja, possui somente exemplos de uma única classe, então o algoritmo retorna a classe majoritária daquele conjunto de treinamento (linha 3). Caso contrário, utilizando o conjunto de exemplos de treinamento, um atributo é escolhido de forma a partitionar os exemplos em subconjuntos, de acordo com valores deste atributo (linha 5). Para cada subconjunto, outro atributo é escolhido para partitionar novamente cada um deles (linhas 7–10). Este processo prossegue, enquanto um dos subconjuntos contenha uma mistura de exemplos pertencendo a classes diferentes. Uma vez obtido um subconjunto uniforme em que todos os exemplos naquele

subconjunto pertencem à mesma classe, um nó folha é criado e rotulado com o mesmo nome da respectiva classe (linha 3).

Quando um novo exemplo deve ser classificado, começando pela raiz da árvore induzida, o classificador testa e desvia para cada nó com o respectivo atributo até que atinja uma folha. A classe deste nó folha será atribuída ao novo exemplo.

Algoritmo 1 Algoritmo de indução de uma árvore de decisão

Entrada: D ▷ conjunto original de exemplos
Saída: T ▷ árvore de decisão resultante

```

1: function buildTree( $D$ )
2: if terminationCondition( $D$ ) then
3:    $T \leftarrow \text{majorityClass}(D)$ 
4: else
5:    $A \leftarrow \text{selectBestAttribute}(D)$  ▷ seleciona o melhor atributo
6:    $T \leftarrow \text{newNode}(A)$ 
7:   for each outcome value  $O_i$  from  $A$  do
8:      $S_i \leftarrow \{\forall z \in D : A = O_i\}$ 
9:      $T.\text{subtree}[i] \leftarrow \text{buildTree}(S_i)$ 
10:  end for
11: end if
12: return  $T$ 

```

1.2.1 Métricas de Impureza

A chave para o sucesso de um algoritmo de aprendizado por árvores de decisão depende do critério utilizado para escolher o atributo que particiona o conjunto de exemplos em cada iteração. O desejável, nesse caso, é obter uma árvore pequena. Portanto, devemos maximizar a separação de classes em cada etapa, fazendo os sucessores de cada nó mais puros possíveis. Isso implica em caminhos mais curtos na árvore. Para isso, métricas de impureza são utilizadas, como descrito a seguir.

Seja T um conjunto de exemplos e p_j as proporções de exemplos em cada classe C_j em T ($j = 1, 2, \dots, k$) representado como $T = [p_1, p_2, \dots, p_k]$ onde $\sum_{j=1}^k p_j = \sum p_j = 1$.

Uma métrica é definida como medida de *impureza* $i(T)$ se satisfaz as seguintes propriedades:

- $i(T)$ é mínima somente quando $p_i = 1$ e $p_j = 0$ para $j \neq i$ (é mínima quando todos os exemplos são da mesma classe)
- $i(T)$ é máxima somente quando $p_j = \frac{1}{k}$ (é máxima quando há exatamente o mesmo número de exemplos de cada classe)

- $i(T)$ é simétrica em relação a p_1, p_2, \dots, p_k (não importa a ordem das classes, desde que as proporções sejam as mesmas, por exemplo, para duas classes $i([p_1, p_2]) = i([p_2, p_1])$).

Quando $i(T)$ atinge seu valor mínimo, T é denominado puro ou homogêneo ou sem mistura entre classes. Para qualquer valor de $i(T)$ acima do valor mínimo (incluindo o valor máximo), T é denominado impuro ou heterogêneo ou contendo mistura entre classes.

Dentre as métricas mais utilizadas de impureza $i(T)$ destacam-se:

- Entropia de Shannon (SHANNON, 1948), dada por $i(T) = -\sum_j p_j \log_b p_j$. Se $b = 2$ então unidade da entropia é em bits. Para k classes, o valor mínimo é zero e o máximo é $\log_b k$;
- Índice Gini (GINI, 1936), onde $i(T) = \sum_j p_j(1 - p_j) = 1 - \sum_j p_j^2$. Para k classes, o valor mínimo é zero e o máximo é $\frac{k-1}{k}$. Como pode ser observado, independentemente do número k de classes, o valor máximo é sempre inferior à unidade, tendendo à 1 quando $k \rightarrow +\infty$ e
- Erro majoritário, onde $i(T) = 1 - \max_j p_j$. Os valores mínimo e máximo são idênticos ao do índice Gini.

Assuma que T foi particionado em r valores do atributo X , ou seja, X assume os valores $X = O_1, X = O_2, \dots, X = O_r$, gerando os subconjuntos T_1, T_2, \dots, T_r , respectivamente, onde T_i é o formado pelos exemplos de T nos quais o atributo $X = O_i$, ou seja, $T_i = \{\forall z \in T : X = O_i\}$. A impureza esperada para este particionamento é a soma ponderada sobre todos os subconjuntos T_i que é dada por $i(T, X) = \sum_{i=1}^r \frac{|T_i|}{|T|} i(T_i)$, onde $|T|$ indica a cardinalidade do conjunto T . Para escolher um atributo para compor a árvore de decisão, essa avaliação é efetuada para todos os atributos X_j ($j = 1, \dots, m$) e escolhendo-se aquele com o menor valor de $i(T, X)$.

Por razões históricas, entretanto, costuma-se utilizar o ganho de impureza ao particionar o conjunto T pelo atributo X dado por $g(T, X) = i(T) - i(T, X)$. Pelo fato de a impureza ser vista como a quantidade de informação, em geral, essa quantidade é também conhecida como ganho de informação pela partição de T de acordo com o atributo X . Nesse caso, o critério de ganho (ou ganho máximo) seleciona o atributo $X \in T$ que maximiza o ganho de informação, $g_{\max}(T, X) = \arg \max_{X \in \{x_1, \dots, x_m\}} g(T, X)$.

1.3 Tabela de Decisão

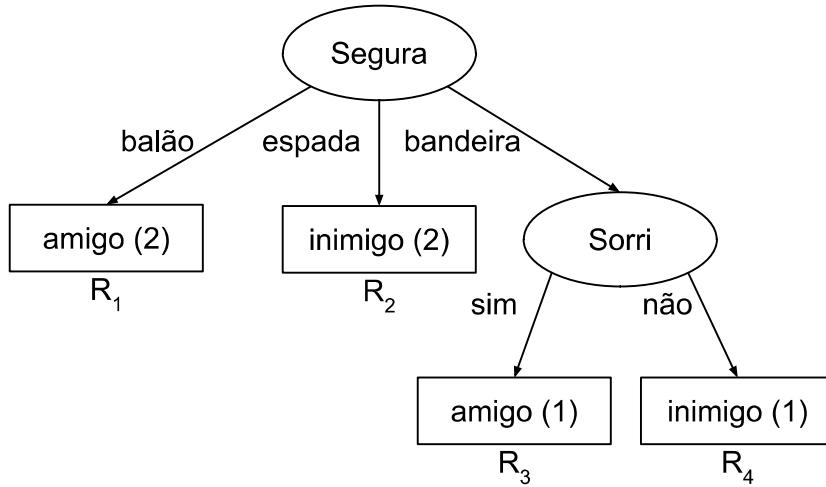
A tabela de decisão é um dos tipos de representação mais simples para hipóteses em aprendizado de máquina supervisionado. Esse tipo de representação permite que relações lógicas de grande complexidade sejam representadas facilmente, contribuindo assim, para a sua compreensão (CRAGUN; STEUDEL, 1987; KOHAVI; SOMMERFIELD, 1998; SANTOS-GOMEZ; DARNELL, 1992; SETHI; CHATTERJEE, 1980; VANTHIENEN; DRIES, 1994). Além disso, essa forma de representação é intercambiável com outros modelos de representação, como por exemplo: árvores de decisão, conjuntos de regras, bases de conhecimento tabulares (HEWETT; LEUCHNER, 2003). Dessa forma, é possível realizar a conversão entre os tipos de representação sem que o resultado seja prejudicado. A escolha do tipo de representação utilizada está relacionada com as necessidades de compreensão do modelo. Dependendo do contexto analisado, certas representações contribuem mais para o entendimento.

A utilização das tabelas de decisão aumentou quando se percebeu que esse tipo de representação poderia ser utilizada para validar a estrutura de uma árvore de decisão, facilitando a verificação da exatidão, consistência, integridade e redundância das árvores, o que contribui diretamente para a validação do conhecimento (HAMILTON; KELLEY; CULBERT, 1991; NONFJALL; LARSEN, 1992; GOUL et al., 1990).

Na Figura 1 é fornecido um exemplo de transformação de uma árvore de decisão para uma tabela de decisão. A árvore de decisão em questão foi gerada a partir de conjunto de exemplos da Tabela 2. Na Figura 1, observa-se que cada caminho da raiz da árvore até uma folha é representada por uma linha na tabela de decisão e corresponde a uma região no espaço dos atributos. Os números entre parênteses indicam o número de exemplos contidos na folha, que é representado na tabela de decisão como a coluna ‘Peso’. O número de folhas na árvore de decisão corresponde ao número de linhas na tabela de decisão. Atributos que não aparecem na árvore são representados pelo símbolo ‘?’ na tabela de decisão.

Tabela 2 – Conjunto exemplos hipotético sobre robôs amigos e inimigos contendo quatro atributos (Cabeça, Corpo, Segura e Sorri) e duas classes (amigo e inimigo).

Exemplo	Cabeça	Corpo	Segura	Sorri	Classe
z_1	redonda	quadrado	bandeira	não	inimigo
z_2	triangular	triangular	balão	sim	amigo
z_3	redonda	redondo	bandeira	sim	amigo
z_4	quadrada	triangular	espada	não	inimigo
z_5	quadrada	quadrado	balão	sim	amigo
z_6	triangular	redondo	espada	sim	inimigo



(a) Árvore de Decisão

Região	Atributos				Classe	Peso
	Cabeça	Corpo	Segura	Sorri		
R_1	?	?	balão	?	amigo	2.00
R_2	?	?	espada	?	inimigo	2.00
R_3	?	?	bandeira	sim	amigo	1.00
R_4	?	?	bandeira	não	inimigo	1.00

(b) Tabela de Decisão

Figura 1 – Exemplo de transformação de Árvore de Decisão (a) para Tabela de Decisão (b).

1.4 Random Trees e Random Forests

Considere um conjunto de treinamento T com m atributos e n exemplos, seja T_l uma amostra *bootstrap* (EFRON; TIBSHIRANI, 1993) do conjunto de treinamento a partir de T com reposição, contendo n exemplos. Uma *Random Tree* é construída da seguinte forma: em cada nó da árvore é escolhido um atributo a partir de a atributos aleatórios, onde $a \leq m$. Desta forma, ao criar diversas *Random Trees*, cada uma delas possui a mesma probabilidade de ser amostrada. A combinação dessa diversidade de *Random Trees* é conhecida como *Random Forest*, detalhada a seguir.

Uma *Random Forest* é definida formalmente como um classificador composto por uma coleção de árvores $\{h_l(x)\}, l = 1, 2, \dots, L$, onde T_l são amostras aleatórias independentes e de distribuição idênticas. Cada árvore prediz a classe da entrada x e, em seguida, a classe mais popular entre as árvores é eleita para esta mesma entrada (BREIMAN, 2001; DUBATH et al., 2011; ZHAO; ZHANG, 2008).

Portanto, *Random Forests* aplicam o mesmo método que o *bagging* (BREIMAN, 1996a) para produzir amostras aleatórias de conjuntos de treinamento (amostras *bootstrap*) para cada árvore. Breiman (BREIMAN, 2001) justifica o uso do método *bagging* em *Random Forests* por duas razões: o uso do *bagging* parece melhorar o desempenho quando atributos aleatórios são usados; *bagging* pode ser usado para fornecer estimativas contínuas do erro de generalização do conjunto combinado de árvores, assim como estimativas para força e correlação, usando o estimador *out-of-bag*.

O erro de classificação da floresta depende da força das árvores individuais da floresta, da correlação entre quaisquer duas árvores na floresta e da importância dos atributos (*variable importance*) (BREIMAN, 2001; BREIMAN, 2004; BREIMAN; CUTLER, 2004; MA; GUO; CUKIC, 2007), a saber:

- Força da árvore individual na floresta: pode ser interpretada como uma medida de desempenho para cada árvore. Uma árvore com uma taxa de erro baixa é um classificador forte. Assim, aumentando a força das árvores individuais, reduz-se a taxa de erro da floresta.
- Correlação entre as árvores da floresta: duas medidas de aleatoriedade (uso do *bagging* e seleção aleatória de atributos) fazem com que as árvores sejam diferentes e, portanto, diminui a correlação entre elas. A baixa correlação tende a diminuir a taxa do erro de classificação.
- Importância dos atributos: Após a construção da floresta, um dos atributos do conjunto de dados tem seus valores permutados nos exemplos *out-of-bag* e, após a permutação, os exemplos são apresentados à respectiva árvore e, por fim, é comparada a taxa de acerto na classificação com e sem a permutação deste atributo. Este processo de permutação é repetido para cada atributo do *dataset*. Quão maior o aumento na taxa de erro gerado pela permutação de um atributo, maior é a importância deste atributo para a representação da classe nesse conjunto de dados.

1.5 Meta-Aprendizado

Quando um algoritmo de aprendizado ou o modelo induzido por um algoritmo de aprendizado é analisado, tal análise está ocorrendo em um meta-nível (GRABCZEWSKI, 2014). Nesta pesquisa o termo meta-aprendizado se refere a toda forma de aprendizado a partir de informações sobre processos e modelos de aprendizado, sendo um termo muito geral.

Formalmente, problemas de meta-aprendizagem se encaixam na definição do problema de aprendizado (Seção 1.1 na página 31). Eles representam apenas um tipo específico de aprendizado (com espaços de representação de dados e modelos focados em outros

processos de aprendizado). Para tornar clara a distinção entre os dois níveis de aprendizado, os processos analisados em meta-nível serão referidos como máquina de aprendizado em meta-objeto ou meta-nível.

O objetivo da análise de meta-nível é aprender como aprender e aplicar o conhecimento induzido em novos aprendizados para obter resultados mais interessantes. Esta é uma etapa fundamental em um processo complexo, cujo o objetivo consiste em encontrar melhores modelos, descrito por

$$P_0 = (D, h, q). \quad (1.1)$$

Resolver o problema P_0 , como descrito em 1.1, consiste em um problema de seleção de hipóteses: encontrar uma hipótese $h \in H$ com objetivo de maximizar $q(h)$, sendo q uma métrica de qualidade.

Segundo (VANSCHOREN, 2018), para resolução do problema de meta-aprendizado é necessário primeiramente coletar os metadados que descrevem as tarefas iniciais \mathcal{T} e as respectivas hipóteses h geradas. Os metadados em questão abrangem todos os aspectos da tarefa de aprendizado, como hiperparâmetros, resultados de avaliações da hipótese ou propriedades mensuráveis da tarefa (*meta-features*). Assim, a partir de metadados, pode-se, aprender com experiencias passadas.

Alguns algoritmos de aprendizado podem ser complexos, por exemplo, sendo compostos por vários outros algoritmos de aprendizado, como no caso de um *ensemble*. Portanto, assume-se que uma combinação de algoritmos de aprendizado complexos também é um algoritmo de aprendizado. Geralmente, algoritmos complexos retornam modelos complexos (modelos compostos de outros modelos). Em termos formais, um algoritmo que faz parte de outro algoritmo é denominado algoritmo filho ou subalgoritmo; o outro algoritmo, portanto, é chamado de algoritmo pai ou superalgoritmo.

Diferentes visões sobre meta-aprendizado podem ser encontradas na literatura científica. Vários processos de aprendizado que adquirem algum conhecimento sobre outros algoritmos ou exploraram esse tipo de conhecimento são referidos como métodos de meta-aprendizado. *Ensembles* simples que utilizam voto majoritário ou outras combinações simples não aprendem em meta-nível, embora contenham o módulo de decisão que lida com os modelos. Algoritmos mais ‘inteligentes’ realizam algumas meta-análises para decidir quais decisões dos componentes do *ensemble* merecem mais confiança e quais devem ser ignorados por estarem provavelmente errados. Essas abordagens são comumente denominadas de empilhamento de modelos (*model stacking*) justamente pelo fato que no topo dos modelos encontra-se o módulo de decisão, que é um meta-algoritmo de aprendizado.

Existem muitas publicações sobre tais modelos, como por exemplo Stolfo et al. (1997), Zenko, Todorovski e Džeroski (2001), Todorovski e Džeroski (2003), Prodromidis,

Chan e Stolfo (2000). Trabalhos que abordam *ensembles* de árvores de decisão têm especial interesse para essa pesquisa, conforme o que se encontra nos próximos capítulos.

1.6 *Explainable Artificial Intelligence*

Interpretabilidade e explicabilidade são frequentemente usadas de forma intercambiável na literatura, mas alguns artigos fazem distinção. Em Montavon, Samek e Müller (2018) a interpretação é o mapeamento do conceito abstrato em um domínio que os humanos podem entender, enquanto a explicação é a coleção de características do domínio interpretável que contribuíram para que um dado exemplo produza uma decisão. Edwards e Veale (2017) dividem as explicações em centradas no modelo e centradas no sujeito, noções que correspondem às definições de interpretabilidade e explicabilidade de Montavon, Samek e Müller (2018). Papéis semelhantes em Doshi-Velez e Kim (2017) assumem interpretabilidade global e local, respectivamente. Nessa visão, podemos ver que o GDPR cobre apenas a explicabilidade. Compreensibilidade é usado na literatura como sinônimo de interpretabilidade (FREITAS, 2014). Transparência é usado como sinônimo de interpretabilidade do modelo, que é algum sentido de compreensão da lógica de funcionamento do modelo.

Existem duas categorias de abordagens para interpretabilidade e explicabilidade: integrada (baseada em transparência) e post-hoc.

- A transparência é uma das propriedades que podem possibilitar a interpretabilidade. A transparência foi um primeiro passo tradicional para a proteção de direitos em instituições de base humana e, por analogia, é transportada para preocupações algorítmicas, como injustiça e discriminação (EDWARDS; VEALE, 2017). Mas, os modelos em IA estão se tornando muito mais complexos do que os baseados em humanos instituições e torna-se difícil encontrar uma explicação significativa que os usuários possam entender. Além disso, o pensamento humano não é transparente e as justificativas na forma de explicações e interpretações podem diferir do mecanismo de decisão real. Além disso, desempenho preditivo e transparência são objetivos conflitantes e devem ser negociados em um modelo (JIN; SENDHOFF, 2008). Em (MARCUS, 2018) afirma-se que não está claro o quanto a transparência importa a longo prazo. Se os sistemas forem robustos e independentes, pode não ser necessário. Mas, se fizerem parte de outros sistemas, a transparência pode ser boa para a depuração.
- A interpretabilidade post-hoc extrai informações do modelo já aprendido e não depende precisamente de como o modelo funciona. A vantagem dessa abordagem é que ela não afeta o desempenho do modelo que é tratado como uma caixa preta (BB). Este é um modo semelhante ao modo como as pessoas justificam suas próprias decisões,

sem conhecer completamente o real funcionamento de seus mecanismos de tomada de decisão. No entanto, cuidados especiais devem ser tomados para evitar sistemas que gerem explicações plausíveis, mas enganosas. Tais explicações podem satisfazer leis como o GDPR, mas há um problema de verificar sua veracidade (DOŠILOVIĆ; BRČIĆ; HLUPIĆ, 2018).

Segundo Gunning (2017), a eXplainable AI (XAI) propõe a criação de um conjunto de técnicas de ML que:

- Produz modelos mais explicáveis, mantendo um alto nível de desempenho de aprendizado
- Permite que os humanos entendam, confiem adequadamente e gerenciem efetivamente a geração emergente de parceiros artificialmente inteligentes.

1.7 Considerações Finais

Neste capítulo foram apresentados conceitos básicos sobre Aprendizado de Máquina, incluindo Meta-Aprendizado. Esses conceitos, juntamente com aqueles apresentados no capítulo seguinte, formam a base da proposta de pesquisa que é apresentada no Capítulo 3.

2

Trabalhos Relacionados

Nas seções seguintes são apresentados os trabalhos encontrados na literatura que estão relacionados ao algoritmo proposto. Os trabalhos mencionados tratam-se de abordagens para combinação de múltiplos modelos, tais como: *Combined Multiple Models* (CMM) na Seção 2.1 e o *Merging Decision Trees* (MDT) na Seção 2.2.

Além das referências básicas utilizadas nesta pesquisa (STRECHT; MENDES-MOREIRA; SOARES, 2014; DOMINGOS, 1997), também foram encontrados na literatura mais trabalhos relacionados ao aqui desenvolvido, tais como: LIME (Local Interpretable Model-Agnostic Explanations) (RIBEIRO; SINGH; GUESTRIN, 2016b) e LORE (Local Rule-based Explanations) (GUIDOTTI et al., 2019).

Segundo o trabalho realizado por Ribeiro, Singh e Guestrin (2016b), o LIME é uma técnica que explica as previsões de qualquer classificador de maneira interpretável e fiel, aprendendo um modelo interpretável localmente em torno da previsão (exemplo de teste). Como resultado deste trabalho, foi possível demonstrar a flexibilidade do método proposto explicando diferentes modelos, como por exemplo *Random Forest* e Redes Neurais na aplicação em diferentes contextos como, interpretação de texto e análise de imagens.

O trabalho de Guidotti et al. (2019) propôs um método de explicação baseado em regras locais, denominado como LORE. O algoritmo tem como objetivo fornecer explicações sobre a decisão tomada por um classificador tipo caixa-preta em um exemplo específico. Os resultados obtidos mostraram que o método proposto supera as abordagens existentes em termos da qualidade das explicações e da precisão em relação aos algoritmos utilizados na abordagem de caixa-preta.

Pesquisas detalhadas sobre a explicabilidade dos modelos podem ser encontradas em (BURKART; HUBER, 2021; GUIDOTTI et al., 2018; LINARDATOS; PAPASTEFA-NOPOULOS; KOTSIANTIS, 2021).

2.1 *Combined Multiple Models*

Para a combinação de classificadores, existem metodologias propostas pelos pesquisadores da área que abordam diferentes formas para realizá-la, como por exemplo: *Stacking* (WOLPERT, 1992), *Bagging* (BREIMAN, 1996b), *Bayesian Average* (BUNTINE, 1992), *Boosting* (FREUND; SCHAPIRE et al., 1996). O principal foco dessas metodologias é reduzir os problemas relacionados a instabilidade e conseguir melhorar precisão dos classificadores. A comprehensibilidade do resultado, para essas abordagens, não foi tratado de forma profunda.

Existem outras metodologias que, além de focar na redução da instabilidade e no aumento do ganho precisão, focam também no aumento do nível de comprehensibilidade do modelo. O algoritmo denominado *Combined Multiple Models* (CMM) criado por Domingos (1997) apresenta-se como uma abordagem para alcançar este objetivo. O CMM concentra-se na combinação de modelos baseados em árvores de decisão. Segundo Domingos (1997), apesar de uma única árvore de decisão ser facilmente compreendida por um humano, desde que não seja muito grande, cinquenta dessas árvores, mesmo que individualmente simples, excedem a capacidade até mesmo de um especialista. Essa foi a principal motivação que sustentou ideia de que a combinação de múltiplas árvores em apenas uma única recuperaria a facilidade de comprehensão por se tratar de apenas uma árvore e não mais de várias árvores.

Segundo Domingos (1997), um modelo é preciso à medida que seu resultado coincide com o resultado real. É estável na medida em que produz um mesmo resultado por meio de conjuntos de treinamento diferentes pertencentes ao mesmo domínio. É comprehensível na medida em que oferece a capacidade de ser humanamente compreendido. Apesar de ainda parecer elusivo alcançar esses três objetivos (precisão, estabilidade e comprehensibilidade) simultaneamente, os resultados foram significativos.

O Algoritmo 2 refere-se a proposta do CMM adaptada à implementação original. A partir das diferentes amostras extraídas do conjunto original de exemplos D , são geradas L árvores de decisão (linhas 2-5). Por se tratar de diferentes amostras, cada árvore criada apresentará uma estrutura exclusiva, diferenciando-se uma da outra. Posteriormente a criação das L árvores, o algoritmo cria, de forma aleatória, Z novos exemplos não classificados, ou seja, exemplos não possuem ainda uma classe definida. Cada novo exemplo gerado é submetido as L árvores criadas. Este procedimento é repetido de forma de que todos os novos exemplos recebam uma classe (linhas 7-11). O resultado da classificação realizada pelas L árvores são agregados em D_{new} por meio de votação, utilizando-se o critério do voto majoritário. Em caso de empate, optou-se por classificar pela classe minoritária (BREIMAN, 1996b). A árvore de decisão resultante será induzida a partir da nova amostra obtida, que é formada pela junção dos conjuntos D (conjunto

original de exemplos) e D_{new} (conjunto dos novos exemplos classificados) (linha 12).

Algoritmo 2 CMM (DOMINGOS, 1997)

Entrada: D ▷ conjunto original de exemplos rotulados
Entrada: L ▷ número de árvores que serão geradas
Entrada: Z ▷ número de exemplos aleatórios que serão gerados
Saída: T ▷ árvore de decisão resultante

```

1: function CMM( $D, L, Z$ )
2: for  $l \in \{1, 2, \dots, L\}$  do
3:    $a_l \leftarrow \text{selectSample}(D)$  ▷ Seleciona uma amostra do conjunto  $D$ 
4:    $M \leftarrow M \cup \{\text{buildTree}(a_l)\}$  ▷ Cria uma árvore de decisão baseada na amostra  $a_l$  usando o Algoritmo 1, adicionando-a em  $M$ 
5: end for
6:  $D_{\text{new}} \leftarrow \emptyset$ 
7: for  $z \in \{1, 2, \dots, Z\}$  do
8:    $x \leftarrow \text{generateRandomExample}()$  ▷ Cria um novo exemplo aleatoriamente
9:    $y \leftarrow M.\text{classify}(x)$  ▷ Classifica o novo exemplo  $x$  utilizando os classificadores contidos em  $M \{M_1, \dots, M_L\}$ 
10:   $D_{\text{new}} \leftarrow D_{\text{new}} \cup \{(x, y)\}$  ▷ Adiciona o novo exemplo rotulado em  $D_{\text{new}}$ 
11: end for
12:  $T \leftarrow \text{buildTree}(D \cup D_{\text{new}})$  ▷ Cria a árvore de decisão final a partir da união dos conjuntos  $D$  e  $D_{\text{new}}$ 
13: return  $T$ 

```

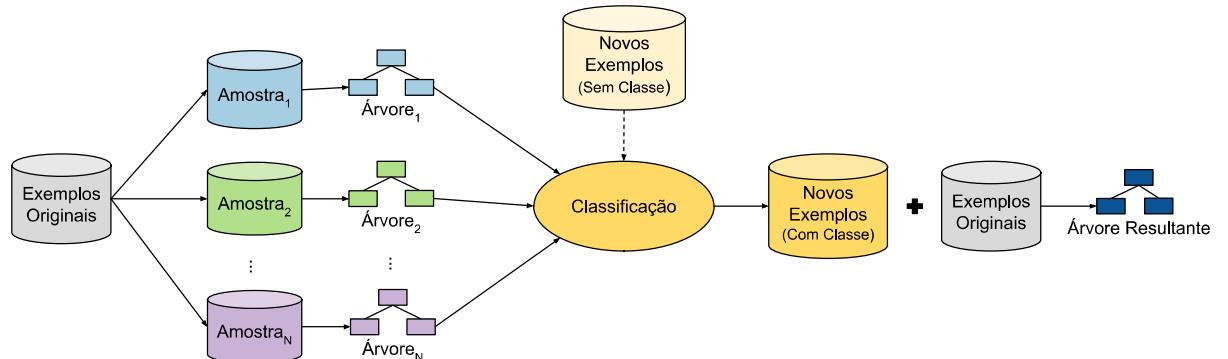


Figura 2 – CMM adaptado de Domingos (1997).

Em correspondência ao Algoritmo 2, a Figura 2 ilustra a implementação do CMM. Da esquerda para a direita temos, inicialmente, a extração das amostras do conjunto de exemplo originais. Para cada amostra extraída, é feita a criação de uma árvore de decisão. Posteriormente, são criados novos exemplos aleatórios que serão classificados pelas árvores de decisão geradas. Feito isso, os novos exemplos aleatórios, agora classificados, são unidos com o conjunto de exemplos originais para dar origem a árvore de decisão resultante.

Em relação aos resultados, segundo os autores, o desempenho do CMM foi comparado com duas outras abordagens, que foram: a utilização de uma única árvore gerada a partir do conjunto original de exemplos e uma abordagem denominada *Bagging* (BREIMAN, 1996b). Foram utilizadas 26 bases de dados (LICHMAN, 2013) para realizar esta

análise. A precisão média obtida pelo CMM foi 2,1% maior comparado ao modelo único. O CMM retém em média 60 % dos ganhos de precisão obtido pela abordagem *Bagging*. Em 6 dos 26 conjuntos o CMM apresentou maior precisão do que a abordagem *Bagging*. A respeito da comprehensibilidade do modelo resultante, em todos os casos o CMM superou a abordagem *Bagging*, onde a métrica utilizada foi o tamanho do modelo gerado. Em todos os 26 conjuntos de dados a CMM é mais estável que o modelo único.

2.2 *Merging Decision Trees*

Outra metodologia proposta é a *Merging Decision Trees* (MDT) criada por Strecht, Mendes-Moreira e Soares (2014). Trata-se, também, de uma metodologia voltada exclusivamente para os casos onde se utilizam classificadores baseados em árvores de decisão. A metodologia foi criada para ser utilizada em um projeto realizado na Universidade de Porto, que tem como objetivo identificar antecipadamente os casos de desistência dos alunos em relação aos cursos oferecidos pela universidade. A finalidade era criar estratégias para evitar essas desistências e por meio dos modelos criados identificar os motivos e razões que levam o aluno desistir do curso. Para a implementação do projeto foram utilizados os dados de cada curso que estão armazenados na base de dados da universidade.

Devido as informações serem distintas para cada curso, as árvores de decisão geradas apresentaram diferenças. Para isso, foi proposta uma abordagem para combinar as diferentes árvores geradas. A abordagem é dividida em duas etapas, a primeira etapa realiza o agrupamento das árvores e a segunda etapa é responsável por fundi-las transformando-as em uma única árvore resultante. O processo de fusão consiste em cruzar as regras de decisão de pares de modelos de um grupo de forma recursiva.

Na Figura 3 é ilustrada a implementação do MDT. O processo ① realiza a extração das amostras de cada curso que estão armazenadas no banco de dados da universidade. O processo ② cria as árvores de decisão com base nas amostras extraídas. O processo ③ realiza a transformação das árvores de decisão em tabelas de decisão. O processo ④ realiza o agrupamento das tabelas de decisão. O processo ⑤ realiza a fusão das tabelas de decisão agrupadas. O processo ⑥ realiza a transformação das tabelas de decisão fundidas novamente em árvores de decisão. Finalmente, o processo ⑦ avalia as árvores de decisão resultantes do ponto de vista de melhoria de desempenho tomando como base as amostras obtidas no início do processo.

No projeto foram analisados os dados correspondentes ao ano letivo de 2012/2013. Foram extraídos 5779 conjuntos de dados de 391 programas oferecidos pela universidade. Para a criação das árvores de decisão, trabalhou-se apenas com os cursos que possuem no mínimo 100 alunos matriculados. Com esse filtro foram criados 730 árvores, uma para

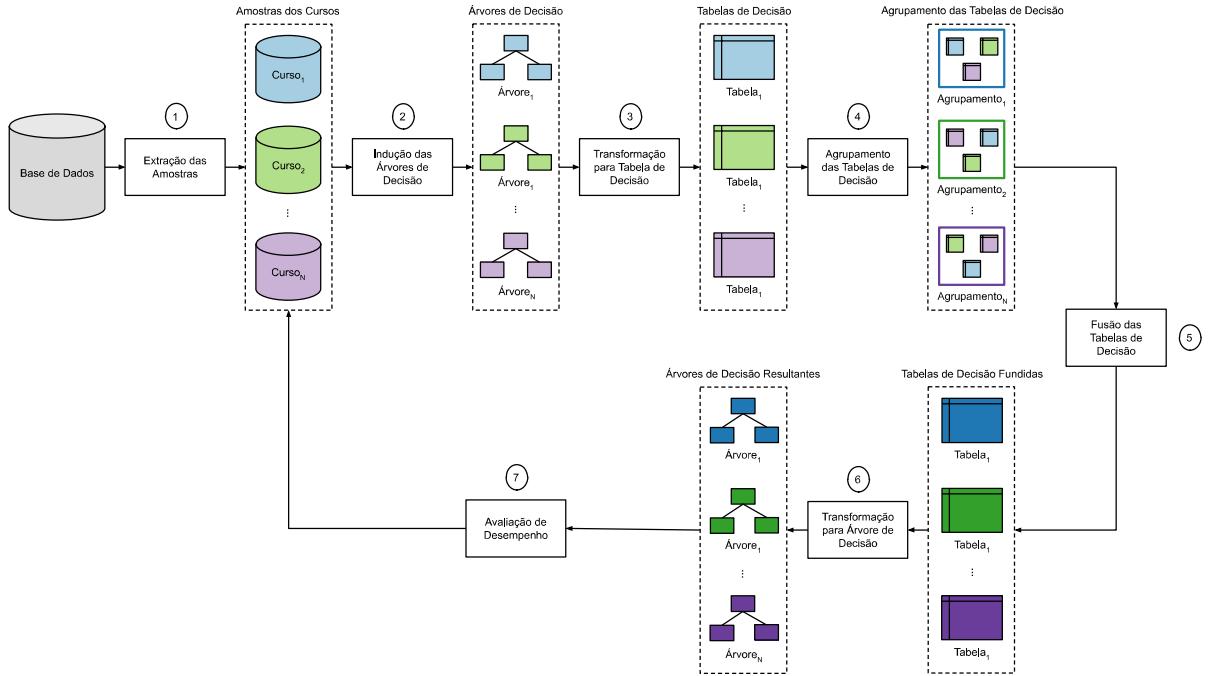


Figura 3 – MDT adaptado de Strecht, Mendes-Moreira e Soares (2014).

cada curso, que corresponde a 12% dos 5779 cursos oferecidos pela universidade. Na fase de agrupamento das árvores de decisão, foram utilizados quatro critérios: agrupamento por áreas científicas, agrupamento pelo número de variáveis, agrupamento pela importância das variáveis e um agrupamento composto por todas árvores criadas. A intenção é identificar qual agrupamento apresenta o melhor desempenho.

Na Figura 4 é ilustrada a fusão das árvores contidas em um agrupamento. O primeiro processo realiza a transformação das árvores de decisão (a_1, \dots, a_L) para as tabelas de decisão (t_1, \dots, t_L). O segundo processo combina as tabelas de decisão agrupadas por meio da intersecção entre elas da seguinte forma:

$$T_1 \leftarrow t_1 \cap t_2$$

$$T_2 \leftarrow T_1 \cap t_3$$

$$T_3 \leftarrow T_2 \cap t_4$$

⋮

$$T_{L-2} \leftarrow T_{L-3} \cap t_{L-1}$$

$$T_{L-1} \leftarrow T_{L-2} \cap t_L$$

Ou seja, a primeira (t_1) e a segunda (t_2) tabelas são fundidas, produzindo uma tabela fusão T_1 . Então a terceira tabela t_3 é fundida com a tabela T_1 dando origem à nova tabela T_2 . Esse processo é repetido até que todas as tabelas sejam fundidas em apenas

uma única tabela final T_{L-1} . O terceiro e último processo é responsável por converter a tabela resultante T_{L-1} em uma árvore de decisão.

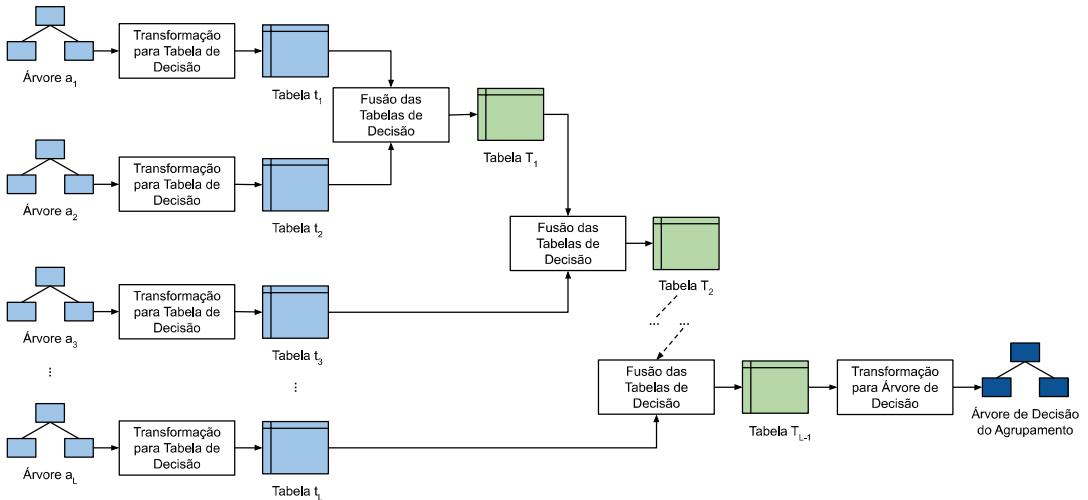


Figura 4 – Fusão do MDT adaptado de Strecht, Mendes-Moreira e Soares (2014).

Em relação aos resultados, segundo os autores, a partir de uma perspectiva de capacidade de fusão, a fusão por áreas científicas não é necessariamente a melhor maneira de agrupar as árvores, o agrupamento pelo número de variáveis e o agrupamento pela importância das variáveis parecem ser abordagens mais adequadas, considerando a média de fusão obtida por cada grupo. Já na perspectiva da melhoria no poder preditivo, o agrupamento composto por todas as árvores criadas teve uma melhora de 0,04%, que foi a maior média obtida. Já o agrupamento por áreas científicas produziu resultados interessantes, obtendo uma melhoria de 0,01% no poder preditivo, além de representar um indicativo de que árvores de decisão obtidas por meio de cursos com conteúdo semelhante são mais suscetíveis a generalizar conhecimento do que aqueles em que a similaridade surge de características dos próprios modelos (como o número de variáveis ou sua importância). No caso dos agrupamentos por número de variáveis e importância das variáveis houve uma piora no poder preditivo.

2.3 Considerações Finais

Neste capítulo foram apresentados trabalhos encontrados na literatura que estão relacionados à proposta de pesquisa que é apresentada no capítulo seguinte.

3

Indução de Meta Árvore de Decisão

Nas seções seguintes é apresentado o projeto de pesquisa na qual foi desenvolvido um algoritmo para induzir uma árvore de decisão em meta-nível, que será referenciada como *Meta-Tree* (MT). Este algoritmo utiliza as árvores de decisão induzidas pelo algoritmo *Random Forest* a partir de um conjunto de exemplos.

Essas árvores de decisão geradas são denominadas de nível zero, representadas por L_0 , na qual o valor zero indica que não ocorreu meta-aprendizado em sua indução. Utilizando meta-aprendizado nas árvores L_0 será gerada árvore final de nível um, representada por L_1 , indicando o primeiro nível de meta-aprendizado.

A proposta encontra-se representada pelos Algoritmos 3 e 4 descritos abaixo. Os detalhes de cada algoritmo são descritos a seguir.

3.1 Visão Geral

A proposta geral de pesquisa encontra-se representada sob a forma do Algoritmo 3 em alto nível. De acordo com o Algoritmo 3, inicialmente, é gerada uma floresta de árvores de decisão F utilizando o algoritmo *Random Forest* contendo L árvores considerando todos os m atributos do conjunto de exemplos D (linha 2). O número mínimo de $L = 128$ árvores é proveniente de resultado de pesquisa anterior realizada por Oshiro, Perez e Baranauskas (2012).

Em seguida, cada árvore da floresta é representada por uma tabela de decisão individual (linhas 4–7). O método `Tree2Table` (linha 5) faz a conversão de cada árvore da floresta para sua respectiva tabela de decisão conforme Algoritmo 4. O Algoritmo 4 é responsável identificar cada folha contida na árvore, considerando as seguintes informações: as ramificações de subárvore que formam o caminho da raiz até a folha, a classe e o número de exemplos contidos na respectiva folha (linhas 4–12). Para cada folha identificada na árvore de decisão, são armazenadas as informações no vetor r (linhas 6–10) que corresponde

Algoritmo 3 Algoritmo proposto para a indução da meta-árvore de decisão

Entrada: D \triangleright conjunto de n exemplos rotulados $\{(\mathbf{x}_i, y_i), i = 1, 2, \dots, n\}$ contendo m atributos $\{X_1, X_2, \dots, X_m\}$

Entrada: L \triangleright número de árvores da floresta (*Random Forest*), onde $L \geq 1$, *default* $L = \max\{128, 3\lfloor \log_2 \frac{n}{m} \rfloor\}$

Saída: T \triangleright meta-árvore de decisão

- 1: **function** MT(D, L)
- 2: $F \leftarrow \text{buildRandomForest}(D, L)$ \triangleright Criação da floresta F contendo L árvores. Sejam $\{F_1, F_2, \dots, F_L\}$ as árvores individuais contidas na floresta F
- 3: $Z \leftarrow \emptyset$
- 4: **for** $i \in \{1, 2, \dots, L\}$ **do**
- 5: $t \leftarrow \text{Tree2Table}(F_i, L, n)$ \triangleright Transformação da árvore de decisão F_i em uma tabela de decisão t
- 6: $Z \leftarrow Z \cup t$ \triangleright Adiciona a tabela de decisão t ao conjunto de tabelas de decisão Z
- 7: **end for**
- 8: $D_{\text{new}} \leftarrow \text{Table2Instances}(Z)$ \triangleright Transformação do conjunto de tabelas de decisão Z em um novo conjunto de exemplos D_{new}
- 9: $T \leftarrow \text{buildTree}(D_{\text{new}})$ \triangleright Algoritmo 1
- 10: **return** T

Algoritmo 4 Algoritmo que converter uma árvore de decisão em uma tabela de decisão

Entrada: T \triangleright árvore de decisão

Entrada: n \triangleright número de exemplos de treinamento

Entrada: L \triangleright número de árvores da floresta

Saída: R \triangleright tabela de decisão resultante

- 1: **function** Tree2Table(T, L)
- 2: $R \leftarrow \emptyset$
- 3: $W \leftarrow \sum_i w_i$ \triangleright Soma dos pesos de todas as folhas em T
- 4: **for** each leaf $l \in T$ with class C and weight w **do**
- 5: $r \leftarrow [?, \dots, ?, ?, ?, ?]$ \triangleright Linha $[X_1, \dots, X_m, Y, \text{Weight}]$ sem preenchimento de valores
- 6: **for** each attribute X_i with threshold O_j from root to leaf l in T **do**
- 7: $r[X_i] \leftarrow O_i$ \triangleright Preenche $r[X_1, \dots, X_m]$ com o valor dos atributos contidos em cada nó
- 8: **end for**
- 9: $r[Y] \leftarrow C$ \triangleright Preenche $r[Y]$ com o valor da classe contido na folha
- 10: $r[\text{Weight}] \leftarrow \frac{nw}{WL}$ \triangleright Preenche $r[\text{Weight}]$ com o peso normalizado
- 11: $R \leftarrow R \cup r$ \triangleright Adiciona linha r em R
- 12: **end for**
- 13: **return** R

a uma linha da tabela de decisão resultante R (linha 11), ocorrendo, desta forma, a transformação de representação de árvore de decisão para tabela de decisão.

Para o entendimento do Algoritmo 4 considere:

- Cada atributo presente no conjunto original de exemplos corresponde a uma coluna da tabela de decisão;
- Cada folha presente na árvore de decisão corresponde a:
 - uma única linha na tabela de decisão, se houver apenas atributos categóricos ou se a estratégia de representação de valores numéricos for escolhida como a média (Seção 3.2) ou

- duas linhas na tabela de decisão, caso a estratégia de representação de valores numéricos for escolhida como os limites inferior e superior do intervalo (uma linha para o valor inferior e outra linha para o valor superior) (Seção 3.2).
- Na representação de uma folha, as colunas da tabela de decisão assumem os valores contidos nas ramificações das subárvore que formam o caminho da raiz até a respectiva folha, considerando que a origem do caminho sempre será a raiz da árvore de decisão;
- Para os atributos que não estiverem presentes na representação da folha, as respectivas colunas assumem o valor ‘?’ , simbolizando a ausência de valor.
- Uma forma direta de representar o peso de cada folha seria utilizar o número de exemplos que chegaram até a respectiva folha. Entretanto, neste trabalho, optou-se por utilizar as métricas descritas na Seção 3.3 como peso da folha.
- Se a árvore é uma única folha, então a tabela contém uma única linha. As colunas que representam os atributos, assumem o valor ‘?’ , a coluna que representa a classe assume a mesma classe da folha
- A normalização dos pesos (linha 10 do Algoritmo 4) é descrita na Seção 3.4.

Considerando que todas L árvores de decisão contidas na floresta foram convertidas para tabelas de decisão e unidas em Z , o método **Table2Instances** (linha 8 do Algoritmo 3) transforma Z em um novo conjunto de treinamento D_{new} , que servirá como entrada para a indução da meta-árvore de decisão. É nesta etapa que ocorre o meta-aprendizado, onde é dado como conjunto de treinamento o aprendizado adquirido por todas as árvores de decisão contidas na floresta (nível L_0) e induz-se uma meta-árvore de decisão (nível L_1), combinando assim, o aprendizado de todas as árvores de decisão da floresta em uma única árvore de decisão. O método **Table2Instances** é responsável por formatar as linhas contidas no conjunto de tabelas de decisão Z no modo esperado pelo algoritmo utilizado para induzir a meta-árvore.

Em correspondência ao Algoritmo 3, na Figura 5 é ilustrado o processo realizado na indução de uma meta-árvore de decisão. No processo ① é realizada a indução da *Random Forest* contendo as L árvores de decisão. No processo ② todas as árvores de decisão contidas na floresta são transformadas para tabelas de decisão (Algoritmo 4). No processo ③, considerando que todas as tabelas de decisão possuem a mesma estrutura, é então realizada a união de todas as tabelas, dando origem a uma única tabela de decisão resultante que contém todas as linhas de todas as tabelas de decisão. No processo ④ é realizada a transformação da tabela de decisão resultante para um novo conjunto de treinamento. Finalmente, no processo ⑤ é realizada a indução da árvore final com base no novo conjunto de exemplos.

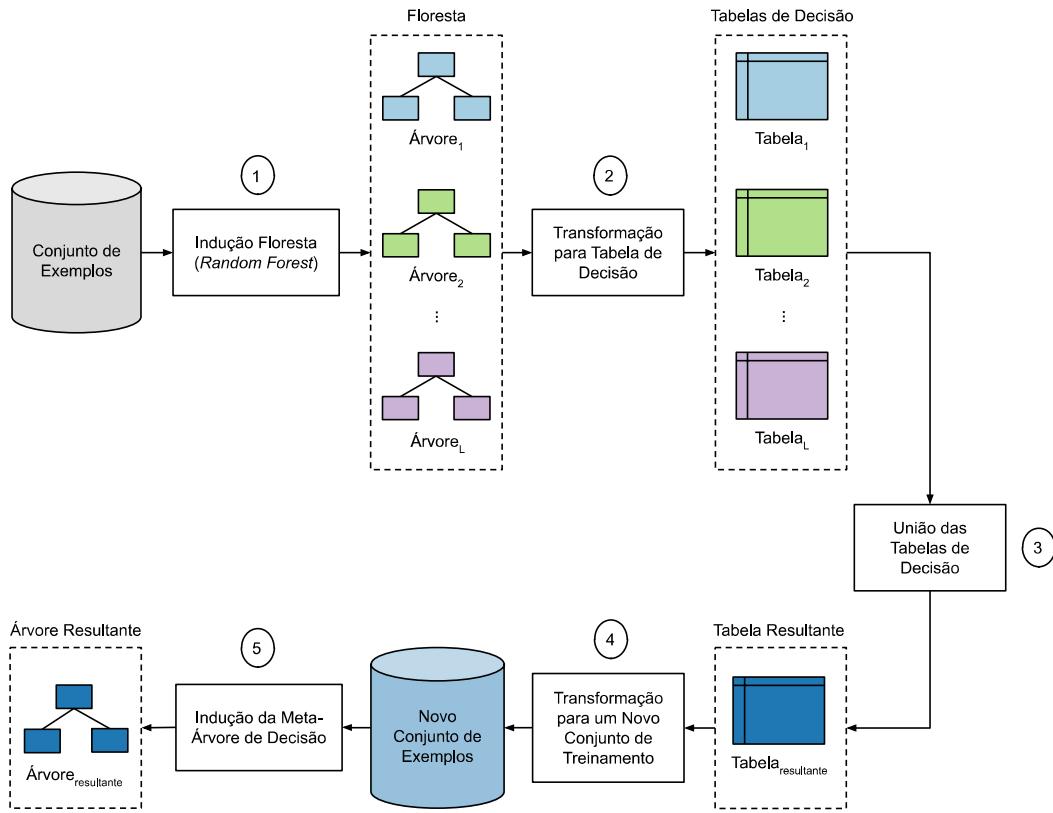


Figura 5 – Processo de Indução da Meta Árvore de Decisão.

3.2 Tratamento de Nós de Decisão de Atributos Numéricos

Na conversão de um caminho da raiz até a folha da árvore de decisão em uma linha na tabela de decisão, fica claro que, no caso de atributos numéricos, o nó de decisão da árvore resulta, em geral, em um intervalo de valores.

Por exemplo, na Figura 6 na página 54 é mostrada uma árvore de decisão contendo nós de decisão em atributos numéricos. Esta árvore foi induzida com base no conjunto de treinamento representado na Tabela 3 na página oposta, adaptado de (QUINLAN, 1993). Como é possível observar, a folha R_1 tem em sua representação o intervalo [65.00, 77.50] para o atributo Umidade. A folha R_2 tem em sua representação o intervalo [77.50, 96.00] para o atributo Umidade. A folha R_4 tem em sua representação o intervalo [64.00, 70.50] para o atributo Temperatura. A folha R_5 tem em sua representação o intervalo [70.50, 73.00] para o atributo Temperatura. Finalmente, a folha R_6 tem em sua representação o intervalo [73.00, 85.00] para o atributo Temperatura. É importante ressaltar que os intervalos definidos se limitam ao maior e menor valores do respectivo atributo numérico contidos no conjunto de treinamento.

Tabela 3 – Conjunto exemplos hipotético sobre jogar ou não jogar com base nas características climáticas contendo quatro atributos (Clima, Temperatura, Umidade e Ventando) e duas classes (sim e não).

Exemplo	Clima	Temperatura	Umidade	Ventando	Classe
z_1	ensolarado	85	85	não	não
z_2	ensolarado	80	90	sim	não
z_3	nublado	83	86	não	sim
z_4	chuvisco	70	96	não	sim
z_5	chuvisco	68	80	não	sim
z_6	chuvisco	65	70	sim	não
z_7	nublado	64	65	sim	sim
z_8	ensolarado	72	95	não	não
z_9	ensolarado	69	70	não	sim
z_{10}	chuvisco	75	80	não	sim
z_{11}	ensolarado	75	70	sim	sim
z_{12}	nublado	72	90	sim	sim
z_{13}	nublado	81	75	não	sim
z_{14}	chuvisco	71	91	sim	não
Mínimo	-	64	65	-	-
Máximo	-	85	96	-	-

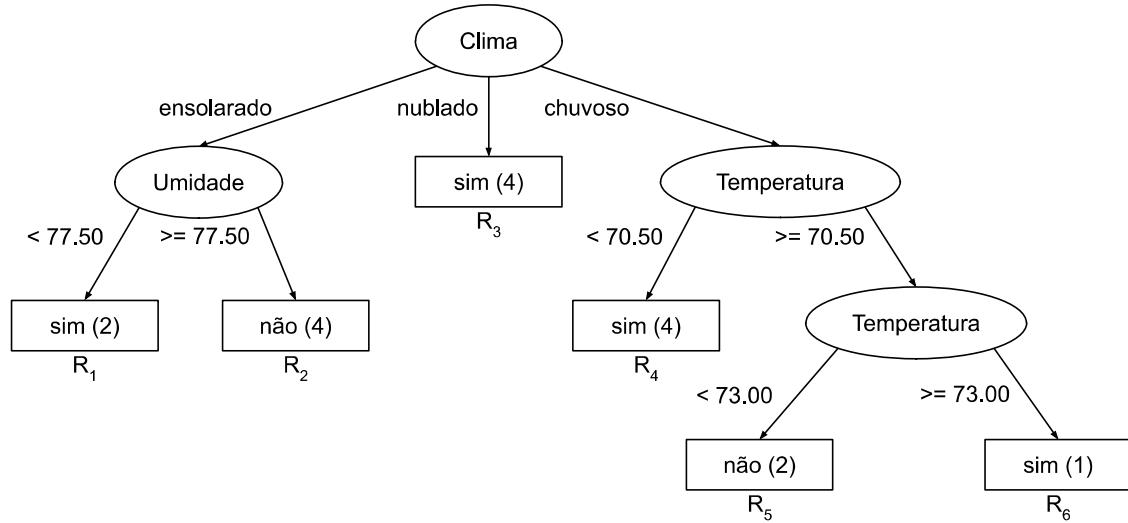
Entretanto, a transformação da tabela de decisão para ser o conjunto de treinamento para o nível de meta-aprendizado há a necessidade de expressar o intervalo de valores por meio de um valor único.

Assim, para representar intervalos na tabela de decisão foram adotadas e implementadas as seguintes estratégias neste trabalho de pesquisa:

- **média:** representar o intervalo por número representativo do mesmo, por exemplo, a média ou mediana dos valores no intervalo da folha. Neste trabalho, optou-se por utilizar a média como estratégia de representação de intervalos. A tabela de decisão correspondente à Figura 6 encontra-se na Figura 7 na página 55.
- **inf-sup:** representar os limites inferior e superior do intervalo na folha, o que resulta em duas linhas na tabela de decisão para cada folha. O peso de cada uma das linhas é dividido igualmente (metade em cada linha) em relação ao peso original. A tabela de decisão correspondente à Figura 6 encontra-se na Figura 8 na página 55.

3.3 Métricas para o Peso de uma Folha

Cada folha corresponde a uma regra que, por sua vez, pode ser vista como tendo dois componentes $L \rightarrow R$. No caso de regras de classificação, R é a classe da folha e L representa as condições (testes nos atributos) até atingir aquela folha.



(a) Árvore de Decisão

Região	Atributos				Classe	Peso
	Clima	Temperatura	Umidade	Ventando		
R ₁	ensolarado	?	[65.00, 77.50)	?	sim	2.00
R ₂	ensolarado	?	[77.50, 96.00]	?	não	4.00
R ₃	nublado	?	?	?	sim	4.00
R ₄	chuvisco	[64.00, 70.50)	?	?	sim	1.00
R ₅	chuvisco	[70.50, 73.00)	?	?	não	2.00
R ₆	chuvisco	[73.00, 85.00]	?	?	sim	1.00

(b) Tabela de Decisão

Figura 6 – (a) Exemplo de árvore de decisão contendo atributos numéricos induzida a partir do conjunto de exemplos da Tabela 3. (b) Tabela de decisão correspondente à árvore de decisão em (a).

Sob esse ponto de vista, é possível definir a correspondente matriz de contingência de uma folha (ou regra), mostrada na Tabela 4. Nesta tabela, L denota o conjunto de exemplos para os quais a condição da regra é verdadeira (os exemplos são cobertos pela regra) e seu complemento \bar{L} denota o conjunto de exemplos para os quais a condição da regra é falsa (exemplos não cobertos pela regra) e analogamente para R e \bar{R} . LR denota o conjunto de exemplos $L \cap R$ no qual L e R são ambos verdadeiros (a regra classifica corretamente os exemplos), $L\bar{R}$ denota o conjunto de exemplos $L \cap \bar{R}$ no qual L é verdadeiro e R é falso (a regra classifica incorretamente os exemplos) e assim por diante.

Por generalidade, denota-se a cardinalidade de um conjunto A por a , ou seja, $a = |A|$. Assim, l denota o número de exemplos no conjunto L , ou seja, $l = |L|$, r denota o número de exemplos no conjunto R , ou seja $r = |R|$, lr denota o número de exemplos no

Região	Atributos				Classe	Peso
	Clima	Temperatura	Umidade	Ventando		
R ₁	ensolarado	?	71.25	?	sim	2.00
R ₂	ensolarado	?	86.75	?	não	4.00
R ₃	nublado	?	?	?	sim	4.00
R ₄	chuvisco	67.25	?	?	sim	1.00
R ₅	chuvisco	71.75	?	?	não	2.00
R ₆	chuvisco	79.00	?	?	sim	1.00

Figura 7 – Exemplo de tabela de decisão para a árvore da Figura 6 utilizando a média para representar o intervalo de valores.

Região	Atributos				Classe	Peso
	Clima	Temperatura	Umidade	Ventando		
R ₁	ensolarado	?	65.00	?	sim	1.00
R ₁	ensolarado	?	77.50	?	sim	1.00
R ₂	ensolarado	?	77.50	?	não	2.00
R ₂	ensolarado	?	96.00	?	não	2.00
R ₃	nublado	?	?	?	sim	4.00
R ₄	chuvisco	64.00	?	?	sim	0.50
R ₄	chuvisco	70.50	?	?	sim	0.50
R ₅	chuvisco	70.50	?	?	não	1.00
R ₅	chuvisco	73.00	?	?	não	1.00
R ₆	chuvisco	73.00	?	?	sim	0.50
R ₆	chuvisco	85.00	?	?	sim	0.50

Figura 8 – Exemplo de tabela de decisão para a árvore da Figura 6 utilizando os limites inferior e superior para representar o intervalo de valores.

conjunto LR com $lr = |LR|$ e assim por diante. Como anteriormente, n indica o número total de exemplos.

Tabela 4 – Matriz de contingência para a folha (regra) $L \rightarrow R$

		L	\bar{L}		
R	lr	\bar{lr}	r		
\bar{R}	$l\bar{r}$	$\bar{l}\bar{r}$	\bar{r}		
		l	\bar{l}	n	

A frequência relativa $|A|/n = a/n$ associada ao subconjunto A é denotada por $p(A)$, onde A é um subconjunto dos n exemplos. Dessa forma, a frequência relativa é usada como uma estimativa de probabilidade. A notação $p(A|B)$ segue sua definição habitual em probabilidade, dada por (3.1), onde A e B são ambos subconjuntos dos n exemplos.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)} = \frac{\frac{|AB|}{n}}{\frac{|B|}{n}} = \frac{\frac{ab}{n}}{\frac{b}{n}} = \frac{ab}{b} \quad (3.1)$$

Várias medidas podem ser usadas para avaliar o desempenho de uma folha sendo a precisão (confiabilidade positiva) a mais comum. Entretanto, com novos problemas a serem tratados, novas medidas considerando novidade, simplicidade e facilidade de compreensão humana podem ser interessantes (TODOROVSKI; FLACH; LAVRAC, 2000). Utilizando como base a matriz de contingência, é possível definir a maioria das medidas sobre regras. De especial interesse neste trabalho serão utilizadas as métricas confiabilidade positiva **prel** (Equação 3.2), novidade **nov4** (Equação 3.4), satisfação **sat** (Equação 3.5) e precisão de Laplace **lacc** (Equação 3.6), descritas a seguir.

A confiabilidade positiva corresponde a razão entre número de exemplos classificados corretamente pela regra e o número total de exemplos cobertos pela regra. Assume valores no intervalo $[0, 1]$.

$$\text{prel}(L \rightarrow R) = P(R|L) = \frac{lr}{l} \quad (3.2)$$

$$\text{nov}(L \rightarrow R) = P(LR) - P(L)P(R) = \frac{lr}{n} - \frac{l \cdot r}{n^2} \quad (3.3)$$

$$\text{nov4}(L \rightarrow R) = 4 \times \text{nov}(L \rightarrow R) \quad (3.4)$$

$$\text{sat}(L \rightarrow R) = \frac{P(\bar{R}) - P(\bar{R}|L)}{P(\bar{R})} = 1 - \frac{n \cdot l\bar{r}}{l \cdot \bar{r}} \quad (3.5)$$

A precisão de Laplace não se enquadra diretamente na notação de frequência/probabilidade proposta por (LAVRAC; FLACH; ZUPAN, 1999b) mas corrige o problema de regras com poucos erros cobrindo muitos exemplos da confiabilidade positiva (CLARK; BOSWELL, 1991). Na Equação 3.6, k representa o número de classes do conjunto de treinamento. Esta métrica assume valores no intervalo $(0, 1)$.

$$\text{lacc}(L \rightarrow R) = \frac{lr + 1}{l + k} \quad (3.6)$$

Considerando L e R , a novidade é definida verificando se LR é independente deles. Isto pode ser obtido comparando o resultado observado lr contra o valor esperado sob a consideração de independência $\frac{l \cdot r}{n}$. Quanto mais o valor observado diferir do valor esperado, maior a probabilidade de que exista uma associação verdadeira e inesperada entre L e R . Pode ser demonstrado que $-0,25 \leq \text{nov} \leq 0,25$: quanto maior um valor positivo (perto de 0,25), mais forte é a associação entre L e R , enquanto que quanto menor um valor negativo (perto de -0,25), mais forte é a associação entre L e \bar{R} . Neste trabalho, o valor

da métrica novidade será multiplicado por 4 para de forma a colocar a métrica no intervalo $[-1, +1]$, o que leva a Equação 3.4.

Já a satisfação é o aumento relativo na precisão entre a regra $L \rightarrow \text{verdade}$ e a regra $L \rightarrow R$. Segundo Lavrac, Flach e Zupan (1999b) esta medida, cujos valores variam no intervalo $[-1, +1]$, é indicada para tarefas voltadas à descoberta de conhecimento, sendo capaz de promover um equilíbrio entre regras com diferentes condições e conclusões.

Como pode ser observado, as métricas novidade e satisfação podem assumir valores negativos. Como o intuito deste trabalho consiste em representar uma árvore de decisão contendo regras cuja conclusão seja a classe (e não seu complemento, ou seja, todas as demais classes), regras com valores negativos para essas duas métricas não serão consideradas, ou seja, serão descartadas do processo de criação da meta-árvore de decisão final. Como a intenção da proposta é produzir uma árvore com a mesma característica/representatividade de uma árvore de decisão convencional. Os valores negativos para essas métricas que buscam descoberta de conhecimento tem um significado diferente, de forma que quando isso ocorre, significa que poderia ser qualquer outra classe que não seja a classe da respectiva folha. Para usufruir desta característica, seria necessário criar nós de dízimo que os exemplos não são (*NOT CLASS*), mas isso, mudaria totalmente a estrutura da árvore de decisão. Uma abordagem muito semelhante foi adotada no trabalho realizado por Lavrac, Flach e Zupan (1999a), onde os mesmos indicaram o valor zero para os casos onde as métricas novidade e satisfação obteram um valor negativo.

3.4 Normalização dos Pesos de cada Árvore

No processo de indução da meta árvore de decisão é esperado que tal árvore reflita a quantidade de exemplos fornecido no conjunto de treinamento, de forma análoga à geração de uma única árvore sem uso de meta-aprendizado. A abordagem adotada para isso, nesta pesquisa, é descrita a seguir.

Seja T_i uma árvore da floresta contendo folhas cujos pesos (sem normalização) são $\{w_{i1}, w_{i2}, \dots\}$. Seja a soma dos pesos da árvore T_i dada por $W_i = \sum_j w_{ij}$. Os pesos para a árvore T_i precisam ser ajustados de forma que sua soma seja igual à unidade, sendo dados por $\{\frac{w_{i1}}{W_i}, \frac{w_{i2}}{W_i}, \dots\}$. Agora, para que a árvore T_i represente o número total n de exemplos do conjunto de treinamento (o tamanho da amostra *bootstrap* n_i para gerar cada árvore é igual ao número n de exemplos no conjunto treinamento) os pesos de cada árvore são dados por $\{n \frac{w_{i1}}{W_i}, n \frac{w_{i2}}{W_i}, \dots\}$.

Considerando que há L árvores na floresta $\{T_1, T_2, \dots, T_L\}$ que é transformada em tabela de decisão para que ocorra o meta-aprendizado então é necessário adequar o peso da meta árvore de decisão como sendo $\{n \frac{w_{i1}}{LW_i}, n \frac{w_{i2}}{LW_i}, \dots\}$ para cada uma das árvores que a

compõe, ou seja, $\{\{n \frac{w_{11}}{LW_1}, n \frac{w_{12}}{LW_1}, \dots\}, \{n \frac{w_{21}}{LW_2}, n \frac{w_{22}}{LW_2}, \dots\}, \dots, \{n \frac{w_{L1}}{LW_L}, n \frac{w_{L2}}{LW_L}, \dots\}\}$.

Na Figura 9 encontra-se, por meio de um exemplo simples, a implementação do algoritmo proposto. Para isto, foi utilizado um conjunto de treinamento que contém somente atributos categóricos. Para facilitar o entendimento, foi utilizado o número de exemplos em cada folha para representar peso da respectiva linha na tabela de decisão. Ou seja, para esta ilustração, não foram aplicadas as métricas propostas para os pesos das folhas com o objetivo de facilitar o entendimento do exemplo.

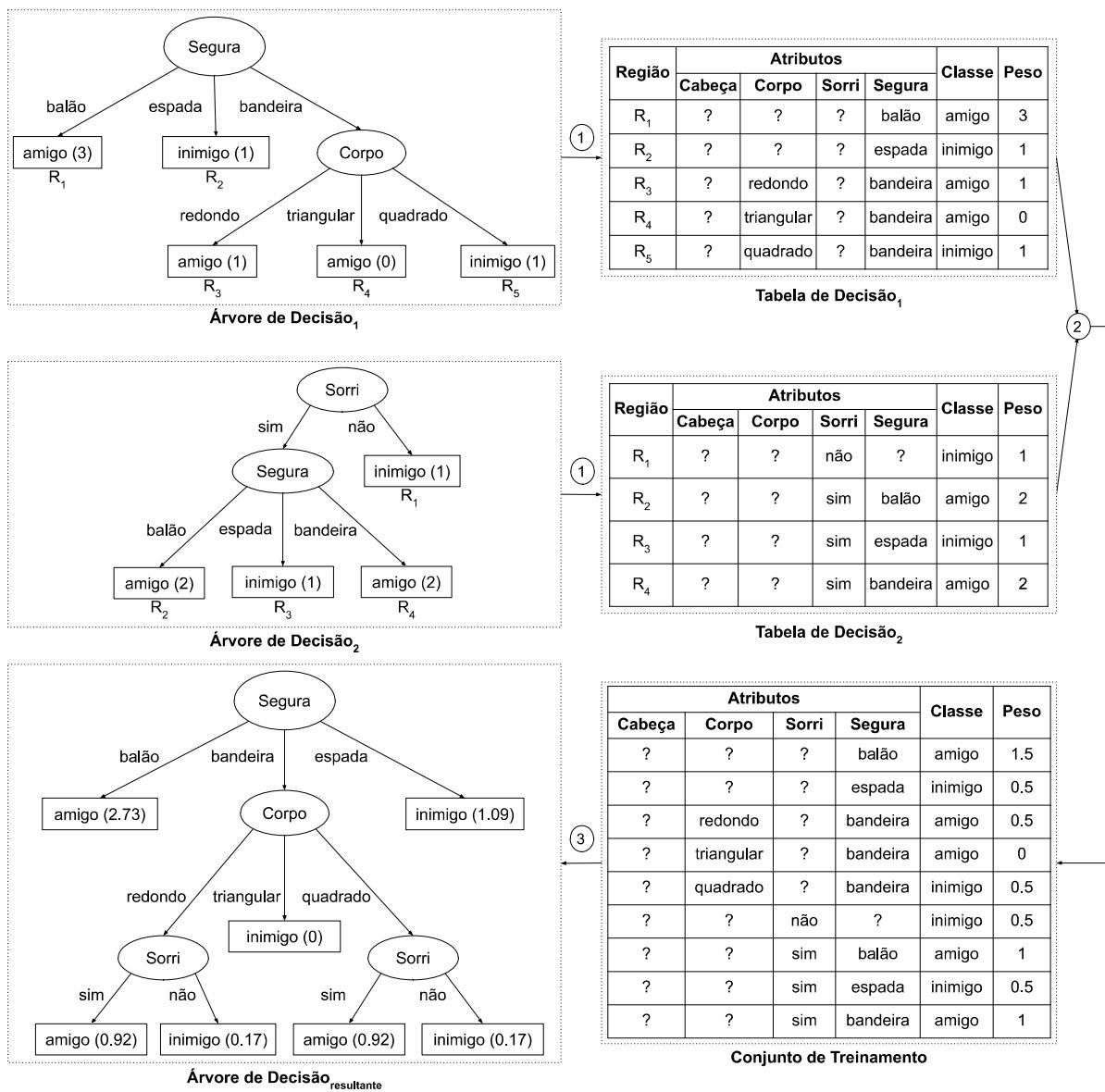


Figura 9 – Ilustração do algoritmo de meta aprendizado proposto e implementado.

Para o início do processo considere que foi realizada a criação de uma *Random Forest* que contém apenas duas árvores. Para a criação desta floresta considere o conjunto de exemplos representados na Tabela 2. O processo (1) é responsável por realizar a transformação das árvores de decisão contidas na floresta em tabelas de decisão (Algoritmo 4).

O processo ② realiza a união das tabelas de decisão, formando assim, após formatação necessária, um novo conjunto de exemplos. O processo ③ realiza a indução da meta-árvore de decisão com base no novo conjunto de exemplos.

3.5 Considerações Finais

Neste capítulo foi descrito o projeto de pesquisa envolvendo meta-aprendizado. No capítulo a seguir é descrita a metodologia e resultados de experimentos realizados utilizando a proposta aqui apresentada.

4

Configuração Experimental

Para medir a eficiência da meta árvore de decisão, a proposta foi avaliada por meio da realização de um estudo experimental conforme descrito a seguir. O estudo experimental comparou uma árvore única (DT), uma floresta (RF) e a meta árvore de decisão (MT), representada pelo Algoritmo 3. Foram analisados o desempenho preditivo entre DT, MT e RF e complexidade das árvores entre DT e MT, já que uma floresta possui complexidade muito elevada em relação a uma única árvore.

O desempenho preditivo foi avaliado usando a extensão multiclasse da medida *Receiver Operating Characteristic (ROC) Area Under Curve (AUC)* que agraga os valores ROC AUC sobre cada par de classes (HAND; TILL, 2001; BOWERS; ZHOU, 2019). A medida ROC AUC se mostrou uma métrica melhor do que a precisão, pois é independente do limite de decisão e invariante para distribuições de probabilidade de classe a priori (BRADLEY, 1997).

A métrica para mensurar a complexidade de uma árvore adotada nesta pesquisa foi o número de folhas.

4.1 Datasets

Neste projeto foram utilizadas como fonte de dados os seguintes repositórios:

- *Machine Learning Data Set Repository*¹;
- *University of California, Irvine* (UCI) (DUA; GRAFF, 2019)² e
- *OpenML*³.

¹ *Machine Learning Data Set Repository* encontra-se disponível em <<http://mldata.org>>.

² UCI encontra-se disponível em <<https://archive.ics.uci.edu/ml/datasets.php>>

³ *OpenML* encontra-se disponível em <<https://www.openml.org>>

Durante o levantamento dos *datasets* verificou-se alguns casos que possuem atributos de identificação dos exemplos. Geralmente esses atributos são descartados já que não contribuem diretamente para a generalização do modelo. Para a execução dos experimentos, esses atributos de identificação foram removidos por meio do pré-processamento dos datasets. O conceito de atributo de identificação pode ser entendido como todo o atributo que possua valores únicos em sua distribuição.

Nas Tabelas 5, 6 e 7 são apresentados os *datasets* selecionados para os experimentos. No total foram selecionados 150 *datasets*. Para facilitar a análise de resultados, os datasets foram particionados em quatro categorias:

- (i) na Tabela 5 são apresentados os 43 *datasets* que contém somente atributos categóricos;
- (ii) na Tabela 6 são apresentados os 51 *datasets* que contém somente atributos numéricos;
- (iii) na Tabela 7, são apresentados os 56 *datasets* que contém atributos categóricos e numéricos;
- (iv) a união de (i), (ii) e (iii) resultando nos 150 *datasets* analisados. Nota-se que as categorias (i), (ii) e (iii) são disjuntas entre si.

4.2 Algoritmos e Parametrização

Para a realização dos experimentos foi utilizada a ferramenta Weka (HALL et al., 2009), onde foram selecionados os algoritmos: J48 (DT) e *RandomForest* (RF). O algoritmo proposto foi inserido na Weka e foi denominado de *MetaTree* (MT). Os parâmetros utilizados foram os seguintes:

1. DT: algoritmo de indução de árvore de decisão J48 executado com seus parâmetros *default*;
2. MT: Algoritmo 3, denotado por $MT(peso, discretização)$ para os *datasets* das categorias (ii) e (iii) da seção anterior ou $MT(peso)$ para os *datasets* da categoria (i) da seção anterior, onde:
 - *peso*: que indica a métrica utilizada na ponderação das linhas da tabela de decisão (*meta-dataset*). Foram utilizadas as métricas de Confiabilidade Positiva (prel), Laplace (lacc), Novidade (nov4) e Satisfação (sat), ou seja, $peso \in \{\text{prel}, \text{lacc}, \text{nov4}, \text{sat}\}$ e
 - *discretização*, que indica a estratégia de discretização utilizada para o tratamento de atributos numéricos. Foram utilizadas as estratégias de Intervalo (*range*) e Média (*mean*) dos valores, ou seja, $discretização \in \{\text{mean}, \text{range}\}$

3. RF: *Random Forest* utilizando 128 árvores. Este valor é proveniente de resultado de pesquisa anterior realizada por (OSHIRO; PEREZ; BARANAUSKAS, 2012);

Com esses parâmetros o Algoritmo 3 foi executado, para cada *dataset* das categorias (ii) e (iii) o total de $2 \times 4 \times 2 = 16$ execuções; para *datasets* da categoria (i) o total foi de $1 \times 4 = 4$ execuções. O total geral é de 20 execuções para cada *dataset*. Os resultados da categoria (iv) correspondem à união dos resultados das categorias disjuntas (i), (ii) e (iii).

Os experimentos foram realizados utilizando validação cruzada com 10-*folds* (REFA-EILZADEH; TANG; LIU, 2009), totalizando $(20+2)$ algoritmos $\times 10$ *folds* $\times 150$ *datasets* = 33.000 execuções.

Tabela 5 – Conjunto com 43 *datasets* somente com atributos categóricos. Atr: número de atributos, Ex: número de exemplos, Cl: número de classes.

Dataset	Atr	Ex	Cl
analcatdata-boxing1	3	120	2
analcatdata-boxing2	3	132	2
analcatdata-dmft	4	797	6
analcatdata-donner	3	28	2
analcatdata-marketing	32	364	5
analcatdata-reviewer	7	379	3
analcatdata-fraud	11	42	2
audiology	69	226	24
audiology-binary	69	226	2
balloons-adult+stretch	4	20	2
balloons-adult-stretch	4	20	2
balloons-yellow-small	4	20	2
blogger	5	100	2
breast-cancer	9	286	2
car-df	6	1728	4
contact-lenses	4	24	3
dbworld-subjects	242	64	2
dbworld-subjects-stemmed	229	64	2
dna	180	3186	3
king-and-rook	6	28056	18
kr-vs-kp	36	3196	2
lung-cancer	56	32	3
molecular-promotor-gene	57	106	2
molecular-splice-junction	60	3190	3
monks-problems	6	601	2
mushroom	22	8124	2
mushroom-expanded	21	8416	2
nursery	8	12960	5
nursery-4class	8	12958	4
phishing	30	11055	2
post-operative-patient	8	90	3
postoperative-patient-data	8	90	3
primary-tumor	17	339	22
qualitative-bankruptcy	6	250	2
servo	4	167	2
shuttle-landing-control	6	15	2
solar-flare-1	12	315	5
solar-flare-2	12	1066	6
soybean	35	683	19
spect	22	267	2
splice	60	3190	3
tic-tac-toe	9	958	2
vote	16	435	2

Tabela 6 – Conjunto com 51 *datasets* somente com atributos numéricos. Atr: número de atributos, Ex: número de exemplos, Cl: número de classes.

Dataset	Atr	Ex	Cl
analcatdata-authorship	70	841	4
balance-scale	4	625	3
banknote-authentication	4	1372	2
blood-transfusion-service	4	748	2
chatfield-figure	12	235	2
column-2C	6	310	2
column-3C	6	310	3
delta-elevators	6	9517	2
disclosure-z	3	662	2
ecoli	7	336	8
eeg-eye-state	14	14980	2
gas-drift	128	13910	6
glass	9	214	7
hayes-roth	4	132	2
heart-statlog	13	270	2
hill-valley	100	1212	2
ionosphere	34	351	2
iris	4	150	3
japanese-vowels	14	9961	9
letter	16	20000	26
madelon	500	2600	2
magic-telescope	10	19020	2
mfeat-factors	216	2000	10
mfeat-fourrier	76	2000	10
mfeat-karhunen	64	2000	10
optdigits	64	5620	10
page-blocks	10	5473	5
pc4	37	1458	2
pendigits	16	10992	10
phoneme	5	5404	2
pima-diabetes	8	768	2
pollen	5	3848	2
prnn-synth	2	250	2
qsar-biodeg	41	1055	2
satimage	36	6430	6
segment	19	2310	7
semeion	256	1593	10
sonar	60	208	2
spambase	57	4601	2
steel-plates-fault	33	1941	2
strikes	6	625	2
transplant	3	131	2
triazines	60	186	2
vehicle	18	846	4
visualizing-galaxy	4	323	2
wdbc	30	569	2
wilt	5	4839	2
wine	13	178	3
wisconsin	32	194	2
wisconsin-breast-cancer	9	699	2
yeast	8	1484	10

Tabela 7 – Conjunto com 56 *datasets* com atributos categóricos e numéricos. Cat: número de atributos categóricos, Num: número de atributos numéricos, Atr: número de atributos, Ex: número de exemplos, Cl: número de classes.

Dataset	Cat	Num	Atr	Ex	Cl
analcatdata-AIDS	2	2	4	50	2
analcatdata-asbestos	2	1	3	83	2
analcatdata-bondrate	6	4	10	57	5
analcatdata-braziltourism	4	4	8	412	7
analcatdata-broadway	4	3	7	95	5
analcatdata-broadwaymult	4	3	7	285	7
analcatdata-creditscore	3	3	6	100	2
analcatdata-cyyoung8092	3	7	10	97	2
analcatdata-cyyoung9302	4	6	10	92	2
analcatdata-homerun	14	12	26	163	4
analcatdata-lawsuit	1	3	4	264	2
analcatdata-vineyard	1	2	3	468	2
analcatdata-votesurvey	1	3	4	48	4
analcatdata-wildcat	2	3	5	163	2
anneal	32	6	38	898	6
arsenic-female-bladder	1	3	4	559	2
autos	10	15	25	205	7
biomed	1	7	8	209	2
bridges-version	8	3	11	105	6
cars	1	6	7	406	3
churn	4	16	20	5000	2
cleveland-14-heart-disease	7	6	13	303	5
cloud	1	6	7	108	2
cmc	7	2	9	1473	3
collins	2	20	22	500	2
credit-rating	9	6	15	690	2
cylinder-bands	19	18	37	540	2
dermatology	33	1	34	366	6
dresses-sales	11	1	12	500	2
electricity	1	7	8	45312	2
eucalyptus	5	14	19	736	5
fruitfly	2	2	4	125	2
german-credit	13	7	20	1000	2
grub-damage	6	2	8	155	4
haberman	1	2	3	306	2
hepatitis	13	6	19	155	2
horse-colic	19	7	26	368	2
hungarian-14-heart-disease	7	6	13	294	5
hypothyroid	22	7	29	3772	4
ilpd	1	9	10	583	2
irish	3	2	5	500	2
lymphography	15	3	18	148	4
musk	2	167	169	6598	2
newton-hema	1	2	3	140	2
pasture-production	1	21	22	36	3
prmn-viruses	8	10	18	61	4
sick	22	7	29	3772	2
speed-dating	63	59	122	8378	2
squash-stored	3	21	24	52	3
squash-unstored	3	20	23	52	3
stress	4	8	12	202	2
tae	2	3	5	151	3
veteran	4	3	7	137	2
visualizing-livestock	1	1	2	130	5
vowel	2	10	12	990	11
zoo	15	1	16	101	7

4.3 Validação

Para a comparação múltipla de algoritmos foi utilizado o teste de Friedman (1940), cuja hipótese nula é que a diferença encontrada nos resultados é dada pelo acaso. Se o teste de Friedman rejeitar a hipótese nula, um teste *post-hoc* é necessário para verificar em que pares de algoritmos as diferenças são realmente significativas (DEMŠAR, 2006; GARCÍA; HERRERA, 2009). O teste *post-hoc* escolhido para este trabalho foi o de Bonferroni-Dunn (DUNN, 1961) realizando a comparação todos-*versus*-todos, ou seja, fazendo todas as comparações possíveis entre os algoritmos. Em ambos os testes (Friedman e *post-hoc*) foi utilizado nível de significância de 5% ($\alpha = 0.05$) que corresponde ao nível de confiança de 95% ($\gamma = 1 - \alpha = 0.95$). Os testes foram realizados usando o software R para computação estatística (R Core Team, 2013).

Os resultados reportados na Seção 5 na próxima página, portanto, fazem uso da estratégia de validação aqui descrita, ou seja, Teste de Friedman com *post-hoc* de Bonferroni-Dunn com nível de confiança de 95%.

4.4 Considerações Finais

Neste capítulo foram apresentados configurações utilizadas para a realização do experimento e os resultados obtidos são reportados no capítulo seguinte.

5

Resultados & Discussão

Neste capítulo são reportados os resultados da execução da proposta do Capítulo 3 com a configuração experimental descrita no Capítulo 4.

Conforme descrito no capítulo anterior, foram analisados quatro recortes nas categorias: (i) 43 *datasets* que contêm somente atributos categóricos; (ii) 51 *datasets* que contém somente atributos numéricos; (iii) 56 *datasets* que contêm atributos categóricos e numéricos e (iv) 150 *datasets*, que correspondem à união de (i), (ii) e (iii). Nota-se, novamente, que as categorias (i), (ii) e (iii) são disjuntas entre si. Foram utilizadas as estratégias de intervalo e média para lidar com atributos numéricos; as métricas utilizadas para o peso da folha, foram Confiabilidade Positiva, Laplace, Novidade e Satisfação.

5.1 Resultados

Na Figura 10 são apresentados, sob a forma de *boxplot*, a média e o desvio padrão dos resultados obtidos pela árvore de decisão (DT), meta-árvores (MT) e floresta (RF) considerando a métrica ROC AUC.

Na Figura 11 são apresentados, sob a forma de *boxplot*, a média e o desvio padrão do número de folhas (em escala logarítmica) obtido pela árvore de decisão (DT) e meta-árvores (MT).

Nas Tabelas 8, 9 e 10 do Apêndice A são apresentados os valores AUC, média e ranking médio da árvore de decisão, meta-árvores e floresta em relação aos conjuntos de datasets categóricos, numéricos e misto, respectivamente. Nas Tabelas 12, 13 e 14 do Apêndice A são apresentados o número de folhas, média e ranking médio da árvore de decisão e meta-árvores em relação aos conjuntos de datasets categóricos, numéricos e misto, respectivamente. Nestas tabelas, cada célula armazena a média e o desvio padrão referente as 10 execuções realizadas na validação cruzada.

O resultado do teste *post-hoc* de Bonferroni-Dunn para a métrica ROC AUC pode

ser encontrado na forma de diagrama de diferença crítica nas Figuras 12 e 13. O resultado do teste *post-hoc* de Bonferroni-Dunn para o número de folhas pode ser encontrado na forma de diagrama de diferença crítica nas Figuras 14 e 15.

5.2 Discussão

5.2.1 Datasets Categóricos

Como pode ser observado nas Figuras 10(a) e 12(a), MT(prel) e MT(sat) obtiveram um desempenho inferior, porém não significativamente, em relação a *RF*. Nota-se que *DT* foi significativamente inferior *RF*. Esse resultado é interessante, pois, o desempenho de uma única árvore induzida pelo algoritmo proposto se assemelhou ao desempenho de uma floresta de 128 árvores, o que apresenta significativa melhoria em relação à comprehensibilidade do modelo final, por tratar-se de apenas uma única árvore de decisão.

Nas Figuras 11(a) e 14(a), é possível observar que MT(prel) e MT(sat) induziram uma árvore de decisão maior, porém não significativamente, do que a *DT*. Este resultado é interessante pois a combinação das 128 árvores da floresta resultou em uma árvore de decisão não significativamente maior que uma árvore de decisão convencional, o que favorece a interpretabilidade do modelo resultante.

Como pode ser observado nas Figuras 10(b) e 13(a), todos os algoritmos utilizados não apresentaram diferença significativa em relação ao desvio padrão de ROC AUC. É possível observar que não houve uma variação significativa de desempenho ao lidar com diferentes conjuntos de dados oriundos da validação cruzada. Neste sentido, o algoritmo proposto manteve desempenho quanto ao desvio padrão semelhante à *DT* e à *RF*.

Nas Figuras 11(b) e 15(a), MT(prel), MT(lacc) e MT(sat) não apresentaram diferença significativa em relação à *DT*, indicando que não houve uma variação significativa do número de folhas em relação à *DT*.

5.2.2 Datasets Numéricos

Como pode ser observado nas Figuras 10(a) e 12(b), o algoritmo proposto em todas as possíveis configurações obteve um desempenho significativamente inferior a *DT* e *RF*. Nota-se que as abordagens propostas para lidar com os algoritmos numéricos (média e intervalo) não conseguiram se assemelhar ao desempenho obtido pela *DT* e *RF*.

Nas Figuras 11(a) e 14(b), é possível observar que MT(prel, mean), MT(sat, mean), MT(prel, range) e MT(sat, range) induziram uma árvore de decisão menor, porém não

significativamente, do que a *DT*. Apesar da combinação das árvores da floresta ter resultado em uma árvore de decisão menor, o desempenho obtido foi significativamente inferior.

Como pode ser observado nas Figuras 10(b) e 13(b), MT(nov4, mean) não obtiveram uma diferença significativa no desvio padrão de ROC AUC em relação a *DT* e *RF*. MT(nov4, range) e MT(lacc, mean) não obteve uma diferença significativa no desvio padrão de ROC AUC em relação a *DT*. Já as demais configurações do algoritmo obtiveram uma diferença inferior significativa no desvio padrão de ROC AUC em relação a *DT* e *RF*.

Nas Figuras 11(b) e 15(b), MT(prel, mean), MT(lacc, mean), MT(sat, mean), MT(prel, range), MT(lacc, range) e MT(sat, range) obtiveram um desempenho do desvio padrão do número de folhas significativamente menor em relação a *DT*.

5.2.3 Datasets Categóricos e Numéricos

Como pode ser observado nas Figuras 10(a) e 12(c), MT(prel, mean), MT(lacc, mean), MT(sat, mean), MT(prel, range), MT(lacc, range) e MT(sat, range) obtiveram um desempenho inferior, porém não significativamente, em relação a *DT*. Nota-se que todas as configurações do algoritmo proposto e a *DT* obtiveram um desempenho significativamente inferior *RF*.

Nas Figuras 11(a) e 14(c), é possível observar que MT(prel, mean), MT(sat, mean), MT(prel, range) e MT(sat, range) induziram uma árvore de decisão sem diferença significativa do número de folhas em relação a *DT*.

Como pode ser observado nas Figuras 10(b) e 13(c), MT(nov4, mean) não obteve uma diferença significativa no desvio padrão de ROC AUC em relação a *DT* e *RF*. MT(prel, mean), MT(lacc, mean), MT(sat, mean), MT(prel, range), MT(lacc, range) e MT(sat, range) não obtiveram uma diferença significativa no desvio padrão de ROC AUC em relação a *DT*.

Nas Figuras 11(b) e 15(c), MT(prel, mean), MT(sat, mean) e MT(prel, range) não apresentaram diferença significativa do desvio padrão do número de folhas em relação a *DT*.

5.2.4 Todos os Datasets

Como pode ser observado nas Figuras 10(a) e 12(d), MT(prel, mean) e MT(sat, mean) obtiveram um desempenho inferior, porém não significativamente, em relação a *DT*. Nota-se que todas as configurações do algoritmo proposto e a *DT* obtiveram um desempenho significativamente inferior *RF*.

Nas Figuras 11(a) e 14(d), é possível observar que MT(prel, mean), MT(sat, mean), MT(prel, range) e MT(sat, range) induziram uma árvore de decisão sem diferença significativa do número de folhas em relação a DT .

Como pode ser observado nas Figuras 10(b) e 13(d), MT(nov4, mean) não obteve uma diferença significativa no desvio padrão de ROC AUC em relação a DT e RF . MT(prel, mean), MT(lacc, mean), MT(sat, mean), MT(prel, range), MT(lacc, range) e MT(sat, range) não obteve uma diferença significativa no desvio padrão de ROC AUC em relação a DT .

Nas Figuras 11(b) e 15(d), MT(prel, mean), MT(lacc, mean), MT(sat, mean), MT(prel, range), MT(lacc, range) e MT(sat, range) obtiveram um desempenho do desvio padrão do número de folhas significativamente menor em relação a DT .

5.2.5 Tratamento de Atributos Numéricos

Em relação as estratégias para os atributos numéricos, é possível observar que a estratégia da média obteve um desempenho melhor que a estratégia de intervalo. Porém, ambas não se mostraram como boas alternativas para lidar com os atributos numéricos. Já para datasets somente com atributos categóricos, o algoritmo proposto obteve um resultado interessante.

5.2.6 Ponderação das Folhas

Em relação as métricas de ponderação, a confiabilidade positiva e satisfação, obtiveram um resultado melhor que as métricas de laplace e novidade. A métrica de novidade foi a que obteve o pior desempenho, o que mostra que sua proposta de busca de conhecimento possivelmente novo não se mostrou interessante para o peso das folhas.

Em relação à Satisfação, que tem uma proposta semelhante à Novidade, ela se mostrou como uma métrica eficiente, obtendo um dos melhores desempenhos em relação as outras métricas. A métrica de Laplace, também não obteve um bom desempenho interessante para a ponderação dos pesos. Já a confiabilidade positiva obteve o melhor resultado.

5.3 Considerações Finais

Neste capítulo foram apresentados os resultados obtidos pelo experimento, como também, a discussão dos mesmos. Um resumo dos principais resultados é fornecido a seguir:

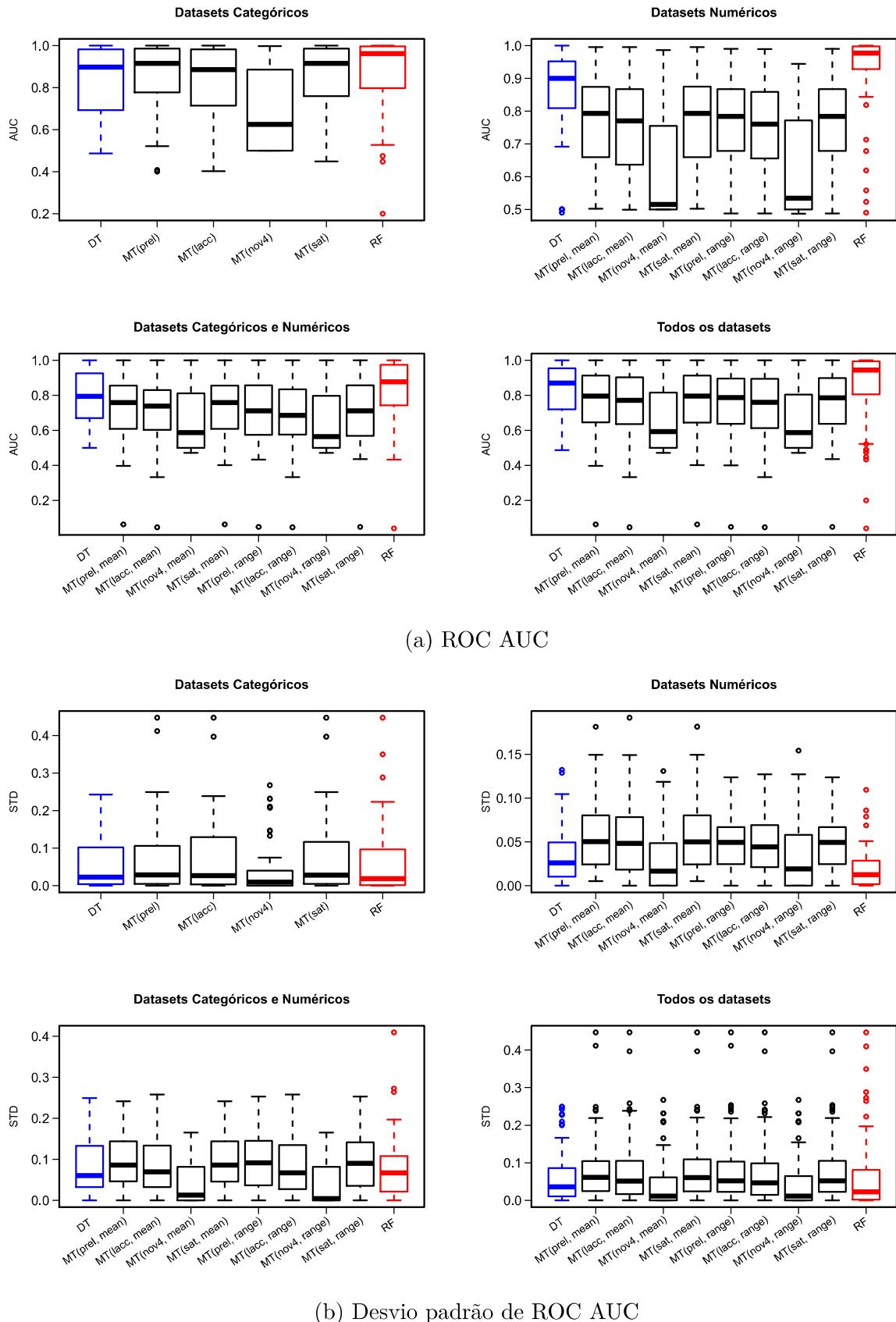
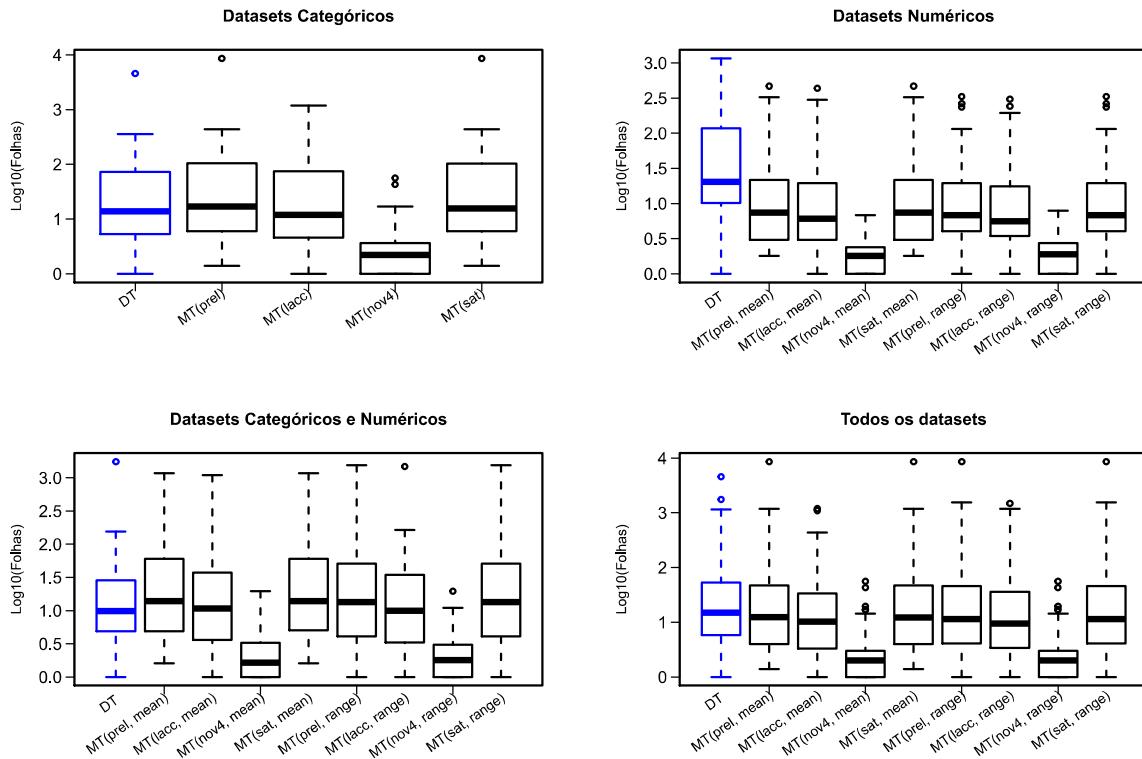
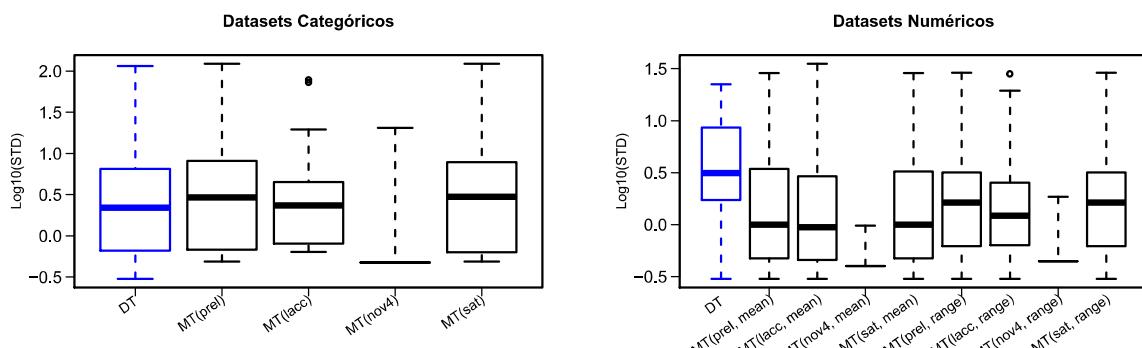


Figura 10 – Resultados de ROC AUC (Tabelas 8-11 do Apêndice A). A cor azul destaca o resultado obtido pela DT, a cor vermelha destaca o resultado obtido pela RF. A cor preta representa as variações do algoritmo proposto.



(a) Número de folhas (escala logarítmica)



(b) Desvio padrão do número de folhas (escala logarítmica)

Figura 11 – Resultados sobre número de folhas. A cor azul destaca o resultado obtido pela DT. A cor preta representa as variações do algoritmo proposto.

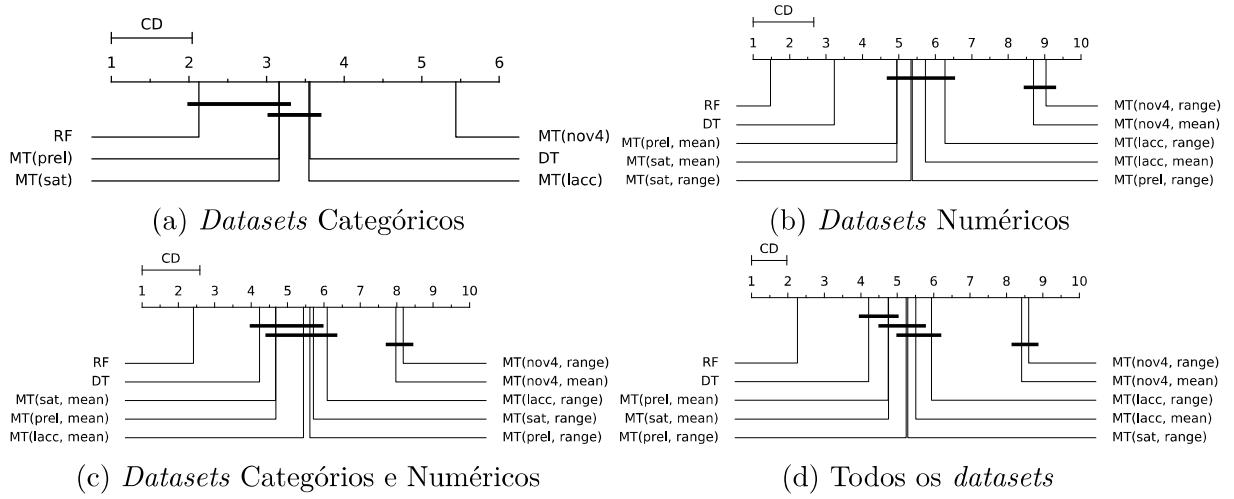


Figura 12 – Diagrama de diferença crítica considerando os resultados obtidos de ROC AUC (Tabelas 8-11 do Apêndice A).

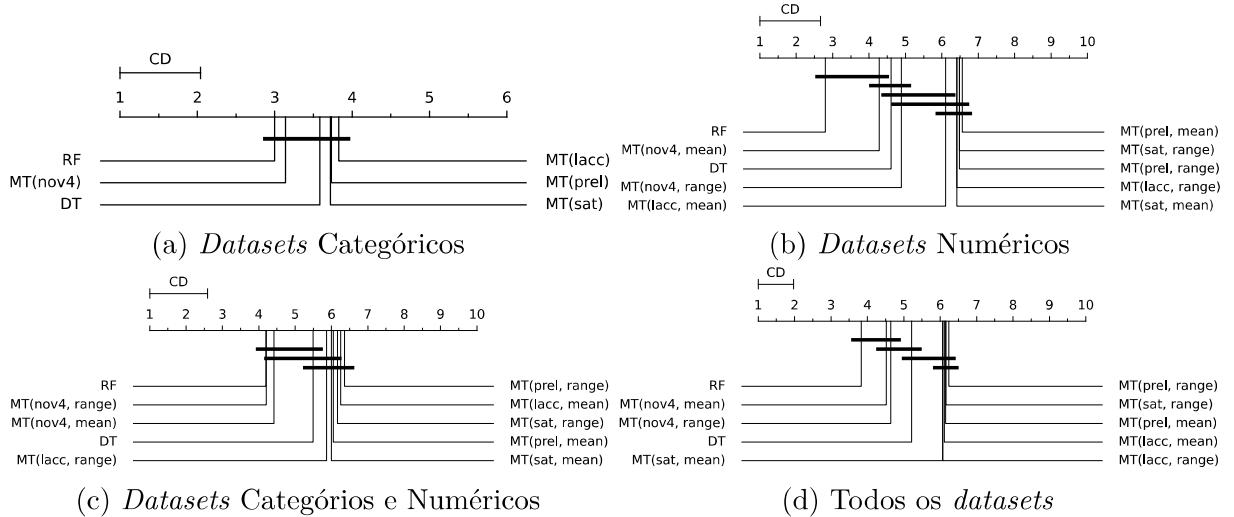


Figura 13 – Diagrama de diferença crítica considerando o desvio padrão da ROC AUC.

- Para os conjuntos: *datasets* numéricos, *datasets* mistos e todos os *datasets*, o algoritmo proposto obteve um desempenho inferior em relação a *DT* e *RF*
- Para o conjunto de *datasets* categóricos, o algoritmo proposto obteve um desempenho inferior, porém, sem diferença significativa em relação a *RF*. Já a *DT* obteve um desempenho significativamente inferior a *RF*, o que significa que o algoritmo proposto superou o desempenho da árvore de decisão convencional.
- Com os resultados obtidos, em relação as estratégias para lidar com atributos numéricos, a estratégia de Média obteve um desempenho superior a estratégia de Intervalo
- Em relação as métricas de ponderação das folhas, a que teve o melhor desempenho foi a Confiabilidade Positiva, seguida pelas métricas: Satisfação, Laplace e Novidade, respectivamente.

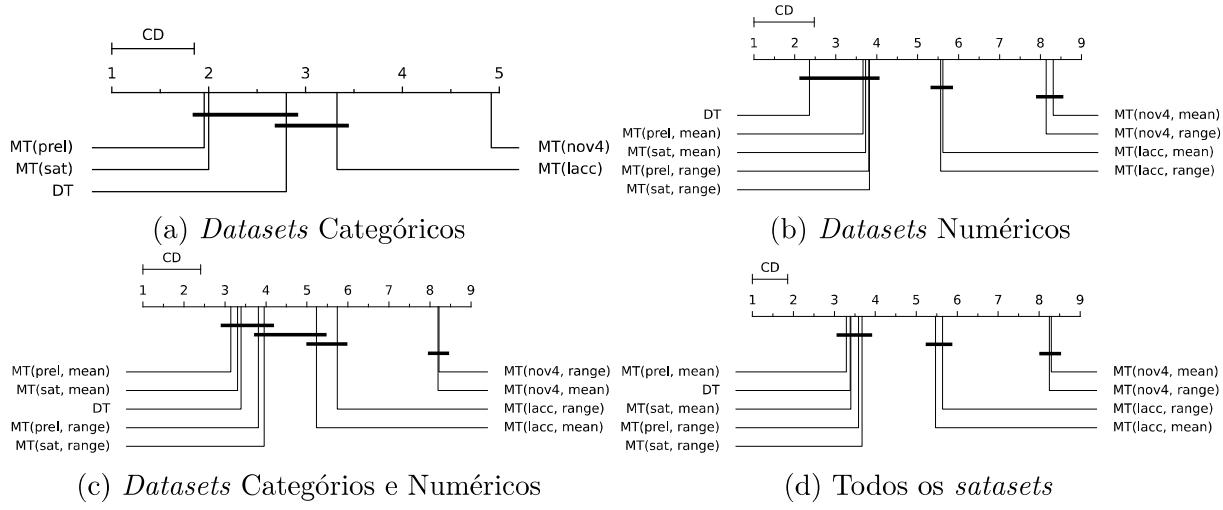


Figura 14 – Diagrama de diferença crítica considerando o número de folhas (escala logarítmica) da DT e das MTs.

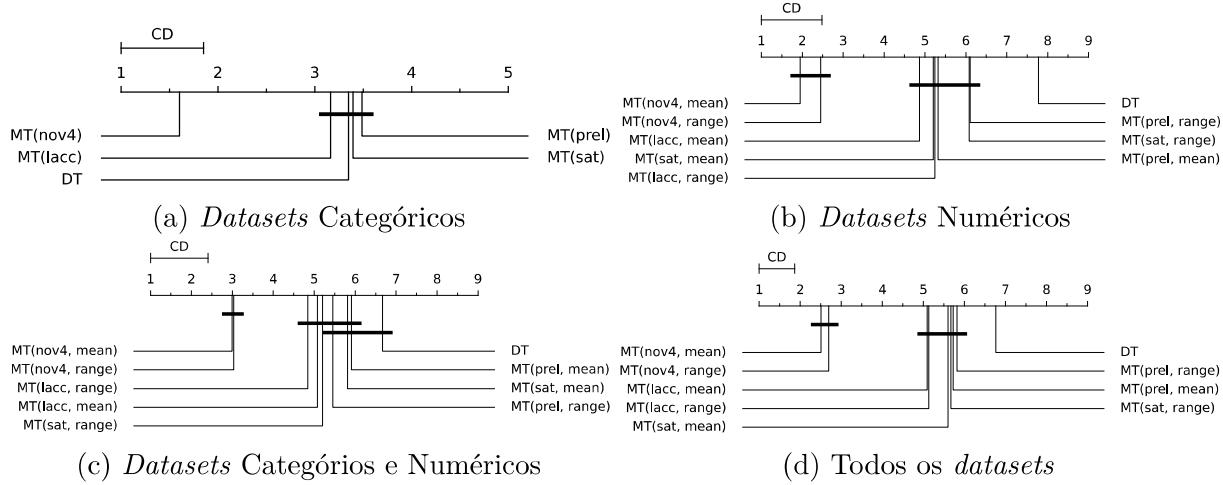


Figura 15 – Diagrama de diferença crítica considerando o desvio padrão do número de folhas (escala logarítmica) da DT e das MTs.

No próximo capítulo serão apresentadas as conclusões obtidas, principais contribuições e sugestões para trabalhos futuros que possam dar continuidade ao aqui desenvolvido.

6

Conclusão

6.1 Considerações Iniciais

Neste trabalho foi apresentada uma proposta para combinar as várias árvores de uma Random Forest em apenas uma árvore de decisão utilizando meta-aprendizado.

No Capítulo foram apresentados a introdução, a motivação e o objetivo da pesquisa. No Capítulo 1 foi apresentada a fundamentação teórica sobre Aprendizado de Máquina, Árvores de Decisão e Meta-Aprendizado.

No Capítulo 2 foram apresentados trabalhos relacionados ao aqui proposto. No Capítulo 3 encontra-se a proposta de desenvolvimento desta pesquisa, bem como sua representação sob a forma dos Algoritmos 3 e 4, descritos em alto nível:

- O Algoritmo 3 induz uma meta-árvore de decisão por meio da combinação das árvores de decisão de uma floresta
- O Algoritmo 4 transforma uma árvore de decisão em uma tabela de decisão. Este processo é uma etapa para a combinação das árvores da floresta.
- Foram propostas duas estratégias diferentes (Média e Intervalo) para lidar com os atributos numéricos.
- Foram utilizadas quatro métricas diferentes (Confiabilidade Positiva, Laplace, Novidade e Satisfação) para a ponderação das folhas.
- Foi definido a estratégia de normalização para os pesos de cada árvore

No Capítulo 4 são apresentadas a metodologia de avaliação experimental:

- A proposta foi analisada de forma empírica em 150 *datasets* de diferentes domínios. Para análise, os *datasets* foram divididos da seguinte forma: *datasets* categóricos, *datasets* numéricos e *datasets* mistos.
- Para avaliação, o método proposto foi comparado ao desempenho de uma árvore de decisão convencional e uma floresta.
- Foi definido o teste de Friedman e *post-hoc* com significância de 5% o que corresponde a 95% de confiança

No Capítulo 5 encontram-se os resultados obtidos dos experimentos realizados:

- Foram analisados os 150 *datasets* nas seguintes perspectivas: *datasets* categóricos, *datasets* numéricos, *datasets* mistos e todos os *datasets*.
- Para os conjuntos: *datasets* numéricos, *datasets* mistos e todos os *datasets*, o algoritmo proposto obteve um desempenho inferior em relação a *DT* e *RF*
- Para o conjunto de *datasets* categóricos, o algoritmo proposto obteve um desempenho inferior, porém, sem diferença significativa em relação a *RF*. Já a *DT* obteve um desempenho significativamente inferior a *RF*, o que significa que o algoritmo proposto superou o desempenho da árvore de decisão convencional.
- Com os resultados obtidos, em relação as estratégias para lidar com atributos numéricos, a estratégia de Média obteve um desempenho superior a estratégia de Intervalo
- Em relação as métricas de ponderação das folhas, a que teve o melhor desempenho foi a Confiabilidade Positiva, seguida pelas métricas: Satisfação, Laplace e Novidade, respectivamente.

6.2 Principais Contribuições

As principais contribuições deste trabalho são:

- Foi desenvolvido um algoritmo para realizar a transformação de uma árvore de decisão para uma tabela de decisão, apoiando-se no conceito de intercambialidade entre as representações do modelo. Por meio deste conceito e considerando a ponderação dos exemplos do dataset, foi possível criar um meta-dataset para indução da árvore de decisão final.

- Para os *datasets* categóricos, MT(prel) e MT(sat) obtiveram um desempenho inferior, porém não significativamente, em relação a *RF*. Já a *DT* obteve um desempenho significativamente inferior à *RF*. Esse resultado é interessante, pois, o desempenho de uma única árvore induzida pelo algoritmo proposto se assemelhou ao desempenho de uma floresta de 128 árvores, o que apresenta significativa melhoria em relação à comprehensibilidade do modelo final, por tratar-se de apenas uma única árvore de decisão.
- O trabalho realizado dá abertura para aplicações onde há o interesse em entender como o modelo chegou ao resultado, como também, identificar quais atributos foram determinantes para a classificação.
- Os resultados deste trabalho foram publicados em conferência internacional (FERREIRA; CANTÃO; BARANAUSKAS, 2022).

6.3 Trabalhos Futuros

Como continuação deste trabalho, alguns pontos que podem ser explorados em trabalhos futuros são os seguintes:

- Avaliar mais profundamente a estabilidade do algoritmo proposto e compará-lo em relação a árvore de decisão convencional.
- Aprimorar o algoritmo de forma que seja possível realizar o processamento utilizando vários níveis de meta-aprendizado.
- Avaliação de estratégias diferentes para o tratamento de atributos numéricos.

Referências

- BARANAUSKAS, J. A. *Extração Automática de Conhecimento Utilizando Múltiplos Indutores*. Tese (Doutorado) — ICMC-USP, 2001. <<http://www.teses.usp.br/teses/disponiveis/55/55134/tde-08102001-112806/>>.
- BOWERS, A. J.; ZHOU, X. Receiver operating characteristic (roc) area under the curve (auc): A diagnostic measure for evaluating the accuracy of predictors of education outcomes. *Journal of Education for Students Placed at Risk (JESPAR)*, Taylor & Francis, v. 24, n. 1, p. 20–46, 2019.
- BRADLEY, A. P. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recogn.*, Elsevier Science Inc., USA, v. 30, n. 7, p. 1145–1159, jul. 1997. ISSN 0031-3203. Disponível em: <[https://doi.org/10.1016/S0031-3203\(96\)00142-2](https://doi.org/10.1016/S0031-3203(96)00142-2)>.
- BREIMAN, L. Bagging predictors. v. 24, n. 2, p. 123–140, 1996.
- BREIMAN, L. Technical note: some properties of splitting criteria. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 24, n. 1, p. 41–47, 1996. ISSN 0885-6125.
- BREIMAN, L. Random forests. *Machine Learning*, v. 45, n. 1, p. 5–32, 2001.
- BREIMAN, L. *Wald Lecture II, Looking Inside the Black Box*. 2004. <<http://www.stat.berkeley.edu/users/breiman>>.
- BREIMAN, L.; CUTLER, A. *Random Forests: Classification/Clustering*. 2004. <<http://www.stat.berkeley.edu/users/breiman/RandomForests>>.
- BREIMAN, L. et al. *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth & Books, 1984.
- BUNTINE, W. L. *A Theory of Learning Classification Rules*. 1992.
- BURKART, N.; HUBER, M. F. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, v. 70, p. 1–74, 2021.
- CLARK, P.; BOSWELL, R. Rule induction with cn2: Some recent improvements. In: KODRATOFF, Y. (Ed.). *Proceedings of the 5th European Conference (EWSL 91)*. [S.l.: s.n.], 1991. p. 151–163.
- CRAGUN, B. J.; STEUDEL, H. J. A decision-table-based processor for checking completeness and consistency in rule-based expert systems. *International Journal of Man-Machine Studies*, Elsevier, v. 26, n. 5, p. 633–648, 1987.

Referências

- DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, JMLR.org, v. 7, p. 1–30, dez. 2006. ISSN 1532-4435. Disponível em: <<http://dl.acm.org/citation.cfm?id=1248547.1248548>>.
- DIETTERICH, T. G. *Machine Learning Research: Four Current Directions*. 1997. <<http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.5440>>.
- DOMINGOS, P. Knowledge acquisition from examples via multiple models. In: MORGAN KAUFMANN PUBLISHERS, INC. *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*. [S.l.], 1997. p. 98–106.
- DOSHI-VELEZ, F.; KIM, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
- DOŠILOVIĆ, F. K.; BRČIĆ, M.; HLUPIĆ, N. Explainable artificial intelligence: A survey. In: IEEE. *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. [S.l.], 2018. p. 0210–0215.
- DUA, D.; GRAFF, C. *UCI Machine Learning Repository*. 2019. Disponível em: <<http://archive.ics.uci.edu/ml>>.
- DUBATH, P. et al. Random forest automated supervised classification of hipparcos periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, Blackwell Publishing Ltd, v. 414, n. 3, p. 2602–2617, 2011. ISSN 1365-2966. Disponível em: <<http://dx.doi.org/10.1111/j.1365-2966.2011.18575.x>>.
- DUNN, O. J. Multiple comparisons among means. *Journal of the American Statistical Association*, [American Statistical Association, Taylor and Francis, Ltd.], v. 56, n. 293, p. 52–64, 1961. ISSN 01621459.
- EDWARDS, L.; VEALE, M. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, HeinOnline, v. 16, p. 18, 2017.
- EFRON, B.; TIBSHIRANI, R. *An Introduction to the Bootstrap*. [S.l.]: Chapman & Hall, 1993.
- FERREIRA, C. A.; CANTÃO, A. H.; BARANAUSKAS, J. A. Decision tree induction through meta-learning. In: MAGLOGIANNIS, I. et al. (Ed.). *Artificial Intelligence Applications and Innovations*. Cham: Springer International Publishing, 2022. p. 101–111. ISBN 978-3-031-08337-2. <https://doi.org/10.1007/978-3-031-08337-2_9>.
- FREITAS, A. A. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter*, ACM New York, NY, USA, v. 15, n. 1, p. 1–10, 2014.
- FREUND, Y.; SCHAPIRE, R. E. et al. Experiments with a new boosting algorithm. In: CITESEER. *Icml*. [S.l.], 1996. v. 96, p. 148–156.
- FRIEDMAN, M. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, JSTOR, v. 11, n. 1, p. 86–92, 1940. ISSN 0003-4851.

- GARCÍA, S.; HERRERA, F. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, v. 9, p. 2677–2694, 2009. Disponível em: <<http://www.jmlr.org/papers/volume9/garcia08a/garcia08a.pdf>>.
- GINI, C. On the measure of concentration with special reference to income and statistics. *Colorado College Publication, General Series*, v. 208, p. 73–79, 1936.
- GOUL, M. et al. Validating expert systems. *IEEE Intelligent Systems*, IEEE, v. 1, n. 3, p. 51–58, 1990.
- GRABCZEWSKI, K. *Meta-Learning in Decision Tree Induction*. Springer, 2014. v. 498. (Studies in Computational Intelligence, v. 498). ISBN 978-3-319-00959-9. Disponível em: <<https://doi.org/10.1007/978-3-319-00960-5>>.
- GUIDOTTI, R. et al. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, IEEE, v. 34, n. 6, p. 14–23, 2019.
- GUIDOTTI, R. et al. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, ACM New York, NY, USA, v. 51, n. 5, p. 1–42, 2018.
- GUNNING, D. Explainable artificial intelligence (xai). *Defense advanced research projects agency (DARPA), nd Web*, v. 2, n. 2, p. 1, 2017.
- HALL, M. et al. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, ACM, v. 11, n. 1, p. 10–18, 2009.
- HAMILTON, D.; KELLEY, K.; CULBERT, C. State-of-the-practice in knowledge-based system verification and validation. *Expert Systems with Applications*, Elsevier, v. 3, n. 4, p. 403–410, 1991.
- HAND, D. J.; TILL, R. J. A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learn.*, Kluwer Academic Publishers, USA, v. 45, n. 2, p. 171–186, out. 2001. ISSN 0885-6125. Disponível em: <<https://doi.org/10.1023/A:1010920819831>>.
- HEWETT, R.; LEUCHNER, J. Restructuring decision tables for elucidation of knowledge. *Data & Knowledge Engineering*, Elsevier, v. 46, n. 3, p. 271–290, 2003.
- JIN, Y.; SENDHOFF, B. Pareto-based multiobjective machine learning: An overview and case studies. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 38, n. 3, p. 397–415, 2008.
- KOHAVI, R.; SOMMERFIELD, D. Targeting business users with decision table classifiers. In: *KDD*. [S.l.: s.n.], 1998. p. 249–253.
- LAVRAC, N.; FLACH, P.; ZUPAN, B. Rule evaluation measures: A unifying view. In: SPRINGER. *International Conference on Inductive Logic Programming*. [S.l.], 1999. p. 174–185.
- LAVRAC, N.; FLACH, P.; ZUPAN, B. Rule evaluation measures: A unifying view. In: DZEROSKI, S.; FLACH, P. (Ed.). *Proceedings of the Ninth International Workshop on Inductive Logic Programming (ILP-99)*. [S.l.: s.n.], 1999. v. 1634, p. 74–185.

Referências

- LICHMAN, M. *UCI Machine Learning Repository*. 2013. <<http://archive.ics.uci.edu/ml>>. Disponível em: <\{<http://archive.ics.uci.edu/m>>.
- LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable ai: A review of machine learning interpretability methods. *Entropy*, v. 23, n. 1, 2021.
- MA, Y.; GUO, L.; CUKIC, B. Statistical framework for the prediction of fault-proneness. In: *Advances in machine learning applications in software engineering*. [S.l.]: Idea Group, 2007.
- MARCUS, G. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.
- MICHALSKI, R. S. A theory and methodology of inductive learning. In: *Machine learning*. [S.l.]: Springer, 1983. p. 83–134.
- MITCHELL, T. M. *Machine Learning*. [S.l.]: WCB McGraw-Hill, 1997.
- MONTAVON, G.; SAMEK, W.; MÜLLER, K.-R. Methods for interpreting and understanding deep neural networks. *Digital signal processing*, Elsevier, v. 73, p. 1–15, 2018.
- NONFJALL, H.; LARSEN, H. L. Detection of potential inconsistencies in knowledge bases. *International Journal of Intelligent Systems*, Wiley Online Library, v. 7, n. 2, p. 81–96, 1992.
- OSHIRO, T. M.; PEREZ, P. S.; BARANAUSKAS, J. A. How many trees in a random forest? In: *Proceedings of the 8th International Conference on Machine Learning and Data Mining in Pattern Recognition, MLDM 2012, Lecture Notes in Computer Science, ISBN 978-3-642-31536-7*. Berlin, Germany: [s.n.], 2012. v. 7376, p. 154–168. <http://dx.doi.org/10.1007/978-3-642-31537-4_13>.
- PRODROMIDIS, A.; CHAN, P.; STOLFO, S. Meta-learning in distributed data mining systems: Issues and approaches. *Advances in distributed and parallel knowledge discovery*, AAAI/MIT Press Menlo Park, v. 3, p. 81–114, 2000.
- QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, p. 81–106, 1986. Reprinted in Shavlik and Dietterich (eds.), 1990. *Readings in Machine Learning*, Morgan Kaufmann Publishers, Inc.
- QUINLAN, J. R. *C4.5: Programs for Machine Learning*. [S.l.]: Morgan Kaufmann, 1993. San Francisco, CA.
- R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2013. ISBN 3-900051-07-0. Disponível em: <<http://www.R-project.org/>>.
- REFAEILZADEH, P.; TANG, L.; LIU, H. Cross-validation. *Encyclopedia of database systems*, Springer, v. 5, p. 532–538, 2009.
- REZENDE, S. O. *Sistemas inteligentes: fundamentos e aplicações*. [S.l.]: Editora Manole Ltda, 2003.
- RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 1135–1144.

- RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "why should i trust you?"explaining the predictions of any classifier. In: *Proc. of the 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining.* [S.l.: s.n.], 2016. p. 1135–1144.
- SANTOS-GOMEZ, L.; DARNELL, M. J. Empirical evaluation of decision tables for constructing and comprehending expert system rules. *Knowledge Acquisition*, Elsevier, v. 4, n. 4, p. 427–444, 1992.
- SETHI, I. K.; CHATTERJEE, B. Conversion of decision tables to efficient sequential testing procedures. *Communications of the ACM*, ACM, v. 23, n. 5, p. 279–285, 1980.
- SHANNON, C. E. A mathematical theory of communication. *Bell system technical journal*, Wiley Online Library, v. 27, n. 3, p. 379–423, 1948.
- SIRIKULVIRIYA, N.; SINTHUPINYO, S. Integration of rules from a random forest. In: *International Conference on Information and Electronics Engineering.* [S.l.: s.n.], 2011. v. 6, p. 194–198.
- STOLFO, S. J. et al. Jam: Java agents for meta-learning over distributed databases. In: *KDD.* [S.l.: s.n.], 1997. v. 97, p. 74–81.
- STRECHT, P.; MENDES-MOREIRA, J.; SOARES, C. Merging decision trees: a case study in predicting student performance. In: SPRINGER. *International Conference on Advanced Data Mining and Applications.* [S.l.], 2014. p. 535–548.
- TODOROVSKI, L.; DŽEROSKI, S. Combining classifiers with meta decision trees. *Mach. Learn.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 50, n. 3, p. 223–249, mar. 2003. ISSN 0885-6125. Disponível em: <<https://doi.org/10.1023/A:1021709817809>>.
- TODOROVSKI, L.; FLACH, P.; LAVRAC, N. Predictive performance of weighted relative accuracy. In: ZIGHED, D. A.; KOMOROWSKI, J.; ZYTKOW, J. (Ed.). *4th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD2000).* [S.l.: s.n.], 2000. p. 255–264.
- UNION, C. O. E. *Council regulation (EU) no 279/2016 - Official website of the European Union.* 2016. Disponível em: <<https://eur-lex.europa.eu/eli/reg/2016/679/oj>>.
- VANSCHOOREN, J. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*, 2018.
- VANTHIENEN, J.; DRIES, E. Illustration of a decision table tool for specifying and implementing knowledge based systems. *International Journal on Artificial Intelligence Tools*, World Scientific, v. 3, n. 02, p. 267–288, 1994.
- WEISS, S. M.; KULIKOWSKI, C. A. *Computer systems that learn: classification and prediction methods from statistics, neural nets, machine learning, and expert systems.* [S.l.]: Morgan Kaufmann Publishers Inc., 1991.
- WOLPERT, D. H. Stacked generalization. *Neural networks*, Elsevier, v. 5, n. 2, p. 241–259, 1992.
- ZENKO, B.; TODOROVSKI, L.; DŽEROSKI, S. A comparison of stacking with meta decision trees to bagging, boosting, and stacking with other methods. In: *ICDM.* [S.l.]: IEEE Computer Society, 2001. p. 669–670.

Referências

- ZHAO, Y.; ZHANG, Y. Comparison of decision tree methods for finding active objects. *Advances in Space Research*, v. 41, p. 1955–1959, 2008.

Apêndices

A

Resultados dos Experimentos

Neste apêndice encontram-se os resultados obtidos dos experimentos reportados Seção 5 na página 67.

Tabela 8 – Valores AUC: árvore de decisão, meta-árvore e floresta - 43 Datasets categóricos

Dataset	MT					
	DT	prel	lacc	nov4	sat	RF
analcatdata-boxing1	0.88±0.12	0.85±0.15	0.85±0.15	0.50±0.00	0.85±0.15	0.89±0.09
analcatdata-boxing2	0.84±0.10	0.86±0.09	0.86±0.09	0.50±0.00	0.86±0.09	0.85±0.11
analcatdata-dmft	0.55±0.02	0.54±0.03	0.55±0.02	0.53±0.02	0.54±0.02	0.53±0.03
analcatdata-donner	0.50±0.00	0.60±0.41	0.58±0.40	0.50±0.00	0.58±0.40	0.20±0.35
analcatdata-marketing	0.62±0.10	0.52±0.07	0.52±0.07	0.50±0.00	0.52±0.07	0.67±0.05
analcatdata-reviewer	0.65±0.08	0.65±0.09	0.65±0.08	0.50±0.00	0.65±0.09	0.69±0.08
analcatdata-fraud	0.69±0.24	0.72±0.25	0.73±0.24	0.73±0.21	0.72±0.25	0.78±0.22
audiology	0.93±0.03	0.92±0.03	0.82±0.05	0.77±0.01	0.92±0.03	0.97±0.02
audiology-binary	0.98±0.04	0.98±0.03	0.98±0.03	0.90±0.06	0.98±0.03	0.99±0.01
balloons-adult+stretch	1.00±0.00	1.00±0.00	1.00±0.00	0.88±0.23	1.00±0.00	1.00±0.00
balloons-adult-stretch	1.00±0.00	1.00±0.00	1.00±0.00	0.75±0.27	1.00±0.00	1.00±0.00
balloons-yellow-small	1.00±0.00	1.00±0.00	1.00±0.00	0.63±0.23	1.00±0.00	1.00±0.00
blogger	0.69±0.21	0.78±0.18	0.77±0.20	0.59±0.21	0.74±0.18	0.90±0.09
breast-cancer	0.63±0.10	0.62±0.14	0.61±0.15	0.50±0.00	0.62±0.12	0.65±0.14
car-df	0.98±0.01	0.99±0.01	0.98±0.01	0.50±0.00	0.99±0.01	0.99±0.00
contact-lenses	0.95±0.11	1.00±0.00	1.00±0.00	0.87±0.07	1.00±0.00	0.90±0.15
dbworld-subjects	0.75±0.10	0.86±0.20	0.68±0.19	0.50±0.00	0.86±0.20	0.95±0.07
dbworld-subjects-stemmed	0.78±0.15	0.85±0.20	0.76±0.19	0.50±0.00	0.85±0.20	0.93±0.10
dna	0.95±0.02	0.97±0.01	0.97±0.01	0.83±0.02	0.97±0.01	0.99±0.00
king-and-rook	0.88±0.00	0.92±0.00	0.89±0.00	0.50±0.00	0.92±0.00	0.96±0.00
kr-vs-kp	1.00±0.00	0.99±0.00	1.00±0.00	0.94±0.01	0.99±0.00	1.00±0.00
lung-cancer	0.68±0.23	0.53±0.15	0.50±0.00	0.50±0.00	0.53±0.15	0.68±0.29
molecular-promotor-gene	0.83±0.14	0.78±0.12	0.73±0.16	0.50±0.00	0.78±0.12	0.98±0.04
molecular-splice-junction	0.96±0.01	0.98±0.01	0.98±0.01	0.61±0.15	0.98±0.01	0.99±0.00
monks-problems	0.54±0.07	0.59±0.07	0.60±0.08	0.50±0.00	0.59±0.07	0.80±0.08
mushroom	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
mushroom-expanded	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
nursery	1.00±0.00	1.00±0.00	1.00±0.00	0.97±0.00	1.00±0.00	1.00±0.00
nursery-4class	1.00±0.00	1.00±0.00	1.00±0.00	0.97±0.00	1.00±0.00	1.00±0.00
phishing	0.98±0.00	0.98±0.00	0.98±0.00	0.95±0.00	0.98±0.00	1.00±0.00
post-operative-patient	0.49±0.03	0.41±0.09	0.40±0.15	0.50±0.00	0.51±0.16	0.47±0.17
postoperative-patient-data	0.49±0.02	0.40±0.10	0.47±0.17	0.50±0.00	0.45±0.12	0.45±0.19
primary-tumor	0.70±0.03	0.79±0.03	0.73±0.03	0.61±0.02	0.79±0.04	0.80±0.03
qualitative-bankruptcy	0.99±0.01	1.00±0.01	1.00±0.01	0.99±0.01	1.00±0.01	1.00±0.00
servo	0.95±0.05	0.98±0.03	0.98±0.03	0.95±0.05	0.98±0.03	0.99±0.02
shuttle-landing-control	0.50±0.00	0.80±0.45	0.70±0.45	0.50±0.00	0.80±0.45	0.80±0.45
solar-flare-1	0.89±0.03	0.90±0.04	0.90±0.03	0.78±0.03	0.90±0.03	0.90±0.04
solar-flare-2	0.92±0.01	0.92±0.01	0.92±0.02	0.92±0.01	0.92±0.02	0.92±0.01
soybean	0.98±0.01	0.97±0.01	0.95±0.03	0.79±0.02	0.97±0.01	1.00±0.00
spect	0.81±0.11	0.79±0.11	0.78±0.11	0.71±0.15	0.78±0.11	0.79±0.11
splice	0.96±0.01	0.98±0.01	0.98±0.01	0.58±0.13	0.98±0.01	0.99±0.00
tic-tac-toe	0.90±0.04	0.88±0.02	0.87±0.03	0.69±0.02	0.88±0.02	1.00±0.00
vote	0.98±0.02	0.98±0.02	0.98±0.02	0.97±0.03	0.98±0.02	0.99±0.01
Média	0.83±0.17	0.84±0.17	0.82±0.17	0.69±0.19	0.84±0.17	0.86±0.18
Rank Médio	3.59	3.10	3.65	5.54	3.17	1.9

Tabela 9 – Valores AUC: árvore de decisão, meta-árvore e floresta - 51 Datasets numéricos

Dataset	MT(mean)						MT(range)			
	DT	prel	lacc	nov4	sat	prel	lacc	nov4	sat	RF
analcdt-data-authorship	0.95±0.02	0.91±0.02	0.90±0.03	0.68±0.12	0.91±0.02	0.72±0.06	0.71±0.05	0.54±0.04	0.72±0.06	1.00±0.00
balance-scale	0.81±0.04	0.79±0.05	0.77±0.05	0.51±0.05	0.79±0.05	0.84±0.04	0.80±0.06	0.50±0.01	0.84±0.04	0.94±0.01
barknote-authentication	0.98±0.01	0.87±0.08	0.87±0.08	0.90±0.02	0.87±0.08	0.64±0.11	0.65±0.13	0.83±0.13	0.64±0.11	1.00±0.00
blood-transfusion-service	0.72±0.05	0.70±0.07	0.71±0.07	0.50±0.00	0.71±0.05	0.72±0.07	0.72±0.06	0.50±0.00	0.73±0.07	0.68±0.05
chatfield-figure	0.91±0.06	0.88±0.08	0.87±0.08	0.85±0.06	0.88±0.08	0.93±0.04	0.90±0.07	0.77±0.15	0.93±0.04	0.96±0.04
column-2C	0.85±0.08	0.81±0.15	0.82±0.15	0.83±0.07	0.81±0.15	0.85±0.05	0.84±0.04	0.77±0.11	0.85±0.05	0.93±0.04
column-3C	0.90±0.05	0.92±0.03	0.92±0.03	0.89±0.02	0.92±0.03	0.89±0.04	0.85±0.05	0.83±0.05	0.89±0.04	0.96±0.02
delta-elevators	0.90±0.01	0.92±0.01	0.92±0.01	0.84±0.02	0.92±0.01	0.90±0.01	0.90±0.01	0.84±0.02	0.90±0.01	0.94±0.01
disclosure-z	0.50±0.00	0.54±0.05	0.55±0.06	0.50±0.00	0.54±0.05	0.55±0.07	0.57±0.06	0.50±0.00	0.55±0.07	0.52±0.07
ecoli	0.92±0.05	0.85±0.09	0.74±0.05	0.50±0.00	0.85±0.09	0.78±0.03	0.76±0.04	0.50±0.00	0.78±0.03	0.97±0.02
eeg-eye-state	0.86±0.01	0.71±0.03	0.70±0.02	0.51±0.02	0.71±0.03	0.72±0.02	0.72±0.02	0.64±0.02	0.72±0.02	0.99±0.00
gas-drift	0.99±0.00	0.74±0.03	0.72±0.04	0.55±0.08	0.74±0.03	0.78±0.05	0.78±0.04	0.52±0.06	0.78±0.05	1.00±0.00
glass	0.81±0.06	0.65±0.07	0.50±0.00	0.50±0.00	0.65±0.07	0.68±0.05	0.50±0.00	0.50±0.00	0.68±0.05	0.95±0.02
hayes-roth	0.78±0.13	0.53±0.10	0.53±0.10	0.50±0.00	0.52±0.09	0.59±0.12	0.59±0.12	0.50±0.00	0.58±0.11	0.89±0.11
heart-statlog	0.76±0.10	0.78±0.07	0.78±0.07	0.50±0.00	0.78±0.07	0.77±0.07	0.76±0.10	0.50±0.00	0.77±0.07	0.91±0.04
hill-valley	0.50±0.00	0.51±0.02	0.50±0.03	0.50±0.00	0.51±0.02	0.54±0.05	0.53±0.05	0.50±0.00	0.54±0.05	0.62±0.04
ionosphere	0.89±0.05	0.66±0.09	0.68±0.10	0.62±0.09	0.66±0.09	0.68±0.12	0.66±0.10	0.59±0.07	0.68±0.12	0.98±0.02
iris	0.98±0.04	0.94±0.06	0.94±0.06	0.89±0.06	0.94±0.06	0.88±0.06	0.83±0.00	0.83±0.00	0.88±0.06	0.99±0.02
japanese-vowels	0.95±0.00	0.91±0.01	0.90±0.01	0.75±0.04	0.91±0.01	0.84±0.01	0.84±0.01	0.71±0.03	0.84±0.01	1.00±0.00
letter	0.95±0.00	0.90±0.01	0.87±0.01	0.50±0.00	0.90±0.01	0.83±0.01	0.82±0.01	0.60±0.04	0.83±0.01	1.00±0.00
madelon	0.69±0.03	0.61±0.04	0.61±0.03	0.50±0.00	0.61±0.04	0.60±0.05	0.60±0.06	0.50±0.00	0.60±0.05	0.71±0.02
magic-telescope	0.87±0.01	0.87±0.01	0.87±0.01	0.70±0.03	0.87±0.01	0.85±0.01	0.85±0.01	0.78±0.04	0.85±0.01	0.94±0.01
mfeat-factors	0.95±0.01	0.63±0.06	0.54±0.06	0.50±0.00	0.63±0.06	0.64±0.07	0.54±0.08	0.50±0.00	0.64±0.07	1.00±0.00
mfeat-fourier	0.89±0.02	0.80±0.01	0.76±0.03	0.50±0.00	0.80±0.01	0.79±0.03	0.72±0.04	0.50±0.00	0.79±0.03	0.98±0.00
mfeat-karhunen	0.92±0.01	0.84±0.02	0.79±0.02	0.50±0.00	0.84±0.02	0.79±0.02	0.74±0.04	0.50±0.00	0.79±0.02	1.00±0.00
optdigits	0.95±0.01	0.86±0.01	0.86±0.01	0.50±0.00	0.86±0.01	0.82±0.03	0.79±0.03	0.50±0.00	0.82±0.03	1.00±0.00
page-blocks	0.94±0.02	0.88±0.05	0.87±0.05	0.79±0.02	0.88±0.04	0.80±0.07	0.77±0.08	0.53±0.10	0.82±0.09	0.99±0.01
pc4	0.78±0.09	0.71±0.18	0.65±0.19	0.50±0.00	0.71±0.18	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00	0.94±0.02
pendigits	0.98±0.00	0.96±0.01	0.96±0.01	0.83±0.02	0.96±0.01	0.87±0.01	0.87±0.02	0.77±0.02	0.87±0.01	1.00±0.00
phoneme	0.89±0.02	0.88±0.02	0.88±0.01	0.74±0.02	0.88±0.02	0.86±0.03	0.86±0.02	0.70±0.11	0.86±0.03	0.97±0.01
pima-diabetes	0.75±0.08	0.79±0.06	0.79±0.06	0.67±0.03	0.79±0.06	0.70±0.08	0.67±0.09	0.70±0.08	0.70±0.08	0.82±0.05
pollen	0.50±0.00	0.50±0.05	0.50±0.03	0.50±0.00	0.50±0.05	0.51±0.05	0.51±0.04	0.49±0.03	0.51±0.05	0.49±0.04

Continua na próxima página

Tabela 9 — Continuação da página anterior

Dataset	MT(mean)						MT(range)			
	DT	prel	lacc	nov4	sat	prel	lacc	nov4	sat	RF
prmn-synth	0.85±0.08	0.87±0.07	0.81±0.08	0.84±0.07	0.87±0.07	0.85±0.08	0.87±0.08	0.80±0.08	0.85±0.08	0.93±0.05
qsar-biodeg	0.82±0.04	0.64±0.10	0.64±0.10	0.50±0.00	0.64±0.10	0.74±0.11	0.73±0.10	0.50±0.00	0.74±0.11	0.93±0.02
satimage	0.93±0.01	0.75±0.04	0.77±0.02	0.72±0.05	0.75±0.04	0.89±0.01	0.88±0.01	0.83±0.03	0.89±0.01	0.99±0.00
segment	0.99±0.00	0.80±0.04	0.80±0.03	0.72±0.02	0.80±0.04	0.87±0.02	0.86±0.02	0.74±0.02	0.87±0.02	1.00±0.00
semeion	0.87±0.02	0.68±0.02	0.50±0.00	0.50±0.00	0.68±0.02	0.68±0.02	0.50±0.00	0.50±0.00	0.68±0.02	1.00±0.00
sonar	0.73±0.08	0.65±0.10	0.59±0.10	0.50±0.00	0.65±0.10	0.54±0.06	0.53±0.06	0.50±0.00	0.54±0.06	0.95±0.03
spambase	0.94±0.01	0.64±0.14	0.64±0.14	0.52±0.01	0.64±0.14	0.89±0.01	0.88±0.01	0.53±0.09	0.89±0.01	0.99±0.00
steel-plates-fault	1.00±0.00	1.00±0.01	1.00±0.01	0.99±0.02	1.00±0.01	0.99±0.02	0.99±0.02	0.94±0.02	0.99±0.02	1.00±0.00
strikes	1.00±0.00	0.90±0.06	0.89±0.06	0.69±0.07	0.90±0.06	0.77±0.06	0.76±0.07	0.72±0.06	0.77±0.06	1.00±0.00
transplant	0.99±0.02	0.76±0.11	0.76±0.11	0.76±0.11	0.76±0.11	0.94±0.09	0.94±0.09	0.94±0.09	0.94±0.09	1.00±0.00
triazines	0.83±0.13	0.57±0.08	0.58±0.08	0.50±0.00	0.57±0.08	0.54±0.04	0.54±0.04	0.50±0.00	0.54±0.04	0.87±0.09
vehicle	0.86±0.03	0.74±0.04	0.73±0.03	0.50±0.00	0.74±0.04	0.73±0.05	0.70±0.04	0.50±0.00	0.73±0.05	0.94±0.01
visualizing-galaxy	0.96±0.04	0.80±0.11	0.80±0.11	0.83±0.06	0.80±0.11	0.89±0.07	0.90±0.07	0.84±0.08	0.89±0.07	0.99±0.01
wdbc	0.93±0.04	0.62±0.13	0.62±0.13	0.66±0.13	0.62±0.13	0.91±0.05	0.91±0.05	0.90±0.05	0.91±0.05	0.99±0.01
wilf2	0.93±0.03	0.82±0.07	0.81±0.06	0.51±0.02	0.82±0.07	0.89±0.07	0.90±0.06	0.50±0.00	0.89±0.07	0.99±0.02
wine	0.96±0.05	0.69±0.11	0.70±0.11	0.50±0.00	0.69±0.11	0.67±0.08	0.67±0.08	0.50±0.00	0.67±0.08	1.00±0.00
wisconsin	0.49±0.09	0.51±0.08	0.50±0.01	0.50±0.00	0.51±0.08	0.49±0.03	0.49±0.03	0.50±0.00	0.49±0.03	0.56±0.08
wisconsin-breast-cancer	0.95±0.05	0.96±0.04	0.96±0.05	0.92±0.04	0.96±0.04	0.95±0.03	0.95±0.03	0.91±0.07	0.95±0.03	0.99±0.01
yeast	0.73±0.03	0.78±0.02	0.76±0.02	0.50±0.00	0.78±0.02	0.75±0.04	0.74±0.04	0.71±0.02	0.75±0.04	0.84±0.02
Média	0.86±0.13	0.76±0.13	0.75±0.14	0.63±0.16	0.76±0.13	0.76±0.13	0.74±0.14	0.63±0.15	0.76±0.13	0.92±0.13
Rank Médio	3.20	4.87	5.83	8.72	4.87	5.35	6.35	9.04	5.31	1.42

Tabela 10 – Valores AUC: árvore de decisão, meta-árvore e floresta - 56 Datasets categóricos e numéricos

Dataset	MT(mean)						MT(range)			RF
	DT	prel	lacc	nov4	sat	prel	lacc	nov4	sat	
analcatdata-AIDS	0.66±0.25	0.43±0.20	0.33±0.24	0.50±0.00	0.43±0.20	0.44±0.24	0.33±0.24	0.50±0.00	0.44±0.24	0.43±0.41
analcatdata-asbestos	0.76±0.15	0.79±0.13	0.79±0.13	0.80±0.14	0.79±0.13	0.80±0.14	0.80±0.14	0.80±0.14	0.80±0.14	0.83±0.15
analcatdata-bondrate	0.54±0.15	0.56±0.22	0.50±0.13	0.50±0.00	0.56±0.22	0.56±0.22	0.50±0.13	0.50±0.00	0.56±0.22	0.63±0.26
analcatdata-braziltourism	0.52±0.05	0.65±0.12	0.66±0.14	0.62±0.13	0.64±0.12	0.65±0.13	0.66±0.14	0.62±0.13	0.65±0.13	0.68±0.12
analcatdata-broadway	0.66±0.11	0.63±0.18	0.64±0.19	0.50±0.00	0.64±0.18	0.63±0.18	0.65±0.19	0.50±0.00	0.63±0.19	0.73±0.12
analcatdata-broadwaymult	0.71±0.06	0.76±0.05	0.76±0.05	0.74±0.10	0.76±0.05	0.76±0.05	0.76±0.05	0.74±0.10	0.76±0.05	0.67±0.08
analcatdata-creditscore	0.99±0.02	0.77±0.10	0.77±0.10	0.77±0.10	0.77±0.10	0.99±0.02	0.99±0.02	0.99±0.02	0.99±0.02	0.99±0.03
analcatdata-cyyoung8092	0.69±0.15	0.59±0.24	0.61±0.24	0.50±0.00	0.59±0.24	0.58±0.25	0.58±0.23	0.50±0.00	0.58±0.25	0.81±0.16
analcatdata-cyyoung9302	0.73±0.20	0.74±0.21	0.70±0.22	0.50±0.00	0.74±0.21	0.75±0.21	0.69±0.22	0.50±0.00	0.75±0.21	0.84±0.16
analcatdata-homerun	0.52±0.08	0.53±0.06	0.50±0.00	0.50±0.00	0.53±0.06	0.53±0.07	0.50±0.00	0.50±0.00	0.53±0.07	0.49±0.11
analcatdata-lawsuit	0.92±0.17	0.89±0.07	0.89±0.07	0.89±0.07	0.89±0.07	0.78±0.18	0.75±0.20	0.64±0.12	0.78±0.18	0.99±0.02
analcatdata-vineyard	0.90±0.03	0.89±0.03	0.84±0.05	0.83±0.04	0.89±0.03	0.86±0.04	0.84±0.05	0.83±0.04	0.86±0.04	0.95±0.02
analcatdata-votesurvey	0.62±0.25	0.63±0.21	0.61±0.22	0.47±0.14	0.63±0.21	0.63±0.21	0.61±0.22	0.47±0.14	0.63±0.21	0.59±0.27
analcatdata-wildcat	0.77±0.09	0.78±0.09	0.74±0.08	0.50±0.00	0.78±0.09	0.76±0.11	0.75±0.10	0.50±0.00	0.76±0.11	0.83±0.08
anneal	0.99±0.01	0.94±0.03	0.93±0.03	0.91±0.03	0.94±0.03	0.93±0.02	0.93±0.03	0.91±0.03	0.93±0.02	1.00±0.00
arsenic-female-bladder	0.50±0.00	0.80±0.08	0.79±0.11	0.50±0.00	0.80±0.08	0.69±0.10	0.68±0.11	0.50±0.00	0.69±0.10	0.79±0.09
autos	0.92±0.03	0.81±0.08	0.81±0.08	0.59±0.10	0.81±0.08	0.81±0.08	0.82±0.08	0.59±0.10	0.80±0.08	0.97±0.03
biomed	0.89±0.04	0.67±0.10	0.66±0.11	0.60±0.09	0.67±0.10	0.61±0.19	0.61±0.19	0.72±0.11	0.61±0.19	0.96±0.03
bridges-version1	0.77±0.11	0.57±0.13	0.72±0.07	0.58±0.13	0.56±0.12	0.56±0.12	0.72±0.07	0.58±0.13	0.55±0.12	0.83±0.09
cars	0.93±0.04	0.86±0.05	0.82±0.06	0.82±0.03	0.86±0.05	0.87±0.03	0.87±0.03	0.83±0.03	0.87±0.03	0.97±0.02
churn	0.83±0.04	0.78±0.05	0.79±0.05	0.78±0.06	0.78±0.05	0.77±0.05	0.76±0.06	0.70±0.04	0.77±0.05	0.91±0.03
cleveland-14-heart-disease	0.80±0.09	0.85±0.06	0.85±0.05	0.76±0.09	0.85±0.06	0.86±0.06	0.85±0.05	0.76±0.09	0.86±0.06	0.90±0.07
cloud	0.92±0.08	0.54±0.18	0.55±0.18	0.50±0.00	0.54±0.18	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00	0.78±0.11
cmc	0.66±0.04	0.71±0.03	0.70±0.03	0.50±0.00	0.71±0.03	0.68±0.04	0.67±0.04	0.62±0.03	0.68±0.03	0.68±0.02
collins	0.99±0.02	1.00±0.00	1.00±0.00	0.62±0.11	1.00±0.00	1.00±0.00	1.00±0.00	0.50±0.00	1.00±0.00	1.00±0.00
credit-rating	0.89±0.03	0.90±0.03	0.91±0.03	0.86±0.04	0.90±0.03	0.90±0.04	0.90±0.04	0.86±0.04	0.90±0.04	0.93±0.03
cylinder-bands	0.77±0.03	0.64±0.10	0.63±0.10	0.50±0.00	0.64±0.10	0.64±0.10	0.63±0.10	0.50±0.00	0.64±0.10	0.88±0.06
dermatology	0.97±0.02	0.94±0.04	0.94±0.03	0.80±0.06	0.95±0.04	0.94±0.04	0.94±0.04	0.80±0.06	0.94±0.04	1.00±0.00
dresses-sales	0.51±0.04	0.56±0.10	0.58±0.09	0.50±0.00	0.56±0.09	0.56±0.10	0.58±0.09	0.50±0.00	0.56±0.09	0.55±0.07
electricity	0.93±0.00	0.79±0.03	0.79±0.03	0.76±0.02	0.79±0.03	0.79±0.01	0.79±0.01	0.73±0.01	0.79±0.01	0.98±0.00
eucalyptus	0.84±0.03	0.76±0.04	0.74±0.04	0.50±0.00	0.76±0.04	0.74±0.03	0.74±0.04	0.50±0.00	0.74±0.03	0.86±0.02
fruityfly	0.52±0.09	0.40±0.10	0.40±0.16	0.50±0.00	0.40±0.13	0.52±0.14	0.56±0.14	0.50±0.00	0.48±0.11	0.52±0.20

Continua na próxima página

Tabela 10 — Continuação da página anterior

Dataset	MT(mean)						RF			
	DT	prel	lacc	nov4	sat	prel	lacc	nov4	sat	
german-credit	0.65±0.07	0.74±0.06	0.74±0.07	0.60±0.07	0.74±0.06	0.73±0.06	0.74±0.06	0.60±0.07	0.73±0.06	0.79±0.06
grub-damage	0.60±0.08	0.66±0.10	0.68±0.10	0.50±0.00	0.66±0.10	0.67±0.09	0.68±0.09	0.50±0.00	0.67±0.09	0.65±0.11
haberman	0.58±0.08	0.56±0.13	0.56±0.13	0.50±0.00	0.56±0.13	0.57±0.15	0.58±0.12	0.50±0.00	0.56±0.12	0.65±0.11
hepatitis	0.70±0.20	0.60±0.16	0.58±0.16	0.50±0.00	0.60±0.16	0.50±0.00	0.50±0.00	0.50±0.00	0.50±0.00	0.87±0.08
horse-colic	0.85±0.04	0.92±0.05	0.92±0.05	0.81±0.04	0.92±0.05	0.92±0.05	0.92±0.05	0.81±0.04	0.92±0.05	0.94±0.04
hungarian-14-heart-disease	0.77±0.15	0.83±0.07	0.82±0.06	0.82±0.07	0.83±0.07	0.86±0.07	0.85±0.06	0.82±0.07	0.86±0.07	0.88±0.07
hypothyroid	1.00±0.01	0.98±0.04	0.98±0.04	0.93±0.06	0.98±0.04	0.89±0.10	0.86±0.10	0.89±0.06	0.90±0.09	1.00±0.00
ilpd	0.68±0.06	0.62±0.06	0.60±0.07	0.50±0.00	0.62±0.06	0.57±0.09	0.52±0.05	0.50±0.00	0.57±0.09	0.75±0.08
iris	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
lymphography	0.79±0.14	0.84±0.10	0.83±0.13	0.82±0.12	0.84±0.10	0.84±0.10	0.83±0.12	0.82±0.12	0.84±0.10	0.93±0.07
musk	1.00±0.00	0.95±0.16	1.00±0.00	0.92±0.01	0.95±0.16	0.65±0.24	0.80±0.26	0.55±0.15	0.65±0.24	1.00±0.00
newton-hema	0.84±0.13	0.83±0.09	0.83±0.09	0.81±0.10	0.83±0.09	0.82±0.10	0.82±0.09	0.81±0.10	0.82±0.10	0.86±0.09
pasture-production	0.84±0.16	0.86±0.15	0.67±0.18	0.50±0.00	0.86±0.15	0.86±0.15	0.67±0.18	0.50±0.00	0.86±0.15	0.94±0.10
prmn-viruses	0.98±0.06	0.54±0.14	0.50±0.00	0.50±0.00	0.54±0.14	0.54±0.14	0.50±0.00	0.50±0.00	0.54±0.14	1.00±0.00
sick	0.95±0.04	0.93±0.05	0.93±0.05	0.90±0.05	0.93±0.05	0.90±0.04	0.89±0.04	0.58±0.13	0.90±0.04	1.00±0.00
speed-dating	1.00±0.00	0.85±0.24	0.80±0.26	1.00±0.00	0.85±0.24	0.60±0.21	0.65±0.24	1.00±0.00	0.65±0.24	0.99±0.00
squash-stored	0.73±0.17	0.44±0.21	0.50±0.00	0.50±0.00	0.44±0.22	0.43±0.20	0.50±0.00	0.50±0.00	0.44±0.19	0.84±0.17
squash-unstored	0.82±0.23	0.48±0.06	0.50±0.00	0.50±0.00	0.48±0.06	0.47±0.08	0.50±0.00	0.50±0.00	0.47±0.08	0.89±0.16
stress	0.83±0.23	0.69±0.17	0.65±0.17	0.58±0.11	0.69±0.17	0.53±0.08	0.53±0.08	0.50±0.00	0.53±0.08	0.96±0.06
tae	0.74±0.10	0.70±0.09	0.69±0.09	0.50±0.00	0.69±0.09	0.65±0.11	0.60±0.08	0.50±0.00	0.65±0.11	0.82±0.07
veteran	0.65±0.10	0.74±0.19	0.76±0.19	0.64±0.17	0.74±0.19	0.67±0.17	0.67±0.17	0.64±0.17	0.67±0.17	0.73±0.14
visualizing-livestock	0.68±0.05	0.06±0.02	0.05±0.03	0.50±0.00	0.06±0.02	0.05±0.03	0.05±0.03	0.50±0.00	0.05±0.03	0.04±0.02
vowel	0.94±0.03	0.81±0.03	0.59±0.03	0.50±0.00	0.81±0.03	0.77±0.02	0.57±0.01	0.50±0.00	0.77±0.02	1.00±0.00
zoo	0.98±0.04	0.98±0.02	0.97±0.02	0.93±0.02	0.98±0.02	0.98±0.02	0.97±0.02	0.93±0.02	0.98±0.02	1.00±0.00
Média	0.79±0.15	0.73±0.18	0.72±0.18	0.65±0.17	0.73±0.18	0.71±0.18	0.70±0.18	0.64±0.16	0.71±0.18	0.83±0.18
Rank Médio	4.29	4.59	5.33	7.94	4.61	5.56	6.25	8.21	5.73	2.44

Tabela 11 – Média dos valores de AUC e o ranking médio: árvore de decisão, meta-árvore e floresta - Todos os 150 Datasets

	DT	MT(mean)				MT(range)			
		prel	lacc	nov4	sat	prel	lacc	nov4	sat
Média	0.82±0.15	0.77±0.17	0.75±0.17	0.65±0.17	0.77±0.16	0.76±0.17	0.75±0.17	0.65±0.17	0.76±0.17
Rank Médio	4.23	4.64	5.55	8.47	4.71	5.17	6.08	8.68	5.28

Tabela 12 – Número de folhas: árvore de decisão e meta-árvore - 43 Datasets categóricos

Dataset	DT	MT			
		prel	lacc	nov4	sat
analecatdata-boxing1	13.90±0.30	12.00±0.00	12.00±0.00	1.00±0.00	12.00±0.00
analecatdata-boxing2	13.60±0.66	12.00±0.00	12.00±0.00	1.00±0.00	12.00±0.00
analecatdata-dmft	113.90±14.15	148.20±11.74	103.20±6.66	3.90±1.14	149.40±15.49
analecatdata-donner	1.00±0.00	8.00±0.00	8.00±0.00	1.00±0.00	8.00±0.00
analecatdata-marketing	91.20±16.84	1.40±1.20	1.40±1.20	1.00±0.00	1.40±1.20
analecatdata-reviewer	5.20±2.09	30.80±3.40	20.00±3.13	1.00±0.00	27.80±2.99
analecatdata-fraud	5.50±3.77	2.80±0.60	2.70±0.64	1.90±0.30	2.80±0.60
audiology	30.20±2.64	37.20±6.08	6.00±3.52	2.00±0.00	37.20±6.08
audiology-binary	8.90±0.94	13.00±1.73	12.40±1.96	2.00±0.00	13.40±1.69
balloons-adult+stretch	3.00±0.00	3.00±0.00	3.00±0.00	2.40±0.49	3.00±0.00
balloons-adult-stretch	3.00±0.00	3.00±0.00	3.00±0.00	2.20±0.40	3.00±0.00
balloons-yellow-small	3.00±0.00	3.00±0.00	3.00±0.00	2.30±0.46	3.00±0.00
logger	9.40±3.83	16.90±4.70	14.50±4.18	3.40±0.80	15.70±5.10
breast-cancer	7.50±7.42	66.30±13.93	50.40±14.57	1.00±0.00	68.20±13.18
car-df	120.40±4.15	132.50±2.46	84.20±3.82	1.00±0.00	132.50±2.46
contact-lenses	3.60±0.49	3.00±0.00	3.00±0.00	2.20±0.40	3.00±0.00
dbworld-subjects	7.60±0.66	3.60±0.49	2.50±0.81	1.00±0.00	3.60±0.49
dbworld-subjects-stemmed	6.60±0.80	3.60±0.66	2.80±0.75	1.00±0.00	3.60±0.66
dna	85.80±5.33	47.10±6.77	33.60±3.44	3.00±0.00	47.10±6.77
king-and-rook	4560.10±115.79	8589.30±123.27	1191.00±19.55	1.00±0.00	8589.30±123.27
kr-vs-skp	29.40±2.20	40.20±2.99	37.20±3.68	4.00±0.00	40.20±2.99
lung-cancer	8.50±2.50	1.60±0.92	1.00±0.00	1.00±0.00	1.60±0.92
molecular-promotor-gene	17.20±1.99	5.80±1.47	4.30±0.90	1.00±0.00	5.80±1.47
molecular-splice-junction	175.70±15.29	155.30±8.72	146.20±5.93	2.60±1.96	155.50±9.02
monks-problems	24.50±28.82	107.70±4.41	99.20±3.68	1.00±0.00	107.40±3.95
mushroom	25.00±0.00	303.40±79.00	299.80±73.00	42.90±16.69	303.40±79.00
mushroom-expanded	24.40±0.49	248.40±73.18	249.40±78.23	56.00±20.49	248.40±73.18
nursery	352.90±7.88	438.70±8.59	314.50±5.85	17.00±0.00	438.70±8.59
nursery-4class	356.70±12.91	433.20±7.67	324.90±10.76	17.00±0.00	433.20±7.67
phishing	175.20±14.47	235.20±14.72	219.30±17.06	14.40±0.80	221.90±14.43
post-operative-patient	1.80±1.60	8.10±4.59	5.20±2.04	1.00±0.00	9.70±3.63
postoperative-patient-data	1.40±1.20	8.40±3.93	4.90±2.34	1.00±0.00	10.20±3.99

Continua na próxima página

Tabela 12 — Continuação da página anterior

Dataset	DT	MT		
		prel	lacc	nov4
primary-tumor	45.30±3.82	44.80±2.93	9.00±1.48	2.00±0.00
qualitative-bankruptcy	4.40±0.92	6.20±2.23	6.00±2.24	3.00±0.00
servo	4.80±2.40	14.00±2.68	10.00±2.00	4.00±0.00
shuttle-landing-control	1.00±0.00	4.00±0.00	2.80±1.47	1.00±0.00
solar-flare-1	21.90±1.76	35.30±4.08	19.70±3.69	5.80±2.40
solar-flare-2	45.00±12.30	101.80±13.53	66.70±6.51	9.00±0.00
soybean	61.30±5.69	87.20±13.62	26.00±2.83	4.40±1.20
spect	8.40±1.96	8.90±0.70	8.50±0.81	1.90±0.30
splice	175.70±15.29	155.10±9.48	144.20±4.89	2.20±1.83
tic-tac-toe	89.00±3.79	71.60±2.69	66.40±4.10	3.00±0.00
vote	5.80±0.40	9.90±2.51	9.10±2.07	3.40±0.92
Média	156.94±684.44	271.19±1288.05	84.72±192.94	5.43±10.55
Rank Médio	2.80	1.95	3.32	4.91
				2.0

Tabela 13 – Número de folhas: árvore de decisão e meta-árvore - 51 Datasets numéricos

Dataset	DT	MT(mean)						MT(range)	
		prel	lacc	nov4	sat	prel	lacc	nov4	sat
analcatdata-authorship	20.40±1.02	4.30±0.78	3.60±0.66	1.80±0.40	4.30±0.78	4.80±0.40	4.70±0.46	1.80±0.40	4.80±0.40
balance-scale	39.60±5.70	7.80±0.60	6.30±0.46	1.10±0.30	7.80±0.60	9.80±1.08	8.60±1.02	1.10±0.30	9.80±1.08
banknote-authentication	14.90±1.76	12.60±1.56	11.20±1.17	3.20±0.60	12.60±1.56	23.30±3.61	23.20±2.23	7.80±1.47	23.30±3.61
blood-transfusion-service	6.00±1.61	12.70±3.61	10.50±2.84	1.00±0.00	12.10±2.70	10.70±2.28	9.70±1.55	1.00±0.00	9.60±2.54
chatfield-figure	6.80±1.33	3.00±0.00	2.90±0.30	2.00±0.00	3.00±0.00	3.20±0.60	3.00±0.63	1.80±0.40	3.20±0.60
column-2C	10.00±2.19	4.60±0.66	4.30±0.46	2.00±0.00	4.60±0.66	4.90±1.04	5.10±2.21	1.90±0.30	4.90±1.04
column-3C	12.30±3.13	4.10±1.14	3.40±0.66	2.00±0.00	4.10±1.14	5.50±1.63	3.50±1.80	2.00±0.00	5.50±1.63
delta-elevators	107.90±20.66	114.90±11.94	98.90±5.84	2.20±0.40	114.90±11.94	114.80±14.08	98.30±11.37	2.20±0.40	114.80±14.08
disclosure-z	1.00±0.00	26.80±3.22	22.40±2.42	1.00±0.00	26.80±3.22	21.00±3.69	17.40±1.69	1.00±0.00	21.00±3.69
ecoli	18.30±1.79	2.70±0.46	2.00±0.00	1.00±0.00	2.70±0.46	2.00±0.00	2.00±0.00	1.00±0.00	2.00±0.00
eeg-eye-state	725.50±22.39	325.20±28.82	299.50±21.29	3.80±0.98	325.20±28.82	263.90±14.69	241.80±12.06	3.00±0.00	263.90±14.69
gas-drift	176.00±8.56	24.30±3.03	23.30±4.38	1.80±0.98	24.30±3.03	17.20±3.25	17.20±2.56	1.20±0.60	17.20±3.25
Glass	22.60±2.62	2.00±0.00	1.00±0.00	1.00±0.00	2.00±0.00	2.00±0.00	1.00±0.00	1.00±0.00	2.00±0.00
hayes-roth	8.60±2.54	2.20±0.40	2.00±0.00	1.00±0.00	2.00±0.00	5.00±1.73	4.20±1.72	1.00±0.00	4.80±1.72
heart-statlog	17.40±1.85	3.40±0.49	3.40±0.49	1.00±0.00	3.40±0.49	4.50±0.50	4.30±0.78	1.00±0.00	4.50±0.50
hill-valley	1.00±0.00	13.70±3.69	12.60±2.80	1.00±0.00	13.70±3.69	7.80±2.40	7.60±1.02	1.00±0.00	7.80±2.40
ionosphere	14.20±1.78	4.00±1.00	3.70±0.90	2.40±0.49	4.00±1.00	3.40±0.49	3.40±0.49	2.50±0.50	3.40±0.49
iris	4.70±0.46	3.00±0.00	3.00±0.00	2.80±0.40	3.00±0.00	2.40±0.49	2.00±0.00	2.00±0.00	2.40±0.49
japanese-vowels	432.10±8.58	45.20±3.87	31.40±2.73	3.00±0.00	45.20±3.87	54.90±1.76	42.50±2.54	3.10±0.30	54.90±1.76
letter	1158.10±11.29	47.00±3.26	28.00±1.41	1.00±0.00	47.00±3.26	42.90±2.34	31.30±1.00	2.20±0.40	42.90±2.34
madelon	165.80±7.01	15.30±5.37	15.00±5.80	1.00±0.00	15.30±5.37	10.50±3.11	10.20±2.52	1.00±0.00	10.50±3.11
magic-telescope	327.50±19.92	467.90±18.87	436.00±35.38	5.20±0.87	467.90±18.87	332.80±19.83	305.20±28.21	6.60±1.02	332.80±19.83
mfeat-factors	71.70±3.47	1.90±0.30	1.30±0.46	1.00±0.00	1.90±0.30	2.00±0.45	1.30±0.46	1.00±0.00	2.00±0.45
mfeat-fourier	124.90±10.08	6.50±0.50	3.80±0.40	1.00±0.00	6.50±0.50	5.60±0.92	3.00±0.63	1.00±0.00	5.60±0.92
mfeat-karhunen	115.40±3.58	7.20±0.40	5.00±0.00	1.00±0.00	7.20±0.40	7.10±0.30	4.60±0.80	1.00±0.00	7.10±0.30
optdigits	208.80±8.73	12.50±1.57	10.90±1.14	1.00±0.00	12.50±1.57	15.50±1.75	9.40±1.36	1.00±0.00	15.50±1.75
page-blocks	45.90±5.61	14.80±4.94	12.30±4.03	4.10±0.54	15.20±3.92	12.20±6.01	14.40±5.18	2.50±1.86	12.90±6.01
pc4	41.70±10.39	4.50±3.11	3.30±3.00	1.00±0.00	4.50±3.11	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
pendigits	188.30±6.00	45.40±2.20	42.00±4.36	6.30±0.46	45.40±2.20	45.80±3.09	41.30±2.53	7.00±0.89	45.80±3.09
phoneme	119.20±17.50	114.50±12.08	103.80±9.04	2.00±0.00	114.50±12.08	110.20±11.80	98.80±7.33	2.50±0.92	110.20±11.80
pima-diabetes	19.20±6.19	19.40±3.90	17.10±2.17	2.00±0.00	19.40±3.90	10.90±3.48	12.70±3.80	2.00±0.00	10.90±3.48
pollen	1.10±0.30	292.10±22.68	258.20±20.99	1.00±0.00	292.10±22.68	236.30±28.93	195.20±19.48	2.90±0.30	236.30±28.93

Continua na próxima página

Tabela 13 — Continuação da página anterior

Dataset	DT	MT(mean)				MT(range)			
		prel	lacc	nov4	sat	prel	lacc	nov4	sat
pnnt-synth	4.20±1.08	3.90±0.70	3.90±1.04	2.00±0.00	3.90±0.70	5.60±1.96	4.80±1.25	2.00±0.00	5.60±1.96
qsar-biodeg	53.40±8.64	6.30±0.90	6.10±1.04	1.00±0.00	6.30±0.90	6.50±1.86	5.50±1.75	1.00±0.00	6.50±1.86
satimage	275.90±13.78	26.20±3.22	22.80±3.74	4.40±0.49	26.20±3.22	22.00±2.28	18.00±2.10	3.80±0.40	22.00±2.28
segment	41.10±2.51	12.20±0.75	11.90±0.94	3.90±0.30	12.20±0.75	12.80±0.87	10.60±0.66	3.10±0.30	12.80±0.87
semeion	137.60±3.88	2.00±0.00	1.00±0.00	1.00±0.00	2.00±0.00	2.10±0.30	1.00±0.00	1.00±0.00	2.10±0.30
sonar	15.10±1.70	1.80±0.40	1.50±0.50	1.00±0.00	1.80±0.40	1.80±0.40	1.50±0.50	1.00±0.00	1.80±0.40
spambase	104.20±8.32	39.20±6.54	36.30±6.15	3.40±0.49	38.90±6.19	39.00±3.69	35.90±3.14	3.40±0.49	39.10±3.62
steel-plates-fault	7.00±0.00	7.40±0.80	7.30±0.78	6.80±0.75	7.40±0.80	18.00±4.43	17.80±4.31	7.90±1.45	18.00±4.43
strikes	11.30±2.28	16.60±1.91	15.30±2.10	2.20±0.40	16.60±1.91	17.00±0.77	16.10±1.22	2.00±0.00	17.00±0.77
transplant	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00	4.00±0.89	4.00±0.89	4.00±0.89	4.00±0.89
triazines	14.20±3.06	2.20±0.40	2.30±0.46	1.00±0.00	2.30±0.46	2.60±0.66	2.40±0.49	1.00±0.00	2.60±0.66
vehicle	68.50±10.18	9.00±1.26	6.50±0.81	1.00±0.00	9.00±1.26	6.80±1.54	5.90±0.83	1.00±0.00	6.80±1.54
visualizing-galaxy	6.30±0.78	2.30±0.64	2.10±0.30	2.00±0.00	2.30±0.64	4.60±0.66	4.60±0.92	3.00±0.89	4.60±0.66
wdbc	11.80±2.52	3.10±0.30	3.10±0.30	2.40±0.49	3.10±0.30	4.10±1.04	4.20±0.98	2.20±0.40	4.10±1.04
wilt2	27.60±4.72	32.20±6.40	28.40±6.05	1.90±0.30	32.20±6.40	38.50±9.19	35.80±12.69	1.00±0.00	38.50±9.19
wine	5.40±0.66	2.00±0.00	2.00±0.00	1.00±0.00	2.00±0.00	2.00±0.00	1.00±0.00	1.00±0.00	2.00±0.00
wisconsin	10.40±5.54	2.50±2.38	2.10±2.07	1.00±0.00	2.50±2.38	1.30±0.64	1.30±0.64	1.00±0.00	1.30±0.64
wisconsin-breast-cancer	12.20±2.86	5.70±0.64	5.60±0.66	2.80±0.40	5.70±0.64	5.60±1.62	5.60±1.62	2.60±0.49	5.60±1.62
yeast	155.60±11.94	8.70±0.78	5.60±0.92	1.00±0.00	8.70±0.78	5.70±1.62	4.10±0.30	3.90±0.54	5.70±1.62
Média	101.78±197.04	36.29±86.61	32.31±79.58	2.05±1.39	36.28±86.61	31.25±66.51	27.63±59.50	2.25±1.73	31.24±66.52
Rank Médio	2.35	3.66	5.61	8.31	3.72	3.80	5.55	8.13	3.82

Tabela 14 – Número de folhas: árvore de decisão e meta-árvore - 56 Datasets categóricos e numéricos

Dataset	DT	MT(mean)						MT(range)		
		prel	lacc	nov4	sat	prel	lacc	nov4	sat	sat
analcatdata-AIDS	3.70±2.61	3.40±1.74	5.00±0.00	1.00±0.00	3.40±1.74	5.00±1.55	5.00±0.00	1.00±0.00	5.00±1.55	5.00±1.55
analcatdata-asbestos	6.10±1.58	5.40±1.69	5.40±1.62	3.00±0.00	5.30±1.73	3.90±0.70	3.70±0.46	3.00±0.00	3.80±0.75	3.80±0.75
analcatdata-bondrate	7.10±7.02	5.90±2.02	4.00±2.45	1.00±0.00	5.90±2.02	5.90±2.02	4.00±2.45	1.00±0.00	5.90±2.02	5.90±2.02
analcatdata-braziltourism	6.40±7.46	13.70±3.61	12.60±4.45	5.40±2.33	13.70±3.61	13.40±3.67	12.60±4.45	5.40±2.33	13.40±3.67	13.40±3.67
analcatdata-broadway	12.10±3.81	8.70±3.41	7.00±0.00	1.00±0.00	8.70±3.41	8.70±3.41	7.00±0.00	1.00±0.00	8.70±3.41	8.70±3.41
analcatdata-broadwaymult	32.90±3.42	3.00±0.00	3.00±0.00	2.80±0.60	3.00±0.00	3.00±0.00	3.00±0.00	2.80±0.60	3.00±0.00	3.00±0.00
analcatdata-creditscore	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00	2.40±0.80	2.00±0.00	2.00±0.00	2.40±0.80	2.40±0.80
analcatdata-cyyoung8092	3.00±1.41	62.00±0.00	1.00±0.00	62.00±0.00	62.00±0.00	62.00±0.00	62.00±0.00	1.00±0.00	62.00±0.00	62.00±0.00
analcatdata-cyyoung9302	3.10±0.83	59.00±0.00	53.20±17.40	1.00±0.00	59.00±0.00	59.00±0.00	53.20±17.40	1.00±0.00	59.00±0.00	59.00±0.00
analcatdata-homerun	28.00±11.55	2.50±2.42	1.00±0.00	1.00±0.00	2.50±2.42	2.50±2.42	1.00±0.00	1.00±0.00	2.50±2.42	2.50±2.42
analcatdata-lawsuit	3.20±0.60	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00	2.30±0.90	2.40±0.80	2.00±0.00	2.30±0.90	2.30±0.90
analcatdata-vineyard	25.00±1.55	14.80±1.08	10.60±0.66	9.00±0.00	14.80±1.08	12.30±0.64	10.00±0.77	9.00±0.00	12.30±0.64	12.30±0.64
analcatdata-voicesurvey	9.60±1.20	2.00±0.00	1.90±0.30	1.20±0.40	2.00±0.00	2.00±0.00	1.90±0.30	1.20±0.40	2.00±0.00	2.00±0.00
analcatdata-wildcat	6.80±1.78	5.40±0.66	4.40±0.92	1.00±0.00	5.40±0.66	4.00±0.00	4.00±0.00	1.00±0.00	4.00±0.00	4.00±0.00
anneal	39.50±5.00	52.50±10.28	38.10±5.72	19.50±6.50	52.50±10.28	50.10±9.44	37.60±6.09	19.50±6.50	50.10±9.44	50.10±9.44
arsenic-female-bladder	1.00±0.00	81.70±12.90	77.40±17.20	1.00±0.00	81.70±12.90	43.00±0.00	43.00±0.00	1.00±0.00	43.00±0.00	43.00±0.00
autos	43.90±7.38	26.00±3.29	22.60±0.49	2.10±2.02	26.00±3.29	26.00±3.29	22.60±0.49	2.10±2.02	26.00±3.29	26.00±3.29
biomed	6.80±2.36	3.30±0.64	3.30±0.64	2.00±0.00	3.30±0.64	5.00±0.77	4.90±0.83	2.00±0.00	5.00±0.77	5.00±0.77
bridges-version1	19.70±20.19	54.00±0.00	3.30±0.90	1.60±0.92	54.00±0.00	54.00±0.00	3.30±0.90	1.60±0.92	54.00±0.00	54.00±0.00
cars	31.00±2.93	8.20±0.40	7.70±1.49	4.80±1.83	8.20±0.40	7.90±1.51	7.80±1.89	3.60±1.96	8.00±1.48	8.00±1.48
churn	84.90±5.73	71.30±10.25	68.60±13.37	12.00±0.00	71.30±10.25	67.10±6.55	63.90±11.18	11.00±0.00	67.10±6.55	67.10±6.55
cleveland-14-heart-disease	26.00±6.56	17.00±4.40	13.70±2.93	4.60±1.43	17.00±4.40	16.10±2.84	13.20±2.36	4.60±1.43	16.10±2.84	16.10±2.84
cloud	8.00±0.00	3.40±0.92	3.20±0.98	1.00±0.00	3.40±0.92	1.70±0.46	1.60±0.49	1.00±0.00	1.70±0.46	1.70±0.46
cmc	155.70±18.90	221.40±19.24	152.10±13.02	1.00±0.00	220.10±17.73	201.80±14.71	140.00±11.86	3.10±0.30	199.00±12.43	199.00±12.43
collins	3.00±0.00	15.00±0.00	15.00±0.00	1.60±0.49	15.00±0.00	15.00±0.00	15.00±0.00	1.00±0.00	15.00±0.00	15.00±0.00
credit-rating	19.20±6.46	83.20±24.07	75.40±22.08	2.00±0.00	83.20±27.13	55.00±19.79	47.70±13.79	2.00±0.00	55.00±19.79	55.00±19.79
cylinder-bands	139.30±61.05	138.10±16.41	129.10±26.91	1.00±0.00	136.80±13.83	138.10±16.41	129.10±26.91	1.00±0.00	136.80±13.83	136.80±13.83
dermatology	26.20±3.60	20.20±2.75	17.50±2.77	7.30±0.90	20.20±2.75	20.20±2.75	17.50±2.77	7.30±0.90	20.20±2.75	20.20±2.75
dresses-sales	9.00±12.50	202.20±24.28	163.60±20.51	1.00±0.00	196.00±17.29	202.20±24.28	163.60±20.51	1.00±0.00	196.00±17.29	196.00±17.29
electricity	1750.80±36.53	1178.30±54.50	1100.10±47.11	8.90±2.34	1178.30±54.50	1553.50±158.24	1482.10±120.74	2.00±0.00	1553.50±158.24	1553.50±158.24
eucalyptus	76.50±26.52	106.20±14.59	126.90±47.06	1.00±0.00	106.20±14.59	133.10±21.10	126.90±47.06	1.00±0.00	133.10±21.10	133.10±21.10
fruityfly	3.20±2.64	12.10±4.64	10.40±3.20	1.00±0.00	11.00±3.61	9.10±2.39	7.10±2.02	1.00±0.00	8.10±2.12	8.10±2.12

Continua na próxima página

Tabela 14 — Continuação da página anterior

Dataset	DT	MT(mean)						MT(range)					
		prel	lacc	nov4	sat	prel	lacc	nov4	sat	prel	lacc	nov4	sat
german-credit	83.10±22.12	73.10±12.81	64.30±9.56	3.60±1.74	73.10±12.81	89.00±15.76	59.40±15.22	3.60±1.74	89.00±15.76	59.40±15.22	3.60±1.74	89.00±15.76	59.40±15.22
grub-damage	47.00±6.66	23.00±2.45	21.00±0.00	1.00±0.00	23.00±2.45	23.00±2.45	21.00±0.00	1.00±0.00	23.00±2.45	21.00±0.00	1.00±0.00	23.00±2.45	21.00±0.00
haberman	17.40±6.17	20.70±5.48	17.20±5.11	1.00±0.00	19.50±5.50	19.30±10.26	19.20±10.07	1.00±0.00	16.70±7.98	19.20±10.07	1.00±0.00	16.70±7.98	19.20±10.07
hepatitis	9.40±2.20	2.80±0.98	2.10±1.14	1.00±0.00	2.70±0.90	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
horse-colic	4.60±1.02	1.26.00±0.00	126.00±0.00	2.00±0.00	126.00±0.00	126.00±0.00	126.00±0.00	2.00±0.00	126.00±0.00	126.00±0.00	2.00±0.00	126.00±0.00	126.00±0.00
hungarian-14-heart-disease	5.40±2.42	10.30±1.68	9.20±1.94	4.20±0.40	10.30±1.68	9.10±0.83	7.90±1.58	4.20±0.40	9.10±0.83	7.90±1.58	4.20±0.40	9.10±0.83	7.90±1.58
hypothyroid	14.60±0.66	18.40±4.05	14.70±2.72	2.00±0.00	17.90±4.13	24.40±9.32	20.90±5.63	2.00±0.00	24.80±8.83	20.90±5.63	2.00±0.00	24.80±8.83	20.90±5.63
ipd	33.60±12.82	14.10±2.12	12.00±1.95	1.00±0.00	14.10±2.12	4.20±3.34	2.00±2.05	1.00±0.00	4.20±3.34	2.00±2.05	1.00±0.00	4.20±3.34	2.00±2.05
irish	10.00±0.00	10.50±0.50	10.10±0.30	10.00±0.00	10.50±0.50	10.00±0.00	10.00±0.00	10.00±0.00	10.00±0.00	10.00±0.00	10.00±0.00	10.00±0.00	10.00±0.00
lymphography	17.30±2.49	18.10±4.44	9.40±3.61	4.00±0.00	18.10±4.44	18.10±4.44	9.40±3.61	4.00±0.00	18.10±4.44	9.40±3.61	4.00±0.00	18.10±4.44	9.40±3.61
musk	2.00±0.00	92.80±30.60	103.00±0.00	2.00±0.00	92.80±30.60	31.60±46.74	62.20±49.97	2.20±0.60	31.60±46.74	62.20±49.97	2.20±0.60	31.60±46.74	62.20±49.97
newton-hema	15.50±1.36	11.00±0.00	11.00±0.00	11.00±0.00	11.00±0.00	11.00±0.00	11.00±0.00	11.00±0.00	11.00±0.00	11.00±0.00	11.00±0.00	11.00±0.00	11.00±0.00
pasture-production	5.20±1.08	4.00±0.00	2.80±1.47	1.00±0.00	4.00±0.00	4.00±0.00	4.00±0.00	4.00±0.00	2.80±1.47	1.00±0.00	4.00±0.00	4.00±0.00	4.00±0.00
prnn-viruses	4.20±0.60	1.60±1.80	1.00±0.00	1.00±0.00	1.60±1.80	1.60±1.80	1.60±1.80	1.60±1.80	1.60±1.80	1.60±1.80	1.60±1.80	1.60±1.80	1.60±1.80
sick	28.60±4.13	12.40±1.62	11.90±1.64	2.00±0.00	12.30±1.62	18.70±1.35	17.80±0.87	2.80±0.98	18.70±1.35	17.80±0.87	2.80±0.98	18.70±1.35	17.80±0.87
speed-dating	3.00±0.00	61.30±102.08	36.60±78.26	3.00±0.00	61.70±101.93	28.30±78.95	31.60±78.25	3.00±0.00	30.90±78.43	31.60±78.25	3.00±0.00	30.90±78.43	31.60±78.25
squash-stored	5.40±2.15	13.60±10.29	1.00±0.00	1.00±0.00	13.60±10.29	13.60±10.29	13.60±10.29	1.00±0.00	13.60±10.29	1.00±0.00	1.00±0.00	13.60±10.29	1.00±0.00
squash-unstored	4.00±0.00	3.30±6.90	1.00±0.00	1.00±0.00	3.30±6.90	3.30±6.90	1.00±0.00	1.00±0.00	3.30±6.90	3.30±6.90	1.00±0.00	3.30±6.90	3.30±6.90
stress	5.80±1.17	2.40±1.20	2.00±0.00	1.70±0.46	2.40±1.20	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00	2.00±0.00
tae	28.60±2.84	4.40±0.49	4.40±0.49	1.10±0.30	4.80±0.75	4.00±0.00	3.30±0.46	1.10±0.30	4.00±0.00	3.30±0.46	1.10±0.30	4.00±0.00	3.30±0.46
veteran	14.40±2.69	5.40±0.49	5.10±0.30	3.70±0.90	5.40±0.49	4.80±0.40	4.80±0.40	3.70±0.90	4.80±0.40	4.80±0.40	3.70±0.90	4.80±0.40	4.80±0.40
visualizing-livestock	3.00±0.00	27.00±0.00	26.00±0.00	1.00±0.00	27.00±0.00	52.00±0.00	26.00±0.00	1.00±0.00	52.00±0.00	26.00±0.00	1.00±0.00	52.00±0.00	26.00±0.00
vowel	140.70±17.87	72.30±5.92	30.00±0.00	1.00±0.00	72.30±5.92	64.50±6.87	30.00±0.00	1.00±0.00	64.50±6.87	30.00±0.00	1.00±0.00	64.50±6.87	30.00±0.00
zoo	8.30±0.90	6.40±1.11	4.40±1.02	3.00±0.00	6.40±1.11	6.40±1.11	4.40±1.02	3.00±0.00	6.40±1.11	4.40±1.02	3.00±0.00	6.40±1.11	6.40±1.11
Média	55.42±231.19	55.69±158.75	48.72±147.70	3.02±3.50	55.49±158.62	59.91±206.60	53.11±196.56	2.90±3.36	59.72±206.49	53.11±196.56	2.90±3.36	59.72±206.49	53.11±196.56
Rank Médio	3.39	3.14	5.23	8.19	3.30	3.81	5.74	8.22	3.95	5.74	8.22	3.95	5.74

Tabela 15 – Média do número de folhas e o ranking médio: árvore de decisão e meta-árvore - Todos os 150 Datasets

	DT	MT(mean)				MT(range)			
		prel	lacc	nov4	sat	prel	lacc	nov4	sat
Média	100.29±411.24	110.87±705.66	53.46±146.32	3.38±6.25	110.71±705.64	110.74±709.60	53.51±163.73	3.41±6.24	110.57±709.58
Rank Médio	3.38	3.29	5.47	8.29	3.4	3.59	5.64	8.24	3.67