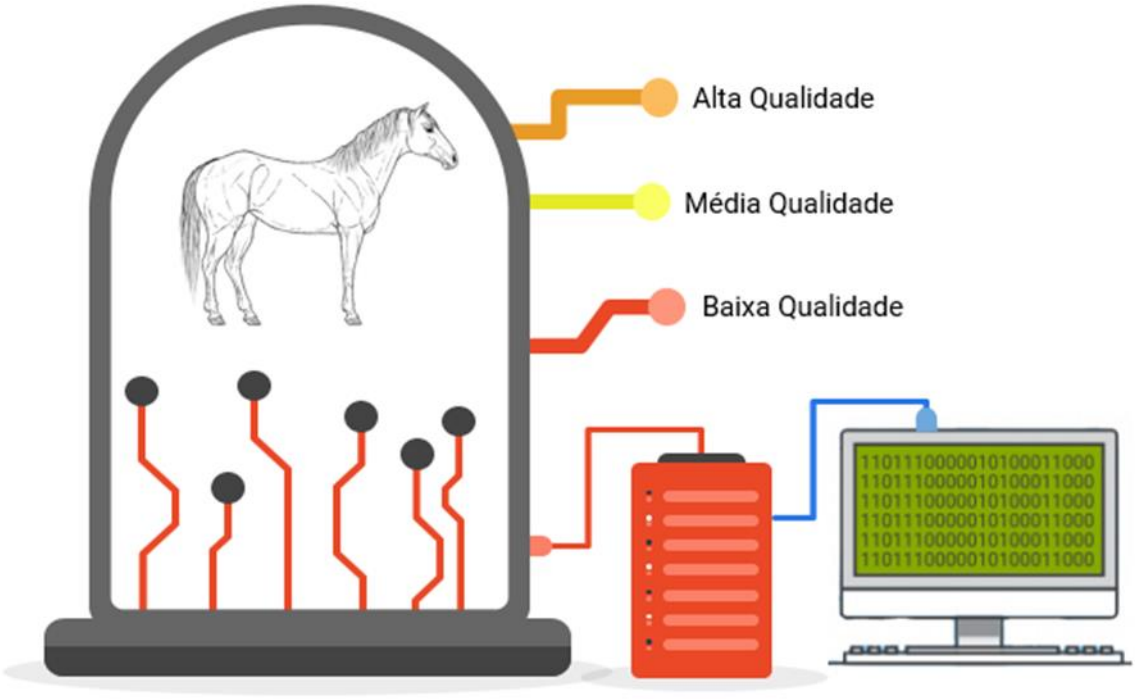
ISSN 1981-7231



COMUNICADO TÉCNICO

123

Aprendizado de máquina aplicado à classificação do Cavalo Pantaneiro usando a combinação do NCA com o algoritmo K-NN

Igor Pinho Souza
Soumaya Ounkhir
Marcel José Soleira Grassi
Claudio Pereira Flores
Otávio Nathanael Campos de Oliveira
Sandra Aparecida Santos
Adriana Mello de Araujo

Corumbá, MS
Outubro, 2023

Embrapa

Aprendizado de máquina aplicado à classificação do Cavalo Pantaneiro usando a combinação do NCA com o algoritmo K-NN

Igor Pinho Souza, graduando em Sistema de Informação Instituto Federal do Mato Grasso do Sul, bolsista CNPq, Corumbá, MS. **igor1_souza@hotmail.com**
Soumaya Ounkhir, Análise e Desenvolvimento de Sistemas, Instituto Federal de Educação de Mato Grosso do Sul, Corumbá, MS. **ounkhir.soumaya@gmail.com**
Marcel José Soleira Grassi, mestre em Ciência da Computação, docente do Instituto Federal de Educação de Mato Grosso do Sul, Corumbá, MS. **marcel.grassi@ifms.edu.br**
Claudio Pereira Flores, mestre em Ciência da Computação, Analista da Embrapa Pantanal, Corumbá, MS. **claudio.flores@embrapa.br**
Otávio Nathanael Campos de Oliveira. Análise e Desenvolvimento de Sistemas, Instituto Federal de Educação de Mato Grosso do Sul, Corumbá, MS. **otavio.nathanael@gmail.com**
Sandra Aparecida Santos, doutora em Zootecnia, pesquisadora da Embrapa Pantanal, Corumbá, MS **sandra.santos@embrapa.br**
Adriana Mello de Araujo, doutora em Genética e Melhoramento, pesquisadora da Embrapa Pantanal, Corumbá, MS. **adriana.araujo@embrapa.br**

Introdução

A raça de cavalo Pantaneiro é reconhecida como uma das mais importantes do Brasil, principalmente por sua multifuncionalidade e excelente adaptação às áreas inundáveis do Pantanal (Santos et al., 1995, 2016). No entanto, a conservação e classificação genética da raça têm sido um desafio devido a muitos anos de cruzamentos inautênticos, resultando em uma grande diversidade de tipos, tanto morfológicos quanto de qualidade.

Para contribuir com a conservação da raça, este estudo propõe a classificação dos cavalos Pantaneiros utilizando o algoritmo de aprendizado de máquina K-NN, que foi treinado com base num conjunto de 222 dados de medidas corporais dos cavalos utilizados pela ABCCP. O pré-processamento dos dados foi realizado com o algoritmo NCA, que utiliza uma técnica de otimização de gradiente para ajustar os pesos das características dos exemplos de treinamento, maximizando a precisão do classificador K-NN.

Além disso, foram realizados testes com diferentes variáveis para identificar quais proporcionam melhores resultados na classificação dos cavalos Pantaneiros. O objetivo geral do trabalho foi permitir que criadores da raça possam identificar mais rapidamente a qualidade de um cavalo com base em suas medidas corporais. Essa técnica é mais uma opção na conservação e seleção genética da raça pantaneira.

A conformação corporal é um dos aspectos mais importantes no julgamento de um animal e pode indicar sua integridade e habilidade física. Animais bem proporcionados reúnem beleza, elegância, distinção e aptidão, qualidades que os tornam adequados para reprodução, trabalho, lazer e práticas esportivas (Balieiro, 1971).

Em resumo, este estudo propõe a utilização do algoritmo K-NN para classificar os cavalos Pantaneiros com base em suas medidas corporais, pré-processando os dados com o algoritmo NCA para maximizar a precisão do classificador. Serão realizados testes com diferentes variáveis para identificar as que proporcionam melhores resultados na classificação dos cavalos Pantaneiros, contribuindo para a conservação e seleção genética da raça.

Os aplicativos desenvolvidos na Embrapa contribuem para o fortalecimento de infraestruturas resilientes e o desenvolvimento da industrialização inclusiva e sustentável, ODS11, pois viabilizam soluções tecnológicas inovadoras e sustentáveis para a agricultura brasileira (Bueno; Torres, 2022).

Material e métodos

Este estudo se baseou em um trabalho anterior de Ounkhir, intitulado "Desenvolvimento de uma

Aplicação para a Classificação de Cavalos Pantaneiros e Estudo Comparativo do Desempenho dos Algoritmos Naive Bayes, K-NN e C 4.5 nesta Classificação" (Ounkhir et al., 2019), onde foi selecionado o algoritmo K-NN e realizado o tratamento dos dados.

Neste trabalho, aprimoramos a metodologia com a incorporação do pré-processamento de dados utilizando o algoritmo NCA (Neighborhood Components Analysis), que é uma técnica que visa melhorar o desempenho do K-NN proposta por Goldberger et al. (2004), onde visa aproximar as distâncias entre amostras através da representação em uma dimensão menor. O NCA maximiza a precisão da classificação por meio da transformação linear, determinada por meio da variante estocástica da precisão de classificação "leave-one-out".

O algoritmo K-Nearest Neighbor (K-NN) é um método de aprendizado supervisionado. Ele trabalha encontrando os K exemplos rotulados mais próximos de um dado não classificado. A partir da avaliação dos rótulos desses exemplos mais próximos, o K-NN toma uma decisão sobre a classe do dado não rotulado (Peterson, 2009).

O funcionamento do K-NN se baseia em três etapas principais: (1) a distância é calculada entre o ponto a ser classificado e todos os pontos no conjunto de treinamento, (2) é selecionado o K número de pontos mais próximos, e (3) a previsão é feita com base na maioria das classes dos K vizinhos. A quantidade de vizinhos a serem considerados é determinada pelo parâmetro K.

A combinação do NCA com o classificador KNeighborsClassifier é uma abordagem atraente para a classificação de dados, sendo implementada pela biblioteca scikit-learn (https://scikit-learn.org/). O NCA tem a capacidade de lidar com problemas de classificação multiclasse sem aumentar o tamanho do modelo ou adicionar novos parâmetros, tornando-se uma opção robusta para a classificação de dados (Zhang, 2016).

De acordo com Langlois et al. (1983), existem dois tipos de critérios de avaliação dos equinos: o critério direto, que avalia a capacidade do animal na execução de tarefas, e o critério indireto, que compara uma característica com outras. No entanto, os critérios utilizados para avaliar a capacidade do animal nem sempre são imparciais.

Em seleção, uma das ferramentas é a avaliação fenotípica, que inclui aspectos visuais mensuráveis como conformação e desempenho (Santos et al., 1995). Neste estudo, faremos uso das medidas de conformação corporal para treinar o algoritmo de aprendizado de máquina K-NN na classificação de cavalos.

Na pesquisa de Ounkhir et al. (2019), foi feita a seleção das características no intuito de eliminar aquelas que não facilitam o desempenho do classificador. A seleção das características foi realizada por meio do algoritmo de árvore de decisão, que identificou as características para serem utilizadas no processo de classificação, sendo 8 medidas lineares e 3 relações. Após consulta com avaliadores especializados, foi adicionada mais uma característica, o perímetro do tórax (PT), elevando o número total de características utilizadas para 12, são elas: altura da garupa (AG), altura da cernelha (AC), altura do dorso (AD), largura das ancas (LA), comprimento dorso lombar (CDL), comprimento do corpo (CC), comprimento da espádua (CE), e largura do peito (LP). O modelo também levou em conta proporções lineares que avaliam o equilíbrio das medidas dos cavalos, tais como as proporções (AD/AG), (LA/AD) e (AC/CC).

O conjunto de dados de treinamento empregado engloba informações de 296 cavalos. Os dados foram previamente rotulados por especialistas da ABCCP com um atributo de classe, que pode ser 0, 1 ou 2, representando, respectivamente, um cavalo de baixa qualidade, média qualidade e alta qualidade.

Utilizamos o método de validação cruzada para o treinamento e teste do modelo, que consiste em dividir os dados em várias partes e realizar o treinamento e teste para cada uma delas. Essa técnica permite utilizar diferentes partes dos dados para teste, aumentando a confiabilidade do modelo (Singh-Miller et al., 2007).

A validação cruzada é um método de reamostragem de dados para avaliar a capacidade de generalização de modelos preditivos e prevenir o superajuste (overfitting) (Hastie et al., 2009, Berrar, 2019).

Uma questão central na aprendizagem supervisionada diz respeito à precisão do

modelo resultante. Aqui, um problema-chave é o sobreajuste. É muito fácil construir um modelo que se adapte perfeitamente ao conjunto de dados em questão, mas que depois não consiga generalizar bem para novos dados não vistos (Berrar; Dubitzky, 2013).

Já com o uso da validação cruzada, o modelo é treinado e testado em diferentes conjuntos de dados, o que auxilia na avaliação de sua capacidade de generalização e reduz o risco de superajuste (Yang et al., 2012).

Para realizar a validação cruzada, os dados foram divididos em 5 partes, sendo que cada parte foi utilizada como conjunto de teste uma vez, enquanto as outras partes foram utilizadas como conjunto de treinamento. O conjunto de treinamento foi composto por 80% dos dados e o conjunto de teste, por 20%.

Para avaliar o desempenho do modelo, empregamos métricas como acurácia, precisão, recall e F1-score, conforme descrito por Sokolova e Lapalme (2009).

- Precisão: refere-se à proporção de exemplos positivos classificados corretamente em relação ao número total de exemplos rotulados pelo sistema como positivos.
- Recall: representa a proporção de exemplos positivos classificados corretamente em relação ao número total de exemplos positivos nos dados.
- F1-score: é uma combinação harmônica das métricas de precisão e recall, proporcionando uma medida única que considera ambos os aspectos.
- Acurácia: indica a eficácia geral de um classificador, medindo a proporção de exemplos corretamente classificados em relação ao total de exemplos.

Usamos o Google Colab como plataforma de desenvolvimento para a criação e treinamento do modelo. Para isso, empregamos a linguagem Python em conjunto com as bibliotecas Scikit-learn, segundo Pedregosa et al. (2011) e Yellowbrick, segundo Bengfort et al. (2018). Os trabalhos consultados apontaram que ambos mostraram ser ferramentas e métodos para o pré-processamento dos dados e treinamento do modelo, além de visualizações gráficas para análise dos resultados.

Dessa forma, a etapa de Materiais e Métodos abrange a seleção e pré-processamento dos dados, a aplicação do algoritmo NCA, a inclusão de uma nova característica sugerida por especialistas e a utilização da validação cruzada para avaliar o desempenho do modelo. Essas etapas são fundamentais para garantir a confiabilidade do estudo e a capacidade do modelo em classificar corretamente os cavalos pantaneiros de acordo com suas características morfológicas.

Resultados

Na Figura 1 é apresentada a acurácia do modelo K-NN para diferentes valores de K, utilizando o conjunto de dados composto por 13 medidas, sem a utilização do NCA para pré-processamento.

A linha azul representa o modelo, sem a validação cruzada. No entanto, seus resultados não são considerados confiáveis, conforme mencionado anteriormente.

Os resultados exibem uma variação considerável, e a média da acurácia para os valores de K igual a 2, 3 e 4 é próxima, situando-se entre 83% e 85%. Para evitar situações de empate, é geralmente recomendado que o valor de K seja ímpar, o que deve ser levado em consideração na escolha do melhor valor para K.

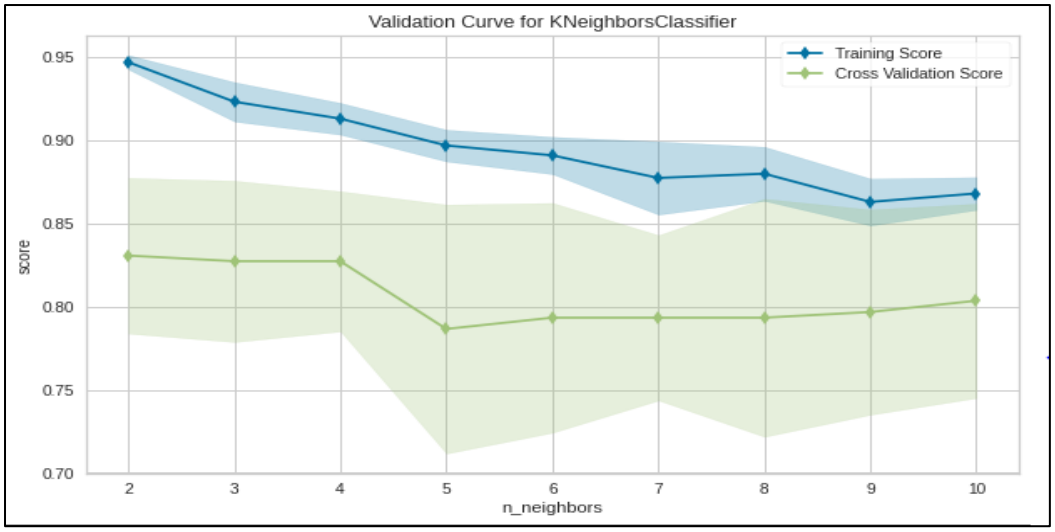


Figura 1. Acurácia do conjunto de dados 13 medidas, sem o pré-processamento.

Observando agora o conjunto de 12 medidas sem pré-processamento (Figura 2), nota-se que a variação na acurácia é menor, indicando que o modelo possui uma melhor capacidade de generalização. Isso sugere que a medida adicional utilizada no treinamento anterior pode estar prejudicando, em vez de ajudar, nas previsões.

Os melhores valores para K continuam sendo 2, 3 e 4, com um leve destaque para K igual a 3, cuja média de acurácia é de 84%.

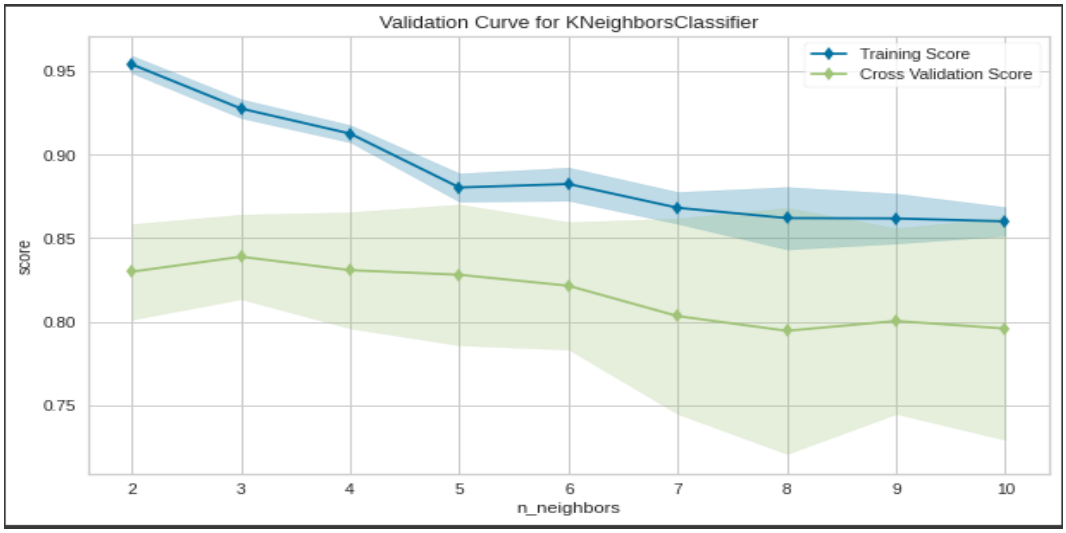


Figura 2. Acurácia do conjunto de dados 12 medidas, sem o pré-processamento.

A aplicação do NCA no conjunto de 13 medidas também resultou em uma melhoria na acurácia geral do modelo após o pré-processamento (Figura 3). Ao analisarmos os quatro gráficos, podemos observar que o valor de K=3 apresenta o melhor desempenho em todos eles, por isso será o valor utilizado. Além disso, é interessante notar que a utilização do NCA apresenta resultados muito próximos para ambos os conjuntos de 12 e 13 medidas.

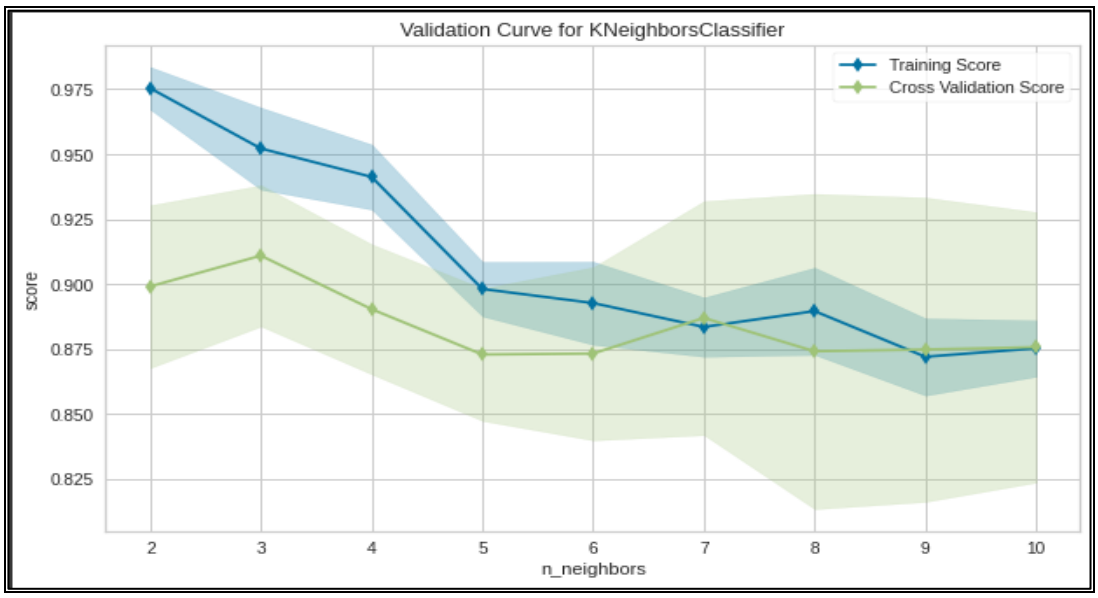


Figura 3. Acurácia do conjunto de dados 13 medidas com o pré-processamento (NCA).

Ao aplicar o NCA como pré-processamento no conjunto de 12 medidas (Figura 4), observamos uma melhora na média da acurácia, que aumentou de 84% para 90%. Além disso, o valor de K igual a 3 continua apresentando o melhor desempenho.

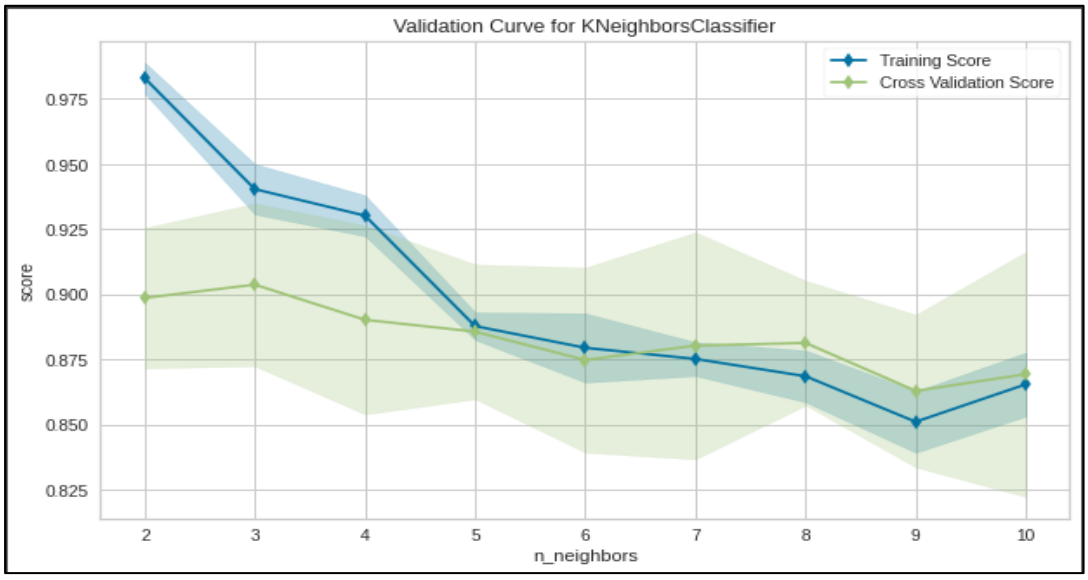


Figura 4. Acurácia do conjunto de dados 12 medidas com o pré-processamento (NCA).

A Tabela 1 extraída a seguir, apresenta a avaliação para as três classes distintas 0, 1 e 2. O “macro médio” é uma métrica usada para avaliar o desempenho geral do modelo, onde calcula a média das medidas de desempenho (como precisão, recall e f1-score) de todas as classes. É útil quando todas as classes são igualmente importantes e quando se deseja ter uma ideia geral do desempenho do modelo em

todas as classes, sem dar peso extra a uma ou outra classe.

No caso do modelo treinado com o conjunto de dados com 13 medidas, podemos notar uma acurácia geral de 81%, com valores de precisão e recall variando entre 0.76 e 0.86. O f1-score é de 0.81, indicando um bom desempenho do modelo em geral. Já para o modelo treinado com o conjunto de dados 12 medidas, notamos uma melhora na acurácia geral, chegando a 85%, e um f1-score médio ponderado de 0.85. Isso indica mais uma vez que a inclusão da medida "perímetro do tórax" no conjunto de 13 medidas não resultou em uma melhora no desempenho da classificação.

Tabela 1 - Precisão, Recall, f1-score e avaliação geral do modelo com 13 e 12 medidas sem pré-processamento e do modelo com 13 e 12 medidas com pré-processamento (NCA).

Sem pré-processamento								
Classe	Precisão		Recall		F1-score		Suporte (n)	
Nº de medidas do modelo	13	12	13	12	13	12	13	12
0	0,85	0,94	0,69	0,84	0,76	0,89	16	19
1	0,76	0,90	0,90	0,78	0,83	0,84	21	23
2	0,86	0,73	0,82	0,94	0,84	0,82	22	17
Acurácia					0,81	0,85	59	59
Macro média	0,82	0,86	0,80	0,86	0,81	0,85	59	59
Com pré-processamento NCA								
Classe	Precisão		Recall		Fi-score		Suporte (n)	
0	0,87	1,00	0,87	0,87	0,87	0,93	15	15
1	1,00	0,95	0,95	0,95	0,98	0,95	22	22
2	0,87	0,88	0,91	0,95	0,89	0,91	22	22
Acurácia					0,92	0,93	59	59
Macro média	0,91	0,94	0,91	0,93	0,91	0,93	59	59

Ao comparar os dois resultados seguintes utilizando o conjunto de 12 e 13 medidas com pré-processamento, observa-se que ambos apresentam resultados muito próximos em termos de métricas de precisão, recall e f1-score. O resultado da avaliação 12 NCA obteve uma precisão de 0,87 para a classe 0, 1,00 para a classe 1 e 0,87 para a classe 2. Na avaliação do 13 NCA obteve uma precisão de 1,00 para a classe 0, 0,95 para a classe 1 e 0,88 para a classe 2. Em termos gerais, a avaliação 13 NCA apresentou resultados ligeiramente melhores, com uma pontuação ligeiramente mais alta em todas as métricas. A macro média para ambas as avaliações é muito próxima, sendo 0,91 e 0,94 para a avaliação NCA, sugerindo que ambos os modelos têm um desempenho semelhante no geral.

Conclusão

Com base nos resultados obtidos no banco de dados, podemos concluir que o algoritmo K-NN (k=3) proporcionou melhor acurácia além de ser a melhor escolha para critérios de empate. O desempenho do algoritmo também foi melhorado através do uso do pré-processamento de dados denominado Neighbourhood Component Analysis (NCA), que permitiu uma distribuição mais homogênea das classes.

Os resultados também provaram que ao invés de 15 medidas lineares e suas relações, os técnicos podem utilizar 9 medidas e suas relações nas avaliações dos cavalos Pantaneiros. Esta conclusão já pode ser adotada pelos técnicos da ABCCP.

É importante destacar que os resultados satisfatórios obtidos nos modelos de classificação por aprendizagem de máquina podem ser melhorados com dados adicionais para avaliar melhor a eficácia do modelo em diferentes cenários, além de validar sua aplicabilidade em situações de vida real.

Agradecimentos

Ao CNPq pela bolsa concedida ao projeto Embrapa 10 20 02 007 00 02 001- Rede de Recursos Genéticos Animais.

À Associação Brasileira de Criadores de Cavalo Pantaneiro – ABCCP

Referências

BALIEIRO, E. de S. **Subsídios ao estudo do cavalo pantaneiro**. São Paulo: Comissão Coordenadora de Criação do Cavalo Pantaneiro, 1971. p. 59-65.

BENGFORT, B.; DANIELSEN, N.; BILBRO, R.; GRAY, L.; McINTYRE, K., RICHARDSON, G.; MILLER, T.; MAYFIELD, G.; SCHAFER, P.; KEUNG, J. **Yellowbrick v0.6**. 2018. Disponível em: https://zenodo.org/record/1206264. Acesso em: 10 nov. 2022.

BERRAR, D. Cross-validation. **Encyclopedia of Bioinformatics and Computational Biology**, v. 1, p. 542-545, 2019. DOI: **https://doi.org/10.1016/B978-0-12-809633-8.20349-X**.

BERRAR, D.; DUBITZKY, W. Overfitting. In: DUBITZKY, W.; WOLKENHAUER, O.; CHO, K.-H.; YOKOTA, H. (ed.). **Encyclopedia of systems biology**. New York: Springer, 2013. p. 1617–1619. DOI: https://doi.org/10.1007/978-1-4419-9863-7_601.

BUENO, A. M. C.; TORRES, D. A. P. **Objetivos de Desenvolvimento Sustentável da agenda 2030 e bioeconomia**: oportunidades e potencialidades para atuação da Embrapa. Brasília, DF: Embrapa, 2022. 103 p. **http://www.alice.cnptia.embrapa.br/alice/handle/doc/1142941.**

GOLDBERGER, J.; HINTON, G. E.; ROWEIS, S.; SALAKHUTDINOV, R. R. Neighborhood components analysis. **Advances in Neural Information Processing Systems**, v. 17, 2004. Disponível em: https://papers.nips.cc/paper_files/paper/2004/file/42fe880812925e520249e808937738d2-Paper.pdf. Acesso em: 10 out. 2022.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The elements of statistical learning**: data mining, inference, and prediction. 2nd ed. New York: Springer, 2009. 745 p.

LANGLOIS, B.; MINKEMA, D.; BRUNS, E. Genetic problems in horse breeding. **Livestock Production Science**, v. 10, n. 1, p. 69-81, Jan. 1983.

OUNKHIR, S.; OLIVEIRA, O. N. C.; KOIKE, C. Y.; GRASSI, M. J. S.; SAQUI, D.; SANTOS, S. A. Análise de algoritmos de aprendizado de máquina para classificação do padrão racial do cavalo pantaneiro. In: EVENTO DE INICIAÇÃO CIENTÍFICA DO PANTANAL, 7., 2019, Corumbá. **Resumos**... Brasília, DF: Embrapa, 2019. p. 8. http://www.alice.cnptia.embrapa.br/alice/handle/doc/1119864.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-Learn: machine learning in Python. **Journal of Machine Learning Research**, v. 12, n. 85, p. 2825-2830, 2011.

PETERSON, L. E. K-nearest neighbor. **Scholarpedia**, v. 4, n. 2, p. 1883, 2009. DOI: **https://doi.org/10.4249/scholarpedia.1883**.

SANTOS, S. A.; MAZZA, M. C. M.; SERENO, J. R. B.; ABREU, U. G. P. de; SILVA, J. A. da. **Avaliação e conservação do cavalo pantaneiro**. Corumbá: EMBRAPA-CPAP, 1995. 40 p. (EMBRAPA-CPAP. Circular técnica, 21). http://www.infoteca.cnptia.embrapa.br/infoteca/handle/doc/784346.

SANTOS, S. A.; SALIS, S. M.; COMASTRI FILHO, J. A. (ed.). **Cavalo Pantaneiro**: rústico por natureza. Brasília, DF: Embrapa, 2016. 603 p.

SINGH-MILLER, N.; COLLINS, M.; HAZEN, T.J. **Dimensionality reduction for speech recognition using neighborhood components analysis**. Interspeech 2007. DOI: https://doi.org/10.21437/ -376.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. **Information Processing and Management**, v. 45, n. 4, p. 427-437, July 2009. DOI: https://doi.org/10.1016/j.ipm.2009.03.002.

YANG, W.; WANG, K.; ZUO, W. Fast neighborhood component analysis. **Neurocomputing**, v. 83, p. 31-37, Apr. 2012. DOI: **https://doi.org/10.1016/j.neucom.2011.10.021.**

ZHANG, Z. Introduction to machine learning: k-nearest neighbors. **Annals of Translational Medicine**, v. 4, n. 11, June 2016. DOI: https://doi.org/10.21037/atm.2016.03.37.

Disponível no endereço eletrônico: https://www.embrapa.br/busca-de-publicacoes/-/publicacao/

Embrapa Pantanal
Ria 21 de Setembro, 1880
Corumbá, MS
Fone: (67) 3234 5800

www.embrapa.br/pantanal
www.embrapa.br/fale-conosco/sac
www.embrapa.br

1ª edição

Comitê Local de Publicações

Presidente
Adriana Mello de Araujo

Membros
Agostinho C. Catella, Ana Helena B Marozi Fernandes, José A. Comastri Filho, Márcia Divina de Oliveira

Supervisão editorial
Adriana M. Araújo

Diagramação de texto
Marcelo Xavier

Normalização bibliográfica
Ana Lucia Delalibera de Faria (CRB-1/324)

Ilustração da capa
Igor Pinho

1ª edição
Publicação digital (2023)