CLASSIFICAÇÃO DE VARIEDADES DE UVAS-PASSAS USANDO APRENDIZADO DE MÁQUINA E VISÃO COMPUTACIONAL

CLASSIFICATION OF RAISIN VARIETIES USING MACHINE LEARNING AND COMPUTER VISION

Diego Ribeiro Araújo[1] (IF Goiano)

Nattane Luiza da Costa[2] (IF Goiano)

Júlio César Ferreira[3] (IF Goiano)

Gabriel da Silva Vieira[4] (IF Goiano)

RESUMO: A diferenciação de produtos alimentícios através de algoritmos de classificação baseados em informações químicas e físicas tem desempenhado um papel importante na agricultura e na ciência de alimentos, permitindo o controle de qualidade, garantia de autenticidade e automação de processos industriais. No entanto, a aplicação desses algoritmos em diferentes produtos e propriedades específicas requer estudos e adaptações específicas. Neste artigo, nosso objetivo é classificar duas variedades de uvas-passas, Kecimen e Besni, reconhecidas como um dos principais produtos agrícolas do país. Para isso, utilizamos uma base de dados pública que descreve propriedades morfológicas dos grãos de uvas-passas obtidas por meio de técnicas avançadas de visão computacional. A classificação é realizada utilizando técnicas de aprendizado de máquina supervisionado, incluindo SVM (Support Vector Machines), LDA (Linear Discriminant Analysis) e ANN (Artificial Neural Networks), além de algoritmos de seleção de variáveis. Nosso objetivo é identificar as características distintivas de cada variedade, contribuindo para a classificação precisa e confiável. Com os experimentos, obtivemos resultados robustos para caracterizar e classificar as variedades de uvas-passas. Esses resultados podem abrir caminho para pesquisas futuras sobre a caracterização e classificação de outros produtos alimentícios.

PALAVRAS-CHAVE: Classificação de produtos alimentícios. Aprendizado de máquina. Seleção de variáveis. Visão computacional. Propriedades morfológicas.

ABSTRACT: *The differentiation of food products through classification algorithms based on chemical and physical information has played an important role in agriculture and food science, enabling quality control, authenticity assurance, and automation of industrial processes. However, applying these*

[1] Acadêmico do curso de Sistemas de Informação - Instituto Federal Goiano (IF Goiano), Campus Urutaí, Goiás, Brasil. E-mail: diego.ribeiro@estudante.ifgoiano.edu.br

[2] Docente do Instituto Federal Goiano (IF Goiano), Campus Urutaí, Goiás, Brasil. Doutorado em Ciência da Computação pela Universidade Federal de Goiás (UFG). E-mail: nattane.luiza@ifgoiano.edu.br

[3] Docente do Instituto Federal Goiano (IF Goiano), Campus Urutaí, Goiás, Brasil. Doutorado em Engenharia Elétrica pela Universidade Federal de Uberlândia (UFU). E-mail: julio.ferreira@ifgoiano.edu.br

[4] Docente do Instituto Federal Goiano (IF Goiano), Campus Urutaí, Goiás, Brasil. Doutor em Ciência da Computação pela Universidade Federal de Goiás (UFG). E-mail: gabriel.vieira@ifgoiano.edu.br

algorithms to different products and specific properties requires targeted studies and adaptations. In this project, our goal is to classify two varieties of raisins, Kecimen e Besni, which are recognized as one of the country's major agricultural products. To achieve this, we use a public dataset describing the morphological properties of raisin grains obtained through advanced computer vision techniques. The classification is performed using supervised machine learning techniques, including SVM (Support Vector Machines), LDA (Linear Discriminant Analysis), and ANN (Artificial Neural Networks), as well as variable selection algorithms. We aim to identify the distinctive features of each variety, contributing to precise and reliable classification. We obtained robust results that not only characterize and classify the raisin varieties but also pave the way for future research on the characterization and classification of other food products.

KEYWORDS*: Food product classification. Machine learning. Variable selection. Computer vision. Morphological properties.*

1 Introdução

O consumo de uvas-passas remonta a tempos pré-históricos e é produzido na maioria das regiões geográficas do mundo. Esse consumo ocorre em todas as culturas e regiões demográficas, como pode ser observado em Williamson and Carughi (2010, p. 2). As diferenças entre os tipos de uva, secas ou não, influenciam na qualidade e no preço de mercado, levando em consideração características como aparência, tamanho, cor e região de produção. Essas características tornam necessário diferenciar os tipos de uvas-passas para controle de autenticidade e qualidade, além de possibilitar a automação do processo de colheita e embalagem por meio de robôs (Akhter; SOFI, 2021, p. 4).

Recentemente, Cinar et al. (2020, p. 3) conduziram um estudo para classificar duas variedades de uvas-passas (Kecimen e Besni) com base em propriedades morfológicas obtidas por meio de visão computacional e aprendizado de máquina. Os autores extraíram 7 variáveis morfológicas de 900 amostras de uvas-passas (450 amostras de cada variedade). Após a extração das variáveis, eles aplicaram técnicas de aprendizado de máquina, como Máquinas de Vetores de Suporte, Redes Neurais Artificiais e Regressão Logística, alcançando uma taxa de acurácia de 86%.

Em Abbasgholipour et al. (2012), os autores não utilizaram seleção de variáveis. O estudo foi realizado para desenvolver a tecnologia de detecção de passas baseada em Visão Computacional utilizando Algoritmo Genético (GA) e Espaço de Cores Matiz-Saturação-Intensidade (HSI). O resultado da aplicação do GA para detectar uma região no espaço HSI para

segmentação de passas mostrou que pode superar os efeitos das condições variáveis com uma taxa de erro aceitável.

Azcarate et al. (2018, p. 3) utilizou PCA como método de seleção de variáveis em um estudo com 3 classes e 41 amostras, cujo objetivo era explorar a viabilidade da discriminação de variedades de vinhos brancos argentinos. O uso da sensibilidade da fluorescência em conjunto com os algoritmos U-PLS-DA e SPA-LDA mostrou-se uma ferramenta útil para a indústria vinícola e instituições governamentais de fiscalização, permitindo a detecção do tipo de uva utilizado em vinhos disponíveis comercialmente ou evitando fraudes no mercado de vinhos brancos.

Em Khojastehnazhand e Ramezani (2020, p. 4), foi investigado as qualidades das passas utilizando técnicas de processamento de imagem. Foram aplicados os algoritmos de classificação SVM e LDA, sendo que o SVM obteve um melhor desempenho, com uma acurácia de 85,55%. No entanto, quando resíduos foram misturados entre as passas analisadas, o desempenho do sistema diminuiu.

Por outro lado, Wang et al. (2012, p. 2) obteve um melhor resultado utilizando o algoritmo LDA, com uma acurácia de 99%. Foram utilizados cinco tipos de algoritmos de classificação: Linear Discriminant Analysis (LDA), Partial Least Squares (PLS), Soft Independent Modelling of Class Analogies (SIMCA), Least-squares Support Vector Machine (LS-SVM) e Radial basis function (RBF). O objetivo do estudo era desenvolver uma aplicação para a classificação rápida de passas, a fim de reduzir a adulteração de passas de baixo preço como sendo de alto preço. Nenhum método de seleção de variáveis foi utilizado, e o estudo analisou 3 classes de passas e 74 amostras.

Mollazade et al. (2012, p. 2) classificou 4 classes de passas e 44 amostras utilizando os algoritmos ANN, SVM, DT e BN para desenvolver uma aplicação que classificasse passas de forma mais eficiente e confiável do que o método manual. O algoritmo ANN apresentou o melhor desempenho, com uma acurácia de 96,33%.

Karimi et al. (2017, p. 8) desenvolveu um sistema para medir e reconhecer a qualidade e pureza de passas mistas usando imagens de passas a granel, através de Machine Vision. Nessa pesquisa, os algoritmos de classificação utilizados foram ANN, TCR e SVM, onde o melhor resultado foi obtido através da SVM com 92,71% de acurácia utilizando 5 classes

de passas.

No entanto, ainda existem outros algoritmos de classificação além de técnicas de seleção de variáveis que podem aumentar a taxa de classificação das variedades de uvas-passas, além de identificar as principais características que diferenciam cada classe de dados.

Neste artigo, propomos classificar duas variedades de uvas-passas, Kecimen e Besni, com base em variáveis obtidas por visão computacional, além de identificar as principais características que diferenciam os tipos de passas utilizando algoritmos de aprendizado de máquina e de seleção de variáveis. O banco de dados resultante da pesquisa de Cinar et al. (2020) está publicamente disponível no Repositório de aprendizado de máquina UCI, o qual utilizamos nesta pesquisa.

2 Materiais e Métodos

2.1 Amostras de Uvas-passas

As amostras de uvas-passas foram obtidas de uma base de dados pública que descrevem características morfológicas de grãos de uvas-passas. Ela é formada por 900 passas divididas igualmente em duas espécies, Besni e Kecimen, e são compostas por 7 variáveis: ConvexArea, Perimeter, Área, MajorAxisLength, MinorAxisLength, Eccentricity, Extent.

2.2 Análise de Componentes Principais

A Análise de Componentes Principais (PCA - Principal Component Analysis) é uma técnica estatística utilizada para reduzir a dimensionalidade dos dados, identificando as principais variáveis ou componentes que explicam a maior parte da variabilidade nos dados. Ao aplicar o PCA, verificamos que duas variáveis - eccentricity e extent - têm menor contribuição na diferenciação das duas uvas-passas. Já as outras variáveis que tiveram maior contribuição estavam mais relacionadas ao tamanho das passas, ou seja, uma era relativamente maior do que a outra.

2.3 BoxPlot

BoxPlot são gráficos que mostram a distribuição dos dados, destacando a mediana, quartis, valores mínimos e máximos, essa visualização permite observar as diferenças entre as duas passas em relação a cada variável individualmente e foi identificado uma maior distinção das passas com variáveis relacionadas ao tamanho.

2.4 Validação Cruzada

Utilizamos o método de validação cruzada (cross-validation) com uma técnica conhecida como "Holdout". Com isso, a base de dados foi dividida em duas partes: uma parte maior, composta por 70% dos dados, que foi usada para treinamento, e uma parte menor, composta por 30% dos dados, usada para teste.

O conjunto de treinamento (70%) é usado para ajustar os parâmetros do modelo, permitindo que ele aprenda com os dados e capture os padrões e relações presentes nos mesmos. O conjunto de teste (30%) é usado para avaliar o desempenho do algoritmo em dados não vistos anteriormente.

2.5 Algoritmo de Classificação

Aplicamos algumas técnicas de *machine learning*, como Artificial Neural Networks (ANN) e Linear Discriminant Analysis (LDA), para diferenciar os tipos de passas Kecimen e Besni por meio de análises morfológicas. E também utilizamos o chi-squared ajudando a identificar as variáveis que têm uma associação mais forte com a variável de interesse ou que contribuem mais para o resultado qualificado.

3 Resultados e Discussões

Primeiramente, classificamos a base de dados para obter resultado de referência para comparar este resultado com os resultados do conjunto de dados após aplicar métodos e técnicas de aprendizado de máquina. O algoritmo de classificação SVM com todas as características das amostras de dados, juntamente com a validação cruzada (cross-validation), resultou em 82,73% de precisão.

Houve uma insatisfação com os resultados desses dados para o estudo, o que nos levou a utilizar o PCA (Principal Component Analysis) como próximo passo. O objetivo era determinar a melhor variável para a análise das uvas-passas. Com isso, observou-se que duas variáveis, *eccentricity* e *extent*, contribuem menos para a diferenciação entre as duas passas. Por outro lado, as demais variáveis apresentaram uma contribuição significativa, estando mais relacionadas ao tamanho das passas. Em outras palavras, uma das passas era relativamente maior que a outra, como ilustrado na Figura 1.

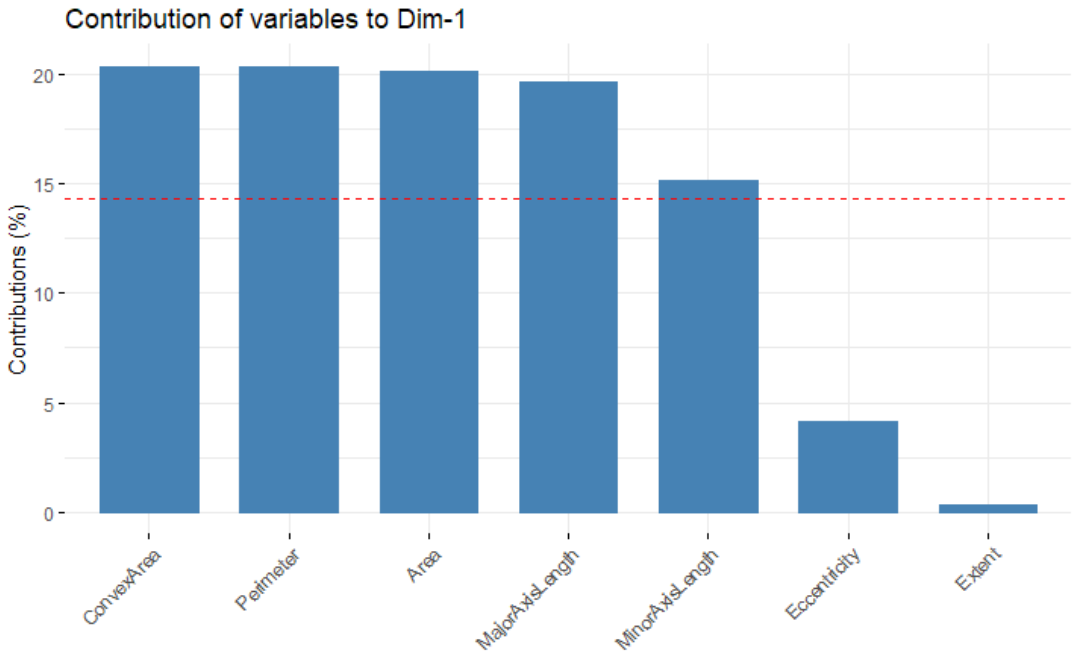

Figura 1: Contribuição das variáveis utilizando Análise de Componentes Principais (PCA).

3.1 Análise da Importância da Variável

Para ajudar na visualização e no entendimento das características essenciais do conjunto de dados utilizamos o boxplot. Ele ofereceu uma representação compacta e informativa da distribuição, identificando os valores discrepantes e permitindo comparações entre grupos ou distribuições diferentes.

Houve uma grande diferença nas medianas dos boxplot entre as espécies das uvas-passas quando a comparação era referente às variáveis de tamanho. Além disso, a extensão vertical dos boxplots, ou seja, a altura das caixas, pode fornecer uma ideia da variabilidade dos tamanhos. Quando um boxplot é mais alto do que o outro, indica que a espécie de uva-passa correspondente possui uma maior dispersão ou variabilidade em seus tamanhos.

Dessa forma, o boxplot permite uma análise visual clara da diferença entre as espécies de uva-passa em relação ao tamanho delas. Ele destaca as diferenças na localização e variabilidade dos tamanhos, ajudando a compreender rapidamente como as espécies se distinguem nessa característica específica conforme mostrado na Figura 2.

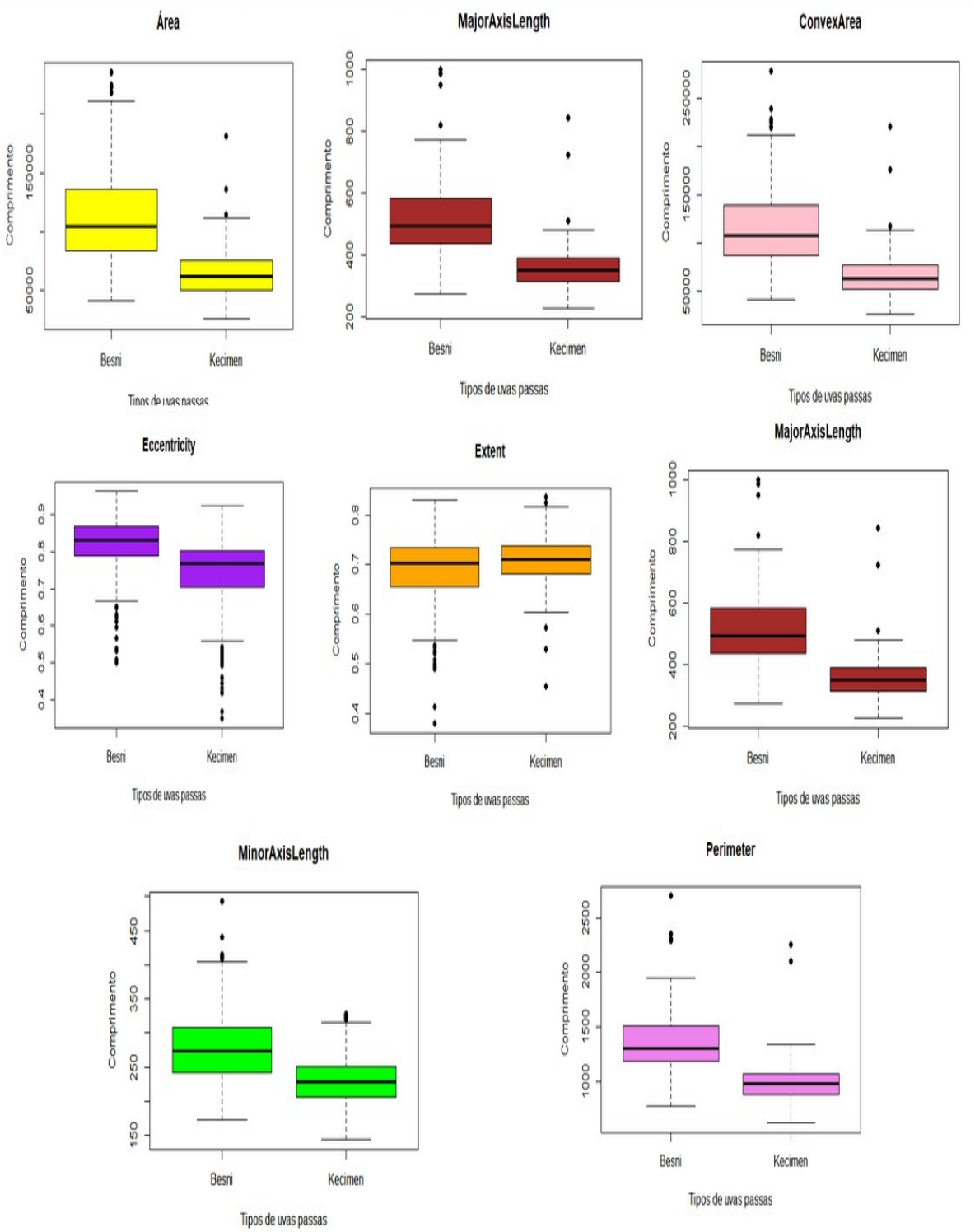

Figura 2: Boxplot das variáveis da base de dados.

3.2 Análise de classificação

Após realizar a análise das variáveis disponíveis, a fim de determinar quais eram as variáveis mais relevantes para o objetivo da pesquisa, utilizamos o teste chi-squared para determinar a melhor sequência dessas variáveis, que proporciona os melhores resultados.

O teste chi-squared é uma ferramenta estatística adequada para analisar a associação entre variáveis categóricas. Nesse caso, ele foi aplicado para avaliar a relação entre as variáveis consideradas e a variável de interesse ou o resultado esperado. Ao calcular o chi-squared, foi possível identificar a dependência ou independência estatística entre as variáveis.

Tabela 1: Resultados gerais da classificação com validação cruzada (cross-validation) utilizando Support Vector Machine (SVM) e o chi-squared para cada variável.

#	Componentes	Acc	Sensitivity	Specificity
#1	MajorAxisLength	84,07%	80,74%	87,41%
#2	MajorAxisLength, Perimeter	85,93%	84,44%	87,41%
#3	MajorAxisLength, Perimeter, ConvexArea	87,41%	88,15%	86,67%
#4	MajorAxisLength, Perimeter, ConvexArea, Área	91,48%	86,67%	96,30%
#5	MajorAxisLength, Perimeter, ConvexArea, Área, MinorAxisLength	87,04%	82,96%	86,67%
#6	MajorAxisLength, Perimeter, ConvexArea, Área, MinorAxisLength, Eccentricity	84,07%	82,96%	84,81%
#7	MajorAxisLength, Perimeter, ConvexArea, Área, MinorAxisLength, Eccentricity, Extent	86,30%	82,96%	84,51%

Com base nessa análise, foi realizado um estudo dos resultados utilizando o algoritmo de classificação SVM em conjunto com a validação cruzada (cross-validation), conforme apresentado na Tabela 1. Foi constatado que o melhor resultado apresentou uma precisão de 91.48%, como pode ser observado na linha 4 da Tabela 1. Desse modo, a sequência das variáveis foi a seguinte: MajorAxisLength, Perimeter, ConvexArea, Área.

Tabela 2: Resultados gerais da classificação com validação cruzada (cross-validation) utilizando Linear Discriminant Analysis (LDA) e o chi-squared para cada variável.

#	Componentes	Acc	Sensitivity	Specificity
#1	MajorAxisLength	83,70%	79,63%	87,41%
#2	MajorAxisLength, Perimeter	85,56%	83,33%	87,78%
#3	MajorAxisLength, Perimeter, ConvexArea	87,04%	85,19%	88,52%
#4	MajorAxisLength, Perimeter, ConvexArea, Área	91,11%	88,15%	92,59%
#5	MajorAxisLength, Perimeter, ConvexArea, Área, MinorAxisLength	87,78%	85,93%	89,26%
#6	MajorAxisLength, Perimeter, ConvexArea, Área, MinorAxisLength, Eccentricity	86,30%	84,44%	88,89%
#7	MajorAxisLength, Perimeter, ConvexArea, Área, MinorAxisLength, Eccentricity, Extent	88,15%	86,67%	89,63%

Tabela 3: Resultados gerais da classificação com validação cruzada (cross-validation) utilizando Artificial Neural Networks (ANN) e o chi-squared para cada variável.

#	Componentes	Acc	Sensitivity	Specificity
#1	MajorAxisLength	83,33%	79,26%	86,67%
#2	MajorAxisLength, Perimeter	85,19%	82,22%	87,41%
#3	MajorAxisLength, Perimeter, ConvexArea	86,67%	84,44%	88,15%
#4	MajorAxisLength, Perimeter, ConvexArea, Área	90,70%	87,41%	92,22%
#5	MajorAxisLength, Perimeter, ConvexArea, Área, MinorAxisLength	86,67%	84,07%	89,63%
#6	MajorAxisLength, Perimeter, ConvexArea, Área, MinorAxisLength, Eccentricity	85,19%	83,33%	87,78%
#7	MajorAxisLength, Perimeter, ConvexArea, Área, MinorAxisLength, Eccentricity, Extent	87,04%	85,19%	88,89%

Com isso, obtivemos como resultado as acurácias de 90.71% para a ANN demonstrado na Tabela 3 e 91.11% para o LDA na Tabela 2, com o seguinte conjunto de variáveis: MajorAxisLength, Perimeter, ConvexArea, Área, respectivamente. Isso indica que os algoritmos treinados foram capazes de distinguir corretamente os dois tipos de passas na maioria dos casos, mesmo quando aplicados a dados não utilizados durante o treinamento.

No entanto, um algoritmo obteve maior destaque do que os dois anteriores, o Support Vector Machine (SVM), que é conhecido por ter um bom desempenho em conjuntos de

dados de médio a grande porte. Isso ocorre porque a função objetivo do SVM depende apenas dos vetores de suporte, o que reduz a complexidade computacional em relação ao tamanho total do conjunto de dados. Ele apresentou precisão de 91.48%, e com base nesses resultados, essa metodologia demonstrou ser eficaz na classificação de diferentes tipos de passas com base em análises morfológicas.

4 Considerações finais

A pesquisa e a análise realizadas mostraram que a aplicação de métodos e técnicas de aprendizado de máquina, juntamente com a utilização do teste chi-squared e do PCA, podem ser eficazes para melhorar a classificação e compreender as características essenciais das uvas-passas estudadas.

Ao realizar a classificação inicial com o algoritmo SVM e a validação cruzada, obteve-se uma precisão de 82,73%, que não foi considerada satisfatória para o estudo. Isso levou à utilização do PCA para identificar as variáveis mais relevantes com o apoio da representação boxplot, que permitiu uma análise clara das diferenças nas medianas e na variabilidade dos tamanhos entre as espécies de uvas-passas.

Através do teste chi-squared, foi possível determinar a sequência de variáveis que proporciona os melhores resultados. Essa abordagem permitiu alcançar uma precisão de 91,48% na classificação, utilizando as variáveis MajorAxisLength, Perimeter, ConvexArea e Área.

Esses resultados evidenciam a importância de uma análise criteriosa das variáveis e a utilização de técnicas estatísticas adequadas, como o teste chi-squared, para otimizar a classificação e obter uma compreensão mais profunda das características das uvas-passas estudadas.

Em suma, este trabalho demonstrou a importância da seleção adequada de variáveis, da análise estatística e da visualização dos dados para aprimorar a classificação. Essas abordagens podem ser aplicadas em futuros estudos e contribuir para o avanço do conhecimento nessa área e os resultados podem ter implicações práticas na indústria de alimentos, por exemplo, na classificação automatizada de passas em processos de produção.

REFERÊNCIAS

ABBASGHOLIPOUR, Mahdi et al. Image processing with genetic algorithm in a raisin sorting system based on machine vision. In: **Fourth International Conference on Digital Image Processing (ICDIP 2012)**. SPIE, 2012. p. 525-530.

AKHTER, Ravesa; SOFI, Shabir Ahmad. Precision agriculture using IoT data analytics and machine learning. **Journal of King Saud University-Computer and Information Sciences**, v. 34, n. 8, p. 5602-5618, 2022.

AZCARATE, Silvana M. et al. Modeling excitation–emission fluorescence matrices with pattern recognition algorithms for classification of Argentine white wines according grape variety. **Food Chemistry**, v. 184, p. 214-219, 2015.

ÇINAR, İlkay; KOKLU, Murat; TAŞDEMIR, Şakir. Classification of raisin grains using machine vision and artificial intelligence methods. **Gazi Mühendislik Bilimleri Dergisi**, v. 6, n. 3, p. 200-209, 2020.

KARIMI, Navab; KONDROOD, Ramin Ranjbarzadeh; ALIZADEH, Tohid. An intelligent system for quality measurement of Golden Bleached raisins using two comparative machine learning algorithms. **Measurement**, v. 107, p. 68-76, 2017.

KHOJASTEHNAZHAND, Mostafa; RAMEZANI, Hamed. Machine vision system for classification of bulk raisins using texture features. **Journal of Food Engineering**, v. 271, p. 109864, 2020.

MOLLAZADE, Kaveh; OMID, Mahmoud; AREFI, Arman. Comparing data mining classifiers for grading raisins based on visual features. **Computers and electronics in agriculture**, v. 84, p. 124-131, 2012.

WANG, Songjing et al. Application of hybrid image features for fast and non-invasive classification of raisin. **Journal of food engineering**, v. 109, n. 3, p. 531-537, 2012.

WILLIAMSON, Gary; CARUGHI, Arianna. Polyphenol content and health benefits of raisins. **Nutrition Research**, v. 30, n. 8, p. 511-519, 2010.

Recebido em 21/10/2024

Aprovado em 19/11/2024