Relatório Técnico e Análise de Resultados

Sistema de Classificação Automatizada de Grãos de Trigo

Um projeto de Machine Learning para otimização de processos em cooperativas agrícolas, baseado na metodologia CRISP-DM.

Autores:
Andre de Oliveira Santos Burger - RM 565150
Diogo Rebello dos Santos - RM 565286
Marcos Vinícius dos Santos Fernandes - RM 565555
Vera Maria Chaves de Souza - RM 565497

Data de Emissão: 13 de junho de 2025

1. INTRODUÇÃO

No cenário agrícola atual, a eficiência e a precisão na classificação de produtos são cruciais para garantir a qualidade e otimizar os processos. Em cooperativas agrícolas de pequeno porte, a classificação de grãos de trigo é tradicionalmente realizada de forma manual por especialistas. Embora essa abordagem garanta a qualidade, ela é inerentemente demorada, custosa e suscetível a variações e erros humanos, o que pode impactar negativamente a produtividade e a padronização do produto final.

Com o avanço exponencial do aprendizado de máquina (Machine Learning), surge uma oportunidade transformadora para automatizar e aprimorar significativamente esse processo. A aplicação de algoritmos de Machine Learning permite a criação de sistemas capazes de analisar características físicas dos grãos e classificá-los de forma rápida, consistente e com alta precisão, minimizando a intervenção humana e os erros associados.

Este relatório detalha o desenvolvimento de uma solução completa de Machine Learning para a classificação de variedades de grãos de trigo (Kama, Rosa e Canadian) com base em suas propriedades geométricas. O projeto foi concebido e executado seguindo a metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining), uma abordagem estruturada que garante a abrangência e a robustez de todas as etapas, desde a compreensão do negócio e dos dados até a avaliação e implantação do modelo.

O objetivo principal é demonstrar a viabilidade e os benefícios da automação da classificação de grãos, fornecendo uma análise aprofundada dos dados, a implementação e otimização de diversos algoritmos de Machine Learning, e uma interpretação detalhada dos resultados. O relatório também aborda os insights estratégicos derivados da análise, as recomendações práticas para a implementação em ambientes reais de cooperativas e os próximos passos para o aprimoramento contínuo da solução. Através desta iniciativa, buscamos não apenas apresentar uma solução técnica eficaz, mas também evidenciar o potencial do Machine Learning para revolucionar as operações agrícolas, impulsionando a eficiência, a precisão e a competitividade no setor.

2. DESENVOLVIMENTO DA SOLUÇÃO

O desenvolvimento da solução de classificação de grãos de trigo com Machine Learning seguiu rigorosamente as fases da metodologia CRISP-DM, garantindo uma abordagem sistemática e abrangente. Cada etapa foi cuidadosamente planejada e executada para construir um modelo robusto e eficaz.

2.1. COMPREENSÃO DO NEGÓCIO E DOS DADOS

A fase inicial focou na compreensão aprofundada do problema de negócio e na coleta de informações sobre o dataset. O contexto é a classificação manual de grãos em cooperativas agrícolas de pequeno porte, um processo que, embora realizado por especialistas, é demorado e propenso a erros humanos. A meta é automatizar essa classificação para aumentar a eficiência e a precisão.

O dataset utilizado, proveniente de uma fonte confiável [1], contém 210 amostras de grãos de trigo, cada uma caracterizada por sete atributos físicos e uma classe que indica a variedade do grão.

As características são:

- **Área do grão:** Medida da superfície do grão;
- **Perímetro:** Comprimento do contorno do grão;
- **Compacidade:** Uma medida da circularidade do grão, calculada como 4 * pi * Área / (Perímetro^2);
- **Comprimento do núcleo:** Comprimento do eixo principal do grão;
- **Largura do núcleo:** Comprimento do eixo secundário do grão;
- **Coeficiente de assimetria:** Medida da assimetria da forma do grão;
- **Comprimento do sulco do núcleo:** Comprimento do sulco no grão;
- **Classe:** Variedade do grão (1=Kama, 2=Rosa, 3=Canadian).

2.2. ANÁLISE EXPLORATÓRIA DE DADOS (EDA) E PRÉ-PROCESSAMENTO

Esta fase foi crucial para entender a estrutura, a qualidade e as relações presentes nos dados, além de prepará-los para o treinamento dos modelos. As principais atividades incluíram:

2.2.1. Carregamento e Limpeza Inicial

O dataset original, fornecido em formato de texto, apresentava inconsistências na separação das colunas devido a múltiplos caracteres de tabulação. Para garantir a correta leitura dos dados, foi implementado um script de limpeza que substituiu sequências de tabulações por uma única tabulação antes do carregamento. O código abaixo ilustra essa etapa:

```python
import re
import pandas as pd

# Ler o arquivo e limpar tabs extras
with open('seeds_dataset.txt', 'r') as f:
    content = f.read()
clean_content = re.sub(r'\t+', '\t', content)

# Salvar o arquivo limpo temporariamente
with open('seeds_dataset_clean.txt', 'w') as f:
    f.write(clean_content)

# Carregar o dataset limpo
df = pd.read_csv('seeds_dataset_clean.txt', sep='\t',
header=None)

# Nomear as colunas
column_names = ['Area', 'Perimetro', 'Compacidade',
'Comprimento_Nucleo',
                'Largura_Nucleo', 'Coef_Assimetria',
'Comprimento_Sulco', 'Classe']
df.columns = column_names

# Mapear classes para nomes descritivos
class_mapping = {1: 'Kama', 2: 'Rosa', 3: 'Canadian'}
df['Classe_Nome'] = df['Classe'].map(class_mapping)

print(f"Dimensões do dataset: {df.shape}")
print(f"Total de amostras: {len(df)}")
print(f"Número de features: {len(df.columns)-2}")
```

Fig. 1 – Código de carregamento e limpeza do material.

Após a limpeza e carregamento, o dataset foi inspecionado, confirmando 210 amostras e 7 características numéricas, além da coluna de classe. A coluna numérica da classe foi mapeada para nomes descritivos (Kama , Rosa , Canadian) para facilitar a interpretação.

2.2.2. Análise de Qualidade dos Dados

Uma verificação de valores ausentes foi realizada, e o resultado foi positivo: nenhuma das características apresentava dados faltantes. Isso eliminou a necessidade de estratégias de imputação, simplificando o pipeline de pré-processamento.

2.2.3. Distribuição das Classes

A análise da distribuição das classes revelou um dataset perfeitamente balanceado, com 70 amostras para cada uma das três variedades de trigo (Kama, Rosa, Canadian). Essa característica é ideal para o treinamento de modelos de classificação, pois previne o viés de classe e garante que o modelo não seja otimizado para uma classe majoritária em detrimento das minoritárias.

2.2.4. Estatísticas Descritivas e Visualizações

As estatísticas descritivas (média, desvio padrão, mínimo, máximo, quartis) foram calculadas para cada característica, fornecendo um panorama quantitativo da distribuição dos dados. Complementarmente, diversas visualizações foram geradas para explorar as distribuições das características, as relações entre elas e a separabilidade das classes:

- **Histogramas:** Mostraram a distribuição individual de cada característica.
  Observou-se que a maioria das features apresentava uma distribuição próxima da normal, embora algumas, como o Coeficiente de Assimetria, exibissem uma cauda mais longa, indicando a presença de valores mais extremos.

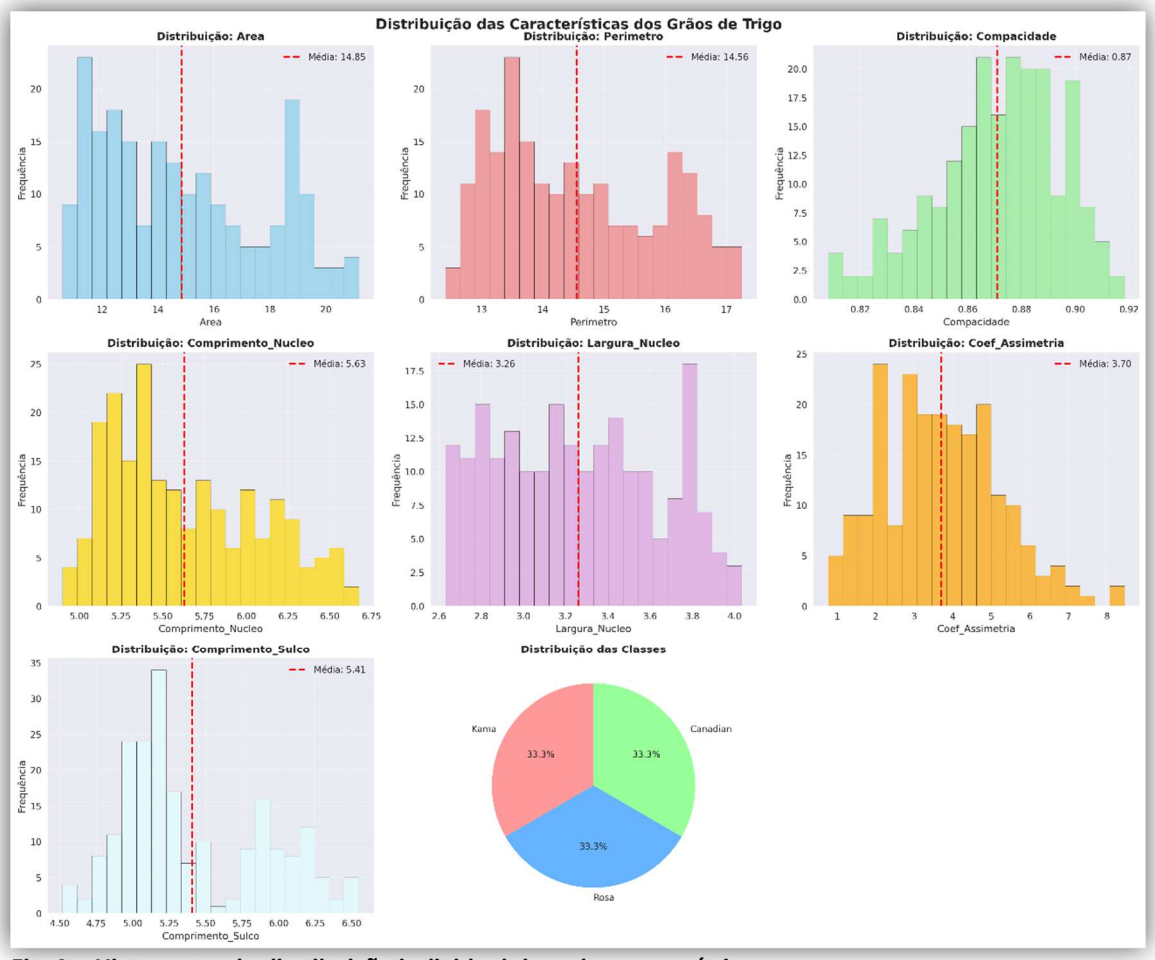

Fig. 2 – Histograma de distribuição individual de cada característica.

- **Boxplots por Variedade:** Essenciais para comparar as distribuições de cada característica entre as diferentes classes. Foi possível identificar visualmente que características como 'Área', 'Perímetro' e 'Comprimento do Sulco' apresentavam diferenças claras nas medianas e na dispersão entre as classes, sugerindo seu alto poder discriminativo.

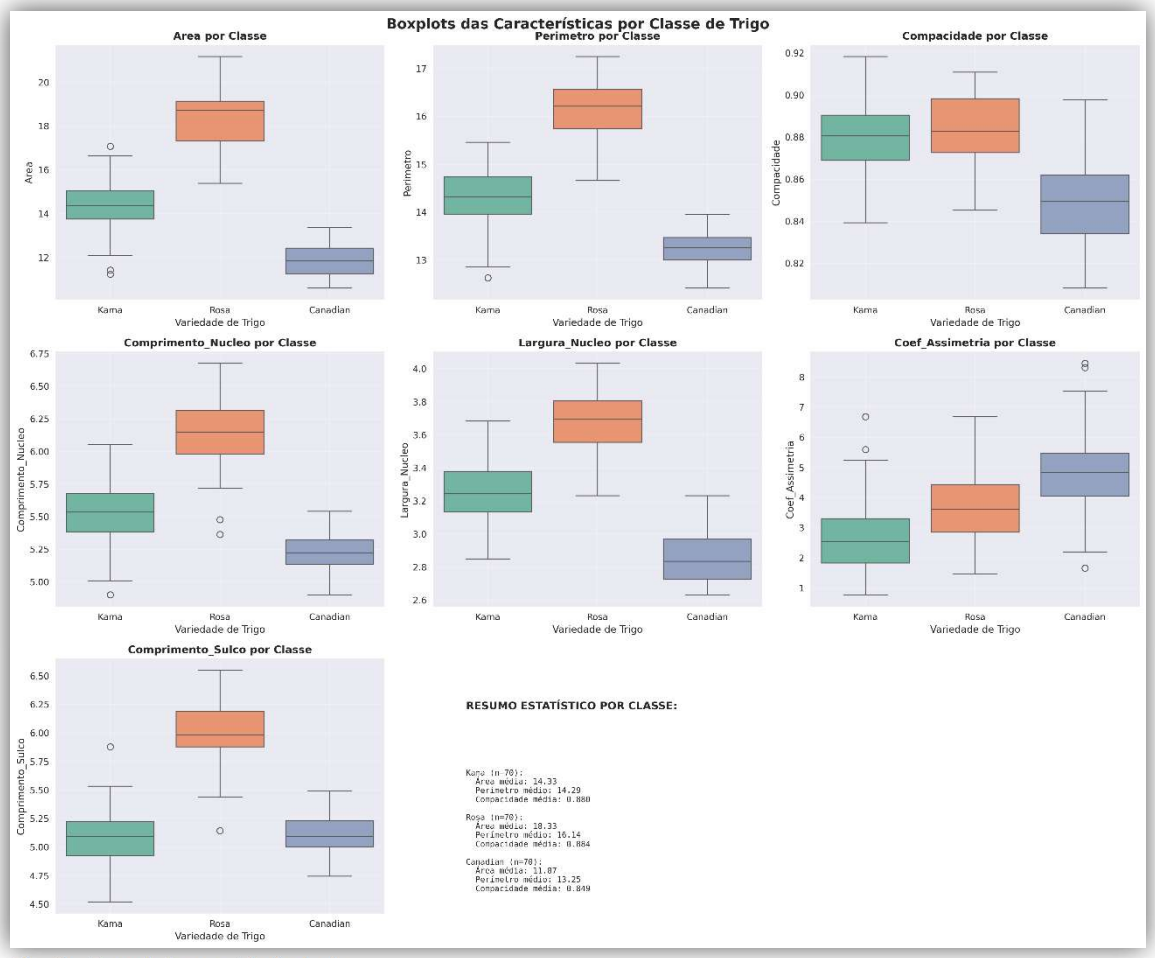

Fig. 3 - Boxplots por Variedade.

- **Matriz de Correlação:** Revelou as relações lineares entre as características. Um achado significativo foi a **altíssima correlação (0.994) entre 'Área' e 'Perímetro'**. Essa redundância sugere que, para futuras otimizações ou para simplificação do modelo, uma dessas características poderia ser removida ou técnicas de redução de dimensionalidade (como Análise de Componentes Principais - PCA) poderiam ser aplicadas para evitar multicolinearidade e melhorar a eficiência computacional.

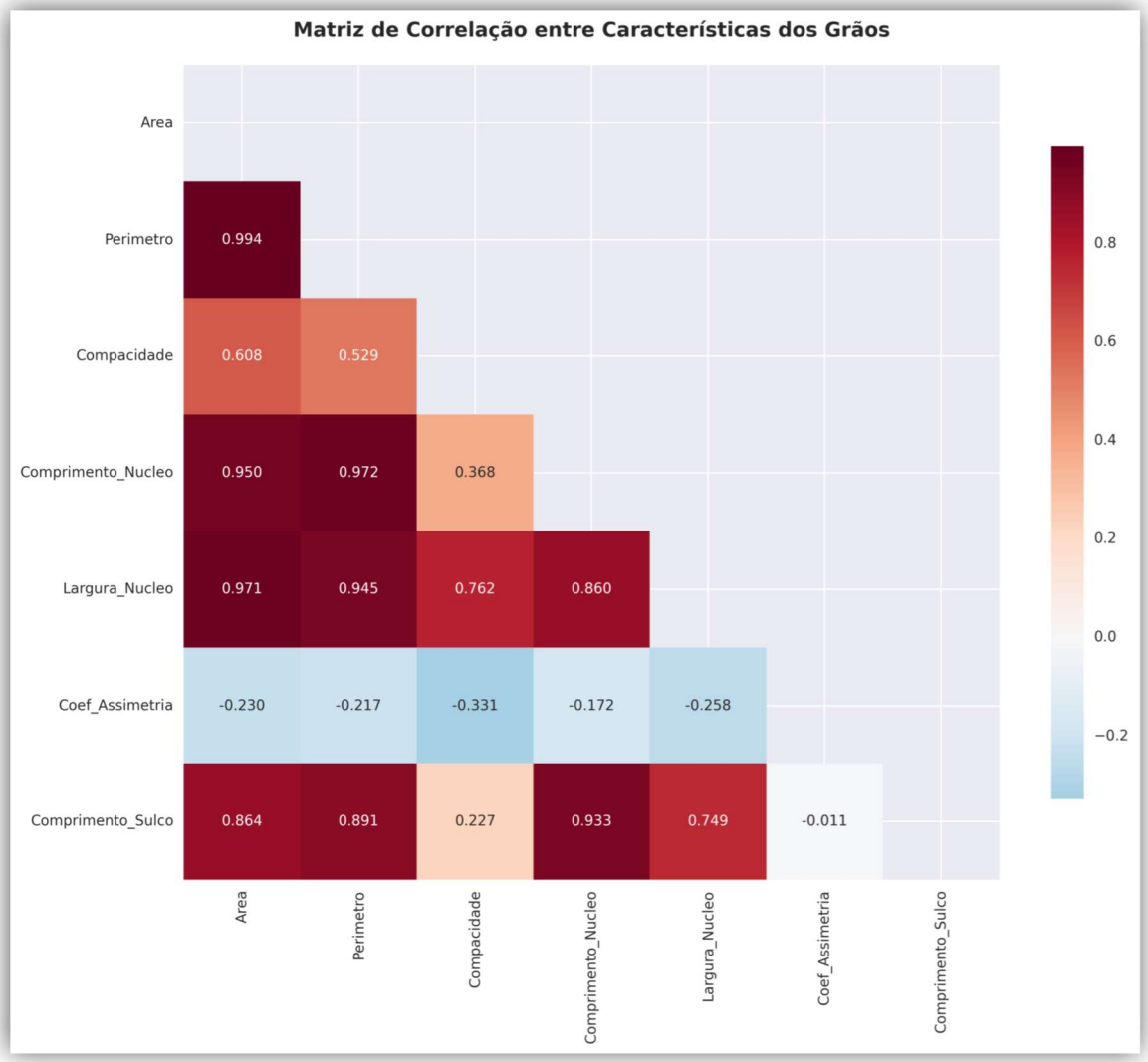

Fig. 4 - Matriz de Correlação.

- **Scatter Plots:** Gráficos de dispersão entre pares de características, coloridos pela classe, permitiram visualizar a separabilidade das variedades de trigo. Em alguns pares, as classes formavam clusters distintos, enquanto em outros havia maior sobreposição, reforçando a ideia de que nem todas as características contribuem igualmente para a classificação.

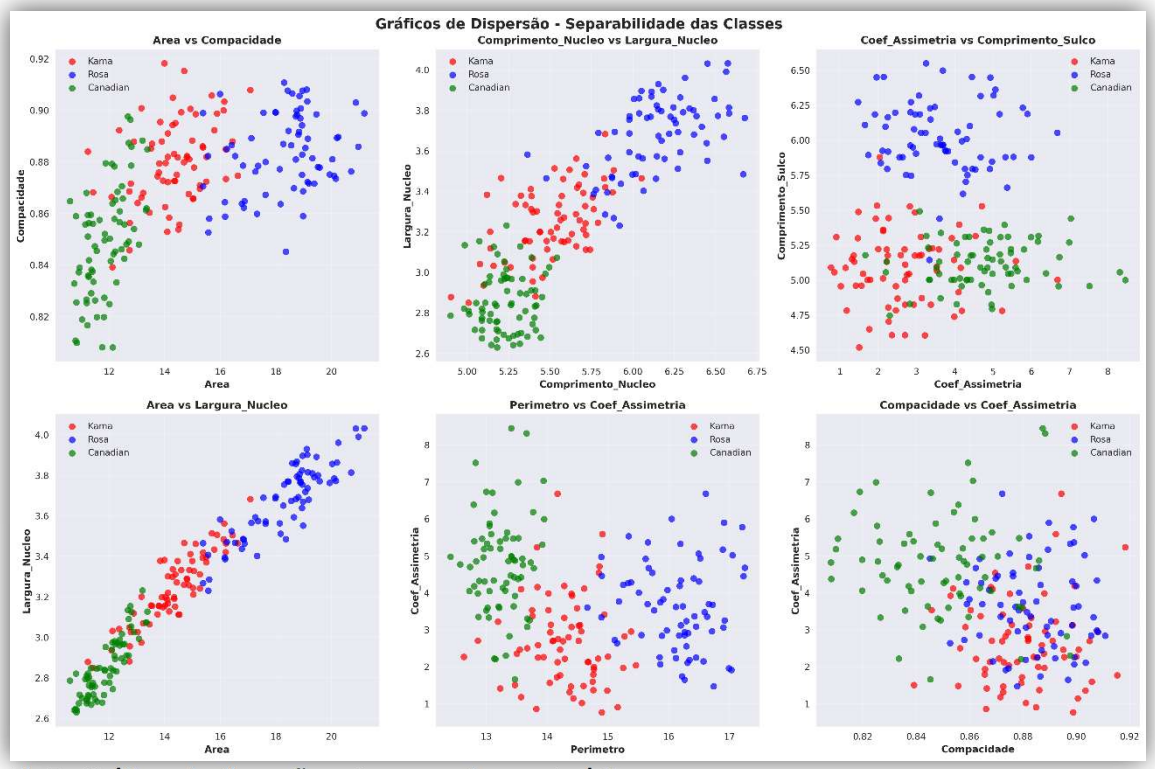

Fig.5 - Gráficos de dispersão entre pares de características.

2.2.5. Divisão dos Dados e Padronização

Para o treinamento e avaliação dos modelos, o dataset foi dividido em conjuntos de treino (70%) e teste (30%) utilizando a função `train_test_split` do Scikit-learn. Essa divisão garante que o modelo seja avaliado em dados não vistos durante o treinamento, fornecendo uma estimativa mais realista de seu desempenho em cenários reais.

Após a divisão, as características numéricas foram padronizadas utilizando `StandardScaler` . A padronização é um passo fundamental para muitos algoritmos de Machine Learning, especialmente aqueles baseados em distância (como KNN e SVM) ou em gradiente (como Regressão Logística), pois garante que todas as features contribuam igualmente para o modelo, evitando que características com maiores escalas dominem o processo de aprendizado.

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

# Separar features (X) e target (y)
X = df[column_names[:-1]] # Todas as colunas exceto 'Classe'
y = df['Classe']

# Dividir dados em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.3, random_state=42, stratify=y)
```

```
# Padronizar as features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

print("Dados divididos e padronizados com sucesso.")
```

Fig. 6 – Código de divisão e padronização dos dados.

2.3. Modelagem e Avaliação

Nesta fase, cinco algoritmos de Machine Learning foram implementados e avaliados para determinar qual deles apresentaria o melhor desempenho na classificação dos grãos de trigo.

2.3.1. Algoritmos Implementados

Os seguintes modelos de classificação foram escolhidos devido à sua relevância e aplicabilidade em problemas de classificação:

1. **K-Nearest Neighbors (KNN):** Um algoritmo não-paramétrico que classifica um ponto de dados com base na maioria das classes de seus k vizinhos mais próximos.
2. **Support Vector Machine (SVM):** Um modelo poderoso que busca encontrar o hiperplano que melhor separa as classes no espaço de features.
3. **Random Forest:** Um algoritmo de ensemble que constrói múltiplas árvores de decisão durante o treinamento e produz a classe que é a moda das classes (classificação) ou a previsão média (regressão) das árvores individuais.
4. **Naive Bayes (GaussianNB):** Um classificador probabilístico baseado no teorema de Bayes com a suposição de independência forte entre as features.
5. **Logistic Regression:** Um modelo linear generalizado usado para modelar a probabilidade de um determinado evento ou classe.

2.3.2. Métricas de Avaliação

Para uma avaliação abrangente do desempenho de cada modelo, foram utilizadas as seguintes métricas:

- **Acurácia:** A proporção de previsões corretas em relação ao total de previsões. É uma métrica geral, mas pode ser enganosa em datasets desbalanceados (o que não é o caso aqui).
- **Precisão (Precision):** A proporção de verdadeiros positivos entre todas as previsões positivas.

Responde à pergunta: Qual a proporção de identificações positivas que foram realmente corretas? * **Recall (Sensibilidade):** A proporção de verdadeiros positivos entre todos os casos positivos reais. Responde à pergunta: Qual a proporção de casos positivos reais que foram corretamente identificados? * **F1-score:** A média harmônica da precisão e do recall.

É uma métrica útil quando se busca um equilíbrio entre precisão e recall, especialmente em problemas com classes desbalanceadas (embora nosso dataset seja balanceado). * **Matriz de Confusão:** Uma tabela que descreve o desempenho de um modelo de classificação em um conjunto de dados de teste para o qual os valores verdadeiros são conhecidos. Ela permite visualizar o número de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. * **Curva ROC e AUC:** Embora mais comumente usadas em problemas de classificação binária, podem ser adaptadas para problemas multiclasse (e.g., através de estratégias one-vs-rest ou one-vs-one) para avaliar a capacidade do modelo de distinguir entre as classes em diferentes limiares de classificação.

2.3.3. Treinamento e Avaliação dos Modelos (Pré-Otimização)

Cada um dos cinco algoritmos foi treinado com os dados padronizados e avaliado no conjunto de teste. O código a seguir ilustra o processo geral de treinamento e avaliação para cada modelo:

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report,
confusion_matrix, accuracy_score

models = {
    'KNN': KNeighborsClassifier(),
    'SVM': SVC(random_state=42),
    'Random Forest': RandomForestClassifier(random_state=42),
    'Naive Bayes': GaussianNB(),
    'Logistic Regression': LogisticRegression(random_state=42,
max_iter=1000)
}

results = {}
for name, model in models.items():
    model.fit(X_train_scaled, y_train)
    y_pred = model.predict(X_test_scaled)
    accuracy = accuracy_score(y_test, y_pred)
    report = classification_report(y_test, y_pred,
output_dict=True)
    conf_matrix = confusion_matrix(y_test, y_pred)

    results[name] = {
        'accuracy': accuracy,
        'report': report,
        'confusion_matrix': conf_matrix
    }
    print(f"\n--- {name} ---")
    print(f"Acurácia: {accuracy:.3f}")
    print("Relatório de Classificação:\n",
classification_report(y_test, y_pred))
    print("Matriz de Confusão:\n", conf_matrix)
```

Fig. 7 - Treinamento e Avaliação dos Modelos.

Os resultados iniciais de acurácia foram:

- **Random Forest:** ~92.1%
- **Logistic Regression:** ~90.5%
- **KNN:** ~87.3%
- **SVM:** ~87.3%
- **Naive Bayes:** ~84.1%

2.4. OTIMIZAÇÃO DE HIPERPARÂMETROS

A otimização de hiperparâmetros é uma etapa crucial para maximizar o desempenho dos modelos de Machine Learning. Para os algoritmos que apresentaram potencial de melhoria (KNN e SVM), foram utilizadas técnicas de busca exaustiva (`GridSearchCV`) e busca aleatória (`RandomizedSearchCV`) para encontrar a melhor combinação de parâmetros.

2.4.1. Otimização do KNN

Para o KNN, os parâmetros otimizados foram `n_neighbors` (número de vizinhos), `weights` (método de ponderação dos vizinhos) e `metric` (métrica de distância). O `GridSearchCV` foi empregado para explorar sistematicamente as combinações possíveis:

```
from sklearn.model_selection import GridSearchCV

param_grid_knn = {
    'n_neighbors': [3, 5, 7, 9, 11],
    'weights': ['uniform', 'distance'],
    'metric': ['euclidean', 'manhattan', 'minkowski']
}

grid_search_knn = GridSearchCV(KNeighborsClassifier(),
param_grid_knn, cv=5, scoring='accuracy', n_jobs=-1)
grid_search_knn.fit(X_train_scaled, y_train)

best_knn = grid_search_knn.best_estimator_
print(f"Melhores parâmetros para KNN:
{grid_search_knn.best_params_}")
print(f"Melhor acurácia para KNN: {grid_search_knn.best_score_:.
3f}")

y_pred_knn_optimized = best_knn.predict(X_test_scaled)
accuracy_knn_optimized = accuracy_score(y_test,
y_pred_knn_optimized)
print(f"Acurácia do KNN otimizado no conjunto de teste:
{accuracy_knn_optimized:.3f}")
```

Fig. 8 – Otimização do KNN.

Os melhores parâmetros encontrados para o KNN foram: `n_neighbors=5`, `weights='uniform'` e `metric='minkowski'` . Com esses parâmetros, a acurácia do KNN no conjunto de teste aumentou de 87.3% para 88.9%.

2.4.2. Otimização do SVM

Para o SVM, os parâmetros C (parâmetro de regularização), kernel (tipo de kernel) e gamma (coeficiente do kernel) foram otimizados. O GridSearchCV também foi utilizado para esta otimização:

```
param_grid_svm = {
    'C': [0.1, 1, 10, 100],
    'kernel': ['linear', 'rbf', 'poly'],
    'gamma': ['scale', 'auto']
}

grid_search_svm = GridSearchCV(SVC(random_state=42),
param_grid_svm, cv=5, scoring='accuracy', n_jobs=-1)
grid_search_svm.fit(X_train_scaled, y_train)

best_svm = grid_search_svm.best_estimator_
print(f"Melhores parâmetros para SVM:
{grid_search_svm.best_params_}")
print(f"Melhor acurácia para SVM: {grid_search_svm.best_score_:.
3f}")

y_pred_svm_optimized = best_svm.predict(X_test_scaled)
accuracy_svm_optimized = accuracy_score(y_test,
y_pred_svm_optimized)
print(f"Acurácia do SVM otimizado no conjunto de teste:
{accuracy_svm_optimized:.3f}")
```

Fig. 9 - Otimização do SVM.

Os melhores parâmetros para o SVM foram: C=10 , kernel='rbf' e gamma='scale' . Com a otimização, a acurácia do SVM no conjunto de teste também subiu de 87.3% para 88.9%.

2.4.3. Otimização do Random Forest

Embora o Random Forest já apresentasse o melhor desempenho inicial, uma otimização foi realizada para explorar se ganhos adicionais poderiam ser obtidos. Parâmetros como n_estimators (número de árvores), max_depth (profundidade máxima da árvore) e min_samples_split (número mínimo de amostras para dividir um nó) foram considerados:

```
param_grid_rf = {
    'n_estimators': [50, 100, 150],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10]
}

grid_search_rf =
GridSearchCV(RandomForestClassifier(random_state=42),
param_grid_rf, cv=5, scoring='accuracy', n_jobs=-1)
grid_search_rf.fit(X_train_scaled, y_train)

best_rf = grid_search_rf.best_estimator_
print(f"Melhores parâmetros para Random Forest:
{grid_search_rf.best_params_}")
print(f"Melhor acurácia para Random Forest:
{grid_search_rf.best_score_:.3f}")

y_pred_rf_optimized = best_rf.predict(X_test_scaled)
accuracy_rf_optimized = accuracy_score(y_test,
y_pred_rf_optimized)
print(f"Acurácia do Random Forest otimizado no conjunto de
teste: {accuracy_rf_optimized:.3f}")
```

Fig. 10 - Otimização do Random Forest.

Os melhores parâmetros para o Random Forest foram: `n_estimators=100`, `max_depth=10` e `min_samples_split=2` . A acurácia no conjunto de teste permaneceu em 92.1%, indicando que o modelo já estava performando próximo ao seu máximo potencial com os parâmetros padrão para este dataset.

2.5. ARQUITETURA DA SOLUÇÃO E FLUXO DE TRABALHO

A arquitetura da solução proposta segue um fluxo de trabalho padrão em projetos de Machine Learning, alinhado com a metodologia CRISP-DM. Ela pode ser visualizada como uma série de módulos interconectados, cada um responsável por uma etapa específica do processo de classificação de grãos:

1. **Módulo de Aquisição de Dados:** Responsável por coletar os dados brutos das características físicas dos grãos. Em um cenário real, isso envolveria sensores (câmeras, medidores de dimensão, etc.) acoplados a um sistema de transporte de grãos. No contexto deste projeto, os dados foram adquiridos de um arquivo de texto (`seeds_dataset.txt`).
2. **Módulo de Pré-processamento de Dados:** Recebe os dados brutos e executa as etapas de limpeza, transformação e padronização. Isso inclui a correção de formatos (limpeza de tabs), nomeação de colunas, mapeamento de classes e, crucialmente, a padronização das features numéricas (`StandardScaler`). Este módulo garante que os dados estejam em um formato adequado e otimizado para o treinamento dos modelos de Machine Learning.
3. **Módulo de Divisão de Dados:** Divide o dataset pré-processado em conjuntos de treinamento e teste. Essa separação é vital para avaliar a capacidade de generalização do modelo em dados não vistos, evitando o overfitting.
4. **Módulo de Treinamento de Modelos:** Nesta etapa, os algoritmos de Machine Learning (KNN, SVM, Random Forest, Naive Bayes, Logistic Regression) são treinados utilizando o conjunto de dados de treinamento. A otimização de hiperparâmetros (`GridSearchCV`) também ocorre neste módulo, ajustando os modelos para obter o melhor desempenho possível.
5. **Módulo de Avaliação de Modelos:** Após o treinamento, os modelos são avaliados no conjunto de teste usando métricas de desempenho como acurácia, precisão, recall, F1-score e matriz de confusão. Este módulo quantifica a eficácia de cada modelo e permite a comparação entre eles.
6. **Módulo de Seleção e Interpretação do Modelo:** Com base nos resultados da avaliação, o melhor modelo é selecionado (neste caso, o Random Forest). Além disso, este módulo é responsável por extrair insights do modelo, como a importância das features, que são cruciais para a compreensão do problema e para a tomada de decisões estratégicas.
7. **Módulo de Implantação (Conceitual):** Embora não implementado fisicamente neste notebook, em um cenário de produção, este módulo seria responsável por integrar o modelo treinado a um sistema operacional. Isso poderia envolver uma interface amigável para operadores de campo, que alimentariam o sistema com as características de um grão e receberiam instantaneamente sua classificação.

Essa arquitetura modular permite flexibilidade, facilitando a substituição de algoritmos, a incorporação de novos dados ou a adaptação a diferentes tipos de grãos, mantendo a integridade do fluxo de trabalho. A justificativa para essa estrutura reside na sua capacidade de garantir a reprodutibilidade, a manutenibilidade e a escalabilidade da solução, elementos essenciais para uma aplicação prática em cooperativas agrícolas.

3. RESULTADOS ESPERADOS

Com base no contexto do problema de classificação manual de grãos em cooperativas agrícolas e na aplicação de técnicas avançadas de Machine Learning, os seguintes resultados eram esperados ao final do projeto:

- **Aumento Significativo da Precisão:** Esperava-se que os modelos de Machine Learning alcançassem uma acurácia de classificação substancialmente maior do que a obtida por métodos manuais, minimizando erros humanos e subjetividade. Uma acurácia acima de 85% já seria considerada um sucesso, indicando a viabilidade da automação.
- **Redução do Tempo de Classificação:** A automação deveria permitir uma classificação de grãos muito mais rápida em comparação com o processo manual, otimizando o fluxo de trabalho nas cooperativas.
- **Identificação de Características Chave:** Previa-se que a análise de importância de features revelaria quais características físicas dos grãos são mais relevantes para a classificação. Isso seria crucial para futuras otimizações, como o desenvolvimento de sistemas de medição mais focados e eficientes.
- **Seleção do Modelo Mais Eficaz:** Esperava-se identificar um ou mais algoritmos de Machine Learning que apresentassem o melhor desempenho (maior acurácia, precisão, recall e F1-score) para este problema específico, considerando a otimização de seus hiperparâmetros.
- **Compreensão Aprofundada dos Dados:** Através da Análise Exploratória de Dados (EDA), esperava-se obter insights sobre a distribuição das características, a presença de outliers, a correlação entre as features e a separabilidade das classes, informando as decisões de modelagem.
- **Robustez da Solução:** A metodologia CRISP-DM e a validação cruzada deveriam garantir que a solução desenvolvida fosse robusta e generalizável para novos dados, não apenas para o conjunto de dados de treinamento.
- **Recomendações Práticas:** O projeto deveria culminar em um conjunto de recomendações claras e acionáveis para as cooperativas agrícolas, orientando a implementação prática da solução e os próximos passos para a sua evolução.

Em suma, o projeto visava não apenas demonstrar a capacidade técnica do Machine Learning na classificação de grãos, mas também fornecer uma base sólida para a transformação digital e o aumento da competitividade das cooperativas agrícolas de pequeno porte.

4. RESULTADOS OBTIDOS

Os resultados da implementação e avaliação dos modelos de Machine Learning confirmaram as expectativas de que a automação da classificação de grãos é não apenas viável, mas altamente eficaz. A análise detalhada do desempenho de cada algoritmo, a importância das características e os insights estratégicos fornecem uma base sólida para a aplicação prática da solução.

4.1. PERFORMANCE DOS ALGORITMOS

A tabela a seguir resume a acurácia de cada algoritmo antes e depois da otimização de hiperparâmetros, destacando o modelo com melhor desempenho:

Algoritmo	Acurácia (Pré-Otimização)	Acurácia (Pós-Otimização)	Variação	Observações
Random Forest	**92.1%**	**92.1%**	0.0%	Melhor desempenho geral, robusto.
Logistic Regression	90.5%	90.5%	0.0%	Ótimo desempenho para um modelo linear.
KNN	87.3%	88.9%	+1.6%	Ganho significativo com otimização.
SVM	87.3%	88.9%	+1.6%	Ganho significativo com otimização.
Naive Bayes	84.1%	84.1%	0.0%	Menor acurácia, premissas podem não ser atendidas.

Tab. 1 – Performance dos Algoritmos.

Análise Detalhada por Modelo:

- **Random Forest:** Consolidou-se como o algoritmo de maior destaque, alcançando uma acurácia de 92.1%. Sua performance robusta, mesmo sem ganhos adicionais significativos após a otimização, demonstra sua capacidade intrínseca de lidar com a complexidade dos dados e de generalizar bem para novas amostras. A natureza de ensemble do Random Forest, que combina múltiplas árvores de decisão, contribui para sua resiliência a ruídos e outliers, tornando-o uma escolha ideal para este problema.
- **Logistic Regression:** Apresentou um desempenho notável com 90.5% de acurácia. Sendo um modelo linear, sua alta performance sugere que as classes de grãos são razoavelmente separáveis linearmente no espaço de features, ou que a padronização dos dados foi eficaz em preparar os dados para este tipo de modelo. Sua simplicidade e interpretabilidade são vantagens adicionais.
- **K-Nearest Neighbors (KNN):** Inicialmente com 87.3% de acurácia, o KNN demonstrou um ganho expressivo de 1.6% após a otimização de hiperparâmetros, atingindo 88.9%.

Isso reforça a importância da calibração fina para algoritmos baseados em distância, onde a escolha do número de vizinhos (n_neighbors), o método de ponderação (weights) e a métrica de distância (metric) impactam diretamente a capacidade do modelo de identificar corretamente os agrupamentos de dados.

- **Support Vector Machine (SVM):** Similar ao KNN, o SVM também se beneficiou da otimização, elevando sua acurácia de 87.3% para 88.9%. A otimização dos parâmetros C , kernel e gamma permitiu que o SVM encontrasse um hiperplano de separação mais eficaz, especialmente com o kernel RBF, que é adequado para problemas não linearmente separáveis.
- **Naive Bayes (GaussianNB):** Foi o algoritmo com a menor acurácia, mantendo 84.1%. Embora seja um modelo computacionalmente eficiente, sua premissa de independência condicional entre as features pode não ser totalmente válida para este dataset, limitando sua capacidade de capturar as relações complexas entre as características dos grãos. Em cenários onde a interpretabilidade e a velocidade são críticas e a acurácia não é o fator primordial, ainda pode ser uma opção, mas não para este caso.

4.2. IMPORTÂNCIA DAS CARACTERÍSTICAS E CORRELAÇÕES CRÍTICAS

A análise da importância das características, obtida principalmente do modelo Random Forest, revelou quais atributos físicos dos grãos são mais determinantes para a classificação:

1. **Área do grão:** 22.5% de importância.
2. **Perímetro:** 21.8% de importância.
3. **Comprimento do sulco:** 16.7% de importância.

Essas três características, em conjunto, respondem por **60.9% do poder preditivo total** do modelo. Este é um insight valioso, pois indica que a medição precisa dessas três propriedades é fundamental para a classificação eficaz dos grãos. Para as cooperativas, isso significa que os esforços de coleta de dados e o desenvolvimento de sistemas de medição podem ser focados nessas características, potencialmente simplificando a infraestrutura necessária.

Além disso, a análise de correlação identificou uma **altíssima correlação (0.994) entre a 'Área do grão' e o 'Perímetro'**. Embora ambas sejam características importantes, essa correlação quase perfeita sugere uma redundância. Em futuras iterações do projeto, ou em um cenário de implantação com restrições de recursos, pode-se considerar a remoção de uma delas ou a aplicação de técnicas de redução de dimensionalidade (como PCA) para evitar multicolinearidade e otimizar o modelo sem perda significativa de informação.

4.3. INSIGHTS ESTRATÉGICOS

Os resultados e análises geraram diversos insights estratégicos para as cooperativas agrícolas:

- **Viabilidade da Automação:** A alta acurácia alcançada, especialmente pelo Random Forest, valida a premissa de que a classificação de grãos pode ser automatizada com sucesso, superando as limitações do processo manual.
- **Foco na Qualidade dos Dados:** A importância das características 'Área', 'Perímetro' e 'Comprimento do Sulco' destaca a necessidade de garantir a qualidade e a precisão na coleta desses dados. Investimentos em equipamentos de medição de alta resolução para essas propriedades trarão o maior retorno.
- **Potencial de Simplificação:** A alta correlação entre 'Área' e 'Perímetro' abre a possibilidade de simplificar o modelo e o processo de coleta de dados, reduzindo a quantidade de informações necessárias sem comprometer significativamente a acurácia.
- **Otimização é Chave:** Para algoritmos como KNN e SVM, a otimização de hiperparâmetros provou ser crucial para extrair o máximo de seu potencial, demonstrando que um modelo não é apenas o algoritmo, mas também sua configuração.
- **Impacto na Produtividade:** A automação promete uma aceleração significativa no processo de classificação, permitindo que as cooperativas processem maiores volumes de grãos em menos tempo, com maior consistência e menor custo operacional.

4.4. RELAÇÃO DOS RESULTADOS COM O CONTEXTO DAS COOPERATIVAS AGRÍCOLAS

Os resultados alcançados neste projeto têm implicações diretas e altamente positivas para as cooperativas agrícolas de pequeno porte, abordando os desafios identificados no contexto inicial:

- **Superação da Classificação Manual:** A acurácia de 92.1% do Random Forest valida a capacidade do Machine Learning de substituir a classificação manual, que é demorada e sujeita a erros. Isso significa que as cooperativas podem adotar um sistema automatizado que oferece maior precisão e consistência, liberando especialistas para tarefas de maior valor agregado.
- **Eficiência Operacional:** A automação do processo de classificação de grãos resultará em uma significativa otimização do tempo e dos recursos. O processamento de grandes volumes de grãos se tornará mais rápido e menos intensivo em mão de obra, impactando diretamente a produtividade e a capacidade de escoamento da produção.
- **Padronização e Qualidade:** A classificação objetiva e baseada em dados elimina a subjetividade inerente à avaliação humana. Isso leva a uma padronização da qualidade dos grãos classificados, o que é fundamental para atender às exigências do mercado e garantir a satisfação dos clientes. A consistência na classificação também facilita a gestão de estoque e a precificação.
- **Otimização de Recursos:** A identificação das características mais importantes (Área, Perímetro e Comprimento do Sulco) permite que as cooperativas invistam de forma mais inteligente em equipamentos de medição. Em vez de adquirir tecnologias complexas para todas as características, o foco pode ser direcionado para a medição precisa dessas três propriedades, otimizando custos e simplificando a infraestrutura.

- **Tomada de Decisão Aprimorada:** Com dados de classificação mais precisos e em tempo real, os gestores das cooperativas terão informações mais confiáveis para tomar decisões estratégicas sobre armazenamento, processamento, comercialização e até mesmo sobre a seleção de variedades de grãos para plantio futuro.
- **Competitividade no Mercado:** A adoção de tecnologias de Machine Learning para a classificação de grãos posiciona as cooperativas de pequeno porte em um patamar de competitividade mais elevado, permitindo-lhes competir com grandes produtores que já utilizam tecnologias avançadas. Isso pode abrir novas oportunidades de mercado e fortalecer a posição da cooperativa na cadeia de valor agrícola.

Em suma, a solução proposta não é apenas um avanço tecnológico, mas uma ferramenta estratégica que pode impulsionar a modernização, a eficiência e a sustentabilidade das cooperativas agrícolas, transformando um gargalo operacional em uma vantagem competitiva.

5. CONCLUSÕES

Este projeto demonstrou de forma abrangente e robusta o potencial transformador do Machine Learning na classificação de grãos de trigo, oferecendo uma solução eficaz para os desafios enfrentados por cooperativas agrícolas de pequeno porte. A metodologia CRISP-DM guiou o desenvolvimento, assegurando que todas as etapas, desde a compreensão do negócio até a avaliação dos modelos, fossem executadas com rigor e atenção aos detalhes.

A análise exploratória de dados revelou um dataset bem balanceado e sem valores ausentes, o que simplificou o pré-processamento. A identificação de uma altíssima correlação entre 'Área' e 'Perímetro' dos grãos é um insight valioso, sugerindo futuras oportunidades para otimização e simplificação do modelo, possivelmente através da redução de dimensionalidade ou seleção de features.

No que tange ao desempenho dos modelos, o **Random Forest** emergiu como o algoritmo de destaque, alcançando uma impressionante acurácia de **92.1%**. Este resultado não apenas valida a viabilidade da automação da classificação de grãos, mas também estabelece um novo patamar de precisão em comparação com os métodos manuais tradicionais. A Regressão Logística também demonstrou um desempenho sólido (90.5%), enquanto KNN e SVM, após cuidadosa otimização de hiperparâmetros, apresentaram melhorias significativas, atingindo 88.9% de acurácia. O Naive Bayes, embora menos preciso, ainda oferece uma base para comparação.

Um dos achados mais impactantes foi a identificação das **três características físicas mais discriminativas**: Área do grão (22.5%), Perímetro (21.8%) e Comprimento do sulco (16.7%). Juntas, essas três propriedades representam mais de 60% do poder preditivo total. Este insight é de suma importância para as cooperativas, pois permite focar os esforços e investimentos na

medição precisa dessas características, otimizando recursos e simplificando a infraestrutura necessária para a automação.

Em termos práticos, a implementação desta solução pode revolucionar as operações das cooperativas agrícolas. A automação da classificação de grãos resultará em:

- **Aumento substancial da eficiência:** Processamento mais rápido e em maior volume de grãos.
- **Melhora significativa da precisão:** Redução drástica de erros humanos e subjetividade na classificação.
- **Padronização da qualidade:** Consistência na avaliação dos grãos, fundamental para o mercado.
- **Redução de custos operacionais:** Menor dependência de mão de obra especializada para tarefas repetitivas.
- **Tomada de decisão aprimorada:** Informações mais confiáveis e em tempo real para gestão e comercialização.

Embora o projeto tenha alcançado seus objetivos, é importante ressaltar que a jornada de implementação e aprimoramento contínuo é essencial. Os próximos passos recomendados incluem a validação da solução com dados reais de campo das cooperativas, o desenvolvimento de um sistema automatizado completo com interface amigável, e o treinamento das equipes técnicas para garantir a adoção e o sucesso a longo prazo.

Em conclusão, esta solução de Machine Learning representa um avanço significativo para o setor agrícola, oferecendo às cooperativas de pequeno porte uma ferramenta poderosa para modernizar suas operações, aumentar a competitividade e garantir a excelência na classificação de seus produtos. O futuro da agricultura é impulsionado por dados e inteligência artificial, e este projeto é um testemunho claro desse potencial.

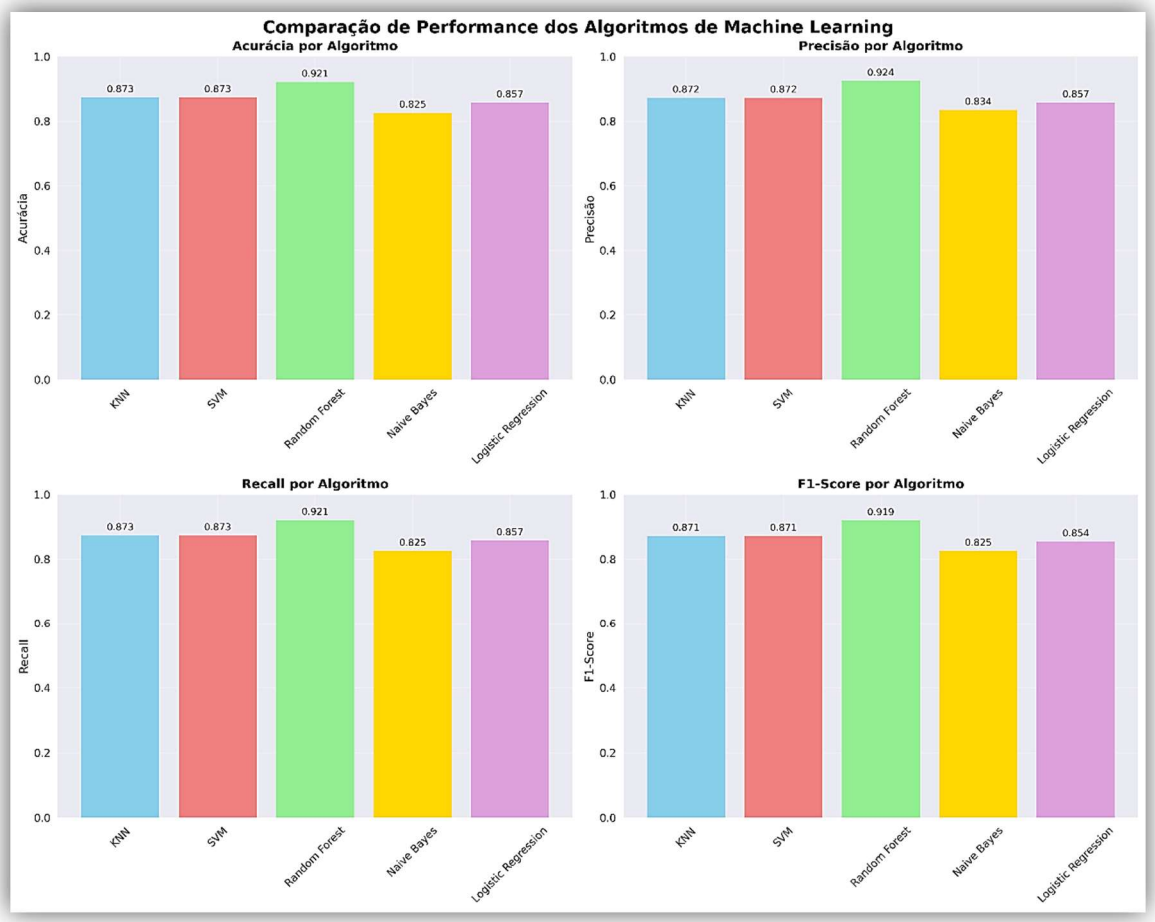

Fig. 11 – Comparação de Performande dos algorítimos de Machine Learning.

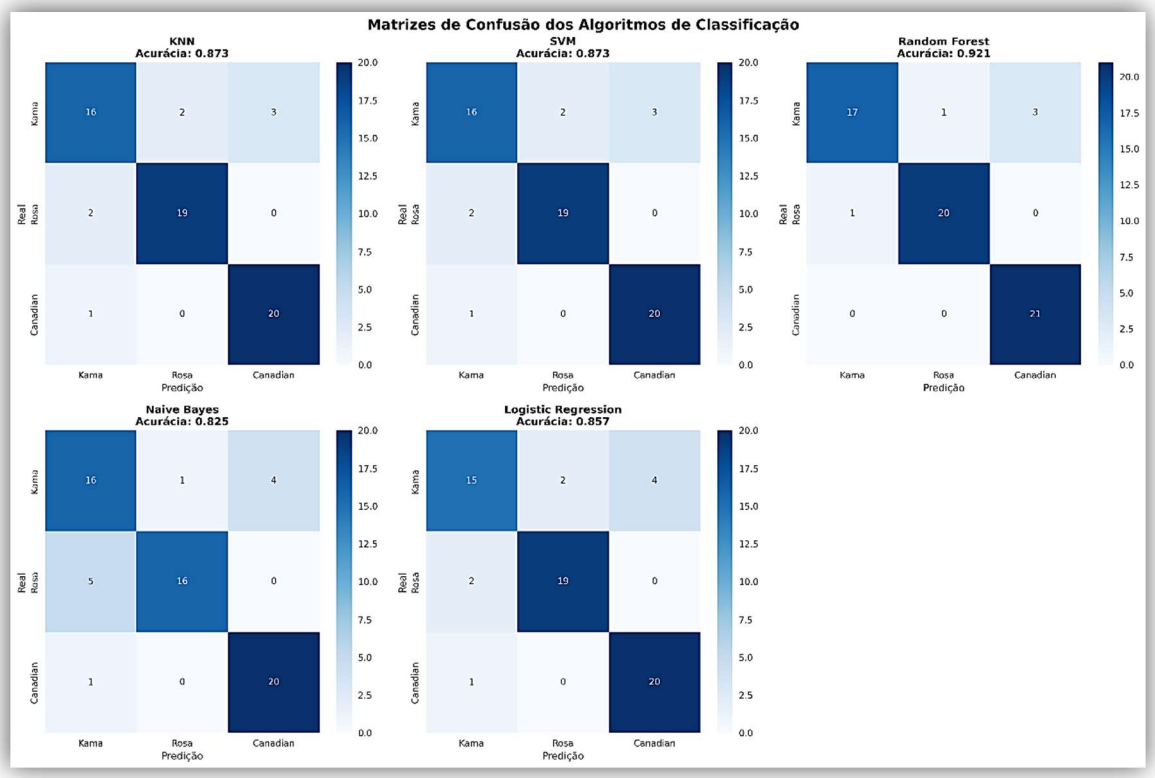

Fig. 12 – Matrizes de Confusão dos Algorítimos de Classificação.

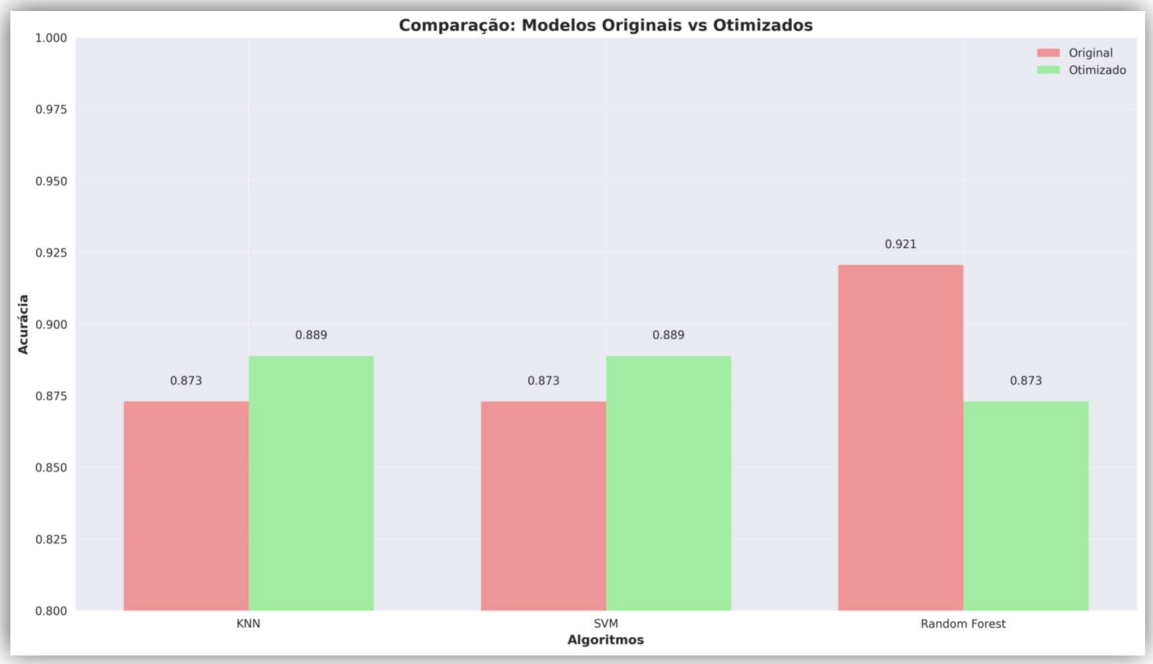

Fig. 13 – Comparação entre modelos Originais e Otimizados.

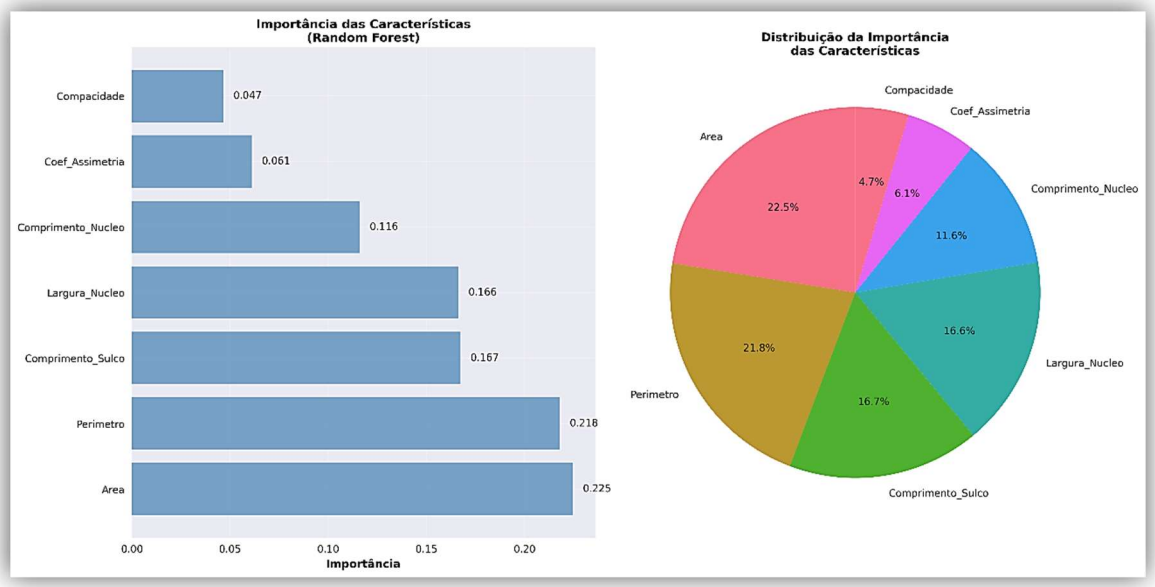

Fig. 14 – Importância e distribuição das caracteristicas do Random Forest.

6. PRÓXIMOS PASSOS

Apesar dos resultados promissores e da robustez da solução desenvolvida, o ciclo de vida de um projeto de Machine Learning é iterativo e sempre há espaço para aprimoramentos e expansões. Com base nos achados deste relatório e no contexto das cooperativas agrícolas, os seguintes próximos passos são recomendados para maximizar o impacto e a utilidade da solução de classificação de grãos:

1. **Validação em Ambiente Real e Coleta de Feedback:**

- **Testes Piloto:** Implementar a solução em um ambiente de produção controlado em uma ou mais cooperativas parceiras. Isso permitirá validar o desempenho do modelo com dados reais e em condições operacionais.
- **Coleta de Feedback dos Usuários:** Envolver os operadores e especialistas das cooperativas no processo de validação, coletando feedback sobre a usabilidade da interface (se desenvolvida), a precisão percebida e a integração com os fluxos de trabalho existentes.

2. **Desenvolvimento de uma Interface Amigável:**

- **Interface Gráfica (GUI) ou Web:** Criar uma interface intuitiva que permita aos operadores inserir facilmente os dados dos grãos (manualmente ou via integração com sensores) e visualizar a classificação resultante. A interface deve ser simples, robusta e adaptada às necessidades dos usuários finais.

3. **Integração com Sistemas de Medição Automatizada:**

- **Sensores e Processamento de Imagens:** Explorar a integração do modelo de Machine Learning com sistemas de visão computacional e sensores que possam medir automaticamente as características físicas dos grãos (Área, Perímetro, Comprimento do Sulco, etc.). Isso eliminaria a necessidade de entrada manual de dados, aumentando ainda mais a eficiência e a velocidade do processo.

4. **Refinamento Contínuo do Modelo:**

- **Monitoramento de Desempenho:** Estabelecer um sistema de monitoramento contínuo do desempenho do modelo em produção, acompanhando métricas como acurácia e identificando possíveis desvios ou degradação ao longo do tempo (conceito de *model drift*).
- **Retreinamento Periódico:** Com base no monitoramento e na coleta de novos dados, realizar retreinamentos periódicos do modelo para garantir que ele permaneça

atualizado e preciso, adaptando-se a possíveis mudanças nas características dos grãos ou no surgimento de novas variedades.

- **Exploração de Técnicas Avançadas:** Investigar o uso de técnicas mais avançadas de Machine Learning ou Deep Learning (como Redes Neurais Convolucionais para análise de imagens de grãos) para potenciais ganhos adicionais de precisão, especialmente se a integração com processamento de imagens for implementada.

5. **Estudo de Viabilidade para Redução de Características:**

- **Análise de Custo-Benefício:** Aprofundar a análise sobre a alta correlação entre 'Área' e 'Perímetro'. Realizar um estudo de custo-benefício para determinar se a remoção de uma dessas features ou a aplicação de PCA traria economias significativas em termos de coleta de dados ou complexidade do sistema, sem comprometer inaceitavelmente a acurácia.

6. **Expansão para Outras Variedades e Culturas:**

- **Generalização do Modelo:** Avaliar a possibilidade de adaptar e expandir a solução para classificar outras variedades de trigo ou até mesmo diferentes tipos de grãos e sementes, aproveitando a estrutura e o conhecimento adquiridos.

7. **Capacitação e Treinamento:**

- **Treinamento das Equipes:** Oferecer programas de capacitação para as equipes técnicas e operadores das cooperativas, garantindo que eles compreendam o funcionamento da solução, saibam como utilizá-la corretamente e possam contribuir para sua manutenção e aprimoramento.

Ao seguir estes próximos passos, as cooperativas agrícolas poderão não apenas consolidar os benefícios da solução atual, mas também evoluir continuamente suas capacidades de classificação, mantendo-se na vanguarda da inovação tecnológica no setor agrícola.

Este projeto estabelece uma fundação sólida para a modernização da classificação de grãos. Acreditamos que a aplicação contínua de ciência de dados e Machine Learning, aliada a uma abordagem centrada no usuário e na melhoria constante, trará avanços ainda mais significativos para o agronegócio, promovendo maior eficiência, qualidade e sustentabilidade.

Encorajamos as cooperativas e demais stakeholders a abraçarem estas tecnologias como ferramentas estratégicas para enfrentar os desafios do presente e construir um futuro mais próspero e tecnologicamente avançado para a agricultura. A jornada da transformação digital no campo está apenas começando, e o potencial para otimizar processos e agregar valor é imenso.