Análise de algoritmos de aprendizado de máquina para classificação do padrão racial do cavalo Pantaneiro[1]

Soumaya Ounkhir[2], Otávio Nathanael Campos de Oliveira[2], Camila Yumi Koike[3], Marcel José Soleira Grassi[4], Diego Saqui[4] e Sandra Aparecida Santos[5]

[1] Financiado pelo projeto "Conservação *in situ* de equídeos" (SEG/Embrapa 21.15.02.003.03.00), com apoio financeiro da Associação Brasileira de Criadores de Cavalo Pantaneiro (ABCCP)
[2] Acadêmicos do curso de Análise e Desenvolvimento de Sistemas, Instituto Federal de Educação de Mato Grosso do Sul, Corumbá, MS
[3] Bacharel em Ciência da Computação, mestre em Ciência da Computação, docente do Instituto Federal de Educação de Mato Grosso do Sul, Jardim, MS
[4] Bacharel em Ciência da Computação, mestre em Ciência da Computação, docente do Instituto Federal de Educação de Mato Grosso do Sul, Corumbá, MS.
[5] Zootecnista, doutora em Zootecnia, pesquisadora da Embrapa Pantanal, Corumbá, MS

O cavalo Pantaneiro é uma raça localmente adaptada à região pantaneira, que ora tem muita água, ora tem muita seca, além de outros fatores como altas temperaturas, insetos, predadores, entre outros. Apesar dessa raça possuir uma população de cavalos registrados ainda baixa, percebe-se, atualmente, um crescente aumento e interesse por estes animais, principalmente pela alta qualidade dos cavalos expostos em feiras e em leilões, assim como também pelo seu desempenho no campo esportivo. Para o registro da raça na Associação Brasileira de Criadores de Cavalos Pantaneiros (ABCCP) são estabelecidos alguns critérios para avaliar se a raça está dentro do padrão racial, entre os quais 15 medidas lineares associado com alguns escores. Embora existam essas métricas, na prática o critério é bastante subjetivo. Por isso, pretende-se utilizar métodos de aprendizado de máquina para encontrar padrões ocultos nos dados por meio de *Clustering* usando o algoritmo *Kmeans*. Portanto, o objetivo desta pesquisa é realizar uma análise de algoritmos de aprendizado de máquina, que é caracterizado como modelo computacional com alta capacidade de aprendizagem capaz de realizar a predição de informações e padrões de comportamento, para classificar o cavalo Pantaneiro em alta, média e baixa qualidade, tornando este processo mais automatizado e mais próximo de uma avaliação de um especialista. Além disso, este trabalho objetiva diminuir o número das medidas lineares utilizadas como base para avaliar a qualidade do animal afim de diminuir o trabalho de medir os cavalos usando a seleção de características. Para o estudo utilizou-se a base de dados de cavalos Pantaneiros com registro na ABCCP, a base de dados de cavalos que foram premiados nos últimos 10 anos e dados de cavalos que não foram registrados por estarem fora do padrão. A partir dessas informações, realizou-se a seleção e pré-processamento dos dados, excluindo os outliers e realizando o tratamento dos dados faltantes assim como a sua correta conversão para dados aceitáveis por meio da linguagem de programação Python e Biblioteca Pandas. Após a limpeza das duas bases de dados, foi utilizado o algoritmo *K-means* que recebeu como entrada a base de dados composta por cavalos de alta e média qualidade. Este algoritmo foi capaz de separar os dados que estavam misturados em duas classes distintas com base na distância euclidiana. Após a separação, atribuiu-se a correta classe para cada dado. Na segunda etapa, foi feita a pesquisa sobre os melhores métodos para realizar a seleção das melhores características (*RandomForest, Kbest, Matriz Correlacional, RFE, RFECV*). Em seguida, foram feitos vários testes classificando a base com todos os seus atributos e depois com as características que foram selecionadas usando diferentes algoritmos: *KNN, Naive Bayes*, Árvore de Decisão, *DummyClassifier e RandomForestClassifier*. Para esses obteve-se uma taxa de acerto de 73%, 76%, 71%, 27% e 82%, respectivamente utilizando todas as 15 medidas. Conclui-se que, com a seleção de oito características com diferentes configurações obtidas, o melhor algoritmo foi o *RandomForestClassifier* com 86,49% de acerto.