

Projeto de Data Mining – Diagnóstico de Doenças Dermatológicas

1. Introdução

Com o avanço das tecnologias de informação, a geração e recolha de dados tornaram-se constantes em praticamente todas as áreas. Neste cenário, a análise de grandes volumes de dados tornou-se essencial para extrair conhecimento útil, apoiar decisões e otimizar processos. A mineração de dados (Data Mining) surge como uma ferramenta fundamental para transformar dados brutos em informação relevante, com impacto real em contextos tão diversos como os negócios, a ciência ou a saúde.

Neste contexto, a área de Data Mining desempenha um papel crucial ao aplicar técnicas estatísticas e computacionais para identificar padrões relevantes em bases de dados complexas. A metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) surge como um padrão amplamente consolidado e utilizado na condução de projetos desta natureza. Esta metodologia estrutura o processo de mineração de dados em seis fases bem definidas:

- 1. Compreensão do Negócio (Business Understanding):** definição do problema do ponto de vista do domínio da aplicação;
- 2. Compreensão dos Dados (Data Understanding):** exploração e caracterização do conjunto de dados;
- 3. Preparação dos Dados (Data Preparation):** limpeza e transformação dos dados para torná-los adequados à modelação;
- 4. Modelação (Modeling):** aplicar algoritmos de aprendizagem automática;
- 5. Avaliação (Evaluation):** verificar se os modelos obtidos cumprem os objetivos definidos;
- 6. Implementação (Deployment):** entregar os resultados de forma útil para o utilizador final.

2. Business Understanding

Incidindo agora no âmbito deste projeto, será utilizado o dataset *Dermatology*. Este conjunto de dados foi construído com o objetivo de apoiar o diagnóstico diferencial de doenças dermatológicas, em específico do grupo *erythemato-squamous*, incluindo diferentes patologias como *psoriasis*, *seboreic dermatitis*, *lichen planus*, *pityriasis rosea*, *cronic dermatitis* e *pityriasis rubra pilaris*.

A realização de um diagnóstico destas doenças representa um verdadeiro desafio clínico, uma vez que partilham sintomas clínicos e características histopatológicas muito semelhantes. Esta semelhança dificulta a distinção precisa entre as diferentes patologias, levando, muitas vezes, à necessidade de exames como biópsias.

No entanto, mesmo com estes exames, nem sempre é possível obter um diagnóstico inequívoco, devido à sobreposição de padrões microscópicos. Além disso, os sintomas podem variar ao longo do tempo, o que aumenta ainda mais a complexidade do processo diagnóstico.

Dada esta realidade, é evidente a necessidade de desenvolver ferramentas auxiliares que apoiem os profissionais de saúde no processo diagnóstico, de forma mais rápida, precisa e não invasiva. A mineração de dados apresenta-se como uma abordagem promissora neste sentido. Utilizando técnicas de aprendizagem automática (*machine learning*), é possível identificar padrões relevantes nos dados clínicos e histopatológicos que permitam prever, com elevada precisão, o tipo de doença presente em cada paciente.

3. Data Understanding

3.1 Descrição do Dataset

O dataset utilizado, denominado *Dermatology*, é composto por **366 instâncias (registos)** correspondentes a diferentes pacientes e por **34 atributos descritivos**, aos quais se junta um **atributo de classe (label)** que representa o diagnóstico final da doença dermatológica.

Os atributos dividem-se em três grandes grupos:

- **Atributos clínicos** (1 a 11)
- **Atributos histopatológicos** (12 a 33)
- **Atributo contínuo: idade** (atributo 34)

3.2 Atributo de Classe (Label)

Valor da Classe	Doença Dermatológica
1	Psoriasis
2	Seboreic Dermatitis
3	Lichen Planus
4	Pityriasis Rosea
5	Cronic Dermatitis
6	Pityriasis Rubra Pilaris

3.3 Caracterização dos Atributos

- Atributos 1 a 33: valores entre 0 e 3 (ordinais)
- Atributo 11 (histórico familiar): binário (0 ou 1)
- Atributo 34 (idade): contínuo, com 8 valores ausentes representados por `?` e um valor a `0` estes problema serão resolvidos no Passo a Segir Data Preparation

4. Data Preparation

4.1 Tratamento de Valores em Falta ou Nulos

- Inicialmente e uma vez que temos apenas um Valor nulo `0` decidimos alterar este manualmente no dataset alterando o para o valor medio nas idades no caso Aproximadamente 40
- Para a Substituição da idade ausente `?` utilizamos os operadores `Declare Missing Values` e `Replace Missing Values` estes vao respetivamente identificar e alterar todos os valores `?` pela **média** das idades no caso Aproximadamente 40

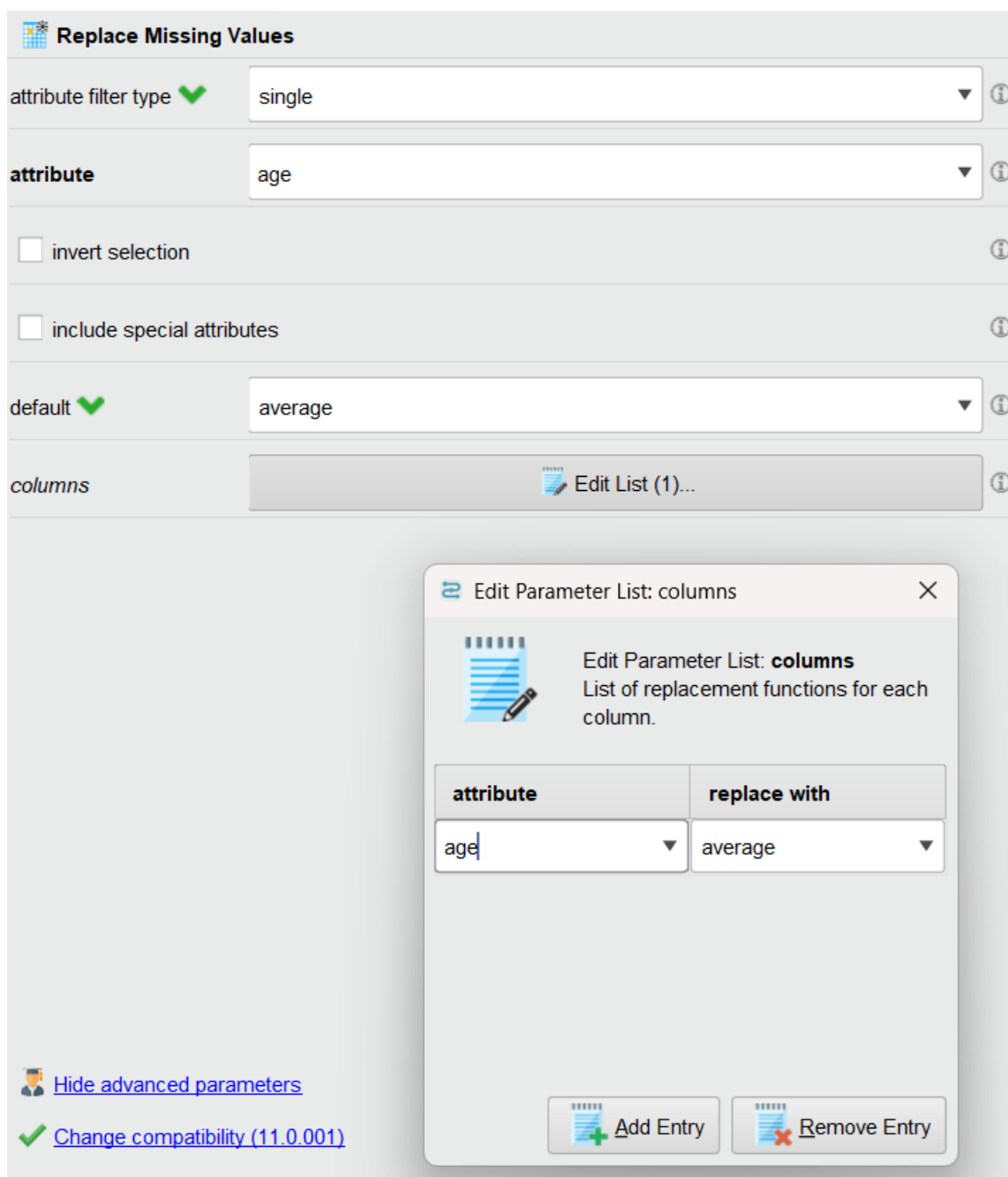
Declare Missing Values - image

The image shows a configuration window titled "Declare Missing Value". It contains several settings for identifying missing values in a dataset:

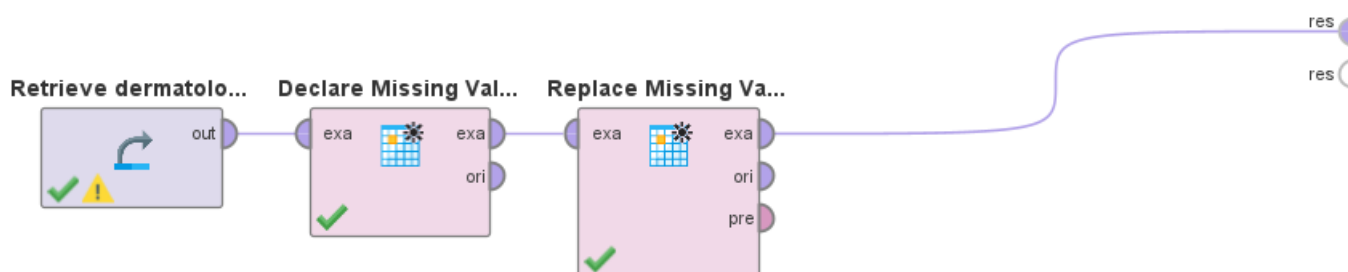
- attribute filter type**: set to "single".
- attribute**: set to "age".
- invert selection**: unchecked checkbox.
- include special attributes**: unchecked checkbox.
- mode**: set to "nominal".
- nominal value**: set to "?".

Each setting has a green checkmark icon indicating it is correctly configured, and an information icon (i) is available for each field.

Replace Missing Values - Image



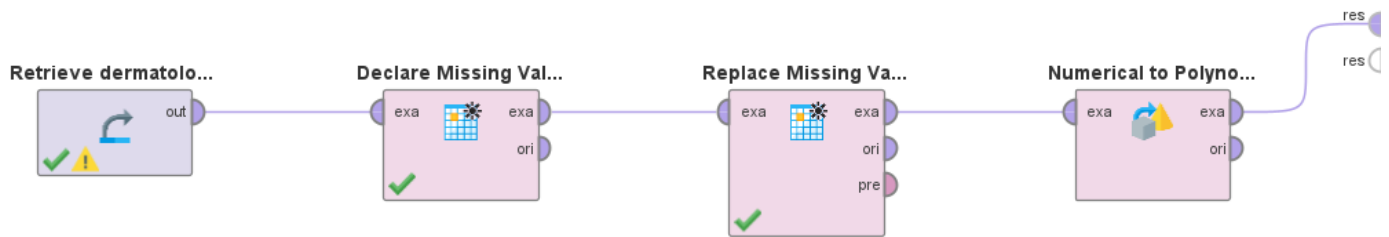
Design - Image



4.2 Conversão de Tipos de Dados

De seguida, como o atributo `class label` se encontrava no formato `integer` ou seja um (valor inteiro), foi necessário convertê-lo para o formato `nominal` ou `polynomial`. Esta conversão era essencial para que, mais tarde, fosse possível aplicar a `cross validation`, uma vez que este obriga a que variável de destino esteja num formato categórico. Para isso, recorremos ao operador `Numerical to Polynomial`.

Design - Image

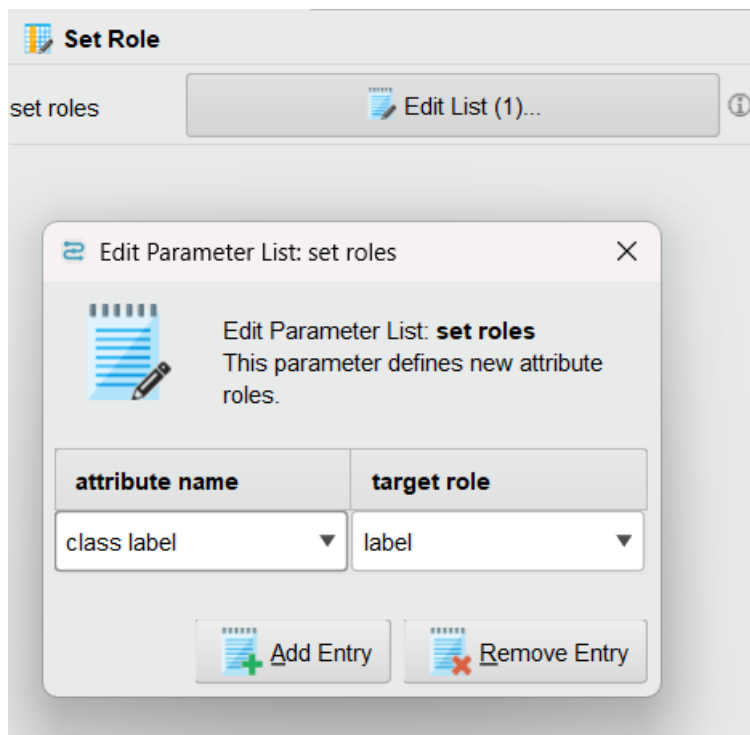
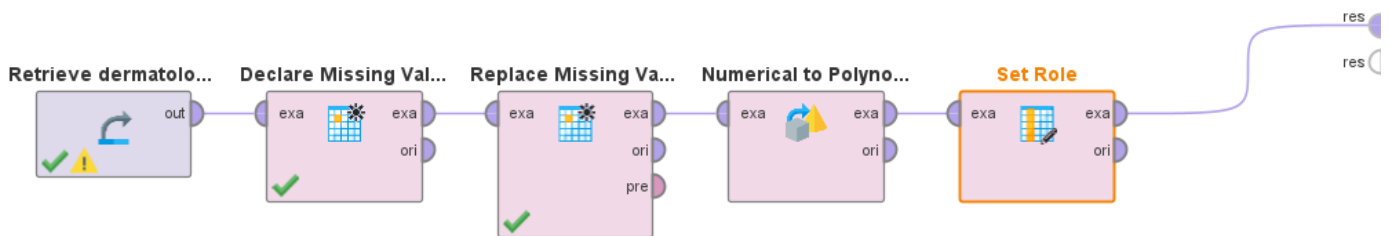


5. Modeling

Entrando agora na fase da modelação e aplicação da técnica de classificação iremos utilizar um Algoritmo C4.5 (Decision Tree) e no Rapid Miner este será feito com recurso a diferentes critérios de divisão de dados (gain_ratio, information_gain, gini_index e accuracy) e utilizando os restantes parametros como default.

5.1 Alteração da Role do Atributo "LABEL"

De forma a iniciar a fase de modelação e antes de Iniciar o Desenvolvimento da Árvore de decisão propriamente dita foi necessario definir um `attribute` no caso `class label` como `role label` através do operador `Set Role`



Este após alterado fica marcado pela cor verde na coluna do atributo

Row No.	class label
1	2
2	1
3	3
4	1
5	3
6	2
7	5
8	3
9	4
10	4
11	1
12	2
13	2
14	1
15	3
16	4
17	2


5.2 Algoritmo C4.5 (Decision Tree)


O algoritmo C4.5 consiste num algoritmo de decision tree (árvore de decisão), desenvolvido por Ross Quinlan, amplamente utilizado em tarefas de classificação. A construção da árvore é realizada de forma recursiva, selecionando a cada divisão o atributo mais informativo com base em critérios como gain ratio, information gain, gini index ou accuracy. Uma das principais vantagens do C4.5 é a sua capacidade de lidar com atributos contínuos e discretos, bem como com valores em falta. Além disso, o algoritmo aplica técnicas de pruning (poda) para reduzir o sobreajuste e melhorar a capacidade de generalização do modelo. O resultado final é uma árvore de decisão robusta, precisa e de fácil interpretação, adequada para a classificação de novos dados.

5.3 Critérios de Divisão de Dados

Pasando agora aos critrios de divisão iremos explicar no que cada um consiste, demostar a sua formula e bem como analisar os seus Resultados no ambito do trabalho

5.3.1 Gain_Ratio

criterion


gain ratio


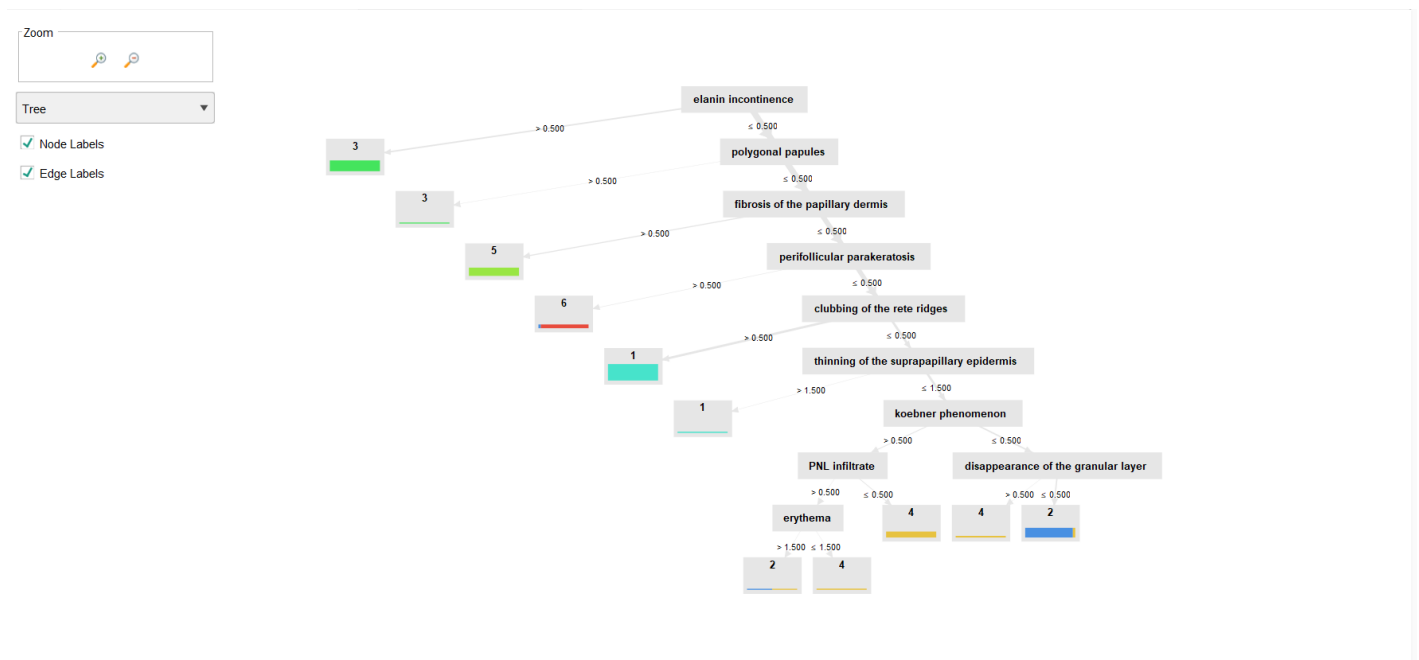
Definição:
O Gain Ratio é uma modificação do critério de Information Gain, que visa penalizar atributos com muitos valores possíveis. Isso ajuda a evitar que o modelo favoreça atributos com muitos valores, mas que podem não ser os melhores para a classificação. Ele é calculado dividindo o ganho de informação pelo Split Info (informação sobre a divisão do atributo).

Fórmula:

Gain_Ratio = \frac{Information_Gain}{Split_Info}

AMBITO DO PROBLEMA - INTERPERTAÇÃO DE RESULTADOS

TREE IMAGE OUTPUT



Percurso:

- Raiz: **melanin incontinence**
Se o valor for menor ou igual a 0.5, avança na análise.
- Passa por **polygonal papules** e **fibrosis of the papillary dermis** – atributos histopatológicos e clínicos relevantes.
- Segue por:
 - **perifollicular parakeratosis**
 - **clubbing of the rete ridges**
 - **thinning of the suprapapillary epidermis**
 - **koebner phenomenon**
- Termina com:
 - **PNL infiltrate**
 - **erythema**
 - **disappearance of the granular layer**


Conclusões:


- As classes terminais (2, 4, 6...) representam diferentes diagnósticos dermatológicos.
 - Classe 2 (**azul**) → pode indicar uma forma leve ou intermediária de dermatose.
 - Classe 6 (**vermelho**) → possível condição inflamatória mais grave.
- A árvore evidencia como a combinação de alterações histológicas com fenómenos clínicos (como o fenómeno de Koebner) permite distinguir eficazmente entre diferentes diagnósticos.

TREE OUTPUT CODE

```
Tree
elanin incontinence > 0.500: 3 {2=0, 1=0, 3=70, 5=0, 4=0, 6=0}
elanin incontinence ≤ 0.500
|   polygonal papules > 0.500: 3 {2=0, 1=0, 3=2, 5=0, 4=0, 6=0}
|   polygonal papules ≤ 0.500
|   |   fibrosis of the papillary dermis > 0.500: 5 {2=0, 1=0, 3=0, 5=52, 4=0, 6=0}
|   |   fibrosis of the papillary dermis ≤ 0.500
|   |   |   perifollicular parakeratosis > 0.500: 6 {2=1, 1=0, 3=0, 5=0, 4=0, 6=20}
|   |   |   perifollicular parakeratosis ≤ 0.500
|   |   |   |   clubbing of the rete ridges > 0.500: 1 {2=0, 1=109, 3=0, 5=0, 4=0, 6=0}
|   |   |   |   clubbing of the rete ridges ≤ 0.500
|   |   |   |   |   thinning of the suprapapillary epidermis > 1.500: 1 {2=0, 1=3, 3=0, 5=0, 4=0, 6=0}
|   |   |   |   |   thinning of the suprapapillary epidermis ≤ 1.500
|   |   |   |   |   |   koebner phenomenon > 0.500
|   |   |   |   |   |   |   PNL infiltrate > 0.500
|   |   |   |   |   |   |   |   erythema > 1.500: 2 {2=1, 1=0, 3=0, 5=0, 4=1, 6=0}
|   |   |   |   |   |   |   |   erythema ≤ 1.500: 4 {2=0, 1=0, 3=0, 5=0, 4=3, 6=0}
|   |   |   |   |   |   |   |   PNL infiltrate ≤ 0.500: 4 {2=0, 1=0, 3=0, 5=0, 4=36, 6=0}
|   |   |   |   |   |   |   |   koebner phenomenon ≤ 0.500
|   |   |   |   |   |   |   |   |   disappearance of the granular layer > 0.500: 4 {2=0, 1=0, 3=0, 5=0, 4=6, 6=0}
|   |   |   |   |   |   |   |   |   disappearance of the granular layer ≤ 0.500: 2 {2=59, 1=0, 3=0, 5=0, 4=3, 6=0}
```

5.3.2 Information_Gain

criterion 

information gain 

**** Definição:****

O Information Gain (IG) é uma métrica que calcula a redução na incerteza (ou entropia) do sistema após uma divisão. Ele é baseado na teoria da informação e mede o quanto um atributo ajuda a reduzir a incerteza sobre a classe ou variável alvo. Quanto maior o ganho de informação, mais eficaz é o atributo na divisão dos dados.

Fórmula:

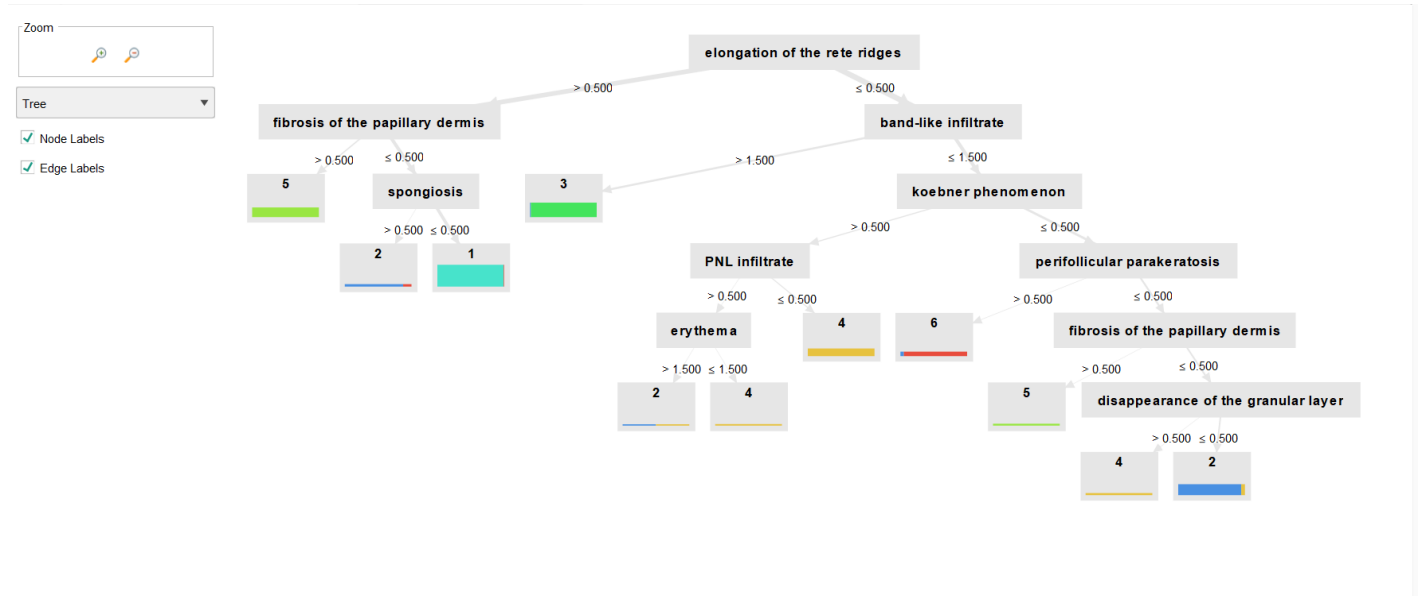
$$\text{Information_Gain} = \text{Entropy}(D) - \sum \left(\frac{|D_v|}{|D|} \times \text{Entropy}(D_v) \right)$$

Onde:

- D é o conjunto de dados.
- D_v são os subconjuntos criados pela divisão.
- Entropy é uma medida da impureza dos dados.

AMBITO DO PROBLEMA - INTERPERTAÇÃO DE RESULTADOS

***TREE IMAGE OUTPUT**



Percurso:

- Raiz: **fibrosis of the papillary dermis**
 - Se presente, analisa-se **spongiosis** (separação celular na epiderme), levando às classes 2 ou 1 conforme a intensidade.
 - Caso ausente, segue para:
 - **elongation of the rete ridges**
 - **band-like infiltrate**
 - **koebner phenomenon**
- Continua por:
 - **PNL infiltrate + erythema**
 - **perifollicular parakeratosis, fibrosis** (novamente), e **disappearance of the granular layer**

Conclusões:

- Lesões estruturais profundas (como a fibrose) revelam-se essenciais para o diagnóstico.
- A árvore **reutiliza atributos em diferentes caminhos**, sugerindo que os mesmos sinais podem indicar doenças distintas, consoante o contexto.
 - **Classe 4 (amarelo)** é prevalente.
 - **Classe 6** surge com infiltrado PNL mas **sem eritema**, podendo apontar para uma doença menos vascular/inflamatória.


TREE OUTPUT CODE

```

Tree
elongation of the rete ridges > 0.500
|  fibrosis of the papillary dermis > 0.500: 5 {2=0, 1=0, 3=0, 5=47, 4=0, 6=0}
|  fibrosis of the papillary dermis ≤ 0.500
|  |  spongiosis > 0.500: 2 {2=7, 1=0, 3=0, 5=0, 4=0, 6=1}
|  |  spongiosis ≤ 0.500: 1 {2=0, 1=112, 3=0, 5=0, 4=0, 6=1}
elongation of the rete ridges ≤ 0.500
|  band-like infiltrate > 1.500: 3 {2=1, 1=0, 3=72, 5=0, 4=0, 6=0}
|  band-like infiltrate ≤ 1.500
|  |  koebner phenomenon > 0.500
|  |  |  PNL infiltrate > 0.500
|  |  |  |  erythema > 1.500: 2 {2=1, 1=0, 3=0, 5=0, 4=1, 6=0}
|  |  |  |  erythema ≤ 1.500: 4 {2=0, 1=0, 3=0, 5=0, 4=3, 6=0}
|  |  |  |  PNL infiltrate ≤ 0.500: 4 {2=0, 1=0, 3=0, 5=0, 4=36, 6=0}
|  |  |  koebner phenomenon ≤ 0.500
|  |  |  |  perifollicular parakeratosis > 0.500: 6 {2=1, 1=0, 3=0, 5=0, 4=0, 6=18}
|  |  |  |  perifollicular parakeratosis ≤ 0.500
|  |  |  |  |  fibrosis of the papillary dermis > 0.500: 5 {2=0, 1=0, 3=0, 5=5, 4=0, 6=0}
|  |  |  |  |  fibrosis of the papillary dermis ≤ 0.500
|  |  |  |  |  |  disappearance of the granular layer > 0.500: 4 {2=0, 1=0, 3=0, 5=0, 4=6, 6=0}
|  |  |  |  |  |  disappearance of the granular layer ≤ 0.500: 2 {2=51, 1=0, 3=0, 5=0, 4=3, 6=0}

```

5.3.3 Gini_Index

criterion 

gini index ▼

Definição: O Gini Index é uma métrica de impureza que mede a desigualdade nas classes dentro de um nó. Ele é utilizado principalmente em árvores de decisão como o algoritmo CART (Classification and Regression Tree). O índice de Gini calcula a probabilidade de uma amostra ser classificada incorretamente se ela fosse rotulada aleatoriamente.

Fórmula:

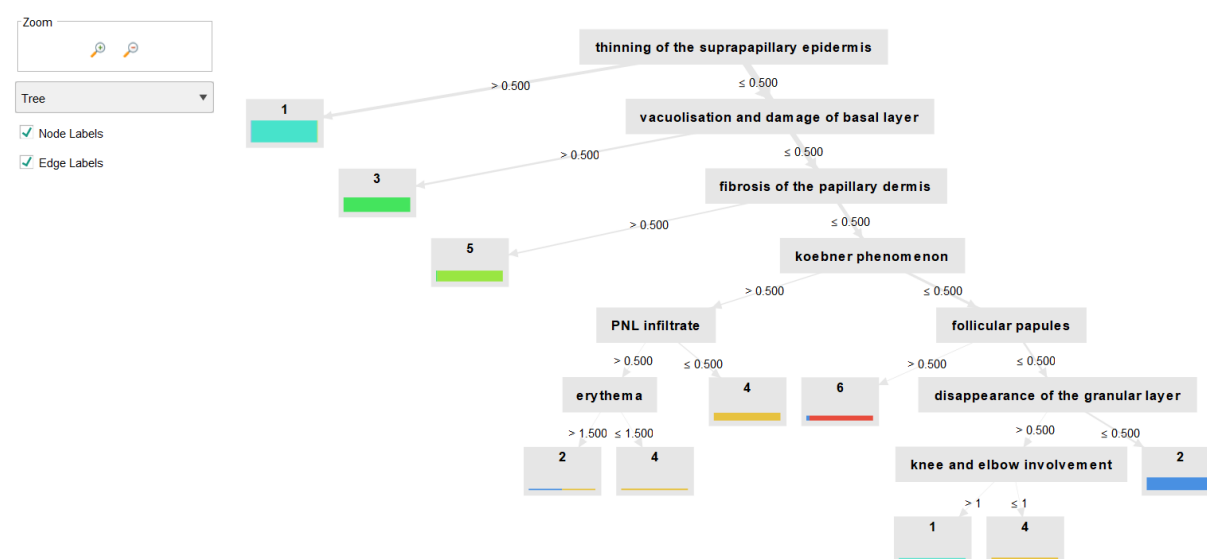
$$\text{Gini}(D) = 1 - \sum_{i=1}^m p_i^2$$

Onde:

- p_i é a probabilidade de um item ser classificado na classe i .
- m é o número de classes.

AMBITO DO PROBLEMA - INTERPERTAÇÃO DE RESULTADOS

TREE IMAGE OUTPUT



Percurso:

- Raiz: **thinning of the suprapapillary epidermis** (atrofia epitelial)
- Em seguida:
 - **vacuolisation and damage of basal layer**
 - **fibrosis of the papillary dermis**
 - **koebner phenomenon**
- Divide-se em dois ramos:
 - **PNL infiltrate + erythema**
 - **follicular papules + disappearance of the granular layer + knee and elbow involvement**

Conclusões:

- A árvore contrapõe **resposta inflamatória generalizada** (infiltrado + eritema) com **sinais cutâneos localizados** (joelhos/cotovelos).
 - **Classe 1** → poucos sinais → forma benigna.
 - **Classe 4** → infiltrado + eritema → doença inflamatória típica.
 - **Classe 2 (azul)** → alterações estruturais com poucos sinais clínicos → diagnóstico diferencial possível.


TREE OUTPUT CODE


```

Tree
thinning of the suprapapillary epidermis > 0.500: 1 {2=1, 1=108, 3=0, 5=1, 4=0, 6=0}
thinning of the suprapapillary epidermis ≤ 0.500
| vacuolisation and damage of basal layer > 0.500: 3 {2=0, 1=0, 3=71, 5=0, 4=0, 6=0}
| vacuolisation and damage of basal layer ≤ 0.500
| | fibrosis of the papillary dermis > 0.500: 5 {2=0, 1=0, 3=1, 5=51, 4=0, 6=0}
| | fibrosis of the papillary dermis ≤ 0.500
| | | koebner phenomenon > 0.500
| | | | PNL infiltrate > 0.500
| | | | | erythema > 1.500: 2 {2=1, 1=0, 3=0, 5=0, 4=1, 6=0}
| | | | | erythema ≤ 1.500: 4 {2=0, 1=0, 3=0, 5=0, 4=3, 6=0}
| | | | PNL infiltrate ≤ 0.500: 4 {2=0, 1=0, 3=0, 5=0, 4=36, 6=0}
| | | koebner phenomenon ≤ 0.500
| | | | follicular papules > 0.500: 6 {2=1, 1=0, 3=0, 5=0, 4=0, 6=20}
| | | | follicular papules ≤ 0.500
| | | | | disappearance of the granular layer > 0.500
| | | | | | knee and elbow involvement > 1: 1 {2=0, 1=3, 3=0, 5=0, 4=0, 6=0}
| | | | | | knee and elbow involvement ≤ 1: 4 {2=0, 1=0, 3=0, 5=0, 4=6, 6=0}
| | | | | disappearance of the granular layer ≤ 0.500: 2 {2=58, 1=1, 3=0, 5=0, 4=3, 6=0}

```

5.3.4 Accuracy

criterion


accuracy


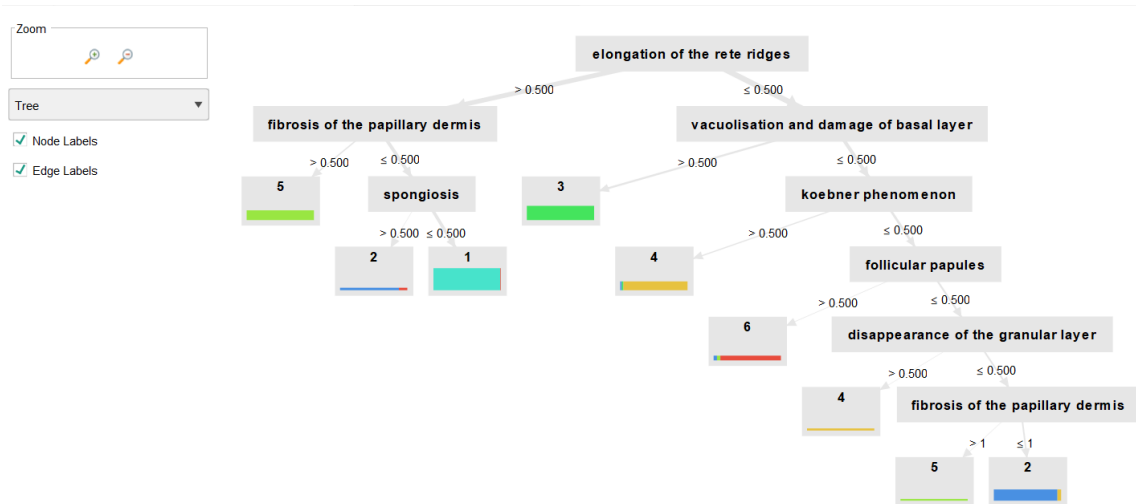
Definição: A Accuracy é a taxa de acerto de um modelo, ou seja, a proporção de previsões corretas em relação ao total de previsões feitas. Ela mede a capacidade do modelo de prever corretamente as instâncias de dados.

Fórmula:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}}$$

AMBITO DO PROBLEMA - INTERPERTAÇÃO DE RESULTADOS

TREE IMAGE OUTPUT



Percurso:

- Inicia-se com **fibrosis of the papillary dermis** (como na Árvore 2).
- Depois segue:
 - **elongation of the rete ridges**
 - **vacuolisation and damage of basal layer**
 - **koebner phenomenon**
 - **follicular papules**
- Prossegue com:
 - **disappearance of the granular layer**
 - repetição de **fibrosis**
 - e finalmente, análise de **atributo quantitativo** ≤ 1

Conclusões:

- Semelhante à Árvore 3, mas com **maior foco em alterações estruturais** do que em fenómenos clínicos.
 - **Classe 5** aparece em ramos com **múltiplas alterações histológicas** → diagnóstico mais avançado.
 - **Classe 2** surge em cenários com **poucos marcadores clínicos**, sugerindo formas menos visíveis ou subclínicas da doença.

TREE OUTPUT CODE

```

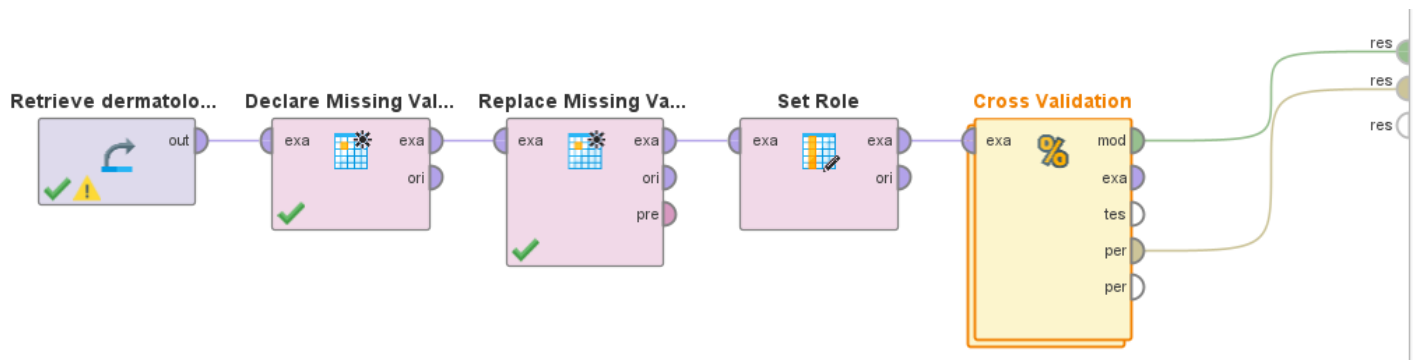
Tree
elongation of the rete ridges > 0.500
|  fibrosis of the papillary dermis > 0.500: 5 {2=0, 1=0, 3=0, 5=47, 4=0, 6=0}
|  fibrosis of the papillary dermis ≤ 0.500
|  |  spongiosis > 0.500: 2 {2=7, 1=0, 3=0, 5=0, 4=0, 6=1}
|  |  spongiosis ≤ 0.500: 1 {2=0, 1=112, 3=0, 5=0, 4=0, 6=1}
elongation of the rete ridges ≤ 0.500
|  vacuolisation and damage of basal layer > 0.500: 3 {2=0, 1=0, 3=71, 5=0, 4=0, 6=0}
|  vacuolisation and damage of basal layer ≤ 0.500
|  |  koebner phenomenon > 0.500: 4 {2=1, 1=0, 3=1, 5=0, 4=40, 6=0}
|  |  koebner phenomenon ≤ 0.500
|  |  |  follicular papules > 0.500: 6 {2=1, 1=0, 3=0, 5=1, 4=0, 6=18}
|  |  |  follicular papules ≤ 0.500
|  |  |  |  disappearance of the granular layer > 0.500: 4 {2=0, 1=0, 3=0, 5=0, 4=6, 6=0}
|  |  |  |  disappearance of the granular layer ≤ 0.500
|  |  |  |  |  fibrosis of the papillary dermis > 1: 5 {2=0, 1=0, 3=0, 5=4, 4=0, 6=0}
|  |  |  |  |  fibrosis of the papillary dermis ≤ 1: 2 {2=52, 1=0, 3=0, 5=0, 4=3, 6=0}

```

6. Evaluation

Por ultimo na Fase de Avaliaçao fase vamos analisar as performances obtidas de cada uma das árvores obtidas. Nesta Vamos Utilizar o Operador `Cross Validation` com k folds 10 para a validação) este divide se em duas partes **Training** Onde é utilizado o operador do Algoritmo C4.5 `Decision Tree` e a escolha dos criterios de divisão de dados (`gain_ratio`, `information_gain`, `gini_index` e `accuracy`) e uma parte de **Testing** onde serão utilizados os operadores de `Apply Model` e `Performance Classification` utilizando os criterios de `Accuracy Classification error` e `Root Mean Squared Error`

6.1 Cross Validation



% Cross Validation

☐ split on batch attribute

☐ leave one out

number of folds ☒ 10

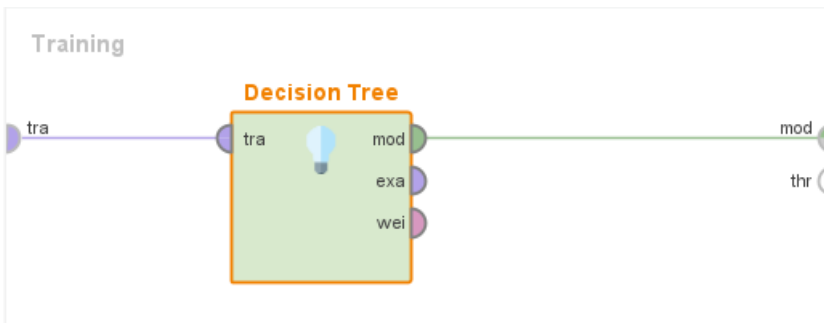
sampling type ☒ automatic

☐ use local random seed ☒

☒ enable parallel execution

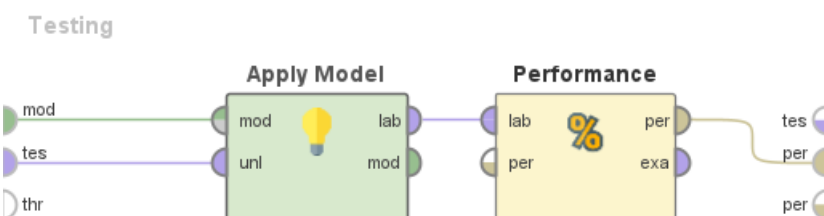
6.2 Training

Na Parte do Training Como dito Anteriormente Apenas temos de inserir o Operador **Decision Tree** e Selecionar o metodo de Divisão de Dados Pretendido



6.3 Testing

Por sua vez na Parte do Training basta selecionar os criterio de Accuracy **Classification error** e **Root Mean Squared Error**



Parameters

Performance (Performance (Classification))

main criterion

first

☒ accuracy

☒ classification error

☐ kappa

☐ weighted mean recall

☐ weighted mean precision

☐ spearman rho

☐ kendall tau

☐ absolute error

☐ relative error

☐ relative error lenient

☐ relative error strict

☐ normalized absolute error

☒ root mean squared error

6.4 Results

6.4.1 Gain_Ratio

Table View

Plot View

accuracy: 85.20% +/- 18.92% (micro average: 85.25%)

	true 2	true 1	true 3	true 5	true 4	true 6	class precision
pred. 2	41	1	0	0	3	0	91.11%
pred. 1	17	111	3	5	15	3	72.08%
pred. 3	1	0	65	0	0	0	98.48%
pred. 5	0	0	1	47	0	0	97.92%
pred. 4	1	0	3	0	31	0	88.57%
pred. 6	1	0	0	0	0	17	94.44%
class recall	67.21%	99.11%	90.28%	90.38%	63.27%	85.00%	

classification_error: 14.80% +/- 18.92% (micro average: 14.75%)

	true 2	true 1	true 3	true 5	true 4	true 6	class precision
pred. 2	41	1	0	0	3	0	91.11%
pred. 1	17	111	3	5	15	3	72.08%
pred. 3	1	0	65	0	0	0	98.48%
pred. 5	0	0	1	47	0	0	97.92%
pred. 4	1	0	3	0	31	0	88.57%
pred. 6	1	0	0	0	0	17	94.44%
class recall	67.21%	99.11%	90.28%	90.38%	63.27%	85.00%	

root_mean_squared_error

root_mean_squared_error: 0.294 +/- 0.237 (micro average: 0.370 +/- 0.000)

```
PerformanceVector
PerformanceVector:
accuracy: 85.20% +/- 18.92% (micro average: 85.25%)
ConfusionMatrix:
True: 2 1 3 5 4 6
2: 41 1 0 0 3 0
1: 17 111 3 5 15 3
3: 1 0 65 0 0 0
5: 0 0 1 47 0 0
4: 1 0 3 0 31 0
6: 1 0 0 0 0 17
classification_error: 14.80% +/- 18.92% (micro average: 14.75%)
ConfusionMatrix:
True: 2 1 3 5 4 6
2: 41 1 0 0 3 0
1: 17 111 3 5 15 3
3: 1 0 65 0 0 0
5: 0 0 1 47 0 0
4: 1 0 3 0 31 0
6: 1 0 0 0 0 17
root_mean_squared_error: 0.294 +/- 0.237 (micro average: 0.370 +/- 0.000)
```

6.4.2 Information_Gain

accuracy: 95.89% +/- 3.50% (micro average: 95.90%)

	true 2	true 1	true 3	true 5	true 4	true 6	class precision
pred. 2	56	0	0	0	3	3	90.32%
pred. 1	2	112	0	0	0	1	97.39%
pred. 3	1	0	70	0	0	0	98.59%
pred. 5	0	0	0	51	0	0	100.00%
pred. 4	1	0	2	0	46	0	93.88%
pred. 6	1	0	0	1	0	16	88.89%
class recall	91.80%	100.00%	97.22%	98.08%	93.88%	80.00%	

classification_error: 4.11% +/- 3.50% (micro average: 4.10%)

	true 2	true 1	true 3	true 5	true 4	true 6	class precision
pred. 2	56	0	0	0	3	3	90.32%
pred. 1	2	112	0	0	0	1	97.39%
pred. 3	1	0	70	0	0	0	98.59%
pred. 5	0	0	0	51	0	0	100.00%
pred. 4	1	0	2	0	46	0	93.88%
pred. 6	1	0	0	1	0	16	88.89%
class recall	91.80%	100.00%	97.22%	98.08%	93.88%	80.00%	

root_mean_squared_error

root_mean_squared_error: 0.182 +/- 0.095 (micro average: 0.202 +/- 0.000)

```
PerformanceVector
PerformanceVector:
accuracy: 95.89% +/- 3.50% (micro average: 95.90%)
ConfusionMatrix:
True: 2 1 3 5 4 6
2: 56 0 0 0 3 3
1: 2 112 0 0 0 1
3: 1 0 70 0 0 0
5: 0 0 0 51 0 0
4: 1 0 2 0 46 0
6: 1 0 0 1 0 16
classification_error: 4.11% +/- 3.50% (micro average: 4.10%)
ConfusionMatrix:
True: 2 1 3 5 4 6
2: 56 0 0 0 3 3
1: 2 112 0 0 0 1
3: 1 0 70 0 0 0
5: 0 0 0 51 0 0
4: 1 0 2 0 46 0
6: 1 0 0 1 0 16
root_mean_squared_error: 0.182 +/- 0.095 (micro average: 0.202 +/- 0.000)
```

6.4.3 Gini_Index

accuracy: 95.08% +/- 3.59% (micro average: 95.08%)

	true 2	true 1	true 3	true 5	true 4	true 6	class precision
pred. 2	55	0	0	0	3	0	94.83%
pred. 1	1	110	0	1	0	1	97.35%
pred. 3	1	0	67	0	0	0	98.53%
pred. 5	0	0	1	51	0	0	98.08%
pred. 4	2	2	4	0	46	0	85.19%
pred. 6	2	0	0	0	0	19	90.48%
class recall	90.16%	98.21%	93.06%	98.08%	93.88%	95.00%	

☒ Table View ☐ Plot View

classification_error: 4.92% +/- 3.59% (micro average: 4.92%)

	true 2	true 1	true 3	true 5	true 4	true 6	class precision
pred. 2	55	0	0	0	3	0	94.83%
pred. 1	1	110	0	1	0	1	97.35%
pred. 3	1	0	67	0	0	0	98.53%
pred. 5	0	0	1	51	0	0	98.08%
pred. 4	2	2	4	0	46	0	85.19%
pred. 6	2	0	0	0	0	19	90.48%
class recall	90.16%	98.21%	93.06%	98.08%	93.88%	95.00%	

root_mean_squared_error

root_mean_squared_error: 0.203 +/- 0.087 (micro average: 0.219 +/- 0.000)

```
PerformanceVector
PerformanceVector:
accuracy: 95.08% +/- 3.59% (micro average: 95.08%)
ConfusionMatrix:
True: 2 1 3 5 4 6
2: 55 0 0 0 3 0
1: 1 110 0 1 0 1
3: 1 0 67 0 0 0
5: 0 0 1 51 0 0
4: 2 2 4 0 46 0
6: 2 0 0 0 0 19
classification_error: 4.92% +/- 3.59% (micro average: 4.92%)
ConfusionMatrix:
True: 2 1 3 5 4 6
2: 55 0 0 0 3 0
1: 1 110 0 1 0 1
3: 1 0 67 0 0 0
5: 0 0 1 51 0 0
4: 2 2 4 0 46 0
6: 2 0 0 0 0 19
root_mean_squared_error: 0.203 +/- 0.087 (micro average: 0.219 +/- 0.000)
```

6.4.4 Accuracy - Results

☒ Table View ☐ Plot View

accuracy: 95.06% +/- 3.85% (micro average: 95.08%)

	true 2	true 1	true 3	true 5	true 4	true 6
pred. 2	55	0	0	0	3	1
pred. 1	2	112	0	0	0	1
pred. 3	1	0	66	0	0	0
pred. 5	0	0	0	51	0	0
pred. 4	1	0	6	0	46	0
pred. 6	2	0	0	1	0	18
class recall	90.16%	100.00%	91.67%	98.08%	93.88%	90.00%

classification_error: 4.94% +/- 3.85% (micro average: 4.92%)

	true 2	true 1	true 3	true 5	true 4	true 6
pred. 2	55	0	0	0	3	1
pred. 1	2	112	0	0	0	1
pred. 3	1	0	66	0	0	0
pred. 5	0	0	0	51	0	0
pred. 4	1	0	6	0	46	0
pred. 6	2	0	0	1	0	18
class recall	90.16%	100.00%	91.67%	98.08%	93.88%	90.00%

root_mean_squared_error

root_mean_squared_error: 0.207 +/- 0.088 (micro average: 0.222 +/- 0.000)

```

PerformanceVector
PerformanceVector:
accuracy: 95.06% +/- 3.85% (micro average: 95.08%)
ConfusionMatrix:
True: 2 1 3 5 4 6
2: 55 0 0 0 3 1
1: 2 112 0 0 0 1
3: 1 0 66 0 0 0
5: 0 0 0 51 0 0
4: 1 0 6 0 46 0
6: 2 0 0 1 0 18
classification_error: 4.94% +/- 3.85% (micro average: 4.92%)
ConfusionMatrix:
True: 2 1 3 5 4 6
2: 55 0 0 0 3 1
1: 2 112 0 0 0 1
3: 1 0 66 0 0 0
5: 0 0 0 51 0 0
4: 1 0 6 0 46 0
6: 2 0 0 1 0 18
root_mean_squared_error: 0.207 +/- 0.088 (micro average: 0.222 +/- 0.000)

```

6.5 COMPARAÇÃO DE RESULTADOS

Critério	Accuracy (%)	Classification Error (%)	RMSE
gain_ratio	85.20 ± 18.92	14.80 ± 18.92	0.294
information_gain	95.89 ± 3.50	4.11 ± 3.50	0.182
gini_index	95.08 ± 3.59	4.92 ± 3.59	0.203
accuracy (critério)	95.06 ± 3.85	4.94 ± 3.85	0.207

7. ANALISE E CONCLUSÕES FINAIS

Após a aplicação integral da metodologia CRISP-DM — desde a compreensão do problema e dos dados, passando pela preparação dos dados, até à fase de modelação com o algoritmo C4.5 — avalíamos quatro critérios de divisão para a construção da árvore de decisão sobre o dataset Dermatology: gain ratio, information gain, gini index e accuracy.

Os resultados obtidos mostraram que o critério information gain proporcionou o melhor desempenho global, atingindo uma accuracy de 95,89%, o menor erro de classificação (4,11%) e o menor RMSE (0,182). Estes valores refletem uma elevada capacidade de generalização do modelo e uma separação eficaz entre as diferentes classes dermatológicas.

Com base nesta análise, concluímos que o melhor modelo foi gerado com o critério information gain, sendo este o mais adequado para a tarefa de classificação em causa. Verificámos que a escolha apropriada do critério de divisão foi determinante para maximizar o desempenho preditivo da árvore de decisão, reforçando o seu potencial de aplicação em contextos médicos, nomeadamente no apoio ao diagnóstico dermatológico.