PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
ESCOLA POLITÉCNICA
BACHARELADO EM MATEMÁTICA

Artur Lima da Silveira

TOOL PREDITIVO EM DIAGNÓSTICO DE ALZHEIMER VIA REGRESSÃO LOGÍSTICA MULTIVARIADA

Porto Alegre
2020

Artur Lima da Silveira

TOOL PREDITIVO EM DIAGNÓSTICO DE ALZHEIMER VIA REGRESSÃO LOGÍSTICA MULTIVARIADA

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do grau de Bacharel em Matemática: Linha de formação em Matemática Empresarial pela Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Orientadora: Profª Drª Eliete Biasotto Hauser

Porto Alegre
2020

ARTUR LIMA DA SILVEIRA

TOOL PREDITIVO EM DIAGNÓSTICO DE ALZHEIMER VIA REGRESSÃO LOGÍSTICA MULTIVARIADA

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do grau de Bacharel em Matemática: Linha de formação em Matemática Empresarial pela Escola Politécnica da Pontifícia Universidade Católica do Rio Grande do Sul.

Aprovada em: 08 de julho de 2020.

BANCA EXAMINADORA:

_____

Rossana Fraga Benites

_____

Wyllians Vendramini Borelli

_____

Augusto Vieira Cardona

AGRADECIMENTOS

Deixo meus agradecimentos aos meus pais, Edilson e Taís, por incentivar e proporcionar minha formação.

Ao meu melhor amigo e irmão Paulo Correia da Silveira Neto que se dispôs prontamente sempre que necessário, assim como sua esposa Sheyla Werner.

A minha namorada Camila Pagliarini, pelo cuidado e carinho diários.

Agradeço a minha orientadora, professora Eliete Biasotto Hauser pelo auxílio e compreensão proporcionados em meio à pandemia, momento em que o trabalho dos professores aumentou consideravelmente.

A minha família e aos amigos do Luzibr pelo apoio emocional.

Aos meus colegas e amigos do trabalho, por todos os conselhos e compreensão da correria.

“A matemática, vista corretamente, possui não apenas verdade, mas também suprema beleza – uma beleza fria e austera, como a de uma escultura”.

Bertrand Russell

RESUMO

O objetivo deste trabalho é o desenvolvimento de uma metodologia para a criação de um instrumento de detecção da doença de Alzheimer através da análise de quarenta e duas variáveis que caracterizam uma amostra de quarenta e um pacientes voluntários do projeto Superidosos desenvolvido no Instituto do Cérebro do Rio Grande do Sul (InsCer). Para tanto, identificou-se quais variáveis da amostra são mais relevantes para explicar o fato de o paciente ter ou não a doença. Foram usados modelos de regressão logística multivariada e análise de componentes principais, assim como as técnicas de validação do modelo desenvolvido. Essa escolha fundamenta-se na literatura disponível a respeito de tools preditivos para diagnósticos de doenças, análise de riscos e outras áreas. A amostra caracteriza-se por testes cognitivos, informações clínicas e hábitos pessoais. Foram selecionados onze testes cognitivos para o modelo de regressão logística multivariada e os resultados obtidos indicam que os mesmos possuem forte correlação entre si. Na sequência, com análise de componentes principais, foi possível utilizar todos os onze testes, tornando-os estatisticamente significativos para estimar a probabilidade de o paciente ter a doença de Alzheimer.

Palavras-chave: Modelagem Matemática. Regressão Logística Multivariada. Análise de Componentes Principais. Alzheimer.

SUMÁRIO

1. INTRODUÇÃO

A doença de Alzheimer é uma doença neurodegenerativa crônica que se apresenta principalmente como demência, conceituada como um processo patológico progressivo e incurável e, até o momento, não possui qualquer estratégia para alteração do curso natural, concessão ou prevenção da doença (BORELLI, 2019).

Muitas vezes relacionada com a idade, o Alzheimer causa deterioração cognitiva da memória de curto prazo podendo causar alterações de comportamento, assim como déficit de aprendizado e memória, sendo a forma mais prevalente de demência. Mesmo não sendo uma parte típica do envelhecimento, o fator de risco aumenta significativamente a partir dos 65 anos de idade, onde atinge um terço dos idosos acima de 85 anos e com aproximadamente 473.000 casos novos por ano no mundo (BORELLI, 2019; HAUSER et al. 2019 - 2020). No início costuma manifestar-se como perda leve de memória evoluindo progressivamente até interferir diretamente na capacidade de comunicação e de reação ao meio ambiente (NITRINI et al., 2005).

Não existe uma razão específica para o despertar da doença, olhando as médias dos casos ocorridos, é possível afirmar que a possibilidade é maior conforme o andar do envelhecimento e geralmente em mulheres. Acredita-se também que há uma relação genética, ou seja, as probabilidades aumentam quando outros integrantes da família apresentam a doença, porém, não é suficientemente relevante para afirmar que uma pessoa terá a doença por este motivo. Alguns fatores clínicos também podem aumentar as chances, tais como insuficiência cardíaca, hipertensão arterial ou até mesmo algum dano à região encefálica (NITRINI et al., 2005).

Esse trabalho de pesquisa tem como principal objetivo utilizar modelagem matemática e análise estatística para o desenvolvimento de um instrumento auxiliar no diagnóstico de Alzheimer, cuja resposta permita estabelecer a probabilidade de o paciente observado pertencer a um grupo previamente determinado como portador da doença. A partir de um conjunto de quarenta e duas variáveis independentes, serão aplicadas técnicas para decidir a melhor forma de observar os fatores relevantes levando em consideração os testes cognitivos aplicados, informações clínicas e hábitos pessoais.

Utilizando linguagem de programação R (VRIES, 2015) e o *Analytics Software Solutions* (SAS) (COLLUM, 2019) serão utilizados os dados do projeto Superidosos (Correlação Entre Neuroimagem Molecular, Estrutural e Funcional em Superidosos) para construção do instrumento de detecção do Alzheimer. Foram analisados referenciais teóricos e para a proposição de uma metodologia de desenvolvimento de um modelo com sua base em regressão logística multivariada, tendo como variável resposta a presença ou ausência de Alzheimer de uma amostra de pacientes.

Nesse estudo, a técnica de regressão logística multivariada foi escolhida fundamentada em tools preditivos para diagnósticos de doenças disponíveis na literatura, dentre os quais destaca-se o Gupta (2019).

No capítulo 2 é apresentado o referencial teórico do trabalho, começando pela regressão logística multivariada e passando brevemente por análise de correlação, multicolinearidade e análise de componentes principais. No terceiro capítulo comparecem os dados utilizados na amostra, a análise de correlação de Pearson com uma breve descrição das variáveis com forte correlação com a variável dependente, os parâmetros calculados pela regressão logística multivariada e uma validação do modelo criado. O capítulo 4 apresenta uma proposta de melhorias para o modelo, fazendo ajustes de multicolinearidade através da análise de componentes e ajustes no ponto de corte. Na conclusão há uma análise comparativa dos tools criados, e as sugestões para continuidade do estudo.

2. FUNDAMENTAÇÃO TEÓRICA

Modelos preditivos estão cada vez mais presentes em nosso cotidiano devido ao aumento de dados históricos e o avanço da tecnologia. A fim de modelar situações teóricas através de números, a matemática é utilizada em muitas áreas e, na medicina, pode auxiliar em decisões clínicas mais quantitativas do que qualitativas, ajudando em diagnósticos e tratamento, como também a prever o crescimento das doenças, análises de epidemias, evolução de diagnósticos por imagem e até mesmo no diálogo com os pacientes quando precisam ser estimadas as chances de complicações e óbitos de procedimentos cirúrgicos (MASSAD et al., 2004).

2.1. Modelo logístico: Análise Multivariada

Na literatura, a regressão logística é uma das mais recomendadas para alocar um objeto em uma classe. É um método estatístico multivariado pelo fato de relacionar um conjunto de variáveis independentes com uma variável resposta categórica (HAIR et al., 2005; HAUSER 2018).

As técnicas de discriminação buscam estimar a probabilidade de uma variável dependente binária assumir um determinado valor em função de outras variáveis, tentando assim, encontrar uma função ou conjunto de funções que discrimine os grupos definidos pela variável binária, minimizando erros, tendo como principal diferença da regressão múltipla, o fato de não pressupor existência de normalidade dos resíduos e homogeneidade de variância (LEMESHOW; HOSMER, 1989).

Há uma infinidade de situações para as quais a regressão logística possui utilidade prática, como por exemplo para a modelagem de inadimplência, ocorrência de uma doença, de um sinistro ou até mesmo de óbito, por isso, se tornou uma técnica muito popular principalmente em bancos para criar modelos de risco de crédito (Credit Scoring) (LEMESHOW; HOSMER, 1989).

O uso do modelo logístico multivariado neste estudo visa estimar a probabilidade de uma pessoa ser portadora da doença de Alzheimer dado um conjunto de informações

composto por variáveis de comportamento, clínicas e resultantes de testes cognitivos. Neste caso, o evento de interesse é possuir a doença (variável dependente), cuja ocorrência é representada por 1, e a não ocorrência por 0. Enquanto isso, um exemplo de variável a idade presente no modelo, a qual se apresenta como uma variável discreta entre 0 e 100. Podem ser introduzidas também outras variáveis no modelo as quais estariam, de alguma forma, relacionadas ao Alzheimer.

Nesse estudo, considera-se a função logística escrita como:

$$f(Z) = \frac{1}{1 + e^{-(Z)}} \qquad (1)$$

Na Eq.(1):

$$Z = \ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad (2)$$

na qual P é a probabilidade de ocorrência do evento de interesse, X determinado $X = [X_1, X_2, \ldots X_k]$ representa o vetor de variáveis independentes e α e β os parâmetros do modelo a serem estimados.

A função $f(Z)$ é entendida como a probabilidade da variável dependente ser 1 ou 0, conhecidas as variáveis independentes X_i , e os parâmetros calculados α, e β_i sendo $i = 1 \ldots k$

O método para estimar α e β_i é o método de máxima verossimilhança, que maximiza a probabilidade de ocorrência do evento. O objetivo principal de estimar estes parâmetros é encontrar a função logística a qual as ponderações das variáveis independentes evidenciem a correlação de cada uma delas com a ocorrência do evento de interesse, bem como calcular a probabilidade de ocorrência do evento (HOSMER; LEMESHOW, 1989).

Para alcançar tal objetivo, inicia-se convertendo a probabilidade de cada evento em razão de chance *odds ratio* (OR), visto como a probabilidade de ocorrência do evento (p) comparada com a probabilidade de não ocorrer $(1 - P)$. Ou seja, a razão entre as chances (OR) da ocorrência do Alzheimer, em relação a não ocorrência do Alzheimer

(MEZZOMO, 2009). Isto permite descobrir a chance de o evento ocorrer, em relação a ele não ocorrer sob as mesmas condições.

Obtendo assim o logaritmo natural da razão de chance:

$$Logit = \ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_k X_k \qquad (3)$$

No início da equação tem-se o logaritmo natural da razão de chance e, após a igualdade, as variáveis independentes e coeficientes que expressam mudanças no *ln* da razão de chance. (HOSMER; LEMESHOW, 1989).

Se X for uma variável categórica binária, assumindo $X = 1$ se tem o impacto da presença de um coeficiente sobre a outra e tem-se o coeficiente exponencial OR $= e^{b_1}$ onde $e \cong 2.718$, chamada constante matemática de Euler, a qual é a base do logaritmo neperiano utilizado no modelo de regressão logística multivariada.

Assim, aplicando a função exponencial em ambos os lados da Eq(3) ao expoente composto dos coeficientes estimados, sendo que o modelo logístico tenha sido ajustado a um conjunto de dados, obtem-se a razão de chance estimada.

$$e^{logit} = \left(\frac{P}{1-P}\right) = e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)} \qquad (4)$$

Com a razão de chance estimada, identifica-se a influência de um determinado evento sob a ocorrência relacionada. Após simplificar o modelo de regressão logística, a equação fica expressa como:

$$P = \frac{e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}{1+e^{(\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}} = \frac{1}{1+e^{-(\ln odds)}} \qquad (5)$$

A partir da função acima é possível entender o significado de β e α, sendo α o logaritmo natural da chance quando todas as variáveis explicativas são nulas e β_k

representa a mudança do logaritmo natural da chance dada a variação de uma unidade na variável X.

Os coeficientes logísticos β_k estimados são medidas das variações na proporção das probabilidades, também conhecidos como razão de desigualdade. Estes são expressos em logaritmos, necessitando serem transformados para melhor interpretação.

A função logística é aplicada não apenas pela simplicidade de suas propriedades teóricas, mas também, devido a sua simples interpretação como o logaritmo da razão de chance (odds) (HOSMER; LEMESHOW, 1989).

A construção do modelo de regressão logística é realizada utilizando o método de máxima verossimilhança. O método consiste em encontrar o conjunto de parâmetros com a maior probabilidade de ter gerado a amostra utilizada na estimação. O processo (WILKS S. S., 1938) consiste em realizar a otimização da seguinte equação:

$$\hat{\beta}_{mle} = \max_{\beta} \ln\left(\prod_{i=1}^{n} f(X_i, \alpha, \beta_i)\right) \qquad (6)$$

Na Eq(6), $\hat{\beta}_{mle}$ é o estimador de máxima verossimilhança (*maximum likelihood estimator – mle*), $f()$ a função de densidade de probabilidade dos dados, α e β os parâmetros dessa função e X_i os valores observados da amostra. No caso da regressão logística utilizada, a variável suposta é a distribuição de Bernoulli (WILKS, 1938), que é definida como:

$$f(X; \sigma) = \sigma^X (1-\sigma)^{1-X}, para\ X \in [0; 1] \qquad (7)$$

2.2. Análise de Correlação

Situado sempre com valores entre -1 e 1, o coeficiente de correlação de Pearson mede o grau de correlação linear entre duas variáveis quantitativas, o índice reflete a força da relação linear entre duas variáveis distintas (MANN, 2010).

Quando o coeficiente é positivo, afirma-se que há uma correlação positiva entre as duas variáveis, ou seja, quando uma aumenta, a outra aumenta também, como por exemplo analisando a correlação entre a idade e a Alzheimer, conforme aumenta a idade, aumenta também a possibilidade de o indivíduo possuir a doença. Já quando o coeficiente é negativo, há uma correlação negativa, isto é, se uma aumenta a outra diminui, por exemplo em um teste cognitivo de memória, quanto maior o escore, menor a probabilidade de possuir Alzheimer.

Quando o coeficiente de Pearson é zerado, significa que as variáveis não dependem linearmente uma da outra, no entanto pode existir uma outra dependência que não seja linear, uma correlação de Pearson superior a 0,7 ou inferior a -0,7 são consideradas correlações fortes (MANN, 2010).

2.3. Multicolinearidade

Um dos possíveis problemas do processo de estimação dos modelos de regressão logística é a multicolinearidade. Tal evento ocorre quando há alta correlação entre as variáveis explicativas do modelo. Na presença de multicolinearidade, variáveis independentes com relações lineares significativas com a variável resposta podem apresentar elevado *p-valor.* O *p-valor* é definido como a probabilidade de se observar um valor da estatística de teste que excede ao encontrado e que o valor de corte usual é e 0,05. A multicolineraidade pode existir apesar do R^2 e do diagnóstico global da regressão indicarem que o modelo explica a variável resposta (Maia, 2019).

A detecção da multicolinearidade pode ser realizada de diversas maneiras. Dentre elas estão a elaboração de uma matriz de correlação das variáveis independentes e o cálculo do VIF (*variance inflation factor*). Na matriz de correlação, verifica-se se não há correlações maiores que 70% entre as variáveis independentes. Já o VIF mede a dependência das variáveis independentes com a seguinte fórmula (Maia, 2019):

$$VIF_i = \frac{1}{1 - R_i^2} \tag{8}$$

Na Eq(8), R_i^2 é o coeficiente de determinação (R^2) da regressão da variável X_i contra todas as outras variáveis independentes. Como regra de bolso, a literatura sugere que um VIF maior que 5 (Maia, 2019) indica presença de multicolinearidade entre as variáveis independentes.

2.4. Análise de Componentes Principais

A análise de componentes principais é uma técnica de redução de dimensionalidade que consiste em encontrar um novo conjunto de variáveis formado por combinações lineares das variáveis coletadas. O novo conjunto é formado necessariamente por variáveis independentes, além disso, contém o mesmo número de variáveis e toda a informação que as variáveis iniciais possuíam (JOHNSON AND WICHERN, 1998).

No processo, a primeira variável é construída de forma a conter o máximo da variabilidade do conjunto de dados inicial, a segunda é formada da mesma maneira, mas com a restrição de ser ortogonal à primeira variável. As próximas variáveis são encontradas tais que sejam independentes a todas as variáveis anteriores. O gráfico (Figura 1) a seguir representa o caso mais simples no qual duas novas variáveis (PC1 e PC2) são compostas como combinações lineares pelas variáveis 1 e 2:

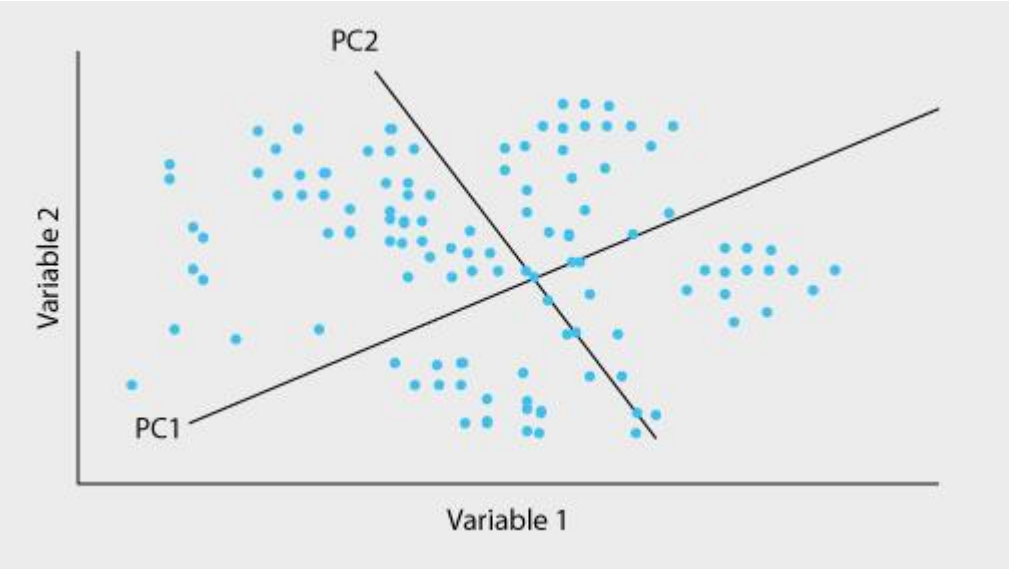

Figura 1 - Análise de componentes principais

A principal vantagem desta técnica é que, como os primeiros componentes possuem a maior parte da variabilidade dos dados, possibilita-se a estimação de modelos com menos variáveis independentes sem que ocorra perda das informações (JOHNSON AND WINCHERN, 1998).

Para Λ_i os autovalores da matriz de covariância X, ordenados em forma decrescente, e V_i os correspondentes auto vetores (ortonormalizados), o j-ésimo componente principal é dado por $CP_j = V_i * X, j = 1,2 \dots k$.

3. APLICAÇÃO: DADOS DO PROJETO SUPERIDOSOS

A amostra feita com 41 indivíduos, provenientes de consultórios, dos ambulatórios do Serviço de Neurologia do Hospital São Lucas da PUCRS, do Centro de Extensão Universitária Vila Fátima da PUCRS e da própria sociedade, com idade superior a 50 anos, alfabetizados e assinantes do Termo de Consentimento Livre e Esclarecido (TCLE)(Superidosos 2019, Alzheimer's Team 2019), está detalhada na tese de doutorado Correlação Entre Neuroimagem Molecular, Estrutural e Funcional em Superidosos (BORELLI, 2019).

O diagnóstico utilizado na amostra para ser considerado portador da doença de Alzheimer foram os critérios médicos de IWG, os quais são compostos basicamente por coleta de informações e exame clínico. Após isso, pacientes enquadram-se na síndrome demencial amnésica (memória que caracteriza a doença de Alzheimer) e a ressonância é típica para essa doença. Pela diversidade de definições da doença de Alzheimer, não é possível descrever um teste que seja específico para diagnosticar a doença (BORELLI, 2019).

3.1. Análise de Correlação

Devido ao tamanho da amostra, primeiramente foi realizada uma análise de correlação de Pearson entre as variáveis independentes e a variável Resposta, a qual foi criada como uma variável binaria que apresenta 1 para pacientes pré-diagnosticados com Alzheimer e 0 para pacientes que não são portadores da doença. Utilizou-se o software SAS (COLLUM,2019) para identificar as variáveis que contém uma correlação forte com a variável resposta. Os resultados obtidos são referenciados no quadro 1.

N	Nome	Coeficiente de Correlação
1	ACER_M	-0,88
2	MOCA	-0,87
3	A7	-0,85
4	RAVLT_S	-0,84
5	MMSE	-0,83
6	ACER_AO	-0,82
7	A6	-0,78
8	ACER_F	-0,74
9	TMT_B	0,74
10	ACER_V	-0,73
11	CFT	-0,7
12	BNT	-0,66
13	B1	-0,55
14	TMT_A	0,53
15	ACER_L	-0,51
16	FAS	-0,48
17	INTERNET	-0,39
18	DS_T	-0,39
19	IDADE	0,36
20	DS_I	-0,35
21	CONVSOCIAL	-0,33
22	IDIOMA	-0,32
23	CARNE	0,23
24	DS_D	-0,2
25	ESC	-0,18
26	CHIMA	-0,17
27	LIVROS	-0,17
28	PESO	-0,15
29	VIAGEM	-0,14
30	MUSICA	0,13
31	EXFISICOQUANT	0,12
32	FILMES	-0,11
33	PCRUZADAS	-0,1
34	SONO	0,07
35	LEGUMES	0,07
36	CAFE	0,07
37	FRUTAS	-0,06
38	CHA	-0,06
39	ALTURA	0,05
40	EDG_15	-0,05
41	RENDAFAM	0,05
42	COHABIT	-0,01

Quadro 1

Das 42 variáveis analisadas, apenas onze apresentaram correlação forte com a variável dependente onde, segundo MALAWI (2012), uma correlação de Pearson superior a 0,7 ou inferior a -0,7 são consideradas correlações fortes.

No Quadro 2 constam as onze variáveis selecionadas, com suas características descritivas: Mínimo, Máximo, Média, Mediana, Moda e Desvio Padrão são referentes aos dados apresentados na amostra. Portanto no Quadro 2, Mínimo é o menor valor apresentado na amostra para aquela variável, Máximo é o maior valor apresentado na amostra para aquela variável e assim sucessivamente.

O detalhamento amplo de todas as variáveis incluindo a pontuação máxima possível, pode ser encontrado na tese de doutorado intitulada Correlação Entre Neuroimagem Molecular, Estrutural e Funcional em Superidosos (BORELLI, 2020).

X_n	Descrição	Nome	Mínimo	Máximo	Média	Mediana	Moda	Desvio Padrão
X_1	Addenbrooke's Cognitive Rvaluation - Revised - Parte de Memória	ACER_M	4	26	18,20	20	20	6,43
X_2	Montreal cognitive assessment (teste geral de funções cognitivas)	MOCA	7	30	23,43	25	27	5,54
X_3	Rey auditory-verbal learning test (teste de memória) - lista A7	A7	0	15	7,37	8	0	4,86
X_4	Rey auditory-verbal learning test (teste de memória) - soma das listas A1 até A5	RAVLT_S	4	60	39,33	41	45	14,72
X_5	Mini Mental State Examination	MMSE	10	30	26,66	29	30	4,70
X_6	Addenbrooke's Cognitive Rvaluation - Revised - Parte de Atenção e Orientação	ACER_AO	5	18	16,17	18	18	3,41
X_7	Rey auditory-verbal learning test (teste de memória) - lista A6	A6	0	15	7,22	8	0	4,52
X_8	Addenbrooke's cognitive evaluation - revised - parte de Fluência	ACER_F	3	14	9,85	10	10	3,24
X_9	Trail Making Test, Parte B	TMT_B	44	300	160,49	141	300	88,85
X_{10}	Addenbrooke's cognitive evaluation - revised - parte Visuoespacial	ACER_V	8	16	14,61	16	16	2,13
X_{11}	Category Fluence Test	CFT	3	26	15,54	16	13	5,87

Quadro 2

3.2. Cálculo dos Parâmetros do Modelo Logístico Multivariado

Para a análise de regressão logística, foi utilizada a linguagem de programação R (VRIES, 2015). Como foram selecionados anteriormente os parâmetros significativos, não se colocou um critério de saída de variáveis independentes por p-valor, mantendo assim todas as variáveis selecionadas anteriormente no modelo e estimando o β de todas elas como é possível ver no Quadro 3:

X_K	$\hat{\beta}_k$	Parâmetro Estimado
X_1	$\hat{\beta}_1$	-0,4434878
X_2	$\hat{\beta}_2$	-0,1413785
X_3	$\hat{\beta}_3$	0,0123871
X_4	$\hat{\beta}_4$	0,0756546
X_5	$\hat{\beta}_5$	0,4344057
X_6	$\hat{\beta}_6$	0,0005082
X_7	$\hat{\beta}_7$	-0,0138291
X_8	$\hat{\beta}_8$	-0,1230365
X_9	$\hat{\beta}_9$	-0,1530377
X_{10}	$\hat{\beta}_{10}$	0,0097521
X_{11}	$\hat{\beta}_{11}$	-0,0917968
Intercepto	$\hat{\alpha}$	-0,1200864

Quadro 3

Dessa forma, fica completamente calculado o modelo logístico que determinará a probabilidade de o indivíduo ter ou não Alzheimer, dados os testes cognitivos que entraram no modelo, descrito na Eq(9).

$$P(X) = \frac{1}{1 + e^{\begin{array}{c}-(-0{,}120+(-0{,}334)X_1+(-0{,}141)X_2+(0{,}012)X_3+(0{,}076)X_4+(0{,}434)X_5+(0{,}001)X_6 \\ +(-0{,}014)X_7+(-0{,}123)X_8+(-0{,}153)X_9+(0{,}010)X_{10}+(-0{,}092)X_{11})\end{array}}} \quad (9)$$

3.3. Validação do modelo: Análise de resultados

Para validação inicial do modelo, foi aplicada a função logística nos dados coletados originalmente.

Verificou-se que o percentual de acerto acumulado foi de 95,1% tendo como ponto de corte P=0,5, “valor padronizado para regressão logística”. Ou seja, se considerar como certa a presença de Alzheimer para probabilidades superiores a 50% e como certa a ausência de Alzheimer para probabilidades inferiores a 50% o modelo acerta trinta e nove entre todos os quarenta e um casos analisados como demonstrados na Figura 2 e na Quadro 4.

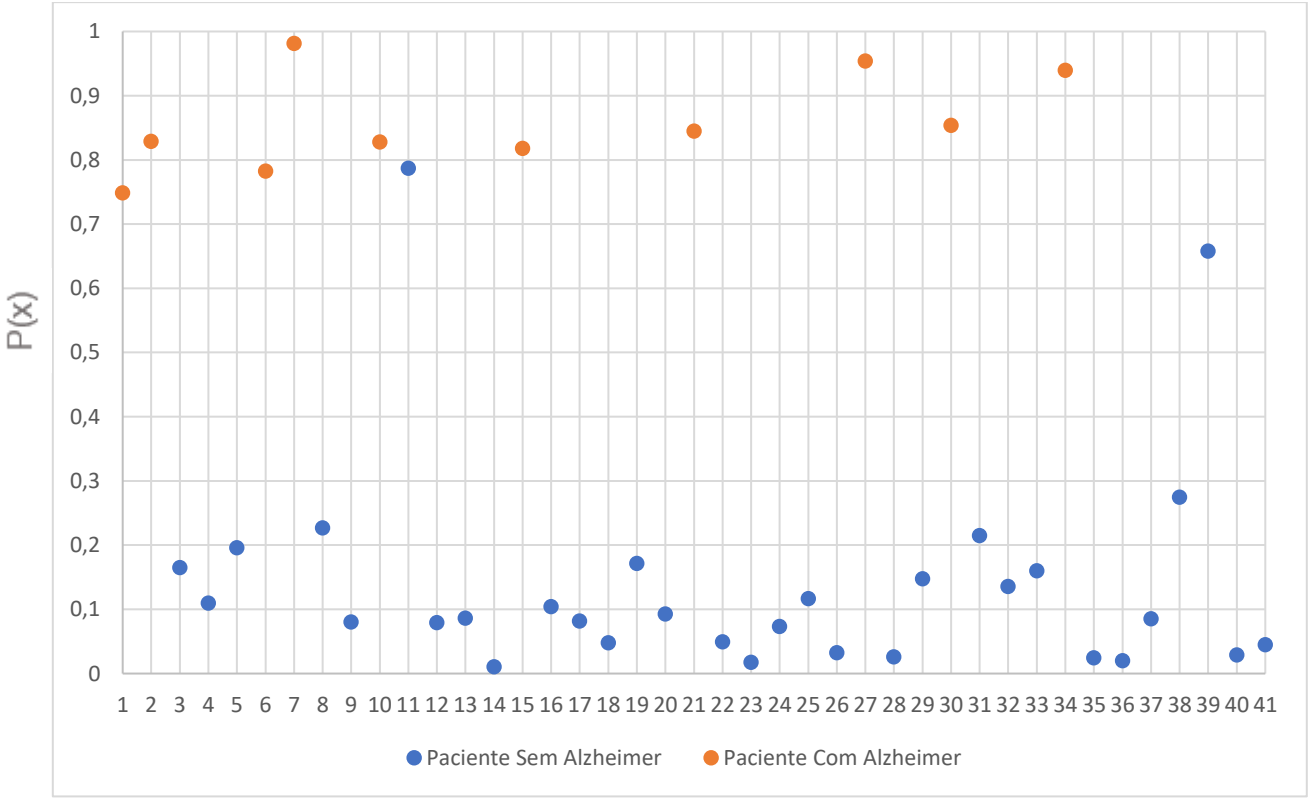

Figura 2

Paciente	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	Variável Resposta	Probabilidade estimada P(X)
1	13	18	0	27	28	18	3	8	103	13	13	1	**0,85**
2	4	14	0	20	15	7	4	7	300	8	6	1	**0,83**
3	18	27	10	36	30	18	9	11	174	16	18	0	**0,17**
4	24	20	6	45	27	17	8	10	182	16	17	0	**0,11**
5	25		13	50	30	18	14	12	57	16	22	0	**0,20**
6	12	18	0	26	18	9	0	8	300	12	12	1	**0,78**
7	5		0	19	18	9	0	3	300	13	3	1	**0,98**
8	25		12	51	30	18	13	11	44	16	16	0	**0,23**
9	19	25	11	56	29	17	11	10	101	14	19	0	**0,08**
10	8	14	1	20	19	10	2	4	300	9	8	1	**0,83**
11	18		3	33	26	17	3	10	143	15	17	0	**0,79**
12	20	26	11	47	30	18	12	14	64	16	23	0	**0,08**
13	23	28	8	37	30	18	10	9	287	15	11	0	**0,09**
14	23	27	15	58	29	18	15	12	100	16	21	0	**0,01**
15	13	17	1	15	24	15	0	9	300	16	13	1	**0,82**
16	23	26	10	38	29	18	5	13	201	16	19	0	**0,10**
17	20	25	10	42	29	18	10	10	118	16	16	0	**0,08**
18	26	29	10	52	30	18	8	14	62	16	24	0	**0,05**
19	22	27	8	51	30	18	5	13	55	16	23	0	**0,17**
20	16	21	10	57	27	18	8	11	163	16	16	0	**0,09**
21	4	7	0	4	10	5	0	3	300	9	8	1	**0,84**
22	20	26	10	52	28	18	10	14	123	16	23	0	**0,05**
23	24	25	11	50	26	16	12	10	176	14	17	0	**0,02**
24	18	29	12	48	29	18	8	14	72	16	24	0	**0,07**
25	19	27	8	38	29	18	9	10	98	15	20	0	**0,12**
26	26	29	10	52	30	18	11	13	82	16	21	0	**0,03**
27	11		0	21	22	13	1	4	300	14	6	1	**0,95**
28	24	29	12	55	30	18	12	12	63	16	18	0	**0,03**
29	21	25	6	39	27	17	5	11	93	16	12	0	**0,15**
30	10	16	0	13	23	14	0	5	239	12	7	1	**0,85**
31	23	25	3	45	29	18	4	13	138	12	19	0	**0,21**
32	21	23	8	24	28	18	5	9	80	14	13	0	**0,14**
33	18	22	9	40	29	18	9	6	141	16	10	0	**0,16**
34	4	14	0	18	22	13	1	6	300	13	8	1	**0,94**
35	20	30	15	60	30	18	13	14	100	16	26	0	**0,02**
36	19	27	13	53	28	18	13	13	89	15	19	0	**0,02**
37	25	29	8	40	30	18	7	12	70	16	14	0	**0,09**
38	16	23	7	39	27	16	7	7	252	14	9	0	**0,27**
39	20		7	45	29	18	7	9	205	16	12	0	**0,66**
40	22	25	14	58	30	18	13	7	162	16	13	0	**0,03**
41	24	27	10	50	29	18	9	13	143	16	21	0	**0,04**

Quadro 4

No processo de regressão logística, para uma melhor análise de desempenho, é importante utilizar amostras diferentes para desenvolvimento e para validação (HAIR, 2005). Bons indicadores de diagnóstico do modelo são: Especificidade - a capacidade do modelo de acertar os pacientes sem Alzheimer. Sensibilidade - a capacidade do modelo de acertar os pacientes que possuem a doença de Alzheimer. Precisão - a capacidade do modelo de acertar tanto quem tem a doença, quanto quem não tem (HAIR, 2005 e GREENHALGH, 1997).

Para fazer tal diagnóstico, utilizando a linguagem de programação R, a amostra total foi separada aleatoriamente em duas, mantendo a proporção de pacientes com Alzheimer e sem em ambas. Para a amostra de desenvolvimento foi separada 70% do total de pacientes, deixando os outros 30% para validação.

Após essa etapa, também utilizando a linguagem de programação R, foi desenvolvido um novo modelo de regressão logística apenas utilizando os 31 pacientes da amostra de desenvolvimento, testado o modelo nos 10 pacientes da amostra de validação e calculados os 3 indicadores citados acima. Como se trata de uma amostra pequena, o que da margem para coincidências, o processo foi replicado 1000 vezes e feita a média dos indicadores, representadas na Figura 3.

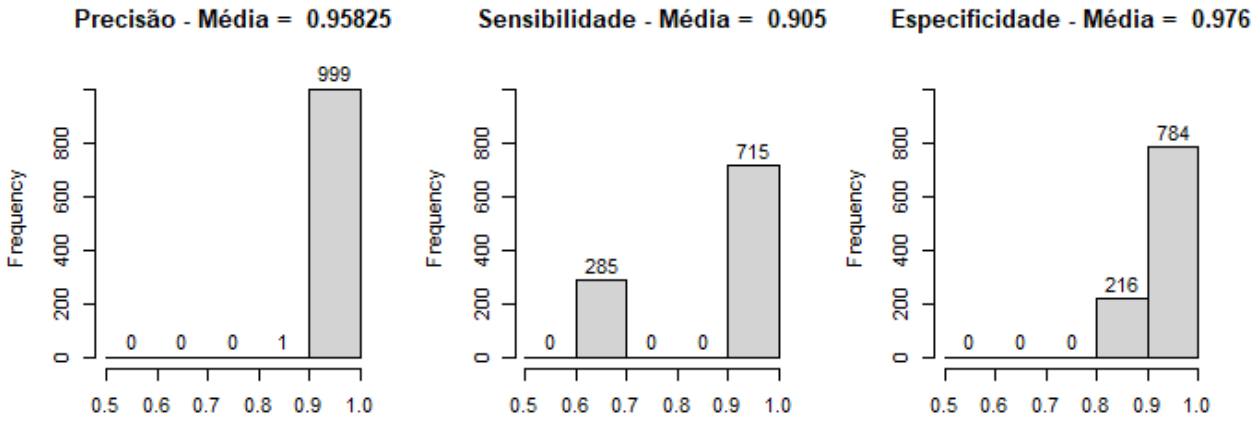

Figura 3

3.4. Tool Preditivo 1

Após o desenvolvimento do modelo, foi construída uma calculadora demonstrada na Figura 4 na ferramenta Microsoft Excel buscando viabilizar o uso do modelo preditivo em novos pacientes.

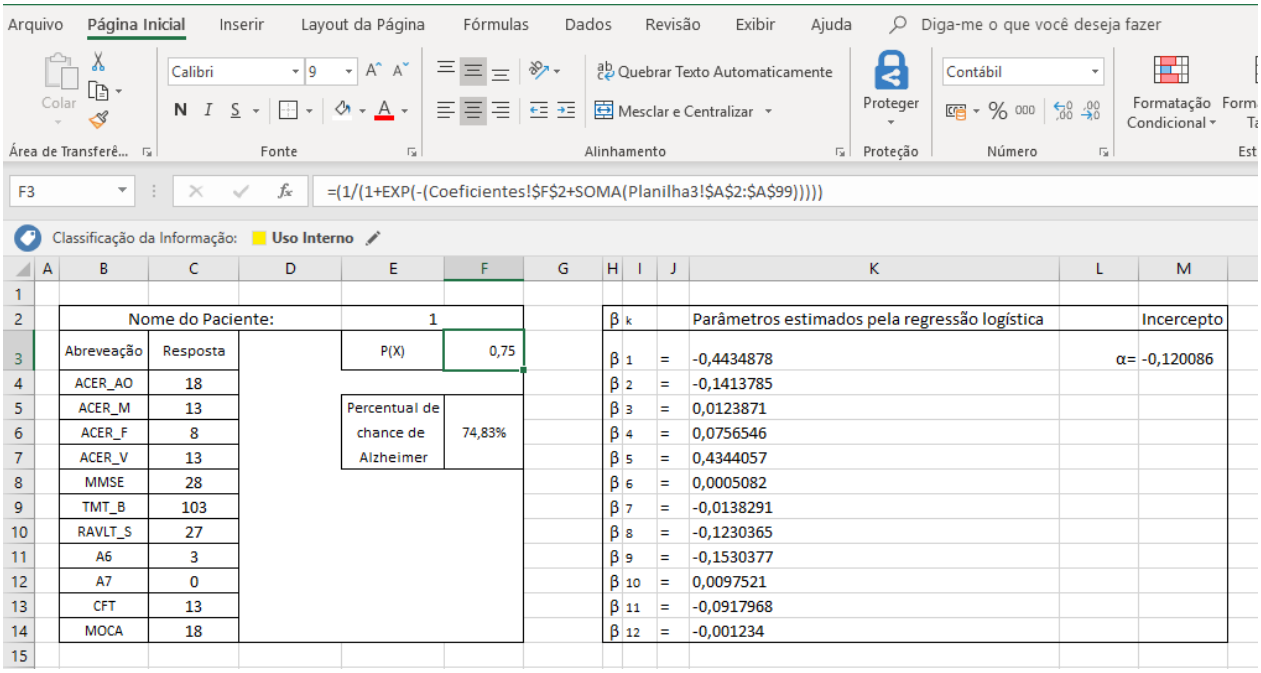

Figura 4

A calculadora serve para retornar P(X) de qualquer regressão logística, portanto foi desenvolvida de maneira com que a mudança dos coeficientes calculados pelo modelo de regressão logística possa ser mudada facilmente, assim como acrescentar novos coeficientes e novas variáveis.

4. PROPOSTA DE MELHORIA DO MODELO

Ao desenvolver o modelo, foi visto que o p-valor das variáveis independentes permaneceu muito alto, o que pode não ter interferido fortemente na performance do modelo, mas pode causar uma descaracterização do significado de algumas variáveis e significar a presença de multicolinearidade.

Utilizando a linguagem de programação R foram calculados os *p-valor* das variáveis na mesma regressão logística apresentada no capítulo 2, os quais estão no Quadro 5.

Parâmetro Estimado		p-valor
α	-0,1200864	0,986
β_1	-0,4434878	0,567
β_2	-0,1413785	0,498
β_3	0,0123871	0,977
β_4	0,0756546	0,870
β_5	0,4344057	0,580
β_6	0,0005082	0,963
β_7	-0,0138291	0,877
β_8	-0,1230365	0,749
β_9	-0,1530377	0,679
β_{10}	0,0097521	0,968
β_{11}	-0,0917968	0,686

Quadro 5

Para verificar se houve multicolinearidade, utilizando a linguagem de programação R, foi calculado o VIF (*variance inflation factor*) e apresentou o resultado presente no Quadro 6.

VIF					
X_1	X_2	X_3	X_4	X_5	
56.93	6.49	8.47	4.60	63.73	
X_6	X_7	X_8	X_9	X_{10}	X_{11}
3.72	5.89	9.02	10.61	7.45	5.28

Quadro 6

Como um VIF superior a 5 indica multicolinearidade entre as variáveis, utilizou-se a linguagem de programação R para fazer uma matriz de correlação entre as variáveis independentes como demonstrada na figura 5. Nota-se que quase todas as correlações são superiores a 0,7, o que indica novamente uma correlação forte entre as variáveis independentes e mostra que não são apenas pares correlacionados e sim que todas apresentam correlação entre si.

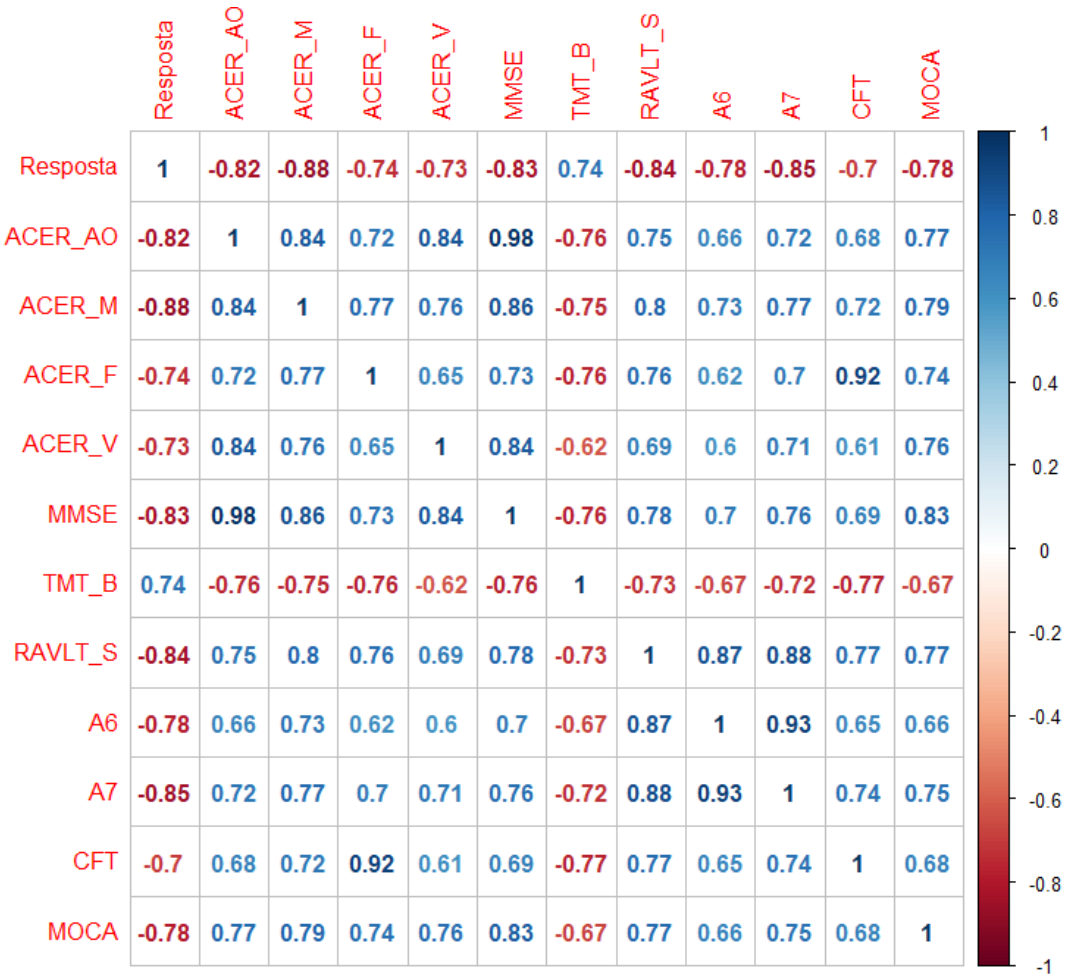

Figura 5

4.1 Análise de Componentes Principais

Para contornar a situação de multicolinearidade utilizou-se Análise de Componentes Principais, através da função *propcomp* presente na linguagem de programação R, e criou-se outro conjunto de onze variáveis originários das onze variáveis correlacionadas. O novo conjunto de variáveis foi criado ortogonalmente evitando correlações. A seguir percebe-se a proporção da variabilidade explicada em cada um dos componentes criados pela técnica utilizada.

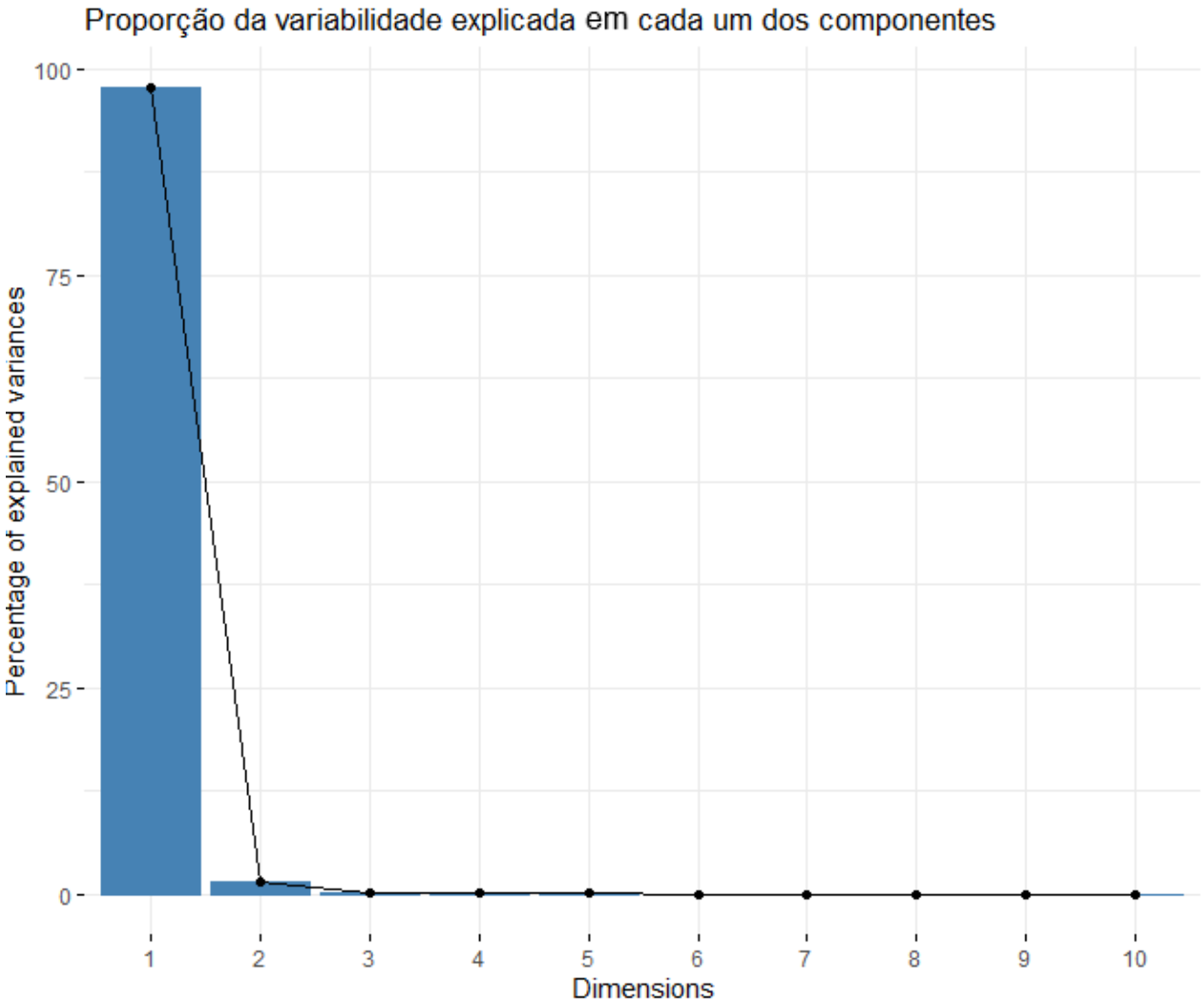

Figura 6

Ao analisar a Figura 6, nota-se que 95% das variáveis independentes iniciais são explicadas pelos primeiros dois componentes ortogonais criados. Os componentes PC_j

onde $j = 1\ e\ 2$ podem ser criados através das onze variáveis que os originaram utilizando a Eq(10) e os parâmetros apresentados no quadro7, onde ʎ é o peso da variável no componente e μ é o centro de ajuste, apresentados a seguir.

$$PC_j = \lambda_1(X_1 - \mu_1) + \lambda_2(X_2 - \mu_2) \dots \lambda_k(X_k - \mu_k) \qquad (10)$$

	PC1		PC2		Centro	
X1	ʎ1=	0,0290	ʎ1=	0,0989	μ1=	16,1707
X2	ʎ2=	0,0543	ʎ2=	0,2370	μ2=	18,1951
X3	ʎ3=	0,0276	ʎ3=	0,0937	μ3=	9,8536
X4	ʎ4=	0,0149	ʎ4=	0,7552	μ4=	14,6097
X5	ʎ5=	0,0402	ʎ5=	0,1560	μ5=	26,6585
X6	ʎ6=	-0,9860	ʎ6=	0,1618	μ6=	160,4878
X7	ʎ7=	0,1207	ʎ7=	0,8346	μ7=	39,6097
X8	ʎ8=	0,0342	ʎ8=	0,2197	μ8=	7,2195
X9	ʎ9=	0,0395	ʎ9=	0,2316	μ8=	7,3658
X10	ʎ10=	0,0505	ʎ10=	0,1643	μ10=	15,5365
X11	ʎ11=	0,0387	ʎ11=	0,2079	μ11=	23,3658

Tabela 7

Ao realizar novamente a regressão logística pelo R, utilizando somente os 2 componentes citados anteriormente, obteve-se os resultados que, como esperado, possuem baixo p-valor conforme demonstrado na quadro 8 e o modelo logístico que determinará a probabilidade de o indivíduo ter ou não Alzheimer na Eq(11).

Parâmetro estimado		P-valor
$\hat{\alpha}$	-3.01335	0.0150
$\hat{\beta}_1$	0.03766	0.0063
$\hat{\beta}_2$	-0.20734	0.0136

Quadro 8

$$f(Z) = \frac{1}{1 + e^{-(-3.01335 + 0.03766 PC_1 - 0.20734 PC_2)}} \qquad (11)$$

Seguindo a mesma lógica da validação feita na seção 3.3, aplicamos o novo modelo logístico nos dados coletados originalmente.

Verificou-se que o percentual de acerto acumulado foi de 100% tendo como ponto de corte p=0,5, valor padronizado para regressão logística. Ou seja, se considerar como certa a presença de Alzheimer para probabilidades superiores a 50% e como certa a ausência de Alzheimer para probabilidades inferiores a 50% o modelo acerta todos os quarenta e um casos analisados como demonstrados na Figura 7 e no quadro 9 a seguir.

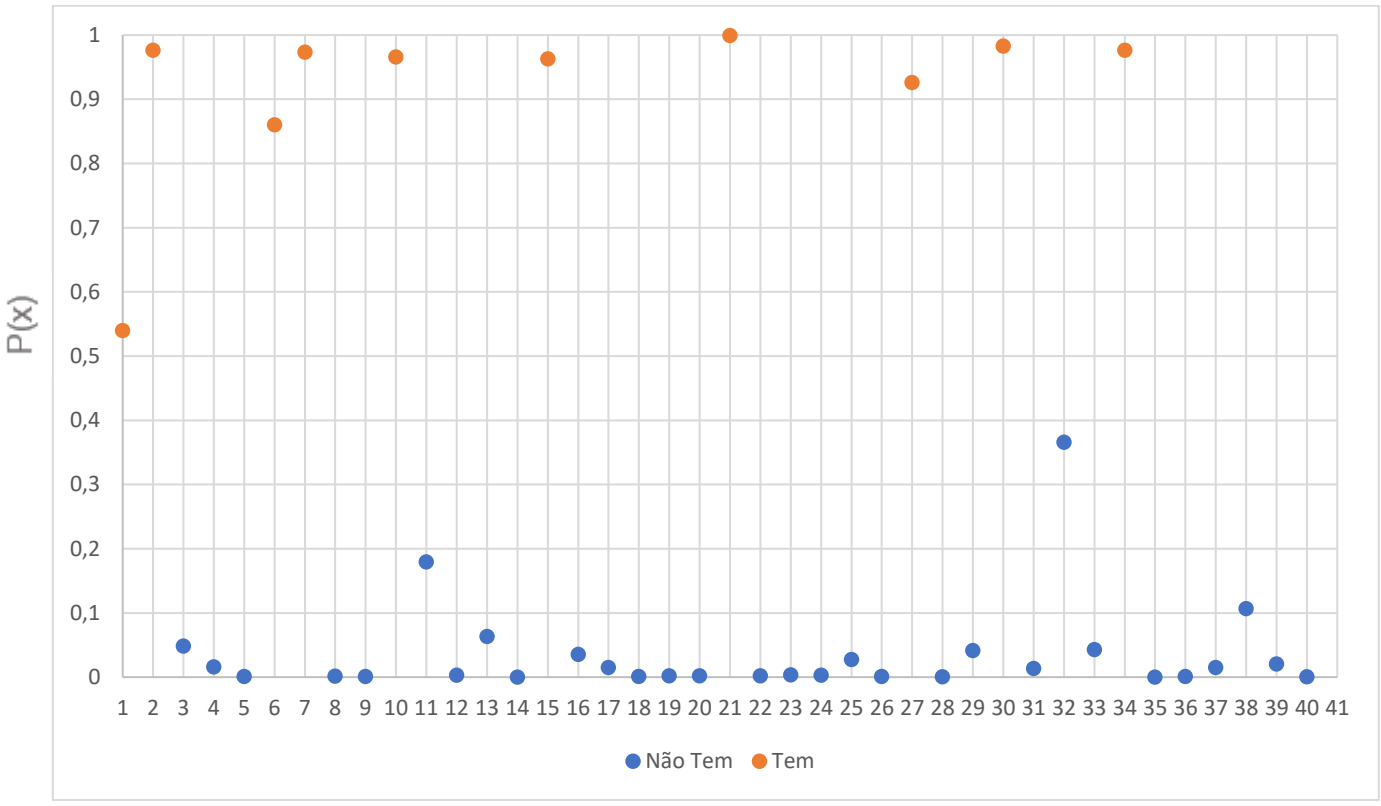

Figura 7

Paciente	PC_1	PC_2	Variável Resposta	Probabilidade Estimada P(X)
1	-54,14	-25,13	1	0,54
2	142,86	-6,57	1	0,98
3	13,10	2,21	0	0,05
4	20,27	9,01	0	0,02
5	-104,70	-1,68	0	0,00
6	141,11	2,33	1	0,86
7	142,72	-6,00	1	0,97
8	-117,23	-4,48	0	0,00
9	-61,30	7,24	0	0,00
10	142,39	-4,79	1	0,97
11	-16,18	-10,15	0	0,18
12	-97,24	-4,16	0	0,00
13	124,56	21,07	0	0,06
14	-63,34	12,68	0	0,00
15	141,83	-4,50	1	0,96
16	39,32	8,54	0	0,04
17	-42,74	-2,15	0	0,02
18	-100,13	0,79	0	0,00
19	-106,36	-3,92	0	0,00
20	0,31	15,21	0	0,00
21	145,45	-23,21	1	1,00
22	-39,48	8,58	0	0,00
23	13,32	14,88	0	0,00
24	-89,39	-2,52	0	0,00
25	-62,07	-8,65	0	0,03
26	-80,33	4,10	0	0,00
27	141,65	-1,01	1	0,93
28	-99,26	3,15	0	0,00
29	-66,47	-11,47	0	0,04
30	82,71	-19,11	1	0,98
31	-23,24	1,83	0	0,01
32	-77,51	-25,96	0	0,37
33	-19,11	-3,01	0	0,04
34	142,60	-6,60	1	0,98
35	-63,81	14,99	0	0,00
36	-73,09	4,41	0	0,00
37	-90,06	-10,68	0	0,01
38	90,87	12,22	0	0,11
39	43,21	11,91	0	0,02
40	-1,46	19,77	0	0,00
41	-19,65	10,82	0	0,00

Quadro 9

Após esta etapa, foi utilizada a mesma metodologia de validação citada na seção 3.3, separando a amostra aleatoriamente em base de desenvolvimento e validação, e repetindo o processo mil vezes e extraindo a média dos indicadores de precisão, especificidade e sensibilidade, conforme demonstrado na Figura 7:

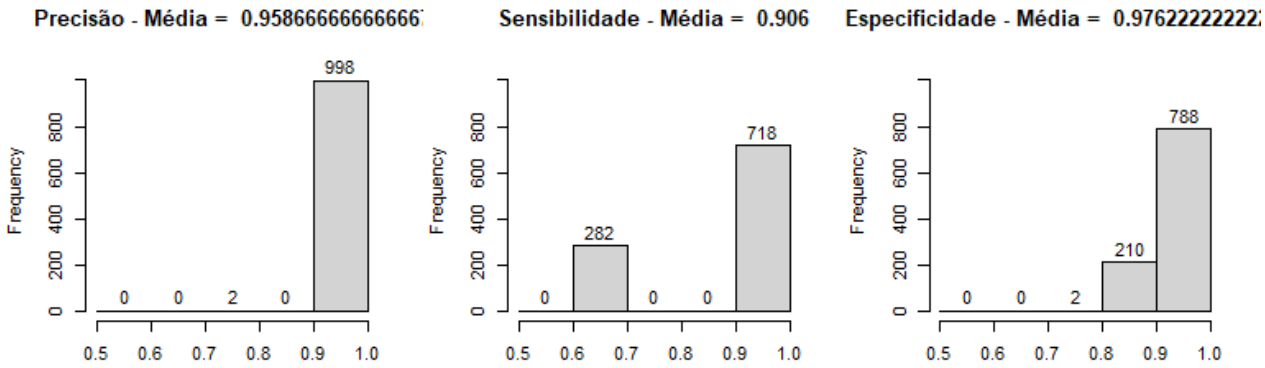

Figura 8

Como visto na Figura 8, a sensibilidade do modelo permanece menor que a especificidade, indicando que, acerta-se percentualmente de forma mais precisa quem não possui o Alzheimer e com menos precisão de quem possui.

Para melhorar o indicador de especificidade do modelo e aumentar a assertividade nos pacientes portadores da doença, alterou-se o ponto de corte para o cálculo dos três indicadores, precisão, sensibilidade e especificidade, aumentando de 5 em 5% para achar o ponto de encontro dos 3 indicadores.

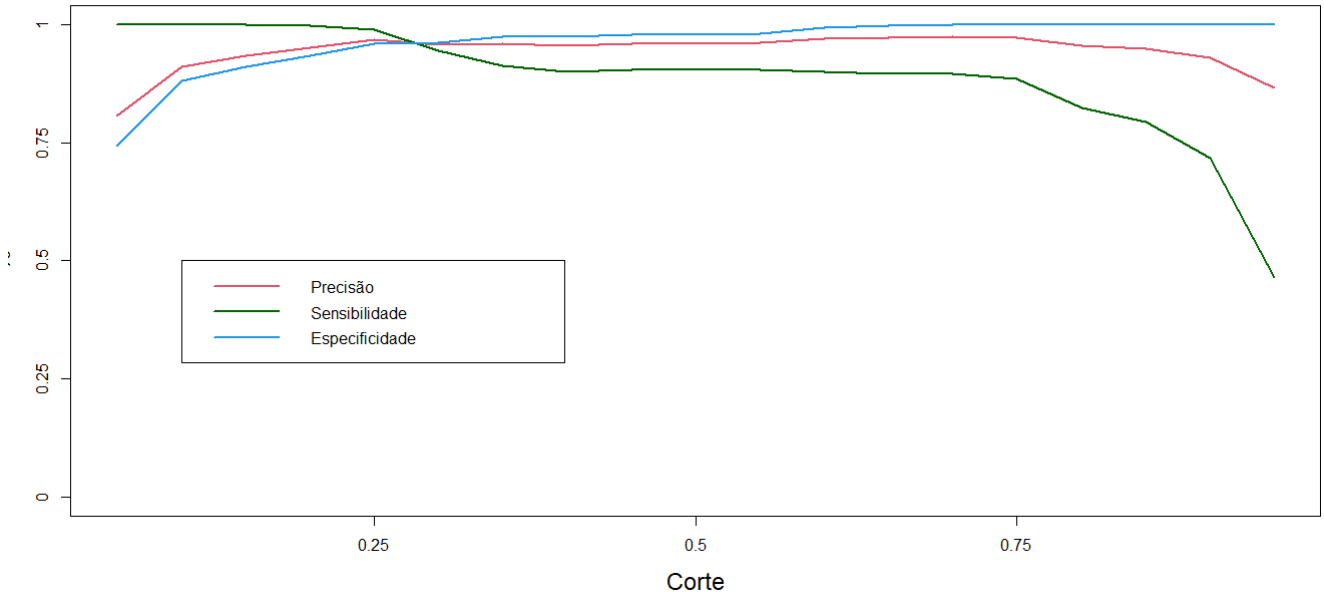

Figura 9

Analisando o gráfico da Figura 8, fica visível que o ponto de corte mais perto do equilíbrio entre os indicadores é em torno de 25%, portanto, assegura-se que, após substituídas as variáveis no modelo logístico obtido, gerado pela técnica descrita no trabalho, é coerente afirmar que o ponto de corte de 0,25 está bem próximo do melhor possível para a caracterização da doença de Alzheimer no modelo desenvolvido.

4.2 Tool Preditivo 2

Para o segundo modelo, também foi desenvolvido um tool preditivo apresentado na figura 10 para auxiliar na aplicação do modelo em dados de novos pacientes.

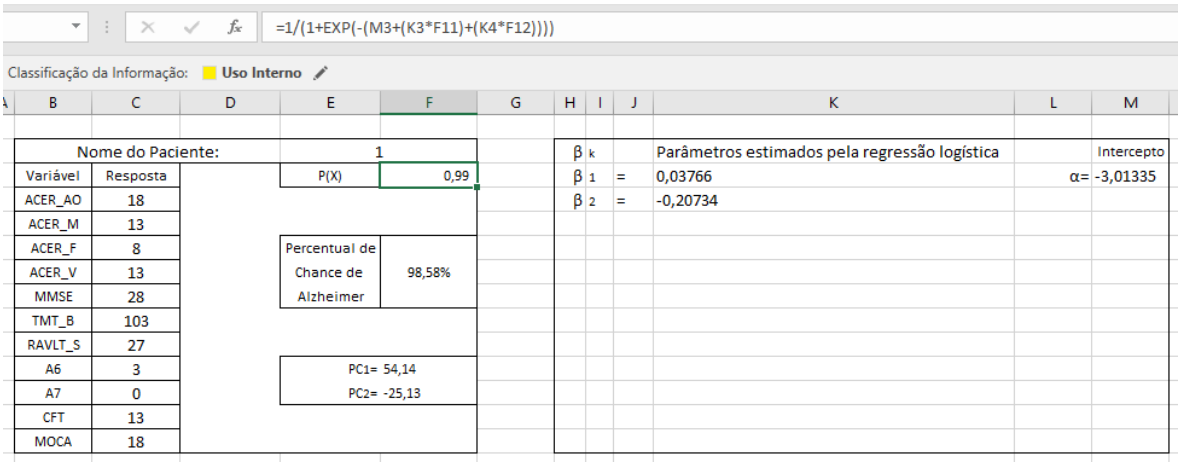

Figura 10

Diferente do primeiro tool, neste é necessário o cálculo dos componentes, o que torna um pouco mais difícil a mudança dos parâmetros por necessitar não só dos $\hat{\alpha}$ e $\hat{\beta}$ estimados, mas também dos pesos das variáveis nos componentes.

5. CONCLUSÃO

O principal objetivo desse trabalho de conclusão de curso foi atingido. Foi apresentada uma metodologia para construção de instrumento de detecção da doença de Alzheimer utilizando dados de testes cognitivos, selecionados dentre uma amostra de quarenta e duas variáveis por apresentarem uma forte correlação com a presença da doença. Após o cálculo dos parâmetros, o modelo logístico Eq(9) obtido mostrou-se adequado, pois reproduziu o diagnóstico conhecido *a priori* de 95% dos pacientes que participaram do estudo.

Utilizando análise de componentes principais foi desenvolvida uma proposta de melhoria onde foi calculado um segundo modelo logístico, Eq(10), o qual também se mostrou adequado, reproduzindo o diagnóstico conhecido em 100% dos casos.

Comparando os dois modelos gerados, a vantagem é que além do aumento da assertividade, com a análise de componentes principais ocorreu a eliminação da multicolinearidade entre as variáveis.

Devido ao tamanho da amostra, se fez necessário utilizar algumas técnicas de contorno para validação do modelo, o qual possivelmente poderia ser avaliado melhor com uma amostra mais significativa.

A partir dos modelos, Eq(9) e Eq(10), foram criados dois tools preditivos, com interface amigável, apresentados nas seções 3.4 e 4.2, respectivamente, e a ser disponibilizados para auxílio em diagnósticos nos estudos desenvolvidos nos projetos Alzheimer’s Teams e Superidosos, desenvolvidos no Instituto do Cérebro do Rio Grande do Sul

Como continuidade do presente estudo, pretende-se aprofundar o estudo da fundamentação teórica matemática, aplicar a metodologia apresentada em uma amostra maior e no desenvolvimento de tools preditivos em outras situações.

REFERÊNCIAS

A Guide to Appropriate use of Correlation Coefficient in Medical Research. Us National Library of Medicine National Institutes of Health, 2012. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/> Acesso em: 20 de Jun. de 2020.

Alzheimer's Team, INSCER Instituto do Cérebro, 2019. Disponível em: <http://www.pucrs.br/inscer/alzheimers-team/> Acesso em: 20 de Jun. de 2020.

BORELLI, WYLLIANS., **CORRELAÇÃO ENTRE NEUROIMAGEM MOLECULAR, ESTRUTURAL E FUNCIONAL EM SUPERIDOSOS**. Tese (Doutorado em medicina) Programa de Pós-Graduação de Medicina e Ciências da Saúde da Pontifícia Universidade Católica do Rio Grande do Sul. Porto alegre, 2019.

COLLUM, ROB.; **Visual Analytics with SAS Viya** special collection. Institute Inc. Cary, NC, USA, 2019.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada**: **para os cursos de administração, ciências contábeis e economia**. São Paulo: Atlas, 2009.

GREENHALGH T. **How to read a paper. Papers that report diagnostic or screening tests**.BMJ. Aug 30;315(7107):540-3. Review 1997.

GUPTA**. Prediction Tool for Favorable Neurologic Outcome, Soi Prediction Tool, 2019.** Disponível em:< https://soipredictiontool.shinyapps.io/GNOScore/> Acesso em: 20 de jun. de 2020.

HAIR Jr., J.F. ANDERSON, R.E.; TATHAM, R.L.; BLACK, W.C. **Análise multivariada de dados**. 5. ed. Porto Alegre: Bookman, 2005.

HAUSER, Eliete B.. **Notas de aula de Cálculo Numérico.** Porto Alegre 2020.

HAUSER, E.B, VENTURINI, G. T, GREGGIO S., BORELLI, W V, COSTA, J.C **Carotid arterial input function as an inverse problem in kinetic modeling of [18]2-fluoro-2 deoxy-D-glucose(FDG)**. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2019

HAUSER, E. B, BORELLI, W V, COSTA, J.C **Biomechanical Model Improving Alzheimer's Disease.** Intech Open .2020

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. Madison: Prentice Hall International, 1998.

LEIBING, **A. Doença de Alzheimer- uma história.** Informativo Psiquiátrico.,v. 17 supl. 1, p.S4-S9, p.1998.

LEMESHOW, S**.;** HOSMER, D. **Applied logistic regression**. New York: John Wiley & Sons, 1989.

LUZARDO, A. R.; GORINI, M. I. P. C.; SILVA, A.P.S.S. **Características de idosos com doença de alzheimer e seus cuidadores: Uma série de casos em um serviço de neurogeriatria,** 2006.

MAIA, ALEXANDRE. **Econometria: Conceitos e Aplicações.** São Paulo Saint Paul 2019.

MANN, PREM S**. Introdutory Statistics 7ª ed.** S.l.: John Wiley & Sons. 2010

MASSAD, Eduardo,.et al. **Métodos quantitativos em medicina**, Barueri, SP. Manole, 2004.

MEYER, Walter J. **Concepts of mathematical modeling**. New York: McGraw-Hill, 1984.

MEZZOMO, M. Estudo da mortalidade Infantil – um estudo de regressão logística múltipla. **Monografia de Especialização em Estatística e Modelagem Quantitativa**. Centro de Ciências Naturais e Exatas – Universidade Federal de Santa Maria. Santa Maria, RS, Brasil, 2009).

Modelagem Matemática e Farmacocinética, inscer Instituto do Cérebro, 2019. Disponível em: <http://www.pucrs.br/inscer/modelagem-matematica-em-farmacocinetica/> Acesso em: 20 de Jun. de 2020.

NITRINI, R.; CARAMELLI, P.; BOTTINO, C. M. C.; DAMASCENO, B. P.; BRUCKI, S. M. D.; ANGHINAH, R. **Diagnóstico de doença de Alzheimer no Brasil: avaliação cognitiva e funcional.** Arquivos de Neuropsiquiatria, 63 (3-A): 720-727, 2005.

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL. Biblioteca Central Ir. José Otão. **Modelo para apresentação de trabalhos acadêmicos, teses e dissertações elaborado pela Biblioteca Central Irmão José Otão**. 2011. Disponível em: <www.pucrs.br/biblioteca/trabalhosacademicos>. Acesso em: abril de 2020.

Superidosos, INSCER Institudo do Cérebro, 2019 Disponível em: <http://www.pucrs.br/inscer/superidosos/> Acesso em: 20 de Jun. de 2020

VRIES, ANDRIE, E.; MEYS JORIS. **R FOR DUMMIES** a wiley brand. 2. Ed. Copyright by John Wiley & Sons. Canada, 2015

WILKS, S. S. **The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses**. Ann. Math. Statist. 9.1938.

ANEXOS

ANEXO A - Arquivo R

```
setwd("C:\\Users\\Artur\\Desktop\\TCC_artur_lima")
dados = read.table("DadosSuperidosos.txt",head=T,sep=';')
names(dados)
names(dados)[1]='Resposta'
names(dados)

#install.packages('e1071')

library(brglm2)
library(corrplot)
library(factoextra)
library(caTools)
library(caret)
library(e1071)

#primeira Regressão
model<-brglm(formula = Resposta ~ . , family = "binomial", data = dados)
model
summary(model)

#Gráfico de correlações

windows()
corrplot(cor(dados), method = 'number', main = 'Correlações Entre as variáveis')

#Análise de Componentes Principais

acp = prcomp(dados[,-1])
componentes = acp$x

#Gráfico com proporção das variabilidades
windows()
fviz_eig(acp, main = 'Proporção da variabilidade explicada por cada um dos
componentes')

#Gráfico com as cargas dos dois primeiros fatores
windows()
fviz_pca_var(acp,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE,     # Avoid text overlapping
             title = 'Cargas dos Dois Primeiros Componentes'
```

```
)


dadospca =as.data.frame(cbind(dados[,1],componentes))
names(dadospca)[1] = 'Resposta'
names(dadospca)

modelo_pca = brglm(formula = Resposta ~ . , family = "binomial", data = dadospca)
modelo_pca
summary(modelo_pca)

#Segunda Regressão
modelo_final = brglm(formula = Resposta ~ PC1 + PC2 , family = "binomial", data =
train)
modelo_final
summary(modelo_final)

indices = sample.split(dadospca$Resposta, SplitRatio = 0.7)
train = dadospca[indices,]
test = dadospca[!(indices),]


modelo_train = brglm(formula = Resposta ~ PC1 + PC2 , family = "binomial", data =
train)
modelo_train
summary(modelo_train)

corte = .5
test_pred = predict(modelo_train, type = "response", newdata = test[,2:3])
test_pred_Attrition <- factor(ifelse(test_pred >= corte, "Yes", "No"))
test_actual_Attrition <- factor(ifelse(test[,1]==1,"Yes","No"))

conf <- confusionMatrix(test_pred_Attrition, test_actual_Attrition, positive = "Yes")

#Gráfico de precisão, sensibilidade e especificidade
acc <- conf$overall[1]
sens <- conf$byClass[1]
spec <- conf$byClass[2]

acc
sens
spec



#Separar base e regressões 10000 vezes
```

```
bootstrap = matrix(NA,nrow=10000,ncol = 3)
names(bootstrap) = c('Precisão','Sensibilidade','Especificidade')

for(i in 1:1000){
  indices = sample.split(dadospca$Resposta, SplitRatio = 0.7)
  train = dadospca[indices,]
  test = dadospca[!(indices),]
  modelo_train = brglm(formula = Resposta ~ PC1 + PC2 , family = "binomial", data =
train)
  corte = .5
  test_pred = predict(modelo_train, type = "response", newdata = test[,2:3])
  test_pred_Attrition <- factor(ifelse(test_pred >= corte, "Yes", "No"))
  test_actual_Attrition <- factor(ifelse(test[,1]==1,"Yes","No"))
  conf <- confusionMatrix(test_pred_Attrition, test_actual_Attrition, positive = "Yes")
  bootstrap[i,1] = conf$overall[1]
  bootstrap[i,2] = conf$byClass[1]
  bootstrap[i,3] = conf$byClass[2]
  if( i %% 100 == 0 ) {
    print(paste('Bootstrap na interação ',i))
    }
}
```

#Gráfico de precisão, sensibilidade e especificidade

```
bootstrap = as.data.frame(bootstrap)
names(bootstrap) = c('Precisão','Sensibilidade','Especificidade')
summary(bootstrap)

windows()
par(mfrow=c(1,3))
h = hist(bootstrap[,1],main = 'Precisão',xlim=c(0.5,1),ylim=c(0,1050),breaks =
c(.5,.6,.7,.8,.9,1))
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

h = hist(bootstrap[,2],main = 'Sensibilidade',xlim=c(0.5,1),ylim=c(0,1050),breaks =
c(.5,.6,.7,.8,.9,1))
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

h = hist(bootstrap[,3],main = 'Especificidade',xlim=c(0.5,1),ylim=c(0,1050),breaks =
c(.5,.6,.7,.8,.9,1))
text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))


cortes= seq(.05,.95, by = .05)
L = length(cortes)
```

```
bootstrap_cortes = array(NA,dim=c(1000,3,L))

#Cálculo da regressão 1000 vezes para diferentes pontos de corte


for(j in 1:L){
  corte = cortes[j]
    for(i in 1:1000){
    indices = sample.split(dadospca$Resposta, SplitRatio = 0.7)
    train = dadospca[indices,]
    test = dadospca[!(indices),]
    modelo_train = brglm(formula = Resposta ~ PC1 + PC2 , family = "binomial", data =
train)
    test_pred = predict(modelo_train, type = "response", newdata = test[,2:3])
    test_pred_Attrition <- factor(ifelse(test_pred >= corte, "Yes", "No"))
    test_actual_Attrition <- factor(ifelse(test[,1]==1,"Yes","No"))
    conf <- confusionMatrix(test_pred_Attrition, test_actual_Attrition, positive = "Yes")
    bootstrap_cortes[i,1,j] = conf$overall[1]
    bootstrap_cortes[i,2,j] = conf$byClass[1]
    bootstrap_cortes[i,3,j] = conf$byClass[2]
    if( i %% 100 == 0 ) {
      print(paste('Bootstrap na interação ',i,'corte ',j))
    }
  }
}

medias_cortes = matrix(NA,L,3)
colnames(medias_cortes) = c('Precisão','Sensibilidade','Especificidade')
rownames(medias_cortes) = cortes

for(j in 1:L){
  medias_cortes[j,1] = mean(bootstrap_cortes[,1,j])
  medias_cortes[j,2] = mean(bootstrap_cortes[,2,j])
  medias_cortes[j,3] = mean(bootstrap_cortes[,3,j])

}

medias_cortes

#Graphics.off()

for(j in 1:L){
  windows()
  par(mfrow=c(1,3))
  h = hist(bootstrap_cortes[,1,j],main = paste('Precisão - corte = ',cortes[j],'\n Média =
',medias_cortes[j,1]),xlim=c(0,1),ylim=c(0,1050),breaks = c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
  text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))
```

```
  h = hist(bootstrap_cortes[,2,j],main = paste('Sensibilidade - corte = ',cortes[j],'\n Média
= ',medias_cortes[j,2]),xlim=c(0,1),ylim=c(0,1050),breaks = c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
  text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

  h = hist(bootstrap_cortes[,3,j],main = paste('Especificidade - corte = ',cortes[j],'\n Média
= ',medias_cortes[j,3]),xlim=c(0,1),ylim=c(0,1050),breaks = c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
  text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))
}


#graphics.off()
windows()
plot(cortes,
medias_cortes[,1],xlab="Corte",ylab="%",cex.lab=1.5,cex.axis=1.5,ylim=c(0,1),type="l",l
wd=2,axes=FALSE,col=2)
axis(1,seq(0,1,length=5),seq(0,1,length=5),cex.lab=1.5)
axis(2,seq(0,1,length=5),seq(0,1,length=5),cex.lab=1.5)
lines(cortes,medias_cortes[,2],col="darkgreen",lwd=2)
lines(cortes,medias_cortes[,3],col=4,lwd=2)
box()
legend(.1,.5,col=c(2,"darkgreen",4,"darkred"),lwd=c(2,2,2,2),c("Precisão","Sensibilidade"
,"Especificidade"))
```