Aplicação de Técnicas de Mineração de Dados e Aprendizado de Máquina na Comparação de Perfis de AVC entre Idosos e Adultos de Meia-Idade: Estudo da PNS 2019

Ligia Ferreira de Carvalho Gonçalves[1], Luis Enrique Zárate Gálvez[1]

[1] Curso Ciência de Dados– Pontifícia Universidade Católica de Minas Gerais (PUC-Minas)
CEP – 30140-100 – Belo Horizonte – MG – Brasil

ligiacarv.goncalves@gmail.com, zarate@pucminas.br

Abstract. *This work aims to explore the use of machine learning techniques to describe the profile of individuals diagnosed with stroke, in order to compare the profile between two distinct age groups: middle-aged adults (40-59) and elderly individuals (60-80). The Decision Tree algorithm was applied to the database provided by the 2019 Brazilian National Health Survey. The conclusions indicate that the rules generated for middle-aged adults are mainly about routine habits, such as work or salt consumption, while for elderly individuals they are more related to intrinsic factors, such as the presence of chronic diseases or gender.*

Resumo. *Esse trabalho tem por objetivo explorar o uso de técnicas de aprendizado de máquina para a descrição do perfil de indivíduos diagnosticados com AVC, a fim de comparar o perfil entre duas faixas etárias distintas: adultos na meia idade(40 – 59) e idosos (60-80). Foi aplicado o algoritmo Árvore de Decisão na base de dados fornecida pela Pesquisa Brasileira Nacional de Saúde de 2019. As conclusões indicam que as regras geradas para adultos na meia-idade são principalmente sobre hábitos rotineiros, como trabalho ou consumo de sal, enquanto para idosos estão mais relacionadas a fatores intrínsecos, como a presença de doenças crônicas ou o gênero.*

1. Introdução

O Acidente Vascular Cerebral (AVC), segundo a Sociedade Brasileira de AVC[1], pode ser caracterizado pelo surgimento de um déficit neurológico súbito causado por um problema nos vasos/artérias ou veias do cérebro. No Brasil, de acordo com o Portal de Transparência do Registro Civil[2], dentre as doenças cardiovasculares, o AVC foi o principal responsável pelos óbitos em 2023. Somente em 2024, no período de janeiro a março, foram registrados cerca de 20 mil óbitos (BRASIL, 2024). Embora a ocorrência do AVC seja maior entre indivíduos com faixas etárias mais avançadas (Rajati et al., 2023), recentemente tem se observado um aumento preocupante no número de casos entre adultos na meia-idade, grupo que até o momento, era acreditado ter baixa predisposição a essa condição. A pesquisa realizada pela *'American Heart Association'*[3], discute o crescimento de casos entre adultos com menos de 49 anos.

[1] https://avc.org.br/
[2] https://transparencia.registrocivil.org.br/painel-registral/especial-covid/
[3] https://www.heart.org/en/

Aprendizado de máquina (AM) é uma área que explora o estudo e desenvolvimento de modelos computacionais que aprendem através de conjuntos de dados, e que tem recebido uma grande atenção na área da medicina (Paixão et al., 2022) sendo utilizado para aprimorar os sistemas de diagnósticos clínicos, e especificamente na cardiologia, tem sido responsável por grande parte das produções científicas no que diz respeito ao auxílio no diagnóstico de doenças cardiovasculares.

Em face dos dados e discussões apresentadas, é essencial a identificação e o entendimento dos fatores que caracterizam o perfil de adultos na meia-idade, de 40 a 59 anos, bem como do perfil de idosos, de 60 a 80 anos, diagnosticados com AVC. Uma maneira de compreender como ambas as populações se assemelham ou diferenciam, é através da aplicação de modelos de aprendizado computacional, que permitem descobrir se as mesmas condições (regras extraídas) que caracterizam a ocorrência do quadro de AVC se aplicam para ambas as populações, e caso sejam diferentes buscar entender onde e quando elas diferem.

O presente trabalho, visa identificar os fatores que melhor caracterizam o perfil de indivíduos que sofreram AVC, utilizando uma metodologia de descoberta de conhecimento e construção de modelos de aprendizado. Além disso, é realizada uma análise comparativa dos resultados dos modelos, visando contribuir para a pesquisa na área de AVC. Como fonte de dados, foi considerado o recente estudo do Instituto Brasileiro de Geografia e Estatística (IBGE), Pesquisa Nacional de Saúde (PNS) 2019[4], pesquisa realizada por meio de questionários no ano de 2019 em todo o território nacional.

2. Trabalhos Relacionados

Dado o impacto que a doença crônica discutida tem nas faixas etárias descritas na seção anterior, foi realizada uma busca em repositórios a partir de termos em inglês ou português, das palavras-chave: AVC, fatores de risco e aprendizado de máquina. Vários estudos investigam os fatores de risco que indicam a vulnerabilidade ao AVC, enquanto outros autores aplicam técnicas de aprendizado de máquina para identificar padrões e prever sua ocorrência.

No trabalho de Yousufuddin e Young (2019), os autores discutem a relação entre envelhecimento e ocorrência de AVC. Eles concluíram que a presença de fatores como hipertensão e diabetes aumenta o risco de AVC à medida que o indivíduo envelhece, enquanto o risco relativo associado a fatores como consumo de cigarro e pressão alta diminui com o tempo. Ademais, em Noche R.B et al., (2020) os autores apontam a lacuna de conhecimento que existe em relação à compreensão dos fatores de risco para adultos na meia-idade, sua pesquisa discute justamente a importância de se entender o fenômeno de recorrência de Acidente Vascular Cerebral nessa faixa etária. Já em um estudo publicado por Howard. G et al., em 2023, é levantada a discussão sobre as diferenças associadas à idade no papel dos fatores de risco para a ocorrência de AVC, os autores trabalham com grupos etários distintos aqueles usados no presente trabalho, além disso, nesse trabalho não foi utilizado nenhum algoritmo de aprendizado de máquina para identificar e comparar os perfis daqueles diagnosticados com AVC em cada grupo.

[4] https://www.pns.icict.fiocruz.br/

Em Dritsas et al. (2022), os autores propuseram o uso modelos de AM para a predição de AVC em uma base adquirida pela plataforma Kaggle, utilizaram os seguintes algoritmos: Naive Bayes, Random Forest, Logistic Regression, K-Nearest Neighbors, Stochastic Gradient Descent, Decision Tree, Multilayer Percepton, Majority Voting e Stacking. Os resultados obtidos com esse trabalho demostraram que a classificação por Stacking obteve um desempenho melhor quando comparada aos outros métodos utilizados, com AUC de 98.9%, F-Measure, Precisão e Recall igual a 97.4% e acurácia de 98%.

Desse modo, diferente dos trabalhos mencionados anteriormente, este artigo busca unificar os pontos abordados em ambos, utilizando fatores de risco e modelos de aprendizado de máquina para descrever o perfil das diferentes faixas etárias. O objetivo é realizar uma análise comparativa dos atributos presentes nas regras mais abrangentes para cada faixa etária.

3. Materiais e Métodos

3.1 Descrição da Base de Dados

Como fonte de dados, foi utilizada para a realização deste trabalho, a Pesquisa Nacional de Saúde (PNS), referente ao ano de 2019, que foi realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em parceria com o Ministério da Saúde. A pesquisa coleta dados sobre a situação geral da saúde e estilo de vida da população brasileira e possui 293726 registros e 1088 atributos. Para este estudo, foi realizado um corte para a doença crônica Acidente Vascular Cerebral (AVC), contendo 88.861 registros referentes a indivíduos que não foram diagnosticadas com AVC e apenas 1.975 para aquelas que de fato apresentaram um diagnóstico positivo para a doença.

3.2 Entendimento do Problema e seleção conceitual de atributos

O entendimento acerca do domínio de problema é uma etapa importante em um processo de descoberta de conhecimento, que visa a construção de modelos de aprendizado mais representativos. O entendimento prévio, permite observar a complexidade de cada problema e a descoberta de conhecimento útil e não obvio, acerca do domínio considerado. Também, devido à alta dimensionalidade da base de dados PNS 2019, a seleção conceitual de atributos é uma estratégia relevante. Na Figura 1, é possível visualizar o modelo conceitual que foi desenvolvido para este estudo. O modelo foi construído utilizando o método CAPTO, recentemente proposto em [Zárate et al., 2023].

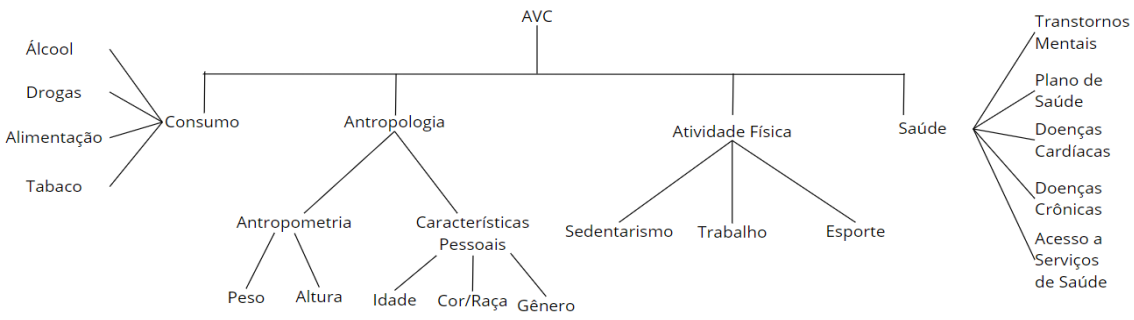

Figura 1 – Modelo Conceitual Verticalizado

O método é uma abordagem que propõe a captura de conhecimento, tanto explícito como tácito, para o entendimento de um domínio do problema prévio a aplicação dos algoritmos de aprendizado de máquina para reduzir a dimensionalidade, selecionando os atributos que melhor representam o domínio do problema. A Tabela 1, mostra os atributos selecionados considerando o modelo conceitual previamente construído.

Tabela 1. Variáveis Selecionadas a Partir do Mapa Conceitual

Variáveis Extraídas da base PNS 2019 a partir do Mapa Conceitual		
Dimensão	**Aspecto**	**Variáveis PNS**
Antropologia	Características Pessoais	C8, C6, C9
	Antropometria	P1a e P4a
Consumo	Alimentação	P9a, P15, P18, P20b, P25a, P26A, P26b
	Álcool	P27, P28a, e P29
	Drogas	Não há informação disponível na base
	Tabagismo	P50, P51, P52, P53, P54, P54a, P54b, P54c, P54d, P54e, P54f, P54g
Saúde	Acesso a serviços de saúde e Plano de saúde	I00102
	Doenças Mentais	Q92 e Q110a
	Doenças Cardíacas	Q2a, Q63a
	Doenças Crônicas	Q30a, Q60, Q63a, Q68
Atividade Física	Sedentarismo	P45a, P45b
	Trabalho	P38, P39, P40, P41, P42, P43, P44, E17, E19
	Esporte	P34, P35, P37, P36

O dicionário de dados pode ser encontrado em https://www.pns.icict.fiocruz.br/

3.3 Pré-Processamento e preparação de dados

Devido à alta taxa de dados ausentes nos atributos selecionados, foi necessário realizar combinações entre atributos a fim de minimizar o impacto causado por eles. Para isto, foram criados 6 novos atributos, são estes:

IMC: Indice de Massa Corporal, a partir dos atributos Peso (P00103) e Altura (P00403). A categorização foi realizada baseado nas faixas definidas pela Organização Mundial da Saúde (OMS). IMC = P00103 /((P00403 / 100) ^ 2).

Se IMC < 18.5: 'Abaixo do peso' *Se 30 <= IMC < 34.99: 'Obesidade grau I'*
Se 18.5 <= IMC < 24.99: 'Peso normal' *Se 35 <= IMC < 39.99: 'Obesidade grau II'*
Se 25 <= IMC < 29.99: 'Sobrepeso' *Se IMC >= 40: 'Obesidade grau III'*

Fumo_Diário: Para a criação do atributo contendo informações sobre o consumo diário de produtos que contém tabaco, foram utilizados os seguintes atributos da PNS: P05402 (Cigarros Industrializados), P05405 (Cigarros de palha ou enrolados a mão), P05408 (Cigarros de cravo ou de Bali), P05411 (Cachimbos), P05414 (Charutos ou cigarrilhas), P05417(Narguilé (sessões)), P05421 (Outro). O procedimento aplicado consistiu na soma de valores resposta entre os atributos (*Fumo_Diario = P05402 + P05405 + P05408 + P05411 + P05414 + P05417+ P05421*). Quando é encontrado um registro vazio (NaN), é considerado como sendo zero (Não-fumante), desse modo foi considerado não fumante o indivíduo cujo resultado da soma de todas as colunas associadas a ele fosse igual a zero. Foi utilizado como base para categorização a padronização proposta pelo Governo do Canadá, intitulada 'Tobacco Use Statistics – Terminology' (https://www.canada.ca/en.html). O procedimento adotado é dado a seguir:

Se Cigarros == 0: 'Não Fumante' *Se 11 <= Cigarros <= 20: 'Fumante Moderado'*
Se 0 < Cigarros <= 10:' Fumante Leve' *Se Cigarros > 20: 'Fumante Pesado'*

Consumo_Alcool_Semanal: A estratégia utilizada para a criação deste atributo combina duas respostas: a frequência semanal de consumo (P02801) e a quantidade de doses consumidas por ocasião (P029). Quando a frequência semanal não está disponível, mas

existe a informação que o indivíduo em questão consome álcool menos de uma vez por mês, a função desenvolvida ajusta a quantidade consumida, dividindo-a por quatro, com o objetivo de estimar uma média semanal de consumo. Utilizou-se para categorizar os padrões definidos pelo National Institute on Alcohol Abuse and Alcoholism (NIAAA). (https://www.niaaa.nih.gov).

Para o gênero feminino	**Para o gênero masculino:**
Se Doses == 0: 'Não bebe'	Se Doses == 0: 'Não bebe'
Se 1 <= Doses <= 7: 'Baixo Consumo'	Se Doses 1 <= Doses <= 14: 'Baixo Consumo'
Se 8 <= Doses <= 14: 'Consumo Moderado'	Se 15 <= Doses <= 28: 'Consumo Moderado'
Se Doses > 14: 'Consumo Alto'	Se Doses > 28: 'Consumo Alto'

Atividade_Física_Moderada_Semanal e Atividade_Física_Intensa_Semanal: Utilizando como referência a definição da OMS, sobre atividades físicas moderadas (*AtividadeM_Semanal = ([P04001*(P04101*60)] + P04102) + ([P042*(P04301*60)] + P04302)*) e intensas (*AtividadeI_Semanal = ([P035*(P03701*60)] + P03702) + ([P03904*(P03905*60)] + P03906)*), foram criados os seguintes atributos referentes aos minutos semanais dedicados a essas atividades. Foram utilizados os seguintes atributos: P035 (Dias por semana que pratica exercício físico ou esporte), P037 (Tempo de duração em horas e minutos – P035), P039 (Dias, horas e Minutos por semana que o entrevistado faz atividades pesadas), P04001 (Dias por semana que realiza trajeto a pé ou de bicicleta), P041 (Tempo de duração em horas e minutos – P04001), P042 (Dias por semana que envolva deslocamento para a realização de atividades habituais) e P043 (Tempo de duração em horas e minutos – P042). Para categorização foi utilizado como referência o guia de atividade publicado pelo Governo Federal em 2021 (https://www.gov.br/saude/pt-br/composicao/saps/ecv/publicacoes/guia-de-atividade-fisica-parapopulacao- brasileira/view).

Para Atividade Física Moderada	***Para Atividade Física Intensa***
Se 0 <= Minutos <= 149: 'Nível 1'	*Se 0 <= Minutos <= 74: 'Nível 1'*
Se 150 <= Minutos <= 299: 'Nível 2''	*Se 75 <= Minutos <= 149: 'Nível 2'*
Se 300 <= Minutos <= 449: 'Nível 3'	*Se 150 <= Minutos <= 224: 'Nível 3'*
Se 450 <= Minutos <= 599: 'Nível 4'	*Se 225 <= Minutos <= 299: 'Nível 4'*
Se Minutos => 600: 'Nível 5'	*Se Minutos => 300: 'Nível 5'*

Jornada_de_Trabalho: É referente a quantidade de horas trabalhadas semanalmente. Foi resultado da soma dos valores entre os atributos E017 (Total de horas trabalhadas por semana no trabalho principal) e E019 (Total de horas trabalhadas por semana em outros trabalhos) caso alguma instância estivesse vazia (NaN), esse valor foi considerado como zero, desse modo o atributo final não foi afetado e não houve alteração nos dados utilizados. Em outras palavras, se todos os atributos utilizados fossem nulos, o atributo final teria um saldo igual a zero, significando que essa pessoa não está empregada. Para sua categorização foi utilizado como base as recomendações da Organização Mundial de Saúde (OMS).

Se Horas == 0: 'Não está empregada'	Se 41 <= Horas <= 54: 'Jornada de Trabalho Excessiva'
Se Horas <= 40: 'Jornada de Trabalho Normal'	Se Horas >= 55: 'Jornada de Trabalho Excessivo de Alto Risco'

Classificação_Alimentação (*score*): Pesos foram atribuídos a cada alimento do respondente e foram baseados na sua importância para a construção de uma dieta que previna a ocorrência de doenças cardiovasculares. Desse modo o consumo regular de frutas (P018) e vegetais (P00901) foi associado à redução do risco de AVC devido ao seu alto conteúdo de antioxidantes, fibras e micronutrientes (Dauchet et al., 2005). O consumo regular de peixe (P015) também é responsável pela saúde cardiovascular devido aos efeitos anti-inflamatórios por ser um alimento rico em ácidos graxos de acordo com (Mozaffarian e Rimm, 2006). Em contrapartida, o consumo excessivo de doces (P02501),

fast food (P02602) e refrigerantes (P02002) está relacionado ao aumento do risco de obesidade, hipertensão e doenças cardiovasculares (Malik et al., 2010). Para finalizar o tratamento desses atributos, foi criado o atributo score, que corresponde à soma dos produtos entre a frequência de consumo de cada alimento e seu respectivo peso. Para a categorização desse atributo foram definidos valores de corte no atributo score para separar as categorias relacionadas ao tipo de alimentação.

Se Score <= 0: 'Alimentação Não Saudável'

Se 0 < Score <= 5: 'Alimentação Pouco Saudável'

Se Score > 5: 'Alimentação Saudável'

Em todos os atributos criados foi aplicado o método de codificação OrdinalEncoder, normalmente utilizado para codificar atributos categóricos em que seus elementos possuam uma hierarquia entre si . O restante dos atributos, já estavam codificados pela forma que as respostas estavam dispostas no questionário da PNS 2019, mantendo-se a codificação original. A única alteração realizada foi na codificação do atributo 'Consumo de Sal'. Foi feita uma inversão na ordem, onde o valor '5', que originalmente correspondia ao nível mais baixo, passou a representar o consumo mais alto, enquanto o valor '1', que antes correspondia ao nível mais alto, agora representa o consumo mais baixo.

Após as etapas descritas anteriormente, a base de dados resultante foi dividida em duas novas bases de dados referentes às faixas etárias consideradas neste estudo: a BaseDados-1, associada à faixa etária de 40 a 59 anos, e a BaseDados-2 correspondente à faixa etária de 60 até 80 anos. Foi aplicada uma análise de correlação e entropia para analisar respectivamente a existência de atributos com pouca variância e informação. Em outras palavras, identificar atributos que apresentem pouca informação e atributos altamente relacionados com o atributo alvo, para evitar classificação direta de um atributo dominante. Após esse procedimento, não foi possível eliminar atributos tendo como bases essas análises, pois os valores de entropia estavam bem próximos dificultando estabelecer um valor de corte e a análise de correlação para ambas as bases não mostrou a existência de nenhum atributo com alta correlação com o atributo classe.

Finalmente foi verificado se existiam inconsistências após a codificação, a fim de identificar instâncias que possuam os valores de todos os atributos idênticos, mas pertencentes a classes distintas. Os registros que atendiam essa condição foram eliminados de ambas as bases. Na Tabela 2, estão descritos os atributos presentes nas bases de dados após o pré-processamento e preparação de dados.

Tabela 2. Atributos presentes na base após as etapas de pré-processamento

Atributo	Descrição do atributo
Gênero	Referente ao gênero do entrevistado.
Cor/Raça	Referente a cor/raça do entrevistado.
Plano de Saúde	Se o entrevistado possui plano de saúde.
Quantidade de Trabalhos	Quantos empregos o entrevistado possui.
Consumo_Sal	Percepção do entrevistado sobre seu consumo de sal.
Trabalho Doméstico Pesado	Se o entrevistado nas suas atividades domésticas, faz faxina pesada (não considerar atividade doméstica remunerada).
Hipertensão	Se o entrevistado foi diagnosticado com hipertensão.
Diabetes	Se o entrevistado foi diagnosticado com diabetes.
Colesterol_Alto	Se o entrevistado foi diagnosticado com colesterol alto.
Doença_Cardíaca	Se o entrevistado foi diagnosticado com alguma doença cardíaca.
AVC	Coluna Alvo: Se o entrevistado foi diagnosticado com AVC.
Depressão	Se o entrevistado foi diagnosticado com hipertensão.
Doenças Mentais	Se o entrevistado foi diagnosticado com algum transtorno mental.

Categoria_IMC	Índice de Massa Corporal do entrevistado.
Jornada de Trabalho	Nível de horas da jornada de trabalho do entrevistado.
Categoria_Fumantes	Nível de consumo de tabaco pelo entrevistado.
Categoria_Alcool_Semanal	Consumo de álcool semanal do entrevistado.
classificacao_alimentacao(score)	Tipo de alimentação do entrevistado.
Categoria_AtividadeM	Nível de atividade física moderada praticada pelo entrevistado.
Categoria_AtividadeI	Nível de atividade física intensa praticada pelo entrevistado.

4. Resultados

4.1 Treinamento e Validação dos Modelos

Para a construção dos modelos de aprendizado, foi utilizado o ambiente *Knime Analytics (*https://www.knime.com/), que é uma plataforma gratuita e de código aberto, que integra as principais técnicas de integração e análise de dados para projetos em ciência de dados. Vale ressaltar que as instâncias do atributo alvo que estão como 1:AVC correspondem aos indivíduos com diagnóstico positivo de AVC e os que estão como valor 2:Não-AVC correspondem aos diagnósticos negativos. A BaseDados-1 (adultos) possui 20894 instâncias e 20 atributos, com 20482 registros para a classe negativa e apenas 412 para a positiva. Já a BaseDados-2 (idosos) é composta por 12127 registros e 20 atributos, sendo que 11561 registros são referentes à classe negativa e 566 associados à positiva.

Ambas as bases tiveram suas instâncias separadas em 70% para o conjunto de treino e 30% para o conjunto de teste (*hold-out*). Para o processo de treinamento e teste, foi aplicado o método de balanceamento *RandomUnderSampling* (RUS) para ambos os conjuntos de treino e teste, uma vez que a alta taxa de desbalanceamento das bases resultava em experimentos injustos com a classe minoritária, pois mascaravam o cálculo das medidas de avaliação, especialmente a precisão e o recall. A fim de exemplificar, utilizando o balanceamento apenas no conjunto de teste, o modelo acertava todas as instâncias para a classe negativa e errava tudo para a negativa independentemente do método de balanceamento utilizado. Foram testados diversos métodos para tentar contornar essa situação, no entanto, não houve melhora.

Tabela 3. Divisão para treinamento/validação e teste para a BaseDados-1

Classe	Original	Treino RUS	Teste RUS
1	412	299	113
2	20482	299	113
Total	20894	598	226

Tabela 4. Divisão para treinamento/validação e teste para a BaseDados-2

Classe	Original	Treino RUS	Teste RUS
1	566	390	176
2	11561	390	176
Total	12127	780	352

Para a construção dos modelos, foi utilizado o algoritmo caixa-branca, Árvore de Decisão, pela sua capacidade interpretativa. Os parâmetros utilizados pela árvore gerada para a BaseDados-1 foram o Critério = gini e número de registros mínimos por nó = 16 e pela árvore gerada para a BaseDados-2 foram o Critério = gini e número de registros mínimos por nó = 22. É possível observar os resultados obtidos nessa etapa na Tabelas 5 que demonstra que os modelos tiveram uma capacidade similar de aprendizado.

Tabela 5. Métricas de Avaliação Referentes aos Conjuntos de Treinamento

Base	Classe	Precisão	Recall	F1-Measure	Acurácia
BaseDados-1	1: AVC	0.657	0.629	0.643	0.651
	2:NÃO AVC	0.644	0.672	0.658	
BaseDados-2	1: AVC	0.633	0.574	0.602	0.621

	2: NÃO AVC	0.610	0.667	0.637

Na Figura 2, é possível observar o fluxograma que foi construído para a extração de conhecimento nas duas bases usando a plataforma Knime.

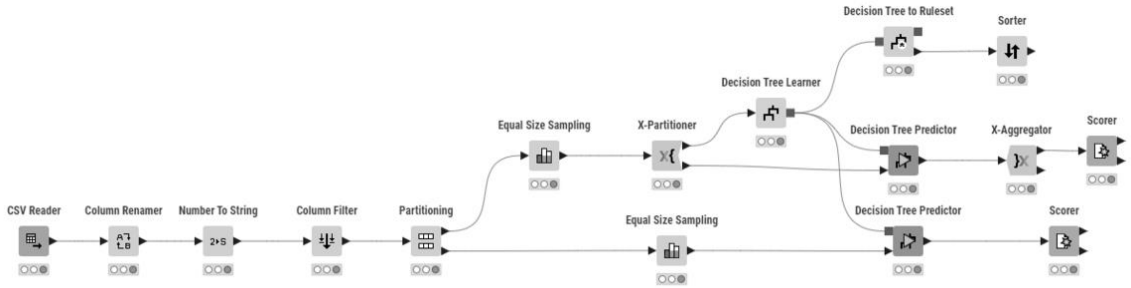

Figura 2. Fluxograma Construído no KNIME

4.2 Análise dos Resultados

Analisando os resultados das métricas para o conjunto de teste, para a Precisão (VP/(VP+FP)) e F-Measure (média harmônica entre a precisão e o recall), é perceptível que o modelo teve uma capacidade adequada para classificar corretamente suas predições. A BaseDados-1 (adultos) obteve uma média para as métricas em torno de 70% em ambas as classes, já para a BaseDados-2 (idosos) o modelo atingiu métricas em torno de 60%. Na Tabela 6, estão dispostas as métricas para cada classe para a BaseDados-1 e BaseDados-2.

Tabela 6. Métricas de Avaliação Referentes aos Conjuntos de Teste.

Base de Dados	Classe	Precisão	Recall	F1-Measure	Acurácia
BaseDados-1	1: AVC	0.678	0.726	0.701	0.690
	2:NÃO AVC	0.705	0.655	0.679	
BaseDados-2	1: AVC	0.637	0.608	0.622	0.631
	2: NÃO AVC	0.625	0.653	0.639	

Vale ressaltar que o recall (VP/(VP+FN)), métrica responsável por verificar a capacidade que o modelo tem em classificar corretamente determinada condição, para casos em que a condição de fato acontece, ou seja, 1:AVC e foram classificados como 1 (um), e os que pertencem à classe 2:Não-AVC e foram classificados como 2, é notório a partir das matrizes de confusão, em que a coluna representa a classe verdadeira e o cabeçalho a classe predita, dispostas abaixo que os modelos acertaram uma quantidade aceitável, para o total de registros de cada classe em ambas as bases de dados consideradas.

Matriz de Confusão para o Teste BaseDados-1

	1:AVC	**2:NÃO-AVC**
1	82	31
2	39	74

Matriz de Confusão para o Teste BaseDados-2

	1:AVC	**2:NÃO-AVC**
1	107	69
2	61	115

4.3 Comparação das Regras Geradas

Para análise comparativa das regras geradas pelas árvores de decisão, correspondentes às duas populações, foi utilizado o módulo *DecisionTree to Ruleset* que fornece um dataset

com as regras geradas pela árvore bem como as métricas *Record count*, que representa o número total de instâncias que correspondem a uma determinada regra derivada da árvore de decisão, e o *Number of correct* que indica o número de registros que são classificados corretamente pela regra. A fim de entender o comportamento de cada faixa etária em relação aos fatores mais relevantes, para cada base de dados foram extraídas as cinco principais regras, aquelas que classificaram os maiores números de instâncias.

Tabela 7. Regras com as Maiores Coberturas Geradas pelas Respectivas Bases

Principais Regras Geradas e Extraídas	
Para a BaseDados-1 (40 – 59 anos)	**Para a BaseDados-2 (60 – 80 anos)**
Se Hipertensão = 2 E Doença Cardíaca = 2 E Quantidade Trabalhos = 1 Categoria_IMC = 2 ENTÃO AVC = 2	Se Doenças cardíacas = 2 E Hipertensão= 2 ENTÃO AVC = 2
Se Hipertensão = 1 E Doença Cardíaca = 1 ENTÃO AVC = 1	Se Doenças cardíacas = 2 E Hipertensão = 1 e Depressão = 1 ENTÃO AVC = 1
Se Hipertensão = 2 E Doença Cardíaca = 2 E Quantidade Trabalhos = 0 E Gênero = 2 E Colesterol Alto = 2 ENTÃO AVC = 2	Se Doenças cardíacas = 2 E Hipertensão = 1 e Depressão = 2 e Etnia = 0 e Quantidade de trabalho = 1 ENTÃO AVC = 2
Se Hipertensão = 1 E Doença Cardíaca = 2 E Categoria_Fumante = 0 E Categoria_IMC = 2 E Consumo de Sal = 2 ENTÃO AVC = 1	Se Doenças cardíacas = 2 E Hipertensão = 1 e Depressão = 2 e Etnia = 1 ENTÃO AVC = 2
Se Hipertensão = 1 E Doença Cardíaca = 2 E Categoria_Fumante = 0 E Categoria_IMC = 2 E Consumo de Sal = 3 ENTÃO AVC = 2	Se Doenças cardíacas = 2 E Hipertensão = 1 e Depressão = 2 e Etnia = 0 e Quantidade de trabalho = 0 e Gênero = 1 ENTÃO AVC = 1

Embora não tenha sido possível identificar a ocorrência de regras iguais para classificar as instâncias presentes em ambas as populações, ao analisar as regras geradas é notória uma similaridade nos atributos escolhidos para ambas as bases, como a presença ou ausência de doenças cardíacas.

Além disso, as regras referentes à população idosa diagnosticada com AVC têm frequentemente em sua composição atributos condicionados a características intrínsecas do indivíduo, como etnia ou gênero. Em contrapartida, os atributos presentes nas regras com maiores coberturas, para o grupo dos adultos na meia-idade, estão associados a hábitos ou rotinas, ou seja, ações que podem ser alteradas com mais facilidade. Sendo assim, é possível concluir que os fatores que caracterizam o perfil da população idosa são mais concretos, ou seja, a chance de reverter ou prevenir a ocorrência do AVC para esse grupo é mais complicada e envolve um estudo mais aprofundado de caso a caso. Já para a faixa etária mais jovem, o acontecimento do Acidente Vascular Cerebral é um cenário mais fácil de ser controlado, uma vez que os fatores que melhor descreveram o perfil desse grupo são altamente suscetíveis a mudanças.

5. Conclusões

Apesar de apresentar resultados satisfatórios, este trabalho possui algumas limitações a serem discutidas. O desbalanceamento da base, como discutido anteriormente, impossibilitou o uso da prática comum de balanceamento apenas do conjunto de treino. Além disso, foi aplicado apenas um algoritmo, ou seja, não há comparativos sobre a eficácia que diferentes modelos de aprendizado de máquina teriam sobre o problema em questão. Sendo assim, para trabalhos futuros é interessante a experimentação de outros algoritmos caixa preta, conhecidos por sua melhor capacidade de classificação e resultados. Para essas análises, é importante acrescentar uma secção que discuta a extração de regras desses algoritmos e a sua interpretabilidade, uma vez que esses modelos não deixam seu processo de tomada de decisão explícito.

Finalizado o estudo, percebeu-se que, embora as faixas etárias não partilhem de regras semelhantes para descrever o perfil dos indivíduos diagnosticados com Acidente Vascular Cerebral, elas não são completamente excludentes. O estudo sobre o grupo entre 40 e 59 anos pode ter ação prescritiva e preditiva, utilizando o perfil de comportamento para identificar indivíduos com alta chance de serem vítimas de AVC ao entrarem na terceira idade e orientando-os sobre como alterar hábitos para evitar isso. Para finalizar, ficou evidente a necessidade de pesquisas nesse campo no Brasil, bem como a coleta de dados mais representativos para o problema discutido.

Agradecimentos

Os autores agradecem o apoio recebido do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Processo Nº 303133/2021-0, e do Fundo de Incentivo à Pesquisa (FIP) da PUC Minas, Processo Nº 30914-1S/2024.

Referências

Dauchet, L., Amouyel, P., Hercberg, S., & Dallongeville, J. (2005). Fruit and vegetable consumption and risk of coronary heart disease: A meta-analysis of cohort studies. *The Journal of Nutrition, 135*(10), 2589-2593.

Dritsas, E., & Trigka, M. (2022). Stroke Risk Prediction with Machine Learning Techniques. *Sensors (Basel)*, 22(13), 4670. doi: 10.3390/s22134670. PMID: 35808172; PMCID: PMC9268898.

Howard, George et al. "Age-Related Differences in the Role of Risk Factors for Ischemic Stroke." *Neurology* vol. 100,14 (2023): e1444-e1453. doi:10.1212/WNL.0000000000206837

Malik, V. S., Schulze, M. B., & Hu, F. B. (2010). Intake of sugar-sweetened beverages and weight gain: a systematic review. *The American Journal of Clinical Nutrition, 84*(2), 274-288.

Mozaffarian, D., & Rimm, E. B. (2006). Fish intake, contaminants, and human health: evaluating the risks and the benefits. *JAMA, 296*(15), 1885-1899.

Noche, Rommell B. et al. "Abstract 156: Recurrent Stroke in Middle-Aged Lacunar Stroke Survivors: Understanding Risk Factors and Vulnerability in an Important Target Population." *Stroke* (2020): n. pag.

Paixão, Gabriela Miana de Mattos et al. "Machine Learning in Medicine: Review and Applicability." "Machine Learning na Medicina: Revisão e Aplicabilidade." *Arquivos brasileiros de cardiologia* vol. 118,1 (2022): 95-102. doi:10.36660/abc.20200596

Rajati, F., Rajati, M., Rasulehvandi, R., & Kazeminia, M. (2023). Prevalence of stroke in the elderly: A systematic review and meta-analysis. *Interdisciplinary Neurosurgery, 32*, 101746, ISSN 2214-7519. doi: 10.1016/j.inat.2023.101746.

Yousufuddin, M., & Young, N. (2019). Aging and ischemic stroke. *Aging (Albany NY), 11*(9), 2542-2544. doi: 10.18632/aging.101931. PMID: 31043575; PMCID: PMC6535078.

Zárate, L., Petrocchi, B., Dias, M. C., Felix, C., & Gomes, M. (2023). CAPTO - A method for understanding problem domains for data science projects: CAPTO - Um método para entendimento de domínio de problema para projetos em ciência de dados. *Concilium, 23*, 922-941. doi: 10.53660/CLM-1815-23M33.