NOÉ MITTERHOFER EITERER PONCE DE LEON DA COSTA

ROBUSTEZ DE CLASSIFICADORES *NAIVE BAYES* HÍBRIDOS QUANTO A QUEBRA DO PRESSUPOSTO DE INDEPENDÊNCIA DAS VARIÁVEIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Moysés Nascimento

Coorientadora: Ana Carolina C. Nascimento

VIÇOSA - MINAS GERAIS
2023

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
C836r
2023

Costa, Noé Mitterhofer Eiterer Ponce de Leon da, 1998-
Robustez de classificadores *Naive Bayes* híbridos quanto a quebra do pressuposto de independência das variáveis / Noé Mitterhofer Eiterer Ponce de Leon da Costa. – Viçosa, MG, 2023.
1 dissertação eletrônica (66 p.): il. (algumas color.).

Inclui apêndices.
Orientador: Moysés Nascimento.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2023.
Referências bibliográficas: f. 49-54.
DOI: https://doi.org/10.47328/ufvbbt.2023.275
Modo de acesso: World Wide Web.

1. Teoria bayesiana de decisão estatística. 2. Simulação (Computadores híbridos). 3. Análise multivariada. 4. Cultivos agrícolas - Melhoramento genético - Métodos estatísticos. I. Nascimento, Moysés, 1979-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.542

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

NOÉ MITTERHOFER EITERER PONCE DE LEON DA COSTA

ROBUSTEZ DE CLASSIFICADORES *NAIVE BAYES* HÍBRIDOS QUANTO A QUEBRA DO PRESSUPOSTO DE INDEPENDÊNCIA DAS VARIÁVEIS

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 16 de fevereiro de 2023.

Assentimento:

Noé Mitterhofer Eiterer Ponce de Leon da Costa
Autor

Moysés Nascimento
Orientador

À minha família.

AGRADECIMENTOS

A Deus.

Aos meus pais e pilares da minha vida, Felipe e Marinês.

A minha irmã, Flora.

Ao meu orientador Moysés Nascimento, pela amizade, conhecimentos, paciência, incentivo e preocupação.

Aos meus amigos de graduação e aos meus professores de graduação pelas amizades, ensinamentos e momentos compartilhados antecedentes ao meu mestrado.

Aos meus amigos de mestrado pela amizade e momentos compartilhados.

Aos professores e funcionários do departamento de Estatística pelos ensinamentos, paciência, amizade e incentivo.

Aos meus amigos do Laboratório de Inteligência Computacional e Aprendizado Estatístico pela amizade e ensinamentos.

A minha coorientadora Ana Carolina Campana Nascimento pela contribuição no meu aprendizado e paciência.

Aos membros da banca, Prof. Doutor Moysés Nascimento, Profa. Doutora Camila Ferreira Azevedo e Prof. Doutor Filipe Ribeiro Formiga Teixeira.

À Universidade Federal de Viçosa, pela oportunidade de realizar a pós-graduação.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

À Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), pela concessão da bolsa de estudos.

A todos que contribuíram direta ou indiretamente para a realização desse trabalho.

RESUMO

COSTA, Noé Mitterhofer Eiterer Ponce de Leon da Costa, M.Sc., Universidade Federal de Viçosa, fevereiro de 2023. **Robustez de classificadores *Naive Bayes* híbridos quanto a quebra do pressuposto de independência das variáveis**. Orientador: Moysés Nascimento. Coorientadora: Ana Carolina Campana Nascimento.

Resumo: O aumento populacional gera uma demanda para o aumento da produção agrícola, principalmente no quesito da produtividade, uma vez que quase todas as áreas agricultáveis já produzem alimentos. Dentro da demanda do aumento de produtividade, o melhoramento genético aliado a estatística é fundamental para alcançar as atuais demandas. A Estatística oferece diversos métodos para análises dos dados agropecuários, entre esses métodos estão os classificadores. Tais métodos são capazes de alocar cada observação em uma das classes de interesse. Entre os métodos disponíveis, o classificador *Naive Bayes* (NB) se destaca pela sua simplicidade e bom desempenho. Entretanto, o mesmo tem como pressuposição a independência entre as variáveis preditoras. Diante do fato de que tal pressuposição é dificilmente alcançada na prática, este trabalho tem por objetivo avaliar métodos híbridos na tentativa de melhorar seu desempenho considerando diferentes níveis de dependência entre variáveis. As metodologias combinadas ao NB foram à análise de componentes principais (PCA + NB), componentes esparsos (SPCA + NB) e análise discriminante (AD + NB). Foram simulados dados com diferentes níveis de correlação (0,10; 0,50 e 0,90) e diferentes vetores de médias. Todos os cenários foram avaliados considerando 2, 4, 8 e 16 variáveis. As metodologias usadas na comparação dos métodos propostos foram *Random Forest*, *Bagging* e Rede Neural Artificial através do cálculo da acurácia média e o respectivo erro padrão da média. A partir dos resultados obtidos por simulação pôde-se concluir 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 a acurácia média.

Palavras-chave: Classificador híbrido. Metodologias combinadas. Simulação.

ABSTRACT

COSTA, Noé Mitterhofer Eiterer Ponce de Leon da Costa, M.Sc., Universidade Federal de Viçosa, February 2023. **Robustness of hybrid Naive Bayes classifiers in breaking variable independence guidelines**. Advisor: Moysés Nascimento. Co-advisor: Ana Carolina Campana Nascimento.

Population growth generates a demand for increased agricultural production, especially in terms of productivity, since almost all arable areas already produce food. Within the demand for increased productivity, plant breeding combined with statistics is essential to meet current demands. Statistics offers several methods for analyzing agricultural data, among these methods are the classifiers. Such methods are capable of allocating each observation into one of the classes of interest. Among the available methods, the Naive Bayes (NB) classifier stands out for its simplicity and good performance. However, it presupposes independence between the predictor variables. Given the fact that such an assumption is difficult to achieve in practice, this work aims to evaluate hybrid methods in an attempt to improve their performance considering different levels of dependence between variables. The methodologies combined with NB were principal component analysis (PCA + NB), sparse components (SPCA + NB) and discriminant analysis (AD + NB). Data with different levels of correlation (0.10; 0.50 and 0.90) and different mean vectors were simulated. All scenarios were evaluated considering 2, 4, 8 and 16 variables. The methodologies used in the comparison of the proposed methods were Random Forest, Bagging and Artificial Neural Network through the calculation of the average accuracy and the respective standard error of the average. From the results obtained by simulation, it can be concluded that the assumption of independence is important, since the increase in correlation always resulted in a reduction in the average accuracy of the classifiers. The hybrid classifiers proposed in the work are presented as good alternatives to the NB, since they presented results similar to or superior to the NB itself and other methods evaluated in terms of average accuracy.

Keywords: Hybrid classifier. Combined methodologies. Simulation.

LISTA DE SIGLAS E ABREVIATURAS

NB *Naive Bayes*

PCA Análise Componentes Principais

SPCA Análise Componentes Esparsos

AD Análise Discriminante

BAG *Bagging*

RF *Random Forest*

RNA Rede Neural Artificial

BSEJ *Backward Sequential Elimination and Joining*

KDB *k-dependence Bayesian classifiers*

TAN *Tree Augmented Naive Bayes*

SP-TAN *SuperParent*

SNNB *Selective Neighborhood based Naive Bayes*

AODE *Averaged One-Dependence Estimators*

HNB *Hidden Naive Bayes*

AODE *Averaged One-Dependence Estimators*

HNB *Hidden Naive Bayes*

SUMÁRIO

1. INTRODUÇÃO GERAL

O Brasil ocupa atualmente posições de destaque no cenário mundial de produção agrícola. Pode-se citar como exemplo o primeiro lugar na produção de soja, terceiro maior produtor de milho grão, quarto em algodão e entre outros produtos agrícolas (USDA, 2022). Porém, a agricultura mundial se vê diante de novos desafios impostos pelo mercado ou pelas contingências. O aumento da população mundial pressiona a produção agrícola para atender essa alta de consumo (Borém, 2013). Também se fará necessário produzir sob condições de escassez hídrica e nutricional visto as mudanças climáticas e o esgotamento dos recursos naturais (Stewart e Lal, 2018). Para atender o aumento da produtividade e sob as condições adversas previstas, o melhoramento genético vegetal e animal são áreas das Ciências Agrárias que estão sendo constantemente desenvolvidas para atender essas e outras necessidades. São áreas multidisciplinares que move instituições públicas e privadas na busca por genótipos mais produtivos e robustos às condições ambientais (Cruz et al., 2011).

A estatística é estratégica para o melhoramento na potencialização de resultados mais rápidos e melhores. Essa área do conhecimento disponibiliza um universo de possíveis análises para os dados agropecuários. Assim, o profissional seleciona, dentro do universo de métodos estatísticos disponíveis, o método mais apropriado para aquele conjunto de dados e com uma finalidade bem definida, como por exemplo, solucionar problemas de predição, classificação, reconhecimento de padrões, realizar inferência, simular dados e entre outras finalidades.

Dentre as abordagens, a solução de um problema de classificação se dá a partir de uma função “classificadora” em que uma nova observação é alocada em uma das classes ou em um dos grupos de interesse. Entre os classificadores mais utilizados estão aqueles com base em árvores de decisão, redes neurais, análise discriminante, máquina de vetores de suporte e o *Naive Bayes* (NB). Cada um desses métodos tem suas pressuposições, procedimentos, formas de implementação em *softwares*, vantagens e desvantagens.

Dentre esses, o NB é um classificador de fácil implementação e que tem apenas a pressuposição de independência entre as variáveis. Comparado aos outros métodos de classificação, o NB tem menor volume de uso em trabalhos publicados (Drury, 2017). Essa baixa exploração do NB constatada em trabalhos frente aos outros métodos é difícil de ser explicada, até mesmo porque o NB tem um uso muito amplo.

Talvez seja pela baixa divulgação ou pelo receio da violação de sua pressuposição, uma vez que a pressuposição de independência entre as variáveis é dificilmente alcançada na prática. Mas ainda é possível encontrar bons trabalhos na área do melhoramento genético que utilizaram o NB (Van der Heide et al. 2019; Xu et al. 2021). Teoricamente, a violação da pressuposição do método sugere que o pesquisador utilize outro método ou até métodos combinados para solução o problema de classificação.

A pressuposição de independência do NB é importante devido a onerosidade ao desdobrar os cálculos das probabilidades condicionais. Assim, em cenários de dependência entre as variáveis, o NB tem baixa acurácia e não deve ser adotado (Lewis, 1998). Existem várias propostas na literatura que tratam do caso de variáveis k-dependentes, especialmente o caso em que $k = 2$ (Sahami, 1996; Friedman et al., 1997). Porém, surgem problemas de precisão devido aproximações que são feitas nos cálculos das probabilidades condicionais. O esforço em encontrar formas de reduzir o efeito da dependência entre as variáveis para manter um bom desempenho do método é explicado principalmente pelo amplo uso e simplicidade do NB.

Uma das possibilidades de manter o bom desempenho do NB em cenários de violação da sua pressuposição é combinando a métodos estatísticos que reduz a correlação entre as variáveis e mantém boa explicação do conjunto de dados. Assim, surgem as possibilidades do uso combinado de NB com as análises de componentes principais e esparsos que garantem, sob normalidade, a independência entre os componentes. Outra metodologia que surge como alternativa de método combinado é a análise discriminante que maximiza a diferença entre o vetor de médias de duas populações.

Assim, este trabalho tem como objetivo propor o uso de classificadores híbridos, isto é, metodologias combinadas ao NB para obter dados transformados visando aumentar a performance classificatória do método. Especificamente, serão avaliados os seguintes classificadores híbridos: i) NB com os escores obtidos pela análise de componentes principais (NB + PCA); ii) NB com os escores obtidos pela análise de componentes esparsos (NB + SPCA); iii) NB com os escores obtidos por análise de discriminante (NB + AD). Finalmente, a eficiência desses classificadores híbridos será comparada com os resultados de acurácia média obtidos por meio do NB, *NBTree*, *Random Forest* (RF), *Bagging* (BAG) e Redes Neurais Artificiais (RNA).

2. REVISÃO DE LITERATURA

2.1. Problemas de classificação no melhoramento genético

Dentro do melhoramento genético surgem problemas de classificação rotineiramente para categorizar plantas ou animas em grupos, tais como saudável ou doente, variedade A ou B, suscetível ou resistente e entre outros interesses. Essas categorizações são importantes para o pesquisador traçar novos ou rever objetivos no programa de melhoramento.

Em um primeiro exemplo no melhoramento animal, Van der Heide et al. (2019) avaliaram a regressão logística múltipla, NB e *Random Forest* na predição de sobrevivência de vacas holandesas a 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, 6 meses após o primeiro parto e 200 dias após o primeiro parto. O NB alcançou a maior acurácia em três dos cinco momentos de avaliação ao longo da vida do animal.

Dentre as aplicações no melhoramento genético vegetal, Xu et al. (2021) avaliou 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 suporte de vetor, 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 suporte de vetor, 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.

2.2. O classificador *Naive Bayes*

O classificador NB é oriundo da Estatística Bayesiana e foi nomeada em homenagem a seu precursor, o reverendo Thomas Bayes (1701-1761). Este classificador é baseado no Teorema de Bayes, aplicado para cálculo de probabilidade

condicionais. Considere os eventos A e B em $(\Omega, \mathbb{F}, \mathbb{P})$. Seja $P(B) > 0$, então o Teorema de Bayes é dado por:

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}. \quad (1)$$

Em que $P(A)$ e $P(B)$ são as probabilidades de ocorrência dos eventos A e B, respectivamente. Já $P(A|B)$ e $P(B|A)$ são as probabilidades condicionais de ocorrer o evento A dado que B ocorreu e de ocorrer o evento B dado que A já ocorreu, respectivamente (Magalhães, 2006). Por meio dessa e outras contribuições de Thomas Bayes, surgiu a Estatística Bayesiana.

Sendo um classificador de fácil implementação no *software* R, 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 Berrar (2018),

$$P(c_k|x_i) = \frac{\prod_{i=1}^{n} P(c_k)P(x_i|c_k)}{P(\mathbf{x})}. \quad (3)$$

Em que o denominador é facilmente calculado através do conjunto de dados. Desta forma, o interesse está apenas no numerador o qual pode ser reescrito em:

$$\prod_{i=1}^{n} P(c_k)P(x_i|c_k) = P(c_k)P(x_1, \ \ldots, x_n|c_k). \quad (4)$$

E segundo Berrar (2018), desdobrando as probabilidades condicionais:

$$= P(c_k)P(x_1|c_k)P(x_2|c_k, x_1)P(x_3, \ \ldots, x_n|c_k, x_1, x_2) \quad (5)$$

$$= P(c_k)P(x_1|c_k)P(x_2|c_k, x_1)P(x_3|c_k, x_1, x_2)P(x_4, \ \ldots, x_n|c_k, x_1, x_2, x_3) \quad (6)$$

$$= P(c_k)P(x_1|c_k)P(x_2|c_k, x_1)P(x_3|c_k, x_1, x_2) \ldots P(x_n|c_k, x_1, x_2, x_3, \ldots, x_{n-1}). \quad (7)$$

Assim, o cálculo das probabilidades condicionais sem pressupor independência é oneroso computacionalmente. Por isso, utiliza-se a pressuposição que a variável x_i é independente da variável x_j para $i \neq j$ e obtém:

$$P(x_i|c_k, x_j) = P(x_i|c_k), \quad (8)$$

para $i \neq j$. Segundo Berrar (2018), reescrevendo a (8) com essa pressuposição é obtido:

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

$$P(c_k, x_i) = P(c_k) \prod_{i=1}^{n} P(x_i|c_k). \quad (10)$$

Apesar de muito importante, a pressuposição de independência entre as variáveis é difícil de ser atingida na prática e justifica o *Naive* ao nome do classificador. Ela é dita ingênua porque é uma pressuposição dificilmente atendida e geralmente as variáveis avaliadas guardam alguma relação entre si. Por meio da equação (10) é obtido a função do classificador dada por:

$$f_i^{NB}(x) = P(c_k) \prod_{j=1}^{n} P\left(x_j|c_k\right) \quad \text{e} \quad (11)$$

$$h(x) = \arg\max_{i\in\{0,\ldots,m-1\}} f_i^{NB}(x). \quad (12)$$

O classificador irá alocar o indivíduo na população i se a probabilidade de boa classificação na população i for máxima (Berrar, 2018).

2.3. *Naive Bayes* e a pressuposição de independência

Lewis (1998) apresentou uma revisão sobre o NB, apresentando a pressuposição de independência, além de a teoria e a aplicação do NB para dados de texto. Ao fim do trabalho, Lewis concluiu que o NB é um bom classificador para recuperação de informações e alcançou bom desempenho em dados de texto. O autor levantou ainda questionamentos das condições necessárias e suficientes para uma boa performance do classificador NB, qual a melhor estratégia para selecionar dados para o NB e quanto o método perde em acurácia ao utilizar dados com variáveis que guardam alguma dependência entre si.

Já Rish (2001) procurou entender características dos dados que afetam o NB. O autor teve como um primeiro resultado o decréscimo na performance do classificador ao aumentar o número de classes. Assim, o número considerado como ótimo foi de duas classes. Outro resultado desse trabalho foi que a acurácia do NB não está diretamente correlacionada com o grau de independência entre as variáveis e sim com a entropia da distribuição das variáveis. Um conjunto de variáveis com baixa entropia possibilita boa performance do classificador. Ainda outro resultado importante desse estudo é que o NB apresentou bom desempenho quando as variáveis são de baixa entropia ou quando existe uma dependência quase determinística entre si. A entropia de uma variável aleatória pode ser interpretada como uma medida da incerteza sobre a variável aleatória, antes de observá-la, ou a

quantidade de informação ganhada após observar a variável aleatória (Shannon, 1948).

Ao fim da década de 90 e início dos anos 2000 surgiram inúmeras propostas na tentativa de melhorar a acurácia do NB diante da dependência entre os atributos de estudo. Entre as várias propostas, o *Semi-Naive Bayes* (Kononenko, 1991) propõe utilizar uma probabilidade condicional com uma dependência de atributo quando constatada. Assim, primeiramente, o algoritmo procura verificar a presença de dependência e quando constatada, é utilizada uma probabilidade condicional que envolva os dois atributos que guardam uma dependência entre si. O autor apresenta uma regra para constatar a dependência entre os atributos e também uma aproximação para a probabilidade condicional entre dois atributos dependentes.

O *NBTree* (Kohavi, 1996) é uma proposta que une as metodologias de árvore de decisão e o NB. Nesta proposta, as folhas da árvore são classificadores NB, ao invés de regras de classificação como em uma árvore comum. O crescimento da árvore é interrompido quando a taxa de erro não reduz ou está em um valor desejado. Importante ressaltar que o *NBTree* sofre com o problema de fragmentação e do *disjunction problem*, ou seja, quando resta poucos dados de treinamento nos ramos mais desenvolvidos ou até nas próprias folhas. Kohavi (1996) obteve resultados promissores quanto ao aumento da acurácia ao avaliar que, em 13 dos 29 conjuntos de dados, o *NBTree* obteve acurácia maior que o NB.

A *Backward Sequential Elimination and Joining* (BSEJ) (Pazzani, 1998) é capaz de gerar novos atributos a partir do produto cartesiano entre dois ou mais atributos existentes com dependência entre si. O novo atributo é geralmente *booleano* ou uma combinação aritmética dos antigos atributos que serão substituídos pelo novo atributo. Assim, faz-se uma avaliação do classificador para avaliar se a substituição foi boa ou não para o aumento da acurácia do NB. Essa avaliação é importante, pois as probabilidades estimadas para os atributos conjuntamente é menos confiável e precisa que a obtida pelo produto das probabilidades individuais. Para tentar tornar o cálculo mais confiável é utilizado a abordagem *wrapper* que une atributos somente quando a nova probabilidade calculada promover o aumento da acurácia avaliada por meio do procedimento de *leave-one-out cross-validation*.

O *k-dependence Bayesian classifiers* (KDB) (Sahami, 1996) flexibiliza a pressuposição de independência entre os atributos ao propor que cada atributo depende de até k outros atributos em estudo. Quando $k = 0$, então o classificador é

o próprio NB e quando k é igual ao número atributos $n-1$, então o cenário é de que nenhum atributo é independente (Sahami, 1996). O KDB necessita de um tempo computacional maior comparado ao NB e classificadores que consideram $k=1$ como, por exemplo, o *SuperParent* (Keogh e Pazzani, 1999) ou o AODE (Zheng et al., 2005).

Na proposta *Selective Bayesian Network Classifiers* (Singh e Provan, 1996) é feito uma seleção de um subconjunto dos dados originais com os atributos que forneçam a maior acurácia. Existe várias formas de encontrar esse subconjunto de atributos. Em uma delas, a K2-AS é a implementada pelos autores e reúne o algoritmo K2 para construção do classificador e o *Attribute Selection* para seleção dos atributos. O K2-AS inicia apenas com a variável de classe e é feito adição dos atributos até que a acurácia do classificador não aumente ao adicionar novos atributos.

O *Tree Augmented Naive Bayes* (TAN) (Friedman et al., 1997) é uma proposta que confere até duas dependências para cada atributo. A primeira dependência é a da classe de interesse e a segunda é para um segundo atributo, chamado de pai, desde que exista a dependência. O pai pode variar entre os atributos e pode existir atributo órfão que é quando aquele atributo não guarda relação com nenhum outro. Para encontrar o pai é utilizado a informação condicional mútua entre os atributos. O *SuperParent* (SP-TAN) (Keogh e Pazzani, 1999) é semelhante ao TAN com a diferença que é utilizado o algoritmo *greedy heuristic* na busca do pai que aumenta a acurácia do classificador e consequentemente aumenta também a complexidade computacional.

A *Lazy Bayesian Rule* (Zheng et al., 1999) é uma proposta que também relaxa a pressuposição de independência entre os atributos. Inicialmente é selecionado um sobconjunto W de variáveis pelo método *heuristic wrapper* que minimiza a taxa de erro do classificador. Portanto, a etapa de seleção é feita antes da classificação e depende dos dados selecionados para treinamento. Assim, assume-se independência entre as variáveis restantes e o subconjunto W de variáveis selecionadas. Porém, ocorre um aumento de complexidade computacional devido a etapa de seleção de variáveis.

A proposta *Selective Neighborhood based Naive Bayes* (SNNB) (Xie et al., 2002) tem um formato parecido com a *Selective Bayesian Network Classifiers*. Esse último procedimento propõe o classificador que utiliza os atributos que retorna a maior acurácia. Já o SNNB seleciona o classificador com maior acurácia selecionando objetos ou indivíduos do conjunto de dados. Portanto, o SNNB produz múltiplos

classificadores com múltiplas observações ou objetos em diferentes raios de vizinhança. Então é feito a seleção do classificador que retorna a maior acurácia em um determinado subconjunto dos dados.

Uma outra abordagem é *Averaged One-Dependence Estimators* (AODE) por Zheng et al. (2005). Essa proposta estima uma probabilidade condicional considerando uma dependência entre 2 variáveis e utiliza uma classificação média com base em todas as estimativas das condicionais de dois atributos. O uso da média é para reduzir a variância do AODE que é geralmente mais alta que o NB. Essa maior variância do AODE é devido a estimativa da probabilidade condicional de duas variáveis ser calculada em um menor conjunto de dados do que comparado a condicional de uma variável que é calculada pelo NB.

Zhang et al. (2005) propôs o método *Hidden Naive Bayes* (HNB). Para o i-ésimo atributo A_i é calculado um *hidden parent* A_{hp_i}, um peso que mede a influência das outras variáveis naquele i-ésimo atributo. Assim, o A_{hp_i} é o *hidden parent* que representa um valor da dependência daquele atributo em relação aos demais. Por fim, Zhang et al. (2005) avaliou a acurácia média em 36 conjuntos de dados, onde o HNB obteve a maior acurácia média em quatro conjuntos de dados e o NB em outros cinco desses conjuntos. Neste estudo, o método que obteve melhor desempenho foi o AODE que apresentou maior acurácia média em doze dos 36 conjuntos de dados avaliados com diferentes quantidades de observações e atributos.

E ainda atualmente surgem propostas para continuar a melhorar o NB frente a cenários que sua pressuposição é violada. Nesse trabalho foram avaliadas metodologias combinadas ao NB que podem manter uma boa eficiência do método quando sua pressuposição não é atendida. Duas dessas metodologias são a análise de componentes principais (PCA) e a análise de componentes esparsos (SPCA) que produzem componentes não correlacionados entre si mesmo que os dados originais tenham uma multicolinearidade elevada. Outra metodologia combinada ao NB foi a análise de discriminante (AD) que para cada observação gera uma combinação linear das características observadas com o maior poder de discriminação entre os grupos. Também foi avaliado o *NBTree*, algoritmo proposto por Kohavi (1996) que combina o NB com árvore de decisão. Finalmente, para fim de comparação foram avaliados os resultados obtidos com o NB, das propostas de híbridos com os resultados de

algoritmos de aprendizado de máquina tais como *Random Forest* (RF), *Bagging* (BAG) e Rede Neural Artificial (RNA).

2.4. Métodos estatísticos para obtenção de variáveis não correlacionadas

Na Estatística Multivariada, as análises de componentes são métodos que permitem obter um conjunto de variáveis latentes (não observáveis e não correlacionadas) de tamanho menor ou igual ao número de variáveis do conjunto de dados originais. Nestes métodos, a propriedade de ortogonalidade entre as variáveis latentes está presente e é interessante pelo fato de atender a pressuposição de independência do NB sob a distribuição normal (Papoulis, 1991). Foram utilizados nesse trabalho, para fim de avaliação, os componentes principais e esparsos.

É importante destacar que a variabilidade total dos dados de p variáveis será explicada por p componentes. Mas, consegue-se alcançar a explicação de bom percentual da variabilidade inicial dos dados por menos do que p componentes. Assim, o pesquisador utiliza um número de componentes necessário para explicar certa quantidade desejada da variação dos dados originais.

2.5. Análise de componentes principais

A análise de componentes principais foi proposta por Pearson em 1901 e desenvolvida individualmente por Hotelling, que publicou em 1933 um trabalho propondo a análise de fatores. A PCA é resumidamente uma técnica da estatística multivariada que busca a redução de dimensionalidade das variáveis originais em um número reduzido de componentes, também denominados variáveis latentes. Assim, a técnica PCA explica a estrutura de variância por meio de combinações lineares de componentes não correlacionados.

A técnica PCA consiste, geometricamente, em uma rotação rígida dos eixos coordenados em novos eixos na direção da máxima variabilidade dos dados. Portanto, torna necessário um menor número de componentes para explicar a mesma quantidade de variabilidade inicial em dados correlacionados comparado a dados com variáveis de baixa correlação.

Para demonstração da obtenção dos componentes no método PCA, considere o vetor aleatório $\mathbf{X}^t = [X_1 \quad X_2 \quad \cdots \quad X_p]$ de uma população com matriz de variâncias e covariâncias dada por $\mathbf{\Sigma}$ com autovalores $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ (Regazzi, 2020). Portanto, os componentes principais serão dados pelas seguintes combinações lineares:

$$CP_1 = c_{11}X_1 + c_{12}X_2 + \cdots + c_{1p}X_p \quad (13)$$

$$CP_2 = c_{21}X_1 + c_{22}X_2 + \cdots + c_{2p}X_p \quad (14)$$

...

$$CP_p = c_{p1}X_1 + c_{p2}X_2 + \cdots + c_{pp}X_p. \quad (15)$$

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

Assim, podemos definir:

$$V(CP_i) = V(\mathbf{c}_i^T\mathbf{X}) = \mathbf{c}_i^T\mathbf{\Sigma}\mathbf{c_i} \quad (16)$$

$$COV(CP_i, CP_j) = COV(\mathbf{c}_i^T\mathbf{X}, \mathbf{c}_j^T\mathbf{X}) = \mathbf{c}_i^T\mathbf{\Sigma}\mathbf{c}_j = \mathbf{0}. \quad (17)$$

Em que $COV(CP_i, CP_j)$ é a covariância do i-ésimo componente com o j-ésimo componente para $i \neq j$, $\mathbf{c}_i^t$ é a matriz transposta dos *loadings* do i-ésimo componente, $\mathbf{X}$ é o vetor de observações e $\mathbf{\Sigma}$ é a matriz de covariâncias (Regazzi, 2020).

Por definição do método, o primeiro componente principal λ_1 deve ser o maior dos componentes. Portanto, o primeiro componente deve ser o máximo da forma quadrática λ dada por

$$\lambda = \frac{\mathbf{c}^T\mathbf{\Sigma}\mathbf{c}}{\mathbf{c}^T\mathbf{c}}. \quad (18)$$

Sob a restrição $\mathbf{c}_i^t\mathbf{c}_i = 1$, os componentes são obtidos a partir da maximização da forma quadrática (18) e é dada pelo sistema de equações homogêneas:

$$(\mathbf{\Sigma} - \lambda_i\mathbf{I})\mathbf{c}_i = \mathbf{0} \quad (19)$$

Encontrando os elementos de $\mathbf{c}_i$, então é obtido o seu componente correspondente:

$$CP_i = c_{i1}X_1 + c_{i2}X_2 + \cdots + c_{ip}X_p. \quad (20)$$

Agora, a partir da equação (19) é obtida a igualdade dada por:

$$\mathbf{\Sigma}\mathbf{c}_i = \lambda_i\mathbf{c}_i \quad (21)$$

Retorna-se em (16) e (17) e obtém os seguintes resultados:

$$V(CP_i) = \mathbf{c}_i^T\mathbf{\Sigma}\mathbf{c_i} = \mathbf{c}_i^T\lambda_i\mathbf{c}_i = \lambda_i\mathbf{c}_i^T\mathbf{c}_i = \lambda_i \quad (22)$$

$$COV(CP_i, CP_j) = \mathbf{c}_i^T\mathbf{\Sigma}\mathbf{c}_j = \mathbf{c}_i^T\lambda_j\mathbf{c}_j = \lambda_j\mathbf{c}_i^T\mathbf{c}_j = \mathbf{0}. \quad (23)$$

Como $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$, então de (22) tem que $V(CP_1) \geq V(CP_2) \geq \ldots \geq 0$. Um segundo resultado importante é que:

$$\sum_{i=1}^{p} V(X_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} V(CP_i) \quad (24)$$

O resultado (23) demonstra que CP_i e CP_j são independentes desde que $\mathbf{c}_i^T\mathbf{c}_j = 0$ e sob normalidade (Papoulis, 1991). E o resultado (24) é útil para demonstrar que os componentes gerados retêm a variabilidade total dos dados. O percentual da contribuição de um componente na explicação da variabilidade total dos dados é fornecido pela importância relativa ($IR(\%)$). Regazzi (2020) define a importância relativa de um componente principal como:

$$IR(\%) = \frac{\lambda_i}{\sum_{i=1}^{p}\lambda_i} \cdot 100 \qquad (25)$$

A importância relativa representa a proporção da variância total explicada pelo componente principal λ_i.

2.6. Análise de componentes esparsos

A interpretação de um componente obtido por PCA pode ser complicada pelo fato de ser uma combinação linear de todas as variáveis originais. Uma forma de contornar o problema de interpretabilidade é desprezar as variáveis que tem *loading* (ou coeficiente) próximo ou igual a zero. Recordando da equação (2.3), o *loading* da primeira variável no primeiro componente, por exemplo, é dado por c_{11}.

Zou, Hastie e Tibshirani (2006) propuseram o método SPCA utilizando a regressão *Elastic Net* (Zou e Hastie, 2005). A regressão *Elastic Net* une conceitos da regressão *Ridge* e *Lasso* (Tibshirani, 1996) (*Least Absolute Shrinkage and Selecion Operator*), sendo aplicada no intuito de reduzir os coeficientes em função da variável latente.

O componente obtido por SPCA é o resultado da combinação linear das variáveis restantes, uma vez que o estimador *Elastic Net* fornece estimativas nulas, descartando as variáveis irrelevantes.

O estimador *Lasso* é dado por:

$$\widehat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\beta}(\|\boldsymbol{y_i} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_1\|\boldsymbol{\beta}\|_1), \qquad (26)$$

sendo $\lambda \geq 0$, em que $\boldsymbol{y_i}$ é o i-ésimo componente principal e $\mathbf{X}$ é a matriz de observações. O estimador *Elastic Net* é dado por:

$$\widehat{\boldsymbol{\beta}}_{en} = (1+\lambda_2)\left\{\arg\min_{\beta}(\|\boldsymbol{y_i} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1)\right\}, \qquad (27)$$

com a restrição $\lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 < t$, para determinado t com λ_1 e λ_2 não negativos. Nota-se que o estimador *Lasso* é um caso especial para quando $\lambda_2 = 0$ do estimador *Elastic Net*. Assim, obtém a seguinte aproximação:

$$\frac{\widehat{\boldsymbol{\beta}}_{en}}{\|\widehat{\boldsymbol{\beta}}_{en}\|} \approx \mathbf{v}_i, \tag{28}$$

sendo $\mathbf{v}_i$ o coeficiente esparso de interesse. É possível variar os valores de t tais que $\mathbf{v}_i$ se aproxime ainda mais de zero.

Uma vez que ocorreu uma modificação nos vetores de *loadings*, esses 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 é apenas aproximada.

2.7. Método estatístico baseado na maximização da diferença entre médias populacionais

Na estatística multivariada, muito frequentemente, tem-se o interesse em classificar indivíduos ou observações com base nas suas variáveis em alguma das p classes da população. E as classes, geralmente, são conhecidas pelo pesquisador. Uma das formas de se fazer essa classificação é utilizando uma função discriminante. Para a separação das classes ou grupos é utilizado uma regra de decisão que irá maximizar a diferença entre as médias das populações. Como, por exemplo, pode-se citar as análises discriminante linear, quadrática, múltipla e a canônica.

A abordagem de um classificador híbrido NB, a partir dos escores obtidos por uma função discriminante merece ser avaliada pelo fato das combinações lineares obtidas maximizar a diferença entre os grupos. A cada observação será gerada uma combinação linear, ou também denominada escore, que será utilizado posteriormente no NB.

2.7.1. Análise discriminante linear de Fisher

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, 2020). Também pode-se dizer que a AD irá buscar alocar um indivíduo em uma população para que a distância generalizada de Mahalanobis seja mínima (Regazzi, 2020).

Suponha um vetor de observações $\mathbf{X}$ proveniente de uma população π_i normalmente distribuída com vetor de médias $\boldsymbol{\mu}_\mathbf{i}$ e matriz de variâncias $\boldsymbol{\Sigma}$. Então sua função densidade de probabilidade, segundo Regazzi (2020), é dada por:

$$f_i(\mathbf{x}) = |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}[(\mathbf{x}-\mu_\mathbf{i})^t \Sigma^{-1}(\mathbf{x}-\mu_\mathbf{i})]}. \quad (29)$$

Em que $D_i^2 = (\mathbf{x} - \mu_\mathbf{i})^t \Sigma^{-1}(\mathbf{x} - \mu_\mathbf{i})$ é a distância generalizada de Mahalanobis. Segundo Regazzi (2020), a Função Discriminante Linear de Fisher é dada por:

$$D(\mathbf{X}) = [\mu_\mathbf{1} - \mu_\mathbf{2}]^t \mathbf{\Sigma}^{-1} \mathbf{X} \quad (30)$$

Para cada observação $\mathbf{x_0}$ inserida em (30) será obtido o escore $D(\mathbf{x_0})$ a ser utilizado no NB. O ponto médio entre as duas médias populacionais, segundo Regazzi (2020), é dado por:

$$m = \frac{1}{2}[\mu_\mathbf{1} - \mu_\mathbf{2}]^t \mathbf{\Sigma}^{-1}[\mu_\mathbf{1} + \mu_\mathbf{2}] = \frac{1}{2}[\mathbf{l}^t \mu_\mathbf{1} + \mathbf{l}^t \mu_\mathbf{2}] = \frac{1}{2}[D(\mu_\mathbf{1}) + D(\mu_\mathbf{2})] \quad (31)$$

A seguinte simplificação da função discriminante é apresentada por Hair (2009):

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \cdots + W_n X_{nk} \quad (32)$$

onde 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. Hair (2009) destaca a semelhança da AD com uma regressão múltipla, salvo que a AD envolve em estabelecer os pesos W_i que melhor discriminará os dois ou mais classes de estudo.

Considerando duas populações π_1 e π_2, uma observação $\mathbf{x_0}$ será alocado na população π_1 em caso:

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} > 1. \quad (33)$$

Equivalente a:

$$D_1^2 < D_2^2. \quad (34)$$

Semelhante, a alocação da observação $\mathbf{x_0}$ se dá na população π_1 se:

$$D(\mathbf{x_0}) \geq m. \quad (35)$$

Uma pressuposição da análise discriminante é a normalidade multivariada. Essa pressuposição é importante para uso do método O (Okamoto, 1963) para cálculo da probabilidade de má classificação. Seja $\widehat{\mathrm{P}}(j|i)$ a probabilidade de classificar uma observação da população π_i em π_j, com $i \neq j$, para duas populações como:

$$\widehat{\mathrm{P}}(2|1) = \widehat{\mathrm{P}}(1|2) = \phi\left(-\frac{D}{2}\right). \quad (36)$$

Em que,

$$\phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}\, dx. \quad (37)$$

Todas as equações apresentadas anteriormente utilizam parâmetros que são dificilmente conhecidos na prática. Portanto, devem ser obtidas estimativas para os vetores de médias e a matriz de covariâncias. Felizmente, as equações e regras de classificação continuam as mesmas ao utilizar os estimadores ao invés dos parâmetros (Ferreira, 2011).

2.9. Algoritmos Híbridos

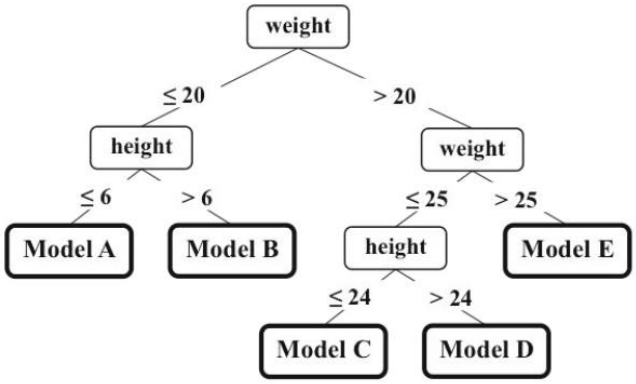
Kohavi (1996) descreveu o mal desempenho do NB em grandes conjuntos de dados como um ponto negativo do classificador. Porém, segundo Kohavi, o NB é robusto a presença de variáveis irrelevantes e mantém um bom desempenho em conjuntos de dados com muitas variáveis. Essa característica é vantajosa para dados que não tem variáveis dominantes sobre as demais. Nos classificadores por árvore de decisão, o autor aponta o problema da partição dos dados que após alguns nós irá esgotar o conjunto de dados de treinamento. Mas, as árvores têm as vantagens, segundo o autor, serem rápidas e robustas a grandes conjuntos de dados.

Assim, Kohavi une NB e árvore de decisão para utilizar de suas vantagens e propor o classificador híbrido *NBTree*. O *NBTree* segue o conceito geral de uma árvore de decisão com a diferença que as folhas são substituídas por classificadores NB. Assim, cada classe que era representada por uma folha, será agora descrita por um NB. Kohavi também afirma que o *NBTree* é mais robusto a ausência de independência que o NB.

No estudo de comparação, Kohavi avaliou a acurácia do NB, *NBTree* e do algoritmo C4.5 em 29 conjuntos de dados. O algoritmo C4.5 foi proposto por Quinlan (2014) e consiste em um algoritmo inicial de árvore de decisão. Por fim, Kohavi avalia que sua proposta de algoritmo é robusta para dados com muitos atributos, mais robusto que o NB para ausência da independência condicional, tem menor número de nós comparado ao C4.5 e tem um ótimo tempo computacional comparado ao NB e C4.5.

Em um exemplo de aplicação, Gajderowicz (2013) utilizou o *NBTree* em seu trabalho de ontologia para classificar felinos (gatos, tigres, panteras e entre outros). Os atributos utilizados na classificação foram altura, peso e *habitat* dos indivíduos. A Figura 3, do trabalho de Gajderowicz, ilustra o funcionamento do *NBTree*. Os modelos de A até E são classificadores NB e em cada nó é mostrado o ponto de corte para tomada de decisão a qual ramo seguir.

Figura 1. Representação do *NBTree* por Gajderowicz (2013), em que o autor discute a utilização de variáveis ordinárias (altura e peso) no modelo de classificação.



2.10. *Bagging* e *Random Forest*

Árvores de decisão possuem alta variabilidade no sentido de que ao utilizar um conjunto de dados de treinamento e obter duas árvores, uma em seguida da outra, a primeira certamente será diferente da segunda. Uma forma de contornar esse problema seria obter várias amostras da população avaliada, obter uma árvore de decisão para cada amostra e obter o valor médio predito por todas as árvores avaliadas. Porém, essa forma de contornar o problema pode ser onerosa em questão de tempo e em recurso computacional. Assim, utiliza-se a técnica *bootstrap* que obtém B amostras com reposição dos dados disponíveis, gera B modelos e finalmente utiliza os B modelos para cálculo do modelo médio dado por:

$$\hat{f}_{médio}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_B(x). \qquad (38)$$

Assim, reduz a variabilidade das árvores construídas. Um alto valor de B não significa *overfitting* e geralmente é utilizado um valor de B tal que a taxa de erro já tenha se estabilizado (James, 2013). Para cálculo do erro utiliza-se geralmente 2/3 das observações para criação do modelo. O restante 1/3 das observações são utilizadas para predição e cálculo da taxa de erro. Dessa forma, é construída a árvore de decisão pelo algoritmo *Bagging* (BAG).

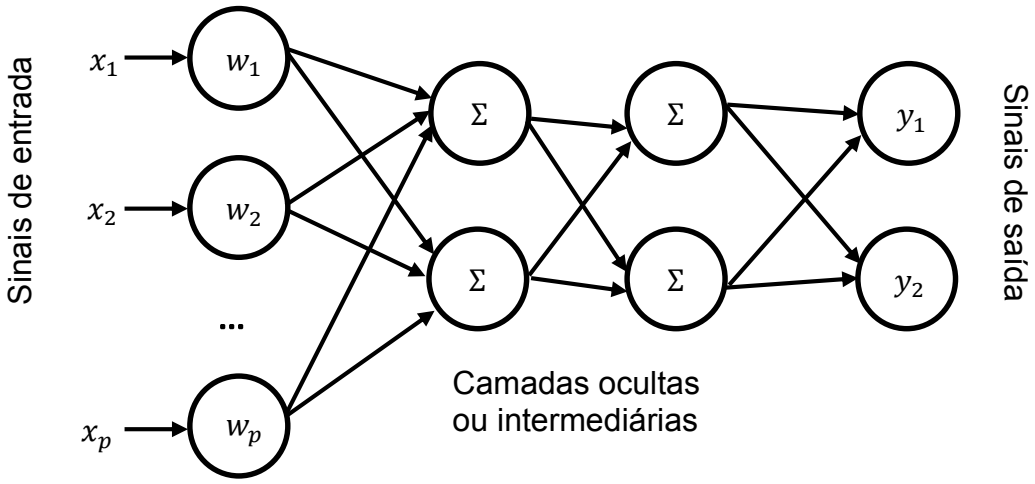
O método *Random Forest* é uma variação do algoritmo BAG. Uma vez que todas as variáveis, geralmente com algum nível de correlação, são utilizadas para produzir o modelo médio. Assim, os valores preditos por BAG são altamente

correlacionados devido as semelhanças nas estruturas das árvores. Para contornar esse problema, pode ser alterado o número de variáveis preditoras m na construção da árvore. Assim, Ho (1995) propõe o RF sugerindo $m = \sqrt{p}$ para problemas de classificação e $m = \frac{p}{3}$ para problemas de regressão, onde p é a quantidade total de variáveis disponíveis no conjunto de dados.

2.11. Rede Neural Artificial

A estrutura de uma rede neural assemelha com o sistema nervoso biológico. A célula do sistema nervoso biológico corresponde a um neurônio, sendo que ele irá emitir um impulso de acordo com o impulso recebido. Semelhante, o neurônio artificial é uma unidade de processamento com uma função de ativação que irá gerar um valor de saída a partir do valor de chegada ao neurônio. Na Figura 2 está ilustrado o modelo básico de uma rede neural artificial.

Figura 2. Representação esquemática de uma Rede Neural Artificial com camada de entrada, duas camadas ocultas e uma camada de saída.



Os sinais de entrada $\{x_1, x_2, \ldots, x_p\}$ são os valores assumidos pelas p variáveis em estudo. Os pesos sinápticos $\{w_1, w_2, \ldots, w_p\}$ representam os valores de ponderação ou de importância para cada sinal de entrada. O combinador linear (Σ) tem como função somar os valores recebidos pela rede ponderados pelos pesos sinápticos. O limiar de ativação (θ) é uma determinada quantidade que o valor gerado pelo combinador linear precisa alcançar para que seja direcionado a saída do neurônio. O potencial de ativação (u) é a diferença entre o valor gerado pelo combinador linear e o limiar de ativação. Em caso de $u \geq 0$, então é gerado um potencial excitatório e o sinal é

transferido. Se $u < 0$, então o potencial gerado é dito inibitório e o sinal não é transferido. A função de ativação (f) é uma função que restringe o intervalo do sinal de saída do neurônio a valores compreendidos no intervalo de sua imagem funcional. As funções de ativação, geralmente, têm imagem em $[0,1]$ ou $[-1,1]$. Na última camada $\{y_1, y_2, \ldots, y_p\}$ existem os sinais de saída que são os valores finais apresentados pela rede após o processamento.

Algumas das funções de ativação utilizadas em redes neurais são:

1. A função sigmoidal definida por:

$$sigmoid(x, \alpha) = \frac{1}{1 + e^{-\alpha x}}. \quad (39)$$

   Em que α é um parâmetro da função sigmoidal. A função sigmoidal é monótona crescente e com imagem no intervalo entre 0 e 1.

2. A função sinal definida por:

$$sgn(x) = \begin{cases} 1 \text{ se } x \geq 0 \\ 0 \text{ se } x < 0 \end{cases}. \quad (40)$$

   A função sinal assumirá valor 0 se o valor x for negativo e o valor 1 em caso de valor x positivo.

3. A tangente hiperbólica definida por

$$tanh\left(\frac{x}{2}\right) = \frac{1 - e^{-x}}{1 + e^{-x}}. \quad (41)$$

   A tangente hiperbólica tem imagem entre -1 e 1, sendo não simétrica em relação a origem. É uma função monótona crescente.

E entre outras.

A rede neural é caracterizada pelo número de camadas ocultas, as conexões entre as camadas, número de neurônios em cada camada, algoritmo de aprendizado e o tipo de conexão entre os neurônios (*feedforward* ou *feedback*).

Quanto ao tipo de conexão entre os neurônios existem redes do tipo *feedforward* e *feedback*. As redes do tipo *feedforward* é quando o fluxo de informação é da entrada para a saída e enquanto nas redes do tipo *feedback* ocorre comunicação da saída com a entrada da rede.

Quanto a sua estrutura, uma rede neural artificial pode ser dividida em camadas de entrada, oculta ou intermediária e de saída (Cruz e Nascimento, 2018). A camada

de entrada é aquela que recebe as informações provenientes dos dados. Já o processamento dessas informações acontece principalmente na camada oculta ou intermediária. Quando uma rede neural possui duas ou mais camadas ocultas, então se diz que a rede é de múltiplas camadas (Cruz e Nascimento, 2018). Em uma rede de múltiplas camadas é possível selecionar a estrutura e quantidade de neurônios que melhor explica a variável resposta. Apesar de existir métodos empíricos para esse dimensionamento (Bullinaria, 2003), a forma mais utilizada para dimensionamento da rede é por tentativa e erro. Por fim, a camada de saída tem neurônios que dá fim ao processamento de informações e apresentam os resultados. Em problemas de classificação, o número de neurônios na camada de saída é igual ao número de classes da população utilizada no problema (Cruz e Nascimento, 2018).

CAPÍTULO 1

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

Resumo: O classificador *Naive Bayes* (NB) se destaca pela sua simplicidade, bom desempenho e sua ingênua pressuposição de independência entre as variáveis preditoras. Tal pressuposição é dificilmente alcançada na prática. Diante disso, esse trabalho tem por objetivo avaliar metodologias combinadas ao NB na tentativa de melhorar seu desempenho considerando diferentes níveis de dependência entre variáveis. As metodologias combinadas com o NB utilizadas foram análise de componentes principais (PCA), componentes esparsos (SPCA) e análise discriminante (AD). Foram simulados dados com diferentes níveis de correlação (0,10; 0,50 e 0,90) e diferentes vetores de médias ($\mu_A = [i \dots i]^T$ e $\mu_B = [j \dots j]^T$, onde $i = 0, 1, 2$ ou 3 e $j = 1, 2$ ou 3, considerando que $i \neq j$). Todos os cenários foram avaliados considerando 2, 4, 8 e 16 variáveis. As metodologias usadas na comparação dos métodos propostos foram *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 a acurácia média.
Palavras-chave: classificador híbrido, metodologias combinadas, simulação.

Abstract: The Naive Bayes (NB) classifier stands out for its simplicity, good performance and also for its naive assumption of independence between the predictor variables. This assumption of independence between the variables is hardly achieved in practice. Therefore, this work aims to evaluate methodologies combined with NB in an attempt to improve its performance in the face of breaking the assumption of independence. The combined methodologies used were principal components analysis (PCA), sparse components (SPCA) and discriminant analysis (AD). The

simulated data had different levels of correlation (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 $i = 0, 1, 2$ or 3 and $j = 1, 2$ or 3, considering $i \neq j$). All scenarios were considered considering 2, 8 and 16 variables. The methodologies used and the comparison of the proposed methods were Random Forest, Bagging and Artificial Neural Network through the calculation of the average accuracy and the standard error of the media. In general, increasing controller variables good average accuracy gains to NB in intuitive downslope scenarios, but little or no gain in accuracy when upadapted. Thus, we conclude that the assumption of independence is important, since the increase in dynamics always resulted in a reduction in the average accuracy of classifieds. The classified hybrids proposed in the work are presented as good alternatives to the NB, since they presented results similar to or superior to the NB itself and other methods evaluated in terms of average accuracy.

Keywords: hybrid assembler, combined methods, simulation.

1. Introdução

A agricultura mundial se vê diante de novos desafios impostos pelo mercado ou pelas contingências. O aumento da população mundial pressiona a produção agrícola para atender essa alta de consumo (Borém, 2013). Para atender o aumento da produtividade e sob as condições adversas previstas, o melhoramento genético vegetal e animal são áreas das Ciências Agrárias que estão sendo constantemente desenvolvidas para atender essas e outras necessidades (Cruz et al., 2011). No melhoramento genético surgem problemas de classificação rotineiramente para categorizar plantas ou animas em grupos, tais como saudável ou doente, variedade A ou B, suscetível ou resistente e entre outros interesses. Essas categorizações são importantes para o pesquisador traçar novos ou rever objetivos no programa de melhoramento.

Existem diversas metodologias para problemas de classificação como, por exemplo, classificadores bayesianos, análise discriminante, árvores de decisão, rede neural artificial, máquina suporte de vetor e entre outros. O *Naive Bayes* se destaca como um classificador simples e eficiente frente as complexas técnicas de aprendizado de máquinas (Cruz e Nascimento, 2018) e até mesmo outros classificadores bayesianos.

Apesar de ter uso amplo (Drury, 2017), o NB frente aos outros métodos de classificação tem menor volume de uso em trabalhos publicados. Essa baixa exploração do NB é difícil de ser explicada. Talvez seja pela sua baixa divulgação ou pelo receio a vioalação de sua pressuposição, uma vez que a pressuposição de independência entre as variáveis é dificilmente atendida na prática.

A pressuposição de independência do NB é importante devido a onerosidade ao desdobrar os cálculos das probabilidades condicionais. Assim, em cenários de dependência entre as variáveis, o NB tem acurácia muita baixa e não deve ser adotado (Lewis, 1998). Existem várias propostas na literatura que tratam do caso de variáveis k-dependentes, especialmente o caso em que $k = 2$ (Sahami, 1996; Friedman et al., 1997;). Porém, surgem problemas de precisão devido aproximações que são feitas nos cálculos das probabilidades condicionais. O esforço em encontrar formas de reduzir o efeito da dependência entre as variáveis para manter um bom desempenho do método é explicado principalmente pelo amplo uso e simplicidade do NB.

Vários estudos sobre sua pressuposição de independência e propostas para contornar o efeito da violação da pressuposição foram publicados no fim da década de 90 e início dos anos 2000. Em um deles, Lewis (1998) levantou questionamentos das condições necessárias e suficientes para uma boa performance do classificador NB, qual a melhor estratégia para selecionar dados para o NB e quanto o método perde em acurácia ao utilizar dados com variáveis que guardam alguma dependência entre si.

O uso de metodologias combinadas ao NB pode ajudar o classificador a manter um bom desempenho em cenários de violação da sua pressuposição. Assim, surge o uso combinado de NB com as análises de componentes principais e esparsos. Nessa combinação, o NB irá utilizar componentes que são não correlacionados oriundos de dados correlacionados por simulação. Outra metodologia que surge como alternativa é a análise discriminante que maximiza a diferença entre o vetor de médias de duas populações. A simulação também precisa levar em consideração a diferença entre os vetores de médias, além dos diferentes níveis de dependência entre as variáveis simulados para avaliar a robustez do NB a sua pressuposição de independência.

Diante do exposto, esse trabalho tem o objetivo de propor o uso de classificadores híbridos, isto é, 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: componentes obtidos por análise de componentes principais (PCA); componentes obtidos por análise de componentes esparsos (SPCA); escores obtidos por análise de discriminante (AD). Por fim, foi feita a comparação da eficiência desses classificadores híbridos com os resultados de acurácia média obtidos por NB, *NBTree*, *Random Forest* (RF), *Bagging* (BAG) e Redes Neurais Artificiais (RNA).

2. Dados simulados

Para avaliar a robustez do NB frente a diferentes níveis de dependência entre variáveis sob a distribuição normal multivariada (DNM) com diferentes níveis de correlação ($\rho = 0{,}10; 0{,}50$ ou $0{,}90$). Destaca-se que sob a DNM, a não correlação implica em independência entre as variáveis (Papoulis, 1991) e torna-se possível avaliar o NB em cenários que atendem ou não a sua pressuposição. O NB foi avaliado quanto a classificação de duas populações com tamanho de amostra $n_A = n_B = 100$ e diferentes números de variáveis ($p = 2, 4, 8$ ou 16) em cada população. As diferentes quantidades de variáveis utilizadas visam entender o efeito do número de variáveis em situações favorável ou não no atendimento da pressuposição do NB.

Os valores paramétricos para os vetores de médias foram:

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

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

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

O valor da variância de todas as variáveis simuladas foi estrategicamente igual a um. Assim, a correlação ρ_{XY} entre duas variáveis será igual a covariância entre elas e o controle da simulação de matrizes de covariâncias no *software* R é bem implementado. Portanto, as matrizes de covariâncias simuladas nesse trabalho tiveram diagonal igual a um e demais elementos da matriz igual a correlação desejada entre as variáveis. Por fim, as matrizes de covariâncias utilizadas nesse trabalho são dadas por:

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

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

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

Destaca-se que as matrizes de covariâncias têm dimensão p x p, onde p é a quantidade de variáveis da população simulada.

Todos os 18 cenários simulados (Tabela 1) foram repetidos para populações com 2, 4, 8 ou 16 variáveis para estudo da importância do número de variáveis para os classificadores.

Tabela 1. Cenários simulados combinando os diferentes níveis de correlação entre variáveis com os diferentes valores paramétricos de média das populações.

Cenário	ρ	μ_1	μ_2	Cenário	ρ	μ_1	μ_2
1	0,1	0	1	10	0,1	1	2
2	0,5	0	1	11	0,5	1	2
3	0,9	0	1	12	0,9	1	2
4	0,1	0	2	13	0,1	1	3
5	0,5	0	2	14	0,5	1	3
6	0,9	0	2	15	0,9	1	3
7	0,1	0	3	16	0,1	2	3
8	0,5	0	3	17	0,5	2	3
9	0,9	0	3	18	0,9	2	3

3. O classificador *Naive Bayes*

Sendo um classificador de fácil implementação no *software* R, 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 Berrar (2018),

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

Em que o denominador é facilmente calculado através do conjunto de dados. Desta forma, o interesse está apenas no numerador o qual pode ser reescrito em:

$$\prod_{i=1}^{n} P(c_k)P(x_i|c_k) = P(c_k)P\left(x_1, \ \ldots, x_n \middle| c_k\right). \qquad (7)$$

E segundo Berrar (2018), desdobrando as probabilidades condicionais:

$$= P(c_k)P(x_1|c_k)P(x_2|c_k, x_1)P\left(x_3, \ \ldots, x_n \middle| c_k, x_1, x_2\right) \qquad (8)$$

$$= P(c_k)P(x_1|c_k)P(x_2|c_k,x_1)P(x_3|c_k,x_1,x_2)P\left(x_4,\ \ldots,x_n\middle|c_k,x_1,x_2,x_3\right) \quad (9)$$

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

Assim, o cálculo das probabilidades condicionais sem pressupor independência é oneroso computacionalmente. Assim, utiliza-se a pressuposição que a variável x_i é independente da variável x_j para $i \neq j$ e obtém:

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

para $i \neq j$. Segundo Berrar (2018), reescrevendo a (8) com essa pressuposição é obtido:

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

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

Apesar de muito importante, a pressuposição de independência entre as variáveis é difícil de ser atingida na prática e justifica o *Naive* ao nome do classificador. Ela é dita ingênua porque é uma pressuposição dificilmente atendida e geralmente as variáveis avaliadas guardam alguma relação entre si. Por meio da equação (13) é obtido a função do classificador dada por:

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

$$h(x) = \arg\max_{i\in\{0,\ldots,m-1\}} f_i^{NB}(x). \quad (15)$$

O classificador irá alocar o indivíduo na população i se a probabilidade de boa classificação na população i for máxima (Berrar, 2018).

Diante do exposto percebe-se a importância do pressuposto de independência na construção do algoritmo e, desta forma, o uso de métodos combinados visando atender tal pressuposto é interessante e merece ser avaliado.

4. Métodos estatísticos combinados ao *Naive Bayes*

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).

A PCA é resumidamente uma técnica da estatística multivariada para redução de dimensionalidade das variáveis originais em poucos componentes ou também denominados variáveis latentes (Regazzi, 2020). Assim, a técnica PCA explica a estrutura de variância por meio de combinações lineares de componentes não correlacionados retendo a variabilidade total dos dados em p componentes (Regazzi, 2020). Os componentes principais serão dados pelas seguintes combinações lineares:

$$CP_p = c_{p1}X_1 + c_{p2}X_2 + \cdots + c_{pp}X_p. \qquad (16)$$

Em que CP_i é o i-ésimo componente ($i = 1,2, \ldots, p$), c_{ij} é o *loading* da j-ésima variável do i-ésimo componente ($j = 1,2, \ldots, p$) e X_j é a j-ésima variável. Por definição do método, o primeiro componente principal λ_1 deve ser o maior dos componentes. Os componentes CP_i e CP_j são independentes desde que $\mathbf{c}_i^t\mathbf{c}_j = 0$ e sob normalidade (Regazzi, 2020).

Zou, Hastie e Tibshirani (2006) propuseram o método SPCA utilizando a regressão *Elastic Net* (Zou e Hastie, 2005). O componente obtido por SPCA é o resultado da combinação linear das variáveis restantes, uma vez que o estimador *Elastic Net* irá fornecer estimativas nulas, descartando aquelas variáveis irrelevantes. O estimador *Elastic Net* é dado por:

$$\widehat{\boldsymbol{\beta}}_{en} = (1 + \lambda_2)\left\{\arg\min_{\beta}(\|\boldsymbol{y_i} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1)\right\} \qquad (17)$$

com a restrição $\lambda_2\|\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 < t$, para determinado t com λ_1 e λ_2 não negativos. Assim, obtém a seguinte aproximação:

$$\frac{\widehat{\boldsymbol{\beta}}_{en}}{\|\widehat{\boldsymbol{\beta}}_{en}\|} \approx \mathbf{v}_i. \qquad (18)$$

Em que $\mathbf{v}_i$ o coeficiente esparso procurado. Uma vez que ocorreu uma modificação nos vetores de *loadings*, esses 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 é apenas aproximada.

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, 2020). Também pode-se dizer que a AD irá buscar alocar um indivíduo em uma população para que a distância generalizada

de Mahalanobis seja mínima (Regazzi, 2020). A seguinte simplificação da função discriminante é apresentada por Hair (2009):

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \cdots + W_nX_{nk}. \quad (19)$$

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.

5. Comparação entre metodologias

Foram utilizados o *NBTree* e os algoritmos de aprendizado de máquina *Bagging*, *Random Forest* e Rede Neural Artificial para comparação com as metodologias híbridas propostas nesse trabalho. Assim, para todos os oito métodos avaliados foi calculada a acurácia média e o desvio padrão da média com base das 50 repetições de cada cenário.

O cálculo da acurácia é fornecido por:

$$A = \frac{VP + VN}{VP + VN + FP + FN},$$

onde VP são verdadeiros positivos, VN são verdadeiros negativos, FP são falsos positivos e FN são falsos negativos. Todos esses valores são obtidos através da matriz de confusão. Por fim, foi calculada a acurácia média através de:

$$\bar{A} = \frac{1}{50}\sum_{i=1}^{50} A_i.$$

O erro padrão da média das 50 repetições foi calculado por:

$$S(\bar{A}) = \frac{S_A}{\sqrt{50}}.$$

Em que S_A é o desvio padrão das 50 repetições avaliadas.

6. Estudo de dimensionamento de rede

Foi feito um estudo de dimensionamento da RNA para estudo da estrutura de rede que fornece o maior valor de acurácia média. Foram avaliadas todas as possíveis estruturas combinando 1 a 20 neurônios na primeira camada oculta com 1 a 20 neurônios na segunda camada oculta, totalizando 400 estruturas de rede avaliadas em todos os cenários. Cada estrutura foi repetida 50 vezes para obtenção da acurácia média e seu erro padrão da média. Na Tabela 2 é apresentado parte das combinações de estrutura avaliadas no cenário 1. Assim, foram avaliadas 20.000 redes em cada

cenário possível da Tabela 1. Considerando todos os cenários possíveis, total de 18 cenários, e os diferentes números de variáveis, total de 4 variáveis, foram avaliadas 1.440.000 redes neurais nesse trabalho.

Tabela 2. Parte das combinações utilizadas no estudo de dimensionamento da estrutura de rede neural realizado no cenário 1. Foram feitas 50 repetições para cálculo da acurácia média e o desvio padrão da média da estrutura.

ρ	μ_1	μ_2	Primeira camada oculta	Segunda camada oculta
0,1	0	1	1	1
0,1	0	1	2	1
0,1	0	1	3	1
0,1	0	1	4	1
0,1	0	1	5	1
0,1	0	1	6	1
0,1	0	1	7	1
0,1	0	1	8	1
0,1	0	1	9	1
0,1	0	1	10	1
...	...	...	...	...
0,1	0	1	20	20

7. Aspectos computacionais

Todo o trabalho de simulação, ajuste e avaliação dos métodos foi conduzido no software R (R Development Core Team, 2022). As populações foram simuladas utilizando a função *rmvnorm* do pacote *mvtnorm* (Genz e Bretz, 2021). Para implementação do NB foi utilizada a função *naiveBayes* do pacote *e1071* (Meyer, 2021). Para obtenção dos componentes principais e esparsos foram utilizadas, respectivamente, as funções *pca* e *spca* do pacote *mixOmics* (Rohart, 2017). Para obtenção dos escores da análise discriminante foi utilizada a função *lda* do pacote *MASS* (Venables, 2002). Para implementação dos classificadores BAG e RF foi utilizada a função *randomForest* do pacote *randomForest* (Liaw e Wiener, 2022). A função *neuralnetwork* do pacote ANN2 (Lammers, 2020) foi a utilizada para implementação da RNA. Ainda foram utilizados os pacotes *caret* (Kuhn, 2021) e

caTools (Tuszynski, 2021) para resolver alguns empecilhos que surgiram na montagem do código. O pacote *ggplot2* (Wickham, 2016) foi utilizado na construção dos gráficos apresentados nos resultados do trabalho.

O classificador híbrido *NBTree* não possui função nativa ou pacote com função disponível para implementação direta na linguagem R. Assim, foi implementado uma conexão da biblioteca *Weka* da linguagem Java no R e a utilização de funções adequadas para se implementar a função. Para isso, foram utilizados os pacotes *rJava* (Urbanek, 2021), *RWekajars* (Hornik, 2019) e *RWeka* (Hornik et al., 2009).

8. Resultados e discussão

Os resultados discutidos serão aqueles obtidos com os cenários de 1 a 9 da Tabela 1, uma vez que os resultados dos cenários 10 a 18 foram muitos semelhantes aos obtidos em 1 a 9 e os seus resultados estão apresentados somente nas tabelas do apêndice do trabalho. Os gráficos *heatmap* apresentam os resultados de acurácia média em que a primeira população tem $\mu_A = 0$ e a média da segunda população varia em $\mu_B = 1,\ 2$ ou 3. As tabelas com os valores de erro padrão da média de cada metodologia também estão apresentadas somente no apêndice.

Na classificação de populações com médias mais discrepantes foi obtido um valor mais elevado de acurácia média comparado a populações de médias mais próximas para qualquer que seja o número de variáveis (Figuras 1, 4, 7 e 10). É observado também que os valores de acurácia média apresentado pelo NB para duas, quatro, oito ou dezesseis variáveis apresentam uma certa proximidade com os outros métodos avaliados e são valores que em geral revelam um bom desempenho dos métodos utilizados (Figuras 3, 4, 5 e 6).

Nas Tabelas 3, 6, 9 e 12 do apêndice é possível avaliar a quantidade de componentes principais e esparsos utilizados para duas, quatro, oito e dezesseis variáveis, respectivamente. O resultado seguiu o esperado de utilizar menor quantidade de componentes em cenários de alta correlação comparado quando em baixa correlação. Nas mesmas tabelas 3, 6, 9 e 12 é possível observar a quantidade de neurônios na primeira e segunda camada da RNA que alcançou a maior acurácia média dentre todas as 400 estruturas avaliadas em cada cenário no estudo de dimensionamento de rede.

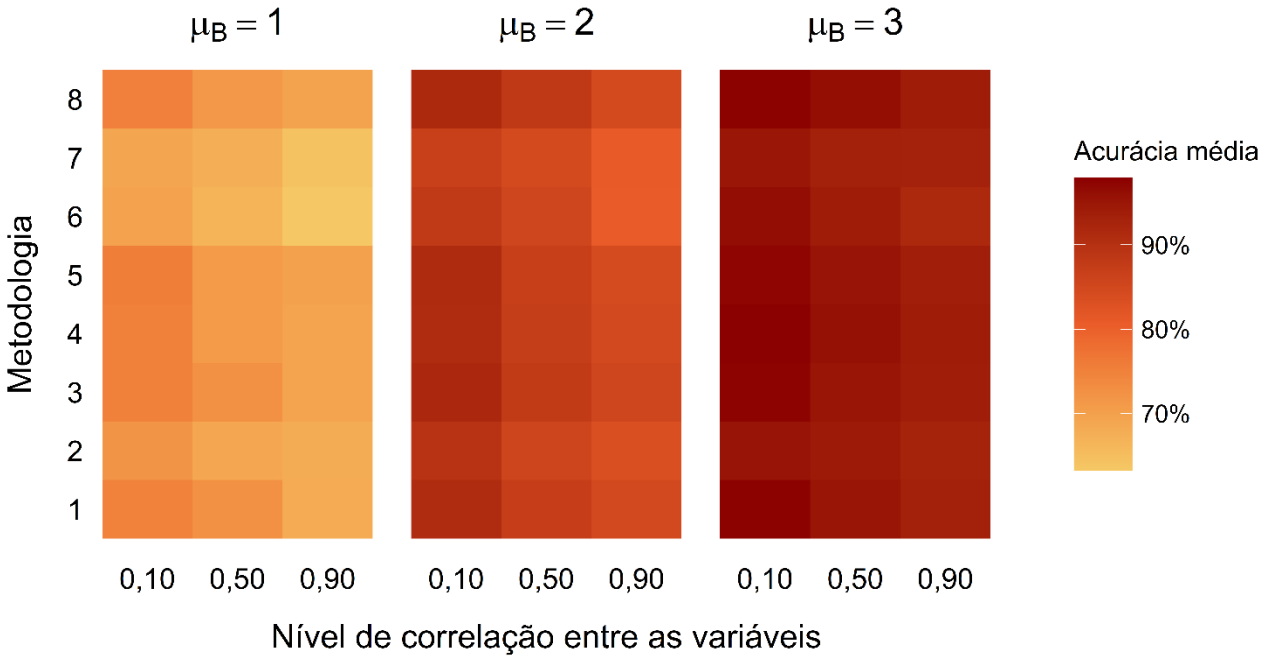
Nos 18 cenários avaliados em populações com duas variáveis (Figura 3), a RNA apresentou o maior valor de acurácia média em 7 cenários, NB + SPCA em 4 cenários, NB + PCA em 3 cenários, NB + AD também em 3 cenários e o NB apenas em 1 cenário avaliado. O classificador NB apresentou acurácia média maior que os classificadores *NBTree*, RF e BAG em todos os cenários avaliados.

Em um mesmo cenário, o maior ganho de acurácia média ao utilizar outro método invés do NB foi de 2,73% no uso de NB + AD no cenário 17. Mantendo as médias constantes, todos as metodologias avaliadas tiveram redução da acurácia média diante do aumento da correlação entre as variáveis (Figura 3).

O menor valor de erro padrão da média considerando duas variáveis e as cinquenta repetições foi de 0,25% utilizando RNA no cenário 3 e o maior foi de 1,11%

com BAG no cenário 18. Esses resultados podem ser conferidos na Tabela 2 do apêndice.

Figura 3 – Gráfico *heatmap* para a acurácia média em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias (8) para classificação de duas populações com duas variáveis normalmente distribuídas em que $\mu_A = 0$. A acurácia média em (8) é a maior acurácia média entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.

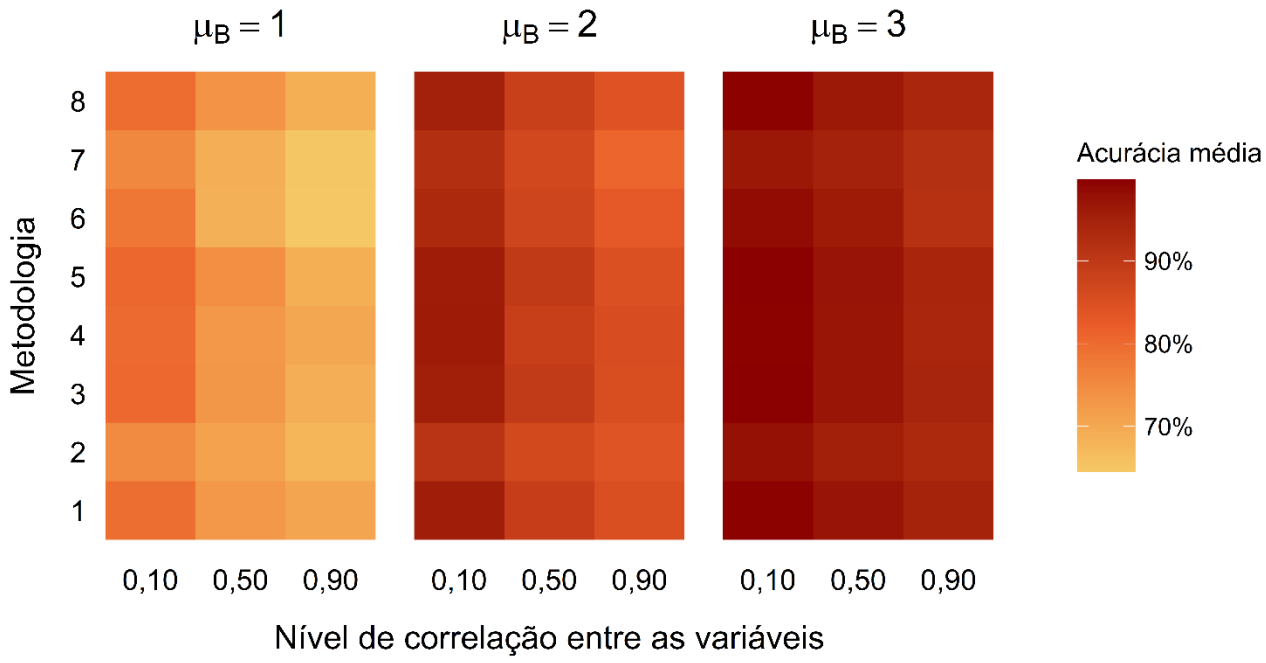


Considerando todos os métodos, a maior perda de acurácia foi de 9,25% quando variou a correlação de 0,10 para 0,90 no classificador RF em populações com médias $\mu_1 = 0$ e $\mu_2 = 1$. Assim, em população com variáveis de médias próximas, aumentar a correlação parece afetar negativamente os métodos de classificação (Figura 3). 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 efeito negativo menor do que quando em médias próximas (Figura 3).

Para os 18 cenários avaliados em populações com quatro variáveis (Figura 4), em 8 cenários o NB obteve a maior acurácia média, em 4 cenários foi o NB + AD,

também em 4 cenários foi o NB + PCA e em 2 cenários foi o NB + SPCA. 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 parecidos com o do presente trabalho com destaque que a proposta do autor demanda maior recurso computacional.

Figura 4 – Gráfico *heatmap* para a acurácia média em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias (8) para classificação de duas populações com quatro variáveis normalmente distribuídas em que $\mu_A = 0$. A acurácia média em (8) é a maior acurácia média entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.



Comparando o NB com os outros classificadores dentro de um mesmo cenário é possível observar que o maior ganho de acurácia foi de 2,06% ao utilizar NB + AD no cenário 2. Esse valor é próximo do 2,73% obtido também com NB + AD na mesma comparação em população com duas variáveis.

O menor valor de erro padrão da média foi de 0,11% ao utilizar RNA no cenário 7 e o maior igual a 1,05% utilizando RF no cenário 3. Novamente um classificador de árvore de decisão obteve o maior erro padrão da média.

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 18,29% utilizando o classificador BAG em populações com médias $\mu_1 = 1$ e $\mu_2 = 2$. Na mesma comparação, a menor redução de acurácia média foi de 4,68% também utilizando BAG e em populações com médias de $\mu_1 = 0$ e $\mu_2 = 3$. Essa comparação mostra novamente que o aumento de correlação em cenário de populações com médias discrepantes tem efeito negativo menor na acurácia média do que quando em populações com médias próximas.

Avaliando populações com oito variáveis (Figura 5), destaca-se que dos 18 cenários avaliados, em 6 cenários o NB + PCA apresentou a maior acurácia média, em 5 cenários o NB + AD, 4 cenários foram NB + SPCA, 2 cenários foram o NB e em 1 cenário foi a RNA.

Ao utilizar oito variáveis, as metodologias híbridas propostas obtiveram valores de acurácia próximos ou melhores que a obtida por Sahami (1996) com sua proposta *k-dependence Bayesian* classifiers. Sahami 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, diferente do que ocorreu em populações de duas e quatro variáveis, apresentaram desempenho melhor que o NB em alguns dos cenários avaliados. A RNA figurou com a melhor acurácia média apenas no cenário 7, aquele em que as populações têm médias discrepantes no mais baixo nível de correlação ($\rho = 0{,}10$). Assim, como aconteceu no resultado com quatro variáveis, a RNA não apresentou resultados muito promissores de acurácia média frente aos outros métodos.

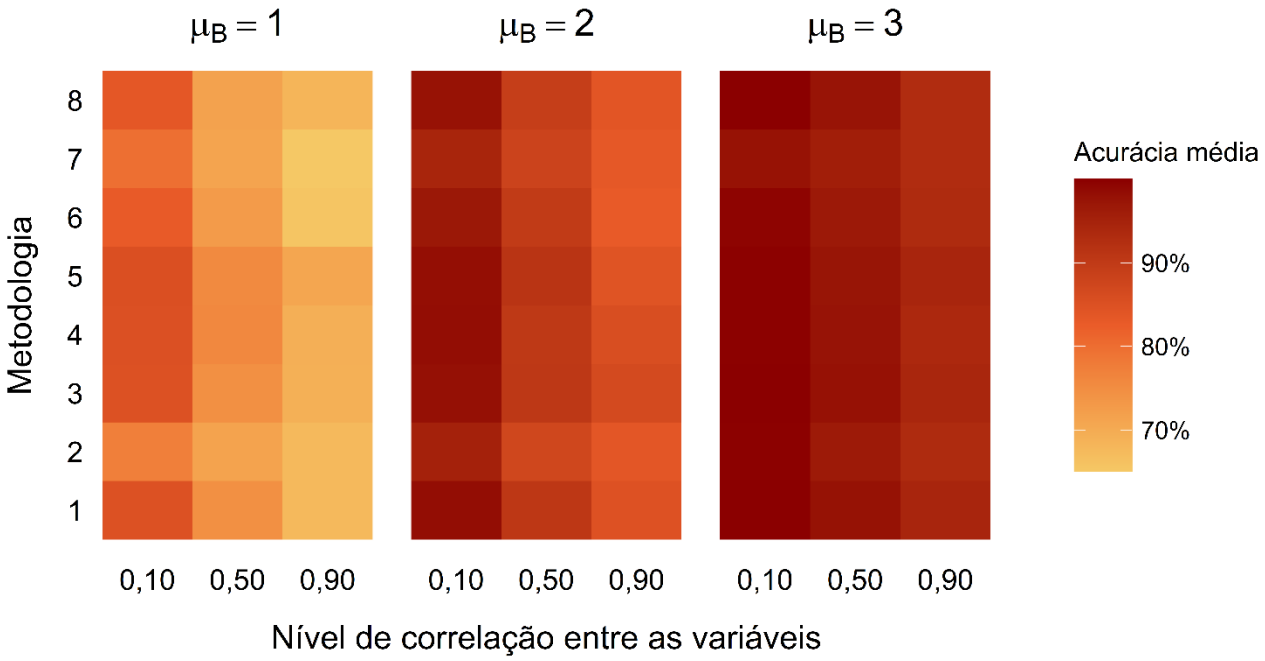
Ressalta-se que o maior ganho de acurácia média foi em 5,54% ao utilizar NB + PCA no cenário 12 na comparação do NB com os outros métodos dentro de um mesmo cenário. Na mesma comparação quando utilizada duas ou quatro variáveis

foram apontados cenários de correlação $\rho = 0{,}50$. Agora com oito variáveis foi possível obter um ganho considerável em cenário de alta correlação ($\rho = 0{,}90$). Esse importante resultado demonstra um ganho de acurácia média quando a pressuposição do NB não é atendida.

Quanto ao erro padrão da média, o menor valor foi de 0,00% ao utilizar RNA no cenário 7 e o maior igual a 1,05% utilizando RF no cenário 3. O primeiro resultado é expressivo pelo fato de que no cenário 7, a RNA obteve acurácia média igual a 100,00% nas cinquenta repetições e que resultou naquele erro padrão da média.

Na comparação mantendo médias constantes e variando a correlação de 0,10 a 0,90, obteve-se a maior redução de acurácia média em 21,65% utilizando RF em populações de médias $\mu_1 = 2$ e $\mu_2 = 3$. E a menor redução foi em 4,84% no uso de BAG em populações com médias $\mu_1 = 0$ e $\mu_2 = 3$. Semelhante aos resultados para duas e quatro variáveis, o aumento da correlação reduziu em maior magnitude a acurácia média em populações com médias próximas do que quando com médias discrepantes.

Figura 5 - Gráfico *heatmap* para a acurácia média em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias (8) para classificação de duas populações com oito variáveis normalmente distribuídas em que $\mu_A = 0$. A acurácia média em (8) é a maior acurácia média entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.



Com dezesseis variáveis (Figura 6), o NB + PCA apresentou a maior acurácia média em 9 cenários, em 3 cenários foi o NB, também em 3 cenários foi o NB + SPCA, em 2 cenários NB + AD e em 1 cenário foi a RNA. 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. Por isso, apresentaram erro padrão da média de 0,00% nesse cenário.

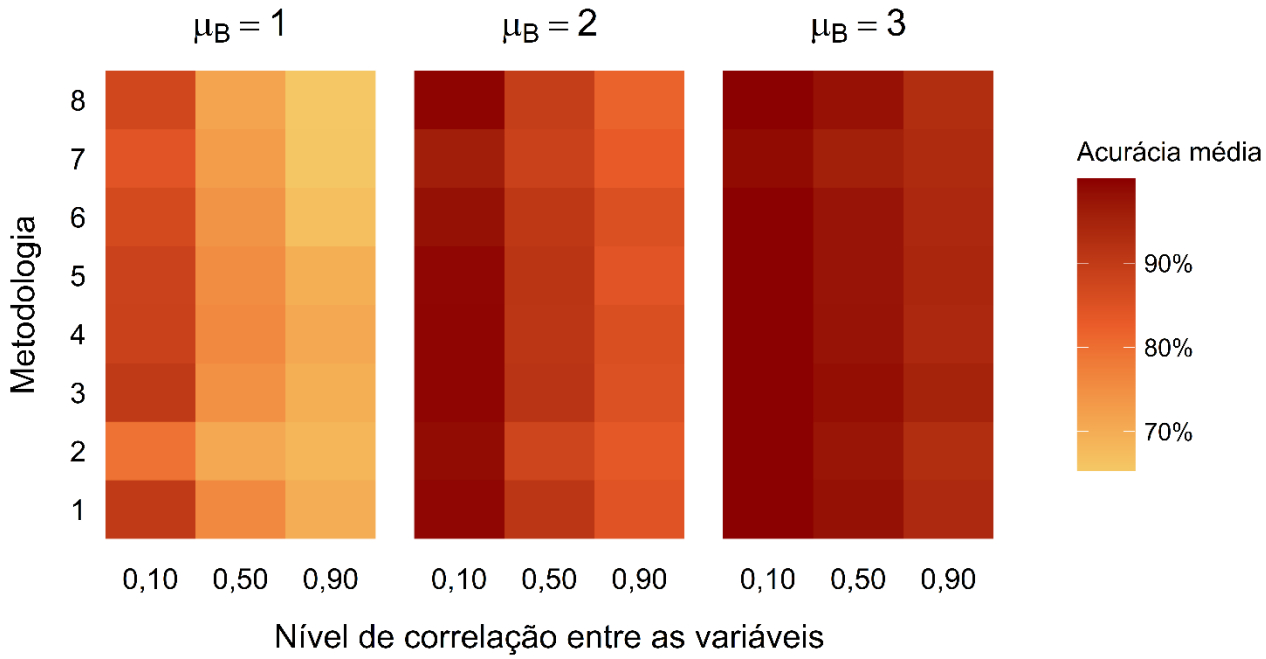
Os valores de acurácia obtidos pelas metodologias propostas com dezesseis variáveis estão próximas ou melhores daqueles obtidos com o *Selective Bayesian Network Classifier* (Singh e Provan, 1996) e o *Hidden Naive Bayes* (Zhang et al., 2005). Porém, os métodos propostos pelos autores apresentam maior 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 cenários. Portanto, os métodos BAG, RF, *NBTree* e RNA não alcançaram bons resultados de forma geral quando comparados aos outros classificadores.

Na avaliação dentro de cada cenário comparando o NB aos outros métodos, alcançou-se o maior ganho de acurácia média em 3,32% utilizando NB + PCA no cenário 18 onde a correlação é elevada ($\rho = 0{,}90$). Essa mesma comparação feita com populações de duas, quatro, oito ou dezesseis variáveis não apresentou nenhum

ganho na acurácia média maior que 4%. Assim, o NB parece ser um classificador tão aceitável quanto os outros avaliados no trabalho.

Figura 6 - Gráfico *heatmap* para a acurácia média em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias (8) para classificação de duas populações com dezesseis variáveis normalmente distribuídas em que $\mu_A = 0$. A acurácia média em (8) é a maior acurácia média entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.



O menor erro padrão da média foi de 0,00% alcançado por diversos métodos anteriormente citados no cenário 7 e o maior valor foi de 0,92% obtido com BAG no cenário 18. É interessante destacar que para qualquer número de variáveis o maior valor de erro padrão da média foi obtido por um classificador em árvore, seja o BAG ou o RF.

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.

9 Resultados e discussão para importância da quantidade de variáveis

Em geral, o aumento de variáveis proporcionou bons ganhos de acurácia média ao NB em cenários de baixa correlação (Tabela 3). Os maiores ganhos de acurácia média foram de 22,15%, 20,72% e 20,43%, e todos esses ganhos obtidos em cenários de correlação igual a 0,10. Esses mesmos três ganhos também foram obtidos em cenários de populações com médias próximas. Portanto, em baixa correlação, parece que o aumento de variáveis melhorou a acurácia do NB. Os três menores ganhos de acurácia média ao aumentar o número de variáveis de duas para dezesseis foram em -2,15%, -0,05% e 0,29%, e todos em cenários de correlação igual a 0,90. Assim, em alta correlação, o aumento do número de variáveis não teve efeito positivo na acurácia média do NB. Resultados similares foram obtidos com os demais classificadores.

Van der Heide et al. (2019) avaliaram a regressão logística múltipla, NB e *Random Forest* na predição de sobrevivência de vacas holandesas a 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, 6 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, a acurácia média obtida pelas metodologias propostas neste trabalho estão próximas ou melhores das obtidas por Van der Heide et al. (2019).

Dentre as aplicações no melhoramento genético vegetal, Xu et al. (2021) avaliou 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 suporte de vetor, 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 suporte de vetor, 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.

Tabela 3. Acurácia média em percentual do classificador *Naive Bayes* para classificação de duas populações com duas ($\overline{A}_{2v}$), quatro ($\overline{A}_{4v}$), oito ($\overline{A}_{8v}$) e dezesseis ($\overline{A}_{16v}$) variáveis normalmente distribuídas.

ρ	μ_1	μ_2	$\overline{A}_{2v}$	$\overline{A}_{4v}$	$\overline{A}_{8v}$	$\overline{A}_{16v}$
0,10	0	1	74,90%	79,55%	85,09%	90,20%
0,50	0	1	72,56%	72,88%	74,68%	75,75%
0,90	0	1	68,24%	70,55%	67,60%	69,93%
0,10	0	2	91,12%	95,73%	98,44%	99,32%
0,50	0	2	87,19%	88,90%	90,70%	91,14%
0,90	0	2	84,75%	85,00%	84,92%	84,71%
0,10	0	3	97,70%	99,50%	99,97%	100,00%
0,50	0	3	95,26%	97,28%	97,87%	98,13%
0,90	0	3	93,47%	94,83%	94,51%	93,74%
0,10	1	2	74,17%	80,18%	84,71%	89,54%
0,50	1	2	72,27%	74,53%	75,03%	75,59%
0,90	1	2	69,36%	69,60%	66,75%	70,79%
0,10	1	3	91,17%	95,75%	98,65%	99,27%
0,50	1	3	86,82%	89,45%	90,57%	91,63%
0,90	1	3	84,99%	85,20%	84,42%	85,27%
0,10	2	3	73,58%	80,63%	85,89%	89,88%
0,50	2	3	70,90%	74,15%	74,93%	75,66%
0,90	2	3	70,50%	69,83%	69,91%	68,63%

10. Conclusões

Assim, pode-se concluir que:

i. A pressuposição de independência é importante, uma vez que o aumento na correlação entre as variáveis sempre resultou em redução da acurácia média dos classificadores;

ii. O NB é um bom classificador quando se compara os valores de acurácia de predição obtidos pelos outros métodos. Nenhum dos classificadores híbridos apresentou resultado muito discrepante do NB;

iii. O aumento do número de variáveis não contornou o problema de não atendimento da pressuposição, uma vez que não proporcionou aumento na acurácia média dos classificadores em cenários de alta correlação;

iv. Ao utilizar menor número de variáveis, redes neurais mais simples apresentaram melhores resultados de acurácia média. Ao aumentar o número de variáveis disponíveis no conjunto de dados houve um aumento na complexidade da rede neural;

v. Os classificadores que utilizam árvore de decisão, BAG ou RF, tem variabilidade na acurácia de predição maior que o NB e os demais classificadores avaliados.

REFERÊNCIAS

BERRAR, D. Bayes' theorem and Naive Bayes classifier. **Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics**, v. 403, p. 412, 2018.

BORÉM, A.; MIRANDA, G. V. **Melhoramento de plantas**. 6nd ed. Viçosa: Universidade Federal de Viçosa, 2013. 523 p.

BULLINARIA, J. A. Apresenta informações sobre Introduction to neural network. Disponível em: http://www.cs.bham.ac.uk/~jxb/inn.html. Acesso em: 13 nov. 2022.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética**. 1st ed. Visconde do Rio Branco: Suprema, 2011. 620 p.

CRUZ, C. D.; NASCIMENTO, M. Inteligência computacional aplicada ao melhoramento genético. Viçosa, MG: Editora UFV, 2018. 414 p.

DRURY, B.; REBAZA-VALVERDE, J.; MOURA, M. F.; LOPES, A. A. A survey of the applications of Bayesian networks in agriculture. **Engineering Applications of Artificial Intelligence**, v. 65, p. 29-42, 2017.

FERREIRA, D. F. **Estatística multivariada**. Lavras: Editora UFLA, 2008. 662 p.

FRIEDMAN, N.; GEIGER, D.; GOLDSZMIDT, M. Bayesian network classifiers. **Machine learning**, v. 29, p. 131-163, 1997.

GAJDEROWICZ, B.; SADEGHIAN, A.; SOUTCHANSKI, M. Ontology enhancement through inductive decision trees. In: **Uncertainty Reasoning for the Semantic Web II: International Workshops URSW 2008-2010 Held at ISWC and UniDL 2010 Held at FLoC, Revised Selected Papers**. Springer Berlin Heidelberg, 2013. p. 262-281.

GENZ, A.; BRETZ F.; MIWA T.; MI X.; LEISCH F.; SCHEIPL, F.; HOTHORN, T. **mvtnorm: Multivariate Normal and t Distributions**, 2021. R package version 1.1-3.

Disponível em: http://CRAN.R-project.org/package=mvtnorm. Acesso em: 13 nov. 2022.

HAIR JR., J. F.; BLACK, W. C.; BABIN, B. J.; Anderson, R. E.; Tathan, R. L. **Análise Multivariada de Dados**. 6nd ed. Porto Alegre: Bookman, 2009. 688 p.

HO, T. K. Random decision forests. In: **Proceedings of 3rd international conference on document analysis and recognition**. IEEE, 1995. p. 278-282.

HORNIK, K. **RWekajars: R/Weka Interface Jars**, 2019. R package version 3.9.3-2. Disponível em: https://CRAN.R-project.org/package=RWekajars. Acesso em: 13 nov. 2022.

HORNIK, K.; BUCHTA, C.; ZEILEIS, A. Open-Source Machine Learning: R Meets Weka. **Computational Statistics**, v. 24, p. 225-232, 2009.

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of educational psychology**, v. 24, n. 6, p. 417, 1933.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to Statistical Learning with Applications in R**. 1st ed. New York: Springer, 2013. 426 p.

KOHAVI, R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. **AAAI Press**, p. 202-207, 1996.
KONONENKO

KONONENKO, I. Semi-naive Bayesian classifier. In: **Machine Learning—EWSL-91: European Working Session on Learning Porto, Portugal, March 6–8, 1991 Proceedings 5**. Springer Berlin Heidelberg, 1991. p. 206-219.

KUHN, M. **caret: Classification and Regression Training**, 2021. R package version 6.0-90. Disponível em: https://CRAN.R-project.org/package=caret. Acesso em: 13 nov. 2022.

LAMMERS, B. **ANN2: Artificial Neural Networks for Anomaly Detection**, 2020. R package version 2.3.4. Disponível em: https://CRAN.R-project.org/package=ANN2. Acesso em: 13 nov. 2022.

LEWIS, D. D. Naive (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.

LIAW, A.; WIENER, M. Classification and Regression by randomForest. **R News**, v. 2, n. 3, p. 18-22, 2002.

MAGALHÃES, M. N. **Probabilidade e variáveis aleatórias**. 2° ed. São Paulo: Editora da Universidade de São Paulo, 2006. 411 p.

MEYER D.; DIMITRIADOU, E.; HORNIK, K.; WEINGESSEL A.; LEISCH, F. **e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien**, 2021. R package version 1.7-9. Disponível em: https://CRAN.R-project.org/package=e1071. Acesso em: 13 nov. 2022.

OKAMOTO, M. An asymptotic expansion for the distribution of the linear discriminant function. **The Annals of Mathematical Statistics**, p. 1286-1301, 1963.

PAPOULIS, A. **Probability, Random Variables and Stochastic Processes**. 3nd ed. New York: McGraw Hill Higher Education, 1991. 624 p.

PAZZANI, M. J. Constructive induction of Cartesian product attributes. **Feature Extraction, Construction and Selection: A Data Mining Perspective**, p. 341-354, 1998.

PEARSON, K. LIII. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin philosophical magazine and journal of science**, v. 2, n. 11, p. 559-572, 1901.

QUINLAN, J. Ross. **C4. 5: programs for machine learning**. Elsevier, 2014. 302 p.

R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2021. Disponível em: https://www.R-project.org/. Acesso em: 13 nov. 2022.

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

RISH, I. An empirical study of the naive Bayes classifier. In: **IJCAI 2001 workshop on empirical methods in artificial intelligence**. 2001. p. 41-46.

ROHART, F.; GAUTIER, B.; SINGH, A.; CAO, L. mixOmics: An R package for 'omics feature selection and multiple data integration. **PLoS computational biology**, v. 13, n. 11, p. e1005752, 2017.

SAHAMI, M. Learning Limited Dependence Bayesian Classifiers. In: **KDD**. 1996. p. 335-338.

SHANNON, C. E. **A Mathematical Theory of Communication**. Bell System Technical Journal, n. 27, p. 379-423, 1948.

SINGH, M.; PROVAN, M.; LANGLEY, P. Induction of selective bayesian network classifiers. **Machine Learning**, v. 2, 1996.

STEWART, B. A.; LAL, R. **Soil and Climate**. 1st ed. Boca Raton: CRC Press, 2018. 448 p.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 1, p. 267-288, 1996.

TUSZYNSKI, J. **caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc.**, 2021. R package version 1.18.2. Disponível em: https://CRAN.R-project.org/package=caTools. Acesso em: 13 nov. 2022.

URBANEK, S. **rJava: Low-Level R to Java Interface**, 2021. R package version 1.0-6. Disponível em: https://CRAN.R-project.org/package=rJava. Acesso em: 13 nov. 2022.

UNITED STATES DEPARTMENT OF AGRICULTURE NATIONAL AGRICULTURAL STATISTICS SERVICE (2022) AGRICULTURAL STATISTICS, USDANASS. Disponível em: https://www.nass.usda.gov. Acesso em: 18 jun. 2022.

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.

VENABLES, W. N.; RIPLEY, B. D. **Modern Applied Statistics with S**. 4nd ed. New York: Springer, 2002. 498 p.

WICKHAM, H. **ggplot2: Elegant Graphics for Data Analysis**, 2016. Disponível em: https://CRAN.R-project.org/package=ggplot2. Acesso em: 13 nov. 2022.

XIE, Z.; HSU, W.; LIU, Z.; LEE, M. L. Snnb: A selective neighborhood based naive Bayes for lazy learning. In: **Advances in Knowledge Discovery and Data Mining: 6th Pacific-Asia Conference, PAKDD 2002 Taipei, Taiwan, May 6–8, 2002 Proceedings 6**. Springer Berlin Heidelberg, 2002. p. 104-114.

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.

ZHANG, Harry; JIANG, Liangxiao; SU, Jiang. Hidden naive bayes. In: **Aaai**. p. 919-924, 2005.

ZHENG, F.; WEBB, G. I. Averaged One-Dependence Estimators, 2010.

ZHENG, Z.; WEBB, G. I. Lazy learning of Bayesian rules. **Machine Learning**, v. 41, p. 53-84, 2000.

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.

ZOU, H.; HASTIE, T.; TIBSHIRANI, R. Sparse principal component analysis. **Journal of Computacional and Grahical Statistics**, v. 15, n. 2, p. 265-286, 2006.

APÊNDICE

Tabela 1. Acurácia média em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias (8) para classificação de duas populações com dezesseis variáveis normalmente distribuídas. A acurácia média em (8) é a maior acurácia média entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.

ρ	μ_1	μ_2	$\overline{A}_1$	$\overline{A}_2$	$\overline{A}_3$	$\overline{A}_4$	$\overline{A}_5$	$\overline{A}_6$	$\overline{A}_7$	$\overline{A}_8$
0,10	0	1	74,90%	72,10%	75,10%	75,16%	75,65%	69,71%	69,30%	75,25%
0,50	0	1	72,56%	69,10%	72,75%	70,99%	70,99%	66,78%	67,69%	71,46%
0,90	0	1	68,24%	68,04%	69,61%	69,50%	69,99%	63,26%	64,38%	69,77%
0,10	0	2	91,12%	89,70%	91,95%	91,07%	91,37%	88,07%	86,90%	91,73%
0,50	0	2	87,19%	85,75%	87,91%	87,35%	87,05%	85,56%	84,57%	88,27%
0,90	0	2	84,75%	83,39%	85,50%	84,80%	84,49%	80,80%	81,00%	84,63%
0,10	0	3	97,70%	95,47%	97,57%	97,90%	97,27%	96,42%	95,05%	97,65%
0,50	0	3	95,26%	94,45%	95,29%	96,01%	95,41%	94,07%	93,36%	96,23%
0,90	0	3	93,47%	93,09%	93,94%	94,09%	93,78%	91,72%	93,14%	94,07%
0,10	1	2	74,17%	70,81%	74,99%	75,36%	74,18%	68,71%	68,97%	75,77%
0,50	1	2	72,27%	68,21%	71,62%	73,33%	71,40%	66,47%	65,17%	72,27%
0,90	1	2	69,36%	68,35%	69,19%	69,30%	68,53%	64,77%	63,98%	69,57%
0,10	1	3	91,17%	88,21%	90,94%	91,64%	90,80%	87,37%	87,67%	90,92%
0,50	1	3	86,82%	85,35%	86,92%	87,59%	87,74%	84,65%	83,96%	88,21%
0,90	1	3	84,99%	83,84%	84,54%	83,94%	85,19%	82,15%	82,15%	86,03%
0,10	2	3	73,58%	70,44%	75,18%	74,37%	74,87%	68,81%	69,07%	75,30%
0,50	2	3	70,90%	68,17%	70,40%	71,38%	72,83%	66,71%	65,45%	71,43%
0,90	2	3	70,50%	68,84%	70,24%	69,08%	68,43%	63,75%	62,85%	69,22%

Tabela 2. Erro padrão da média da acurácia em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias não regularizadas (8) para classificação de duas populações com duas variáveis normalmente distribuídas. A $S(\bar{X}_8)$ é o erro padrão da média da rede com maior acurácia entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.

ρ	μ_1	μ_2	$S(\bar{X}_1)$	$S(\bar{X}_2)$	$S(\bar{X}_3)$	$S(\bar{X}_4)$	$S(\bar{X}_5)$	$S(\bar{X}_6)$	$S(\bar{X}_7)$	$S(\bar{X}_8)$
0,10	0	1	0,70%	0,75%	0,69%	0,72%	0,73%	0,69%	0,86%	0,82%
0,50	0	1	0,79%	0,85%	0,75%	0,62%	0,84%	0,83%	0,72%	0,72%
0,90	0	1	0,51%	0,72%	0,78%	0,80%	0,84%	0,97%	0,85%	0,74%
0,10	0	2	0,52%	0,50%	0,50%	0,58%	0,52%	0,47%	0,55%	0,52%
0,50	0	2	0,64%	0,61%	0,63%	0,53%	0,58%	0,61%	0,72%	0,54%
0,90	0	2	0,52%	0,63%	0,67%	0,72%	0,72%	0,75%	0,63%	0,54%
0,10	0	3	0,27%	0,28%	0,24%	0,34%	0,31%	0,50%	0,37%	0,24%
0,50	0	3	0,33%	0,26%	0,30%	0,40%	0,33%	0,53%	0,42%	0,31%
0,90	0	3	0,43%	0,36%	0,42%	0,52%	0,38%	0,48%	0,53%	0,46%
0,10	1	2	0,84%	0,79%	0,80%	0,88%	0,67%	0,92%	0,69%	0,76%
0,50	1	2	0,65%	0,77%	0,74%	0,86%	0,59%	0,79%	0,81%	0,72%
0,90	1	2	0,76%	0,91%	0,75%	0,91%	0,74%	0,88%	0,92%	0,77%
0,10	1	3	0,49%	0,55%	0,88%	0,63%	0,44%	0,63%	0,64%	0,42%
0,50	1	3	0,62%	0,62%	0,84%	0,66%	0,59%	0,62%	0,66%	0,52%
0,90	1	3	0,73%	0,75%	0,82%	0,60%	0,60%	0,74%	0,84%	0,48%
0,10	2	3	0,81%	0,88%	0,47%	0,82%	0,68%	0,94%	0,77%	0,72%
0,50	2	3	0,67%	0,76%	0,54%	0,84%	0,68%	1,05%	0,77%	0,78%
0,90	2	3	0,70%	0,80%	0,57%	0,90%	0,71%	1,11%	0,90%	0,74%

Tabela 3. Quantidade de componentes utilizados na PCA (CP_{PCA}) e SPCA (CE_{SPCA}) para alcançar o mínimo de 80% da variabilidade dos dados originais explicada pelos componentes em cenários de duas variáveis normalmente distribuídas. Quantidade de neurônios na primeira (C_1) e segunda camada (C_2), respectivamente, da rede neural artificial com maior acurácia entre todas as 400 estruturas avaliadas para cada cenário.

ρ	μ_1	μ_2	CP_{PCA}	CE_{SPCA}	C_1/C_2	ρ	μ_1	μ_2	CP_{PCA}	CE_{SPCA}	C_1/C_2
0,10	0	1	2	2	3/1	0,10	1	2	2	2	1/3
0,50	0	1	1	2	1/6	0,50	1	2	1	2	1/6
0,90	0	1	1	2	1/4	0,90	1	2	1	2	1/5
0,10	0	2	2	1	2/1	0,10	1	3	2	1	1/1
0,50	0	2	1	1	2/6	0,50	1	3	1	1	1/8
0,90	0	2	1	1	2/13	0,90	1	3	1	1	1/3
0,10	0	3	1	1	1/13	0,10	2	3	2	2	1/1
0,50	0	3	1	1	1/8	0,50	2	3	1	2	2/2
0,90	0	3	1	1	11/2	0,90	2	3	1	2	3/1

Tabela 4. Acurácia média em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias (8) para classificação de duas populações com quatro variáveis normalmente distribuídas. A acurácia média em (8) é a maior acurácia média entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.

ρ	μ_1	μ_2	$\overline{A}_1$	$\overline{A}_2$	$\overline{A}_3$	$\overline{A}_4$	$\overline{A}_5$	$\overline{A}_6$	$\overline{A}_7$	$\overline{A}_8$
0,10	0	1	79,55%	74,93%	80,45%	80,35%	80,58%	78,15%	75,55%	79,87%
0,50	0	1	72,88%	71,13%	73,10%	72,85%	74,38%	68,68%	69,03%	73,60%
0,90	0	1	70,55%	67,83%	69,15%	70,30%	68,93%	64,55%	64,93%	68,97%
0,10	0	2	95,73%	91,10%	95,63%	96,03%	95,88%	93,45%	92,23%	95,22%
0,50	0	2	88,90%	86,55%	89,50%	88,73%	89,75%	87,45%	86,68%	88,52%
0,90	0	2	85,00%	84,08%	85,40%	85,58%	84,78%	83,13%	80,95%	84,32%
0,10	0	3	99,50%	97,88%	99,70%	99,48%	99,75%	98,40%	96,60%	99,55%
0,50	0	3	97,28%	95,45%	96,98%	97,10%	97,25%	96,15%	95,08%	96,50%
0,90	0	3	94,83%	93,45%	94,18%	93,88%	94,13%	91,70%	92,08%	93,90%
0,10	1	2	80,18%	74,68%	81,30%	80,35%	79,13%	77,33%	75,88%	79,55%
0,50	1	2	74,53%	71,25%	74,08%	73,68%	73,93%	68,75%	69,65%	72,45%
0,90	1	2	69,60%	68,10%	70,88%	70,35%	67,38%	63,58%	62,00%	69,27%
0,10	1	3	95,75%	92,80%	96,10%	96,00%	96,08%	93,63%	92,18%	95,70%
0,50	1	3	89,45%	86,35%	89,25%	89,18%	89,38%	87,50%	87,10%	89,07%
0,90	1	3	85,20%	83,75%	85,10%	84,75%	84,83%	83,40%	81,73%	85,07%
0,10	2	3	80,63%	74,25%	80,45%	79,48%	80,43%	75,83%	75,08%	80,42%
0,50	2	3	74,15%	70,25%	73,53%	73,00%	72,08%	70,25%	67,98%	72,37%
0,90	2	3	69,83%	68,65%	70,08%	68,95%	69,28%	64,10%	62,58%	68,42%

Tabela 5. Erro padrão da média da acurácia em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias não regularizadas (8) para classificação de duas populações com quatro variáveis normalmente distribuídas. A $S(\bar{X}_8)$ é o erro padrão da média da rede com maior acurácia entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.

ρ	μ_1	μ_2	$S(\bar{X}_1)$	$S(\bar{X}_2)$	$S(\bar{X}_3)$	$S(\bar{X}_4)$	$S(\bar{X}_5)$	$S(\bar{X}_6)$	$S(\bar{X}_7)$	$S(\bar{X}_8)$
0,10	0	1	0,59%	0,71%	0,42%	0,90%	0,62%	0,64%	0,66%	0,75%
0,50	0	1	0,73%	0,75%	0,69%	0,80%	0,59%	0,81%	0,86%	0,67%
0,90	0	1	0,63%	0,72%	0,78%	0,76%	0,74%	0,82%	1,05%	0,72%
0,10	0	2	0,26%	0,32%	0,34%	0,56%	0,35%	0,62%	0,34%	0,37%
0,50	0	2	0,54%	0,48%	0,57%	0,56%	0,43%	0,51%	0,56%	0,47%
0,90	0	2	0,48%	0,52%	0,55%	0,48%	0,55%	0,74%	0,68%	0,65%
0,10	0	3	0,11%	0,09%	0,11%	0,28%	0,09%	0,44%	0,15%	0,11%
0,50	0	3	0,29%	0,24%	0,24%	0,33%	0,27%	0,50%	0,27%	0,37%
0,90	0	3	0,36%	0,35%	0,37%	0,42%	0,38%	0,48%	0,53%	0,33%
0,10	1	2	0,56%	0,53%	0,63%	0,72%	0,62%	0,75%	0,71%	0,64%
0,50	1	2	0,67%	0,75%	0,57%	0,70%	0,67%	0,81%	0,84%	0,77%
0,90	1	2	0,69%	0,74%	0,75%	0,79%	0,73%	1,01%	0,84%	0,72
0,10	1	3	0,31%	0,35%	0,31%	0,43%	0,29%	0,44%	0,32%	0,34%
0,50	1	3	0,51%	0,48%	0,41%	0,52%	0,53%	0,59%	0,55%	0,49%
0,90	1	3	0,57%	0,47%	0,55%	0,62%	0,58%	0,67%	0,66%	0,66%
0,10	2	3	0,72%	0,57%	0,63%	0,63%	0,54%	0,90%	0,67%	0,65%
0,50	2	3	0,67%	0,69%	0,66%	0,74%	0,76%	0,90%	0,83%	0,74%
0,90	2	3	0,81%	0,76%	0,68%	0,71%	0,75%	0,93%	0,93%	0,74%

Tabela 6. Quantidade de componentes utilizados na PCA (CP_{PCA}) e SPCA (CE_{SPCA}) para alcançar o mínimo de 80% da variabilidade dos dados originais explicada pelos componentes em cenários de quatro variáveis. Quantidade de neurônios na primeira (C_1) e segunda camada (C_2), respectivamente, da rede neural artificial com maior acurácia entre todas as avaliadas para cada cenário.

ρ	μ_1	μ_2	CP_{PCA}	CE_{SPCA}	C_1/C_2	ρ	μ_1	μ_2	CP_{PCA}	CE_{SPCA}	C_1/C_2
0,10	0	1	3	3	1/1	0,10	1	2	3	3	1/2
0,50	0	1	2	2	1/1	0,50	1	2	2	2	1/1
0,90	0	1	1	1	2/7	0,90	1	2	1	1	1/1
0,10	0	2	2	3	1/13	0,10	1	3	2	3	1/6
0,50	0	2	1	1	1/6	0,50	1	3	1	1	1/1
0,90	0	2	1	1	2/3	0,90	1	3	1	1	1/10
0,10	0	3	1	2	4/7	0,10	2	3	3	3	1/1
0,50	0	3	1	1	1/12	0,50	2	3	2	2	1/4
0,90	0	3	1	1	1/4	0,90	2	3	1	1	1/4

Tabela 7. Acurácia média em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias (8) para classificação de duas populações com oito variáveis normalmente distribuídas. A acurácia média em (8) é a maior acurácia média entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.

ρ	μ_1	μ_2	$\overline{A}_1$	$\overline{A}_2$	$\overline{A}_3$	$\overline{A}_4$	$\overline{A}_5$	$\overline{A}_6$	$\overline{A}_7$	$\overline{A}_8$
0,10	0	1	85,09%	77,27%	85,00%	85,14%	85,42%	82,96%	79,78%	83,56%
0,50	0	1	74,68%	71,42%	74,62%	75,83%	75,66%	72,76%	71,19%	71,64%
0,90	0	1	67,60%	67,59%	69,15%	69,34%	70,94%	65,65%	65,06%	68,29%
0,10	0	2	98,44%	95,26%	98,11%	98,44%	98,32%	96,89%	94,60%	97,80%
0,50	0	2	90,70%	87,30%	90,64%	90,42%	91,38%	89,85%	88,06%	89,28%
0,90	0	2	84,92%	83,80%	86,59%	85,69%	84,42%	82,90%	83,34%	84,04%
0,10	0	3	99,97%	99,82%	99,94%	99,88%	99,85%	99,46%	97,78%	100,00%
0,50	0	3	97,87%	96,61%	97,90%	97,63%	97,45%	96,79%	95,86%	97,54%
0,90	0	3	94,51%	93,40%	94,33%	93,88%	94,53%	93,35%	93,04%	93,11%
0,10	1	2	84,71%	77,63%	85,84%	84,94%	86,06%	82,35%	80,24%	84,59%
0,50	1	2	75,03%	71,94%	74,40%	73,70%	74,19%	72,05%	70,31%	72,70%
0,90	1	2	66,75%	67,89%	70,45%	69,82%	70,00%	66,76%	65,01%	66,75%
0,10	1	3	98,65%	95,74%	98,59%	98,44%	98,44%	97,00%	93,71%	97,95%
0,50	1	3	90,57%	87,78%	91,36%	91,07%	90,32%	88,29%	88,49%	89,83%
0,90	1	3	84,42%	84,43%	85,65%	85,69%	85,29%	83,31%	82,93%	83,67%
0,10	2	3	85,36%	77,75%	85,89%	85,81%	85,33%	82,50%	80,36%	84,89%
0,50	2	3	73,58%	69,10%	74,93%	74,28%	73,97%	72,34%	70,71%	71,88%
0,90	2	3	70,77%	68,83%	69,91%	70,87%	68,76%	64,64%	64,62%	67,12%

Tabela 8. Erro padrão da média da acurácia em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias não regularizadas (8) para classificação de duas populações com oito variáveis normalmente distribuídas. A $S(\bar{X}_8)$ é o erro padrão da média da rede com maior acurácia entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.

ρ	μ_1	μ_2	$S(\bar{X}_1)$	$S(\bar{X}_2)$	$S(\bar{X}_3)$	$S(\bar{X}_4)$	$S(\bar{X}_5)$	$S(\bar{X}_6)$	$S(\bar{X}_7)$	$S(\bar{X}_8)$
0,10	0	1	0,60%	0,61%	0,63%	0,83%	0,52%	0,64%	0,66%	0,78%
0,50	0	1	0,87%	0,68%	0,71%	0,85%	0,86%	0,81%	0,86%	0,79%
0,90	0	1	0,80%	0,84%	0,78%	0,79%	0,69%	0,82%	1,05%	0,95%
0,10	0	2	0,21%	0,21%	0,18%	0,45%	0,23%	0,62%	0,34%	0,28%
0,50	0	2	0,51%	0,48%	0,57%	0,68%	0,49%	0,51%	0,56%	0,51%
0,90	0	2	0,65%	0,51%	0,58%	0,60%	0,74%	0,74%	0,68%	0,71%
0,10	0	3	0,03%	0,04%	0,06%	0,07%	0,06%	0,44%	0,15%	0,00%
0,50	0	3	0,25%	0,21%	0,25%	0,37%	0,21%	0,50%	0,27%	0,27%
0,90	0	3	0,45%	0,40%	0,45%	0,50%	0,45%	0,48%	0,53%	0,49%
0,10	1	2	0,69%	0,72%	0,71%	0,74%	0,55%	0,75%	0,71%	0,60%
0,50	1	2	0,77%	0,65%	1,00%	0,94%	0,81%	0,81%	0,84%	0,91%
0,90	1	2	0,83%	0,76%	0,95%	0,84%	0,97%	1,01%	0,84%	0,83
0,10	1	3	0,22%	0,20%	0,23%	0,40%	0,18%	0,44%	0,32%	0,25%
0,50	1	3	0,42%	0,54%	0,48%	0,57%	0,54%	0,59%	0,55%	0,57%
0,90	1	3	0,62%	0,62%	0,70%	0,70%	0,61%	0,67%	0,66%	0,58%
0,10	2	3	0,68%	0,76%	0,65%	0,84%	0,65%	0,90%	0,67%	0,72%
0,50	2	3	0,76%	0,75%	0,59%	0,75%	0,84%	0,90%	0,83%	0,77%
0,90	2	3	0,83%	0,84%	0,78%	0,96%	0,97%	0,93%	0,93%	0,86%

Tabela 9. Quantidade de componentes utilizados na PCA (CP_{PCA}) e SPCA (CE_{SPCA}) para alcançar o mínimo de 80% da variabilidade dos dados originais explicada pelos componentes em cenários de oito variáveis. Quantidade de neurônios na primeira (C_1) e segunda camada (C_2), respectivamente, da rede neural artificial com maior acurácia entre todas as avaliadas para cada cenário.

ρ	μ_1	μ_2	CP_{PCA}	CE_{SPCA}	C_1/C_2	ρ	μ_1	μ_2	CP_{PCA}	CE_{SPCA}	C_1/C_2
0,10	0	1	6	6	1/7	0,10	1	2	6	6	1/3
0,50	0	1	4	4	1/7	0,50	1	2	4	4	1/11
0,90	0	1	1	1	1/4	0,90	1	2	1	1	1/4
0,10	0	2	4	4	2/10	0,10	1	3	4	4	9/9
0,50	0	2	2	2	1/5	0,50	1	3	2	2	1/6
0,90	0	2	1	1	1/13	0,90	1	3	1	1	2/1
0,10	0	3	2	2	1/8	0,10	2	3	6	6	1/3
0,50	0	3	1	1	12/2	0,50	2	3	4	4	1/1
0,90	0	3	1	1	1/2	0,90	2	3	1	1	1/17

Tabela 10. Acurácia média em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias (8) para classificação de duas populações com dezesseis variáveis normalmente distribuídas. A acurácia média em (8) é a maior acurácia média entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.

ρ	μ_1	μ_2	$\overline{A}_1$	$\overline{A}_2$	$\overline{A}_3$	$\overline{A}_4$	$\overline{A}_5$	$\overline{A}_6$	$\overline{A}_7$	$\overline{A}_8$
0,10	0	1	90,20%	79,56%	90,28%	88,75%	88,38%	86,76%	84,44%	87,31%
0,50	0	1	75,75%	70,78%	74,79%	75,72%	75,29%	74,14%	72,63%	71,56%
0,90	0	1	69,93%	68,55%	69,69%	70,84%	69,66%	66,90%	65,75%	65,35%
0,10	0	2	99,32%	98,61%	99,40%	99,35%	99,32%	98,18%	95,80%	99,48%
0,50	0	2	91,14%	87,86%	91,50%	91,05%	91,40%	90,61%	88,72%	89,42%
0,90	0	2	84,71%	83,74%	85,16%	85,67%	84,36%	85,27%	82,96%	81,67%
0,10	0	3	100,00%	100,00%	100,00%	100,00%	99,97%	99,97%	98,76%	100,00%
0,50	0	3	98,13%	97,38%	98,33%	97,64%	97,62%	97,62%	95,60%	97,79%
0,90	0	3	93,74%	92,86%	95,13%	93,90%	94,25%	93,91%	93,44%	92,79%
0,10	1	2	89,54%	80,25%	88,83%	89,29%	89,53%	86,19%	84,15%	86,82%
0,50	1	2	75,59%	70,07%	75,41%	74,84%	76,28%	73,83%	71,65%	71,44%
0,90	1	2	70,79%	67,49%	71,54%	69,44%	69,96%	68,30%	66,89%	66,03%
0,10	1	3	99,27%	98,53%	99,57%	99,40%	99,24%	98,27%	96,17%	99,46%
0,50	1	3	91,63%	87,66%	91,07%	91,20%	91,28%	90,04%	88,63%	89,38%
0,90	1	3	85,27%	83,32%	85,78%	84,54%	85,00%	83,41%	84,16%	82,05%
0,10	2	3	89,88%	80,27%	90,08%	89,63%	90,22%	85,61%	82,62%	86,62%
0,50	2	3	75,66%	69,07%	75,15%	75,70%	75,54%	73,21%	73,08%	71,42%
0,90	2	3	68,63%	66,42%	70,91%	68,96%	70,81%	65,81%	66,46%	65,77%

Tabela 11. Erro padrão da média da acurácia em percentual do classificador *Naive Bayes* (1), *NBTree* (2), classificador híbrido NB + PCA (3), NB + SPCA (4), NB + AD (5), *Random Forest* (6), *Bagging* (7) e redes neurais artificias não regularizadas (8) para classificação de duas populações com dezesseis variáveis normalmente distribuídas. A $S(\bar{X}_8)$ é o erro padrão da média da rede com maior acurácia entre todas as 400 estruturas de redes avaliadas no estudo de dimensionamento.

ρ	μ_1	μ_2	$S(\bar{X}_1)$	$S(\bar{X}_2)$	$S(\bar{X}_3)$	$S(\bar{X}_4)$	$S(\bar{X}_5)$	$S(\bar{X}_6)$	$S(\bar{X}_7)$	$S(\bar{X}_8)$
0,10	0	1	0,54%	0,52%	0,68%	0,80%	0,57%	0,68%	0,63%	0,55%
0,50	0	1	0,55%	0,66%	0,85%	0,84%	0,79%	0,77%	0,74%	0,96%
0,90	0	1	0,61%	0,79%	0,68%	0,73%	0,70%	0,71%	0,83%	0,71%
0,10	0	2	0,14%	0,12%	0,12%	0,17%	0,14%	0,74%	0,23%	0,15%
0,50	0	2	0,48%	0,46%	0,43%	0,57%	0,46%	0,50%	0,52%	0,51%
0,90	0	2	0,52%	0,57%	0,59%	0,58%	0,57%	0,69%	0,65%	0,71%
0,10	0	3	0,00%	0,00%	0,00%	0,00%	0,03%	0,27%	0,03%	0,00%
0,50	0	3	0,21%	0,21%	0,24%	0,26%	0,24%	0,46%	0,23%	0,29%
0,90	0	3	0,43%	0,37%	0,40%	0,47%	0,40%	0,40%	0,46%	0,50%
0,10	1	2	0,48%	0,61%	0,50%	0,71%	0,54%	0,69%	0,48%	0,51%
0,50	1	2	0,75%	0,69%	0,76%	0,91%	0,67%	0,75%	0,82%	0,86%
0,90	1	2	0,77%	0,71%	0,75%	0,97%	0,79%	0,84%	0,79%	0,86%
0,10	1	3	0,14%	0,12%	0,14%	0,22%	0,15%	0,48%	0,27%	0,13%
0,50	1	3	0,45%	0,50%	0,58%	0,67%	0,52%	0,63%	0,49%	0,51%
0,90	1	3	0,64%	0,65%	0,57%	0,60%	0,61%	0,49%	0,60%	0,69%
0,10	2	3	0,46%	0,63%	0,49%	0,70%	0,57%	0,63%	0,64%	0,61%
0,50	2	3	0,79%	0,71%	0,72%	0,76%	0,62%	0,65%	0,86%	0,84%
0,90	2	3	0,67%	0,71%	0,80%	0,80%	0,79%	0,92%	0,74%	0,88%

Tabela 12. Quantidade de componentes utilizados na PCA (CP_{PCA}) e SPCA (CE_{SPCA}) para alcançar o mínimo de 80% da variabilidade dos dados originais explicada pelos componentes em cenários de dezesseis variáveis. Quantidade de neurônios na primeira (C_1) e segunda camada (C_2), respectivamente, da rede neural artificial com maior acurácia entre todas as avaliadas para cada cenário.

ρ	μ_1	μ_2	CP_{PCA}	CE_{SPCA}	C_1/C_2	ρ	μ_1	μ_2	CP_{PCA}	CE_{SPCA}	C_1/C_2
0,10	0	1	10	11	11/1	0,10	1	2	10	10	17/1
0,50	0	1	7	7	1/7	0,50	1	2	7	7	1/4
0,90	0	1	1	1	1/4	0,90	1	2	1	1	1/7
0,10	0	2	8	8	20/2	0,10	1	3	8	8	20/4
0,50	0	2	3	3	15/1	0,50	1	3	3	3	9/1
0,90	0	2	1	1	1/19	0,90	1	3	1	1	1/17
0,10	0	3	4	4	1/1	0,10	2	3	10	10	17/2
0,50	0	3	1	1	13/15	0,50	2	3	7	7	1/18
0,90	0	3	1	1	15/1	0,90	2	3	1	1	1/3