Análise de Fatores de Risco para Acidente Vascular Cerebral (AVC)

Pesquisa Nacional de Saúde 2019

Iury Mota Santos, iury.santos@sga.pucminas.br

Thiago Andrade Monteiro, o.thiagoamonteiro@gmail.com

Iago Paiva Faria, iago.faria.1625316@sga.pucminas.br

Professor: **Gabriel Fonseca**

Curso de Ciência de Dados, Instituto de Informática e Ciências Exatas
Pontifícia Universidade de Minas Gerais (PUC MINAS), Belo Horizonte – MG – Brasil

Resumo Executivo

O Acidente Vascular Cerebral (AVC), conhecido como derrame, é uma condição neurológica grave causada pela interrupção do fluxo sanguíneo no cérebro, seja por obstrução (isquêmico) ou rompimento de um vaso (hemorrágico). Essa interrupção gera falta de oxigênio e nutrientes, resultando em sequelas motoras, cognitivas ou até óbito.

Globalmente, o AVC está entre as principais causas de morte e incapacidade. No Brasil, é uma das maiores causas de mortalidade, relacionado a fatores como hipertensão, diabetes, tabagismo, sedentarismo e envelhecimento. Além dos impactos individuais, gera elevados custos sociais e econômicos.

Sobre a Pesquisa

A Pesquisa Nacional de Saúde (PNS 2019), realizada pelo IBGE e Ministério da Saúde, fornece dados representativos sobre condições de saúde e prevalência de doenças crônicas.

Este projeto propõe analisar a PNS 2019 para identificar relações entre variáveis demográficas, clínicas e comportamentais associadas ao AVC. O objetivo é compreender padrões relevantes para a saúde pública e apoiar a conscientização social.

Introdução e Contextualização

Foco do Trabalho

Análise do Acidente Vascular Cerebral (AVC) utilizando dados da PNS 2019, identificando fatores de risco e padrões associados à doença através de variáveis demográficas, clínicas e comportamentais.

Contexto Nacional

O AVC é responsável por cerca de 100 mil mortes anuais no Brasil, gerando custos sociais e econômicos significativos. É uma das principais causas de morte e incapacidade no país.

Base de Dados

A PNS 2019, realizada pelo IBGE, constitui uma base sólida para análises quantitativas sobre prevalência e fatores de risco, permitindo identificar padrões relevantes na população brasileira.

Problema e Objetivos

O Problema

Apesar da elevada incidência do AVC no Brasil, os fatores de risco associados frequentemente não são devidamente identificados ou monitorados pela população, dificultando ações preventivas eficazes. A ausência de análises integradas limita a compreensão dos determinantes dessa condição.

Objetivo Geral

Desenvolver uma análise baseada na PNS 2019 para investigar quais variáveis demográficas, clínicas e comportamentais estão associadas à ocorrência de AVC, possibilitando a identificação de padrões relevantes para a saúde pública.

> **Pergunta Orientada a Dados:** Quais fatores demográficos, clínicos e comportamentais estão mais associados à prevalência de Acidente Vascular Cerebral (AVC) na população brasileira, de acordo com os dados da PNS 2019?

01

Analisar variáveis demográficas

Investigar a relação entre características demográficas e a ocorrência do AVC

02

Avaliar fatores de risco

Examinar a influência de fatores clínicos e comportamentais como hipertensão, diabetes, tabagismo e sedentarismo

03

Construir visualizações

Desenvolver gráficos e tabelas para compreender os padrões identificados nos dados

04

Desenvolver modelo preditivo

Construir e otimizar modelo utilizando variáveis selecionadas

05

Avaliar desempenho

Validar métricas do modelo para assegurar análise robusta e interpretável

Justificativas e Público-Alvo

Justificativas

O desenvolvimento deste trabalho é motivado pela necessidade de compreender melhor os fatores associados à ocorrência do AVC na população brasileira. A escolha de analisar os dados da PNS 2019 permite explorar uma base confiável e representativa, possibilitando a identificação de padrões demográficos, clínicos e comportamentais relacionados à doença.

A análise contribui para estratégias de prevenção e conscientização em saúde pública, fornecendo insights baseados em evidências para formuladores de políticas e profissionais da saúde.

Público-Alvo

Pacientes

Indivíduos diagnosticados com AVC que buscam compreender melhor sua condição

Grupos de Risco

Idosos, hipertensos, diabéticos, pessoas com histórico familiar de AVC, sedentários e fumantes

Profissionais da Saúde

Envolvidos no acompanhamento, diagnóstico e reabilitação de pacientes com AVC

Gestores Públicos

Formuladores de políticas públicas em saúde que necessitam de dados para tomada de decisão

Análise Exploratória dos Dados

293.726

Amostra Total

Indivíduos com 15 anos ou mais participantes da pesquisa

96,5%

Taxa de Resposta

Refletindo alta qualidade e confiabilidade dos dados coletados

108.525

Domicílios

Distribuídos em 2.000 municípios de todas as unidades da federação

A Pesquisa Nacional de Saúde (PNS) 2019 apresenta representatividade nacional, garantindo a qualidade das análises. Para iniciar o estudo, foi necessário preparar o ambiente e os dados disponíveis através da leitura do layout SAS e carregamento do arquivo de dados em DataFrame do pandas.

Variáveis e Dicionário

```
dicionario_variaveis = {
        "V0001": "Unidade da Federação (UF)",
        "V0022": "Peso do morador selecionado",
        "V0031": "Região geográfica",
        "V0026": "Situação do domicílio (urbano/rural)",
        "A00601": "Forma como a água chega ao domicílio (canalizada, sem canalização etc.)",
        "Q068": "Diagnóstico médico de AVC (acidente vascular cerebral)",
        "C006": "Sexo do morador (masculino/feminino)",
        "C008": "Idade do morador (em anos)",
        "Q002": "Diagnóstico médico de hipertensão arterial",
        "Q030": "Diagnóstico médico de diabetes",
        "Q060": "Diagnóstico médico de colesterol alto",
        "P050": "Atualmente, o(a) Sr(a) fuma algum produto do tabaco",
        "Q092": "Diagnóstico médico de depressão",
        "W00101": "Peso do morador (em kg)",
        "W00201": "Altura do morador (em cm)",
        "I001": "Autoavaliação do estado de saúde (muito bom, bom, regular etc.)",
        "P035": "Prática de atividade física (tempo semanal)",
        "F00101": "Rendimento domiciliar per capita",
}
```

Diagrama Conceitual dos Fatores de Risco

O desenho amostral possibilita uma análise detalhada das condições de saúde, do acesso a serviços e dos comportamentos da população. A partir dele, realizamos análise de correlação para medir a força da associação entre fatores e a ocorrência de AVC.

Ajuste Metodológico

Após exploração inicial, identificou-se que algumas variáveis não estavam presentes no conjunto de dados final. Foi necessário realizar ajuste no escopo, removendo variáveis indisponíveis e focando nos dados efetivamente validados.

Resultados Amostrais e Qualidade dos Dados

A análise da qualidade dos dados revela padrões importantes sobre a cobertura e completude das variáveis. Enquanto algumas possuem cobertura praticamente completa, outras apresentam valores faltantes significativos que devem ser considerados nas análises.

Dados Ausentes

A análise de dados ausentes revela que algumas variáveis apresentam alta proporção de registros faltantes. Destacam-se como mais problemáticas as variáveis relacionadas a condições de saúde e estilo de vida:

- Diagnóstico médico de problema no coração: 98,39% ausentes
- Altura do indivíduo: 97,77% ausentes
- Peso do indivíduo: 97,76% ausentes

Variáveis demográficas básicas como situação do domicílio, região geográfica e unidade da federação não apresentam registros ausentes, garantindo base consistente para análises agregadas.

Detecção de Outliers

Foi realizada a detecção de outliers nas principais variáveis numéricas utilizando o método do IQR (Intervalo Interquartil):

Variável	Outliers	Limite Inferior	Limite Superior
Peso do morador	4.457	0.00	8.00
Altura (cm)	21	135.70	192.50
Idade (anos)	7	-35.50	104.50
Atividade física	0	-2.50	9.50

Análise de Correlações e Padrões

O algoritmo calcula a correlação de todas as variáveis numéricas com a variável alvo AVC, substituindo códigos pelos respectivos nomes e classificando como positiva ou negativa.

Desafio Metodológico: Correlação de Pearson

A primeira abordagem utilizou correlação linear de Pearson, mas o resultado indicou correlação negativa contraintuitiva, sugerindo que o aumento da idade estaria associado a menor ocorrência de AVC.

Para superar essa limitação, optou-se pelo teste **Qui-Quadrado** de independência, ideal para verificar associação entre variáveis categóricas, tratando cada faixa etária como categoria distinta.

Validação Estatística

O teste Qui-Quadrado confirma que existe associação estatisticamente significativa entre faixa etária e diagnóstico de AVC, validando a hipótese inicial de que as variáveis estão relacionadas.

Evidência Visual: Idade e AVC

O gráfico serve como evidência visual definitiva que corrobora o teste Qui-Quadrado, demonstrando que o risco aumenta progressivamente com o envelhecimento.

Padrões Contraintuitivos e Fatores de Confusão

Atividade Física

Demonstrou aparente correlação positiva com AVC. Isso ocorre porque pessoas idosas ou aposentadas, com risco naturalmente elevado, frequentemente mantêm ritmo de atividade física maior.

Tabagismo

Indivíduos que fumam menos apresentam maior incidência de AVC nos dados. Pessoas mais velhas têm maior risco e muitos deixam de fumar por questões de saúde.

Depressão

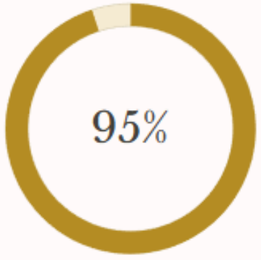
Aparece fortemente associada ao AVC. Parte é real, mas outra é explicada pela idade: pessoas idosas tendem a apresentar tanto mais AVC quanto mais diagnósticos de depressão.

Indução de Modelos Preditivos

Modelo 1: K-Nearest Neighbors (KNN)

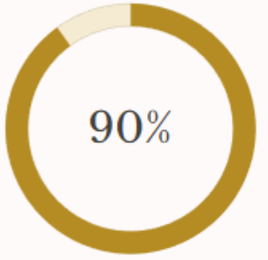
O KNN é um algoritmo supervisionado baseado na similaridade entre observações. Sua premissa é que instâncias próximas no espaço de características tendem a pertencer à mesma classe.

Para preparação dos dados, realizou-se normalização das variáveis numéricas e aplicou-se SMOTE para balancear as classes da variável target, devido à forte desproporção entre indivíduos com e sem diagnóstico de AVC.



Acurácia

Desempenho elevado após balanceamento e ajuste de hiperparâmetros



Precision/Recall

Valores acima de 0,9 para ambas as classes

O modelo KNN apresentou desempenho superior após balanceamento, conseguindo diferenciar adequadamente indivíduos com e sem histórico de AVC. O bom desempenho está relacionado à capacidade do KNN de capturar padrões locais e relações não lineares nos dados.

Modelo 2: Regressão Logística

A Regressão Logística é um método estatístico de classificação que estima a probabilidade de uma instância pertencer a uma classe. Funciona ajustando uma função sigmoide que transforma valores previstos em probabilidades entre 0 e 1.

Os resultados mostraram bom desempenho geral na capacidade de prever casos de AVC, com métricas de precisão e recall equilibradas.

Vantagem Interpretativa

A Regressão Logística fornece forma de interpretar a influência de cada variável no desfecho. Cada coeficiente indica o quanto uma variável aumenta ou diminui a probabilidade de AVC, permitindo compreender quais fatores têm maior impacto.

Modelo 3: Rede Neural (MLP)

A Rede Neural Multi-Layer Perceptron (MLP) é capaz de aprender padrões complexos e não lineares através de múltiplas camadas de neurônios conectadas. Foram testadas duas versões: sem e com SMOTE.

1

MLP sem SMOTE

Acurácia de 97,8%, mas recall muito baixo para classe minoritária (AVC). O modelo aprendeu a classificar indivíduos sem AVC, mas praticamente ignorou casos positivos.

2

MLP com SMOTE

Acurácia de 92,4%, com melhora no recall da classe minoritária. O balanceamento permitiu que a MLP aprendesse padrões de AVC, aumentando recall de 0% para 22%.