Universidade de Brasília
Faculdade de Educação Física
Programa de Pós-Graduação em Educação Física

Machine Learning em Saúde Pública: Variáveis-Chave Associadas ao Acidente Vascular Cerebral na População Brasileira.

Ian Caetano Quadrado

Brasília
2025

Machine Learning em Saúde Pública: Variáveis-Chave Associadas ao Acidente Vascular Cerebral na População Brasileira.

Linha de pesquisa: “Aspectos Biológicos Relacionados ao Desempenho e à Saúde”

Tema: “Biomecânica e Análise de Sinais Fisiológicos”

Tese apresentada ao Programa de Pós graduação *Stricto Sensu* em Educação Física – FEF/UnB como requisito para obtenção do título de Doutor em Educação Física.

Orientador:

Prof. Dr. Jake do Carmo

Coorientador:

Prof. Dr. F. Assis Nascimento

Para minha esposa, Anna, e minha filha, Isabel.

Agradecimentos

Meu doutorado e minha filha chegaram praticamente ao mesmo tempo. Lembro claramente do dia em que descobrimos a gravidez e disse à minha esposa:

— Amor, se achar que vai ficar pesado levar os dois juntos, adio o doutorado para quando tivermos mais tempo.

Mesmo sabendo da energia que seria necessária para conciliar esses dois grandes projetos, além das demandas do dia a dia, decidimos encarar o desafio. Sem cobranças e com leveza, as coisas foram acontecendo. Eu chegava tarde do trabalho e ia estudar; nos fins de semana, quando precisava me dedicar à pesquisa, ela cuidava e ia passear com a Bebel. Foi uma caminhada de muita parceria, apoio e companheirismo.

Quero agradecer profundamente ao meu orientador, Jake. Já havia tido aula com ele na graduação, mas nosso primeiro contato mais próximo foi na banca de seleção para o mestrado. Antes de iniciar a arguição, ele disse:

— Não tenho prazer em colocar alguém contra a parede para fazer questionamentos, mas, já que essa é a nossa tarefa, vamos em frente.

Naquele momento, eu já estava bastante nervoso, mas ouvir isso, vindo de alguém que não via o processo como uma disputa nem um momento de exposição indesejada, trouxe calma e clareza. Anos depois, escolhi tê-lo como orientador no doutorado, certo de que encontraria nele a rara combinação de excelência técnica com sensibilidade humana.

Agradeço também ao meu coorientador, pelos insights valiosos e pelas discussões técnicas que ajudaram a moldar este trabalho.

À minha família, que, perto ou longe, sempre me apoiou.

À minha filha, que sempre imaginei assim: quando comecei este doutorado, pensei que terminaria com ela aos quatro anos, me assistindo na defesa. E assim está sendo.

Estou muito feliz por toda essa jornada que, ao contrário de ser um ponto final, para mim representa uma porta que se abre para uma nova fase de aprendizado e conhecimento.

Sou grato também aos amigos, da UnB e de fora dela, que de formas diferentes ajudaram a tornar essa caminhada mais leve — compartilhando momentos de estudo,

palavras de incentivo ou simplesmente estando presentes quando precisei de companhia e apoio.

Aos professores da banca, agradeço a generosidade em dedicar tempo e atenção à leitura deste trabalho, pelas críticas construtivas e pelas contribuições que o tornaram mais consistente e completo.

E agradeço a Deus, pela força nos dias difíceis, pela serenidade nas incertezas e por abrir caminhos onde eu não via possibilidades, renovando minha fé de que cada passo tinha um propósito maior.

Resumo

O Acidente Vascular Cerebral (AVC) permanece como uma das principais causas de mortalidade e incapacidade no Brasil, com forte heterogeneidade territorial e determinantes que extrapolam o clínico-biológico. Métodos tradicionais capturam parcialmente essa complexidade; por outro lado, abordagens de ciência de dados — em especial o aprendizado de máquina — permitem integrar múltiplas dimensões e revelar padrões latentes úteis à vigilância e à gestão.

Nesse contexto, o objetivo desta tese foi verificar a viabilidade de hierarquizar e predizer a presença autorreferida de AVC a partir de marcadores sociodemográficos, clínicos, funcionais, comportamentais e de uso de serviços, utilizando técnicas de aprendizado de máquina aplicadas aos microdados da PNS-2019.

A base analítica compreendeu 293.727 respondentes e 1.114 variáveis/categorias após limpeza, recodificação e padronização; o desfecho foi definido pela pergunta de diagnóstico médico prévio de AVC. As preditoras cobriram blocos de características individuais e domiciliares (idade, renda per capita, composição e densidade domiciliar), condições e comportamentos de saúde (HAS/DM autorreferidos, tabagismo, álcool, alimentação), funcionalidade e reabilitação (fisioterapia, limitações nas atividades, uso de dispositivos), uso de serviços e prevenção (consultas, exames de colesterol/glicemia, medicamentos contínuos) e temas contemporâneos (sexualidade/reprodução, violência/cuidado informal, digitalização do cotidiano).

O treinamento foi conduzido por Unidade da Federação para capturar especificidades locais, tendo a Random Forest como modelo principal para estimar importância de variáveis; A avaliação interna baseou-se no erro out-of-bag (OOB) e na estabilidade do ranking em rodagens repetidas. Dois níveis analíticos foram considerados: top-10 variáveis por estado e top-30 para leitura contextual ampliada.

Os resultados evidenciaram um “núcleo nacional” comum, demonstrando a estabilidade do processo — idade, renda domiciliar per capita e perguntas relacionadas diretamente com AVC (faz dieta por conta do AVC? Faz fisioterapia por conta do AVC?) — e, simultaneamente, peculiaridades locais que refletem contextos regionais: padrões alimentares e acesso (Norte), climatério/vida sexual e planejamento reprodutivo (Sul/Sudeste), organização do cuidado e reabilitação

(Sudeste/Sul), violência e cuidado informal (Centro-Oeste/Nordeste) e determinantes associados à digitalização do cotidiano (tempo de telas, telessaúde). Esses achados sugerem vias diretas (cardiometabólicas e de manejo pós-evento) e indiretas (acesso e continuidade do cuidado, tempo e trabalho, estrutura domiciliar) na explicação do desfecho, reforçando a natureza multifatorial e socialmente mediada do AVC. Do ponto de vista translacional, a abordagem mostrou capacidade de priorizar preditores com valor programático; iluminar desigualdades territoriais com granularidade estadual; e aproximar vigilância, clínica e gestão, oferecendo insumos para prevenção e linhas de cuidado pós-AVC (APS forte com controle de HAS/DM, reabilitação oportuna, suporte ao cuidador e estratégias sobre determinantes sociais).

Como limitações, destacam-se o desenho transversal e o desfecho autorreferido, que limitam inferências causais e podem incorporar viés de informação; além disso, parte das variáveis de alta importância reflete condições pós-evento, sendo particularmente úteis à vigilância e ao planejamento, mais do que à prognosticação individual.

Conclusão: é viável e útil empregar aprendizado de máquina sobre a PNS-2019 para mapear fatores associados ao AVC com granularidade estadual, produzindo evidências acionáveis para o SUS e abrindo agenda para validação externa (bases clínicas/administrativas), análises longitudinais e monitoramento temporal em novas edições de inquéritos.

Palavras-chave: AVC; Pesquisa Nacional de Saúde; aprendizado de máquina; Random Forest; desigualdades regionais; importância de variáveis; vigilância em saúde; linhas de cuidado.

Abstract

Stroke (Cerebrovascular Accident, CVA) remains one of the leading causes of mortality and disability in Brazil, marked by strong territorial heterogeneity and determinants that go beyond clinical-biological factors. Traditional methods capture this complexity only partially; on the other hand, data science approaches—especially machine learning—allow for the integration of multiple dimensions and the unveiling of latent patterns that are useful for surveillance and management.

In this context, the objective of this thesis was to assess the feasibility of ranking and predicting self-reported stroke based on sociodemographic, clinical, functional, behavioral, and healthcare utilization markers, using machine learning techniques applied to microdata from the 2019 National Health Survey (PNS-2019).

The analytical base comprised 293,727 respondents and 1,114 variables/categories after cleaning, recoding, and standardization; the outcome was defined by the question on prior medical diagnosis of stroke. Predictors spanned blocks of individual and household characteristics (age, per capita income, household composition and density), health conditions and behaviors (self-reported hypertension/diabetes, smoking, alcohol, diet), functionality and rehabilitation (physical therapy, activity limitations, device use), healthcare utilization and prevention (consultations, cholesterol/glucose testing, continuous medication), and contemporary issues (sexuality/reproduction, violence/informal care, digitalization of daily life).

Training was conducted by Federative Unit to capture local specificities, with Random Forest as the main model to estimate variable importance. Internal evaluation was based on out-of-bag (OOB) error and ranking stability across repeated runs. Two analytical levels were considered: top-10 variables per state and top-30 for broader contextual interpretation.

The results revealed a common “national core,” demonstrating process stability—age, per capita household income, and stroke-specific questions (e.g., “Do you follow a diet because of stroke?” “Do you attend physical therapy because of stroke?”)—while also highlighting local peculiarities reflecting regional contexts: dietary patterns and access (North), menopause/sexual life and reproductive planning (South/Southeast), organization of care and rehabilitation (Southeast/South), violence and informal care (Midwest/Northeast), and determinants related to the digitalization of daily life (screen time, telehealth). These findings suggest both direct pathways

(cardiometabolic and post-event management) and indirect ones (access and continuity of care, time and work demands, household structure) in explaining the outcome, reinforcing the multifactorial and socially mediated nature of stroke. From a translational perspective, the approach demonstrated the ability to prioritize predictors with programmatic value; shed light on territorial inequalities with state-level granularity; and bridge surveillance, clinical practice, and management, providing insights for prevention and post-stroke care pathways (strong PHC with hypertension/diabetes control, timely rehabilitation, caregiver support, and strategies addressing social determinants).

As limitations, the cross-sectional design and self-reported outcome restrict causal inferences and may introduce information bias; moreover, some of the high-importance variables reflect post-event conditions, being particularly useful for surveillance and planning rather than for individual prognostication.

Conclusion: It is feasible and useful to apply machine learning to PNS-2019 data to map stroke-associated factors with state-level granularity, generating actionable evidence for the Brazilian Unified Health System (SUS) and opening an agenda for external validation (clinical/administrative datasets), longitudinal analyses, and temporal monitoring in future survey editions.

Keywords: Stroke; National Health Survey; machine learning; Random Forest; regional inequalities; variable importance; health surveillance; care pathways.

Lista de Tabelas

Lista de Figuras

Lista de Abreviaturas e Siglas

AC – Acre
AHA – American Heart Association
AIT – Ataque Isquêmico Transitório
AL - Alagoas
AM - Amazonas
AP – Amapá
APS – Atenção Primária à Saúde
AUC-PR - Área sob a Curva de Precisão-Revocação (Precision-Recall), útil em contextos com classes desbalanceadas
AUC-ROC - Área sob a Curva ROC (Receiver Operating Characteristic), avalia a capacidade discriminatória do modelo
AVC - Acidente Vascular Cerebral
AVD – Atividades da Vida Diária
AVE – Acidente Vascular Encefálico
BA – Bahia
BALANCEDACC - Acurácia Balanceada, média entre sensibilidade e especificidade
BRIER - Brier Score, avalia a calibração probabilística das predições
CE – Ceará
DF – Distrito Federal
DRC – Doença Renal Crônica
ES – Espírito Santo
ESF – Estratégia Saúde da Família
F1-Score - Média harmônica entre Precisão e Recall, mede o equilíbrio entre os dois indicadores
GO - Goiás
IA - Inteligência Artificial
IMC – Índice de Massa Corporal
IST – Infecções Sexualmente Transmissíveis
JACCARD - Índice de Jaccard, mede a similaridade entre conjuntos de variáveis selecionadas em diferentes execuções
K - Número fixo de variáveis de maior importância hierarquizadas em cada execução do modelo

MA - Maranhão
MCC - Coeficiente de Correlação de Matthews, mede a qualidade da classificação considerando todas as categorias da matriz de confusão
MG – Minas Gerais
ML - Machine Learning
MS – Mato Grosso do Sul
MT – Mato Grosso
OMS – Organização Mundial da Saúde
OOB – Out-of-Bag
P.A. – Pressão Arterial
PA – Pará
PB - Paraíba
PE – Pernambuco
PI – Piauí
PNS – Pesquisa Nacional de Saúde
PR - Paraná
PRECISION - Precisão, proporção de verdadeiros positivos entre todos os classificados como positivos
RECALL - Recall (ou Sensibilidade), proporção de verdadeiros positivos corretamente identificados entre todos os positivos reais
RJ – Rio de Janeiro
RN – Rio Grande do Norte
RO – Rondônia
RR – Roraima
RS – Rio Grande do Sul
SC – Santa Catarina
SE – Sergipe
SISAB – Sistema de Informação em Saúde para a Atenção Básica
SP – São Paulo
TO - Tocantins
Top-10 – 10 variáveis mais importantes na correlação com AVC para tal estado
Top-30 – 30 variáveis mais importantes na correlação com AVC para tal estado

Sumário

1.Introdução

O Acidente Vascular Cerebral (AVC) - atualmente chamado de AVE (Acidente Vascular Encefálico) - se configura como um fenômeno patológico de elevada prevalência e gravidade, ocupando uma posição de destaque no panorama das questões de saúde pública no mundo todo (Boehme; Esenwa; Elkind, 2017; Feigin; Norrving; Mensah, 2017; Katan; Luft, 2018).

O AVC se destaca pela gravidade com que afeta a saúde e a capacidade funcional dos indivíduos, levando a elevados níveis de incapacidade e mortalidade em variadas localidades e contextos sociodemográficos (Donkor, 2018; Kuriakose; Xiao, 2020; Malaeb *et al*., 2021). Este evento crítico pode ser categorizado com base na alteração do fluxo sanguíneo cerebral: o AVC isquêmico resulta do bloqueio ou estreitamento das artérias, enquanto o AVC hemorrágico é causado por rupturas arteriais (Boehme; Esenwa; Elkind, 2017; Hurford *et al*., 2020). Adicionalmente, existem os Ataques Isquêmicos Transitórios (AITs), que são episódios temporários com sintomas semelhantes aos do AVC. Geralmente não causam sequelas permanentes, mas indicam alto risco de um AVC futuro e exigem avaliação médica imediata. (Boehme; Esenwa; Elkind, 2017; Ortiz-Garcia *et al*., 2022).

No âmbito territorial brasileiro, o AVC emerge como uma das principais causas de óbito e incapacidade, imprimindo um ônus significativo sobre o arcabouço do sistema de saúde e, por extensão, sobre a sociedade (De Santana *et al*., 2018), (Santos *et al*., 2022a). Tal realidade se traduz em desafios multifacetados, englobando desde a demanda por recursos financeiros para o tratamento e reabilitação dos acometidos, até o impacto psicossocial experimentado pelos pacientes e seus núcleos familiares (Oliveira-Kumakura *et al*., 2023a).

A intrincada natureza etiopatogênica do AVC, aliada à sua manifestação clínica abrupta e frequentemente severa, sublinha a imperiosa necessidade de estratégias proativas de rastreamento e intervenção precoce (Fernandes, 2015; Minelli *et al*., 2022a). A otimização dos protocolos de atendimento emergencial e a disseminação de conhecimento acerca dos sinais e sintomas premonitórios do AVC constituem pilares fundamentais para a mitigação dos danos neurológicos e para a potencialização dos desfechos positivos na jornada de recuperação dos pacientes (Maniva *et al*., 2018).

De um ponto de vista global, o AVC responde por aproximadamente 11% do total de óbitos registrados, consolidando-se como a segunda maior causa de morte e a terceira maior causa de anos de vida perdidos por incapacidade, conforme elucida o estudo conduzido por Feigin, Norrving e Mensah (2017). Estas estatísticas reiteram a magnitude e a urgência da problemática, demandando a mobilização de esforços coletivos para a prevenção, o diagnóstico precoce e o tratamento eficaz do AVC (Katan; Luft, 2018; Nogueira *et al*., 2021).

Focando no contexto nacional brasileiro, evidencia-se uma tendência de transição nas taxas de incidência do AVC ao longo das últimas décadas, delineando um quadro epidemiológico em constante mutação (Martins *et al*., 2013; Santos *et al*., 2022a).

Nota-se uma diminuição nos índices de mortalidade associada ao AVC, um fenômeno possivelmente relacionado às melhorias na infraestrutura de saúde e na qualidade do atendimento médico-hospitalar (Silva *et al*., 2022). Contudo, em contrapartida, observa-se um aumento na prevalência de sobreviventes do AVC que convivem com sequelas e limitações funcionais (Minelli *et al*., 2022b).

Esta realidade ressalta a necessidade premente de estratégias preventivas eficazes e de programas de reabilitação robustos, que visem não apenas a sobrevivência, mas também a recuperação da qualidade de vida e funcionalidade dos indivíduos afetados (Gomes *et al*., 2016; Oliveira-Kumakura *et al*., 2023a).

Dessa forma, o AVC se estabelece como um desafio iminente e complexo para a saúde pública no Brasil e mundialmente, exigindo investimentos em pesquisa, políticas públicas e práticas clínicas que se orientem para a prevenção, a detecção precoce e a intervenção terapêutica adequada, com o intuito de minimizar o impacto devastador dessa condição sobre os indivíduos e a coletividade (Feigin; Norrving; Mensah, 2017; Katan; Luft, 2018; Yan *et al*., 2016).

O ônus associado ao AVC no território brasileiro manifesta-se de maneira heterogênea, evidenciando notórias discrepâncias que são influenciadas por uma série de determinantes socioeconômicos e regionais (Minelli *et al*., 2022a; Santos *et al*., 2022a). A distribuição desigual da incidência e dos desfechos pós-AVC no Brasil é um reflexo palpável das disparidades existentes entre diferentes estratos da população, com um ônus acentuadamente maior recaído sobre aqueles em situação de vulnerabilidade socioeconômica (Do Carmo Ferreira; Sarti; De Azevedo Barros, 2022; Kuper *et al*., 2007; Lotufo, 2005; Silva *et al*., 2022).

Em áreas caracterizadas por baixa renda e por um acesso precário a serviços de saúde de qualidade, as populações experimentam uma taxa de incidência de AVC significativamente mais elevada, além de enfrentarem desfechos clínicos substancialmente piores após a ocorrência de um evento vascular cerebral (Tetzlaff *et al*., 2020; Thompson *et al*., 2022). Santos *et al*., (2022) corroboram essa afirmação, enfatizando a vulnerabilidade dessas comunidades e a necessidade urgente de estratégias de intervenção que sejam capazes de mitigar tais desigualdades.

A heterogeneidade na incidência de AVC no Brasil é também marcada por variações geográficas pronunciadas. Estudos como o conduzido pela (Paiva *et al*., 2018) ressaltam a existência de discrepâncias regionais significativas, com as regiões Norte apresentando índices de mortalidade por AVC superiores aos observados nas regiões Sul do país. Essas diferenças geográficas, por sua vez, são reflexo de uma complexa interação de fatores, incluindo aspectos socioeconômicos, a disponibilidade e o acesso a serviços de saúde, bem como peculiaridades culturais e de estilo de vida (Da Silva Paiva *et al*., 2018; Goulart, 2016).

Diante desse cenário, torna-se evidente a necessidade de intervenções direcionadas e de políticas públicas adaptadas, que levem em consideração as especificidades das diversas populações afetadas pelo AVC no Brasil (Meschia *et al*., 2014). A implementação de estratégias de prevenção e tratamento que sejam culturalmente sensíveis e acessíveis é um passo crucial para reduzir as disparidades existentes e para melhorar os desfechos associados ao AVC em todo o território nacional (Fernandes, 2015; Nugem *et al*., 2020).

A etiologia do AVC é intrinsecamente multifatorial, englobando uma vasta gama de fatores genéticos, ambientais e relacionados ao estilo de vida. O'Donnell *et al*. (2016) destacam a hipertensão arterial como o fator de risco modificável mais preeminente para o AVC, salientando o papel crucial de outros elementos, tais como diabetes, fibrilação atrial, tabagismo e obesidade, na patogênese desta condição (O'Donnell *et al*., 2016).

Especificamente no contexto brasileiro, observa-se uma tendência ascendente na prevalência desses fatores de risco, um fenômeno que contribui para a exacerbação do risco de AVC na população (Da Luz *et al*., 2020; Malta *et al*., 2021a). Essa realidade reforça a necessidade imperativa de estratégias de prevenção e educação em saúde que sejam capazes de abordar de maneira efetiva esses determinantes de risco, promovendo estilos de vida mais saudáveis e contribuindo

para a redução da incidência de AVC no país (Boden-Albala; Quarles, 2013; De Oliveira *et al*., 2024a).

Em suma, o AVC no Brasil é um fenômeno complexo e multifacetado, marcado por desigualdades e desafios significativos (Bensenor *et al*., 2015a; Minelli *et al*., 2022b). A compreensão aprofundada das dimensões socioeconômicas e regionais que influenciam a incidência e os desfechos do AVC é fundamental para a formulação de estratégias de intervenção e políticas públicas eficazes, visando a mitigação das disparidades existentes e a melhoria da saúde cardiovascular da população brasileira (Ribeiro *et al*., 2018; Vincens; Stafström, 2015a).

Diante do intrincado cenário imposto pelo AVC, o Machine Learning (ML) tem se destacado como um instrumento revolucionário, dotado de um imenso potencial transformador na esfera da predição e manejo dessa condição clínica (Abedi *et al*., 2022; Mainali; Darsie; Smetana, 2021). O ML, sendo uma ramificação especializada da inteligência artificial (IA), abarca uma vasta gama de algoritmos e modelos computacionais, os quais são habilmente projetados para analisar, interpretar e extrair significado de conjuntos de dados de grande complexidade, viabilizando assim a identificação de padrões subjacentes e a realização de previsões assertivas (Gkantzios *et al*., 2023; Li *et al*., 2023).

No tocante à predição de eventos de AVC, modelos baseados em ML emergem como ferramentas de vanguarda, apresentando a capacidade de integrar e processar uma multiplicidade de variáveis, incluindo, mas não se limitando a dados clínicos, demográficos e históricos dos pacientes (Heo *et al*., 2019; Sirsat; Fermé; Câmara, 2020). Tal integração propicia a geração de modelos preditivos robustos, capazes de estimar o risco de AVC de um indivíduo com elevada precisão e confiabilidade, conforme demonstrado por Wang *et al*. (2020).

A aplicação de técnicas de ML na predição do AVC no Brasil ganha importância devido à diversidade e complexidade da população brasileira, além das conhecidas disparidades na incidência e nos desfechos do AVC (Daidone *et al*., 2024; Parro *et al*., 2021). O ML oferece aos profissionais de saúde uma ferramenta poderosa, capaz de gerar insights valiosos sobre os perfis de risco dos pacientes, permitindo intervenções mais rápidas, precisas e personalizadas (Lotufo; Bensenor, 2013; Su *et al*., 2022).

Além da aplicação no contexto do AVC, diversos algoritmos de ML já demonstraram sua eficácia na predição de outras doenças. Por exemplo, algoritmos

como redes neurais artificiais, máquinas de suporte vetorial e florestas aleatórias foram amplamente utilizados para prever condições como diabetes, doenças cardiovasculares e câncer (Delpino *et al*., 2022; Saridena; Saridena; Kethar, 2023).

Essas técnicas não apenas permitiram a detecção precoce e a estratificação do risco, mas também contribuíram para a personalização de estratégias de tratamento. A adaptação de tais algoritmos ao contexto do AVC potencializa sua capacidade preditiva, considerando a complexidade multifatorial dessa condição clínica (Al Kuwaiti *et al*., 2023; Chaki; Wozniak, 2024a).

A escolha de modelos de ML, como redes neurais ou florestas aleatórias, em relação a opções mais simples, como a regressão logística, baseia-se em sua habilidade de capturar interações complexas entre variáveis clínicas, demográficas e históricas, essenciais para prever o risco de AVC (Awasthi *et al*., 2024). Esses modelos apresentam precisão preditiva superior, especialmente quando aplicados a grandes conjuntos de dados heterogêneos (Talwar *et al*., 2023).

Ademais, a escolha do modelo também leva em conta a necessidade de um equilíbrio entre interpretabilidade e performance. Por exemplo, enquanto redes neurais oferecem alta precisão, podem carecer de transparência na tomada de decisões, o que é um aspecto crítico na área médica. Já modelos como florestas aleatórias oferecem uma melhor explicabilidade, o que pode ser preferível para a comunicação de resultados aos profissionais de saúde e para a implementação de estratégias de intervenção (Luo *et al*., 2019).

A decisão por um modelo específico, portanto, busca não apenas maximizar a acurácia preditiva, mas também garantir que os insights gerados sejam compreensíveis e aplicáveis na prática clínica (Shipley *et al*., 2022).

Os modelos de ML não se restringem apenas à avaliação de risco, mas também são um instrumentos na descoberta e na análise de novos fatores de risco, assim como na exploração das complexas e multifacetadas interações entre os diversos determinantes do AVC (Dritsas; Trigka, 2022b; Shao *et al*., 2022).

Esta capacidade de elucidar as intrincadas dinâmicas que permeiam a etiologia do AVC coloca o ML em uma posição privilegiada para contribuir significativamente para o avanço do conhecimento científico no campo da pesquisa sobre AVC, impulsionando, assim, o desenvolvimento de estratégias preventivas mais eficazes e de abordagens terapêuticas inovadoras (Alanazi; Abdou; Luo, 2021; Ouriques Martins *et al*., 2019).

Dessa forma, a adoção e a integração do ML nos protocolos de predição e manejo do AVC no Brasil representam um passo crucial em direção à personalização e à otimização do cuidado ao paciente, bem como à ampliação da compreensão dos mecanismos subjacentes a esta condição clínica (Sahu; Mishra; Kushwaha, 2022). Através da habilidade única do ML em processar e interpretar vastas quantidades de dados, torna-se possível não apenas antecipar eventos de AVC, mas também identificar oportunidades de intervenção precoce, maximizando assim as chances de recuperação e minimizando o impacto a longo prazo do AVC sobre a vida dos indivíduos (Ghassemi *et al*., 2020).

É importante sublinhar que o AVC continua a se apresentar como um desafio sanitário de magnitudes consideráveis no Brasil, ostentando uma influência marcante e preocupante nas esferas da mortalidade, morbidade, e na qualidade de vida dos indivíduos afetados (Lotufo, 2005). Este quadro se configura ainda mais complexo quando se observa as notórias disparidades presentes na incidência do AVC e nos resultados clínicos subsequentes ao evento, o que culmina em um apelo veemente por uma mobilização estratégica e integrada visando o desenvolvimento e a implementação de estratégias preditivas e preventivas eficazes (Pitchai *et al*., 2022).

É neste cenário que o ML emerge com destaque, apresentando-se como uma ferramenta revolucionária e inovadora, dotada da capacidade singular de processar, analisar e extrair conhecimentos valiosos a partir de vastos conjuntos de dados (Mainali; Darsie; Smetana, 2021).

Esta capacidade analítica avançada, intrinsecamente ligada ao ML, posiciona esta tecnologia como um vetor de transformação, oferecendo novas perspectivas e possibilidades na jornada em busca da redução do impacto avassalador do AVC (Su *et al*., 2022).

Adentrando o universo das aplicações práticas do ML na predição do AVC, observa-se um vasto campo de pesquisa e inovação. As seções subsequentes deste trabalho têm como objetivo explorar de forma aprofundada e minuciosa os diferentes modelos e algoritmos de ML aplicados especificamente à predição do AVC. Foi realizada uma avaliação criteriosa de sua eficácia, robustez e aplicabilidade no contexto clínico, buscando, assim, demarcar e elucidar as implicações diretas e indiretas destes achados para a prática na área da saúde, bem como para as estratégias de prevenção e manejo do AVC em solo brasileiro.

Neste processo investigativo, foi dada ênfase especial às nuances e peculiaridades associadas ao contexto brasileiro, considerando as variáveis sociodemográficas, econômicas e de saúde que conferem ao Brasil um perfil único e desafiador no que tange à luta contra o AVC. Foi contemplada a discussão sobre como o ML pode ser adaptado e personalizado para atender às necessidades específicas das diversas regiões do país, promovendo, assim, uma abordagem mais inclusiva e equitativa no combate ao AVC.

Assim, este trabalho se propõe a ser não apenas uma exposição teórica sobre a aplicação do ML na predição do AVC, mas também uma fonte de inspiração e orientação para futuras pesquisas, práticas clínicas e políticas de saúde pública. Busca-se, através da disseminação do conhecimento e da reflexão crítica, contribuir de maneira significativa para a transformação do panorama atual do AVC no Brasil, pavimentando o caminho para um futuro em que a prevenção e o manejo desta condição sejam mais efetivos, acessíveis e personalizados.

2.Objetivos

2.1 Objetivo geral

Identificar, por meio de técnicas de ML aplicadas aos microdados da PNS-2019, os fatores mais relevantes associados à ocorrência de AVC na população brasileira, analisando suas variações regionais e destacando especificidades locais. O estudo busca compreender como determinantes clínicos, demográficos, funcionais, comportamentais e sociais se articulam no risco de AVC, de modo a subsidiar estratégias de prevenção, vigilância epidemiológica e organização das linhas de cuidado no Sistema Único de Saúde (SUS).

2.2 Objetivos Específicos

Caracterizar a epidemiologia do AVC no Brasil

- Analisar padrões de incidência, prevalência e mortalidade associados ao AVC no país, com ênfase em desigualdades regionais.
- Identificar os principais fatores de risco e determinantes sociais vinculados ao AVC na população brasileira.

Analisar a influência dos fatores de risco na ocorrência do AVC

- Determinar quais variáveis apresentam maior impacto sobre a probabilidade de ocorrência do AVC em cada região do Brasil.
- Comparar a relevância relativa dos fatores entre regiões, buscando identificar padrões compartilhados e desigualdades regionais.
- Explorar a interação entre variáveis clínicas, sociodemográficas e funcionais no desfecho AVC.

Aplicar modelagem estatística e aprendizado de máquina na análise dos fatores de risco

- Utilizar algoritmos de ML e métodos estatísticos para hierarquizar a importância das variáveis associadas ao risco de AVC.
- Avaliar a robustez, estabilidade e interpretabilidade dos modelos empregados, considerando tanto análises nacionais quanto estaduais.
- Diferenciar variáveis preditoras de variáveis consequentes (pós-evento), discutindo suas implicações para vigilância, prevenção e prognóstico individual.

3.Hipótese

Os fatores de risco para AVC apresentam variações significativas entre as diferentes regiões do Brasil, sendo influenciados por características socioeconômicas, demográficas e clínicas específicas de cada localidade. A utilização de abordagens estatísticas e de aprendizado de máquina pode revelar padrões regionais distintos e contribuir para estratégias mais eficazes de prevenção e controle da doença.

4.Justificativa e Relevância do Estudo

O AVC representa uma das principais causas de morte e incapacidade no mundo, com uma incidência particularmente alarmante no Brasil, onde se observam disparidades substanciais em termos de acesso a cuidados de saúde e desfechos clínicos associados à condição (Silva *et al*., 2022). Essas diferenças regionais destacam a necessidade de uma abordagem que não apenas preveja o risco de AVC, mas que também identifique os fatores mais relevantes que contribuem para o aumento da sua probabilidade em cada região do país.

Diante desse cenário, torna-se essencial uma investigação aprofundada sobre os determinantes clínicos, demográficos e socioeconômicos do AVC no Brasil, a fim de compreender como esses fatores interagem e impactam diferentes populações (De Oliveira *et al*., 2024a). A identificação desses fatores pode subsidiar estratégias de prevenção mais eficazes e contribuir para a formulação de políticas públicas que reduzam as desigualdades na atenção à saúde. (Ogunpola *et al*., 2024).

A relevância deste estudo reside na sua capacidade de explorar, comparar e analisar as variações regionais dos fatores de risco para AVC, considerando as peculiaridades sociodemográficas e econômicas que caracterizam o país (Bensenor *et al*., 2015a). Compreender como esses fatores se distribuem e influenciam a incidência do AVC permitirá o desenvolvimento de estratégias direcionadas a populações específicas, maximizando o impacto das ações preventivas e de intervenção (Sunny *et al*., 2024).

Além disso, ao abordar as disparidades existentes nos fatores de risco e nos desfechos do AVC entre diferentes regiões do Brasil, o estudo assume um caráter de justiça social, fornecendo subsídios para políticas de saúde que promovam maior equidade no acesso e na qualidade dos cuidados prestados (Do Carmo Ferreira; Sarti; De Azevedo Barros, 2022). Identificar os determinantes mais relevantes em cada localidade pode auxiliar na alocação mais eficiente de recursos, otimizando o planejamento e a implementação de ações preventivas.

A pesquisa também se justifica pela necessidade de superar lacunas no conhecimento sobre a epidemiologia do AVC no Brasil, especialmente no que diz respeito às desigualdades regionais. Ao integrar abordagens estatísticas e de aprendizado de máquina, este estudo pretende revelar padrões ocultos e fornecer uma análise mais detalhada da influência de múltiplos fatores no risco de AVC (Olawade *et al*., 2023).

Portanto, diante do impacto devastador do AVC no Brasil e da necessidade de uma compreensão mais precisa dos fatores que contribuem para sua incidência, este estudo se justifica não apenas pela sua relevância científica e acadêmica, mas também pelo seu potencial de influenciar positivamente práticas clínicas, políticas de saúde e, consequentemente, a qualidade de vida da população brasileira (Fernandes, 2015).

5.Referencial Teórico

5.1 Epidemiologia do AVC no Brasil

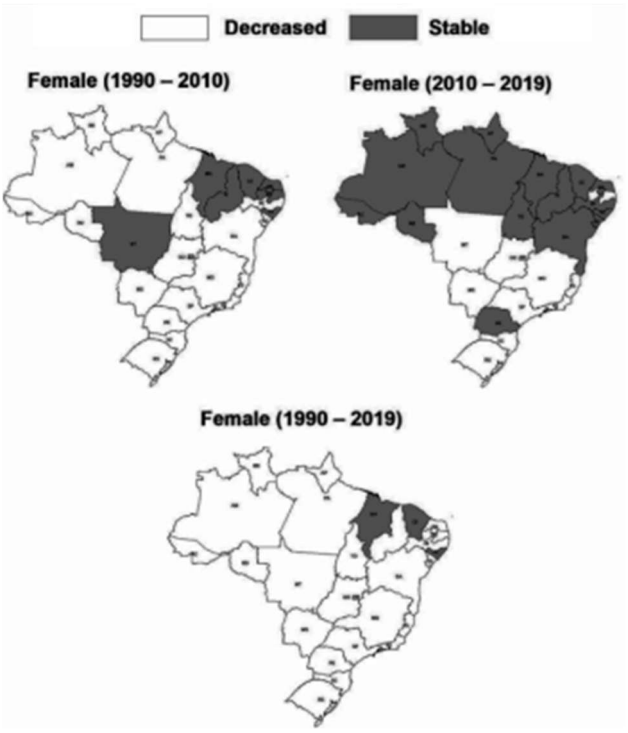
A incidência do AVC no Brasil permanece uma preocupação significativa para a saúde pública, especialmente devido às profundas desigualdades regionais e socioeconômicas (Cabral *et al*., 2009). O AVC apresenta taxas alarmantes de ocorrência, particularmente em regiões onde o acesso a cuidados médicos preventivos e de emergência é limitado (Lotufo; Benseñor, 2009). Estudos epidemiológicos indicam que a incidência é maior nas regiões Norte e Nordeste do Brasil, onde barreiras no acesso a cuidados primários e especializados dificultam tanto a prevenção quanto o diagnóstico precoce (Martins *et al*., 2022).

O impacto dos fatores de risco modificáveis, como hipertensão arterial, tabagismo, diabetes, obesidade e sedentarismo, é amplamente reconhecido como um dos principais determinantes do aumento da incidência do AVC (Rochmah *et al*., 2021).

A baixa atividade física, em particular, contribui para o agravamento de condições como hipertensão e obesidade, que são fatores de risco diretamente associados ao AVC (Coelho; Burini, 2009). Além disso, a interação desses fatores com determinantes sociais e ambientais, como baixo nível educacional e infraestrutura precária de saúde, exacerba ainda mais as disparidades regionais (Pontes-Neto *et al*., 2008).

Embora haja uma tendência de redução na incidência global de AVC em estados mais desenvolvidos, como São Paulo e Rio Grande do Sul, esta redução não é uniforme, refletindo uma transição epidemiológica incompleta em nível nacional (Lotufo; Benseñor, 2009). Esse cenário é ilustrado pela análise da mudança na taxa de mortalidade padronizada por idade (por 100.000 habitantes) atribuível à baixa atividade física em mulheres (≥ 25 anos) no Brasil nos períodos de 1990-2010, 2010-2019 e 1990-2019, conforme apresentado por SILVA *et al*., (2022) na figura 1:

Figura 1 - Mudança na taxa de mortalidade padronizada por idade (por 100.000 habitantes) devido a AVC atribuível à baixa atividade física em mulheres (≥ 25 anos) no Brasil (1990-2010; 2010-2019; e 1990-2019).

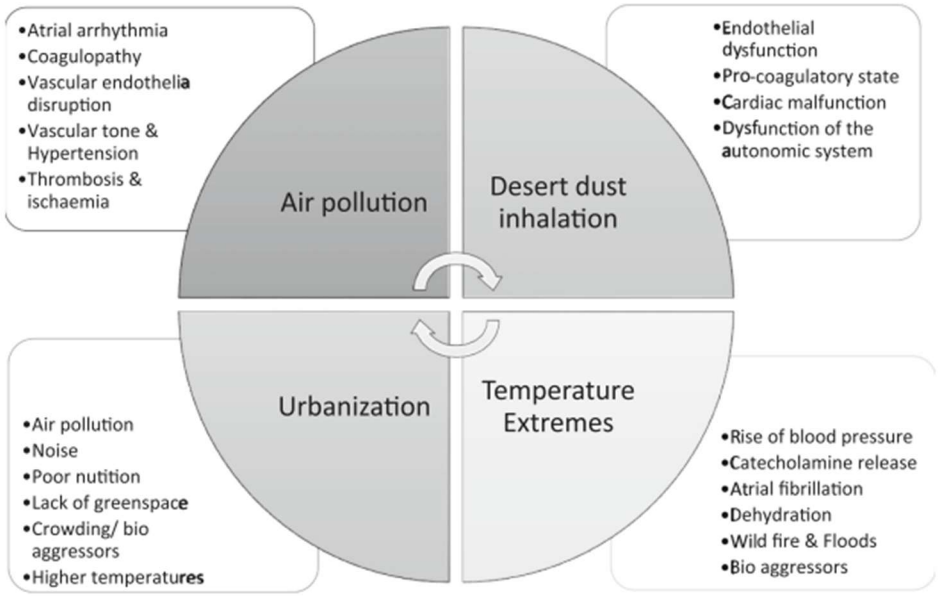


Fonte: Silva *et al*., 2022.

Além disso, a urbanização acelerada em algumas regiões está diretamente relacionada ao aumento de fatores de risco cardiovasculares, como a exposição à poluição do ar, temperaturas extremas e más condições urbanas, os quais impactam significativamente as taxas de AVC.

Esses fatores ambientais contribuem para condições como arritmia atrial, disfunção endotelial e aumento da pressão arterial (PA), intensificando o risco de eventos cerebrovasculares. Por exemplo, a poluição do ar está associada a alterações na coagulação e no tônus vascular, enquanto a urbanização promove o surgimento de condições como má nutrição, ruído excessivo e falta de áreas verdes. Adicionalmente, a inalação de poeira agrava disfunções endoteliais e do sistema autonômico, e os temperatura extremas intensificam crises hipertensivas, desidratação e outros fatores predisponentes. Esse cenário reforça a urgência de ações integradas que combinem programas educativos, promoção de hábitos saudáveis e melhorias estruturais no sistema de saúde (Malta *et al*., 2021b) , como ilustrado na figura de Ranta *et al*., (2023) :

Figura 2 - Mecanismos e interações dos fatores de risco ambientais para AVC.



Fonte: (Ranta *et al*., 2023)

A prevalência de AVC no Brasil tem crescido ao longo dos anos, impulsionada pela redução da mortalidade aguda e pelo aumento da expectativa de vida. Essa situação resulta em um número crescente de indivíduos sobreviventes ao evento agudo, muitos dos quais apresentam sequelas que impactam negativamente sua qualidade de vida. Dados recentes indicam que, enquanto o número de novos casos de AVC pode ter diminuído em algumas regiões, o número total de pessoas vivendo com a condição aumentou substancialmente (Cabral *et al*., 2009; Feigin; Norrving; Mensah, 2017).

Esse declínio observado na mortalidade geral pode ser atribuído, em parte, ao avanço em intervenções emergenciais, como o uso de trombólise em casos isquêmicos e a melhoria na capacitação das equipes médicas. No entanto, essas melhorias não se refletem de maneira uniforme em todo o território nacional (Martins *et al*., 2022). Fatores como a distância dos centros de atendimento, a ausência de serviços especializados e as barreiras culturais no reconhecimento precoce dos sintomas contribuem para taxas mais altas de mortalidade em determinadas localidades (Tereza *et al*., 2022).

Adicionalmente, os determinantes sociais de saúde desempenham um papel crítico e multifacetado na mortalidade por AVC no Brasil, refletindo a profunda interação entre condições socioeconômicas e os desfechos clínicos (Barreto, 2017).

A pobreza, um dos fatores mais evidentes, impõe barreiras significativas ao acesso a cuidados preventivos, diagnóstico precoce e tratamentos adequados. Indivíduos em situação de vulnerabilidade econômica frequentemente enfrentam dificuldades em adquirir medicamentos essenciais, realizar consultas regulares e acessar serviços de saúde de qualidade, o que contribui diretamente para o agravamento de condições subjacentes como hipertensão e diabetes, ambos fatores de risco primários para o AVC (Lessa, 1985; Malta *et al*., 2021a).

O baixo nível educacional amplifica ainda mais essa disparidade, pois limita a capacidade de compreender informações médicas e reconhecer sinais de alerta (Berkman *et al*., 2011). A alfabetização em saúde, ou seja, a habilidade de interpretar e aplicar orientações relacionadas à saúde, é crucial para buscar atendimento em tempo hábil, especialmente em casos de emergência como o AVC, onde cada minuto é decisivo (De Jesus *et al*., 2024). No entanto, essa competência é muitas vezes insuficiente em comunidades marginalizadas, resultando em atrasos na procura por tratamento e, consequentemente, piores prognósticos (Paasche-Orlow; Wolf, 2007).

Além disso, a desigualdade no acesso a serviços de emergência é um exemplo claro e alarmante de como as condições estruturais perpetuam diferenças nos índices de mortalidade (Pandian *et al*., 2020). Em regiões menos desenvolvidas, como o Norte e Nordeste do Brasil, a escassez de unidades especializadas, e a infraestrutura insuficiente dificultam não apenas o atendimento inicial, mas também o acompanhamento e a reabilitação. A falta de transporte adequado e a distância até os centros de saúde especializados também desempenham um papel significativo, criando barreiras adicionais que afetam desproporcionalmente as populações rurais e periféricas (Martins *et al*., 2022).

Diante desse cenário, a necessidade de políticas públicas que ampliem a cobertura e a acessibilidade dos serviços de saúde torna-se evidente e urgente (Facchini; Tomasi; Dilélio, 2018). Essas políticas devem incluir desde a expansão de unidades especializadas em regiões de maior vulnerabilidade até a implementação de programas de educação em saúde voltados para comunidades de baixa renda, que promovam o conhecimento sobre prevenção e a identificação de sinais de alerta do AVC (Marmot; Bell, 2012). Somente por meio de uma abordagem integrada e equitativa será possível mitigar os efeitos dos determinantes sociais na mortalidade por AVC, reduzindo as disparidades e melhorando os desfechos clínicos em todo o país (Malta *et al*., 2021b).

5.2 Etiologia e Fatores Preditivos do AVC

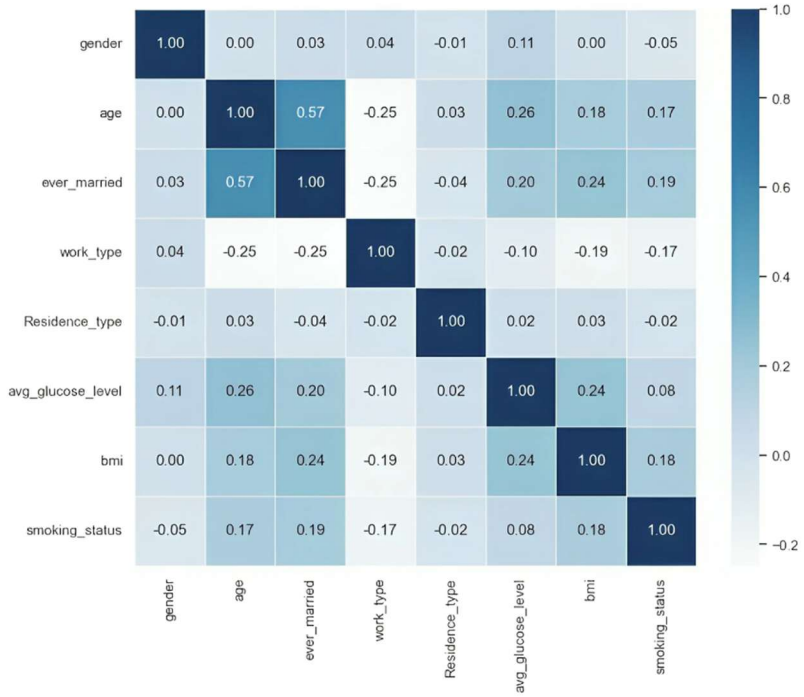
A predição de AVC é um desafio complexo que exige a integração de múltiplas fontes de informações para alcançar maior precisão e utilidade clínica. Nesse contexto, os dados clínicos e demográficos emergem como pilares fundamentais para a construção de modelos preditivos robustos, especialmente em um país como o Brasil, onde as disparidades regionais e socioeconômicas impactam significativamente os desfechos de saúde.

Estudos recentes têm explorado a eficácia de técnicas de aprendizado de máquina na predição de ocorrências de AVC, considerando fatores demográficos, clínicos e de estilo de vida. Por exemplo, CHAKRABORTY *et al*., (2024) investigaram a aplicação de técnicas de ML para prever ocorrências de AVC, combinando métodos que reduzem a complexidade dos dados e realçam os padrões mais importantes com estratégias que integram diferentes modelos preditivos em uma solução única e mais robusta. Essa abordagem conjunta permitiu alcançar uma precisão de 98,6% na predição de casos de AVC, demonstrando como a simplificação dos dados, aliada à combinação de múltiplos modelos, pode aumentar a capacidade de identificar corretamente os indivíduos em risco. O estudo também apresenta, na Figura 3, a matriz de correlação das variáveis no conjunto de dados de AVC, destacando as relações entre os fatores analisados.

Essa matriz permite visualizar como diferentes variáveis interagem entre si, evidenciando, por exemplo, a forte correlação entre idade e estado civil, a relação positiva entre níveis médios de glicose e idade, e a influência do IMC nos níveis glicêmicos. Além disso, observa-se que algumas variáveis apresentam correlações fracas ou inexistentes, sugerindo que outros fatores podem ser mais determinantes na previsão do risco de AVC. Esses achados reforçam a importância de abordagens analíticas robustas para compreender melhor os fatores que contribuem para a doença e aprimorar estratégias preventivas.

Além disso, a pesquisa de (Shobayo *et al*., 2023) desenvolveu um modelo de predição de AVC utilizando dados demográficos e comportamentais, aplicando o algoritmo de árvore de decisão (Figura 4). Os resultados indicaram que a idade e o índice de massa corporal (IMC) são preditores significativos de incidência de AVC.

Figura 3 - Matriz de correlação das variáveis no conjunto de dados de AVC.

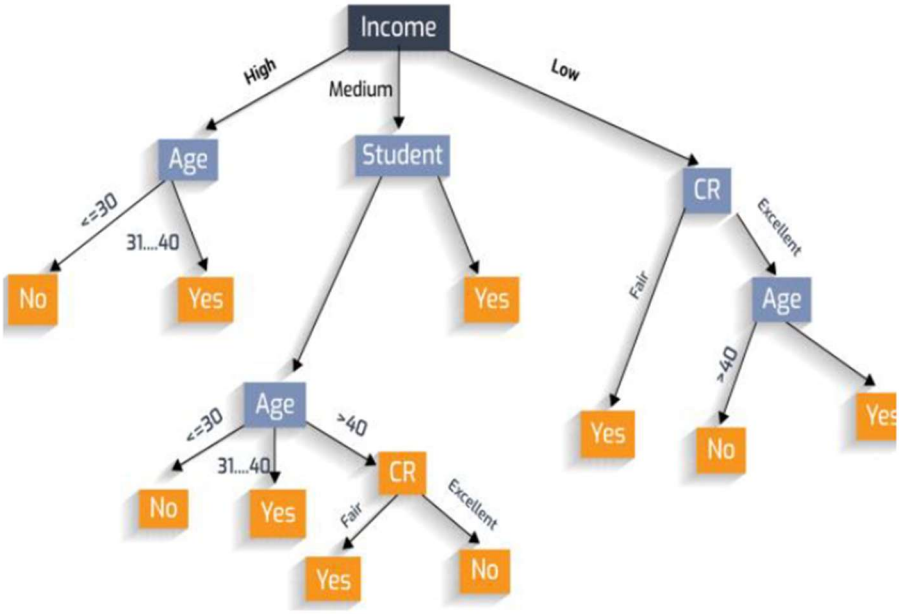


Fonte: (Chakraborty et al., 2024)

A estrutura da árvore de decisão ilustra como diferentes variáveis influenciam a previsão do risco, segmentando os indivíduos com base em características como faixa etária, status socioeconômico e fatores comportamentais. Esse modelo permite visualizar de maneira clara a hierarquia de importância das variáveis e a forma como elas se combinam para determinar o risco de AVC, contribuindo para uma melhor interpretação dos fatores envolvidos e auxiliando na formulação de estratégias preventivas mais eficazes.

No contexto brasileiro, Kurtz *et al*. (2022) propuseram uma metodologia baseada em dados para desenvolver modelos de predição de tempo de internação e mortalidade em 30 dias em um coorte multicêntrico de UTIs no Brasil. A análise incluiu dados de 130 UTIs de 43 hospitais brasileiros, destacando a importância de considerar variáveis clínicas e demográficas na construção de modelos preditivos.

Figura 4 - Uma ilustração da árvore de decisão.



Fonte: (Charbuty; Jijo; Abdulazeez, 2021)

Esses estudos enfatizam a relevância de integrar dados clínicos e demográficos na construção de modelos preditivos de AVC, especialmente em contextos com disparidades regionais e socioeconômicas, como o Brasil.

Os dados clínicos oferecem insights diretos sobre as condições fisiológicas que predispõem ao AVC. Variáveis como P.A. elevada, glicemia alterada, perfil lipídico desfavorável e histórico de fibrilação atrial são reconhecidamente os principais marcadores de risco para essa condição.

A hipertensão arterial, por exemplo, é o fator mais prevalente associado ao AVC (Li *et al*., 2022), sendo responsável por mais da metade dos casos registrados (Virani *et al*., 2021). A inclusão de tais variáveis em modelos preditivos permite uma estratificação precisa de risco, possibilitando intervenções direcionadas e oportunas (Sheer *et al*., 2022). Além disso, o avanço nos registros eletrônicos de saúde, como o Sistema de Informação em Saúde para a Atenção Básica (SISAB), oferece uma rica fonte de dados que pode ser utilizada para análises mais abrangentes e personalizadas.

Os dados demográficos, por sua vez, complementam a análise ao incorporar aspectos individuais e sociais que influenciam a suscetibilidade ao AVC. Informações como idade, sexo e nível socioeconômico têm mostrado forte correlação com o risco de ocorrência de AVC (Santos *et al*., 2022a). No Brasil, por exemplo, a idade

avançada está associada a um aumento expressivo na probabilidade de eventos vasculares, enquanto mulheres na pós-menopausa apresentam risco ampliado devido a alterações hormonais e metabólicas (Copstein; Fernandes; Bastos, 2013). Além disso, populações de baixa renda enfrentam maior exposição a fatores de risco modificáveis e menor acesso a cuidados preventivos e emergenciais, o que agrava os desfechos clínicos (Avan *et al*., 2019).

A combinação de dados clínicos e demográficos em modelos preditivos é especialmente relevante em um contexto marcado por desigualdades sociais e regionais. No entanto, a implementação eficaz dessas abordagens no Brasil enfrenta desafios significativos relacionados à qualidade e à disponibilidade dos dados (Victora, 2016). Problemas como subnotificação, fragmentação dos sistemas de informação e falta de padronização comprometem a acurácia dos modelos e limitam seu uso em larga escala. Além disso, a ausência de interoperabilidade entre os sistemas de saúde dificulta a integração de dados provenientes de diferentes fontes, restringindo o potencial de análises mais complexas (Barbalho *et al*., 2022).

Apesar desses desafios, avanços tecnológicos na área de ML têm possibilitado o desenvolvimento de modelos preditivos mais sofisticados, que integram variáveis clínicas e sociais para uma análise holística. Técnicas como redes neurais, florestas aleatórias e máquinas de suporte vetorial demonstram capacidade de identificar interações complexas entre variáveis, proporcionando insights valiosos sobre os fatores que aumentam o risco de AVC (Olabanjo *et al*., 2024). No entanto, para maximizar o impacto desses modelos, é essencial investir na qualidade e acessibilidade dos dados, promovendo maior padronização nos registros e capacitação das equipes de saúde responsáveis pela coleta e gestão dessas informações (Al-Hgaish *et al*., 2025; Mainali; Darsie; Smetana, 2021).

Em síntese, os dados clínicos e demográficos desempenham um papel indispensável na predição do AVC, especialmente em contextos de alta heterogeneidade populacional como o Brasil. A integração eficiente dessas informações em modelos de predição tem o potencial de transformar o manejo preventivo e emergencial do AVC, contribuindo para uma redução significativa dos índices de morbidade e mortalidade associados a essa condição. Contudo, para que esse potencial seja plenamente alcançado, é necessário superar as barreiras relacionadas à qualidade dos dados e à infraestrutura de saúde, garantindo que a tecnologia esteja alinhada às necessidades e especificidades da população brasileira.

5.3 Machine Learning na Saúde

O uso de ML na área da saúde representa uma das mais notáveis transformações tecnológicas no campo médico, revolucionando desde o diagnóstico precoce de doenças até a personalização de tratamentos (Davenport; Kalakota, 2019).

O ML, uma subárea da IA, utiliza algoritmos e técnicas estatísticas para identificar padrões complexos em grandes volumes de dados e fazer previsões baseadas em informações previamente coletadas. Diferentemente das abordagens tradicionais, que dependem de modelos predefinidos ou hipóteses específicas, o ML adapta-se continuamente aos dados, tornando-se uma ferramenta poderosa para lidar com a complexidade e diversidade presentes no setor de saúde (Weissler *et al*., 2021).

O conceito de ML baseia-se na ideia de que máquinas podem aprender a partir de dados, identificando correlações e tendências sem a necessidade de programação explícita para cada tarefa (Davenport; Kalakota, 2019). Em vez de seguir regras fixas, os algoritmos analisam padrões e fazem ajustes para melhorar continuamente sua acurácia e desempenho. Na saúde, essa capacidade é particularmente útil em contextos em que grandes conjuntos de dados clínicos, genômicos e demográficos precisam ser analisados para detectar relações que seriam impossíveis de identificar por métodos tradicionais (Weissler *et al*., 2021).

As aplicações de ML na saúde são diversas e abrangem desde a detecção precoce de doenças crônicas, como diabetes e doenças cardiovasculares, até o auxílio no tratamento de condições complexas, como câncer e doenças raras (An *et al*., 2023). Na área de imagens médicas, por exemplo, algoritmos de ML têm demonstrado alto desempenho na identificação de anomalias em exames de raios-X, tomografias e ressonâncias magnéticas, muitas vezes superando especialistas humanos em termos de precisão. Além disso, sistemas baseados em ML estão sendo usados para prever a resposta de pacientes a determinados medicamentos, ajudando a personalizar tratamentos e melhorar os desfechos clínicos (Das *et al*., 2024; Miotto *et al*., 2018).

No contexto do AVC, o ML tem mostrado resultados promissores na predição de risco, integrando variáveis clínicas, demográficas e até genéticas (Dritsas; Trigka, 2022a). Modelos baseados em ML podem avaliar dados de múltiplas fontes

simultaneamente, como histórico médico, níveis de glicemia, P.A. e fatores sociodemográficos, oferecendo uma visão abrangente e precisa do risco individual de AVC (Olabanjo *et al*., 2024). Essa abordagem permite intervenções preventivas mais eficazes, otimizando o uso de recursos e reduzindo o impacto da doença em populações de risco (Chakraborty *et al*., 2024).

As abordagens tradicionais de análise de dados, como regressões lineares e análise de variância, têm limitações significativas ao lidar com os desafios contemporâneos da saúde (Obermeyer; Emanuel, 2016). Essas metodologias são eficazes para conjuntos de dados menores e variáveis com relações lineares, mas tornam-se insuficientes diante da complexidade, volume e heterogeneidade dos dados clínicos e populacionais modernos (Chen; Asch, 2017; Rajkomar; Dean; Kohane, 2019). Nesse cenário, o ML destaca-se como uma alternativa superior por diversas razões.

Primeiramente, o ML é capaz de processar grandes volumes de dados multidimensionais, integrando informações de diferentes fontes, como registros médicos eletrônicos, dados de sensores vestíveis e características demográficas. Essa capacidade é crucial em condições como o AVC, onde múltiplos fatores de risco interagem de maneira complexa, tornando os métodos tradicionais inadequados para capturar toda a extensão dessas interações (Langner, 2024; Shishehbori; Awan, 2024).

Além disso, o ML pode identificar padrões não lineares e relações sutis entre variáveis que não seriam detectadas por técnicas estatísticas convencionais (Dritsas; Trigka, 2022b; Olabanjo *et al*., 2024). Por exemplo, enquanto um modelo de regressão linear pode identificar um aumento progressivo do risco de AVC com o aumento da PA, um algoritmo de ML pode detectar interações entre hipertensão, diabetes e fatores genéticos que amplificam exponencialmente esse risco em determinados subgrupos populacionais (Orfanoudaki *et al*., 2020).

Outro diferencial do ML é sua capacidade de aprendizado contínuo. À medida que novos dados são coletados, os modelos podem ser atualizados para refletir mudanças nos padrões de saúde e nos fatores de risco, mantendo sua relevância e acurácia ao longo do tempo (Armstrong; Clifton, 2021; Singh *et al*., 2023). No caso do AVC, isso significa que os modelos podem ser ajustados para incorporar novas descobertas científicas ou mudanças no perfil epidemiológico da população brasileira, garantindo maior eficácia nas estratégias de prevenção (Rajagopal *et al*., 2024).

Por fim, o ML promove o avanço da medicina personalizada, permitindo que os riscos e tratamentos sejam adaptados às características únicas de cada paciente. Essa abordagem personalizada é particularmente relevante no manejo do AVC, onde intervenções preventivas direcionadas podem reduzir significativamente a morbidade e a mortalidade.

A implementação de ML na saúde pública enfrenta obstáculos significativos. Do ponto de vista técnico, um dos maiores desafios é a qualidade e a disponibilidade dos dados. No Brasil, como em outros países, a coleta de dados de saúde enfrenta problemas como subnotificação, inconsistências nos registros e falta de padronização entre diferentes sistemas (Silva; Sanine, 2020). Isso afeta diretamente a capacidade dos modelos de ML de gerar previsões precisas e confiáveis. Além disso, a fragmentação dos sistemas de informação e a ausência de interoperabilidade dificultam a integração de dados provenientes de diferentes fontes, limitando o uso de análises mais robustas.

Outro desafio técnico é a transparência e a interpretabilidade dos modelos de ML. Muitas das abordagens mais avançadas, como redes neurais profundas, são frequentemente descritas como "caixas-pretas", devido à dificuldade de compreender como as decisões são tomadas (Banegas-Luna *et al*., 2020). Em um ambiente como a saúde, onde as decisões têm implicações críticas para a vida dos pacientes, a falta de clareza pode gerar resistência por parte de profissionais e instituições de saúde (Stiglic *et al*., 2020).

No âmbito ético, a privacidade e a segurança dos dados são preocupações centrais. Os algoritmos de ML dependem de grandes volumes de informações sensíveis, como dados genéticos e históricos médicos, o que aumenta o risco de vazamentos e usos indevidos (Da Nunes; Guimarães; Dadalto, 2022). Além disso, a potencial presença de vieses nos modelos, muitas vezes derivados de preconceitos embutidos nos dados de treinamento, pode perpetuar desigualdades e levar a decisões que desfavorecem grupos vulneráveis (Elias *et al*., 2024). Esses problemas são particularmente preocupantes em um país como o Brasil, onde desigualdades socioeconômicas e regionais já são marcantes.

Por fim, a implementação do ML na saúde pública requer infraestrutura tecnológica avançada, como armazenamento em nuvem e capacidade computacional robusta, além de equipes treinadas para operar e interpretar os resultados dos modelos. Muitas regiões brasileiras, especialmente as mais remotas, não possuem

os recursos necessários para adotar essas tecnologias, o que cria barreiras adicionais à sua disseminação.

5.4 Aplicações de Machine Learning no AVC

Modelos de ML têm sido amplamente aplicados à predição de risco de AVC, conforme evidenciado por diversas revisões de estudos anteriores (Chaki; Wozniak, 2024b; Li *et al*., 2023; Olabanjo *et al*., 2024; Shao *et al*., 2022; Sirsat; Fermé; Câmara, 2020; Wang *et al*., 2020b). Esses modelos utilizam algoritmos sofisticados para analisar grandes volumes de dados, identificando padrões e relações que seriam difíceis de detectar por métodos convencionais (Davenport; Kalakota, 2019).

Redes neurais artificiais, florestas aleatórias, máquinas de suporte vetorial e até mesmo a regressão logística são frequentemente empregados para prever a ocorrência de AVC com base em variáveis como histórico médico, fatores de risco modificáveis (hipertensão, diabetes, tabagismo), características demográficas (idade, sexo) e fatores socioeconômicos (Vu *et al*., 2024). Estudos destacam que o desempenho desses modelos pode ser significativamente superior às abordagens tradicionais, especialmente em cenários onde os dados são complexos e heterogêneos.

As redes neurais artificiais, por exemplo, são conhecidas por sua capacidade de capturar interações não lineares e padrões complexos em grandes conjuntos de dados. No entanto, sua natureza de "caixa-preta" pode limitar sua aplicação em contextos onde a transparência é essencial. Em contrapartida, as florestas aleatórias oferecem um equilíbrio valioso entre precisão e interpretabilidade, fornecendo informações claras sobre a importância das variáveis utilizadas no modelo. Máquinas de suporte vetorial, por sua vez, são eficazes em problemas de classificação binária, como distinguir indivíduos com alto risco de AVC daqueles com baixo risco, mas podem apresentar dificuldades em conjuntos de dados de alta dimensionalidade. A regressão logística, embora menos sofisticada, continua sendo uma referência por sua simplicidade e interpretabilidade, sendo muitas vezes utilizada como base para comparação com algoritmos mais avançados (Asadi *et al*., 2024; Fan *et al*., 2020).

Apesar dos avanços significativos, a implementação de modelos de ML na predição de AVC enfrenta desafios técnicos e éticos consideráveis. A qualidade dos

dados é uma questão central, uma vez que dados incompletos, inconsistentes ou desatualizados podem comprometer a confiabilidade das previsões.

Os desafios éticos também não podem ser ignorados. Questões como privacidade dos dados e vieses algorítmicos são críticas em um país marcado por profundas desigualdades sociais e regionais (Naik *et al*., 2022). Modelos de ML, quando treinados em dados que refletem essas desigualdades, podem perpetuar ou até ampliar disparidades no cuidado em saúde, favorecendo grupos já privilegiados em detrimento de populações vulneráveis (Gurevich; El Hassan; El Morr, 2022). Portanto, é fundamental que a implementação dessas tecnologias seja acompanhada de estratégias para mitigar esses riscos, garantindo que os benefícios do ML sejam distribuídos de forma equitativa.

5.5 Integração de Dados em Modelos de Machine Learning

A integração de dados clínicos e demográficos em ML desempenha um papel essencial na construção de sistemas preditivos robustos para o AVC. A combinação de informações detalhadas, como níveis de glicemia, P.A. e histórico de hipertensão, com variáveis demográficas, como idade, sexo e nível socioeconômico, permite capturar a complexidade multifatorial dessa condição. Essa abordagem melhora a acurácia dos modelos, possibilita intervenções personalizadas e identifica interações entre fatores que, de outra forma, poderiam ser negligenciadas (Chakraborty *et al*., 2024; Olabanjo *et al*., 2024).

Enquanto os dados clínicos fornecem a base para entender os determinantes biológicos do risco de AVC, os dados demográficos adicionam contexto, ajustando os modelos às características individuais e populacionais. Fatores como idade avançada e condições socioeconômicas adversas não apenas influenciam diretamente o risco, mas também interagem com variáveis clínicas de formas complexas. A inclusão estratégica dessas variáveis, com o uso de técnicas de seleção e redução para evitar redundância, é crucial para maximizar o potencial dos modelos preditivos, equilibrando precisão e eficiência computacional (Dritsas; Trigka, 2022b; Olabanjo *et al*., 2024).

Entre as técnicas mais utilizadas está a análise de importância das variáveis, que identifica quais fatores têm maior influência no desempenho do modelo. Métodos como "feature importance" em florestas aleatórias ou análise baseada em permutação

quantificam o impacto de cada variável individualmente, permitindo que os desenvolvedores priorizem aquelas que mais contribuem para a acurácia do modelo (Gerstorfer; Hahn-Klimroth; Krieg, 2023; Iranzad; Liu, 2024).

Além disso, a redução de dimensionalidade, por meio de técnicas como Análise de Componentes Principais e métodos de seleção sequencial, é frequentemente aplicada para simplificar os conjuntos de dados sem perder informações cruciais. Essas abordagens são especialmente úteis em contextos onde há alta correlação entre variáveis, como ocorre frequentemente com dados clínicos (por exemplo, P.A. sistólica e diastólica) e dados socioeconômicos (Destrero *et al*., 2009; Pudjihartono *et al*., 2022).

Ao aplicar essas técnicas, é possível não apenas otimizar o desempenho computacional dos modelos, mas também aumentar sua interpretabilidade. Em contextos médicos, onde a transparência e a justificativa das decisões são essenciais, a capacidade de explicar por que uma variável específica foi incluída no modelo é um diferencial significativo.

Vários estudos têm demonstrado o valor da integração de dados clínicos e demográficos na predição de desfechos de AVC utilizando modelos de aprendizado de ML. Por exemplo, HEO *et al*. (2019) desenvolveram modelos preditivos baseados em aprendizado profundo, floresta aleatória e regressão logística, utilizando 38 variáveis, incluindo dados demográficos (como idade e sexo), clínicos (como glicemia e PA) e históricos médicos. Os resultados indicaram que o modelo de aprendizado profundo apresentou desempenho superior ao escore ASTRAL (Acute Stroke Registry and Analysis of Lausanne), uma ferramenta tradicional usada para prever desfechos funcionais após AVC com base em seis variáveis clínicas simples, como idade e glicemia. O estudo demonstrou que a integração de um conjunto mais amplo de variáveis por meio de aprendizado de máquina pode melhorar significativamente a acurácia na predição de desfechos de longo prazo em pacientes com AVC isquêmico.

Outro exemplo é o trabalho de SHOBAYO *et al*. (2023), que explorou a aplicação do algoritmo de Random Forest para prever a incidência de AVC. Nesse estudo, foram integrados dados demográficos e comportamentais, como idade, IMC e padrões de atividade física, para construir um modelo preditivo robusto. Os resultados indicaram que o Random Forest superou outros modelos, como árvores de decisão e regressão logística. O estudo destacou a relevância de variáveis como

idade e IMC como preditores significativos, demonstrando o potencial do aprendizado de máquina para melhorar a acurácia na predição de risco de AVC, embora não tenha abordado diretamente variáveis socioeconômicas, como nível de escolaridade ou acesso a cuidados médicos.

No Brasil, estudos como o de MALTA *et al*. (2021a) destacaram o papel das desigualdades socioeconômicas na prevalência e impacto das doenças crônicas não transmissíveis, incluindo fatores de risco para o AVC. Utilizando dados da PNS, os autores identificaram disparidades significativas entre regiões e estratos socioeconômicos, mostrando que indivíduos com menor escolaridade, sem plano de saúde privado e residentes em regiões menos desenvolvidas enfrentam maior prevalência de doenças crônicas e suas limitações associadas. Esses achados reforçam a importância de considerar fatores contextuais, como desigualdades regionais e socioeconômicas, ao desenvolver modelos preditivos adaptados à realidade brasileira.

5.6 Lacunas na Literatura

Apesar dos avanços significativos no uso de ML para análise do risco de AVC, ainda existem lacunas importantes na literatura que limitam a aplicabilidade e a generalização dos modelos desenvolvidos. Essas lacunas são particularmente evidentes no contexto brasileiro, onde a heterogeneidade populacional, as disparidades regionais e os desafios infraestruturais exigem abordagens específicas e adaptadas.

Uma das principais limitações dos estudos existentes é a concentração em populações de países desenvolvidos, onde os fatores de risco e as condições de saúde diferem significativamente daqueles observados no Brasil. Muitos modelos preditivos são baseados em dados de alta qualidade, provenientes de registros médicos eletrônicos bem integrados, com acesso amplo a tecnologias avançadas de diagnóstico e tratamento. No entanto, essa realidade não reflete o cenário brasileiro, onde os dados frequentemente apresentam fragmentação, subnotificação e inconsistências.

Além disso, grande parte dos modelos preditivos de AVC não considera variáveis socioeconômicas e contextuais, focando principalmente em fatores clínicos, como P.A. e glicemia. No entanto, aspectos como renda, escolaridade e acesso a

serviços de saúde desempenham um papel crucial no risco de AVC e seus desfechos, especialmente em populações vulneráveis. A ausência dessas variáveis limita a capacidade dos modelos de capturar as nuances do risco em contextos socioeconomicamente diversos.

Outra limitação significativa é o uso de bases de dados restritas, com amostras reduzidas e pouco representativas da diversidade populacional. Modelos desenvolvidos a partir de dados homogêneos podem apresentar viés ao serem aplicados em populações heterogêneas, como a brasileira, onde há grandes variações epidemiológicas e demográficas. Esse fator compromete a generalização dos resultados e reduz a aplicabilidade dos modelos em diferentes realidades regionais.

Além disso, os métodos de validação utilizados em muitos estudos nem sempre refletem cenários reais, como a presença de dados ausentes ou a aplicação em sistemas de saúde com recursos limitados. A falta de testes em ambientes mais próximos da realidade brasileira reduz a robustez e aplicabilidade prática dos modelos no contexto nacional.

Por fim, a diversidade populacional do Brasil, incluindo diferenças culturais, comportamentais e ambientais, demanda abordagens mais adaptadas à realidade local. Variáveis que podem ser irrelevantes em outros países podem ter influência significativa no risco de AVC no Brasil, como fatores associados ao acesso a cuidados preventivos, padrões alimentares e uso de medicamentos. No entanto, essas variáveis ainda são subexploradas nos modelos atuais, comprometendo a precisão das análises e a relevância das predições.

6.Metodologia

Este estudo adota uma abordagem metodológica sistemática e rigorosa, com o intuito de explorar a aplicabilidade e eficácia de modelos de ML na predição do AVC no Brasil. A investigação é ancorada em um extenso banco de dados proveniente da PNS 2019, que engloba 1.114 categorias (perguntas feitas aos entrevistados) e 293.727 respostas de questionários, fornecendo uma rica fonte de dados para análise e modelagem.

6.1 Preparação dos Dados

A fase inicial do estudo foi dedicada à preparação e limpeza dos dados da PNS-2019. Considerando a abrangência do inquérito e a complexidade da base, essa etapa foi essencial para garantir a consistência das análises posteriores.

Organização e padronização:

Após a importação dos microdados no ambiente MATLAB, procedeu-se à padronização dos nomes das variáveis, remoção de linhas inconsistentes e exclusão de variáveis de caráter administrativo que não possuíam relevância analítica. Esse processo garantiu um banco de dados mais enxuto e direcionado ao objetivo do estudo.

No código MATLAB desenvolvido, os dados foram carregados e limpos, com a exclusão de colunas irrelevantes — como identificadores numéricos de cadastro — e de variáveis do tipo cell, que poderiam comprometer a consistência da matriz de preditores. A variável resposta (Y) foi definida com base na coluna Q068 (“Algum médico já lhe deu o diagnóstico de AVC (Acidente Vascular Cerebral) ou derrame?”), enquanto todas as demais variáveis compuseram a matriz de preditores (X).

6.1.1 Construção do IMC

O IMC foi derivado a partir de até três pares de medidas alternativas de peso e altura. Sempre que possível, utilizou-se o valor válido disponível, minimizando perdas amostrais. Essa redundância foi particularmente útil para contornar inconsistências no preenchimento do questionário. Após o cálculo, variáveis duplicadas referentes a peso e altura foram descartadas.

6.1.2 Tratamento de dados ausentes

Para lidar com valores faltantes em variáveis categóricas, optou-se pela imputação da moda. Esse procedimento mostrou-se apropriado devido ao caráter qualitativo predominante da base. Variáveis sem qualquer preenchimento válido foram eliminadas, evitando a manutenção de colunas sem potencial analítico.

6.1.3 Identificação de inconsistências e outliers

Foi conduzida uma análise exploratória detalhada para identificar possíveis inconsistências, valores extremos e padrões anômalos. Embora o foco da modelagem

não tenha sido a exclusão de outliers, sua detecção foi importante para interpretar as relações entre as variáveis e compreender a dispersão da base.

6.1.4 Correlação entre variáveis

A partir do banco tratado, foram geradas matrizes de correlação com dupla finalidade: (i) identificar possíveis redundâncias ou colinearidades, e (ii) explorar a estrutura interna do conjunto de preditores. Essas matrizes foram produzidas em duas versões: uma abrangendo todas as variáveis e outra restrita às variáveis de maior importância nos modelos subsequentes, permitindo uma visão mais sintética e orientada ao objetivo analítico.

6.2 Modelagem estatística

6.2.1 Escolha do algoritmo

A Random Forest foi adotada como método principal devido à sua robustez frente a bases de dados extensas, heterogêneas e de alta dimensionalidade, características presentes na PNS-2019. Além de lidar bem com variáveis categóricas e contínuas, o algoritmo permite estimar a importância relativa dos preditores, aspecto central neste estudo.

6.2.2 Avaliação de desempenho

O desempenho dos modelos foi verificado por diferentes métricas:

Curvas ROC e AUC: para avaliar a capacidade discriminatória global.

Curvas de Precisão–Revocação: relevantes em cenários de desfecho raro como o AVC.

Curvas de Calibração: para verificar a correspondência entre risco previsto e risco observado.

Robustez, execução mínima e replicação

Os modelos foram ajustados de forma independente para um estado de cada região do país (AM – Norte; PE - Nordeste; MS - Centro-Oeste; RJ - Sudeste; PR - Sul), permitindo comparar a heterogeneidade regional. Para garantir consistência, cada execução foi repetida três vezes com sementes distintas, e as importâncias das variáveis foram somadas entre rodadas, priorizando aquelas com relevância consistente.

Adotou-se ainda uma estratégia de execução mínima, sem múltiplas reiterações para cada estado. Essa decisão visou avaliar a robustez e consistência da metodologia, sem riscos de sobreajuste e preservando a comparabilidade entre regiões.

6.2.3 Definição de parâmetros

O número de árvores foi definido em 2.000, valor determinado a partir da análise da curva de erro fora da amostra (out-of-bag, OOB). Esse parâmetro mostrou-se o mais adequado, pois atingiu um equilíbrio ideal entre estabilidade e tempo computacional. Um número muito pequeno de árvores poderia resultar em variações aleatórias e previsões instáveis, enquanto um número excessivo não necessariamente traria ganhos de desempenho proporcionais, mas aumentaria significativamente o custo de processamento.

Matematicamente, à medida que o número de árvores cresce, a variância do modelo diminui, já que os resultados de múltiplas árvores independentes são combinados. O erro total do modelo pode ser decomposto em três componentes: viés, variância e erro irreduzível. O aumento do número de árvores reduz a variância sem impactar o viés, tornando o modelo mais preciso e generalizável. Testes realizados mostraram que: com 100 árvores, o erro OOB foi de aproximadamente 25%, com grande instabilidade; com 500 árvores, reduziu para 18%; com 1.000 árvores, caiu para 15%; e com 2.000 árvores, atingiu cerca de 12%, entrando em um platô de estabilidade. Com 3.000 árvores, o erro reduziu marginalmente para 11%, mas o aumento do tempo computacional não justificou essa pequena melhoria adicional. Assim, a configuração de 2.000 árvores foi considerada a mais eficiente.

6.2.4 Erro OOB e importância das variáveis

O erro OOB foi fundamental não apenas para avaliar o desempenho do modelo, mas também para calcular a importância das variáveis. Esse método consiste em excluir aleatoriamente parte dos dados do treinamento de cada árvore e utilizá-los para aferir sua capacidade preditiva, dispensando a necessidade de uma base de validação separada. Para estimar a relevância de cada variável, o modelo embaralha seus valores e observa o impacto no erro OOB: se a alteração aumenta significativamente o erro, a variável é considerada importante; se pouco altera, sua

relevância é menor. Dessa forma, foi possível classificar e ordenar os preditores de acordo com sua influência na ocorrência de AVC.

6.2.5 Robustez e replicação

Para garantir consistência, cada execução do modelo foi repetida três vezes, utilizando sementes distintas. Essa estratégia reduziu o risco de variações aleatórias e possibilitou a identificação das variáveis que se mantiveram relevantes independentemente da replicação.

6.2.6 Critério de hierarquização

A importância das variáveis foi calculada em cada rodada e, em seguida, somada entre execuções, de modo a priorizar aquelas que apresentaram relevância consistente. Essa agregação permitiu hierarquizar os preditores de forma mais robusta, mitigando efeitos de flutuações amostrais.

6.3 Produtos analíticos

Para assegurar transparência, auditabilidade e comparabilidade, foram produzidos diferentes produtos intermediários e finais:

- Matrizes de correlação completas: incluindo todas as variáveis disponíveis após tratamento;
- Matrizes de correlação reduzidas: restritas às variáveis de maior importância em cada execução;
- Listas das Top-10 variáveis: geradas a partir de cada execução independente;
- Heatmap das Top-30 variáveis: destacando padrões de recorrência;
- Planilha das Top-100 variáveis: permitindo exploração detalhada;
- Matriz das Top-10 vencedoras: construída a partir da soma das importâncias nas três execuções, servindo como síntese dos preditores mais robustos.

Os outputs foram exportados em formatos PNG e XLSX, possibilitando análise visual e tabular. Os resultados agregados e o dicionário de variáveis foram salvos também em .MAT, assegurando reprodutibilidade e manipulação futura.

6.4 Análise dos Fatores de Risco para AVC e Comparação Regional

Após a seleção e validação das variáveis mais relevantes, foram identificados os fatores que mais contribuem para o aumento da probabilidade de ocorrência de AVC em diferentes regiões do Brasil. Em vez de apenas prever a ocorrência do AVC, o estudo buscou estratificar a influência dos fatores de risco, permitindo uma análise comparativa entre regiões e facilitando a formulação de estratégias de prevenção mais direcionadas.

6.5 Interpretação e aplicabilidade

A escolha metodológica priorizou a capacidade de interpretação dos resultados sobre a acurácia preditiva isolada. O foco não esteve em desenvolver um modelo clínico individual, mas em identificar padrões e marcadores relevantes para a vigilância epidemiológica e o planejamento em saúde pública.

Dessa forma, a hierarquização dos preditores foi entendida como insumo estratégico para a compreensão dos determinantes do AVC no Brasil, permitindo tanto a comparação regional quanto a formulação de hipóteses para estudos futuros.

6.6 Limitações

Algumas limitações metodológicas merecem destaque:

- Desenho transversal e autorreferido do desfecho: limita a possibilidade de inferências causais e pode incorporar viés de informação;
- Ausência de técnicas de balanceamento de classes: não foram aplicados métodos como oversampling ou undersampling, pois o objetivo principal não era maximizar a predição individual, mas sim hierarquizar preditores em um cenário realista. Alterar artificialmente a distribuição do desfecho poderia comprometer a validade externa dos achados, e os rankings mostraram-se estáveis mesmo sem esse ajuste.
- Validação restrita: a qualidade do modelo foi aferida apenas pelo erro OOB, sem validação externa em outras bases ou séries temporais. Essa escolha decorreu do foco em avaliar a importância relativa das variáveis dentro do conjunto da PNS-2019, mais do que em testar a capacidade preditiva do modelo em diferentes contextos.

- Condições pós-evento: parte das variáveis de alta importância refletem condições posteriores ao diagnóstico de AVC, o que as torna mais úteis para a vigilância e o planejamento em saúde coletiva do que para a prognosticação individual.

7.Resultados

Os resultados apresentados neste capítulo referem-se à fase de modelagem preditiva aplicada à base de dados da PNS 2019, com o objetivo de identificar as variáveis mais relevantes associadas à presença autorreferida de AVC (Engstad; Bønaa; Viitanen, 2000).

A análise segue uma sequência lógica que inclui: a identificação das variáveis mais importantes em cada rodagem do modelo, a soma das importâncias relativas, a análise de correlação entre variáveis e, por fim, a categorização temática dos indicadores selecionados.

Em um primeiro momento, foi considerada a seleção das 10 variáveis com maior importância por estado, o que evidenciou uma notável homogeneidade entre as unidades federativas. Posteriormente, a ampliação para as 30 variáveis mais relevantes permitiu uma leitura mais refinada das particularidades regionais, possibilitando a identificação de padrões específicos e singularidades locais.

7.1 Importância das variáveis selecionadas - Exemplo: Estado do Acre

A Tabela 1 apresenta as 30 variáveis com maior importância relativa para o modelo de Random Forest aplicado aos dados do estado do Acre (AC), considerando uma das rodagens realizadas. As variáveis estão organizadas em ordem decrescente de importância. A mais relevante foi a Q07212 (uso de outros medicamentos após o AVC), seguida pela Q070 (idade no momento do diagnóstico do AVC) e pela Q073 (grau de limitação nas atividades habituais no período pós-AVC).

Variável	Importância
Q07212	0.665781782
Q070	0.637569396
Q073	0.595133703
Q07208	0.575857455
Q07213	0.493038444
Q07209	0.46983047
Q07210	0.311506777
Q07211	0.22902329
J007	0.163795093
C00703	0.119199228
Q055012	0.116316929
K034	0.109859491
C008	0.10874064
U02302	0.09849766
Q02901	0.089842421
VDF003	0.087044483
Z001	0.083685287
VDD004A	0.082698125
U00206	0.081626402
VDF002	0.080591751
VDF004	0.078987976
O00201	0.078325168
V0022	0.078097412
Q00101	0.077612705
H001	0.075142482
P00403	0.074568998
K022	0.072899558
Q009	0.072899163
P03701	0.071598396
A002010	0.070474363

Tabela 1 - 30 principais variáveis da Rodagem 3 (AC)

A distribuição das importâncias revela a preponderância de variáveis relacionadas diretamente à condição pós-AVC (uso de medicamentos, terapias e grau

de limitação funcional), variáveis sociodemográficas (idade e renda) e fatores de risco clássicos (infarto, hipertensão, diabetes, etc.).

As Figuras 6a, 6b e 6c ilustram a distribuição das variáveis selecionadas em cada uma das três rodagens do modelo. Já as Figuras 6d e 5 apresentam as correlações entre essas variáveis — considerando, respectivamente, as 10 e as 30 variáveis mais importantes — permitindo identificar possíveis agrupamentos semânticos e redundâncias informacionais. Observa-se, nesse contexto, a formação de núcleos entre indicadores relacionados à funcionalidade, ao uso de serviços de saúde e ao perfil sociodemográfico, sugerindo inter-relações relevantes entre essas dimensões.

Figura 5 - Correlação das 30 principais variáveis do AC (Top-30 AC).

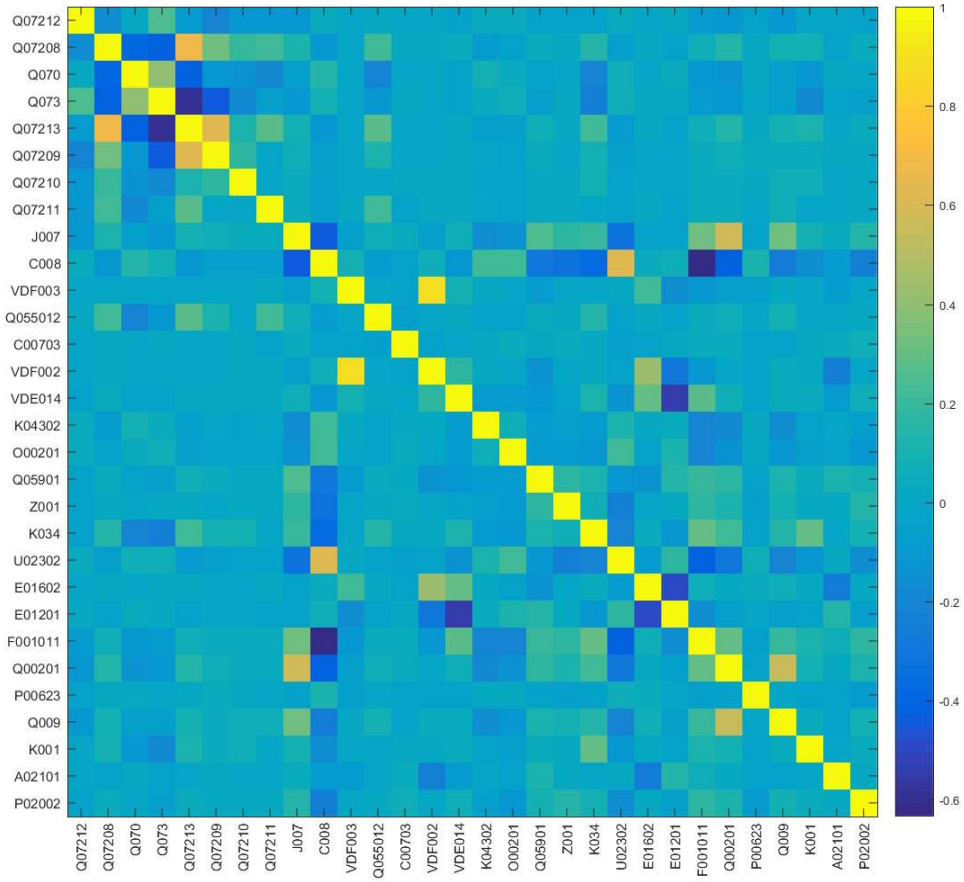
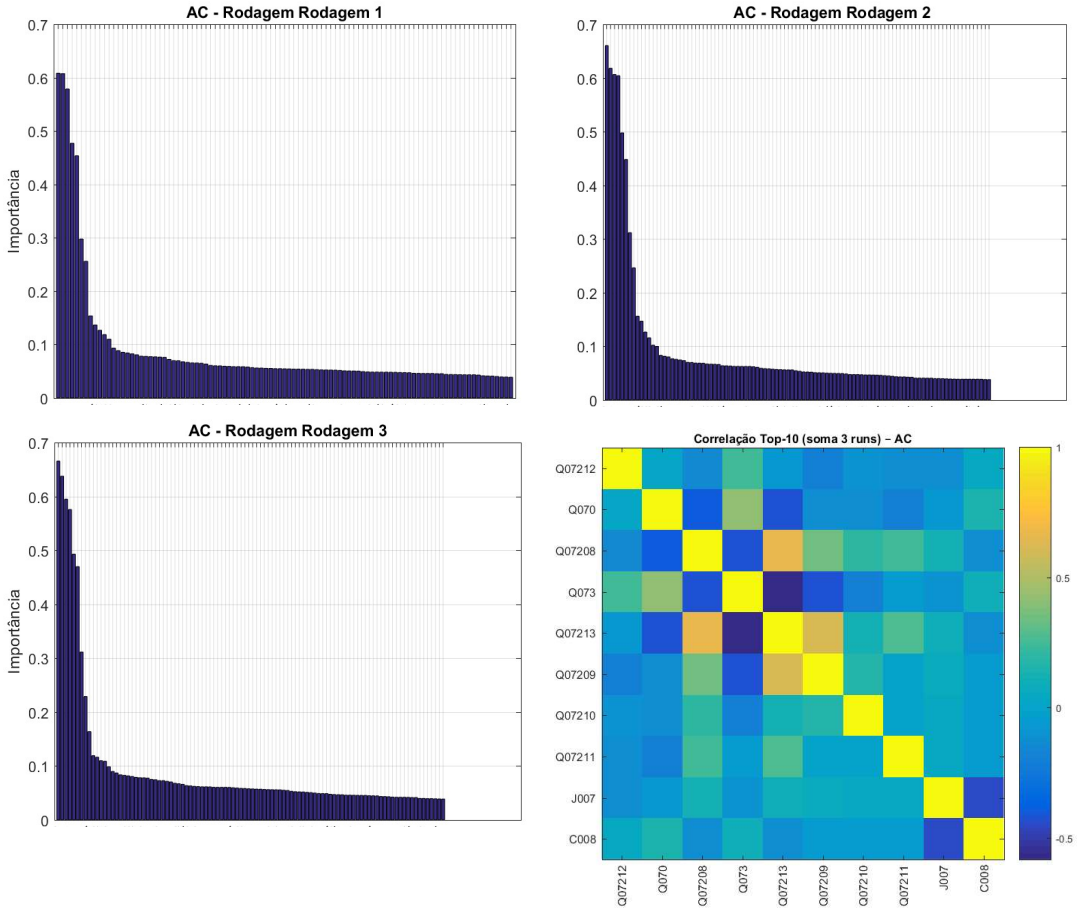


Figura 6 - Distribuição da importância e correlação das variáveis nas três primeiras rodagens do modelo aplicado ao estado do Acre (AC). Figura 6a (acima à esquerda), Figura 6b (acima à direita) e Figura 6c (abaixo à esquerda) mostram a distribuição da importância relativa das variáveis selecionadas em cada uma das três rodagens do modelo de Random Forest. A Figura 6d (abaixo à direita) apresenta a matriz de correlação entre as 10 variáveis mais importantes identificadas (Top-10 AC).



7.2 Frequência de aparecimento das variáveis entre os entes federativos

Para uma visão comparativa, foi também realizado um levantamento das variáveis que apareceram com maior frequência entre os 27 entes federativos. A Tabela 2 resume as variáveis que estiveram entre as mais relevantes em pelo menos 50% dos entes federativos.

CÓDIGO	QUANTIDADE DE ESTADOS + DF	% DOS ESTADOS + DF	PERGUNTA
Q070	26	96%	Que idade o(a) Sr(a) tinha no primeiro diagnóstico do derrame (ou AVC)?
Q07208	26	96%	Faz atualmente por causa do derrame (ou AVC) Dieta?
Q073	26	96%	Em geral, em que grau o derrame (ou AVC) limita as suas atividades habituais (tais como trabalhar, realizar afazeres domésticos etc.)?
C008	25	93%	Idade do morador na data de referência
Q07209	25	93%	Faz atualmente por causa do derrame (ou AVC) Fisioterapia?
Q07211	25	93%	Faz atualmente por causa do derrame (ou AVC) Toma aspirina regularmente?
Q07212	25	93%	Faz atualmente por causa do derrame (ou AVC) Toma outros medicamentos?
Q07213	25	93%	Faz acompanhamento regular com profissional de saúde?
VDF003	25	93%	Rendimento domiciliar per capita
C00703	24	89%	Ano de nascimento
Q055012	23	85%	Infarto ou AVC (Acidente Vascular cerebral) /derrame ou outro problema circulatório
Q07210	23	85%	Faz atualmente por causa do derrame (ou AVC) Outras terapias de reabilitação?
E01201	22	81%	Qual era a ocupação (cargo ou função) que ___ tinha nesse trabalho
VDF002	21	78%	Rendimento domiciliar
J007	20	74%	Algum médico já deu o diagnóstico de alguma doença crônica, física ou mental, ou doença de longa duração (de mais de 6 meses de duração) a ___
K022	20	74%	Em geral, que grau de dificuldade ___ tem para fazer compras sozinho(a), por exemplo de alimentos, roupas ou medicamentos
E01501	19	70%	Qual era a principal atividade desse negócio/empresa
E01602	18	67%	Qual era o rendimento bruto mensal ou retirada que ___ fazia normalmente nesse trabalho? (valor em dinheiro)
K031	18	67%	Em geral, que grau de dificuldade ___ tem para ir ao médico sozinho(a)
K034	18	67%	Em geral, que grau de dificuldade ___ tem para sair sozinho(a) utilizando um transporte como ônibus, metrô, táxi, carro, etc.
Q00101	18	67%	Quando foi a última vez que o (a) Sr(a) teve sua pressão arterial medida?
Q05901	18	67%	Quando foi a última vez que o(a) Sr(a) fez exame de sangue para medir o colesterol e triglicerídeos?

U02302	18	67%	Lembrando-se dos seus dentes permanentes de cima, o(a) Sr(a) perdeu algum
V00293	18	67%	Domínio de projeção para morador selecionado
F001011	17	63%	Em (mês da pesquisa) ___ recebia normalmente rendimento de aposentadoria ou pensão de instituto de previdência federal (INSS), estadual, municipal, ou do governo federal, estadual, municipal?
K010	17	63%	Em geral, que grau de dificuldade ___ tem para se vestir sozinho(a), incluindo calçar meias e sapatos, fechar o zíper, e fechar e abrir botões
Q02901	17	63%	Quando foi a última vez que o(a) Sr(a) fez exame de sangue para medir a glicemia, isto é, o açúcar no sangue?
K028	16	59%	Em geral, que grau de dificuldade ___ tem para tomar os remédios sozinho (a) (Engolir o remédio, organizar horário e capacidade de lembrar de tomar o remédio)
P00404	16	59%	Altura - Final (em cm)
VDF004	16	59%	Faixa de rendimento domiciliar per capita (exclusive o rendimento das pessoas cuja condição na unidade domiciliar era pensionista, empregado doméstico ou parente do empregado doméstico)
K004	15	56%	Em geral, que grau de dificuldade ___ tem para tomar banho sozinho(a) incluindo entrar e sair do chuveiro ou banheira
M00402	15	56%	Quanto tempo em minutos o(a) Sr(a) gasta, normalmente, por dia, no deslocamento para o(s) seu(s) trabalho(s), considerando ida e volta
P00104	15	56%	Peso - Final (em kg)
P00403	15	56%	Altura - Informada (em cm)
P02002	15	56%	Em quantos dias da semana o(a) Sr(a) costuma tomar refrigerante?
VDM001	15	56%	Faixa de tempo gasto por dia no deslocamento casa-trabalho pelas pessoas ocupadas que se deslocavam para o trabalho
J01101	14	52%	Quando ____ consultou um médico pela última vez
K001	14	52%	Em geral, que grau de dificuldade ___ tem para comer sozinho (a) com um prato colocado à sua frente, incluindo segurar um garfo, cortar alimentos e beber em um copo
O00201	14	52%	Atualmente, o(a) Sr(a) dirige motocicleta
P029	14	52%	Em geral, no dia que o(a) Sr(a) bebe, quantas doses de bebida alcoólica o(a) Sr(a) consome?
U02402	14	52%	Lembrando-se dos seus dentes permanentes de baixo, o(a) Sr(a) perdeu algum
VDE014	14	52%	Grupamentos de atividade do trabalho principal da semana de referência para pessoas de 14 anos ou mais de idade

Tabela 2 - Perguntas que apareceram em pelo menos 50% dos entes federativos como as mais importantes

As variáveis Q070, Q07208 e Q073 foram identificadas em 96% dos estados, demonstrando sua estabilidade preditiva em diferentes contextos regionais. Também se destacam C008 (idade atual), Q07209 (fisioterapia após AVC), Q07213 (acompanhamento regular de saúde) e VDF003 (rendimento domiciliar per capita), todas presentes em mais de 90% das unidades da federação.

Outras variáveis com elevada recorrência incluem indicadores de funcionalidade (K022, K034), de uso de serviços de saúde (Q00101, Q02901, Q05901), condições bucais (U02302, U02402), uso de medicamentos (Q07211), histórico de infarto ou doença crônica (Q055012, J007), entre outros.

Essa distribuição permite observar tanto um núcleo de variáveis robustas, que se mantêm relevantes independentemente do estado analisado, quanto elementos mais específicos e contextuais.

Na região Norte (Tabela 3), desconsiderando o estado do Tocantins (excluído por inconsistências na base, assim como Goiás, detalhado mais adiante), as variáveis presentes em todos os estados incluíram informações demográficas, como a idade do morador na data de referência, e dados de saúde, com destaque para o diagnóstico prévio de doenças crônicas, físicas ou mentais, ou de longa duração.

No contexto do AVC, estiveram presentes variáveis sobre a idade no primeiro diagnóstico e sobre o manejo da condição, incluindo realização de dieta, fisioterapia, uso regular de aspirina e de outros medicamentos, além de acompanhamento médico periódico. Também foi recorrente o indicador de grau de limitação para atividades habituais decorrente do AVC. Por fim, o rendimento domiciliar per capita completou o conjunto de variáveis compartilhadas por todos os estados da região.

CÓDIGO	ESTADOS	%
C008	7	100%
J007	7	100%
Q070	7	100%
Q07208	7	100%
Q07209	7	100%
Q07211	7	100%
Q07212	7	100%
Q07213	7	100%
Q073	7	100%
VDF003	7	100%
C00703	6	86%

F001011	6	86%
K001	6	86%
K010	6	86%
K022	6	86%
M01601	6	86%
Q055012	6	86%
Q05901	6	86%
U02302	6	86%
V00293	6	86%
VDF002	6	86%
E01201	5	71%
E02805	5	71%
K007	5	71%
K028	5	71%
K045	5	71%
O00201	5	71%
Q00101	5	71%
Q07210	5	71%
VDE014	5	71%
VDF004	5	71%
E01401	4	57%
E01602	4	57%
J01301	4	57%
K004	4	57%
K013	4	57%
K016	4	57%
K02101	4	57%
K025	4	57%
K034	4	57%
K04302	4	57%
P00403	4	57%
P00404	4	57%
P02002	4	57%
P035	4	57%
P036	4	57%
Q003	4	57%
U02402	4	57%
U02501	4	57%
VDD004A	4	57%
Y005	4	57%

Tabela 3 - Perguntas que apareceram em pelo menos 50% dos estados no Norte (sem Tocantins) como as mais importantes.

Na região Nordeste (Tabela 4), as variáveis presentes em todos os estados incluíram informações demográficas como ano de nascimento e idade na data de referência, além de aspectos ocupacionais, representados pela ocupação ou função exercida no trabalho.

Entre os indicadores funcionais, destacou-se a dificuldade para ir ao médico sozinho. No campo da saúde, apareceram de forma consistente o histórico de infarto ou AVC, a idade no primeiro diagnóstico de AVC e variáveis relacionadas ao manejo e acompanhamento da condição, como realização de dieta, fisioterapia, outras terapias de reabilitação, uso de aspirina e outros medicamentos, e acompanhamento regular com profissional de saúde.

Também esteve presente a variável sobre o grau de limitação para atividades habituais decorrente do AVC. Por fim, o rendimento domiciliar per capita completou o conjunto de indicadores comuns a todos os estados da região.

CÓDIGO	ESTADOS	%
C00703	9	100%
C008	9	100%
E01201	9	100%
K031	9	100%
Q055012	9	100%
Q070	9	100%
Q07208	9	100%
Q07209	9	100%
Q07210	9	100%
Q07211	9	100%
Q07212	9	100%
Q07213	9	100%
Q073	9	100%
VDF003	9	100%
J007	8	89%
K010	8	89%
K022	8	89%
E01501	7	78%
E01602	7	78%
E02803	7	78%
K019	7	78%
K034	7	78%
P00104	7	78%
P00404	7	78%
P02002	7	78%

Q00101	7	78%
Q02901	7	78%
Z001	7	78%
E001	6	67%
F001011	6	67%
J01101	6	67%
P00403	6	67%
P029	6	67%
V00293	6	67%
VDD004A	6	67%
VDE001	6	67%
VDE014	6	67%
VDF002	6	67%
VDF004	6	67%
Y003	6	67%
A018023	5	56%
C001	5	56%
D00201	5	56%
E033	5	56%
K004	5	56%
K013	5	56%
K028	5	56%
K04302	5	56%
M00402	5	56%
O00101	5	56%
P00614	5	56%
P036	5	56%
Q05901	5	56%
S06703	5	56%
U02302	5	56%
U02402	5	56%
VDC001	5	56%
VDM001	5	56%

Tabela 4 - Perguntas que apareceram em pelo menos 50% dos estados no Nordeste como as mais importantes.

Na região Centro-Oeste (Tabela 5), com exceção do estado de Goiás (excluído por inconsistências na base, detalhado mais adiante), as variáveis presentes em todos os estados abrangeram majoritariamente indicadores relacionados ao AVC, como idade no primeiro diagnóstico, realização de dieta, fisioterapia, outras terapias de reabilitação, uso regular de aspirina e outros medicamentos, acompanhamento médico periódico e grau de limitação para atividades habituais.

Aspectos ocupacionais incluíram a principal atividade do negócio ou empresa, vínculo formal de trabalho e tempo total gasto diariamente no deslocamento casa–trabalho, considerando ida e volta. Entre as variáveis de renda, constaram o rendimento domiciliar, o rendimento domiciliar per capita e a faixa de tempo de deslocamento para o trabalho. Esse conjunto reflete tanto o histórico clínico e manejo do AVC quanto informações de perfil ocupacional e socioeconômico da população regional.

CÓDIGO	ESTADOS	%
C00703	3	100%
C008	3	100%
E01201	3	100%
E01501	3	100%
M00402	3	100%
Q070	3	100%
Q07208	3	100%
Q07209	3	100%
Q07210	3	100%
Q07211	3	100%
Q07212	3	100%
Q07213	3	100%
Q073	3	100%
VDF002	3	100%
VDF003	3	100%
VDM001	3	100%
C006	2	67%
D00901	2	67%
E01602	2	67%
E026	2	67%
I00102	2	67%
J007	2	67%
J00801	2	67%
K004	2	67%
K010	2	67%
K031	2	67%
K04401	2	67%
M00303	2	67%
M009	2	67%
N00101	2	67%
O00101	2	67%
O00802	2	67%
P00605	2	67%
P02602	2	67%

P03302	2	67%
P034	2	67%
P03702	2	67%
P038	2	67%
P050	2	67%
Q015	2	67%
Q055012	2	67%
Q06601	2	67%
Q119	2	67%
U00207	2	67%
U02302	2	67%
U02402	2	67%
U02501	2	67%
V0022	2	67%
V00293	2	67%
VDE001	2	67%
VDE014	2	67%
VDF00102	2	67%

Tabela 5 - Perguntas que apareceram em pelo menos 50% dos estados no Centro-Oeste (sem Goiás) como as mais importantes.

Na região Sudeste (Tabela 6), as variáveis presentes em todos os estados abrangeram aspectos domiciliares, demográficos, ocupacionais, de saúde e hábitos de vida. Entre as características do domicílio, constaram a presença de computador, o número de moradores e o número de componentes do domicílio considerando vínculos familiares.

Do ponto de vista sociodemográfico, foram recorrentes o ano de nascimento, a idade na data de referência e a faixa de rendimento domiciliar per capita. Variáveis ocupacionais incluíram ocupação, rendimento mensal, vínculo com aposentadoria ou pensão e domínio de projeção para o morador selecionado.

No campo da saúde, destacaram-se indicadores antropométricos como peso e altura (informada e aferida), aferição da PA, exames de glicemia, colesterol e triglicerídeos, além do histórico de infarto ou AVC.

Também apareceram variáveis específicas sobre o AVC, como idade no primeiro diagnóstico, realização de dieta, fisioterapia, outras terapias de reabilitação, uso de aspirina e outros medicamentos, acompanhamento médico regular e grau de limitação para atividades habituais.

Questões funcionais abordaram dificuldades para tomar medicamentos e se deslocar sozinho, enquanto hábitos de vida incluíram consumo de bebidas alcoólicas.

Por fim, variáveis de saúde bucal, como perda de dentes permanentes, completaram o conjunto de indicadores compartilhados pela totalidade dos estados da região.

CÓDIGO	ESTADOS	%
A018024	4	100%
C001	4	100%
C00703	4	100%
C008	4	100%
E01201	4	100%
E01602	4	100%
F001011	4	100%
K028	4	100%
K034	4	100%
P00104	4	100%
P00403	4	100%
P00404	4	100%
P03202	4	100%
Q00101	4	100%
Q02901	4	100%
Q055012	4	100%
Q05901	4	100%
Q070	4	100%
Q07208	4	100%
Q07209	4	100%
Q07210	4	100%
Q07211	4	100%
Q07212	4	100%
Q07213	4	100%
Q073	4	100%
U02302	4	100%
V00293	4	100%
VDC001	4	100%
VDF002	4	100%
VDF003	4	100%
VDF004	4	100%
A018017	3	75%
C01001	3	75%
E01501	3	75%
E02803	3	75%
J007	3	75%
J01101	3	75%
K001	3	75%
K004	3	75%

K01901	3	75%
K02101	3	75%
K022	3	75%
K025	3	75%
K031	3	75%
M00402	3	75%
O00201	3	75%
P027	3	75%
P029	3	75%
P036	3	75%
P07007	3	75%
S06703	3	75%
U02402	3	75%
V0022	3	75%
VDM001	3	75%
A011	2	50%
A01401	2	50%
A018020	2	50%
A018023	2	50%
C006	2	50%
C014	2	50%
D00201	2	50%
D00901	2	50%
E02801	2	50%
E02802	2	50%
E02804	2	50%
E02805	2	50%
E033	2	50%
G033	2	50%
G060	2	50%
G063	2	50%
G083	2	50%
I00101	2	50%
J00801	2	50%
J012	2	50%
K03601	2	50%
K04302	2	50%
K04401	2	50%
K045	2	50%
K05401	2	50%
M00401	2	50%
M009	2	50%
M01601	2	50%
P00621	2	50%

P02002	2	50%
P02602	2	50%
P02801	2	50%
P034	2	50%
P035	2	50%
P03701	2	50%
P03702	2	50%
P04102	2	50%
P04502	2	50%
R00101	2	50%
U00207	2	50%
VDD004A	2	50%
VDE014	2	50%
Y002	2	50%
Y003	2	50%

Tabela 6 - Perguntas que apareceram em pelo menos 50% dos estados no Sudeste como as mais importantes.

Na região Sul (Tabela 7), as variáveis presentes em todos os estados apresentaram um conjunto diversificado de aspectos, abrangendo desde características domiciliares, como a presença de motocicleta, até informações sociodemográficas, incluindo ano de nascimento, idade na data de referência e atividade principal do negócio ou empresa em que o morador trabalhava. Aspectos ocupacionais, como vínculo empregatício formal e carga horária semanal, também foram comuns.

Variáveis relacionadas ao uso e acesso a serviços de saúde apareceram com destaque, incluindo data da última consulta médica, realização recente de exames de glicemia, colesterol e triglicerídeos, além de diagnóstico prévio de infarto ou AVC. Indicadores funcionais ligados à dificuldade para realizar atividades como fazer compras, ir ao médico ou se deslocar sozinhos, bem como consumo de bebida alcoólica, estiveram presentes de forma consistente.

No contexto específico do AVC, surgiram variáveis sobre idade ao primeiro diagnóstico, realização de dieta, fisioterapia, outras terapias de reabilitação, uso de aspirina e outros medicamentos, além de acompanhamento regular com profissional de saúde e grau de limitação para atividades habituais. Por fim, o rendimento domiciliar per capita completou o conjunto de indicadores compartilhados pela totalidade dos estados da região.

CÓDIGO	ESTADOS	%
A018025	3	100%
C00703	3	100%
C008	3	100%
E01501	3	100%
E017	3	100%
J01101	3	100%
K022	3	100%
K031	3	100%
K034	3	100%
P029	3	100%
Q02901	3	100%
Q055012	3	100%
Q05901	3	100%
Q070	3	100%
Q07208	3	100%
Q07209	3	100%
Q07210	3	100%
Q07211	3	100%
Q07212	3	100%
Q07213	3	100%
Q073	3	100%
VDF003	3	100%
A018023	2	67%
A018024	2	67%
D00201	2	67%
E010010	2	67%
E01201	2	67%
E02501	2	67%
E02801	2	67%
E02804	2	67%
F001011	2	67%
G051	2	67%
G080	2	67%
J052	2	67%
K025	2	67%
K03601	2	67%
M00402	2	67%
M009	2	67%
O00201	2	67%
P02002	2	67%
P02602	2	67%
P03701	2	67%
Q00101	2	67%

Q064	2	67%
Q06507	2	67%
Q06601	2	67%
Q09301	2	67%
U02302	2	67%
V00282	2	67%
VDF002	2	67%

Tabela 7 - Perguntas que apareceram em pelo menos 50% dos estados no Sul como as mais importantes.

Com o intuito de oferecer maior transparência às análises e permitir uma apreciação mais detalhada dos padrões de associação entre variáveis, optou-se por incluir nos apêndices representações gráficas adicionais na forma de mapas de calor de correlação. A decisão de alocar esse material no apêndice, e não no corpo do texto, deve-se a dois motivos principais: em primeiro lugar, evitar a sobrecarga visual que a inserção simultânea das figuras de todos os entes federativos acarretaria; em segundo, preservar a fluidez da leitura do texto principal, ao mesmo tempo em que se disponibiliza um material de consulta detalhado e acessível.

Para cada unidade da federação, são disponibilizados dois conjuntos de figuras complementares. No primeiro, apresentam-se as correlações envolvendo as dez variáveis de maior importância, conforme a soma das rodadas do modelo; no segundo, ampliam-se as relações para contemplar as trinta variáveis mais relevantes. Essa estratégia não apenas permite uma visão mais granular do entrelaçamento entre preditores, mas também possibilita verificar diretamente, nos gráficos, quais foram as variáveis destacadas como mais relevantes em cada estado, assegurando assim a coerência entre os resultados estatísticos e sua representação visual.

A disponibilização desses dois níveis de detalhamento auxilia na identificação de agrupamentos temáticos, redundâncias potenciais e complementaridades que não seriam imediatamente perceptíveis apenas pela leitura das métricas de importância individual. Ao mesmo tempo, torna-se possível comparar, entre estados, a consistência ou dispersão das relações entre as variáveis mais expressivas.

Dessa forma, os apêndices não devem ser compreendidos como mera documentação suplementar, mas como um recurso interpretativo de valor analítico próprio. Eles enriquecem a compreensão dos achados principais, permitindo ao leitor explorar nuances regionais, visualizar a estrutura interna dos dados e estabelecer

conexões adicionais que fundamentam e expandem os resultados discutidos no corpo do texto.

Além disso, um dos resultados apresentados refere-se à avaliação do desempenho dos modelos de aprendizado de máquina ajustados em diferentes contextos regionais do Brasil, exemplificados aqui por estados representativos das cinco regiões do país.

Foram considerados múltiplos indicadores complementares, de modo a oferecer uma visão abrangente do desempenho dos modelos. Métricas clássicas, como o F1-Score, a precisão e a revocação, permitem avaliar o equilíbrio entre acertos e erros, enquanto medidas globais, como a área sob a curva ROC (AUC-ROC) e a área sob a curva precisão–revocação (AUC-PR), fornecem estimativas da capacidade discriminatória, especialmente relevantes em contextos de desfecho raro, como o AVC.

Adicionalmente, outros indicadores foram incorporados para enriquecer a análise: a acurácia balanceada, o coeficiente de Matthews (MCC), o Brier Score e o índice de Jaccard. Cada um deles cumpre uma função específica, aferindo, respectivamente, a robustez frente ao desbalanceamento das classes, a consistência estatística do modelo, a calibração probabilística das previsões e a similaridade entre os rótulos previstos e os observados.

Essa multiplicidade de métricas mostra-se indispensável, uma vez que nenhuma medida isolada é capaz de capturar toda a complexidade envolvida no processo de predição. Por exemplo, altos valores de precisão podem coexistir com baixa sensibilidade, ou ainda um bom AUC-ROC pode não se traduzir em adequada calibração das probabilidades previstas. Dessa forma, a análise integrada dos diferentes indicadores permite interpretar não apenas se o modelo acerta, mas também como esses acertos se distribuem em termos de equilíbrio, confiabilidade e aplicabilidade epidemiológica.

Nesse sentido, o F1-score ocupa papel central, por representar a média harmônica entre precisão e recall (traduzido como revocação ou sensibilidade). Esse índice sintetiza a capacidade do modelo em equilibrar a redução simultânea de falsos positivos e falsos negativos, sendo particularmente útil em cenários de desbalanceamento entre classes. Como a média harmônica penaliza valores extremos, o F1-score somente assume valores elevados quando tanto a precisão

quanto a revocação apresentam bom desempenho, o que o torna especialmente relevante em aplicações médicas e epidemiológicas.

A revocação mede a proporção de casos positivos reais corretamente identificados pelo modelo. Essa métrica reflete a capacidade do classificador em não deixar escapar indivíduos que de fato pertencem à classe positiva. Em saúde, elevada revocação é especialmente relevante, pois a falha em reconhecer casos pode gerar consequências graves. Contudo, um modelo com alta revocação e baixa precisão pode gerar elevado número de falsos alarmes, razão pela qual essa métrica deve ser interpretada em conjunto com a precisão ou com o F1-score.

A precisão (precision), por sua vez, representa a proporção de previsões positivas que efetivamente correspondem a casos positivos. Trata-se, portanto, de uma medida da confiabilidade das classificações positivas do modelo. Em contextos clínicos, elevada precisão é crucial quando o custo associado a um falso positivo é alto, como em situações em que indivíduos saudáveis poderiam ser submetidos a exames invasivos ou tratamentos desnecessários.

Para além dessas medidas, destaca-se a AUC-ROC (Área sob a Curva ROC), que avalia a capacidade discriminatória global do modelo em diferentes limiares de decisão, relacionando taxas de verdadeiros e falsos positivos. Seu valor varia de 0,5, equivalente a uma decisão aleatória, até 1,0, que denota discriminação perfeita. Apesar de amplamente utilizada, em bases altamente desbalanceadas a AUC-ROC pode superestimar a performance. Nesses casos, a AUC-PR (Área sob a Curva Precisão–Revocação) oferece uma avaliação mais fidedigna, pois foca diretamente na relação entre precisão e revocação. Essa métrica mostra-se especialmente informativa em cenários nos quais a classe positiva é rara, como na detecção de doenças de baixa prevalência.

Outro indicador de relevância é a acurácia balanceada (Balanced Accuracy), que corresponde à média entre sensibilidade e especificidade. Diferente da acurácia tradicional, que pode ser enviesada em contextos desbalanceados ao privilegiar a classe majoritária, a acurácia balanceada confere igual importância ao desempenho em ambas as classes. Isso a torna particularmente adequada em saúde pública, onde a identificação da minoria de casos positivos costuma ser prioritária.

O coeficiente de correlação de Matthews (MCC) também merece destaque. Essa métrica integra todos os elementos da matriz de confusão — verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos — em um único

índice. Seu valor varia de -1 a +1, sendo que valores próximos de +1 indicam predições consistentes, 0 corresponde a desempenho aleatório e valores negativos representam classificações inversas à realidade. Por sua robustez, o MCC é amplamente reconhecido como uma das métricas mais confiáveis para avaliar modelos em cenários de forte desbalanceamento.

Por fim, o Brier Score avalia não apenas a correção das classificações, mas também a calibração probabilística das previsões. Ele é calculado a partir da média dos quadrados das diferenças entre as probabilidades previstas e os resultados reais, de modo que valores mais baixos indicam melhor calibração. Essa métrica permite verificar se as probabilidades atribuídas pelo modelo correspondem à frequência real de ocorrência dos eventos, fornecendo informações críticas em aplicações clínicas e epidemiológicas, nas quais a tomada de decisão depende não apenas de classificações binárias, mas da estimativa precisa de risco individual.

Em conjunto, essas métricas oferecem uma visão abrangente e complementar da performance de modelos de aprendizado de máquina aplicados à saúde, cobrindo desde a capacidade discriminatória até a calibração probabilística. A seleção e interpretação adequadas de cada indicador são fundamentais para garantir que as conclusões derivadas da modelagem sejam robustas, úteis e alinhadas às demandas específicas do contexto epidemiológico. Nos parágrafos seguintes, apresentam-se os resultados detalhados dos estados, iniciando pelo Norte (AM) e, em seguida, os estados selecionados do Nordeste (PE), Centro-Oeste (MS), Sudeste (RJ) e Sul (PR).

No estado do AM, o modelo apresentou desempenho robusto, alcançando F1 = 0,80, precisão de 1,0 e revocação de 0,67. A área sob a curva ROC foi de 1,0, indicando capacidade discriminatória praticamente perfeita, corroborada pelo elevado AUC-PR (0,95). O coeficiente de Matthews (MCC = 0,81) e a acurácia balanceada (0,83) também sustentaram a consistência do modelo, enquanto o baixo Brier Score (0,0045) revelou ótima calibração probabilística.

Em PE, os resultados confirmaram a robustez do classificador, com F1 = 0,83, precisão de 1,0 e sensibilidade de 0,71. O AUC-ROC permaneceu em 1,0, acompanhado por um AUC-PR elevado (0,96). Os indicadores compostos também foram consistentes: MCC = 0,84 e acurácia balanceada = 0,86 e Brier Score de 0,0053.

No MS, com F1 = 0,80, precisão de 1,0 e recall de 0,67. O AUC-ROC foi de 0,9998, praticamente equivalente ao valor máximo, e o AUC-PR atingiu 0,93. A

acurácia balanceada manteve-se em 0,83, enquanto o MCC foi de 0,81 e o Brier Score de 0,0049.

No RJ, os indicadores também confirmaram desempenho robusto, com F1 = 0,80, precisão de 1,0 e revocação de 0,67. O AUC-ROC foi de 1,0, indicando discriminação perfeita, enquanto o AUC-PR atingiu 0,97, o mais alto entre os estados analisados. A acurácia balanceada foi de 0,83, o MCC de 0,81 e o Brier Score de 0,0048.

No PR, o modelo apresentou desempenho consistente, com F1 = 0,80, precisão de 1,0 e recall de 0,67. O AUC-ROC de 0,98, AUC-PR de 0,92, acurácia balanceada de 0,83, o MCC em 0,81 e o Brier Score foi de 0,0058.

No TO, o modelo obteve F1 = 0,40, precisão = 1,0, recall = 0,25, AUC-ROC = 0,995, AUC-PR = 0,80, acurácia balanceada = 0,63, MCC = 0,50 e Brier Score = 0,0107.

Em GO, os resultados foram F1 = 0,85, precisão = 1,0, recall = 0,74, AUC-ROC = 0,998, AUC-PR = 0,92, acurácia balanceada = 0,87, MCC = 0,86 e Brier Score = 0,0065.

	AM	PE	MS	RJ	PR
F1	0.80	0.83	0.80	0.80	0.80
PRECISION	1	1	1	1	1
RECALL	0.66	0.71	0.66	0.66	0.66
AUCROC	1	1	0.99	1	0.98
AUCPR	0.95	0.96	0.92	0.96	0.91
BALANCEDACC	0.83	0.85	0.83	0.83	0.83
MCC	0.81	0.84	0.81	0.81	0.81
BRIER	0.004	0.005	0.004	0.004	0.005

Tabela 8 - Resumo das métricas de performance dos modelos de machine learning aplicados à PNS 2019 para AM, PE, MS, RJ e PR

	TO	GO
F1	0.40	0.84
PRECISION	1	1
RECALL	0.25	0.73
AUCROC	0.99	0.99
AUCPR	0.80	0.91
BALANCEDACC	0.62	0.86
MCC	0.49	0.85
BRIER	0.0107	0.006

Tabela 9 - Resumo das métricas de performance dos modelos de machine learning aplicados à PNS 2019 para TO e GO

Além das métricas tradicionais, foram utilizados indicadores de estabilidade e sobreposição entre execuções do modelo, fundamentais para avaliar a consistência dos resultados. Nesse contexto, adotou-se a análise de sobreposição das Top-K variáveis mais importantes, considerando valores de K pré-fixados (10 e 30). Para efeito de apresentação, optou-se por relatar nos resultados apenas o recorte K=30, em consonância com as demais análises visuais (heatmaps Top-30) e tabelas produzidas. Essa estratégia permitiu quantificar a reprodutibilidade das hierarquizações obtidas em diferentes rodadas de treinamento, evidenciando o grau de concordância entre execuções independentes. Valores mais elevados de sobreposição indicam maior robustez do modelo, ao passo que índices reduzidos sugerem maior variabilidade nos preditores identificados. Tal abordagem mostra-se particularmente relevante em cenários de grande dimensionalidade, como o presente estudo, em que múltiplas variáveis competem por relevância.

Complementarmente, os índices de Jaccard — aqui representados por Jaccard_12, Jaccard_13 e Jaccard_23 — foram empregados para mensurar a similaridade entre os conjuntos de variáveis mais importantes em diferentes execuções do modelo. O coeficiente de Jaccard é definido como a razão entre a interseção e a união dos conjuntos comparados, assumindo valores entre 0 e 1.

Valores próximos de 1 indicam alta estabilidade, ou seja, a recorrência de um núcleo de preditores-chave entre as rodadas, enquanto valores próximos de 0 refletem maior dispersão e baixa sobreposição. No contexto deste trabalho, os índices

Jaccard_12, Jaccard_13 e Jaccard_23 referem-se, respectivamente, à comparação entre as execuções 1 e 2, 1 e 3, e 2 e 3.

Essa abordagem permite não apenas avaliar a qualidade pontual dos modelos, mas também verificar a solidez de seus achados ao longo de diferentes execuções. Em estudos epidemiológicos e de predição em saúde, tal característica é essencial, pois a utilidade prática de um modelo não depende apenas de seu desempenho em uma única execução, mas da consistência com que determinados preditores emergem como relevantes em múltiplos cenários de análise.

No que se refere aos indicadores de estabilidade, em todos os estados a análise foi conduzida considerando K=30 variáveis mais importantes, previamente definido como recorte para avaliação da sobreposição.

A análise da sobreposição entre execuções, medida pelos índices de Jaccard, mostra que no AM, os valores foram de 0,54 (Jaccard_12), 0,62 (Jaccard_13) e 0,54 (Jaccard_23), enquanto em PE observaram-se 0,58, 0,54 e 0,62, respectivamente. No MS, por outro lado, os índices oscilaram entre 0,40 e 0,46, no RJ, os valores de sobreposição mantiveram-se em torno de 0,54 em todas as combinações, e no PR, os coeficientes variaram entre 0,46 e 0,58.

	AM	PE	MS	RJ	PR
K	30	30	30	30	30
Jaccard_12	0.53	0.57	0.42	0.53	0.57
Jaccard_13	0.62	0.53	0.46	0.53	0.46
Jaccard_23	0.53	0.62	0.39	0.53	0.50

Tabela 10 - Métricas de estabilidade estrutural dos modelos por unidade federativa

No TO, os resultados foram de 0,05 (Jaccard_12), 0,00 (Jaccard_13) e 0,03 (Jaccard_23). Já em GO, os valores corresponderam a 0,20, 0,07 e 0,18, respectivamente.

	TO	GO
K	30	30
Jaccard_12	0.05	0.20
Jaccard_13	0	0.07
Jaccard_23	0.03	0.17

Tabela 11 - Métricas de estabilidade estrutural dos modelos de TO e GO.

Além dos indicadores clássicos de desempenho, foram incorporadas medidas de concordância entre execuções, com o objetivo de avaliar a estabilidade das hierarquizações produzidas pelo modelo. Nesse sentido, empregou-se a correlação de Spearman, que compara a ordenação relativa das variáveis entre pares de execuções independentes. Esse coeficiente expressa a força da associação monotônica entre rankings, variando de −1 a 1, onde valores mais próximos da unidade denotam maior alinhamento entre as posições ocupadas pelas variáveis em cada rodada.

De forma complementar, utilizou-se o coeficiente de concordância de Kendall (Kendall's W), voltado para a análise simultânea das três execuções do modelo. Diferentemente do Spearman, que se restringe a comparações par-a-par, o Kendall W resume em um único índice o grau de homogeneidade global das classificações. Seus valores oscilam entre 0 (ausência de concordância) e 1 (concordância perfeita), permitindo aferir a robustez do conjunto de rankings de maneira integrada. Essa abordagem fornece uma visão mais abrangente da consistência dos resultados, especialmente em cenários de alta dimensionalidade, nos quais múltiplas variáveis disputam relevância.

No Amazonas, os coeficientes de Spearman variaram de 0,49 a 0,52 entre as execuções, enquanto o Kendall's W atingiu 0,65. Em Pernambuco, observaram-se valores de Spearman entre 0,53 e 0,56, acompanhados de Kendall's W igual a 0,68. No Mato Grosso do Sul, os índices de Spearman oscilaram de 0,37 a 0,44, com Kendall's W de 0,58. No Rio de Janeiro, os coeficientes de Spearman situaram-se entre 0,54 e 0,59, e o Kendall's W foi de 0,70. Resultados próximos foram encontrados no Paraná, com Spearman variando de 0,53 a 0,59 e Kendall's W igualmente em 0,70.

	AM	PE	MS	RJ	PR
Spearman_rho12	0.514	0.546	0.371	0.577	0.578
Spearman_rho13	0.518	0.529	0.442	0.589	0.594
Spearman_rho23	0.489	0.563	0.373	0.536	0.532
Kendall_W	0.652	0.684	0.577	0.702	0.696

Tabela 12 - Estabilidade dos rankings de variáveis entre execuções do Random Forest, expressa pelos coeficientes de Spearman (ρ) e Kendall's W nos estados do AM, PE, MS, RJ e PR.

No TO, os coeficientes de Spearman variaram de 0,76 a 0,79 entre as execuções, enquanto o Kendall's W atingiu 0,85. Em Goiás, observaram-se valores de Spearman entre 0,78 e 0,78, acompanhados de Kendall's W de 0,85.

	TO	GO
Spearman_rho12	0.773	0.779
Spearman_rho13	0.765	0.781
Spearman_rho23	0.789	0.782
Kendall_W	0.850	0.853

Tabela 13 - Estabilidade dos rankings de variáveis entre execuções do Random Forest, expressa pelos coeficientes de Spearman (ρ) e Kendall's W nos estados do TO e GO

8.Discussão

A interpretação dos achados deste estudo requer atenção especial aos aspectos metodológicos que influenciam diretamente a robustez e a aplicabilidade dos modelos preditivos desenvolvidos. Em análises baseadas em grandes bases populacionais, como a PNS 2019, não apenas a escolha das variáveis importa, mas também a compreensão da natureza de cada uma delas e de sua relação temporal e causal com o desfecho estudado (James Ezeh *et al*., 2024; Patharkar *et al*., 2024).

Essa reflexão é particularmente relevante quando se trata de eventos complexos como o AVC, cujo diagnóstico e consequências podem estar associados a múltiplos fatores(Feigin; Norrving; Mensah, 2017). Entre os cuidados necessários está a distinção entre variáveis que atuam como potenciais preditoras e aquelas que, na realidade, são consequência ou parte da definição do próprio evento (Shmueli, 2010). Esse aspecto ganha relevância ao considerar variáveis relacionadas à idade, frequentemente incluídas em análises de risco de AVC, mas que exigem uma avaliação criteriosa quanto à sua adequação no contexto preditivo.

A análise do desempenho dos modelos em diferentes estados revelou não apenas métricas elevadas de acurácia, mas também nuances importantes sobre a aplicação prática do aprendizado de máquina em epidemiologia populacional. A adoção da estratégia de execução mínima — sem reexecuções destinadas à otimização individual — mostrou-se acertada, pois garantiu maior comparabilidade entre as regiões e evitou o risco de sobreajuste, comum em análises excessivamente calibradas a um contexto/estado específico. Essa decisão metodológica permitiu

observar com mais clareza a robustez da abordagem, destacando que os bons resultados obtidos não dependem de ajustes finos, mas sim da consistência da técnica utilizada e da qualidade dos dados da PNS-2019.

No conjunto dos estados analisados, observou-se desempenho elevado e consistente: em todos os casos, a precisão atingiu 1,0, indicando ausência de falsos positivos. Essa característica, aliada a valores de F1 variando entre 0,80 e 0,83, sugere que as diferenças se concentraram sobretudo na capacidade de sensibilidade, que oscilou entre 0,67 e 0,71. Os resultados de AM, RJ e PR, por exemplo, mostraram padrão semelhante (F1 = 0,80; sensibilidade = 0,67), enquanto em PE houve um leve ganho de sensibilidade (0,71), refletindo-se no F1 mais alto (0,83).

A incorporação de TO e GO, entretanto, evidenciou contrastes importantes. No TO, o desempenho foi substancialmente inferior, com F1 = 0,40 e sensibilidade de apenas 0,25, em que pese a precisão de 1,0. Esse desequilíbrio demonstra maior fragilidade do modelo nesse estado, mesmo com valores elevados de AUC-ROC (0,995) e AUC-PR (0,80). Já em GO, o classificador obteve um dos melhores desempenhos da análise, com F1 = 0,85, sensibilidade de 0,74 e discriminação praticamente perfeita (AUC-ROC = 0,998; AUC-PR = 0,92), situando-se entre os estados mais robustos.

As curvas de desempenho discriminatório reforçam esse quadro. O AUC-ROC manteve-se em valores praticamente perfeitos em todos os estados, chegando a 1,0 em AM, PE e RJ, e muito próximo disso em MS (0,9998), GO (0,998) e TO (0,995), enquanto no PR o valor foi ligeiramente menor (0,98). O AUC-PR apresentou variações mais expressivas: no RJ atingiu 0,97, o maior entre os avaliados, seguido de perto por GO (0,92), enquanto no TO registrou 0,80, o valor mais baixo. Esses resultados demonstram que, embora a discriminação entre classes tenha se mantido em patamares excelentes, nuances relacionadas ao equilíbrio entre precisão e sensibilidade afetam o desempenho preditivo em contextos distintos.

Os indicadores compostos — MCC (≈0,81–0,86, com exceção do TO = 0,50), acurácia balanceada (≈0,83–0,87, contra 0,63 em TO) e Brier Score (≈0,0045–0,0065, contra 0,0107 em TO) — corroboraram a robustez geral da técnica, destacando boa calibração probabilística em quase todos os cenários. Ainda assim, o desempenho destoante do TO sugere que a limitação não se deve à técnica em si, mas a particularidades da base de dados, que podem incluir baixa representatividade amostral ou inconsistências de preenchimento. Em contrapartida, GO se sobressaiu

como um dos estados de maior consistência, com métricas superiores e comportamento estável, reforçando que a robustez do modelo pode emergir sempre que as condições de qualidade e estruturação dos dados são mais favoráveis.

No que se refere à estabilidade da hierarquização de preditores, a avaliação foi realizada considerando o recorte fixo das Top-30 variáveis mais relevantes, aplicado de forma idêntica em todos os estados. A análise dos coeficientes de Jaccard revelou contrastes interessantes: enquanto AM, PE e RJ mostraram sobreposições em torno de 0,54 a 0,62, sugerindo consistência moderada entre execuções, o MS apresentou valores mais baixos (0,40–0,46), apontando para maior variabilidade. Já no PR, os índices oscilaram entre 0,46 e 0,58, situando-se em posição intermediária.

No TO, a baixa performance geral refletiu-se também em menor reprodutibilidade do ranking, sugerindo instabilidade estatística estrutural. Em GO, por sua vez, a estabilidade elevada reforçou o bom desempenho observado, embora isso não elimine possíveis limitações ligadas ao contexto amostral.

Complementarmente, a análise baseada nos coeficientes de Spearman e no Kendall's W trouxe evidências adicionais sobre a consistência dos rankings de importância. Em AM, PE e RJ, os valores de Spearman variaram entre 0,49 e 0,59, acompanhados de Kendall's W entre 0,65 e 0,70, configurando estabilidade moderada. No MS, observaram-se os menores índices (Spearman entre 0,37 e 0,44; Kendall's W = 0,58), sinalizando maior instabilidade na ordenação das variáveis. Em contrapartida, RJ e PR se destacaram por apresentarem os maiores níveis de concordância (Kendall's W = 0,70).

Além disso, os resultados de TO e GO reforçaram, em um primeiro momento, a robustez da abordagem: nesses estados, os coeficientes de Spearman permaneceram consistentemente elevados (≈0,76–0,79), acompanhados dos maiores valores de Kendall's W observados no estudo (≈0,85). Contudo, essa aparente estabilidade deve ser interpretada de forma distinta em cada caso. No TO, a estabilidade dos rankings esconde fragilidades conceituais, já que as variáveis selecionadas apresentaram correlações próximas de zero e ausência de agrupamentos funcionais, sugerindo que o problema está na qualidade e estrutura da base de dados. Em GO, por outro lado, o alto nível de concordância parece refletir de fato um ambiente de dados mais bem estruturado, ainda que passível de algumas limitações amostrais.

Assim, enquanto no TO os resultados apontam para um cenário de estabilidade aparente, mas de baixa relevância substantiva, no GO a consistência reflete um desempenho sólido e representativo. Esse contraste evidencia que os problemas identificados são de natureza distinta: no primeiro caso, relacionados a fragilidades estatísticas e informacionais; no segundo, a possíveis restrições pontuais de representatividade, sem comprometer a robustez geral do modelo.

Os resultados evidenciam que a aplicação do Random Forest aos microdados da PNS-2019 se mostrou consistente e robusta em diferentes regiões do Brasil, com padrões de desempenho elevados e replicáveis. As variações observadas entre os estados não comprometem a solidez da abordagem, mas acrescentam camadas de interpretação importantes, sugerindo que, mesmo em um quadro de estabilidade metodológica, há espaço para investigações futuras sobre como fatores regionais modulam a previsibilidade dos modelos em saúde populacional.

Além disso, a escolha por trabalhar com um estado de cada região foi essencial para captar a heterogeneidade territorial do Brasil. Essa estratégia mostrou como diferenças em perfis sociodemográficos, clínicos e de acesso a serviços podem influenciar marginalmente as métricas de desempenho, sem comprometer a estabilidade geral dos modelos. Assim, torna-se possível identificar não apenas a validade global da metodologia, mas também levantar hipóteses sobre variações regionais que poderão ser aprofundadas em análises futuras.

8.1 Diferenciação entre variáveis preditoras e consequentes: o caso das variáveis relacionadas à idade no estudo de AVC

Em estudos preditivos baseados em dados populacionais, como os da PNS 2019, é fundamental distinguir variáveis que antecedem o desfecho daquelas que são consequência ou fazem parte da definição do próprio evento (James Ezeh *et al*., 2024; Patharkar *et al*., 2024; Shmueli, 2010).

Essa separação é especialmente importante quando o objetivo é construir modelos de aprendizado de máquina, como Random Forest, para prever desfechos clínicos como o diagnóstico de AVC (Marketou *et al*., 2022).

A utilização indevida de variáveis consequentes pode gerar data leakage — ou vazamento de informação — comprometendo a validade preditiva do modelo (Starcke *et al*., 2025a). Entre as variáveis frequentemente relacionadas ao risco de AVC estão

aquelas associadas à idade (Li *et al*., 2017; Soto-Cámara *et al*., 2020). Contudo, nem todas as variáveis de idade disponíveis no banco de dados são apropriadas para uso como preditoras. Abaixo detalhamos a avaliação de três variáveis distintas:

- Idade atual do entrevistado

A idade atual do entrevistado representa a idade cronológica no momento da coleta dos dados. Trata-se de uma variável anterior ao desfecho, o que a caracteriza como uma variável preditora legítima. Epidemiologicamente, a idade é um fator de risco amplamente estabelecido para AVC, com aumento significativo da incidência a partir da sexta década de vida (Soto-Cámara *et al*., 2020).

Além de seu respaldo clínico, a idade apresenta ampla variabilidade entre os indivíduos, o que favorece sua utilização em modelos baseados em árvore de decisão, como Random Forest, permitindo uma maior capacidade de divisão das amostras em subgrupos informativos (Venkatasubramaniam *et al*., 2017). Assim, sua inclusão como variável independente é recomendada tanto do ponto de vista técnico quanto interpretativo.

- Ano de nascimento

O ano de nascimento também representa uma informação coletada antes do desfecho. No entanto, ele costuma ser redundante em relação à idade, uma vez que a idade geralmente é derivada diretamente a partir do ano de nascimento subtraído da data da entrevista. O uso simultâneo das duas variáveis pode causar colinearidade e distorcer a análise de importância relativa no modelo, especialmente quando variáveis contínuas com maior amplitude numérica dominam os critérios de divisão das árvores.

Ainda assim, o ano de nascimento pode ter utilidade analítica em contextos específicos — por exemplo, quando o objetivo for estudar efeitos de coorte ou de geração, ao agrupar indivíduos por períodos históricos de nascimento. Fora desse contexto, sua inclusão como preditor numérico não oferece vantagem adicional em relação à idade.

- Idade ao diagnóstico de AVC

Diferentemente das duas anteriores, esta variável refere-se à idade que o indivíduo tinha quando recebeu o diagnóstico de AVC. Por definição, ela só é informada por respondentes que já relataram ter tido um AVC, permanecendo

ausente ou nula para todos os demais. Isso a torna uma variável subsequente ao desfecho, ou seja, uma consequência, e não uma causa (Chowdhury; Turin, 2020).

Utilizar essa variável como preditora em um modelo de classificação (teve ou não teve AVC) resulta em um erro metodológico grave: o modelo pode aprender que a simples presença ou ausência de valor nessa coluna já indica o resultado da variável dependente (Gorelick, 2006). Essa violação do princípio de temporalidade dos dados compromete totalmente a validade do modelo, pois ele passa a fazer predições baseadas em informações que, na prática, não estariam disponíveis no momento da tomada de decisão.

Essa situação caracteriza um típico caso de data leakage, e, portanto, a variável "idade ao diagnóstico de AVC" não deve ser incluída em modelos preditivos. Seu uso é mais adequado em análises descritivas da população que já sofreu AVC, ajudando a identificar perfis clínicos ou padrões etários associados ao evento (Starcke *et al*., 2025b).

A correta distinção entre variáveis preditoras e variáveis consequentes é essencial para a construção de modelos válidos e interpretáveis (Kapoor; Narayanan, 2023). Embora variáveis relacionadas à idade sejam fundamentais na análise do risco de AVC, é necessário compreender suas diferentes funções dentro da base de dados.

Enquanto a idade atual é uma preditora válida e informativa, a idade ao diagnóstico deve ser excluída dos modelos preditivos por representar uma informação posterior ao evento. Já o ano de nascimento pode ser utilizado de forma alternativa ou contextual, desde que sua relação com a idade seja considerada. Essa avaliação criteriosa fortalece a consistência do modelo, evita viés técnico e contribui para resultados mais confiáveis (Patharkar *et al*., 2024).

8.2 Relação entre Explicação, Predição e o Princípio da Temporalidade das Variáveis

A distinção entre variáveis preditoras e consequentes, previamente discutida, é reforçada por Shmueli (2010), que diferencia de forma sistemática os objetivos de explicação e predição. Enquanto abordagens explicativas buscam compreender relações causais e mecanismos subjacentes, modelos preditivos priorizam a capacidade de antecipar eventos futuros com base em padrões extraídos dos dados.

Em análises preditivas na área da saúde, o respeito ao princípio da temporalidade das variáveis é fundamental para a validade metodológica. A utilização de informações obtidas apenas após a ocorrência do evento — como medidas clínicas ou funcionais registradas exclusivamente em indivíduos já diagnosticados — caracteriza data leakage, situação na qual o modelo é treinado com dados indisponíveis no momento real de tomada de decisão (ocasião do AVC). Essa prática gera métricas de desempenho artificialmente elevadas e compromete a capacidade de generalização do modelo, resultando em predições que, na prática, não são prospectivas, mas retrospectivas.

Esse risco se intensifica quando a base de dados apresenta lacunas relevantes em variáveis coletadas antes do desfecho, como exames laboratoriais, hábitos de vida ou histórico de doenças crônicas, levando o algoritmo a priorizar variáveis temporariamente posteriores ao evento. Essa dependência de informações pós-desfecho desloca o foco preditivo para um reconhecimento de casos já consolidados, alterando a natureza da análise.

A observação de padrões regionais demonstra que, em determinadas localidades, variáveis pós-desfecho assumem papel central na estrutura do modelo, sinalizando a necessidade de avaliação crítica não apenas do desempenho estatístico, mas também da relevância temporal e clínica das variáveis utilizadas. Essa análise criteriosa assegura que a acurácia reportada seja reflexo de uma verdadeira capacidade de previsão antecipada, e não de um ajuste enviesado a dados retrospectivos.

8.3 Análise Comparativa das Variáveis Mais Relevantes para Predição de AVC nas Regiões Brasileiras

8.3.1 Região Norte

Principais variáveis recorrentes (sem TO)

- Idade atual e idade no diagnóstico do AVC
- Rendimento domiciliar per capita
- Grau de limitação funcional pós-AVC
- Acompanhamento de saúde e reabilitação (dieta, fisioterapia, medicamentos)

Na Região Norte, a análise dos modelos revela um padrão marcadamente influenciado por variáveis associadas ao período posterior ao evento do AVC. Predominam entre as mais relevantes a idade ao diagnóstico, o grau de limitação funcional e o uso de estratégias de reabilitação e medicação. Esses elementos sugerem que o modelo está aprendendo a partir de registros consolidados da condição clínica, o que limita sua capacidade de prever o evento de forma antecipada. Em outras palavras, o modelo opera mais como um classificador de casos já ocorridos do que como uma ferramenta preditiva em sentido estrito.

A ausência quase sistemática de informações sobre exames laboratoriais, hábitos de vida e histórico de doenças crônicas, esperadas em análises preditivas de saúde, aponta para lacunas importantes na base de dados da região (Bensenor *et al*., 2015b). Essa ausência pode ser explicada por duas possibilidades complementares: a baixa taxa de preenchimento desses campos ou a baixa variabilidade real dessas informações dentro da população, o que afeta sua significância estatística nas análises automatizadas.

Contudo, ao se excluir o estado do Tocantins (TO), observa-se uma estrutura mais coesa e consistente entre os demais estados da Região Norte. As variáveis emergentes mantêm coerência temática, combinando dados sociodemográficos com condições clínicas e funcionais associadas ao AVC, compondo um núcleo analítico robusto. Nota-se, por exemplo, a recorrência de indicadores ligados à gravidade funcional do derrame e às práticas de tratamento, como dieta específica e uso de aspirina, sugerindo uma leitura fiel do fenômeno a partir das variáveis disponíveis.

8.3.1.1 Impacto da inclusão do Tocantins (TO)

A inclusão dos dados do TO, no entanto, altera sensivelmente esse panorama. Nenhuma variável atinge o limiar de presença em mais de 90% dos estados da Região Norte quando o TO é considerado. Essa quebra de consistência estatística reflete-se diretamente na performance do modelo: há diluição dos padrões robustos previamente observados, instabilidade na lista de variáveis mais relevantes e enfraquecimento da capacidade do modelo de capturar sinais consistentes em nível regional.

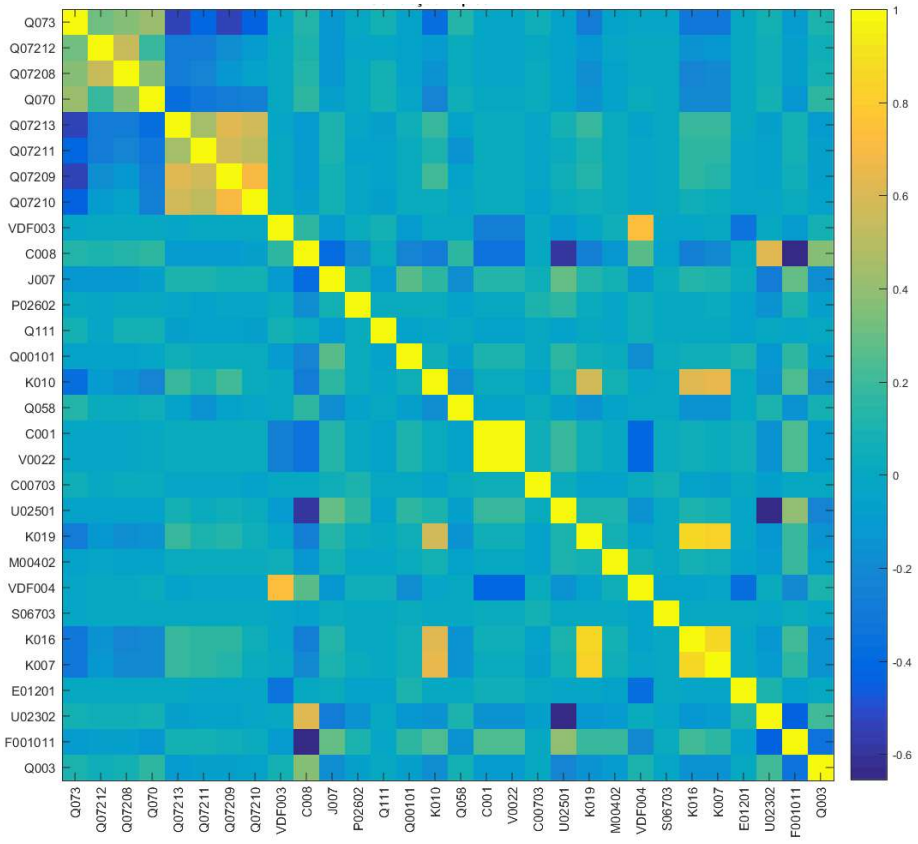
Essa instabilidade é confirmada ao se observar a matriz de correlação das 30 variáveis mais importantes para o TO (Figura 8). Diferente do que se esperaria de uma base de dados bem estruturada como o do AM (Figura 7) — com blocos ou

clusters de variáveis correlacionadas representando domínios temáticos coerentes (como saúde funcional, renda, ocupação etc.) —, a matriz revela um padrão de correlação disperso e desorganizado.

Predomina um mar de correlações próximas de zero, com escassa presença de relações fortes ou moderadas entre variáveis. A maioria das correlações paira entre 0,1 e -0,1, indicando que as variáveis selecionadas como “mais importantes” não compartilham informações estatisticamente relevantes entre si. Além disso, aparecem correlações negativas pontuais e fracas, muitas vezes entre variáveis sem relação teórica clara, o que reforça a hipótese de ruído ou inconsistência na base.

Outro ponto crítico é a ausência de qualquer estrutura de agrupamento. Em estados com dados consistentes, é comum observar agrupamentos funcionais entre variáveis relacionadas ao mesmo eixo de análise (ex: autonomia funcional, condição laboral, comorbidades).

Figura 7 - Correlação das 30 variáveis mais significativas do AM.

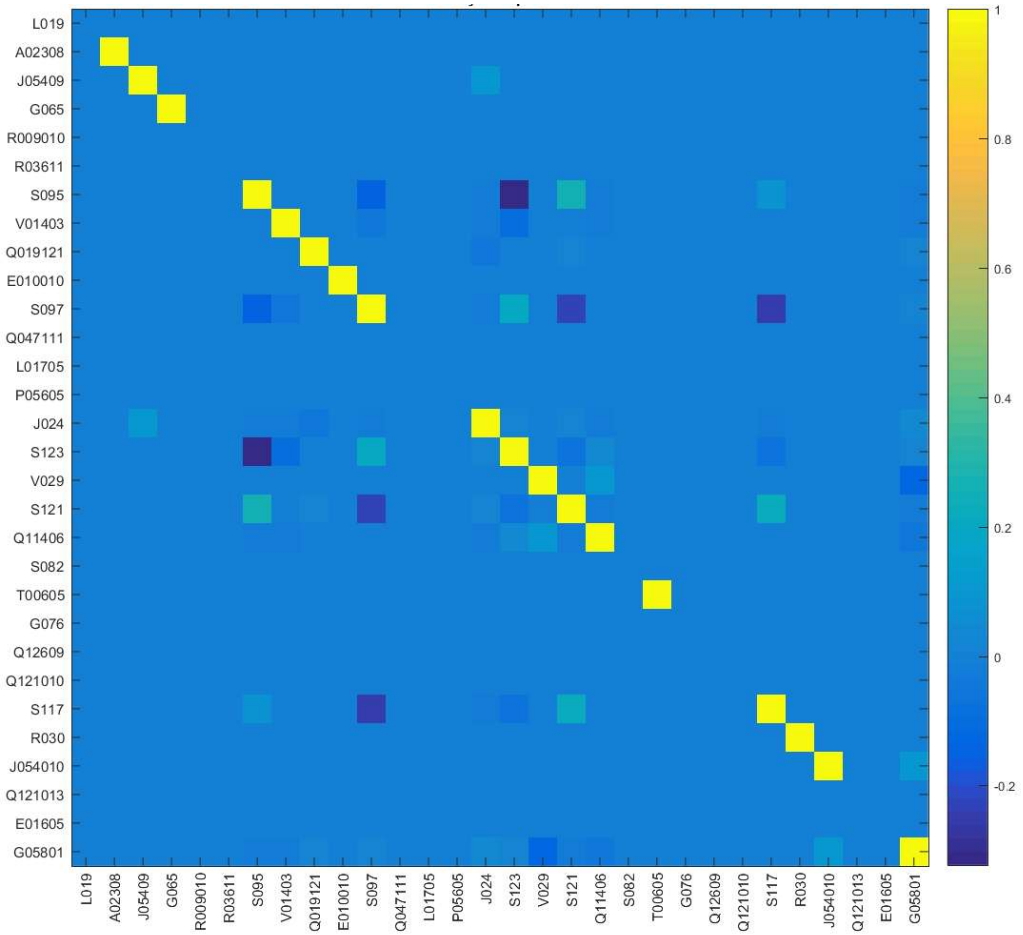


No TO, tais agrupamentos são inexistentes. Cada variável parece atuar isoladamente, sem formar um núcleo de informações que reflita um fenômeno clínico

ou sociodemográfico robusto. Isso compromete tanto a confiabilidade dos resultados quanto a reprodutibilidade dos achados (Vigneau, 2020).

Esses problemas sugerem causas técnicas e estruturais importantes, como baixa variabilidade nas respostas, excesso de valores constantes (como zeros), falhas na codificação ou baixo preenchimento de campos (Nijman *et al*., 2022). Tais características reduzem drasticamente a efetividade dos algoritmos de importância relativa, como aqueles empregados nas abordagens de Random Forest, e comprometem a legitimidade estatística da seleção das variáveis.

Figura 8 - Correlação das 30 variáveis mais significativas de TO.



Diante desse cenário, os modelos construídos para a Região Norte sem o Tocantins tendem a apresentar melhor desempenho e maior robustez, ainda que se restrinjam majoritariamente à identificação de casos confirmados, não à predição antecipada. Isso indica a necessidade de ações estruturais voltadas à melhoria da coleta de dados em saúde populacional, com ênfase na atenção primária, na

informatização dos registros e na ampliação da captação de informações clínicas e preventivas (Yuen *et al*., 2023).

Para o caso específico do Tocantins, recomenda-se uma análise técnica aprofundada da estrutura de codificação e preenchimento da base de dados, uma vez que sua inclusão compromete de maneira significativa os padrões estatísticos da região como um todo.

8.3.2 Região Nordeste

Principais variáveis recorrentes

- Ano de nascimento e idade atual
- Ocupação
- Diagnóstico de infarto/AVC prévio
- Limitação funcional
- Reabilitação: fisioterapia, dieta, medicamentos
- Acompanhamento com profissional de saúde

A Região Nordeste apresenta um perfil de variáveis semelhante ao observado na Região Norte, com destaque para informações relacionadas ao evento do AVC e ao seu manejo clínico e funcional (Feigin; Norrving; Mensah, 2017). Idade, diagnóstico de infarto/AVC e grau de limitação funcional seguem como pilares informacionais centrais nos modelos treinados para os estados nordestinos.

No entanto, o Nordeste diferencia-se ao incorporar, de forma mais evidente, marcadores socioeconômicos, especialmente os relacionados ao trabalho e à ocupação (Pereira; Queiroz, 2023).

Variáveis como “cargo ou função exercida” emergem com frequência relevante, refletindo não apenas a diversidade laboral da região, mas também possíveis vulnerabilidades associadas ao perfil ocupacional.

Essa presença recorrente pode estar associada à alta informalidade e à prevalência de atividades autônomas ou de baixa proteção trabalhista — características históricas do mercado de trabalho nordestino. Como consequência, o modelo parece captar, de maneira indireta, aspectos de risco social que não estão

necessariamente associados a condições clínicas, mas sim à estrutura socioeconômica do indivíduo (Pereira; Queiroz, 2023).

O cruzamento entre variáveis funcionais (limitação, reabilitação) e ocupacionais oferece um retrato mais complexo e contextualizado da condição de saúde da população (Da Silveira *et al*., 2010). Isso amplia ligeiramente a profundidade analítica do modelo, ainda que a maioria das variáveis identificadas siga ancorada em informações retrospectivas — ou seja, após a ocorrência do AVC.

O modelo da Região Nordeste demonstra uma leve expansão em relação ao modelo do Norte, incorporando dimensões socioeconômicas relevantes, mas ainda depende, em larga medida, de variáveis que descrevem o evento já ocorrido. A presença de dados sobre ocupação e reabilitação reforça a necessidade de integração entre bases de dados de saúde e trabalho, permitindo análises mais abrangentes e intervenções preventivas mais eficazes.

Adicionalmente, a ausência de marcadores laboratoriais e de estilo de vida entre as variáveis mais importantes evidencia uma lacuna crítica na coleta de dados preventivos (Da Silveira *et al*., 2010). Para ampliar o potencial preditivo da modelagem, seria fundamental investir em melhorias na cobertura e qualidade dessas informações, promovendo uma maior articulação entre vigilância epidemiológica, atenção primária e indicadores sociais de risco.

8.3.3 Região Centro-Oeste

Principais variáveis recorrentes

- Idade no diagnóstico do AVC
- Grau de limitação funcional
- Adoção de dieta como forma de reabilitação

A análise das variáveis mais importantes na Região Centro-Oeste revela um padrão limitado de recorrência estatística. A maioria dos modelos é sustentada por variáveis diretamente ligadas ao pós-evento — como idade ao diagnóstico, limitação funcional e estratégias de reabilitação.

Esse conjunto reduzido pode ser parcialmente explicado pela influência desproporcional do estado de Goiás (GO), responsável por cerca de 25% da base regional (um dos quatro entes federativos da região).

A matriz de correlação referente às 30 variáveis mais relevantes identificadas para GO (Figura 9) evidencia a origem dessa limitação. O gráfico revela um padrão irregular, com múltiplas áreas sem correlação significativa entre variáveis relevantes. Há grande concentração de valores baixos ou próximos de zero fora da diagonal principal, indicando fraca co-ocorrência entre as variáveis destacadas como importantes. Além disso, a distribuição das correlações aparenta ser ruidosa e aleatória, sem agrupamentos tematicamente coerentes — o que sugere ausência de estrutura ou coerência interna nos dados.

Observam-se ainda alguns blocos de correlação local elevada (ex: entre Q070, Q073 e Q07208), o que pode indicar colinearidade entre variáveis derivadas de um mesmo domínio (idade e condição pós-AVC). Porém, fora esse pequeno núcleo, as demais variáveis mostram relações pouco expressivas entre si, reforçando a hipótese. Essa instabilidade interna compromete a repetibilidade dos resultados. Entre diferentes execuções do modelo, há baixa permanência das mesmas variáveis quando avaliadas as 10 variáveis mais importantes, o que dificulta análises mais robustas e comparações interregionais. A inclusão de GO no cálculo das recorrências regionais distorce os achados dos demais estados, reduzindo a diversidade informacional e mascarando padrões consistentes.

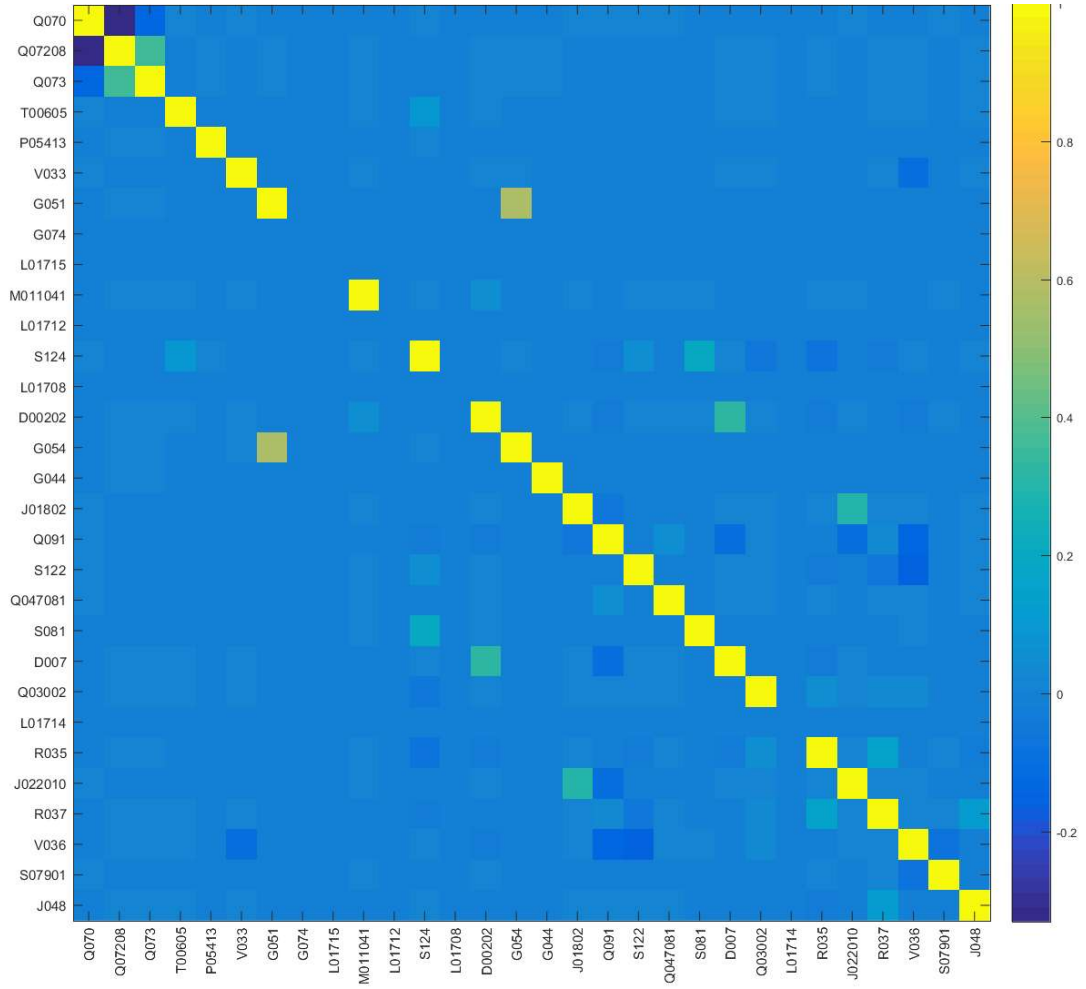
8.3.3.1 Comparação com Mato Grosso (MT)

Ao observar a matriz de correlação do estado do MT (Figura 10), nota-se uma estrutura significativamente mais estável e coerente:

- Há agrupamentos temáticos bem definidos. Por exemplo, o bloco de variáveis Q07209 a Q07213 (reabilitação e medicações) apresenta correlações cruzadas fortes e consistentes.
- Variáveis clínicas e funcionais aparecem integradas a domínios como comorbidades (J007), condições socioeconômicas (ex: K010, K022) e dados de mobilidade ou hábitos.
- Correlações intermediárias entre domínios diferentes sugerem que os dados estão captando relações reais e multidimensionais, em vez de padrões aleatórios ou ruído.

- A distribuição das correlações é mais rica e densa que em GO, com menor incidência de valores nulos ou negativos extremos.

Figura 9 - Correlação das 30 variáveis mais significativas do GO.



Esse contraste destaca como MT apresenta uma base estatística mais madura, permitindo ao modelo captar relações complexas entre diferentes aspectos da vida do indivíduo e sua condição pós-AVC. A estrutura da matriz de correlação indica que o modelo aplicado no estado consegue dialogar com múltiplas dimensões — clínica, funcional, ocupacional e socioeconômica.

Diferenças observadas ao remover Goiás (GO)

A exclusão de GO dos cálculos regionais gera melhorias evidentes:

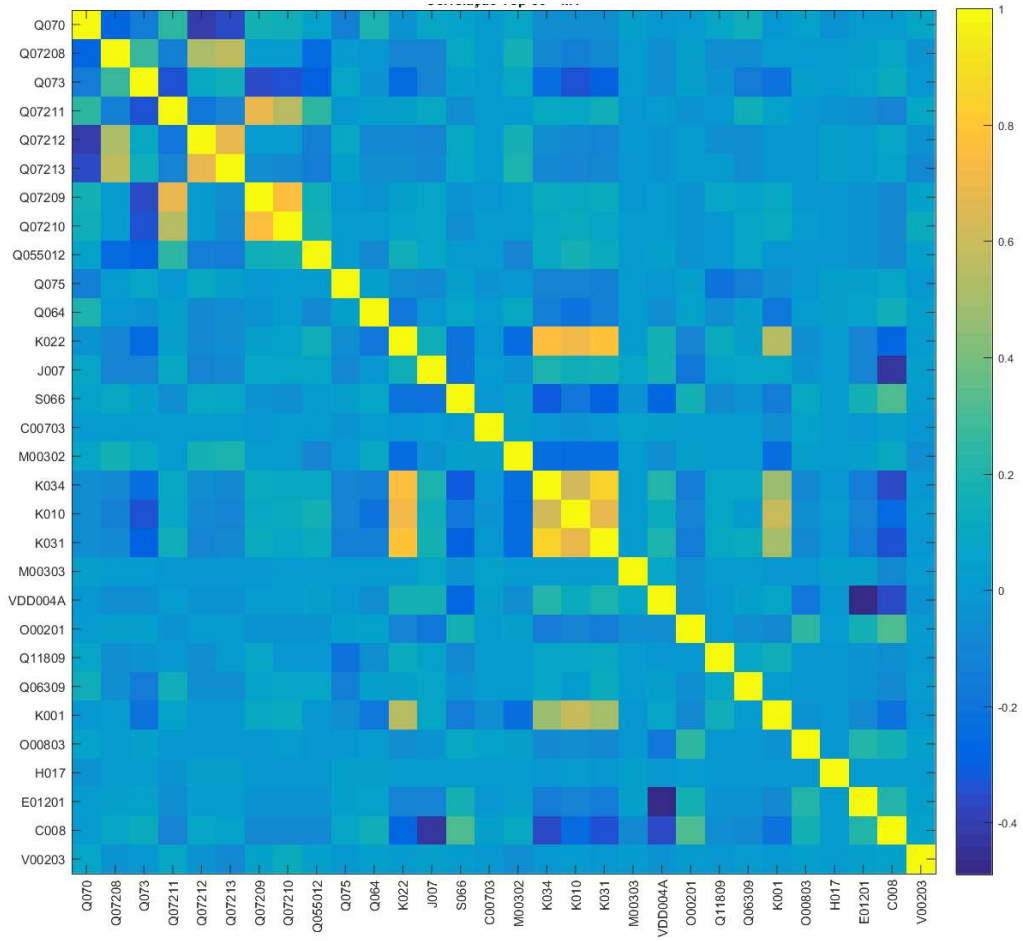
- Maior estabilidade entre execuções: As variáveis-chave se repetem com maior regularidade.
- Aumento da diversidade temática: Além das variáveis clínicas, surgem indicadores de renda, trabalho, mobilidade e saúde preventiva.

- Eliminação de ruídos técnicos: As anomalias observadas na matriz de GO, como correlações infladas artificialmente ou colunas quase planas, desaparecem.
- A exclusão de GO dos cálculos regionais gera melhorias evidentes:
- Maior estabilidade entre execuções: As variáveis-chave se repetem com maior regularidade.

Exemplos de variáveis que surgem com a remoção de GO

- Indicadores ocupacionais (ex: atividade do negócio, vínculo formal)
- Tempo de deslocamento ao trabalho
- Renda domiciliar per capita
- Acompanhamento contínuo com profissional de saúde
- Medidas associadas a reabilitação funcional

Figura 10 - Correlação das 30 variáveis mais significativas do MT.



A comparação entre GO e MT evidencia como diferentes padrões estruturais dos dados impactam diretamente os resultados analíticos. Em MT, observa-se o comportamento mais esperado: um pequeno núcleo de variáveis de alta relevância sustentando o modelo, seguido por uma longa cauda de preditores secundários, o que confere clareza na hierarquização e maior confiabilidade interpretativa. Já em GO, o modelo atribuiu importâncias elevadas e relativamente homogêneas a muitas variáveis, refletindo um cenário de redundância excessiva, no qual diferentes preditores captam informações muito semelhantes, dificultando a distinção de quais fatores são, de fato, mais centrais. Esse “achatamento” compromete a robustez da análise regional, uma vez que a estabilidade observada nos rankings não se traduz necessariamente em maior capacidade explicativa.

Na Região Centro-Oeste, portanto, enquanto MT se apresenta como um caso de hierarquização consistente, GO revela limitações ligadas à redundância e sobreposição informacional. Situação distinta ocorre em TO, onde o problema não é a redundância, mas a fragilidade informacional: a ausência de estrutura estatística e de variabilidade adequada leva a importâncias artificiais e homogêneas, resultando em estabilidade ilusória e baixo desempenho preditivo. Esses dois estados, por motivos diferentes, foram os únicos identificados no estudo em que o comprometimento estrutural da base de dados afetou de forma significativa tanto a robustez quanto a interpretabilidade dos modelos regionais.

8.3.4 Região Sudeste

Principais variáveis recorrentes

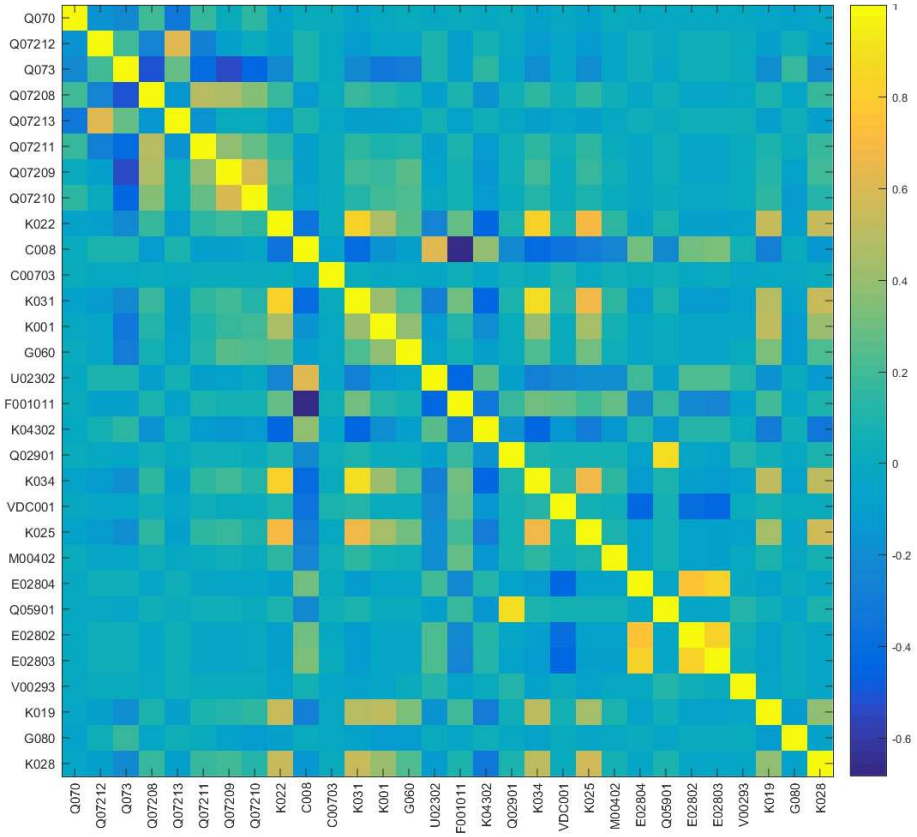
A Região Sudeste apresentou a maior diversidade de variáveis relevantes dentre todas as regiões analisadas, conforme evidenciado na Figura 11, que ilustra a matriz de correlação das 30 variáveis de maior importância identificadas para o estado de MG.

Dentre os itens mais frequentes, observam-se desde indicadores sociodemográficos amplos até variáveis clínicas e funcionais altamente específicas. Entre os principais fatores que figuraram nos modelos estão:

- Rendimento domiciliar per capita e quantidade de moradores, marcando o pano de fundo socioeconômico do domicílio;

- Posses do domicílio, como a presença de computador, funcionando como indicador indireto de renda e acesso à informação;
- Altura, peso e exames laboratoriais (colesterol, glicemia, PA), que refletem uma base de dados com bom preenchimento de informações de saúde preventiva;
- Grau de dificuldade funcional e uso de transporte ou ida ao médico, indicando aspectos da autonomia e do uso de serviços de saúde;
- Condições pré-existentes, como infarto, diabetes e hipertensão, compondo o perfil clínico de risco do indivíduo;
- Ocupação e rendimento do trabalho, conectando a análise com a inserção socioeconômica mais ampla;
- Idade ao diagnóstico e grau de limitação funcional pós-AVC, marcadores diretos do desfecho de interesse.

Figura 11 - Correlação das 30 variáveis mais significativas de MG.



Ao contrário de outras regiões onde o modelo tende a depender quase exclusivamente de variáveis pós-evento, como reabilitação e grau de limitação, o Sudeste oferece ao modelo insumos mais ricos e diversificados. Essa característica

permite à inteligência artificial detectar fatores de risco reais e antecipatórios, em vez de apenas reconhecer a existência consolidada de um AVC.

A estrutura do modelo treinado com os dados do Sudeste é composta por diferentes dimensões de análise:

1. Eixo socioeconômico: A presença de variáveis como renda per capita, número de moradores e posse de bens demonstra uma forte influência das condições estruturais da vida cotidiana na construção do risco.
2. Eixo clínico-funcional: A recorrência de exames laboratoriais, doenças prévias e dados funcionais sugere que o modelo consegue capturar o histórico de saúde de forma robusta.
3. Eixo comportamental e de acesso à saúde: A frequência de uso de transporte, ida ao médico e inserção ocupacional expande ainda mais a capacidade de previsão, conectando hábitos, mobilidade e serviço de saúde.

Essa pluralidade de indicadores garante um modelo mais sensível e generalizável, com potencial para identificar sinais precoces de vulnerabilidade, e não apenas suas consequências.

A Região Sudeste demonstra uma estrutura de dados mais completa, coerente e padronizada, o que impacta diretamente a qualidade dos resultados analíticos. A uniformidade nos registros — tanto nos domínios clínicos quanto nos socioeconômicos — sugere:

- Maior efetividade na capacitação dos agentes de coleta, indicando domínio sobre o instrumento da pesquisa;
- Maior familiaridade da população com os itens do questionário, o que pode estar relacionado a níveis mais altos de escolaridade e acesso à informação;
- Menor incidência de valores ausentes ou mal preenchidos, permitindo ao modelo encontrar padrões reais e não artificiais.

8.3.5 Região Sul

Principais variáveis recorrentes

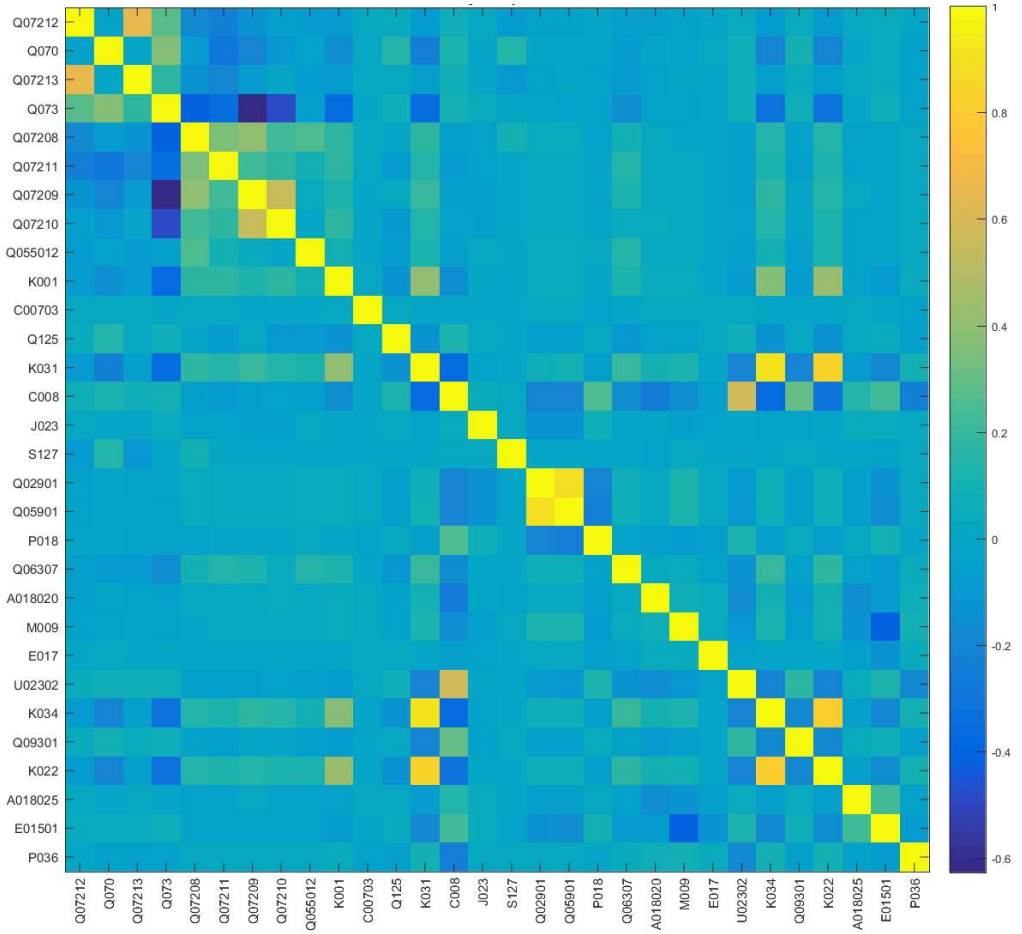
A análise das variáveis mais relevantes na Região Sul evidencia uma composição equilibrada entre fatores clínicos, ocupacionais e de estilo de vida, conforme ilustrado

na Figura 12, que apresenta as 30 variáveis de maior importância identificadas para o estado do RS.

Entre os itens que mais frequentemente figuram nas primeiras posições dos modelos regionais, destacam-se:

- Idade atual e ano de nascimento, que permanecem como pilares do perfil de risco;
- Características ocupacionais, como formalização do trabalho (carteira assinada), número de horas trabalhadas e função exercida, compondo uma camada de análise socioeconômica com especial riqueza;
- Indicadores clínicos preventivos, como exames laboratoriais (níveis de glicemia, colesterol e PA);
- Hábitos de vida, notadamente o consumo de álcool e a frequência de consultas médicas regulares;
- Variáveis funcionais, ligadas à autonomia e capacidade de realizar tarefas do dia a dia — como sair de casa sozinho, vestir-se ou fazer compras.

Figura 12 - Correlação das 30 variáveis mais significativas de RS.



O Sul apresenta uma estrutura de dados mais integrada e prospectiva, quando comparado às outras regiões. O modelo consegue captar com maior sensibilidade indicadores comportamentais, funcionais e ocupacionais, compondo um retrato mais holístico do risco populacional.

A capacidade de mapear variáveis como tipo e frequência de consumo de álcool, autonomia para tomar medicamentos ou habilidade para se locomover sem ajuda aponta para uma leitura mais refinada do cotidiano do indivíduo, permitindo antecipar quadros de fragilidade ou risco social.

Soma-se a isso a forte presença de variáveis de saúde ocupacional, que ajudam a entender as condições laborais como possíveis gatilhos ou fatores protetores para eventos vasculares.

O padrão identificado na Região Sul reforça a ideia de um modelo com maior maturidade em termos de vigilância funcional e saúde preventiva. Entre os principais desdobramentos, destacam-se:

- Presença sistemática de variáveis laboratoriais e de rotina médica, sinalizando bom acesso ao sistema de saúde e preenchimento regular dos campos clínicos;
- Integração entre dimensões sociais, clínicas e comportamentais, permitindo identificar grupos em situação de vulnerabilidade não apenas física, mas também funcional e social;
- Potencial de aplicação mais preditiva e menos tautológica, já que o modelo depende menos de indicadores diretamente derivados do diagnóstico de AVC e mais de sinais de risco presentes antes da ocorrência do evento.

8.4 Lições Regionais sobre Dados e Predição de AVC no Brasil

A análise regional dos dados relacionados ao AVC revela disparidades significativas na qualidade, completude e natureza das variáveis mais relevantes nos modelos de correlação. A forma como cada região se destaca — ou falha — na entrega de dados confiáveis oferece importantes aprendizados sobre os limites e potencialidades da vigilância em saúde populacional no Brasil.

8.5 Diferenças entre regiões: entre ruídos e riqueza informacional

Enquanto regiões como Sudeste e Sul se sobressaem pela riqueza e diversidade de variáveis — abrangendo desde aspectos clínicos até socioeconômicos e de funcionalidade —, outras como Norte e Centro-Oeste enfrentam graves desafios estruturais nos dados. O Nordeste aparece em uma posição intermediária, revelando sinais promissores, mas ainda muito ancorados em variáveis retrospectivas (pós-evento).

As regiões Sudeste e Sul oferecem um retrato mais completo da realidade dos pacientes e seus contextos, com variáveis como altura, peso, histórico clínico (hipertensão, diabetes), frequência de exames laboratoriais, grau de dificuldade para atividades do dia a dia, posse de bens, e condições de trabalho. Isso não apenas aumenta o poder preditivo dos modelos, mas também sugere que essas populações têm acesso mais consolidado a sistemas de saúde e educação, além de apresentarem melhor estrutura na coleta dos dados (Gaspar *et al.*, 2021).

Já no caso do Centro-Oeste, a presença desproporcional de problemas técnicos relacionados ao estado de GO compromete severamente os resultados. GO, responsável por 25% da base de dados da região, apresentou anomalias estatísticas, como colunas de correlação inteiramente zeradas e baixa repetição de variáveis entre execuções, o que reduz a estabilidade e a confiabilidade dos modelos.

A exclusão de GO trouxe ganhos imediatos: maior diversidade de variáveis, maior repetição entre execuções, e desaparecimento de artefatos estatísticos. Fenômeno semelhante é observado no Norte com relação ao TO. Quando excluído da análise, os modelos da região Norte ganham clareza e coesão.

Variáveis como idade no diagnóstico do AVC, grau de limitação funcional e práticas de reabilitação se tornam consistentes entre os estados. A inclusão de TO, contudo, dilui esses padrões, impedindo que qualquer variável atinja o limiar de presença em mais de 90% dos estados. Tal como em GO, a hipótese mais plausível é que o estado apresente uma estrutura de respostas ruidosa, possivelmente associada a erros técnicos, baixa variabilidade de respostas ou preenchimento incompleto dos questionários.

O Nordeste, por sua vez, apesar de compartilhar com o Norte a predominância de variáveis associadas ao evento e à reabilitação, distingue-se por incorporar elementos ocupacionais e socioeconômicos relevantes.

A presença recorrente de perguntas sobre cargo, função, informalidade e rendimento sugere que o modelo está aprendendo, ainda que indiretamente, sobre vulnerabilidades sociais associadas à saúde. Isso pode ser reflexo da diversidade de realidades econômicas na região e do peso das desigualdades estruturais na determinação dos riscos de AVC (Vincens; Stafström, 2015b).

8.6 Variáveis pós-evento e o limite preditivo

Um dos principais achados transversais às regiões menos desenvolvidas em termos de dados é a predominância de variáveis retrospectivas, ou seja, associadas ao que ocorre depois do AVC: fisioterapia, dieta pós-AVC, grau de limitação, uso de medicamentos, acompanhamento de saúde, entre outras. Essa característica revela uma limitação grave: os modelos estão sendo treinados mais como ferramentas de identificação de casos já consolidados do que como sistemas de predição precoce, capazes de antever o risco de um AVC antes de sua ocorrência.

A ausência recorrente de variáveis como histórico detalhado de doenças crônicas, frequência de exames preventivos, prática de atividade física ou indicadores nutricionais abrangentes (para além de peso e altura) compromete de maneira significativa a capacidade dos modelos de operarem sob uma lógica verdadeiramente preventiva. Em epidemiologia, esses elementos não apenas servem como preditores independentes, mas também permitem a construção de perfis compostos que capturam interações complexas entre fatores biológicos, comportamentais e sociais.

No campo dos indicadores nutricionais, informações detalhadas sobre padrões alimentares — como consumo de frutas, verduras, legumes, leguminosas, peixes, cereais integrais e oleaginosas, bem como ingestão de sódio, açúcares adicionados e gorduras saturadas — são fundamentais para compreender o risco cardiovascular. Evidências apontam que dietas ricas em vegetais e pobres em sódio reduzem significativamente a P.A. e a rigidez arterial, enquanto padrões alimentares de alta densidade energética e pobres em fibras estão associados à inflamação sistêmica e à aterosclerose subclínica, importantes precursores do AVC. Além disso, o acompanhamento longitudinal desses hábitos permite identificar transições alimentares em nível populacional, como a substituição de refeições tradicionais por alimentos ultraprocessados, fenômeno em expansão nas últimas décadas no Brasil.

A prática de atividade física, por sua vez, constitui um dos pilares mais consistentes na prevenção primária e secundária do AVC. A ausência de dados detalhados sobre frequência, duração, intensidade e modalidade de exercício físico limita a capacidade de diferenciar indivíduos que, apesar de inativos no lazer, apresentam altos níveis de atividade ocupacional ou de deslocamento ativo (como caminhar ou pedalar para o trabalho). Esses nuances são críticos, pois diferentes tipos de atividade física produzem impactos distintos no perfil cardiovascular: exercícios aeróbicos regulares melhoram a função endotelial e reduzem a PA, enquanto o treinamento resistido contribui para o controle glicêmico e a manutenção da massa muscular, reduzindo o risco de sarcopenia e fragilidade.

Outro ponto relevante é que, tanto no caso da alimentação quanto da atividade física, há interações sinérgicas com outros fatores de risco. Por exemplo, uma dieta rica em fibras e antioxidantes potencializa os efeitos benéficos do exercício sobre o metabolismo lipídico, enquanto o sedentarismo pode neutralizar parcialmente os benefícios de um padrão alimentar saudável. Ignorar essas variáveis nos modelos significa subestimar a influência de redes complexas de fatores que determinam o risco de AVC.

Infelizmente, em muitas regiões analisadas, mesmo quando essas informações constam nos instrumentos de coleta, elas são prejudicadas por baixa taxa de preenchimento, categorização simplista (por exemplo, “pratica exercício: sim ou não”) e falta de padronização nas unidades de medida. Essa fragilidade metodológica reduz a variabilidade estatística, tornando as variáveis irrelevantes para algoritmos baseados em importância relativa, como o Random Forest.

Como consequência, perpetua-se uma lógica de modelagem centrada no passado clínico, sem capacidade de antecipar riscos e orientar intervenções precoces. A incorporação sistemática e qualificada de indicadores nutricionais e de atividade física, com granularidade suficiente para captar hábitos e mudanças ao longo do tempo, poderia transformar esses modelos em ferramentas mais próximas de um sistema de vigilância epidemiológica inteligente, capaz de direcionar recursos de prevenção para grupos de maior vulnerabilidade antes da ocorrência do evento agudo.

8.7 Boas práticas, oportunidades e recomendações

As regiões Sudeste e Sul despontam como referências em estrutura de dados, revelando caminhos importantes:

- Integração de dimensões clínicas, sociais e funcionais: A junção de informações de saúde, trabalho, transporte, posse de bens e rotinas diárias permite uma visão mais holística da população, essencial para qualquer modelo de risco.
- Padronização e qualidade de preenchimento: A repetição consistente das variáveis mais relevantes entre diferentes execuções é um indicativo de que os dados seguem um padrão de coleta e preenchimento mais confiável.
- Indícios de bom acesso ao sistema de saúde: A frequência de exames preventivos e diagnósticos laboratoriais como glicemia, colesterol e P.A. revela que essas populações têm mais oportunidades de contato com serviços de atenção primária.

Já para as demais regiões, especialmente Norte, Centro-Oeste e em parte o Nordeste, os dados sugerem:

- Urgência na qualificação da coleta: Treinamento de entrevistadores, revisão de formulários e reforço na supervisão dos dados coletados são medidas cruciais para reduzir ruídos e inconsistências.
- Fortalecimento da atenção primária: A ausência de dados sobre exames e diagnósticos prévios pode indicar baixa frequência de consultas ou de coleta sistemática dessas informações.
- Melhoria nos sistemas de codificação e armazenamento: Estados como GO e TO podem estar enfrentando problemas não apenas de coleta, mas de estruturação interna dos dados, o que exige suporte técnico direto.

As desigualdades regionais no Brasil, já bem conhecidas em termos de saúde, infraestrutura e economia, se refletem também na qualidade e coerência dos dados de saúde pública.

A análise aqui conduzida mostra que não basta a existência de instrumentos de coleta — é preciso garantir preenchimento adequado, padronização e diversidade

informacional para que ferramentas baseadas em aprendizado de máquina possam de fato contribuir com a predição e prevenção de agravos como o AVC.

O avanço em regiões como Sudeste e Sul demonstra o que é possível alcançar. Por outro lado, os desafios das regiões Norte, Centro-Oeste e parte do Nordeste reforçam a necessidade de investimento em vigilância em saúde, capacitação técnica, e melhoria na gestão da informação, como pilares fundamentais para uma política pública baseada em dados.

8.8 Análise das Recorrências de Variáveis por Estado: Convergência Nacional, Heterogeneidade Regional e Implicações Analíticas

A análise da importância relativa das variáveis selecionadas por modelos preditivos de AVC a partir de dados da PNS 2019, por unidade da federação, evidencia uma expressiva consistência nacional em torno de determinadas dimensões clínicas e funcionais. Em praticamente todos os estados, destacam-se, entre as dez variáveis mais recorrentes, os seguintes elementos:

- Q070: Idade ao diagnóstico do AVC;
- Q073: Grau de limitação funcional pós-AVC;
- Q07208 a Q07213: Conjunto de estratégias terapêuticas e de acompanhamento (incluindo dieta, fisioterapia, outras terapias de reabilitação, uso de aspirina, uso de outros medicamentos e acompanhamento com profissional de saúde).

Essa convergência temática evidencia que os modelos construídos com base nesses dados estão fortemente ancorados em variáveis que descrevem o desfecho consolidado do AVC e suas implicações clínicas.

Ainda que, em um primeiro momento, a repetição dessas variáveis possa ser interpretada como um possível indício de *data leakage*, a análise detalhada sugere que se trata de um reflexo da própria natureza da base de dados, construída para captar a condição de saúde no momento da entrevista.

Assim, é mais adequado interpretar esse padrão como uma leitura coerente do fenômeno, e não como um viés técnico. A alta relevância estatística dessas variáveis indica, na verdade, que os modelos estão identificando os principais elementos

estruturantes da trajetória de pacientes que vivenciaram um AVC, alinhando-se, portanto, às expectativas epidemiológicas para tal desfecho.

8.9 Homogeneidade Estrutural e Uniformidade Funcional

A uniformidade observada na importância atribuída a variáveis associadas ao evento AVC e suas repercussões funcionais ocorre mesmo diante da heterogeneidade sociodemográfica entre as regiões brasileiras. O padrão recorrente de variáveis pós-evento (como reabilitação e limitação funcional) sugere que, independentemente do estado, a resposta institucional e individual ao AVC se estrutura em torno de eixos similares.

Esse comportamento revela uma homogeneidade estrutural no preenchimento das informações, possivelmente decorrente de protocolos padronizados de coleta da PNS e de uma compreensão comum, por parte dos entrevistados, sobre os elementos centrais da vivência com o AVC.

Em reforço a essa constatação, observou-se que em todos os entes federativos analisados — com exceção de GO e TO, cujos dados apresentaram inconsistências técnicas — as oito primeiras variáveis mais relevantes foram idênticas em todas as execuções e rodadas de modelo. São elas:

- Faz atualmente por causa do derrame (ou AVC) toma outros medicamentos?
- Faz acompanhamento regular com profissional de saúde?
- Que idade o(a) Sr(a) tinha no primeiro diagnóstico do derrame (ou AVC)?
- Em geral, em que grau o derrame (ou AVC) limita as suas atividades habituais (tais como trabalhar, realizar afazeres domésticos etc.)?
- Faz atualmente por causa do derrame (ou AVC) dieta?
- Faz atualmente por causa do derrame (ou AVC) toma aspirina regularmente?
- Faz atualmente por causa do derrame (ou AVC) fisioterapia?
- Faz atualmente por causa do derrame (ou AVC) outras terapias de reabilitação?

A única exceção pontual refere-se aos estados de Rondônia (RO) e Roraima (RR), onde a variável relativa às “outras terapias de reabilitação” não apareceu de

forma sistemática em todas as execuções. Ainda assim, as demais sete variáveis permaneceram, reforçando o padrão nacional.

Essa constância indica que os modelos estão captando não apenas o impacto clínico do AVC, mas também aspectos centrais da resposta terapêutica e institucional ao evento — elementos que compõem a espinha dorsal da vivência pós-AVC no Brasil. A presença reiterada dessas variáveis sinaliza um alto grau de completude e padronização da coleta de dados nestes campos, tornando-os referências confiáveis na análise do fenômeno.

Estados como Acre, Amazonas, Ceará, Maranhão, Minas Gerais, Paraná, Rio Grande do Sul e São Paulo compartilham exatamente as mesmas dez variáveis mais relevantes, o que reforça a consistência nacional na forma como o fenômeno é retratado.

8.10 Casos de Variabilidade e Padrões Regionais

Em contrapartida, unidades da federação com menor densidade populacional, como Roraima e Amapá, apresentaram maior variabilidade nas variáveis selecionadas. Nesses casos, observa-se a emergência de itens associados à autonomia funcional (como dificuldades para realizar tarefas diárias: alimentar-se, tomar banho, fazer compras), que não aparecem com a mesma frequência em estados maiores (Rosa *et al*., 2018; Vaz *et al*., 2020).

Essa oscilação pode decorrer de dois fatores:

- Substituição dos campos principais por variáveis funcionais análogas, frente à baixa incidência de casos confirmados de AVC;
- Presença de perfis mais severos de limitação funcional, que amplificam o peso estatístico dessas variáveis nas correlações analisadas.

8.11 Contextualização de Variáveis de Base

Além das variáveis diretamente relacionadas ao evento AVC, determinados campos surgem de forma recorrente nos modelos como elementos estruturantes do perfil individual. Entre eles, destacam-se:

J007: presença de doenças crônicas — como hipertensão e diabetes, condições reconhecidas como importantes fatores de risco modificáveis para AVC e

que influenciam não apenas a probabilidade de ocorrência do evento, mas também a gravidade e o prognóstico.

C008: idade do morador — indicador demográfico fundamental, pois a idade é um dos preditores mais consistentes para AVC em diferentes populações, além de se correlacionar com a prevalência de múltiplos fatores de risco.

C00703: ano de nascimento — variável diretamente associada à idade, mas que também permite análises por coortes geracionais, possibilitando identificar padrões históricos e contextuais que podem influenciar a exposição a riscos.

Embora essas variáveis não descrevam diretamente o desfecho final (ocorrência do AVC), sua presença constante entre as mais relevantes nos modelos indica seu valor como variáveis de pano de fundo. Elas funcionam como pilares na estratificação de risco, auxiliando a segmentar a população em subgrupos com diferentes níveis de vulnerabilidade e permitindo compreender como fatores estruturais — como envelhecimento populacional, acúmulo de comorbidades e transições epidemiológicas — moldam o cenário de saúde.

Na prática, o uso combinado dessas variáveis com outras de natureza clínica, funcional ou socioeconômica fortalece a capacidade dos modelos preditivos, pois oferece um quadro contextual mais robusto. Isso é particularmente relevante em estratégias preventivas, onde a identificação de grupos de maior risco antes do evento possibilita ações direcionadas, como intensificação do rastreamento, aconselhamento sobre hábitos de vida e monitoramento contínuo de condições crônicas.

8.12 Implicações Metodológicas e Estratégicas

A análise conjunta da estrutura dos dados e do desempenho dos modelos evidencia um cenário de potencialidades e limitações distintas. De forma geral:

- Os modelos atuais atuam predominantemente como classificadores de casos já confirmados de AVC, em vez de ferramentas preditivas baseadas em fatores antecedentes ao evento;
- A ausência de variáveis preventivas — como histórico familiar, hábitos de vida ou resultados de exames laboratoriais prévios — compromete a capacidade de estimar o risco futuro de forma acurada;

- Por outro lado, a padronização nacional das respostas pós-evento viabiliza análises consistentes sobre o impacto funcional do AVC e sobre a resposta institucional subsequente.

Alguns estados, como GO e TO, apresentam particularidades que merecem destaque. A inclusão de GO nos modelos do Centro-Oeste e de TO nos modelos do Norte gerou efeitos disruptivos: redução na consistência das variáveis selecionadas, aumento do ruído estatístico e ocorrência de anomalias nas matrizes de correlação. Esses achados reforçam a necessidade de auditoria técnica aprofundada para avaliar a completude, a coerência e a qualidade da base de dados nesses estados.

A análise desagregada por unidade federativa confirma a robustez estrutural dos dados da PNS 2019 no que diz respeito aos desfechos pós-AVC, especialmente em variáveis de funcionalidade, idade e reabilitação. Embora os modelos não se mostrem ideais para previsão prospectiva de risco, sua utilidade na caracterização e categorização dos perfis pós-evento permanece inequívoca, fornecendo subsídios relevantes para políticas de reabilitação e acompanhamento.

8.13 Análise Detalhada dos Padrões nas Questões com Maior Pontuação por Estado Brasileiro

Na análise inicial das dez variáveis mais relevantes por estado na PNS 2019 (Top-10), observou-se que as primeiras posições tendem a ser altamente uniformes entre os estados brasileiros, com forte predominância de variáveis diretamente relacionadas ao AVC, como idade ao diagnóstico, uso de medicamentos, acompanhamento de saúde e grau de limitação funcional.

Essa homogeneidade, embora esperada diante do foco clínico do levantamento, limita a capacidade de identificar padrões regionais mais amplos e elementos contextuais que influenciam o impacto e a gestão do AVC em diferentes realidades sociais, econômicas e culturais.

Por isso, para uma visão mais rica e localizada, optou-se por ampliar a análise para as trinta variáveis com maior pontuação por estado (Top-30). Essa abordagem permite revelar características singulares de cada unidade da federação, incluindo hábitos de vida, condições domiciliares, comportamentos preventivos, acesso a serviços de saúde, saúde bucal, consumo alimentar, práticas familiares de cuidado e

até situações de vulnerabilidade social. É a partir dessa lente mais abrangente que se constrói a análise detalhada a seguir.

A análise das 30 variáveis com maior pontuação por estado revela uma teia complexa e multifacetada da realidade brasileira no que se refere à saúde, funcionalidade, hábitos de vida e contexto socioeconômico.

Embora o foco da pesquisa esteja centrado no AVC, muitas variáveis destacadas extrapolam esse eixo, refletindo condições estruturais e comportamentais que podem influenciar direta ou indiretamente a saúde da população. A seguir, discutem-se os principais padrões identificados, suas possíveis causas e interpretações.

8.14 Saúde Bucal: Padrão Universal com Intensidade Regional

Entre as 30 variáveis com maior pontuação em praticamente todos os estados, destacam-se consistentemente indicadores ligados à condição bucal dos entrevistados, especialmente:

- Perda de dentes permanentes superiores (U02302) e inferiores (U02402);
- Uso de prótese dentária (U02501), incluindo dentaduras, pontes e implantes;
- Frequência de escovação com escova de dentes (U00101) ou uso de enxaguantes bucais (U00207);
- Local da última consulta odontológica (U01002).

Essas variáveis, embora não façam parte direta do manejo clínico do AVC, têm aparecido sistematicamente como altamente pontuadas nas análises por estado, especialmente nas regiões Norte, Nordeste e Centro-Oeste.

A inclusão recorrente dessas questões indica que a saúde bucal, apesar de sua aparente “periferia” em relação a doenças cardiovasculares, funciona como um marcador robusto de condições estruturais de saúde: acesso, autocuidado, educação em saúde, renda e organização do sistema de atenção primária (Cheng *et al*., 2018).

Além disso, a perda dentária e a necessidade de próteses não são eventos isolados, mas frequentemente refletem processos cumulativos de negligência sistêmica, falta de acesso a prevenção e tratamentos e ausência de políticas públicas eficazes ao longo da vida (Cheng *et al*., 2018; Shiga *et al*., 2020). Essas perdas

também podem comprometer a mastigação, a nutrição e, indiretamente, a saúde cardiovascular — estabelecendo, portanto, um elo funcional com desfechos como o AVC.

A forte presença das variáveis U02302, U02402 e U02501 nestas regiões pode ser explicada por uma série de fatores interligados:

- Baixa cobertura de atenção básica odontológica – embora o SUS ofereça atenção odontológica, a cobertura é muito desigual. Estados como Acre, Roraima e Maranhão ainda enfrentam carência de equipes de Saúde Bucal na Estratégia Saúde da Família (Schmidt *et al*., 2018).
- Menor escolaridade média e menor renda domiciliar per capita, o que reduz a priorização de cuidados preventivos e dificulta o acesso a serviços particulares (Egídio de Souza *et al*., 2015).
- Prevalência de práticas alimentares ricas em açúcar e pouca conscientização sobre higiene bucal, especialmente em áreas rurais e ribeirinhas (Egídio de Souza *et al*., 2015).

Esses fatores levam não apenas à perda dentária precoce, como também à normalização da perda de dentes como parte do envelhecimento, o que perpetua o ciclo.

No Centro-Oeste, particularmente em estados como Mato Grosso e Mato Grosso do Sul, a presença dessas variáveis pode estar associada à combinação de:

- Áreas rurais vastas e de difícil acesso;
- Grupos populacionais indígenas ou de comunidades tradicionais com pouca inserção no modelo urbano de atenção à saúde;
- Presença de serviços públicos com cobertura limitada fora dos centros urbanos (Da Silva *et al*., 2020).

Curiosamente, essas variáveis também aparecem entre as mais pontuadas em estados com infraestrutura de saúde mais robusta, como São Paulo e o Distrito Federal.

Nesses estados, a presença das questões pode não refletir a precariedade, mas sim o peso da saúde bucal como componente importante no modelo de atenção primária à saúde. A boa cobertura do SUS e o acesso à informação podem fazer com que a saúde bucal seja mais bem registrada e mais discutida durante os atendimentos.

Além disso, é possível que populações mais idosas e com histórico de acesso desigual ao longo da vida ainda carreguem o ônus da perda dentária, o que reforça o aparecimento dessas variáveis, mesmo em contextos mais desenvolvidos (Bomfim; Cascaes; De Oliveira, 2021).

Outro aspecto relevante é que a saúde bucal impacta diretamente a autoestima, a comunicação e o convívio social. A perda de dentes ou o uso de próteses precárias pode causar constrangimento, isolamento e até abandono de atividades profissionais ou sociais — elementos que, por sua vez, podem agravar fatores de risco indiretos para AVC, como sedentarismo, depressão e má alimentação (Matsuyama *et al*., 2021).

É possível que os respondentes atribuam alta relevância às questões bucais por perceberem um vínculo claro entre “sentir-se saudável” e “ter a boca saudável”, especialmente em contextos em que a saúde é avaliada mais pelo bem-estar funcional do que por indicadores clínicos abstratos (Kusama *et al*., 2025; Reissmann *et al*., 2013).

A saúde bucal também se correlaciona com doenças crônicas. Já há evidências científicas robustas associando doença periodontal com inflamação sistêmica crônica, o que pode aumentar o risco de aterosclerose e eventos vasculares. Assim, a presença sistemática dessas variáveis pode indicar, mesmo que de forma indireta, um marcador biológico de risco cardiovascular (Carrizales-Sepúlveda *et al*., 2018; Priyamvara *et al*., 2020; Sanz *et al*., 2020).

8.15 Autonomia Funcional e Atividades da Vida Diária (AVDs): Reflexo da Sobrevida, Suporte Social e Expectativas Regionais

Um dos conjuntos mais recorrentes entre as 30 principais variáveis por estado está ligado à autonomia funcional e à realização das chamadas atividades da vida diária (AVDs). Isso inclui desde tarefas básicas como:

- Comer sozinho (K001)
- Tomar banho (K004)
- Vestir-se (K010)
- Sentar-se ou levantar-se da cadeira (K019)
- Ir ao banheiro (K007)
- Sair de casa utilizando transporte (K034)

- Fazer compras (K022)
- Administrar finanças (K025)
- Tomar medicações sem ajuda (K028)

Essas variáveis aparecem com destaque em todos os estados da Região Sul e Sudeste, em vários do Norte (RR, AP, AM) e, em menor escala, em outros do Centro-Oeste e Nordeste. Sua recorrência sugere que a funcionalidade não é apenas um desfecho clínico importante pós-AVC, mas também um ponto de convergência entre envelhecimento, infraestrutura de apoio, saúde familiar e cultura de cuidado (Souto *et al*., 2024; Vieira *et al*., 2023).

Sul e Sudeste – Sobrevida, Envelhecimento e “Expectativas Funcionais”: Estados como RS, SC, PR, SP, MG e RJ registram com frequência indicadores relacionados à autonomia funcional. Uma leitura inicial pode associar isso ao perfil demográfico dessas regiões, que concentram uma população proporcionalmente mais idosa. Isso se deve a:

- Maior expectativa de vida;
- Menores taxas de natalidade;
- Melhor controle e tratamento das doenças crônicas, inclusive o AVC.

Nesses locais, com maior acesso a serviços médicos e reabilitação, os indivíduos sobrevivem a eventos graves como o AVC com maior frequência — mas também com maior risco de sequelas motoras, sensoriais ou cognitivas. Portanto, a presença das variáveis de AVDs indica prevalência de limitações funcionais na população com sobrevida prolongada (Francisco *et al*., 2025a).

Além disso, populações mais escolarizadas e urbanizadas tendem a ter maior percepção e menor tolerância à perda de autonomia. Enquanto em contextos mais rurais ou precários a dependência funcional pode ser incorporada como parte do envelhecer, em regiões mais ricas e urbanas ela é vivenciada como uma ruptura crítica da qualidade de vida. Isso também contribui para maior destaque dessas variáveis em estados mais desenvolvidos (Pekçetin *et al*., 2025).

Norte – Ausência de suporte: A presença marcante dessas variáveis em estados como RR, AP e AM exige uma interpretação distinta. Nesses casos, a prevalência de idosos é menor, mas os eventos agudos tendem a gerar maior impacto funcional (De Oliveira Cacho *et al*., 2022), por motivos como:

- Atraso no diagnóstico e tratamento de AVCs (menor rede de urgência);
- Baixa oferta de fisioterapia e reabilitação ambulatorial;
- Barreiras geográficas ao acesso a serviços (rios, florestas, longas distâncias);
- Ausência de dispositivos auxiliares (cadeira de rodas, andador, prótese, etc.).

A presença dessas variáveis no Norte pode indicar que mesmo eventos cerebrovasculares de menor gravidade geram perda funcional significativa e sustentada, justamente porque não há suporte suficiente para reabilitação precoce (Silva *et al*., 2024). Além disso, há menor infraestrutura para apoiar pessoas com limitações, como transporte público adaptado, calçadas acessíveis e cuidadores especializados (Alves *et al*., 2025).

Nesse contexto, a avaliação das AVDs ganha ainda mais relevância, pois não se limita ao plano clínico: ela traduz o impacto real da doença no cotidiano e na dignidade dos indivíduos (Fatema *et al*., 2022; Kim; Kim; Kim, 2014). Perder a capacidade de sair de casa sozinho, cuidar das próprias finanças ou realizar a higiene pessoal representa um abalo profundo à autonomia e ao senso de identidade — sobretudo em regiões onde depender de ajuda não é uma escolha, mas uma imposição estrutural (Muren; Hütler; Hooper, 2008).

É por isso que essas variáveis aparecem com tanta força: elas são o elo entre a saúde e a vida vivida. Além disso, muitos pacientes e familiares percebem a funcionalidade como o principal critério de recuperação, mais do que indicadores laboratoriais ou diagnósticos (Nguyen *et al*., 2024b; Yao *et al*., 2023).

Outro elemento que emerge da análise é a influência da rede de apoio familiar e comunitário. Em estados onde há forte presença de variáveis como K02101 (quem presta ajuda para realizar atividades) ou K03601/02 (se essa ajuda é paga), nota-se uma preocupação com quem cuida do paciente funcionalmente dependente. Isso pode indicar, por exemplo:

- Sul e Sudeste: redes formais de cuidado ou cuidadores remunerados mais comuns (Predebon *et al*., 2021);
- Nordeste e Norte: cuidado prestado por familiares, geralmente mulheres, sem remuneração — o que gera sobrecarga e perpetua desigualdades de gênero (Morais *et al*., 2012).

Há também uma associação indireta entre dificuldades em AVDs e renda domiciliar per capita (VDF003)(Ghoneem *et al*., 2022; Oliveira-Kumakura *et al*., 2023b; Santos *et al*., 2022b). Indivíduos com menor renda tendem a:

- Iniciar o tratamento de forma mais tardia;
- Ter menor acesso a dispositivos auxiliares (bengala, prótese, cadeira de rodas);
- Depender exclusivamente do SUS para cuidados pós-alta;
- Viver em moradias com menor acessibilidade física (banheiros inadequados, escadas, ausência de corrimãos).

Em regiões com maior desigualdade interna (como RJ, BA ou DF), a funcionalidade aparece como marcador de exclusão, pois a perda de AVDs é mais incapacitante quando não há recursos financeiros ou estruturais para compensá-la (Coté *et al*., 2025; Seleme *et al*., 2024).

Esses achados mostram que, para além de tratar o AVC, o sistema de saúde precisa oferecer suporte duradouro para manutenção da funcionalidade. Isso inclui:

- Reabilitação precoce e domiciliar;
- Apoio a cuidadores familiares;
- Adaptação da infraestrutura urbana e domiciliar;
- Promoção de tecnologias assistivas com acesso público;
- Programas de reinserção social e ocupacional.

A ausência de tais medidas leva à cronificação da dependência e à desestruturação familiar — um ciclo que começa com o AVC, mas se mantém pela ausência de políticas públicas efetivas (Day *et al*., 2023; De Campos *et al*., 2017; Oliveira-Kumakura *et al*., 2023c).

8.16 Alimentação e Comportamento Alimentar: Entre a Praticidade Econômica e o Estilo de Vida Urbano

Entre as variáveis com maior pontuação por estado, destaca-se a recorrência de questões relacionadas aos hábitos alimentares cotidianos (Stolses *et al*., 2023), especialmente:

- Consumo regular de refrigerantes (P02002)

- Frequência de ingestão de alimentos doces, como chocolates, balas e biscoitos recheados (P02501)
- Substituição de refeições principais por lanches rápidos como salgados, pizzas ou sanduíches (P02602)
- Quantidade de doses de bebida alcoólica consumidas por ocasião (P029)

Essas variáveis aparecem de forma marcante em estados do Nordeste (como BA, AL e PE), Sudeste (SP e RJ) e Norte (RR, AM e PA), evidenciando que a alimentação, enquanto fator de risco indireto para doenças crônicas, também reflete padrões econômicos, culturais e estruturais profundamente distintos entre as regiões (Da Costa Louzada *et al*., 2023; Stolses *et al*., 2023).

Nordeste – Nos estados do Nordeste, especialmente em contextos de maior vulnerabilidade socioeconômica, observa-se um aumento expressivo no consumo de refrigerantes e alimentos ultraprocessados, o que pode estar relacionado à busca por alimentos de baixo custo, alto valor calórico e fácil acesso, cuja presença vem se expandindo nas periferias urbanas. Entre 2008 e 2018, a participação de ultraprocessados no total de calorias aumentou em mais de 3 pontos percentuais na região Nordeste, justamente entre os grupos de menor renda e escolaridade, apontando para a expansão desse padrão alimentar em segmentos mais vulneráveis da população (Da Costa Louzada *et al*., 2023).

A substituição de refeições completas por alimentos industrializados ou lanches rápidos reflete tanto as restrições financeiras quanto a insuficiência de políticas públicas de segurança alimentar voltadas para adultos, como restaurantes populares ou programas equivalentes. Esse cenário evidencia uma transição nutricional incompleta, em que a população passa a consumir produtos industrializados em maior proporção, mas sem acessar a variedade e o equilíbrio característicos de dietas saudáveis (Da Costa Louzada *et al*., 2023).

Sudeste – Ritmo Urbano, Oferta de Ultraprocessados e Cultura de Conveniência: Nos grandes centros urbanos do Sudeste, como São Paulo e Rio de Janeiro, o consumo elevado de ultraprocessados reflete menos uma limitação de recursos e mais o ritmo de vida acelerado e uma cultura de conveniência. Estudos demonstram que o Sudeste apresenta uma das maiores participações de calorias provenientes de alimentos ultraprocessados — como refrigerantes, biscoitos e outros

produtos industrializados — chegando a mais de 20 % em muitos domicílios urbanos e áreas metropolitanas, incluindo São Paulo (22,38 %) (Da Costa Louzada *et al*., 2023).

A urbanização intensa (com cerca de 97 % da população urbana na região) e a forte presença desses produtos, aliados ao marketing agressivo e à falta de tempo para preparar refeições caseiras, promovem a normalização da substituição de refeições por lanches rápidos e bebidas açucaradas. Nesse contexto, a alimentação deixa de ser apenas uma necessidade fisiológica e passa a ser moldada pela pressão do tempo, pelo consumo impulsivo e por ambientes que favorecem escolhas pouco saudáveis (De Paula Costa *et al*., 2021).

Nos estados do Norte, como RR, AM e PA, a substituição de refeições completas por lanches também aparece com destaque, mas por dinâmicas que incluem desigualdades regionais de acesso e padrões de aquisição. Evidências da Pesquisa de Orçamentos Familiares 2017–2018 mostram que, no Norte, pequenos mercados são mais utilizados do que supermercados para comprar alimentos — e essa preferência se acentua entre famílias com insegurança alimentar moderada/grave; nesses contextos, além de arroz/feijão e proteínas, ultraprocessados também entram mais na cesta, o que ajuda a entender a conveniência e a praticidade como motores dessas escolhas (De Oliveira *et al*., 2024b). Em TO, por exemplo, 63,3% dos domicílios urbanos apresentaram insegurança alimentar e menor disponibilidade calórica nos domicílios mais vulneráveis, reforçando que a restrição material e a baixa diversidade de oferta impactam diretamente o padrão alimentar (Schott *et al*., 2020).

Estudos sobre o ambiente alimentar e sobre segurança alimentar e nutricional mostram que as escolhas não dependem só do indivíduo, mas também de onde e como os alimentos são ofertados (tipo de comércio, distância, preço, propaganda) e das políticas públicas que garantem ou não o acesso regular a comida saudável; no Norte, a maior dependência de pequenos comércios e a distribuição territorial menos densa tornam o acesso a alimentos saudáveis mais difícil, o que pode favorecer itens prontos/de longa duração quando o acesso a frescos é limitado (De Oliveira *et al*., 2024b).

Em paralelo, análises nacionais (VIGITEL, 2012–2022) mostram que as desigualdades regionais não se manifestam do mesmo modo para todos os marcadores.

Para frutas e hortaliças, Norte e Nordeste mantêm as menores prevalências de consumo regular (≥ 5 dias/semana) — ~40% no Norte (40,4% em 2012–2020; 40,2% em 2021–2022) e ~40% no Nordeste (39,9%; 40,0%) — versus ~51% no Sudeste (50,8%; 51,2%) e ~59–61% no Sul (61,1%; 59,2%).

Já nas bebidas açucaradas, o desenho é distinto: no período pré-pandemia, as maiores prevalências ocorreram no Sul (até 23,2%), com menor probabilidade de consumo regular no Norte e maior no Centro-Oeste, Sudeste e Sul. Durante a pandemia, houve queda geral e o Nordeste apresentou o menor valor. Esses contrastes indicam que o baixo consumo de frutas e hortaliças no Norte/Nordeste convive com prevalências mais altas de bebidas açucaradas no Sul/Sudeste, compondo “perfis distintos” por região e refletindo condições estruturais de acesso e ambiente alimentar. (Vieira *et al*., 2025).

Reforçando o papel do ambiente alimentar, dados da Pesquisa sobre Orçamento Familiar 2017–2018 mostram que, no Norte (e no Nordeste), os domicílios — sobretudo com insegurança alimentar moderada/grave — dependem mais de pequenos mercados do que de supermercados para adquirir alimentos, o que afeta disponibilidade, preço e variedade de itens frescos e pode empurrar escolhas para produtos prontos/de longa duração (De Oliveira *et al*., 2024b).

Esses padrões não se reduzem à “escolha individual”: eles espelham como o acesso ao alimento e o tempo são socialmente estruturados. Em contextos de pobreza, há uma busca por calorias acessíveis e pontos de venda próximos; nos centros urbanos, a pressão do tempo e a conveniência normalizam lanches rápidos e bebidas prontas. Em ambos os casos, trata-se de um modelo que favorece doenças crônicas não-transmissíveis — como hipertensão, obesidade e dislipidemias —, relação também destacada em estudos de disponibilidade domiciliar e consumo no país (De Oliveira *et al*., 2024b; Schott *et al*., 2020). Durante a pandemia, registrou-se ainda aumento do consumo de bebidas alcoólicas em meses de quarentena, sugerindo interação entre estresse/ambiente urbano e padrões de consumo — mais um fator de risco cardiovascular relevante no cenário regional (Vieira *et al*., 2025).

8.17 Padrões Reprodutivos, Gênero e Saúde Sexual: Variáveis que Refletem Ciclos de Vida, Cultura e Políticas Públicas

Em meio às 30 variáveis com maior pontuação por estado, um grupo específico chama atenção por tratar de aspectos frequentemente negligenciados nas análises epidemiológicas tradicionais: as questões reprodutivas, sexuais e de gênero. São variáveis que abordam:

- Número total de filhos vivos (P040)
- Uso de preservativo nas últimas relações sexuais (Q029)
- Frequência de atividade sexual (Q011)
- Se ainda menstrua (P051)
- Se teve relação sexual nos últimos 12 meses (Q010)

Esses indicadores são capturados pela PNS e têm sido usados para descrever padrões e desigualdades de comportamento sexual no Brasil, incluindo baixa prevalência de uso consistente de preservativos em adultos sexualmente ativos (Gomes; De Souza Lopes, 2022).

Essas variáveis aparecerem com destaque em estados como PA, RN, PR e RO é coerente com a heterogeneidade demográfica do país: a transição da fecundidade ocorreu em ritmos e momentos distintos entre regiões e municípios, com atrasos persistentes no Norte e Nordeste em relação ao Sudeste e Sul (Quaresma *et al*., 2023). Além disso, análises históricas mostram que o declínio da fecundidade no Brasil foi desigual no tempo e no espaço, o que ajuda a entender por que paridade (número de filhos) continua variando muito entre subpopulações (Potter *et al*., 2010).

Norte e Nordeste – Fecundidade Histórica e Diversidade Etária: Nos estados do Norte e Nordeste, a presença dessas variáveis pode refletir taxas de fecundidade historicamente mais altas e uma transição demográfica mais recente/heterogênea — especialmente em municípios fora dos grandes centros — contribuindo para maior variação de paridade e idade reprodutiva (Quaresma *et al*., 2023). Essa história reprodutiva importa porque maior paridade tem sido associada, em mulheres na pós-menopausa, a maior risco de AVC em estudos populacionais (De Havenon *et al*., 2021a; Zhang *et al*., 2015).

Além da paridade, complicações obstétricas (p. ex., pré-eclâmpsia) deixam risco residual de doenças cardiovasculares e AVC que persiste décadas após a

gestação. Achados de coortes de longo prazo sugerem maior probabilidade de AVC na vida adulta tardia em mulheres com antecedente de pré-eclâmpsia, mesmo após ajuste de fatores ao longo do curso de vida (De Havenon *et al*., 2021a; Wu *et al*., 2017a). Sínteses sistemáticas mais amplas confirmam que os transtornos hipertensivos da gestação elevam o risco de AVC isquêmico e de qualquer AVC na vida futura (Poorthuis *et al*., 2017).

A variável “ainda menstrua” sinaliza a posição da mulher no pico da transição menopausal, um período em que biomarcadores metabólicos e vasculares se deterioram de forma acelerada, independentemente do envelhecimento cronológico. A American Heart Association (AHA) destaca a transição da menopausa como janela de aceleração do risco cardiovascular e um momento oportuno para prevenção (El Khoudary *et al*., 2020). Além do estágio, idade na menopausa também importa: meta-análise recente mostrou que menopausa precoce (<43–45 anos) se associa a maior risco de AVC isquêmico, e coortes europeias observaram mais que o dobro do risco de AVC total quando a menopausa ocorre <40 anos em comparação a 50–54 anos (Welten *et al*., 2021).

Em contextos de vulnerabilidade social, a baixa adoção do preservativo é um marcador de exposição a Infecções Sexualmente Transmissíveis (ISTs) e de barreiras a políticas de saúde sexual e reprodutiva; a PNS descreve esses padrões e sua variação por sexo, escolaridade e região, validando o uso de preservativos e a atividade sexual recente como variáveis relevantes para estratificar risco e acesso a cuidado (Gomes; De Souza Lopes, 2022).

Sul e Sudeste – Saúde da Mulher, Planejamento Familiar e Atividade Sexual na Maturidade: Em estados como o PR, com fecundidade mais baixa e transição demográfica avançada, o destaque para variáveis reprodutivas e sexuais indica um foco crescente na saúde da mulher na meia-idade e no envelhecimento. Isso inclui o monitoramento do climatério/menopausa, a promoção do uso de preservativos — inclusive em parcerias estáveis — e o acompanhamento da vida sexual de mulheres mais velhas, um grupo sub-representado em inquéritos tradicionais, mas abrangido na PNS 2019 (que analisou inclusive pessoas idosas e estratificou o uso de preservativos por sexo e situação conjugal) (Gomes; De Souza Lopes, 2022).

Nesses contextos, variáveis como P051 (“ainda menstrua?”) e Q010 (“teve relação nos últimos 12 meses?”) são úteis para captar dimensões de saúde sexual na maturidade e mudanças hormonais relevantes ao risco cardiovascular. Evidências

sintetizadas em revisões e coortes mostram que a transição menopausal e o tipo/timing da menopausa se associam a desfechos vasculares — por exemplo, meta-análises e estudos de coorte investigaram idade à menopausa (natural versus cirúrgica) e risco de AVC, com sinal para diferenças por tipo de menopausa e para riscos específicos (p.ex., maior risco de AVC hemorrágico em menopausa mais tardia; efeitos distintos após menopausa cirúrgica) (Welten *et al*., 2021). Além disso, sínteses de fatores sexo-específicos relatam que complicações hipertensivas da gestação e ooforectomia elevam o risco de AVC ao longo da vida, reforçando a importância do histórico reprodutivo na avaliação de risco em mulheres na pós-menopausa (Poorthuis *et al*., 2017).

O uso de preservativo (Q029) aparece como marcador programático e de comportamento: na PNS 2019, a não utilização no último ato sexual foi substancialmente maior entre pessoas casadas/que coabitam (≈75% em ambos os sexos) e cresce com a idade, o que sustenta a necessidade de campanhas voltadas também a parcerias estáveis e à população mais velha (Gomes; De Souza Lopes, 2022). Estudos nacionais com a PNS 2019 mostram ainda baixa prevalência de uso consistente de preservativos na população adulta, com piores indicadores entre mulheres, pessoas com maior idade e menor escolaridade — além de desigualdades regionais —, apontando lacunas de educação em saúde e de alcance das ações preventivas (Trindade *et al*., 2021).

Esse conjunto de variáveis também captura dimensões sutis, porém cruciais, de gênero e autonomia. Evidências brasileiras indicam maior prevalência de atividade sexual desprotegida em mulheres, especialmente as mais velhas, e discutem a necessidade de empoderamento para negociação do uso do preservativo — um componente que conecta Q029 a aspectos de poder de decisão e educação em saúde (De Souza *et al*., 2022).

Por fim, o histórico reprodutivo mais amplo permanece relevante mesmo em regiões de baixa fecundidade: a literatura observa associações entre alto número de gestações e risco aumentado de AVC em populações fora do Brasil, sugerindo que paridade ao longo do curso de vida pode influenciar a carga vascular na maturidade — um racional que justifica considerar “número de filhos vivos (P040)” em modelos explicativos e de vigilância em saúde (Zhang *et al*., 2015).

Assim, nas regiões Sul/Sudeste, o aparecimento de P051, Q010, Q011 e Q029 entre variáveis relevantes é coerente com uma agenda de saúde da mulher voltada

ao climatério/menopausa, à sexualidade ativa na meia-idade e ao reforço do uso de preservativos também em relacionamentos estáveis — aspectos que têm implicações diretas para prevenção de ISTs e para o manejo do risco cardiovascular e cerebrovascular ao longo do envelhecimento feminino (Gomes; De Souza Lopes, 2022).

8.18 Renda e Estrutura Familiar: Condições Sociais como Variáveis-Chave na Determinação da Saúde

Entre as variáveis com maior pontuação nos estados, poucas aparecem com tanta frequência e regularidade quanto as que refletem condições sociais básicas, como:

- Rendimento domiciliar per capita (VDF003)
- Ano de nascimento (C00703)
- Idade atual (C008)
- Número de moradores no domicílio (D001)
- Número de pessoas por dormitório (D005)

A literatura mostra, de forma consistente, que determinantes sociais moldam risco, acesso e desfechos em saúde, inclusive para AVC e seus fatores de risco, o que justifica a presença dessas variáveis nos modelos em praticamente todas as UF brasileiras (Francisco *et al*., 2025b; Pantoja-Ruiz *et al*., 2025).

A VDF003 desponta entre as mais relevantes porque o gradiente socioeconômico no Brasil se traduz em diferenças de risco e de mortalidade por AVC, independentemente do PIB per capita: em análise longitudinal por estados (2002–2009), maior desigualdade de renda associou-se a maior mortalidade por AVC; reduzir 10 pontos no índice de Gini estimou-se ligado a 18% de queda na mortalidade por AVC (Vincens; Stafström, 2015c).

Além disso, desigualdades de acesso a serviços e medicamentos persistem mesmo sob o SUS, com maior necessidade não atendida entre grupos de menor renda e escolaridade; nesse contexto, a expansão qualificada da Atenção Primária à Saúde (APS), especialmente por meio da Estratégia Saúde da Família (ESF) – principal modelo de organização da APS – em áreas pobres do Rio de Janeiro

mostrou-se capaz de reverter parte desse quadro, estando associada a reduções substanciais de mortalidade e iniquidades. (Coube *et al*., 2023a, 2023b).

Essa mesma clivagem socioeconômica aparece em perfis comportamentais e de risco cardiovascular quando se comparam populações de favelas (aglomerados subnormais, segundo classificação do IBGE) versus não-favelas em grandes centros (Chan *et al*., 2022).

O rendimento familiar influencia, de forma mensurável:

- Alimentação e escolhas de risco (maior exposição a padrões alimentares não saudáveis em contextos socialmente vulneráveis) (Chan *et al*., 2022).
- Consultas e exames preventivos e obtenção de medicamentos (necessidades não atendidas maiores entre os mais pobres) (Coube *et al*., 2023a, 2023b).
- Capacidade de reorganizar a rotina após eventos agudos (como AVC), dada a dependência de cuidadores familiares e o impacto sobre sua carga e qualidade de vida (Da Silva; Boery, 2021; Predebon *et al*., 2021b).

Em suma, renda não é apenas um dado econômico: é um determinante direto da prevenção, tratamento e recuperação (Pantoja-Ruiz *et al*., 2025).

Variáveis como D001 (moradores) e D005 (pessoas/dormitório) são proxies (variável que não mede diretamente o fenômeno de interesse, mas que serve como substituto ou indicador indireto dele) de condições domiciliares. Evidências no Brasil indicam que superlotação aumenta risco de doenças infecciosas (p.ex., respiratórias e tuberculose) e potencializa transmissão em domicílios vulneráveis; em São Paulo, maior lotação associou-se a pior saúde respiratória infantil, e, em estudos nacionais, a superlotação mediu parte do efeito da pobreza/inequidade na incidência de tuberculose. Em populações vulneráveis, COVID-19 também se propagou mais em domicílios superlotados (Cardoso *et al*., 2004; Coelho *et al*., 2024; Pelissari; Diaz-Quijano, 2017).

Para além de infecções, a superlotação relaciona-se a pior sono/descanso e pior saúde mental em diferentes faixas etárias, fatores com repercussão cardiovascular indireta (Johnson *et al*., 2015; Sharifi *et al*., 2024).

Ao mesmo tempo, mais moradores podem significar rede de apoio para o cuidado pós-AVC; no Brasil, coortes e ensaios com cuidadores familiares mostram que suporte e intervenções educativas reduzem carga e melhoram a qualidade de vida de quem cuida, com impacto potencial no manejo domiciliar (Da Silva; Boery, 2021; Predebon *et al*., 2021b).

Essa dupla face (risco por superlotação vs. apoio social) explica por que D001 e D005 devem ser interpretadas em conjunto com idade, vínculos e composição familiar (Predebon *et al*., 2021b).

C00703 (ano de nascimento) e C008 (idade) não importam apenas por capturarem o risco basal crescente de AVC; elas identificam coortes expostas a ambientes sanitários, educacionais e econômicos muito distintos ao longo da vida. A população brasileira com mais de 50 anos viveu mudanças rápidas no perfil epidemiológico, na escolaridade e no acesso a serviços, o que se reflete nos padrões de risco atuais; estudos nacionais também ligam desvantagem socioeconômica acumulada ao longo do curso de vida a maior risco cardiovascular (Camelo *et al*., 2015; Lima-Costa *et al*., 2018).

A presença maciça de variáveis sociais nos modelos estatísticos reforça um princípio da epidemiologia contemporânea: a saúde é construída (ou deteriorada) em camadas, sendo a camada social uma das mais determinantes. Essas camadas — renda, moradia, estrutura familiar, trajetória geracional — organizam oportunidades (ou ausências) de informação, escolhas alimentares, tempo para autocuidado, estabilidade emocional e continuidade do tratamento/reabilitação; quando persistentemente desfavoráveis, acumulam-se e emergem como doenças crônicas não-transmissíveis, incapacidades e eventos agudos como o AVC (Hone *et al*., 2020a; Pantoja-Ruiz *et al*., 2025).

8.19 Uso de Serviços de Saúde, Reabilitação e Diagnóstico Precoce: Acesso, Continuidade e Disparidade Regional

Entre as variáveis com maior pontuação por estado, um conjunto de questões se destaca por abordar o uso de serviços de saúde, o diagnóstico precoce de doenças crônicas e o acesso à reabilitação. São variáveis que abrangem:

- Última consulta médica ou odontológica (U00801, U01002)

- Uso de medicamentos contínuos, como aspirina (U07211) ou outros (U07212)
- Realização de exames preventivos: colesterol (Q030) e glicemia (Q032)
- Participação em fisioterapia, psicoterapia, terapias de reabilitação (U07209, U07210)
- Uso de dispositivos auxiliares como bengala, cadeira de rodas, andador (U07401, U07402, U07403)

Evidências nacionais mostram que, no Brasil, utilização e acesso variam fortemente por região, renda e cobertura por plano de saúde, com desigualdades persistentes na prevenção e no tratamento — o que explica o peso estatístico dessas variáveis em praticamente todas as Unidades Federativas (Coube *et al*., 2023b; Palmeira *et al*., 2022).

Norte e Nordeste – Dependência do SUS e Barreiras de Oferta: Em estados como AM e MA, a recorrência dessas variáveis é compatível com maior dependência do SUS e menor cobertura por planos de saúde, além de barreiras geográficas e logísticas típicas de áreas rurais/remotas (distância, transporte), que restringem consultas, exames e reabilitação.

Estudos com a PNS 2013/2019 documentam que a cobertura privada é menor no Norte/Nordeste e que a ter um plano de saúde e a posição socioeconômica estão diretamente associadas à maior utilização de serviços; paralelamente, há “necessidades não atendidas” (serviços e medicamentos) com maior intensidade entre grupos de menor renda. Em reabilitação, estudos multicêntricos e análises nacionais apontam acesso desigual e dificuldades de deslocamento como entraves à adesão (Coube *et al*., 2023b; De Souza Júnior *et al*., 2021).

Nesses contextos, o uso de serviços tende a ser mais reativo (e episódico), orientado pela gravidade e não pela prevenção, o que se traduz em menor probabilidade de controle oportuno de colesterol/glicemia e de início precoce de reabilitação. Estudos com PNS mostram iniquidades na adoção de medidas preventivas e de controle de fatores de risco conforme escolaridade e renda (Patriota; Ko Maung; Marques-Vidal, 2023).

Além disso, a presença de variáveis sobre dispositivos (bengala, andador, cadeira de rodas) pode refletir sequelas residuais associadas a reabilitação tardia/insuficiente. Em amostras brasileiras de sobreviventes de AVC, o uso de

dispositivos de marcha é frequente e correlaciona-se à dependência funcional, reforçando a necessidade de reabilitação adequada e contínua (Caro; Costa; Cezar da Cruz, 2018).

Sudeste e Sul – Envelhecimento, Continuidade do Cuidado e Reabilitação Estruturada: Nos estados do Sudeste/Sul (por exemplo, SP e PR), as mesmas variáveis surgem com destaque, porém em cenário de maior acesso a consultas, exames e terapias regulares (SUS e suplementar). A cobertura por plano é mais elevada nessas regiões, fator associado a mais utilização e menor necessidade não atendida; com populações mais idosas e maior longevidade pós-AVC, continuidade do cuidado e adesão a planos terapêuticos (incluindo reabilitação) tornam-se eixos centrais (Coube *et al*., 2023b; De Souza Júnior *et al*., 2021).

Nesse arranjo, fisioterapia e psicoterapia tendem a integrar projetos terapêuticos ativos (não apenas intervenções tardias). Meta-análises e revisões indicam que intervenções psicossociais e outras não farmacológicas reduzem sintomas depressivos pós-AVC, favorecendo participação e adesão à reabilitação; por sua vez, reabilitação iniciada nas primeiras semanas associa-se a melhores desfechos funcionais, especialmente quando oferecida em unidades de AVC e serviços multiprofissionais organizados (Coleman *et al*., 2017; Deng *et al*., 2017; Langhorne *et al*., 2018; Yi *et al*., 2024).

A regularidade de consultas e exames (como colesterol e glicemia) é um importante marcador de rastreamento e controle de fatores de risco, incluindo hipertensão, diabetes e dislipidemia. Análises da PNS 2019 apontam prevalências relevantes de colesterol alto autorreferido e evidenciam associações com fatores socioeconômicos e demográficos, reforçando a necessidade de acompanhamento contínuo (De Sá *et al*., 2022).

Em amostras com diabetes, estudos nacionais utilizando a PNS documentam lacunas no chamado controle ABC — conjunto de metas que envolve a glicemia, a P.A. e o colesterol. O alcance simultâneo desses indicadores reduz complicações cardiovasculares e representa um parâmetro de qualidade no cuidado ao diabetes, destacando ainda o papel dos determinantes sociais nesse processo (De Sá *et al*., 2022).

Quanto a medicamentos contínuos, há evidências brasileiras de subutilização de prevenção secundária entre pessoas com história de AVC (p. ex., estatinas e antiagregantes) — um indicador indireto de falhas na organização do cuidado e no

acesso — e de benefício consistente do tratamento com estatinas na redução de recorrência e melhora funcional após AVC isquêmico (De Abreu *et al*., 2018; Rodrigues *et al*., 2020; Tramacere *et al*., 2019; Vitturi; Gagliardi, 2020).

8.20 Violência, Vulnerabilidade Social e Cuidado Familiar: Duas Faces de uma Rede Fragilizada

Entre as variáveis com maior pontuação por estado, destacam-se duas frentes ligadas ao ambiente doméstico, mas que revelam realidades contrastantes: de um lado, exposição à violência e ameaça; de outro, envolvimento familiar no cuidado de idosos ou pessoas doentes.

Em MT e RO, a presença de itens sobre agressões físicas/psicológicas e ameaças é coerente com evidências nacionais de violência em domicílio, majoritariamente perpetrada por parceiros íntimos e frequentemente subnotificada; análises com a PNS 2019 estimam que 19,4% das mulheres brasileiras relataram violência em 2019, sobretudo psicológica e dentro de casa, e estudos comparando fontes apontam discrepâncias entre registros (sinal de subcaptação) (De Vasconcelos *et al*., 2025; Stein *et al*., 2023).

Norte e Centro-Oeste – violência “invisível” e redes de proteção quebradas: a inclusão de perguntas sobre violência interpessoal pode refletir vazios de proteção social, baixa vigilância comunitária em áreas rurais/remotas e barreiras à denúncia, especialmente entre mulheres com menor escolaridade e autonomia — padrão observado em recortes da PNS 2019 para áreas rurais (Stochero; Pinto, 2024).

A violência não é apenas origem de sofrimento psíquico: revisões sistemáticas e coortes associam a exposição a violência (incluindo violência por parceiro íntimo) a pior perfil cardiovascular, maior probabilidade de hipertensão, depressão e distúrbios do sono, além de comportamentos de risco (tabagismo, baixa adesão terapêutica). Esses mecanismos psicossociais e neuroendócrinos (eixo estresse-inflamação) ajudam a explicar o elo entre violência crônica e doença cardiovascular/AVC (Suglia; Sapra; Koenen, 2015).

Especificamente para o sono, meta-análises recentes ligam privação/insônia a maior risco de eventos cardiovasculares, sugerindo que medo crônico e hipervigilância em ambientes violentos podem agravar o risco por vias comportamentais e biológicas (Pan *et al*., 2023).

Sudeste e Sul – cuidado familiar como solução e sobrevivência: em estados como MG, PR e BA, ganham peso variáveis sobre quem cuida e quanto a família se envolve nos cuidados diários. A literatura brasileira mostra que o cuidado informal é frequentemente sustentado por familiares (majoritariamente mulheres), com carga física e emocional relevante e impacto na renda e na saúde do próprio cuidador. Estudos com cuidadores de pessoas com AVC no Brasil (Sul/Sudeste) descrevem sobrecarga moderada a elevada, pior qualidade de vida e sintomas de ansiedade/depressão; revisões indicam que maior sobrecarga do cuidador se associa também a piores desfechos do paciente (Bierhals *et al*., 2023; Caro; Costa; Da Cruz, 2018; Murayama *et al*., 2024).

Essa sobrecarga é a outra face da “rede fragilizada”: quando políticas formais de apoio (ex.: centros-dia, suporte domiciliar, benefícios ao cuidador) são insuficientes, o cuidado recai na família, com risco de adoecimento do cuidador e queda da qualidade do cuidado — um ciclo já descrito em coortes brasileiras e revisões internacionais (Bierhals *et al*., 2023).

Variáveis de violência e de cuidado domiciliar sofrem sub-registro por constrangimento/medo e por diferenças entre fontes de dados (inquéritos populacionais vs. registros), o que implica viés de resposta; o fato de ainda assim emergirem como preditores fortes sugere que o fenômeno real é maior do que o mensurado (Stein *et al*., 2023b).

8.21 Padrões Tecnológicos e de Lazer: Digitalização da Vida Cotidiana e Seus Efeitos Sobre a Saúde

Entre as variáveis que aparecem entre as 30 mais pontuadas por estado, um bloco emergente trata do uso de tecnologias digitais, redes sociais e dispositivos móveis (p.ex., tempo diário em redes sociais/celular; uso da internet para lazer, comunicação ou busca de informações de saúde). Em SP, DF, PI e MA, esses indicadores capturam um determinante comportamental contemporâneo: a vida mediada por telas, com potenciais riscos e oportunidades para a saúde.

SP e DF — urbanização, sedentarismo e digitalização do cotidiano: nos grandes centros, o acesso à internet é mais alto e majoritariamente móvel, o que amplia a exposição a telas no trabalho e no lazer (≥ 86% nas regiões

Sul/Sudeste/Centro-Oeste em 2019; urbano 88% vs. rural 53%; celular como principal dispositivo) (Nakayama *et al*., 2023).

Esse padrão associa-se a mais tempo sedentário (teletrabalho e lazer em tela) e a pior sono quando há uso noturno de mídias, ambos ligados a desfechos cardiometabólicos: metanálises mostram que o trabalho remoto durante a pandemia aumentou comportamento sedentário e reduziu atividade física; tempo sedentário elevado eleva o risco de AVC; e o uso de telas antes de dormir relaciona-se a pior qualidade do sono (Chaudhary *et al*., 2024; Hooker *et al*., 2022; Wang *et al*., 2022b; Wilke *et al*., 2022).

Além disso, o uso intensivo de redes sociais está associado a sintomas de ansiedade, depressão e pior sono (efeitos geralmente modestos, porém consistentes em sínteses recentes) (Ahmed *et al*., 2024; Karim *et al*., 2020).

PI e MA — digitalização em contextos desiguais: em estados com maior vulnerabilidade socioeconômica, as variáveis tecnológicas podem refletir a chegada da conectividade via celular onde serviços presenciais ainda falham. Estudos sobre o fosso digital no Brasil documentam menor acesso no Norte/Nordeste, forte desigualdade urbano-rural e dependência do celular como porta de entrada — fatores que condicionam o quanto a internet pode servir de via de informação em saúde, teleatendimento e suporte social (Nakayama *et al*., 2023).

Nesses cenários, a internet pode ser ferramenta de enfrentamento (busca de informação confiável; telessaúde) ou marcador de vulnerabilidade (uso predominantemente recreativo e passivo em ambientes com poucas alternativas de lazer ativo) (Guimarães *et al*., 2021; Lamas *et al*., 2025).

Entre os mecanismos comportamentais plausíveis que mediam a relação entre o uso de telas e desfechos em saúde destacam-se:

- Sedentarismo prolongado (trabalho e lazer em tela) relaciona-se a maior risco de AVC e pior perfil cardiometabólico (Joundi *et al*., 2021; Wang *et al*., 2022b).
- Uso noturno de telas está associado a pior qualidade do sono e redução de duração do sono, o que impacta PA, metabolismo e humor (Han; Zhou; Liu, 2024; Zhong *et al*., 2025).
- Alimentação distraída (com telas) favorece ingestão não atenta (“mindless eating”) e pior autorregulação alimentar em adultos, com

efeitos demonstrados em ensaios e meta-análises (Robinson *et al*., 2013).

- Redes sociais: metanálises apontam associações pequenas, porém significativas, com pior bem-estar/sono; para solidão, os achados são heterogêneos e dependem do modo de uso (p.ex., uso passivo vs. interativo), sugerindo que o efeito pode ser bidirecional (Ahmed *et al*., 2024; Fam; Männikkö, 2025; Hall, 2025; Karim *et al*., 2020).

Potenciais efeitos protetores: Em paralelo, a digitalização também amplia o acesso à informação de saúde e suporte terapêutico para grupos com barreiras de mobilidade (idosos, cuidadores, pessoas pós-AVC).

No Brasil, instrumentos de eHealth literacy foram validados e estudos mostram que maior letramento digital melhora a análise crítica de informações online; e telereabilitação pós-AVC apresenta eficácia semelhante à reabilitação presencial para domínios como função motora, equilíbrio e atividades de vida diária, especialmente útil onde a oferta presencial é limitada (De Oliveira Collet *et al*., 2024; Hao *et al*., 2023; Mialhe *et al*., 2022; Tchero *et al*., 2018).

8.22 Peculiaridades por Estado – Exemplos Notáveis e o Retrato da Singularidade Local

Apesar da presença de padrões amplos e repetitivos em diversos estados, a análise das 30 variáveis mais pontuadas por unidade federativa revela também características únicas que escapam às tendências nacionais.

Essas variáveis específicas não se destacam por frequência, mas por valor explicativo contextualizado — revelando traços da cultura, estrutura social ou fragilidades regionais que merecem atenção especial.

A seguir, são mostrados exemplos notáveis, cada um representando uma janela para o entendimento mais refinado da realidade de cada estado:

AC: Empregados Domésticos em Domicílios

A saliência da presença de empregados domésticos como variável em um estado de baixa renda média pode ser interpretada como marcador de estratificação socioeconômica intraurbana, dado que o trabalho doméstico remunerado no Brasil é amplamente descrito como prática atravessada por gênero, raça e classe, historicamente concentrada entre mulheres negras em posições de maior

precariedade — logo, útil como proxy de concentração de renda e hierarquias locais (Acciari, 2021; Bernardino-Costa, 2014).

Essa estratificação é relevante para o risco e a mortalidade por AVC: em análise longitudinal por estados brasileiros (2002–2009), a desigualdade de renda mostrou-se associada independentemente a maiores taxas de mortalidade por AVC; estimou-se que redução de 10 pontos no índice de Gini correspondeu a queda de 18% na mortalidade por AVC, mesmo após controle por PIB per capita e demais covariáveis (Vincens; Stafström, 2015a).

Evidências de maior escopo reforçam o gradiente social: revisões sistemáticas e meta-análises indicam que baixo status socioeconômico associa-se a maior mortalidade por AVC e a piores desfechos funcionais pós-evento, sugerindo vias que combinam exposição desigual a fatores de risco, acesso tardio/insuficiente a prevenção secundária e reabilitação, e constrangimentos materiais para continuidade do cuidado (Avan *et al*., 2019; Nguyen *et al*., 2024a; Wang *et al*., 2020a).

AL: Frequência a Cursos – Valorização da Formação Educacional

A participação em cursos educacionais ou profissionalizantes pode ser entendida como um indicador de mobilidade social e de estratégias de qualificação ao longo da vida. Isso se alinha à literatura que identifica a educação como um determinante social crucial da saúde: níveis mais elevados de escolaridade estão associados a menor risco de doenças cardiovasculares, incluindo AVC, em parte devido ao melhor acesso à informação, condições de trabalho mais seguras e maior capacidade de utilizar serviços de saúde. Estudos prospectivos mostram que indivíduos com menor escolaridade têm maior risco de incidência de AVC, especialmente do tipo isquêmico (Xiuyun *et al*., 2020).

Além disso, evidências indicam que níveis mais baixos de educação se correlacionam com maior risco vitalício de doenças cardiovasculares (Magnani *et al*., 2024). No contexto brasileiro, diferenças educacionais explicam parte substancial das desigualdades em autoavaliação da saúde e comportamentos de risco que contribuem para doenças crônicas (Marió; Woolcock, 2008). Assim, a frequência a cursos educacionais não impacta diretamente o risco de AVC, mas funciona como marcador de um contexto educacional e social que modula consistentemente a distribuição de risco e mortalidade por doenças cerebrovasculares.

AM: Substituição do Almoço por Lanches – Informalidade Urbana

A variável “substituir o almoço por lanches” em contextos urbanos do AM é congruente com três conjuntos de evidências (Da Mata; Neves; De Medeiros, 2022; Da Silva Oliveira *et al*., 2024; Davies; Frausin; Parry, 2017):

- maior disponibilidade e consumo de ultraprocessados em capitais da Amazônia (p.ex., Boa Vista e Macapá, com aumentos significativos de snacks, biscoitos e produtos prontos entre 2019–2021);
- vulnerabilidade alimentar em áreas urbanas amazônicas;
- organização do trabalho e do tempo (jornadas extensas, pausas curtas, ausência de refeitórios) que favorece padrões de alimentação de conveniência.

Do ponto de vista comportamental e de ambiente alimentar, estudos qualitativos na amazônia brasileira mostram que custo, praticidade e vida útil longa dos produtos industrializados orientam escolhas, sobretudo quando o acesso a alimentos frescos é incerto — um quadro compatível com a “troca” de refeições completas por lanches rápidos durante a jornada de trabalho informal/precária (Sato *et al*., 2020).

Em paralelo, análises nacionais recentes documentam padrões de refeições e lanches e sugerem que a redução do número de refeições diárias/omissão de refeições associa-se a pior perfil cardiometabólico, reforçando a plausibilidade de risco quando o almoço é sistematicamente substituído por lanches (Rodrigues *et al*., 2024; Selingardi *et al*., 2024).

No plano fisiopatológico, esse arranjo alimentar — com maior participação de ultraprocessados, irregularidade de horários e menor densidade nutricional — relaciona-se a hipertensão, dislipidemia, obesidade e diabetes, todos determinantes proximais do AVC. Revisões sistemáticas e meta-análises indicam que maior consumo de ultraprocessados associa-se a pior desfecho cardiometabólico e maior risco cardiovascular, incluindo AVC, enquanto meta-análise específica aponta maior risco de hipertensão em adultos com alto consumo de ultraprocessados (Lv *et al*., 2024; Mendoza *et al*., 2024; Vitale *et al*., 2023).

BA/MG: Tarefas de Cuidado com Familiares – Rede de Cuidado Informal

Nos estados de BA e MG, a proeminência de variáveis que mensuram quem cuida e quanto a família se envolve no cuidado cotidiano de idosos/doentes é consistente com a centralidade do cuidado informal no pós-AVC no Brasil e com a

heterogeneidade de oferta/coordenação dos serviços territoriais. Em Belo Horizonte, por exemplo, estudo qualitativo em serviços comunitários voltados a idosos pobres descreve a estratégia de “cuidado baseado na comunidade” como resposta a pressões sobre a rede formal, evidenciando que a casa e a família são nós organizadores do cuidado quando a atenção domiciliar pública é irregular ou insuficiente (Lloyd-Sherlock *et al*., 2023).

Do ponto de vista dos desfechos, a literatura brasileira mostra que sobreviventes de AVC dependem majoritariamente de cuidadores familiares, quase sempre mulheres, e que a sobrecarga do cuidador se associa a maior dependência funcional do paciente e a piores indicadores psicossociais do próprio cuidador (ansiedade, depressão, pior qualidade de vida). Ensaios e estudos observacionais no país documentam que intervenções educativas domiciliares de enfermagem melhoram dimensões de qualidade de vida (relações sociais e autonomia) dos cuidadores ao longo de 12 meses; já em análises transversais, a sobrecarga aumenta com as horas de cuidado e com a baixa independência do idoso após o AVC (Bierhals *et al*., 2023; Pereira *et al*., 2013).

A insuficiência de acesso à reabilitação após a alta — componente crucial para reduzir incapacidade e dependência — reforça a centralidade dessas redes informais. Protocolo multicêntrico brasileiro destaca a lacuna histórica de informações e, ao sintetizar evidências nacionais, cita coorte hospitalar em São Paulo na qual 70% de 665 pacientes não receberam reabilitação após a alta, além da percepção, por neurologistas, de acesso inadequado na rede pública; tais barreiras tendem a intensificar a carga transferida às famílias. Estudos recentes também mostram acesso desigual a profissionais de reabilitação no primeiro mês pós-alta, variando por região e condições do usuário (De Oliveira Cacho *et al*., 2022b; Magalhães *et al*., 2023).

Por conseguinte, em BA e MG o destaque para “tarefas de cuidado” pode ser lido, em chave epidemiológica, como um marcador de organização social do cuidado: onde o envelhecimento populacional avança e a integração efetiva entre atenção primária, atenção domiciliar e reabilitação é incompleta, a família absorve maior parcela do trabalho terapêutico, com efeitos em cascata sobre a recuperação funcional pós-AVC (paciente) e sobrecarga/saúde mental (cuidador). Evidências nacionais entre idosos confirmam que experiências piores na atenção primária e barreiras de acesso modulam a utilização de serviços e podem comprometer a gestão

de condições crônicas, contexto no qual o papel do cuidador informal se torna ainda mais determinante (Macinko *et al*., 2018).

PR: Incentivo no Pré-Natal – Políticas de Atenção Primária Atuantes

As variáveis que mensuram oferta/recebimento de incentivos no pré-natal é compatível com APS capilarizada e com o uso de transferências condicionadas como alavancas programáticas do cuidado pré-natal. Evidências nacionais mostram que a qualidade e a cobertura do pré-natal na rede básica melhoraram, ainda que com desigualdades sociais remanescentes, o que torna plausível interpretar esse indicador como proxy de APS operante em territórios com maior organização preventiva (Tomasi *et al*., 2017).

No plano materno-infantil, estudos de coorte e análises populacionais indicam que a participação no Programa Bolsa Família associa-se a melhores desfechos perinatais (menor baixo peso ao nascer; menos prematuridade), e a mudanças favoráveis em estado nutricional/consumo alimentar durante a gestação — efeitos que dependem da condicionalidade de acompanhamento pré-natal e da integração com a estratégia de saúde da família (Falcão *et al*., 2023; Lisboa *et al*., 2022; Ortelan *et al*., 2024; Santana *et al*., 2022).

Em populações urbanas pobres, a utilização da ESF associa-se a menor mortalidade por todas as causas, destacando o papel estruturante da APS na redução de iniquidades (Hone *et al*., 2020b).

O vínculo com o AVC decorre do controle oportuno dos transtornos hipertensivos da gestação e do diabetes gestacional, cujo histórico eleva de forma consistente o risco cardiovascular e cerebrovascular ao longo da vida. Evidências de meta-análises e estudos de coorte indicam que a ocorrência de pré-eclâmpsia ou de outros transtornos hipertensivos da gestação pode duplicar o risco futuro de AVC e de doenças cardiovasculares. Nesse sentido, declaração científica da AHA recomenda vigilância cardiovascular ao longo de todo o curso de vida em mulheres com tais desfechos obstétricos adversos. Assim, sistemas que ampliam e qualificam o pré-natal tendem, indiretamente, a reduzir o risco cerebrovascular de longo prazo por meio da prevenção e da gestão precoce desses agravos (De Havenon *et al*., 2021b; Parikh *et al*., 2021; Wu *et al*., 2017b).

PE: Insuficiência Renal Crônica – Alerta para Comorbidades

O destaque para a variável diagnóstico de Doença Renal Crônica (DRC) sugere carga expressiva e possivelmente subdetectada do agravo em PE, em linha

com evidências locais: em Cabrobó (povo Truká), estudo de base populacional estimou alta prevalência de DRC, maior entre mulheres, apontando para a urgência de estratégias de detecção precoce em grupos vulneráveis do estado (Gomes *et al*., 2024).

Em perspectiva nacional, a mortalidade por DRC apresenta tendência ascendente com desigualdades regionais, com incrementos particularmente marcados nas regiões Norte e Nordeste — cenário compatível com a relevância da variável em PE (Gouvêa *et al*., 2023).

Do ponto de vista prognóstico, a DRC é reconhecida como um importante determinante de risco cerebrovascular. Evidências consistentes demonstram que a redução da função renal está associada a maior ocorrência de AVC. Adicionalmente, a presença de alterações urinárias, como a albuminúria, contribui para identificar indivíduos sob maior vulnerabilidade cardiovascular. Esses achados reforçam a relevância da triagem combinada da função renal e da excreção urinária de albumina na atenção primária, ampliando a capacidade preditiva para além dos fatores de risco tradicionais (Lee *et al*., 2010; Matsushita *et al*., 2015).

Possíveis vetores locais para a carga de DRC incluem hipertensão e diabetes insuficientemente controlados e padrões dietéticos ricos em sódio: no semiárido de PE, inquérito populacional estimou prevalência de hipertensão de 27,4% em adultos, com forte gradiente por idade, escolaridade e estado nutricional; em coortes brasileiras, razões Na/K urinárias muito acima do recomendado pela OMS (Organização Mundial da Saúde) reforçam a necessidade de intervenções populacionais em sal e ultraprocessados (Pereira *et al*., 2019; Santiago *et al*., 2019).

Ademais, o uso disseminado de anti-inflamatórios não esteroidais — muitas vezes sem supervisão — constitui um fator iatrogênico relevante de lesão renal, com plausível agravamento em cenários de desidratação e multimorbidade (Lucas *et al*., 2019).

À luz desse conjunto, a DRC como variável importante em PE funciona como marcador de comorbidades cardio-metabólicas e de risco de AVC ao longo do curso de vida, justificando estratégias integradas: rastreio ativo em grupos de risco, controle agressivo de P.A. e glicemia, promoção de redução de sódio e revisão do uso de nefrotóxicos na atenção primária e especializada (Lee *et al*., 2010; Matsushita *et al*., 2015).

RR: Consumo Frequente de Doces – Hábitos Alimentares de Risco

A proeminência da variável “ingestão frequente de doces” é compatível com o padrão alimentar observado na capital Boa Vista, onde análise do Vigitel (2019–2021) mostrou aumento estatisticamente significativo no consumo de ≥5 subgrupos de ultraprocessados, com destaque para biscoitos/doces e snacks; o estudo atribui parte desse comportamento a barreiras logísticas para alimentos frescos na Amazônia, levando a maior adesão a produtos de longa vida de prateleira e baixo custo (Da Silva Oliveira *et al*., 2024).

Sob a ótica de risco, a alta ingestão de açúcar e doces associa-se a pior perfil cardiometabólico e a maior risco de AVC em sínteses prospectivas (dose–resposta), enquanto o maior consumo de ultraprocessados se relaciona adversamente a doença cardiovascular (com evidência para AVC variando de baixa a moderada, conforme metanálises recentes). Esses achados sustentam a plausibilidade de que um padrão com doces frequentes contribua para o risco cerebrovascular por vias de hipertensão, dislipidemia, obesidade e diabetes (Qu *et al*., 2024; Wang *et al*., 2022a).

Além disso, estudos domiciliares na Amazônia Ocidental documentam insegurança alimentar em áreas urbanas, reforçando que, em contextos de menor acesso físico e econômico a in natura, doces e ultraprocessados tornam-se alternativas mais previsíveis e disponíveis — ainda que metabolicamente desfavoráveis (Da Mata; Sanudo; De Medeiros, 2024; Da Silva Oliveira *et al*., 2024).

Do ponto de vista regulatório, revisões sobre a região apontam a necessidade de medidas estruturais (rotulagem de advertência, tributação de bebidas açucaradas) para mitigar o consumo de produtos de baixo valor nutricional — agenda particularmente pertinente em estados com desafios logísticos como RR (Da Silva Oliveira *et al*., 2024).

RJ: Prevenção do Câncer de Colo do Útero – Política de Rastreio Eficiente

O destaque de variáveis ligadas ao Papanicolau pode ser interpretado como indicador de rede de atenção primária funcional, com campanhas regulares e boa capilaridade nos serviços básicos. Em termos nacionais — parâmetro pertinente para grandes capitais como o Rio de Janeiro — a cobertura do exame aumentou entre 2013 e 2019, com diferenças segundo características sociodemográficas e tempo de entrega do resultado entre SUS e setor privado, sugerindo capacidade instalada e coordenação do cuidado nas áreas mais estruturadas do país (Azevedo; Silva *et al*., 2023).

Em série de capitais brasileiras (2016–2021), observou-se avanço (com queda em 2020 e recuperação subsequente), o que reforça a leitura de política de rastreio organizada em centros urbanos, compatível com a realidade fluminense (Costa *et al*., 2024).

A relevância para AVC decorre do papel da APS na prevenção e no manejo de fatores de risco (hipertensão, diabetes, dislipidemia). Evidências mostram que a expansão/uso da ESF se associa a reduções de mortalidade por doenças do aparelho circulatório, incluindo cerebrovasculares, e a menor risco de óbito em populações urbanas vulneráveis — efeito plausivelmente mediado por maior acesso a rastreamento, cuidado longitudinal e coordenação terapêutica (Rasella *et al*., 2014).

Assim, no RJ, a proeminência de marcadores de rastreamento ginecológico não é apenas um sinal de política oncológica eficiente, mas um traçador de desempenho sistêmico da APS que se conecta, por vias bem estabelecidas, à redução do risco e da mortalidade por AVC na população (Hone *et al*., 2020a; Rasella *et al*., 2014).

SP: Binge Drinking – Padrão Grave de Consumo de Álcool

Em SP, a proeminência de variáveis associadas ao consumo episódico excessivo de álcool é coerente com evidências locais de alta exposição a contextos de vida noturna e “pre-drinking” (beber antes de ingressar em bares/boates), ambos associados a intoxicação na saída dos estabelecimentos e a comportamentos de risco. Estudos com amostragem por portal em casas noturnas da cidade documentaram a prática de pre-drinking e sua relação com maior nível de álcool expirado e desfechos adversos; análises complementares identificaram fatores ambientais dos clubes correlacionados ao binge drinking (padrão de consumo episódico excessivo de álcool em um curto intervalo de tempo) entre frequentadores (Santos *et al*., 2015).

Em paralelo, coortes universitárias brasileiras mostram que o pre-drinking é comum e se associa a danos relacionados ao álcool, indicando subgrupos urbanos particularmente vulneráveis (jovens e estudantes) (Santos *et al*., 2022c).

Do ponto de vista cerebrovascular, o padrão de ingestão importa tanto quanto a dose média. Meta-análise prospectiva indica que o beber pesado associa-se a maior risco de todos os tipos de AVC, com associação mais forte para hemorrágico, ao passo que supostos efeitos protetores da ingestão leve/moderada são inconsistentes e sensíveis a vieses; evidências genéticas e de grandes coortes sugerem relação

aproximadamente linear entre consumo de álcool, P.A. e risco de AVC, sem benefício cardiovascular líquido (Larsson *et al*., 2016; Millwood *et al*., 2019).

Em nível agudo, estudos caso-cruzado mostram que a ingestão alcoólica pode desencadear AVC isquêmico nas horas subsequentes e elevar o risco imediato de eventos cardiovasculares, enquanto padrões de heavy episodic drinking e altas doses estão ligados a maior risco de AVC isquêmico e hemorrágico em análises recentes (Mostofsky *et al*., 2010, 2016; Smyth *et al*., 2023).

Na dimensão da saúde mental se reforça a necessidade de políticas integradas: revisão sistemática e meta-análise em população adulta mostram que transtornos mentais comuns (ansiedade/depressão/fobias) duplicam as chances de transtorno por uso de álcool; em subgrupos urbanos sob estresse ocupacional e social, normas de consumo e cultura do trabalho relacionam-se ao uso problemático. Esses achados apoiam a leitura do binge drinking como marcador de sofrimento psíquico não detectado e de vulnerabilidade comportamental em grandes centros (Puddephatt *et al*., 2022; Thørrisen *et al*., 2022).

9.Conclusão

A presente tese teve como propósito central examinar a viabilidade de utilização de técnicas de aprendizado de máquina aplicadas aos microdados da PNS 2019 para a predição e hierarquização da presença autorreferida de AVC. Ao integrar diferentes blocos de variáveis — sociodemográficas, clínicas, funcionais, comportamentais e de uso de serviços —, buscou-se não apenas avaliar a acurácia de modelos supervisionados, mas sobretudo explorar o potencial dessas ferramentas para revelar padrões latentes, interações complexas e desigualdades estruturais que moldam o perfil epidemiológico do AVC no Brasil.

Os resultados demonstraram que modelos como a Random Forest foram capazes de reproduzir com consistência associações já consolidadas na literatura — a exemplo da hipertensão arterial, do diabetes mellitus e da idade avançada como determinantes de risco —, ao mesmo tempo em que atribuíram relevância a marcadores menos tradicionais, como indicadores de funcionalidade, condições domiciliares e barreiras de acesso a serviços de saúde. Essa constatação reforça a pertinência de abordagens analíticas que transcendam o escopo estritamente

biomédico, reconhecendo o AVC como fenômeno multifatorial e fortemente condicionado por determinantes sociais da saúde.

Cumpre salientar que a heterogeneidade identificada entre regiões e grupos populacionais traduz não apenas diferenças biológicas, mas, sobretudo, desigualdades sociais, econômicas e territoriais, que se expressam em distintos perfis de risco, de diagnóstico e de reabilitação.

Tal achado alinha-se a uma literatura crescente que concebe o AVC não apenas como desfecho clínico, mas também como marcador de iniquidades sociais persistentes. Destarte, políticas públicas voltadas à prevenção e ao cuidado do AVC devem necessariamente articular estratégias biomédicas a intervenções intersetoriais que enfrentem os condicionantes estruturais que perpetuam tais disparidades.

No plano metodológico, o emprego de aprendizado de máquina revelou-se profícuo por possibilitar a análise de um conjunto massivo e heterogêneo de variáveis, incorporando relações não lineares e interações complexas, frequentemente subdimensionadas em análises convencionais. Tais técnicas não devem ser concebidas como substitutivas das abordagens epidemiológicas clássicas, mas como ferramentas complementares, capazes de ampliar o horizonte analítico e fortalecer a capacidade preditiva e explicativa de estudos populacionais.

Assim, a presente investigação aporta contribuições em três dimensões fundamentais: no campo metodológico, ao evidenciar a aplicabilidade e robustez de modelos de aprendizado de máquina em bases nacionais de saúde, oferecendo alternativas para o aprimoramento da vigilância epidemiológica; no plano empírico, ao reafirmar o peso das desigualdades sociais e territoriais na configuração do risco e das consequências do AVC, conferindo-lhes centralidade no debate sobre equidade em saúde; e no plano prático, ao fornecer subsídios para a formulação de políticas públicas mais sensíveis à complexidade do fenômeno, em particular no que se refere à qualificação do pré-natal, à ampliação da cobertura e da resolutividade da atenção primária e à redução das barreiras de acesso aos serviços de prevenção e reabilitação.

Em síntese, a análise aqui desenvolvida sugere que o emprego de ciência de dados aplicada à epidemiologia do AVC possui o potencial de transformar grandes volumes de dados em inteligência estratégica para a gestão em saúde, favorecendo a formulação de respostas mais precisas, oportunas e equitativas. Ao integrar dimensões clínicas, sociais e estruturais em uma mesma matriz analítica, esta tese

reafirma a importância de uma abordagem holística e multidimensional do AVC, posicionando-se como contribuição ao esforço coletivo de construção de sistemas de saúde mais justos, eficazes e sustentáveis.

10.Referências Bibliográficas

ABEDI, Vida *et al.* Editorial: Machine Learning in Action: Stroke Diagnosis and Outcome Prediction. **Frontiers in neurology**, v. 13, 20 jul. 2022.

ACCIARI, Louisa. Practicing Intersectionality: Brazilian Domestic Workers'Strategies of Building Alliances and Mobilizing Identity. **Latin American Research Review**, v. 56, n. 1, p. 67–81, 9 mar. 2021.

AHMED, Oli *et al.* Social media use, mental health and sleep: A systematic review with meta-analyses. **Journal of Affective Disorders**, v. 367, p. 701–712, 15 dez. 2024.

AL KUWAITI, Ahmed *et al.* A Review of the Role of Artificial Intelligence in Healthcare. **Journal of Personalized Medicine**, v. 13, n. 6, p. 951, 1 jun. 2023.

ALANAZI, Eman M.; ABDOU, Aalaa; LUO, Jake. Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models. **JMIR formative research**, v. 5, n. 12, 1 dez. 2021.

AL-HGAISH, Areen *et al.* Enhancing Performance of Machine Learning Models in Healthcare: An Analytical Framework for Assessing and Improving Data Quality. **Lecture Notes in Networks and Systems**, v. 1075 LNNS, p. 137–153, 2025.

ALVES, Gleica Soyan Barbosa *et al.* Barriers and facilitators to accessing healthcare services among elderly people living in a rural Amazonian community, Brazil. **BMC Health Services Research**, v. 25, n. 1, p. 1–14, 1 dez. 2025.

AN, Qi *et al.* A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. **Sensors**, v. 23, n. 9, p. 4178, 22 abr. 2023.

ARMSTRONG, J.; CLIFTON, D. Continual learning of longitudinal health records**, Symposium Proceedings**, 22 dez. 2021.

ASADI, Farkhondeh *et al.* The most efficient machine learning algorithms in stroke prediction: A systematic review. **Health science reports**, v. 7, n. 10, 1 out. 2024.

AVAN, Abolfazl *et al.* Socioeconomic status and stroke incidence, prevalence, mortality, and worldwide burden: an ecological analysis from the Global Burden of Disease Study 2017. **BMC medicine**, v. 17, n. 1, 24 out. 2019.

AWASTHI, Raghav *et al.* Artificial Intelligence in Healthcare: 2023 Year in Review. **medRxiv**, p. 2024.02.28.24303482, 2024.

AZEVEDO E SILVA, Gulnar *et al.* Papanicolaou test in Brazil: analysis of the National Health Survey of 2013 and 2019. **Revista de Saúde Pública**, v. 57, n. 1, p. 55, 2023.

BANEGAS-LUNA, Antonio-Jesús *et al.* When will the mist clear? On the Interpretability of Machine Learning for Medical Applications: a survey. **Int. J. Mol. Sci,** 22(9), 4394, 2020.

BARBALHO, Ingridy M. P. *et al.* Electronic health records in Brazil: Prospects and technological challenges. **Frontiers in Public Health**, v. 10, p. 963841, 3 nov. 2022.

BARRETO, Mauricio Lima. Health inequalities: a global perspective. **Ciencia & saude coletiva**, v. 22, n. 7, p. 2097–2108, 2017.

BENSENOR, Isabela M. *et al.* Prevalence of stroke and associated disability in Brazil: National Health Survey--2013. **Arquivos de neuro-psiquiatria**, v. 73, n. 9, p. 746–750, 1 set. 2015a.

BENSENOR, Isabela M. *et al.* Prevalence of stroke and associated disability in Brazil: National Health Survey - 2013. **Arquivos de Neuro-Psiquiatria**, v. 73, n. 9, p. 746–750, 1 set. 2015b.

BERKMAN, Nancy D. *et al.* Low health literacy and health outcomes: An updated systematic review. **Annals of Internal Medicine**, v. 155, n. 2, p. 97–107, 2011.

BERNARDINO-COSTA, Joaze. Intersectionality and female domestic workers' unions in Brazil. **Women's Studies International Forum**, v. 46, 72–80, 2014.

BIERHALS, Carla Cristiane Becker Kottwitz *et al.* Quality of life in caregivers of aged stroke survivors in southern Brazil: Arandomized clinical trial. **Revista Latino-Americana de Enfermagem**, v. 31, p. e3657, 2023.

BODEN-ALBALA, Bernadette; QUARLES, Leigh W. Education strategies for stroke prevention. **Stroke**, v. 44, n. 6 Suppl 1, jan. 2013.

BOEHME, Amelia K.; ESENWA, Charles; ELKIND, Mitchell S. V. Stroke Risk Factors, Genetics, and Prevention. **Circulation research**, v. 120, n. 3, p. 472–495, 3 fev. 2017.

BOMFIM, Rafael Aiello; CASCAES, Andreia Morales; DE OLIVEIRA, Cesar. Multimorbidity and tooth loss: the Brazilian National Health Survey, 2019. **BMC Public Health**, v. 21, n. 1, p. 1–8, 1 dez. 2021.

CABRAL, N. L. *et al.* Incidence of stroke subtypes, prognosis and prevalence of risk factors in Joinville, Brazil: a 2 year community based study. **Journal of neurology, neurosurgery, and psychiatry**, v. 80, n. 7, p. 755–761, jul. 2009.

CAMELO, Lidyane V. *et al.* Associations of life course socioeconomic position and job stress with carotid intima-media thickness. The Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). **Social Science & Medicine**, v. 141, p. 91–99, 1 set. 2015.

CARDOSO, Maria Regina Alves *et al.* Crowding: Risk factor or protective factor for lower respiratory disease in young children? **BMC Public Health**, v. 4, n. 1, p. 1–8, 3 jun. 2004.

CARO, Camila Caminha; COSTA, Jacqueline Denubila; CEZAR DA CRUZ, Daniel Marinho. The use of mobility assistive devices and the functional independence in stroke patients. **Cadernos Brasileiros de Terapia Ocupacional**, v. 26, n. 3, p. 558–568, 2018.

CARO, Camila Caminha; COSTA, Jacqueline Denubila; DA CRUZ, Daniel Marinho Cezar. Burden and Quality of Life of Family Caregivers of Stroke Patients. **Occupational Therapy in Health Care**, v. 32, n. 2, p. 154–171, 3 abr. 2018.

CARRIZALES-SEPÚLVEDA, Edgar Francisco *et al.* Periodontal Disease, Systemic Inflammation and the Risk of Cardiovascular Disease. **Heart Lung and Circulation**, v. 27, n. 11, p. 1327–1334, 1 nov. 2018.

CHAKI, Jyotismita; WOZNIAK, Marcin. Deep Learning and Artificial Intelligence in Action (2019-2023): A Review on Brain Stroke Detection, Diagnosis, and Intelligent Post-Stroke Rehabilitation Management. **IEEE Access**, v. 12, p. 52161–52181, 2024a.

CHAKI, Jyotismita; WOZNIAK, Marcin. Deep Learning and Artificial Intelligence in Action (2019-2023): A Review on Brain Stroke Detection, Diagnosis, and Intelligent Post-Stroke Rehabilitation Management. **IEEE Access**, v. 12, p. 52161–52181, 2024b.

CHAKRABORTY, Pritam *et al.* Predicting stroke occurrences: a stacked machine learning approach with feature selection and data preprocessing. **BMC Bioinformatics**, v. 25, n. 1, p. 1–23, 1 dez. 2024.

CHAN, Jasper J. L. *et al.* Inequalities in the prevalence of cardiovascular disease risk factors in Brazilian slum populations: A cross-sectional study. **PLOS Global Public Health**, v. 2, n. 9. e0000990, 1 set. 2022.

CHARBUTY, Bahzad; JIJO, Bahzad Taha; ABDULAZEEZ, Adnan Mohsin. Classification Based on Decision Tree Algorithm for Machine Learning. **Journal of Applied Science and Technology Trends**, v. 2, n. 01, p. 20–28, 24 mar. 2021.

CHAUDHARY, Nicole *et al.* Transitioning to Working from Home Due to the COVID-19 Pandemic Significantly Increased Sedentary Behavior and Decreased Physical Activity: A Meta-Analysis. **International Journal of Environmental Research and Public Health**, v. 21, n. 7, p. 851, 1 jul. 2024.

CHEN, Jonathan H.; ASCH, Steven M. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. **The New England journal of medicine**, v. 376, n. 26, p. 2507–2509, 29 jun. 2017.

CHENG, Fei *et al.* Tooth loss and risk of cardiovascular disease and stroke: A dose-response meta analysis of prospective cohort studies. **PLOS ONE**, v. 13, n. 3, p. e0194563, 1 mar. 2018.

CHOWDHURY, Mzi; TURIN, T. C. Variable selection strategies and its importance in clinical prediction modelling. **Fam Med Com Health**, v. 8, p. 262, 2020.

COELHO, Christianne de Faria; BURINI, Roberto Carlos. Atividade física para prevenção e tratamento das doenças crônicas não transmissíveis e da incapacidade funcional. **Revista de Nutrição**, v. 22, n. 6, p. 937–946, 2009.

COELHO, Lara E. *et al.* SARS-CoV-2 transmission in a highly vulnerable population of Brazil: a household cohort study. **The Lancet Regional Health - Americas**, v. 36, p. 100824, 1 ago. 2024.

COLEMAN, Elisheva R. *et al.* Early Rehabilitation After Stroke: a Narrative Review. **Current atherosclerosis reports**, v. 19, n. 12, p. 59, 1 dez. 2017.

COPSTEIN, Leslie; FERNANDES, Jefferson Gomes; BASTOS, Gisele Alsina Nader. Prevalence and risk factors for stroke in a population of Southern Brazil. **Arquivos de Neuro-Psiquiatria**, v. 71, n. 5, p. 294–300, 2013.

COSTA, Annielson de Souza *et al.* Pap Smear Cancer Coverage in Brazilian Capitals including the Pandemic Period Caused by the SARS-CoV-2 Virus: Ecological Study. **International Journal of Environmental Research and Public Health**, v. 21, n. 3, 1 mar. 2024.

COTÉ, Kathryn E. *et al.* Neighborhood income inequality associated with functional independence after ischemic stroke: a cohort study. **Journal of Stroke and Cerebrovascular Diseases**, v. 34, n. 1, 1 jan. 2025.

COUBE, Maíra *et al.* Inequalities in unmet need for health care services and medications in Brazil: a decomposition analysis. **Lancet Regional Health - Americas**, v. 19, p. 100426, 1 mar. 2023a.

COUBE, Maíra *et al.* Persistent inequalities in health care services utilisation in Brazil (1998–2019). **International Journal for Equity in Health**, v. 22, n. 1, p. 1–15, 1 dez. 2023b.

DA COSTA LOUZADA, Maria Laura *et al.* Consumo de alimentos ultraprocessados no Brasil: distribuição e evolução temporal 2008–2018. **Revista de Saúde Pública**, v. 57, p. 12, 14 abr. 2023.

DA LUZ, Tamires Conceição *et al.* Cardiovascular risk factors in a Brazilian rural population. **Ciencia & saude coletiva**, v. 25, n. 10, p. 3921–3932, 1 out. 2020.

DA MATA, Mayline Menezes; NEVES, José Anael; DE MEDEIROS, Maria Angélica Tavares. Hunger and its associated factors in the western Brazilian Amazon: a population-based study. **Journal of Health, Population, and Nutrition**, v. 41, n. 1, p. 36, 1 dez. 2022.

DA MATA, Mayline Menezes; SANUDO, Adriana; DE MEDEIROS, Maria Angélica Tavares. Insegurança alimentar e insegurança hídrica domiciliar: um estudo de base populacional em um município da bacia hidrográfica do Rio Amazonas, Brasil. **Cadernos de Saúde Pública**, v. 40, n. 4, p. e00125423, 2024.

DA NUNES, Heloá Conceição; GUIMARÃES, Rita Miranda Coessens; DADALTO, Luciana. Desafios bioéticos do uso da inteligência artificial em hospitais. **Revista Bioética**, v. 30, n. 1, p. 82–93, 9 maio 2022.

DA SILVA, Hellen Carla Alves *et al.* Cárie dentária e fatores associados aos 12 anos na Região Centro-Oeste do Brasil em 2010: um estudo transversal. **Ciência & Saúde Coletiva**, v. 25, n. 10, p. 3981–3988, 28 set. 2020.

DA SILVA, Jaine Kareny; BOERY, Rita Narriman Silva de Oliveira. Effectiveness of a support intervention for family caregivers and stroke survivors. **Revista Latino-Americana de Enfermagem**, v. 29, p. e3482, 2021.

DA SILVA OLIVEIRA, Elyecleyde Katiane *et al.* Consumption of Ultra-Processed Foods in the Brazilian Amazon during COVID-19. **Nutrients**, v. 16, n. 13, p. 2117, 1 jul. 2024.

DA SILVA PAIVA, Laércio *et al.* Regional differences in the temporal evolution of stroke: a population-based study of Brazil according to sex in individuals aged 15-49 years between 1997 and 2012. **BMC research notes**, v. 11, n. 1, 21 maio 2018.

DA SILVEIRA, Denise Silva *et al.* Gestão do trabalho, da educação, da informação e comunicação na atenção básica à saúde de municípios das regiões Sul e Nordeste do Brasil. **Cadernos de Saúde Pública**, v. 26, n. 9, p. 1714–1726, 2010.

DAIDONE, Mario *et al.* Machine learning applications in stroke medicine: advancements, challenges, and future prospectives. **Neural regeneration research**, v. 19, n. 4, p. 769–773, 1 abr. 2024.

DAS, Surajit *et al.* Machine Learning in Healthcare Analytics: A State-of-the-Art Review. **Archives of Computational Methods in Engineering**, v. 31, n. 7, p. 3923–3962, 4 abr. 2024.

DAVENPORT, Thomas; KALAKOTA, Ravi. The potential for artificial intelligence in healthcare. **Future Healthcare Journal**, v. 6, n. 2, p. 94, jun. 2019.

DAVIES, Gemma; FRAUSIN, Gina; PARRY, Luke. Are There Food Deserts in Rainforest Cities? **Annals of the American Association of Geographers**, v. 107, n. 4, p. 794–811, 4 jul. 2017.

DAY, Carolina Baltar *et al.* A longitudinal study of burden among spouse and non-spouse caregivers of older adults with stroke-induced-dependency. **Revista Brasileira de Enfermagem**, v. 76, n. 6, 2023.

DE ABREU, Fernanda Gabriela *et al.* Stroke at baseline of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil): a cross-sectional analysis. **São Paulo Medical Journal**, v. 136, n. 5, p. 398, 1 set. 2018.

DE CAMPOS, Lívia Mizuki *et al.* How Many Patients Become Functionally Dependent after a Stroke? A 3-Year Population-Based Study in Joinville, Brazil. **PLOS ONE**, v. 12, n. 1, p. e0170204, 1 jan. 2017.

DE HAVENON, Adam *et al.* Association of Preeclampsia With Incident Stroke in Later Life Among Women in the Framingham Heart Study. **JAMA Network Open**, v. 4, n. 4, p. e215077–e215077, 1 abr. 2021a.

DE HAVENON, Adam *et al.* Association of Preeclampsia With Incident Stroke in Later Life Among Women in the Framingham Heart Study. **JAMA Network Open**, v. 4, n. 4, p. e215077, 26 abr. 2021b.

DE JESUS, Patricia Romualdo *et al.* The low health literacy in Latin America and the Caribbean: a systematic review and meta-analysis. **BMC public health**, v. 24, n. 1, 1 dez. 2024.

DE OLIVEIRA CACHO, Roberta *et al.* Access to rehabilitation after stroke in Brazil (AReA study): multicenter study protocol. **Arquivos de Neuro-Psiquiatria**, v. 80, n. 10, p. 1067–1074, 1 out. 2022a.

DE OLIVEIRA CACHO, Roberta *et al.* Access to rehabilitation after stroke in Brazil (AReA study): multicenter study protocol. **Arquivos de Neuro-Psiquiatria**, v. 80, n. 10, p. 1067, 1 out. 2022b.

DE OLIVEIRA COLLET, Giulia *et al.* Influence of digital health literacy on online health-related behaviors influenced by internet advertising. **BMC Public Health**, v. 24, n. 1, p. 1949, 1 dez. 2024.

DE OLIVEIRA, Gláucia Maria Moraes *et al.* Estatística Cardiovascular – Brasil 2023. **Arquivos Brasileiros de Cardiologia**, v. 121, n. 2, p. e20240079, 17 jun. 2024a.

DE OLIVEIRA, Roberta Teixeira *et al.* Food Acquisition Locations and Food Groups Acquired According to Levels of Food Insecurity in Brazil. **International Journal of Environmental Research and Public Health**, v. 21, n. 12, p. 1577, 27 nov. 2024b.

DE PAULA COSTA, Danielle Vasconcellos *et al.* Diferenças no consumo alimentar nas áreas urbanas e rurais do Brasil: Pesquisa Nacional de Saúde. **Ciência & Saúde Coletiva**, v. 26, p. 3805–3813, 2021.

DE SÁ, Ana Carolina Micheletti Gomide Nogueira *et al.* Prevalence and factors associated with self-reported diagnosis of high cholesterol in the Brazilian adult population: National Health Survey 2019. **Epidemiologia e Serviços de Saúde**, v. 31, n. Special Issue 1, p. e2021380, 2022.

DE SANTANA, Nathalia Matos *et al.* The burden of stroke in Brazil in 2016: an analysis of the Global Burden of Disease study findings. **BMC research notes**, v. 11, n. 1, 16 out. 2018.

DE SOUZA JÚNIOR, Paulo Roberto Borges *et al.* Cobertura de plano de saúde no Brasil: análise dos dados da Pesquisa Nacional de Saúde 2013 e 2019. **Ciência & Saúde Coletiva**, v. 26, p. 2529–2541, 2021.

DE SOUZA, Tiago Odilio *et al.* Prevalence of unprotected sexual activity in the Brazilian population and associated factors: National Health Survey, 2019. **Epidemiologia e Servicos de Saude**, v. 31, n. 2, 2022.

DE VASCONCELOS, Nádia Machado *et al.* Who are the Adult Women Exposed to Violence in Brazil? **Revista de Saude Publica**, v. 59, p. 1–15, 2025.

DELPINO, F. M. *et al.* Machine learning for predicting chronic diseases: a systematic review. **Public health**, v. 205, p. 14–25, 1 abr. 2022.

DENG, Linghui *et al.* Interventions for management of post-stroke depression: A Bayesian network meta-analysis of 23 randomized controlled trials. **Scientific Reports**, v. 7, n. 1, p. 1–12, 1 dez. 2017.

DO CARMO FERREIRA, Maria; SARTI, Flavia Mori; DE AZEVEDO BARROS, Marilisa Berti. Social inequalities in the incidence, mortality, and survival of neoplasms in women from a municipality in Southeastern Brazil. **Cadernos de Saúde Pública**, v. 38, n. 2, p. e00107521, 7 mar. 2022.

DONKOR, Eric S. Stroke in the 21st Century: A Snapshot of the Burden, Epidemiology, and Quality of Life. **Stroke research and treatment**, v. 2018, 2018.

DRITSAS, Elias; TRIGKA, Maria. Stroke Risk Prediction with Machine Learning Techniques. **Sensors**, v. 22, n. 13, p. 4670, 21 jun. 2022a.

DRITSAS, Elias; TRIGKA, Maria. Stroke Risk Prediction with Machine Learning Techniques. **Sensors**, v. 22, n. 13, 1 jul. 2022b.

EGÍDIO DE SOUZA, Mara *et al.* Relação entre fatores socioeconômicos, clínicos e saúde bucal em escolares da zona rural: um estudo longitudinal. **RFO UPF**, v. 20, n. 2, p. 208–215, 2015.

EL KHOUDARY, Samar R. *et al.* Menopause Transition and Cardiovascular Disease Risk: Implications for Timing of Early Prevention: A Scientific Statement from the American Heart Association. **Circulation**, v. 142, n. 25, p. E506–E532, 22 dez. 2020.

ELIAS, Mariele Abadia *et al.* Artificial intelligence in health and bioethical implications: a systematic review. **Revista Bioética**, v. 31, p. e3542PT, 5 abr. 2024.

ENGSTAD, Torgeir; BØNAA, Kaare H.; VIITANEN, Matti. Validity of self-reported stroke: The Tromso study. **Stroke**, v. 31, n. 7, p. 1602–1607, 2000.

FACCHINI, Luiz Augusto; TOMASI, Elaine; DILÉLIO, Alitéia Santiago. Qualidade da Atenção Primária à Saúde no Brasil: avanços, desafios e perspectivas. **Saúde em Debate**, v. 42, n. spe1, p. 208–223, set. 2018.

FALCÃO, Ila R. *et al.* Participation in Conditional Cash Transfer Program During Pregnancy and Birth Weight-Related Outcomes. **JAMA Network Open**, v. 6, n. 11, 28 nov. 2023.

FAM, Jia Yuin; MÄNNIKKÖ, Niko. Loneliness and Problematic Media Use: Meta-Analysis of Longitudinal Studies. **Journal of medical Internet research**, v. 27, n. 1, p. e60410, 14 ago. 2025.

FAN, Fenglei *et al.* On Interpretability of Artificial Neural Networks: A Survey. 8 jan. 2020.

FATEMA, Zareen *et al.* 'Quality of life at 90 days after stroke and its correlation to activities of daily living': A prospective cohort study. **Journal of Stroke and Cerebrovascular Diseases**, v. 31, n. 11, p. 106806, 1 nov. 2022.

FEIGIN, Valery L.; NORRVING, Bo; MENSAH, George A. Global Burden of Stroke. **Circulation research**, v. 120, n. 3, p. 439–448, 3 fev. 2017.

FERNANDES, Jefferson Gomes. Stroke prevention and control in Brazil: missed opportunities. **Arquivos de neuro-psiquiatria**, v. 73, n. 9, p. 733–735, 1 set. 2015.

FRANCISCO, Priscila Maria Stolses Bergamo *et al.* Stroke in older people in Brazil: prevalence, associated factors, limitations and care practices. A cross-sectional study. **Sao Paulo Medical Journal**, v. 143, n. 3, 2025a.

FRANCISCO, Priscila Maria Stolses Bergamo *et al.* Stroke in older people in Brazil: prevalence, associated factors, limitations and care practices. A cross-sectional study. **São Paulo Medical Journal**, v. 143, n. 3, p. e2024132, 2025b.

GASPAR, Renato Simões *et al.* Income inequality and non-communicable disease mortality and morbidity in Brazil States: a longitudinal analysis 2002-2017. **The Lancet Regional Health - Americas**, v. 2, 1 out. 2021.

GERSTORFER, Yannick; HAHN-KLIMROTH, Max; KRIEG, Lena. A Notion of Feature Importance by Decorrelation and Detection of Trends by Random Forest Regression. **Data Science Journal**, v. 22, n. 1, 2023.

GHASSEMI, Marzyeh *et al.* A Review of Challenges and Opportunities in Machine Learning for Health. **AMIA Summits on Translational Science Proceedings**, v. 2020, p. 191, 2020.

GHONEEM, Ahmed *et al.* Association of Socioeconomic Status and Infarct Volume With Functional Outcome in Patients With Ischemic Stroke. **JAMA Network Open**, v. 5, n. 4, p. e229178–e229178, 1 abr. 2022.

GKANTZIOS, Aimilios *et al.* From Admission to Discharge: Predicting National Institutes of Health Stroke Scale Progression in Stroke Patients Using Biomarkers and Explainable Machine Learning. **Journal of personalized medicine**, v. 13, n. 9, 1 set. 2023.

GOMES, Ana Beatriz Ayroza Galvão Ribeiro *et al.* Popular stroke knowledge in Brazil: A multicenter survey during “World Stroke Day”. **eNeurologicalSci**, v. 6, p. 63, 1 mar. 2016.

GOMES, Nayara Lopes; DE SOUZA LOPES, Claudia. Panorama of risky sexual behaviors in the Brazilian adult population – PNS 2019. **Revista de Saude Publica**, v. 56, p. 61, 2022.

GOMES, Orlando Vieira *et al.* Prevalence and associated factors of chronic kidney disease among Truká Indigenous adults in Cabrobó, Brazil: a population-based study. **The Lancet Regional Health - Americas**, v. 38, 1 out. 2024.

GORELICK, Marc H. Bias arising from missing data in predictive models. **Journal of Clinical Epidemiology**, v. 59, n. 10, p. 1115–1123, 1 out. 2006.

GOULART, Alessandra Carvalho. “EMMA Study: a Brazilian community-based cohort study of stroke mortality and morbidity”. **Sao Paulo medical journal = Revista paulista de medicina**, v. 134, n. 6, p. 543–554, 1 nov. 2016.

GOUVÊA, Ellen de Cassia Dutra Pozzetti *et al.* Mortality trend due to chronic kidney disease in Brazil: an ecological study. **Epidemiologia e Serviços de Saúde : Revista do Sistema Unico de Saúde do Brasil**, v. 32, n. 3, p. e2023313, 2023.

GUIMARÃES, Vinícius Henrique Almeida *et al.* Knowledge about COVID-19 in Brazil: Cross-sectional web-based study. **JMIR Public Health and Surveillance**, v. 7, n. 1, p. e24756, 1 jan. 2021.

GUREVICH, Emma; EL HASSAN, Basheer; EL MORR, Christo. Equity within AI systems: What can health leaders expect? **Healthcare Management Forum**, v. 36, n. 2, p. 119, 1 mar. 2022.

HALL, Jeffrey A. Loneliness and social media. **Annals of the New York Academy of Sciences**, v. 1543, n. 1, p. 5–16, 1 jan. 2025.

HAN, Xiaoning; ZHOU, Enze; LIU, Dong. Electronic Media Use and Sleep Quality: Updated Systematic Review and Meta-Analysis. **Journal of Medical Internet Research**, v. 26, n. 1, p. e48356, 1 jan. 2024.

HAO, Jie *et al.* Effects of virtual reality-based telerehabilitation for stroke patients: A systematic review and meta-analysis of randomized controlled trials. **Journal of Stroke and Cerebrovascular Diseases**, v. 32, n. 3, 1 mar. 2023.

HEO, Joon Nyung *et al.* Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. **Stroke**, v. 50, n. 5, p. 1263–1265, 1 maio 2019.

HONE, Thomas *et al.* Primary healthcare expansion and mortality in Brazil's urban poor: A cohort analysis of 1.2 million adults. **PLoS Medicine**, v. 17, n. 10, p. e1003357, 30 out. 2020a.

HONE, Thomas *et al.* Primary healthcare expansion and mortality in Brazil's urban poor: A cohort analysis of 1.2 million adults. **PLoS Medicine**, v. 17, n. 10, p. e1003357, 30 out. 2020b.

HOOKER, Steven P. *et al.* Association of Accelerometer-Measured Sedentary Time and Physical Activity With Risk of Stroke Among US Adults. **JAMA Network Open**, v. 5, n. 6, p. e2215385–e2215385, 1 jun. 2022.

HURFORD, Robert *et al.* Diagnosis and management of acute ischaemic stroke. **Practical neurology**, v. 20, n. 4, p. 306–318, 1 ago. 2020.

IRANZAD, Reza; LIU, Xiao. A review of random forest-based feature selection methods for data science education and applications. **International Journal of Data Science and Analytics**, p. 1–15, 3 fev. 2024.

JAMES EZEH, Chinedu *et al.* The role of predictive analytics in enhancing public health surveillance: Proactive and data-driven interventions. **World Journal of Advanced Research and Reviews**, v. 2024, n. 03, p. 3059–3077, 2024.

JOHNSON, Dayna A. *et al.* Influence of neighbourhood-level crowding on sleep-disordered breathing severity: mediation by body size. **Journal of Sleep Research**, v. 24, n. 5, p. 559–565, 1 out. 2015.

JOUNDI, Raed A. *et al.* Association between Excess Leisure Sedentary Time and Risk of Stroke in Young Individuals. **Stroke**, v. 52, n. 11, p. 3562–3568, 1 nov. 2021.

KAPOOR, Sayash; NARAYANAN, Arvind. Leakage and the reproducibility crisis in machine-learning-based science. **Patterns**, v. 4, n. 9, p. 100804, 8 set. 2023.

KARIM, Fazida *et al.* Social Media Use and Its Connection to Mental Health: A Systematic Review. **Cureus**, v. 12, n. 6, p. e8627, 15 jun. 2020.

KATAN, Mira; LUFT, Andreas. Global Burden of Stroke. **Seminars in neurology**, v. 38, n. 2, p. 208–211, 1 abr. 2018.

KIM, Kyung; KIM, Young Mi; KIM, Eun Kyung. Correlation between the Activities of Daily Living of Stroke Patients in a Community Setting and Their Quality of Life. **Journal of Physical Therapy Science**, v. 26, n. 3, p. 417, 2014.

KUPER, Hannah *et al.* The socioeconomic gradient in the incidence of stroke: a prospective study in middle-aged women in Sweden. **Stroke**, v. 38, n. 1, p. 27–33, jan. 2007.

KURIAKOSE, Diji; XIAO, Zhicheng. Pathophysiology and Treatment of Stroke: Present Status and Future Perspectives. **International journal of molecular sciences**, v. 21, n. 20, p. 1–24, 2 out. 2020.

KURTZ, Pedro *et al.* Hospital Length of Stay and 30-Day Mortality Prediction in Stroke: A Machine Learning Analysis of 17,000 ICU Admissions in Brazil. **Neurocritical Care**, v. 37, n. 2, p. 313–321, 1 ago. 2022.

KUSAMA, Taro *et al.* Bidirectional longitudinal associations between subjective oral health and subjective well-being. **Archives of Gerontology and Geriatrics**, v. 131, p. 105740, 1 abr. 2025.

LAMAS, Celina de Almeida *et al.* Telehealth Initiative to Enhance Primary Care Access in Brazil (UBS+Digital Project): Multicenter Prospective Study. **Journal of Medical Internet Research**, v. 27, p. e68434, 29 abr. 2025.

LANGHORNE, Peter *et al.* Very early versus delayed mobilisation after stroke. **Cochrane Database of Systematic Reviews**, v. 2018, n. 10, 16 out. 2018.

LANGNER, Taro. Machine Learning Techniques for MRI Data Processing at Expanding Scale. **arXiv**, 2404.14326, 2024.

LARSSON, Susanna C. *et al.* Differing association of alcohol consumption with different stroke types: A systematic review and meta-analysis. **BMC Medicine**, v. 14, n. 1, p. 1–11, 24 nov. 2016.

LEE, Meng *et al.* Low glomerular filtration rate and risk of stroke: meta-analysis. **BMJ**, v. 341, n. 7776, p. 767, 30 set. 2010.

LESSA, I. Epidemiologia dos acidentes vasculares encefálicos na cidade do salvador: aspectos clínicos. **Arquivos de Neuro-Psiquiatria**, v. 43, n. 2, p. 133–139, 1985.

LI, An le *et al.* Risk probability and influencing factors of stroke in followed-up hypertension patients. **BMC Cardiovascular Disorders**, v. 22, n. 1, p. 1–10, 1 dez. 2022.

LI, Linxin *et al.* Articles Age-specific risks, severity, time course, and outcome of bleeding on long-term antiplatelet treatment after vascular events: a population-based cohort study. **The Lancet**, v. 390, p. 490–499, 2017.

LI, Xiao Sheng *et al.* Machine learning in the prediction of post-stroke cognitive impairment: a systematic review and meta-analysis. **Frontiers in neurology**, v. 14, 2023.

LIMA-COSTA, M. Fernanda *et al.* The Brazilian Longitudinal Study of Aging (ELSI-Brazil): Objectives and Design. **American Journal of Epidemiology**, v. 187, n. 7, p. 1345, 1 jul. 2018.

LISBOA, Cinthia Soares *et al.* Socioeconomic and nutritional aspects of pregnant women assisted by Programa Bolsa Família: cohort NISAMI. **Ciência & Saúde Coletiva**, v. 27, n. 1, p. 315–324, 2022.

LLOYD-SHERLOCK, Peter *et al.* Addressing pressures on health services in Belo Horizonte, Brazil through community-based care for poor older people: a qualitative study. **Lancet Reg Health Am.,** 27:100619, 2023.

LOTUFO, Paulo A.; BENSEÑOR, Isabela M. Stroke mortality in Brazil: one example of delayed epidemiological cardiovascular transition. **International journal of stroke : official journal of the International Stroke Society**, v. 4, n. 1, p. 40–41, 2009.

LOTUFO, Paulo Andrade. Stroke in Brazil: a neglected disease. **Sao Paulo medical journal = Revista paulista de medicina**, v. 123, n. 1, p. 3–4, 2 jan. 2005.

LOTUFO, Paulo Andrade; BENSENOR, Isabela Judith Martins. Race and stroke mortality in Brazil. **Revista de Saúde Pública**, v. 47, n. 6, p. 1201, dez. 2013.

LUCAS, Guillherme Nobre Cavalcanti *et al.* Pathophysiological aspects of nephropathy caused by non-steroidal anti-inflammatory drugs. **Brazilian Journal of Nephrology**, v. 41, n. 1, p. 124–130, 1 mar. 2019.

LUO, Yi *et al.* Balancing accuracy and interpretability of machine learning approaches for radiation treatment outcomes modeling. **BJR open**, v. 1, n. 1, 4 jul. 2019.

LV, Jia Le *et al.* Ultra-processed food consumption and metabolic disease risk: an umbrella review of systematic reviews with meta-analyses of observational studies. **Frontiers in Nutrition**, v. 11, p. 1306310, 2024.

MACINKO, James *et al.* Primary care and healthcare utilization among older Brazilians (ELSI-Brazil). **Revista de Saúde Pública**, v. 52, n. Suppl 2, p. 6s, 2018.

MAGALHÃES, Jordana P. *et al.* Access to rehabilitation professionals by individuals with stroke one month after hospital discharge from a stroke unit in Brazil is insufficient regardless of the pandemic. **Journal of Stroke and Cerebrovascular Diseases**, v. 32, n. 8, p. 107186, 1 ago. 2023.

MAGNANI, Jared W. *et al.* Educational Attainment and Lifetime Risk of Cardiovascular Disease. **JAMA Cardiology**, v. 9, n. 1, p. 45–54, 10 jan. 2024.

MAINALI, Shraddha; DARSIE, Marin E.; SMETANA, Keaton S. Machine Learning in Action: Stroke Diagnosis and Outcome Prediction. **Frontiers in neurology**, v. 12, 6 dez. 2021.

MALAEB, Diana *et al.* Effect of Sociodemographic Factors, Concomitant Disease States, and Measures Performed in the Emergency Department on Patient Disability in Ischemic Stroke: Retrospective Study from Lebanon. **Stroke research and treatment**, v. 2021, 2021.

MALTA, Deborah Carvalho *et al.* Prevalence of high risk for cardiovascular disease among the Brazilian adult population, according to different risk calculators: a comparative study. **Ciencia & saude coletiva**, v. 26, n. 4, p. 1221–1231, 1 abr. 2021a.

MALTA, Deborah Carvalho *et al.* Desigualdades socioeconômicas relacionadas às doenças crônicas não transmissíveis e suas limitações: Pesquisa Nacional de Saúde, 2019. **Revista Brasileira de Epidemiologia**, v. 24, p. e210011, 10 dez. 2021b.

MANIVA, Samia Jardelle Costa de Freitas *et al.* Educational technologies for health education on stroke: an integrative review. **Revista brasileira de enfermagem**, v. 71, n. suppl 4, p. 1724–1731, 2018.

MARIÓ, Estanislao Gacitúa; WOOLCOCK, Michael. Social Exclusion and Mobility in Brazil. **Washington, DC: World Bank,** 42486, 2005.

MARKETOU, Maria E. *et al.* Stroke Risk Prediction with Machine Learning Techniques. **Sensors**, v. 22, n. 13, p. 4670, 21 jun. 2022.

MARMOT, M.; BELL, R. Fair society, healthy lives. **Public health**, v. 126 Suppl 1, n. SUPPL.1, p. S4, 1 set. 2012.

MARTINS, Sheila Cristina Ouriques *et al.* Past, present, and future of stroke in middle-income countries: the Brazilian experience. **International journal of stroke : official journal of the International Stroke Society**, v. 8 Suppl A100, n. 100 A, p. 106–111, 2013.

MARTINS, Sheila Cristina Ouriques *et al.* Disparities in Stroke Patient-Reported Outcomes Measurement Between Healthcare Systems in Brazil. **Frontiers in Neurology**, v. 13, p. 857094, 6 maio 2022.

MATSUSHITA, Kunihiro *et al.* Estimated glomerular filtration rate and albuminuria for prediction of cardiovascular outcomes: A collaborative meta-analysis of individual participant data. **The Lancet Diabetes and Endocrinology**, v. 3, n. 7, p. 514–525, 1 jul. 2015.

MATSUYAMA, Y. *et al.* Causal effect of tooth loss on depression: evidence from a population-wide natural experiment in the USA. **Epidemiology and Psychiatric Sciences**, v. 30, p. e38, 2021.

MENDOZA, Kenny *et al.* Ultra-processed foods and cardiovascular disease: analysis of three large US prospective cohorts and a systematic review and meta-analysis of prospective cohort studies. **The Lancet Regional Health – Americas**, v. 37, 1 set. 2024.

MESCHIA, James F. *et al.* Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. **Stroke**, v. 45, n. 12, p. 3754–3832, 11 dez. 2014.

MIALHE, Fábio Luiz *et al.* Evaluating the psychometric properties of the eHealth Literacy Scale in Brazilian adults. **Revista Brasileira de Enfermagem**, v. 75, n. 1, p. e20201320, 2022.

MILLWOOD, Iona Y. *et al.* Conventional and genetic evidence on alcohol and vascular disease aetiology: a prospective study of 500 000 men and women in China. **The Lancet**, v. 393, n. 10183, p. 1831–1842, 4 maio 2019.

MINELLI, Cesar *et al.* Brazilian Academy of Neurology practice guidelines for stroke rehabilitation: part I. **Arquivos de Neuro-psiquiatria**, v. 80, n. 6, p. 634–652, 1 jun. 2022a.

MINELLI, Cesar *et al.* Brazilian practice guidelines for stroke rehabilitation: Part II. **Arquivos de Neuro-psiquiatria**, v. 80, n. 7, p. 741–758, 1 jul. 2022b.

MIOTTO, Riccardo *et al.* Deep learning for healthcare: review, opportunities and challenges. **Briefings in Bioinformatics**, v. 19, n. 6, p. 1236–1246, 27 nov. 2018.

MORAIS, Huana Carolina Cândido *et al.* Burden and modifications in life from the perspective of caregivers for patients after stroke. **Revista Latino-Americana de Enfermagem**, v. 20, n. 5, p. 944–953, set. 2012.

MOSTOFSKY, Elizabeth *et al.* Alcohol and Acute Ischemic Stroke Onset: The Stroke Onset Study. **Stroke; a journal of cerebral circulation**, v. 41, n. 9, p. 1845, set. 2010.

MOSTOFSKY, Elizabeth *et al.* Alcohol and immediate risk of cardiovascular events. **Circulation**, v. 133, n. 10, p. 979–987, 8 mar. 2016.

MURAYAMA, Luis Henrique Vallesquino *et al.* Caregiver burden, hopelessness, and anxiety: Association between sociodemographic and clinical profiles of patients with stroke. **Journal of Stroke and Cerebrovascular Diseases**, v. 33, n. 11, p. 107905, 1 nov. 2024.

MUREN, Marie Almkvist; HÜTLER, Matthias; HOOPER, Julie. Functional capacity and health-related quality of life in individuals post stroke. **Topics in Stroke Rehabilitation**, v. 15, n. 1, p. 51–58, jan. 2008.

NAIK, Nithesh *et al.* Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility? **Frontiers in Surgery**, v. 9, p. 862322, 14 mar. 2022.

NAKAYAMA, Luis Filipe *et al.* The Digital Divide in Brazil and Barriers to Telehealth and Equal Digital Health Care: Analysis of Internet Access Using Publicly Available Data. **Journal of Medical Internet Research**, v. 25, p. e42483, 2023.

NGUYEN, Mai T. H. *et al.* Influence of Socioeconomic Status on Functional Outcomes After Stroke: A Systematic Review and Meta-Analysis. **Journal of the American Heart Association**, v. 13, n. 9, p. 33078, 7 maio 2024a.

NGUYEN, Thi Nguyet Que *et al.* Multi-task learning for predicting quality-of-life and independence in activities of daily living after stroke: a proof-of-concept study. **Frontiers in Neurology**, v. 15, p. 1449234, 27 set. 2024b.

NIJMAN, S. W. J. *et al.* Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review. **Journal of Clinical Epidemiology**, v. 142, p. 218–229, 1 fev. 2022.

NOGUEIRA, Raul G. *et al.* Global impact of COVID-19 on stroke care. **International journal of stroke : official journal of the International Stroke Society**, v. 16, n. 5, p. 573–584, 1 jul. 2021.

NUGEM, Rita *et al.* Stroke Care in Brazil and France: National Policies and Healthcare Indicators Comparison. **Journal of multidisciplinary healthcare**, v. 13, p. 1403–1414, 2020.

OBERMEYER, Ziad; EMANUEL, Ezekiel J. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. **The New England journal of medicine**, v. 375, n. 13, p. 1216–1219, 29 set. 2016.

O'DONNELL, Martin J. *et al.* Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. **Lancet (London, England)**, v. 388, n. 10046, p. 761–775, 20 ago. 2016.

OGUNPOLA, Adedayo *et al.* Machine Learning-Based Predictive Models for Detection of Cardiovascular Diseases. **Diagnostics**, v. 14, n. 2, p. 144, 1 jan. 2024.

OLABANJO, Olusola *et al.* Stroke Risk Factor Prediction Using Machine Learning Techniques: A Systematic Review. **Journal of Applied Sciences**, v. 24, n. 1, p. 1–15, 15 jan. 2024.

OLAWADE, David B. *et al.* Using artificial intelligence to improve public health: a narrative review. **Frontiers in Public Health**, v. 11, p. 1196397, 2023.

OLIVEIRA-KUMAKURA, Ana Railka de Souza *et al.* Functionality and quality of life in Brazilian patients 6 months post-stroke. **Frontiers in neurology**, v. 14, 2023a.

OLIVEIRA-KUMAKURA, Ana Railka de Souza *et al.* Functionality and quality of life in Brazilian patients 6 months post-stroke. **Frontiers in Neurology**, v. 14, p. 1020587, 2023b.

OLIVEIRA-KUMAKURA, Ana Railka de Souza *et al.* Functionality and quality of life in Brazilian patients 6 months post-stroke. **Frontiers in Neurology**, v. 14, p. 1020587, 20 abr. 2023c.

ORFANOUDAKI, Agni *et al.* Machine learning provides evidence that stroke risk is not linear: The non-linear Framingham stroke risk score. **PloS one**, v. 15, n. 5, 1 maio 2020.

ORTELAN, Naiá *et al.* Evaluating the relationship between conditional cash transfer programme on preterm births: a retrospective longitudinal study using the 100 million Brazilian cohort. **BMC Public Health**, v. 24, n. 1, p. 1–16, 1 dez. 2024.

ORTIZ-GARCIA, Jorge *et al.* Recent advances in the management of transient ischemic attacks. **Faculty reviews**, v. 11, 22 jul. 2022.

OURIQUES MARTINS, Sheila C. *et al.* Priorities to reduce the burden of stroke in Latin American countries. **The Lancet. Neurology**, v. 18, n. 7, p. 674–683, 1 jul. 2019.

PAASCHE-ORLOW, Michael K.; WOLF, Michael S. The causal pathways linking health literacy to health outcomes. **American journal of health behavior**, v. 31 Suppl 1, n. SUPPL. 1, 2007.

PALMEIRA, Nathalia Campos *et al.* Análise do acesso a serviços de saúde no Brasil segundo perfil sociodemográfico: Pesquisa Nacional de Saúde, 2019. **Epidemiologia e Serviços de Saúde**, v. 31, n. 3, p. e2022966, 2022.

PAN, Yuan *et al.* The association between sleep deprivation and the risk of cardiovascular diseases: A systematic meta-analysis. **Biomedical Reports**, v. 19, n. 5, p. 78, 1 nov. 2023.

PANDIAN, Jeyaraj D. *et al.* Stroke systems of care in low-income and middle-income countries: challenges and opportunities. **The Lancet**, v. 396, n. 10260, p. 1443–1451, 31 out. 2020.

PANTOJA-RUIZ, Camila *et al.* Socioeconomic Status and Stroke: A Review of the Latest Evidence on Inequalities and Their Drivers. **Stroke**, v. 56, n. 3, p. 794–805, 1 mar. 2025.

PARIKH, Nisha I. *et al.* Adverse Pregnancy Outcomes and Cardiovascular Disease Risk: Unique Opportunities for Cardiovascular Disease Prevention in Women: A Scientific Statement From the American Heart Association. **Circulation**, v. 143, n. 18, p. E902–E916, 4 maio 2021.

PARRO, V. C. *et al.* Predicting COVID-19 in very large countries: The case of Brazil. **PloS one**, v. 16, n. 7, 1 jul. 2021.

PATHARKAR, Abhidnya *et al.* Predictive modeling of biomedical temporal data in healthcare applications: review and future directions. **Frontiers in Physiology**, v. 15, p. 1386760, 2024.

PATRIOTA, Pollyanna; KO MAUNG, Ko; MARQUES-VIDAL, Pedro. Reported recommendations to address cardiovascular risk factors differ by socio-economic status in Brazil. Results from the Brazilian National Health Survey 2019. **Preventive Medicine Reports**, v. 36, p. 102527, 1 dez. 2023.

PEKÇETIN, Emel *et al.* Urban versus rural older adults: occupational balance and quality of life comparison. **BMC Geriatrics**, v. 25, n. 1, p. 49, 1 dez. 2025.

PELISSARI, Daniele Maria; DIAZ-QUIJANO, Fredi Alexander. Household crowding as a potential mediator of socioeconomic determinants of tuberculosis incidence in Brazil. **PLOS ONE**, v. 12, n. 4, p. e0176116, 1 abr. 2017.

PEREIRA, Antonia Jaine da Silva; QUEIROZ, Silvana Nunes de. GERAÇÃO QUE NEM ESTUDA NEM TRABALHA NO NORDESTE BRASILEIRO. **Revista Econômica do Nordeste**, v. 54, n. 1, p. 67–86, 13 mar. 2023.

PEREIRA, Roberta Amorim *et al.* Sobrecarga dos cuidadores de idosos com acidente vascular cerebral. **Revista da Escola de Enfermagem da USP**, v. 47, n. 1, p. 185–192, 2013.

PEREIRA, Taísa Sabrina Silva *et al.* Effect of urinary sodium-to-potassium ratio change on blood pressure in participants of the longitudinal health of adults study - ELSA-Brasil. **Medicine**, v. 98, n. 28, p. e16278, 1 jul. 2019.

PITCHAI, R. *et al.* An Artificial Intelligence-Based Bio-Medical Stroke Prediction and Analytical System Using a Machine Learning Approach. **Computational intelligence and neuroscience**, v. 2022, 2022.

PONTES-NETO, Octávio Marques *et al.* Stroke awareness in Brazil: alarming results in a community-based study. **Stroke**, v. 39, n. 2, p. 292–296, fev. 2008.

POORTHUIS, Michiel H. F. *et al.* Female- and Male-Specific Risk Factors for Stroke: A Systematic Review and Meta-analysis. **JAMA Neurology**, v. 74, n. 1, p. 75–81, 1 jan. 2017.

POTTER, Joseph E. *et al.* Mapping the timing, pace, and scale of the fertility transition in Brazil. **Population and Development Review**, v. 36, n. 2, p. 283–307, jun. 2010.

PREDEBON, Mariane Lurdes *et al.* The capacity of informal caregivers in the rehabilitation of older people after a stroke. **Investigacion y Educacion en Enfermeria**, v. 39, n. 2, p. e03, 2021a.

PREDEBON, Mariane Lurdes *et al.* The capacity of informal caregivers in the rehabilitation of older people after a stroke. **Investigacion y Educacion en Enfermeria**, v. 39, n. 2, p. e03, 2021b.

PRIYAMVARA, Aditi *et al.* Periodontal Inflammation and the Risk of Cardiovascular Disease. **Current Atherosclerosis Reports**, v. 22, n. 7, 1 jul. 2020.

PUDDEPHATT, Jo Anne *et al.* Associations of common mental disorder with alcohol use in the adult general population: a systematic review and meta-analysis. **Addiction**, v. 117, n. 6, p. 1543–1572, 1 jun. 2022.

QU, Yang *et al.* Ultra-processed food consumption and risk of cardiovascular events: a systematic review and dose-response meta-analysis. **eClinicalMedicine**, v. 69, 1 mar. 2024.

QUARESMA, Guilherme *et al.* Transición de la fecundidad en municipios brasileños: un análisis exploratorio de datos transversales en 1991, 2000 y 2010. **Revista Latinoamericana de Población**, v. 17, p. e202219–e202219, 2 maio 2023.

RAJAGOPAL, Anjali *et al.* Machine Learning Operations in Health Care: A Scoping Review. **Mayo Clinic Proceedings: Digital Health**, v. 2, n. 3, p. 421–437, 1 set. 2024.

RAJKOMAR, Alvin; DEAN, Jeffrey; KOHANE, Isaac. Machine Learning in Medicine. **The New England journal of medicine**, v. 380, n. 14, p. 1347–1358, 4 abr. 2019.

RANTA, Annemarei *et al.* Environmental factors and stroke: Risk and prevention. **Journal of the neurological sciences**, v. 454, 15 nov. 2023.

RASELLA, Davide *et al.* Impact of primary health care on mortality from heart and cerebrovascular diseases in Brazil: a nationwide analysis of longitudinal data. **The BMJ**, v. 349, p. g4014, 3 jul. 2014.

REISSMANN, Daniel R. *et al.* Association between perceived oral and general health. **Journal of Dentistry**, v. 41, n. 7, p. 581–589, 1 jul. 2013.

RIBEIRO, Ícaro J. S. *et al.* Determinants of Stroke in Brazil: A Cross-Sectional Multivariate Approach from the National Health Survey. **Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association**, v. 27, n. 6, p. 1616–1623, 1 jun. 2018.

ROBINSON, Eric *et al.* Eating attentively: a systematic review and meta-analysis of the effect of food intake memory and awareness on eating1. **The American Journal of Clinical Nutrition**, v. 97, n. 4, p. 728, 1 abr. 2013.

ROCHMAH, Thinni Nurul *et al.* Economic burden of stroke disease: A systematic review. **International Journal of Environmental Research and Public Health**, v. 18, n. 14, p. 7552, 2 jul. 2021.

RODRIGUES, João Paulo Vilela *et al.* Use of statins for the secondary prevention of stroke: are we respecting the scientific evidences? **Journal of Stroke and Cerebrovascular Diseases**, v. 29, n. 8, 1 ago. 2020.

RODRIGUES, Paulo Rogério Melo *et al.* How many meals and snacks do Brazilians eat a day? Findings from the 2017-2018 Brazilian National Dietary Survey. **Cadernos de Saúde Pública**, v. 40, n. 2, p. e00009923, 19 fev. 2024.

ROSA, Ana *et al.* IMPACTO SOCIOECONÔMICO DO ACIDENTE VASCULAR CEREBRAL NO ESTADO DE RORAIMA: UM ESTUDO DE COORTE DE BASE HOSPITALAR. **Revista Brasileira de Neurologia e Psiquiatria**, v. 22, n. 2, p. 124–141, 2018.

SAHU, Adarsh; MISHRA, Jyotika; KUSHWAHA, Namrata. Artificial Intelligence (AI) in Drugs and Pharmaceuticals. **Combinatorial chemistry & high throughput screening**, v. 25, n. 11, p. 1818–1837, 8 dez. 2022.

SANTANA, Jerusa da Mota *et al.* Influence of conditional cash transfer program on prenatal care and nutrition during pregnancy: NISAMI cohort study. **São Paulo Medical Journal**, v. 140, n. 4, p. 595, 2022.

SANTIAGO, Emerson Rogério Costa *et al.* Prevalence of systemic arterial hypertension and associated factors among adults from the semi-arid region of Pernambuco, Brazil. **Arquivos Brasileiros de Cardiologia**, v. 113, n. 4, p. 687–695, 1 out. 2019.

SANTOS, Emily dos *et al.* Incidence, lethality, and post-stroke functional status in different Brazilian macro-regions: The SAMBA study (analysis of stroke in multiple Brazilian areas). **Frontiers in neurology**, v. 13, 15 set. 2022a.

SANTOS, Emily dos *et al.* Incidence, lethality, and post-stroke functional status in different Brazilian macro-regions: The SAMBA study (analysis of stroke in multiple Brazilian areas). **Frontiers in neurology**, v. 13, 15 set. 2022b.

SANTOS, Mariana G. R. *et al.* Factors associated with pre-drinking among nightclub patrons in the city of São Paulo. **Alcohol and alcoholism (Oxford, Oxfordshire)**, v. 50, n. 1, p. 95–102, 1 jan. 2015.

SANTOS, Mariana G. R. *et al.* Pre-drinking, alcohol consumption and related harms amongst Brazilian and British university students. **PLOS ONE**, v. 17, n. 3, p. e0264842, 1 mar. 2022c.

SANZ, M. *et al.* Periodontitis and cardiovascular diseases. Consensus report. **Global Heart**, v. 15, n. 1, 3 fev. 2020.

SARIDENA, Abhaya; SARIDENA, Ananya; KETHAR, Jothsna. Machine Learning for Risk Prediction of Cardiovascular Disease: Current Advances and Future Prospects. **Journal of Student Research**, v. 12, n. 4, 30 nov. 2023.

SATO, Priscila de Morais *et al.* Mothers' food choices and consumption of ultra-processed foods in the Brazilian Amazon: A grounded theory study. **Appetite**, v. 148, p. 104602, 1 maio 2020.

SCHMIDT, Alissa *et al.* Distribuição dos cursos de Odontologia e de cirurgiões-dentistas no Brasil: uma visão do mercado de trabalho. **Revista da ABENO**, v. 18, n. 1, p. 63–73, 28 mar. 2018.

SCHOTT, Eloise *et al.* Food availability and food insecurity in households in the state of Tocantins, Northern Brazil. **Revista de Nutricao**, v. 33, p. 1–12, 2020.

SELEME, Ana Luísa Goncalves Gomes Coelho *et al.* DESIGUALDADES NO TRATAMENTO DO AVC: UMA REVISÃO INTEGRATIVA DOS DETERMINANTES SOCIAIS DA SAÚDE. **ARACÊ**, v. 6, n. 2, p. 1283–1302, 7 out. 2024.

SELINGARDI, Sara de Almeida *et al.* Temporal patterns of food consumption and their association with cardiovascular risk in rotating shift workers. **Clinical Nutrition ESPEN**, v. 62, p. 95–101, 1 ago. 2024.

SHAO, Huiling *et al.* The feasibility and accuracy of machine learning in improving safety and efficiency of thrombolysis for patients with stroke: Literature review and proposed improvements. **Frontiers in neurology**, v. 13, 20 out. 2022.

SHARIFI, Sina *et al.* Dwelling characteristics and mental well-being in older adults: A systematic review. **Heliyon**, v. 10, n. 18, 30 set. 2024.

SHEER, Richard *et al.* Predictive Risk Models to Identify Patients at High-Risk for Severe Clinical Outcomes With Chronic Kidney Disease and Type 2 Diabetes. **Journal of Primary Care & Community Health**, v. 13, p. 21501319211063730, 1 jan. 2022.

SHIGA, Yuji *et al.* Effect of tooth loss and nutritional status on outcomes after ischemic stroke. **Nutrition**, v. 71, 1 mar. 2020.

SHIPLEY, Emily *et al.* Bridging the Gap Between Artificial Intelligence Research and Clinical Practice in Cardiovascular Science: What the Clinician Needs to Know. **Arrhythmia & Electrophysiology Review**, v. 11, n. 1, p. e03, 1 abr. 2022.

SHISHEHBORI, Farnoush; AWAN, Zainab. Enhancing Cardiovascular Disease Risk Prediction with Machine Learning Models. 29 jan. 2024.

SHMUELI, Galit. To Explain or to Predict? **Statistical Science**, v. 25, n. 3, p. 289–310, 2010.

SHOBAYO, Olamilekan *et al.* Prediction of Stroke Disease with Demographic and Behavioural Data Using Random Forest Algorithm. **Analytics 2023, Vol. 2, Pages 604-617**, v. 2, n. 3, p. 604–617, 2 ago. 2023.

SILVA, Marcelo Alves da; SANINE, Patricia Rodrigues. Interoperabilidade entre os Sistemas de Informação em Saúde Brasileiros: uma revisão integrativa. **Revista de Saúde Pública de Mato Grosso do Sul**, v. 3, n. 2, p. 17–29, 2020.

SILVA, Diego Augusto Santos *et al.* Physical activity to prevent stroke mortality in Brazil (1990-2019). **Revista da Sociedade Brasileira de Medicina Tropical**, v. 55, n. suppl 1, 2022.

SILVA, Luana Karoline Castro *et al.* Acidente vascular cerebral no Brasil: prevalência, limitações em atividade, acesso à saúde e tratamento fisioterapêutico. **Arquivos de neuro-psiquiatria**, v. 82, n. 12, p. 1–11, 1 dez. 2024.

SINGH, Amritpal *et al.* Class-Incremental Continual Learning for General Purpose Healthcare Models. 7 nov. 2023.

SIRSAT, Manisha Sanjay; FERMÉ, Eduardo; CÂMARA, Joana. Machine Learning for Brain Stroke: A Review. **Journal of stroke and cerebrovascular diseases : the official journal of National Stroke Association**, v. 29, n. 10, 1 out. 2020.

SMYTH, Andrew *et al.* Alcohol Intake as a Risk Factor for Acute Stroke: The INTERSTROKE Study. **Neurology**, v. 100, n. 2, p. E142–E153, 10 jan. 2023.

SOTO-CÁMARA, Raúl *et al.* Age-related risk factors at the first stroke event. **Journal of Clinical Medicine**, v. 9, n. 7, p. 1–12, 1 jul. 2020.

SOUTO, Rayone Moreira Costa Veloso *et al.* Prevalence of disability and associated functional limitations among older adults in Brazil. **PLOS Global Public Health**, v. 4, n. 11, p. e0003225, 14 nov. 2024.

STARCKE, Jonathan *et al.* The Effect of Data Leakage and Feature Selection on Machine Learning Performance for Early Parkinson's Disease Detection. **Bioengineering 2025, Vol. 12, Page 845**, v. 12, n. 8, p. 845, 6 ago. 2025a.

STARCKE, Jonathan *et al.* The Effect of Data Leakage and Feature Selection on Machine Learning Performance for Early Parkinson's Disease Detection. **Bioengineering 2025, Vol. 12, Page 845**, v. 12, n. 8, p. 845, 6 ago. 2025b.

STEIN, Caroline *et al.* Comparing estimates of intimate-partner violence against women across different data sources in brazil. **Population Medicine**, v. 5, n. Supplement, 26 abr. 2023a.

STEIN, Caroline *et al.* Comparing estimates of intimate-partner violence against women across different data sources in brazil. **Population Medicine**, v. 5, n. Supplement, 26 abr. 2023b.

STIGLIC, Gregor *et al.* Interpretability of machine learning based prediction models in healthcare. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 10, n. 5, 20 fev. 2020.

STOCHERO, Luciane; PINTO, Liana Wernersbach. Prevalence of violence against women living in rural areas and associated factors: a cross-sectional study based on the 2019 National Health Survey. **Ciência & Saúde Coletiva**, v. 29, n. 1, p. e20452022, 2024.

STOLSES, Priscila Maria *et al.* Prevalência e fatores associados ao acidente vascular cerebral em idosos no Brasil, 2019. 6 jun. 2023.

SU, Po Yuan *et al.* Machine Learning Models for Predicting Influential Factors of Early Outcomes in Acute Ischemic Stroke: Registry-Based Study. **JMIR medical informatics**, v. 10, n. 3, 1 mar. 2022.

SUGLIA, Shakira F.; SAPRA, Katherine J.; KOENEN, Karestan C. Violence and Cardiovascular Health: A Systematic Review. **American journal of preventive medicine**, v. 48, n. 2, p. 205, 1 fev. 2015.

SUNNY, Md Nagib Mahfuz *et al.* Optimizing Healthcare Outcomes through Data-Driven Predictive Modeling. **Journal of Intelligent Learning Systems and Applications**, v. 16, n. 4, p. 384–402, 9 set. 2024.

TALWAR, Ashna *et al.* Performance of advanced machine learning algorithms overlogistic regression in predicting hospital readmissions: A meta-analysis. **Exploratory Research in Clinical and Social Pharmacy**, v. 11, p. 100317, 1 set. 2023.

TCHERO, Huidi *et al.* Telerehabilitation for stroke survivors: Systematic review and meta-analysis. **Journal of Medical Internet Research**, v. 20, n. 10, p. e10867, 26 out. 2018.

TEREZA, Denise M. *et al.* Stroke epidemiology in southern Brazil: Investigating the relationship between stroke severity, hospitalization costs, and health-related quality of life. **Anais da Academia Brasileira de Ciências**, v. 94, n. 2, p. e20211492, 13 jun. 2022.

TETZLAFF, Juliane *et al.* Income inequalities in stroke incidence and mortality: Trends in stroke-free and stroke-affected life years based on German health insurance data. **PloS one**, v. 15, n. 1, 1 jan. 2020.

THOMPSON, Stephanie G. *et al.* Geographic Disparities in Stroke Outcomes and Service Access: A Prospective Observational Study. **Neurology**, v. 99, n. 4, p. E414–E426, 26 jul. 2022.

THØRRISEN, Mikkel Magnus *et al.* Are workplace factors associated with employee alcohol use? The WIRUS cross-sectional study. **BMJ Open**, v. 12, n. 10, p. e064352, 13 out. 2022.

TOMASI, Elaine *et al.* Qualidade da atenção pré-natal na rede básica de saúde do Brasil: indicadores e desigualdades sociais. **Cadernos de Saúde Pública**, v. 33, n. 3, p. e00195815, 2017.

TRAMACERE, Irene *et al.* Comparison of statins for secondary prevention in patients with ischemic stroke or transient ischemic attack: A systematic review and network meta-analysis. **BMC Medicine**, v. 17, n. 1, p. 1–12, 26 mar. 2019.

TRINDADE, Raquel Elias da *et al.* Contraception use and family planning inequalities among Brazilian women. **Ciencia e Saude Coletiva**, v. 26, n. suppl 2, p. 3493–3504, 2021.

VAZ, Davis Wilker Nascimento *et al.* Perfil epidemiológico do Acidente Vascular Cerebral no Estado do Amapá, Brasil. **Research, Society and Development**, v. 9, n. 8, p. e938986642–e938986642, 2 ago. 2020.

VENKATASUBRAMANIAM, Ashwini *et al.* Decision trees in epidemiological research. **Emerg Themes Epidemiol**, v. 14, p. 11, 2017.

VICTORA, Cesar. Socioeconomic inequalities in Health: Reflections on the academic production from Brazil. **International Journal for Equity in Health**, v. 15, n. 1, p. 1–3, 17 nov. 2016.

VIEIRA, Marina Mendes Lopes *et al.* Functional limitation in the older Brazilian adults: Association with multimorbidity and socioeconomic conditions. **PLOS ONE**, v. 18, n. 11, p. e0294935, 1 nov. 2023.

VIEIRA, Yohana Pereira *et al.* Inequities in the food consumption of the Brazilian population in the face of the COVID-19 pandemic. **Discover public health**, v. 22, n. 1, p. 1–12, 1 dez. 2025.

VIGNEAU, Evelyne. Clustering of variables for enhanced interpretability of predictive models. **Informatica (Slovenia)**, v. 45, n. 4, p. 507–516, 18 ago. 2020.

VINCENS, Natalia; STAFSTRÖM, Martin. Income Inequality, Economic Growth and Stroke Mortality in Brazil: Longitudinal and Regional Analysis 2002-2009. **PloS one**, v. 10, n. 9, 9 set. 2015a.

VINCENS, Natalia; STAFSTRÖM, Martin. Income Inequality, Economic Growth and Stroke Mortality in Brazil: Longitudinal and Regional Analysis 2002-2009. **PLoS ONE**, v. 10, n. 9, p. e0137332, 9 set. 2015b.

VINCENS, Natalia; STAFSTRÖM, Martin. Income Inequality, Economic Growth and Stroke Mortality in Brazil: Longitudinal and Regional Analysis 2002-2009. **PLoS ONE**, v. 10, n. 9, p. e0137332, 9 set. 2015c.

VIRANI, Salim S. *et al.* Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association. **Circulation**, v. 143, n. 8, p. E254–E743, 23 fev. 2021.

VITALE, Marilena *et al.* Ultra-Processed Foods and Human Health: A Systematic Review and Meta-Analysis of Prospective Cohort Studies. **Advances in Nutrition**, v. 15, n. 1, p. 100121, 1 jan. 2023.

VITTURI, Bruno Kusznir; GAGLIARDI, Rubens José. Effects of statin therapy on outcomes of ischemic stroke: a real-world experience in Brazil. **Arquivos de Neuro-Psiquiatria**, v. 78, n. 8, p. 461–467, 1 set. 2020.

VU, Thien *et al.* Machine Learning Approaches for Stroke Risk Prediction: Findings from the Suita Study. **Journal of cardiovascular development and disease**, v. 11, n. 7, 1 jul. 2024.

WANG, Siping *et al.* Socioeconomic status predicts the risk of stroke death: A systematic review and meta-analysis. **Preventive Medicine Reports**, v. 19, p. 101124, 1 set. 2020a.

WANG, Wenjuan *et al.* A systematic review of machine learning models for predicting outcomes of stroke with structured data. **PloS one**, v. 15, n. 6, 1 jun. 2020b.

WANG, Yuanxin *et al.* The Dose-Response Associations of Sugar-Sweetened Beverage Intake with the Risk of Stroke, Depression, Cancer, and Cause-Specific Mortality: A Systematic Review and Meta-Analysis of Prospective Studies. **Nutrients**, v. 14, n. 4, p. 777, 1 fev. 2022a.

WANG, Zhongting *et al.* Sedentary behavior and the risk of stroke: A systematic review and dose-response meta-analysis. **Nutrition, Metabolism and Cardiovascular Diseases**, v. 32, n. 12, p. 2705–2713, 1 dez. 2022b.

WEISSLER, E. Hope *et al.* The role of machine learning in clinical research: transforming the future of evidence generation. **Trials**, v. 22, n. 1, p. 1–15, 1 dez. 2021.

WELTEN, Sabrina J. G. C. *et al.* Age at Menopause and Risk of Ischemic and Hemorrhagic Stroke. **Stroke**, v. 52, n. 8, p. 2583–2591, 1 ago. 2021.

WILKE, Jan *et al.* Physical Activity During Lockdowns Associated with the COVID-19 Pandemic: A Systematic Review and Multilevel Meta-analysis of 173 Studies with 320,636 Participants. **Sports Medicine - Open**, v. 8, n. 1, p. 125, 1 dez. 2022.

WU, Pensée *et al.* Preeclampsia and future cardiovascular health. **Circulation: Cardiovascular Quality and Outcomes**, v. 10, n. 2, 1 fev. 2017a.

WU, Pensée *et al.* Preeclampsia and future cardiovascular health. **Circulation: Cardiovascular Quality and Outcomes**, v. 10, n. 2, 1 fev. 2017b.

XIUYUN, Wen *et al.* Education and stroke: evidence from epidemiology and Mendelian randomization study. **Scientific Reports**, v. 10, n. 1, p. 1–11, 1 dez. 2020.

YAN, Lijing L. *et al.* Prevention, management, and rehabilitation of stroke in low- and middle-income countries. **eNeurologicalSci**, v. 2, p. 21–30, 1 mar. 2016.

YAO, Shu Chin *et al.* Physical function, depressive symptoms, and quality of life with post-acute stroke care. **Collegian**, v. 30, n. 3, p. 475–482, 1 jun. 2023.

YI, Yunhao *et al.* Effectiveness of non-pharmacological therapies for treating post-stroke depression: A systematic review and network meta-analysis. **General Hospital Psychiatry**, v. 90, p. 99–107, 1 set. 2024.

YUEN, Thomas *et al.* Digital Transformation in Healthcare: Technology Acceptance and Its Applications. **International Journal of Environmental Research and Public Health 2023, Vol. 20, Page 3407**, v. 20, n. 4, p. 3407, 15 fev. 2023.

ZHANG, Yanmei *et al.* Parity and Risk of Stroke among Chinese Women: Cross-sectional Evidence from the Dongfeng-Tongji Cohort Study OPEN. **Nature Publishing Group**, 2015.

ZHONG, Charlie *et al.* Electronic Screen Use and Sleep Duration and Timing in Adults. **JAMA Network Open**, v. 8, n. 3, p. e252493–e252493, 3 mar. 2025.

11.Apêndices

Considerando a natureza abrangente desta pesquisa e a riqueza do material obtido ao longo do processo investigativo, foi reunida uma quantidade expressiva de imagens, tabelas e documentos complementares. A inserção integral desse conteúdo no corpo do trabalho resultaria em um volume excessivo de páginas, o que poderia comprometer a fluidez da leitura e dificultar a interpretação dos resultados principais.

Com o intuito de preservar a clareza e a objetividade do texto, bem como de oferecer ao leitor um acesso mais ágil e organizado a esse conjunto de informações, optou-se por disponibilizar o material de forma digital. Para tanto, foi criada uma pasta específica em ambiente eletrônico, que concentra todo o conteúdo suplementar reunido e devidamente categorizado.

O recurso de disponibilização por meio de link de acesso tem por finalidade ampliar a facilidade de consulta e interpretação, permitindo que o leitor explore o material complementar de maneira autônoma, navegando pelas imagens, quadros, planilhas e registros visuais conforme sua necessidade ou interesse. Tal estratégia, além de atender a critérios de praticidade, garante maior transparência e acessibilidade, possibilitando que o exame das evidências visuais seja realizado em sua integralidade, sem as limitações físicas do formato impresso ou mesmo do documento digital em PDF.

Dentro dessa pasta eletrônica, organizada de forma padronizada para cada Unidade Federativa analisada, encontram-se os seguintes tipos de arquivos:

- Arquivos de correlação Top-10 (soma das 3 execuções – “runs”): Mapas de calor (heatmaps) das 10 variáveis mais relevantes identificadas pelo modelo Random Forest, considerando a soma das três execuções (run1, run2 e run3). Estão disponíveis em formato PNG (visualização gráfica) e XLSX (planilha de correlações numéricas). Exemplo: corr_top10_SUM_AL.png / corr_top10_SUM_AL.xlsx.

- Arquivos de correlação Top-30: Apresentam as correlações entre as 30 variáveis mais relevantes de cada estado, em PNG e XLSX, permitindo observar não apenas os núcleos centrais, mas também variáveis secundárias. Exemplo: correlacao_top30_AL.png.

- Arquivos de correlação Top-100: Disponíveis em XLSX, reúnem as 100 variáveis mais importantes, possibilitando análises estatísticas detalhadas e replicações adicionais. Exemplo: correlacao_top100_AL.xlsx.

- Arquivos de correlação completa (run1, run2, run3): Mapas de calor que representam a correlação entre todas as variáveis do banco para o estado em questão. São arquivos extensos, úteis para análises mais exploratórias e de consistência. Disponíveis em PNG e XLSX. Exemplo: corr_full_run1_AL.png / corr_full_run1_AL.xlsx.

- Arquivos de correlação Top-10 por execução individual (run1, run2, run3): Matrizes de correlação considerando apenas uma execução específica do Random Forest, úteis para avaliar a estabilidade entre runs. Disponíveis em PNG e XLSX. Exemplo: corr_top10_run1_AL.png / corr_top10_run1_AL.xlsx.

- Arquivos em formato MATLAB (.mat): Contêm os resultados brutos das execuções do Random Forest para cada estado, possibilitando reanálises ou ajustes diretamente em ambiente MATLAB. Exemplo: RF_runs_AL.mat.

Em todos os casos, quando houver indicação de “run1” no nome do arquivo, também estarão disponíveis os equivalentes run2 e run3, assegurando a transparência e a replicabilidade dos resultados.

Essa organização foi planejada para que o leitor possa, conforme seu interesse, aprofundar-se progressivamente, partindo das análises mais sintéticas (Top-10) até as matrizes completas de correlação, preservando a integridade e a riqueza do material obtido. Dessa forma, o presente trabalho mantém sua organização estrutural de acordo com as normas acadêmicas da ABNT, ao mesmo tempo em que recorre a um suporte eletrônico como recurso auxiliar, garantindo tanto o rigor científico quanto a acessibilidade das informações.

Segue o link:

https://drive.google.com/file/d/1F4EAC03IGdOnSASSUOdAbgvKSoOK0n1M/view?usp=drive_link