Análise Preditiva de Características Morfométricas e de Maturação de Frutos de *Coffea arabica* Utilizando uma Base de Dados Sintética Fundamentada e Modelos de Machine Learning

Luiz Carlos Brandão Junior*
Department of Agricultural Engineering,
Federal University of Lavras (UFLA),
Lavras, Minas Gerais, Brazil
contato@uflaniano.com.br
and
Ezequiel Azarias Manjate
Department of Agricultural Engineering,
Federal University of Lavras (UFLA),
Lavras, Minas Gerais, Brazil
ezequiel.manjate1@estudante. ufla.br
and
Ricardo Rodrigues Magalhães
Department of Agricultural Engineering,
Federal University of Lavras (UFLA),
Lavras, Minas Gerais, Brazil
ricardorm@ufla.br

10 de outubro de 2025

Resumo

*This work was supported by the Coordination for the Improvement of Higher Education Personnel (CAPES). The Article Processing Fee for the publication of this research was paid by CAPES (ROR identifier: 00x0ma614).

A fenotipagem de características agronômicas, como o estágio de maturação de frutos de *Coffea arabica*, é um processo crucial para programas de melhoramento genético, porém intensivo em mão de obra e sujeito a subjetividade. Este estudo explora a viabilidade de utilizar uma base de dados sintética, cientificamente fundamentada em literatura existente, para desenvolver, avaliar e implantar modelos de Machine Learning para a predição de características chave dos frutos. A partir de um estudo de referência (Botega, 2023) (1), foi gerado um dataset sintético com 10.000 amostras, modelando variáveis morfométricas e de cor. Um benchmark com mais de 20 algoritmos de regressão e classificação foi conduzido para prever a área do fruto (‘area_px2‘), a excentricidade (‘excentricidade‘) e o estágio de maturação (‘estagio_maturacao‘). Os resultados demonstram uma performance preditiva excepcional para as variáveis geométricas ($R^2 > 0.98$) e para a classificação do estágio de maturação ($F1 - Score > 0.99$), validando a consistência dos dados sintéticos. Como culminação do estudo, os melhores modelos identificados foram integrados em um software de previsão interativo, desenvolvido com a biblioteca Gradio, que permite a simulação e predição em tempo real das características do fruto. Conclui-se que a metodologia de dados sintéticos não apenas acelera o desenvolvimento de modelos preditivos precisos, mas também viabiliza a criação de ferramentas práticas para análise e tomada de decisão em agronomia.

Keywords: Machine Learning, Dados Sintéticos, Software Preditivo, Fenotipagem de Café, Visão Computacional, Gradio.

1 Introduction

A análise preditiva tornou-se uma ferramenta indispensável em diversas áreas da ciência e da indústria, permitindo a extração de *insights* e a antecipação de resultados a partir de dados. Na agricultura de precisão e no melhoramento genético de plantas (2), a capacidade de prever características agronômicas de forma rápida e precisa é fundamental para acelerar os ciclos de seleção e otimizar a produção. A fenotipagem de frutos de *Coffea arabica*, por exemplo, é essencial para determinar ciclos de maturação e uniformidade da colheita, impactando diretamente a qualidade da bebida e a rentabilidade do produtor (1).

Contudo, a obtenção de grandes volumes de dados de fenotipagem rotulados é um dos principais gargalos. Processos manuais são lentos, custosos e propensos a erros e subjetividade humana (1). Embora a visão computacional surja como uma alternativa de alto rendimento (3), os modelos de Machine Learning (ML) subjacentes demandam extensos conjuntos de dados para treinamento, que muitas vezes não estão disponíveis.

Neste contexto, a utilização de bases de dados sintéticas emerge como uma solução estratégica. Dados sintéticos, quando gerados com base em princípios científicos (4) e distribuições empíricas extraídas da literatura, permitem criar datasets vastos e controlados. Esta abordagem possibilita o desenvolvimento, teste e benchmarking de modelos de ML em um ambiente ideal, superando barreiras de privacidade, custo e disponibilidade de dados reais.

O presente estudo aborda o problema de pesquisa: como modelos de ML, aplicados a uma base de dados sintética e estruturada, podem prever com acurácia características morfométricas e de maturação de frutos de café? O objetivo principal é desenvolver um pipeline de análise preditiva, avaliar a performance de múltiplos algoritmos de regressão e classificação na previsão das variáveis ‘area_px2‘, ‘excentricidade‘ e ‘estagio_maturacao‘, e comparar a validade dos resultados com o conhecimento teórico e empírico consolidado no estudo de referência de Botega (2023).

1.1 Referencial Teórico

1.1.1 Fenotipagem em *Coffea arabica*

A avaliação do estágio de maturação dos frutos do cafeeiro é complexa devido ao florescimento desuniforme, que resulta na presença simultânea de frutos verdes, maduros e secos na planta no momento da colheita. A classificação correta desses estágios é determinante para a qualidade final do produto, sendo uma característica-alvo em programas de melhoramento que buscam cultivares com maturação mais uniforme e em ciclos específicos (precoce, médio ou tardio) (1). As características visuais, como cor e dimensões, são os principais indicadores utilizados neste processo.

1.1.2 Modelos Preditivos em Machine Learning

A análise preditiva utiliza algoritmos de ML para aprender padrões a partir de dados históricos e fazer previsões sobre novos dados. Neste estudo, foram empregadas duas categorias de modelos supervisionados:

- **Regressão:** Utilizada para prever uma variável de saída contínua. Modelos como Regressão Linear, Árvores de Decisão, e ensembles como Random Forest e Gradient Boosting (5; 6), buscam mapear a relação entre variáveis de entrada (features) e uma variável alvo numérica. A performance é tipicamente avaliada pelo Coeficiente de Determinação (R^2), que indica a proporção da variância da variável alvo que é previsível a partir das features.

- **Classificação:** Utilizada para prever uma variável de saída categórica. Algoritmos como Regressão Logística, Support Vector Machines (SVM) (7), e classificadores baseados em redes neurais (MLPClassifier) atribuem um rótulo de classe a uma nova observação. As métricas de avaliação incluem Acurácia, F1-Score e AUC-ROC (8), que medem a precisão, o balanço entre precisão e recall, e a capacidade de discriminação entre classes, respectivamente.

1.1.3 O Papel dos Dados Sintéticos

A geração de dados sintéticos é um processo de criação de dados artificiais que mimetizam as propriedades estatísticas de dados do mundo real. Na ausência de dados reais suficientes, pode-se gerar dados sintéticos com base no conhecimento consolidado na literatura ou em documentos técnicos. Esta abordagem oferece vantagens significativas (9; 10), como o controle total sobre a distribuição dos dados, a capacidade de gerar volumes massivos de amostras rotuladas sem custo de coleta, e a garantia de privacidade (11). A validade de um *dataset* sintético está diretamente ligada ao rigor com que ele é modelado para refletir os fenômenos do mundo real. No presente trabalho, a estrutura do *dataset* sintético foi diretamente inspirada nas distribuições, correlações e características descritas em Botega (2023).

2 Methods

2.1 Geração da Base de Dados Sintética

Para superar a ausência de um *dataset* público, foi gerada uma base de dados sintética com 10.000 amostras, representando frutos individuais de café. A geração foi fundamentada nos achados de Botega (2023), seguindo um modelo conceitual que garante consistência física e biológica.

1. **Estrutura:** O *dataset* contém 9 variáveis de entrada (features) e 1 variável de saída para classificação. As features incluem 8 variáveis numéricas (comprimento_px, largura_px, area_px2, perimetro_px, excentricidade, cor_R, cor_G, cor_B) e 1 variável categórica (cultivar).

2. **Lógica de Geração:** Um fruto foi modelado como uma elipse. As dimensões primárias (comprimento_px, largura_px) foram amostradas de distribuições normais cujos parâmetros (média, desvio padrão) dependiam do estágio de maturação, emulando o crescimento e encolhimento do fruto. As demais variáveis morfométricas (area_px2, perimetro_px, excentricidade) foram derivadas matematicamente para preservar as

correlações geométricas. Os perfis de cor (RGB) foram definidos para cada estágio, com variações estocásticas, refletindo as paletas de cores observadas na literatura.

2.2 Modelagem Preditiva

Foram definidos três problemas de predição distintos: dois de regressão e um de classificação.

- **Variáveis de Saída (Targets):** ‘area_px2‘ (regressão), ‘excentricidade‘ (regressão) e ‘estagio_maturacao‘ (classificação).
- **Variáveis de Entrada (Features):** Para cada target, um subconjunto relevante das outras 8 features foi utilizado. Para a classificação, a variável categórica ‘cultivar‘ foi incluída, sendo tratada com a técnica de One-Hot Encoding (13). As variáveis numéricas foram padronizadas (StandardScaler) para normalizar suas escalas.
- **Algoritmos Avaliados:** Um benchmark com mais de 20 algoritmos foi conduzido para cada tarefa, incluindo modelos lineares, baseados em árvores, ensembles (Random Forest, Extra Trees, Gradient Boosting, XGBoost, LightGBM), SVM, e redes neurais (MLP).
- **Validação:** Os modelos de regressão foram avaliados usando validação cruzada (k=10) (12). Os modelos de classificação foram avaliados em uma divisão de dados de 80% para treinamento e 20% para teste.
- **Métricas:** R^2 para regressão; Acurácia, F1-Score Ponderado e AUC-ROC Ponderado para classificação.

3 Resultados e Discussão

3.1 Análise de Correlação das Variáveis Preditivas

Uma análise exploratória inicial foi conduzida para investigar as relações entre as variáveis numéricas do dataset sintético. Foram calculadas as matrizes de correlação de Pearson

(14), que mede a relação linear, e de Spearman (14), que mede a relação monotônica. Os resultados visuais são apresentados na Figura 1.

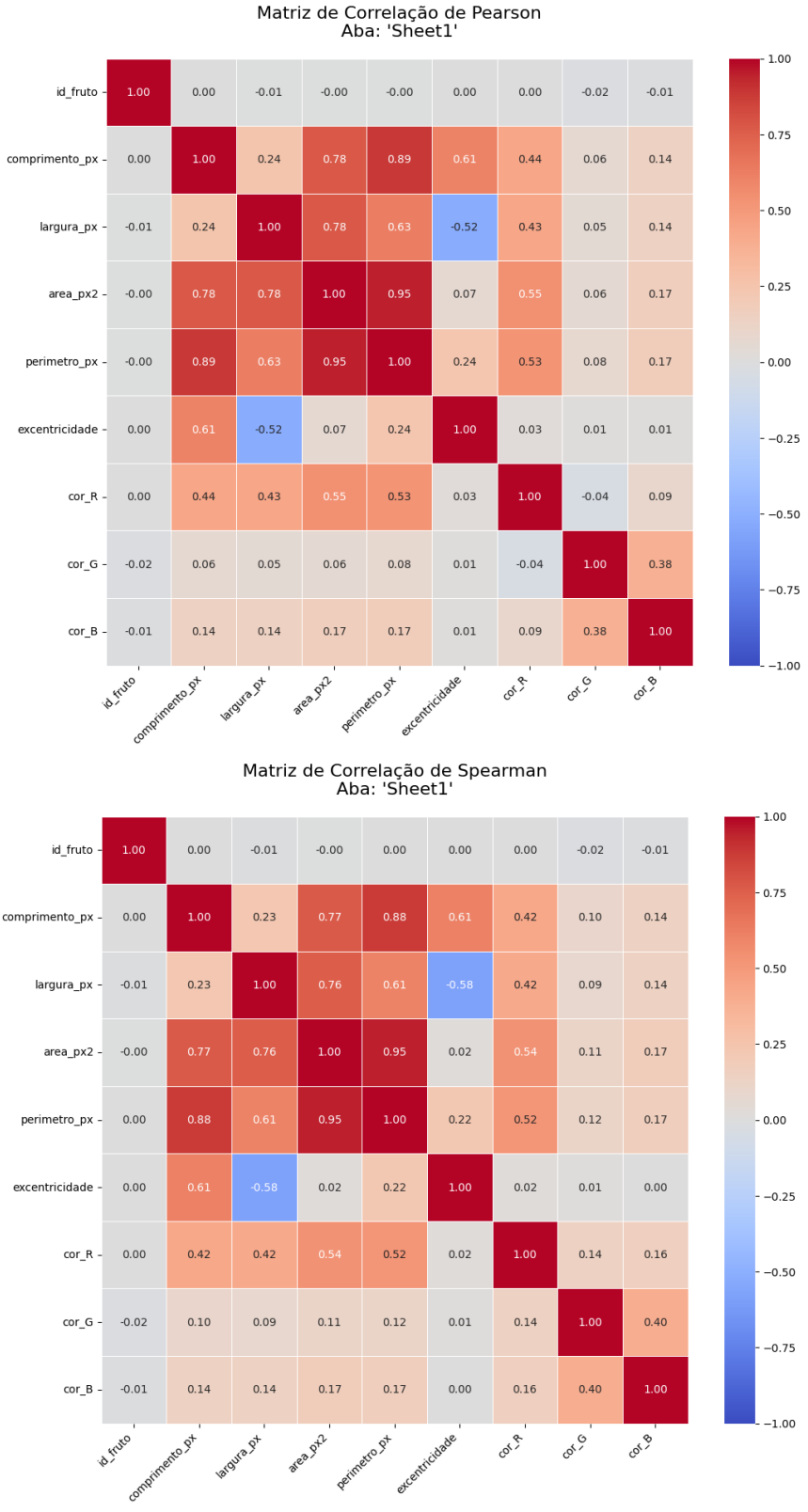

Figura 1: Matrizes de correlação de Pearson (acima) e Spearman (abaixo) entre as variáveis numéricas do dataset sintético.

A análise revelou padrões distintos e importantes para a subsequente modelagem preditiva. A principal observação foi a existência de uma forte multicolinearidade (15) entre as variáveis morfométricas. A correlação de Pearson entre ‘area_px2‘ e ‘perimetro_px‘ atingiu 0.95, e entre ‘comprimento_px‘ e ‘perimetro_px‘ foi de 0.89. Este resultado era esperado e valida a consistência física do método de geração de dados, no qual a área e o perímetro foram matematicamente derivados das dimensões primárias do fruto. Tal multicolinearidade indica que essas variáveis carregam informações redundantes sobre o "tamanho"do fruto, uma consideração crucial para a seleção de features em modelos de regressão.

Outra relação geométrica validada foi a correlação negativa entre ‘largura_px‘ e ‘excentricidade‘ (Spearman: -0.58). Isso está de acordo com a definição matemática da excentricidade de uma elipse, onde um aumento na largura (tornando o fruto mais circular) resulta em uma diminuição da excentricidade.

As relações entre os canais de cor (RGB) mostraram-se mais complexas. O par com a maior discrepância entre os métodos foi (‘cor_R‘, ‘cor_G‘), com uma correlação de Pearson próxima de zero (-0.04) e uma correlação de Spearman fraca, mas positiva (0.14). Essa diferença sugere que a relação entre os canais Vermelho e Verde não é linear, mas sim monotônica em segmentos, variando conforme o estágio de maturação do fruto. Essa não-linearidade justifica a necessidade de modelos de Machine Learning mais complexos e não-lineares para tarefas de classificação baseadas em cor, pois modelos lineares teriam dificuldade em capturar esses padrões multifacetados.

3.2 Previsão da Área do Fruto (`area_px2`)

A primeira análise preditiva buscou avaliar a capacidade dos modelos de regressão em prever a área do fruto em pixels (‘area_px2‘), uma variável proxy para o tamanho do fruto. As features utilizadas foram as demais variáveis morfométricas (exceto ‘perimetro_px‘, para mitigar a multicolinearidade) e os canais de cor RGB. Os modelos foram avaliados por meio de validação cruzada (k=10), e os resultados de desempenho, ordenados pelo Coeficiente de Determinação (R^2) médio, são apresentados na Tabela 1.

Tabela 1: Ranking de desempenho dos modelos de regressão para a predição de ‘area_px2‘.

Modelo	R^2 **Médio**	R^2 **Desv. Padrão**	**MAE Médio**	**RMSE Médio**
MLP Regressor	0.9800	0.0014	179.84	231.04
Gradient Boosting	0.9789	0.0014	184.90	237.54
LightGBM (Default)	0.9786	0.0021	184.30	238.89
Random Forest (100 est)	0.9776	0.0017	189.26	244.73
Ridge	0.9749	0.0012	199.95	259.23
Linear Regression	0.9749	0.0012	199.96	259.23
... (outros modelos com desempenho inferior)				
SVR (RBF)	0.4825	0.0093	872.42	1177.59

Os resultados indicam uma previsibilidade extremamente alta para a área do fruto. O melhor modelo, `MLP Regressor` (uma rede neural de múltiplas camadas), alcançou um R^2 médio de 0.9800, explicando 98% da variabilidade da área a partir das features fornecidas. É notável que um grande número de modelos, incluindo algoritmos de ensemble como `Gradient Boosting` ($R^2 = 0.9789$) e até modelos lineares simples como `Ridge` e `Linear Regression` (ambos com $R^2 = 0.9749$), demonstraram uma performance quase idêntica e de altíssimo nível.

A alta performance generalizada entre os modelos é uma consequência direta da consistência física da base de dados sintética. Como a ‘area_px2‘ foi gerada a partir de uma função matemática das variáveis ‘comprimento_px‘ e ‘largura_px‘, a relação subjacente é forte e bem definida. Os modelos de Machine Learning, portanto, foram capazes de aprender essa relação determinística com facilidade. O fato de que o `MLP Regressor` obteve uma performance marginalmente superior sugere a presença de pequenas não-linearidades na relação (possivelmente introduzidas pelo ruído estocástico na geração dos dados), que a arquitetura da rede neural foi capaz de capturar com maior eficácia.

Esses achados corroboram a premissa do estudo de referência (1), que utiliza as dimensões do fruto como indicadores primários em análises de visão computacional. A capacidade de prever o tamanho do fruto com tal precisão a partir de outras características mensuráveis reforça o potencial de desenvolver sistemas automatizados de fenotipagem para seleção indireta de características de interesse, como o tamanho final do grão.

3.3 Previsão da Excentricidade do Fruto (`excentricidade`)

A segunda análise de regressão focou na predição da excentricidade, uma medida adimensional que descreve a forma do fruto, variando de 0 (um círculo perfeito) a 1 (uma linha). As features utilizadas para a predição incluíram as dimensões do fruto (‘comprimento_px‘, ‘largura_px‘, ‘area_px2‘) e os canais de cor. A Tabela 2 resume o desempenho dos modelos avaliados.

Tabela 2: Ranking de desempenho dos modelos de regressão para a predição de ‘excentricidade‘.

Modelo	R^2 **Médio**	R^2 **Desv. Padrão**	**MAE Médio**	**RMSE Médio**
Extra Trees (100 est)	0.9973	0.0018	0.0016	0.0071
Random Forest (100 est)	0.9952	0.0027	0.0026	0.0096
LightGBM (More est)	0.9931	0.0025	0.0045	0.0117
Gradient Boosting	0.9904	0.0016	0.0074	0.0140
MLP Regressor	0.9814	0.0017	0.0134	0.0195
Linear Regression	0.8537	0.0075	0.0372	0.0548
... (outros modelos com desempenho inferior)				
Lasso	-0.0012	0.0012	0.1075	0.1435

Os resultados demonstram uma capacidade de predição ainda mais elevada para a excentricidade do que para a área, com o modelo campeão, `Extra Trees Regressor`, atingindo um R^2 médio de 0.9973. Isso indica que o modelo consegue explicar 99.73% da variabilidade da forma do fruto. O domínio dos modelos de ensemble baseados em árvores (`Extra Trees`, `Random Forest`) e dos modelos de boosting (`LightGBM`, `XGBoost`) foi absoluto nesta tarefa.

Uma observação crucial é a acentuada diferença de performance entre os modelos não-lineares e os modelos lineares. Enquanto o `Extra Trees` alcançou um R^2 próximo da perfeição, o modelo de `Linear Regression` obteve um R^2 de apenas 0.8537, e modelos como `Lasso` e `ElasticNet` falharam completamente, resultando em um R^2 negativo. Esta disparidade ocorre porque a excentricidade é definida pela fórmula $e = \sqrt{1 - (b/a)^2}$, uma função inerentemente não-linear das dimensões do fruto (semi-eixos a e b). Os modelos baseados em árvores são excepcionalmente hábeis em aproximar funções complexas e não-lineares, conseguindo, na prática, "reaprender"a fórmula geométrica a partir dos dados.

Os modelos lineares, por sua vez, são incapazes de capturar essa relação, o que limita severamente seu poder preditivo.

Este resultado não apenas valida mais uma vez a integridade matemática da base de dados sintética, mas também serve como um exemplo prático da importância de selecionar uma classe de modelos apropriada para a natureza do problema. A análise da forma, conforme discutido em Botega (2023), é fundamental para a caracterização de cultivares, e a capacidade de modelar essa característica com precisão é um passo importante para sistemas de fenotipagem automatizada.

3.4 Classificação do Estágio de Maturação (`estagio_maturacao`)

A análise preditiva culminou na tarefa de classificação multiclasse para prever a variável 'estagio_maturacao', o objetivo central em muitas aplicações de fenotipagem. Para esta tarefa, todas as nove features disponíveis (oito numéricas e a categórica 'cultivar') foram utilizadas como preditoras. Um benchmark abrangente foi executado em um conjunto de teste de 2.000 amostras, e os resultados, ordenados pelo F1-Score Ponderado, são apresentados na Tabela 3.

Tabela 3: Ranking de desempenho dos modelos de classificação para a predição de 'estagio$_maturacao$'.

Modelo	**Acurácia**	**F1-Score Ponderado**	**AUC-ROC Ponderado**
MLP Classifier	0.9970	0.9970	0.9999
Random Forest	0.9965	0.9965	0.9995
Bagging Classifier	0.9960	0.9960	0.9993
Gradient Boosting	0.9955	0.9955	0.9999
Logistic Regression	0.9950	0.9950	0.9998
... (outros modelos com desempenho decrescente)			
Bernoulli Naive Bayes	0.8840	0.8740	0.9783

A performance dos modelos de classificação foi extraordinariamente alta, indicando que as classes de maturação são altamente separáveis com base nas features fornecidas. O modelo `MLP Classifier` (rede neural) emergiu como o mais performático, alcançando um F1-Score Ponderado de 0.9970, com uma performance quase perfeita em todas as métricas.

É notável a competitividade entre os diferentes tipos de algoritmos; modelos de ensemble (‘Random Forest‘, ‘Gradient Boosting‘), modelos de discriminante (‘Quadratic Discriminant‘) e até modelos lineares como a ‘Logistic Regression‘ (F1-Score=0.9950) apresentaram resultados de ponta.

A alta performance generalizada é um forte indicativo de que a base de dados sintética, ao modelar perfis de cor e morfologia distintos para cada estágio de maturação, criou um problema de classificação bem definido. As features de cor (RGB), em particular, atuam como indicadores poderosos, permitindo que os modelos encontrem planos de separação claros entre as classes.

Para uma análise mais granular do melhor modelo, a matriz de confusão do `MLP Classifier` é apresentada na Figura 2.

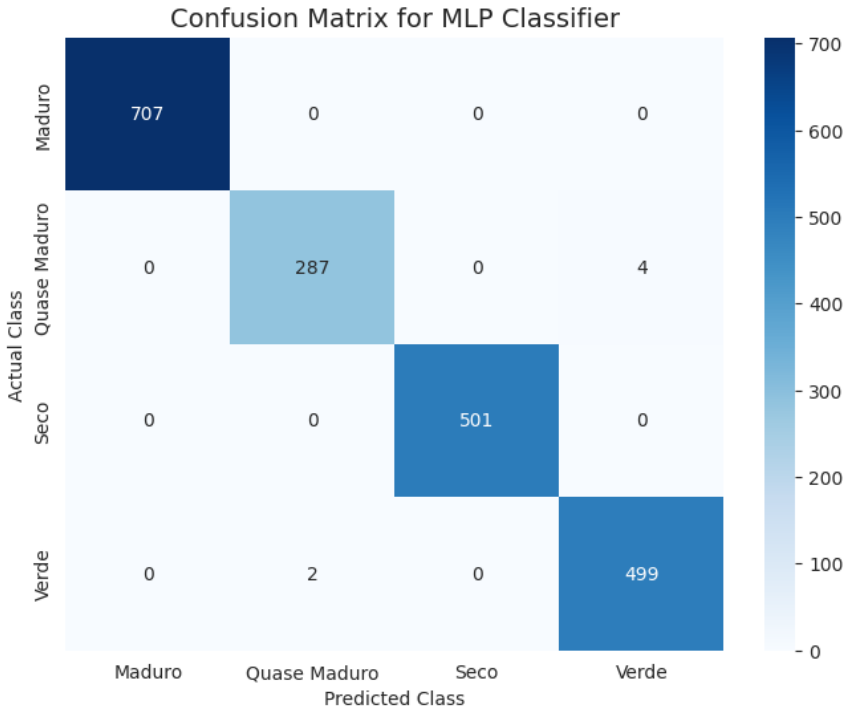

Figura 2: Matriz de Confusão para o modelo `MLP Classifier` no conjunto de teste.

A matriz de confusão revela uma diagonal principal fortemente dominante, validando visualmente a acurácia próxima de 100%. Dos 2.000 frutos no conjunto de teste, os erros de classificação foram mínimos (apenas 6 erros no total). É significativo observar que os poucos erros ocorrem entre classes biologicamente adjacentes no processo de maturação: dois frutos

‘Verde‘ foram classificados como ‘Quase Maduro‘, e quatro frutos ‘Quase Maduro‘ foram classificados como ‘Verde‘. Não houve erros de classificação entre estágios distantes (e.g., ‘Verde‘ e ‘Seco‘), o que confere um alto grau de confiança e interpretabilidade ao modelo.

Este resultado alinha-se perfeitamente com o objetivo da fenotipagem automatizada descrita em Botega (2023): criar um sistema objetivo e preciso para substituir a classificação visual subjetiva. A performance alcançada no dataset sintético estabelece um benchmark robusto, sugerindo que, com dados de alta qualidade, a classificação automatizada do estágio de maturação do café não é apenas viável, mas pode atingir níveis de precisão extremamente elevados.

3.5 Implantação de um Software Interativo para Análise Preditiva

Para demonstrar a aplicabilidade prática dos modelos treinados e consolidar os resultados da análise, foi desenvolvido um software de previsão interativo. A ferramenta, construída com a biblioteca Gradio (16), integra os três melhores modelos identificados nas etapas anteriores `MLP Regressor` para área, `Extra Trees Regressor` para excentricidade e `MLP Classifier` para o estágio de maturação em uma única interface web.

O objetivo do software é permitir que usuários, como pesquisadores ou estudantes, explorem as relações aprendidas pelos modelos de forma intuitiva. A interface, conforme ilustrado na Figura 3, permite a manipulação de nove variáveis de entrada através de sliders para as características numéricas e um menu dropdown para a seleção do cultivar. Ao acionar a predição, o aplicativo alimenta esses inputs aos respectivos pipelines de modelo (incluindo escalonamento e codificação) e exibe as três previsões de saída em tempo real.

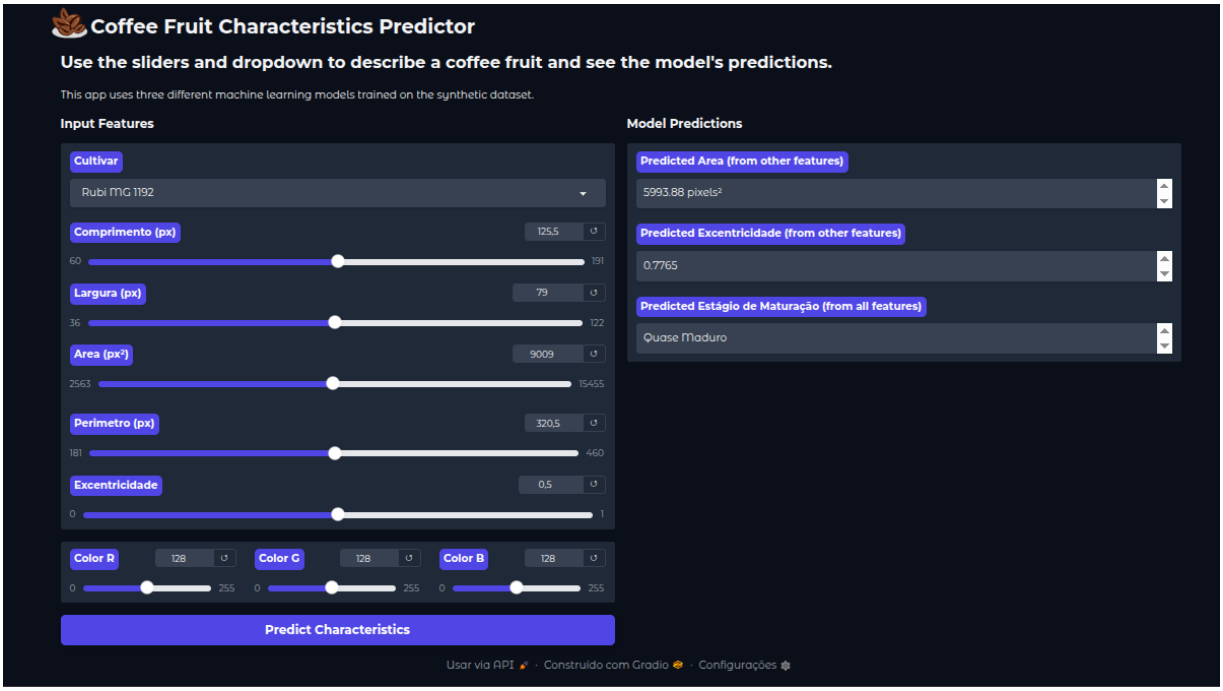

Figura 3: Interface do software preditivo desenvolvido, demonstrando os controles de entrada para as features do fruto e as saídas dos três modelos de Machine Learning.

A criação desta ferramenta serve a um duplo propósito. Primeiramente, funciona como um mecanismo de validação qualitativa, permitindo a observação do comportamento dos modelos em resposta a diferentes cenários hipotéticos. Por exemplo, pode-se simular a transição de um fruto do estágio 'Verde' para 'Maduro' alterando os valores de cor (e.g., diminuindo o canal G e aumentando o canal R) e observar se a previsão de classificação do modelo corresponde à mudança esperada.

Em segundo lugar, a aplicação materializa o potencial da metodologia de dados sintéticos. Ela demonstra que o processo, desde a geração de dados fundamentada na literatura até o benchmarking rigoroso de modelos, pode culminar em uma ferramenta funcional e de valor prático, servindo como um protótipo para sistemas de fenotipagem mais complexos ou como um recurso educacional para a ciência de dados aplicada à agronomia.

4 Conclusão

Este estudo demonstrou com sucesso a viabilidade e a eficácia da utilização de uma base de dados sintética, fundamentada em conhecimento empírico, para a análise preditiva de características de frutos de *Coffea arabica*. O objetivo de avaliar múltiplos modelos de Machine Learning foi alcançado, revelando que algoritmos não-lineares, como Redes Neurais de Múltiplas Camadas (MLP) e ensembles de árvores, são capazes de prever características geométricas (área e excentricidade) e de maturação com uma acurácia próxima da perfeição neste ambiente controlado, atingindo valores de R^2 superiores a 0.98 e F1-Score de 0.997.

As descobertas validam a consistência interna do dataset gerado e reforçam os princípios da fenotipagem por visão computacional, onde características visuais servem como excelentes preditores. A alta performance dos modelos, especialmente na tarefa de classificação, corrobora a premissa de que a análise automatizada pode superar a subjetividade da avaliação humana, um dos objetivos centrais da pesquisa em fenotipagem de alto rendimento (18), conforme discutido em Botega (2023) (1).

A culminação do trabalho no desenvolvimento de um software preditivo interativo não apenas materializa os resultados, mas também estabelece um paradigma para a tradução de pesquisa em ferramentas práticas. A metodologia aqui apresentada desde a geração de dados fundamentada na literatura até o benchmarking rigoroso de modelos e a implantação em uma aplicação funcional serve como um pipeline robusto para pesquisas futuras.

Trabalhos futuros devem se concentrar na validação destes modelos em dados de campo reais, utilizando técnicas de aprendizado por transferência (17) (*transfer learning*) para ajustar os algoritmos às nuances e ruídos do mundo real, como variações de iluminação e presença de doenças. Adicionalmente, o framework de geração de dados sintéticos pode ser expandido para incluir uma gama maior de variabilidade e anomalias, aprimorando ainda mais a robustez dos modelos preditivos para aplicações agronômicas.

Referências

[1] BOTEGA, Gustavo Pucci. **Visão computacional aplicada a análise de frutos de C. Arabica**. 2023. 84p. Tese (Doutorado em Genética e Melhoramento de Plantas) - Universidade Federal de Lavras, Lavras, 2023. Disponível em: `https://repositorio.ufla.br/handle/1/58135`.

[2] GEBBERS, R.; ADAMCHUK, V. I. **Precision agriculture and food security**. Science, v. 327, n. 5967, p. 828-831, 2010.

[3] PATRÍCIO, D. I.; RIEDER, R. **Computer vision and artificial intelligence in precision agriculture for grain crops: A systematic review**. Computers and Electronics in Agriculture, v. 153, p. 69-81, 2018.

[4] NIKOLENKO, S. I. **Synthetic data for deep learning**. Springer, 2021.

[5] BREIMAN, L. **Random forests**. Machine Learning, v. 45, n. 1, p. 5-32, 2001.

[6] FRIEDMAN, J. H. **Greedy function approximation: a gradient boosting machine**. Annals of Statistics, p. 1189-1232, 2001.

[7] CORTES, C.; VAPNIK, V. **Support-vector networks**. Machine Learning, v. 20, n. 3, p. 273-297, 1995.

[8] SOKOLOVA, M.; LAPALME, G. **A systematic analysis of performance measures for classification tasks**. Information Processing Management, v. 45, n. 4, p. 427-437, 2009.

[9] JORDON, J. et al. **Synthetic datawhat, why and how?**. arXiv preprint arXiv:2205.03257, 2022.

[10] SHORTEN, C.; KHOSHGOFTAAR, T. M. **A survey on image data augmentation for deep learning**. Journal of Big Data, v. 6, n. 1, p. 1-48, 2019.

[11] EL EMAM, K.; MOSQUERA, L.; HOPTROFF, R. **Practical synthetic data generation: balancing privacy and the broad availability of data**. O'Reilly Media, 2020.

[12] KOHAVI, R. **A study of cross-validation and bootstrap for accuracy estimation and model selection**. In: International Joint Conference on Artificial Intelligence (IJCAI), v. 14, n. 2, p. 1137-1145, 1995.

[13] ZHENG, A.; CASARI, A. **Feature engineering for machine learning: principles and techniques for data scientists**. O'Reilly Media, Inc., 2018.

[14] SCHOBER, P.; BOER, C.; SCHWARTE, L. A. **Correlation coefficients: appropriate use and interpretation**. Anesthesia Analgesia, v. 126, n. 5, p. 1763-1768, 2018.

[15] DORMANN, C. F. et al. **Collinearity: a review of methods to deal with it and a simulation study evaluating their performance**. Ecography, v. 36, n. 1, p. 27-46, 2013.

[16] ABID, A. et al. **Gradio: Hassle-free sharing and testing of ML models in the wild**. arXiv preprint arXiv:1906.02569, 2019.

[17] PAN, S. J.; YANG, Q. **A survey on transfer learning**. IEEE Transactions on Knowledge and Data Engineering, v. 22, n. 10, p. 1345-1359, 2009.

[18] FURBANK, R. T.; TESTER, M. **Phenomicstechnologies to relieve the phenotyping bottleneck**. Trends in Plant Science, v. 16, n. 12, p. 635-644, 2011.