

# Relatório Modelagem Estatística

Kayo Yokoyama

June 2023

## 1 Introdução

Este relatório apresenta uma análise de modelagem estatística utilizando dados de ressonância magnética (MRI) longitudinal de 150 indivíduos, com o objetivo de classificar entre pessoas com e sem Transtorno Neurocognitivo/Cognitivo Maior (TCM/TNM), ou mais conhecido como Demência. Estas condições neurológicas resultam em declínio progressivo das funções cognitivas, como memória, pensamento, linguagem, percepção e habilidades motoras, ela é dividida em alguns subtipos como a doença de Alzheimer, a Demência Vascular, a Demência por Corpos de Lewy, a Demência Associada à Doença de Parkinson, etc. Para aquilo, aplicaremos técnicas de modelagem estatística e machine learning, explorando variáveis como idade, gênero, educação, status socioeconômico e medidas específicas obtidas a partir das imagens de MRI, como volume intracraniano estimado (eTIV) e volume cerebral total normalizado (nWBV).

Os dados utilizados neste estudo são provenientes do projeto Open Access Series of Imaging Studies (OASIS), que disponibiliza conjuntos de dados de MRI do cérebro de forma gratuita para a comunidade científica. O principal propósito do projeto OASIS é facilitar descobertas futuras nas áreas de neurociência básica e clínica, ao compilar e distribuir esses conjuntos de dados de MRI.

### 1.1 Motivação

Aos longos dos anos tive contato pessoal com pessoas que sofrem de algum subtipo de TCM/TNM, e sempre me perguntei como seria possível detectar esse tipo de doença em pessoas de várias idades, para que medidas pudessem ser tomadas o mais rápido possível, assim amenizando os danos causados na pessoa e diminuindo os gastos com o sistema de saúde, com cuidados e tratamentos. Com isso, mais inicialmente em minha graduação me deparei com este dataset, mas nunca lembrei nem tive a oportunidade de analisá-lo mais a fundo. Agora, espero conseguir sanar minimamente a minha curiosidade e aprender mais sobre o assunto.

### 1.2 Dados

Este conjunto de dados consiste em uma coleção longitudinal de 150 indivíduos com idades entre 60 e 96 anos. Cada indivíduo foi submetido a dois ou mais exames, separados por pelo menos um ano, totalizando 373 sessões de imagem. Para cada indivíduo, estão incluídas 3 ou 4 varreduras individuais de ressonância magnética ponderada em T1 obtidas em sessões de exame únicas. Os indivíduos são todos destros e incluem homens e mulheres. Dos indivíduos, 72 foram classificados como não dementes ao longo do estudo. Dos indivíduos incluídos, 64 foram classificados como dementes no momento de suas visitas iniciais e permaneceram assim nas varreduras subsequentes, incluindo 51 indivíduos com doença de Alzheimer leve a moderada. Outros 14 indivíduos foram classificados como não dementes no momento de sua visita inicial e posteriormente foram classificados como dementes em uma visita posterior. [1] Além disso, os casos de demência analisados são relacionados mais especificamente ao subtipo da doença de Alzheimer.

- Subject ID: ID para cada paciente
- MRI ID: O ID para cada ressonância de um determinado paciente (pode haver mais de um ID por ter feitos múltiplas visitas)
- Group: Demented, Nondemented ou Converted, classifica se a pessoa possui, não possui, ou se ela desenvolveu demência ao longo das sessões, seja ele o grau que for.

- Visit: Número da visita feita pelo paciente
- MR Delay: Atraso do tempo de repetição (TR, do inglês Time of Repetition) durante a sequência de aquisição de imagens. Esse atraso pode ser usado para obter informações adicionais sobre o tecido ou a região sendo examinada, como para suprimir o sinal de determinados tecidos ou para melhorar a visualização de estruturas específicas.
- M/F: Gênero (M/Masculino or F/Feminino).
- EDUC (Education): Anos de educação.
- SES (Socioeconomic Status): Status socioeconômico avaliado pelo Hollingshead Index of Social Position. Classificado em categorias de 1 (Maior status) até 5 (Menor status) (Hollingshead, 1957)
- MMSE (Mini-Mental State Exam): Pontuação do Exame de Estado Mini-Mental. Varia entre 0 (pior) e 30 (melhor) (Folstein, Folstein, & McHugh, 1975)
- CDR (Clinical Dementia Rating): Avaliação clínica do grau de Demência (0 = não possui, 0.5 = Alzheimer bem leve, 1 = Alzheimer leve, 2 = Alzheimer Moderado) (Morris, 1993) — Há o grau 3 = severo, porém não foi documentado em nenhum paciente.
- eTIV (Estimated Total Intracranial Volume): Estimativa do volume intracraniano total (cm<sup>3</sup>) (Buckner et al., 2004)
- nWBV (Normalized Whole Brain Volume): Expressa a porcentagem de todos os voxels na imagem da forma atlas, que estão marcados como matéria branca ou cinza pelo processo automatizado de segmentação do tecido, isto com base no eTIV (Fotenos et al., 2005)
- ASF (Atlas Scaling Factor): Fator Atlas de escala (sem unidade). Fator de escala computado que transforma o espaço nativo do cérebro e do crânio em um objeto atlas (i.e. o determinante da matrix de transformação) (Buckner et al., 2004)

## 2 Abordagem e desenvolvimento geral

### 2.1 Implementação

Todos os modelos e ferramentas foram feitos usando Python [2], juntamente com Pandas [3], Numpy [4], Scikitlearn [5] e Statsmodels [6].

### 2.2 Tratamento dos dados

Inicialmente tinha-se a disposição 373 instâncias para análise e dessas 37 tinham a tag "Converted" na variável "Group", assim, para que ficasse melhor o tratamento e desenvolvimento do projeto, caso uma pessoa fosse "Group = Converted" e "CDR = 0" ela será mudada para "Group = Nondemented", caso contrário se ela for do mesmo grupo e tiver "CDR > 0" ela será agora "Group = Demented". Além disso, cerca de 19 delas haviam dados faltantes com valores *NaN*, mais especificamente nas colunas "SES" e "MMSE". Então, a princípio para que não se perdesse nenhuma instância para futuras análises tentaram-se algumas formas para preencher os dados faltantes. Com isso, foi feita uma análise de como os dados se correlacionam e como alguns deles estão distribuídos para melhor embasamento. Exemplo de dados faltantes na tabela 1 e gráficos na figure

1

Index	Subject ID	MRI ID	Group	Visit	MR Delay	M/F	Age
2	OAS2.0002	OAS2.0002_MR1	1	1	0	0	75
3	OAS2.0002	OAS2.0002_MR2	1	2	560	0	76
4	OAS2.0002	OAS2.0002_MR3	1	3	1895	0	80

EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
12	NaN	23.0	0.5	1678	0.736	1.046
12	NaN	28.0	0.5	1738	0.713	1.010
12	NaN	22.0	0.5	1698	0.701	1.034

Table 1: Exemplo de linhas com dados faltantes

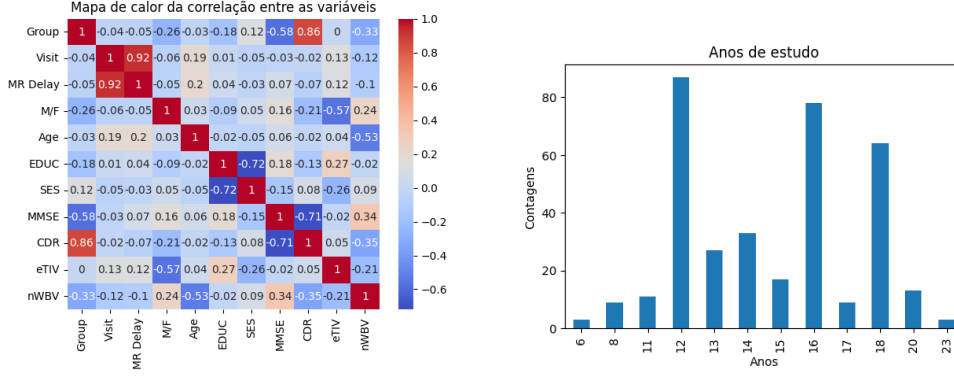


Figure 1: Mapa de Correlações e contagem de idades

Dado que a correlação entre SES e EDUC ser inversamente proporcional e relativamente alta, números que correspondem aos dados foram alocados para os dados faltantes, como  $SES = 3$  ou  $SES = 2$ , após isso também foi feita uma pequena regressão linear para tentar outra forma de preenchimento. Mas, no final os modelos se saíram piores do que caso esses dados fossem removidos completamente do dataset, e é o que foi usado pelo resto do relatório.

## 2.3 Métodos

Primeiramente, foram usadas múltiplas abordagens de modelo para análises e testes rápidos, randomizando a disposição do dataset a cada teste, além disso a abordagem de 5-fold cross-validation foi usada para escolher os melhores parâmetros. Mas, por fim optei por realizar os testes com apenas 4 algoritmos distintos, sendo eles a regressão logística, a árvore de decisão, floresta aleatória e o naive bayes. Dito isso, tendo a acurácia média dos modelos após 100 sessões de treino, tivemos o seguinte rendimento para acurácia [7], cross-entropy loss [8] e Brier score [9] médio para cada um, dado a Table 2:

Table 2: Resultados de rendimento pelo Scikitlearn

Modelo	Acurácia Média	Cross-entropy Loss	Brier Score
Regressão Logística	0.82356	15.17139	0.15594
Árvore de Decisão	0.77103	15.17139	0.21443
Floresta Aleatória	0.81323	18.07656	0.1547
Naive Bayes	0.80028	15.49418	0.17636

Veja que nenhum modelo se saiu extremamente mal, mas, um modelo que parece ser justo para uso é o de regressão logística. Onde, além de ser relativamente simples de interpretar para uso ele é flexível pois pode acomodar várias variáveis dependentes ou independentes e explorar relações entre as mesmas através do uso de termos de interação, fora que permite realizar testes de hipóteses e intervalos de confiança para os coeficientes, auxiliando na avaliação da significância estatística das relações entre as variáveis.

Uma boa forma de verificar suas predições e pela análise da curva de predições (irônico):

Observe que em alguns momentos o modelo está fazendo uma superestimação, curva acima da diagonal, e em outros uma subestimação das predições, curva abaixo da diagonal, no geral ela tende a se manter perto do que seria a probabilidade adequada para a previsão. Complementando, alguns dos motivos para que essas

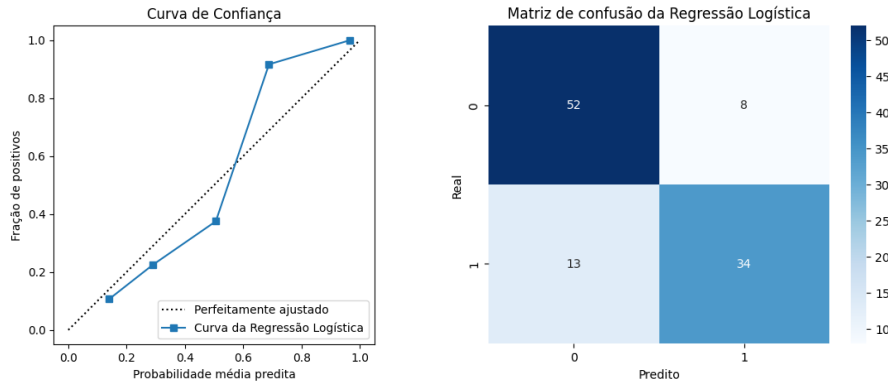


Figure 2: Curva de predição e matriz de confusão da regressão logística

questões podem envolver várias razões como o fato da regressão logística assumir certas condições subjacentes, como a linearidade das relações entre as variáveis independentes e dependentes, ou então tamanho insuficiente da amostra, ou simplesmente uma escolha errada dos melhores parâmetros. Dessa forma, mais a frente discutiremos motivos e maneiras de futuras melhoras.

Além disso, olhando para a matrix de confusão não há uma discrepância muito grande com relação as classificações do modelo e as reais, ele consegue dizer muito bem quando alguém tende ter ou não demência. Assim, calculando as taxas atreladas temos um outro indicativo de um bom modelo:

- Taxa de Verdadeiros Positivos (TVP): 0.8
- Taxa de Falsos Positivos (TFP): 0.19
- Taxa de Falsos Negativos (TFN): 0.2
- Taxa de Verdadeiros Negativos (TVN): 0.81

Em vista disso, vamos tomar outra perspectiva de análise para complementar o que já foi dito. Então, usando outra biblioteca a regressão logística não será otimizada com "Limited-memory BFGS" (padrão do scikit-learn) mas sim com "Newton-Raphson" (padrão do statsmodel.api).

Table 3: Sumário da Regressão Logística pelo Statsmodel

Variável	Coefficiente	Erro Padrão	Estatística t	Valor P	Intervalo de Confiança
Constante	49.9418	24.397	2.047	0.041	(2.125, 97.758)
MR Delay	0.0004	0.000	1.391	0.164	(-0.000, 0.001)
M/F	-1.0490	0.532	-1.971	0.049	(-2.092, -0.006)
Age	-0.1358	0.037	-3.650	0.000	(-0.209, -0.063)
EDUC	-0.2323	0.117	-1.979	0.048	(-0.462, -0.002)
SES	-0.5552	0.277	-2.004	0.045	(-1.098, -0.012)
MMSE	-1.0671	0.167	-6.392	0.000	(-1.394, -0.740)
eTIV	0.0038	0.008	0.500	0.617	(-0.011, 0.019)
nWBV	-26.3682	7.317	-3.604	0.000	(-40.709, -12.027)
ASF	7.2786	10.096	0.721	0.471	(-12.508, 27.066)

A partir da tabela podemos ver algumas estatísticas adicionais com relação a testes de hipótese e as variáveis. Assim, percebe-se que no geral as variáveis têm um "Valor P" ou " $P > |t|$ " abaixo de 0.05, o que aponta ser significativo o uso da variável para designação de demência, e erros padrões aceitáveis, tendo também em mente os intervalos de confiança. Além disso, há um notável destaque em uma das variáveis para a classificação do Alzheimer's, ela sendo o "nWBV", indicando que uma diminuição no volume relativo de tecido cerebral está associada a um aumento na probabilidade de uma pessoa ser classificada como tendo demência. Apesar disso tudo, há variáveis que parecem não ser tão relevantes para o modelo, como por exemplo MR Delay, eTIV e

ASF, nenhuma delas apresenta um valor P muito satisfatório para o modelo, apontando assim mais espaço para futuras mudanças que discutiremos depois.

### 3 Análise exploratória dos dados e visualização

Seguindo com o projeto, partiremos para algumas análises adicionais com relação aos dados. Primeiro vamos dar uma olhada em um pairplot/splom/gráfico de relações, Figure 3 no final do arquivo.

Nele, juntamente com o heatmap da Figure 2, é possível observar algumas tendências interessantes com relação a algumas das variáveis, como nWBV x Group, nWBV x Age, MMSE x CDR e o fato de mulheres apresentarem ter menos ocorrência da doença (apesar de ter  $H = 150$  ou 42.37% e  $F = 204$  ou 57.63%).

Mais à frente, neste mesmo relatório, também haverá outra tentativa de ajustar o modelo com base nas informações coletadas ao longo do estudo.

Continuando, vamos primeiro dar atenção para como nWBV x Group se relacionam.

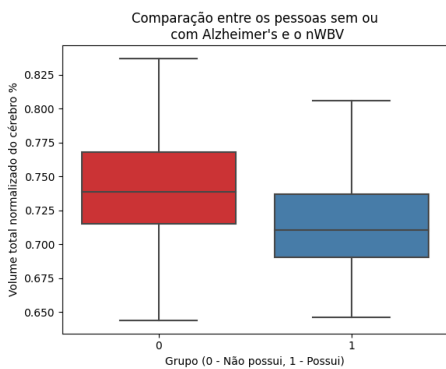


Figure 3: Boxplot comparativo

Nessa comparação, olhe como pessoas que possuem Alzheimer's tendem a ter uma média mais baixa do volume total normalizado do cérebro. Isso é decorrente de que na doença de Alzheimer, à medida que os neurônios são lesionados e morrem em todo o cérebro, as conexões entre redes de neurônios podem se desfazer e muitas regiões cerebrais começam a encolher. Como consequência, nas fases finais do Alzheimer, esse processo, chamado de atrofia cerebral, é generalizado, causando uma perda significativa do volume cerebral. [10]

Outra análise interessante é a de nWBV x Age:

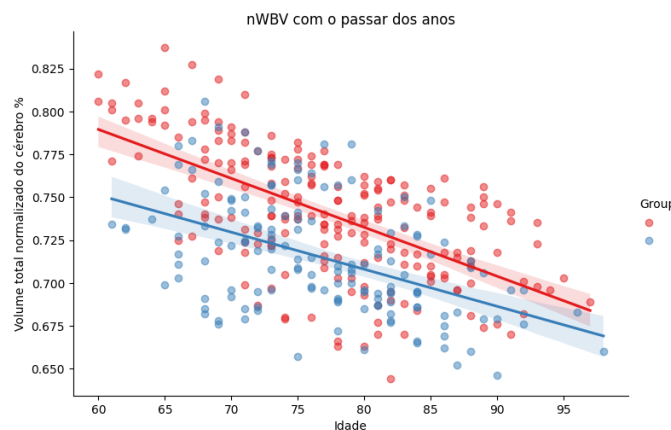


Figure 4: Volume total ao longo dos anos

Mais claramente as duas linhas tem os seguintes formas:

- nWBV for nondemented =  $-0.0029 * Age + 0.9606$
- nWBV for demented =  $-0.0022 * Age + 0.8808$

O decaimento no geral não difere tanto ao longo dos anos para cada grupo. Mas, em algum momento parecer haver um declínio rápido em pessoas de desenvolvem Alzheimer's, isso dado o fato de que o a linha azul de inicia

bem abaixo da vermelha. Além disso, essa diferença entre os grupos não parece ser causada pela quantidade da amostra, dado que temos 146 instâncias do grupo 1 no dataset (41.24%).

Citando outro estudo feito em São Paulo, contendo dados semelhantes, mas com mesma finalidade desta análise em específico, têm-se que "A taxa mediana de redução do volume cerebral foi de 2,4% por década após os 60 anos de idade. Entre os intervalos de 5 anos (de 60 a 105 anos), a taxa variou de 0,2% a 4,6%, embora as diferenças não tenham sido estatisticamente significativas. Também observamos que não houve diferença significativa entre a matéria cinzenta e branca em termos da taxa de redução volumétrica." [11]

Seguindo com os estudos, temos agora a análise de MMSE x CDR:

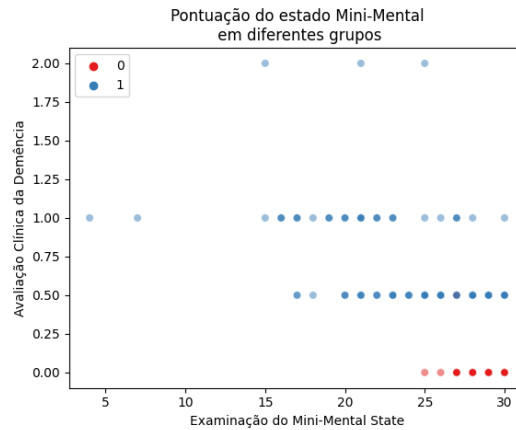


Figure 5: Pontuação do MMSE em diferentes grupos

Hmmm, os pontos vermelhos aqui chamam bastante atenção dado o fato de que todos eles estão acima da pontuação 25. Assim, como aponta um estudo que trata especificamente esta análise com dados similares "O MMSE discriminou bem entre os estágios CDR 0.5, 1, 2 e 3, mas teve um desempenho ruim na separação entre os estágios zero e 0.5 do CDR. As faixas do MMSE foram 30 para ausência de demência, 26-29 para dúvida/questionável, 21-25 para leve, 11-20 para moderada e 0-10 para demência grave." [12]. Assim sendo, isso pode representar um bom indicativo, não o melhor de todos, para a presença de Alzheimer's ou não, dado também que no meu dataset 302 pessoas tiveram pontuação acima ou igual à 25, onde 205 delas tem CDR 0 (67,8%).

Por fim, antes de partir para um possível novo modelo vamos analisar como se dá a distribuição entre gêneros para cada grupo.

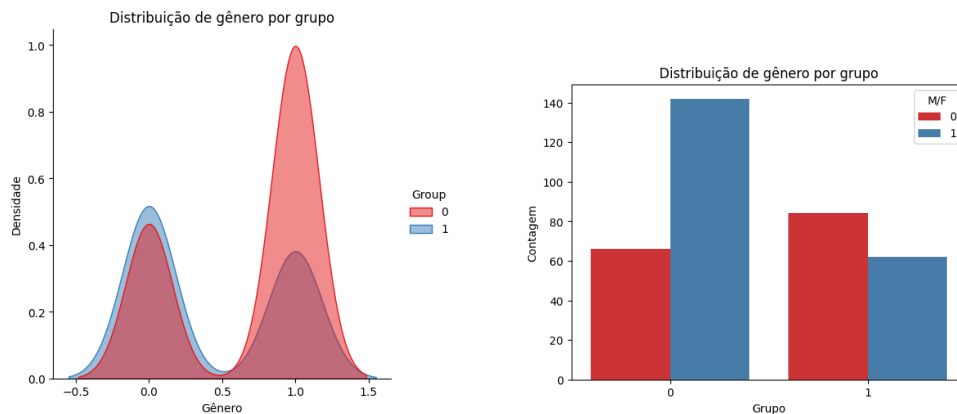


Figure 6: Distribuições de gênero por grupo

Repare que é evidente uma diferença maior na porcentagem de mulheres que tem Alzheimer's com relação a todas no dataset do que com relação a homens. Têm-se 204 instâncias para F, sendo 142 no grupamento

0 e 62 no 1, e 152 instâncias para M, sendo 66 no grupamento 0 e 84 em 1. Assim, cerca de 30% das mulheres apresentaram algum quadro ao longo do estudo contra 55% dos homens apresentaram, uma diferença relativamente significativa para uma desigualdade não muito grande entre homens e mulheres. Porém, "There ain't no such thing as a free lunch" pois apesar de que no meu dataset aparentar bem na cara que mulheres tendem a ter desenvolver menos Alzheimer's a história é outra. Um estudo feito na Suécia com 16,926 pessoas reportou o seguinte "...as taxas de demência foram relativamente semelhantes até depois dos 80 anos, sendo que as mulheres apresentaram taxas significativamente mais altas do que os homens nos últimos anos da década de 80. Para a doença de Alzheimer (DA) isoladamente, as mulheres apresentaram taxas significativamente mais altas de DA a partir do início dos anos 80." [13] e "A sobrevivência das mulheres até idades mais avançadas pode ser uma explicação para a maior incidência de demência e DA entre as mulheres, possivelmente." [13]

Agora, vamos tentar remodelar nossa regressão logística com todos os fatos coletados até agora. Para isso, vamos adicionar algumas análises e fazer outras mudanças na modelagem.

Primeiro, vamos retirar algumas variáveis da análise, estas julgadas anteriormente como supérfluas, estas serão MR Delay, CDR, eTIV e ASF, IDs e MRI IDs, além disso um 10-fold cv foi feito. Fora isso, outras métricas foram adotadas para maior acurácia, agora teremos acurácia, log loss, F1 score, brier score, roc auc score, precision e recall.

Como visto anteriormente o modelo se sai bem usando um otimizador como Newton-raphson, mas para fins de teste usaremos uma outra alternativa disponível no scikit-learn o "Newton-Cholesky", a grande diferença é que este utiliza a decomposição Cholesky para resolver o sistema linear que surge no passo de atualização .

Table 4: Sumário da Regressão Logística feita a mão

Variável	Coefficiente	Erro Padrão	Estatística t	Valor P	[0.025	0.975]
const	48.915079	2.122763	23.04311	0.00000	44.733	53.097
M/F	-0.765199	0.137518	-5.56437	0.00000	-1.036	-0.494
Age	-0.090947	0.010450	-8.70303	0.00000	-0.112	-0.070
EDUC	-0.141632	0.032643	-4.33887	0.00002	-0.206	-0.077
SES	-0.096954	0.082487	-1.17539	0.24101	-0.259	0.066
MMSE	-0.831804	0.018620	-44.67266	0.00000	-0.868	-0.795
nWBV	-22.007691	2.296492	-9.58318	0.00000	-26.532	-17.484

Veja que com as mudanças feitas tivemos bons avanços. Por exemplo, os intervalos de confiança e os erros padrões estão menores, porém ainda há uma variável que parece não ser tão benéfica para o nosso estudo em comparação com as restantes, a EDUC. Com isso, faremos o último ajuste de uso das variáveis e veremos os resultados.

Table 5: Sumário da Regressão Logística feita a mão

Variável	Coefficiente	Erro Padrão	Estatística t	Valor P	[0.025	0.975]
const	48.207998	2.092604	23.03732	0.00000	44.086	52.330
M/F	-0.748816	0.134450	-5.56948	0.00000	-1.014	-0.484
Age	-0.099205	0.010406	-9.53352	0.00000	-0.120	-0.079
EDUC	-0.075854	0.022932	-3.30777	0.00108	-0.121	-0.031
MMSE	-0.809437	0.019703	-41.08163	0.00000	-0.848	-0.771
nWBV	-22.673412	2.300431	-9.85616	0.00000	-27.205	-18.142

Juntamente da Table 5 têm-se:

Table 6: Medidas adicionais

Modelo	Acurácia	Log Loss	F1 Score	Brier Score	Roc Auc Score	Precision	Recall
Média geral	0.8186	0.3799	0.7631	0.1814	0.8041	0.819	0.7196
Melhor modelo	0.9159	0.2604	0.8861	0.08411	0.9041	0.92105	0.8537

## 4 Conclusão

Finalmente, pelo que é possível notar pelas Table 5 e Table 6, resultados bons mas não tão satisfatórios podem ser alcançados usando a regressão logística. Assim, acredito que ela seja uma possível boa candidata para o problema proposto, porém com algumas ressalvas, melhoradas e ajustes a serem feitos. Além de que a regressão logística possui limitações intrínsecas a ela, como ser inadequada para lidar com variáveis dependentes contínuas, como no caso de nWBV, ou a presunção de que sempre há uma relação linear entre as variáveis independentes e a variável dependente. Ademais, outro modelo que parecia ser útil era o de Random Forest, talvez este pudesse ser um próximo candidato mais adequado para este projeto, dado que os resultados finais foram bem medianos no geral.

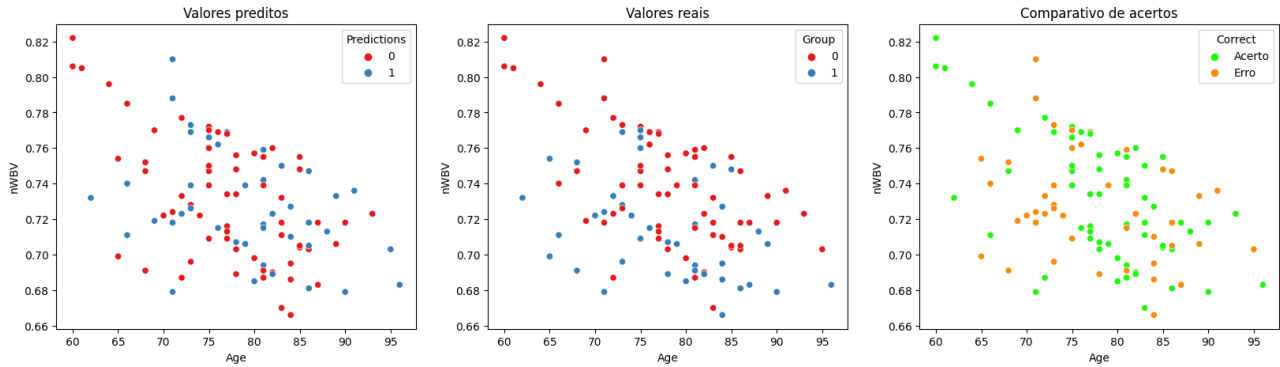


Figure 7: Comparativo final do modelo

Além disso, fiquei limitado ao meu dataset enquanto no site haviam outros dados que eu poderia ser explorado, apesar de que alguns deles tinham muitos dados faltantes, mas pode ser que com a minha eliminação de variáveis supérfluas elas se tornassem viáveis a final de contas, fazendo assim com que eu tivesse um modelo talvez mais robusto. Não só, mas também fiquei limitado com relações as ferramentas de uso pois o scikitlearn não tinha todas as estatísticas necessárias para análise e o statsmodel não tinha todo o poder ou variedade que o outro tinha em questões de modelagem e alternativas para uso, fora a documentação nada intuitiva e um tanto precária do statsmodel.

Outro ponto é que na melhor comparação deve ser feita para propor de fato um modelo muito bom para este problema, com maior revisão na literatura focado em informações médicas e estatísticas, porém, tentei focar em algum modelo que tivesse um background estatístico já intrínseco em si, embora outros modelos possam ser calibrados e ajustados para o mesmo efeito. Acredito que com mais tempo, experiência e poder computacional melhores resultados possam ser adquiridos.

Visando também uma abrangência de análise outros fatos interessantes surgiram ao longo do estudo que podem ser examinados com mais afinco e podem, ou não, estar relacionados de forma direta ou indireta com a causa geral de Alzheimer's. Por exemplo, o fato de que pessoas com maior status socioeconômico tendem a ter mais anos de educação, ou vice-versa, o que pode dar espaço de pensamento para inferir que pessoas assim tem mais oportunidades e maior acesso a uma medicina de qualidade. Dito isso, corroborando para minha análise, em um estudo há a seguinte observação "Uma coorte de 931 indivíduos não demenciados com idade  $\geq 75$  anos do Projeto Kungsholmen, Estocolmo, Suécia, foi acompanhada por 3 anos entre 1987 e 1993. Um total de 101 casos incidentes de demência, sendo 76 deles de Alzheimer, foram detectados. Indivíduos com menor nível educacional apresentaram um risco relativo ajustado de desenvolver Alzheimer de 3,4 (intervalo de confiança de 95%: 2,0, 6,0), e indivíduos com menor status socioeconômico apresentaram um risco relativo ajustado de 1,6 (intervalo de confiança de 95%: 1,0, 2,5)." [14]



## 5 Referências

### References

- [1] Daniel S. Marcus, Anthony F. Fotenos, John G. Csernansky, John C. Morris, Randy L. Buckner; Open Access Series of Imaging Studies: Longitudinal MRI Data in Nondemented and Demented Older Adults. *J Cogn Neurosci* 2010; 22 (12): 2677–2684. doi: <https://doi.org/10.1162/jocn.2009.21407>
- [2] Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace. <https://www.python.org>
- [3] McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56). <https://pandas.pydata.org>
- [4] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, 585, 357–362. <https://doi.org/10.1038/s41586-020-2649-2>
- [5] Pedregosa, F., Varoquaux, Gaël, Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830. <https://scikit-learn.org/stable/>
- [6] Seabold, S., & Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.
- [7] Precisão e revocação. (2023, June 24). In Wikipedia. [https://pt.wikipedia.org/wiki/Precis%C3%A3o\\_e\\_revoca%C3%A7%C3%A3o#Precis%C3%A3o](https://pt.wikipedia.org/wiki/Precis%C3%A3o_e_revoca%C3%A7%C3%A3o#Precis%C3%A3o)
- [8] Entropia cruzada. (2023, June 24). In Wikipedia. [https://pt.wikipedia.org/wiki/Entropia\\_cruzada#Estimativa](https://pt.wikipedia.org/wiki/Entropia_cruzada#Estimativa)
- [9] Brier score. (2023, June 24). In Wikipedia. [https://en.wikipedia.org/wiki/Brier\\_score#Definition](https://en.wikipedia.org/wiki/Brier_score#Definition)
- [10] National Institute on Aging. (1988). National Institute on Aging. Bethesda, Md. :National Institutes of Health, National Institute on Aging, <https://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease>
- [11] Rodrigues MAS, Rodrigues TP, Zatz M, Lebrão ML, Duarte YA, Naslavsky MS, do Nascimento FBP, Amaro Junior E. Quantitative evaluation of brain volume among elderly individuals in São Paulo, Brazil: a population-based study. *Radiol Bras.* 2019 Sep-Oct;52(5):293-298. doi: 10.1590/0100-3984.2018.0074. PMID: 31656345; PMCID: PMC6808618.
- [12] Perneczky R, Wagenpfeil S, Komossa K, Grimmer T, Diehl J, Kurz A. Mapping scores onto stages: minimal state examination and clinical dementia rating. *Am J Geriatr Psychiatry.* 2006 Feb;14(2):139-44. doi: 10.1097/01.JGP.0000192478.82189.a8. PMID: 16473978.
- [13] Beam CR, Kaneshiro C, Jang JY, Reynolds CA, Pedersen NL, Gatz M. Differences Between Women and Men in Incidence Rates of Dementia and Alzheimer’s Disease. *J Alzheimers Dis.* 2018;64(4):1077-1083. doi: 10.3233/JAD-180141. PMID: 30010124; PMCID: PMC6226313.
- [14] Anita Karp and others, Relation of Education and Occupation-based Socioeconomic Status to Incident Alzheimer’s Disease, *American Journal of Epidemiology*, Volume 159, Issue 2, 15 January 2004, Pages 175–183, <https://doi.org/10.1093/aje/kwh018>



Figure 8: Pairplot imensamente grande de todas as variáveis