Cola para se dar bem no Mercado de Trabalho e Academia

Exercício I Prático Obrigatório – IAI Supervisionada para Predição e Regressão Multipla Robusta

Programa SAS

data cliente;

input Bu_Unit Sales Price Qu_level Claims NPS Satisfac;

datalines;

1 65.98 97.8 96.77 13.58 98.9 97.??

2 15.84 98.9 98.39 12.35 97.8 98.91

3 8.89 100 100 11.11 100 100

4 12.46 98.9 95.16 12.35 96.7 96.74

5 80.67 21.98 19.35 100 2.2 21.59

6 32.17 23.08 22.58 97.53 3.3 23.91

7 23.45 24.18 24.19 96.3 2.75 25

8 89.96 24.18 19.35 95.06 2.2 26.09

```
9 31.43 64.84 56.45 50.62 65.93 65.22
10 11.23 65.93 51.61 49.38 71.43 66.3
11 77.46 70.33 53.23 46.91 63.74 68.48
12 23.9 68.13 51.61 45.68 61.54 67.39
13 7.4 86.81 80.65 25.93 90.11 86.96
14 0.29 87.91 79.03 24.69 85.71 85.87
15 83.42 87.91 77.42 22.22 90.11 88.04
16 100 86.81 75.81 25.93 84.62 84.78
17 15.84 98.9 98.39 12.35 97.8 28.91
;
proc print; run;
/* input Bu_Unit Sales Price Qu_level
  Claims NPS Satisfac;*/;
proc robustreg;
  model Satisfac = Sales Price Qu_level
  Claims NPS /diagnostics;
run;
```

Resultado

Estimaciones de parámetro						
Parámetro	DF	Estimación	Error estándar	Límites de confianza al 95%		Chi-cuadrado
Intercept	1	56.9427	25.9621	6.0579	107.8275	4.8
Sales	1	0.0020	0.0070	-0.0118	0.0158	0.0
Price	1	0.2437	0.2473	-0.2410	0.7283	0.9
Qu_level	1	0.1104	0.0442	0.0237	0.1970	6.2
Claims	1	-0.4243	0.2109	-0.8377	-0.0109	4.0
NPS	1	0.1180	0.0570	0.0062	0.2297	4.2
Scale	1	0.8244				

Vemos que as variáveis preditoras significativas foram Nível de Qualidade (margem de confiança- mc 98,7%), atuando positivamente (coeficiente de regressão-cr 0,1104); Reclamações (mc 95,58%) atuando negativamente (cr -0,4243), sendo a variável que mais impacta na satisfação do cliente (maior valor absoluto), quando aumenta 1% a reclamação cai 0,4243% a satisfação dos clientes. Por último NPS (mc 96,15%).

Diagnósticos		
Obs	Residual robusto estandarizado	Valor extremo
17	-76.7228	*

Veja que a observação 17 foi um outlier.

Estimaciones de parámetro					
Variable	DF	Estimación de parámetros	Error estándar	t valor	Pr > |t|
Intercept	1	133.83449	572.99783	0.23	0.8196
Sales	1	0.04179	0.15544	0.27	0.7930
Price	1	-0.40530	5.45778	-0.07	0.9421
Qu_level	1	-0.45918	0.97620	-0.47	0.6473
Claims	1	-0.97117	4.65490	-0.21	0.8385
NPS	1	0.48696	1.25822	0.39	0.7061

Perceba que se rodarmos proc reg, teremos o resultado do Excel que sempre esta errado (99% das vezes), veja que nada deu significativo por que temos um outlier.

Acabamos com a discussão de ciência de dados (estatística) agora iremos para a IA

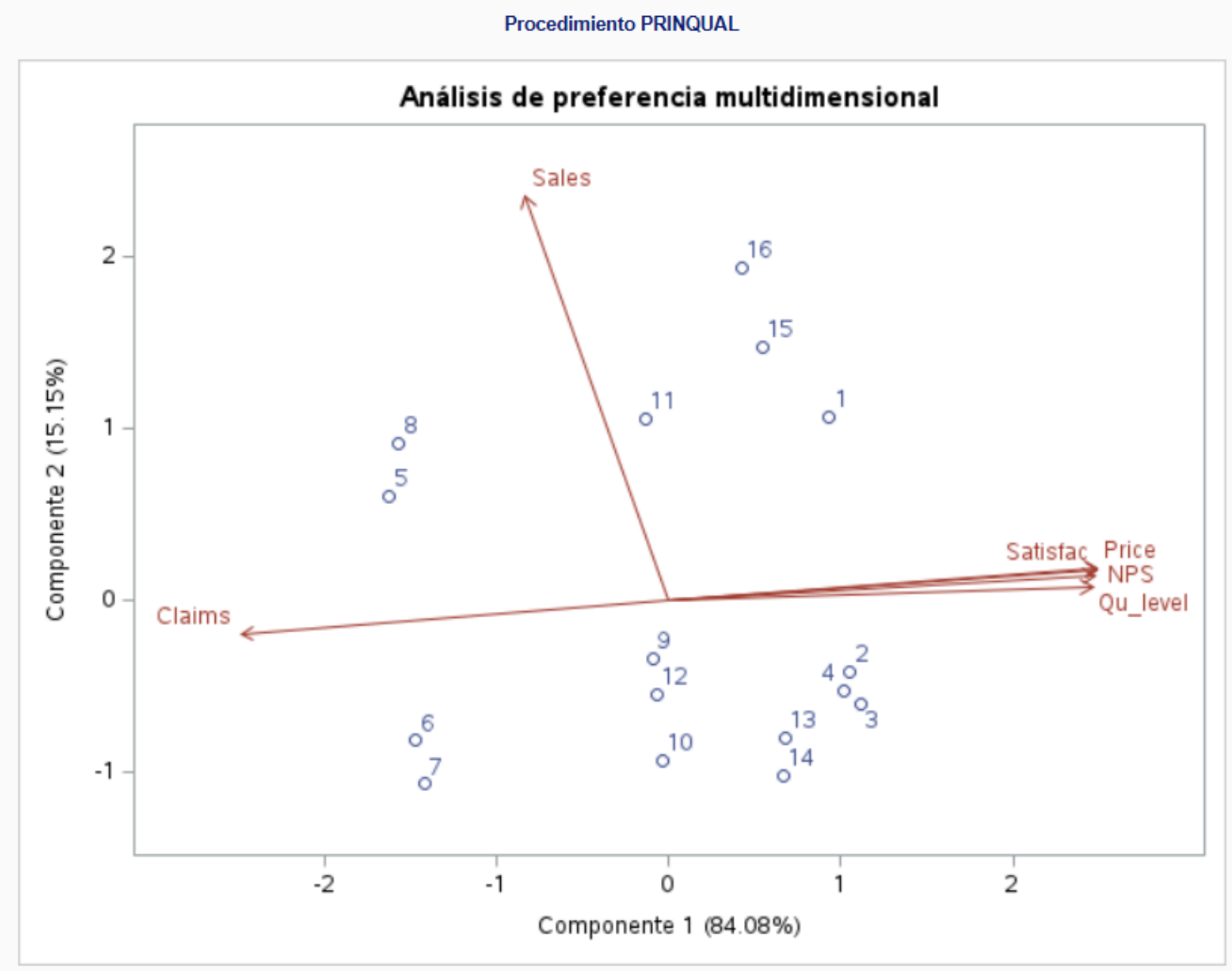
Temos que eliminar o dado 17 por que é um outlier

Programa SAS

```
data pca_reg;
input Sales Price Qu_level Claims NPS Satisfac;
cards;
65.98 97.8 96.77 13.58 98.9 97.83
15.84 98.9 98.39 12.35 97.8 98.59
8.89 100 100 11.11 100 100
12.46 98.9 95.16 12.35 96.7 96.74
```

```
80.67 21.98 19.35 100 2.2 21.59
32.17 23.08 22.58 97.53 3.3 23.91
23.45 24.18 24.19 96.3 2.75 25
89.96 24.18 19.35 95.06 2.2 26.09
31.43 64.84 56.45 50.62 65.93 65.22
11.23 65.93 51.61 49.38 71.43 66.3
77.46 70.33 53.23 46.91 63.74 68.48
23.9 68.13 51.61 45.68 61.54 67.39
7.4 86.81 80.65 25.93 90.11 86.96
0.29 87.91 79.03 24.69 85.71 85.87
83.42 87.91 77.42 22.22 90.11 88.04
100 86.81 75.81 25.93 84.62 84.78
;
proc print; run;
proc prinqual plots=(MDPref);
  transform identity(Sales Price Qu_level Claims NPS Satisfac);
  ods select MDPrefPlot;
run;
```

Resultado



Podemos observar na ferramenta PCA-Biplo (IAI ou ML Não Supervisionado para Redução de Dimensão e Asociação) que a variável Price (precificação) foi a que apresentou menor ângulo com a Satisfação, assim ela é mais importante que NPS e Nível de Qualidade, que apresentam vetores apontando na mesma direção. Resultado diferente ao da estatística, ou seja a estatística erra no resultado da análise.

As reclamações aparecem formando um ângulo de 180 graus aproximadamente, ou seja quando aumentam as reclamações diminui a satisfação de clientes.

Se rodássemos Random Forest (IAI ou ML supervisionado para Predição) a principal variável preditora seria Precificação.

Exercício II Pratico Obrigatório – ANOVA – MANOVA – Winsorização – IA

IA:

- IAI ou ML Não Supervisionada para Agrupamentos e Distancias Multivariados:

- Cluster Analysis

- IAI Não Supervisionada para Redução de Dimensão e Associação:

- PCA-Biplot

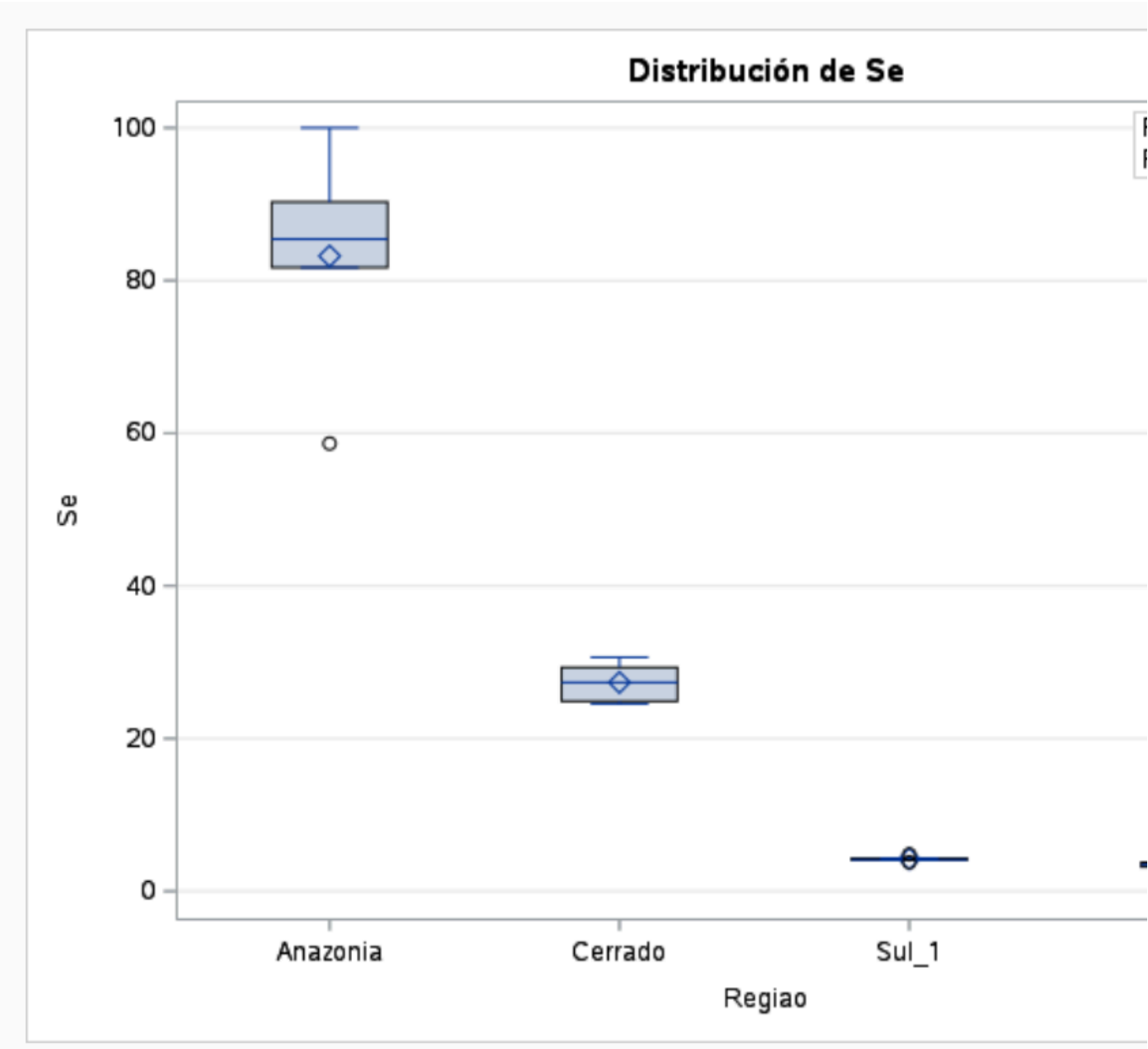
- IAI Supervisionada para Classificação: Rede Neural, Random Forest, SVM, N Bayes etc.

Programa SAS

```
 data rastreab;
input Registro Regiao $ Rb Se Cs Br;
/* Rastreabilidade da Soja Brasileira */;
datalines;
1 Anazonia 73.99 100.00 96.97 91.??
2 Anazonia 78.02 90.28 93.94 80.00
3 Anazonia 70.70 81.71 94.50 85.71
4 Anazonia 77.29 58.65 100.00 68.57
5 Anazonia 75.00 85.42 93.94 57.14
6 Cerrado 82.05 24.55 78.79 94.29
7 Cerrado 80.22 29.32 81.82 60.00
8 Cerrado 87.18 30.64 96.97 71.43
9 Cerrado 88.28 24.88 100.00 97.14
10 Cerrado 84.43 27.35 89.39 77.14
11 Sul_1 100.00 4.12 81.82 76.40
12 Sul_1 93.40 4.10 79.00 77.14
13 Sul_1 92.31 3.79 78.79 71.43
14 Sul_1 96.70 4.28 81.60 71.43
15 Sul_1 93.68 4.82 81.82 72.86
16 Sul_2 95.97 3.20 81.30 65.71
17 Sul_2 88.64 3.46 81.82 97.14
18 Sul_2 93.04 3.13 80.30 100.00
19 Sul_2 89.74 3.79 81.21 57.14
20 Sul_2 91.85 4.28 80.53 60.00
```

```
;
proc print; run;
/* input Registro Regiao $ Rb Rb Se Cs Br; */;
proc anova;
  class Regiao;
  model Rb Se Cs Br = Regiao;
  run;
```

Resultado



Podemos observar que existe um outlier para a variável preditora Se (selênio), se encontra na região Amazônica e apresenta um valor menor a 60%. Temos 2 opções para levar esse banco de dados para a IA.

A primeira é eliminar o outlier (*Triming),* isso é mal visto pelos clientes, financiadores e parceiros. Para cortar esse outlier simplesmente devemos trocar o valor do outlier por um ponto no banco de dados do SAS.

Obs	Registro	Regiao	Rb	Se	Cs	Br
1	1	Amazonia	73.99	100.00	96.97	91.59
2	2	Amazonia	78.02	90.28	93.94	80.00
3	3	Amazonia	70.70	81.71	94.50	85.71
4	4	Amazonia	77.29	58.65	100.00	68.57
5	5	Amazonia	75.00	85.42	93.94	57.14
6	6	Cerrado	82.05	24.55	78.79	94.29
7	7	Cerrado	80.22	29.32	81.82	60.00
8	8	Cerrado	87.18	30.64	96.97	71.43
9	9	Cerrado	88.28	24.88	100.00	97.14
10	10	Cerrado	84.43	27.35	89.39	77.14
11	11	Sul_1	100.00	4.12	81.82	76.40
12	12	Sul_1	93.40	4.10	79.00	77.14
13	13	Sul_1	92.31	3.79	78.79	71.43
14	14	Sul_1	96.70	4.28	81.60	71.43
15	15	Sul_1	93.68	4.82	81.82	72.86
16	16	Sul_2	95.97	3.20	81.30	65.71
17	17	Sul_2	88.64	3.46	81.82	97.14

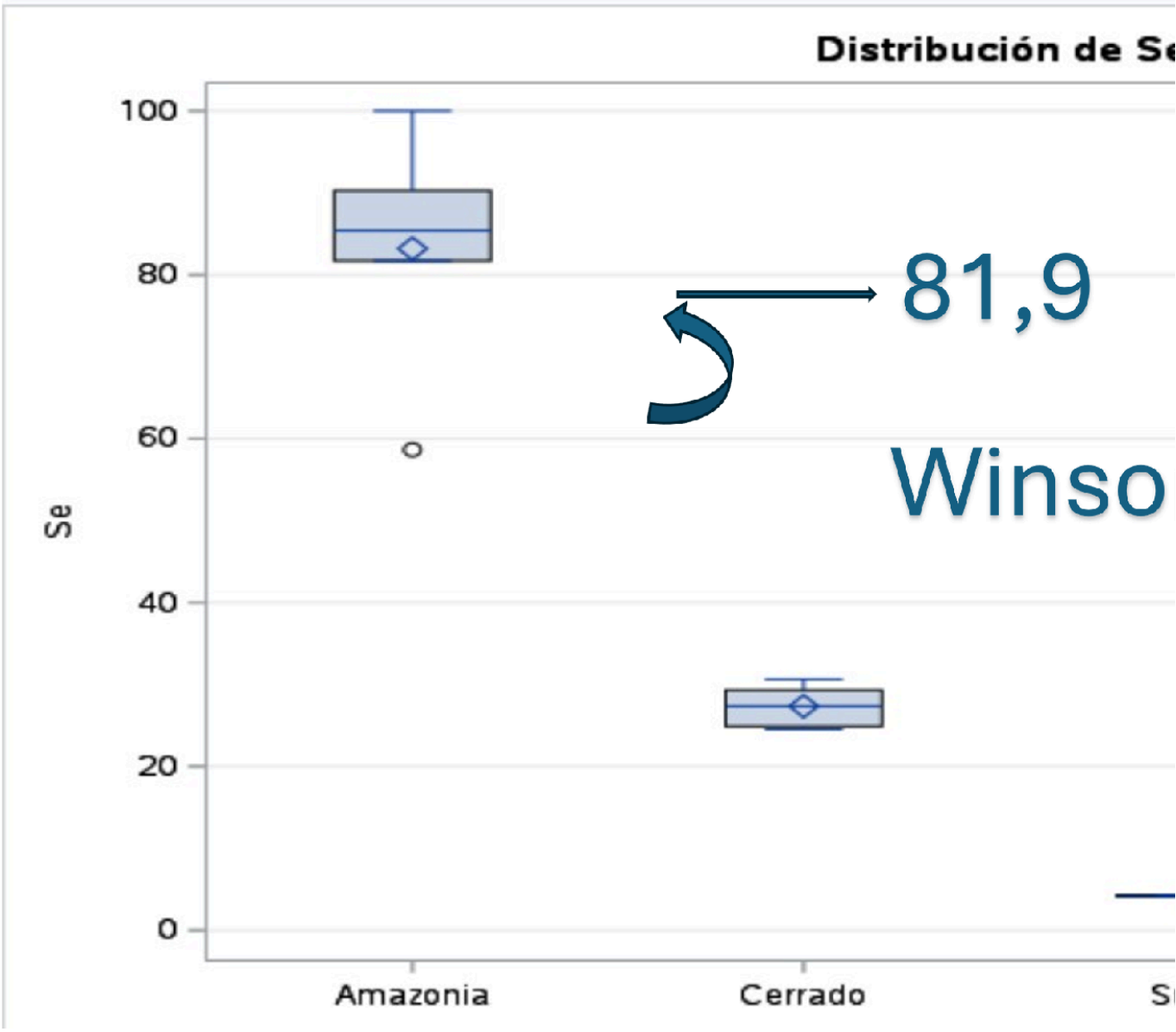
5 re

4

Rast

Vemos que o outlier é 58.65, que deve ser substituído por um ponto. Também vemos que se cortarmos esse outlier também perdemos os demais dados da línea, 77.29 100 e 68.57 quando a gente roda IA.

A alternativa e Winsorizar (*Winsorizing*), assim podemos substituir o valor do outlier (dado extremo fora de contexto) por um dado extremo dentro do contexto, por exemplo colocar o outlier inferior no bigode inferior do box plot.

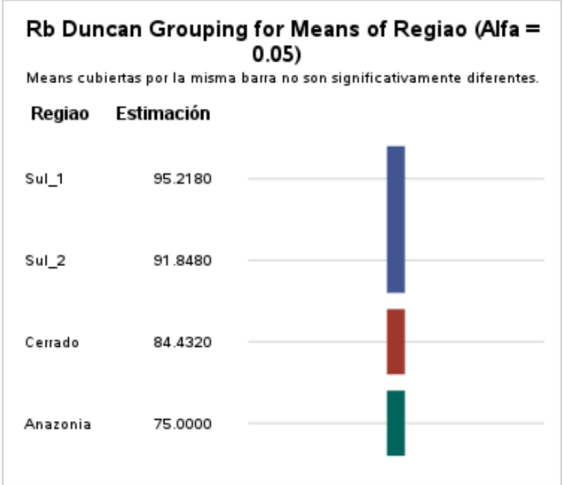


Assim troco o valor 58.65 pelo valor 81,9, bigode inferior que coincide neste casso com o piso da caixa.

Programa SAS para Teste de Comparação de Médias de Duncan

```
proc anova;
    class Regiao;
    model Rb Se Cs Br = Regiao;
```

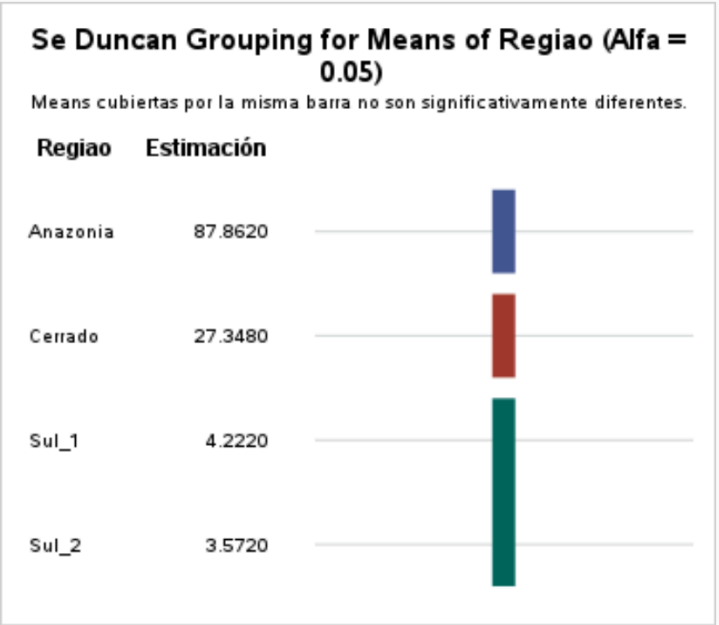
```
means Regiao / duncan lines;
run;
```



Podemos observar que Sul_1 e Sul_2 apresentaram as maiores medias aritméticas, diferiram das outras regiões com 95% confiança. Porem Sul_1 e Sul_2 não apresentaram diferenças estatisticamente significativas (as duas regiões estão unidas por uma barra azul).

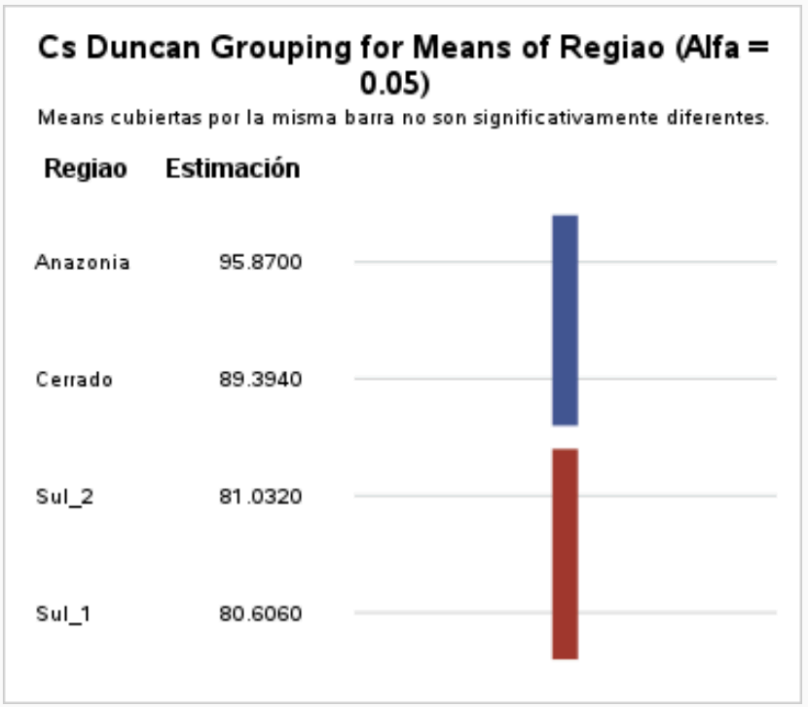
As regiões Sul diferiram estatisticamente do Cerrado e da Amazonia.

O Cerrado diferiu da Amazonia.



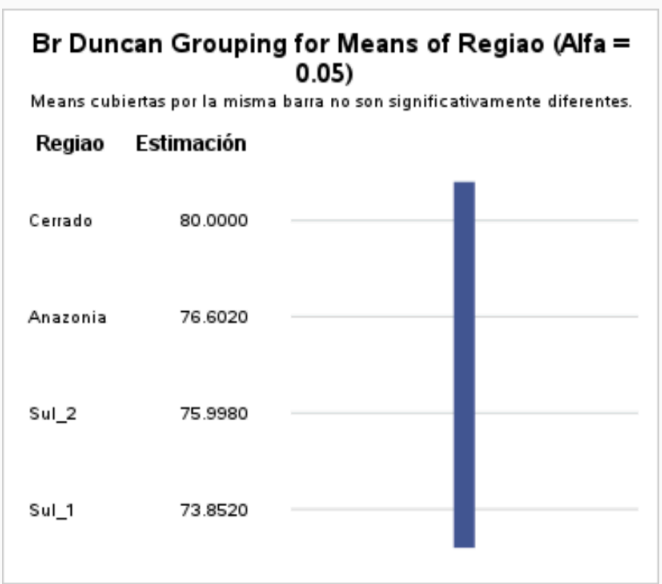
A maior média para a variável preditora Se: Selênio foi a da Amazonia, que diferiu estatisticamente das outras regiões com 95% de confiança pelo Teste de Duncan.

Em seguida temos a média do Cerrado, que diferiu das régios Sul, que não apresentaram diferenças, mesmo comportamento observado para o Rb.



Para o Cs: Césio, podemos observar que se formaram dois grupos, Amazonia e Cerrado com maiores médias aritméticas e Sul_1 e Sul_2 com as menores médias.

Analogamente ao observado para Rb e Se as regiões Sul não diferem entre si, assim possivelmente não exista diferença estatística entre elas o que indica que as duas regiões deverão ser unidas para a IA, essa tendencia será confirmada na MANOVA.



Para o elemento químico Br: Bromo, não foram detectas diferenças estatísticas entre as regiões.

A confiança para detectar diferenças para o Br foi muito baixa, 7% (p < 0,9296), assim o Br não deve ir para a IA. Conferimos essa probabilidade através de RobustANOVA, quadro a seguir:

Test de Kruskal-Wallis		
Chi-cuadrado	DF	Pr > ChiSq
0.4254	3	0.9349

Programa SAS para IAI Não Supervisionada - Agrupamentos e Distancias Multivariados, Cluster Analysis

Para obter esses dados utilizamos a opção Tabela Dinâmica do Excel o Libre Office Calc no banco de dados original. Tabela Dinâmica é a principal ferramenta de Data Crunching.

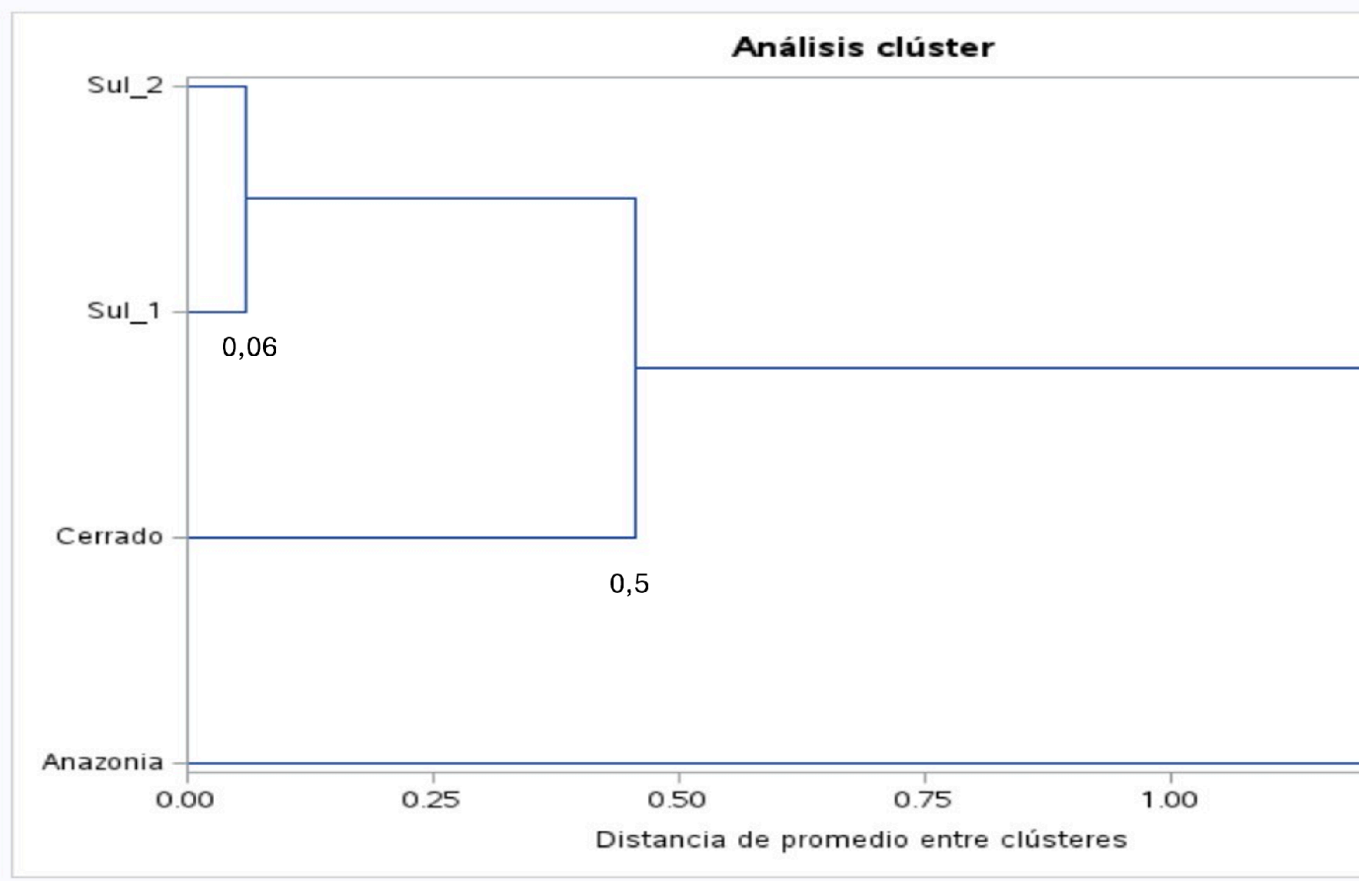
Observe que agora trabalharemos com as médias aritméticas obtidas por Tabela Dinâmica em Excel ou Libre Office Calc.

```
data cluster;
input Regioes $ Rb Se Cs Br;
datalines;
Anazonia 75 87.862 95.87 76.602
Cerrado 84.432 27.348 89.394 80
Sul_1 95.218 4.222 80.606 73.852
Sul_2 91.848 3.572 81.032 75.998
;
proc print; run;
/* IAI ou ML Não Supervisioando para Agrupamentos e Distancias
   Multivariados - Cluster Analysis */;
/* input Regioes $ Rb Se Cs Br; */;
proc cluster outtree = arvore method = average;
var Rb Se Cs;
id Regioes;
```

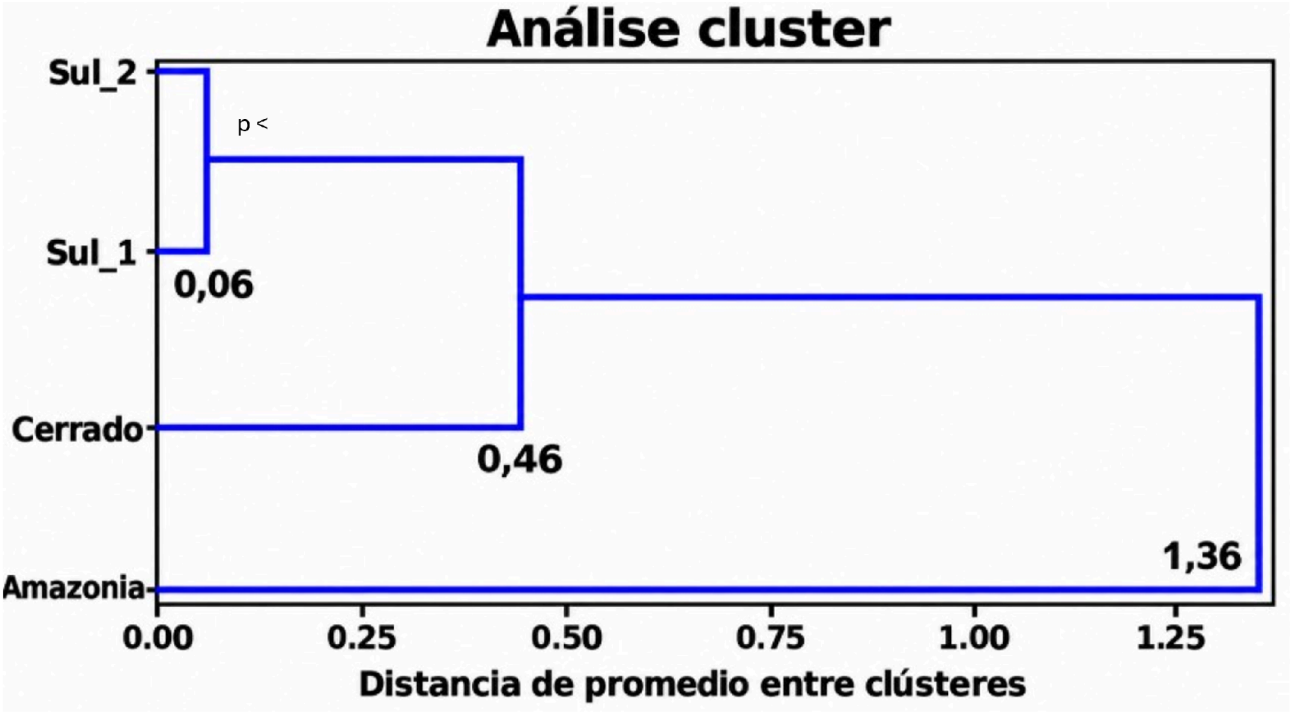
run;

PROC TREE DATA = arvore;

RUN;



Cluster com alta resolução com ajuda do GPT



Historial de clúster					
Número de clústers	Clústeres unidos		Frec	Norm RMS Distance	Enlace
3	Sul_1	Sul_2	2	0.0592	
2	Cerrado	CL3	3	0.456	
1	Anazonia	CL2	4	1.3639	

Podemos observar que a menor distancia multivariada aconteceu entre o Sul_2 e Sul_1, possivelmente termos que juntar essas duas regiões como somente uma região Sul, para isso precisamos fazer um contraste de MANOVA, para calcularmos os p valores (confiança para empresas). Isso é

indispensável para rodar IAI ou ML Supervisionado para Classificação, o mais importante para elaborar robôs de software ou aplicativos utilizando algoritmos como Redes Neurais, Random Forest, Support Vector Machine, Navies Bayes, etc. Esse tipo de IA rodaremos no Weka na disciplina.

Também esse tipo de IA é o mais importante do mundo, para substituir tarefas dos seres humanos.

Programa SAS para rodar MANOVA e

Contrastes Multivariados

data rastr_soja;

input Registro Regiao $ Rb Se Cs Br;

datalines;

1 Anazonia 73.99 100 96.97 91.59

2 Anazonia 78.02 90.28 93.94 80

3 Anazonia 70.7 81.71 94.5 85.71

4 Anazonia 77.29 81.9 100 68.57

5 Anazonia 75 85.42 93.94 57.14

6 Cerrado 82.05 24.55 78.79 94.29

7 Cerrado 80.22 29.32 81.82 60

8 Cerrado 87.18 30.64 96.97 71.43

9 Cerrado 88.28 24.88 100 97.14

10 Cerrado 84.43 27.35 89.39 77.14

11 Sul_1 100 4.12 81.82 76.4

12 Sul_1 93.4 4.1 79 77.14

13 Sul_1 92.31 3.79 78.79 71.43

14 Sul_1 96.7 4.28 81.6 71.43

15 Sul_1 93.68 4.82 81.82 72.86

16 Sul_2 95.97 3.2 81.3 65.71

17 Sul_2 88.64 3.46 81.82 97.14

18 Sul_2 93.04 3.13 80.3 100

19 Sul_2 89.74 3.79 81.21 57.14

20 Sul_2 91.85 4.28 80.53 60

```
;
proc print; run;
/* input Registro Regiao $ Rb Se Cs Br; */
/* Ordem Alfabetica dos Niveis do Fator Regiao: [Amazonia Cerrado Sul_1
Sul_2] */;

title "Contrates em MANOVA";

proc glm;
  class Regiao;
  model  Rb Se Cs Br = Regiao;
 contrast "Sul_1 Vs Sul_2" Regiao  0 0 1 -1;
 contrast " Cerrado Vs Sul_1 e Sul_2" Regiao 0 2 -1 -1;
 contrast " Amazonia Regia Vs Cerrado, Sul_1 e Sul_2" Regiao 3 -1 -1 -1;
 manova h=_all_ / printe printh;
  contrast "Sul_1 Vs Sul_2" Regiao  0 0 1 -1;
  contrast " Cerrado Vs Sul_1 e Sul_2" Regiao 0 2 -1 -1;
  contrast " Amazonia Regia Vs Cerrado, Sul_1 e Sul_2" Regiao 3 -1 -1 -1;
run;
```

Resultados MANOVA Contrastes

Criterio de test MANOVA y estadísticos F exactos para la hipótesis de efectos Sul_1 Vs Sul_2 no generales
H = Contraste Matriz SSCP para Sul_1 Vs Sul_2
Matriz SSCP de error E =

S=1 M=0.5 N=6

Estadístico	Valor	Valor F	Num DF	Den DF	Pr > F
Lambda de Wilks	0.76244112	1.45	3	14	0.2696
Traza de Pillai	0.23755888	1.45	3	14	0.2696
Traza Hotelling-Lawley	0.31157668	1.45	3	14	0.2696
Raíz más grande de Roy	0.31157668	1.45	3	14	0.2696

Podemos observar que a margem de confiança do contraste Sul_1 Vs Sul_2 foi menor que 95% (73%) que é o valor mínimo para falar que tem diferença entre categorias, assim não tem diferença entre as duas regiões Sul e elas devem ser fusionadas (juntadas) para a IAI ou ML Supervisionado para Classificação.

Agora faremos a comparação entre Cerrado e regiões Sul

Criterio de test MANOVA y estadísticos F exactos para la hipótesis de efectos Cerrado Vs Sul_1 e Sul_2 no gene
H = Contraste Matriz SSCP para Cerrado Vs Sul_1 e Sul_2
Matriz SSCP de error E =

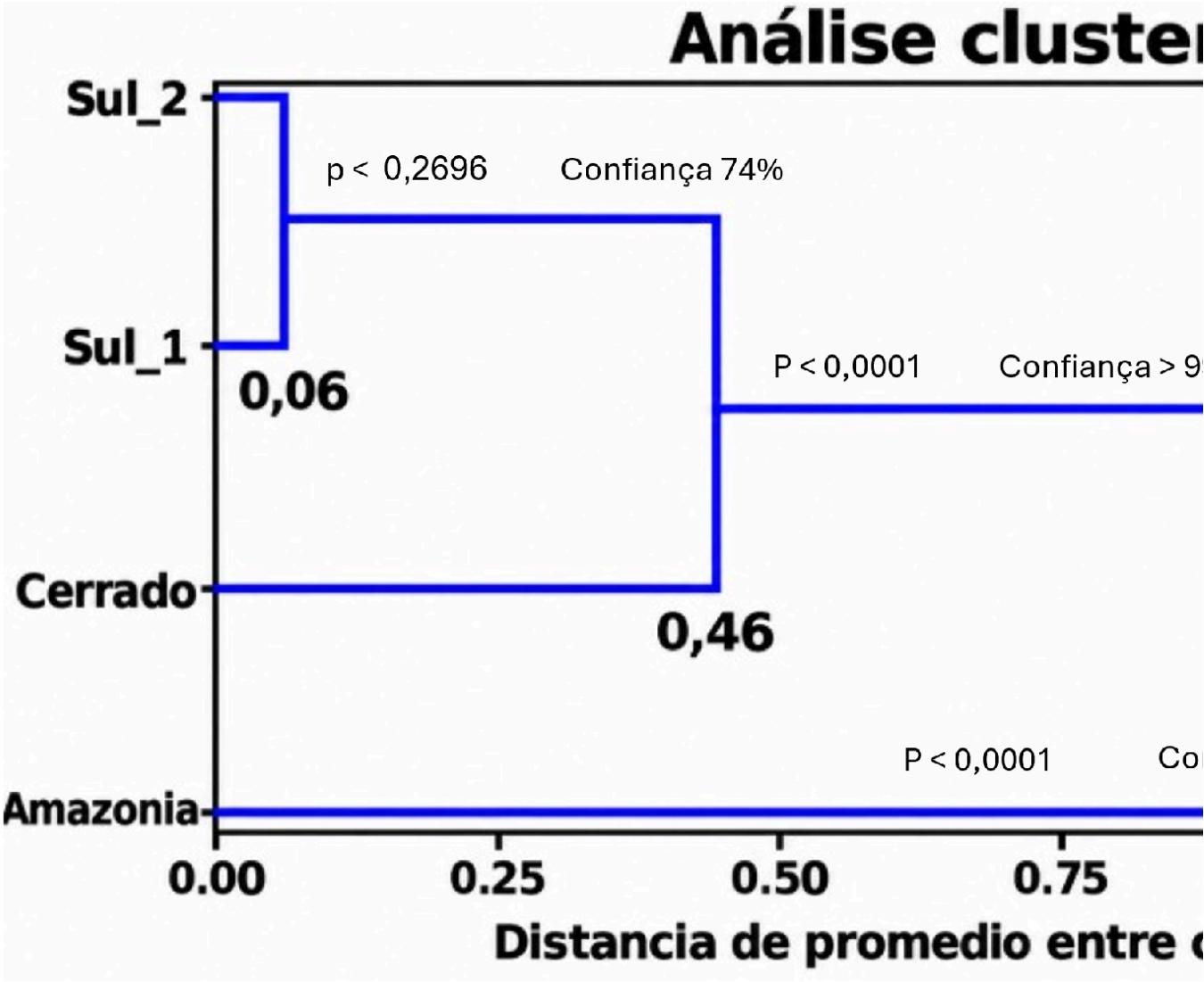
S=1 M=0.5 N=6

Estadístico	Valor	Valor F	Num DF	Den DF	F
Lambda de Wilks	0.07500024	57.56	3	14	<.
Traza de Pillai	0.92499976	57.56	3	14	<.
Traza Hotelling-Lawley	12.33329129	57.56	3	14	<.
Raíz más grande de Roy	12.33329129	57.56	3	14	<.

Podemos observar que o Cerrado é diferente das regiões Sul, confiança maior que ((1-0,0001)*100= 99, 99%). Assim concluímos que Cerrado difere das regiões Sul do ponto de vista multivariado.

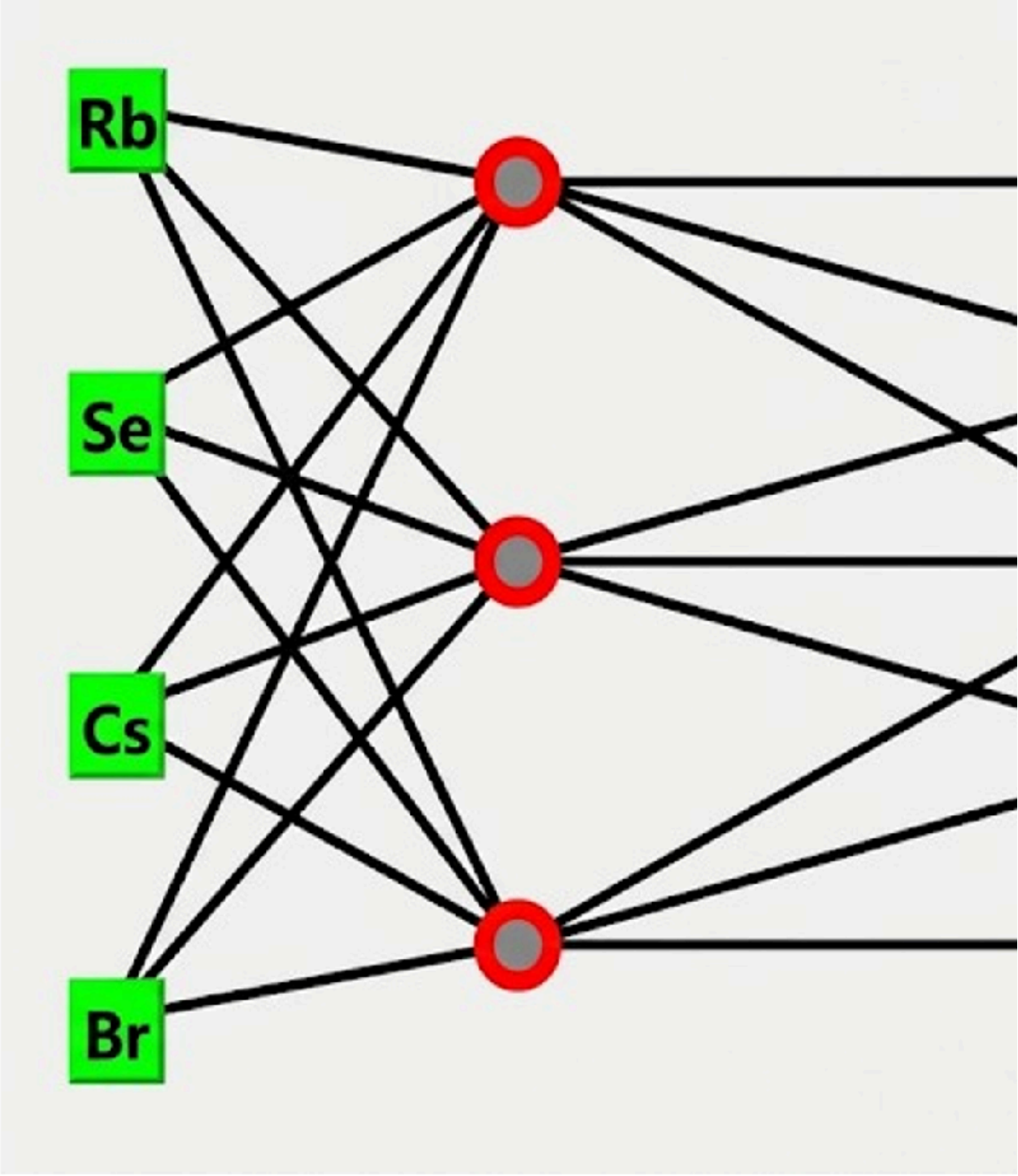
erente das Cerrado e Sul (p < 0,0001 ou confiança de 99,99%).

Agora vamos passar os p valores de Contrastes em MANOVA para o dendrograma de Cluster Analysis.



Agora o próximo passo é rodar IAI ou ML Supervisionado para Classificação, que faremos no Weka.

Rede Neu



Rede Neura

```
=== Confusio

   a   b   c
   5   0   0 |
   0   4   1 |
   0   0  10 |
```

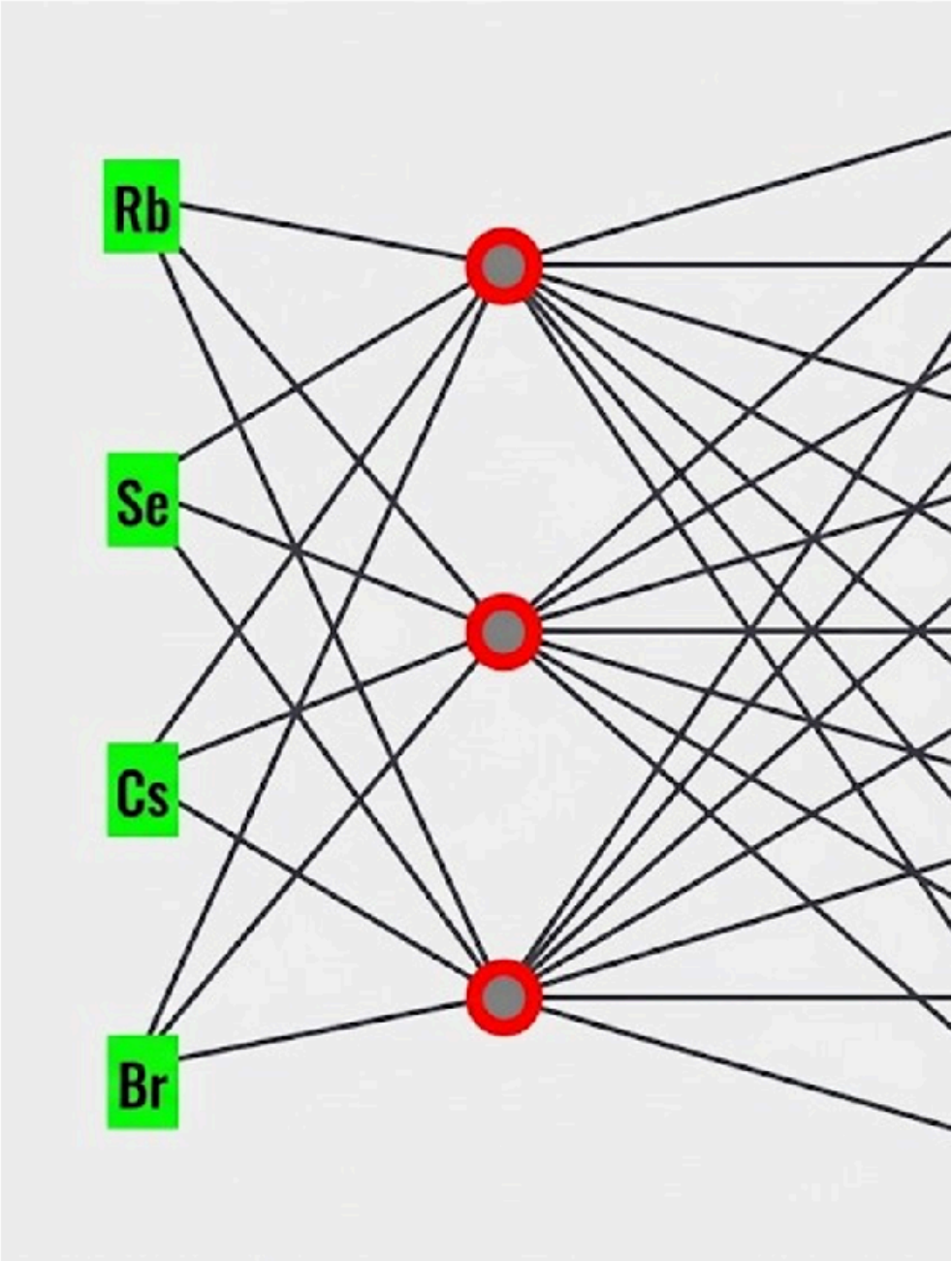
Margem de Erro= 1/20*1

Função

Novo conhecime

Inputs
Threshold
Attrib Rb
Attrib Se
Attrib Cs
Attrib Br

Rede Neura



Rede Neura

```
=== Confusion

  a   b   c    <
  5   0   0 |
  0   5   0 |
  0   0  10 |
```

Margem de Erro= 0/20*100

```
=== Confusion Matrix ===

  a  b  c   <-- classified as
  5  0  0 |  a = Amazonia
  0  5  0 |  b = Cerrado
  0  0 10 |  c = Sul
```

Rede N
Camada
Cross
Sp

```
=== Confusion Matrix ===

 a  b  c   <-- classi
 5  0  0 |   a = Amazo
 0  5  0 |   b = Cerra
 0  0 10 |   c = Sul
```