Redes Bayesianas para Previsão de Doadores de Sangue

Cristina Zayra de Nobrega Romani

Universidade Federal de Uberlândia
Faculdade de Computação
Bacharelado em Sistemas de Informação

Monte Carmelo - MG
2021

Cristina Zayra de Nobrega Romani

Redes Bayesianas para Previsão de Doadores de Sangue

Trabalho de Conclusão de Curso apresentado à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como requisito exigido parcial à obtenção do grau de Bacharel em Sistemas de Informação.

Área de concentração: Inteligência Artificial

Orientador: Profa. Dra. Fernanda Maria da Cunha Santos

Monte Carmelo - MG
2021

Gostaria de dedicar esse trabalho principalmente a minha família por sempre estarem comigo, e também aos meus amigos e professores que fizeram parte dessa jornada.

Agradecimentos

Agradeço primeiramente a Deus pela constante presença em minha vida me dando força e sabedoria nos momentos de dificuldades.

Agradeço aos meus pais Aguinaldo Sebastião Romani e Aurea Maria Marques de Nobrega e meus irmãos Gustavo de Nobrega Romani e Rodnei de Nobrega Romani, pelo o amor, incentivo e apoio incondicional.

A minha orientadora Fernanda Maria da Cunha Santos pelos ensinamentos, apoio e paciência para a realização desse trabalho, e ao professor Marcos Luiz de Paula Bueno por também participar do desenvolvimento.

Aos professores Carlos Cesar Mansur Tuma e Leandro Nogueira Couto por aceitarem o convite para compor a banca e pela presença em minha vida acadêmica.

Meu amigo Rafael Luan, pelo companheirismo e amizade nesses anos de graduação.

E por fim Diego Silva Siqueira, meu gestor e mentor pelos conhecimentos transmitidos e pela ajuda que tem me dado todos esses anos.

‘"A tecnologia ensinou uma lição à humanidade: nada é impossível."
(Lewis Mumford)

Resumo

Os hemocentros são entidades responsáveis por administrar bancos de sangue para transfusões. Em diversas situações, o hemocentro necessita prever doadores de sangue regulares para garantir o estoque e a rotatividade exigida pelos Hospitais. Diante disso, este trabalho apresenta uma solução baseada no aprendizado de máquina utilizando o modelo probabilístico das Redes Bayesianas para predição de doadores regulares. Os classificadores empregados são o Naive Bayes e o Tree Augmented Naive Bayes (TAN). A base de dados analisada foi "Conjunto de dados do Centro de Serviços de Transfusão de Sangue", retirada do repositório *UCI Machine Learning Repository*. Em primeiro instante aplicou-se as medidas de associação, risco relativo e odds ratio, para encontrar as variáveis que afetam na escolha do doador regular. A base possui classes desbalanceadas sendo necessário aplicar as técnicas SMOTE e k-fold estratificado para neutralizar o problema. Para avaliar os modelos empregados utilizou-se as métricas acurácia, precisão e sensibilidade. Os resultados encontrados demonstraram que o TAN possui maiores taxas de acerto para encontrar um doador de sangue regular. Porém, não foi possível obter melhores resultados devido ao desbalanceamento da base de dados.

Palavras-chave: Doação de sangue, Rede Bayesiana, Modelo probabilístico, Aprendizado de máquina.

Lista de ilustrações

Lista de tabelas

Sumário

Capítulo 1

Introdução

O sangue é responsável por transportar substâncias pelo corpo dos seres vivos com grande importância para a preservação da vida. Em cenários adversos, como por exemplo, em casos de doenças e acidentes é necessário a transfusão de sangue (MOURA et al., 2006). Para atender à essa procura é necessário que os estoques de sangue dos hemocentros estejam em um nível considerável para que seja possível atender a todos.

Para manter os bancos de sangue em níveis estáveis de abastecimento é fundamental que aconteçam doações periódicas e isso se torna um dos principais problemas enfrentados pelos hemocentros (CARLESSO et al., 2017). Em alguns casos as doações ocorrem de forma altruísta, voluntária e humanística (CUNHA; DIAS, 2008) quando a pessoa se predispõe a doar sangue. Entretanto, o cenário mais comum é a ausência desse elemento.

Em busca de evitar que essa situação se agrave causando a diminuição das bolsas de sangue, os hemocentros se tornam responsáveis por buscarem em sua base de dados indivíduos que são julgados como mais prováveis doadores de sangue regulares. A busca é realizada através de vários parâmetros e variáveis o que pode causar demanda de tempo desnecessária. Assim, com o propósito de auxiliar e aprimorar a pesquisa, o uso de alguma ferramenta computacional para encontrar esse doador torna-se uma solução eficiente, além de identificar previsões de doadores regulares para o próximo mês (SILVA, 2018).

1.1 Motivação

Com o grande avanço computacional surgiu o ramo de pesquisa conhecido como inteligência artificial. Segundo Kaplan e Haenlein (2018) a inteligência artificial é "definida como a capacidade de um sistema de interpretar corretamente dados externos, aprender com esses dados e usá-los para atingir objetivos e tarefas específicos por meio de adaptação flexível [...]". Essa área busca utilizar mecanismos computacionais para simular a inteligência e a capacidade de resolver problemas complexos.

De acordo com Korb e Nicholson (2011) a inteligência artificial está presente na medicina desde 1960 utilizando seus esforços para o diagnóstico de doenças. Ao longo do

tempo, essa área se tornou cada vez mais presente nesse cenário com a utilização de sistemas que são, por exemplo, capazes de identificar em imagens digitais lesões na pele, a utilização de probabilidades para previsão do diagnóstico de doenças (LOBO, 2017).

A atuação de algoritmos da Inteligência Artificial na tarefa de previsão usa os valores de atributos do sistema em estudo para classificar um determinado atributo conhecido como variável alvo (TAN; STEINBACH; KUMAR, 2009). Diversos algoritmos foram usados com esse propósito na medicina: redes neurais artificiais, árvores de decisão, máquinas de vetores de suporte, classificadores Bayesianos, classificadores de vizinho mais próximo e outros. Com o objetivo de auxiliar na compreensão e percepção do problema de predição de doadores regulares, propõe-se o uso de Redes Bayesianas.

As Redes Bayesianas são modelos gráficos probabilísticos compostos por nós, que representam as variáveis utilizadas, e os arcos responsáveis por demostrar ligações e conexões entre elas, sendo capaz de demonstrar sua dependência probabilística (SATO; SATO, 2015). Além disso, são modelos interpretáveis capazes de permitirem melhor compreensão a interação probabílistica através das variáveis do problema.

Como exemplo, cita-se a dissertação de Silva (2018) que utilizou um classificador Naive Bayes com a mesma base de dados aplicada à este trabalho, denominada "Conjunto de dados do Centro de Serviços de Transfusão de Sangue" que se encontra no Repositório de Aprendizado de Máquina (UCI) (YEH, 2008). Nesta dissertação, o autor comparou diferentes classificadores de Aprendizado de Máquina, mas não testou a topologia de rede Bayesiana Tree Augmented Naive Bayes, a qual será analisada neste estudo.

1.2 Objetivos

1.2.1 Objetivo Geral

O trabalho tem como objetivo principal avaliar se as redes Bayesianas Naive Bayes e Tree Augmented Naive Bayes são modelos relevantes para classificar doadores regulares de sangue e, desta forma, auxiliar os Hemocentros na previsibilidade de doadores.

1.2.2 Objetivos Específicos

- ❏ Analisar diferentes topologias de uma Rede Bayesiana como é o caso do Naive Bayes e Tree Augmented Naive Bayes;
- ❏ Avaliar se a base de dados "Conjunto de dados do Centro de Serviços de Transfusão de Sangue" da UCI está de fato adequada para esse trabalho;
- ❏ Avaliar quantitativamente a previsão de redes bayesianas na base de dados por meio de medidas como acurácia, precisão e sensibilidade;

- ❑ Encontrar quais variavéis são mais preditivas para a escolha de um doador utilizando as medidas de associação risco relativo e odds ratio;

1.3 Hipótese

A hipótese desse trabalho é comprovar que as redes bayesianas são modelos probabilísticos apropriados para resolver o problema de previsão de um doador de sangue regular, incluindo, previamente, a validação dos atributos da base de dados pelas medidas de associação.

1.4 Organização da Monografia

O Capítulo 2 descreve os conceitos necessários para a compreensão deste trabalho, além de conter a exposição dos principais trabalhos científicos relacionados ao tema proposto e que auxiliaram o desenvolvimento deste. O Capítulo 3 apresenta as etapas da metodologia empregada no classificador, bem como os resultados alcançados. Por fim, o Capítulo 4 expõe as conclusões obtidas.

Capítulo 2

Fundamentação Teórica

Neste capítulo, apresenta os conceitos de redes bayesianas, e as definições de conteúdos relevantes para o entendimeto deste trabalho. Ademais, descreve-se a revisão bibliográfica dos trabalhos correlatos.

2.1 Redes Bayesianas

As redes bayesianas são modelos gráficos capazes de representar o domínio em estudo e as incertezas envolvendo suas variáveis, facilitando a percepção do conhecimento.

A sua representação é realizada por um modelo matemático conhecido como grafo direcionado acíclico que busca apresentar as relações que existem em uma determinada situação. Ela é formada por nós que representam as variáveis utilizadas no domínio e arcos que conectam cada nó, que são reponsáveis por representar a interação probabilística entre cada variável.

Para a sua representação é definida uma dependência que usualmente se denomina de nó pai e nó filho. As variáveis que originam os arcos são chamadas de pais e as que recebem os arcos são chamadas de filhos. Para exemplificar, na Figura 1 é apresentado um exemplo de rede bayesiana simples onde o nó B é pai de D e E que são seus filhos, por sua vez, o nó A é pai de B que é seu filho.

A representação de rede bayesiana é definida através da relação de nós filhos com o seus pais, onde cada nó possui uma distribuição de probabilidade específica. Para se calcular a probabilidade conjunta de n nós usando a rede é definida a seguinte expressão:

$$P\left(X_1 = x_1, \cdots, X_n = x_n\right) = \prod_{i=1}^{n} P[X_i = x_i \mid pais(X_i)] \tag{1}$$

onde $x_1, \cdots, X_n$ representam os nós da rede bayesiana, $pais(Xi)$ é o nó pai do nó em análise e $P[X_i = x_i \mid pais(X_i)]$ é a probabilidade condicional do nó analisado.

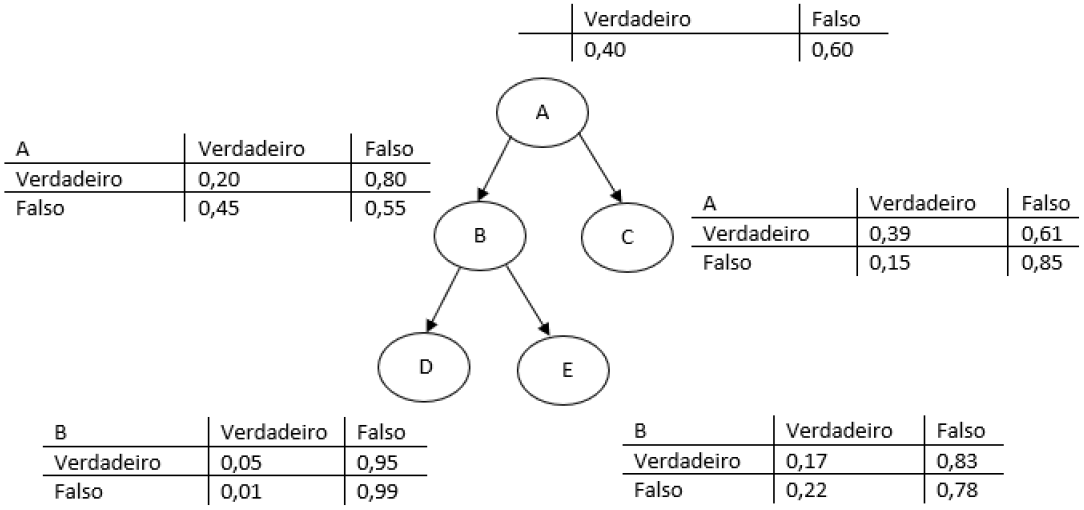

Figura 1 – Representação de uma Rede Bayesiana.

Fonte: Autora.

2.2 Classificadores Bayesianos

2.2.1 Teorema de Bayes

O Teorema de Bayes é um teorema proposto pelo pastor Thomas Bayes (1702-1761) muito estudado e utilizado na área da estatística para definir a probabilidade de eventos acontecerem mediante a outro evento já ter acontecido (ARA-SOUZA, 2010). O teorema deseja encontrar a conexão entre dois eventos distintos, buscando encontrar as suas relações e a probabilidade de acontecimento de um evento fundamentado que o outro já tenha acontecido.

O teorema de Bayes é representado pelo cálculo de probabilidade à posteriori de um padrão $x = (x_1, x_2, ..., x_d) \in \Re^d$ estar associado a uma classe $\omega_1, \omega_2, ..., \omega_c$. Assim, a abordagem Bayesiana supõe que $P(\omega_i|x)$ é obtido pelas probabilidades de cada classe $P(\omega_i)$ multiplicado pelas densidades de probabilidade condicionais $p(x|\omega_i)$ de x com respeito a cada uma das classes $\omega_i, i = 1, 2, ..., c$, como descrito na Equação 2.

$$P(\omega_i \mid x) = \frac{P(x \mid \omega_i) * P(\omega_i)}{P(x)} \tag{2}$$

2.2.2 Naive Bayes

O Naive Bayes é uma técnica utilizada para delimitar, em um conjunto de variáveis, classes específicas. Essa divisão é realizada para agrupar elementos que possuem as mes-

mas características levando em consideração que não existem mais ligações para que essa mesma característica aconteça, ou seja, independente de cenários que existam mais ligações externas. Por causa disso, o Naive Bayes é conhecido como ingênuo (SCB, 2016). O Naive Bayes pode ser visto como um caso especial de Redes Bayesianas.

É possível citar como exemplo essa situação:

> Um fruto pode ser considerado como uma maçã se é vermelho, redondo, e tiver cerca de 3 polegadas de diâmetro. Mesmo que esses recursos dependam uns dos outros ou da existência de outras características, todas estas propriedades contribuem de forma independente para a probabilidade de que este fruto é uma maçã (RAY, 2016).

O Naive Bayes é baseado no Teorema de Bayes. Para se prever a variável classe C em função das demais n variáveis, calcula-se:

$$P(C|x_1, \ldots, x_n) = \alpha P(C) \prod_{i=1}^{n} P(x_i|C) \tag{3}$$

onde $\alpha = P(x_1, \ldots, x_n)$ e é constante para todas as classes e $P(C)$ é a distribuição a priori da variável classe. O valor da variável C que resultar na maior probabilidade na Equação 3 é escolhido como a classe prevista.

2.2.2.1 Naive Bayes Gaussiana

O algoritmo Naive Bayes pode ter variações no cálculo da probabilidade dos atributos condicionados a classe $P(x_i|C)$. Dentre essas, destaca-se a função densidade de probabilidade normal:

$$P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} exp(-\frac{(x_i - \mu_C)^2}{2\sigma_C^2}) \tag{4}$$

Na Equação 4, a média μ é o valor médio de x_i considerando as observações da classe C, e a variância σ^2 é a variância da classe x_i considerando as observações da classe C.

2.2.3 Tree Augmented Naive Bayes

Tree Augmented Naive Bayes (TAN) é um classificador desenvolvido por Friedman e Goldszmidt em 1997, sendo considerado uma versão atualizada do Naive Bayes por apresentar melhorias na sua forma e no seu desempenho (ZHANG; JIANG; SU, 2005; SCHEUNEMANN, 2015). Nele é possível representar dependências entre as variáveis presentes no cenário estudado e demonstrar as suas correlações (PADMANABAN, 2014).

No modelo TAN, diferente do Naive Bayes, são encontradas as relações entre as variáveis do domínio, sendo possível identificar as possíveis correlações existentes entre cada

evento. Esse cálculo é em função da informação mútua, onde os atributos mais correlacionados são conectados. A informação mútua é calculada através da Equação 5:

$$Ip(X;Y) = \sum_{x,y} P(x,y) log(\frac{P(x,y)}{P(x)P(y)}) \tag{5}$$

onde X e Y são as váriaveis analisadas e o valor Ip(X;Y) é a informação que X exerce sobre Y ou vice-versa.

No final, o grafo do modelo é formado pela estrutura de uma árvore, que é calculado pela informação mútua entre pares de variáveis, dado pela Equação 6:

$$Ip(X;Y|Z) = \sum_{x,y,z} P(x,y,z) log(\frac{P(x,y|z)}{P(x|z)P(y|z)}) \tag{6}$$

onde X, Y e Z são as variáveis do estudo e Ip(X;Y|Z) é referente a quanta informação Y fornece para X onde Z é conhecido.

No TAN, a classe é o nó conectado a todos os outros, já o nó raiz é o único que possui um pai, a classe, diferente dos demais que são filhos da classe e de outro atributo.

A probabilidade a posteriori é calculado pela Equação 7:

$$P(C|x_1,\ldots,x_n) = P(C) \cdot P(X_{root}|C) \prod_{i=1}^{n} P(x_i|C, X_{parent}) \tag{7}$$

onde $P(C)$ é a distribuição a priori da variável classe, $P(X_{root}|C)$ é a probabilidade do nó raiz da árvore dado que a classe aconteceu e $P(x_i|C, X_{parent})$ é a probabilidade do nó analisado dado que a classe e seu pai aconteceram.

2.3 Medidas de Associação

As medidas de associação são utilizadas para quantificar o grau de interferência de uma variável em um determinado ambiente. Definem a chance de uma pessoa desenvolver determinado evento baseado em características específicas, o que justifica suas aplicações em estudos epidemiológicos. No caso deste trabalho, foi aplicado as medidas de associação para avaliar os atributos da base de dados escolhida para prever um perfil de um possível doador de sangue regular.

Os estudos são desenvolvidos em uma amostra da população identificando no ambiente em questão o desfecho e o fator de risco. O desfecho é o evento de interesse da pesquisa, e o fator de risco são as variáveis que influenciam para que esse desfecho aconteça (WAGNER; CALLEGARI-JACQUES, 1998).

As medidas de associação destacadas nesse trabalho são risco relativo e odds ratio.

Afim de exemplificar o conceito das medidas pode-se utilizar a Figura 2. As variáveis de a,b,c e d representam os valores de cada cenário retrado como a presença e a ausência do desfecho e fator de risco.

	Desfecho Presente	**Desfecho Ausente**	**Total**
Fator de Risco Presente	a	b	a+b
Fator de Risco Ausente	c	d	c+d
Total	a+c	b+d	n

Figura 2 – Representação das Medidas de Associação.

Fonte: Autora.

2.3.1 Risco Relativo

O risco relativo (RR) é considerado uma medida de associação em coortes. Essa medida trata da análise de um grupo específico baseado em um determinado tempo, define o grau de exposição de um fator, um hábito ou uma condição sobre um determinado evento. Ele é definido como a razão da incidência dos casos entre os expostos e a incidência de casos entre os não expostos (FRANCO; PASSOS, 2011).

$$RR = \frac{incidência\ dos\ casos\ entre\ os\ expostos}{incidência\ de\ casos\ entre\ os\ não\ expostos} = \frac{a/(a+b)}{c/(c+d)} \tag{8}$$

Como os estudos do risco relativo utilizam amostras relacionadas a uma população, se faz necessário obter um intervalo de confiança (IC) para justificar se o valor encontrado deve influenciar na população, além de confirmar se a variável analisada interfere no desfecho. Para a demonstração do cálculo do intervalo de 95% de confiança do RR, será utilizado o método da transformação logarítmica descrito por Gardner Altman, 1989. A fórmula para o cálculo do intervalo de confiança do RR é descrito na Equação 9.

$$ICRR = exp[ln(RR) + -Z\alpha \cdot EPln(RR)] \tag{9}$$

$$EPln(RR) = \sqrt{\left[\frac{1}{a} - \frac{1}{a+b}\right]\left[\frac{1}{c} - \frac{1}{c+d}\right]} \tag{10}$$

Na Equação 9 ln(RR) é o logaritmo do risco relativo e $Z\alpha$ é o limite crítico bi-caudal para distribuição normal.

2.3.2 Odds Ratio

O odds ratio (OR) é uma medida de caso-controle que define a probabilidade da ocorrência de um evento em dois tipos de grupos: os que foram expostos ao fator determinante para causar o desfecho e os que não foram expostos.

Segundo Franco e Passos (2011), em estudos de caso-controle, a amostra é dividida em grupos de casos e grupos de controles. Nos grupos de casos acontece de fato o desfecho e sua análise é baseada na razão daqueles que foram expostos ou não ao fator determinante.

$$caso = \frac{casos\ com\ a\ exposição}{casos\ sem\ a\ exposição} = \frac{a}{a+c} \tag{11}$$

Nos grupos de controle não acontece o desfecho e sua análise é baseada na razão daqueles que foram expostos ou não ao fator determinante.

$$controle = \frac{controle\ com\ a\ exposição}{controle\ sem\ a\ exposição} = \frac{b}{b+d} \tag{12}$$

Portanto, através dessas duas situações, é possível determinar a fórmula:

$$OR = \frac{casos}{controle} = \frac{casos\ com\ a\ exposição\ \cdot controle\ sem\ a\ exposição}{controle\ com\ a\ exposição\ \cdot casos\ sem\ a\ exposição} = \frac{a \cdot d}{b \cdot c} \tag{13}$$

Semelhante ao Risco Relativo avalia-se Odds Ratio através do intervalo de confiança. A fórmula para o cálculo do intervalo de confiança do OR é descrito na Equação 14.

$$ICOR = exp[ln(OR) + -Z\alpha \cdot EPln(OR)] \tag{14}$$

$$EPln(OR) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \tag{15}$$

Na mesma maneira a a Equação 14 o ln(OR) é o logaritmo do odds ratio e $Z\alpha$ é o limite crítico bi-caudal para distribuição normal.

2.3.3 Exemplo

O Franco e Passos (2011) apresenta um exemplo sobre o tabagismo na gravidez. Foram utilizadas quinhentas gestantes tabagistas e quinhentas gestantes não tabagistas. Os filhos de cinquenta fumantes nasceram com o peso abaixo de 2.500 gramas e nas não fumantes em apenas cinco. A Tabela 1 apresenta os dados encontrados.

Através desses dados é possível calcular as medidas de associação descritas nesse texto.

O valor do risco relativo é:

$$RR = \frac{\frac{50}{500}}{\frac{5}{500}} = \frac{0,1}{0,01} = 10 \tag{16}$$

Tabela 1 – Exemplo das Medidas de Associação

	Com baixo peso	**Sem baixo peso**	**Total**
Tabagismo presente	50	450	500
Tabagismo ausente	5	495	500
Total	55	945	1.000

Fonte: Franco e Passos (2011)

Com esse valor do risco relativo é possível afirmar que existe uma chance 10 vez maior para mulher que fuma ter um filho de baixo peso do que aquela que não fuma.

E por fim, o valor do odds ratio é:

$$OR = \frac{50 \ * 495}{5 \ * 450} = \frac{24.750}{2.250} = 11 \tag{17}$$

Através do valor do odds ratio é possível deduzir que as gestantes que são tabagistas tem um risco de 11 vezes mais que as mulheres que não são tabagistas de ter um filho que é de baixo do peso.

2.4 Validação Cruzada Estratificada

A validação cruzada é uma técnica para avaliar classificadores por meio do treinamento de vários subconjuntos de dados de entrada disponíveis, além da avaliação no subconjunto complementar dos dados, denominado conjunto de teste.

A validação cruzada evita conjuntos de teste com interseção não vazia. A base de dados é dividida em K conjuntos de mesmo cardinal, e em cada conjunto cria-se subconjuntos dividindo-os entre grupos de dados para treinamento e um para teste.

Ao subdividir os conjuntos, a validação cruzada estratificada mantém a proporção de classes da base completa, para que possa garantir essa mesma proporção nas subdivisões nos grupos de dados destinados para treino e teste. A Figura 3 demonstra um exemplo de validação cruzada estratificada com *k-fold* igual a 5.

A taxa de erro global é obtida pela média das taxas de erro calculadas em cada etapa.

2.5 Medidas de Desempenho

Os classificadores propostos foram avaliados segundo as métricas de desempenho acurácia, precisão e sensibilidade (tradução da palavra *recall*). Antes de definir as métricas é necessário entender os valores que compõem a matriz de confusão, representada na Figura 4, que são os valores que indicam os erros e acertos do modelo ao comparar com os resultados esperados.

Os valores da matriz de confusão são: VP (Verdadeiro Positivo), VN (Verdadeiro Negativo), FP (Falso positivo) e FN (Falso Negativo).

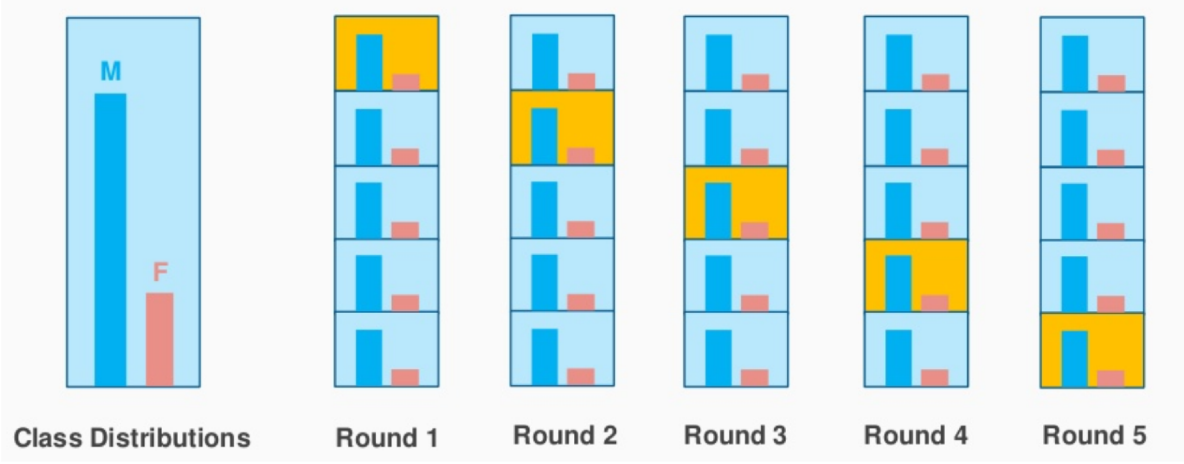

Figura 3 – Representação de k-fold estratificado igual a 5.

Fonte: https://stats.stackexchange.com/questions/49540/understanding-stratified-cross-validation.

		Verdade	
		Condição Positiva	Condição Negativa
Predito	Condição positiva prevista	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	Condição negativa prevista	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Figura 4 – Representação da Matriz de Confusão.

Fonte: Autora.

Realizado a contagem dos valores da matriz de confusão, as métricas de avaliação de performance podem ser calculadas:

- Acurácia: corresponde ao percentual de acertos entre os valores preditos e o resultado real.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{18}$$

- Precisão: define dentre os valores previstos como positivos quantos realmente são

verdadeiros.

$$Precisão = \frac{VP}{VP + FP} \tag{19}$$

- ❑ Sensibilidade: relaciona os valores previstos como positivos de todas as situações que são realmentes positivas.

$$Sensibilidade = \frac{VP}{VP + FN} \tag{20}$$

2.6 Trabalhos Relacionados

2.6.1 Métodos para a predição de doadores de sangue

O primeiro trabalho é a dissertação de Silva (2018), na qual o seu objetivo é encontrar soluções computacionais para os problemas do Hemocentro Regional de Catalão (HEMOCAT), mais especificamente na forma em que é realizada a busca para encontrar possíveis doadores de sangue em situações de emergências. Para isso o autor utilizou uma base de dados encontrada na UCI.

Primeiramente, foi avaliado que a base de dados utilizada apresentava um desbalanceamento, visto que, os dados na distribuição de doadores eram irregulares. Isso aconteceu devido a mensuração das pessoas que doaram sangue ou não doaram na campanha realizada em março no ano de 2007, calculou-se que o valor era de 24% (os que doaram sangue) e de 76% (os que não doaram sangue). Em busca de ajustar e resolver o problema apresentado foi utilizado o método SMOTE.

Após os ajustes necessários a base de dados foi avaliada, nas duas versões estabelecidas, com algoritmos classificadores Naive Bayes, Support VectorMachine, Multi-Layer Perceptron, Radial Basis Function Network, k-Nearest Neighbor (utilizando valores ímpares de 1 a 33 para os k-vizinhos), e as árvores de decisão C4.5 e CART. A avaliação desse algoritmos foram baseadas em três medidas recall, precisão e acurácia.

O algoritmo classificador com o melhor resultado foi o CART, que obteve uma acurácia de 77,46% na base de dados original e 73,80% na base de dados balanceada, porém ao longo do experimento os valores encontrados não foram os ideais e devido a essa circunstância o autor buscou uma nova abordagem para avaliação o uso das recomendações dos Top-K.

Nas recomendações dos Top-K uma porção de 30% da base de dados original foi definida como base para treinamento e para a aplicação de três tipos de heurística: baseada nos vizinhos mais próximos, no Teorema de Bayes e no hiperplano de separação das classes, e estas foram utilizadas para encontrar o grau de confiança dos que pertencem a classe doar.

Para avaliação foram utilizados as recomendações de top-1 até top-10 e os classificadores com as melhores médias nos resultados do primeiro experimento foram utilizados com

as heurística apresentadas como é o caso do Support Vector Machine, k-Nearest Neighbor e Naive Bayes. O resultado obtido foi que a Support Vector Machine obteve os melhores resultados em que para o top-1 chegou a 99,90%.

2.6.2 Diagnóstico de doenças cardiovasculares utilizando Redes Bayesianas

Na dissertação de Saheki (2005) o seu foco é a utilização de redes bayesianas para o diagnótico de doenças cardiovasculares, onde se busca encontrar o grau do problema respiratório apresentado em pacientes atendidos em postos de saúde. Através disso, deve ser possível definir quais pessoas que precisam ser encaminhadas aos hospitais para tratamento mais específico e aquelas que podem permanecer no próprio posto para a assistência devido ao baixo grau periculosidade de seu estado.

A primeira modelagem da rede bayesiana ocorreu de discussões em reuniões com especialistas da área e com fontes bibliográficas. Em seguida, o autor buscou refinar a rede da mesma maneira, incluindo que o nó de maior importância na rede é o Cardiomiopatia. Na Figura 5 é apresentado o modelo final da rede bayesiana utilizada pelo autor.

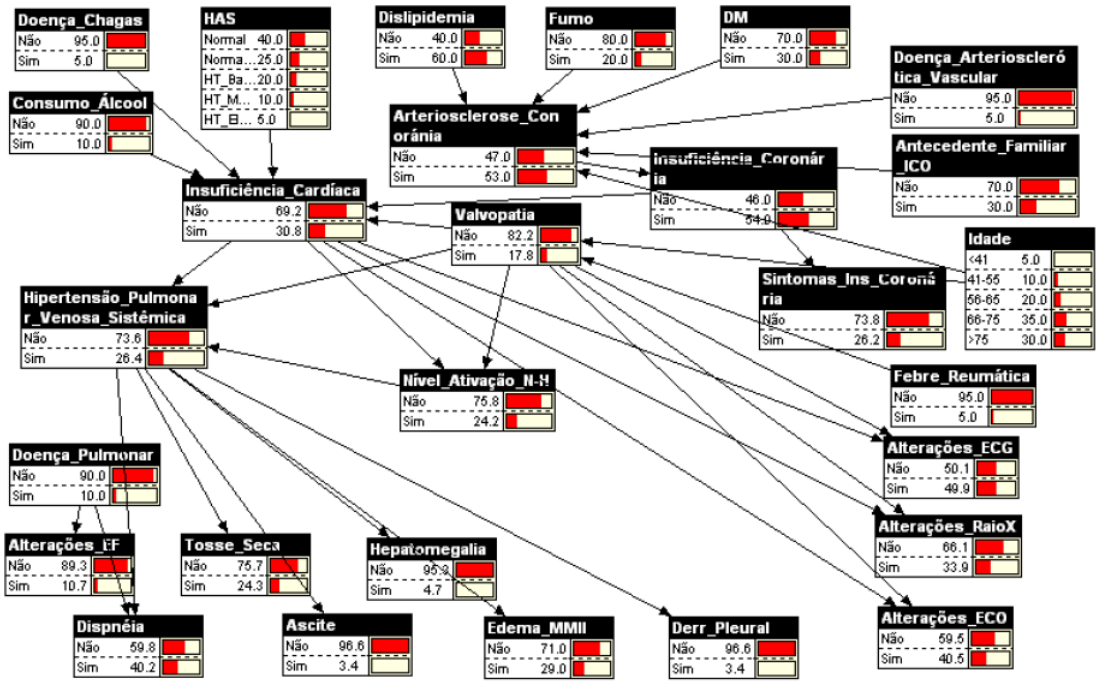

Figura 5 – Modelo Final da Rede Bayesiana

Fonte: Saheki (2005).

Para a definição das probabilidades de cada nó foi necessário a presença de um especialista , com base em seu conhecimento, para definir cada valor, incluindo a correção de

possíveis divergências entre os valores e em algumas situações as correções se utilizou o modelo Noisy-or.

A avaliação da rede foi divida em duas fases: na primeira os especialistas que auxiliariam em sua elaboração analisaram as probabilidades encontradas em dois testes propostos e na segunda cardiologistas que não participaram da elaboração da rede avaliaram a sua forma estrutural.

Na primeira fase algumas conclusões foram propostas relacionadas com o nó de maior importância na rede (o cardiomiopatia). Em situações de que a descrição do caso seja de alto nível é possível que o paciente detenha 94,9% de chance de portar cardiomiopatia, já em exames laboratoriais, como é o caso de ECG, Raio-X e ECO, o resultado seja positivo existe a probabilidade de 83,2% do paciente ter cardiomiopatia.

2.6.3 Diagnóstico de Alzheimer utilizando Redes Bayesianas

Na tese de Seixas (2012) e em seu artigo (SEIXAS et al., 2014) o seu objetivo principal é desenvolver um sistema de apoio a tomada de decisão médica para auxiliar o diagnóstico de demência, doença de alzheimer e transtorno cognitivo leve com seu principal foco na última enfermidade. Para a criação do sistema é utilizado as Redes Bayesianas.

Para aplicar a sua pesquisa o autor utilizou duas bases de dados a primeira da própria Universidade Federal do Rio de Janeiro do seu Instituto de Psiquiatria e a segunda de uma associação entre clínicas e laboratórios de pesquisas sobre doença de alzheimer (CERAD).

A fim de modelar a rede bayesiana foi utilizado a estrutura multinível na qual o primeiro nível está relacionado com os fatores de pré-disposição de determinada doença, o segundo a doença em si, e o terceiro nível os sintomas apresentados. A representação da rede bayesiana multinível é apresentada na Figura 6.

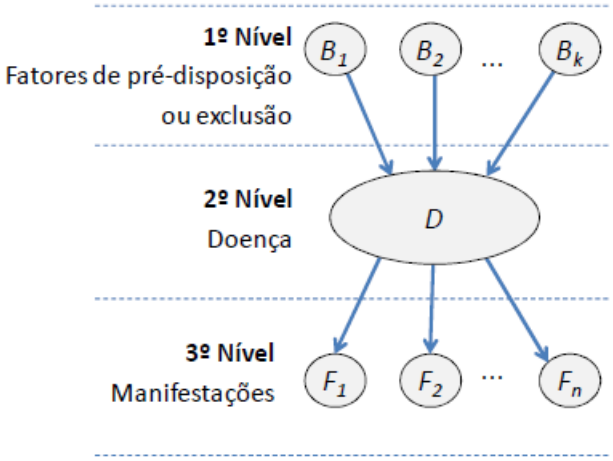

Figura 6 – Estrutura da Rede Bayesiana de Multinível

Fonte: Seixas et al. (2014).

No aprendizado da rede, a base de dados CERAD foi fragmentada em uma base de treinamento para encontrar variáveis correspondentes com as variáveis definidas no terceiro nível da rede, e em casos de apresentar mais de uma variável se utilizaria uma função de priorização. Em casos da variável ser númerico utilizou o metódo de discretização CDR (Clinical Dementia Rating).

As variáveis aleatórias presentes na rede bayesiana foram relacionadas com a medida de entropia para calcular o grau de incerteza presente no diagnótico. Somente após décima sétima variável aleatória (Teste de fluência verbal semântica) que foi possível diminuir o valor da entropia (0,001).

Capítulo 3

Experimentos e Análise dos Resultados

Este capítulo tem como objetivo descrever a base de dados utilizada, as principais etapas da metodologia do classificador para prever doadores regulares de sangue e, principalmente, os resultados atingidos. A metodologia do classificador, assim como a organização deste capítulo, está esboçado na Figura 7.

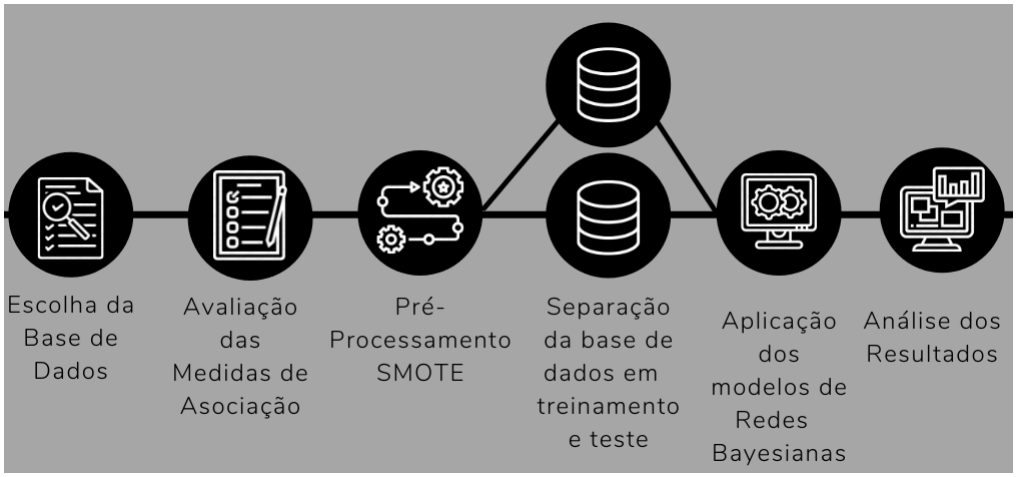

Figura 7 – Representação da estrutura lógica do classificador.

Fonte: Autora.

3.1 Base de Dados

Esse trabalho utilizou a base de dados denominada "Conjunto de dados do Centro de Serviços de Transfusão de Sangue" disponível no site *UCI Machine Learning Repository* (YEH, 2008). O conjunto de dados possui 748 doadores escolhidos de forma aleatória da base de dados do Centro de Serviços de Transfusão de Sangue da cidade de Hsin-Chu em Taiwan.

A base de dados é composta por cinco atributos, sendo eles:

- ❏ Recência: quantidade de meses desde a última doação;
- ❏ Frequência: total de doações realizadas;
- ❏ Quantidade: valor total de sangue doado em mililitros (ml);
- ❏ Tempo: quantidade de meses desde a primeira doação realizada;
- ❏ Classe: uma variável binária, definida como a classe, que indica se o doador doou ou não sangue na campanha realizada em março de 2007, sendo utilizado 0 para indicar que não doou sangue e 1 para indicar que doou sangue.

Para auxiliar na classificação final dessa base de dados e, da maneira semelhante à outros trabalhos da literatura que também utilizaram esta base (SANTHANAM; SUNDARAM, 2010; SILVA, 2018), seguiu as denominações:

- ❏ Novo Doador Voluntário (NDV): doador de sangue voluntário não remunerado, que nunca doou sangue antes;
- ❏ Doador Voluntário Irregular (DVI): doador de sangue voluntário não remunerado que doou sangue no passado, mas não cumpre os critérios de um doador regular;
- ❏ Doador Voluntário Regular (DVR): é um doador de sangue voluntário não remunerado que doa sangue regularmente, conforme uma frequência pré-definida.

As frequências pré-definidas para mensurar um DVR são descritas na Tabela 2 (YEH; YANG; TING, 2009; SANTHANAM; SUNDARAM, 2010) de acordo com os atributos contidos na base de dados. As mesmas medidas estabelecidas na Tabela 2 foram consideradas neste trabalho.

Tabela 2 – Critérios para definir um DVR.

Atributo	**Condição**	**Valores**
Recência	<=	6 meses
Frequência	>=	4 meses
Quantidade	>=	2000 ml
Tempo	>	24 meses

3.2 Avaliação das Medidas de Associação

Em busca de encontrar os atributos da base de dados mais preditivos para definir o perfil de um possível dodador de sangue, este trabalho baseou-se nas medidas de associação. Para avaliar e quantificar cada atributo, considerou o critério de doador voluntário regular descrito na Tabela 2.

O desfecho, ou seja, o evento de interesse da pesquisa, definido neste trabalho foi se a pessoa doou ou não doou sangue em Março de 2007. Já os fatores de risco foram os atributos: recência, frequência, quantidade e tempo. Desse modo, enumerou os doadores e não doadores de sangue para cada atributo da base de dados segundo o critério da Tabela 2. A Tabela 3 exibe os valores encontrados para a varíavel recência.

Tabela 3 – Variável Recência.

	Recência <= 6	**Recência > 6**	**Total**
Doou Sangue	137	41	178
Não Doou Sangue	230	340	570
Total	367	381	748

De acordo com a Tabela 3, gerou-se os valores do risco relativo e odds ratio para o atributo recência, como demonstrado pelos cálculos das equações 21 e 22.

$$RR = \frac{\frac{137}{137+41}}{\frac{230}{230+340}} = 1,907425501 \tag{21}$$

$$OR = \frac{\frac{137}{230}}{\frac{41}{340}} = 4,939554613 \tag{22}$$

Para o atributo frequência, a soma de doadores e não doadores segundo o critério DVR são apresentados na Tabela 4.

Tabela 4 – Variável Frequência.

	Frequência <= 4	**Frequência > 4**	**Total**
Doou Sangue	125	53	178
Não Doou Sangue	266	304	570
Total	391	357	748

Os valores do risco relativo e odds ratio do atributo frequência, foram calculados a partir dos números da Tabela 4, como pode ser conferido pelas equações 23 e 24.

$$RR = \frac{\frac{125}{125+53}}{\frac{266}{304+266}} = 1,504815409 \tag{23}$$

$$OR = \frac{\frac{125}{266}}{\frac{53}{304}} = 2,69541779 \tag{24}$$

Já, a Tabela 5 exibe a quantidade de doadores e não doadores em relação ao atributo quantidade.

Os valores das medidas risco relativo e odds ratio do atributo quantidade, foram estimados nas equações 25 e 26, segundo os dados apresentados na Tabela 5.

$$RR = \frac{\frac{63}{63+115}}{\frac{109}{461+109}} = 1,85084012 \tag{25}$$

Tabela 5 – Variável Quantidade.

	Quantidade >= 2000	**Quantidade < 2000**	**Total**
Doou Sangue	63	115	178
Não Doou Sangue	109	461	570
Total	172	576	748

$$OR = \frac{\frac{63}{109}}{\frac{115}{461}} = 2,316952533 \tag{26}$$

Para o atributo tempo, os valores obtidos na divisão entre doadores e não doares são demonstrados na Tabela 6.

Tabela 6 – Variável Tempo

	Tempo > 24	**Tempo <= 24**	**Total**
Doou Sangue	105	73	178
Não Doou Sangue	327	243	570
Total	432	316	748

Os valores do risco relativo e odds ratio para a variável tempo foram calculados pelas equações 27 e 28.

$$RR = \frac{\frac{105}{105+73}}{\frac{327}{243+327}} = 1,028244511 \tag{27}$$

$$OR = \frac{\frac{105}{327}}{\frac{73}{243}} = 1,068870177 \tag{28}$$

Conforme os valores das medidas de associação obtidos para cada atributo da base de dados, destacados na Tabela 7, foi possível calcular o intervalo de confiança e encontrar o limite inferior e superior de cada medida. Os intervalos de confiança estão descritos na Tabela 8 para a medida risco relativo e na Tabela 9 para a medida odds ratio.

Tabela 7 – Valores das Medidas de Associação das Variáveis

	Medidas de Associação	
	RR	**OR**
Recência	1,907425501	4,939554613
Frequência	1,504815409	2,69541779
Quantidade	1,85084012	2,316952533
Tempo	1,028244511	1,068870177

Segundo os dados apresentados nas Tabelas 8 e 9, foi possível identificar que todas as variáveis da base estão dentro do intervalo de confiança, o que expressa que todas as variáveis interferem no cenário para encontrar um possível doador de sangue regular.

Tabela 8 – Limite Inferior e Superior do Risco Relativo

	Risco Relativo			
	Recência	**Frequência**	**Quantidade**	**Tempo**
Limite Inferior	1.67801	1.17683	1.62929	0.83564
Limite Superior	2.16821	1.92421	2.10252	1.26525

Tabela 9 – Limite Inferior e Superior do Odds Ratio

	Odds Ratio			
	Recência	**Frequência**	**Quantidade**	**Tempo**
Limite Inferior	3.35453	1.87872	1.59811	0.75949
Limite Superior	7.2735	3.86715	3.35913	1.50428

3.3 Desbalanceamento na base de dados

Ao examinar a base de dados constatou o desbalanceamento entre as classes, ou seja, as instâncias definidas como doador que não doou sangue naquele mês foram 570 amostras e já aqueles que doaram foram 178 amostras.

Desta forma, em busca de suavizar este problema foi necessário aplicar a Técnica de Sobreamostragem de Minoria Sintética (SMOTE - *Synthetic Minority Oversampling Technique*). Essa função tem o objetivo de criar valores artificiais, isto é, duplicar exemplos da classe minoritária do conjunto de dados de treinamento antes de ajustar um modelo Bayesiano. Isso equilibra a distribuição de classes e não fornece nenhuma informação adicional ao modelo.´

Na aplicação dessa técnica no algoritmo desenvolvido utilizou-se a biblioteca **Imbalanced** que disponibiliza a função SMOTE().

3.4 Aplicação dos Modelos de Redes Bayesianas

Na implementação dos modelos computacionais propostos, foi utilizado a validação cruzada estratificada, definindo em 3, 5 e 10 o número de *folds*. Foram realizados testes com essas três possibilidades de *folds* visando encontrar qual a melhor divisão do conjunto de dados de treinamento e teste.

Os classificadores Bayesianos foram implementados pela linguagem de programação Python, os quais utilizaram recursos da linguagem para criar os algoritmos clássicos definidos na literatura do Naive Bayes e do TAN. Em cada classificador destaca uma particularidade:

- Naive Bayes Com a finalidade de preparar o modelo, dividiu-se em listas, os dados de acordo com a sua classe, ou seja, separou aqueles que pertencem a classe com o valor 0 daquelas com o valor 1. Em seguida, aplicou-se a média e o desvio padrão. Na sequência, calculou-se a probabilidade dos atributos, utilizando a distribuição

gaussiana, em relação a cada um dos valores da classe. Com isso, resultou prever a qual classe pertence um determinado conjunto de valores.

❏ Tree Augmented Naive Bayes

O algoritmo foi implementado com a ajuda da biblioteca **PGMPY**, desenvolvida para trabalhar com modelos gráficos probabilísticos, e de funções disponíveis pelas classes *BayesianModel* e *VariableElimination*. O grafo foi construído pela classe *Directed Acyclic Graph (DAG)*.

.

3.5 Resultados

A execução dos classificadores Naive Bayes e Tree Augmented Naive Bayes sob a base de dados em estudo, resultou nas representações gráficas, respectivamente, demonstrado pelas Figuras Figura 8 e Figura 9.

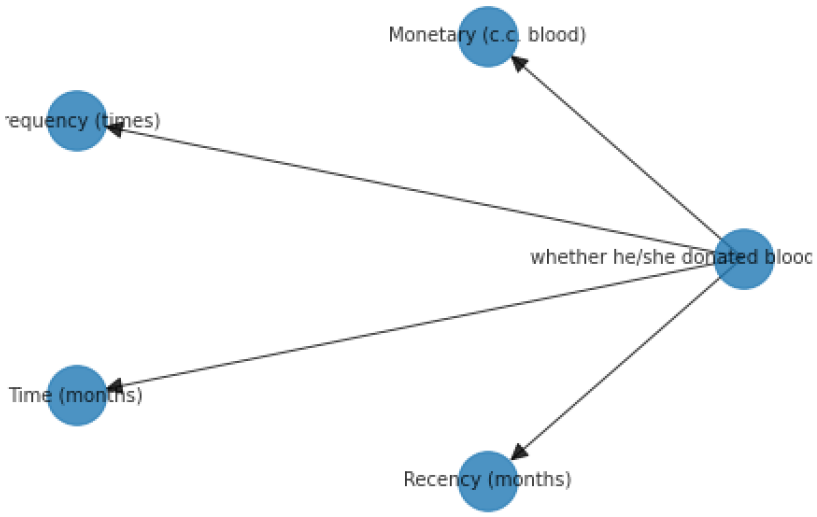

Figura 8 – Representação do Classificador Naive Bayes gerado sob a base de dados em estudo.

Fonte: Autora.

A Figura 8 refere-se ao grafo do classificador Naive Bayes. Nota-se, que o nó classe é o único que possui arestas direcionadas para os demais nós, confirmando que nesse classificador as relações, ou dependências, entre cada variável é desprezada.

A Figura 9 é referente ao grafo do classificador TAN. Nota-se a correlação entre cada variável e a sua importância para o cenário. O nó recência é considerado a raiz da árvore,

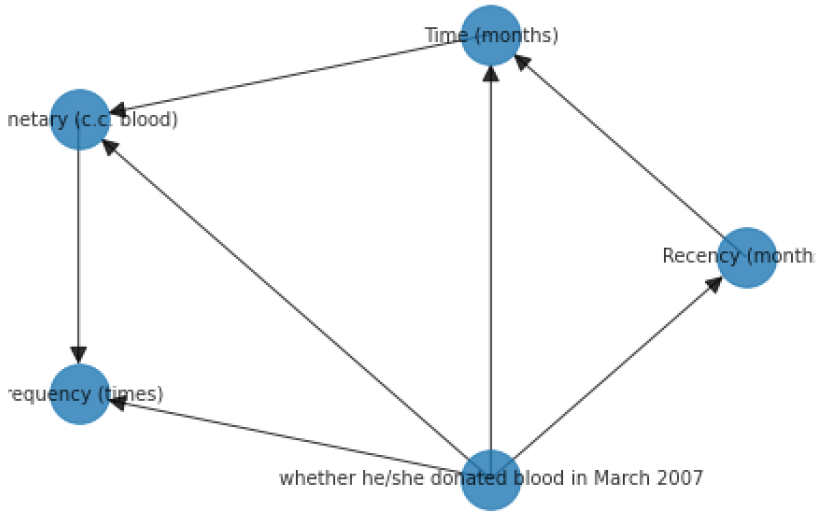

Figura 9 – Representação do Classificador Tree Augmented Naive Bayes gerado sob a base de dados em estudo.

Fonte: Autora.

portanto ele interfere na probabilidade dos demais nós. Também é importante destacar a estrutura da árvore se mantém mesmo depois de excluir o nó classe.

A Tabela 10 expõe os resultados das medidas acurácia, precisão e sensibilidade encontrados para cada modelo Bayesiano. Ademais, nesta tabela há a presença de três diferentes valores de *k-fold*, com o propósito de alcançar os melhores resultados nas divisões dos conjuntos de treinamento e teste sem repetição das amostras.

Tabela 10 – Valores de Acurácia, Precisão e Sensibilidade

K-Fold	**Rede Bayesiana**	**Acurácia**	**Precisão**	**Sensibilidade**
3	NB	50.0%	83.3%	33.3%
	TAN	58.4%	72.4%	36.1%
5	NB	50.0%	90.0%	20.0%
	TAN	59.0%	69.8%	40.4%
10	NB	50.0%	95.0%	10.0%
	TAN	58.3%	69.6%	39.3%

A medida acurácia é importante para quantificar a frequência com que a classificação foi realizada corretamente, sendo assim é observado na Tabela 10 que o TAN obteve melhores valores de acurácia do que o NB para os diferentes k-fold. Esses valores reforçam que a melhoria proposta pelo algoritmo TAN comparado ao Naive Bayes interfere positivamente nos resultados finais do algoritmo. No entanto, o desbalanceamento da base de dados é o motivador do percentual baixo da acurácia, necessitando de outras medidas para validar a exatidão das redes Bayesianas.

A medida precisão demonstra melhores porcentagens de acertos do que a sensibilidade, pois a quantidade de FP (falsos positivos) foi bem menor que a quantidade de FN (falsos negativos), isto é, a quantidade de amostras que classificou não doadores como doadores foi inferior a quantidade que classificou doadores como não doadores. Isto justifica-se pela desproporção entre as amostras, dado que quantidade de doadores é menor que a de não-doadores.

Os percentuais acima de 69% de acertos da precisão demonstra que as amostras classificadas à classe doadores, efetivamente, pertencem à estas classes. Nesta medida, o *k-fold* igual a 10 apresentou valores mais significativos.

Capítulo 4

Conclusão

A alta demanda na procura por transfusões de sangue e a falta de doações frequentes motivam os hemocentros a procurarem em suas bases de dados prováveis doadores. Este trabalho com o objetivo de encontrar melhorias na metodologia de busca desses doadores, aplicou o modelo probabilistico dos classificadores Bayesianos, destacando as tipologias Naive Bayes e Tree Augmented Naive Bayes. Primeiramente, os dados foram analisados pelas medidas de associação risco relativo e odds ratio, para identificar se os atributos interferem na busca por doadores regulares. Na sequência, os dados foram manipulados visando melhor balanceamento entre as classes doadores e não doares de sangue. Posteriormente, as redes Bayesianas foram executadas e a classificação gerada foi avaliada pelas medidas acurácia, precisão e sensibilidade.

As redes Bayesianas são modelos capazes de demonstrar conhecimentos de maneira simplificada através dos grafos, tornando o cenário estudado entendível para qualquer tipo de profissional, dado que a maioria do referencial teórico deste trabalho é composto pela área médica. Nas análises dos experimentos desenvolvidos verificou-se que o TAN apresenta melhor desempenho comparado ao NB, dado que os valores obtidos pela medida acurácia foram maiores que 58%.

As dificuldades encontradas na manipulação da base de dados no desenvolvimento do trabalho estão relacionados à insuficiência de informações e ao desbalanceamento das classes. Devido a esses fatores, utilizou as técnicas SMOTE e k-fold estratificado. Os resultados das medidas de avaliação foram abaixo da expectativa, dado que a proposta do trabalho é provar que as redes bayesianas são capazes de predizer a classificação em estudo. Assim, esperava-se resultados mais consideráveis.

Por se tratar de um assunto inserido num ambiente que carece de soluções emergentes, esperava-se encontrar diversidade em trabalhos internacionais que analisassem bases de dados contendo informações sobre doadores de sangue ou estudos envolvidos na busca por doadores regulares.

Para trabalhos futuros, sugere-se buscar outra base de dados com mais atributos, junto aos Hemocentros vinculados às instituições de ensino superior, contendo informações

capazes de validar as melhorias que o modelo TAN propõe, além de aplicar outros métodos de classificação para comparação. Atrelado a uma nova base de dados, definir um perfil específico para o doador regular.

Referências

ARA-SOUZA, A. L. **REDES BAYESIANAS: UMA INTRODUÇÃO APLICADA A CREDIT SCORING**. Dissertação (Mestrado) — Universidade Federal de São Carlos, São Carlos, Julho 2010. Citado na página 14.

CARLESSO, L. et al. Doador de sangue habitual e fidelizado: fatores motivacionais de adesão ao programa. **Revista Brasileira em Promoção da Saúde**, v. 30, n. 2, p. 213–220, 2017. Citado na página 10.

CUNHA, B. G. F. da; DIAS, M. R. Doador de sangue habitual e fidelizado: fatores motivacionais de adesão ao programa. **Cadernos de Saúde Pública**, v. 6, n. 24, p. 1407–1418, 2008. Citado na página 10.

FRANCO, L. J.; PASSOS, A. D. C. **Fundamentos de Epidemiologia**. Brasil: Manole Ltda, 2011. Citado 3 vezes nas páginas 17, 18 e 19.

KAPLAN, A.; HAENLEIN, M. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. **Business Horizons**, v. 62, p. 15–25, 2018. Citado na página 10.

KORB, K. B.; NICHOLSON, A. E. **Bayesian Artificial Intelligence**. London, UK: Chapman Hall/CRC, 2011. Citado na página 10.

LOBO, L. C. Inteligência artificial e medicina. **Revista Brasileira de Educação Médica**, v. 41, p. 185–193, 2017. Citado na página 11.

MOURA, A. S. de et al. Doador de sangue habitual e fidelizado: fatores motivacionais de adesão ao programa. **Revista Brasileira em Promoção da Saúde**, v. 19, n. 2, p. 61–67, 2006. Citado na página 10.

PADMANABAN, H. Comparative analysis of naive bayes and tree augmented naïve bayes models. 2014. Citado na página 15.

RAY, S. **6 passos fáceis para aprender o algoritmo Naive Bayes (com o código em Python)**. [S.l.], 2016. Disponível em: <https://www.vooo.pro/insights/6-passos-faceis-para-aprender-o-algoritmo-naive-bayes-com-o-codigo-em-python/>. Acesso em: 10.11.2019. Citado na página 15.

SAHEKI, A. H. **Construção de uma Rede Bayesiana Aplicada ao Diagnóstico de Doenças Cardíacas**. Dissertação (Dissertação de Mestrado) — Escola Politécnica da Universidade de São Paulo, 2005. Citado na página 22.

SANTHANAM, T.; SUNDARAM, S. Application of cart algorithm in blood donors classification. **Journal of Computer Science**, v. 6, n. 5, p. 548–552, 2010. Citado na página 26.

SATO, R. C.; SATO, G. T. K. Modelos probabilísticos gráficos aplicados à identificação de doenças. **Revista Einstein**, v. 13, n. 2, p. 330–333, 2015. Citado na página 11.

SCHEUNEMANN, F. **Mineração de Dados para Descoberta de Conhecimento na Área de Oncologia**. Dissertação (Mestrado) — Centro de Tecnologia da Informação do Centro Universitário UNIVATES, Lajeado, Dezembro 2015. Citado na página 15.

SEIXAS, F. L. **Sistema de Apoio à Decisão aplicado ao Diagnóstico de Demência, Doença de Alzheimer e Transtorno Cognitivo Leve**. Dissertação (Tese de Doutorado) — Universidade Federal Fluminense, 2012. Citado na página 23.

SEIXAS, F. L. et al. A bayesian network decision model for supporting the diagnosis of dementia, alzheimer's disease and mild cognitive impairment. **Computersin Biology and Medicine**, v. 51, p. 140–158, 2014. Citado na página 23.

SILVA, F. H. **Estudo e desenvolvimento de métodos para predição de doadores de sangue**. Dissertação (Dissertação de Mestrado) — Universidade Federal de Goías, 2018. Citado 4 vezes nas páginas 10, 11, 21 e 26.

SOCIEDADE BRASILEIRA DE COMPUTAÇÃO. : Análise de classificadores para avaliação automática em fóruns educacionais. Recife, 2016. 12 p. Citado na página 15.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introdução ao Data Mining Mineração de Dados**. Brasil: Ciência Moderna, 2009. Citado na página 11.

WAGNER, M. B.; CALLEGARI-JACQUES, S. M. Medidas de associação em estudos epidemiológicos : risco relativo e odds ratio. **Jornal de Pediatria**, v. 74, p. 247–251, 1998. Citado na página 16.

YEH, I.-C. **Blood Transfusion Service Center Data Set**. Hsin Chu, Taiwan, 2008. Disponível em: <https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center>. Acesso em: 19.05.2021. Citado 2 vezes nas páginas 11 e 25.

YEH, I.-C.; YANG, K.-J.; TING, T.-M. Knowledge discovery on rfm model using bernoulli sequence. **Expert Systems with Applications**, v. 36, n. 3, p. 5866–5871, 2009. Citado na página 26.

ZHANG, H.; JIANG, L.; SU, J. Hidden naive bayes. **Proceeding AAAI'05 Proceedings of the 20th national conference on Artificial intelligence**, v. 2, p. 919–924, 2005. Citado na página 15.