AVALIAÇÃO DO MÉTODO NAIVE BAYES GAUSSIANA NA CLASSIFICAÇÃO DE AMOSTRAS DA BASE DE DADOS BREAST TISSUE

Lucilla Vieira Carneiro [1]
Vitória Polliany de Oliveira Silva [2]
Ronei Marcos de Moraes [3]

RESUMO

Naive Bayes (NB) é um método classificador probabilístico e pode ser denominado como algoritmos bayesianos. Estes mostram-se independentes entre si, pois cada atributo possui sua informação e acaba não influenciando nas informações dos demais. Nesse sentido, destaca-se a relevância desse método, sendo considerado um dos que apresentam mais precisão e eficiência. Diante do exposto, este artigo tem como objetivo avaliar se o método Naive Bayes Gaussiano se adequa a base de dados Breast Tissue através de testes realizados no software Weka. Trata-se de um estudo do tipo exploratório-descritivo, com abordagem quantitativa, onde foi realizada a aplicação do método Naive Bayes Gaussiano (Naive Bayes Updateable) através do Weka com a utilização de parâmetros de cross-validation (folds) e percentagem split sobre os dados fornecidos pelo banco de dados Breast Tissue. De acordo com os testes realizados, o que apresentou melhor resultado foi o cross-validation/10 folds com valor de Kappa = 0,9204 e 93.3962% de precisão de acertos. Observou-se que o método demonstrou uma ótima performance, por isso, se apresenta como algoritmo confiável para este tipo de análise. Conclui-se que o método Naive Bayes Gaussiano se adequa a base de dados Breast Tissue e que o algoritmo Naive Bayes Updateable se apresenta como um ótimo classificador de dados. Assim, torna-se possível predizer que se trata de um modelo apropriado para a detecção de alterações pré-malignas no tecido mamário.

Palavras-Chave: Redes Bayesianas. Naive Bayes Gaussiano. Base de dados Breast Tissue.

INTRODUÇÃO

Atualmente, várias abordagens utilizadas para a análise de dados têm ganhado importância no cenário mundial. Os tipos de dados e distribuição podem exigir uma abordagem específica, tendo em vista que além de extrair informação útil é imprescindível fazê-lo de forma eficiente (MORAES *et al.*, 2020).

[1] Doutoranda do Curso de Modelos de Decisão e Saúde, da Universidade Federal da Paraíba - UFPB, lucilla.vc@hotmail.com;
[2] Mestranda do Curso de Modelos de Decisão e Saúde, da Universidade Federal da Paraíba - UFPB, vitoriapolliany1@gmail.com;
[3] Professor orientador: Doutor em Computação Aplicada, Docente da Graduação e Pós-graduação no Centro de Ciências Exatas e da Natureza da Universidade Federal da Paraíba - UFPB, ronei@de.ufpb.br;

A tomada de decisão é um processo difícil onde se deve escolher uma ação entre tantas outras possíveis, visando minimizar, mitigar ou prevenir um problema. Para isso é importante que haja uma escolha de informações adequadas, dados precisos e completos, baseados em critérios científicos. Assim, mesmo podendo sofrer influência pela qualidade da informação ou recursos disponíveis, é imprescindível analisar criteriosamente as alternativas apresentadas para propor a mais viável (MARTINS; COELHO, 2014).

Naive Bayes (NB) é um método classificador probabilístico, que adveio do teorema de Bayes (Thomas Bayes), e pode ser denominado como algoritmos bayesianos. Estes mostram-se independentes entre si, pois cada atributo possui sua informação e acaba não influenciando nas informações dos demais. Os algoritmos bayesianos são considerados como um dos mais precisos e eficientes, pois dado um conjunto onde este algoritmo é formado por grupos desmembrados entre atributos e valores, é possível identificar qual grupo pertence uma nova instância (NEAPOLITAN, 2003).

Nesse sentido, destaca-se a relevância desse método, sendo considerado um dos que apresentam mais precisão e eficiência. A principal diferença entre os distintos classificadores do método Naive Bayes é que o Gaussiano considera como gaussiana a distribuição, ou seja, assume a normalidade dos dados, enquanto o classificador Naive Bayes Multinomial considera que os dados são distribuídos multinomialmente (SHIKHA, 2019).

Desse modo, o método Naive Bayes proporciona alto desempenho, pois como dito anteriormente, o algoritmo considera os atributos independente, que se torna uma vantagem, quando utilizado para trabalhar com dados incompletos e inconsistentes. Por isso, o método fornece resultados promissores, a partir de um conjunto disponível, mas em relação ao Gaussiano, o processo utiliza algumas distribuições de probabilidade para estimar a classe de um ponto de inferência Bayesiana (ONTIVERO-ORTEGA *et al.*, 2017).

Um estudo realizado por Alam e Pachauri (2017), fez uma comparação entre o Naive Bayes e outros algoritmos. O referido trabalho concluiu que este algoritmo, apesar de simplificado, proporciona melhores resultados em situações complexas da nossa realidade. Além disso, apresenta como principal vantagem o fato de que o Naive

Bayes utiliza um quantitativo menor de dados de treinamento para estimar os parâmetros.

Destarte, a rede NB tem sido bastante utilizada pelos pesquisadores em diversas áreas como, por exemplo, estimação de risco operacional, diagnóstico médico, projeto de jogos computacionais, imputação de dados, entre outras. O método apresenta diversas características vantajosas, tais como: facilidade na construção de seu algoritmo; processo de classificação eficaz quando os atributos são independentes entre si; rapidez na aprendizagem e predição; lida com dados reais e contínuos, possui habilidade de gerenciar incertezas. Entretanto, como desvantagem do método, destaca-se que os parâmetros por serem amplos podem influenciar no desempenho, caso o número da amostra não seja representativo; e a independência pode modificar os resultados finais (ONTIVERO-ORTEGA et al, 2017).

Nesse ínterim, o banco de dados Breast Tissue foi doado e disponibilizado em 2010 pela UCI Machine Learning Repository, sendo de grande importância, tendo em vista que, de acordo com Silva, Marques de Sá e Jossinet (2000), realiza medições de impedância elétrica de amostras de tecido recém-excisado da mama para detecção de câncer. Ressalta-se que utilizando métodos de aprendizado de máquina é possível diagnosticar precocemente algumas doenças como o câncer de mama, um tipo de neoplasia que acomete milhares de pessoas em todo o mundo. Desse modo, estes métodos contribuem para uma tomada de decisão assertiva na perspectiva de salvar vidas.

Assim, configura-se como problema a ser superado com base nos resultados desta pesquisa, a identificação da adequação do método Naive Bayes Gaussiano à análise dos dados do banco de dados Breast Tissue, na perspectiva de que estes resultados contribuam para a tomada de decisão que auxilie no diagnóstico precoce de câncer de mama. Portanto, é imprescindível uma tomada de decisão baseada em métodos científicos a partir de dados e/ou informações que permitem extrair conhecimento de informações armazenadas.

Diante do exposto, este artigo tem como objetivo avaliar se o método Naive Bayes Gaussiano se adequa a base de dados Breast Tissue através de testes realizados no software Weka.

MÉTODOLOGIA

Trata-se de um estudo do tipo exploratório-descritivo, com abordagem quantitativa, onde foi realizada da aplicação do método Naive Bayes Gaussiana sobre os dados fornecidos pelo banco de dados Breast Tissue, disponível em um repositório internacional de banco de dados que pode ser acessado por meio do link: https://archive.ics.uci.edu/ml/datasets/Breast+Tissue. Este banco de dados tem 106 instâncias, 10 atributos e 6 classes, conforme apresentado na Tabela 01.

Os dados foram obtidos a partir da espectroscopia de impedância elétrica realizada em tecido mamário recentemente excisado de 64 mulheres com idade entre 18 e 72 anos de idade, onde foi possível classificar os tecidos e realizar a detecção de câncer de mama (JOSSINET, 1996).

Segundo Silva, Marques de Sá e Jossinet (2000), a espectroscopia de impedância elétrica é uma técnica minimamente invasiva que tem vantagens claras para a caracterização de tecidos vivos devido ao seu baixo custo e facilidade de uso. A análise estatística é usada para derivar um conjunto de regras com base em recursos extraídos da representação gráfica de espectros de impedância elétrica. Essas regras são usadas hierarquicamente para discriminar várias classes de tecido mamário, podendo detectar alterações pré-malignas.

Tabela 01 - Atributos da base de dados

ATRIBUTOS DA BASE DE DADOS BREAST TISSUE
1. I0: Impedividade (ohm) em frequência zero
2. PA500: ângulo de fase em 500 KHz
3. HFS: inclinação de alta frequência do ângulo de fase
4. DA: Distância de impedância entre as extremidades espectrais
5. Área sob o espectro
6. A / DA área normalizada por DA
7. MAX IP: máximo do espectro
8. DR: distância entre I0 e parte real da frequência máxima
9. P: comprimento da curva espectral
10. Classes: carcinoma (car), fibro-adenoma (fad), mastopatia (mas), tecido adiposo (adi), tecido conjuntivo (con) e glândula mamária (gla)

Fonte: Base de dados Breast Tissue, 2010.

O Teorema de Bayes apresenta a seguinte relação, dada uma classe Y em vetor de fatores X. Seja X um vetor de dados de treinamento, onde X é um vetor com n características distintas X = (X_1, X_2, ..., X_n) de acordo com a amostra de dados, Y é a classe de desempenho no espaço de decisão para o vetor X (MITCHELL, 2017):

$$P(y \mid x_1,...,x_N) = \frac{P(y)P(x_1,...,x_N \mid y)}{P(x_1,...,x_N)} \quad (1)$$

Segundo Moraes *et al.* (2012), a rede Naive Bayes (NB) usa a hipótese ‘‘ingênua’’ que cada variável X é condicionalmente independente de qualquer outra variável.

Assim, de acordo com a equações 1, sendo Y uma variável aleatória e X seja um vetor, tal formação irá definir o valor de P que se refere a probabilidade (MITCHELL, 2017).

Dito isso, ressalta-se que diversas abordagens foram desenvolvidas para usar o Método Naive Bayes com variáveis, uma vez que vários métodos de discretização foram utilizados no primeiro estágio para permitir o uso do método posteriormente. No entanto, esta forma pode afetar o viés de classificação e variância do mesmo (MITCHELL, 2017).

A rede NB pode ser modificada para permitir o uso contínuo de variáveis aleatórias no vetor X, como a distribuição Gaussiana. Atrelado a esse fato, destaca-se o método Gaussiano para a distribuição de X e para calcular seus parâmetros, ou seja, o vetor médio e matriz de covariância (MORAES *et al*., 2020).

De acordo com Moraes *et al.* (2020), diante da equação (2) e usando alguns cálculos matemáticos simplificados, é possível reduzir a complexidade computacional dessa equação. Assumindo a premissa de que X segue uma variável n com distribuição gaussiana, com vetor médio μ_j e matriz de covariância $\sum j$, com j = 1; . . .; n, para a classe de desempenho w_i com i € Ω, então:

$$\log [P(w_i \mid X_1, X_2, \ldots, X_n)] = \log [(1/S)\, P(w_i)\, \Pi^{n}_{k=1} \; P(X_k \setminus w_i)]$$
$$= \log (1/S) + \log P(w_i) + \sum^{n}_{k=1} \; \log[P(X_k \setminus w_i)] \quad (2)$$

Como S é um fator de escala, não é necessário computar na regra de classificação para o método Naive Bayes Gaussiano (MORAES *et al*., 2020), então:

$$X \in w_i \quad \text{if} \quad \{\log P(w_i) + \sum^{n}_{k=1} \log[P(X_k \setminus w_i)]\} > \{\log P(w_j) + \sum^{n}_{k=1} \log[P(X_k \setminus w_j)]\} \quad \text{for all } i \neq j \text{ and } i, j \in \Omega \tag{3}$$

Com base no mesmo espaço de decisão com classes M, um método Gaussiano calcula probabilidades de classe condicional e, em seguida, prevê a mais provável de um vetor de dados de treinamento X, de acordo com dados de amostra D. Os parâmetros do método são atrelados com os dados e as probabilidades condicionais e estimadas usando a equação (2) e a decisão final sobre o vetor de dados de treinamento X é feita pela equação (3) (MORAES *et al*., 2020).

O software utilizado para aplicação do método Naive Bayes Gaussiano sobre o banco de dados foi o Weka (Waikato Environment for Knowledge Analysis), versão 3.8.5, desenvolvido na Universidade de Waikato na Nova Zelândia, sendo escrito em Java. De acordo com Bouckaert *et al.* (2016), o software Weka está disponível para acesso de forma gratuita, além de ser considerado um sistema de referência, tendo em vista que fornece vários algoritmos diferentes de modelos de decisão e regressão.

Ademais, o Weka faz relatórios com dados analíticos e estatísticos além de ser uma ferramenta com maior simplicidade e facilidade de uso sem perder funcionalidades relevantes para a mineração de dados (BOUCKAERT *et al*., 2016).

Para realizar os testes no Weka foram utilizados os parâmetros "cross-validation" ("Folds") e "percentage split". Segundo Witten; Frank; Hall (2011), o cross-validation é uma técnica muito utilizada para avaliação de desempenho de modelos de aprendizado de máquina. Durante a classificação, o conjunto de dados é dividido de acordo com o número de dobras (folds) especificado. Essas dobras correspondem a fração do conjunto que será utilizada para teste, enquanto o restante será aplicado ao treinamento.

O parâmetro percentage split utiliza porcentagens definidas pelo pesquisador e traz os valores de porcentagens para treinamento e para testes, onde os dados são previamente reordenados aleatoriamente (BOUCKAERT *et al*., 2016).

Assim, foram realizados 20 testes, sendo 10 testes utilizando o parâmetro cross-validation e 10 testes com percentage split. Os testes do cross-validation, foram realizados utilizando 10, 15, 20, 25, e 30 folds. Os testes com o parâmetro percentage

split foram realizados empregando os seguintes valores: 66%, 70%, 75%, 80% e 85%. Os dados foram analisados mediante os parâmetros de Instâncias Corretas (Exatidão Global), Coeficiente Kappa e número de erros e acertos trazidos nas Matrizes de Confusão.

O coeficiente Kappa é largamente utilizado em várias áreas do conhecimento, apresentando uma medida de concordância na classificação dos dados em relação a um padrão de referência. O referido 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. O coeficiente pondera aspectos relevantes presentes nos dados, sendo necessário o uso de outras abordagens que complementem a análise de uma maneira geral (LANDIS; KOCH, 1977).

Segundo Landis e Koch (1977), a fórmula do coeficiente Kappa é dada pela equação (4), 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).

$$k = \frac{p_o - p_e}{1 - p_e} \tag{4}$$

Além disso, conforme a nomenclatura Landis e Koch (1997), o coeficiente Kappa pode ser interpretado de acordo com as informações apresentadas na Tabela 02.

Tabela 02. 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.

RESULTADOS E DISCUSSÃO

Os resultados obtidos a partir dos testes foram organizados nas tabelas 3 e 4 de acordo com o pior e melhor resultado do método Naive Bayes Gaussiano (Naive Bayes Updateable) testando diferentes parâmetros para cross-validation e percentage split. Na avaliação do algoritmo Naive Bayes Updateable foram utilizados cinco parâmetros do cross-validation (folds) e cinco do percentage split. Para cada grupo foram informados o valor do Kappa encontrado seguido da classificação correspondente.

Tabela 3. Resultados dos testes com cross-validation (Naive Bayes Updateable).

FOLDS	COEFICIENTE KAPPA	CLASSIFICAÇÃO DO KAPPA
10	0,9204	Quase Perfeito
15	0,9204	Quase Perfeito
20	0,909	Quase Perfeito
25	0,8977	Quase Perfeito
30	0,909	Quase Perfeito

Fonte: Dados da pesquisa, 2021.

Tabela 4. Resultados dos testes com Percentage Split (Naive Bayes Updateable).

SPLIT	COEFICIENTE KAPPA	CLASSIFICAÇÃO DO KAPPA
66	0,9664	Quase Perfeito
70	1	Quase Perfeito
75	0,9528	Quase Perfeito
80	1	Quase Perfeito
85	0,92	Quase Perfeito

Fonte: Dados da pesquisa, 2021.

De acordo com os testes realizados, o coeficiente Kappa apresentou classificação quase perfeito para todos os resultados. O maior valor do Kappa foi observado no parâmetro percentage split com 70% e 80%, os quais utilizaram respectivamente, 32 e 21 instâncias para teste obtendo o valor Kappa 1 e a classificação quase perfeito. No entanto, estes podem não refletir o melhor resultado encontrado, tendo em vista que o percentage split não leva em consideração o quantitativo total de instâncias para teste e isso compromete um pouco o modelo. Ademais, podem ser resultados de um sobreajuste gerado pelo modelo utilizando essa base de dados, a qual possui um banco de dados considerado pequeno.

Desse modo, pode-se afirmar que o melhor resultado obtido foi para as simulações que utilizaram o cross-validation/10 folds onde foi encontrado o valor do Kappa = 0,9204, tendo em vista que esse parâmetro não deixa de considerar as instâncias, só vai dividir as informações.

Na matriz de confusão do melhor resultado para o algoritmo Naive Bayes Updateable, apresentado na Figura 1, é possível observar que o modelo é adequado, onde foi detectado na sua diagonal principal 93.3962% de precisão de acertos, utilizando o parâmetro cross-validation/10 folds.

Figura 1. Matriz de confusão Naive Bayes Updateable (cross-validation/10 folds).

```
=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
 21  0  0  0  0  0 |  a = car
  0 15  0  0  0  0 |  b = fad
  1  1 15  1  0  0 |  c = mas
  0  0  1 15  0  0 |  d = gla
  0  0  0  0 13  1 |  e = con
  0  0  0  0  2 20 |  f = adi
```

Fonte: Dados da pesquisa, 2021.

CONCLUSÕES

A partir desse trabalho, conclui-se que o método Naive Bayes Gaussiano se adequa a base de dados Breast Tissue através de testes realizados no software Weka. Destaca-se que o algoritmo Naive Bayes Updateable se apresenta como um ótimo classificador de dados, haja vista que os testes executados apresentaram excelentes resultados, além de ter demonstrado celeridade na execução. Assim, torna-se possível predizer que se trata de um modelo apropriado para a detecção de alterações pré-malignas no tecido mamário.

Destarte, o conjunto de dados utilizado neste estudo é de uma base de dados da saúde e, portanto, recomendam-se novos estudos para que essa técnica possa ser empregada em várias áreas do conhecimento, em especial na área médica em situações diagnósticas e no processo de tomada de decisão, uma vez que o algoritmo em questão traz confiabilidade em seus resultados.

REFERÊNCIAS

ALAM, F.; PACHAURI, S. Comparative Study of J48, Naive Bayes and One-R Classification Technique for Credit Card Fraud Detection using WEKA. **Advances in Computational Sciences and Technology.** v.10, n.6, p. 1731-1743, 2017. Disponível em:< http://www.ripublication.com/acst17/acstv10n6_19.pdf> Acesso em: 10 ago. 2021.

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

JOSSINET, J. Variability of impedivity in normal and pathological breast tissue. **Med. & Biol. Eng. & Comput**, v.34, p.346-350, 1996.

LANDIS, J. R.; KOCH, G. G. The measurement of observer agreement for categorical data. **Biometrics**. v.33, p.159-75, 1977.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining**: Practical Machine Learning Tools and Techniques. Elsevier, 3rd Ed., 2011.

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.

NEAPOLITAN, R. E. **Learning Bayesian Networks**. Prentice-Hall, 2003.

MITCHELL, T. M. Generative and discriminative classifiers: naive bayes and logistic regression. **Carnegie mellon university**, 2017. Disponível em: https://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf. Acesso em: 01 ago 2021.

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.

MORAES, R.M.; MACHADO, L.S.; SOUZA, L.C. Skills Assessment of Users in Medical Training Based on Virtual Reality Using Bayesian NetWorks. **Lecture Notes in Computer Science.** v.7441, p.805-812, 2012.

ONTIVERO-ORTEGA, M., LAGE-CASTELLANOS, A.; VALENTE, A., GOEBEL, R.; VALDES-SOSA, M. FAST Gaussian Naive Bayes for searchlight classification analysis, **NeuroImage,** v. 163, 2017. Disponível em https://doi.org/10.1016/j.neuroimage.2017.09.001. Acesso em: 01 ago 2021.

SHIKHA, A.; BALMUKUMD, J.; TISU K.; MANISH K.; PRABHAT R. Hybrid of Naive Bayes and Gaussian Naive Bayes for Classification: A Map Reduce Approach. **International Journal of Innovative Technology and Exploring Engineering (IJITEE)**. v. 8, 2019. Disponível em: https://www.ijitee.org/wp-content/uploads/papers/v8i6s3/F10540486S319.pdf. Acesso em: 01 ago 2021.

SILVA, J.E.; MARQUES DE SÁ, J.P.; JOSSINET, J. Classification of Breast Tissue by Electrical Impedance Spectroscopy. **Med & Bio Eng & Computing**, v.38, p. 26-30, 2000.