PUC - SP

ÁRVORE DE CLASSIFICAÇÃO E INDICADORES DE GOVERNANÇA MUNDIAIS

ALDAIR ALMEIDA FONSECA
MESTRADO EM ADMINISTRAÇÃO DE EMPRESAS
MÉTODOS QUANTITATIVOS
PROF. DR. ARNOLDO HOYOS

SÃO PAULO
2010

INTRODUÇÃO

O trabalho apresentado tem como principal objetivo comparar os resultados obtidos na Análise discriminante e na Regressão logística do Mintab com a análise de Árvore de Classificação do SPSS, essa análise é denominada de técnica de classificação supervisionada.

1. ENTENDENDO OS DADOS

A base utilizada nesse trabalho se refere aos Indicadores de Governança Mundiais (Worldwide Governance Indicators - WGI).

O banco de dados inicial contempla 211 países de todos os continentes do mundo analisados por 6 variáveis. Através de análise de Amostragem, Componentes Principais e Conglomerados trabalharemos com 48 países, já excluindo-se dessa análise os "outliers", de acordo com tabela abaixo:

Pais	VA	OS	GGE	RQ	RL	CC	Cluster para 60	Princ. Comp.	Nível
NIC	-0,14	-0,39	-0,96	-0,36	-0,86	-0,81	1	-0,62	a
MMR	-2,24	-1,56	-1,68	-2,24	-1,48	-1,69	2	0,33	a
BEN	0,34	0,35	-0,52	-0,46	-0,54	-0,42	2	-1,22	a
SAU	-1,74	-0,39	0,01	0,17	0,33	0,11	2	-2,5	a
SAM	0,63	1,11	-0,07	-0,43	0,74	0,24	2	-2,3	a
IND	0,45	-0,99	-0,03	-0,21	0,12	-0,37	2	-0,88	a
URY	1,02	0,83	0,48	0,08	0,5	1,12	3	-2,92	a
KEN	-0,16	-1,25	-0,6	-0,07	-0,98	-1,01	4	4,01	b
UZB	-1,9	-0,91	-0,68	-1,41	-1,18	-1,08	4	3,54	b
ASM	1	1	0,49	0,36	1,17	0,36	4	4,82	b
AUS	1,36	1,08	1,9	1,78	1,76	2,03	4	2,21	b
VCT	1,11	0,81	0,74	0,4	0,87	1	3	-3,01	a
LBN	-0,4	-1,94	-0,64	-0,2	-0,73	-0,83	4	2,52	b
YUG	0,19	-0,5	-0,28	-0,21	-0,46	-0,16	2	-1,2	a
KGZ	-0,72	-0,68	-0,7	-0,32	-1,26	-1,06	2	-1,55	a
SLE	-0,28	-0,23	-1,13	-0,86	-1,03	-1,07	2	-0,58	a
WBG	-0,94	-1,76	-1,36	-1,12	-0,81	-1,13	3	-2,62	a
KIR	0,71	1,4	-0,58	-1,22	0,39	0,01	2	-0,43	a
HRV	0,48	0,57	0,52	0,5	0,08	0,12	1	-0,21	a
SWE	1,53	1,13	1,99	1,68	1,9	2,24	4	4,45	b
NGA	-0,6	-2,01	-0,98	-0,62	-1,12	-0,92	1	0,03	a
MDA	-0,27	-0,38	-0,76	-0,2	-0,46	-0,64	2	-0,42	a
FRA	1,24	0,58	1,54	1,25	1,4	1,43	4	4,13	b
RWA	-1,24	-0,14	-0,2	-0,49	-0,5	0,03	2	-0,96	a
KAZ	-1,01	0,51	-0,47	-0,37	-0,78	-0,95	2	-0,58	a
SVK	0,89	0,92	0,76	1,14	0,52	0,43	5	-1,23	a

PRK	-2,21	0,35	-2,12	-2,28	-1,06	-1,74	6	-1,21	a
GMB	-0,97	0,14	-0,77	-0,44	-0,25	-0,78	1	0,04	a
VEN	-0,62	-1,23	-0,85	-1,44	-1,59	-1,13	3	-2,75	a
LAO	-1,71	-0,01	-0,84	-1,25	-0,9	-1,23	4	1,99	b
GUF	0,35	0,08	0,76	0,85	0,62	0,84	4	3,49	b
QAT	-0,77	1,01	0,68	0,66	0,86	1,24	2	0,83	a
SDN	-1,77	-2,44	-1,41	-1,36	-1,5	-1,49	2	-1,43	a
GBR	1,33	0,56	1,74	1,79	1,68	1,77	4	2,04	b
COL	-0,26	-1,66	0,13	0,24	-0,5	-0,25	4	2,34	b
MAR	-0,7	-0,47	-0,09	-0,03	-0,11	-0,26	2	-0,81	a
EST	1,03	0,57	1,15	1,47	1,05	0,94	2	-0,58	a
NPL	-0,79	-1,69	-0,75	-0,66	-0,76	-0,68	2	-0,62	a
TUR	-0,19	-0,73	0,2	0,22	0,09	0,1	4	1,73	b
AZE	-1,23	-0,48	-0,64	-0,32	-0,76	-1	2	-1,05	a
BGD	-0,61	-1,54	-0,77	-0,82	-0,7	-1,1	2	-2,21	a
OMN	-1,07	0,95	0,42	0,65	0,82	0,59	4	3,38	b
ABW	1	1,38	1,29	0,85	0,89	1,32	2	-1,9	a
UGA	-0,47	-0,88	-0,51	-0,08	-0,51	-0,79	2	-2,64	a
TZA	-0,09	0,01	-0,45	-0,39	-0,28	-0,51	2	-0,79	a
BHS	1,14	0,74	1,1	0,99	1,2	1,38	4	2,76	b
CHE	1,45	1,23	2,06	1,66	1,86	2,15	3	-3,46	a
BGR	0,6	0,39	0,1	0,75	-0,12	-0,17	2	-1,96	a

2. ANÁLISE DISCRIMINANTE

Discriminant Analysis: Cod 60 versus CC60; GE60; PS60

```
Linear Method for Response: Cod 60

Predictors: CC60; GE60; PS60

Group          1          2

Count         34         14


Summary of classification

                    True Group

Put into Group      1      2

1                  34      0

2                   0     14

Total N            34     14
```

```
N correct          34     14

Proportion      1,000  1,000

N = 48           N Correct = 48          Proportion Correct = 1,000


Squared Distance Between Groups

         1        2

1   0,0000  17,4402

2  17,4402   0,0000


Linear Discriminant Function for Groups

                1        2

Constant  -1,0227  -3,8500

CC60      -2,3442   2,4591

GE60      -0,9070   3,5854

PS60      -0,0587   0,4570
```

Através da análise discriminante podemos observar 100% de acerto da análise conforme destacado em vermelho, alguns países foram remanejados de seus grupos originais para efeito de ajuste.

3. REGRESSÃO LOGÍSTICA

Ordinal Logistic Regression: code para 60 versus CC60; GE60; PS60

```
Link Function: Logit

Response Information

Variable      Value  Count

code para 60  ABW        1

              ASM        1

              AUS        1

              AZE        1

              BEN        1

              BGD        1
```

BGR	1
BHS	1
CHE	1
COL	1
EST	1
FRA	1
GBR	1
GMB	1
GUF	1
HRV	1
IND	1
KAZ	1
KEN	1
KGZ	1
KIR	1
LAO	1
LBN	1
MAR	1
MDA	1
MMR	1
NGA	1
NIC	1
NPL	1
OMN	1
PRK	1
QAT	1
RWA	1
SAM	1
SAU	1
SDN	1
SLE	1
SVK	1
SWE	1
TUR	1

TZA	1
UGA	1
URY	1
UZB	1
VCT	1
VEN	1
WBG	1
YUG	1
Total	48

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Const(1)	-4,11304	1,01100	-4,07	0,000			
Const(2)	-3,38094	0,729081	-4,64	0,000			
Const(3)	-2,93581	0,606233	-4,84	0,000			
Const(4)	-2,61717	0,535941	-4,88	0,000			
Const(5)	-2,36824	0,489708	-4,84	0,000			
Const(6)	-2,15962	0,456132	-4,73	0,000			
Const(7)	-1,98025	0,430672	-4,60	0,000			
Const(8)	-1,81897	0,410257	-4,43	0,000			
Const(9)	-1,67300	0,393665	-4,25	0,000			
Const(10)	-1,54077	0,380095	-4,05	0,000			
Const(11)	-1,41694	0,368585	-3,84	0,000			
Const(12)	-1,30062	0,358787	-3,63	0,000			
Const(13)	-1,18885	0,350266	-3,39	0,001			
Const(14)	-1,08145	0,342881	-3,15	0,002			
Const(15)	-0,978243	0,336509	-2,91	0,004			
Const(16)	-0,879368	0,331057	-2,66	0,008			
Const(17)	-0,785856	0,326481	-2,41	0,016			
Const(18)	-0,695300	0,322581	-2,16	0,031			
Const(19)	-0,603366	0,319151	-1,89	0,059			
Const(20)	-0,511344	0,316250	-1,62	0,106			
Const(21)	-0,421326	0,313925	-1,34	0,180			

Const(22)	-0,330257	0,312087	-1,06	0,290			
Const(23)	-0,235910	0,310729	-0,76	0,448			
Const(24)	-0,140368	0,309919	-0,45	0,651			
Const(25)	-0,0452804	0,309679	-0,15	0,884			
Const(26)	0,0520719	0,310021	0,17	0,867			
Const(27)	0,149737	0,310962	0,48	0,630			
Const(28)	0,244092	0,312443	0,78	0,435			
Const(29)	0,337526	0,314464	1,07	0,283			
Const(30)	0,434530	0,317151	1,37	0,171			
Const(31)	0,535582	0,320588	1,67	0,095			
Const(32)	0,640728	0,324861	1,97	0,049			
Const(33)	0,750092	0,330061	2,27	0,023			
Const(34)	0,862664	0,336224	2,57	0,010			
Const(35)	0,979455	0,343493	2,85	0,004			
Const(36)	1,10088	0,352001	3,13	0,002			
Const(37)	1,22968	0,362096	3,40	0,001			
Const(38)	1,36717	0,374102	3,65	0,000			
Const(39)	1,51680	0,388635	3,90	0,000			
Const(40)	1,68440	0,406774	4,14	0,000			
Const(41)	1,87394	0,429737	4,36	0,000			
Const(42)	2,09267	0,459641	4,55	0,000			
Const(43)	2,34385	0,498844	4,70	0,000			
Const(44)	2,65128	0,554755	4,78	0,000			
Const(45)	3,03475	0,638896	4,75	0,000			
Const(46)	3,52813	0,776542	4,54	0,000			
Const(47)	4,28347	1,07431	3,99	0,000			
CC60	-0,569538	0,806438	-0,71	0,480	0,57	0,12	2,75
GE60	0,443083	0,747635	0,59	0,553	1,56	0,36	6,74
PS60	1,09956	0,474693	2,32	0,021	3,00	1,18	7,61

Log-Likelihood = -181,738

Test that all slopes are zero: G = 8,159, DF = 3, P-Value = 0,043

Goodness-of-Fit Tests

Method	Chi-Square	DF	P

```
Pearson      2417,26  2206  0,001

Deviance      363,48  2206  1,000

Measures of Association:

(Between the Response Variable and Predicted Probabilities)


Pairs       Number  Percent  Summary Measures

Concordant     701     62,1  Somers' D              0,25

Discordant     417     37,0  Goodman-Kruskal Gamma  0,25

Ties            10      0,9  Kendall's Tau-a        0,25

Total         1128    100,0
```

Podemos observar que entre os métodos avaliados através do Minitab, o que apresenta o maior grau de acerto na distribuição das amostras é a Análise discriminante, de acordo com os resultados abaixo:

Análise Discriminante : 100 %

Regressão Logística: 62,1%

Dessa forma podemos concluir que entre as duas análises, para esse caso a Análise discriminante é mais conveniente para a formação dos grupos pois de acordo com os resultados, a probabilidade de êxito é maior.

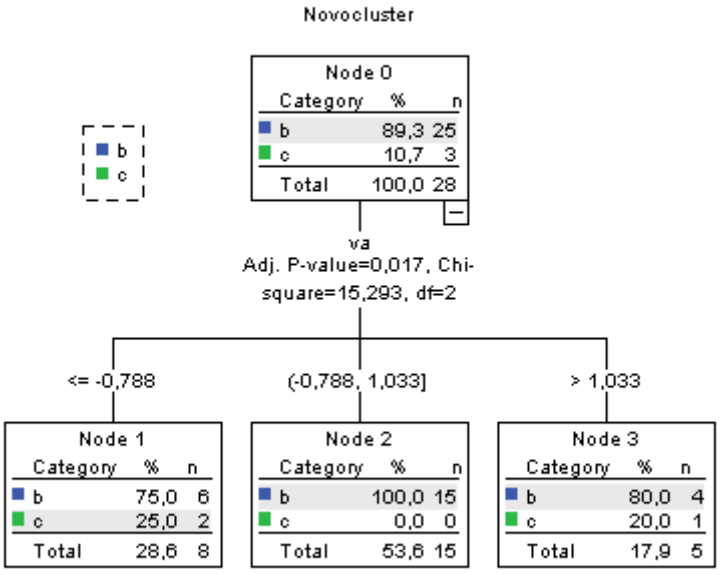
4. ÁRVORE DE CLASSIFICAÇÃO

Para a análise de árvore de classificação, utilizaremos como variável dependente os novos clusters e os dados extraídos do WGI serão as variáveis independentes, sem forçar a primeira variável de análise para ramificação, com o método de crescimento CHAID.

Classification Tree

Model Summary

Specifications	Growing Method	CHAID
	Dependent Variable	Novocluster
	Independent Variables	va, ps, gge, rq, rl, cc
	Validation	None
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	5
	Minimum Cases in Child Node	3
Results	Independent Variables Included	va
	Number of Nodes	4
	Number of Terminal Nodes	3
	Depth	1



Risk

Estimate	Std. Error
,250	,082

Growing Method: CHAID

Dependent Variable: Novocluster

Classification

Observed	Predicted		
	b	c	Percent Correct
b	19	6	76,0%
c	1	2	66,7%
Overall Percentage	71,4%	28,6%	75,0%

Growing Method: CHAID

Dependent Variable: Novocluster

5. CONSIDERAÇÕES FINAIS

Ao analisarmos os resultados obtidos através da Análise discriminante e a Regressão Logística utilizando-se o Minitab e se compararmos esses resultados com a análise de Árvore de classificação através do SPSS, observamos uma nova redistribuição, a partir do critério de diferenciação. Os resultados obtidos demonstram que mesmo com o uso da análise de Árvore de classificação, entre as três análises realizadas, a mais indicada nesse caso ainda é **Análise Discriminante** por obter uma resultado com 100%.

Observamos, porém que a árvore produziu poucos ramos, isso ocorreu, porém provavelmente devido ao número reduzido de amostras analisados nesse estudo, porém para efeito didático e conhecimento da ferramenta esse trabalho mostrou-se bastante proveitoso.