

CLASSIFICADORES HÍBRIDOS BASEADOS NO NAIVE BAYES AVALIADOS EM DIFERENTES NÍVEIS DE DEPENDÊNCIA ENTRE VARIÁVEIS

Noé Mitterhofer Eiterer Ponce de Leon da Costa¹, Moysés Nascimento², Ana Carolina Campana Nascimento², Camila Ferreira Azevedo³, Matheus Massariol Suela⁴

¹ Doutorando em Estatística Aplicada e Biometria pela Universidade Federal de Viçosa, Viçosa-MG.

² Professor(a) associado do Departamento de Estatística da Universidade Federal de Viçosa, Viçosa-MG.

³ Professora adjunta do Departamento de Estatística da Universidade Federal de Viçosa, Viçosa-MG.

⁴ Doutorando em Genética e Melhoramento pela Universidade Federal de Viçosa, Viçosa-MG.

E-mail: noe.costa@ufv.br

Recebido em: 15/02/2024 – Aprovado em: 15/03/2024 – Publicado em: 30/03/2024

DOI: 10.18677/EnciBio_2024A4

RESUMO

O classificador *Naive Bayes* (NB) se destaca pela sua simplicidade e eficiência, porém, a pressuposição de independência entre as variáveis preditoras pode limitar o seu desempenho em cenários onde a independência não é observada. Diante disto, objetivou-se avaliar a combinação do NB com três diferentes metodologias: análise de componentes principais (PCA), componentes esparsos (SPCA) e análise discriminante (AD), na tentativa de melhorar seu desempenho. Foram simulados dados com diferentes níveis de dependência entre variáveis (correlação igual a 0,10; 0,50 e 0,90) e diferentes vetores de médias ($\mu_A = [i \dots i]^T$ e $\mu_B = [j \dots j]^T$, em que $i = 0$ e $j = 1, 2$ ou 3). Todos os cenários foram avaliados considerando 2, 4, 8 e 16 variáveis. Ademais, comparou-se os resultados observados com os métodos *NBTree*, *Random Forest*, *Bagging* e Rede Neural Artificial através do cálculo da acurácia média e o respectivo erro padrão da média. Em geral, o aumento de variáveis proporcionou bons ganhos de acurácia média ao NB em cenários de baixa correlação, porém pouco ou nenhum ganho na acurácia quando em alta correlação. Assim, concluímos que a pressuposição de independência é importante, uma vez que o aumento na correlação sempre resultou em redução da acurácia média dos classificadores. Os classificadores híbridos propostos no trabalho apresentaram-se como boas alternativas ao NB, uma vez que apresentaram resultados semelhantes ou superiores ao próprio NB e demais métodos avaliados quanto à acurácia média.

PALAVRAS-CHAVE: Acurácia de Classificação, Metodologias Combinadas, Simulação.

ROBUSTNESS OF HYBRID NAIVE BAYES CLASSIFIERS IN UNDER VIOLATIONS OF VARIABLE INDEPENDENCE ASSUMPTIONS

ABSTRACT

The Naive Bayes (NB) classifier is known for its simplicity, good performance, and the naive assumption of independence between predictor variables. However, this assumption may limit its effectiveness in scenarios where independence is not

observed. This study aims to evaluate methodologies combined with NB to enhance its performance in scenarios where the independence assumption is violated. The combined methodologies include principal components analysis (PCA), sparse components (SPCA) and discriminant analysis (AD). Simulated data featured varying levels of correlation (correlation equal to 0.10; 0.50 and 0.90) and different mean vectors ($\mu_A = [i \dots i]^T$ and $\mu_B = [j \dots j]^T$, where $i = 0$ and $j = 1, 2$ or 3 , with $i \neq j$). Scenarios were with 2, 4, 8 and 16 variables. The methodologies were compared with NBTree, Random Forest, Bagging and Artificial Neural Network, using the calculation of average accuracy and standard error of the media. While increasing the number of variables led to significant average accuracy gains for NB in scenarios, but with low correlation, there was little or no improvement in accuracy with increased correlation. Therefore, the study concludes that the assumption of independence is crucial, as an increase in correlation consistently resulted in reduced average accuracy of classifiers. The hybrids classifiers proposed in the study emerge as viable alternatives to NB, exhibiting results similar to or even superior to NB and other evaluated methods in terms of average classification accuracy.

KEYWORDS: Classification Accuracy, Combined Methods, Simulation.

INTRODUÇÃO

Diversas metodologias são utilizadas para solucionar problemas de classificação, abrangendo classificadores bayesianos, análise discriminante, *Random Forest*, *Bagging*, Redes Neurais Artificiais, Máquinas de Vetores de Suporte, entre outros (YOU, 2020; FARHADIAN, 2020; PAN *et al.*, 2022, FATHY, 2023). Entre essas abordagens, destaca-se o *Naive Bayes* (NB) como um classificador simples e eficiente diante das complexas técnicas de aprendizado de máquinas (CRUZ; NASCIMENTO, 2019), superando até mesmo outros classificadores bayesianos. No entanto, a pressuposição do NB de independência entre as variáveis pode ser bastante restritiva. Especificamente, em sua formulação o NB pressupõe que as variáveis utilizadas na construção do classificador são independentes.

Conforme apontado por Lewis (1998), em cenários de dependência entre as variáveis, o NB apresenta acurácia muito baixa e não deve ser adotado. Entretanto, existem várias propostas na literatura que tratam do caso de variáveis k -dependentes, especialmente o caso quando $k = 2$ (SAHAMI, 1996; FRIEDMAN *et al.*, 1997.; WANG *et al.*, 2021). Nesses métodos busca-se formas de reduzir o efeito da dependência entre as variáveis visando manter um bom desempenho do método (SAHAMI, 1996; FRIEDMAN *et al.*, 1997; WANG *et al.*, 2021).

O uso de metodologias combinadas ao NB pode auxiliar o classificador a manter um bom desempenho em cenários nos quais a pressuposição de independência é desafiada. Assim, surge a possibilidade do uso combinado de NB com as análises de componentes principais (PCA) e esparsos (SPCA), que geram componentes não correlacionados entre si (REGAZZI; CRUZ, 2020; SILVA, 2023). Além destes, uma alternativa é a análise discriminante (AD) que maximiza a diferença entre o vetor de médias de duas populações (REGAZZI; CRUZ, 2020; SILVA, 2023).

Alguns estudos demonstraram um aumento na acurácia média ao combinar NB e PCA, onde a PCA foi empregada para simplificar conjuntos de dados com alta dimensionalidade (CHEN; KIM, 2013; HERLAMBANG *et al.*, 2021). Wang *et al.* (2022) observaram melhora na acurácia média ao combinar NB e *auto-encoder*. Assim, dada a escassez de literatura que aborda estudos envolvendo metodologias

combinadas, em que métodos são utilizados para mitigar a falta de atendimento da pressuposição do NB em cenários de dependência entre as variáveis, torna-se evidente a necessidade de conduzir um estudo comparativo do NB + PCA, NB + SPCA e NB + AD com as metodologias comumente utilizadas em problemas de classificação.

Diante do exposto, este trabalho tem o objetivo de propor o uso de classificadores híbridos, ou seja, metodologias combinadas ao NB para obter dados transformados visando aumentar a performance preditiva do método. Para isso, avaliou-se os seguintes classificadores híbridos combinados com o NB: i) NB + PCA; ii) NB + SPCA e iii) NB + AD. Por fim, foi realizada a comparação da acurácia desses classificadores híbridos com os resultados de acurácia média obtidos por NB, *NBTree*, *Random Forest* (RF), *Bagging* (BAG) e Redes Neurais Artificiais (RNA).

MATERIAL E MÉTODOS

Dados simulados

A robustez do NB foi avaliada considerando diferentes níveis de dependência entre variáveis sob a distribuição normal multivariada (DNM) e com diferentes níveis de correlação ($\rho = 0,10; 0,50$ ou $0,90$). É importante ressaltar que sob a DNM, a não correlação implica em independência entre as variáveis (REGAZZI; CRUZ, 2020), possibilitando a avaliação do NB em cenários que atendem ou não à sua pressuposição.

A avaliação do NB foi realizada para a classificação de duas populações com o mesmo tamanho de amostra ($n_A = n_B = 100$) e diferentes números de variáveis ($p = 2, 4, 8$ ou 16) em cada população. A variação na quantidade de variáveis utilizadas tem como objetivo compreender o efeito do número de variáveis em situações favoráveis ou desfavoráveis ao atendimento da pressuposição do NB.

Os valores paramétricos para os vetores de médias foram definidos como:

$$\mu_A = [i \dots i]^T \quad (1)$$

$$\mu_B = [j \dots j]^T, \quad (2)$$

em que $i = 0$ e $j = 1, 2$ ou 3 . A diferença entre os valores paramétricos dos vetores de média visa verificar o efeito da diferença entre as médias populacionais no problema de classificação.

As matrizes de covariâncias simuladas neste estudo tiveram diagonal igual a um e os demais elementos da matriz foram ajustados para refletir a correlação desejada entre as variáveis. Por fim, as matrizes de covariâncias utilizadas neste trabalho foram representadas por:

$$\Sigma_A = \begin{bmatrix} 1 & \dots & 0,1 \\ \dots & \dots & \dots \\ 0,1 & \dots & 1 \end{bmatrix} \quad (3)$$

$$\Sigma_B = \begin{bmatrix} 1 & \dots & 0,5 \\ \dots & \dots & \dots \\ 0,5 & \dots & 1 \end{bmatrix} \quad (4)$$

$$\Sigma_C = \begin{bmatrix} 1 & \dots & 0,9 \\ \dots & \dots & \dots \\ 0,9 & \dots & 1 \end{bmatrix}. \quad (5)$$

Todos os nove cenários simulados (Tabela 1) foram repetidos para populações com 2, 4, 8 ou 16 variáveis, proporcionando uma compreensão da importância do número de variáveis para os classificadores. Todo trabalho de simulação e análise estatística foi desenvolvido no *software* R (R CORE TEAM, 2024).

TABELA 1. Descrição dos cenários simulados combinando os diferentes níveis de correlação entre variáveis (ρ) com os diferentes valores paramétricos de média das duas populações (μ_1 e μ_2).

Cenário	ρ	μ_1	μ_2
1	0,10	0	1
2	0,50	0	1
3	0,90	0	1
4	0,10	0	2
5	0,50	0	2
6	0,90	0	2
7	0,10	0	3
8	0,50	0	3
9	0,90	0	3

O classificador Naive Bayes

O classificador *Naive Bayes* é um classificador de fácil implementação computacional, e o NB tem como pressuposição a independência entre as variáveis avaliadas. Para demonstrar a importância dessa pressuposição, suponha que c_k representa pertencer a k-ésima classe de interesse C e x_i representa a i-ésima observação da variável X . Assim, segundo Salmi e Rustam (2019), a probabilidade de pertinência na k-ésima classe dado a observação x_i é dada por:

$$P(c_k | x_i) = \frac{\prod_{i=1}^n P(c_k)P(x_i | c_k)}{P(x)} \quad (6)$$

O denominador da expressão (6) pode ser facilmente calculado através do conjunto de dados. Desta forma, o interesse está apenas no numerador, que pode ser reescrito como:

$$\prod_{i=1}^n P(c_k)P(x_i | c_k) = P(c_k)P(x_1, \dots, x_n | c_k). \quad (7)$$

Segundo Salmi e Rustam (2019), desdobrando as probabilidades condicionais:

$$P(c_k)P(x_i | c_k) = P(c_k)P(x_1 | c_k)P(x_2 | c_k, x_1)P(x_3, \dots, x_n | c_k, x_1, x_2) \quad (8)$$

$$P(c_k)P(x_i | c_k) = P(c_k)P(x_1 | c_k)P(x_2 | c_k, x_1)P(x_3, \dots, x_n | c_k, x_1, x_2, x_3) \quad (9)$$

$$P(c_k)P(x_i | c_k) = P(c_k)P(x_1 | c_k)P(x_2 | c_k, x_1) \dots P(x_n | c_k, x_1, x_2, x_3, \dots, x_{n-1}). \quad (10)$$

Dessa forma, o cálculo das probabilidades condicionais sem pressupor independência é computacionalmente oneroso. Portanto, utiliza-se a pressuposição de que a variável x_i é independente da variável x_j para $i \neq j$, obtendo:

$$P(x_i | c_k, x_j) = P(x_i | c_k), \quad (11)$$

para $i \neq j$.

Segundo Salmi e Rustam (2019), reescrevendo a equação (8) com essa pressuposição, é obtido:

$$P(c_k, x_i) = P(c_k)P(x_1 | c_k)P(x_2 | c_k) \dots P(x_n | c_k) \quad (12)$$

$$P(c_k, x_i) = P(c_k) \prod_{i=1}^n P(x_i | c_k). \quad (13)$$

Por meio da equação (13) é obtida a função do classificador dada por:

$$f_i^{NB}(x) = P(c_k) \prod_{j=1}^n P(x_j|c_k) \quad e \quad (14)$$

$$h(x) = \operatorname{argmax}_{i \in \{0, \dots, m-1\}} f_i^{NB}(x), \quad (15)$$

em que argmax é o argumento que retorna o máximo da função do classificador NB dada por $f_i^{NB}(x)$. Assim, o classificador irá alocar o indivíduo na população i se a probabilidade de boa classificação na população i for máxima. A função naiveBayes do pacote e1071 (MEYER *et al.*, 2023) foi utilizada para ajuste do NB no *software* R (R CORE TEAM, 2024).

Métodos estatísticos combinados ao Naive Bayes

A análise de componentes principais (*Principal Component Analysis* - PCA) é uma técnica da estatística multivariada que visa a redução da dimensionalidade das variáveis originais em poucos componentes, também conhecidos como variáveis latentes (REGAZZI; CRUZ, 2020; SILVA, 2023). Dessa forma, a PCA explora a estrutura de variância por meio de combinações lineares de componentes não correlacionados, retendo uma porcentagem da variabilidade total dos dados em p componentes (REGAZZI; CRUZ, 2020; SILVA, 2023). Os componentes principais são dados pelas seguintes combinações lineares:

$$CP_p = c_{p1}X_1 + c_{p2}X_2 + \dots + c_{pp}X_p, \quad (16)$$

em que CP_i é o i -ésimo componente ($i = 1, 2, \dots, p$), c_{ij} é o *loading* da j -ésima variável do i -ésimo componente ($j = 1, 2, \dots, p$) e X_j é a j -ésima variável.

Zou *et al.* (2006) propuseram o método SPCA utilizando a regressão *elastic net* (ZOU; HASTIE, 2005) para selecionar variáveis que estão relacionadas a cada componente. O componente obtido por SPCA é o resultado da combinação linear das variáveis selecionadas, uma vez que o estimador *elastic net* irá fornecer estimativas de *loadings* nulas (c_{ij}), descartando aquelas variáveis irrelevantes. O estimador *elastic net* $\tilde{\beta}_{en}$ para um modelo de regressão da forma $y_i = X\beta + \varepsilon$ é:

$$\tilde{\beta}_{en} = (1 + \lambda_2) \left\{ \operatorname{argmin}_{\beta} (\|y_i - X\beta\|^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1) \right\}, \quad (17)$$

em que X é a matriz de incidência, y_i é o vetor de observações, β é o vetor de parâmetros e ε é o erro aleatório, λ_1 e λ_2 são parâmetros de penalização. Geralmente, utiliza-se a restrição $\lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \leq t$, para determinado t com λ_1 e λ_2 não negativos. Assim, obtém a seguinte aproximação:

$$\frac{\tilde{\beta}_{en}}{\|\tilde{\beta}_{en}\|} \approx v_i, \quad (18)$$

em que v_i é o *loading*. Uma vez que ocorreu uma modificação nos vetores de *loadings*, estes não são mais os autovetores da matriz de covariâncias como em PCA. Assim, as propriedades de ortogonalidade e de independência dos vetores de *loadings* em SPCA são apenas aproximadas.

A análise discriminante linear de Fisher (AD) classifica as observações buscando minimizar a probabilidade de má classificação, ou seja, a classificação equivocada daquela observação (REGAZZI; CRUZ, 2020). A AD irá buscar alocar um indivíduo em uma população para que a distância generalizada de Mahalanobis seja mínima (REGAZZI; CRUZ, 2020). A função discriminante pode ser apresentada de forma simplificada por:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \dots + W_n X_{nk} \quad (19)$$

em que Z_{jk} é o escore da função discriminante j para o objeto k , a é o intercepto, W_i é o peso discriminante para a variável independente i e X_{ik} é a variável independente i para o objeto k .

Assim, os métodos combinados ao NB foram escolhidos por fornecer variáveis latentes não correlacionadas no caso da análise de componentes principais (PCA) ou esparsos (SPCA), ou por maximizar a diferença entre a média de dois grupos como a análise discriminante (AD). Os componentes principais ou esparsos, e os escores da AD obtidos foram as variáveis utilizadas para as metodologias combinadas neste trabalho (NB + PCA, NB + SPCA e NB + AD). As funções dos pca e spca do pacote mixOmics (ROHART *et al.*, 2017) foram utilizadas na obtenção dos componentes principais e esparsos, respectivamente, no *software* R. O número mínimo de componentes foi aquele necessário para reter pelo menos 80% da variabilidade total dos dados originais.

Comparação entre metodologias

Foram utilizados o *NBTree* e os algoritmos de aprendizado de máquina BAG, RF e RNA (BREIMAN, 2001; CRUZ; NASCIMENTO, 2019) para comparação com as metodologias híbridas propostas neste trabalho. Assim, para todos os oito métodos avaliados foi calculada a acurácia média e o desvio padrão da média com base nos resultados obtidos em 50 repetições de cada cenário.

O classificador híbrido *NBTree* não possui implementação direta por função ou pacote na linguagem R. Portanto, foi necessário utilizar uma conexão com a biblioteca *Weka* da linguagem Java no R e empregar funções adequadas para implementar o *NBTree*. Para este propósito, foram utilizados os pacotes *rJava* (URBANEK, 2024), *RWekajars* (HORNIK, 2022) e *RWeka* (HORNIK *et al.*, 2023). As metodologias BAG e RF foram implementadas a partir da função *randomForest* do pacote *randomForest* (BREIMAN, 2001; BREIMAN *et al.*, 2022). A quantidade de variáveis utilizada no RF foi igual a raiz quadrada da quantidade total de variáveis, enquanto o BAG utilizou todas as variáveis disponíveis para o ajuste (JAMES *et al.*, 2013).

Foi realizado um estudo de dimensionamento da RNA para compreender a estrutura de rede que proporciona o maior valor de acurácia média. Foram avaliadas todas as possíveis estruturas redes com duas camadas, variando o número de neurônios de 1 a 20 em cada camada oculta, totalizando 400 estruturas de rede avaliadas para cada cenário. Cada estrutura foi repetida 50 vezes para obtenção da acurácia média e seu erro padrão da média. Para a discussão deste trabalho, considerou-se a maior acurácia média obtida dentro de cada cenário. Esse estudo de dimensionamento é relativamente comum e puramente empírico (COSTA *et al.*, 2022; SOUSA *et al.*, 2022). A função *neuralnetwork* do pacote ANN2 (LAMMERS, 2022) no *software* R foi empregada para a implementar as RNAs do tipo Perceptron de Múltiplas Camadas que é uma rede do tipo *feedforward*. Cada rede foi ajustada com 500 iterações, utilizando a função de ativação tangente hiperbólica nas camadas ocultas e a *softmax* na camada de saída.

RESULTADOS E DISCUSSÃO

Os gráficos de *heatmap* exibindo os resultados da acurácia média foram gerados utilizando o pacote *ggplot2* (WICKHAM, 2016) e estão representados nas Figuras 1, 2, 3 e 4. Nessas figuras, a primeira população tem média fixa em $\mu_A = 0$,

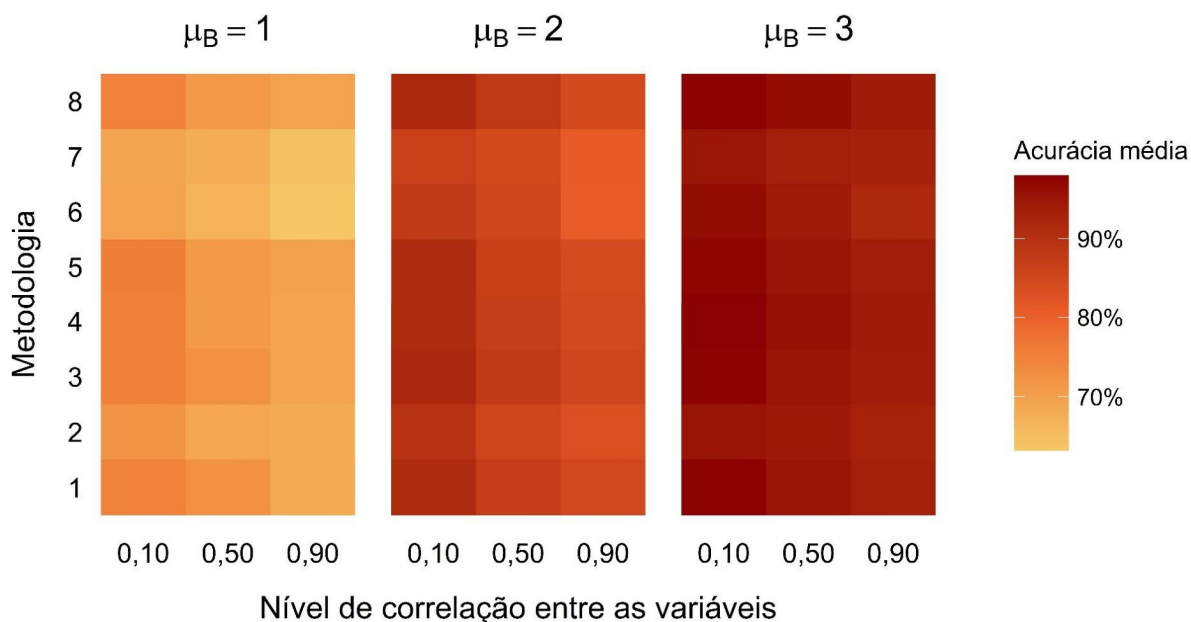
enquanto na segunda população, a média varia em $\mu_B = 1, 2$ ou 3 , conforme descrito na Tabela 1.

Na classificação de populações com médias mais discrepantes foi obtido um valor mais elevado de acurácia média em comparação com populações de médias mais próximas, independentemente do número de variáveis (Figuras 1, 2, 3 e 4). Observa-se também que os valores de acurácia média apresentados pelo NB para duas, quatro, oito ou dezesseis variáveis apresentam certa proximidade com os outros métodos avaliados e são valores que, em geral, indicam um bom desempenho dos métodos utilizados (Figuras 1, 2, 3, e 4). As metodologias propostas (NB + AD, NB + PCA e NB + SPCA) predominaram no ranking das acurácias médias mais altas em comparação com metodologias amplamente consolidadas na literatura (NB, *NBTree*, BAG, RNA e RF), independentemente do número de variáveis da população. O aumento de acurácia média ao utilizar PCA na redução da dimensionalidade de dados reais também foi observado em trabalhos com dados reais (CHEN; KIM, 2013 ; HERLAMBANG *et al.*, 2021).

Considerando cenários de baixa correlação, uma quantidade maior de componentes principais ou esparsos foi necessária para reter pelo menos 80% da variabilidade, em comparação com cenários de alta correlação. Geralmente, um único componente foi suficiente para reter o valor mínimo estabelecido de variação total em alta correlação ($\rho = 0,90$) e com médias discrepantes ($\mu_1 = 0$ e $\mu_2 = 3$).

Nos nove cenários avaliados em populações com duas variáveis (Figura 1), a RNA apresentou o maior valor de acurácia média em dois cenários (cenários 5 e 9), NB + SPCA em dois (7 e 9), NB + PCA em três cenários (2, 4 e 6) e NB + AD em dois cenários (1 e 3). O classificador NB apresentou acurácia média maior que os classificadores *NBTree*, RF e BAG em todos os cenários avaliados. Mantendo as médias constantes, todas as metodologias avaliadas tiveram redução da acurácia média diante do aumento da correlação entre as variáveis (Figura 1).

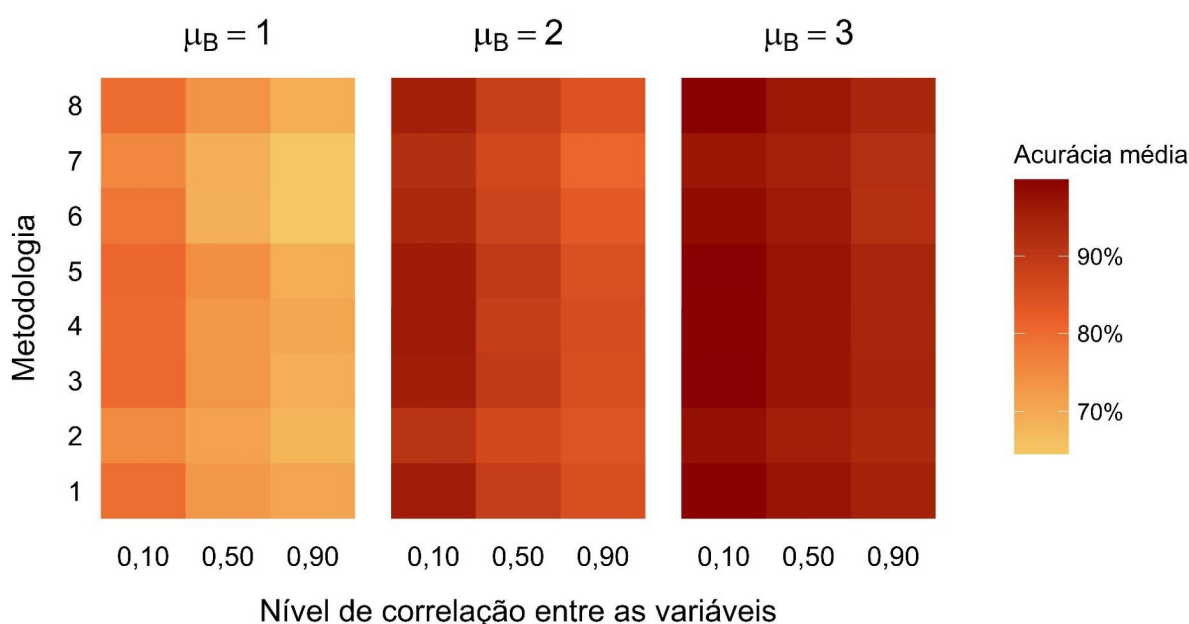
FIGURA 1. Gráfico de *heatmap* para a acurácia média em percentual do classificador NB (1), *NBTree* (2), NB + PCA (3), NB + SPCA (4), NB + AD (5), RF (6), BAG (7) e RNA (8) para classificação de duas populações com duas variáveis normalmente distribuídas em que $\mu_A = 0$.



Considerando todos os métodos, o maior decréscimo de acurácia foi de 9,25% ao variar a correlação de 0,10 para 0,90 no classificador RF em populações com médias $\mu_1 = 0$ e $\mu_2 = 1$. A menor perda na acurácia média, mantendo as médias constantes e variando a correlação de 0,10 para 0,90 foi de 2,01% utilizando BAG em populações com médias $\mu_1 = 0$ e $\mu_2 = 3$. O aumento da correlação em cenários de população com variáveis de médias discrepantes teve um efeito negativo menor do que em médias próximas (Figura 1).

Para os cenários avaliados em populações com quatro variáveis (Figura 2), em três cenários o NB obteve a maior acurácia média (cenários 3, 8 e 9), em 4 cenários foi o NB + AD (1, 2, 5 e 7) e em dois cenários foi o NB + SPCA (4 e 6). Novamente, os classificadores *NBTree*, RF e BAG não superaram a acurácia média obtida por NB em nenhum dos cenários avaliados. Ao contrário do que aconteceu com duas variáveis, a RNA não apresentou o melhor desempenho em nenhum dos cenários. Comparando ao trabalho de Zhang *et al.* (2005), os valores de acurácia obtidos pela proposta *Hidden Naive Bayes* foram semelhantes aos do presente trabalho, com destaque que a proposta do autor demanda maior recurso computacional (ZHANG *et al.*, 2005).

FIGURA 2. Gráfico de *heatmap* para a acurácia média em percentual do classificador NB (1), *NBTree* (2), NB + PCA (3), NB + SPCA (4), NB + AD (5), RF (6), BAG (7) e RNA (8) para classificação de duas populações com quatro variáveis normalmente distribuídas em que $\mu_A = 0$.



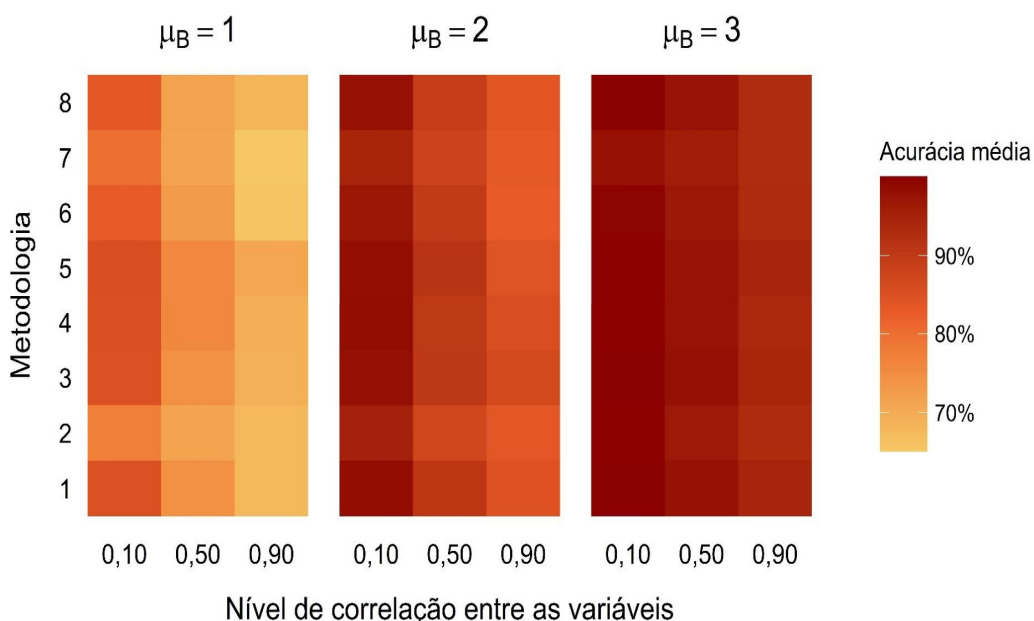
Mantendo as médias constantes e variando a correlação de 0,10 para 0,90, observa-se que a maior perda de acurácia média foi de 17,40% utilizando RF em populações com médias $\mu_1 = 0$ e $\mu_2 = 1$. Na mesma comparação, a menor redução de acurácia média foi de 4,52% utilizando *NBTree* em populações com médias de $\mu_1 = 0$ e $\mu_2 = 3$. Essa comparação reitera que o aumento da correlação em cenários de populações com médias discrepantes tem um impacto negativo menor na acurácia média em comparação com populações que possuem médias próximas.

Avaliando populações com oito variáveis (Figura 3), destaca-se que dos cenários avaliados, em quatro cenários o NB + AD apresentou a maior acurácia média (cenários 1, 3, 5 e 9), em dois cenários foi o NB + CP (6 e 8) e apenas um cenário cada foi NB (4), RNA (7) e NB + SPCA (2). Ao utilizar oito variáveis, as metodologias híbridas propostas alcançaram valores de acurácia próximos ou melhores que os obtidos por Sahami (1996) com sua proposta *k-dependence Bayesian* classifiers. Sahami (1996) propôs o NB utilizando probabilidades condicionais em que as variáveis avaliadas são k-dependentes entre si, fato que demanda maior recurso computacional e dificuldade na implementação. Comparado ao método *Hidden Naive Bayes* proposto por Zhang *et al.* (2005), as metodologias híbridas propostas neste trabalho também alcançaram valores próximos ou melhores aos obtidos pelos autores.

O classificador BAG, novamente, obteve valores de acurácia média menor que o NB em todos os cenários. Enquanto o *NBTree* e o RF, igualmente ocorreu em populações de duas e quatro variáveis, não apresentaram desempenho melhor que o NB nos cenários avaliados. Assim, como aconteceu no resultado com quatro variáveis, a RNA não apresentou resultados muito promissores de acurácia média frente aos outros métodos.

Na comparação mantendo médias constantes e variando a correlação de 0,10 a 0,90, obteve-se a maior redução de acurácia média em 20,55% utilizando NB em populações de médias $\mu_1 = 0$ e $\mu_2 = 1$. E a menor redução foi em 4,83% no uso de BAG em populações com médias $\mu_1 = 0$ e $\mu_2 = 3$. Assim como observado nos resultados para duas e quatro variáveis, o aumento da correlação teve maior impacto na redução da acurácia média em populações com médias próximas do que em populações com médias discrepantes.

FIGURA 3. Gráfico *heatmap* para a acurácia média em percentual do classificador NB (1), *NBTree* (2), NB + PCA (3), NB + SPCA (4), NB + AD (5), RF (6), BAG (7) e RNA (8) para classificação de duas populações com oito variáveis normalmente distribuídas em que $\mu_A = 0$.

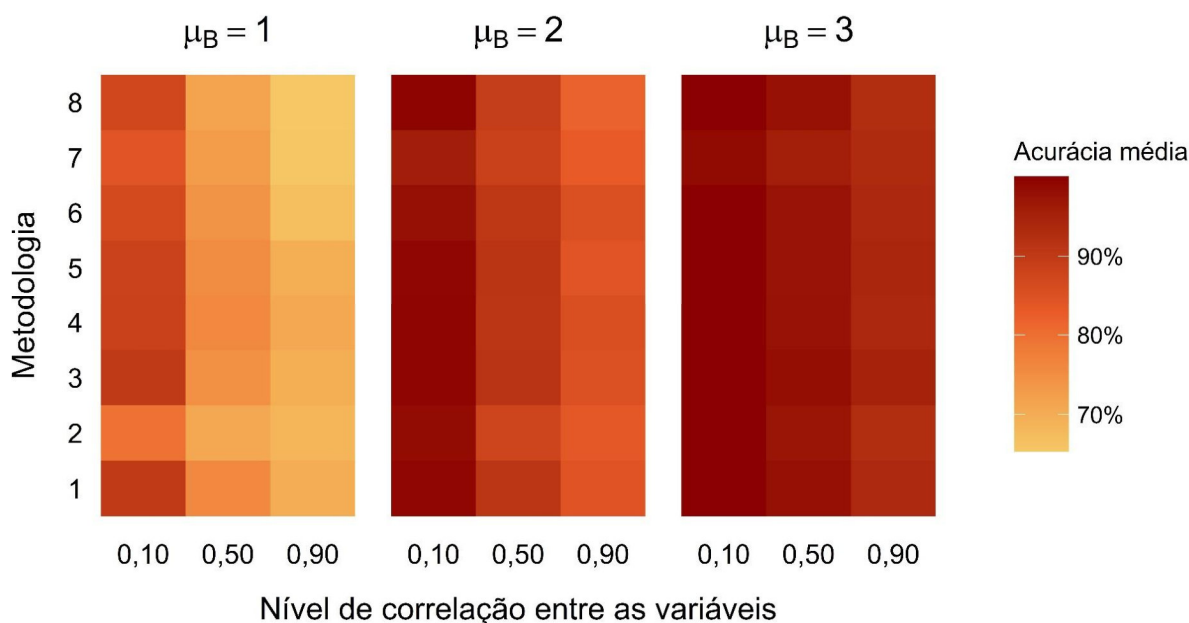


Em populações com dezesseis variáveis (Figura 4), o NB + PCA apresentou a maior acurácia média em três cenários (1, 5, 8 e 9), em 3 cenários foi o NB + SPCA (3 e 6), RNA e NB obtiveram maior acurácia em 1 cenário cada (4 e 2, respectivamente). Ressalta-se que no cenário 7 ($\rho = 0,10$; $\mu_1 = 0$ e $\mu_2 = 3$) os métodos NB, NB + PCA, NB + SPCA, *NBTree* e RNA apresentaram acurácia média igual a 100% nas 50 repetições avaliadas.

Os valores de acurácia obtidos pelas metodologias propostas com dezesseis variáveis estão próximos ou melhores daqueles obtidos com o *Selective Bayesian Network Classifier* (SINGH *et al.*, 1996) e o *Hidden Naive Bayes* (ZHANG *et al.*, 2005). Porém, os métodos propostos pelos autores apresentam maior complexidade de implementação computacional.

Os classificadores *NBTree* e BAG não obtiveram acurácia média maior que o NB em nenhum dos cenários. Os métodos RF e RNA apresentaram resultado de acurácia média maior que o NB apenas em dois (6 e 9) e um cenário (4), respectivamente. Portanto, os métodos BAG, RF, *NBTree* e RNA não alcançaram bons resultados de forma geral quando comparados aos outros classificadores.

FIGURA 4. Gráfico *heatmap* para a acurácia média em percentual do classificador NB (1), *NBTree* (2), NB + PCA (3), NB + SPCA (4), NB + AD (5), RF (6), BAG (7) e RNA (8) para classificação de duas populações com 16 variáveis normalmente distribuídas em que $\mu_A = 0$.



Mantendo as médias constantes e variando a correlação de 0,10 a 0,90, obteve-se a maior redução de acurácia média em 25,15% utilizando RNA com populações de médias $\mu_1 = 0$ e $\mu_2 = 1$. E a menor redução foi em 4,87% no uso de NB + PCA em populações com médias $\mu_1 = 0$ e $\mu_2 = 3$. Resultado comum ao obtido para duas, quatro, oito ou com dezesseis variáveis é uma maior redução da acurácia média frente ao aumento da correlação em cenários de populações com médias próximas e menor perda da acurácia quando as médias são discrepantes.

Van der Heide *et al.* (2019) avaliaram a regressão logística múltipla, NB e *Random Forest* na predição de sobrevivência de vacas holandesas em fase de

segunda lactação. Para isso, foram utilizadas 65 variáveis fenotípicas avaliadas ao longo do tempo e 50 variáveis genômicas avaliadas no nascimento do animal. Os momentos que ocorreram as predições de sobrevivência foram no nascimento, 18 meses após o nascimento, primeiro parto, seis meses após o primeiro parto e 200 dias após o primeiro parto. Os valores de acurácia do NB foram próximos de 80%, alcançando a maior acurácia em três dos cinco momentos de avaliação ao longo da vida do animal. Assim, as acurácias médias obtidas pelas metodologias propostas neste trabalho estão próximas ou melhores que as obtidas por Van der Heide *et al.* (2019).

Dentre as aplicações no melhoramento genético vegetal, Xu *et al.* (2021) avaliaram a acurácia de diversos métodos para descartar genótipos com clorose por deficiência de ferro (IDC) em soja. Foram obtidos os escores IDC em 38.803 linhagens experimentais de soja em 48 localidades ao longo dos anos de 2013 a 2016. A partir desse conjunto de dados foram utilizados 10 métodos de classificação para descartar os genótipos susceptíveis. Os métodos avaliados foram o NB, Rede Neural Artificial, *Random Forest*, k-vizinho mais próximo, máquina de vetores de suporte, máquina de aumento de gradiente, regressão logística, regressão logística penalizada, regressão *Ridge* e regressão linear generalizada bayesiana. O NB obteve a sexta maior acurácia, valor superior aos obtidos com máquina de vetores de suporte, k-vizinho mais próximo, regressão linear generalizada bayesiana e regressão logística. Esses trabalhos mostraram o potencial do NB na obtenção de melhores resultados no melhoramento genético frente a métodos mais complexos computacionalmente.

CONCLUSÕES

Os resultados do estudo destacam a importância da pressuposição de independência para o desempenho do NB, uma vez que o aumento na correlação entre as variáveis resultou consistentemente em redução da acurácia média dos classificadores. Apesar disso, o NB demonstrou ser um bom classificador, quando seus resultados são comparados com os valores de acurácia de predição obtidos por outros métodos. Os classificadores híbridos, embora não tenham apresentado resultados muito superiores do NB, oferecem alternativas interessantes para lidar com a dependência entre variáveis. O aumento do número de variáveis não contornou o problema do não atendimento da pressuposição, uma vez que não proporcionou aumento na acurácia média dos classificadores em cenários de alta correlação.

REFERÊNCIAS

BREIMAN, L. Random Forests. **Machine Learning**. v. 45, 5-32, 2001. Disponível em: <<https://link.springer.com/article/10.1023/A:1010933404324>>. DOI: 10.1023/A:1010933404324.

BREIMAN, L.; CUTLER, A.; LIAW, A.; WIENER, M. **Breiman and Cutler's Random Forests for Classification and Regression**. R package versão 4.7-1.1, 2022. Disponível em: <<https://cran.r-project.org/web/packages/randomForest/index.html>>.

CHEN, Z.G.; KIM, S.R. Combining principal component analysis, decision tree and naïve Bayesian algorithm for adaptive intrusion detection. **Association for Computing Machinery**. p. 312-316, 2013. Disponível em:

<<https://dl.acm.org/doi/abs/10.1145/2513228.2513281>>.
10.1145/2513228.2513281.

DOI:

COSTA, W. G.; CELERI, O. M.; BARBOSA, P. I.; SILVA, G. N.; AZEVEDO, C. F.; BORÉM, A.; NASCIMENTO, N.; CRUZ, C. D. Genomic prediction through machine learning and neural networks for traits with epistasis. **Computacional and Structural Biotechnology Journal**. v. 20, 5490-5499, 2022. Disponível em: <[https://www.csbj.org/article/S2001-0370\(22\)00434-2/fulltext](https://www.csbj.org/article/S2001-0370(22)00434-2/fulltext)>. DOI: 10.1016/j.csbj.2022.09.029.

CRUZ, C. D.; NASCIMENTO, M. **Inteligência computacional aplicada ao melhoramento genético**. Viçosa, MG: Editora UFV, 2019. 414 p.

FARHADIAN, M.; SHOKOUHI, P.; TORKZABAN, P. A decision support system based on support vector machine for diagnosis of periodontal disease. **BMC Res Notes**. v. 13, n. 337, 2020. Disponível em: <<https://bmresnotes.biomedcentral.com/articles/10.1186/s13104-020-05180-5>>. DOI: 10.1186/s13104-020-05180-5.

FATHY, A.; HASSAN, R.A.; SOLIMAN, E.S. Productive and reproductive traits and clinical epidemiological measures in discriminant analysis between baladi buffaloes and crossbred. **Advances in Animal and Veterinary Sciences**. v. 11, n. 1, 2023. Disponível em: <<http://researcherslinks.com/current-issues/Productive-and-Reproductive-Traits-and-Clinical-Epidemiological-Measures-in-Discriminant-Analysis-between-Baladi-Buffaloes-and-Crossbred/33/1/5667/html>>. DOI: 10.17582/journal.aavs/2023/11.1.45.56.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. **Machine learning**. v. 29, p. 131-163, 1997. Disponível em: <<https://link.springer.com/article/10.1023/A:1007465528199>>. DOI: 10.1023/A:1007465528199.

HERLAMBANG, W.D.; LAKSITOWENING, K.A.; ASROR, I. Prediction of Graduation with Naïve Bayes Algorithm and Principal Component Analysis (PCA) on Time Series Data. **9th International Conference on Information and Communication Technology (ICoICT)**. Yogyakarta, Indonesia, p. 645-649, 2021. Disponível em: <<https://ieeexplore.ieee.org/document/9527443/authors#authors>>. DOI: 10.1109/ICoICT52021.2021.9527443.

HORNIK, K.; BUCHTA, C.; HOTHORN, T.; KARATZOGLOU, A.; MEYER, D.; ZEILEIS, A. **R/Weka Interface**. R package versão 0.4-46, 2023. Disponível em: <<https://cran.r-project.org/web/packages/RWeka/index.html>>.

HORNIK, K. **R/Weka Interface Jars**. R package versão 3.9.3-2, 2022. Disponível em: <<https://cran.r-project.org/web/packages/RWekajars/index.html>>.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to Statistical Learning with Applications in R**. New York: Springer, 2013. 426 p.

LAMMERS, B. **Artificial Neural Networks for Anomaly Detection**. R package versão 2.3.4, 2022. Disponível em: <<https://cran.r-project.org/web/packages/ANN2/index.html>>.

LEWIS, D. D. Naïve (Bayes) at Forty: The Independence Assumption in Information Retrieval. **Proceedings of ECML-98, 10th European Conference on Machine Learning**. V. 1398, p. 4-15, 1998. Disponível em: <<https://link.springer.com/chapter/10.1007/BFb0026666>>.

MEYER, D.; DIMITRIADOU, E.; HORNIK, K.; WEINGESSEL, A.; LEISCH, F. **Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien**. R package versão 1.7-14, 2023. Disponível em: <<https://CRAN.R-project.org/package=e1071>>.

PAN, Y.; JIN, H.; GAO, J.; RAUF, H.T. Identification of Buffalo Breeds Using Self-Activated-Based Improved Convolutional Neural Networks. **Agriculture**. v. 12, 2022. Disponível em: <<https://www.mdpi.com/2077-0472/12/9/1386>>. DOI: 10.3390/agriculture12091386.

R CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2023. Disponível em: <<https://www.R-project.org/>>. Acesso em: 16 jan. 2024.

REGAZZI, A. J.; CRUZ, C. D. **Análise Multivariada Aplicada**. 1a ed. Viçosa: Universidade Federal de Viçosa, 2020. 408 p.

ROHART, F.; GAUTIER, B.; SINGH, A.; CAO, K.A.L. mixOmics: An R package for 'omics feature selection and multiple data integration. **PLoS One**. v. 13, n. 11, 2017. Disponível em: <<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005752>>. DOI: 10.1371/journal.pcbi.1005752.

SAHAMI, M. **Learning Limited Dependence Bayesian Classifiers**. KDD. p. 335-338, 1996. Disponível em: <<http://robotics.stanford.edu/users/sahami/papers-dir/kdd96-learn-bn.pdf>>.

SILVA, A. R. **Métodos de análise multivariada em R**. Piracicaba, SP: FEALQ, 2023. 192 p.

SALMI, N.; RUSTAM, Z. Naïve Bayes classifier models for predicting the colon cancer. **IOP conference series: materials science and engineering**. v. 546, 2019. Disponível em: <<https://iopscience.iop.org/article/10.1088/1757-899X/546/5/052068>>. DOI: 10.1088/1757-899X/546/5/052068.

SOUSA, I. C.; NASCIMENTO, M.; SANT'ANNA, I. C.; CAIXETA, E. T.; AZEVEDO, C. F.; *et al.*; *Marker effects and heritability estimates using additive-dominance genomic architectures via artificial neural networks in Coffea canephora*. **PLoS ONE**. v. 17, n. 1, 2022. Disponível em: <<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0262055>>. DOI: 10.1371/journal.pone.0262055.

SINGH, M.; PROVAN, M.; LANGLEY, P. Induction of selective bayesian network classifiers. **Machine Learning**. v. 2, 1996. Disponível em: <<https://arxiv.org/ftp/arxiv/papers/1302/1302.6828.pdf>>.

URBANEK, S. **Low-Level R to Java Interface**. R package versão 1.10-11, 2024. Disponível em: <<https://cran.r-project.org/web/packages/rJava/index.html>>.

VAN DER HEIDE, E. M. M.; VERRKAMP, R. F.; VAN PELT, M. L.; KAMPHUIS, C.; ATHANASIADIS, I.; DUCRO, B. J.; Comparing regression, naive Bayes, and random forest methods in the prediction of individual survival to second lactation in Holstein cattle. **Journal of dairy science**. v. 102, n. 10, p. 9409-9421, 2019. Disponível em: <[https://www.journalofdairyscience.org/article/S0022-0302\(19\)30715-5/fulltext](https://www.journalofdairyscience.org/article/S0022-0302(19)30715-5/fulltext)>. DOI: 10.3168/jds.2019-16295.

WANG, L.; QI, S.; LIU, Y.; LOU, H.; ZUO, X.; Bagging k-dependence Bayesian network classifiers. **Intelligent Data Analysis**. v. 25, n. 3, p. 641-667, 2021. Disponível em: <<https://content.iospress.com/articles/intelligent-data-analysis/ida205125>>. DOI: 10.3233/IDA-205125.

WANG, S.; REN, J.; LIAN, X.; BAI, R.; JIANG, X.; Boosting the Discriminant Power of Naive Bayes. **2022 26th International Conference on Pattern Recognition (ICPR)**. p. 4906-4912, 2022. Disponível em: <<https://ieeexplore.ieee.org/abstract/document/9956358>>. DOI: 10.1109/ICPR56361.2022.9956358.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**. 2a ed. New York: Springer, 2016. Disponível em: <<https://ggplot2.tidyverse.org>>.

XU, Z.; KUREK, A.; CANNON, S. B.; BEAVIS, W. D.; Predictions from algorithmic modeling result in better decisions than from data modeling for soybean iron deficiency chlorosis. **Plos one**. v. 16, n. 7, p. e0240948, 2021. Disponível em: <<https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0240948>>. DOI: 10.1371/journal.pone.0240948.

YOU, J.; VAN DER KLEIN, S.A.S.; LOU, E.; ZUIDHOF, M.J. Application of random forest classification to predict daily oviposition events in broiler breeders fed by precision feeding system. **Computers and Electronics in Agriculture**. v. 175, 1-8, 2020. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0168169920303355?via%3Dihub>>. DOI: 10.1016/j.compag.2020.105526.

ZHANG, Harry; JIANG, Liangxiao; SU, Jiang. Hidden naive bayes. **Aaai**. p. 919-924, 2005. Disponível em: <<https://dl.acm.org/doi/10.5555/1619410.1619480>>. DOI: 10.5555/1619410.1619480.

ZOU, H.; HASTIE, Y.; Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society**. v. 67, n. 2, p. 301-320, 2005. Disponível em: <<https://www.jstor.org/stable/3647580>>.

ZOU, H.; HASTIE, T.; TIBSHIRANI, R. Sparse principal component analysis. **Journal of Computational and Graphical Statistics**. v. 15, n. 2, p. 265-286, 2006. Disponível em: <<https://www.jstor.org/stable/27594179>>.