UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA E CIÊNCIAS ATUARIAIS

Raquel Santos Evangelista

Previsão de AVC em Sergipe utilizando modelos de regressão logísticos com dados da Pesquisa Nacional de Saúde - PNS.

São Cristóvão - SE

2024

Raquel Santos Evangelista

Previsão de AVC em Sergipe utilizando modelos de regressão logísticos com dados da Pesquisa Nacional de Saúde - PNS.

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Cleber Martins Xavier

São Cristóvão - SE

2024

Raquel Santos Evangelista

Previsão de AVC em Sergipe utilizando modelos de regressão logísticos com dados da Pesquisa Nacional de Saúde - PNS.

Trabalho de Conclusão de Curso apresentado ao Departamento de Estatística e Ciências Atuariais da Universidade Federal de Sergipe, como parte dos requisitos para obtenção do grau de Bacharel em Estatística.

Aprovado em 29/10/2024.

Prof. Dr. Cleber Martins Xavier
Orientador

Prof. Dr. José Rodrigo Santos Silva
Membro da Banca

Prof. Dr. Luiz Henrique Gama Dore de Araújo
Membro da Banca

São Cristóvão - SE
2024

Dedico este trabalho à minha avó, minha melhor amiga, que sempre foi meu porto seguro. Perder parte dela despertou em mim o desejo de entender o AVC e buscar formas de preveni-lo.

Agradecimentos

Agradeço, em primeiro lugar, a Deus, cuja graça e amor tornaram este momento possível. Seu cuidado e direção foram fundamentais para que eu seguisse em frente, e é a Ele que dedico todas as coisas, pois é por Ele e para Ele que tudo existe. Reconheço que foi Sua imensa bondade que me guiou até aqui, encorajando-me a continuar e alcançar o fim deste caminho.

Expresso minha profunda gratidão à minha mãe, uma verdadeira guerreira que, sendo mãe solo, provou-se ainda mais incrível, jamais medindo esforços para me ajudar a realizar meus sonhos. Ela sempre me incentivou a sonhar alto, esteve ao meu lado em todos os momentos, acreditou em mim e me amou incondicionalmente. Agradeço também à minha avó, que representa para mim amor e compaixão, sendo o grande amor da minha vida e a razão pela qual escolhi estudar o AVC, que a afetou, mas não a derrotou. Sou igualmente grata ao Igor Estefano, que sempre me incentivou a nunca desistir, acreditando no meu potencial e me motivando com seu apoio constante, sendo uma presença fundamental na minha vida.

Sou imensamente grata pelas amizades que construí ao longo da graduação, que tornaram essa jornada mais leve com seu companheirismo e carinho, são momentos que nunca serão esquecidos. Em especial, agradeço a Lauryane e Gilmara, os maiores presentes que recebi neste curso, e que foram essenciais para minha formação.

Externo meus agradecimentos também ao corpo docente do departamento de Estatística, por toda a contribuição na minha trajetória acadêmica. São profissionais excepcionais, que inspiram e despertam nossa paixão pelo curso, especialmente ao meu orientador, Cleber Xavier, cuja paciência, orientação, disponibilidade e motivação foram essenciais. Seu profissionalismo é algo que admiro profundamente.

Agradeço a todos que, de alguma forma, fizeram parte dessa trajetória e contribuíram para que este momento se tornasse realidade.

“É justo que muito custe
o que muito vale.
(It’s fair that a lot costs
what a lot is worth.)”
(Santa Teresa d’Ávila)

Resumo

O Acidente Vascular Cerebral (AVC) é uma das principais causas de mortalidade e incapacidade no Brasil, tornando-se fundamental a criação de ferramentas de predição que permitam a intervenção precoce e a redução dos impactos dessa condição na saúde pública. O presente trabalho tem como objetivo geral desenvolver um modelo preditivo utilizando regressão logística para prever a ocorrência de Acidente Vascular Cerebral (AVC) com base em dados da Pesquisa Nacional de Saúde (PNS) para o Estado de Sergipe. Realizado em 2019 pelo Ministério da Saúde em parceria com o IBGE, o PNS tem o objetivo de caracterizar a situação de saúde e os estilos de vida da população. Para a construção do modelo preditivo, a metodologia adotada seguiu uma abordagem quantitativa. Dada a natureza desbalanceada dos dados, com uma maior quantidade de pacientes sem histórico de AVC em comparação àqueles com histórico da doença, foram aplicadas técnicas de balanceamento, como *up-sampling*, *down-sampling* e *smote*. Esses métodos foram essenciais para equilibrar a base de dados antes da aplicação da regressão logística, minimizando possíveis vieses e maximizando a precisão do modelo. O desenvolvimento e a análise foram realizados no ambiente de programação R, que permitiu a implementação das técnicas de balanceamento e a avaliação do desempenho do modelo. Os resultados identificaram que o modelo com *down-sampling* conseguiu indentificar melhor os portadores da doença, por ter uma boa acurácia e sensibilidade. As três variáveis mais relevantes para o modelo escolhido para a predição de AVC foram a idade, cor branca e cor parda. O modelo desenvolvido visa auxiliar na identificação de indivíduos com maior risco, promovendo um suporte valioso para estratégias preventivas e tomadas de decisão informadas na área de saúde em Sergipe.

Palavras-chave: AVC, PNS, Regressão Logística e Desbalanceamento.

Abstract

Stroke is one of the main causes of mortality and disability in Brazil, making it essential to create prediction tools that allow early intervention and reduce the impacts of this condition on public health. The general objective of this study is to develop a predictive model using logistic regression to predict the occurrence of Stroke based on data from the National Health Survey (PNS) for the State of Sergipe. Conducted in 2019 by the Ministry of Health in partnership with IBGE, the PNS aims to characterize the health situation and lifestyles of the population. To build the predictive model, the methodology adopted followed a quantitative approach. Given the unbalanced nature of the data, with a greater number of patients without a history of stroke compared to those with a history of the disease, balancing techniques such as over-sampling, under-sampling, and smote were applied. These methods were essential to balance the database before applying logistic regression, minimizing possible biases and maximizing the accuracy of the model. The development and analysis were performed in the R programming environment, which allowed the implementation of balancing techniques and the evaluation of the model's performance. The results showed that the down-sampling model was able to better identify individuals with the disease, as it had good accuracy and sensitivity. The three most relevant variables for the model chosen to predict stroke were age, white skin color, and brown skin color. The model developed aims to help identify individuals at higher risk, providing valuable support for preventive strategies and informed decision-making in the health area in Sergipe.

Keywords: Stroke, PNS, Logistic Regression and Imbalance

Lista de ilustrações

Lista de quadros

Lista de tabelas

Sumário

1 INTRODUÇÃO

O Acidente Vascular Cerebral (AVC) é uma das principais causas de mortalidade e incapacidade no mundo, sendo caracterizado pela interrupção do fluxo sanguíneo para o cérebro, o que pode resultar em danos neurológicos permanentes. Existem dois tipos principais de AVC: o isquêmico, causado pela obstrução de vasos sanguíneos, e o hemorrágico, provocado pela ruptura desses vasos. Em ambos os casos, o diagnóstico e o tratamento rápidos são essenciais para minimizar as sequelas, que podem variar de acordo com a extensão e localização da lesão cerebral. As consequências incluem comprometimentos na fala, nos movimentos e nas funções cognitivas, impactando significativamente a qualidade de vida dos sobreviventes. Devido à sua alta prevalência e ao impacto severo na saúde pública, o AVC é um dos maiores desafios globais na área da saúde (OLIVEIRA, 2010).

No contexto brasileiro, a Pesquisa Nacional de Saúde (PNS) se destaca como uma ferramenta essencial para entender as condições de saúde da população, coletando informações detalhadas sobre doenças crônicas e fatores de risco, como o AVC. Coordenada pelo IBGE em parceria com o Ministério da Saúde, a PNS fornece dados valiosos que apoiam o desenvolvimento de políticas públicas voltadas para a prevenção e o tratamento de condições de saúde de grande impacto, como o AVC (IBGE, 2021).

Considerando que o AVC é uma doença silenciosa, frequentemente manifestando-se sem sinais prévios, e que suas primeiras 24 horas são cruciais para limitar danos cerebrais (ROLIM; MARTINS, 2011), este trabalho visa desenvolver um modelo preditivo capaz de, com base no histórico de saúde do paciente, identificar aqueles com maior risco de sofrer um AVC. Esse tipo de classificação antecipada permite que intervenções preventivas sejam iniciadas a tempo, possibilitando um acompanhamento mais eficaz e adequado às necessidades de cada indivíduo.

Assim, através de técnicas de análise de dados, balanceamento e aprendizado de máquina, busca-se construir uma ferramenta que auxilie na identificação precoce de indivíduos com maior probabilidade de sofrer um AVC, permitindo intervenções preventivas mais eficazes e personalizadas. O modelo preditivo será desenvolvido utilizando a regressão logística, uma técnica amplamente empregada em problemas de classificação binária, e contará com o balanceamento de classes através de técnicas de *down-sampling*, *up-sampling* e *smote* para garantir maior precisão e evitar viéses de predição nos resultados.

O presente trabalho está organizado em 6 capítulos. No Capítulo 2, são apresentados os objetivos geral e específico do estudo. O Capítulo 3 traz a revisão de literatura, abordando o Acidente Vascular Cerebral (AVC), a base de dados da Pesquisa Nacional de Saúde (PNS), os conceitos de aprendizado de máquina, incluindo o modelo logístico, as métricas de validação do modelo e trabalhos correlatos. No Capítulo 4, são descritos o delineamento da pesquisa,

o tratamento dos dados, as variáveis utilizadas, os métodos de balanceamento e os suportes computacionais adotados. O Capítulo 5 apresenta os resultados e discussões, finlizando com o Capítulo 6, onde se encontram a conclusão do trabalho e sugestões para estudos futuros.

2 OBJETIVOS

2.1 Geral

O objetivo geral deste estudo é desenvolver um modelo preditivo baseado em regressão logística para prever a ocorrência de Acidente Vascular Cerebral (AVC), utilizando dados da Pesquisa Nacional de Saúde (PNS) referentes ao estado de Sergipe.

2.2 Específicos

- Descrever as características presentes no conjunto de dados, analisando variáveis relacionadas aos fatores de risco para o AVC.
- Aplicar técnicas para lidar com o desbalanceamento dos dados, como *Down-Sampling*, *Up-Sampling* e *Smote*, visando equilibrar as classes da variável de interesse.
- Identificar o modelo preditivo mais adequado para a previsão do risco de AVC utilizando diferentes medidas de desempenho.
- Realizar previsões utilizando o modelo selecionado comparando os resultados com os dados de teste para avaliar a sua precisão e desempenho.

3 REVISÃO LITERÁRIA

3.1 Acidente Vascular Cerebral - AVC

Segundo a Organização Mundial da Saúde (OMS), o acidente vascular cerebral (AVC) é caracterizado pelo surgimento rápido de distúrbios focais da função cerebral, com sintomas que podem durar 24 horas ou mais, de origem vascular, afetando aspectos sensoriais, motores e cognitivos, dependendo da extensão da lesão. O sinal mais comum do AVC, que ocorre com maior frequência em adultos, é a fraqueza ou dormência súbita no rosto, braço ou perna, podendo afetar todo o corpo ou apenas um lado. Outros sintomas frequentes incluem dificuldade para falar ou compreender, perda de consciência, dor de cabeça intensa, perda auditiva, falta de coordenação e equilíbrio, alterações cognitivas, tontura e confusão mental. Em casos graves, pode ocorrer até morte súbita (POMPERMAIER et al., 2020).

Os acidentes vasculares cerebrais (AVCs) estão entre as principais causas de mortalidade e incapacidade física nos países desenvolvidos. De acordo com dados do Ministério da Saúde - MS, o AVC afeta 16 milhões de pessoas anualmente em todo o mundo. No Brasil, apenas em 2018, foram registrados 197 mil atendimentos relacionados ao AVC no Sistema Único de Saúde - SUS (OLIVEIRA; ANDRADE, 2001).

Segundo projeções da Organização Mundial da Saúde (OMS), até 2030, o AVC deverá continuar sendo a segunda principal causa de morte no mundo, representando 12,2% dos óbitos estimados para esse ano (ANDRADE et al., 2024). Sergipe é visto como oitavo maior estado do Brasil em número de pessoas conviventes com as sequelas do AVC (LOTUFO et al., 2017).

3.1.1 Tipos de AVC

Existem dois tipos principais de AVC, o isquêmico e o hemorrágico. No AVC isquêmico, há uma interrupção do fluxo sanguíneo cerebral, geralmente superior a 24 horas, devido à obstrução de um vaso, seja por embolia ou aterosclerose, causando disfunção neurológica nas áreas afetadas (POMPERMAIER et al., 2020). Por outro lado, o AVC hemorrágico ocorre quando um vaso sanguíneo se rompe, provocando um derrame no espaço intracerebral, seja de forma intraparenquimatosa ou subaracnoide, também resultando em alterações neurológicas significativas (FIGUEIREDO; PEREIRA; MATEUS, 2020) .

O AVC isquêmico (AVCI) é caracterizado por um déficit neurológico focal persistente, resultante de uma isquemia que leva ao infarto, causado pela obstrução proximal de uma artéria. Essa obstrução pode ser provocada por um trombo, êmbolo ou compressão por tumor. Com a interrupção do fornecimento de glicose aos neurônios na área afetada, o quadro clínico se desenvolve rapidamente. Alguns minutos após a isquemia, ocorre o infarto, ou seja, a morte do tecido cerebral atingido (LONGO et al., 2013).

Já o AVC hemorrágico (AVCH) é caracterizado por um sangramento dentro do tecido cerebral, cerebelo ou tronco cerebral, geralmente causado pela ruptura de um pequeno vaso penetrante. Esse extravasamento de sangue sob pressão no tecido cerebral provoca manifestações clínicas, cuja intensidade e características dependem da localização e do volume da hemorragia (MACHADO et al., 2010).

Mamed et al. (2019) revelou que, em 2017, dos 174 óbitos por AVC não especificado em Sergipe, 120 (69%) foram investigados, e mais da metade desses casos (50,8%) tiveram a causa básica de óbito revisada. Esses dados apontam uma falha significativa no reconhecimento do tipo específico de AVC antes do óbito, um fato preocupante, já que as intervenções a serem adotadas dependem diretamente dessa diferenciação. A imprecisão no diagnóstico pode estar associada ao manejo inadequado dos pacientes com AVC isquêmico (AVCi). Um estudo observacional que avaliou a qualidade do atendimento ao AVCi no SUS mostrou que a realização de pelo menos um exame de tomografia computadorizada (TC), etapa crucial no tratamento do AVC agudo, ocorreu em apenas 28% das internações. Entre as hospitalizações onde a TC não foi realizada, 59,6% ocorreram em unidades que possuíam o equipamento disponível pelo SUS. A região Nordeste registrou a maior taxa de não realização do exame em locais com o aparelho disponível (85,3%).

Rolim e Martins (2011) também demonstraram que o grupo de pacientes que realizou a tomografia nas primeiras 24 horas apresentou uma mortalidade hospitalar bruta de 30,7%, enquanto entre aqueles que não fizeram o exame, essa taxa foi de 95,4%. Esses dados evidenciam as falhas no atendimento às vítimas de AVC na região e mostram as consequências severas dessa situação.

3.1.2 Fatores de risco

O AVC está, na maioria dos casos, associado a fatores cardiovasculares que podem ser prevenidos. Diversos fatores elevam o risco de desenvolver um AVC, dividindo-se em categorias imutáveis e mutáveis. Os fatores imutáveis, que não podem ser alterados ou tratados, incluem predisposições hereditárias ou congênitas, idade e sexo. Já os fatores mutáveis, sobre os quais é possível intervir, abrangem hipertensão arterial sistêmica (HAS), altos níveis de colesterol, aterosclerose, condições que aumentam a coagulabilidade sanguínea, diabetes mellitus (DM), tabagismo, obesidade, estresse, sedentarismo, consumo de álcool e uso de contraceptivos hormonais (GUS; FISCHMANN; MEDINA, 2002).

A agregação familiar de doenças cardiovasculares (DCV) pode estar associada à presença de comportamentos específicos (como tabagismo e consumo de álcool) ou a fatores de risco (como hipertensão, diabetes mellitus e obesidade) influenciados tanto por elementos ambientais quanto genéticos. Diferentemente dos fatores de risco genéticos clássicos de herança mendeliana, em que geralmente uma única mutação causa diretamente uma doença, os fatores genéticos de características complexas podem aumentar o risco sem necessariamente desencadear a condição em todos os casos (MEMBERS et al., 2012).

Alguns dos principais fatores de risco para AVC já estão bem estabelecidos, como hipertensão arterial sistêmica, tabagismo, dislipidemias, obesidade, sedentarismo, diabetes mellitus e histórico familiar. É fundamental compreender a prevalência desses fatores de risco, isoladamente ou em combinação, pois a redução desses riscos e a implementação de programas de prevenção primária e secundária são essenciais para a efetividade de qualquer iniciativa em saúde (GUS; FISCHMANN; MEDINA, 2002).

3.2 Pesquisa Nacional de Saúde - PNS

A Pesquisa Nacional de Saúde (PNS) é um estudo domiciliar de abrangência nacional, realizado em 2013 e 2019 pelo Ministério da Saúde em parceria com o Instituto Brasileiro de Geografia e Estatística (IBGE). A Pesquisa Nacional de Saúde, fruto de convênio com o Ministério da Saúde, foi realizada, pela primeira vez, em 2013, com o propósito de ampliar o escopo temático dos Suplementos de Saúde da Pesquisa Nacional por Amostra de Domicílios - PNAD investigados pelo IBGE até 2008. No ano de 2019, a PNS realizou, então, mais um processo de coleta de dados, também em parceria com o Ministério da Saúde, com o intuito de promover a comparação dos indicadores divulgados anteriormente e fornecer aportes à resposta do Sistema Único de Saúde - SUS (IBGE, 2021).

Seu objetivo é caracterizar a situação de saúde e os estilos de vida da população, além de avaliar o acesso e uso dos serviços de saúde, ações preventivas, continuidade dos cuidados e o financiamento da assistência. A amostra inclui 80.000 domicílios, permitindo a estimativa de diversos indicadores nas Unidades Federativas, capitais e regiões metropolitanas. O questionário da pesquisa é dividido em três partes: as duas primeiras são respondidas por um residente e abordam características do domicílio e a situação socioeconômica e de saúde de todos os moradores. A terceira parte, o questionário individual, é respondido por um morador adulto (18 anos ou mais), selecionado aleatoriamente entre os residentes. Esse questionário foca em morbidade e estilos de vida. Para esse indivíduo, também foram realizadas medições de peso, altura, circunferência da cintura, pressão arterial e exames laboratoriais para avaliar o perfil lipídico, nível de glicose no sangue e teor de sódio na urina (SZWARCWALD et al., 2014).

O principal objetivo da PNS 2019 foi fornecer ao país informações detalhadas sobre os determinantes, condicionantes e necessidades de saúde da população brasileira, possibilitando a implementação de medidas consistentes que auxiliem na formulação de políticas públicas e aumentem a eficácia das intervenções em saúde (STOPA et al., 2020).

3.3 Aprendizado de máquina

O Aprendizado de Máquina é um ramo da Inteligência Artificial (IA) voltado para o desenvolvimento de técnicas computacionais que permitem o aprendizado e a construção de sistemas capazes de adquirir conhecimento de forma autônoma. Um sistema de aprendizado é

um programa de computador que toma decisões com base nas experiências acumuladas por meio da resolução bem-sucedida de problemas anteriores. Esses sistemas apresentam características específicas e compartilhadas que permitem sua classificação de acordo com a linguagem de descrição, o modo de operação, o paradigma e a forma de aprendizado utilizada (MONARD; BARANAUSKAS, 2003).

O objetivo do Aprendizado de Máquina (AM) é desenvolver programas que aprimorem seu desempenho com base em exemplos. Para que isso ocorra, é essencial dispor de uma quantidade significativa de exemplos, que servem como base para que o computador gere conhecimento na forma de hipóteses construídas a partir dos dados (CH; MAP, 1997). Existem três principais tipos de Aprendizado de Máquina: Supervisionado, Não Supervisionado e por Reforço.

No Aprendizado Supervisionado, para cada exemplo apresentado ao algoritmo, é fornecida a resposta desejada, ou seja, um rótulo indicando a que classe o exemplo pertence. O objetivo do algoritmo é construir um classificador capaz de identificar corretamente a classe de novos exemplos não rotulados. Quando os rótulos de classe são discretos, o problema é conhecido como classificação; quando envolvem valores contínuos, é chamado de regressão, sendo o método de aprendizado mais amplamente utilizado (LUDERMIR, 2021).

No Aprendizado Não Supervisionado, os exemplos são fornecidos ao algoritmo sem rótulos. O algoritmo agrupa os exemplos com base nas similaridades de seus atributos, formando clusters ou agrupamentos. Após a formação dos clusters, é necessária uma análise para interpretar o significado de cada grupo no contexto do problema sendo estudado. Já no Aprendizado por Reforço, o algoritmo não recebe a resposta correta, mas sim um sinal de reforço, como recompensa ou punição. O algoritmo faz hipóteses com base nos exemplos e avalia se suas decisões foram boas ou ruins. Essa técnica é amplamente utilizada em jogos e robótica (LUDERMIR, 2021).

A aplicação do Aprendizado de Máquina para resolver problemas exige alguns pré-requisitos. É necessário ter um bom conjunto de exemplos, que muitas vezes precisa ser construído e constantemente atualizado. Como os dados nem sempre são de alta qualidade, é essencial aplicar técnicas que melhorem sua confiabilidade. Além disso, nem todos os algoritmos de Aprendizado de Máquina são adequados para todos os problemas, sendo necessária uma seleção cuidadosa dos algoritmos mais apropriados. Após essa escolha, é importante definir os parâmetros do algoritmo. Durante o treinamento, é crucial avaliar se o algoritmo está resolvendo o problema com precisão. Por fim, o sistema deve ser atualizado regularmente, pois mudanças nos dados podem comprometer seu desempenho (LUDERMIR, 2021).

Algoritmos de aprendizado de máquina foram criados para prever com precisão valores futuros ($\hat{E}$) com base em um conjunto de características (X), buscando não apenas ajustar-se bem aos dados passados, mas também generalizar para novos dados. Para avaliar essa capacidade de generalização, os dados são divididos em conjunto de treinamento e teste. O conjunto de treinamento é utilizado para desenvolver e ajustar o modelo, enquanto o conjunto de teste fornece

uma estimativa imparcial do desempenho, conhecido como erro de generalização. Normalmente, a divisão dos dados segue proporções como 70% – 30% ou 80% – 20%, mas é necessário equilibrar para evitar *overfitting* ou *underfitting* dos parâmetros do modelo (GREENWELL, 2022).

Para avaliar o desempenho de generalização de um modelo, é comum utilizar a validação, que envolve dividir o conjunto de treinamento em duas partes: uma para treinar o modelo e outra para estimar seu desempenho. Os dois principais métodos de reamostragem são a validação cruzada k-fold CV e o bootstrap. Na validação cruzada k-fold CV, os dados são divididos em k partes, e cada uma é usada como conjunto de teste uma vez, sendo comum usar $k = 5$ ou $k = 10$. À medida que k aumenta, a estimativa de desempenho se aproxima mais do real, embora isso traga maior custo computacional. Repetir o k-fold CV, como sugerido por (KIM, 2009), melhora a precisão, especialmente em conjuntos de dados menores (menos de 10.000 observações).

A Figura 1 apresenta o conjunto de dados original (mostrado em amarelo) particionado em dados de treinamento (*training*) e em dados de teste (*testing*). Os dados de treinamento são particionados aleatoriamente em K folds. Em seguida, $K - 1$ partes são usadas para treinar um modelo (mostrado em azul) e a parte restante é usada para avaliação (mostrada em laranja). Este processo é repetido K vezes para todas as escolhas possíveis do conjunto de teste, produzindo erros de teste. O desempenho final é relatado pela média dos erros de cada iteração.

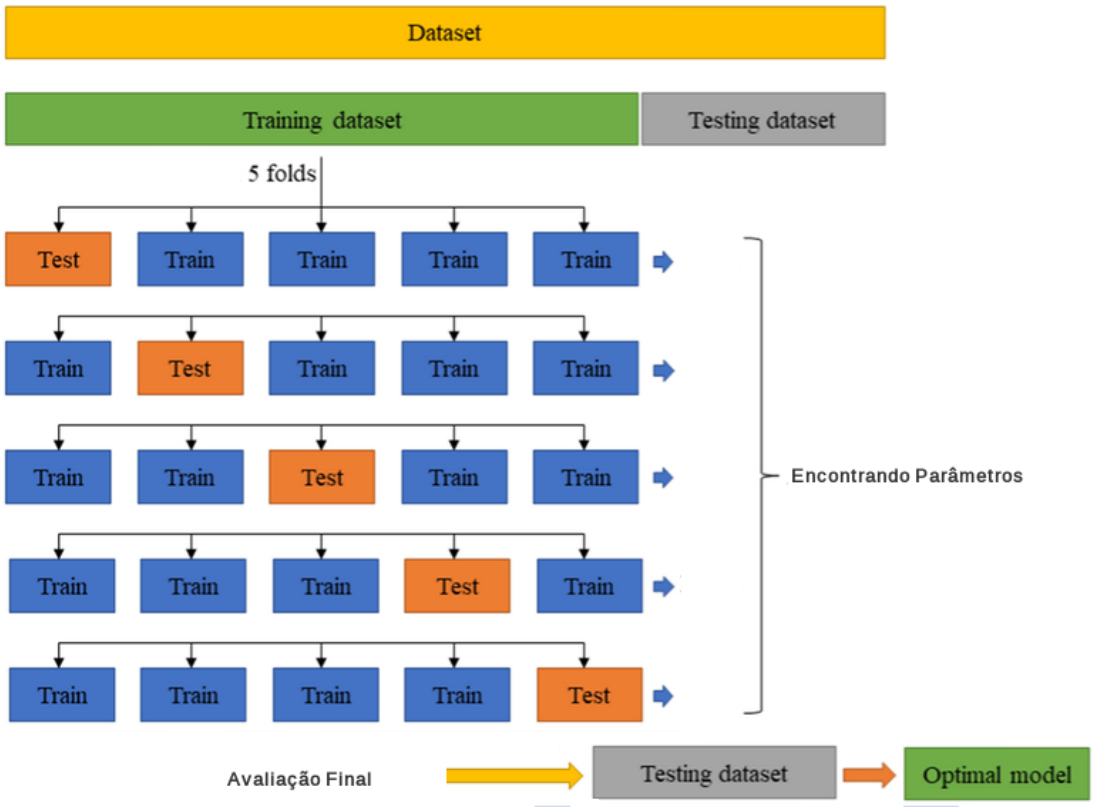
Segundo Efron e Tibshirani (1986), no bootstrap as amostras são feitas com substituição, o que significa que um ponto de dados pode ser selecionado mais de uma vez. Uma porcentagem da amostra original aparece em cada subconjunto bootstrap, e os dados que não são selecionados são chamados de "out-of-bag"(OOB). O modelo é treinado com as amostras bootstrap e validado nos dados OOB. Embora o bootstrap tenha menor variabilidade nas estimativas de erro em comparação com a validação cruzada *k-fold*, ele pode aumentar o viés da estimativa do erro, o que geralmente não é um problema para conjuntos de dados maiores (GREENWELL, 2022).

3.4 Modelo de regressão logística

A técnica de regressão logística foi desenvolvida por volta da década de 1960, como uma resposta ao desafio de prever ou explicar a ocorrência de certos fenômenos quando a variável dependente era de natureza binária. Um dos primeiros estudos a conferir notoriedade a essa técnica foi o "Framingham Heart Study", realizado em parceria com a Universidade de Boston, com o objetivo de identificar fatores associados a doenças cardiovasculares. Esse estudo analisou uma amostra de 5.209 indivíduos, com idades entre 30 e 60 anos, residentes na cidade de Framingham, Massachusetts. Usando a regressão logística, foram identificados fatores de risco como hipertensão arterial, níveis de colesterol, tabagismo, obesidade, diabetes e sedentarismo (FÁVERO et al., 2009).

A regressão logística é uma técnica estatística utilizada para descrever o comportamento

Figura 1 – Ilustração esquemática da validação cruzada *K-fold* para $K = 5$ folds.



Fonte: Adaptado de Tran et al. (2022).

entre uma variável dependente binária e variáveis independentes que podem ser quantitativas ou qualitativas. Ou seja, ela é usada para investigar o efeito de várias variáveis pelas quais indivíduos, objetos ou sujeitos estão expostos, analisando a probabilidade de ocorrência de um determinado evento de interesse (FÁVERO et al., 2009).

A popularidade dessa técnica deriva não apenas de sua capacidade de prever a ocorrência de eventos de interesse, mas também por fornecer a probabilidade associada a essa ocorrência. Portanto, o objetivo da regressão logística é tanto calcular a probabilidade de ocorrência de um evento quanto identificar as características dos indivíduos em cada grupo da variável categórica. A curva logística tem um formato sigmoidal (em "S") e os valores que ela assume variam entre 0 e 1, representando, assim, a probabilidade de o evento de interesse ocorrer (FÁVERO et al., 2009).

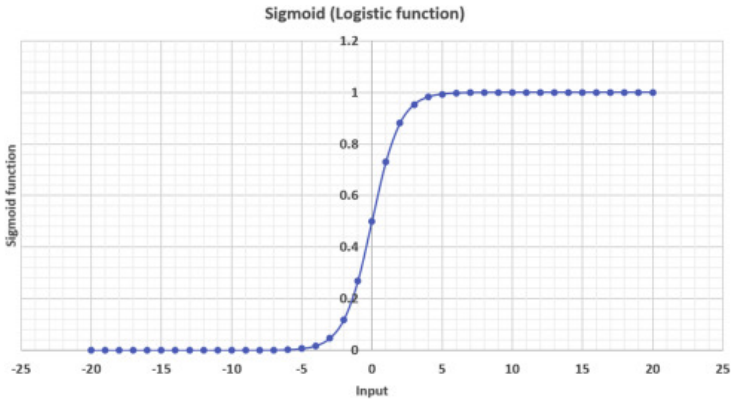
A função logística é dada por

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

e assume valores entre 0 e 1 para qualquer valor de Z entre $-\infty$ e $+\infty$ conforme ilustrado na Figura 2.

Considere uma variável resposta binária Y que pode-se modelar a probabilidade de sucesso

Figura 2 – Representação gráfica da função sigmoide.



Fonte: Belyadi e Haghighat (2021).

$P(Y = 1)$. No caso de uma única variável explicativa, considere o valor de $P(Y = 1) = \pi(x)$ para enfatizar que essa probabilidade depende do valor da variável x. O modelo de regressão logístico tem uma forma linear para o *logit* da probabilidade de sucesso, ou seja, o logaritmo da razão das probabilidades de sucesso e fracasso (AGRESTI, 2007),

$$\text{logit}\left[\pi(x)\right] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x.$$

A regressão logística tem uma fórmula correspondente para $\pi(x)$, usando a função exponencial que é dada por

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}},$$

em que parâmetro de efeito β determina a taxa de aumento ou diminuição da curva para $\pi(x)$. O sinal de β indica quando se a curva sobe ($\beta > 0$) ou diminui ($\beta < 0$).

O modelo de regressão logística geral com múltiplas variáveis explicativas que podem ser quantitativas, categóricas ou ambas é dado por (AGRESTI, 2007)

$$\text{logit}\left[\pi(x)\right] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

em que o parâmetro β_j reflete o efeito da variável x_j nas log odds de $Y = 1$.

Considere que para cada ponto dos dados de treinamento têm-se um vetor de características x_i e sua classificação observada y_i. Logo, pode-se definir a função de verossimilhança para o modelo de regressão logística por

$$L(\alpha, \beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1-\pi(x_i))^{1-y_i}$$

A log-verossimilhança será dada por

$$\begin{aligned}\ell(\alpha,\beta) &= \log\left(L(\alpha,\beta)\right)\\ &= \sum_{i=1}^{n} -\log\left(1+e^{\alpha+\beta x_i}\right) + \sum_{i=1}^{n} y_i(\alpha+\beta x_i)\end{aligned}$$

Para encontrar as estimativas de máxima verossimilhança, deriva-se a função de log-verossimilhança com relação aos parâmetros e igualando-as a zero (AGRESTI, 2007). Entretanto, não há solução de forma fechada para o sistema de equações. Pode-se, no entanto, resolvê-las numericamente de forma aproximada.

3.5 Métricas de avaliação do modelo

Nesse trabalho foi construído um aprendizado supervisionado utilizando o modelo de classificação logístico sendo essencial realizar uma avaliação adequada de seu desempenho. Esse processo de avaliação permite identificar possíveis problemas, como sobreajuste ou falta de generalização, além de fornecer uma visão clara sobre o comportamento do modelo em um ambiente prático, quando colocado em produção.

Existem diversas métricas para avaliar modelos classificadores, cada uma focada em diferentes aspectos, como a quantidade de erros e acertos, ou as classificações de falsos positivos e falsos negativos. Neste trabalho, foram utilizadas as métricas para a avaliação do modelo: Acurácia, Matriz de Confusão, Sensibilidade, Especificidade e a Área Sob a Curva da Característica de Operação do Receptor (AUC ROC). Essas métricas oferecem uma visão abrangente do desempenho do classificador, considerando tanto a qualidade das predições quanto a taxa de erros, (PROVOST; FAWCETT, 2016).

A acurácia é uma métrica simples e amplamente utilizada para avaliar classificadores, sendo calculada pela razão entre o número de acertos e o total de casos analisados. No entanto, por sua simplicidade, pode mascarar deficiências do modelo e gerar resultados enganosos, não refletindo adequadamente o desempenho em casos de desbalanceamento ou erros específicos. Por isso, é recomendado combiná-la com outras métricas que considerem os tipos de erros, como falsos positivos e falsos negativos, para uma avaliação mais completa e precisa (CASTRO; BRAGA, 2011). Sua formulação é dada por

$$\text{acurácia} = \frac{\text{acerto}}{\text{acerto + erro}}.$$

Na fase de avaliação de um modelo, é importante analisar os erros específicos, e a matriz de confusão facilita essa visualização como pode ser visto na Tabela 1. A matriz de confusão exibe dois tipos de acertos (verdadeiros positivos e negativos) e dois tipos de erros (falsos positivos e negativos). Valores verdadeiros correspondem às classificações corretas, enquanto valores falsos indicam erros de classificação (LUQUE et al., 2019).

Avaliar as taxas de falsos positivos e negativos ajuda a identificar o tipo de erro predominante e como o modelo pode se comportar em situações reais. Para simplificar, utilizam-se as siglas: FN (falso negativo), VP (verdadeiro positivo), VN (verdadeiro negativo) e FP (falso positivo). A matriz de confusão é amplamente utilizada na avaliação de classificadores por oferecer uma visão clara sobre a confusão entre as classes, sendo representada por uma matriz $n \times n$, onde n é o número de classes do problema (CASTRO; BRAGA, 2011).

Tabela 1 – Matriz de confusão relacionando os termos verdadeiro positivo e verdadeiro negativo.

Previstos	**Classe Positiva**	**Classe Negativa**
Positivo	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Negativo	Falso Postivo (FP)	Verdadeiro Negativo (VN)

Fonte: Elaborada pela autora.

Luque et al. (2019) destaca a importância das métricas que podem ser extraídas da matriz de confusão e fornece formas de calcula-lás dadas por

- Sensibilidade:
$$\frac{VP}{VP + FN}$$
- Especificidade:
$$\frac{VN}{VN + FP}$$
- Acurácia ou Taxa de Casos Corretamente Classificados:
$$\frac{VN + VP}{VN + VP + FN + FP}$$
- Precisão ou Taxa de Positivos Previstos:
$$\frac{VP}{VP + FP}$$
- Acurácia balanceada:
$$\frac{\text{Sensibilidade} + \text{Especificidade}}{2}$$

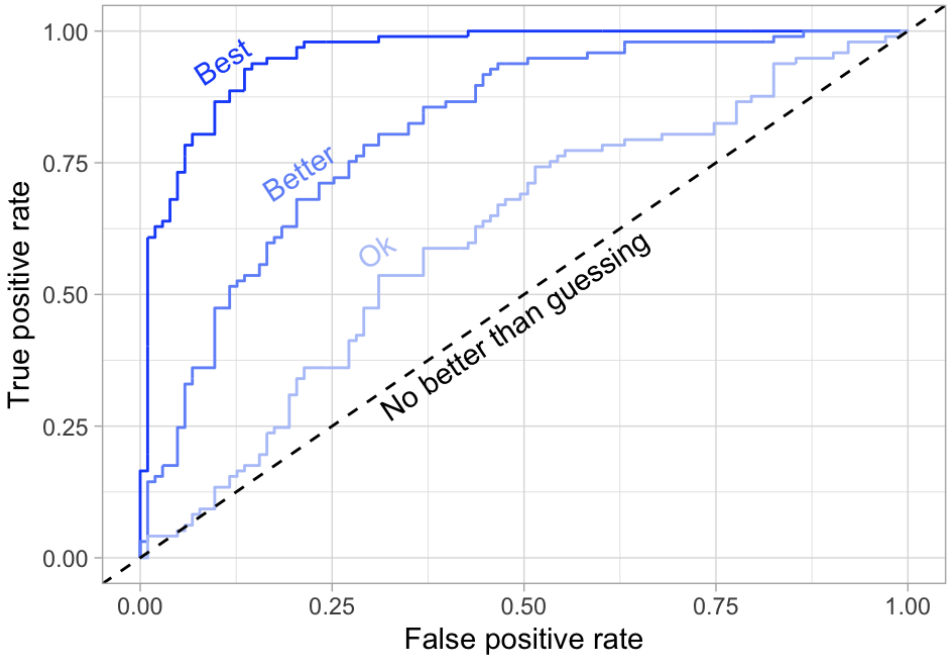
A Sensibilidade mede a proporção de exemplos positivos corretamente classificados como positivos, enquanto a Especificidade avalia a proporção de exemplos negativos corretamente classificados como negativos. A Precisão, por sua vez, refere-se à proporção de casos corretamente classificados como positivos em relação ao total de casos previstos como positivos (incluindo corretos e incorretos). A partir da Precisão, temos a métrica Prevalência Indicada, que representa o número total de casos positivos (verdadeiros e falsos) indicados pelo modelo (CASTRO; BRAGA, 2011). Já a acurácia balanceada é a média aritmética da sensibilidade e especificidade, sendo usada quando os dados são desbalanceados (CRUZ-ROA et al., 2014) .

A área sob a curva (AUC) é uma métrica usada para avaliar a performance de classificadores binários, que devem apresentar alta precisão e sensibilidade. Um bom classificador minimiza falsos positivos e falsos negativos, garantindo previsões corretas tanto para a ocorrência quanto

para a ausência de um evento. Para medir esse equilíbrio, utiliza-se a curva ROC, que representa a taxa de falsos positivos no eixo x e a taxa de verdadeiros positivos no eixo y. Uma linha diagonal do canto inferior esquerdo ao superior direito indica um palpite aleatório, enquanto quanto mais a curva se aproximar do canto superior esquerdo, melhor o desempenho do classificador. O AUC calcula a área sob essa curva, sendo o objetivo maximizar esse valor.

A Figura 3 apresenta a comparação de três curvas ROC em que cada uma reflete o comportamento de um modelo específico. A curva *OK* representa um modelo em que sua área AUC está acima de um modelo ao qual seleciona-se um modelo que classificará as repostas ao acaso (*no better than guessing*). Em seguida, a curva *Better* representa a área AUC de um modelo com melhor desempenho que o modelo *OK*, apresentando um ganho de performance quando comporados os dois modelos. Por fim, a curva *Best* representa o modelo com melhor performance, pois sua área AUC é maior quando comparado aos modelos *Better* e *Ok*.

Figura 3 – Comparação de três modelos considerando a área sob a curva AUC.



Fonte: Boehmke e Greenwell (2019)

3.6 Trabalhos correlatos

Lucena (2013) desenvolveu um estudo cujo objetivo foi propor um modelo para auxiliar a tomada de decisão relacionada à necessidade de encaminhamento dos pacientes acometidos por AVC para os serviços de reabilitação no município de João Pessoa - PB. Para o estudo foi utilizado um questionário contendo itens relacionados ao acesso do paciente ao tratamento de reabilitação, além da dimensão Funções do Corpo da Classificação Internacional de Funcionalidade, Incapacidade e Saúde (CIF). O modelo de regressão logística foi utilizado para subsidiar

a tomada de decisão onde os resultados mostram que a deficiência nas funções relacionadas ao Tônus muscular, ao controle voluntário, emocionais e sexuais maximizam a probabilidade de encaminhar o paciente com AVC para a reabilitação. Além disso, por meio da curva ROC verificou-se que a taxa de acerto do modelo ajustado foi de 73%.

No trabalho de Lima et al. (2016), foi realizado um estudo nas escolas da capital Fortaleza no estado do Ceará no Brasil para analisar os fatores associados ao conhecimento dos adultos jovens sobre o histórico familiar de Acidente Vascular Cerebral (AVC). Foram selecionados 579 adultos jovens de escolas públicas, com coleta de variáveis sociodemográficas, clínicas e de fatores de risco através da aplicação de formulário, sendo analisados utilizando-se regressão logística (*backward elimination*). Os autores observaram uma associação estatística do conhecimento sobre histórico familiar de AVC com situação conjugal com companheiro, circunferência abdominal e pressão arterial normal.

Para medir os impactos de pessoas que sofreram acidente vascular cerebral resultando em alguma deficiência física com foco no desempenho da marcha, Marinho et al. (2018) propuseram um modelo de regressão logística multivariada aplicando o método stepwise com preditores que avaliavam a qualidade de vida comprometida em 124 indivíduos com idade média de 66 anos. Anormalidades de marcha afetam o estilo de vida dos indivíduos, bem como as funcionalidades da vida diária e bem-estar após o acidente vascular cerebral. Nessa análise, os autores observaram que as variáveis TC6M (teste de caminhada de 6 minutos), capacidade funcional e idade estiveram associadas com a qualidade de vida comprometida.

Adicionalmente, Pinheiro et al. (2022) elaboraram um estudo para conhecer fatores que contribuem para a qualidade de vida relacionada a saúde (QVRS) em individuos que sofreram AVC. Neste trabalho, 100 individuos foram acompanhados por um período de um ano e seus dados socieconômicos e clínicos foram coletados. O modelo de regressão logística foi utilizado para avaliar a significância dessas variáveis. Após esse período de acompanhamento, 60% dos indivíduos avaliados apresentaram QVRS favorável e os preditores identificados foram o nível de funcionalidades nas atividades de vida diária, território vascular da lesão e tempo desde o início do AVC.

Pilan (2023) aplicou técnicas de Aprendizado de Máquina e Aprendizagem profunda para auxiliar o diagnóstico do AVC. O autor utilizou os modelos de regressão logística e floresta aleatória como modelos classificadores de fatores de risco e Redes Neurais Convolucionais (CNN) para a classificação de imagens de tomografia computadorizada da região cerebral. O autor observou um desbalanceamento da variável de interesse com 4861 casos de não ocorrência de AVC e apenas 249 casos para ocorrência da doença. Para realizar o balanceamento dessas classes, o autor realizou uma subamostragem da classe não AVC para 275 casos. Em seguida, os modelos foram treinados e a acurácia do modelo de regressão logística foi de 73.33% e do modelo floresta aleatória foi de 73.14%.

Os estudos apresentados sobre o AVC mostram o uso de diferentes abordagens, princi-

palmente utilizando modelos de regressão logística, para prever casos da doença, melhorar o diagnóstico, acompanhamento e reabilitação de pacientes acometidos pela doença. Em conjunto, esses estudos demonstram a importância de modelos estatísticos e técnicas avançadas de aprendizagem de máquinas para a compreensão dos fatores associados ao AVC, melhorando tanto o diagnóstico quanto o tratamento e a reabilitação dos pacientes.

4 METODOLOGIA

O presente trabalho caracteriza-se como uma pesquisa de natureza quantitativa, tendo como objetivo a predição de eventos de interesse em um conjunto de dados desbalanceados, utilizando o modelo de regressão logística. A base de dados utilizada é proveniente da Pesquisa Nacional de Saúde (PNS), com os dados limitados ao estado de Sergipe, garantindo o foco regional necessário para a análise. Neste capítulo serão apresentadas características presentes na base de dados (seção 4.1), definição de classes desbalanceadas (seção 4.2) e suporte computacional utilizado no trabalho (seção 4.3).

4.1 Descrição das variáveis presentes na base de dados

A base de dados utilizada foi extraída do site https://www.pns.icict.fiocruz.br/bases-de-dados/ e corresponde à Pesquisa Nacional de Saúde de 2019. A base original contém 1087 variáveis e 293.726 registros. Após filtrar os dados especificamente para o estado de Sergipe, restaram 8140 respostas. Como a base apresentava muitas respostas em branco, foi necessário realizar uma limpeza nos dados, focando principalmente na nossa variável de interesse, o diagnóstico de AVC. Para isso, foram mantidas apenas as entradas com respostas válidas sobre o diagnóstico de AVC, resultando em 2610 registros. Inicialmente, 19 variáveis foram selecionadas por serem consideradas relevantes para a construção do modelo preditivo. No entanto, ao longo do processo, 7 dessas variáveis foram excluídas por apresentarem respostas em branco, restando 12 variáveis finais. Essas variáveis foram consideradas potencialmente importantes para o modelo, selecionando-se as entradas com informações completas. Assim, ao final dessse processo, a base final utilizada para construção do modelo preditivo continha 2476 registros e 12 variáveis.

No Quadro 1 são apresentadas as seções do questionário da PNS, que é dividido em quatro partes, com perguntas organizadas em módulos. A primeira parte trata da identificação e controle dos entrevistados, a segunda aborda as informações sobre o domicílio, já a terceira parte refere-se às características gerais dos moradores e por fim, a quarta parte é o questionário do morador selecionado, que inclui questões sobre apoio social, estilo de vida, entre outros aspectos (para mais detalhes IBGE (2021)).

A Tabela 2 apresenta os códigos das variáveis selecionadas para a análise, acompanhados de suas respectivas descrições, que correspondem às perguntas do questionário, bem como as categorias, que representam as opções de resposta. Quatro das perguntas pertencem ao Módulo C, que corresponde à Parte 3 do questionário, relacionado às características gerais dos moradores. As outras oito perguntas são da Parte 4, destinada aos moradores selecionados (pessoas com 15 anos ou mais), e fazem parte do Módulo Q, que aborda doenças crônicas.

Quadro 1 – Quadro referente às seções do questionário da PNS.

Parte 1 - Identificação de controle
Parte 2 - Domicílio
Módulo A - Informações do domicílio
Módulo B - Visitas domicíliares de equipe de saúde da família e agentes de endemias
Parte 3 - Questionário do morador
Módulo C - Características gerais dos moradores
Módulo D - Caratcterísticas de educação dos moradores
Módulo E - Características de trabalho das pessoas de 14 anos ou mais
Módulo F - Rendimentos de outras fontes
Módulo G - Pessoas com deficiências
Módulo I - Cobertura de plano de saúde
Módulo J - Utilização de serviços de saúde
Módulo K - Saúde dos indivíduos com 60 anos ou mais
Módulo L - Criança com menos de 2 anos de idade
Parte 4 - Questionário do morador selecionado (pessoas de 15 anos ou mais)
Módulo M - Características do trabalho e apoio social
Módulo N - Percepção do estado de saúde
Módulo O - Acidentes
Módulo P - Estilo de vida
Módulo Q - Doenças crônicas
Módulo R - Saúde da mulher (para mulheres de 15 anos ou mais)
Módulo S - Atendimento pré-natal (para mulheres de 15 anos ou mais)
Módulo U - Saúde bucal
Módulo Z - Paternidade e pré-natal do parceiro (para homens de 15 anos ou mais)
Módulo V - Violência (pessoas de 18 anos ou mais)
Módulo T - Doenças transmissíveis
Módulo Y - Atividade sexual (pessoasde 18 anos ou mais)
Módulo H - Atendimento médico (pessoasde 18 anos ou mais)
Módulo W - Antropometria

Fonte: IBGE (2021)

Tabela 2 – Dicionário das variáveis selecionadas presentes na base de dados.

Código	**Descrição**	**Categorias**
C006	Sexo	1 - Homem, 2- Mulher
C008	Idade do morador	000 a 130
C009	Cor ou Raça	1-Branca, 2-Preta, 3-Amarela, 4-Parda, 5-Indígena
C011	Estado Civil	1-Casado(a), 2-Divorciado, 3-Viúvo(a), 4-Solteiro(a)
Q00201	Algum médico já lhe deu um dignóstico de hipertensão arterial?	1- Sim, 2-Não
Q03001	Algum médico já lhe deu um dignóstico de diabetes?	1-Sim, 2-Não
Q068	Algum médico já lhe deu um dignóstico de AVC?	1-Sim, 2-Não
Q074	Algum médico já lhe deu um dignóstico de asma?	1-Sim, 2-Não
Q092	Algum médico já lhe deu um dignóstico de depressão?	1-Sim, 2-Não
Q11006	Algum médico já lhe deu um dignóstico de transtorno mental?	1-Sim, 2-Não
Q120	Algum médico já lhe deu um dignóstico de câncer?	1-Sim, 2-Não
Q124	Algum médico já lhe deu um dignóstico de insuficiência renal crônica?	1-Sim, 2-Não

Fonte: IBGE (2021)

4.2 Classes desbalanceadas

Ao modelar variáveis repostas binárias, as frequências de ocorrências das classes podem ter um impacto significativo na eficácia do modelo proposto. Classes desbalanceadas ocorrem quando uma classe têm sua proporção muito baixa nos dados de treinamento em comparação com a outra classe. O desequilíbrio de classes pode estar presente em qualquer conjunto de dados e, portanto, deve-se estar ciente das implicações da modelagem desses tipos de dados (KUHN; JOHNSON, 2013).

Uma das abordagens mais diretas para lidar com classes desbalanceadas é ajustar a distribuição dessas classes, tornando o conjunto de dados mais equilibrado (OLIVEIRA et al., 2023). Os métodos utilizados para balancear artificialmente os conjuntos de dados empregados neste trabalho foram o *down-sampling*, *up-sampling* e *smote*.

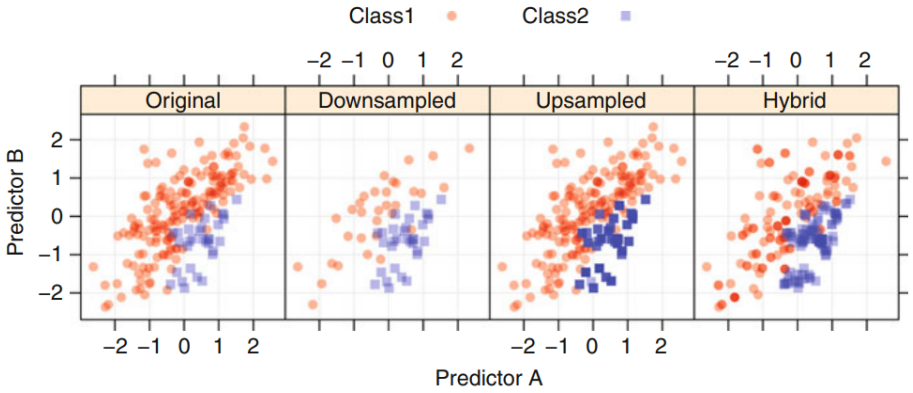
O *down-sampling* seleciona pontos dos dados da classe majoritária para que a classe majoritária tenha aproximadamente o mesmo tamanho que a classe minoritária. Existem várias abordagens para a realização do *down-sampling*. Uma abordagem básica e que foi utilizada nesse trabalho é amostrar aleatoriamente a classe majoritária para que as duas classes tenham aproximadamente o mesmo tamanho. Outra abordagem seria selecionar uma amostra *bootstrap* considerando todos os casos de modo que as classes sejam balanceadas no conjunto *bootstrap* (KUHN; JOHNSON, 2013).

A abordagem *up-sampling* caracteriza-se pela amostragem com substituição dos casos da classe minoritária até que cada classe tenha aproximadamente o mesmo número de elementos. Desta forma, algumas amostras da classe minoritária podem aparecer no conjunto de treinamento com uma frequência bastante alta, enquanto cada amostra na classe majoritária tem uma única realização nos dados (LING; LI, 1998).

Já a técnica *smote* (*synthetic minority over-sampling technique*) proposta por Chawla et al. (2002) é um procedimento de amostragem de dados que usa *up-sampling* e *down-sampling* tendo três parâmetros operacionais: a quantidade de *up-sampling*, a quantidade de *down-sampling* e o número de vizinhos que são usados para imputar novos casos. Para fazer *up-sampling* da classe minoritária, o *smote* sintetiza novos casos. Desta forma, um ponto dos dados é selecionado aleatoriamente da classe minoritária e seus K-vizinhos mais próximos (KNNs) são determinados. O novo ponto dos dados será uma combinação aleatória dos preditores do ponto da base de dados selecionado aleatoriamente e seus vizinhos. Enquanto o algoritmo *smote* adiciona novas amostras à classe minoritária por meio de *up-sampling*, ele também realiza *down-sampling* de casos da classe majoritária por meio de amostragem aleatória para ajudar a equilibrar o conjunto de treinamento. A Figura 4 apresenta uma comparação dos métodos de desbalanceamento.

O exemplo da Figura 4 proposto por Kuhn e Johnson (2013) mostra a comparação dos métodos de balanceamento em um conjunto de dados simulados. Os dados originais (*original*) contêm 168 amostras da primeira classe e 32 da segunda, representando um total de 200 casos. Ao

Figura 4 – Comparação dos métodos de balanceamento dos dados *down-sampling (Downsampled)*, *up-sampling (Upsampled)* e *smote (Hybrid)*.



Fonte: Kuhn e Johnson (2013)

realizar o *down-sampling* (*Downsampled*) os dados foram reduzidos e o tamanho total da amostra ficou com 64 casos divididos uniformemente entre as duas classes. Os dados com *up-sampling* (*Upsampled*) possuem 336 casos com 168 eventos para cada classe da base de dados. Para o método *smote* (*Hybrid*) os dados resultantes foram 128 amostras da primeira classe e 96 da segunda, sendo um total de 224 casos.

4.3 Suporte Computacional

Toda a preparação, limpeza, manipulação e análise dos dados neste trabalho foi realizada utilizando o software livre R Core Team (2024), versão 4.2.2, que fornece uma ampla variedade de técnicas estatísticas (modelagem linear e não linear, testes estatísticos clássicos, análise de séries temporais, classificação, agrupamento, etc) e gráficas, sendo altamente extensível. Para a parte de preparação e limpeza dos dados foi utilizado o pacote *tidyverse* (WICKHAM et al., 2019), uma coleção de pacotes *R* que compartilham uma filosofia de *design* e gramática de alto nível cujas tarefas abrangem: importação de dados, arrumação, manipulação, visualização e programação.

O pacote *caret* (Kuhn, 2008), abreviação de *classification and regression training* e também presente no *software R*, foi utilizado para a construção dos modelos preditivos. O pacote possui ferramentas para divisão dos dados, seleção de variáveis, ajuste dos modelos usando reamostragem, estimativas de importâncias das variáveis, dentre outros. Além do ajuste para modelos de regressão logística, o pacote é dedicado ao treinamento e ajuste de uma ampla variedade de modelos e técnicas de modelagem.

Por fim, também foi utilizado nesse trabalho o *RStudio Cloud* (RStudio Team, 2024) que é um serviço gratuito que permite ao usuário utlizar o *R* e o *RStudio* de forma on-line. O serviço pode ser utilizado para colaboração em projetos com outros usuários através do site https://posit.cloud/. Além disso, uma grande vantagem desse serviço é o armazenamento de

todos os projetos do usuário na nuvem em servidor do próprio *RStudio*, sendo disponibilizados *on-line* e acessados a qualquer momento sem necessidade de uso de computador específico.

5 RESULTADOS E DISCUSSÃO

Antes de iniciar o processo de criação de um modelo preditivo, é necessário preparar adequadamente os dados. Ao analisar a distribuição da variável alvo foi identificado um desbalanceamento entre as classes especificadas, onde dos 2476 entrevistados, apenas 69 relataram ter recebido um diagnóstico de AVC, o que poderia enviesar o classificador em favor dos casos de não AVC, resultando em previsões tendenciosas. Portanto, o balanceamento dos dados é uma etapa essencial para garantir a imparcialidade do modelo. Para corrigir esse desbalanceamento, foram utilizadas as técnicas de *down-sampling*, *up-sampling* e *smote*.

As perguntas selecionadas foram as presentes no Módulos C que abordam características gerais dos moradores, como sexo, idade, estado civil e cor e as presentes no Módulo Q que incluem questões sobre doenças crônicas, como diagnósticos médicos de hipertensão arterial, diabetes, AVC, asma, depressão, transtornos mentais, câncer e insuficiência renal crônica.

5.1 Análise descritiva das variáveis presentes na base de dados.

Na Tabela 3 pode-se observar as variáveis utilizadas para a construção do modelo preditivo, suas descrições e os tipos de variáveis. A variável idade se caracteriza por ser a única variável numérica, Estado civil e Cor/Raça como categóricas e as demais como variáveis binárias.

Tabela 3 – Características das variáveis selecionadas, com suas respectivas descrições e classificações.

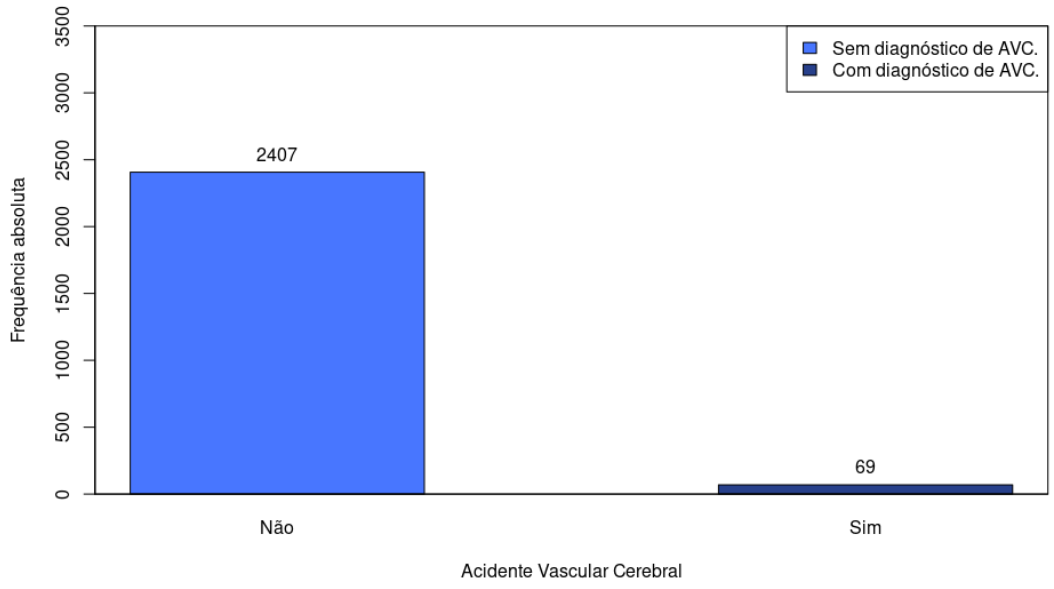
Variáveis	Descrição	Classificação
AVC	Sim/Não	Binária
Hipertensão	Sim/Não	Binária
Diabetes	Sim/Não	Binária
Sexo	Feminino / Masculino	Binária
Estado civil	Casado, Divorciado, Viúvo e Solteiro	Categórica
Cor/Raça	Branca,Preta, Amarela, Parda, Indígena	Categórica
Transtorno mental	Sim/Não	Binária
Insuficienia cardiaca	Sim/Não	Binária
Câncer	Sim/Não	Binária
Depressão	Sim/Não	Binária
Asma	Sim/Não	Binária
Idade	0 a 130 anos	Numérica

Fonte: Elaborada pela autora.

A Figura 5 exibe a proporção de pessoas entrevistadas que receberam ou não o diagnóstico médico de AVC, evidenciando que apenas 2% obtiveram esse diagnóstico, o que gera um desbalanceamento nos dados e pode impactar a construção do modelo preditivo. Segundo Mamed et al. (2019) em 2017, dos 174 óbitos por AVC não especificados em Sergipe, apenas 69% foram investigados, e 50,8% tiveram a causa básica de óbito revisada o que pode demonstrar uma falha

no reconhecimento do tipo específico de AVC antes do óbito. Além disso, conforme destacado por Schmidt et al. (2019), a maioria dos sobreviventes de AVC enfrentam algum tipo de sequela, seja física, comunicacional, funcional, sensorial, mental ou emocional. Para minimizar esses danos, é essencial que essa população tenha acesso a serviços de reabilitação com tratamento integral.

Figura 5 – Frequência dos entrevistados com ou sem diagnóstico de Acidente Vascular Cerebral da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.

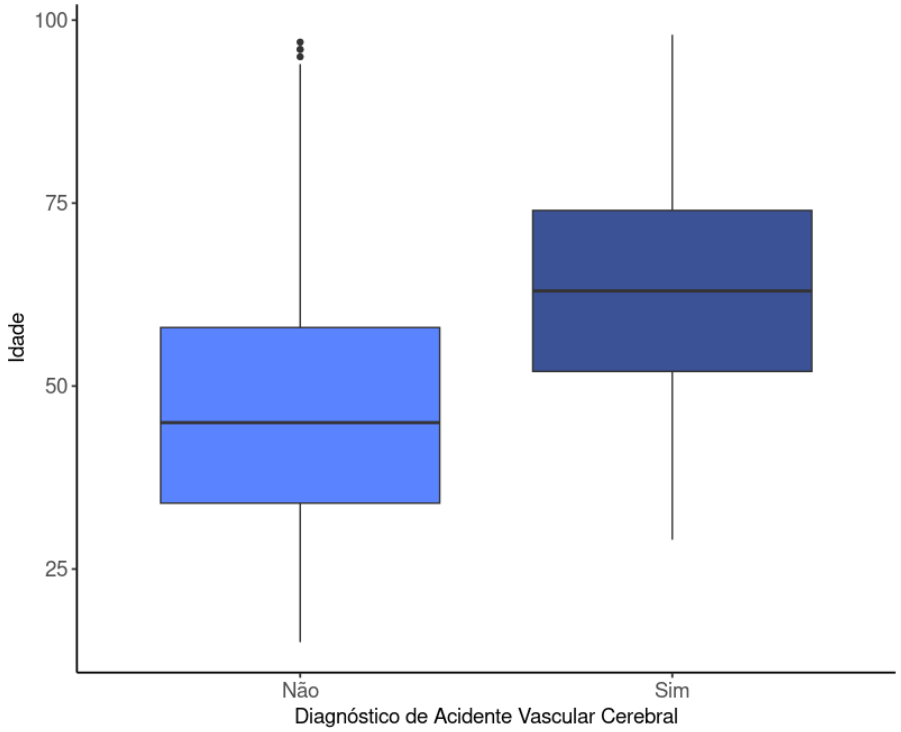


Fonte: Elaborada pela autora.

A Figura 6 apresenta um boxplot que compara a idade dos entrevistados com e sem diagnóstico de AVC. Entre os indivíduos sem diagnóstico de AVC, a mediana da idade está abaixo de 50 anos, com a maioria das idades concentradas entre aproximadamente 35 e 60 anos, e alguns outliers acima de 75 anos. Já entre os entrevistados com diagnóstico de AVC, a mediana de idade é maior do que comparado com os entrevistados sem AVC, em torno de 70 anos, com a maioria das idades situadas entre 60 e 80 anos. Isso indica que indivíduos com diagnóstico de AVC tendem a ser mais velhos em comparação com aqueles sem o diagnóstico, sugerindo uma possível relação entre idade avançada e maior risco de AVC. O boxplot, portanto, aponta a idade como um fator relevante, com maior ocorrência de AVC entre pessoas mais velhas no conjunto de dados analisado.

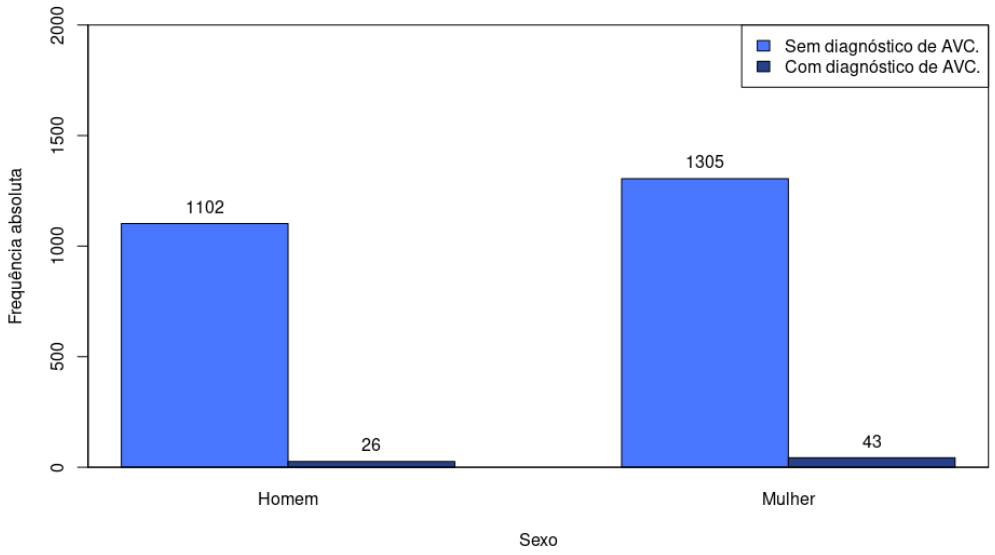
A Figura 7 apresenta a distribuição de entrevistados por sexo, com 1.348 mulheres e 1.128 homens. Entre os diagnosticados com AVC, observa-se uma predominância de casos no sexo feminino. Esse dado está alinhado com o que é descrito por Marinho, Passos e França (2016), indicando que, embora o sexo masculino seja mais acometido pelo AVC antes dos 85 anos, as

Figura 6 – Boxplot da idade dos entrevistados com ou sem diagnóstico de AVC pela Pesquisa Nacional de Saúde para o estado de Sergipe no ano de 2019.



Fonte: Elaborada pela autora.

Figura 7 – Frequência dos entrevistados com ou sem diagnóstico de AVC em relação ao sexo da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.

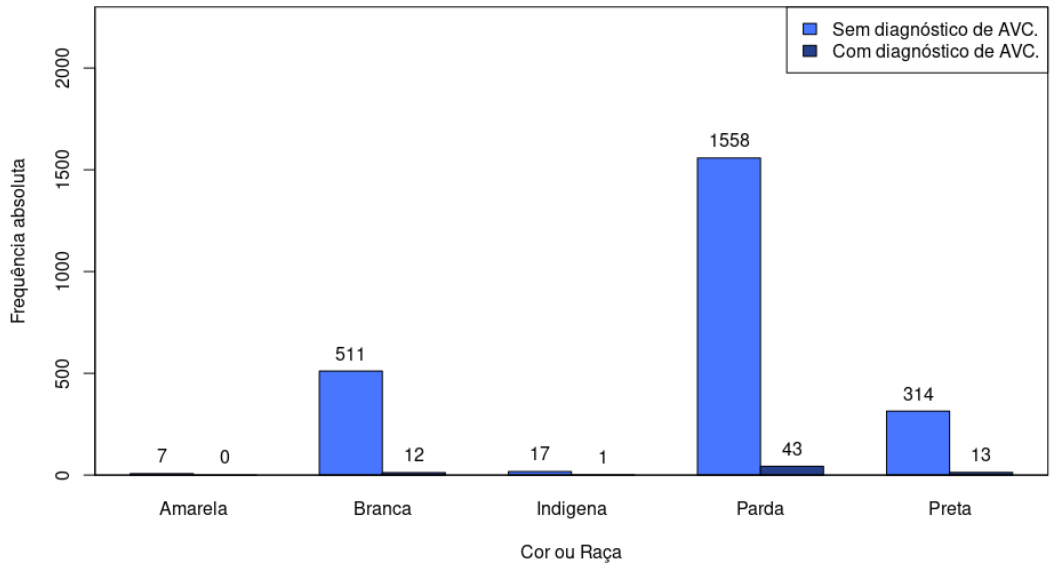


Fonte: Elaborada pela autora

mulheres se tornam mais vulneráveis a partir dessa idade, devido à maior expectativa de vida, e que as chances de AVC dobram após os 55 anos, e como em 2010, a expectativa de vida no Brasil era de 70,5 anos para os homens e 77,7 anos para as mulheres. Isso reforça a necessidade de atenção especial à saúde das mulheres na prevenção de AVC em idades mais avançadas.

A Figura 8 apresenta a frequência absoluta de entrevistados diagnosticados com ou sem AVC, categorizados por cor ou raça. Observa-se que a maior parte dos indivíduos diagnosticados com AVC pertence à cor/raça parda, que também é a maior em número de entrevistados, o que pode estar relacionado ao tamanho dessa população na amostra. Em contraste, as proporções de diagnósticos entre indígenas e amarelos são relativamente pequenas. Embora a Pesquisa Nacional de Saúde (PNS) aponte maior prevalência de casos de AVC entre pessoas de cor branca, estudos revelam que a mortalidade cerebrovascular no Brasil é maior entre pardos e negros, devido à maior taxa de hipertensão e condições econômicas adversas. Essa diferença étnica na mortalidade é mais acentuada entre homens do que entre mulheres (LOTUFO; BENSENOR, 2013).

Figura 8 – Frequência dos entrevistados com ou sem diagnóstico de AVC em relação a cor/raça da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.

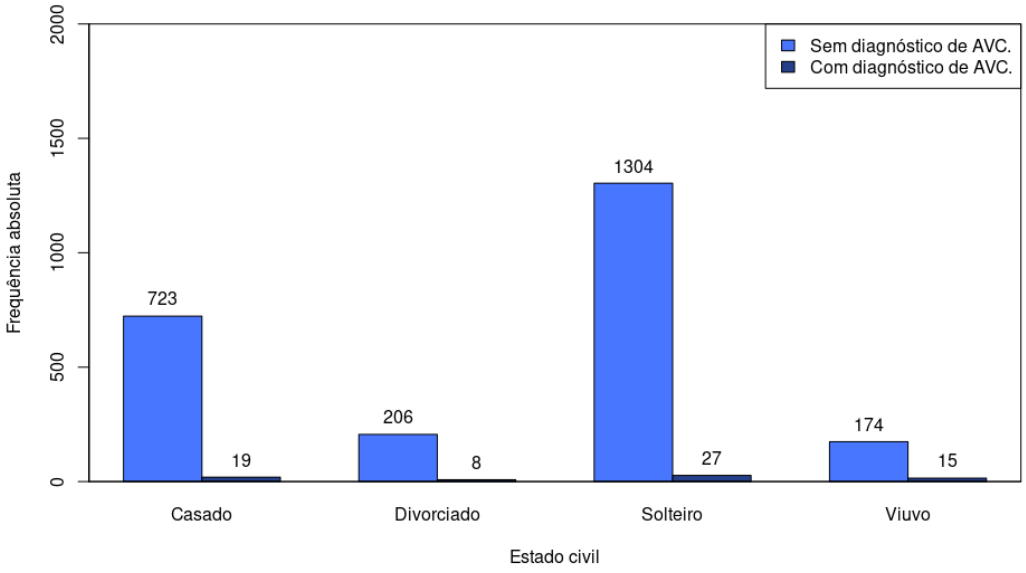


Fonte: Elaborada pela autora

Na Figura 9 é apresentada a distribuição de entrevistados diagnosticados ou não com AVC categorizados de acordo com o estado civil. Observa-se que a maior parte dos indivíduos, é solteira com 1331, seguida pelos casados 742. No entanto, o número de solteiros diagnosticados com AVC é 27 e o de casados com 19. Já as categorias divorciado e viúvo apresenta números menores, tanto de indivíduos quanto de baixo diagnóstico de AVC.

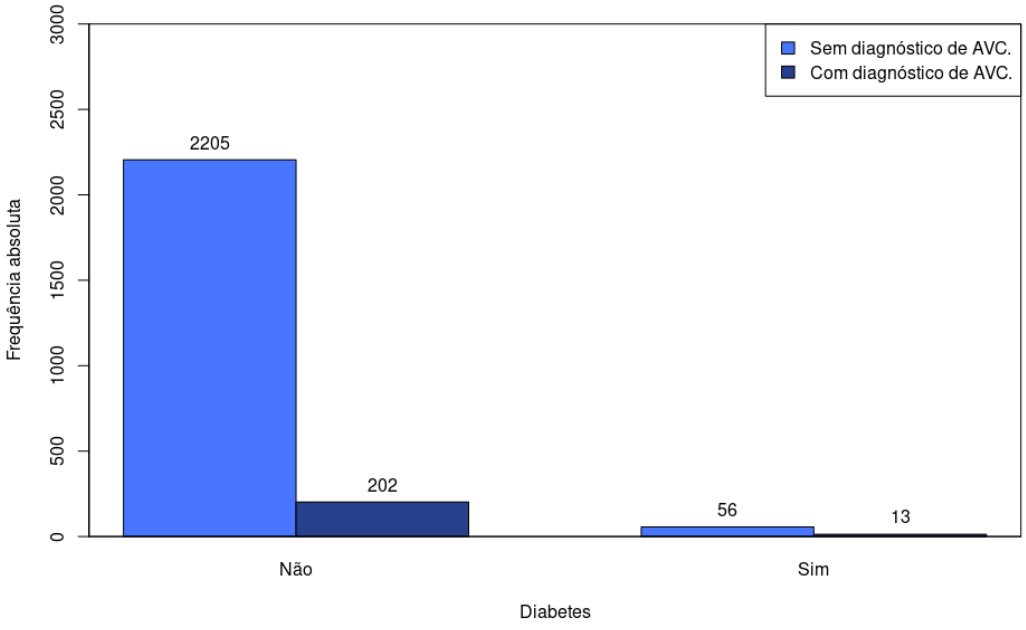
A Figura 10 mostra a distribuição dos entrevistados diagnosticados ou não com AVC,

Figura 9 – Frequência dos entrevistados com ou sem diagnóstico de AVC em relação ao estado civil da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.



Fonte: Elaborado pela autora.

Figura 10 – Frequência dos entrevistados com ou sem diagnóstico de AVC em relação a diabetes da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.

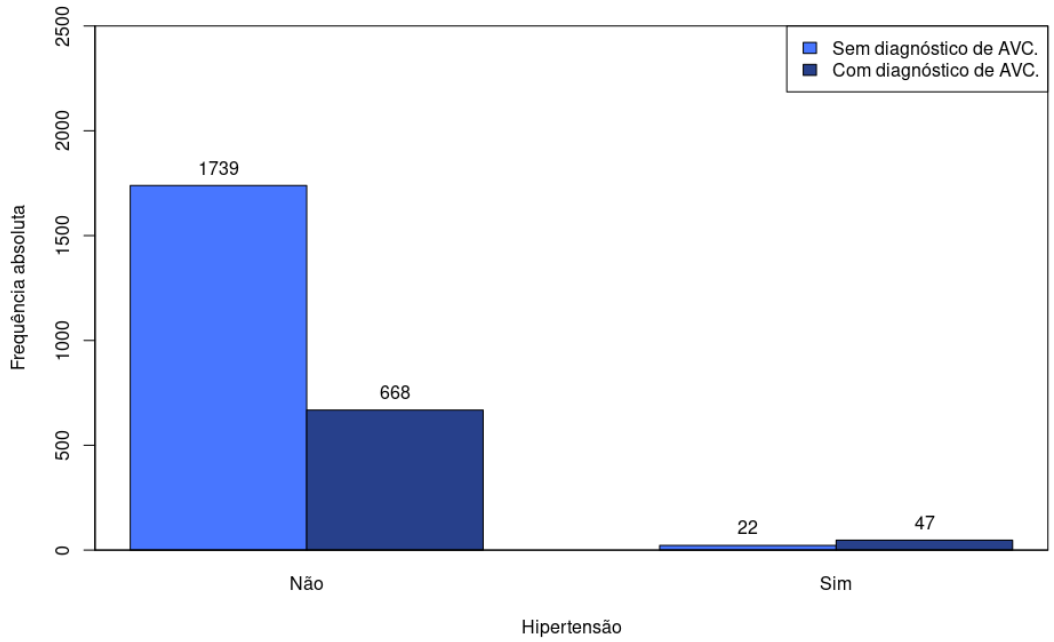


Fonte: Elaborada pela autora.

conforme a presença de diabetes. A maioria dos indivíduos não possui diabetes, com 2205 sem diagnóstico de AVC e 202 diagnosticados. Entre os que têm diabetes, 56 não foram diagnosticados com AVC, enquanto 13 receberam o diagnóstico. Conforme apontado por Wolosker e Faustino (2018), o diabetes é um dos fatores de risco cardiovasculares mais importantes para o desenvolvimento e progressão da aterosclerose carotídea que é uma doença que ocorre quando as artérias carótidas ficam estreitas e/ou obstruídas por placas de gordura que ficaram depositadas no local, e o controle não apenas do diabetes, mas também de outros fatores de risco, é essencial para prevenir o AVC.

A Figura 11 mostra a distribuição dos entrevistados com diagnóstico de AVC em relação à presença de hipertensão, revelando que entre os indivíduos sem hipertensão, 668 foram diagnosticados com AVC, enquanto 1739 não possuem o diagnóstico. Por outro lado, entre aqueles que têm hipertensão, 47 possuem o diagnóstico de AVC, e 22 não foram diagnosticados. Isso pode sugerir que a hipertensão é um fator de risco significativo, já que uma proporção considerável de pessoas hipertensas foi diagnosticada com AVC, embora o número total de hipertensos seja menor em comparação aos indivíduos sem hipertensão.

Figura 11 – Frequência dos entrevistados com ou sem diagnóstico de AVC em relação a hipertensão da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.

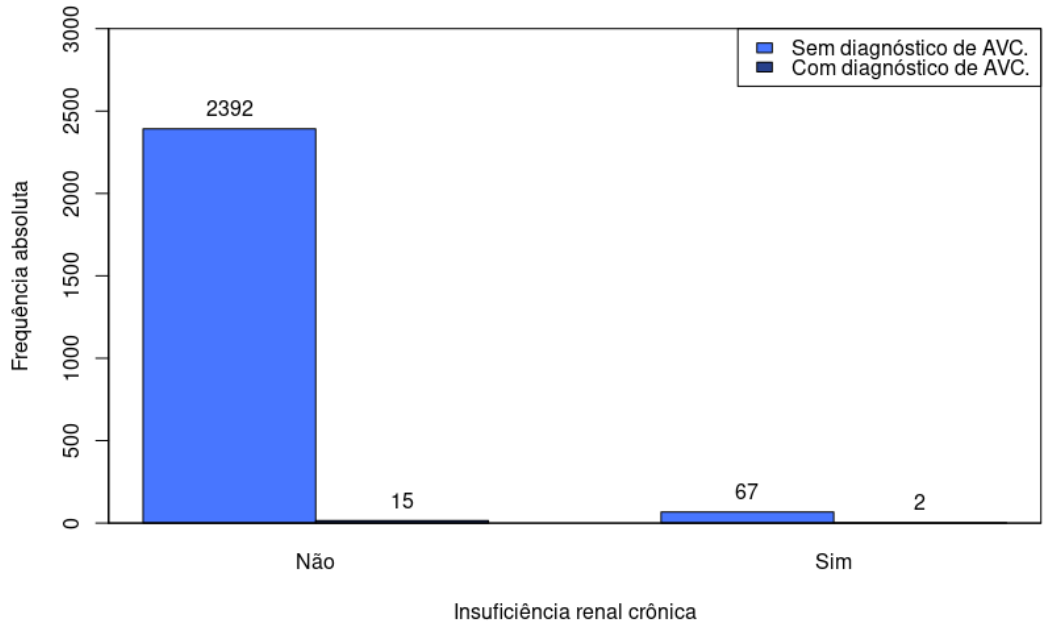


Fonte: Elaborada pela autora

A Figura 12 apresenta a distribuição dos entrevistados com diagnóstico de AVC em relação à presença de doença renal. Entre os indivíduos que não possuem doença renal, 15

foram diagnosticados com AVC e 2392 não. Já entre aqueles com doença renal, apenas 2 foram diagnosticados com AVC, enquanto 67 não têm o diagnóstico. Esses dados sugerem que, nesta amostra, a prevalência de AVC é relativamente baixa entre os portadores de doença renal, o que pode indicar que a presença de doença renal, por si só, não está fortemente associada ao diagnóstico de AVC nessa população.

Figura 12 – Frequência dos entrevistados com ou sem diagnóstico de AVC em relação a insuficiência cardíaca crônica da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.

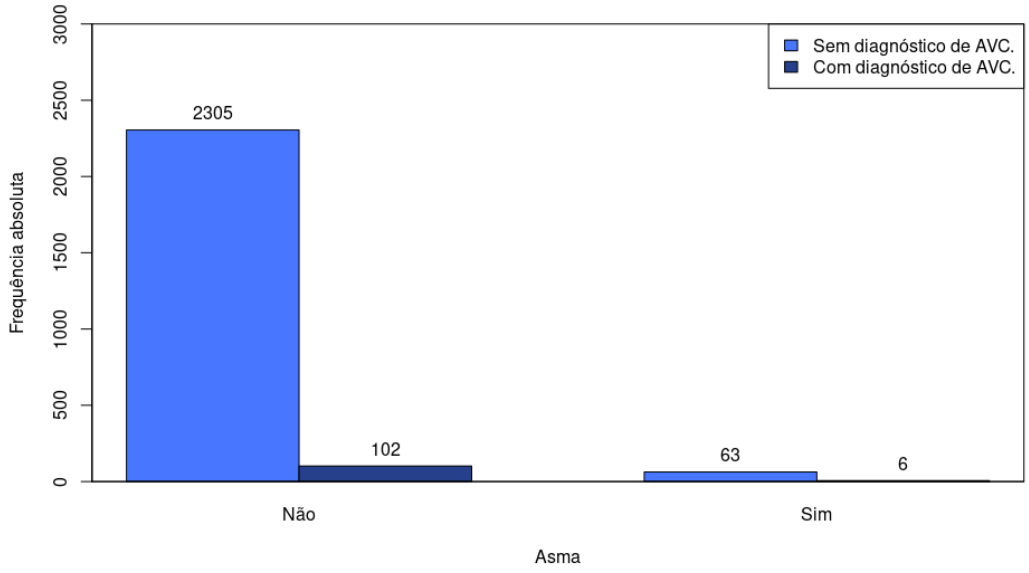


Fonte: Elaborada pela autora.

Na Figura 13 é apresentada a distribuição dos entrevistados diagnosticados ou não com AVC, conforme a presença de asma, é possível observar que entre os indivíduos que não possuem asma, 102 foram diagnosticados com AVC e 2305 não receberam o diagnóstico. Entre os que possuem asma, apenas 6 foram diagnosticados com AVC, enquanto 63 não têm o diagnóstico. Esses números indicam que a prevalência de AVC é baixa tanto entre os portadores quanto entre os não portadores de asma, o que pode indicar que não há uma associação clara entre a presença de asma e o diagnóstico de AVC nesta amostra.

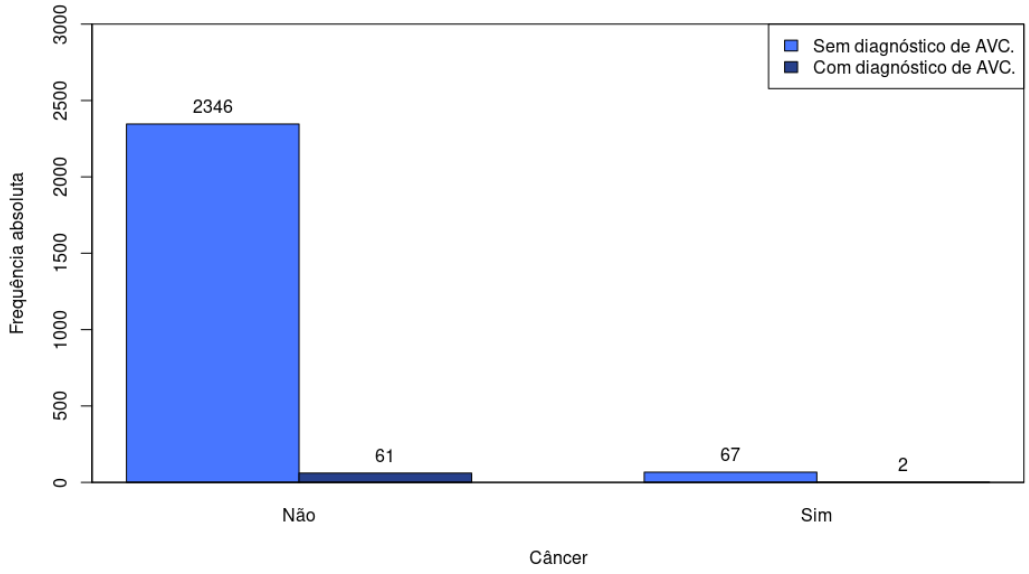
É apresentada na Figura 14 a distribuição dos entrevistados diagnosticados ou não com AVC conforme a presença de câncer. Observa-se que entre os indivíduos que não possuem câncer, 61 foram diagnosticados com AVC e 2346 não receberam tal diagnóstico. Já entre aqueles que possuem câncer, apenas 2 entrevistados foram diagnosticados com AVC, enquanto 67 não possuem o diagnóstico. Estas informações podem sugerir, então, que nesta amostra a presença de

Figura 13 – Frequência dos entrevistados com ou sem diagnóstico de AVC em relação a asma da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.



Fonte: Elaborada pela autora

Figura 14 – Frequência dos entrevistados com ou sem diagnóstico de AVC em relação ao câncer da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.



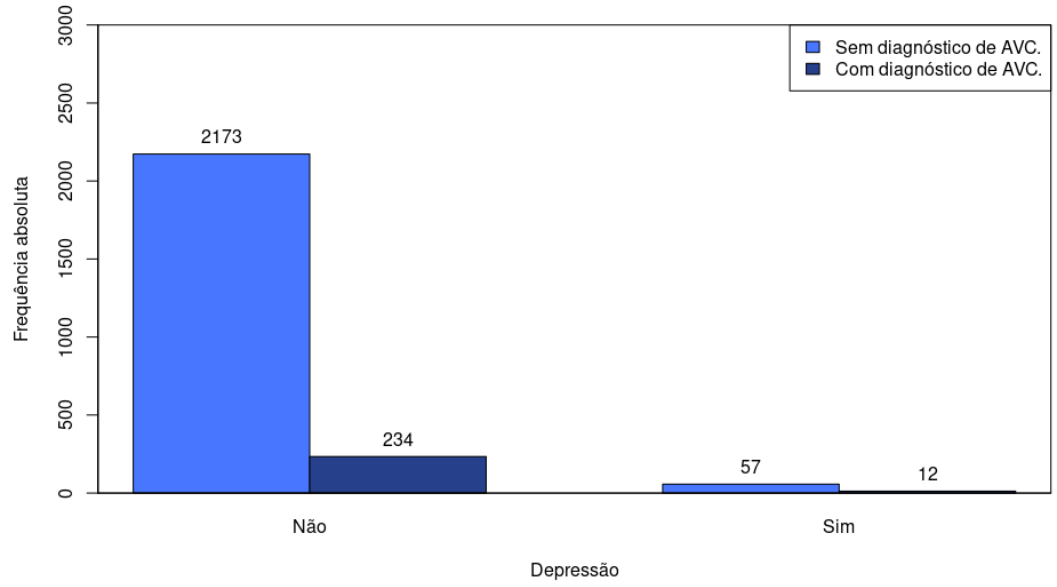
Fonte: Elaborada pela autora

câncer não parece estar fortemente associada ao diagnóstico de AVC, visto que o número de casos de AVC entre os indivíduos com câncer é bastante reduzido.

As Figuras 15 e 16 mostram a relação entre o diagnóstico de AVC e a presença de depressão e transtornos mentais dos entrevistados. Na Figura 15, observa-se que a maioria das pessoas sem AVC não apresenta depressão sendo representada por 2173 entrevistados, enquanto 234 com diagnóstico de AVC também não têm depressão. Entre os que têm depressão, 57 não foram diagnosticados com AVC e 12 foram. Já na Figura 16, 2234 dos indivíduos sem AVC também não apresenta transtornos mentais, enquanto 173 apresenta. Entre os diagnosticados com AVC, 67 não apresentaram transtornos mentais e apenas 2 alegaram ter.

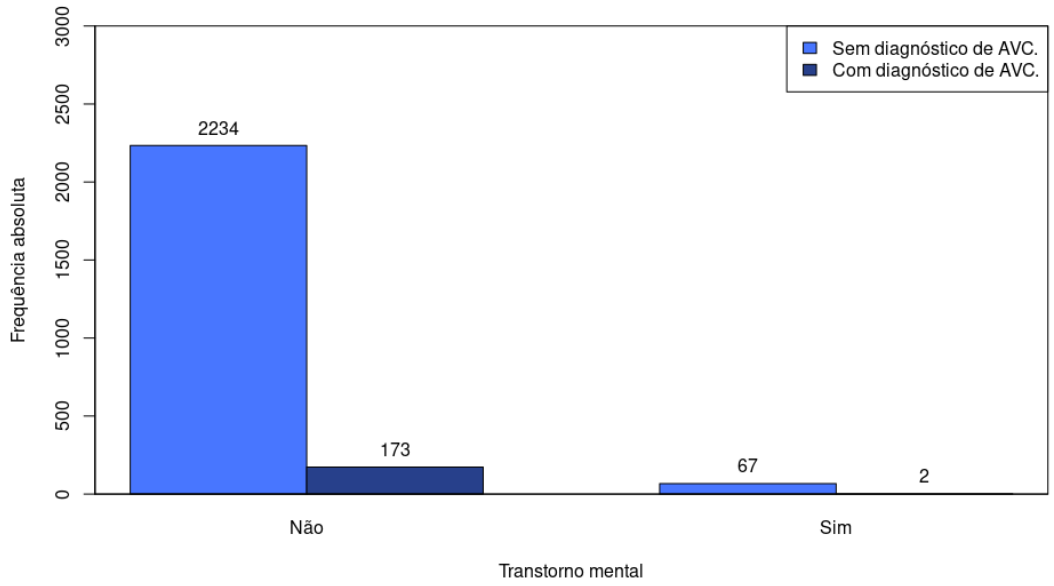
Esses resultados podem sugerir que a maioria dos indivíduos com depressão ou transtornos mentais não possuem diagnóstico de AVC. Porém, segundo o trabalho de Terroni et al. (2009) a depressão torna-se muito comum em pacientes que já possuem o AVC a uma média de 5 anos após o diagnóstico, o que dificulta a adesão ao tratamento, compromete a percepção geral de saúde, reduz os níveis de energia, e diminui tanto a interação social quanto a motivação, destacando-se como uma variável importante na baixa qualidade de vida.

Figura 15 – Frequência dos entrevistados com o diagnóstico de AVC em relação a depressão da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.



Fonte: Elaborada pela autora

Figura 16 – Frequência dos entrevistados com o diagnóstico de AVC em relação a transtorno mental da Pesquisa Nascional de Saúde para o estado de Sergipe no ano de 2019.



Fonte: Elaborada pela autora

5.2 Comparação dos modelos com e sem balanceamento

Para aplicação dos métodos de balanceamentos e escolha do modelo preditivo foi necessário separar a base de dados em treino e teste. Neste trabalho a divisão realizada foi 70% para a base de treino e 30% para a base de teste.

Na Tabela 4 apresenta as medidas descritivas da acurácia dos modelos de predição de AVC, comparando as abordagens de balanceamento, durante o processo de validação cruzada. Observa-se que o modelo sem balanceamento obteve a maior média de acurácia, com 0,9717, possivelmente porque está acertando mais os casos negativos (sem AVC), embora o objetivo seja melhorar a predição dos casos positivos (com AVC). Por outro lado, o método *Down-Sampling* apresentou a menor média, 0,6912, sugerindo que o balanceamento pode ter impactado negativamente a performance em termos de acurácia. Os métodos *Up-Sampling* e *Smote* também tiveram um impacto menor na acurácia quando comparados ao modelo sem balanceamento, mas um comportamento considerável ao se comparar com o *Down-Sampling* .

Para um modelo de regressão logística bem ajustado, espera-se que a maior parte das contagens estejam nas células verdadeiros positivos e verdadeiros negativos, e poucas nas células falsos positivos e falsos negativos. Assim, maiores valores de sensibilidade e especificidade refletem um melhor ajuste do modelo. A Tabela 5 apresenta as métricas de desempenho das três técnicas de balanceamento analisadas, comparadas ao modelo sem balanceamento, utilizando

Tabela 4 – Medidas descritivas da acurácia dos modelos com e sem balanceamento durante a validação cruzada.

Modelos	**Mínimo**	**1st Qu.**	**Mediana**	**Média**	**3rd Qu.**	**Máximo**	**NA's**
Sem Balanceamento	0,9625360	0,9710983	0,9711816	0,9699972	0,9711816	0,9739884	0
Down-Sampling	0,6531792	0,6589595	0,6974063	0,6912462	0,7002882	0,7463977	0
Up-Sampling	0,7080925	0,7262248	0,7485549	0,7414736	0,7608069	0,7636888	0
Smote	0,7225434	0,7262248	0,7630058	0,7547467	0,7752161	0,7867435	0

Fonte: Elaborada pela autora.

o conjunto de dados de treino. Observa-se que sem a aplicação de balanceamento, o modelo alcança a maior acurácia de 0,9717, porém uma sensibilidade de 0,0000 e uma especificidade de 1,0000, indicando que o modelo não consegue identificar nenhum dos casos positivos, no caso o diagnósticos de AVC, revelando que modelo falha completamente em reconhecer os casos de interesse.

Ao aplicar as técnicas de balanceamento, o *Down-Sampling* apresentou uma acurácia de 0,6751, e a especificidade de 0,6716 com sensibilidade de 0,7959 que indica uma melhor capacidade do modelo de identificar os casos de interesse, pois demontra um bom equilíbrio entre a indentificação de casos positivos e negativos. O *Up-Sampling* alcançou uma acurácia de 0,738, uma sensibilidade mais alta de 0,8367, e também uma boa especificidade, indicando que o modelo está capturando bem os casos de AVC. Já o *Smote*, apresenta uma acurácia de 0,7611, com sensibilidade de 0,7755 e especificidade de 0,7606, oferecendo um bom equilíbrio entre a capacidade de detectar casos de AVC e a precisão geral.

Tabela 5 – Medidas de desempenho para os modelos com e sem balanceamento utilizando os dados de treino.

Medidas	**Sem Balanceamento**	***Down-Sampling***	***Up-Sampling***	***Smote***
Acurácia	0,9717	0,6751	0,738	0,7611
95% CI	(0,9628; 0,979)	(0,6525; 0,6972)	(0,7166; 0,7586)	(0,7403; 0,781)
No Information Rate	0,9717	0,9717	0,9717	0,9717
P-Value [Acc >NIR]	0,5379	1	1	1
Sensibilidade	0,00000	0,79592	0,83673	0,77551
Especificidade	1,00000	0,67162	0,73515	0,76069
Acurácia Balanceada	0,50000	0,73377	0,78594	0,76810

Fonte: Elaborada pela autora.

Na Tabela 6 é apresentada as métricas de desempenho das técnicas de balanceamento em comparação ao modelo sem balanceamento, utilizando o conjunto de dados do teste. É possível observar que o modelo sem balanceamento apresenta uma excelente acurácia e especificidade, mas uma baixíssima sensibilidade, conluindo que o modelo está enviesado para os casos negativos e sem utilidade na detecção de AVC. Nas técnicas de balanceamento o *Down-Sampling* apresentou uma acurácia de 0,716, uma sensibilidade de 0,6000 e especificidade de 0,7192, sugerindo que, ao melhorar a identificação dos positivos, ele perde um pouco da capacidade de distinguir corretamente os casos negativos. Já o *Up-Sampling* alcança uma acurácia de 0,7577, sensibilidade

de 0,5500, indicando que o modelo está capturando mais da metade dos casos, e especificidade de 0,7634, com isso ele acaba perdendo um pouco a detecção dos casos positivos. A técnica *Smote*, apresenta a melhor acurácia entre os métodos de balanceamento, com 0,7672, mas com a menor sensibilidade de 0,5000, e uma maior especificidade, mostrando que o modelo é eficiente na identificação dos casos negativos.

Tabela 6 – Medidas de desempenho para os modelos com e sem balanceamento utilizando os dados de teste.

Medidas	**Sem Balanceamento**	***Down-Sampling***	***Up-Sampling***	***Smote***
Acurácia	0,9731	0,716	0,7577	0,7672
95% CI	(0,9587; 0,9835)	(0,6821; 0,7482)	(0,7253; 0,7881)	(0,7351; 0,7971)
No Information Rate	0,9731	0,9731	0,9731	0,9731
P-Value [Acc >NIR]	0,5591	1	1	1
Sensibilidade	0,00000	0,60000	0,55000	0,50000
Especificidade	1,00000	0,71923	0,76349	0,77455
Acurácia Balanceada	0,50000	0,65961	0,65674	0,63728

Fonte: Elaborada pela autora.

A Tabela 7 apresenta a matriz de confusão que compara os valores previstos e reais para o modelo sem balanceamento e para os modelos com técnicas de balanceamento utilizando a base de teste. No modelo sem balanceamento, não houve previsão correta de nenhum caso da classe com AVC, e 20 casos foram previstos incorretamente, evidenciando a dificuldade do modelo em prever os casos da classe minoritária classificando-os como sem AVC. Já com a técnica de *Down-Sampling*, 12 casos foram corretamente previstos com AVC, porém 203 casos de sem AVC foram incorretamente classificados com AVC. Isso indica uma melhora na previsão da classe com AVC, mas com uma redução significativa na precisão dos casos sem AVC gerando muitos falsos positivos.

Tabela 7 – Matriz de confusão comparando os valores previstos e os valores reais para os modelos com e sem balançeamento utilizando a base de teste.

Sem Balanceamento			***Down-Sampling***		
	Referência			Referência	
Predição	Sim	Não	Predição	Sim	Não
Sim	0	0	Sim	12	203
Não	20	723	Não	8	520
Up-Sampling			***Smote***		
	Referência			Referência	
Predição	Sim	Não	Predição	Sim	Não
Sim	11	171	Sim	10	163
Não	9	552	Não	10	560

Fonte: Elaborada pela autora.

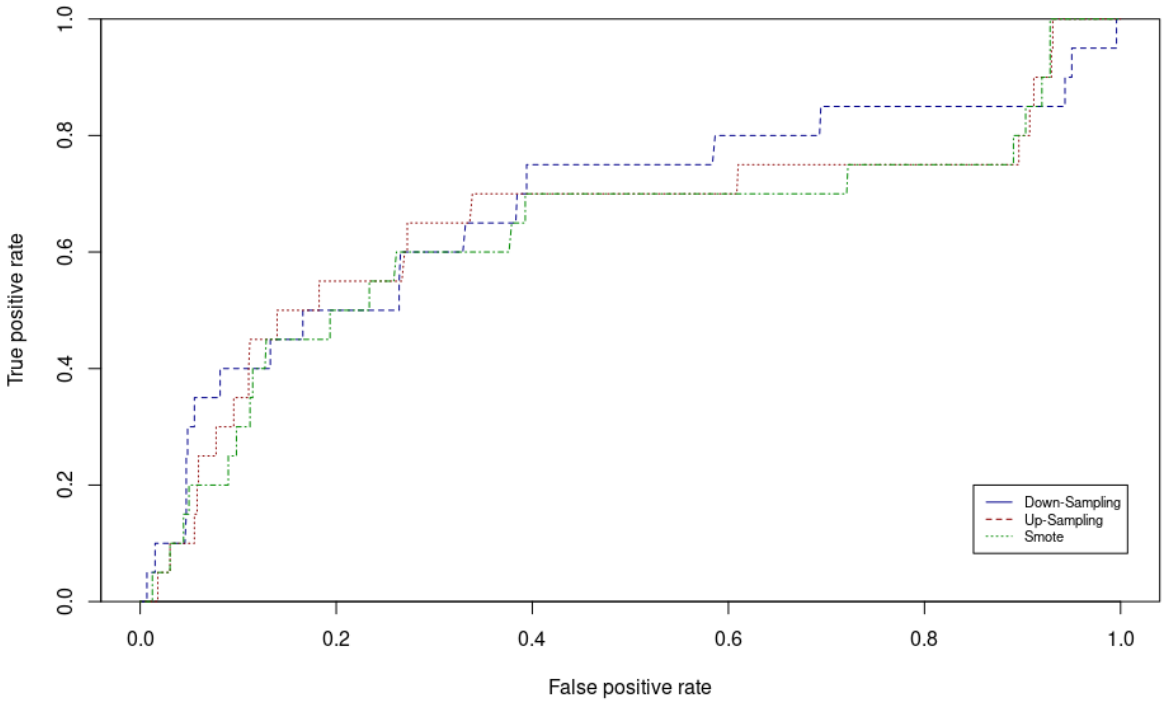
No *Up-Sampling* 11 casos com AVC foram previstos corretamente, mas 171 erros ocorreram ao classificar sem AVC como com AVC. Embora melhore a previsão da classe

minoritária com AVC, também resulta em um número expressivo de falsos positivos. Utilizando o *Smote*, 10 casos de com AVC foram corretamente previstos, enquanto 163 casos de sem AVC foram erroneamente classificados com AVC. Isso mostra uma maior quantidade de falsos positivos, embora a previsão para a classe com AVC tenha tido bons resultados.

Em resumo, as técnicas de balanceamento aumentam a capacidade do modelo de prever a classe com AVC, que são os casos minoritários, mas também elevam o número de falsos positivos, ou seja, classificações incorretas de sem AVC como com AVC. O modelo sem balanceamento apresentou baixa performance na previsão dos casos com AVC ressaltando a importância de utilizar técnicas de balanceamento ao lidar com classes desbalanceadas.

Na Figura 17 as curvas ROC para as técnicas *Down-Sampling*, *Up-Sampling* e *Smote* se comportam de formas muito parecidas. Entretanto, em alguns pontos a curva para técnica *Down-Sampling* se destaca com relação as outras técnicas indicando um ganho de performance preditiva.

Figura 17 – Curva ROC dos modelos logísticos considerando as técnicas de balanceamente e utilizando a base de teste.



Fonte: Elaborada pela autora.

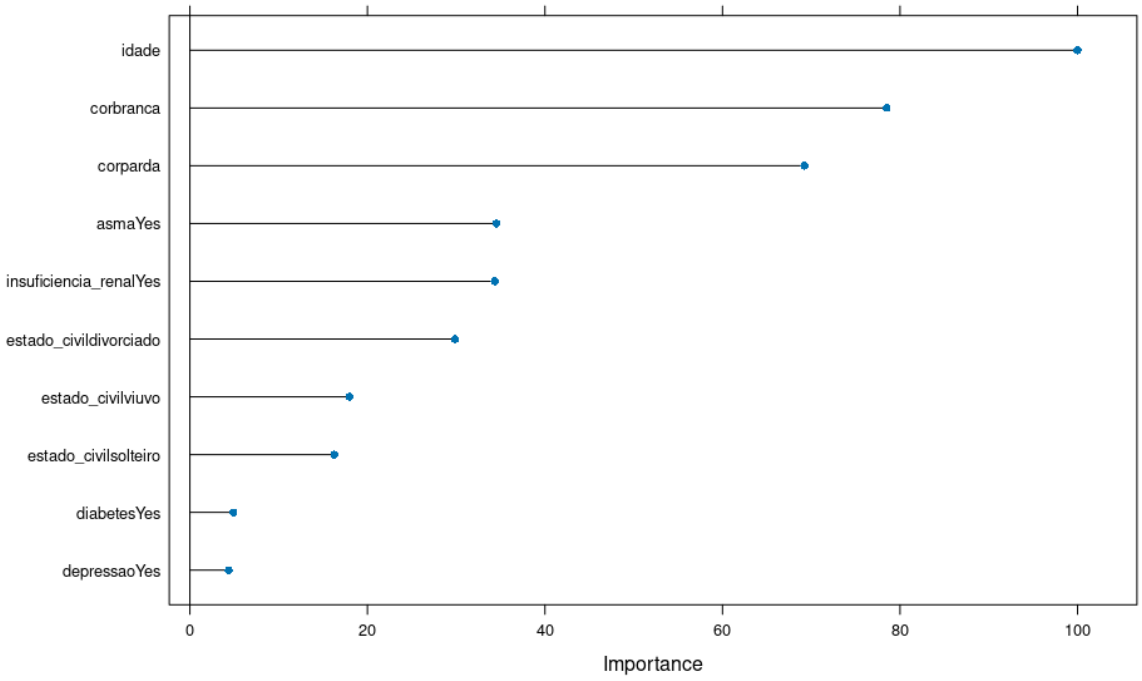
A comparação entre as técnicas de balanceamento sugere que, embora elas melhorem a capacidade preditiva em comparação à ausência de balanceamento, ainda não alcançam uma performance otimizada. Todas as técnicas resultam em curvas acima da diagonal, o que reflete uma

predição razoável, mas nenhuma delas se destaca significativamente, pois as curvas permanecem bastante próximas entre si.

Segundo Kuhn e Johnson (2013) os resultados da comparação das técnicas de balanceamento podem ser misturados. Embora as melhorias na acurácia possam ser modestas, as abordagens de amostragem têm o benefício de permitir melhores compensações entre sensibilidade e especificidade. Considerando as medidas de desempenho dos modelos com balanceamento, todos apresentaram resultados similares. No entanto, a técnica de *Down-Sampling* destacou-se com o melhor desempenho devido à sua maior sensibilidade, alcançando 0,6000 em comparação com as demais, além de uma acurácia balanceada de 0,6596.

Na Figura 18 é apresentada as 10 variáveis mais importantes para o modelo, destacando suas respectivas contribuições para a capacidade preditiva. A variável com maior importância é "idade", seguida por "cor branca" e "cor parda", indicando que essas características possuem um papel significativo na previsão dos resultados do modelo. Variáveis como "asma", "insuficiência renal", e diferentes estados civis também se destacam, mas com menor contribuição em comparação com as três principais. As variáveis "diabetes" e "depressão", embora ainda importantes, possuem menor impacto na predição geral do modelo.

Figura 18 – Contribuição das 10 variáveis mais importantes para o modelo considerando o modelo logístico e a técnica de balanceamento *Down-Sampling*.



Fonte: Elaborada pela autora.

6 CONCLUSÃO

Este trabalho teve como principal intuito desenvolver um modelo preditivo baseado em regressão logística para prever a ocorrência de acidente vascular cerebral (AVC) utilizando dados da Pesquisa Nacional de Saúde. A análise das características do conjunto de dados revelou que, entre os 2476 entrevistados, apenas 69 haviam recebido diagnóstico de AVC, evidenciando um desbalanceamento da base em relação à variável de interesse. Para lidar com esse desbalanceamento, foram aplicadas três técnicas de balanceamento, as quais são *Down-Sampling*, *Up-Sampling* e *Smote*, visando equilibrar as classes e melhorar a capacidade preditiva do modelo.

Para identificar o modelo mais adequado para a previsão de risco de AVC, foi utilizada a base de treino, considerando as métricas de acurácia, sensibilidade, especifidade e acurácia balanceada. Entre as técnicas testadas, o *Smote* apresentou o melhor desempenho considerando sua acurácia de 0,7611 mas apresentou a menor sensibilidade na base de treinamento. Nessa etapa, fica evidenciado que o modelo de regressão logística sem considerar balanceamento apresenta a maior acuracia, entretanto apresentam uma sensibilidade igual a zero, característica de um modelo que não consegue identificar a classe minoritária.

Em seguida, as previsões foram realizadas considerando a base de teste e utilizando a matriz de confusão e a curva ROC para avaliar a precisão do modelo. Observou-se que as medidas de desempenho dos modelos com balanceamento foram similares. No entanto, a técnica de Down-Sampling foi escolhida devido o melhor desempenho para sua sensibilidade de 0,6 e acurácia balanceada de 0,6596. Com isso, foi possível desenvolver um modelo com potencial para auxiliar na identificação de indivíduos com maior risco de AVC, contribuindo para ações preventivas e melhorias na saúde pública.

6.1 Trabalhos futuros

Para trabalhos futuros, recomenda-se:

- explorar outros modelos de aprendizado de máquina, como *random forest* e *knn*;
- aplicar novas técnicas de balanceamento da variável de interesse;
- testar mais modelos considerando a inclusão de variáveis preditoras adicionais;
- utilizar modelos de classificação que já considerem o desbalanceamento dos dados, sem necessiade de utilização de técnicas de balanceamento.

REFERÊNCIAS

AGRESTI, A. *An introduction to categorical data analysis*. Wiley-Blackwell, 2007. Disponível em: <http://scholar.google.de/scholar.bib?q=info:zgZR_0-o5cUJ:scholar.google.com/&output=citation&hl=de&as_sdt=0,5&ct=citation&cd=0>. Citado 2 vezes nas páginas 22 e 23.

ANDRADE, P. N. de et al. Análise epidemiológica do acidente vascular cerebral nos estados do nordeste (2020-2022). *Brazilian Journal of Health Review*, v. 7, n. 3, p. e69525–e69525, 2024. Citado na página 16.

BELYADI, H.; HAGHIGHAT, A. *Machine Learning Guide for Oil and Gas Using Python*. [S.l.]: Gulf Professional Publishing, 2021. 169-295 p. Citado na página 22.

BOEHMKE, B.; GREENWELL, B. *Hands-On Machine Learning with R*. [S.l.]: Chapman and Hall/CRC., 2019. Citado na página 25.

CASTRO, C. L. d.; BRAGA, A. P. Supervised learning with imbalanced data sets: an overview. *Sba: Controle & Automação Sociedade Brasileira de Automatica*, SciELO Brasil, v. 22, p. 441–466, 2011. Citado 2 vezes nas páginas 23 e 24.

CH, R.; MAP, M. Bayesian learning. *book: Machine Learning. McGraw-Hill Science/Engineering/Math*, p. 154–200, 1997. Citado na página 19.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, v. 16, p. 321–357, 2002. Citado na página 30.

CRUZ-ROA, A. et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In: SPIE. *Medical Imaging 2014: Digital Pathology*. [S.l.], 2014. v. 9041, p. 904103. Citado na página 24.

EFRON, B.; TIBSHIRANI, R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical science*, JSTOR, p. 54–75, 1986. Citado na página 20.

FÁVERO, L. P. L. et al. Análise de dados: modelagem multivariada para tomada de decisões. 2009. Citado 2 vezes nas páginas 20 e 21.

FIGUEIREDO, A. R. G. d.; PEREIRA, A.; MATEUS, S. Acidente vascular cerebral isquémico vs hemorrágico: taxa de sobrevivência. *Higeia: Revista Científica da Escola Superior de Saúde Dr. Lopes Dias*, IPCB. ESALD, 2020. Citado na página 16.

GREENWELL, B. M. *Tree-based methods for statistical learning in R*. [S.l.]: Chapman and Hall/CRC, 2022. Citado na página 20.

GUS, I.; FISCHMANN, A.; MEDINA, C. Prevalência dos fatores de risco da doença arterial coronariana no estado do rio grande do sul. *Arq. bras. cardiol*, p. 478–490, 2002. Citado 2 vezes nas páginas 17 e 18.

IBGE. *Pesquisa nacional de saúde: 2019: ciclos de vida: Brasil*. [S.l.]: IBGE, 2021. 139p p. Citado 4 vezes nas páginas 13, 18, 28 e 29.

KIM, J.-H. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, Elsevier, v. 53, n. 11, p. 3735–3745, 2009. Citado na página 20.

Kuhn, M. Building predictive models in r using the caret package. *Journal of Statistical Software*, v. 28, n. 5, p. 1–26, 2008. Citado na página 31.

KUHN, M.; JOHNSON, K. *Applied Predictive Modeling*. [S.l.]: Springer, 2013. ISBN 978-1-4614-6848-6. Citado 3 vezes nas páginas 30, 31 e 46.

LIMA, M. J. M. R. et al. Fatores associados ao conhecimento dos adultos jovens sobre histórico familiar de acidente vascular cerebral. *Revista Latino-Americana de Enfermagem*, 2016. Citado na página 26.

LING, C. X.; LI, C. Data mining for direct marketing: Problems and solutions. *Proceedings of International Conference on Knowledge Discovery from Data (KDD 98)*, p. 73–79, 1998. Citado na página 30.

LONGO, D. L. et al. Medicina interna de harrison. In: *Medicina interna de Harrison*. [S.l.: s.n.], 2013. p. 1796–1796. Citado na página 16.

LOTUFO, P. A.; BENSENOR, I. J. M. Raça e mortalidade cerebrovascular no brasil. *Revista de Saúde Pública*, SciELO Public Health, v. 47, p. 1201–1204, 2013. Citado na página 36.

LOTUFO, P. A. et al. Cerebrovascular disease in brazil from 1990 to 2015: Global burden of disease 2015. *Revista Brasileira de Epidemiologia*, SciELO Brasil, v. 20, p. 129–141, 2017. Citado na página 16.

LUCENA, E. M. d. F. *Modelo de regressão logística para auxiliar a tomada de decisão quanto à necessidade de reabilitação em pacientes com Acidente Vascular Encefálico.* Dissertação (Dissertação de Mestrado) — Universidade Federal da Paraíba, 2013. Citado na página 25.

LUDERMIR, T. B. Inteligência artificial e aprendizado de máquina: estado atual e tendências. *Estudos Avançados*, SciELO Brasil, v. 35, p. 85–94, 2021. Citado na página 19.

LUQUE, A. et al. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, Elsevier, v. 91, p. 216–231, 2019. Citado 2 vezes nas páginas 23 e 24.

MACHADO, A. S. et al. Acidente vascular cerebral hemorrágico associado à acidente ofídico por serpente do gênero bothrops: relato de caso. *Revista da Sociedade Brasileira de Medicina Tropical*, SciELO Brasil, v. 43, p. 602–604, 2010. Citado na página 17.

MAMED, S. N. et al. Perfil dos óbitos por acidente vascular cerebral não especificado após investigação de códigos garbage em 60 cidades do brasil, 2017. *Revista Brasileira de Epidemiologia*, SciELO Brasil, v. 22, n. Suppl 3, p. e190013–supl, 2019. Citado 2 vezes nas páginas 17 e 33.

MARINHO, C. et al. Desempenho da marcha e qualidade de vida nos sobreviventes de avc: um estudo transversal. *Revista Pesq Fisio.*, v. 8, p. 79–87, 2018. Citado na página 26.

MARINHO, F.; PASSOS, V. M. d. A.; FRANÇA, E. B. Novo século, novos desafios: mudança no perfil da carga de doença no brasil de 1990 a 2010. *Epidemiologia e Serviços de Saúde*, SciELO Brasil, v. 25, p. 713–724, 2016. Citado na página 34.

MEMBERS, W. G. et al. Heart disease and stroke statistics—2012 update. *Circulation*, v. 125, n. 1, p. e2–e220, 2012. Disponível em: <https://www.ahajournals.org/doi/abs/10.1161/CIR.0b013e31823ac046>. Citado na página 17.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. *Sistemas inteligentes-Fundamentos e aplicações*, v. 1, n. 1, p. 32, 2003. Citado na página 19.

OLIVEIRA, C. G. et al. Abordagens para tratamento de dados desbalanceados utilizadas no aprendizado de máquina na área da saúde: Um estudo de caso na predição de acidente vascular cerebral (avc). Galoá, 2023. Citado na página 30.

OLIVEIRA, F. Afasia semântica como uma única manifestação de avc agudo. *ArqNeuropsiquiatr*, v. 68, n. 1, p. 119–30, 2010. Citado na página 13.

OLIVEIRA, R. d. M. C. de; ANDRADE, L. A. F. de. Acidente vascular cerebral. *Rev Bras Hipertens*, v. 8, p. 280–290, 2001. Citado na página 16.

PILAN, V. d. P. *Técnicas de Inteligência Artificial para diagnóstico de Acidente Vascular Cerebral através de imagens e dados textuais sobre possíveis vítimas*. Monografia (Trabalho de conclusão de curso) — Universidade Estadual Paulista - UNESP, 2023. Citado na página 26.

PINHEIRO, L. O. R. et al. Preditores de qualidade de vida relacionada à saúde em indivíduos após acidente vascular cerebral (avc) residentes na comunidade: estudo longitudinal prospectivo. *Revista Pesquisa em Fisioterapia*, v. 12, p. e4911, 2022. Disponível em: <https://journals.bahiana.edu.br/index.php/fisioterapia/article/view/4911>. Citado na página 26.

POMPERMAIER, C. et al. Fatores de risco para o acidente vascular cerebral (avc). *Anuário Pesquisa e Extensão Unoesc Xanxerê*, v. 5, p. e24365–e24365, 2020. Citado na página 16.

PROVOST, F.; FAWCETT, T. *Data Science para negócios*. [S.l.]: Alta Books, 2016. Citado na página 23.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2024. Disponível em: <https://www.R-project.org/>. Citado na página 31.

ROLIM, C. L. R. C.; MARTINS, M. Qualidade do cuidado ao acidente vascular cerebral isquêmico no sus. *Cadernos de Saúde Pública*, SciELO Public Health, v. 27, p. 2106–2116, 2011. Citado 2 vezes nas páginas 13 e 17.

RStudio Team. *RStudio: Integrated Development Environment for R*. Boston, MA, 2024. Disponível em: <http://www.rstudio.com/>. Citado na página 31.

SCHMIDT, M. H. et al. Acidente vascular cerebral e diferentes limitações: uma análise interdisciplinar. *Arquivos de Ciências da Saúde da UNIPAR*, v. 23, n. 2, 2019. Citado na página 34.

STOPA, S. R. et al. Pesquisa nacional de saúde 2019: histórico, métodos e perspectivas. *Epidemiologia e Serviços de Saúde*, SciELO Brasil, v. 29, p. e2020315, 2020. Citado na página 18.

SZWARCWALD, C. L. et al. Pesquisa nacional de saúde no brasil: concepção e metodologia de aplicação. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 19, p. 333–342, 2014. Citado na página 18.

TERRONI, L. d. M. N. et al. Depressão pós-avc: aspectos psicológicos, neuropsicológicos, eixo hha, correlato neuroanatômico e tratamento. *Archives of Clinical Psychiatry (São Paulo)*, SciELO Brasil, v. 36, p. 100–108, 2009. Citado na página 41.

TRAN, V. et al. Investigation of ann architecture for predicting residual strength of clay soil. *Neural Comput & Applic*, Springer, v. 34, p. 19253–19268, 2022. Citado na página 21.

WICKHAM, H. et al. Welcome to the tidyverse. *Journal of Open Source Software*, v. 4, n. 43, p. 1686, 2019. Citado na página 31.

WOLOSKER, N.; FAUSTINO, C. B. Abordagem do paciente com diabetes mellitus e doença ateromatosa em outros territórios: membros inferiores. *Rev. Soc. Cardiol. Estado de Säo Paulo*, p. 187–192, 2018. Citado na página 38.