Predição de Acidente Cardiovascular Cerebral

Marcéli Melchiors

Fundação Getúlio Vargas

Resumo
Esse trabalho irá apresentar uma modelagem estatística para a predição de Acidente Cardiovascular Cerebral (AVC), utilizando regressão logística e estatística frequentista.

Abstract
This work will present a statistical modeling for the prediction of Cerebral Cardiovascular Accident (CVA), using logistic regression and frequent statistics.

Palavras-chave: AVC, Modelagem Estatística, Predição de doenças
Keywords: Stroke, Statistical Modeling, Disease Prediction

Introdução

O Acidente Vascular Cerebral (AVC), também conhecido como derrame cerebral, é uma condição médica de extrema gravidade que ocorre quando há uma interrupção no fornecimento de sangue para uma região específica do cérebro. Essa interrupção pode ser resultado da obstrução de um vaso sanguíneo, conhecido como AVC isquêmico, ou do rompimento de um vaso sanguíneo, conhecido como AVC hemorrágico.

Os dados do Sistema de Informações sobre Mortalidade (SIM) do Ministério da Saúde revelam uma preocupante realidade: o AVC é a doença que mais mata no Brasil há mais de uma década. Apenas no ano de 2020, quase 100 mil vidas foram perdidas devido a essa terrível condição. Além da alta taxa de mortalidade, o AVC também é responsável por um considerável número de hospitalizações, o que demanda recursos significativos do sistema de saúde. Em 2021, foram registradas 164.200 internações por AVC, totalizando um custo anual de aproximadamente 250 milhões de reais, de acordo com os dados do Sistema de Informações Hospitalares. É importante destacar que os gastos com internações por AVC têm apresentado um crescimento alarmante nas últimas décadas, considerando que em 2000, o custo totalizou 36 milhões de reais.

Diante desse cenário preocupante, fica evidente que o tratamento do AVC não pode ser abordado de forma isolada e restrito apenas à área da saúde. É necessário compreender que esse problema possui dimensões sociais e econômicas significativas, dada a elevada carga financeira que acarreta aos cofres públicos.

Além do tratamento médico, uma abordagem fundamental para combater o AVC é a prevenção. Compreender os fatores de risco que podem influenciar a ocorrência dessa condição é de suma importância. Nesse sentido, a identificação de um perfil de propensão ao AVC, que englobe características que aumentam a probabilidade de ocorrência dessa condição, pode ser de grande auxílio. Através de métodos de pesquisa adequados, é possível explorar esse campo e buscar estratégias preventivas eficazes.

Assim, a implementação de medidas preventivas e a compreensão dos fatores de risco associados ao AVC são essenciais para enfrentar esse problema de saúde pública. Investir em pesquisas e políticas de prevenção é muito importante para reduzir a incidência e impacto do AVC, bem como para diminuir os custos financeiros e sociais que essa doença acarreta para o Estado brasileiro.

O trabalho consistirá em analisar um conjunto de dados que nos possibilitarao trabalhar com as variáveis de sexo, idade, estado civil, trabalho, residência, glicose, se posssui hipertensão ou doença cardíaca e se o individuo é fumante. Será utilizada regressão logística para conseguir os coeficientes e assim fazer uma análise dos fatores que mais impactam no Acidente Vascular Cerebral, desenvolvendo um perfil de propensão à AVC que poderá melhor direcionar os métodos preventivos de combate à doença.

O conjunto de dados foi retirado de https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Metodologia

1. Modelo

A regressão logística é uma técnica estatística utilizada para compreender a relação entre uma variavel de interesse e um conjunto de variáveis explicativas. Ela é utilizada em estudos que buscam prever a ocorrência de um evento, como determinar os fatores que podem influenciar o surgimento de uma doença. Por isso, é o ideal para prever e determinar o que influencia no Acidente Vascular Cerebral.

No contexto da regressão logística, a variável de interesse é denominada variável dependente ou variável resposta, enquanto as variáveis explicativas são conhecidas como variáveis independentes. O objetivo é encontrar uma relação entre essas variáveis que permitaa fazer previsões sobre a ocorrência do evento em questão.

A regressão logística difere da regressão linear tradicional, pois é adequada para lidar com variáveis dependentes binárias, ou seja, variáveis que possuem apenas dois possíveis valores, como sim ou nao, ocorreu ou não ocorreu, exatamente como o conjunto de dados utilizado nesse trabalho, que tem 1 para ocorrência de AVC e 0 para a não ocorrência. Ela utiliza uma função matemática chamada função logit para modelar a probabilidade de ocorrência do evento em função das variáveis independntes.

Ao aplicar a regressão logística, são obtidos coeficientes para cada uma das variáveis independentes, que indicam a direção e a magnitude da influência que cada variável tem sobre a variável dependente. Esses coeficientes são interpretados como o logaritmo da razão de chances (odds ratio), que ajuda a compreender como as chances de ocorrência do evento são afetadas pelas variáveis independentes.

Com base nos resultados da regressão logistica, será possível identificar quais variáveis têm um grande impacto significativo na ocorrência de AVC. Essas variáveis podem ser consideradas como potenciais fatores de risco ou indicadores relevantes. Essas informações podem ser úteis para tomar decisões informadas, implementar medidas preventivas e desenvolver estratégias que visem reduzir a probabilidade de ocorrência, como mencionado na introdução.

2. Ajustes

A regressão logística é implementada por meio de modelos lineares generalizados. Por isso, será utilizada a função glm em R.

Os modelos lineares generalizados permitem lidar com uma variedade de tipos de dados e distribuições de probabilidade. Eles foram desenvolvidos para superar as limitações dos modelos lineares clássicos, que assumem normalidade e independência.

Os GLMs são compostos por três componentes principais: função de ligação, distribuição de probabilidade e função de variância. A função de ligação relaciona a média da variável resposta às variáveis explicativas, para que as estimativas dos parâmetros sejam interpretáveis. Exemplos de funções de ligação comuns incluem a função identidade, logaritmo, logit e probit.

A distribuição de probabilidade é escolhida de acordo com a natureza dos dados e pode ser qualquer distribuição da família exponencial, como a distribuição normal, poisson, binomial ou outras.

A função logit será escolhida como função de ligação na regressão logística porque permite modelar a relação entre as variáveis independentes e a probabilidade de ocorrência de um evento. Ela transforma a escala linear das variáveis independentes em uma escala de probabilidades, garantindo que a resposta esteja restrita entre 0 e 1.

A distribuição de probabilidade para a regressão logística será a binomial porque a variável dependente é binária, assumindo apenas dois valores possíveis. A distribuição binomial descreve a probabilidade de ocorrência de um evento em um determinado número de tentativas independentes.em relação às váriaveis independentes.

Ao utilizar a função de ligação logit e a distribuição binomial na regressão logística, será possível estimar os coeficientes dos preditores e interpretá-los como log-odds ratios. Esses coeficientes indicam a mudança percentual na probabilidade de ocorrência do evento para cada unidade de mudança nas variáveiss independentes, mantendo as outras variáveis constantes.

A estimação dos parâmetros em um GLM é realizada por meio da Estimativa de Máxima Verossimilhança (MLE). Esse método busca encontrar os valores dos parâmetros que

maximizam a probabilidade de observar os dados observados, dado o modelo estatístico especificado e suas suposições. A estimação por máxima verossimilhança é uma abordagem amplamente utilizada na estatística frequentista e é considerada um dos principais métodos de estimação de parâmetros.

3. Avaliação

A bondade do modelo será estimada por meio do AIC e ROC-AUC

A - Akaike Information Criterion (AIC):

O AIC é utilizado para comparar diferentes modelos. O objetivo é encontrar um equilíbrio entre a complexidade do modelo e a capacidade de ajuste aos dados observados . Ele leva em consideração tanto o quanto o modelo se adequa aos dados, quanto o número de parâmetros estimados pelo modelo.

A fórmula do AIC é:

$$AIC = 2p - 2l,$$

sendo p: quantidade de parâmetros e l: logaritmo da função de verossimilhança maximizada do modelo. Quanto menor o AIC, melhor o modelo é ajustado em relação aos dados e à complexidade.

B - Área sobre a curva(AUC) e Característica Operacional do Receptor(ROC)

A curva ROC é um gráfico que mostra a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos.

O AUC é uma métrica numérica que resume a curva ROC em um valor numérico. Ele representa a área sob a curva ROC e varia de 0 a 1. Quanto maior o valor do AUC, melhor é o desempenho do modelo em classificar corretamente. Um AUC de 0,5 indica um desempenho aleatório, enquanto um AUC de 1 representa um modelo perfeito, capaz de diferenciar muito as classes positiva e negativa.

A curva ROC e o AUC apresentam vantagem pois são independentes do ponto de corte escolhido para a classificação e permitem comparar diferentes modelos de forma objetiva. Além disso, eles são bons quando tem desequilíbrios de classes, sendo úteis mesmo quando as proporções de amostras positivas e negativas são diferentes, como é o caso do conjunto de dados utilizado nesse trabalho, que possui 250 pessoas que tiveram AVC e quase 5000 que não tiveram.

Resultados

Nesta seção, será apresentada os resultados obtidos utilizando os dados e a metodologia já mencionada. Primeiramente, será mostrada uma breve análise visual e exploratória dos dados, como base para o ajuste do modelo.

1. Análise Exploratória dos Dados

A matriz de correlação é uma tabela que mostra as relações entre variáveis. Assim, cada célula contém o coeficiente de correlação entre duas variáveis. Esses coeficientes variam de -1 a 1 e indicam a força e direção da relação. É útil para identificar padrões e relações entre variáveis e ajuda na seleção de variáveis e evita multicolinearidade. A figura a seguir, apresenta a matriz de correlação entre as váriaveis do conjunto de dados usado para este trabalho.

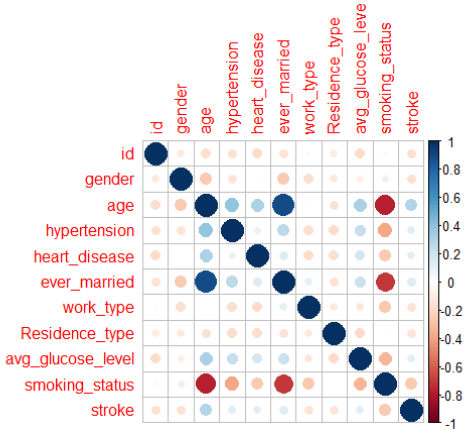

Figura 1: Matriz de Correlação

Pode-se notar que as váriaveis mais correlacionadas com a váriavel stroke são: idade, hipertensão, doença cardíaca, estado civil e status de fumante. Também pode ser percebido a alta correlação entre as variáveis de idade e estado civil, além de estado civil e status de fumante.

Os seguintes gráficos foram plotados para ter uma clareza melhor e escolher as melhores variáveis para os modelos.

O primeiro, trata-se de uma matriz de gráficos, gerando os gráficos entre todas as váriaveis preditoras. No entanto, como o conjunto de dados possui mais de 5000 linhas e 12 colunas,além de algumas variáveis categóricas e binárias, é inviável compreender e tirar alguma conclusão apenas com ele.

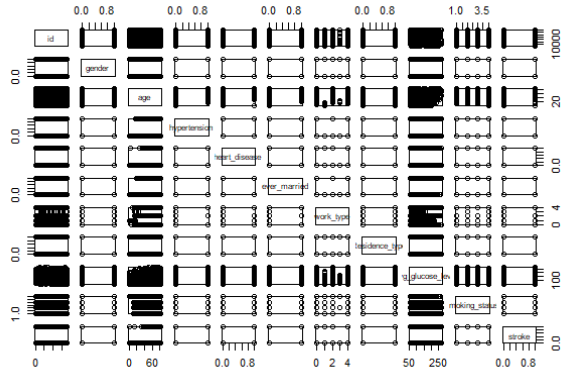

Figura 2: Matriz de Gráficos

Então, para as variáveis categóricas foram criados gráficos de barras individuais e comparadas a diferença entre as pessoas que tiveram AVC ou não tiveram AVC.

As variáveis de tipo de residência e tipo de trabalho se destacaram por apresentar apenas uma leve diferença.

No seguinte gráfico, pode ser visto a comparação de AVC com tipo de residência.

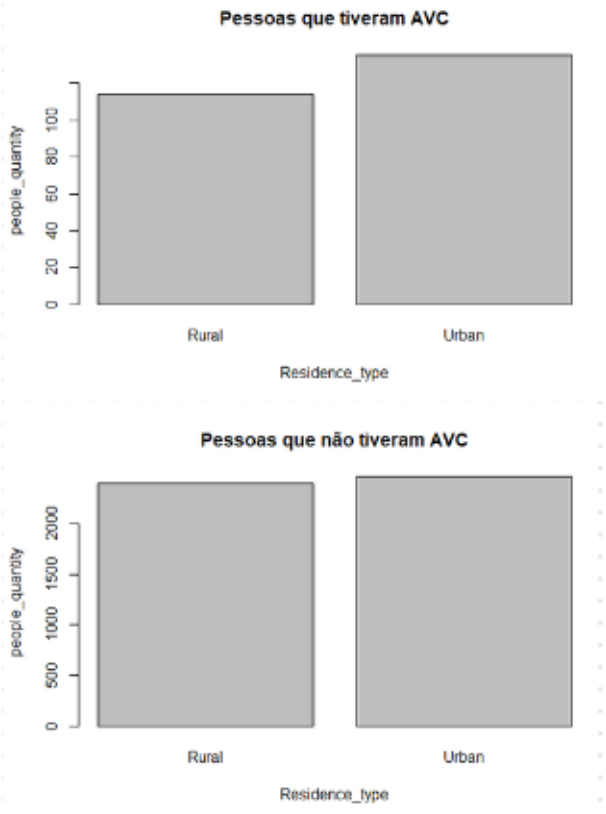

Figura 3: Relação entre AVC e tipo de residência

Na próxima figura pode ser visto a relação entre AVC e tipo de trabalho.

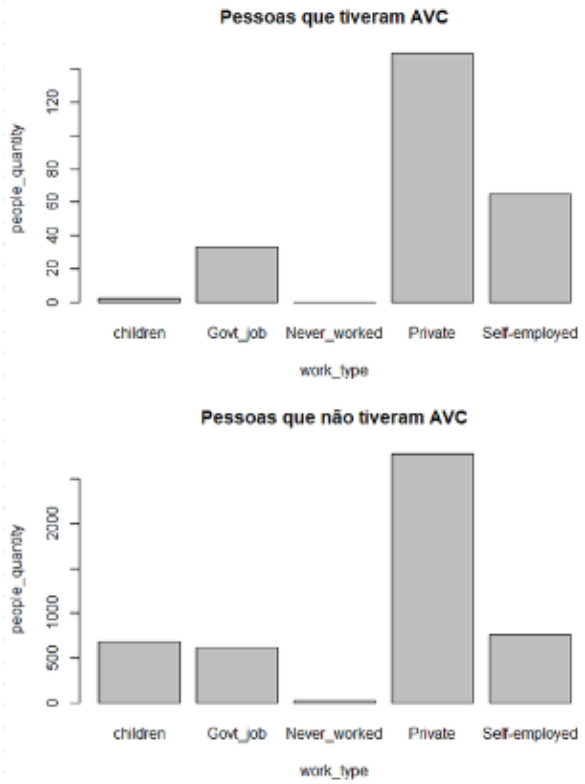

Figura 4: Relação entre AVC e tipo de trabalho.

Já as seguintes variáveis apresentam uma forte diferença entre pessoas que tiveram ou não AVC.

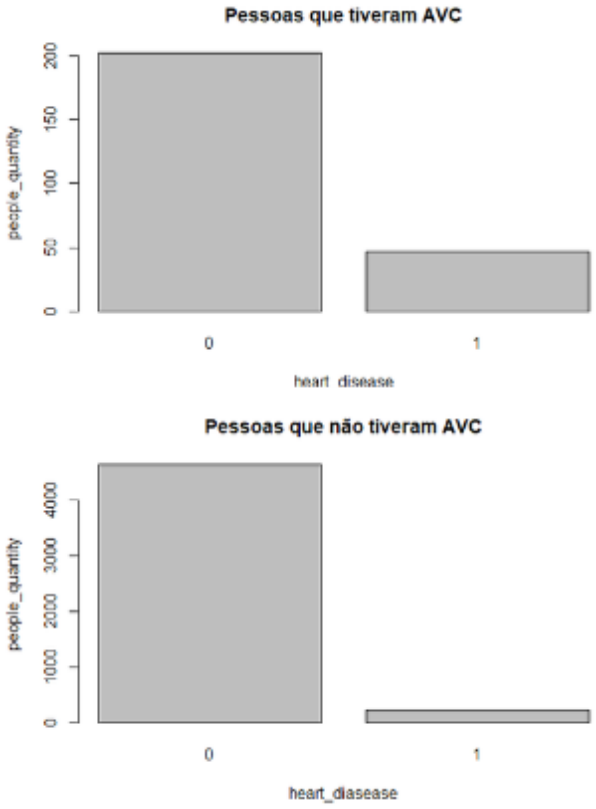

Figura 5: Relação entre AVC e doença cardíaca.

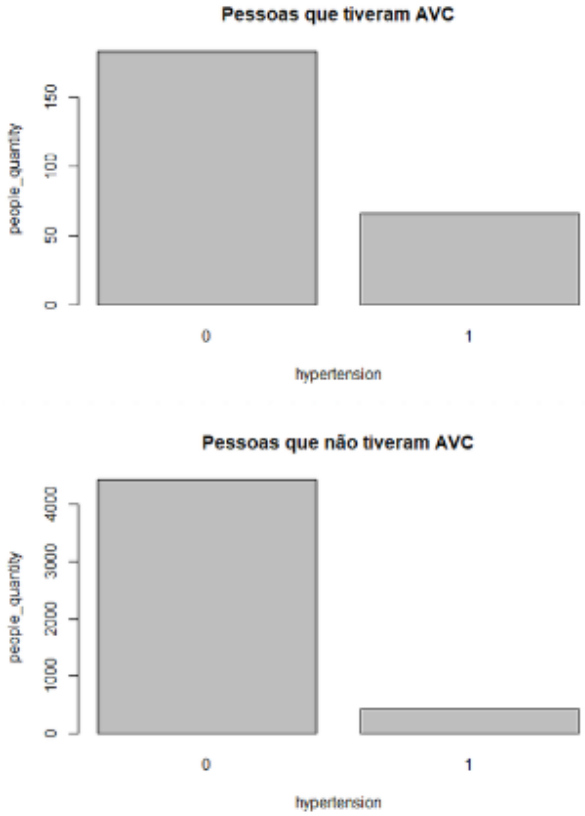

Figura 6: Relação entre AVC e hipertensão

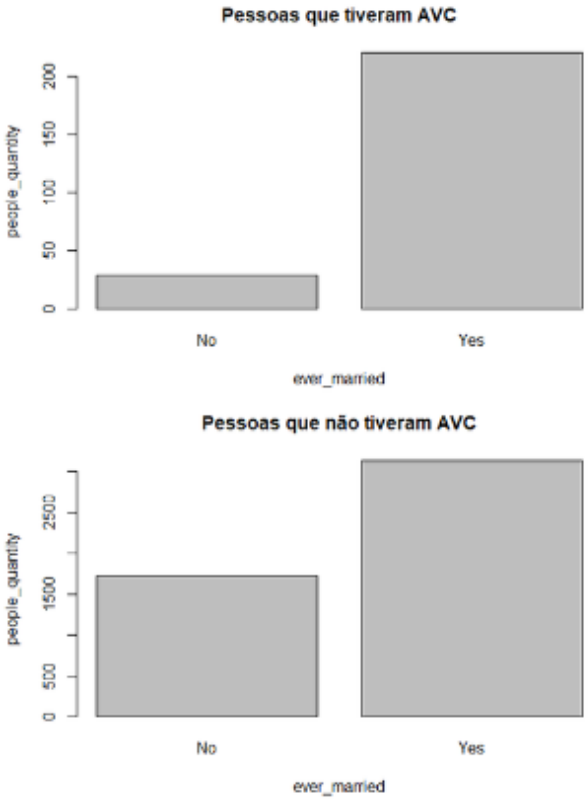

Figura 7: Relação entre AVC e estado civil.

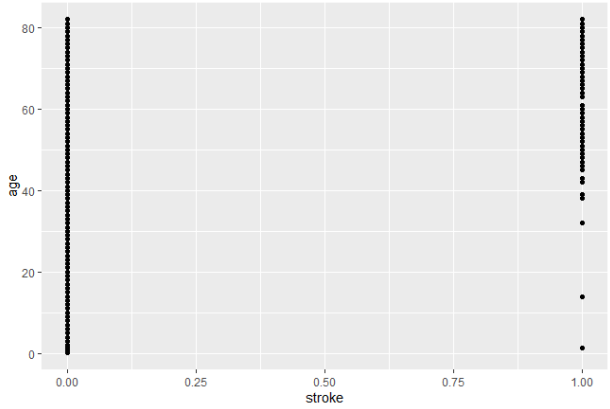

Figura 8: Relação entre AVC e idade.

Quando é plotado o gráfico entre AVC e o status de fumante, pode-se ver que há muitos dados desconhecidos. Então esse fator não será determinante para a conclusão final.

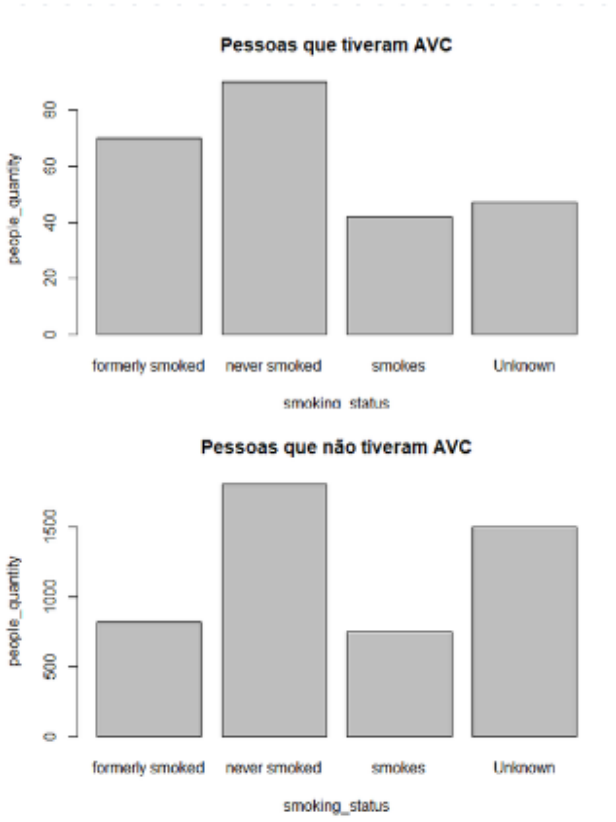

Figura 9: Relação entre AVC e tabagismo.

2. Ajustes do modelo.

Foram testados os seguintes modelos:

Modelo 1 - Com todas as váriaveis preditores: gênero, idade, hipertensão, estado civil, tipo de trabalho, tipo de residência, glicose e status de fumante.

Modelo 2 - Com as variáveis numéricas: idade,

hipertensão e glicose.

Modelo 3 - Com as variáveis numéricas, adicionando tabagismo para ver seu impacto: idade, hipertensão, glicose e status de fumante.

Modelo 4 - Com as variáveis numéricas, adicionando estado civil para ver seu impacto: idade, hipertensão, estado civil e glicose.

Modelo 5 - Sem idade, para perceber o quanto essa variável influencia: hipertensão, coração, estado civil e glicose.

Modelo 6 - Com as váriaveis que mais tem relação de acordo com a análise visual dos dados: idade, hipertensão, estado civil, doença cardíaca e glicose.

Modelo 7 - Modelo 6 adicionando status de fumante: idade, hipertensão, estado civil, doença cardíaca, glicose e status de fumante.

Modelo 8 - Retirando a multicolinearidade do modelo 6:idade, hipertensão, estado civil, doença cardíaca, glicose, idade*estado_civil.

Modelo 9 - Retirando a colinearidade do modelo 7: idade, hipertensão, estado civil, doença cardíaca, glicose, idade, idade*estado_civil, idade*status_de_fumante, status_de_fumante*estado_civil.

Como são muitos modelos, será apresentados os coeficientes e os intervalos de confiança =dos melhores modelos.

Tabela 1: Legenda da tabela

Avaliação dos modelos		
Modelo	AIC	AUC
1	1613,2	0,846
2	1602,4	0,843
3	1605	0,845
4	1603,6	0,844
5	1821,5	0,735
6	1602,8	0,844
7	1605,8	0,846
8	1603,5	0,845
9	1609,2	0,849

Os modelos 2 e 6 que possuem o menor AIC. Com relação ao AUC, ambos os modelos 2 e 6 estão muito próximos ao valor máximo de AUC dentre os 9 modelos.

Primeiramente, os resultados do modelo 2:

Tabela 2: Modelo 2

Coeficientes	Estimativa	Desvio-Padrão
Intercepto	-7.578833	0.355584
idade	0.070585	0.005062
hipertensão	0.384460	0.162331
glicose	0.004354	0.001152

Tabela 3: Intervalos de confiança - modelo 2

Coeficientes	2.5%	97.5%
Intercepto	-8.275764501	-6.881900889
idade	0.060664038	0.080506303
hipertensão	0.066297778	0.702622748
glicose	0.002095396	0.006611995

Agora, com o modelo 6:

Tabela 4: Modelo 6

Coeficientes	Estimativa	Desvio-Padrão
Intercepto	-7.381653	0.374736
idade	0.069677	0.005170
hipertensão	0.380479	0.162745
doença cardiaca	0.322838	0.188053
estado civil (casado)	-0.183790	0.218495
glicose	0.004178	0.001165

Tabela 5: Intervalos de confiança - modelo 6

Coeficientes	2.5%	97.5%
Intercepto	-8.116121643	-6.647184353
idade	0.059543758	0.079809910
hipertensão	0.061505921	0.699452796
doença cardiaca	-0.045738785	0.691415506
estado civil (casado)	-0.612033367	0.244452820
glicose	0.001894166	0.006461581

Então, a seguir será feita a interppretação do modelo 6, pois possui mais váriaveis.

O coeficiente de interceptação (-7.381653) representa o log-odds da variável dependente quando todas as variáveis independentes são iguais a zero.

Para cada aumento de uma unidade na idade, espera-se um aumento de 0.069677 na log-odds da

variável dependente, mantendo todas as outras variáveis constantes

Ter hipertensão esstá associado a um aumento de 0.380479 na log-odds da variável dependente, em comparação com não ter hipertensão, mantendo as outras variáveis constantes.

Ter doença cardíaca está associado a um aumento de 0.322838 na log-odds da variável dependente, em comparação com não ter doença cardíaca, mantendo as outras variáveis constantes

Ser casado está associado a uma diminuição de 0.183790 na log-odds da variável dependente, mantendo as outras váriaveis constantes.

Um aumento de uma unidade no nível de glicose está associado a um aumento de 0.004178 na log-odds da variável dependente, mantendo as outras variáveis constantes.

Discussão/ Conclusão

Em resumo, neste trabalho foi desenvolvido uma regressão logística para a predição de Acidente Vascular Cerebral (AVC), utilizando estatística frequentista. Os resultados mostraram que idosos, pessoas com hipertensão alta, nível de glicose elevado ou/e histórico de doença cardíaca devem preocupar-se com AVC. Além disso, ser casado implica menos chance de ter a doença.

A seguir, será discutido alguns pontos que fazem esses fatores influenciarem em ter essa doença ou não.

Dado que à medida que um ser humana envelhece, seu sistema vascular sofre mudanças naturais, como o estreitamento das artérias e o acúmulo de placas de gordura, o que pode levar à obstrução do fluxo sanguíneo para o cérebro.

Já a principal forma de AVC associada à hipertensãoo é o AVC isquêmico. Nesse tipo de AVC, um coágulo ou placa de gordura bloqueia uma artéria que fornece sangue ao cérebro, resultando em uma diminuição ou interrupção do fluxo sanguíneo. A hipertensão arterial é um fator de risco significativo para o desenvolvimento dessas obstruções arteriais.

A estreita relação entre doença cardíaca e AVC se deve ao fato de que as artérias obstruídas ou danificadas no sistema cardiovascular também podem afetar o fornecimento de sangue ao cérebro. Quando um coágulo sanguíneo ou uma placa de gordura se desprende e bloqueia uma artéria cerebral, ocorre um AVC isquêmico. Além disso, a doença cardíaca pode causar anormalidades no ritmo cardíaco, aumentando o risco de formação de coágulos sanguíneos no coração que podem se desprender e se deslocar para o cérebro, causando um AVC embólico.

Com relaçãoo ao estado civil, poderia haver um estudo mais aprofundado a respeito desse resultado, para entender se há relação causal direta.

Apesar da maioria das implicações dos resultados serem lógicas, foi desenvolvido um modelo de regressão logística com base em um conjunto com uma grande diferença de pessoas que tiveram AVC ou não. Além disso, muitos outros fatores que podem causar AVC não fazem parte do conjunto de dados, fazendo com que o modelo seja limitado. A falta de dados com relação ao tabagismo também limita algumas conclusões, dado que pode ser uma área muito explorada.

Futuramente, poderá ser feito um estudo mais aprofundado sobre a relação do tabagismo com o Acidente Vascular Cerebral. Além disso, estudar as relações causais das variáveis independentes com a variável dependente pode ser útil para investir em saúde pública e reduzir os altos índices de AVC.

Referências

[1] Autor: Andrew Gelman, J. Hill, and Aki Vehtari, Regression and other stories. Editorial: Cambridge Cambridge University Press, 2021

[2] https://www.r-project.org/other-docs.html

[3] https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[4] https://www.gov.br/saude/pt-br/assuntos/saude-de-a-a-z/a/avc