Análise da Aplicação de Metadados na Detecção de Câncer de Pele

Gabriel Mariano Lobato Pereira

[1]Universidade Tecnológica Federal do Paraná (UTFPR)
– –Campo Mourão – PR – Brazil

[2]Departamento Acadêmico de Computação - Universidade Tecnológica Federal do Paraná
Campo Mourão, Brazil.

3

{gabrielmarianopereira}@alunos.utfpr.edu.br

Abstract. *Skin cancer represents the neoplasm with the highest incidence in Brazil. Among its subtypes, melanoma, although less frequent, presents the highest lethality rates. However, the prognosis is favorable, with survival rates of up to 99% with early diagnosis. In this context, automated detection at early stages is crucial, providing benefits to patients and optimizing public health system resources by reducing the demand for complex and costly procedures. The methodology of this study was based on classic machine learning algorithms—specifically Decision Tree and Random Forest—and on the application of feature engineering, such as one-hot encoding, to enhance the interpretability and performance of the models. The experiments conducted on the ISIC 2024 database resulted in a Macro F1-Score of 0.69 for the Random Forest and 0.65 for the Decision Tree.*

Resumo. *O câncer de pele representa a neoplasia de maior incidência no Brasil. Dentre seus subtipos, o melanoma, embora menos frequente, apresenta as maiores taxas de letalidade. Entretanto, o prognóstico é favorável, com taxas de sobrevivência de até 99% mediante diagnóstico precoce. Nesse contexto, a detecção automatizada em estágios iniciais é crucial, proporcionando benefícios aos pacientes e otimizando os recursos do sistema público de saúde ao reduzir a demanda por procedimentos complexos e onerosos. A metodologia deste estudo baseou-se em algoritmos clássicos de aprendizado de máquina, especificamente Árvore de Decisão e Random Forest, e na aplicação de engenharia de características, como one-hot encoding, para aprimorar a interpretabilidade e o desempenho dos modelos. Os experimentos realizados na base de dados ISIC 2024 resultaram em um F1-Macro de 0.69 para o Random Forest e de 0.65 para a Árvore de Decisão.*

1. Materiais

A base de dados utilizada para os experimentos, ISIC 2024 possui 401.062 imagens de lesões cutâneas tanto benignas quanto casos de câncer de pele. Sua coleção de imagens foi captada de diversas instituições apoiadoras do projeto ISIC, incluindo: *Hospital Clínic de Barcelona, Memorial Sloan Kettering Cancer Center, Hospital of Basel, FNQH*

Cairns, The University of Queensland, Melanoma Institute Australia, Monash University and Alfred Health, University of Athens Medical School, and Medical University of Vienna. Essa diversidade de fontes garante uma representação global, reduzindo o viés regional e permitindo a realização de experimentos mais generalizáveis. Além das imagens, a base de dados também fornece metadados. Estas informações foram anotadas por dermatologistas e são conhecidas por serem indicativas de malignidade. A Tabela 1 mostra os metadados fornecidos.

Tabela 1. Metadados da base de dados ISIC 2024

Metadado	**Descrição**
Target	Classificação binária: benigna ou maligna
Lesion-id	Identificador único de lesão
iddx-full	Classificação completa de diagnóstico de lesão
iddx1	Primeira classificação de diagnóstico de lesão
iddx2	Segunda classificação de diagnóstico de lesão
iddx3	Terceira classificação de diagnóstico de lesão
iddx4	Quarta classificação de diagnóstico de lesão
iddx5	Quinta classificação de diagnóstico de lesão
mel-mitotic-index	Índice mitótico dos melanomas malignos invasivos
mel-thick-mm	Espessura em profundidade da invasão melanoma
tbp-lv-dnn-lesion-confidence	Confiança de pontuação da lesão (0–100)
isic-id	Identificador único do caso
patient-id	Identificador único do paciente
age-approx	Idade aproximada do paciente
sex	Sexo do paciente
anatom-site-general	Localização da lesão no corpo do paciente
clin-size-long-diam-mm	Diâmetro máximo da lesão
image-type	Tipo da imagem
tbp-tile-type	Tipo de iluminação da imagem
tbp-lv-A	A, dentro da lesão
tbplv-Aex	A, fora da lesão
tbp-lv-B	B, dentro da lesão
tbp-lv-Bext	B, fora da lesão
tbp-lv-C	Croma dentro da lesão
tbp-lv-Cext	Croma fora da lesão
tbp-lv-H	Matiz dentro da lesão
tbp-lv-Hext	Matiz fora da lesão
tbp-lv-L	L, dentro da lesão
tbp-lv-Lext	L, fora da lesão
tbp-lv-areaMM2	Área da lesão (mm²)
tbp-lv-area-perim-ratio	Proporção entre o perímetro e área da lesão
tbp-lv-color-std-mean	Irregularidade das cores
tbp-lv-deltaA	Média A, de contraste (Interno e Externo)
tbp-lv-deltaB	Média B, de contraste (Interno e Externo)

(continua)

Tabela 1. Metadados da base de dados ISIC 2024 (continuação)

Metadado	Descrição
tbp-lv-deltaL	Média L, de contraste (Interno e Externo)
tbp-lv-deltaLBnorm	Contraste entre lesão e sua pele ao redor
tbp-lv-eccentricity	Excentricidade
tbp-lv-location	Classificação anatômica: braços, pernas, superiores, inferiores, tronco
tbp-lv-location-simple	Classificação anatômica de localização simples
tbp-lv-minorAxisMM	Menor diâmetro da lesão (mm)
tbp-lv-nevi-confidence	Probabilidade estimada de ser um Nevus (0–100)
tbp-lv-norm-border	Irregularidade da borda (0–10)
tbp-lv-norm-color	Variação das cores (0–10)
tbp-lv-perimeterMM	Perímetro da lesão (mm)
tbp-lv-radial-color-std-max	Assimetria de cores
tbp-lv-stdL	Desvio padrão de L (interno)
tbp-lv-stdLExt	Desvio padrão de L (externo)
tbp-lv-symm-2axis	Assimetria de borda
tbp-lv-symm-2axis-angle	Ângulo da assimetria da borda
tbp-lv-x	Coordenada X da lesão
tbp-lv-y	Coordenada Y da lesão
tbp-lv-z	Coordenada Z da lesão
attribution	Atribuição da imagem, fonte
copyright-license	Licença de copyright

Fonte: Autoria própria.

2. Problemas e Perguntas

A detecção do câncer de pele, é algo oneroso para o estado, necessita de muitas etapas como remoção de amostra e exame histopatológico, utilizando da mão de obra, tempo e dinheiro. Então ao aplicar um sistema auxiliador, que facilite a tomada de decisão para que um paciente tome todo este caminho, pode ser algo beneficial para a sociedade.

Ao utilizar a base de dados ISIC, é possível com que haja esta possibilitar, devido a natureza de seus dados similares a regras para o auxílio da identificação do câncer sem o uso de equipamentos modernos. Então gera a questão de que se os metadados realmente possuem a capacidade de descoberta do câncer de pele.

2.1. Hipótese

Para compreensão da dimensão do problema, foi gerada uma hipótese que pode corroborar com a interpretação do problema e suas possíveis soluções.

- H1: ”A área em que a pinta se encontra, influencia de ser um câncer?”

3. Metodologia e Limitações

Esta seção começará lidando com a limpeza dos dados, seguindo para uma análise exploratória, procurando entender como os dados se comportam nesta base de dados através de

buscas em *SQL*, análise da hipótese gerada para a compreensão do problema, e por último a criação de modelos de aprendizado de máquina.

3.1. Limpeza de Dados

Para começar, foram removidas colunas que não contribuiam para a análise, então as colunas "attribution" e "copyright-license" foram removidas.

Remoção de colunas que possuem acima de 50% dos valores, como valores nulos, pois não agregam a análise neste caso, além de possibilitar com que haja um viés, "iddx-2", "iddx-3", "iddx-4", "iddx-5", "mel-mitotic-index", "mel-thick-mm", "lesion-id".

Por ter removido colunas de possibilitaria viés, serão retiradas características que possuem baixo valor análitico, por conterem apenas um valor, ou por já possuirem ideias similares, porém mais simples, ou tendo muitos valores categóricos, que aumentam a complexidade da interpretação por serem muitos valores distintos com poucas amostras. Facilitando então a intepretação do modelo de classificação no final, então as colunas "image-type", "tbp-lv-location" e "iddx-full" foram excluidas.

Além das colunas que foram removidas anteriormente, as colunas "sex", "age-approx" e "anatom-site-general" possuem 11517, 2798 e 5756 valores nulos respectivamente. Para resolver o problema de valores nulos da coluna de idade, analisando a imagem Figura 1 os valores que não existem foram substituiídos pela mediana, por conta da distribuição ter uma cauda.

E para os outros valores, que se tratam de dados categóricos, para não excluir-los, foram substituídos pela moda.

3.2. Análise Exploratória

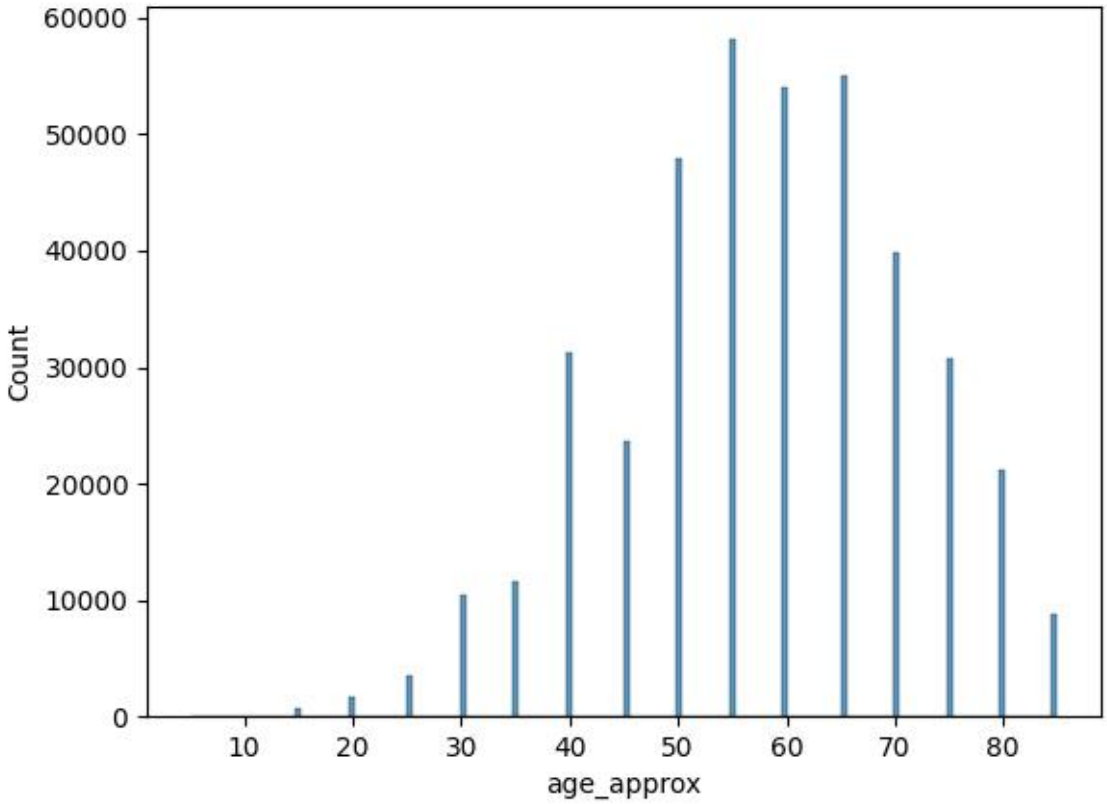
A análise exploratória dos dados buscou compreender padrões comportamentais através de seis questões principais. Observou-se que a mediana de pintas é de 241 em indivíduos sem câncer, contra apenas 1 naqueles com a doença. As lesões malignas concentram-se majoritariamente no tronco, com maior incidência aos 60, 65 e 70 anos e predominância no sexo masculino (frequência 50% superior). A distribuição corporal das pintas por sexo é detalhada na Figura 2.

Na Figura 2 apresenta que para o sexo masculino, a parte posterior do tronco há maior incidência e para o sexo feminino a também, contudo para o segunda parte do corpo que mais possui pintas cancerígenas, são distintas, para os homens ela se encontra na cabeça ou pescoço e para as mulheres, na extremidade inferior.

3.3. Engenharia de Características e Modelagem

Para realizar a modelagem e a engenharia das características, foi necessário realizar uma subamostragem dos dados, para permitir com que seja possível um treinamento com busca exaustiva e que não leve muito tempo. Garantir a aleatoriedade da amostra é importante, então para isso, os dados foram separados em 3 grupos, pintas de pessoas sem câncer, pintas cancerígenas de pessoas e pintas não cancerígenas de pessoas sem câncer. Dos grupos das pintas que não são cancerígenas, foram selecionados aleatoriamente 20 pacientes, resultando então em uma base para treino e teste de 20835 linhas.

Figura 1. Distribuição de idades



Autoria: Própria

Tabela 2. Hiperparâmetros e Valores para a Busca em Grade da Árvore de Decisão

Hiperparâmetro	Valores Testados
criterion	gini, entropy
max_depth	None, 3, 5, 10, 50, 100
min_samples_split	2, 5, 10, 20

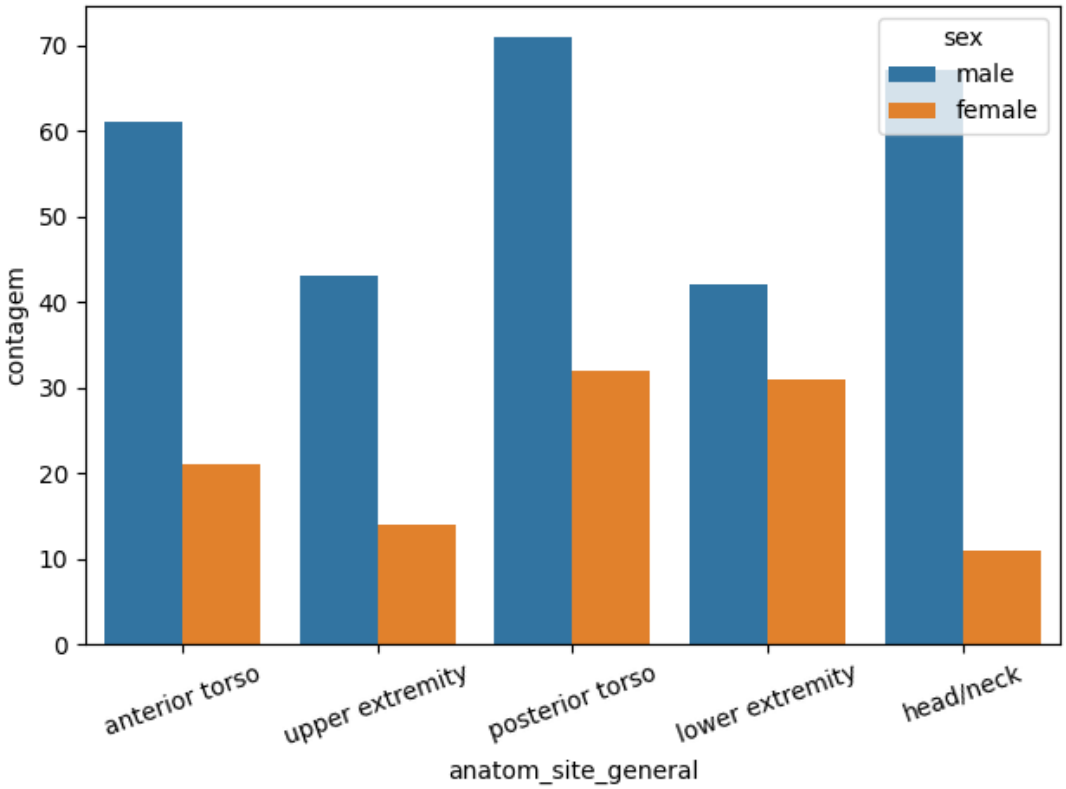
Fonte: Autoria própria.

O treinamento foi realizado utilizando Árvore de Decisão com a busca exaustiva utilizando o parâmetros em Tabela 2.

E *Random Forest* usando os parâmetros em Tabela 3, para obter resultados sólidos, e modelos clássicos de aprendizado de máquina, com a possibilidade de verificar os principais (10 melhores) características para a avaliação do modelo em seu caso teste.

Visando maximizar os resultados, foi necessário aplicar uma engenharia de características, na qual dados categóricos passaram por um *One Hot Encoder*, que distribui seus valores para distintas colunas afirmando a existência desse valor, como foi feito para

Figura 2. Distribuição de Câncer por parte do Corpo (dividido por gênero)



Autoria: Própria

Tabela 3. Hiperparâmetros e Valores para a Busca em Grade do Random Forest

Hiperparâmetro	Valores Testados
n_estimators	50, 100, 200, 300
max_depth	None, 10, 20
min_samples_split	2, 5, 10, 15
class_weight	balanced

Fonte: Autoria própria.

as colunas “sex” e “anatom_site_general”, e para colunas de valor numérico, passaram por um padronizador utilizando *z-score*, exceto a coluna alvo que indica a qual classe a instância pertence.

Para o treino e teste, foi realizada uma validação cruzada em 5 partes, na qual 4 partes serão utilizadas para treinar e 1 para testar, e para a avaliação foi utilizado o F1 Score Macro, devido o desbalanceamento entre classes, com isso será calculada a média dos resultados junto com o desvio padrão.

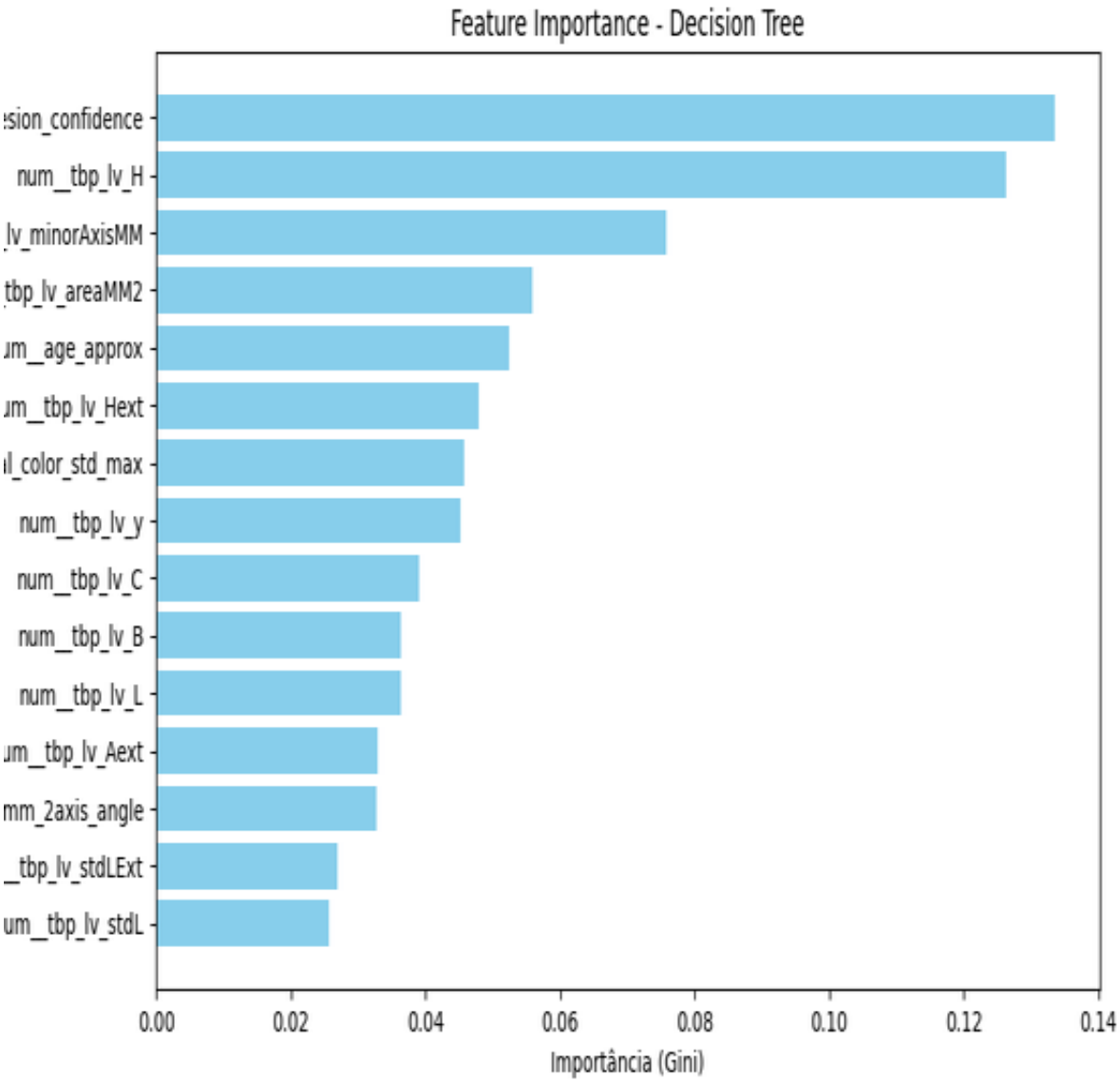
4. Resultados e Discussão

Esta seção será utilizada para avaliar a taxa de acerto dos modelos de classificação e discutir seus resultados, e qual impacto elas geram para casos futuros.

Para o classificador Árvore de Decisão, obteve-se o resultado 0.6564 ± 0.0084, e para o *Random Forest* obteve-se 0.6968 ± 0.0013, mostrando que o segundo modelo obteve um resultado melhor que o primeiro.

Mas o que é possível verificar, é a importância das características utilizadas para realizar a avaliação do primeiro modelo (Figura 4)

Figura 3. 10 Características mais importantes para a Árvore de Decisão



Autoria: Própria

e para o segundo modelo, se distingue para o caso anterior (Figura 4), o que mostra que cada modelo aprende de uma forma distinta.

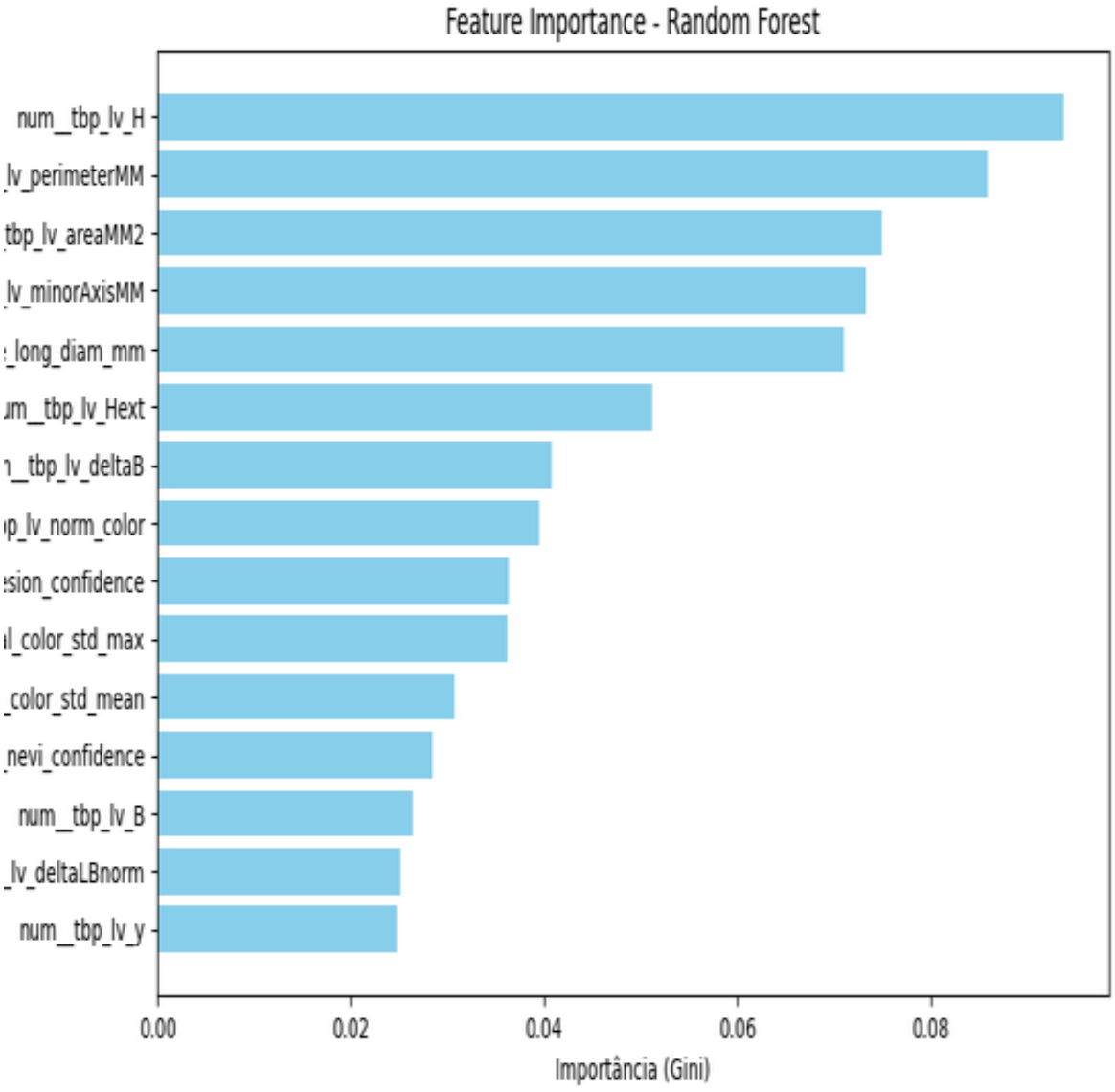
E no primeiro modelo, houve maior importância para características relacionadas

a cor das amostras, como matiz e croma, e para o segundo modelo, obteve-se maior importância, características que representavam medidas de comprimento, como área e perímetros. Mostrando a diferença entre ambos algoritmos e como sua complexidade pode mostrar qual características eles podem priorizar.

5. Trabalhos Futuros

Neste caso, foram analisadas cerca de 40 características, porém algumas demonstraram baixa relevância, e algo que pode ser feita para mitigar esse uso indevido das características, é a aplicação de um algoritmo genético, verificando a melhor combinação de características. E por lidar com mais de um algoritmo de classificação, também pode ser realizada a fusão tardia de características, propostas por **?**, que permitem um resultado mais robusto em troca de um maior custo computacional.

Figura 4. 10 Características mais importantes para *Random Forest*



Autoria: Própria