Bayesian networks for blood donor prediction

Fernanda Maria C. Santos[1], **Cristina Zayra de N. Romani**[2]

[1]Faculdade de Computação (FACOM) – Universidade Federal de Uberlândia (UFU)
Monte Carmelo – MG – Brasil

[2]Faculdade de Ciências Agrárias e Veterinárias – Universidade Estadual Paulista (UNESP)
Jaboticabal – SP – Brasil

fmcsantos@ufu.br,criszayra@gmail.com

Abstract. *Blood centers are responsible for managing blood stocks so that they satisfy at a considerable level, in addition to guaranteeing a quality standard with the blood collected. Both factors are possible if there is a control of regular donors. Thus, this article proposes a computational model that predicts regular blood donors, whose methodology joins the results of association measures to determine the most likely characteristics of a donor, with the Naive Bayes algorithm. The proposed model presented results superior to 68% accuracy and 73% precision in predicting a regular blood donor.*

Resumo. *Os hemocentros são responsáveis por administrarem estoques de sangue para que permaneçam em um nível considerável, além de garantir um padrão de qualidade com o sangue coletado. Ambos fatores são possíveis se existisse um controle dos doadores regulares. Assim, esse artigo propõe um modelo computacional que prediz os doadores de sangue regulares, cuja metodologia une os resultados das medidas de associação para determinar as características mais prováveis de um doador, com o algoritmo Naive Bayes. O modelo proposto apresentou resultados superiores à 68% de acurácia e 73% de precisão na predição de um doador de sangue regular.*

1. Introdução

Os hemocentros são instituições responsáveis pela gestão e o controle de bolsas de sangue para atender possíveis casos de transfusão de sangue para as unidades terapêuticas. Em busca de manter os estoques num nível suficiente para atender as demandas e, principalmente, de garantir a progressão do padrão de qualidade do sangue coletado e transfundido, os hemocentros precisam receber doações de sangue frequentemente. Esse cenário enfrenta impasses como a baixa frequência de pessoas fidelizadas à doação e o aumento de doadores de última hora, os quais não são ideais porque não se pode acompanhar o seu estado de saúde [Barboza and Costa 2014].

Com o intuito de resolver essa adversidade, os hemocentros são responsáveis por buscar a fidelização de doadores de sangue ou pelos doadores de repetição, de modo que possam ser monitorados quanto ao estado de sua saúde, seus hábitos e a sua satisfação em relação ao serviço prestado. Desta forma, possibilitaria aos hemocentros um maior fornecimento de dados para a tomada de ações que possibilitem um número cada vez maior de doadores [Giacomini and Lunardi Filho 2010].

Assim, os hemocentros necessitam definir métodos de busca em sua base de dados que sugestione doadores potenciais. Essa pesquisa é uma avaliação sob diversos parâmetros, a qual poderia ser aplicado algoritmos de aprendizado de máquina (AM) para que de forma eficiente e congruente faça a previsão de futuros doadores de sangue [Silva 2018].

O uso de Inteligência Artificial na Medicina começou na década de sessenta sendo aplicado no diagnóstico de doenças [Korb and Nicholson 2011]. Ao longo do tempo, as ferramentas computacionais inteligentes tornaram-se cada vez mais presentes na área médica, capazes de identificar nas imagens digitais as lesões na pele e classificá-las, além, de também atuarem eficientemente na predição do diagnóstico de doenças [Lobo 2017] e de eventos em estudos epidemiológicos. Similarmente, em estudos envolvendo Hemocentros e seus doadores, encontram-se trabalhos como [T and Shyam 2010] que fez uso da Árvore de Decisão para identifica o comportamento de doadores, assim como [Patil et al. 2015] que aplicou *Support Vector Machine* em dados coletados pelos doadores de sangue antes e após a doação para abstrair e identificar as diferentes reações dos dadores e as medidas preventivas contra as mesmas. No estudo feito em [Alajrami et al. 2019] utilizou de arquiteturas de Redes Neurais Artificiais (RNA) para prever se uma pessoa registrada na base de dados de doadores de sangue da cidade Hsin-Chu, Tailândia, seria um doador potencial. Também, no artigo [Boonyanusith and Jittamai 2012] aplicou as técnicas de RNA e Árvore de Decisão, a fim de prever a partir de uma série de dados comportamentais individuais se pessoas tornariam doadores de sangue, independentemente de já serem ou não doadoras.

O objetivo primordial deste artigo é analisar as Redes Bayesianas, em especial o algoritmo Naive Bayes, na predição de potenciais doadores de sangue. Para isso, será quantificado o grau de influência dos atributos presentes na base de dados em estudo pelas medidas de associação para que, posteriormente, provenham dados relevantes às Redes Bayesianas.

Assim sendo, o artigo está estruturado da seguinte maneira. Na seção 2 é apresentado o referencial teórico sobre as redes bayesianas e as medidas de associação. Na seção 3 está descrito as etapas do modelo computacional proposto, a base de dados e os critérios para medir a performance do modelo. Na seção 4 são apresentados os resultados e suas análises. Por fim, na seção 5 destaca a conclusão e os trabalhos futuros.

2. Fundamentação Teórica

2.1. Redes Bayesianas

As redes bayesianas são modelos gráficos capazes de representar o domínio em estudo e as incertezas envolvendo suas variáveis. Sua representação é pelo modelo matemático conhecido como grafo direcionado acíclico.

O grafo é formado por nós que representam as variáveis utilizadas no domínio e arcos que conectam cada nó são os responsáveis por demonstrar a interação probabilística entre cada variável. As variáveis que originam os arcos são chamadas de pais e as que recebem os arcos são chamadas de filhos.

A representação da rede bayesiana é definida através da relação dos 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:

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

2.2. Naive Bayes

O Naive Bayes é uma técnica baseada no Teorema de Bayes e pode ser visto como um caso especial de Redes Bayesianas. Ele delimita variáveis de acordo com a classe e agrupa elementos que possuem as mesmas características levando em consideração que não é possível ligações entre as demais variáveis do domínio em estudo.

O modelo computacional proposto neste artigo implementou três diferentes tipos de Naive Bayes. São eles: *Gaussian Naive Bayes, Multinomial Naive Bayes* e *Complement Naive Bayes*.

2.3. Medidas de Associação

As medidas de associação são utilizadas em estudos epidemiológicos e são capazes de quantificar o grau de interferência de uma variável em um determinado ambiente e fator [Wagner and Callegari-Jacques 1998]. Duas medidas serão abordadas neste artigo, o risco relativo e *odds ratio*.

O risco relativo (RR) é considerado uma medida de associação em cortes, e é definido como a razão da incidência dos casos entre os expostos e a incidência de casos entre os não expostos [Franco and Passos 2011]. Faz referência de uma análise de um grupo específico baseado num determinado período de tempo, e definem o grau de exposição de um fator, ou de um hábito ou de uma condição sobre uma doença estipulada.

O *odds ratio* (OR) é uma medida de caso-controle que define a probabilidade da ocorrência de uma doença em dois tipos de grupos os que foram expostos ao fator determinante para causar doença e os que não foram expostos. Segundo [Franco and Passos 2011], nos estudos de caso-controle a amostra é dividida em grupos de casos e grupos de controles.

Nos grupos de casos os indivíduos possuem de fato a doença e sua análise é baseada na razão daqueles que foram expostos ou não ao fator determinante. Nos grupos de controle, os indivíduos não possuem a doença e sua análise é baseada na razão daqueles que foram expostos ou não ao fator determinante. Portanto, através dessas duas situações, é possível determinar o cálculo de OR pela razão entre o grupo de casos pelo grupo de controle.

As medidas de associação baseadas em razões, como o risco relativo e o *odds ratio*, fornecem dados sobre a associação entre o fator de risco e o desfecho, permitindo um entendimento sobre uma relação de causalidade. Sob a temática em estudo, os fatores de risco seriam os atributos da base de dados, e o desfecho seria a identificação de um doador promissor.

3. Etapas do Modelo Computacional para Predição de Doadores

As etapas da implementação do modelo computacional proposto para predizer se uma pessoa já cadastrada na base de dados será um doador regular no período seguinte, está ilustrado na Figura 1.

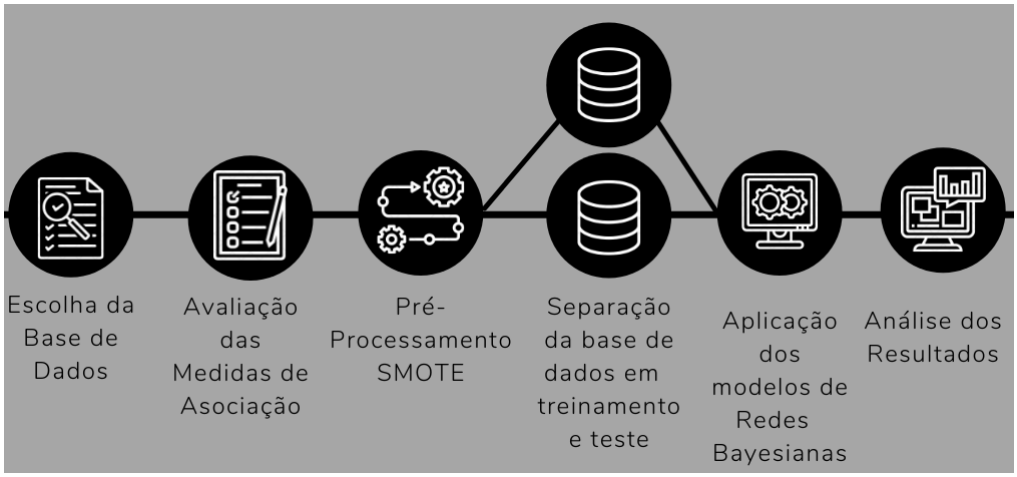


Figura 1. Representação das etapas do modelo computacional.
Fonte: Autora.

Nas subseções a seguir, há uma descrição mais detalhada de cada etapa da metodologia exibida na Figura 1: características da base de dados, parametrização das medidas de associação RR e OD para os atributos da base de dados e, na sequência, os critérios matemáticos para medir a performance dos algoritmos Bayesianos.

3.1. Base de Dados Experimental

A base de dados em estudo é denominada ”Conjunto de dados do Centro de Serviços de Transfusão de Sangue”, disponibilizada no repositório UCI Machine Learning Repository do professor I-Cheng Yeh. Os dados totalizaram 748 doadores de sangue do Centro de Serviços de Transfusão de Sangue da cidade de Hsin-Chu em Taiwan.

A base de dados é formada por cinco variáveis, sendo elas:

- Recência: quantidade de meses desde a última doação;
- Frequência: total de doações realizadas;
- Quantidade: valor total de sangue doado em ml;
- Tempo: quantidade de meses desde a primeira doação realizada;
- E uma variável binária que indica se o doador doou ou não sangue na campanha realizada em março de 2007 (os possíveis valores são 0 para indicar que não doou sangue e 1 para indicar que doou sangue).

3.2. Medidas de Associação

Em busca de encontrar quais os parâmetros são relevantes para prever um possível doador, estimou os valores das medidas de associação para cada atributo da base de dados. Os valores de referência para parametrizar os atributos foram retirados do artigo [T and Shyam 2010] que definiu a classe **Doadores Voluntários Regulares**. A partir desta parametrização, quantificou-se quem seria um doador de sangue regular e quem não seria um doador de sangue regular. Veja a Tabela 1.

Com os valores da Tabela 1, foi possível encontrar as medidas de associação risco relativo e *odds ratio*, os quais são descritos na Tabela 2.

Tabela 1. Quantificação dos atributos da base de dados entre doadores e não doadores de sangue regular.

	Doador de Sangue Regular	**Doador de Sangue Não Regular**
Recência <= 6	137	230
Recência > 6	41	340
Frequência >= 3	125	266
Frequência < 3	53	304
Quantidade >= 2.000	63	109
Quantidade < 2.000	115	461
Tempo > 24	105	327
Tempo < 24	0	243

Tabela 2. Valores das medidas de associação de RR e OR

	Risco Relativo	***Odds Ratio***
Recência	3,4689	4,9395
Frequência	2,1534	2,6954
Quantidade	1,8346	2,3169
Tempo	–	–

Pela análise feita com as medidas da Tabela 2 conclui que apenas as variáveis de recência, frequência e quantidade de sangue doado interferem para na predição de uma pessoa ser doadora voluntária regular.

3.3. Critérios de Medidas de Performance

A precisão dos algoritmos Bayesianos propostos serão avaliados segundo as métricas aritméticas acurácia, precisão, sensibilidade (tradução da palavra *recall*) e F1 score. As métricas são definidas pelos valores que compõem a matriz de confusão, que são: VP (Verdadeiro Positivo), VN (Verdadeiro Negativo), FP (Falso positivo) e FN (Falso Negativo).

4. Experimentos e Análise dos Resultados

A linguagem de programação utilizada neste trabalho foi o Python [1], devido a grande quantidade de bibliotecas disponíveis, sendo a principal utilizada a Scikit-learn, para a definição dos tipos de Redes Bayesianas.

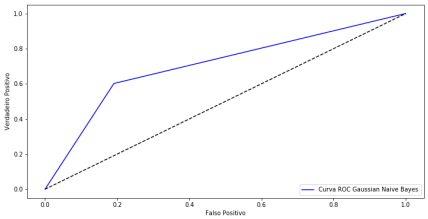
Para a realização do experimento foi considerado três atributos da base de dados que interferem na escolha de um possível doador, que são a recência, a frequência e a quantidade de sangue doado, além da classe que determina se aconteceu ou não a doação na campanha em questão. A base de dados foi dividida em conjunto de teste e em conjunto de treinamento, na qual a divisão foi de 20% e 80%, respectivamente.

Os conjuntos de treinamento e de teste foram testados para os três tipos de classificadores Naive Bayes: *Gaussian, Multinomial* e *Complement*. Após a execução do conjunto de teste, calculou-se as métricas de avaliação para cada um dos classificadores, e os resultados foram apresentados na Tabela 3.

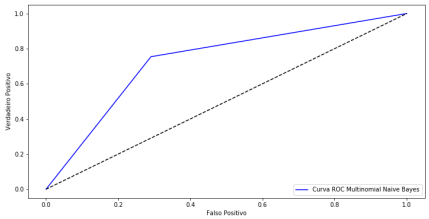
[1] https://www.python.org/

Tabela 3. Métricas de desempenho dos classificadores Naive Bayes.

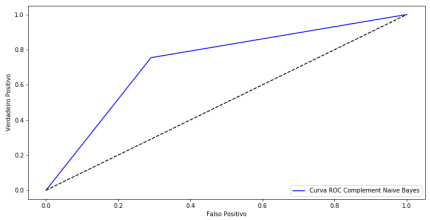
	Gaussian NB	Multinomial NB	Complement NB
Acurácia	0.60%	0.68%	0.68%
Precisão	0.77%	0.73%	0.73%
Sensibilidade	0.60%	0.75%	0.75%
F1 Score	0.69%	0.73%	0.73%



(a) Naive Bayes Gaussian.



(b) Naive Bayes Multinomial.



(c) Naive Bayes Complement.

Figura 2. Curva ROC dos resultados obtidos com o três tipos de Naive Bayes.

Conclui-se pelos resultados obtidos na Tabela 3, que os valores do Multinomial e *Complement* são idênticos, e ao comparar com os valores resultantes da *Guassian* as métricas acurácia, sensibilidade e F1 Score são melhores. Em adição as métricas de avaliação, as curvas ROC dos três classificadores Naive Bayes foram criados, como pode ser visto na Figura 2. Semelhantemente às outras métricas, as curvas ROC resultantes do Multinomial e do *Complement* são idênticas, porém todas as três curvas qualificam por uma maior taxa de verdadeiros positivos e uma menor taxa de falsos positivos.

5. Conclusão e Trabalhos Futuros

Os Hemocentros buscam manter os estoques de sangue num nível suficiente, além de sempre desejar a progressão do padrão de qualidade do sangue coletado e transfundido, através da presença de doadores regulares. Para isso, definiu um modelo computacional constituído pelo algoritmo Naive Bayes para predizer potenciais doadores de sangue.

Os cálculos das medidas de associação RR e OR definiram os atributos adequados da base de dados que estimularam os modelos de Naive Bayes a identificar os doadores apropriados. A base de dados utilizada foi essencial para desenvolvimento deste estudo e de outras pesquisas, dado a importância de tais dados para a área de saúde. Entretanto, poderia existir mais bases de dados públicas com informações sobre doadores de sangue, com o intuito de aprimorar estudos e, consequentemente, melhorar a qualidade de serviços

emmpregados aos hemocentros.

Como trabalho futuro pretende-se aplicar outras técnicas de AM para comparar com os resultados obtidos neste artigo, além de encontrar uma base de dados com maior quantidade de atributos para que seja possível identificar o perfil completo do doador de sangue regular.

Referências

Alajrami, E., Abu-Nasser, B. S., Khalil, A. J., Musleh, M. M., Barhoom, A. M., and Naser, S. S. A. (2019). Blood donation prediction using artificial neural network. *International Journal of Academic Engineering Research (IJAER)*, 3:1–7.

Barboza, S. I. S. and Costa, F. J. A. d. (2014). Marketing social para doação de sangue: análise da predisposição de novos doadores. *Cadernos de SaÃPÃ*, 30:1463 – 1474.

Boonyanusith, W. and Jittamai, P. (2012). Blood donor classification using neural network and decision tree techniques. In *World Congress on Engineering and Computer Science 2012 (WCECS 2012)*.

Franco, L. J. and Passos, A. D. C. (2011). *Fundamentos de Epidemiologia*. Manole Ltda., Brasil.

Giacomini, L. and Lunardi Filho, W. D. (2010). Estratégias para fidelização de doadores de sangue voluntários e habituais. *Acta Paulista de Enfermagem*, 23(Acta paul. enferm., 2010 23(1)):65–72.

Korb, K. B. and Nicholson, A. E. (2011). *Bayesian Artificial Intelligence*. Chapman Hall/CRC, London, UK.

Lobo, L. C. (2017). Inteligência artificial e medicina. *Revista Brasileira de Educação Médica*, 41:185–193.

Patil, R., Poi, M., Pawar, P., Patil, T., and Ghuse, N. (2015). Blood donor's safety using data mining. In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pages 500–505.

Silva, F. H. (2018). Estudo e desenvolvimento de métodos para predição de doadores de sangue. Dissertação de mestrado, Universidade Federal de Goías.

T, S. and Shyam, S. (2010). Application of cart algorithm in blood donors classification. *Journal of Computer Science*, 6.

Wagner, M. B. and Callegari-Jacques, S. M. (1998). Medidas de associação em estudos epidemiológicos : risco relativo e odds ratio. *Jornal de Pediatria*, 74:247–251.