

PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO
Programa de Estudos Pós-Graduados em Administração

CLASSIFICAÇÃO DO BRASIL

**Focando principalmente indicadores relacionados a habitação,
educação, trabalho e muito particularmente SAÚDE**

MÉTODOS QUANTITATIVOS NA PESQUISA EMPÍRICA

Professor: Dr. Arnaldo Jose de Hoyos

Elaine Pinheiro Palmeira

CAP I ANÁLISE DE CONGLOMERADOS

1. INTRODUÇÃO

O presente trabalho tem por objetivo efetuar uma análise comparativa de médias, intervalos de confiança e regressões dos dados da Pesquisa Firjan/FGV sobre o Desenvolvimento dos Municípios nos períodos de 2000 e 2010. Iniciamos com o entendimento dos dados, incluindo a definição dos indivíduos e das variáveis, suas classificações em variáveis categóricas ou quantitativas, os significados e unidades de medida, além da apresentação da tabela de dados.

Na sequência serão geradas análise comparativas dos dados de Saúde agrupado por Estado, excluindo o Distrito Federal por ter apenas um Município. Será calculada a Anova do ISDM e da Saúde por Estado, serão gerados vários gráficos com as diversas variáveis de Saúde. Comparando-se os resultados das médias por estado, poderemos agrupar as linhas de dados pelo nível de desigualdade dos fatores ISDM e Saúde.

Por fim, fazemos as considerações finais. O software estatístico utilizado é o **MINITAB16**.

2. ENTENDENDO OS DADOS

2.1 Os Indivíduos

Esta pesquisa ilustra dois rankings lançados no final de 2012, e chegaram a conclusões diferentes sobre quais cidades de maior desenvolvimento do país.

Os indivíduos desta análise são os 5565 municípios brasileiros. Os dados analíticos foram extraídos do IBGE, e possibilitam uma comparação entre os dados colhidos em 2000 com 2010.

2.2 As Variáveis

As variáveis desta pesquisa incluem os 3 principais índices sintéticos que são ISDM, IFDM e IFDF, que são médias ponderadas dos dados analíticos globais da pesquisa, e variáveis analíticas, referente à educação do ensino pré escola, fundamental e médio.

Tabela 1. Comparativo entre as Variáveis ISDM e IFDM

O QUE O ISDM (FGV) MEDE	O QUE O IFDM (Firjan) MEDE
Educação: taxa de analfabetismo e taxa de crianças e jovens que frequentam a escola em cada etapa, desempenho na Prova Brasil (MEC)	Educação: taxa de matrícula infantil, abandono, distorção idade-série, desempenho no Ideb, taxa de docentes com ensino superior
Saúde e Segurança: taxa de mortalidade infantil, gravidez precoce e mortalidade por causas evitáveis; homicídios	Saúde: número de consultas pré-natal, óbitos por causa mal definidas e óbitos infantis evitáveis
Renda: presença de pobreza e extrema pobreza	Emprego e renda: geração, estoque e salários médios dos empregos formais
Trabalho: taxa de ocupação e formalização	
Habitação: coleta de lixo, energia elétrica, água canalizada, esgotamento sanitário, domicílio próprio	

Tabela 2. A definição das Variáveis

VARIÁVEL	SIGNIFICADO	TIPO	UNIDADE DE MEDIDA
Município	Nome dos municípios	Variável categórica	N/A
Indicador Social de Desenvolvimento dos Municípios (ISDM)	Média ponderada dos indicadores das dimensões Habitação, Renda, Trabalho, Saúde e Segurança e Educação (H, R, T, S e E) padronizada pela média do Brasil.	Variável Quantitativa	Numérico
IFDM	Índice Firjan de Desenvolvimento Municipal – IFDM. O índice é resultado da média ponderada de diferentes indicadores: Emprego & Renda, Educação e Saúde.	Variável Quantitativa	Numérico
IFGF	Composto por cinco indicadores: Receita Própria, Pessoal, Investimentos, Liquidez e Custo da Dívida, o índice tem como base de dados as estatísticas oficiais disponibilizadas anualmente pela Secretaria do Tesouro Nacional, constituídas por informações orçamentárias e patrimoniais prestadas pelos próprios municípios.	Variável Quantitativa	Numérico
Indicador da dimensão Saúde e Segurança (S)	Média ponderada dos indicadores da dimensão Saúde e Segurança (S1_1, S1_2, S1_3, S2_1, S2_2 e S3_1) padronizada pela média do Brasil.	Variável Quantitativa	Numérico
Taxa de mortalidade infantil, por mil nascidos vivos. (S1_1)	Taxa de sobrevivência infantil no primeiro ano de vida, representada pela diferença entre o número de nascidos vivos e o número de óbitos até um ano de idade.	Variável Quantitativa	Numérico
Mortalidade proporcional por doenças com causas evitáveis de menores de 5 anos. (S1_2)	Percentual de óbitos de menores de cinco anos de idade por causas evitáveis, em relação ao total de óbitos com causa definida na população residente na faixa etária. Define-se como causas evitáveis de menores de cinco anos todas aquelas que sejam reduzíveis por: ações de imunização; adequada atenção à mulher na gestação; adequada atenção à mulher no parto; adequada atenção ao recém-nascido; ações de diagnóstico e tratamento	Variável Quantitativa	Numérico

	adequado; ou ações de promoção à saúde vinculadas a ações de atenção.		
Proporção de nascidos vivos com baixo peso ao nascer. (S1_3)		Variável Quantitativa	Numérico
Proporção das adolescentes (10 a 19 anos) que já tiveram filho. (S2_1)	Percentual das adolescentes de 10 a 19 anos que declaram nunca ter tido algum filho nascido vivo ou filho nascido morto.	Variável Quantitativa	Numérico
Mortalidade proporcional por doenças com causas evitáveis na população de 5 a 74 anos. (S2_2)		Variável Quantitativa	Numérico
Taxa de homicídio, por cem mil habitantes. (S3_1)	Percentual da população residente que foi vítima de homicídio. Define-se como homicídio os óbitos por causas externas correspondentes aos seguintes grandes grupos da Classificação Internacional de Doenças CID-10: X85-Y09 Agressões, Y35-Y36 Intervenções legais e operações de guerra.	Variável Quantitativa	Numérico

3. ANÁLISE DAS VARIÁVEIS

3.1 VARIÁVEIS CATEGÓRICAS

Este tipo de variável indica que o foco de concentração deve ser a análise de gráficos do tipo *pie chart* e barras.

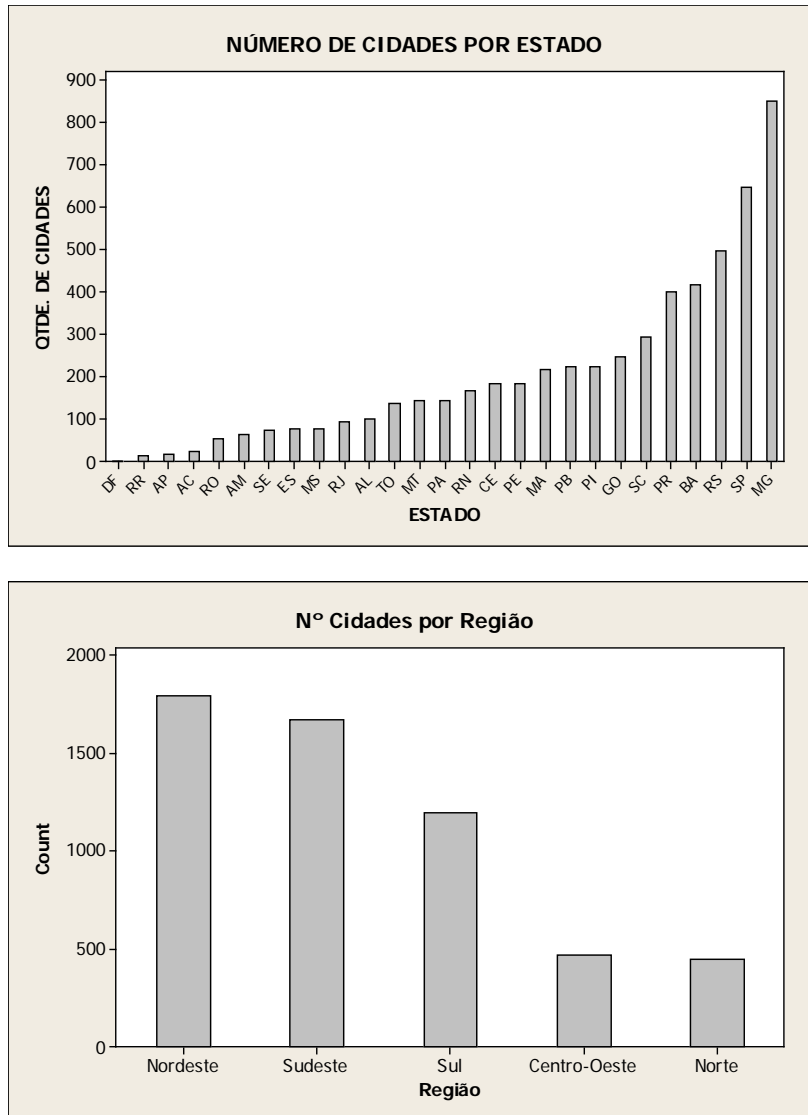
3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

Figura 3. Número de Cidades por Estado e Região do Brasil



Nos gráficos ao lado podemos ter uma dimensão do número de cidades por estado e por região. A região do Brasil com o maior número de cidades é a Nordeste (1790), seguida pela região Sudeste (1669) e pela região Sul, com 1191 cidades. As regiões com menor número de cidades

3.2 VARIÁVEIS QUANTITATIVAS

A análise deste tipo de variável permite a utilização de uma maior gama de ferramentas de análise como histogramas, curvas de densidade, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling.

3.2.1. DENDOGRAMA DE SAÚDE POR ESTADO (-DF)

O Dendograma permite uma análise do grau de similaridade dos dados para uma determinada variável. Em seguida geramos o Dendograma de Saúde por Estado.

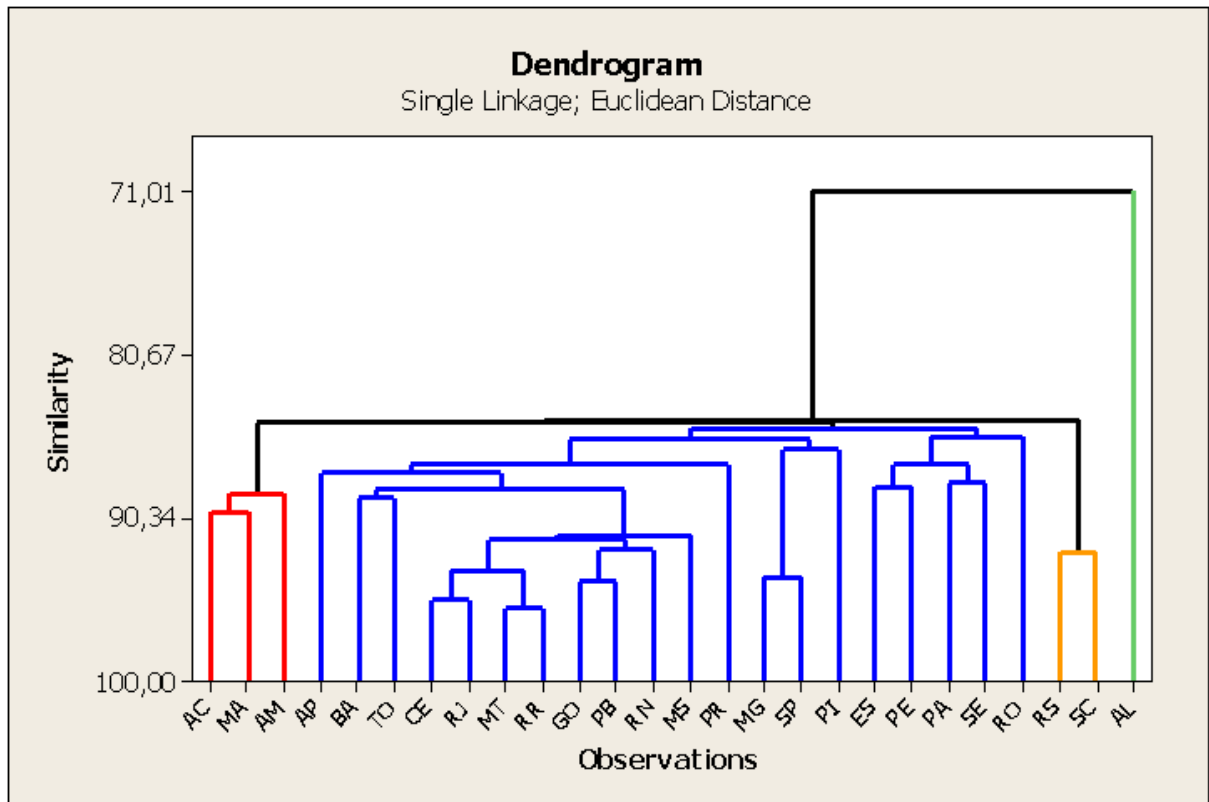


Figura 2. Dendograma da variável Saúde por estados do Brasil (classificação não supervisionada)

Na figura acima podemos verificar um grande grupo de variáveis, agrupadas pela similaridade dos dados, dois grupos menores e um estado isolado (Alagoas). Os estados que possuem maior similaridade são Mato Grosso e Roraima. O nível de similaridade dos dados de saúde destes estados está por volta de 71%, conforme indicado na escala apresentada no eixo Y do gráfico.

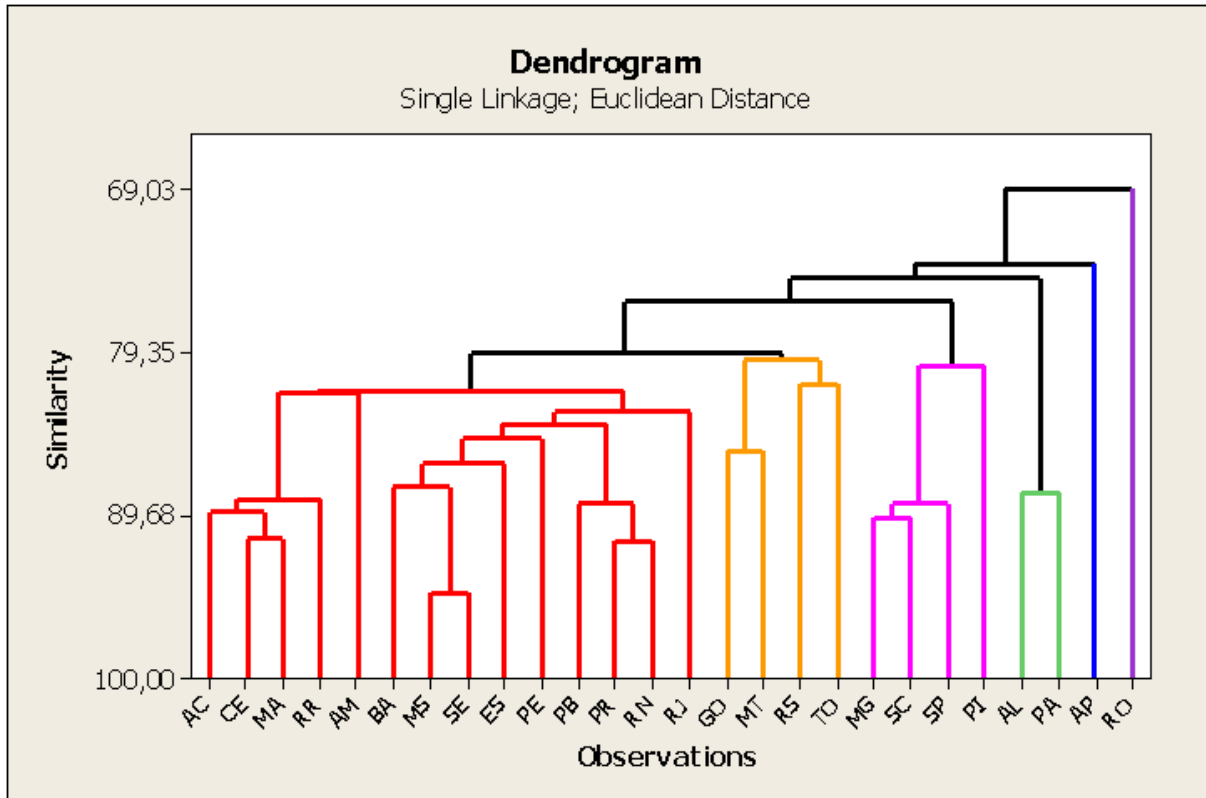


Figura 3. Dendrograma de desigualdade da Saúde dos municípios por estado

No gráfico acima, podemos verificar 6 grupos de dados (vermelho, amarelo, lilás e os outros 3), compostos pelos estados do Brasil, com dois estados, Amapá e Roraima, mais afastados da média de similaridade. No geral, os dados de Saúde possuem similaridade em nível médio, sendo que os estados cujos índices de saúde estão mais próximos são Mato Grosso do Sul e Sergipe, com quase 88% de similaridade. Na classificação não supervisionada não se tem informações prévias sobre estes grupos e não se tem informações sobre os por quês ou os critérios de agrupamento utilizados neste agrupamento.

Quando o nível de desigualdade é baixo poderíamos erroneamente dizer que a situação é boa, mas isso não é verdade. Vale ressaltar que baixa desigualdade não significa que as coisas vão bem, como é o caso da saúde no Brasil, que apresenta más condições generalizadas. Isto significa que existe um padrão nos municípios do estado em termos de saúde, com uma maior similaridade entre tais municípios, e não é possível responder se esta similaridade é boa ou não.

3.2.2. ANÁLISE DAS VARIÂNCIAS DE ISDM E EDUCAÇÃO POR ESTADO – DF

A análise das variâncias permite a verificação e visualização das médias e desvios padrões da variável a ser analisada. O gráfico BOXPLOT ilustra os agrupamentos, o seu tamanho varia de acordo com a quantidade de dados de cada grupo, e também é possível visualizar as ocorrências de *outliers* dentro de um grupo de dados.

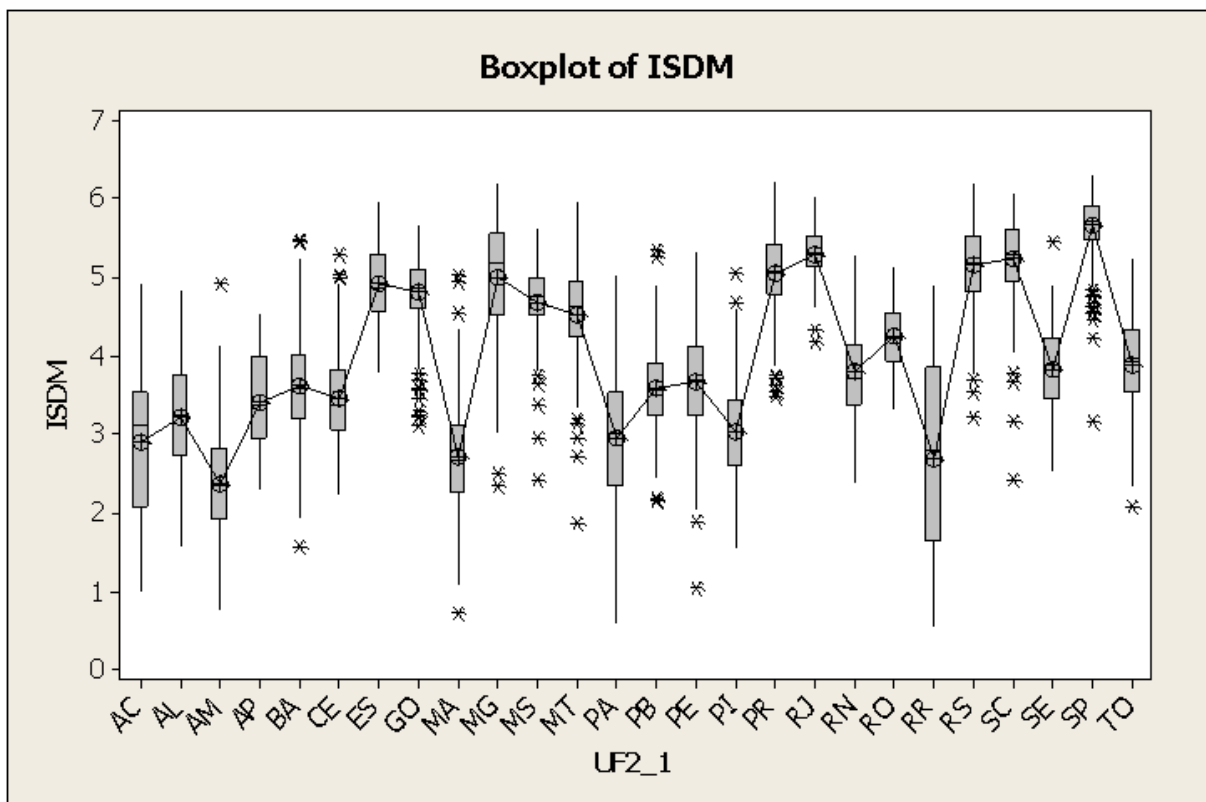


Figura 4. Gráfico BOXPLOT de ISDM por estado

Podemos visualizar na figura acima uma grande variabilidade sobre as médias de ISDM por estado, sendo que o estado com maior variabilidade dos dados é Roraima. Já São Paulo apresenta uma baixa variabilidade dos dados de ISDM, embora tenha muitos *outliers* que são os dados muito distantes das médias.

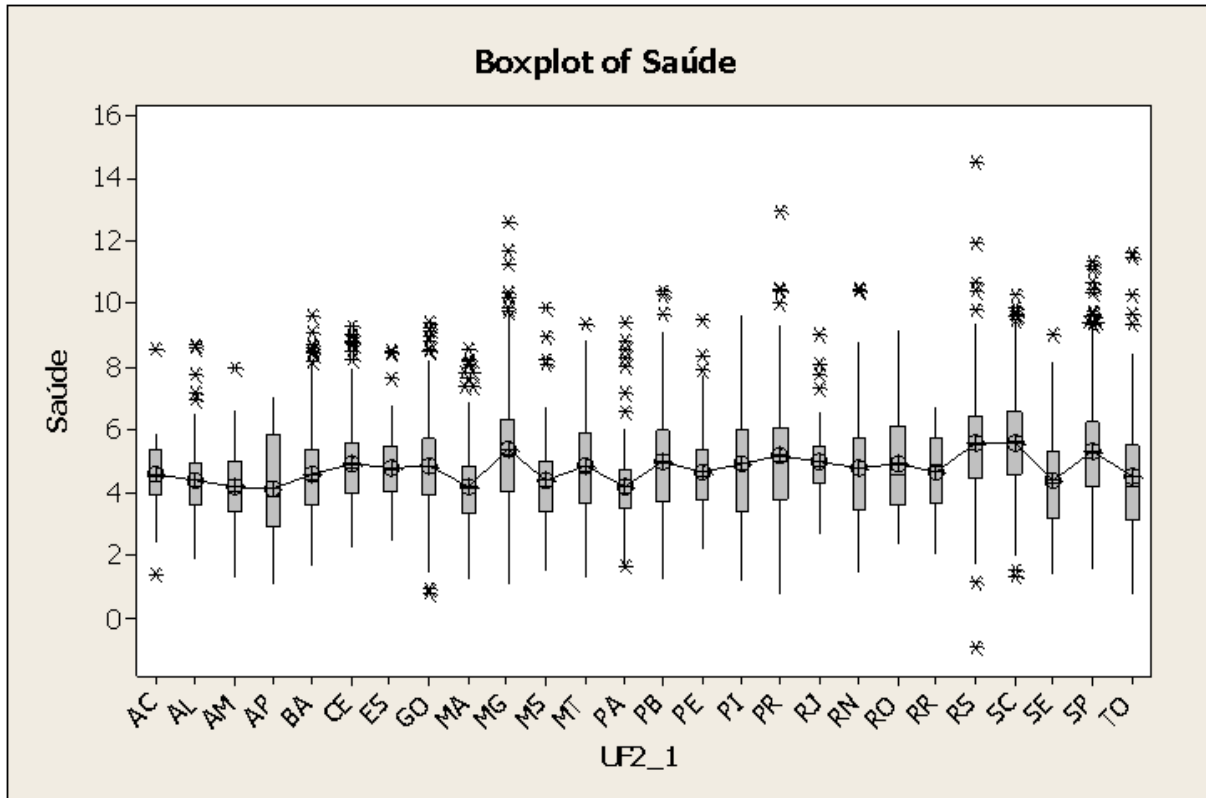


Figura 5. Gráfico BOXPLOT de saúde por estado

A partir da análise do gráfico acima é possível dizer que não existe uma variação grande entre as médias dos estados do Brasil, no que diz respeito à saúde, e também pelo tamanho das caixas de cada estado é possível afirmar que a variância desses dados de saúde não é tão expressiva, embora haja muitos *outliers*, isto é, dados muito acima ou abaixo da média dos dados do estado. O estado que apresenta a maior média de saúde é Santa Catarina (5,60), e o estado que apresenta a menor média é Sergipe, com média de 4,42.

Abaixo podemos visualizar os dados descritivos gerados pelo comando, para a variável ISDM.

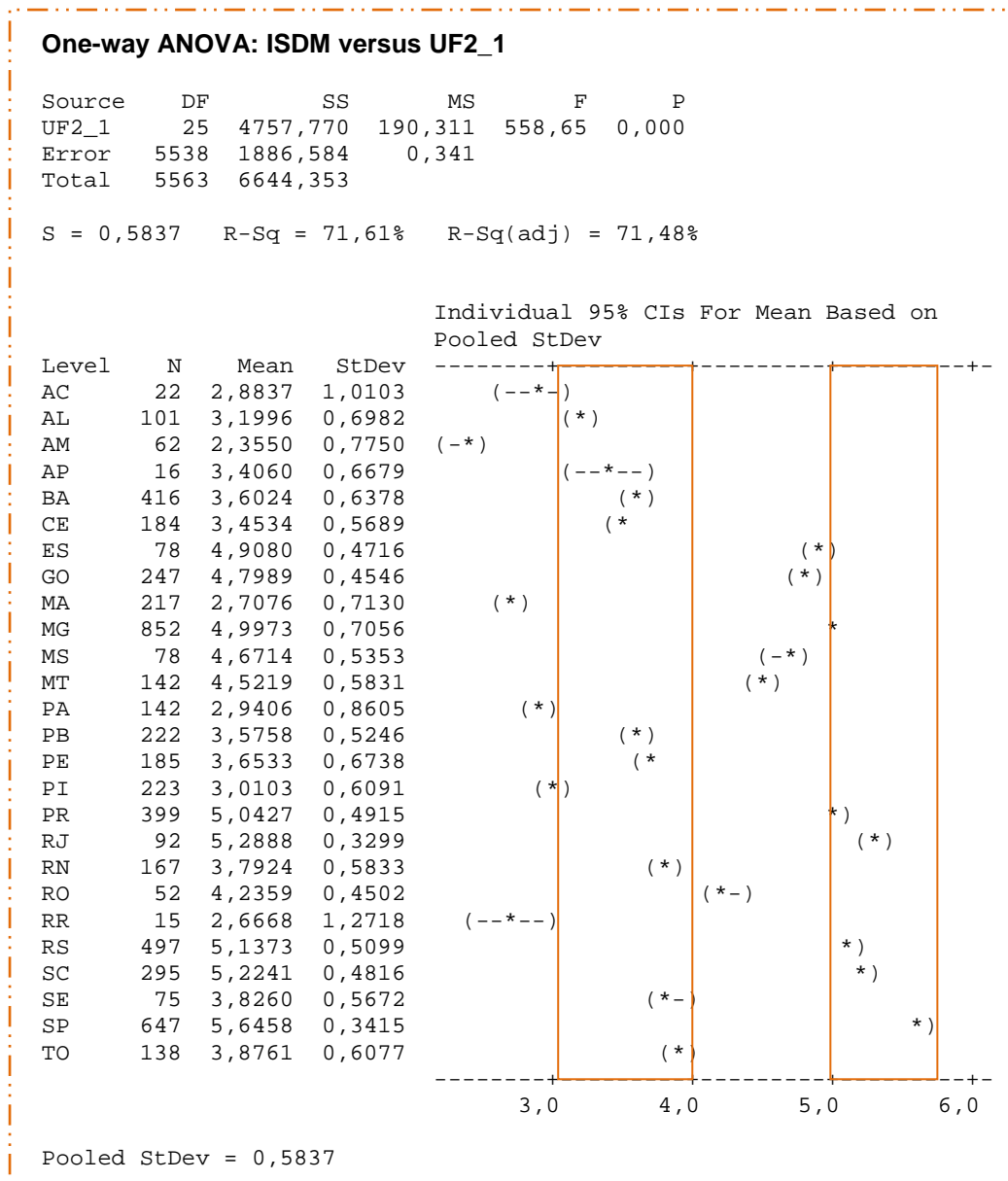


Figura 6. Análise de variância de ISDM por estado

É possível separar os dados das médias de cada estado por quartil. Desta forma teríamos 4 tipos de regiões no Brasil, separados pelos Índices de Desenvolvimento de seus Municípios (ISDM).

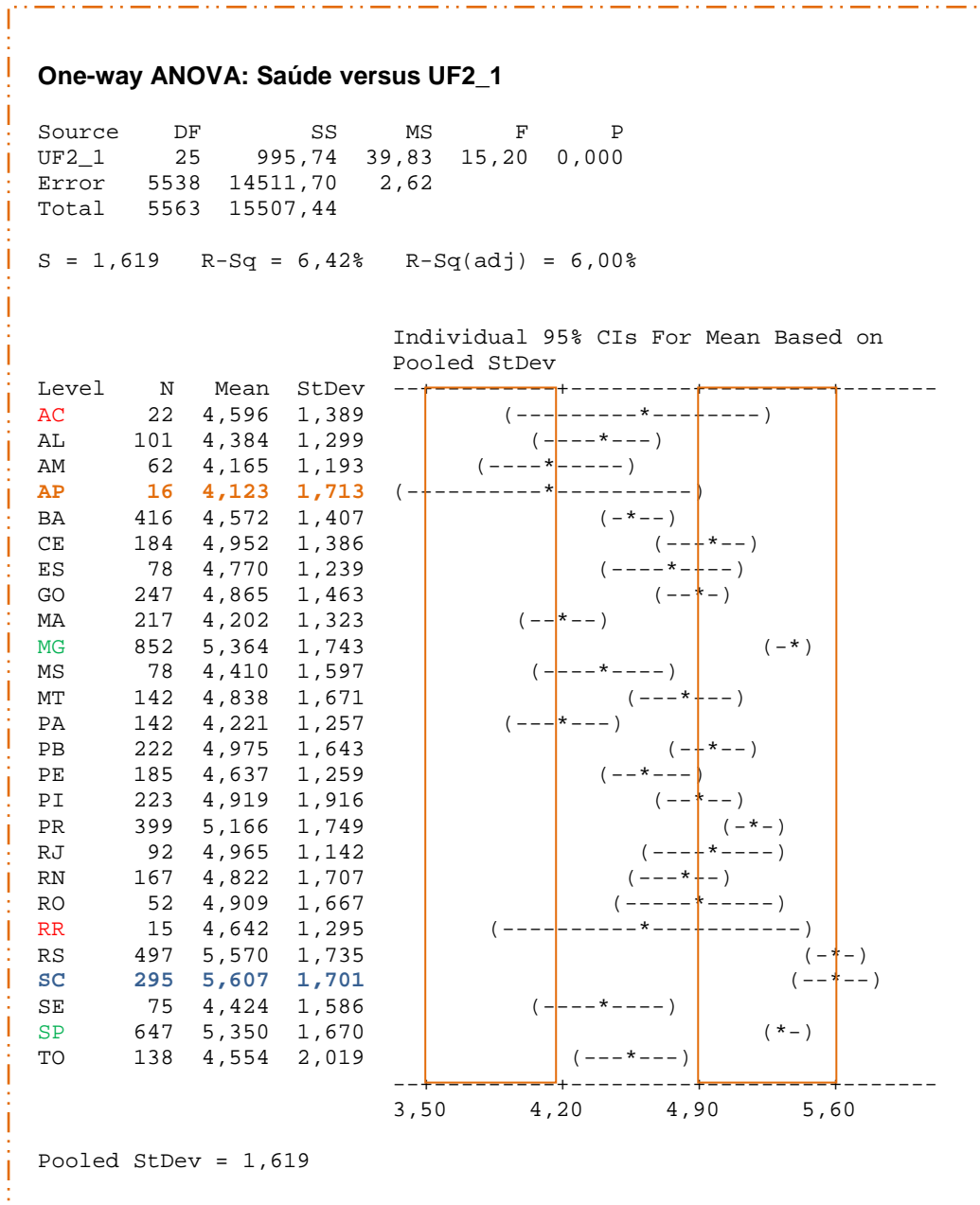


Figura 7. Análise de variância de Saúde por estado

A variabilidade dos dados de saúde é bastante acentuada em relação à média nos estados brasileiros. Roraima, Amapá e Acre são os estados com maior variabilidade dos dados em relação à média, ao passo que Minas Gerais e São Paulo apresentam menor desvio padrão com menor variabilidade. A maior média de saúde entre os estados pertence ao estado de Santa Catarina (5,607) e a menor média é apresentada pelo estado do Pará com 4,123.

3.2.3. DENDOGRAMA DOS DADOS AGRUPADOS PELO RESULTADO DAS MÉDIAS

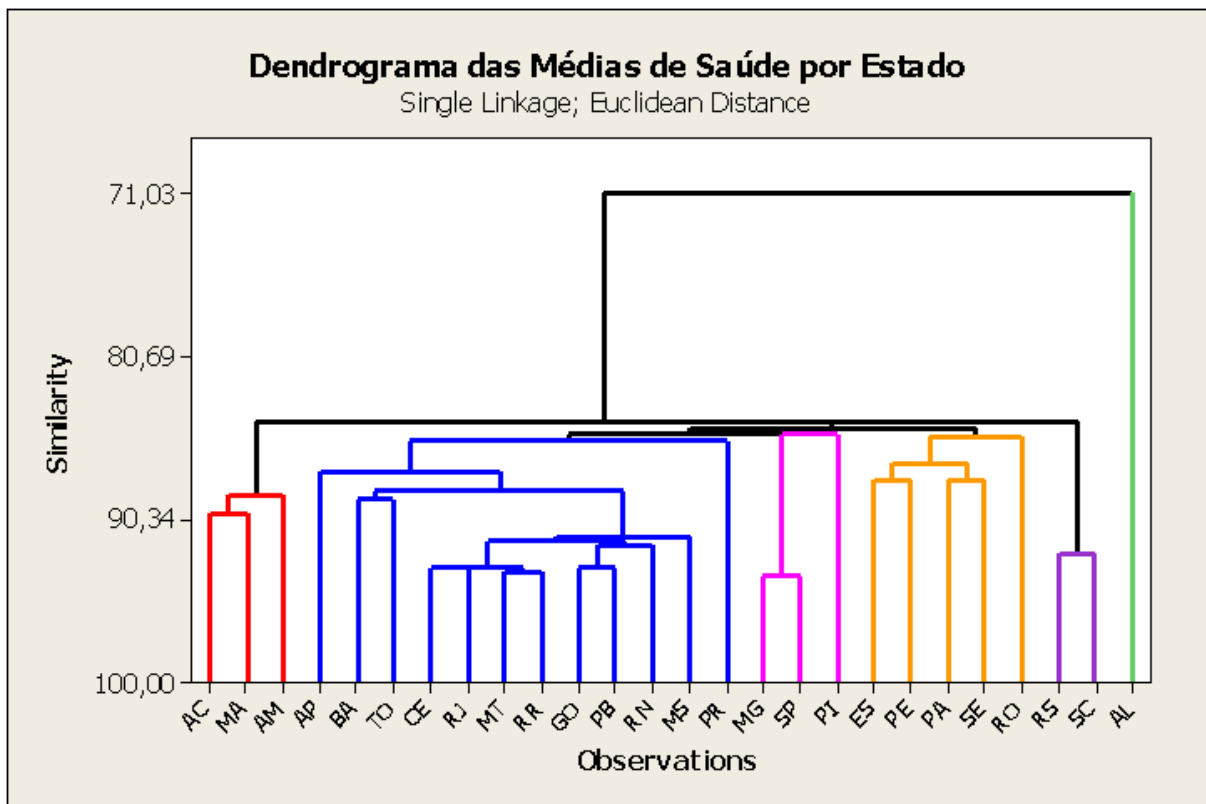


Figura 8. Dendrograma dos dados das médias de ISDM, Saúde S1_1, S1_2 e S1_3 dos municípios dos estados.

Podemos observar que existem muitos grupos (no mínimo 6) de similaridade. E que estes agrupamentos não correspondem às regiões onde os municípios se encontram, podemos obter portanto, algo como 6 “brasis”, conforme podemos ver na figura 9 abaixo. O estado do Alagoas está um pouco mais isolado que os demais, pois não possui dados similares aos demais estados do Brasil. Observe-se que Piauí na realidade está mais para azul e/ou amarelo do que lilas como MG e SP !



Figura 9. Mapa do Brasil dividido em 6 regiões de acordo com o Dendrograma das Médias de Saúde por Estado.

3.2.5. CONSIDERAÇÕES FINAIS

As análises comparativas dos dados nos permitem um resumo dos dados através de cálculos específicos como médias e desvios padrões, tornando a análise dos dados mais fácil e simples. Os gráficos de Boxplot e Dendrograma são excelentes figuras visuais para podermos analisar e interpretar os diferentes comportamentos dos dados. No dendrograma podemos analisar as similaridades dos dados e no Boxplot podemos ver as relações entre as médias e as variâncias dos agrupamentos analisados. Trata-se de ferramentas úteis para análise de grandes volumes de dados.

CAP II ANALISE DISCRIMINANTE

1. INTRODUÇÃO

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos separando-os em uma população de duas ou classes. A discriminação ou separação é a primeira etapa, sendo a parte exploratória da análise e consiste em procurar características capazes de serem utilizadas para alocar objetos em diferentes grupos previamente definidos. A classificação ou alocação pode ser definida como um conjunto de regras que serão usadas para alocar novos objetos.

O presente trabalho tem por objetivo efetuar uma análise comparativa de médias, intervalos de confiança e regressões de dados de indicadores relacionados ao desenvolvimento humano dos municípios do Brasil (ISDM). Utilizamos a análise discriminante para tentar prever ou explicar os indicadores relacionados ao desenvolvimento da saúde dos municípios do Brasil.

Os dados são originários da pesquisa da FGV / FIRJAM sobre o desenvolvimento dos municípios do Brasil. O software estatístico utilizado é o **MINITAB16**.

2. ENTENDENDO OS DADOS

2.1 Os Indivíduos

Esta pesquisa ilustra dois rankings lançados no final de 2012, e chegaram a conclusões diferentes sobre quais cidades de maior desenvolvimento do país.

Os indivíduos desta análise são os 5565 municípios brasileiros. Os dados analíticos foram extraídos do IBGE, e possibilitam uma comparação entre os dados colhidos em 2000 com 2010.

2.2 As Variáveis

As variáveis desta pesquisa incluem os 3 principais índices sintéticos que são ISDM, IFDM e IFDF, que são médias ponderadas dos dados analíticos globais da pesquisa, e variáveis analíticas, referente à educação do ensino pré escola, fundamental e médio. Esta pesquisa não utiliza variáveis do ensino superior.

Tabela 1. Comparativo entre as Variáveis ISDM e IFDM

O QUE O ISDM (FGV) MEDE	O QUE O IFDM (Firjan) MEDE
Educação: taxa de analfabetismo e taxa de crianças e jovens que frequentam a escola em cada etapa, desempenho na Prova Brasil (MEC)	Educação: taxa de matrícula infantil, abandono, distorção idade-série, desempenho no Ideb, taxa de docentes com ensino superior
Saúde e Segurança: taxa de mortalidade infantil, gravidez precoce e mortalidade por causas evitáveis; homicídios	Saúde: número de consultas pré-natal, óbitos por causa mal definidas e óbitos infantis evitáveis
Renda: presença de pobreza e extrema pobreza	Emprego e renda: geração, estoque e salários médios dos empregos formais
Trabalho: taxa de ocupação e formalização	
Habitação: coleta de lixo, energia elétrica, água canalizada, esgotamento sanitário, domicílio próprio	

Tabela 2. A definição das Variáveis

VARIÁVEL	SIGNIFICADO	TIPO	UNIDADE DE MEDIDA
Município	Nome dos municípios	Variável categórica	N/A
Indicador Social de Desenvolvimento dos Municípios (ISDM)	Média ponderada dos indicadores das dimensões Habitação, Renda, Trabalho, Saúde e Segurança e Educação (H, R, T, S e E) padronizada pela média do Brasil.	Variável Quantitativa	Numérico
IFDM	Índice Firjan de Desenvolvimento Municipal – IFDM. O índice é resultado da média ponderada de diferentes indicadores: Emprego & Renda, Educação e Saúde.	Variável Quantitativa	Numérico
IFGF	Composto por cinco indicadores: Receita Própria, Pessoal, Investimentos, Liquidez e Custo da Dívida, o índice tem como base de dados as estatísticas oficiais disponibilizadas anualmente pela Secretaria do Tesouro Nacional, constituídas por informações orçamentárias e patrimoniais prestadas pelos próprios municípios.	Variável Quantitativa	Numérico
Indicador da dimensão Saúde e Segurança (S)	Média ponderada dos indicadores da dimensão Saúde e Segurança (S1_1, S1_2, S1_3, S2_1, S2_2 e S3_1) padronizada pela média do Brasil.	Variável Quantitativa	Numérico
Taxa de mortalidade infantil, por mil nascidos vivos. (S1_1)	Taxa de sobrevivência infantil no primeiro ano de vida, representada pela diferença entre o número de nascidos vivos e o número de óbitos até um ano de idade.	Variável Quantitativa	Numérico
Mortalidade proporcional por doenças com causas evitáveis de menores de 5 anos. (S1_2)	Percentual de óbitos de menores de cinco anos de idade por causas evitáveis, em relação ao total de óbitos com causa definida na população residente na faixa etária. Define-se como causas evitáveis de menores de cinco anos todas aquelas que sejam reduzíveis por: ações de imunização; adequada atenção à mulher na gestação; adequada atenção à mulher no parto; adequada atenção ao recém-nascido; ações de diagnóstico e tratamento adequado; ou ações de promoção à saúde vinculadas a ações de atenção.	Variável Quantitativa	Numérico
Proporção de nascidos vivos com baixo peso ao nascer. (S1_3)	Percentual de nascidos vivos com peso ao nascer de até 2.5	Variável Quantitativa	Numérico
Proporção das adolescentes (10 a 19 anos) que já tiveram filho. (S2_1)	Percentual das adolescentes de 10 a 19 anos que declaram nunca ter tido algum filho nascido vivo ou filho nascido morto.	Variável Quantitativa	Numérico
Mortalidade proporcional por doenças com causas evitáveis na população de 5 a 74 anos. (S2_2)		Variável Quantitativa	Numérico

	doenças de causas infecciosas; ações adequadas de promoção à saúde, prevenção, controle e atenção às doenças não transmissíveis; ações adequadas de prevenção, controle e atenção às causas de morte materna; ações intersetoriais adequadas de promoção à saúde, prevenção e atenção às causas externas.		
Taxa de homicídio, por cem mil habitantes. (S3_1)	Percentual da população residente que foi vítima de homicídio. Define-se como homicídio os óbitos por causas externas correspondentes aos seguintes grandes grupos da Classificação Internacional de Doenças CID-10: X85-Y09 Agressões, Y35-Y36 Intervenções legais e operações de guerra.	Variável Quantitativa	Numérico

ANÁLISE DAS VARIÁVEIS

3.1 VARIÁVEIS CATEGÓRICAS

Este tipo de variável indica que o foco de concentração deve ser a análise de gráficos do tipo *pie chart* e barras.

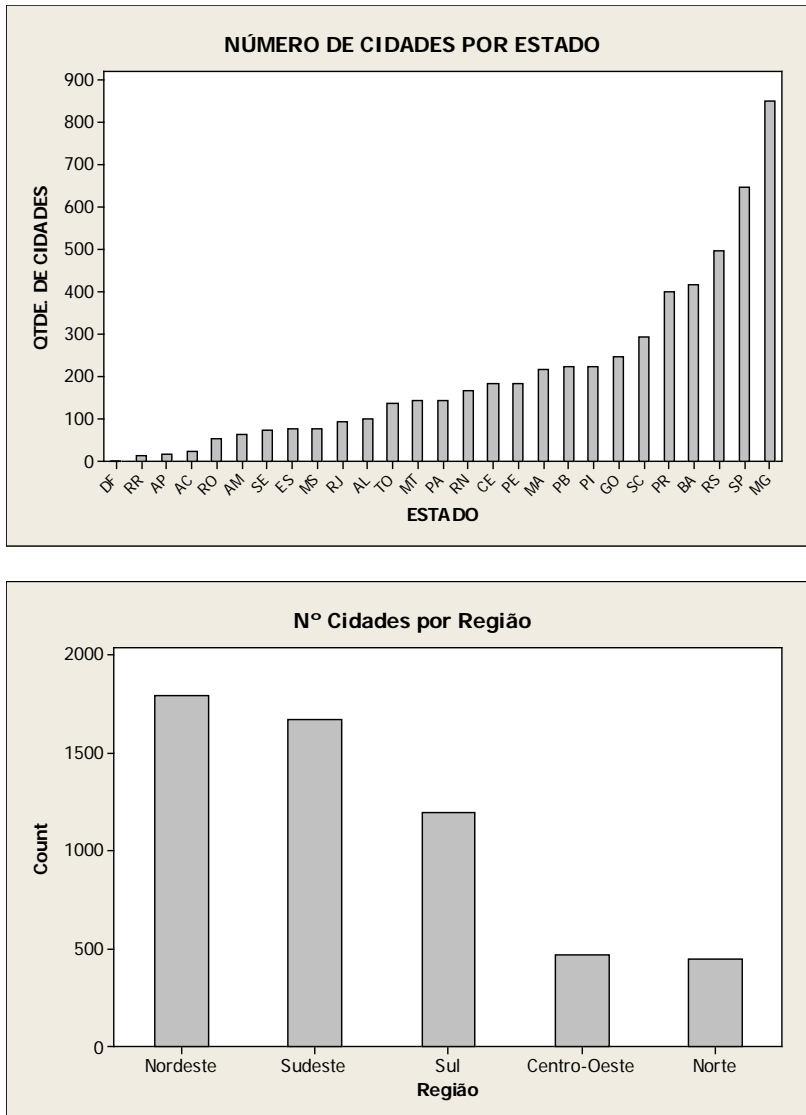
3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

Figura 3. Número de Cidades por Estado e Região do Brasil



Nos gráficos ao lado podemos ter uma dimensão do número de cidades por estado e por região. A região do Brasil com o maior número de cidades é a Nordeste (1790), seguida pela região Sudeste (1669) e pela região Sul, com 1191 cidades. As regiões com menor número de cidades

Figura 1. Número de Cidades por Estado e Região do Brasil

3.2 VARIÁVEIS QUANTITATIVAS

A análise deste tipo de variável permite a utilização de uma maior gama de ferramentas de análise como histogramas, curvas de densidade, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling. Também podemos fazer classificações supervisionadas das variáveis quantitativas, através da análise discriminante.

3.2.1. ANÁLISE DISCRIMINANTE LINEAR DOS MUNICIPIOS POR REGIÃO

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos. Nesta análise é estudada a separação dos objetos de uma população em duas ou mais classes. Neste caso iremos discriminar os valores de saúde dos municípios do Brasil e utilizaremos inicialmente a variável categórica Região. Para a geração de análise discriminante utilizaremos o comando do Minitab:

STAT >> MULTIVARIATE >> DISCRIMINANT ANALYSIS

Discriminant Analysis: Região versus ISDM; S; S1_1; S1_2; S1_3; S3_1						
Linear Method for Response: Região						
Predictors: ISDM; S; S1_1; S1_2; S1_3; S3_1						
Group	Centro-Oeste	Nordeste	Norte	Sudeste	Sul	
Count	467	1790	447	1669	1191	
Summary of classification						
		True Group				
Put into Group	Centro-Oeste	Nordeste	Norte	Sudeste	Sul	
Centro-Oeste	258	254	106	273	285	
Nordeste	23	791	105	94	24	
Norte	30	706	224	21	17	
Sudeste	72	17	5	946	475	
Sul	84	22	7	335	390	
Total N	467	1790	447	1669	1191	
N correct	258	791	224	946	390	
Proportion	0,552	0,442	0,501	0,567	0,327	
N = 5564		N Correct = 2609		Proportion Correct = 0,469		

A região que acertou mais foi a Sudeste (0,567) e a que errou mais é a Sul (0,327). O gráfico exibe o cruzamento de dados entre as regiões. A região Sudeste, por exemplo, possui 1669 municípios e apenas 946 correspondem a região, sendo que 335 são semelhantes aos dados da região Sul. O nome desta matriz é *confusion matrix* ou matriz de confusão. Podemos concluir que o agrupamento por região não é uma boa escolha segundo esta avaliação, pois só tem uma proporção de acerto de 0,469.

3.2.2. ANÁLISE DISCRIMINANTE LINEAR DOS ESTADOS POR “5 BRASIS”

Nesta análise nos interessa verificar os possíveis agrupamento de dados utilizando a variável 5 Brasis (vide figura 2) que demonstra os agrupamentos do Brasil segundo sua proximidade de dados de saúde.



Figura 2. Mapa do Brasil dividido em 6 regiões de acordo com o Dendograma das Médias de Saúde por Estado.

Discriminant Analysis: 5 Brasis versus ISDM m; SAÚDE m; ...

Linear Method for Response: 5 Brasis

Predictors: ISDM m; SAÚDE m; S1_1 m; S1_2 m; S1_3 m; S3_1 m

Group	B1	B2	B3	B4	B5
Count	3	12	4	4	3

Summary of classification

Put into Group	True Group				
	B1	B2	B3	B4	B5
B1	3	0	0	0	0
B2	0	8	1	1	0
B3	0	1	2	0	0
B4	0	0	1	3	1
B5	0	3	0	0	2
Total N	3	12	4	4	3
N correct	3	8	2	3	2
Proportion	1,000	0,667	0,500	0,750	0,667

N = 26

N Correct = 18

Proportion Correct = 0,692

Podemos observar que o agrupamento dos estados do Brasil em 5 regiões apresenta melhores resultados, uma vez que a proporção de acertos se elevou para 0,692. O grupo com maior número de acertos é o B4 (0,750).

3.2.3. ANÁLISE DISCRIMINANTE LINEAR DOS MUNICÍPIOS POR “3 BRASIS”

Discriminant Analysis: 3 Brasis versus ISDM; S; S1_1; S1_2; S1_3; S3_1

```

Linear Method for Response: 3 Brasis

Predictors: ISDM; S; S1_1; S1_2; S1_3; S3_1

Group  Centro-Oeste    NN    SS
Count          467  2237  2860

Summary of classification

Put into Group      True Group
                    Centro-Oeste    NN    SS
Centro-Oeste        288    389    654
NN                   48   1813    150
SS                   131    35   2056
Total N              467   2237   2860
N correct            288   1813   2056
Proportion           0,617  0,810  0,719

N = 5564           N Correct = 4157           Proportion Correct = 0,747

```

Analisando a divisão do Brasil em 3 regiões, sendo Centro-Oeste, a união dos estados das regiões Sul e Sudeste, formando a região SS, e a união dos estados das regiões Norte e Nordeste, formando a região NN, podemos observar que a proporção de acerto é maior (0,747). Se olharmos individualmente as três regiões formados veremos que a região NN (Norte+Nordeste) é a região com maior proporção de acerto(mais homogeneamente ruim...).

3.2.4. ANÁLISE DISCRIMINANTE QUADRÁTICA POR “3 BRASIS”

Uma boa classificação deve resultar em pequenos erros, isto é, deve haver pouca probabilidade de má classificação. Quando a regra de classificação assume que as variâncias das populações são iguais, as funções discriminantes são ditas lineares e quando não são, as funções discriminantes são ditas quadráticas. Vamos agora verificar a função quadrática para 3 Brasis.

Discriminant Analysis: 3 Brasis versus ISDM; S; S1_1; S1_2; S1_3; S3_1

Quadratic Method for Response: 3 Brasis

Predictors: ISDM; S; S1_1; S1_2; S1_3; S3_1

Group	Centro-Oeste	NN	SS
Count	467	2237	2860

Summary of classification

Put into Group	True Group		
	Centro-Oeste	NN	SS
Centro-Oeste	288	314	633
NN	53	1876	181
SS	126	47	2046
Total N	467	2237	2860
N correct	288	1876	2046
Proportion	0,617	0,839	0,715

N = 5564

N Correct = 4210

Proportion Correct = 0,757

No modelo quadrático a proporção foi alterada em apenas 1% (de 0,74 para 0,75). Seguindo a premissa da parcimônia, vamos escolher o método linear pois é o mais simples.

3.2.5. ANÁLISE DISCRIMINANTE LINEAR PARA DADOS AGRUPADOS

Neste exemplo abaixo vamos através do dendograma pesquisar o grau de similaridade em relação as variáveis de desvio padrão(disparidades) nos estados do Brasil. Com base na similaridade poderemos definimos agrupamento de dados e após utilizamos a análise discriminante para verificar a proporção correta dos agrupamentos.

Discriminant Analysis: 5 Brasis versus ISDM sdn; SAÚDE sdn; ...

Linear Method for Response: 5 Brasis

Predictors: ISDM sdn; SAÚDE sdn; S1_1 sdn; S1_2 sdn; S1_3 sdn; S3_1 sdn

Group	B1	B2	B3	B4	B5
Count	3	12	4	4	3

Summary of classification

Put into Group	True Group				
	B1	B2	B3	B4	B5
B1	3	1	0	0	0
B2	0	7	0	1	0
B3	0	2	3	0	1
B4	0	1	0	3	1
B5	0	1	1	0	1
Total N	3	12	4	4	3
N correct	3	7	3	3	1
Proportion	1,000	0,583	0,750	0,750	0,333

N = 26 N Correct = 17 Proportion Correct = 0,654

Neste caso a proporção correta é de 65%, ou seja, os agrupamentos gerados anteriormente pelo agrupamento em 5 Brasis gerou a mesma proporção do método linear utilizado na análise discriminante, neste caso tratando-se de disparidades.

4. CONSIDERAÇÕES FINAIS

A tarefa da análise discriminante é encontrar a melhor função discriminante linear de um conjunto de variáveis que reproduza, tanto quanto possível, um agrupamento de casos considerados.

No Minitab utilizamos o procedimento em passos, no qual cada passo introduz a variável mais poderosa na função discriminante. A função critério para selecionar a próxima variável depende do número de grupos especificados (o número de grupos varia de 2 a 20).

Os cálculos podem ser realizados em toda a população ou em amostra de dados ou mesmo em dados previamente agrupados, como é o caso deste trabalho que agrupou as variáveis de municípios brasileiros em 5 e 3 regiões diferentes de acordo com a proximidade dos dados de saúde.

Para a análise dos dados de saúde utilizamos a análise discriminante linear e conseguimos um resultado de 0,75 de proporção correta para um Brasil agrupado em 3 regiões e tanto para as medias quanto para as disparidades referentes aos estados .

CAP III REGRESSÃO LOGÍSTICA

1. INTRODUÇÃO

A regressão logística é uma técnica estatística que tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias^{1 2}. A regressão logística é amplamente usada em ciências médicas e sociais, e tem outras denominações, como modelo logístico e classificador de máxima entropia.

Enquanto método de predição para variáveis categóricas, a regressão logística é comparável às técnicas supervisionadas propostas em aprendizagem automática (árvores de decisão, redes neurais, etc.), ou ainda a análise discriminante preditiva em estatística exploratória. Trata-se de um modelo de regressão para variáveis dependentes ou de resposta binomialmente distribuídas. É útil para modelar a probabilidade de um evento ocorrer como função de outros fatores. Os dados são originários da pesquisa da FGV / FIRJAM sobre o desenvolvimento dos municípios do Brasil. Neste trabalho abordaremos as variáveis referentes à saúde dos municípios. O software estatístico utilizado é o **MINITAB16**.

2. ENTENDENDO OS DADOS

2.1 Os Indivíduos

Esta pesquisa ilustra dois rankings lançados no final de 2012, e chegaram a conclusões diferentes sobre quais cidades de maior desenvolvimento do país.

Os indivíduos desta análise são os 5.565 municípios brasileiros. Os dados analíticos foram extraídos do IBGE, e possibilitam uma comparação entre os dados colhidos em 2000 com 2010.

2.2 As Variáveis

As variáveis desta pesquisa incluem os 3 principais índices sintéticos que são ISDM, IFDM e IFDF, que são médias ponderadas dos dados analíticos globais da pesquisa, e variáveis analíticas, referente à educação do ensino pré escola, fundamental e médio.

Tabela 1. Comparativo entre as Variáveis ISDM e IFDM

O QUE O ISDM (FGV) MEDE	O QUE O IFDM (Firjan) MEDE
Educação: taxa de analfabetismo e taxa de crianças e jovens que frequentam a escola em cada etapa, desempenho na Prova Brasil (MEC)	Educação: taxa de matrícula infantil, abandono, distorção idade-série, desempenho no Ideb, taxa de docentes com ensino superior
Saúde e Segurança: taxa de mortalidade infantil, gravidez precoce e mortalidade por causas evitáveis; homicídios	Saúde: número de consultas pré-natal, óbitos por causa mal definidas e óbitos infantis evitáveis
Renda: presença de pobreza e extrema pobreza	Emprego e renda: geração, estoque e salários médios dos empregos formais
Trabalho: taxa de ocupação e formalização	
Habitação: coleta de lixo, energia elétrica, água canalizada, esgotamento sanitário, domicílio próprio	

Tabela 2. A definição das Variáveis

VARIÁVEL	SIGNIFICADO	TIPO	UNIDADE DE MEDIDA
Município	Nome dos municípios	Variável categórica	N/A
Indicador Social de Desenvolvimento dos Municípios (ISDM)	Média ponderada dos indicadores das dimensões Habitação, Renda, Trabalho, Saúde e Segurança e Educação (H, R, T, S e E) padronizada pela média do Brasil.	Variável Quantitativa	Numérico
IFDM	Índice Firjan de Desenvolvimento Municipal – IFDM. O índice é resultado da média ponderada de diferentes indicadores: Emprego & Renda, Educação e Saúde.	Variável Quantitativa	Numérico
IFGF	Composto por cinco indicadores: Receita Própria, Pessoal, Investimentos, Liquidez e Custo da Dívida, o índice tem como base de dados as estatísticas oficiais disponibilizadas anualmente pela Secretaria do Tesouro Nacional, constituídas por informações orçamentárias e patrimoniais prestadas pelos próprios municípios.	Variável Quantitativa	Numérico
Indicador da dimensão Saúde e Segurança (S)	Média ponderada dos indicadores da dimensão Saúde e Segurança (S1_1, S1_2, S1_3, S2_1, S2_2 e S3_1) padronizada pela média do Brasil.	Variável Quantitativa	Numérico
Taxa de mortalidade infantil, por mil nascidos vivos. (S1_1)	Taxa de sobrevivência infantil no primeiro ano de vida, representada pela diferença entre o número de nascidos vivos e o número de óbitos até um ano de idade.	Variável Quantitativa	Numérico
Mortalidade proporcional por doenças com causas evitáveis de menores de 5 anos. (S1_2)	Percentual de óbitos de menores de cinco anos de idade por causas evitáveis, em relação ao total de óbitos com causa definida na população residente na faixa etária. Define-se como causas evitáveis de menores de cinco anos todas aquelas que sejam reduzíveis por: ações de imunização; adequada atenção à mulher na gestação; adequada atenção à mulher no parto; adequada atenção ao recém-nascido; ações de diagnóstico e tratamento adequado;	Variável Quantitativa	Numérico

	ou ações de promoção à saúde vinculadas a ações de atenção.		
Proporção de nascidos vivos com baixo peso ao nascer. (S1_3)	Percentual de nascidos vivos com peso ao nascer de até 2.500 gra	Variável Quantitativa	Numérico
Proporção das adolescentes (10 a 19 anos) que já tiveram filho. (S2_1)	Percentual das adolescentes de 10 a 19 anos que declaram nunca ter tido algum filho nascido vivo ou filho nascido morto.	Variável Quantitativa	Numérico
Mortalidade proporcional por doenças com causas evitáveis na população de 5 a 74 anos. (S2_2)		Variável Quantitativa	Numérico
Taxa de homicídio, por cem mil habitantes. (S3_1)	Percentual da população residente que foi vítima de homicídio. Define-se como homicídio os óbitos por causas externas correspondentes aos seguintes grandes grupos da Classificação Internacional de Doenças CID-10: X85-Y09 Agressões, Y35-Y36 Intervenções legais e operações de guerra.	Variável Quantitativa	Numérico

3. ANÁLISE DAS VARIÁVEIS

3.1 VARIÁVEIS CATEGÓRICAS

Este tipo de variável indica que o foco de concentração deve ser a análise de gráficos do tipo *pie chart* e barras.

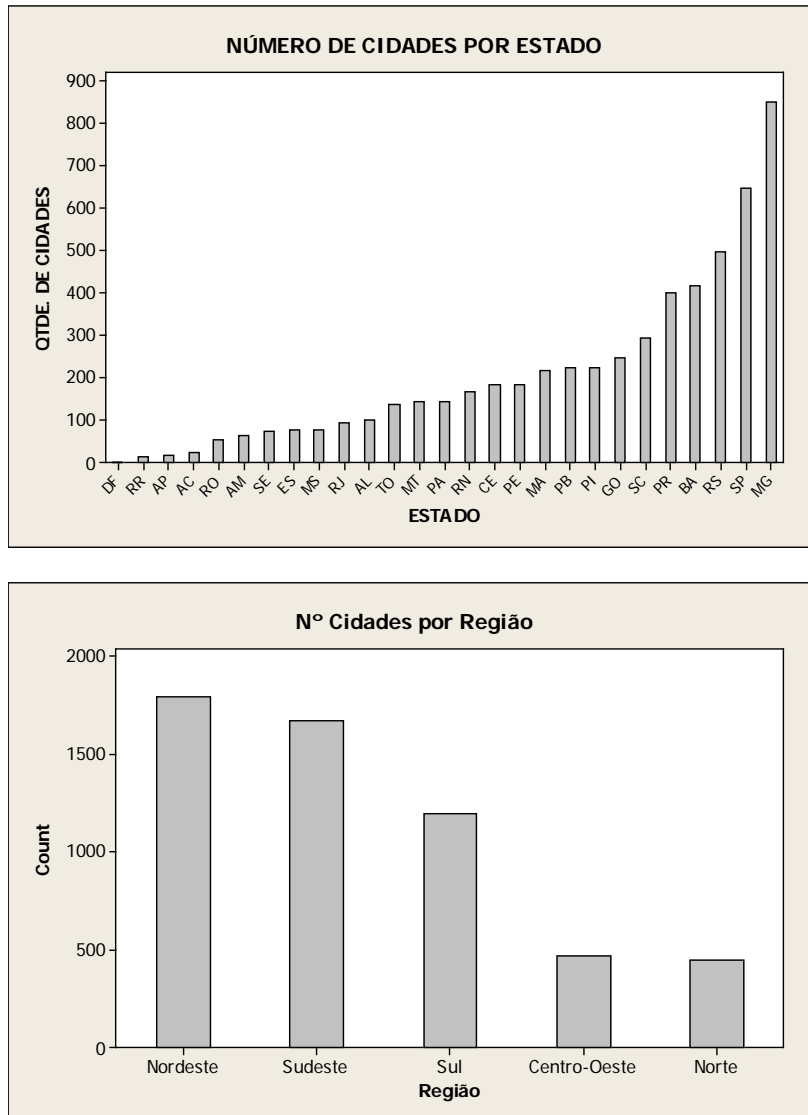
3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exibe o número de cidades por estado. A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”

3.1.2 Variável: “REGIÃO”

Figura 1. Número de Cidades por Estado e Região do Brasil



Nos gráficos ao lado podemos ter uma dimensão do número de cidades por estado e por região. A região do Brasil com o maior número de cidades é a Nordeste (1790), seguida pela região Sudeste (1669) e pela região Sul, com 1191 cidades. As regiões com menor número de cidades

3.2 VARIÁVEIS QUANTITATIVAS

A análise deste tipo de variável permite a utilização de uma maior gama de ferramentas de análise como histogramas, curvas de densidade, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling. Também podemos fazer classificações supervisionadas das variáveis quantitativas, através da análise discriminante.

3.2.1. REGRESSÃO LOGÍSTICA

Stat >> Regression >> Ordinal Logistical Regression

Ordinal Logistic Regression: Região versus ISDM; S; ...

Link Function: Logit

Response Information

Variable	Value	Count
Região	Centro-Oeste	467
	Nordeste	1790
	Norte	447
	Sudeste	1669
	Sul	1191
	Total	5564

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Const(1)	-119,215	131,387	-0,91	0,364			
Const(2)	-116,872	131,387	-0,89	0,374			
Const(3)	-116,350	131,387	-0,89	0,376			
Const(4)	-114,481	131,387	-0,87	0,384			
ISDM	-0,880135	0,0290018	-30,35	0,000	0,41	0,39	0,44
S	7,79536	8,54929	0,91	0,362	2429,31	0,00	4,59886E+10
S1_1	0,0441298	0,0435016	1,01	0,310	1,05	0,96	1,14
S1_2	0,397636	0,434753	0,91	0,360	1,49	0,63	3,49
S1_3	0,356917	0,434789	0,82	0,412	1,43	0,61	3,35
S2_1	0,669376	0,651945	1,03	0,305	1,95	0,54	7,01
S2_2	0,592068	0,652140	0,91	0,364	1,81	0,50	6,49
S3_1	0,0113428	0,0018573	6,11	0,000	1,01	1,01	1,02

Log-Likelihood = -7170,302

Test that all slopes are zero: G = 1979,097, DF = 8, P-Value = 0,000

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	26486,5	22244	0,000
Deviance	14340,6	22244	1,000

Measures of Association:

(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures
Concordant	8543177	73,9	Somers' D 0,48
Discordant	2980325	25,8	Goodman-Kruskal Gamma 0,48
Ties	42526	0,4	Kendall's Tau-a 0,36
Total	11566028	100,0	

4. CONSIDERAÇÕES FINAIS

Enquanto método de predição para variáveis categóricas, a regressão logística é comparável às técnicas supervisionadas propostas em aprendizagem automática (árvores de decisão, redes neurais, etc.), ou ainda a análise discriminante preditiva em estatística exploratória. É possível as colocar em concorrência para escolha do modelo mais adaptado para um certo problema preditivo a resolver. **Considerando então 5 Brasis a % de acertos utilizando Regressão Logística foi de 74%.**

CAP IV ÁRVORES DE CLASSIFICAÇÃO

1. INTRODUÇÃO

A técnica “árvore de classificação” alcançou o segmento de negócios através da utilização em pesquisas de mercado. Tendo como pontos fortes a simplicidade de sua representação gráfica baseado em árvores e a facilidade de entender as regras e perfis derivados de cada segmento (nós), rapidamente foi adotada por outras áreas de marketing, sobretudo aquelas formadas por gestores com menor grau de sofisticação analítica.

Os dados são originários da pesquisa da FGV / FIRJAM sobre o desenvolvimento dos municípios do Brasil. Neste trabalho abordaremos as variáveis referentes à saúde dos municípios. O software estatístico utilizado é o **SPSS21**.

2. ENTENDENDO OS DADOS

2.1 Os Indivíduos

Esta pesquisa ilustra dois rankings lançados no final de 2012, e chegaram a conclusões diferentes sobre quais cidades de maior desenvolvimento do país.

Os indivíduos desta análise são os 5565 municípios brasileiros. Os dados analíticos foram extraídos do IBGE, e possibilitam uma comparação entre os dados colhidos em 2000 com 2010.

2.2 As Variáveis

As variáveis desta pesquisa incluem os 3 principais índices sintéticos que são ISDM, IFDM e IFDF, que são médias ponderadas dos dados analíticos globais da pesquisa, e variáveis analíticas, referente à saúde.

Tabela 1. Comparativo entre as Variáveis ISDM e IFDM

O QUE O ISDM (FGV) MEDE	O QUE O IFDM (Firjan) MEDE
Educação: taxa de analfabetismo e taxa de crianças e jovens que frequentam a escola em cada etapa, desempenho na Prova Brasil (MEC)	Educação: taxa de matrícula infantil, abandono, distorção idade-série, desempenho no Ideb, taxa de docentes com ensino superior
Saúde e Segurança: taxa de mortalidade infantil, gravidez precoce e mortalidade por causas evitáveis; homicídios	Saúde: número de consultas pré-natal, óbitos por causa mal definidas e óbitos infantis evitáveis
Renda: presença de pobreza e extrema pobreza	Emprego e renda: geração, estoque e salários médios dos empregos formais
Trabalho: taxa de ocupação e formalização	
Habitação: coleta de lixo, energia elétrica, água canalizada, esgotamento sanitário, domicílio próprio	

Tabela 2. A definição das Variáveis

VARIÁVEL	SIGNIFICADO	TIPO	UNIDADE DE MEDIDA
Município	Nome dos municípios	Variável categórica	N/A
Indicador Social de Desenvolvimento dos Municípios (ISDM)	Média ponderada dos indicadores das dimensões Habitação, Renda, Trabalho, Saúde e Segurança e Educação (H, R, T, S e E) padronizada pela média do Brasil.	Variável Quantitativa	Numérico
IFDM	Índice Firjan de Desenvolvimento Municipal – IFDM. O índice é resultado da média ponderada de diferentes indicadores: Emprego & Renda, Educação e Saúde.	Variável Quantitativa	Numérico
IFGF	Composto por cinco indicadores: Receita Própria, Pessoal, Investimentos, Liquidez e Custo da Dívida, o índice tem como base de dados as estatísticas oficiais disponibilizadas anualmente pela Secretaria do Tesouro Nacional, constituídas por informações orçamentárias e patrimoniais prestadas pelos próprios municípios.	Variável Quantitativa	Numérico
Indicador da dimensão Saúde e Segurança (S)	Média ponderada dos indicadores da dimensão Saúde e Segurança (S1_1, S1_2, S1_3, S2_1, S2_2 e S3_1) padronizada pela média do Brasil.	Variável Quantitativa	Numérico
Taxa de mortalidade infantil, por mil nascidos vivos. (S1_1)	Taxa de sobrevivência infantil no primeiro ano de vida, representada pela diferença entre o número de nascidos vivos e o número de óbitos até um ano de idade.	Variável Quantitativa	Numérico
Mortalidade proporcional por doenças com causas evitáveis de menores de 5 anos. (S1_2)	Percentual de óbitos de menores de cinco anos de idade por causas evitáveis, em relação ao total de óbitos com causa definida na população residente na faixa etária. Define-se como causas evitáveis de menores de cinco anos todas aquelas que sejam reduzíveis por: ações de imunização; adequada	Variável Quantitativa	Numérico

	atenção à mulher na gestação; adequada atenção à mulher no parto; adequada atenção ao recém-nascido; ações de diagnóstico e tratamento adequado; ou ações de promoção à saúde vinculadas a ações de atenção.		
Proporção de nascidos vivos com baixo peso ao nascer. (S1_3)		Variável Quantitativa	Numérico
Proporção das adolescentes (10 a 19 anos) que já tiveram filho. (S2_1)	Percentual das adolescentes de 10 a 19 anos que declaram nunca ter tido algum filho nascido vivo ou filho nascido morto.	Variável Quantitativa	Numérico
Mortalidade proporcional por doenças com causas evitáveis na população de 5 a 74 anos. (S2_2)		Variável Quantitativa	Numérico
Taxa de homicídio, por cem mil habitantes. (S3_1)	Percentual da população residente que foi vítima de homicídio. Define-se como homicídio os óbitos por causas externas correspondentes aos seguintes grandes grupos da Classificação Internacional de Doenças CID-10: X85-Y09 Agressões, Y35-Y36 Intervenções legais e operações de guerra.	Variável Quantitativa	Numérico

3. ANÁLISE DAS VARIÁVEIS

3.1 VARIÁVEIS CATEGÓRICAS

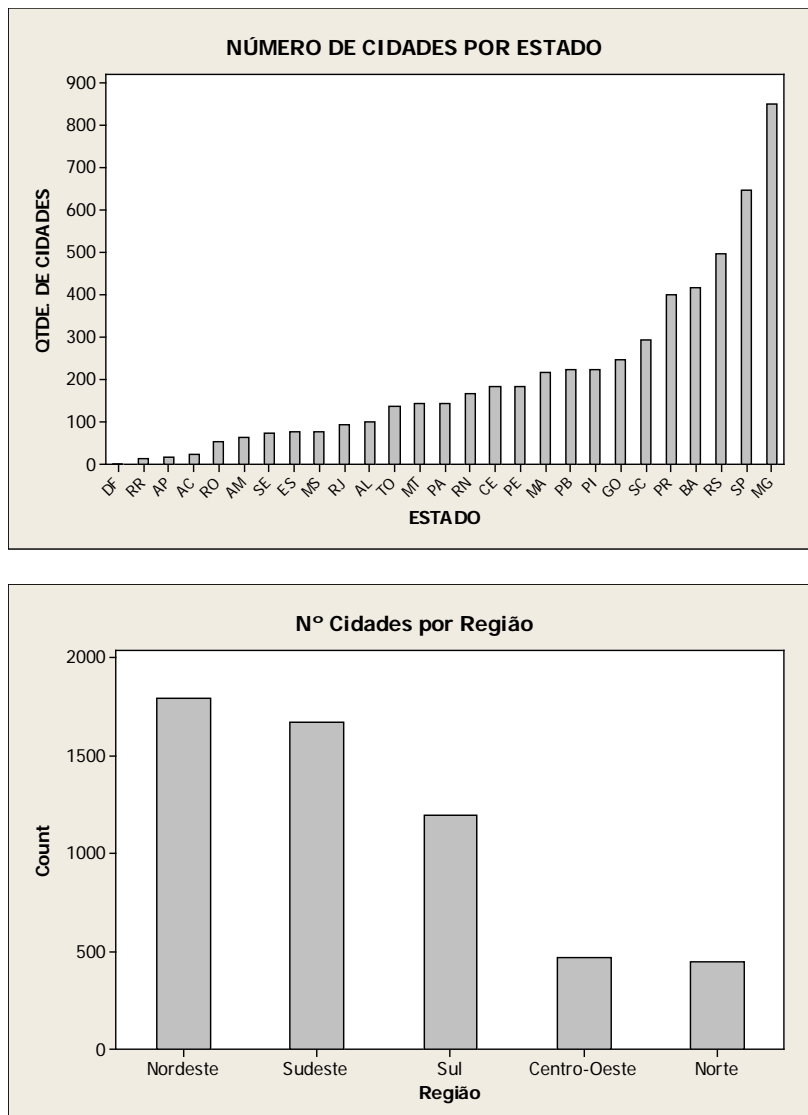
Este tipo de variável indica que o foco de concentração deve ser a análise de gráficos do tipo *pie chart* e barras.

3.1.1 Variável: “Estado”

Fazem parte desta pesquisa os 27 estados brasileiros e suas cidades. O gráfico abaixo exhibe o número de cidades por estado.

A variação no número de cidades por estado é acentuada. Considerando que o Distrito Federal é um estado brasileiro, é o estado com o menor número de cidades (1), enquanto o Mato Grosso possui mais de 852 cidades.

3.1.2 Variável: “REGIÃO”



Nos gráficos ao lado podemos ter uma dimensão do número de cidades por estado e por região. A região do Brasil com o maior número de cidades é a Nordeste (1790), seguida pela região Sudeste (1669) e pela região Sul, com 1191 cidades. As regiões com menor número de cidades é a região Centro-Oeste (468) e a região Norte, que possui 447 cidades.

Figura 3. Número de Cidades por Estado e Região do Brasil

3.2 VARIÁVEIS QUANTITATIVAS

A análise deste tipo de variável permite a utilização de uma maior gama de ferramentas de análise como histogramas, curvas de densidade, gráfico de ramos, box-plot e dot-plot, além de informações numéricas como média, desvio-padrão, mediana, quartis, 5 números, intervalo de confiança e teste de normalidade de Anderson-Darling. Também podemos fazer classificações supervisionadas das variáveis quantitativas, através da análise discriminante.

3.2.1. ÁRVORES DE CLASSIFICAÇÃO DAS VARIÁVEIS DE SAÚDE

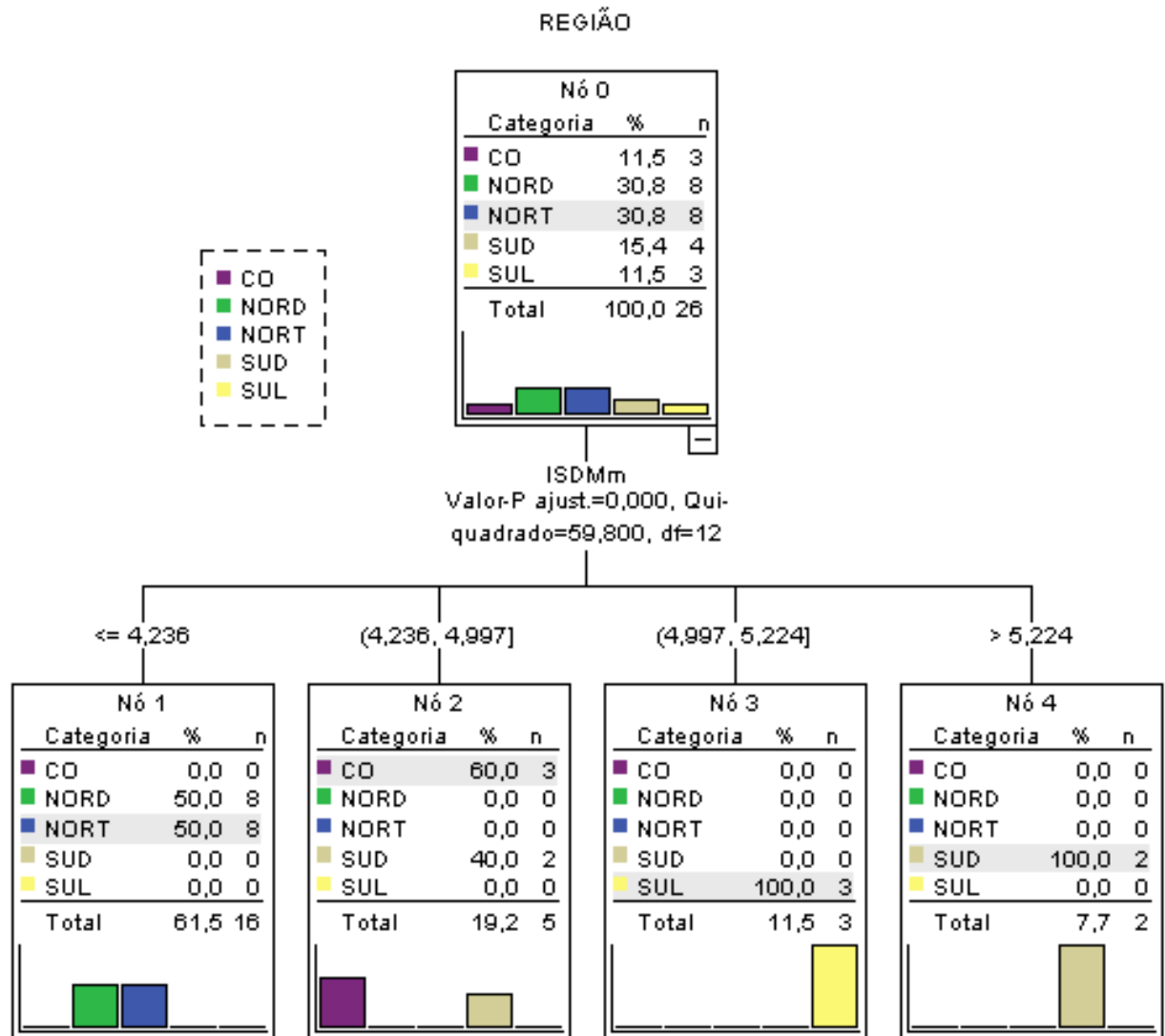
COMANDO SPSS:

ANALISAR >> CLASSIFICAR >> ARVORE

Este resultado se refere à variável dependente REGIAO e as variáveis independente ISDMm, SAÚDEm, S1_1m, S1_2M, S1_3m, S3_1m.

Resumo do modelo

	Método de crescimento	CHAID	
	Variável dependente	REGIÃO	
	Variáveis independentes	ISDMm, SAÚDEm, S1_1m, S1_2M, S1_3m, S3_1m	
Especificações	Validação	Nenhum	
	Profundidade de árvore máxima		3
	Casos mínimos em nó pai		2
	Casos mínimos em nó filho		1
	Variáveis independentes incluídas	ISDMm	
Resultados	Número de nós		5
	Número de nós de terminal		4
	Profundidade		1



Risco	
Estimativas	Modelo padrão
,385	,095

Método de crescimento: CHAID

Variável dependente: REGIÃO

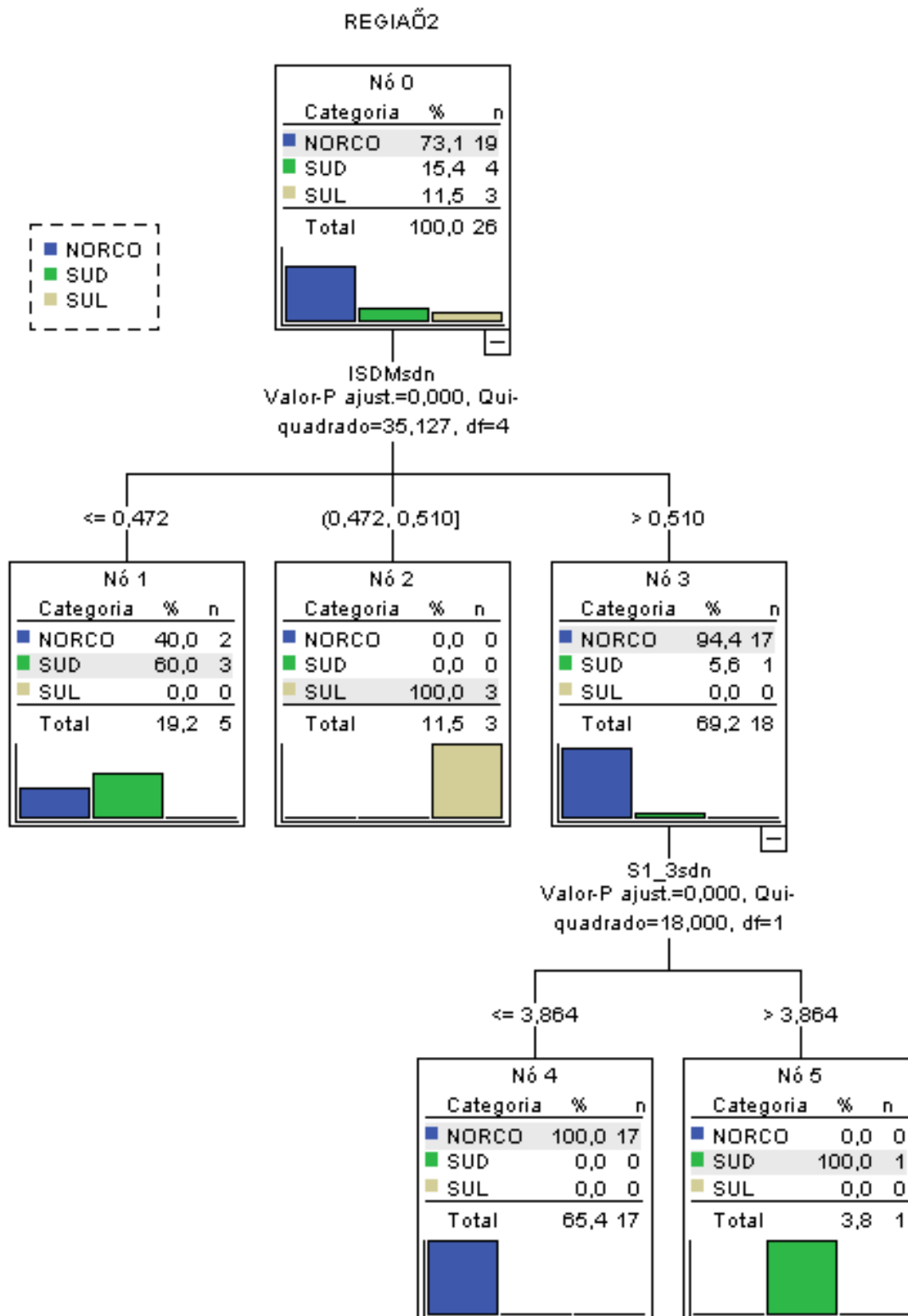
Observado	Posto					
	CO	NORD	NORT	SUD	SUL	Porcentagem Correta
CO	3	0	0	0	0	100,0%
NORD	0	0	8	0	0	0,0%
NORT	0	0	8	0	0	100,0%
SUD	2	0	0	2	0	50,0%
SUL	0	0	0	0	3	100,0%
Porcentagem global	19,2%	0,0%	61,5%	7,7%	11,5%	61,5%

Método de crescimento: CHAID

Variável dependente: REGIÃO

Este resultado de 61,5 % de acertos se refere à 5 Brasis (Nordeste, Norte, Centro-Oeste, Sudeste e Sul) e as variáveis referentes a disparidades nos estados: ISDMsdn, SAÚDEsdn, S1_1sdn, S1_2sdn, S1_3sdn, S3_1sdn.

Resumo do modelo		
	Método de crescimento	CHAID
	Variável dependente	REGIÃO2
	Variáveis independentes	ISDMsdn, SAÚDEsdn, S1_1sdn, S1_2sdn, S1_3sdn, S3_1sdn
Especificações	Validação	Nenhum
	Profundidade de árvore máxima	3
	Casos mínimos em nó pai	2
	Casos mínimos em nó filho	1
	Variáveis independentes incluídas	ISDMsdn, S1_3sdn
Resultados	Número de nós	6
	Número de nós de terminal	4
	Profundidade	2



Risco	
Estimativas	Modelo padrão
,077	,052

Método de crescimento: CHAID

Variável dependente: REGIAÕ2

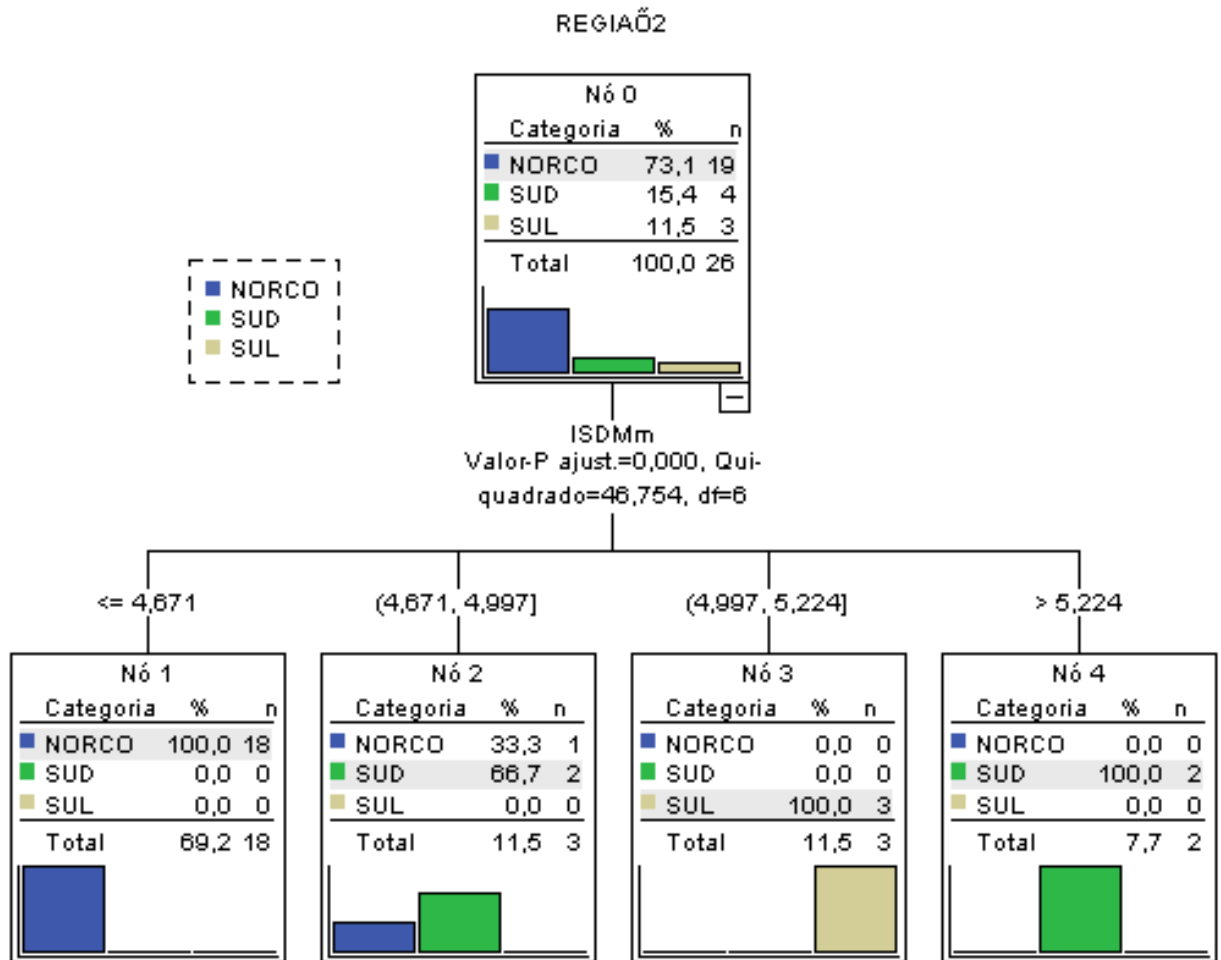
Observado	Posto			
	NORCO	SUD	SUL	Porcentagem Correta
NORCO	17	2	0	89,5%
SUD	0	4	0	100,0%
SUL	0	0	3	100,0%
Porcentagem global	65,4%	23,1%	11,5%	92,3%

Método de crescimento: CHAID

Variável dependente: REGIAÕ2

Este resultado de 92,3 % de acertos se refere à 3 Brasis (Nordeste + Norte + Centro-Oeste, Sudeste e Sul) e as variáveis de disparidades nos estados referentes a : ISDMsdn, SAÚDEsdn, S1_1sdn, S1_2sdn, S1_3sdn, S3_1sdn

Resumo do modelo		
	Método de crescimento	CHAID
	Variável dependente	REGIAÕ2
	Variáveis independentes	ISDMm, SAÚDEm, S1_1m, S1_2M, S1_3m, S3_1m
Especificações	Validação	Nenhum
	Profundidade de árvore máxima	3
	Casos mínimos em nó pai	2
	Casos mínimos em nó filho	1
	Variáveis independentes incluídas	ISDMm
Resultados	Número de nós	5
	Número de nós de terminal	4
	Profundidade	1



Risco

Estimativas	Modelo padrão
,038	,038

Método de crescimento: CHAID

Variável dependente: REGIAÕ2

Posto

Observado	Previsto			
	NORCO	SUD	SUL	Porcentagem Correta
NORCO	18	1	0	94,7%
SUD	0	4	0	100,0%
SUL	0	0	3	100,0%
Porcentagem global	69,2%	19,2%	11,5%	96,2%

Método de crescimento: CHAID

Variável dependente: REGIÃO2

Este resultado de 96,2 % de acertos se refere à 3 Brasis (Nordeste + Norte + Centro-Oeste, Sudeste e Sul) e as medias das variáveis nos estados referentes a : ISDMm, SAÚDEm, S1_1m, S1_2M, S1_3m, S3_1m