

Perfil de Adultos que Sofreram AVC: Uma análise comparativa de técnicas de Aprendizado de Máquina

Rafael Cabral Campos Pinheiro¹, Beethoven M. Andrade¹,
Cristiane Neri Nobre¹, Luis Enrique Zárate¹

¹Pontifical Catholic University of Minas Gerais (PUC-Minas)

{Rafael.pinheiro26@hotmail.com; nobre@pucminas.br; zarate@pucminas.br}

Abstract. *This study applied Machine Learning to characterize Stroke patients using 2019 National Health Survey (PNS) data. Focusing on the 55–77 age group (defined by cumulative frequency), the research involved rigorous pre-processing and hyperparameter tuning. The models effectively identified cases and generated relevant rules for future diagnosis.*

Resumo. *Este trabalho aplicou Machine Learning para caracterizar indivíduos com AVC utilizando dados da PNS 2019. O estudo focou na faixa etária de 55 a 77 anos, definida via frequência acumulada. Após etapas de pré-processamento e otimização de hiperparâmetros, os modelos demonstraram alta eficácia na detecção dos casos e forneceram regras importantes para o diagnóstico da doença.*

Palavras chaves: *Data mining, public health, data balancing.*

1. INTRODUÇÃO

O Acidente Vascular Cerebral (AVC) constitui um dos desafios mais críticos da saúde pública, com impactos severos tanto no cenário brasileiro quanto no panorama global. Dados da Sociedade Brasileira de Doenças Cerebrovasculares (SBDCV) indicam que, apenas no ano de 2024, a patologia foi responsável por aproximadamente 85.427 óbitos no Brasil, evidenciando a urgência de estratégias preventivas eficazes.

Com isso, apresente-se alguns autores importantes para o trabalho. Rodrigues et al. [Rodrigues et al. 2017] analisa os principais fatores de risco associados ao AVC, destaca-se a influência de doenças crônicas e hábitos de vida. Oliveira e Maciel [Oliveira and Maciel 2021] discutem a importância da prevenção por meio da adoção de hábitos saudáveis. Santos et al. [Santos et al. 2020] investigam a relação entre alimentos ultraprocessados e fatores que atrapalham o sistema cardiovascular e metabólico. Dritsas e Trigka [Dritsas and Trigka 2022] aplicam técnicas de aprendizado de máquina para a predição do risco de AVC, comparando diferentes algoritmos.

Este trabalho tem como objetivo a aplicação de técnicas de aprendizado de máquina para a caracterização de perfis de risco na população brasileira, com ênfase na faixa etária entre 55 e 77 anos. A delimitação etária justifica-se pelos altos eventos vasculares neste corte. O procedimento metodológico pauta-se na otimização de modelos e no ajuste de hiperparâmetros, visando à identificação de padrões críticos de comportamento social e estilo de vida.

A análise propõe, ainda, uma investigação aprofundada sobre a correlação entre a patologia e o consumo de alimentos saudáveis. Destaca-se a inclusão do atributo "Consumo de sódio" como indicador representativo de hábitos de risco, devido à sua baixa entropia e elevada relevância preditiva. O estudo utiliza microdados da Pesquisa Nacional de Saúde (PNS 2019), base populacional que oferta informações sobre o perfil epidemiológico do país.

2. METODOLOGIA

2.1. Métodos

O pipeline do processamento para descoberta de conhecimento foi estruturado nas seguintes etapas:

Tabela 1. Resumo das etapas da metodologia.

Etapa	Descrição
1	Definição da faixa etária de maior incidência de AVC (55 a 77 anos) por meio da análise da variável idade da PNS.
2	Construção do modelo conceitual com base na literatura, contemplando hábitos alimentares, saúde mental, condições socioeconômicas, histórico familiar e antropometria.
3	Seleção de 25 atributos relevantes, incluindo variáveis demográficas, clínicas, hábitos de vida e indicadores socioeconômicos.
4	Comparação entre KNN Imputer e MissForest para tratamento de valores ausentes, sendo selecionado o MissForest.
5	Identificação e remoção de outliers utilizando o método IQR ($1,5 \times IQR$), removendo 602 registros.
6	Engenharia de atributos com cálculo do IMC e consolidação do consumo semanal de proteínas.

3. EXPERIMENTOS E ANÁLISE DE RESULTADOS

3.1. Definição do domínio de problema

Para dar início ao estudo, foi primeiro definida a faixa etária-alvo, para o escopo da população considerada. Para isso, foi calculada a frequência relativa acumulada para o atributo "idade do entrevistado". Sendo assim, a faixa etária de inflexão da curva entre 55 e 77 anos, correspondente a 20% e 80% de indivíduos que tiveram AVC.

3.2. Dimensões escolhidas com o modelo conceitual

Tabela 2. Resumo das etapas da metodologia.

Variáveis extraídas na base de dados da PNS		
Dimensão	Aspecto	Variável
Hábitos alimentares	A Dieta e ingestão de alimentos e bebidas:	Módulo P – Estilos de Vida: P6a até P26a, P6b até P24a
Hábitos de saúde	Tabagismo	Módulo P – Estilos de Vida: P36a até P49a
	Atividade física	Módulo P – Estilos de Vida: P30a até P35a
Condições físicas e mentais	Doenças associadas	Módulo Q – Doenças Crônicas: Q30a, Q32a, Q38a
	Doenças pós AVC	Módulo G – Pessoas com Deficiência
	Saúde mental	Módulo Q – Doenças Crônicas: Q33a (BD – PNS). Módulo P – Estilos de Vida: questões sobre percepção de saúde
Condições socioeconômicas	Plano de saúde	Módulo J – Utilização de serviços de saúde
	Renda domiciliar	Módulo F – Rendimentos
	Trabalho	Módulo E – Características do trabalho
Características do indivíduo	Sexo e idade	Módulo C – Características gerais dos moradores
Genética / histórico familiar	Doenças crônicas familiar	Módulo Q – Doenças Crônicas
Registros corporais	Peso, altura e IMC	Módulo W – Antropometria: P1a, P4a, W000201, W000202

Fonte: <https://www.pns.icict.fiocruz.br/>

3.3. Combinação de atributos

Após o tratamento de outliers e de dados ausentes, foi aplicado um procedimento para fusão de atributos.

“Quantos dias come proteínas?”: Para reduzir a dimensionalidade do conjunto de dados, foram combinados atributos que remetessem ao consumo de proteínas, são eles: “Quantos dias come carne vermelha”, “Quantos dias comem peixe?” e “Quantos dias comem frango?”.

Para ajustar a escala, foi utilizada a soma agregada, realizando o somatório direto dos valores de frequência dos três atributos para cada indivíduo, representando o número total de dias de consumo de proteína. Após isso, considerando o contexto da frequência dietética, que é o número de dias em uma semana, o resultado da soma agregada foi limitado a um valor máximo de 7. Assim, qualquer soma resultante maior que 7 foi saturada em 7, enquanto somas menores ou iguais a 7 mantiveram seu valor original.

“IMC”: O cálculo, e categorização do Índice de Massa Corporal IMC foi ajustado de acordo com a escala da Organização Mundial da Saúde - OMS.

$$IMC = \frac{\text{peso}}{\text{altura}^2}$$

- Se $IMC < 18.5$: ‘Abaixo do peso’
- Se $18.5 \leq IMC < 24.99$: ‘Peso normal’
- Se $25 \leq IMC < 29.99$: ‘Sobrepeso’
- Se $30 \leq IMC < 34.99$: ‘Obesidade grau I’
- Se $35 \leq IMC < 39.99$: ‘Obesidade grau II’
- Se $IMC \geq 40$: ‘Obesidade grau III’

3.4. Balanceamento do conjunto de dados

Para lidar com o desbalanceamento do conjunto de dados, adotou-se uma estratégia híbrida. Sendo assim, foi aplicado um procedimento de subamostragem seletiva, conhecida como (*under sampling*) da classe majoritária. Após isso, foi aplicado a técnica SMOTE (*Synthetic Minority Oversampling Technique*) à classe minoritária, que gera dados sintéticos através da interpolação linear entre os vizinhos próximos.

3.5. Otimização de Hiperparâmetros e Validação Cruzada

O refinamento do hiperparâmetro de Mínimo de Instâncias por Nó (MNIN) permitiu avaliar o impacto da topologia do modelo tanto na precisão global quanto na complexidade das regras geradas. Conforme ilustrado nas análises das Figuras 1 e 2, o aumento do MNIN induz uma tendência de crescimento na medida da precisão do diagnóstico de AVC.

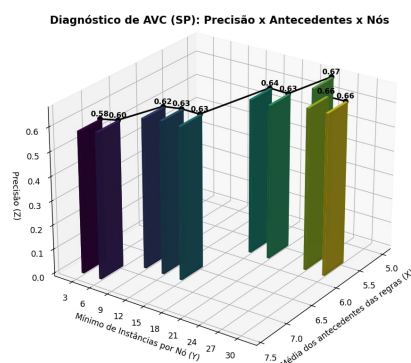


Figura 1. Relação tridimensional entre a Precisão, o Mínimo de Instâncias por Nó e a Média dos Antecedentes das Regras.

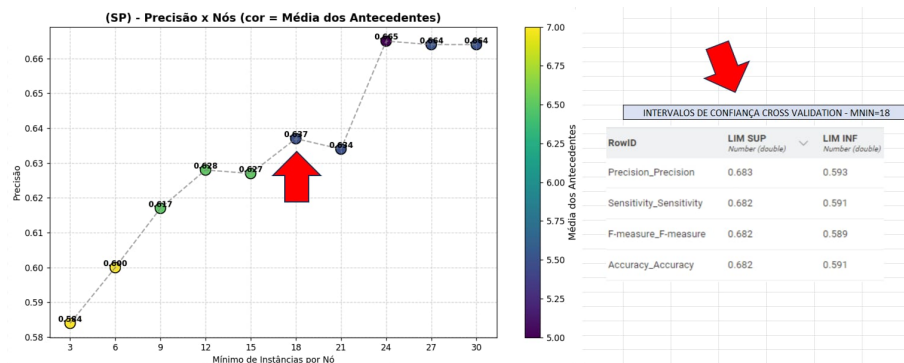


Figura 2. Análise de Precisão por Mínimo de Instâncias por Nó e Intervalos de Confiança obtidos via Cross-Validation para MNIN = 18.

O ponto de parametrização da árvore (MNIN) foi ratificado via validação cruzada (*Cross-Validation*). Os intervalos de confiança gerados para a precisão (limite inferior de 0,593 e superior de 0,683) atestam a estabilidade local do modelo. Essa confiabilidade garante que as regras de associação derivadas dessa configuração representam de forma estável a distribuição da população analisada.

3.6. Análise dos Resultados

A avaliação comparativa dos classificadores clássicos, desenvolvidos na plataforma KNIME, permitiu mensurar a capacidade de generalização de diferentes arquiteturas diante do perfil epidemiológico estruturado. Os resultados consolidados de Acurácia (ACC), Precisão (PRE), Sensibilidade (Recall), Especificidade (SPEC), F1-Score e o Coeficiente Kappa estão dispostos na Figura 5, acompanhados de suas respectivas matrizes de confusão.

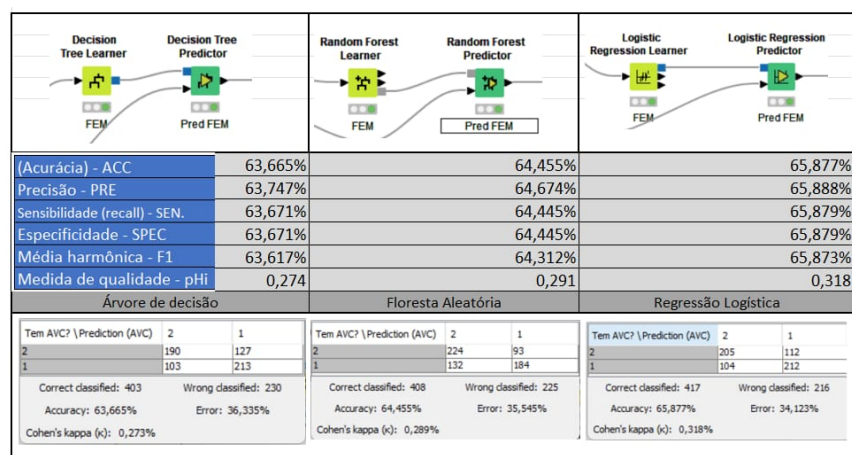


Figura 3. Fluxo de trabalho, métricas de desempenho e matrizes de confusão para os modelos preditivos clássicos.

Com base nos dados consolidados, segue análise do comportamento de cada algoritmo:

- **Árvore de Decisão:** Apresentou o desempenho mais modesto do experimento,

registrando Acurácia de (63,66%) e a menor medida de qualidade $pHi = (0,274)$. Embora tenha demonstrado um equilíbrio razoável ao capturar 213 verdadeiros positivos da classe de risco (Sofreu AVC), gerou 127 falsos positivos, o que reduziu sua Precisão para (63,74%).

- **Floresta Aleatória (Random Forest):** Demonstrou melhor desempenho, elevando a Acurácia para (64,45%) e o índice Cohen's kappa para 0,289. Contudo, o modelo de ensemble exibiu um comportamento mais conservador: aumentou a Especificidade ao classificar corretamente 224 indivíduos saudáveis, mas reduziu a Sensibilidade para (64,44%), falhando no rastreamento primário ao elevar os falsos negativos para 132 casos.
- **Regressão Logística:** Obteve o melhor desempenho global absoluto entre as abordagens clássicas testadas. O modelo linear alcançou a maior Acurácia (65,87%), Precisão (65,88%) e Sensibilidade (65,87%), além do maior Coeficiente Kappa (0,318). A análise de sua matriz de confusão revela um balanço superior na distribuição dos erros, minimizando os falsos negativos para 104 instâncias e isolando com sucesso 212 verdadeiros positivos. Essa superioridade indica que as fronteiras de decisão não lineares conseguiram segregar os perfis de risco de forma mais robusta para o espaço de atributos selecionado.

3.7. Interpretabilidade dos resultados

Com base na análise dos dados, foram extraídas regras de associação de pessoas que tem AVC. Foram extraídas de 18 nós da Árvore de Decisão.

Tabela 3. Principais regras extraídas pela Árvore de Decisão.

Classe	Regra	Cobertura
AVC	Consome sal 4 ou mais vezes por semana, consome proteínas diariamente, não realiza trabalho doméstico pesado, realiza trabalho remunerado pesado e possui hipertensão.	69%
AVC	Consome álcool, possui obesidade, é sedentário e consome verduras apenas um dia por semana.	78,5%
Sem AVC	Consome verduras em cinco dias da semana, é mulher, não realiza trabalho doméstico pesado, não possui hipertensão nem problema cardíaco.	90,4%
Sem AVC	Realiza trabalho doméstico pesado, não consome álcool, não possui hipertensão nem problema cardíaco.	90,2%

A validação médica deste estudo é fundamental. Para os médicos, isso significa o acesso a ferramentas que antecipam o risco de AVC com base em fatores acumulativos, como hipertensão, hábitos alimentares e sedentarismo, possibilitando diagnósticos precoces e intervenções preventivas mais assertivas.

Já para os planos de saúde, a aplicação desses modelos permitiria a criação de sistemas de alerta em grupos de risco, o que reduz custos hospitalares e eleva a qualidade de vida dos pacientes.

4. Conclusões

Este estudo mostrou processo baseado em aprendizado de máquina para descrever o perfil do brasileiro que sofre AVC na faixa etária de 55 a 77 anos. Diante de dados altamente desbalanceados, a abordagem híbrida com SMOTE e a imputação via *MissForest* mostraram-se escolhas metodológicas mais eficazes, garantindo a robustez estatística e a estabilidade local necessárias à validação das análises preditivas do modelo.

No cenário comparativo, a Regressão Logística obteve o melhor desempenho absoluto, alcançando maior acurácia (65,87%) e sensibilidade. Mais do que métricas

isoladas, as regras de associação extraídas comprovaram a tese de que fatores clínicos crônicos, como a hipertensão, atuam em forte sinergia com o consumo excessivo de sal e perfis socioeconômicos para determinar os gatilhos vasculares do AVC.

O trabalho fornece uma estrutura interpretável com aplicabilidade direta na saúde coletiva e na medicina preventiva. Como limitações, apontam-se o caráter transversal dos dados e a impossibilidade de recortes regionais, sugerindo-se para trabalhos futuros o uso de arquiteturas de *Stacking*, redes neurais com KerasMLP.

Referências

- Dritsas, E. and Trigka, M. (2022). Stroke risk prediction with machine learning techniques. *Sensors*, 22(13):4670.
- Oliveira, M. S. and Maciel, R. P. (2021). A importância da prevenção dos fatores de riscos no acidente vascular encefálico: revisão integrativa da literatura. *Research, Society and Development*, 10(16):e353101624294.
- Rodrigues, M. S., Santana, L. F., and Galvão, I. M. (2017). Fatores de risco modificáveis e não modificáveis do AVC isquêmico: uma abordagem descritiva. *Revista de Medicina*, 96(3):187–192.
- Santos, F. d. S., Dias, M. S., Mintem, G. d. C., Oliveira, I. O., and Gigante, D. P. (2020). Processamento de alimentos e fatores de risco cardiometabólicos: revisão sistemática. *Revista de Saúde Pública*, 54:70.