# Uso de Técnicas de Mineração de Dados na Prevenção de Acidente Vascular Cerebral

Luiza Gabriela[1], Aline Layze[1], André Pantoja[1], Harold de Mello Jr.[2],

Karla Figueiredo[2], Antonio Pereira[3]

[1] Universidade Federal do Rio Grande do Norte, Natal RN, Brasil
[2] Universidade do Estado do Rio de Janeiro, Rio de Janeiro RJ, Brasil
[3] Instituto de Tecnologia, Universidade Federal do Pará, 66075-110 Belém (PA), Brasil
luizafonseca94@gmail.com

**Abstract.** The present work compares the results of data mining processes on data from questionnaires answered by the public during campaigns to raise awareness about stroke held in 2015 and 2016. Stroke is one of the leading causes of death worldwide and the use of data mining techniques can help uncover associated risk factors and help prevent the occurrence of more cases. Four traditional classification algorithms were used in the database, which contained information of 592 individuals on the following parameters: socioeconomic, anthropometric, medical history, and knowledge of risks associated with stroke. The results show that classification improves when the participants' knowledge of stroke, including risks, physiopathology, signs and symptoms, etc. are included in the database. The random forest and C4.5 algorithms provided the best classification outcomes about stroke risk with perfect 100% scores, followed by neural network and part with 95% and 97.66%, respectively.

**Keywords:** Cerebral Vascular Accident (CVA), Risk Factors, Health Education, Data Mining, Classification Algorithms.

## 1 Introdução

As doenças cardiovasculares são uma das principais causas de morte em todo o mundo. A Organização Mundial de Saúde estima que 17,5 milhões de pessoas morreram de doenças cardiovasculares em 2015, representando 31% de todas as mortes globais. Destas mortes, cerca de 6,7 milhões ocorreram devido ao Acidente Vascular Cerebral (AVC) [1]. O AVC é causado pela interrupção do fornecimento de sangue para o cérebro, devido ao rompimento de um vaso sanguíneo ou, mais frequentemente, à oclusão vascular por um coágulo.

A manifestação clínica do AVC depende da região cerebral afetada, mas os sinais e sintomas mais frequentes são: diminuição de força e/ou perda de sensibilidade contralateral, afasia, apraxia, disartria, hemianopsia parcial ou completa, alteração de consciência e confusão mental, diplopia, vertigem, nistagmo e ataxia [2]. Apesar das campanhas educativas, o conhecimento do público leigo sobre os sinais, sintomas e fatores de

risco do AVC ainda é inadequado ou insuficiente. Pesquisas anteriores com base em populações saudáveis revelam, por um exemplo, um desconhecimento generalizado dos principais sinais de alerta na detecção precoce do AVC [3] [4].

Em comemoração ao Dia Mundial do AVC (29 de outubro), são organizadas ações de esclarecimento e educação sobre AVC dirigidas para o público leigo em todo o mundo. No Brasil, as campanhas do Dia Mundial do AVC são organizadas pela Rede Brasil AVC e têm sido um enorme sucesso. As ações são realizadas em espaços públicos e envolvem atividades de avaliação dos fatores de risco da população.

No presente trabalho, utilizaram-se técnicas computacionais de mineração de dados em informações coletadas durante a Campanha do Dia Mundial do AVC dos anos de 2015 e 2016 no Estado do Rio Grande do Norte. O objetivo é automatizar a identificação de indivíduos propensos a sofrer um AVC a partir da base de dados coletada. Os resultados do presente estudo podem auxiliar no entendimento dos fatores de risco associados com o AVC e contribuir para a recomendação de medidas preventivas, reduzindo as despesas com saúde pública. Uma busca na literatura revelou a existência de dois trabalhos de automação de diagnóstico de AVC [5] e [6]. No entanto, estes trabalhos utilizam dados de imagens de tomografia e aplicam métodos distintos aos empregados no presente artigo.

Este artigo está organizado da seguinte forma: a seção 2 discute alguns aspectos gerais relacionados à mineração de dados. A seção 3 apresenta a metodologia adotada. A seção 4 apresenta e discute os resultados encontrados. Finalmente, a seção 5 articula algumas conclusões gerais.

## 2 Fundamentos de Mineração de Dados

### 2.1 Processo de Descoberta de Conhecimento em Base de Dados

A mineração de dados é vista como uma das principais etapas do processo interativo de descoberta de informação em base de dados (KDD – *Knowledge Discovery in Databases*). As técnicas utilizadas nesta metodologia têm auxiliado pesquisadores e analistas em processos de tomada de decisão com base na identificação de correlações, padrões e tendências em conjunto de dados.

Os processos de KDD [7] incluem as seguintes fases: i) pré-processamento, na qual é realizada a limpeza e integração de múltiplas fontes de dados em um única fonte; ii) seleção de dados, na qual os dados mais relevantes são escolhidos e, em seguida, transformados em um formato adequado para a mineração; iii) mineração de dados, na qual diversas técnicas, incluindo as ferramentas de inteligência computacional, são aplicadas; iv) avaliação dos padrões identificados; e v) representação visual dos resultados da análise.

Os tipos de padrões que podem ser descobertos dependem das tarefas de mineração de dados empregadas. De um modo geral, existem dois tipos de tarefas de mineração de dados: descritivas e preditivas [8]. Na categoria preditiva, as tarefas envolvem: classificação, predição, agrupamento, detecção de *outliers* e anomalias.

Uma vez que os usuários não possuem um conhecimento *a priori* do tipo de padrão que a base de dados apresenta, é necessário dispor de sistemas versáteis que permitam

analisar dados em diferentes níveis de abstração, de modo a comportar a interatividade do processo KDD.

No presente trabalho, foram utilizados algoritmos de classificação e de análise de relevância para selecionar atributos com a finalidade de melhorar o desempenho dos algoritmos na identificação de indivíduos mais suscetíveis a ter um AVC.

## 2.2 Algoritmos de Classificação

Os algoritmos de classificação organizam os dados em classes pré-definidas. As técnicas de classificação geralmente utilizam um conjunto de treinamento com dados previamente classificados. A partir do aprendizado das características deste conjunto de treinamento, o algoritmo de classificação constrói um modelo e o utiliza para predizer a classe a que um novo registro pertence.

A avaliação da performance do classificador é baseada na precisão da predição, isto é, na proporção de erros obtidos sobre um conjunto completo de amostras. Frequentemente, divide-se o conjunto disponível de dados em três partes: treinamento, validação e teste, determinando-se, então, três parâmetros de erro para aferir o desempenho do classificador.

O desempenho da classificação é apresentado na forma de uma matriz de confusão ou através da curva de característica de operação do receptor (ROC – *Receiver Operating Characteristic*) [9]. A matriz de confusão oferece uma medida da eficácia do modelo de classificação, mostrando o número de classificações corretas pelo número de classificações indicadas pelo modelo, nos conjuntos de treinamento, validação e teste. Na curva ROC é possível inspecionar a taxa de positivos verdadeiros versus a taxa de falsos positivos e o valor correspondente da área sob a curva. Quanto mais próxima essa curva estiver do canto superior esquerdo, melhor será o desempenho do classificador, isto é, maior será a taxa de positivos verdadeiros e menor a de falsos positivos. Curvas próximas à diagonal do gráfico são característica de classificadores que tendem a fazer previsões aleatórias.

A seguir, apresentamos alguns classificadores clássicos.

**Árvores de decisão.** Também conhecidas como árvores de classificação e regressão generalizadas, são uma forma de representação de um conjunto de regras que seguem uma hierarquia de classes ou valores, expressando uma lógica condicional simples. Essas regras podem ser expressas em linguagem natural, facilitando o entendimento dos usuários.

Um dos principais algoritmos para indução de árvores de decisão é o C4.5 [10]. Inspirado no algoritmo pioneiro ID3 [11], o C4.5 é baseado em busca gulosa e, através de uma abordagem de partição recursiva, produz árvores de decisão de um conjunto de dados de treinamento e a cada nó o algoritmo escolhe um atributo que melhor subdivide o conjunto das amostras em subconjuntos homogêneos e caracterizados por sua classe. Na escolha do atributo para subdivisão, o C4.5 utiliza a medida de razão de ganho que permite a geração de árvores mais precisas e menos complexas que o ID3. Adicionalmente, o C4.5 supera limitações do ID3 ao tratar atributos contínuos e ignora valores

desconhecidos, além de utilizar método pós-poda de árvore que melhora a capacidade de generalização do modelo construído.

***Random forest.*** *Random forest* é um termo geral que abrange classificadores baseados em comitês de árvores de decisão. O método básico [12] segue os passos do algoritmo *bagging* [13] para construir os classificadores base, isto é, as árvores de decisão. Além da amostragem *bootstrap*, na qual amostras aleatórias do conjunto de treinamento são geradas com reposição a partir do conjunto original, e da votação por maioria, utilizadas no *bagging*, o *random forest* aplica ainda métodos de subespaços aleatórios [14] na construção do conjunto de treinamento, o que promove diversidade nos classificadores base, podendo melhorar o desempenho de generalização do comitê.

**PART.** O algoritmo PART produz um conjunto de regras do tipo SE-ENTÃO a partir de uma árvore de decisão construída através do algoritmo C4.5. Do mesmo modo que o J48, o PART também utiliza a técnica dividir-para-conquistar. O algoritmo constrói uma árvore de decisão C4.5 parcial a cada iteração, colocando a melhor folha (classe) dentro de uma regra [15]. As regras são induzidas a partir de uma árvore e posteriormente são refinadas. Para cada regra criada é estimada a sua cobertura das instâncias da base de dados. Isso acontece repetidas vezes até que todas as instâncias estejam cobertas. As regras com coberturas mais altas são mantidas e apresentadas para o usuário e as demais são descartadas [16]. A diferença de um algoritmo de indução de regras de decisão para um algoritmo de árvore de decisão, reside no fato de que as regras de decisão são induzidas para cobrir um conjunto de exemplos e, dessa maneira, pode haver sobreposição das regras construídas no espaço de descrição dos exemplos [17]. Assim, os conceitos aprendidos com esses diferentes algoritmos podem ser bastante distintos.

**Redes Neurais Artificiais.** Redes neurais são construções matemáticas relativamente simples, que foram inspiradas em modelos biológicos da conectividade sináptica do sistema nervoso central [18]. A representação de uma rede neural envolve unidades altamente interconectadas com capacidade de adquirir, armazenar e utilizar informação. Em tarefas de classificação, os parâmetros livres ou pesos sinápticos de uma rede neural são adaptados através de um processo contínuo de interação com o ambiente. A rede é treinada através do fornecimento dos valores de entrada e dos respectivos valores desejados de saída, no treinamento supervisionado.

As redes neurais podem ser usadas para criar um mapeamento não linear das características dos dados, de modo a permitir a obtenção de novas características de dimensionalidade reduzida e desempenho aprimorado.

Nas redes neurais perceptron de multicamadas (MLP), os neurônios são arranjados em camadas, os sinais propagam-se da entrada para a camada de saída, passando por pelo menos uma camada intermediária e um algoritmo de retropropagação de erros permite o ajuste automático dos pesos. Parâmetros da função de ativação de cada neurônio e o número de neurônios na camada intermediária também podem ser ajustados durante a fase de treinamento. Redes MLP com apenas uma camada intermediária apresentam capacidade de aproximação universal [19].

# 3 Metodologia

## 3.1 Base de dados

A base de dados original continha registros de 592 indivíduos obtidos de um questionário de avaliação aplicado em participantes dos eventos do Dia Nacional de combate ao AVC realizados na Região Metropolitana de Natal (RN) nos anos de 2015 e 2016.

Inicialmente foram levantados os seguintes dados de cada indivíduo: escolaridade, estado civil, sexo observado pelo entrevistador, cor da pele, idade em anos completos e os demais itens relacionados na tabela 1. Em seguida, foram realizadas perguntas para determinar o nível de conhecimento (NC) do indivíduo sobre o AVC, definindo-se quatro novos atributos e valores de acordo com a quantidade de respostas corretas:

- NC-FISO (Nível de conhecimento sobre a fisiopatologia): 1 – Suficiente, para 5 ou mais acertos; 2 – Regular, para 2 a 4 acertos; 3 – Insuficiente, para 1 ou nenhum acerto.
- NC-FR (Nível de conhecimento sobre fatores de risco): 1 – Suficiente, para 5 ou mais acertos; 2 – Regular, para 2 a 4 acertos; 3 – Insuficiente, para 1 ou nenhum acerto.
- NC-SSS (Nível de conhecimento sobre sinais, sintomas e sequelas): 1 – Suficiente, para 3 ou mais acertos; 2 – Regular, para 2 acertos; 3 – Insuficiente, para 1 ou nenhum acerto.
- NC-CON (Nível de conhecimento sobre conduta imediata): 1 – Suficiente, para os que responderam com a opção de hospitalização; 2 – Insuficiente, para quaisquer outras respostas diferentes de hospitalização.

A tabela 2 apresenta as características dos dados obtidos nesta parte da entrevista.

Os 592 indivíduos foram categorizados em três classes de Risco Cardiovascular (RCV): baixo, com 127 indivíduos; intermediário, com 272 indivíduos; e alto, com 193 indivíduos. A Secretaria de Atenção à Saúde do Ministério da Saúde [20] estabelece que indivíduos com RCV baixo têm menos que 10% de chance de ir a óbito por causa do AVC nos próximos 10 anos. Por outro lado, indivíduos com RCV alto têm 20% de chance de óbito por AVC no mesmo período.

**Tabela 1.** Atributos da base de dados

| Num. | Atributo | Descrição | Conteúdo |
|---|---|---|---|
| 1 | ID | Código identificador | Num. Inteiro |
| 2 | ANO | Ano de realização da entrevista | Num. Inteiro |
| 3 | AGE | Idade | Num. Inteiro |
| 4 | FAX_ET | Faixa etária | Num. Inteiro: 0 a 6 |
| 5 | GENDER | Sexo observado do entrevistado | Num. Inteiro: 1 ou 2 |
| 6 | CITY | Capital do Estado ou interior | Num. Inteiro: 1 ou 2 |
| 7 | SABE | Declara saber o que é AVC | Num. Inteiro: 1 ou 2 |
| 8 | EMER | Conhece o número de emergência (192)? | Num. Inteiro: 1 ou 2 |
| 9 | ETNIA | Etnia autodeclarada | Num. Inteiro: 1 a 7 |
| 10 | SCHOL | Escolaridade autodeclarada | Num. Inteiro: 1 a 7 |

| | | | |
|---|---|---|---|
| 11 | HISTORY | Histórico familiar de AVC | Num. Inteiro: 1 ou 2 |
| 12 | P_NEAR | Algum conhecido (socialmente) que teve AVC? | Num. Inteiro: 1 ou 2 |
| 13 | WEIGHT | Peso autodeclarado | Num. Real |
| 14 | HEIGHT | Altura autodeclarada | Num. Real |
| 15 | FUMATE | Fuma atualmente ou já fumou no passado? | Num. Inteiro: 1 ou 2 |
| 16 | DRINK | Bebe mais de uma dose de bebida alcoólica/dia? | Num. Inteiro: 1 ou 2 |
| 17 | DIET | Come pelo menos 6 porções de frutas e/ou vegetais/dia? | Num. Inteiro: 1 ou 2 |
| 18 | PHY_ACT | Pratica pelo menos 2:30 h de ativ. física/semana? | Num. Inteiro: 1 ou 2 |
| 19 | ESTRESS | Teve grande estresse mental/emocial no último ano? | Num. Inteiro: 1 ou 2 |
| 20 | FAMILY | Seus pais tiveram AVC ou infarto antes dos 65 anos? | Num. Inteiro: 1 ou 2 |
| 21 | PAS | Pressão arterial sistólica | Num. Inteiro |
| 22 | PAD | Pressão arterial diastólica | Num. Inteiro |
| 23 | HIP | Utiliza medicamento para reduzir a pressão sanguínea? | Num. Inteiro: 1 ou 2 |
| 24 | DM | Já foi informado por algum médico que tem diabetes? (Sim <12 meses, Sim >12meses, Não) | Num. Inteiro:1, 2 ou 3 |
| 25 | VENOUS_D | Já foi informado por algum médico que tem doença cardíaca ou doença arterial periférica? (Sim <12 meses, Sim >12meses, Não) | Num. Inteiro:1, 2 ou 3 |
| 26 | COGNITIVE_D | Já foi informado por algum médico que você tem distúrbio cognitivo ou demência? (Sim <12 meses, Sim >12meses, Não) | Num. Inteiro:1, 2 ou 3 |
| 27 | MEMORY | Você ou alguém próximo acredita que você tem uma memória prejudicada? | Num. Inteiro: 1 ou 2 |
| 28 | TCE | Já foi informado por algum médico que você teve traumatismo craniano? (Sim <12 meses, Sim >12meses, Não) | Num. Inteiro:1, 2 ou 3 |
| 29 | AVC-P | Já foi informado por algum médico que você teve AVC ou ataque isquêmico transitório? (Sim <12 meses, Sim >12meses, Não) | Num. Inteiro:1, 2 ou 3 |

**Tabela 2.** Atributos da base de dados obtidos de perguntas sobre os níveis de conhecimento.

| **Num.** | **Atributo** | **Descrição** | **Conteúdo** |
|---|---|---|---|
| 30 | NC-FISIO | Nível de conhecimento: dos itens 30 a 40 quais fazem parte da fisiopatologia do AVC? | Num. Inteiro: 1 a 3 |
| 31 | OBSTRU_VEN | Obstrução venosa? | Num. Inteiro: 1 ou 2 |
| 32 | OBSTRU_ART | Obstrução arterial? | Num. Inteiro: 1 ou 2 |
| 33 | OBSTRU_ANY | Obstrução de um vaso qualquer? | Num. Inteiro: 1 ou 2 |
| 34 | PSYCH_OUTBREAK | Surto psicótico? | Num. Inteiro: 1 ou 2 |
| 35 | BREAK_ART | Ruptura arterial? | Num. Inteiro: 1 ou 2 |
| 36 | BREAK_VEN | Ruptura venosa? | Num. Inteiro: 1 ou 2 |
| 37 | HEART_D | Doença cardíaca? | Num. Inteiro: 1 ou 2 |
| 38 | BRAIN_D | Doença cerebral? | Num. Inteiro: 1 ou 2 |
| 39 | PARASITOSE | Parasitose? | Num. Inteiro: 1 ou 2 |
| 40 | SYNCOPE | Síncope? | Num. Inteiro: 1 ou 2 |
| 41 | ACUTE_ART_D | Doença arterial aguda? | Num. Inteiro: 1 ou 2 |

| | | | |
|---|---|---|---|
| 42 | NC-FR | Nível de conhecimento: dos itens 42 a 48 quais são fatores de risco do AVC? | Num. Inteiro: 1 a 3 |
| 43 | AGE>40 | Idade maior que 40 anos? | Num. Inteiro: 1 ou 2 |
| 44 | MALE | Ser do sexo masculino? | Num. Inteiro: 1 ou 2 |
| 45 | DIET_LOW | Dieta inadequada? | Num. Inteiro: 1 ou 2 |
| 46 | HIPERTENSION | Hipertensão? | Num. Inteiro: 1 ou 2 |
| 47 | SEDENTARY | Sedentarismo? | Num. Inteiro: 1 ou 2 |
| 48 | ESTRESS | Estresse? | Num. Inteiro: 1 ou 2 |
| 49 | DRUG | Consumo de drogas ilícitas? | Num. Inteiro: 1 ou 2 |
| 50 | NC-SSS | Nível de conhecimento: dos itens 50 a 59 quais são sinais, sintomas e sequelas do AVC? | Num. Inteiro: 1 a 3 |
| 51 | MOTOR_D | Déficit motor? | Num. Inteiro: 1 ou 2 |
| 52 | COMUNICATION_D | Distúrbios de linguagem? | Num. Inteiro: 1 ou 2 |
| 53 | EYE_D | Distúrbios de visão? | Num. Inteiro: 1 ou 2 |
| 54 | GIDDINESS | Vertigem? | Num. Inteiro: 1 ou 2 |
| 55 | DYSPNEA | Dispneia? | Num. Inteiro: 1 ou 2 |
| 56 | URINE_ACHE | Disúria? | Num. Inteiro: 1 ou 2 |
| 57 | HEART_ACHE | Dor no peito | Num. Inteiro: 1 ou 2 |
| 58 | VOMIT | Vômitos? | Num. Inteiro: 1 ou 2 |
| 59 | EDEMA | Edema? | Num. Inteiro: 1 ou 2 |
| 60 | MENTAL_CONFUSION | Confusão mental? | Num. Inteiro: 1 ou 2 |
| 61 | NC-CON | Nível de conhecimento: dos itens 61 a 64 quais são de conduta imediata no AVC? | Num. Inteiro: 1 a 2 |
| 62 | HOSPITALIZATION | Hospitalização? | Num. Inteiro: 1 ou 2 |
| 63 | ANTIBIOTIC_USE | Uso de antibióticos? | Num. Inteiro: 1 ou 2 |
| 64 | TEA_USE | Uso de chás? | Num. Inteiro: 1 ou 2 |
| 65 | AHIPERTENSIVE | Uso de anti-hipertensivos? | Num. Inteiro: 1 ou 2 |
| 66 | RCV | Risco cardiovascular | Num. Inteiro: 1 a 3 |

### 3.2 Pré-processamento da Base de Dados e Classificação

Na primeira fase do processo de KDD, foram realizadas tarefas de limpeza dos dados, seleção de variáveis, preenchimento de valores ausentes, tratamento de ruídos, entre outras, para melhorar a qualidade dos dados para extração de padrões. Utilizou-se o software de código aberto WEKA [21] para realizar todas as etapas da mineração de dados.

Dos 66 atributos presentes na base de dados original, alguns deles foram excluídos da base devido à baixa ou à ausência total de utilidade para a descoberta de conhecimento. Esses atributos incluem: o código identificador do indivíduo, a cidade e o ano da realização da entrevista.

Adicionalmente, a fim de reduzir a dimensionalidade dos dados, foram aplicados métodos de seleção de variáveis, avaliando o valor dos atributos com a medição da razão de ganho e da informação de ganho relativamente a cada classe.

Em seguida, dividiram-se os 592 registros em 537 para fase de treinamento dos modelos de classificação e 55 para fase de teste, constituindo a primeira linha da tabela 3.

Esta mostra o número de registros e de atributos de entrada para outras nove bases que foram construídas a partir de processamentos da base 1. A diminuição da quantidade de registros observada a partir da base 5 é justificada pela exclusão dos registros com valores faltantes nos campos NC-FISIO, NC-FR, NC-SSS e NC-CON.

Posteriormente, os modelos de classificação de árvores de decisão (C4.5), *random forest*, Part e rede neural MLP foram aplicados nas bases de dados descritas na tabela 3 para escolher o de melhor desempenho.

**Tabela 3.** Descrição das bases resultantes do pré-processamento.

| **Base Treino** | **Num de reg.** | **Descrição** | **Base Teste** | **Descrição** | **Num de reg.** | **Num de atrib. de entrada** |
|---|---|---|---|---|---|---|
| 1 | 537 | Atributos das tabelas 1 e 2, exceto: ID, City e Ano | 1 | Atributos das tabelas 1 e 2, exceto: ID, City e Ano | 55 | 62 |
| 2 | 537 | Igual à base 1, excluindo: NC-FISIO, NC-FR, NC-SSS, NC-CON | 2 | Igual à base 1, excluindo: NC-FISIO, NC-FR, NC-SSS, NC-COM | 55 | 58 |
| 3 | 537 | Igual à base 2, com os faltantes de todos os atributos preenchidos com a média ou moda | 3 | Igual à base 2 | 55 | 58 |
| 4 | 537 | Igual à base 3 | 4 | Igual à base 2, com os faltantes de todos os atributos preenchidos com a média ou moda | 55 | 58 |
| 5 | 486 | Atributos da tabela 1, exceto: ID, City e Ano, incluindo a soma dos NCs | 5 | Atributos da tabela 1, exceto: ID, City e Ano, incluindo a soma dos NCs | 39 | 27 |
| 6 | 486 | Igual à base 5, com os faltantes de todos os atributos preenchidos com a média ou moda | 6 | Igual à base 5 | 39 | 27 |
| 7 | 486 | Igual à base 6 | 7 | Igual à base 5, com os faltantes de todos os atributos preenchidos com a média ou moda | 39 | 27 |
| 8 | 486 | Atributos da tabela 1, exceto: ID, City e Ano, incluindo NC-FISIO, NC-FR, NC-SSS, NC-CON | 8 | Atributos da tabela 1, exceto: ID, City e Ano, incluindo NC-FISIO, NC-FR, NC-SSS, NC-CON | 39 | 30 |
| 9 | 486 | Igual à base 8, com os faltantes de todos os atributos preenchidos com a média ou moda | 9 | Igual à base 8 | 39 | 30 |
| 10 | 486 | Igual à base 9 | 10 | Igual à base 8, com os faltantes de todos os atributos preenchidos com a média ou moda | 39 | 30 |

# 4 Resultados

A tabela 4 mostra a acurácia dos quatro classificadores para cada uma das bases de treinamento e teste, destacando-se as configurações que obtiveram melhor desempenho. A Figura 1 apresenta a árvore de decisão do C4.5 obtida com a base 7.

**Tabela 4.** Percentuais de acerto dos quatro modelos com diversas bases de treinamento e teste.

| % de acertos no Treinamento | | | | | % de acertos no Teste | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Base** | **C4.5** | ***Random forest*** | **Part** | **Rede Neural** | **Base** | **C4.5** | ***Random forest*** | **Part** | **Rede Neural** |
| 1 | 97,02 | 100,00 | 97,21 | 98,88 | 1 | 92,73 | 92,73 | 94,55 | 94,55 |
| 2 | 97,02 | 100,00 | 97,21 | 98,70 | 2 | 92,73 | 92,73 | 96,36 | 96,36 |
| 3 | 98,88 | 100,00 | 97,39 | 97,21 | 3 | 90,91 | 94,55 | 94,55 | 94,55 |
| 4 | 98,88 | 100,00 | 97,39 | 97,21 | 4 | 90,91 | 94,55 | 94,55 | 94,55 |
| 5 | 95,88 | 95,88 | 97,12 | 98,77 | 5 | 97,44 | 97,44 | 97,44 | 94,87 |
| 6 | 97,74 | 100,00 | 97,53 | 98,97 | 6 | 100,00 | 97,44 | 94,87 | 94,87 |
| 7 | 97,74 | 100,00 | 97,53 | 98,97 | 7 | 100,00 | 97,44 | 94,87 | 94,87 |
| 8 | 96,71 | 100,00 | 96,71 | 98,97 | 8 | 95,00 | 100,00 | 95,00 | 90,00 |
| 9 | 98,77 | 100,00 | 97,94 | 98,77 | 9 | 95,00 | 95,00 | 90,00 | 90,00 |
| 10 | 98,77 | 100,00 | 97,94 | 98,77 | 10 | 97,44 | 92,31 | 94,87 | 97,44 |

Relativamente às bases, nota-se que os dados dos NCs são muito relevantes para o desempenho de generalização dos classificadores. Muito embora o *random forest* consiga aprender bem com quase todas as bases de treinamento, o melhor desempenho de teste ocorre com a base 8, que contém menos da metade de atributos da base original, mas contém os NCs.

O modelo C4.5 quando aplicado às bases de teste 6 e 7 alcança 100% de acerto de classificação, na configuração com a menor quantidade de atributos, incluindo um único da tabela 2 que considera a soma dos NCs.

A árvore de decisão da Figura 1 é construída da seguinte forma: utilizando o método de seleção de variáveis intrínseco do C4.5, taxa de ganho, o atributo "DM" (diabetes) é selecionado como mais vantajoso (DM) e é colocado na raiz da árvore, conforme pode ser observado na Figura 1. Nesta condição, 91 registros da base de dados de treinamento são classificados corretamente como risco alto de AVC ('RCV = 3') dado que o valor do atributo DM é igual a 1 ou 2, isto é, 'Sim (>12 meses)' ou 'Sim (<12 meses)'. Caso o valor de "DM" seja igual a 3, ou seja, 'Não', o resultado gerado pela árvore demonstra que o número de registros apresenta grande quantidade de erros, uma vez que o algoritmo inseriu ramos abaixo do 'DM = 3'.

```
DM = 1: 3 (67.0)
DM = 2: 3 (24.0)
DM = 3
|  VENOUS_D = 1: 3 (39.0)
|  VENOUS_D = 2: 3 (13.0)
|  VENOUS_D = 3
| |  TCE = 1: 3 (9.0)
| |  TCE = 2: 3 (1.0)
| |  TCE = 3
| | |  AVC-P = 1: 3 (3.0)
| | |  AVC-P = 2: 3 (1.0)
| | |  AVC-P = 3
| | | |  HISTORY = 1
| | | | |  GENDER = 1: 2 (58.0)
| | | | |  GENDER = 2
| | | | | |  PHY_ACT = 1
| | | | | | |  FAMILY = 1: 2 (8.0)
| | | | | | |  FAMILY = 2
| | | | | | | |  HIP = 1: 2 (5.0)
| | | | | | | |  HIP = 2
| | | | | | | | |  P_NEAR = 1: 1 (19.0/3.0)
| | | | | | | | |  P_NEAR = 2: 2 (2.0)
| | | | | | | |  HIP = 3: 1 (0.0)
| | | | | | |  FAMILY = 3: 2 (0.0)
| | | | | |  PHY_ACT = 2: 2 (49.0/1.0)
| | | | | |  PHY_ACT = 3: 2 (0.0)
| | | | |  HISTORY = 2
| | | | |  GENDER = 1
| | | | | |  PHY_ACT = 1
| | | | | | |  FUMATE = 1: 2 (15.0)
| | | | | | |  FUMATE = 2
| | | | | | | |  FAMILY = 1: 2 (4.0)
| | | | | | | |  FAMILY = 2
| | | | | | | | |  AGE <= 63
| | | | | | | | | |  HIP = 1: 2 (2.0)
| | | | | | | | | |  HIP = 2
| | | | | | | | | | |  WEIGHT <= 82: 1 (19.0/1.0)
| | | | | | | | | | |  WEIGHT > 82
| | | | | | | | | | | |  C_STATUS = 1: 2 (2.0)
| | | | | | | | | | | |  C_STATUS = 2
| | | | | | | | | | | | |  HEIGHT <= 1.71: 2 (2.0)
| | | | | | | | | | | | |  HEIGHT > 1.71: 1 (2.0)
| | | | | | | | | | | |  C_STATUS = 3: 2 (0.0)
| | | | | | | | | | | |  C_STATUS = 4: 2 (0.0)
| | | | | | | | | |  HIP = 3: 1 (0.0)
| | | | | | | | |  AGE > 63: 2 (10.0)
| | | | | | |  FAMILY = 3: 2 (0.0)
| | | | | |  PHY_ACT = 2: 2 (38.0/1.0)
| | | | | |  PHY_ACT = 3: 2 (0.0)
| | | | |  GENDER = 2
| | | | | |  HIP = 1
| | | | | | |  PAD <= 85: 2 (15.0)
| | | | | | |  PAD > 85: 1 (2.0)
| | | | | |  HIP = 2
| | | | | | |  FUMATE = 1
| | | | | | | |  PHY_ACT = 1: 1 (9.0/2.0)
| | | | | | | |  PHY_ACT = 2: 2 (7.0)
| | | | | | | |  PHY_ACT = 3: 2 (0.0)
| | | | | | |  FUMATE = 2
| | | | | | | |  DIET = 1: 1 (29.0)
| | | | | | | |  DIET = 2
| | | | | | | | |  C_STATUS = 1: 1 (14.0/1.0)
| | | | | | | | |  C_STATUS = 2
| | | | | | | | | |  PAS <= 120: 1 (10.0/1.0)
| | | | | | | | | |  PAS > 120: 2 (5.0)
| | | | | | | | |  C_STATUS = 3: 2 (1.0)
| | | | | | | | |  C_STATUS = 4: 1 (2.0/1.0)
| | | | | | | |  DIET = 3: 1 (0.0)
| | | | | |  HIP = 3: 1 (0.0)
```

**Fig. 1.** Árvore de decisão do C4.5 para a base 7.

Neste caso, a avaliação da taxa de ganho é novamente calculada para todos os registros restantes da base de dados, com exceção dos 91 já classificados e, então, o atributo "VENOUS_D" (doença cardíaca ou doença arterial periférica) é selecionado como o mais promissor da lista de atributos apresentada. Para este atributo, 52 registros da base de dados foram classificados como de alto risco de AVC ('RCV = 3') para "VENOUS_D" igual a 1 ou 2, isto é, 'Sim (>12 meses)' ou 'Sim (<12 meses)'. Quando VENOUS_D é igual a 3, ou seja, 'Não', a subrede apresentada indica novamente que o erro da classificação é alto e por isso inseriu o atributo TCE abaixo do ramo VENOUS_D igual a 3. Este processo se repete até que os registros da base de treinamento tenham sido todos avaliados ou a construção de novos ramos se torne inviável por restrição dos parâmetros de construção da árvore.

## 5 Conclusão

Neste trabalho, foram realizados diferentes pré-processamentos em base de dados, contendo dados relacionados ao conhecimento de indivíduos sobre AVC, a fim de construir bases com atributos de maior relevância, permitindo melhor desempenho de quatro técnicas clássicas de mineração de dados. Os resultados de todas as técnicas foram promissores na classificação do risco de AVC, com acurácia mínima de 90%. Os métodos baseados em árvores de decisão obtiveram melhor precisão, notadamente o *random forest* e o C4.5. Este último com 100% de acertos. Além disso, demonstramos que o desempenho de todos os classificadores melhora quando o nível de conhecimento sobre o AVC, incluindo fatores de risco, fisiopatologia, sinais e sintomas e conduta imediata, é incluído no banco de dados.

A base de teste foi selecionada dos dados coletados e restrita a cerca de 10% da base original devido ao número limitado de registros. O sucesso obtido na mineração é devido ao pré-processamento dos dados, principalmente à seleção de variáveis, que reduziu o número de atributos em mais de 50%. De outro modo, o aumento da base de teste implica na redução da base de treinamento, aumentando as taxas de erros no treinamento e/ou no teste. Neste sentido, verificou-se que ao aumentar a base de teste para o dobro dos dados avaliados, a taxa de erro aumentou para 7,63%, isto é, seis registros classificados erroneamente utilizando-se o método J48.

Em um trabalho futuro, há a necessidade de realizar um estudo mais aprofundado em relação ao aprendizado destes classificadores para cada uma das classes, incluindo validação cruzada, aumento da base de dados e a comparação dos resultados com testes estatísticos.

## Referências


1. World Health Organisation. Cardiovascular diseases (CVDs). Fact Sheet no 317. Updated May 2017, http://www.who.int/mediacentre/factsheets/fs317/en/.
2. World Health Organisation. Stroke, Cardiovascular Accident. Health topics., http://www.who.int/topics/cerebrovascular_accident/en/.
3. Fernandes, A., Azevedo, A., Magalhães, C., Antão, C., Anes, E.: Avaliação do conhecimento referente à deteção precoce e prevenção do Acidente Vascular Cerebral: Dilemas atuais e desafios futuros. In: I Congresso de Cuidados Continuados da Unidade de Longa Duração e Manutenção de Santa Maria Maior. , Bragança (2012).
4. Costa, F., Oliveira, S., Magalhães, P., Costa, B., Papini, R., Silveira, M., Lang, M.: Nível de conhecimento da população adulta sobre acidente vascular cerebral (AVC) em Pelotas-RS, https://www.abnc.org.br/jbnc_art_down.php?id=572. Acesso em: 10 de agosto de 2017.
5. Hajimani, E., Ruano, M.G., Ruano, A.E.: MOGA design for neural networks based system for automatic diagnosis of Cerebral Vascular Accidents. In: 2015 IEEE 9th International Symposium on Intelligent Signal Processing (WISP) Proceedings. pp. 1–6 (2015).
6. Ruano, M.G., Hajimani, E., Ruano, A.E.: A Radial Basis Function classifier for the automatic diagnosis of Cerebral Vascular Accidents. In: 2016 Global Medical Engineering Physics Exchanges/Pan American Health Care Exchanges. pp. 1–4 (2016).
7. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. Int Conf Knowl. Discov. Data Min. 82–88 (1996).
8. Han, J., Kamber, M., Pei, J.: Data mining: concepts and techniques. Morgan Kaufmann Publishers, Waltham, Mass. (2012).
9. Egan, J.P.: Signal detection theory and {ROC} analysis. Academic Press, New York, NY (1975).
10. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1993).
11. Quinlan, J.R.: Induction of decision trees. Mach. Learn. 1, 81–106 (1986).
12. Breiman, L.: Using Iterated Bagging to Debias Regressions. Mach. Learn. 45, 261–277 (2001).
13. Breiman, L.: Bagging Predictors. Mach. Learn. 24, 123–140 (1996).
14. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844 (1998).
15. Witten, I.H., Frank, E.: Data Mining: practical machine learning tools and techniques with Java implementations. Morgan Kaufmann Publishers, San Francisco, CA, USA (2000).
16. Frank, E., Witten, I.H.: Generating accurate rule sets without global optimization., Hamilton, New Zealand: University of Waikato, Department of Computer Science (1998).
17. Baranauskas, J.A., Monard, M.C.: Reviewing some Machine Learning Concepts and Methods. (2000).
18. Haykin, S.: Neural Networks: A Comprehensive Foundation (3rd Edition). Prentice-Hall, Inc., Upper Saddle River, NJ, USA (2007).
19. Cybenko, G.: Approximation by superpositions of a sigmoidal function. Math. Control. Signals Syst. 2, 303–314 (1989).
20. Ministério da Saúde. Secretaria de Atenção à Saúde. Departamento de Atenção Básica.: Rastreamento – série A. Normas e Manuais Técnicos. Cadernos de Atenção Primária, n. 29., Brasília, DF (2010).
21. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, Second Edition. Morgan Kaufmann Publishers, San Francisco, CA, USA (2005).