PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Instituto de Ciências Exatas e de Informática

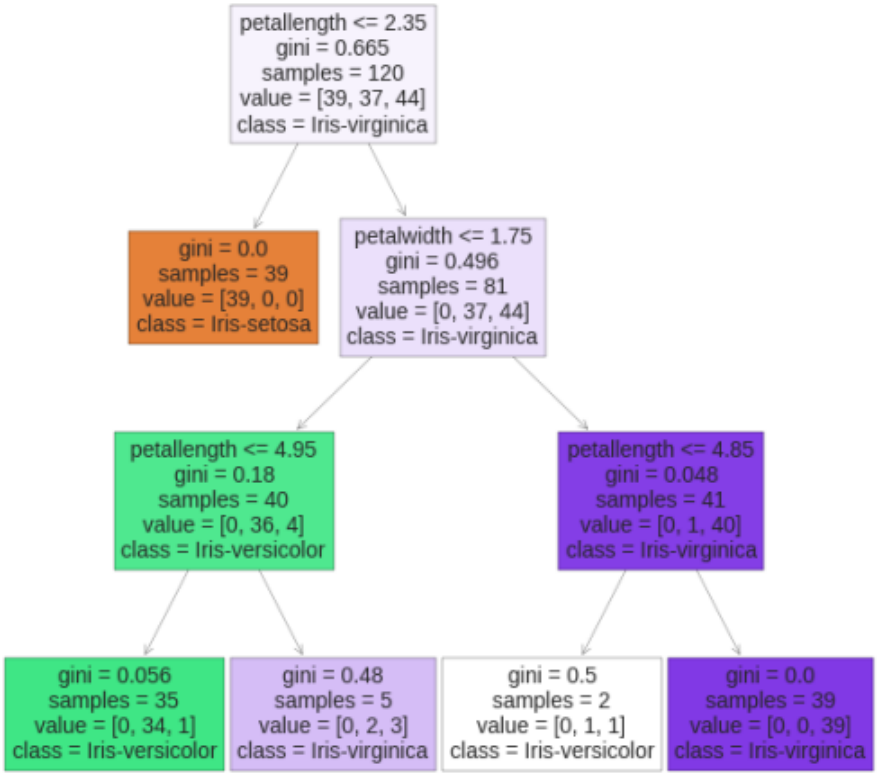
Marcos Antonio Lommez Candido Ribeiro[1]

Lista #3

Inteligência Artificial

[1]Aluno de Graduação em Ciência da Computação – tonilommez@hotmail.com

1) Com base nestas informações, qual as saídas da árvore para os seguintes registros de teste, respectivamente?



Inst.	Tam. da Pétala	Lar. da Pétala	Tam. da Sépala	Lar. da Sépala
1	3.46	0.87	2.45	1.78
2	1.67	1.89	0.78	1.32
3	2.56	2.34	2.45	1.78
4	6.67	2.34	2.45	1.78

c/d) Iris_Versicolor, íris_Setosa, Iris_Versicolor, Iris_Virgínica

2) Considerando a árvore da questão anterior, e as seguintes afirmações:

I. Esta árvore possui 5 regras de classificação

II. Das regras geradas, há apenas uma com cobertura por classe de 100%

III. A menor cobertura por classe é de 6.8% e corresponde à classe Iris_Virgínica

Resposta: c) I e II, apenas.

3) Considere a seguinte matriz de confusão obtida por meio do classificador, Árvore de decisão, para um problema de quatro classes:

Era da classe	Foi classificado como			
	A	B	C	D
A	10	4	2	1
B	1	15	2	0
C	2	3	20	5
D	4	1	2	50

Quais os valores para as métricas abaixo para cada uma das classes A, B, C e D?

	Precisão	Recall	F1Score	TVP	TFN	TFP	TVN
A	58.8%	58.8%	58.8%	58.8%	41.2%	6.7%	93.3%
B	65.2%	83.3%	73.1%	83.3%	16.7%	7.7%	92.3%
C	76.9%	66.6%	71.3%	66.6%	33.4%	6.8%	93.2%
D	89.2%	87.7%	88.4%	87.7%	12.3%	9.3%	90.7%

4) Considerando a base de dados abaixo. Mostre o primeiro e o segundo nível da árvore gerada.

História do crédito	Dívida	Garantias	Renda Anual	Risco
Ruim	Alta	Nenhuma	<15000	Alto
Desconhecida	Alta	Nenhuma	≥15000 a ≤35000	Alto
Desconhecida	Baixa	Nenhuma	≥15000 a ≤35000	Moderado
Desconhecida	Baixa	Nenhuma	>35000	Alto
Desconhecida	Baixa	Nenhuma	>35000	Baixo
Desconhecida	Baixa	Adequada	>35000	Baixo
Ruim	Baixa	Nenhuma	<15000	Alto
Ruim	Baixa	Adequada	>35000	Moderado
Boa	Baixa	Nenhuma	>35000	Baixo
Boa	Alta	Adequada	>35000	Baixo
Boa	Alta	Nenhuma	<15000	Alto
Boa	Alta	Nenhuma	≥15000 a ≤35000	Moderado
Boa	Alta	Nenhuma	>35000	Baixo
Ruim	Alta	Nenhuma	≥15000 a ≤35000	Alto

O primeiro nível da árvore será decidido a partir da "Renda Anual", sendo este o atributo de maior impacto. Já o segundo nível será baseado na "História de Crédito", tanto para a renda > \$35000 como para a renda ≥ \$15000 e ≤ \$35000. Por fim, o nível de renda < \$15000 define todos os seus valores em uma única categoria, logo não terá um segundo nível.

5) Investigue a biblioteca 'chefboost' e implemente os algoritmos ID3 e C45 para gerar a árvore da base de dados da questão anterior.

*Código fonte usado para gerar os valores anexado ao final do PDF

Para o treinamento da arvore separei o set em 11 dados para treinamento e 3 dados para validação.

ID3 e C.45 (mesma saída):

```
# obj[0]: Historia do credito
# obj[1]: Divida
# obj[2]: Garantias
# obj[3]: Renda Anual
def findDecision(obj):
    if obj[3] == '>35000':
        if obj[0] == 'Desconhecida':
            if obj[2] == 'Nenhuma':
                if obj[1] == 'Baixa':
                    return 'Alto'
                else: return 'Alto'
            elif obj[2] == 'Adequada':
                return 'Baixo'
            else: return 'Baixo'
        elif obj[0] == 'Boa':
            return 'Baixo'
        elif obj[0] == 'Ruim':
            return 'Moderado'
        else: return 'Moderado'
    elif obj[3] == '<15000':
        return 'Alto'
    elif obj[3] == '>=15000 a <=35000':
        if obj[1] == 'Alta':
            return 'Alto'
        elif obj[1] == 'Baixa':
            return 'Moderado'
        else: return 'Moderado'
    else: return 'Alto'
```

6) Investigue como é o funcionamento do algoritmo CART. Mostre todos os cálculos necessários para a geração da árvore.

O algoritmo CART se diferencia levemente dos outros algoritmos ao usar uma abordagem um pouco diferente e mais avançada. Entre suas diferenças está a utilização do método de medida de impureza GINI para cálculos de classificação e o uso do erro quadrático médio para problemas de regressão. Embora o algoritmo C4.5 possa trabalhar com valores numéricos, sua natureza ainda é primariamente classificatória, o que o diferencia do CART, onde existe a possibilidade de trabalhar com valores de regressão também. Também é válido dizer que a saída do algoritmo CART é binária, o que difere dos demais que podem possuir diversos filhos. (embora a biblioteca chefboost não implementa como binário)

$$\text{Gini}(t) = 1 - \sum_{i=1}^{c} p(i|t)^2 \tag{1}$$

$$\text{Erro Quadrático Médio}(t) = \frac{1}{|t|} \sum_{i \in t} (y_i - \bar{y}_t)^2 \tag{2}$$

7) Utilizando a biblioteca acima, rode o CART e veja a árvore gerada. Compare as 3 árvores geradas a partir da mesma base de dados.

*Código fonte usado para gerar os valores anexado ao final do PDF

Para o treinamento da arvore separei o set em 11 dados para treinamento e 3 dados para validação.

CART:

```
#obj[0]: Historia do credito
# obj[1]: Divida
# obj[2]: Garantias
# obj[3]: Renda Anual
def findDecision(obj):
    if obj[3] == '>35000':
        if obj[0] == 'Desconhecida':
            if obj[2] == 'Nenhuma':
                if obj[1] == 'Baixa':
                    return 'Alto'
                else: return 'Alto'
            elif obj[2] == 'Adequada':
```

```
13                  return 'Baixo'
14              else: return 'Baixo'
15          elif obj[0] == 'Boa':
16              return 'Baixo'
17          elif obj[0] == 'Ruim':
18              return 'Moderado'
19          else: return 'Moderado'
20      elif obj[3] == '<15000':
21          return 'Alto'
22      elif obj[3] == '>=15000 a <=35000':
23          if obj[1] == 'Alta':
24              return 'Alto'
25          elif obj[1] == 'Baixa':
26              return 'Moderado'
27          else: return 'Moderado'
28      else: return 'Alto'
```

As árvores geradas pelos 3 algoritmos nessa pequena base de dados foram idênticas utilizando a configuração de 11/3. A única diferença reside no atributo "metric_value", especificamente no CART. Acredita-se que essa diferença foi gerada devido à utilização do método Gini em vez do cálculo de entropia, mas ainda resultou na criação da mesma árvore no fim.