UMA ANÁLISE COMPARATIVA DE MODELOS PARA CLASSIFICAÇÃO E PREVISÃO DE SOBREVIVÊNCIA OU ÓBITO DE CRIANÇAS NASCIDAS NO RIO DE JANEIRO EM 2006 NO PRIMEIRO ANO DE VIDA

Mariana Pereira Nunes
Escola Nacional de Ciências Estatísticas, ENCE/IBGE
End. eletrônico: maripnunes@msn.com
Daniel Takata Gomes
Escola Nacional de Ciências Estatísticas, ENCE/IBGE
End. eletrônico: daniel.gomes@ibge.gov.br

RESUMO

A Taxa de Mortalidade Infantil (TMI) é muito importante para avaliar a qualidade de vida de uma população. A partir das informações do Ministério da Saúde (MS) provenientes do Sistema de Informação de Nascidos Vivos (SINASC) e do Sistema de Informações de Mortalidade (SIM) foi possível verificar como diversos fatores, como biológicos, socioeconômicos e assistenciais, estão relacionados à mortalidade de crianças no primeiro ano de vida. Assim, técnicas de Análise Multivariada e Redes Neurais foram utilizadas com o objetivo de prever o óbito (ou não) da criança no primeiro ano de vida a partir da Declaração de Nascido Vivo (DN). Em seguida essas duas metodologias foram comparadas quanto à sua eficácia na previsão dos óbitos.

Palavras-chave: Mortalidade Infantil, Análise Multivariada, Redes Neurais.

INTRODUÇÃO

O objetivo deste trabalho é conhecer o perfil das crianças que nasceram no Estado do Rio de Janeiro no ano de 2006 e morreram antes de completar um ano de vida e avaliar métodos de reconhecimento de padrões quanto a capacidade de prever estes óbitos. Para isso foram utilizados dados provenientes do DATASUS, Ministério da Saúde.

A motivação que levou ao estudo deste tema é a importância de se conhecer os principais fatores que contribuem para a Mortalidade Infantil com a finalidade de, eventualmente, identificar as variáveis a serem tratadas para a melhoria das taxas.

A mortalidade infantil (MI) corresponde ao risco de morte durante o primeiro ano de vida de uma geração. É medida através do quociente de óbitos ocorridos entre o nascimento e o primeiro aniversário e o número de nascidos vivos durante determinado ano e local [BANDEIRA, 2004].

A taxa de Mortalidade Infantil é considerada indicador da qualidade de vida e de desenvolvimento de uma população. Taxas maiores que 50‰ (50 óbitos infantis por mil nascidos vivos) são consideradas altas

e indicam níveis precários de saúde, condições de vida e de desenvolvimento. Já taxas menores que 20‰ indicam maiores índices de desenvolvimento [DUARTE, 2007].

A mortalidade infantil no país tem diminuído nas últimas décadas. A partir da década de 1970 políticas associadas à expansão da rede assistencial e ampliação da estrutura de saneamento básico, assim como campanhas de vacinação e programas de aleitamento materno, foram determinantes para esta redução. Associado a isso, as quedas no nível de fecundidade explicam a queda da mortalidade, principalmente a partir da década de 80 [IBGE - Evolução e Perspectivas da Mortalidade Infantil no Brasil, 1999].

No entanto, estes índices continuam muito elevados (23,3‰) se comparados com o de países desenvolvidos como a Suécia (2,7‰) e o Canadá (5,1‰) e até com outros países com menor nível de desenvolvimento que o Brasil, como Chile (7,9‰) e Cuba (5,9‰).

BASE DE DADOS

O Banco de Dados em estudo foi criado a partir da Declaração de Nascido Vivo (DN) relativa aos nascimentos do ano de 2006 no Estado do Rio de Janeiro, e da Declaração de Óbito (DO), relativa aos óbitos nos anos de 2006 e 2007 provenientes de nascimentos em 2006, também no Estado do Rio de Janeiro.

A DN possui informações socioeconômicas, assistenciais e biológicas, como: local de nascimento e de residência da mãe, idade da mãe, estado civil da mãe, grau de escolaridade da mãe, ocupação da mãe, duração da gestação, tipo de gravidez, tipo de parto, quantidade de consultas pré-natais, presença ou não de anomalias, peso, índice de Apgar 1 e 5 (indicadores do estado de saúde do recém nascido que será explicado mais à frente) e quantidade de filhos tidos (vivos ou mortos), além de características como sexo e raça / cor, data do nascimento e um código de identificação do nascimento, que é único. Em relação à DO o interesse é em informações relativas ao nascimento da criança que morreu.

Baseado em estudos anteriores [OLIVEIRA, 2006 e KOZU et al.], as variáveis consideradas neste trabalho são importantes na classificação de um nascido vivo, quanto ao óbito, ou não, antes de completar o primeiro ano de vida, pois fornecem informações biológicas, socioeconômicas e assistenciais, que são determinantes no estudo da mortalidade infantil.

Uma análise exploratória dos dados mostra que variáveis como o peso, a duração da gestação e os índices de Apgar têm grande influência na chance de sobrevivência de um recém nascido.

ANÁLISE DA MORTALIDADE INFANTIL NO RIO DE JANEIRO EM 2006

O objetivo desta parte do trabalho é descobrir o quanto é possível separar os recém nascidos quanto à chance de sobrevivência ou óbito infantil. Para isso utilizou-se um método de análise multivariada (análise discriminante) e outro de redes neurais.

No método de análise discriminante, utilizou-se a Função de Discriminação de Fisher para criar uma regra de classificação da propensão da criança ao óbito antes do primeiro ano de vida ou não.

As variáveis utilizadas para a estimação da Função Discriminante foram: Sexo, Tipo de Parto, Duração da Gestação, Tipo de Gravidez, Quantidade de Consultas, Índices de Apgar no primeiro e no quinto minuto, Indicador de Anomalia, Peso ao Nascer, Idade, Estado Civil e Escolaridade da Mãe. Todas as variáveis utilizadas são do tipo categórico. As do tipo contínuo, como Peso, Idade da Mãe e os Índices de Apgar, foram transformadas em variáveis categóricas, pois a função discriminante pode perder poder quando existem variáveis independentes de tipos contínuas e categóricas, como é o caso [JOHNSON, 1998].

A seleção das variáveis foi feita através do procedimento Stepwise no software SPSS. Apesar de o teste M-Box rejeitar a hipótese de igualdade das matrizes de covariância (p-valor < 0,001), o teste é muito sensível à hipótese de normalidade, e por isso as diferenças podem ser devido a não normalidade e não à heterogeneidade das covariâncias.

Dado que sabemos de qual população cada observação é proveniente, um bom método para avaliar se o poder discriminatório da função é bom é verificar se a observação foi alocada corretamente ou não. Para isso a tabela 1 mostra os resultados da classificação.

TABELA 1 – RESULTADOS DA CLASIFICAÇÃO POR ANÁLISE DISCRIMINANTE.

Resultados da Classificação[a,b]

				Grupo Predito		
			FL_MI	SOBREVIVENTE	ÓBITO	Total
Casos Selecionados	Original	Quantidade	SOBREVIVENTE	103923	7498	111421
			ÓBITO	508	998	1506
		%	SOBREVIVENTE	93,3	6,7	100,0
			ÓBITO	33,7	66,3	100,0
Casos Não Selecionados	Original	Quantidade	SOBREVIVENTE	69783	4891	74674
			ÓBITO	353	665	1018
		%	SOBREVIVENTE	93,5	6,5	100,0
			ÓBITO	34,7	65,3	100,0

a. 92,9% dos casos selecionados classificados corretamente
b. 93,1% dos casos não selecionados classificados corretamente

Os dados foram divididos em duas amostras aleatórias. Uma composta por 60% dos dados, os casos selecionados da tabela 1, é a amostra de análise, pois foi utilizada para desenvolver a função discriminante. A segunda amostra é composta pelos 40% restantes dos dados, os casos não selecionados da tabela 1, e é chamada de amostra de teste e é relativa às observações e é utilizada para avaliar o desempenho da função discriminante. Uma vez que uma observação é usada para calcular o centróide do grupo, a probabilidade desta observação ser alocada neste grupo aumenta, e há uma tendência a alocar mais objetos corretamente na amostra de análise. O percentual de acertos total é maior na amostra de análise do que na amostra de teste, contrariando o que foi visto acima. Entretanto, no grupo dos óbitos, que é o objeto de estudo, o percentual de acertos é ligeiramente maior na amostra de análise. Por isso a

análise da amostra de teste é mais importante para analisar o poder da função. A tabela 2 mostra o percentual de erros de má classificação da amostra de teste.

TABELA 2 - ERROS DE MÁ CLASSIFICAÇÃO DA AMOSTRA DE TESTE (%).

AMOSTRA DE TESTE		
	GRUPO PREDITO	
GRUPO ORIGINAL	SOBREVIVENTE	ÓBITO
SOBREVIVENTE	93,5%	6,5%
ÓBITO	34,7%	65,3%

O percentual total de acerto de classificação na amostra de teste, em geral, é alto 93,1%. A classificação dos óbitos também se mostrou satisfatória (65%), se mostrando de acordo com a literatura de análise discriminante aplicada a estes tipos de dados [OLIVEIRA, 2006].

Através do teste Q de Press [HAIR et al. 1998], rejeita-se a hipótese nula de que a matriz de classificação do modelo é a mesma que a classificação feita com 50% de chances.

Será feita uma segunda análise utilizando Redes Neurais para verificar se é possível encontrar uma forma de separar melhor os grupos de sobreviventes e de óbitos infantis. Caso haja alguma estrutura não-linear inerente ao processo, é esperado que Redes Neurais tenham desempenho superior a modelos lineares [HAYKIN, 2001].

As variáveis foram as mesmas utilizadas na análise anterior. A rede utilizada é do tipo *feedforward*, pois não há retroalimentação, o processamento é feito somente no sentido da entrada para a saída. Além disso, devido à necessidade do problema foram utilizadas duas camadas: uma camada oculta e uma de saída. A camada oculta possui oito neurônios (número ótimo de neurônios encontrado pelo software SPSS) e a camada de saída possui apenas um neurônio, pois se trata de uma variável resposta binária: 1 óbito; 0 sobrevivente.

A tabela 3 mostra os resultados da classificação estimada pela rede neural. Ela está dividida em amostra de treinamento e de teste, mostrando seu respectivo desempenho.

A amostra de treinamento é usada no aprendizado da rede neural, enquanto que a amostra de teste avalia a real desempenho da rede. Desta maneira a amostra de treinamento pode ficar viciada e deixar de ser capaz de prever corretamente o comportamento dos dados. Um exemplo disso é que o percentual de acerto na amostra de treinamento é ligeiramente maior do que o da amostra de teste. Por isso será interessante analisar a amostra de teste, conforme constante na tabela 4.

O percentual total de acerto é de 98,8%, entretanto a taxa de erro para o grupo de óbitos é muito grande, cerca de 70%. Esta taxa é tão grande que chega a ser inadmissível, uma vez que menos de um terço do total de óbitos foi classificado corretamente.

TABELA 3 – RESULTADO DA CLASSIFICAÇÃO POR REDES NEURAIS.

Classificação

Amostra	FL_MI	Grupo Predito		
		SOBREVIVENTE	ÓBITO	Percentual Correto
Treinamento	SOBREVIVENTE	111.195	163	99,9%
	ÓBITO	1.051	486	31,6%
	Percentual geral	99,4%	0,6%	98,9%
Teste	SOBREVIVENTE	73.977	116	99,8%
	ÓBITO	684	295	30,1%
	Percentual geral	99,5%	0,5%	98,9%

TABELA 4 - ERROS DE MÁ CLASSIFICAÇÃO DA AMOSTRA DE TESTE (%).

AMOSTRA DE TESTE

	GRUPO PREDITO	
GRUPO ORIGINAL	SOBREVIVENTE	ÓBITO
SOBREVIVENTE	99,8%	0,2%
ÓBITO	69,9%	30,1%

COMPARAÇÃO DOS RESULTADOS

Uma maneira de avaliar a eficiência de diferentes técnicas de reconhecimento de padrões é verificar a taxa de classificação correta de cada modelo. Além disso, na análise discriminante foram propostas taxas de classificação para verificar a qualidade da discriminação estimada. Assim, a tabela 5 mostra as taxas de classificação correta, total e para os grupos, para cada um dos métodos.

TABELA 5 – COMPARAÇÃO DA TAXA DE ERRO DE CLASSIFICAÇÃO POR DIFERENTES MÉTODOS DE CLASSIFICAÇÃO.

	Taxa de Erro de Classificação (%)		
	Chance Máxima	Análise Discriminante	Rede Neural
Sobreviventes	0,0	6,5	0,2
Óbitos	100,0	34,7	69,9
Total	1,3	6,9	0,1

Analisando a tabela percebe-se que a Rede Neural foi responsável pela menor taxa de erro de classificação total (0,1%). Entretanto, seu poder explicativo para o grupo de óbitos é muito baixo uma vez

que a taxa de erro foi de 70%. Já a análise discriminante possui a menor taxa de erro para o grupo de óbitos. Dessa maneira, apesar de possuir a maior taxa de erro total, este é o método que se mostrou mais eficiente na classificação dos óbitos infantis no Estado do Rio de Janeiro no ano de 2006.

CONCLUSÕES

Os objetivos deste trabalho foram alcançados. Agora se sabe com mais detalhes como cada fator influencia na chance de sobrevivência de um recém nascido e se conhecem dois métodos de classificação da criança quanto ao óbito infantil ou não.

A análise Discriminante foi capaz de classificar corretamente 65% dos óbitos. A Rede Neural, entretanto teve uma classificação muito pobre, de apenas 30%. O que levou à conclusão que, neste caso, a Análise Discriminante foi mais eficiente para prever os óbitos infantis.

A Rede Neural teve uma performance muito pobre em relação aos óbitos apesar de ter tido uma baixa taxa de erro dos dados em geral. Entretanto, como já foi dito anteriormente, existem diferentes tipos de estruturas de redes neurais, que podem ser usadas para a resolução de diversos problemas complexo. Por isso, em trabalhos futuros, seria interessante testar outras arquiteturas de redes neurais para verificar se há uma melhora no desempenho. Outra alternativa para testar uma melhor performance seria utilizar um Modelo de Regressão Logística que também é largamente utilizado para reconhecimento de padrões.

REFERÊNCIAS BIBLIOGRÁFICAS

BANDEIRA, Mário L.. *Demografia: Objecto, teorias e métodos*. Lisboa: Escolar Editora, 2004

DUARTE, Cristina M. R. Reflexos das políticas de saúde sobre as tendências da mortalidade infantil no Brasil: revisão da literatura sobre a última década**.** Rio de Janeiro, 2007. Disponível em: http://www.scielo.br/pdf/csp/v23n7/02.pdf

HAIR, Joseph F.; TATHAM, Ronald L.; ANDERSON, Rolph E.; BLACK William. *Multivariate Data Analysis*. 5th ed. Prentice-Hall Inc, 1998.

HAYKIN, Simon. *Redes Neurais: Princípios e Práticas*. 2ªed. Porto Alegre: BOOKMAN, 2001.

IBGE. Departamento de População e indicadores sociais. *Evolução e Perspectivas da Mortalidade Infantil no Brasil.* Rio de Janeiro, 1999. Disponível em:
http://www.ibge.gov.br/home/estatistica/populacao/evolucao_perspectivas_mortalidade/evolucao_mortalidade.pdf

JOHNSON, Richard A.; WICHERN, Deam W. *Applied multivariate statistical analysis*. 4th ed. New Jersey : Prentice-Hall Inc, 1998.

KOZU, Kátia T; GODINHO, L.T.; MUNIZ ; M.V.F.; CHIARIONI, P. *Mortalidade Infantil: Causas e Fatores de Risco. Um Estudo Bibliográfico.* Disponível em:
http://www.medstudents.com.br/original/original/mortinf/mortinf.htm

OLIVEIRA, Ivan de. *Função discriminante quadrática aplicada no reconhecimento e classificação de nascidos vivos quanto à sobrevivência ou óbito no primeiro ano de vida.* Curitiba, 2006. 117f. Dissertação (Mestrado em Ciências - Programa de Pós-Graduação em Métodos Numéricos em Engenharia dos departamentos de Matemática e de Construção Civil), Universidade Federal do Paraná.