Comparação dos Métodos Fuzzy Triangular Naive Bayes e Random Forest para Tomada de Decisão

Marciele da Silva[1], Alberlene Baracho[2], Tayna Bernardino Gomes[3], Tiago Mota[2]

[1]PPMDS-CCS- Universidade Federal da Paraíba (UFPB)
58051-900 – João Pessoa – PB – Brasil

[2]PPGCR - CE – Universidade Federal da Paraíba (UFPB)
58.051-900 – João Pessoa – PB – Brasil

[3]PPGFIS - Universidade Federal da Paraíba (UFPB)
58.015-020 – João Pessoa – PB – Brasil

marcieledelsilva@gmail.com, alberlenebaracho@hotmail.com,
tayna.gomes@academico.ufpb.br, tiagomotacr@gmail.com

Abstract. *This study aims to compare the Fuzzy Triangular Naive Bayes and Random Forest methods if they fit the Hepatitis database through the tests performed. This is a descriptive, exploratory study of comparison between the methods, on the data provided by the UCI with the Hepatitis database. With the tests carried out, the result of the Random Forest method was presented as the most accurate and effective method, presenting excellent results, in addition to having demonstrated ease and speed in its execution using this bank. It is concluded that the Fuzzy Triangular Naive Bayes method was not suitable for this base.*

Resumo. *Este estudo tem como objetivo comparar os métodos Fuzzy Triangular Naive Bayes e Random Forest se eles se adequam a base de dados de Hepatites através dos testes realizados. Trata-se de um estudo descritivo, exploratório de comparação entre os métodos, sobre os dados fornecidos pelo UCI com o banco de dados de Hepatites. Com os testes realizados, o resultado do método Random Forest se apresentou como método mais preciso e eficaz, apresentando resultados excelentes, além de ter demostrado facilidade e rapidez na sua execução usando esse banco. Conclui-se que o método Fuzzy Triangular Naive Bayes não se adequou para essa base.*

1. Introdução

A tomada de decisão tem o objetivo de minimizar ou prevenir um problema, se apresenta como um processo onde devemos escolher uma ação entre tantas outras possíveis. Sendo fundamental a escolha de informações adequadas, dados precisos e completos, baseados em critérios científicos. Desse modo, a prioridade será sempre analisar as alternativas apresentadas para propor a mais viável (MARTINS; COELHO, 2014).

Nesse sentido, a realização da tomada de decisão em alguns métodos podendo ser usados nos sistemas de avaliação podem ser baseados em diversas abordagens, como modelos de lógica clássica e lógica fuzzy, modelos de aprendizado de máquina, entre outros (MACHADO; MORAES, 2010).

No entanto, esses métodos de avaliação costumam mostrar melhor resultado quando aplicado a certos tipos de dados. Quando a distribuição estatística dos dados é conhecida, podemos usar um método que foi desenvolvido baseado nessa distribuição estatística e esses métodos costumam apresentar melhor performance (FERREIRA et al. 2015).

Então, Zadeh (1965), estabeleceu um novo conceito de conjuntos, conceituado como Conjuntos Fuzzy, onde os seus elementos não formam um conjunto matemático. A aplicação desse conjunto vem da incerteza acerca dos elementos que o compõe e da necessidade de uma quantificação dessa incerteza que é dado por um grau de pertinência.

Logo o método Fuzzy Triangular Naive Bayes foi criado tendo como base a distribuição estatística triangular e foi estruturado em cima das redes bayesianas do tipo Naive Bayes, sendo capaz de avaliar dados provenientes de eventos fuzzy (MORAES et al. 2020). Esse método foi testado no software R.

A distribuição triangular é definida por três parâmetros: um valor mínimo, um valor máximo e um valor mais provável. Essa distribuição pode ser utilizada para modelar diversos tipos de aplicações e costuma ser usada em situações onde se identifica a ausência de dados detalhados (Jannat; Greenwood 2012).

Diante do exposto, este artigo tem como objetivo comparar os métodos Fuzzy Triangular Naive Bayes e Random Forest se eles se adequam a base de dados Hepatitis através de testes realizados nos softwares R e Weka.

2. Método

Trata-se de um estudo descritivo, exploratório de comparação entre os métodos Fuzzy Triangular Naive Bayes e Random Forest, sobre os dados fornecidos pelo banco de dados Hepatitis, disponível em um repositório internacional de banco de dados que pode ser acessado por meio do endereço eletrônico: https://archive.ics.uci.edu/ml/datasets/Hepatitis.

A Hepatitis é uma base de dados médica que contém informações clínicas de um grupo de pessoas e seus diagnósticos acerca da patologia hepatite, onde as pessoas possuem ou não a doença. Nesse sentido, na base de dados, 86 casos não possuem a patologia e 69 casos a apresentam (ALSHAMRANI; OSMAN, 2017).

A base de dados possui 155 instâncias no total, consistindo em 80 instâncias após a eliminação dos casos vazios, 19 atributos em que 13 são binários e 6 são discretos. Os atributos são categorizados em duas classes, sendo essas: morrer e viver. Existem 13 atributos na classe morrer e 67 na classe viver. Os atributos são demonstrados na Quadro 01.

Quadro 01. Atributos da base de dados.

ATRIBUTOS DA BASE DE DADOS HEPATITIS
1. IDADE: 10, 20, 30, 40, 50, 60, 70, 80
2. SEXO: masculino, feminino
3. ESTERÓIDE: não, sim
4. ANTIVIRAIS: não, sim
5. FADIGA: não, sim
6. MAL-ESTAR: não, sim
7. ANOREXIA: não, sim
8. FÍGADO AUMENTADO: não, sim
9. CONSISTÊNCIA DO FÍGADO: não, sim
10. ESPLENOMEGALIA: não, sim
11. VASOS DILATADOS: não, sim
12 ASCITE: não, sim
13. VARIZES: não, sim
14. BILIRRUBINA: 0,39, 0,80, 1,20, 2,00, 3,00, 4,00
15. FOSFATASE ALCALINA: 33, 80, 120, 160, 200, 250
16. SGOT: 13, 100, 200, 300, 400, 500,
17. ALBUMINA: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
18. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
19. HISTOLOGIA: não, sim

Fonte: Base de dados Hepatitis, 1988.

Assim, os dados foram analisados mediante os parâmetros de Classificação Correta, Coeficiente Kappa e número de erros e acertos trazidos nas Matrizes de Confusão.

De acordo com Landis e Koch (1977), O coeficiente Kappa é largamente utilizado em estudos de confiabilidade, apresentando-se como uma medida que testa o grau de concordância, confiabilidade e precisão na classificação dos dados. Esse coeficiente é demonstrado por meio da porcentagem de acerto, traduzindo a capacidade do algoritmo classificar de maneira correta as instâncias em relação ao número total. Quanto mais próximo de 1 for seu valor, maior será a concordância analisada. a fórmula do coeficiente Kappa é dada pela equação (5), onde Po é a proporção observada de concordâncias (soma das respostas concordantes dividida pelo total) e Pe é a proporção esperada de concordâncias (soma dos valores esperados das respostas concordantes dividida pelo total).

$$k = \frac{p_o - p_e}{1 - p_e}$$

Ainda o coeficiente Kappa pode ser interpretado de acordo com as informações apresentadas na Tabela 01.

Tabela 1. Interpretação do Coeficiente Kappa.

Coeficiente Kappa	Grau de Concordância
< 0.0	Pobre
0.00 |- 0.20	Leve
0.20 |- 0.40	Bom
0.40 |- 0.60	Moderado
0.60 |- 0.80	Considerável
0.80 |- 1.00	Quase perfeito

Fonte: LANDIS; KOCH, 1977.

A Acurácia (ACUR) mede a taxa de classificação correta global, sendo representada por:

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

Enfim, a distribuição triangular para uma variável aleatória X que se encontra em a ≤ X ≤ b pode ser escrita da forma da Equação (Forbes et al. 2011).

$$P(X|a,b,c) = \begin{cases} \frac{2(X-a)}{(b-a)(c-a)}, & \text{se } a \leq X \leq c \\ \frac{2(b-X)}{(b-a)(b-c)}, & \text{se } c \leq X \leq b \\ 0, & \text{caso contrário} \end{cases}$$

Onde a, b, c ∈ IR, c ∈ [a, b] e a e um parâmetro que corresponde ao limite inferior, b e um parâmetro que corresponde ao limite superior e c e o parâmetro de forma equivalente a moda. Essa distribuição inicia no valor mínimo, aumenta linearmente até o pico na moda, e em seguida diminui linearmente até o valor máximo. Vale ressaltar que essa distribuição não e necessariamente simétrica em relação ao ponto modal c.

Assim, a rede Fuzzy Triangular Naive Bayes é dada pela equação (Moraes et al., 2020).

$$P(w_i|X) = \begin{cases} P(w_i) \times \prod_{k=1}^{n} \left(\frac{2(X_k - a_i)}{(b_i - a_i)(c_i - a_i)} \times \mu_i(X_k) \right), & \text{se } a_i \leq X \leq c_i \\ P(w_i) \times \prod_{k=1}^{n} \left(\frac{2(b_i - X_k)}{(b_i - a_i)(b_i - c_i)} \times \mu_i(X_k) \right), & \text{se } c_i \leq X \leq b_i \end{cases}$$

As funções de pertinência μi(Xk) podem ser estimadas usando histogramas dos dados de treinamento. A melhor estimação para a classe de performance do vetor de dados X e obtido a partir dos valores mais altos da função gf. Então, a regra de avaliação para a FTriangNB e dada pela Equação (Moraes et al., 2020).

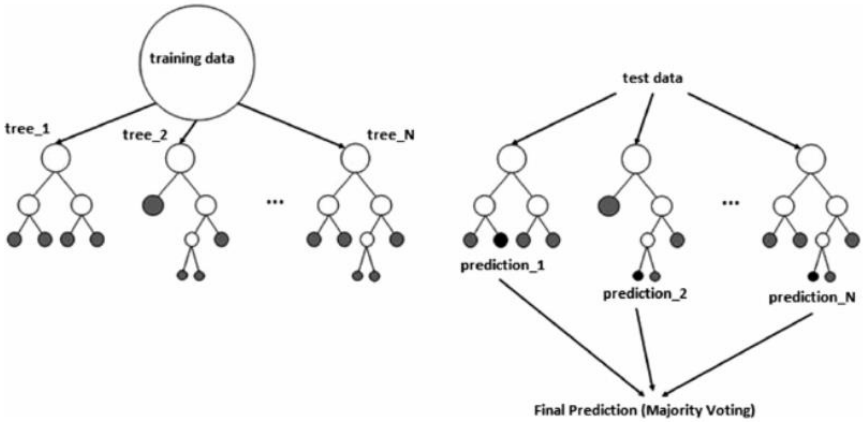
$$X \in w_i \text{ se } g_f(w_i, X) > g_f(w_j, X),$$

Para todo i 6= j ∈ Ω e as funções gf.

$$g_f(w_i, X) = \begin{cases} log[P(w_i)] + 2\sum_{k=1}^{n} log(X_k - a_i) - \\ -n[log(b_i - a_i) + log(c_i - a_i)] + \\ +\sum_{k=1}^{n} log[\mu_i(X_k)], \qquad \text{se } a_i \leq X \leq c_i \\ \\ log[P(w_i)] + 2\sum_{k=1}^{n} log(b_i - X_k) - \\ -n[log(b_i - a_i) + log(b_i - c_i)] + \\ +\sum_{k=1}^{n} log[\mu_i(X_k)], \qquad \text{se } c_i \leq X \leq b_i \end{cases}$$

Por esse viés, outro método para tomada de decisão foi usado nesse trabalho o Random Forest usado no software Weka. O Random Forest se originou das arvores de decisão, se apresentando como um método rápido e robusto. É uma versão estendida do algoritmo de ensacamento com aleatoriedade propriedade injetada. Esse método divide cada nó em ramos usando o melhor de variáveis selecionadas aleatoriamente em cada nó em vez de usar a melhor ramificação através de todas as variáveis. A seleção aleatória de recursos é usada para estender árvores. As árvores estendidas não são podadas. Essa estratégia torna esse método ainda mais preciso, podendo conter qualquer quantidade de árvores (LEITE; DE MORAES; LOPES, 2021).

Dessa maneira, o algoritmo Random Forest requer dois parâmetros para definir antes de iniciar. O primeiro um é o número de variáveis (m) usadas em cada nó para determinar a divisão. A segunda é o número de árvores (N) a serem estendidas. Random Forest usa o algoritmo CART (Classification and Regression Tree) para produzir árvores. Em cada nó, as ramificações são geradas de acordo com o GINI índice de Critérios do algoritmo CART. O índice GINI mede a classe homogeneidade. Se o índice GINI aumenta, a heterogeneidade de classe também aumentará. Se o índice GINI diminuir, a classe a homogeneidade aumentará. Então, quando todas as árvores (N árvores) são geradas, a classe da entrada é determinada segundo todas as previsões da árvore como mostrado (LEITE; DE MORAES; LOPES, 2021). Sendo assim, o algoritmo de classificação de floresta aleatória.



O meu banco de dados usado foi o da Hepatitis que foi doado em 1988 pela UCI Machine Learning Repository, sendo de grande importância, contribuindo para uma melhor tomada de decisão com a finalidade de salvar vidas.

3. Resultados e discussão

O software usado nesse estudo para aplicação do método Fuzzy Triangular Naive Bayes, foi o Sistema R que é um software de uso estatístico porque sua execução depende de comandos e funções oriundas de equações econométricas, esse sistema propõe testes de hipóteses e análise de dados, sejam eles constituídos por padrões numéricos ou nominais. Possui um pacote de dados para funções específicas, o que não impede a inclusão de novos pacotes ou alteração de funções para melhor adequação às necessidades do pesquisador ou do estudo. Também se adequa como ferramenta de apoio aos testes, fortalecendo ou refutando as inferências elaboradas pelo pesquisador (DE ANDRADE, 2018). Nesse sentido, a imagem 1 abaixo, mostra um dos resultados realizados com modelo Fuzzy Triangular Naive Bayes no software R, sendo que foram realizados 15 testes.

Imagem 1. Um dos resultados realizados com modelo Fuzzy Triangular Naive Bayes no software R.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2
         1  1 10
         2  4 10

               Accuracy : 0.44
                 95% CI : (0.244, 0.6507)
    No Information Rate : 0.8
    P-Value [Acc > NIR] : 1.0000

                  Kappa : -0.2069

 Mcnemar's Test P-Value : 0.1814

            Sensitivity : 0.20000
            Specificity : 0.50000
         Pos Pred Value : 0.09091
         Neg Pred Value : 0.71429
             Prevalence : 0.20000
         Detection Rate : 0.04000
   Detection Prevalence : 0.44000
```

Fonte: Dados da pesquisa, 2023.

Como mostra a imagem acima, com um dos testes que foram realizados no software R utilizando como base o pacote do FuzzyClass, com o método Fuzzy Triangular Naive Bayes, porém o banco não se adequou, resultando assim em resultados não sendo satisfatórios.

O outro software utilizado foi o Weka (Waikato Environment for Knowledge Analysis), versão 3.8.5, para aplicação do método Random Forest. O Weka desenvolvido na Universidade de Waikato na Nova Zelândia, sendo escrito em Java, está disponível gratuitamente, um sistema de referência, tendo em vista que fornece vários algoritmos diferentes de modelos de decisão e regressão (BOUCKAERT et al., 2016). Para realizar os testes no Weka foram utilizados os parâmetros cross-validation (muito usado na avaliação de desempenho de modelos de aprendizado de máquinas), Folds (conjunto de dados é dividido de acordo com o número de dobras especificado) e percentage split (são usadas porcentagens definidas pelo pesquisador e traz os valores de porcentagens para treinamento e para testes).

Já para realizar o teste no Weka, baixei o arquivo em Arff pela internet, ao baixar abri esse arquivo diretamente no Weka, o método escolhido por mim foi o Random Forest para ser testado, também fiz ajustes nos parâmetros cross-validation, percentage split e Folds, o primeiro resultado foi positivo, o segundo não tão bom e o terceiro ruim. A imagem 2 apresenta o melhor resultado realizado com método Random Forest pelo teste no Weka.

Quadro 2. Resultados dos testes com Folds usando Random Forest.

FOLDS	KAPPA
9	0.4012
10	0.6419
12	0.5181

Fonte: Dados da pesquisa, 2023.

Quadro 3. Resultados dos testes com Split usando Random Forest.

SPLIT	KAPPA
66	0.5181
70	0.5106
72	0.5075

Fonte: Dados da pesquisa, 2023.

De acordo com os testes realizados, o coeficiente Kappa apresentou classificação moderado para quase todos os resultados. O maior valor do Kappa foi observado no parâmetro percentage split com 66% e 70%. No entanto, estes resultados podem não refletir no melhor resultado encontrado, porém estes podem ser resultados de um ajuste gerado pelo modelo utilizando essa base de dados. Esses resultados quando comparados ao Fuzzy Triangular Naive Bayes, apresentaram diferença quando comparados os valores do Kappa, demonstrou uma performance melhor.

4. Considerações finais

Com esse trabalho, conclui-se que o método Fuzzy Triangular Naive Bayes não se adequa a base de dados Hepatitis através de testes realizados no software R. Destaca-se que o algoritmo Random Forest se apresenta como um método melhor, tendo em vista que os testes executados apresentaram excelentes resultados, além de ter demonstrado facilidade e rapidez na sua execução. Por fim, a base de dados desse estudo pode ser usada em várias áreas do conhecimento assim, se faz necessário novos estudos com a finalidade de melhorar a tomada de decisão para trazer novos resultados.

Referências

Knuth, D. E. (1984), The TeXbook, Addison Wesley, 15th edition.

Smith, A. and Jones, B. (1999). On the complexity of computing. In *Advances in Computer Science*, pages 555–566. Publishing Press. ALSHAMRANI BS, Osman AH (2017) Investigação da doença da hepatite diagnóstico usando diferentes tipos de algoritmos de redes neurais. Int J Comput Sci Netw Mód (IJCSNS) 17: 242.

BOUCKAERT, R. R. et al. WEKA manual for version 3-8-1. University of Waikato, New Zealand, (2016).

DE ANDRADE, Mariana Dionísio. A utilização do sistema R-studio e da jurimetria como ferramentas complementares à pesquisa jurídica. **Revista Quaestio Iuris**, v. 11, n. 2, p. 680-692, (2018).

FERREIRA, J. A., SOARES, E. A., MACHADO, L. S., e MORAES, R. M. (2015). Assessment of fuzzy gaussian naive bayes for classification tasks. **PATTERNS 2015**, page 73.

FORBES, C., EVANS, M., HASTINSGS, N., e PEACOCK, B. (2011). Statistical distributions. John Wiley & Sons.

JANNAT, S. e GREENWOOD, A. G. (2012). Estimating parameters of the triangular distribution using nonstandard information. In Proceedings of the Winter Simulation Conference, pages 1–2.

LEITE, Danilo Rangel Arruda; DE MORAES, Ronei Marcos; LOPES, Leonardo Wanderley. Método de Aprendizagem de Máquina para Classificação da intensidade do desvio vocal utilizando Random Forest. **Journal of Health Informatics**, v. 12, (2021).

MACHADO, L. S. e MORAES, R. M. (2010). Intelligent decision making in training based on virtual reality. In **Computational intelligence in complex decision systems**, pages 85–123. Springer.

MARTINS, F. G.; COELHO, L. S. Aplicação do método de análise hierárquica do processo para o planejamento de ordens de manutenção em dutovias. **Revista GEPROS**, n. 1, p. 65, (2014).

MORAES, R. M., SILVA, I. L. A., e MACHADO, L. S. (2020). Online skills assessment in training based on virtual reality using a novel fuzzy triangular naive bayes network. In **The 14th International FLINS Conference on Robotics and Artificial Intelligence and the 15th International Conference on Intelligent Systems and Knowledge Engineering (FLINS 2020)**. No prelo.

MORAES, R.M.; FERREIRA, J.A.; MACHADO, L.S. A New Bayesian Network Based on Gaussian Naive Bayes with Fuzzy Parameters for Training Assessment in Virtual Simulators. Int. J. Fuzzy Syst, (2020).

ZADEH, L. A. Fuzzy sets. Information and control, Elsevier, v. 8, n. 3, p. 338–353, (1965).