



ALGORITMO KNN APLICADO À CLASSIFICAÇÃO DE ESPÉCIES DE COGUMELOS

Fortunato F. RONCHOLETA¹; Kevin W. C. de OLIVEIRA²; Diego SAQUI³

RESUMO

Na natureza há inúmeros tipos de cogumelos, entretanto, alguns podem ser venenosos e semelhantes entre si, tornando o reconhecimento e o manuseio destes um trabalho oneroso e perigoso. Dessa forma, o estudo e aplicação da classificação de cogumelos fazendo-se uso de técnicas de inteligência artificial, é relevante. Nesse cenário, o objetivo deste trabalho é discutir a classificação final de uma base de dados com informações de cogumelos, fazendo-se uso do algoritmo de *K-Nearest Neighbors* (KNN) e utilizando-se de gráficos, a fim de visualizar os dados que foram obtidos de forma mais clara. Após esse procedimento, na qual tratamos primeiramente a base de dados para depois utilizar o classificador, o modelo obteve cerca de 98,78% de acurácia.

Palavras-chave:

K-Nearest Neighbors; Cogumelos tóxicos; Inteligência artificial; Aprendizado de máquina.

1. INTRODUÇÃO

O presente trabalho tem como objetivo apresentar a aplicação do algoritmo *K-Nearest Neighbors* (KNN) na classificação de espécies de cogumelos venenosos. A identificação correta das espécies de cogumelos é fundamental para evitar acidentes e prevenir danos à saúde humana (DAS, 2020).

O algoritmo KNN é um dos muitos algoritmos de aprendizagem supervisionada, muito utilizado no campo de *data mining* e *Machine Learning*, podendo ser usado tanto para problemas de classificação como de regressão (TAUNK, 2019). Em termos simples, o KNN é um algoritmo que tenta classificar cada ponto de dados examinando a proximidade desses pontos em relação aos seus vizinhos mais próximos. Se a maioria desses vizinhos mais próximos pertencer a uma determinada categoria, o ponto de dados em questão também será classificado nessa categoria. Em suma, o KNN usa a ideia de "vizinhança" para determinar a classe de um ponto com base na classe dos pontos próximos a ele.

Nesta pesquisa, foram utilizados dados previamente coletados de diferentes espécies de cogumelos, contendo informações como cor, odor, habitat, entre outras (SCHLIMMER, 2023). Esses dados foram processados e utilizados como base de treinamento para o algoritmo KNN. O algoritmo analisou as características dos cogumelos a serem classificados e identificou os vizinhos

¹Fortunato Figueiredo Roncholeta, IFSULDEMINAS – Campus Muzambinho. E-mail: 12171002067@muz.ifsuldeminas.edu.br.

²Kevin Willians Casagrande de Oliveira, IFSULDEMINAS – Campus Muzambinho. E-mail: 12182001298@muz.ifsuldeminas.edu.br

³Diego Saqui, IFSULDEMINAS – Campus Muzambinho. E-mail: diego.saqui@muz.ifsuldeminas.edu.br.

mais próximos com base nessas características e tentou agrupá-los em dois grupos, cogumelos venenosos e cogumelos não venenosos.

2. MATERIAL E MÉTODOS

A metodologia foi conduzida conforme as etapas de: Pré-processamento, Avaliação e Interpretação.

Os dados foram coletados por autoria de Schlimmer 2023, que obteve a partir de um registros de cogumelos que foram extraídos do livro 'The Audubon Society Field Guide to North American Mushrooms', de G. H. Lincoff (pres.), Nova York: Alfred A. Knopf, com 8125 instâncias.

O pré-processamento foi a fase responsável pelo tratamento e identificação dos dados a serem analisados, o nome das colunas foi renomeado que estavam apenas com números de 0 a 22, "Forma_tampa", "Superficie_tampa", "cor_tampa", "Hematomas", "odor", "fixacao_braquial", "espaco_braquial", "tamanho_branquia", "cor_branquia", "forma_caule", "raiz", "caule_acima_anel", "caule_abaixo_anel", "cor_acima_anel", "cor_abaixo_anel", "tipo_de_veu", "cor_veu, numero_anel", "tipo_anel", "cor_esporo", "population", "habitat", "venenoso".

O próximo passo foi retirar a linha 0 que estava com os nomes das tabelas. Em seguida os dados contidos no dataset foram analisados e percebeu-se que muitas colunas possuíam valores booleanos, o que poderia prejudicar a classificação. Por esse motivo, as seguintes colunas foram retiradas da base de dados "Hematomas", "fixacao_braquial", "espaco_braquial", "tamanho_branquia", "forma_caule", "tipo_de_veu".

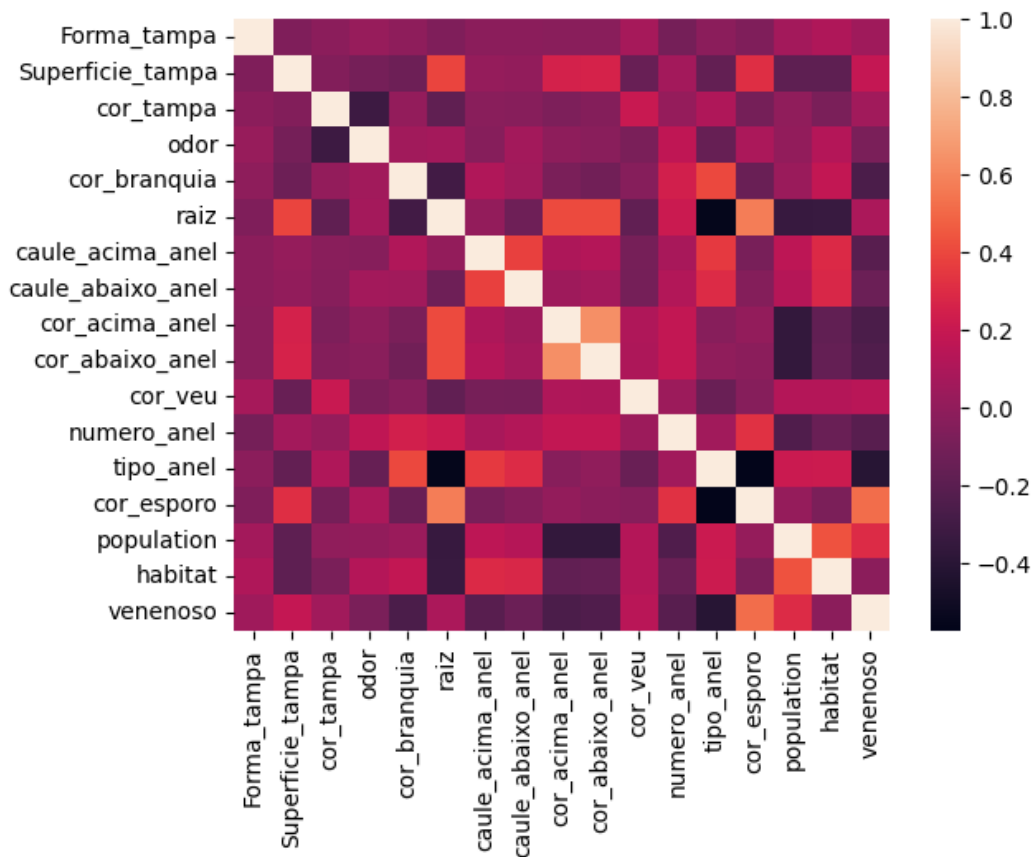
A próxima etapa é a de separar a variável dependente, neste caso, está contido no "venenoso", das demais e depois transformar todos os dados categóricos e numéricos, utilizando a função do *label encoder*. Por fim, são separados os dados de teste e os dados de treino divididos em 50% do total de forma aleatória sem validação cruzada, para então executar o algoritmo KNN.

O classificador foi testado com $K=3$ e 5 , ou seja, três e cinco vizinhos mais próximos, a métrica de distância utilizada foi a Euclidiana que é a padrão do KNN. A técnica definida para o *label encoder* das classes foram codificadas como 0 e 1, onde 0 representa "não venenoso" e 1 representa "venenoso".

Para a avaliação e interpretação foi realizada a observação da acurácia obtida, além da correlação utilizando um mapa de calor.

3. RESULTADOS E DISCUSSÃO

Depois de separados os dados e com o modelo treinado, obteve-se os seguintes resultados. Utilizando o KNN para a classificação de cogumelos, com o $K=3$ foi obtida uma acurácia de 94%, um resultado bem satisfatório. Já com um modelo treinado com o $K=5$ conseguiu-se uma acurácia de 98% o que também é bem satisfatório. Um teste no dataset sem que fosse removido as colunas com dados booleanos, e obteve-se os resultados de acurácia 1.0, o que significa que o modelo teria classificado 100% correto mostrando uma certa dúvida se realmente seria um resultado confiável. Foi realizado um mapa de calor com a correlação de cada classe, é possível observar que a cor do esporo tem um grau considerável de correlação em ser venenoso (Fig. 1).



4. CONCLUSÃO

A presente pesquisa propõe a utilização do algoritmo KNN na classificação de espécies de cogumelos, nesse cenário conclui-se que sim, é possível realizar a classificação de cogumelos com uma precisão considerável utilizando métodos de IA como o KNN.

REFERÊNCIAS

DAS, Bodhaditya et al. Infra-red spectra of different species of cultivated oyster mushrooms: Possible tool for identifying bioactive compounds and establishing taxonomic linkage. Def. Life Sci. J, v. 5, p. 118-124, 2020.

SCHLIMMER, Jeff. Mushroom_classification [Base de dados]. Kaggle, 2023.

TAUNK, Kashvi et al. A brief review of nearest neighbor algorithm for learning and classification. In: 2019 INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING AND CONTROL SYSTEMS (ICCS), 2019, [S.l.]. Anais... [S.l.]: IEEE, 2019. p. 1255-1260.