Módulo 19 - Análise Discriminante

Análise de Tabelas

Inicialmente, vamos relembrar a situação problema

Situação Problema

Um banco deseja classificar seus clientes de acordo com seu perfil de investimento: investidor conservador, investidor moderado ou investidor arrojado. Sabe-se que este perfil depende da renda do investidor (mensal em reais), do valor de seus investimentos (total acumulado em reais) e do tempo que ele é cliente do banco (em meses). O investidor conservador é aquele avesso aos riscos, que investe em modalidades de baixo risco, mesmo que estas ofereçam baixa remuneração ao capital; os moderados são aqueles que não são avessos ao risco, mas também não privilegiam apenas investimentos de risco elevado, fazem alguns investimentos de risco elevado e outros de baixo risco, ou preferem investir em modalidades de risco médio; os arrojados são os investidores que possuem pouca aversão ao risco, preferem investimentos arriscados com alta remuneração do capital.

O banco, então, montou um banco de dados com estas informações de 60 clientes, e deseja, a partir deles obter uma equação que permita classificar casos futuros nas categorias propostas (investidor conservador, moderador e arrojado).

A técnica indicada para resolver este problema é a análise discriminante.

Parte 2 – Análise dos resultados

Analysis Case Processing Summary

Unweighted Cases		N	Percent
Valid		42	70,0
Excluded	Missing or out-of-range group codes	0	,0
	At least one missing discriminating variable	0	,0
	Both missing or out-of-range group codes and at least one missing discriminating variable	0	,0
	Unselected	18	30,0
	Total	18	30,0
Total		60	100,0

Tabela 1: Informações sobre os casos da amostra.

A tabela 1 mostra os casos válidos da análise, que são utilizados para a amostra de construção. Neste caso, 70% (42 clientes do banco) fazem parte da amostra de construção.

Group Statistics

STATUS		Mean	Std. Deviation	Valid N (listwise) Unweighted	Valid N (listwise) Weighted
conservador	renda individual declarada	18864,29	7514,0491	14	14,000
	valor total dos investimentos no banco	17907,14	13050,6978	14	14,000
	tempo em que é cliente do banco	30,3571	12,5122	14	14,000
moderado	renda individual declarada	13242,86	5123,9579	14	14,000
	valor total dos investimentos no banco	13728,57	9924,7498	14	14,000
	tempo em que é cliente do banco	16,7143	9,4741	14	14,000
arrojado	renda individual declarada	14564,29	8468,7305	14	14,000
	valor total dos investimentos no banco	14850,00	17702,4444	14	14,000
	tempo em que é cliente do banco	10,9286	4,6652	14	14,000
Total	renda individual declarada	15557,14	7407,2523	42	42,000
	valor total dos investimentos no banco	15495,24	13703,7818	42	42,000
	tempo em que é cliente do banco	19,3333	12,3677	42	42,000

Tabela 2: Estatísticas descritivas de cada grupo.

A tabela 2 mostra:

“Mean”: média de cada categoria para cada variável independente. Observe que os conservadores possuem renda média individual mais elevada, maior valor médio total de investimentos no banco e são clientes do banco há 30 meses em média. Os arrojados são clientes do banco a menos tempo que os moderados e possuem valor total dos investimentos mais elevado;

“Std. deviation”: desvio padrão para cada categoria em cada variável. Não basta analisar as médias, é importante verificar se a variabilidade dos dados pode ser considerada pequena ou grande, para tanto, podemos calcular o coeficiente de variação dado pela razão desvio- padrão/ média;

“Valid N”: dentre os casos válidos quantos fazem parte de cada categoria, por exemplo, 14 são conservadores, 14 são moderados e 14 são arrojados.

Tests of Equality of Group Means

	Wilks' Lambda	F	df1	df2	Sig.
renda individual declarada	,892	2,350	2	39	,109
valor total dos investimentos no banco	,983	,337	2	39	,716
tempo em que é cliente do banco	,556	15,591	2	39	,000

Tabela 3: Teste de igualdade de médias.

A tabela 3 mostra testes de igualdade de médias para as categorias:

a) Renda individual declarada

H_0: a renda média dos grupos de investidores é igual;

H_1: existe diferença na renda média em pelo menos um grupo de investidores.

Considerando um nível de significância $\alpha = 5\%$, temos o nível de significância observado $sig = 10{,}9\% > 5\%$, então não podemos rejeitar a hipótese nula, ou seja, não existe diferença significativa na renda média dos grupos .

b) Valor total dos investimentos no banco

H_0: o valor médio dos investimentos dos grupos de investidores é igual;

H_1: existe diferença no valor médio dos investimentos em pelo menos um grupo de investidores.

Considerando um nível de significância $\alpha = 5\%$, temos o nível de significância observado $sig = 71{,}6\% > 5\%$, então não podemos rejeitar a hipótese nula, ou seja, não existe diferença significativa no valor médio dos investimentos entre os grupos.

c) Tempo em que é cliente do banco

H_0: o tempo médio em que são clientes do banco é igual para os grupos;

H_1: existe diferença no tempo médio em que são clientes do banco em pelo menos um grupo de investidores.

Considerando um nível de significância $\alpha = 5\%$, temos o nível de significância observado $sig = 0\% < 5\%$, então rejeitamos a hipótese nula, ou seja, existe diferença significativa no tempo médio em que são clientes do banco para pelo menos dois dos três grupos.

A estatística Wilk's Lambda é também chamada de estatística U. Seu valor varia entre 0 e 1 e valores próximos de 0 indicam que a variabilidade dentro dos grupos é menor que a variabilidade total, ou seja, muito da variabilidade pode ser atribuída às diferenças entre as médias dos grupos; valores próximos de 1 indicam que as médias dos grupos são iguais.
Neste caso, temos para a variável "renda individual" U=0,892, para "investimentos" U=0,983 e para "tempo de cliente" U=0,556. Ao que parece, os grupos possuem médias iguais na variável "investimento".

Pooled Within-Groups Matrices

		renda individual declarada	valor total dos investimentos no banco	tempo em que é cliente do banco
Correlation	renda individual declarada	1,000	,599	,153
	valor total dos investimentos no banco	,599	1,000	,365
	tempo em que é cliente do banco	,153	,365	1,000

Tabela 4: Matriz de correlações pooled entre as variáveis.

A tabela 4 mostra as correlações entre as variáveis independentes. Observe que a correlação só é perfeita entre a variável com ela própria. Observamos que a renda individual tem correlação de 0,599 (média) com o valor dos investimentos e de 0,153 (baixa) com o tempo de cliente; o valor dos investimentos possui correlação de 0,365 (baixa) com o tempo de cliente.

Log Determinants

STATUS	Rank	Log Determinant
conservador	3	41,375
moderado	3	39,214
arrojado	3	39,973
Pooled within-groups	3	40,738

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Tabela 5: Determinantes.

Test Results

Box's M		21,478
F	Approx.	1,588
	df1	12
	df2	7371,000
	Sig.	,087

Tests null hypothesis of equal population covariance matrices.

Tabela 6: Resultados do teste Box M.

A tabela 6 mostra o teste Box M's que testa o pressuposto de igualdade das matrizes de covariância.

H_0: as matrizes de covariância das populações são iguais;

H_1: as matrizes de covariância são diferentes.

Considerando um nível de significância $\alpha = 5\%$, temos o nível de significância observado $sig = 8{,}7\% > 5\%$, então não podemos rejeitar a hipótese nula, ou seja, não existe diferença significativa entre as matrizes de covariância das populações.

Eigenvalues

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	,995[a]	97,1	97,1	,706
2	,030[a]	2,9	100,0	,170

a. First 2 canonical discriminant functions were used in the analysis.

Tabela 7: Matriz de autovalores.

A tabela 7 mostra os autovalores das funções. A função 1 explica 97,1% da variância.

Wilks' Lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 2	,487	27,369	6	,000
2	,971	1,118	2	,572

Tabela 8: Resultados teste Wilk´s Lambda.

A tabela 8 apresenta o teste U para as funções discriminantes. Como a variável dependente possui três categorias, foram criadas duas funções discriminantes. O teste das funções "1 through 2" apresenta o teste de significância das duas funções ao mesmo tempo. O p- valor encontrado, $sig = 0\% < \alpha = 5\%$, indica que as duas funções conjuntamente conseguem diferenciar os grupos. O teste das funções "2" (última linha da tabela) apresenta o teste de significância para a função 2 separadamente. O p- valor encontrado, $sig = 57{,}2\% > \alpha = 5\%$,indica que a função 2 não consegue classificar os casos quando considerada sozinha.

Standardized Canonical Discriminant Function Coefficients

	Function	
	1	2
renda individual declarada	,480	1,051
valor total dos investimentos no banco	-,545	-,130
tempo em que é cliente do banco	1,021	-,346

Tabela 9: Coeficientes padronizados da função discriminante.

A tabela 9 mostra as variáveis que mais influenciam o status do investidor.

Na função 1 temos que "o tempo em que é cliente do banco" é a variável que mais influencia (possui o maior valor) o status do investidor, seguida pela "renda individual declarada". Na função 2 a renda individual é a que mais influencia o status do investidor seguida pelo tempo em que é cliente do banco.

Structure Matrix

	Function	
	1	2
tempo em que é cliente do banco	,895*	-,233
renda individual declarada	,309	,920*
valor total dos investimentos no banco	,115	,373*

Pooled within-groups correlations between discriminating variables and standardized canonical discriminant functions
Variables ordered by absolute size of correlation within function.

*. Largest absolute correlation between each variable and any discriminant function

Tabela 10: Matriz estrutura.

A tabela 10 mostra as correlações de cada variável com as funções discriminantes. Percebemos que a variável "tempo em que é cliente do banco" possui maior correlação com a função 1 e as variáveis "renda individual" e "valor total dos investimentos no banco" possuem maior correlação com a função 2. As variáveis com maior correlação são as que possuem poder discriminatório naquela função. Temos, assim, que a função 1 mede o tempo em que o investidor é cliente do banco e a função 2 mede a renda e o valor dos investimentos.

Canonical Discriminant Function Coefficients

	Function	
	1	2
renda individual declarada	,000	,000
valor total dos investimentos no banco	,000	,000
tempo em que é cliente do banco	,108	-,037
(Constant)	-2,522	-1,426

Unstandardized coefficients

Tabela 11: Coeficientes das funções discriminantes não padronizados.

A tabela 11 apresenta os coeficientes das funções discriminantes.Temos:

$D_1 = -2{,}522 + 0{,}0001 renda + 0{,}0001 investimento + 0{,}108 tempo$;

$D_2 = -1{,}426 + 0{,}0001 renda + 0{,}0001 investimento - 0{,}037 tempo$.

Os casos futuros serão classificados de acordo com estas funções. Utilizam-se os coeficientes não padronizados, pois os valores das variáveis são não padronizados.

Functions at Group Centroids

	Function	
STATUS	1	2
conservador	1,318	5,805E-02
moderado	-,369	-,227
arrojado	-,949	,169

Unstandardized canonical discriminant functions evaluated at group means

Tabela 12: Centróides dos grupos.

A tabela 12 mostra os centróides dos grupos em cada uma das funções discriminantes. A classificação é feita comparando-se os valores obtidos pela função 1 e função 2 quando os valores são substituídos em relação aos valores dos centróides de cada grupo em cada função (1 e 2). Uma forma de classificar os casos é analisando o mapa territorial (gráfico 4), apresentado mais adiante.

Classification Processing Summary

Processed		60
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		60

Tabela 13: Casos utilizados na análise.

A tabela 13 mostra os casos processados na análise. Nossa amostra possui 60 casos e todos foram considerados na análise.

Prior Probabilities for Groups

STATUS	Prior	Cases Used in Analysis	
		Unweighted	Weighted
conservador	,333	14	14,000
moderado	,333	14	14,000
arrojado	,333	14	14,000
Total	1,000	42	42,000

Tabela 14: Tabela de probabilidades de ocorrência para os grupos.

A tabela 14 mostra a probabilidade de os casos serem classificados em cada grupo. Como temos três categorias e a probabilidade de um caso ser classificado em cada um deles é igual, temos uma probabilidade de 33,333% de um caso ser classificado em cada grupo. Em outras palavras, estamos considerando que as probabilidades são iguais nos três grupos (conservadores, moderados e arrojados).

Classification Function Coefficients

	STATUS		
	conservador	moderado	arrojado
renda individual declarada	4,528E-04	2,984E-04	3,175E-04
valor total dos investimentos no banco	-1,368E-04	-6,815E-05	-4,91E-05
tempo em que é cliente do banco	,361	,189	,112
(Constant)	-9,622	-4,187	-3,657

Fisher's linear discriminant functions

Tabela 15: Coeficientes que compõem as funções discriminantes.

A tabela 15 mostra as funções discriminantes de Fisher para cada grupo, ou seja, cada categoria tem uma relação linear com as variáveis do estudo. Temos:

$$Dconservador = 0{,}0004528renda - 0{,}0001368investimento + 0{,}361tempo - 9{,}622$$

$$D\,\mathrm{mod}\,erado = 0{,}0002984renda - 0{,}00006815investimento + 0{,}189tempo - 4{,}187$$

$$Darrojado = 0{,}0003175renda - 0{,}0000491investimento + 0{,}112tempo - 3{,}657$$

Diferentemente das funções discriminantes com coeficientes não padronizados, neste caso, classifica-se um determinado caso de acordo com o maior valor em uma dessas três funções discriminantes.

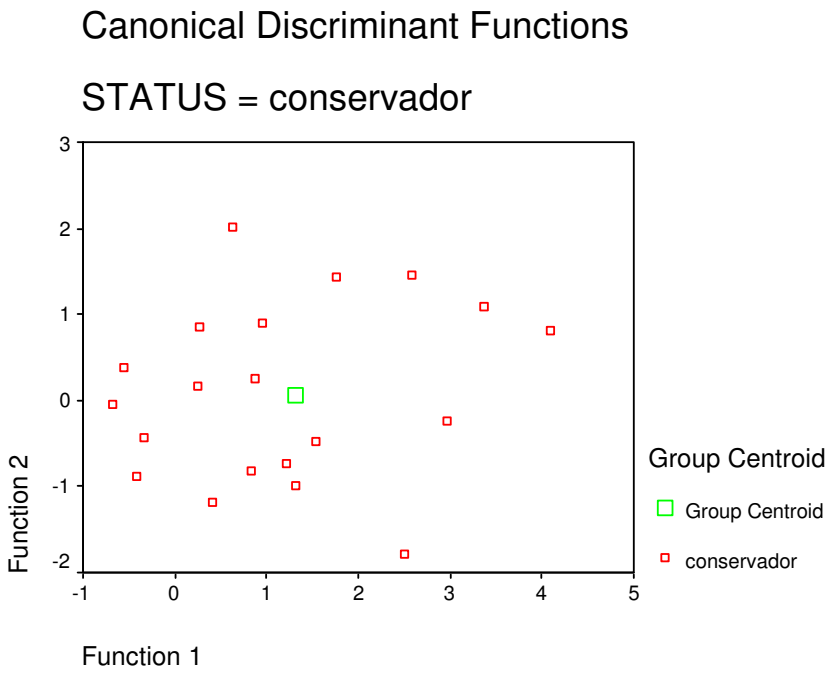


Gráfico 1: Categoria- Conservadores.

O gráfico 1 mostra a dispersão dos casos conservadores e o centróide do grupo.

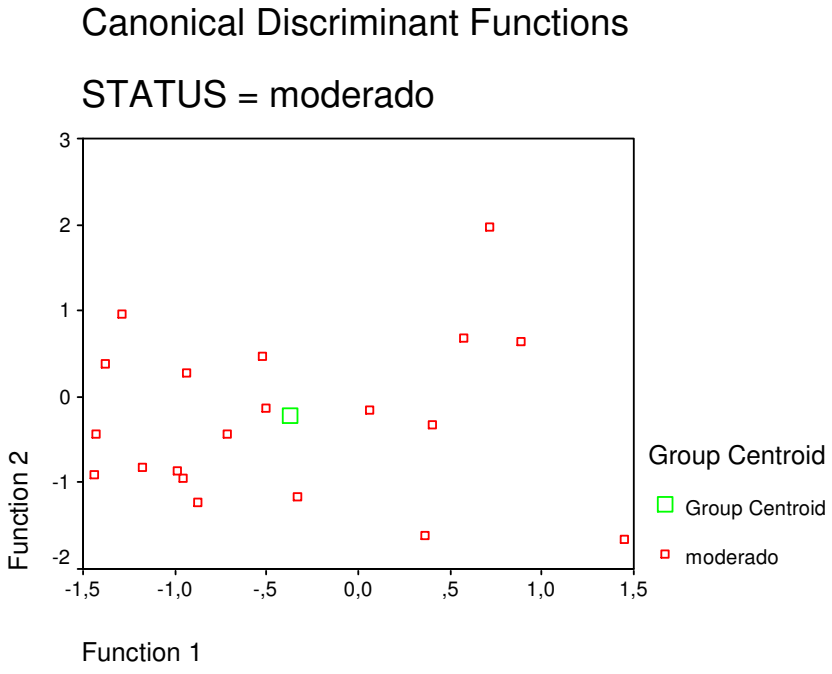


Gráfico 2: Categoria- Moderados.

O gráfico 2 mostra a dispersão dos casos moderados e o centróide dos grupos.

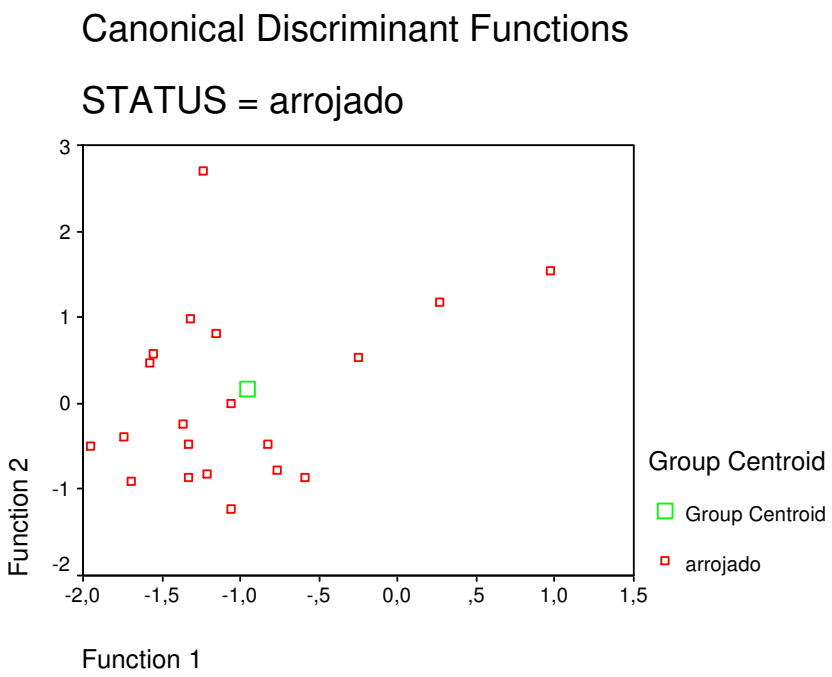
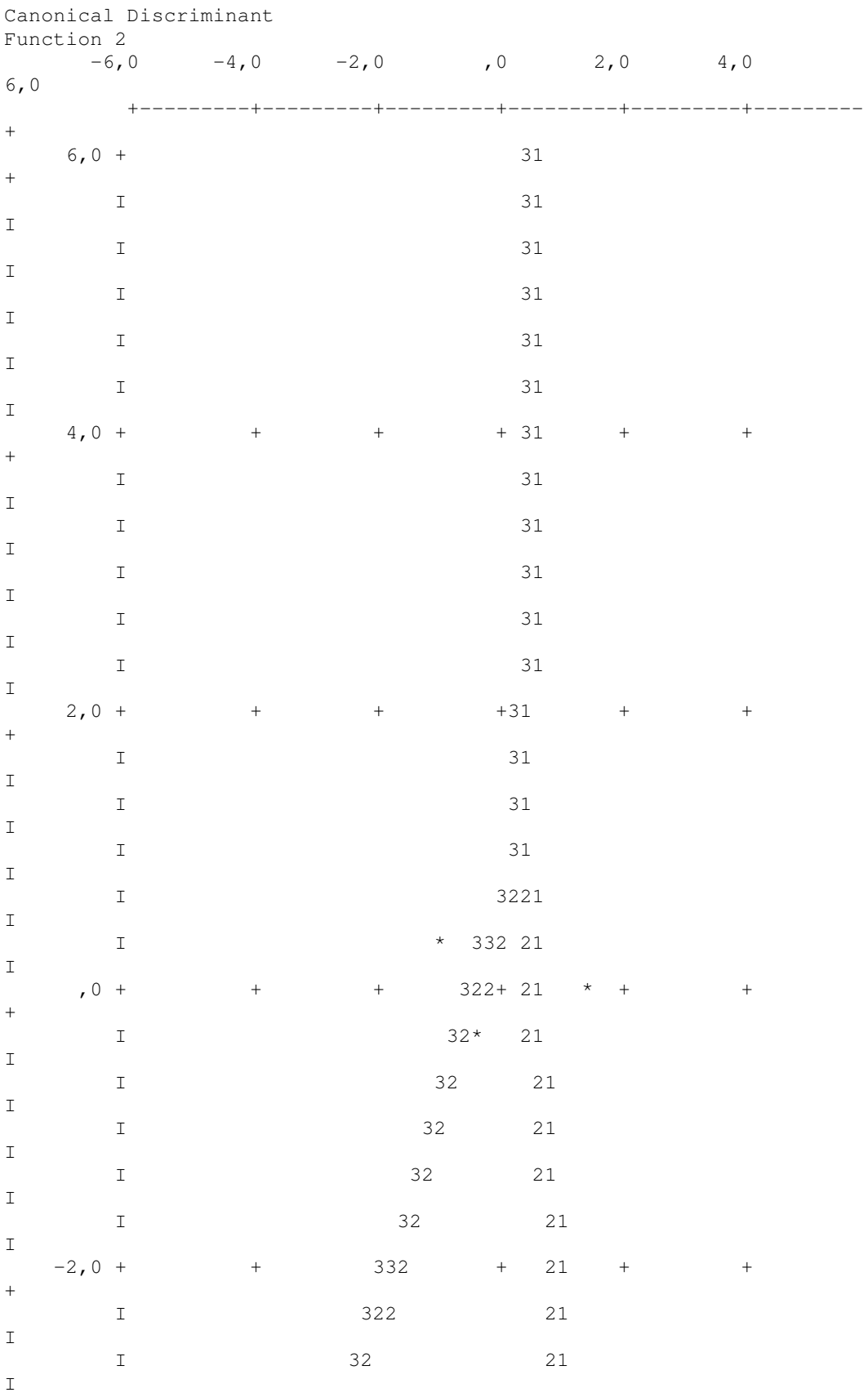
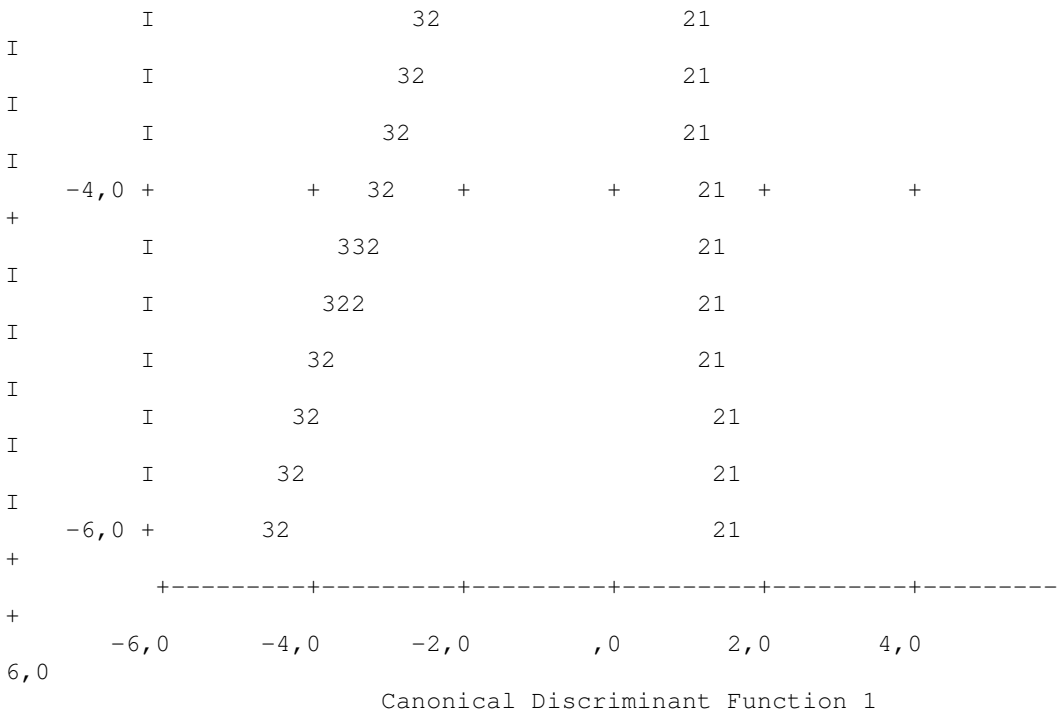


Gráfico 3: Categoria- Arrojados.

O gráfico 3 mostra a dispersão dos casos arrojados e o centróide do grupo.

Territorial Map


Canonical Discriminant
Function 2
-6,0
-4,0
-2,0
,0
2,0
4,0
6,0
6,0
4,0
2,0
,0
-2,0



Symbols used in territorial map

Symbol	Group	Label
1	0	conservador
2	1	moderado
3	2	arrojado
*		Indicates a group centroid

Gráfico 4: Mapa territorial.

O gráfico 4 apresenta o mapa territorial para os três grupos. Apresenta graficamente a disposição dos grupos em relação às funções discriminantes e o centróide de cada grupo.

Classification Results[b,c,d]

				Predicted Group Membership			
			STATUS	conservador	moderado	arrojado	Total
Cases Selected	Original	Count	conservador	10	3	1	14
			moderado	3	7	4	14
			arrojado	1	5	8	14
		%	conservador	71,4	21,4	7,1	100,0
			moderado	21,4	50,0	28,6	100,0
			arrojado	7,1	35,7	57,1	100,0
	Cross-validated[a]	Count	conservador	10	3	1	14
			moderado	3	7	4	14
			arrojado	1	7	6	14
		%	conservador	71,4	21,4	7,1	100,0
			moderado	21,4	50,0	28,6	100,0
			arrojado	7,1	50,0	42,9	100,0
Cases Not Selected	Original	Count	conservador	3	2	1	6
			moderado	1	3	2	6
			arrojado	1	0	5	6
		%	conservador	50,0	33,3	16,7	100,0
			moderado	16,7	50,0	33,3	100,0
			arrojado	16,7	,0	83,3	100,0

a. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case.

b. 59,5% of selected original grouped cases correctly classified.

c. 61,1% of unselected original grouped cases correctly classified.

d. 54,8% of selected cross-validated grouped cases correctly classified.

Tabela 16: Resultados da classificação feitas pela ferramenta.

A tabela 16 mostra a efetividade do modelo, isto é, quanto de acerto o modelo obteve em suas classificações. Duas análises são feitas: 1) classificação com a amostra de construção e 2) classificação com a amostra de validação.

1) Classificação com a amostra de construção.

Neste caso, o modelo conseguiu acertar a classificação em 59,5% dos casos (observação "b" da tabela). Para que possamos considerar que o modelo tenha uma classificação razoável, no mínimo é necessário que esta porcentagem de acerto seja maior que a probabilidade de o caso ser classificado em cada categoria mais 25%. Temos $59{,}5\% > 33{,}3333\% + 25\%$ *ou* $59{,}5\% > 58{,}3333\%$, portanto o modelo possui uma classificação razoável.

2) Classificação com a amostra de validação

Segue o mesmo procedimento realizado no item 1, mas considerando a porcentagem mostrada na observação "c" da tabela. Temos $61{,}1\% > 33{,}3333\% + 25\%$ *ou* $61{,}1\% > 58{,}3333\%$, portanto o modelo possui uma classificação razoável.