CAP 3 - FASE 4: AUTOMATIZANDO A CLASSIFICAÇÃO DE GRÃOS COM MACHINE LEARNING

Análise inicial do dataset:

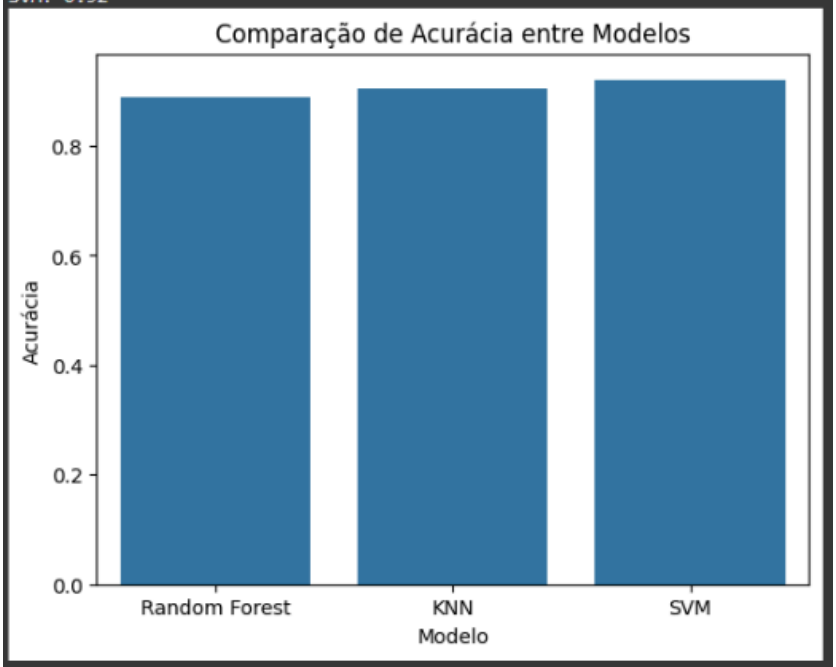
```
Primeiras linhas do conjunto de dados:
    Area  Perimetro  Compacidade  Comprimento_Nucleo  Largura_Nucleo  \
0  15.26      14.84       0.8710               5.763           3.312
1  14.88      14.57       0.8811               5.554           3.333
2  14.29      14.09       0.9050               5.291           3.337
3  13.84      13.94       0.8955               5.324           3.379
4  16.14      14.99       0.9034               5.658           3.562

   Coef_Assimetria  Comprimento_Sulco  Classe
0            2.221              5.220       1
1            1.018              4.956       1
2            2.699              4.825       1
3            2.259              4.805       1
4            1.355              5.175       1

Resumo estatístico dos dados:
              Area   Perimetro  Compacidade  Comprimento_Nucleo  \
count  210.000000  210.000000   210.000000          210.000000
mean    14.847524   14.559286     0.870999            5.628533
std      2.909699    1.305959     0.023629            0.443063
min     10.590000   12.410000     0.808100            4.899000
25%     12.270000   13.450000     0.856900            5.262250
50%     14.355000   14.320000     0.873450            5.523500
75%     17.305000   15.715000     0.887775            5.979750
max     21.180000   17.250000     0.918300            6.675000

       Largura_Nucleo  Coef_Assimetria  Comprimento_Sulco      Classe
count      210.000000       210.000000         210.000000  210.000000
mean         3.258605         3.700201           5.408071    2.000000
std          0.377714         1.503557           0.491480    0.818448
min          2.630000         0.765100           4.519000    1.000000
25%          2.944000         2.561500           5.045000    1.000000
50%          3.237000         3.599000           5.223000    2.000000
75%          3.561750         4.768750           5.877000    3.000000
max          4.033000         8.456000           6.550000    3.000000

Valores ausentes:
Area                0
Perimetro           0
Compacidade         0
Comprimento_Nucleo  0
Largura_Nucleo      0
Coef_Assimetria     0
Comprimento_Sulco   0
Classe              0
dtype: int64
```

Comparação dos Modelos

Desempenho Geral:

- **SVM** teve o melhor desempenho, com uma acurácia de **92%**, seguido por **KNN** com **90%**, e por fim, **Random Forest** com **89%**.
- O **SVM** obteve a melhor pontuação de F1-score.

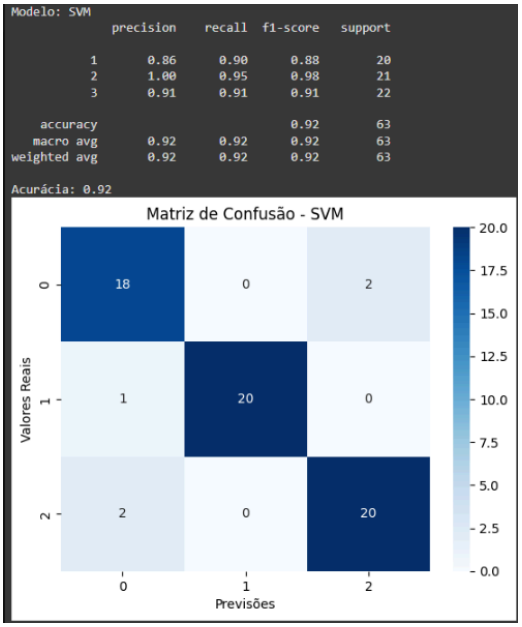
```
Comparação de Acurácia entre os modelos:
Random Forest: 0.89
KNN: 0.90
SVM: 0.92
```

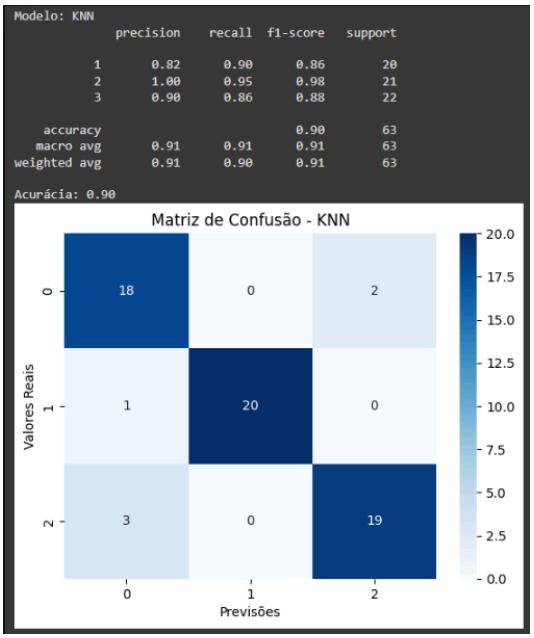


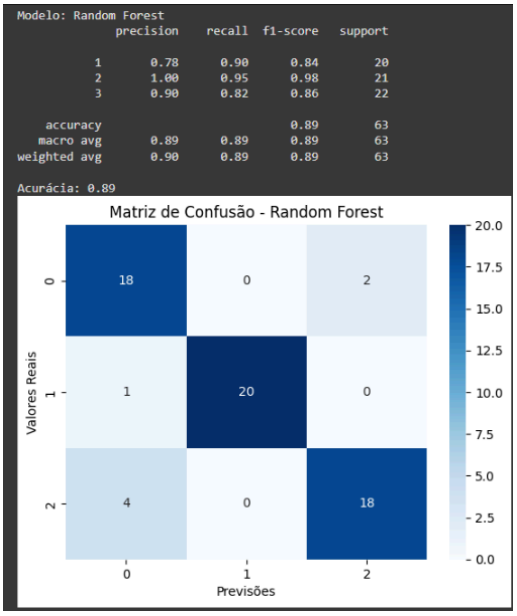

```
Melhores parâmetros encontrados para Random Forest:
{'max_depth': None, 'n_estimators': 100}

Desempenho do modelo otimizado (Random Forest):
              precision    recall  f1-score   support

           1       0.78      0.90      0.84        20
           2       1.00      0.95      0.98        21
           3       0.90      0.82      0.86        22

    accuracy                           0.89        63
   macro avg       0.89      0.89      0.89        63
weighted avg       0.90      0.89      0.89        63
```

Matriz de Confusão:

- Os modelos conseguiram classificar a maioria das amostras de forma correta, porém pequenas diferenças nos erros destacam o melhor desempenho do SVM.
- **SVM** teve menos erros de classificação nas classes 2 e 3.





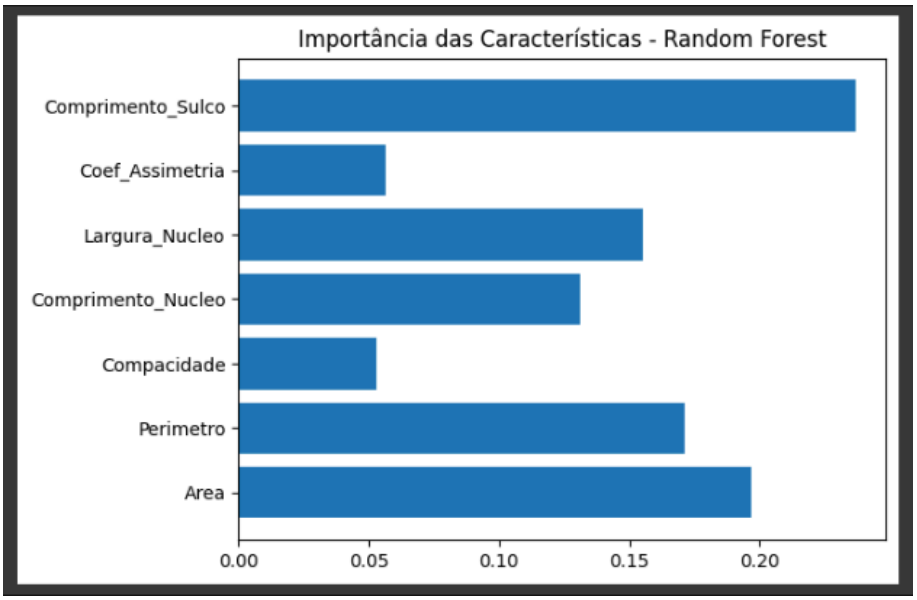


Importância das Características:

- No **Random Forest**, as características mais importantes foram:
    1. **Comprimento_Sulco**
    2. **Área**

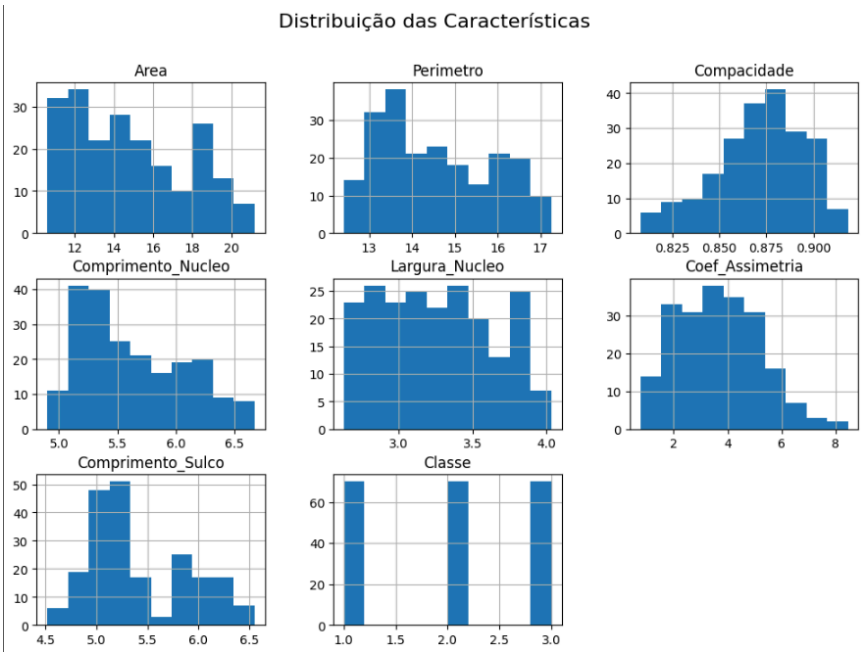
3. **Perímetro**

- Isso sugere que essas características são discriminativas entre as classes.



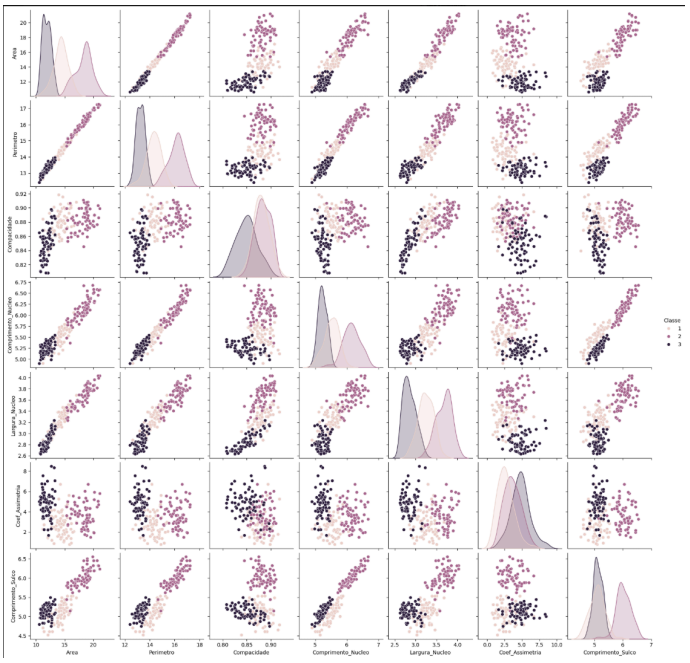
Visualizações

- **Distribuição das Características**:
  - A visualização por histogramas mostra uma boa separação entre classes, especialmente nas características **Área**, **Perímetro** e **Comprimento_Sulco**.
  - As características **Coef_Assimetria** e **Compacidade** apresentam sobreposição maior entre as classes.

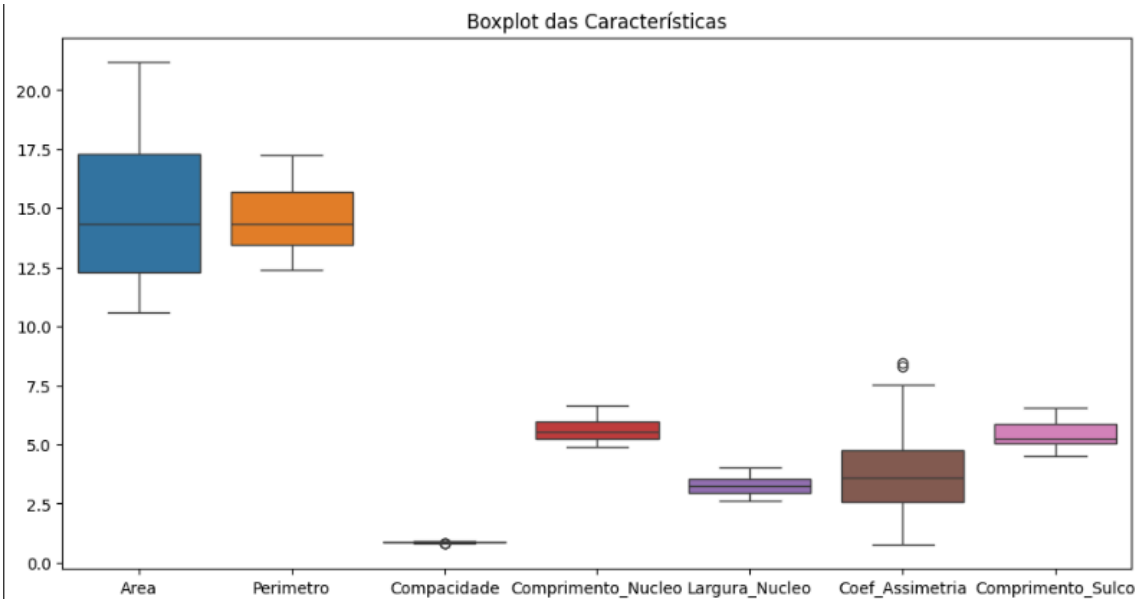


- **Relações entre Características (Pairplot)**:

- O gráfico de dispersão evidencia forte correlação entre algumas características, como **Área** e **Perímetro**, e agrupamentos claros para as diferentes classes.



- **Boxplots**:
  - As diferenças de medianas confirmam a relevância de características como **Comprimento_Sulco** e **Área** na discriminação das classes.



Conclusões

A abordagem automatizada foi eficaz para classificar as variedades de grãos de trigo, alcançando altas taxas de acurácia e métricas robustas. Isso demonstra que o aprendizado de máquina pode substituir ou complementar métodos manuais, reduzindo o tempo e o risco de erros.

O modelo **SVM** foi a melhor escolha para este dataset, devido à sua alta acurácia e equilíbrio entre precisão e recall.