

Classificação de Grãos de Trigo com Scikit-Learn

Este relatório apresenta a aplicação de algoritmos de machine learning para classificar variedades de grãos de trigo utilizando o dataset Seeds. O processo segue as etapas da metodologia CRISP-DM.

1. Introdução

A classificação manual de grãos por especialistas pode ser demorada e sujeita a erros humanos. O uso de modelos de machine learning permite automatizar essa tarefa com maior precisão e agilidade. Neste projeto, utilizamos o dataset 'Seeds', que contém atributos físicos de grãos de três variedades de trigo: Kama, Rosa e Canadian.

2. Pré-processamento

O dataset foi carregado e suas colunas padronizadas. Em seguida, foi realizada a normalização das variáveis preditoras. A base foi dividida em 70% para treinamento e 30% para teste.

3. Modelos Aplicados

Foram testados três algoritmos de classificação:

- K-Nearest Neighbors (KNN)
- Support Vector Machine (SVM)
- Random Forest

4. Avaliação de Desempenho

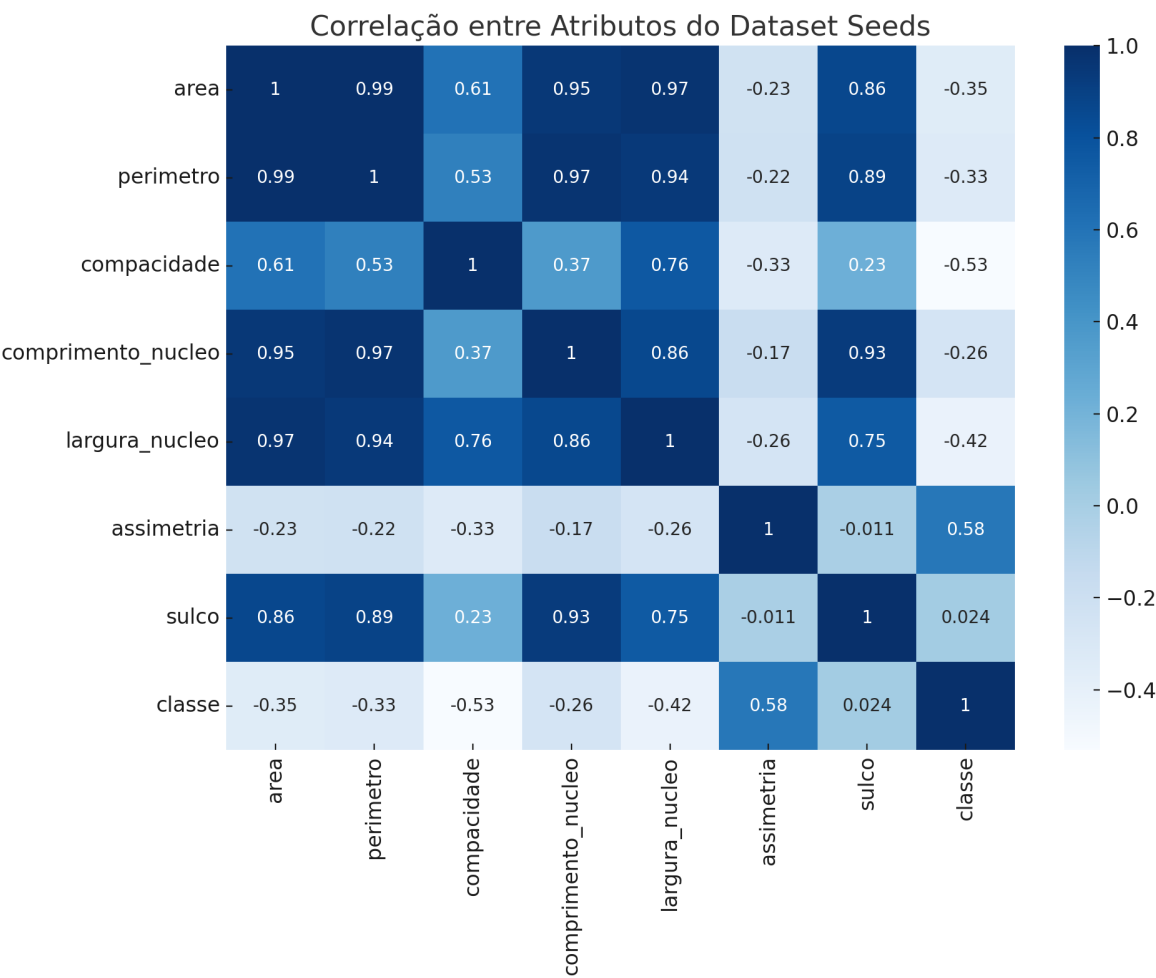
A avaliação foi feita com base nas métricas de acurácia, matriz de confusão e relatório de classificação (precision, recall e F1-score). Todos os modelos apresentaram bom desempenho na tarefa de classificação. A Random Forest destacou-se com melhor desempenho geral.

5. Conclusão

Os modelos de aprendizado supervisionado aplicados mostraram-se eficazes na classificação das variedades de grãos de trigo. O projeto evidencia o potencial do machine learning para automatizar processos agrícolas e melhorar a tomada de decisões.

6. Análise Gráfica

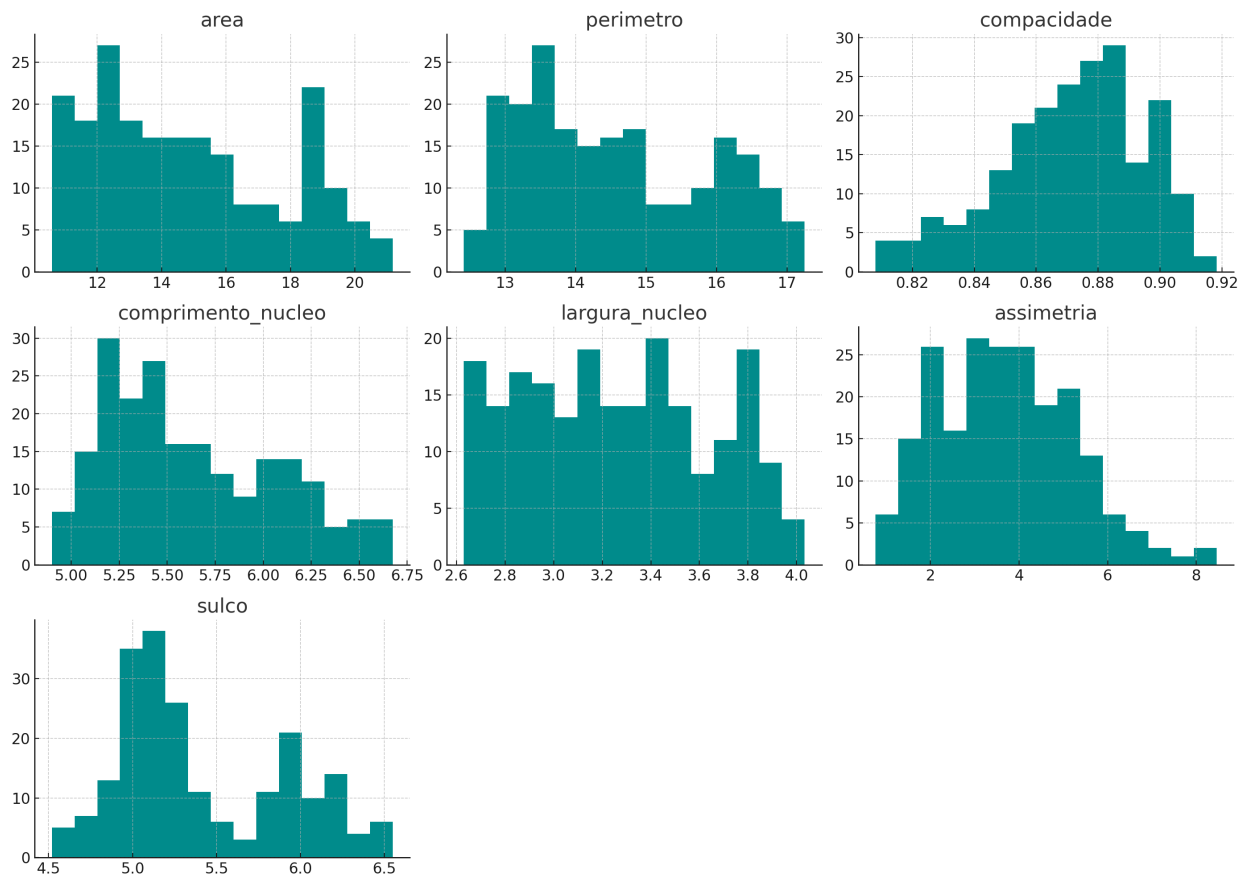
Abaixo, apresentamos a matriz de correlação entre os atributos físicos dos grãos:



Fonte: Análise gerada com Seaborn e Matplotlib.

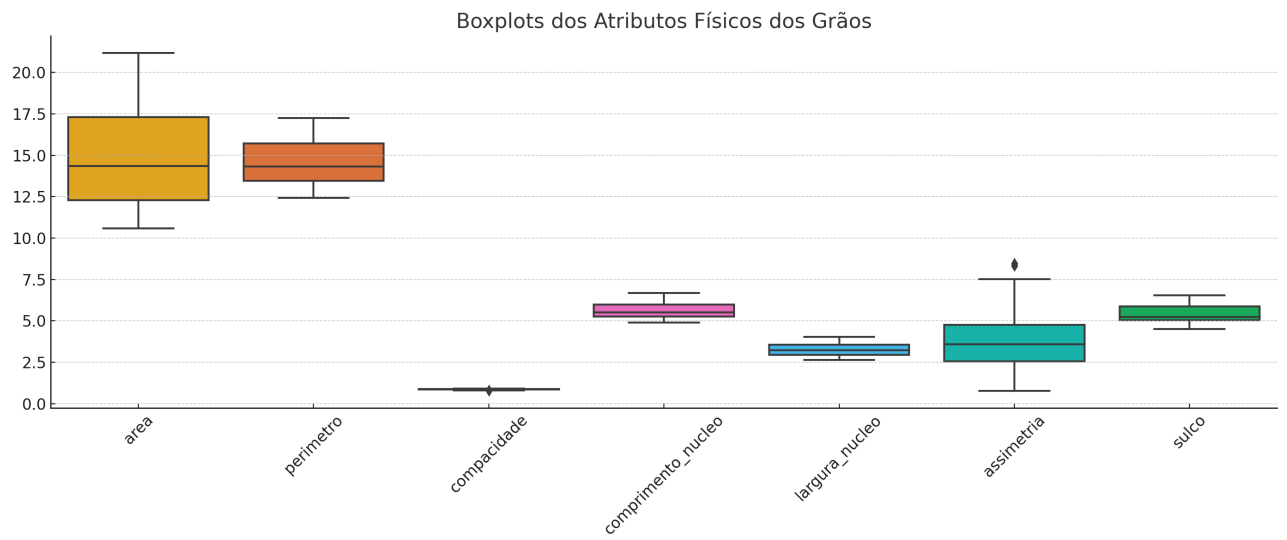
7. Histogramas

Distribuição dos atributos do conjunto de dados.



8. Boxplots

Boxplots para análise de dispersão e identificação de possíveis outliers.



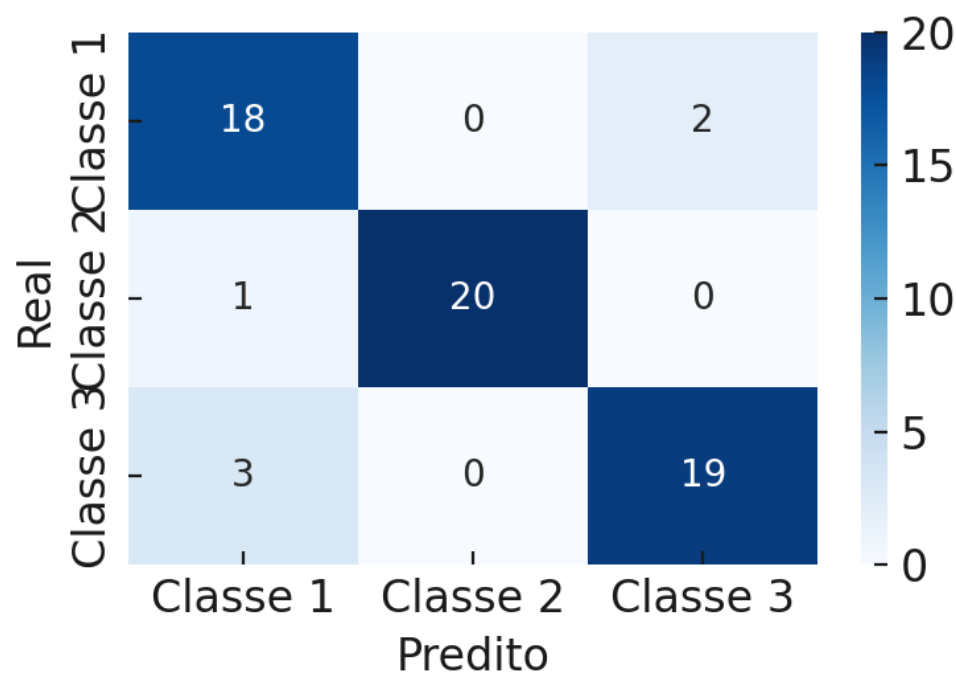
9. Avaliação dos Modelos

KNN

Acurácia: 90.48%

F1-score (macro): 90.55%

Matriz de Confusão - KNN



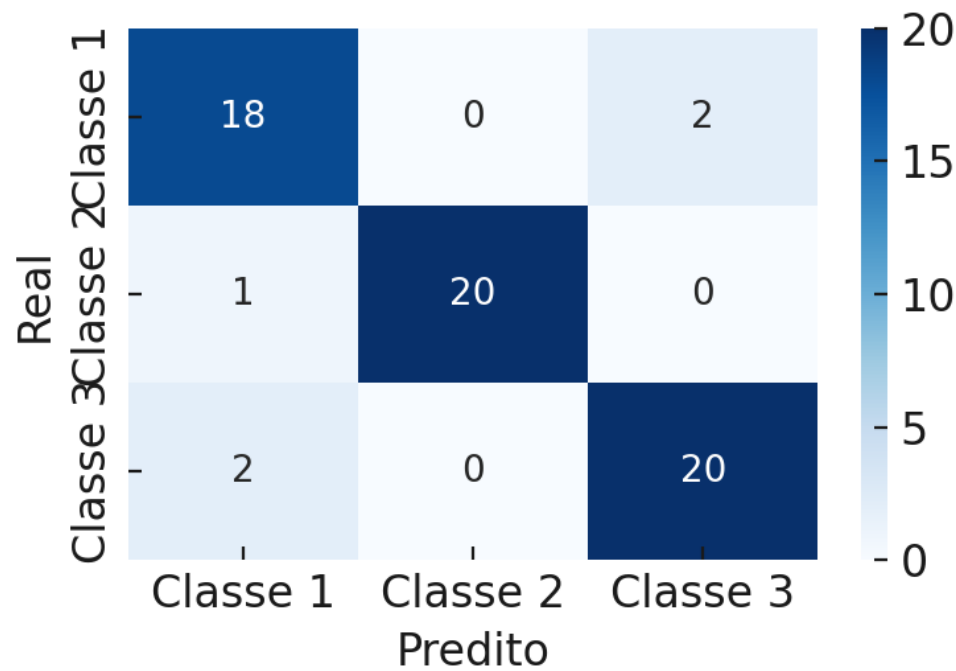
Fonte: Resultado gerado a partir dos dados de teste.

SVM

Acurácia: 92.06%

F1-score (macro): 92.09%

Matriz de Confusão – SVM



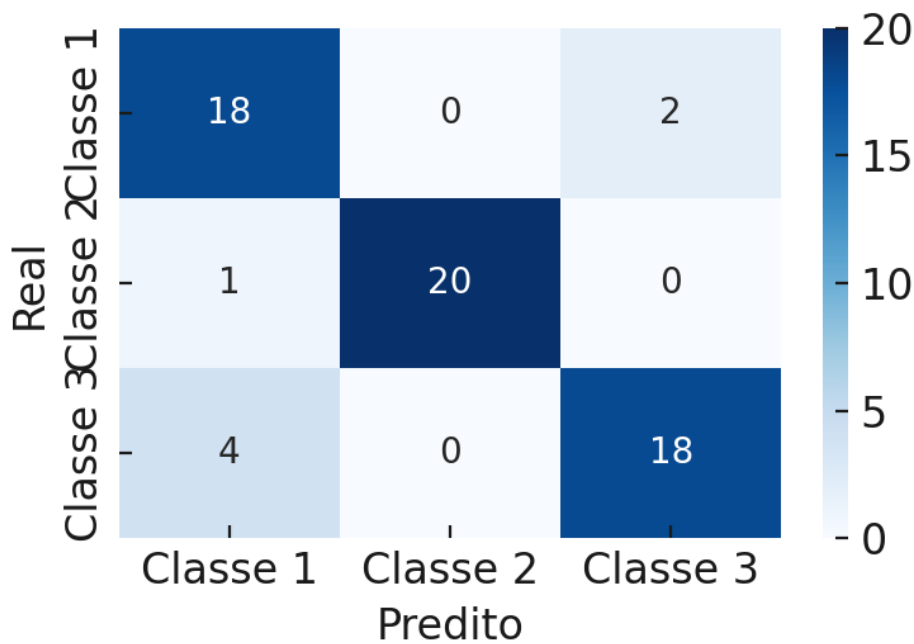
Fonte: Resultado gerado a partir dos dados de teste.

Random Forest

Acurácia: 88.89%

F1-score (macro): 89.00%

Matriz de Confusão – Random Forest



Fonte: Resultado gerado a partir dos dados de teste.

10. SVM Otimizado (GridSearchCV)

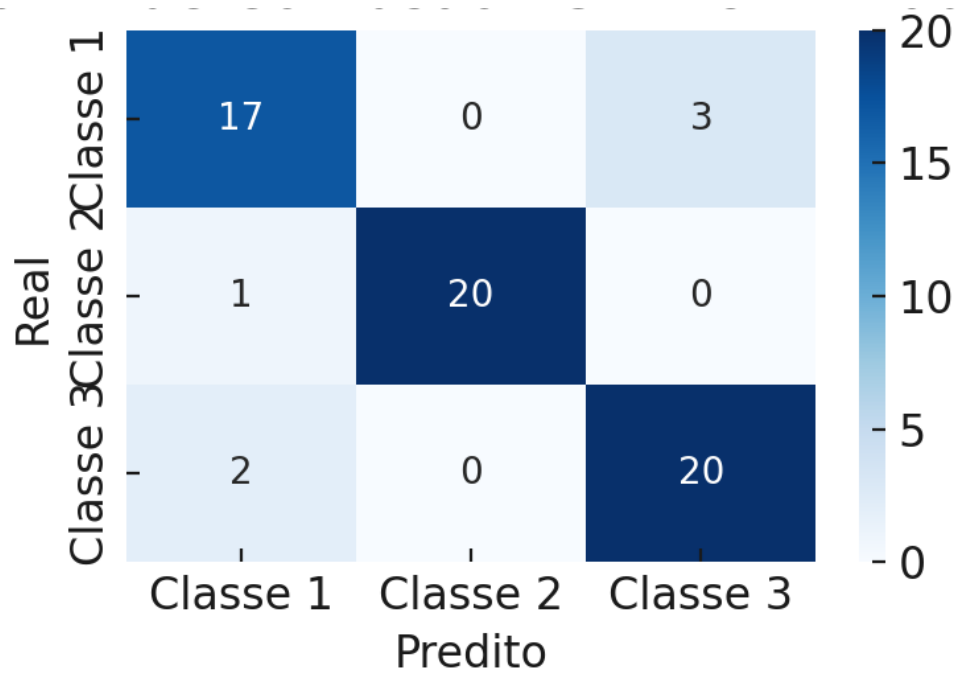
Melhores parâmetros encontrados:

{'C': 1, 'gamma': 'scale', 'kernel': 'linear'}

Acurácia: 90.48%

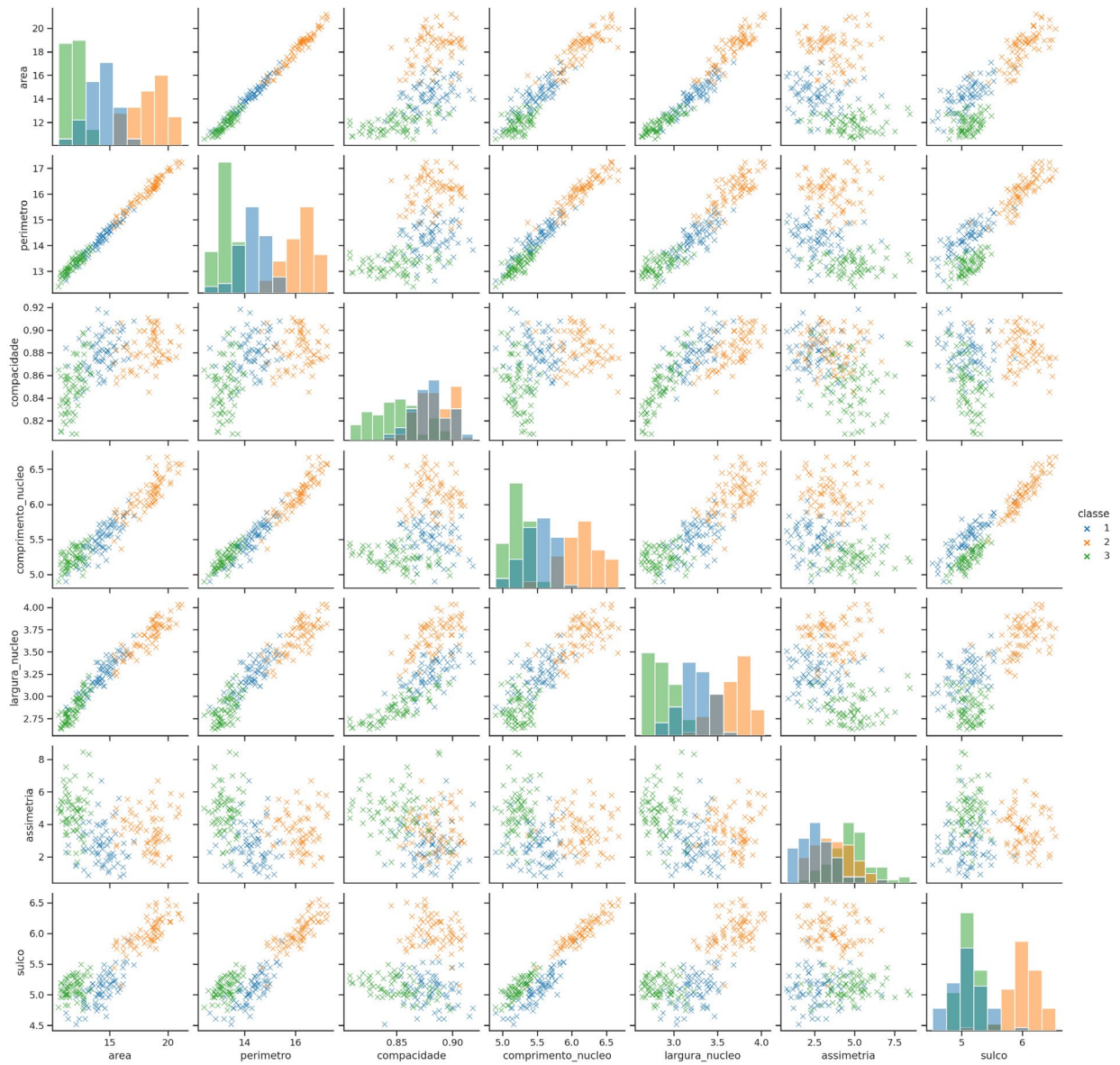
F1-score (macro): 90.48%

Matriz de Confusão – SVM Otimizado



11. Gráfico de Dispersão (Pairplot)

Visualização das relações entre todos os atributos do conjunto de dados Seeds, diferenciando as três classes de grãos (Kama, Rosa e Canadian):



Fonte: Análise gerada com Seaborn e Matplotlib.