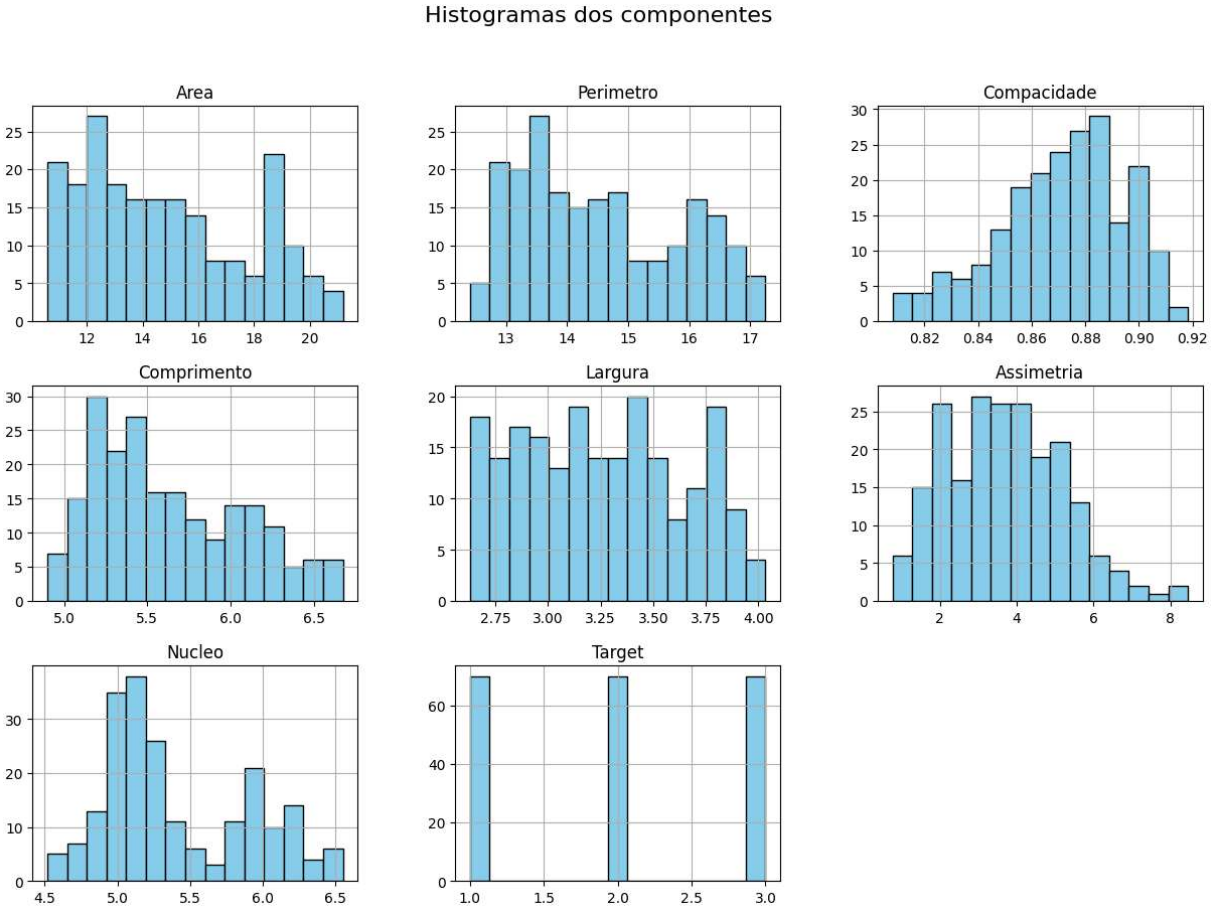
In [1]:

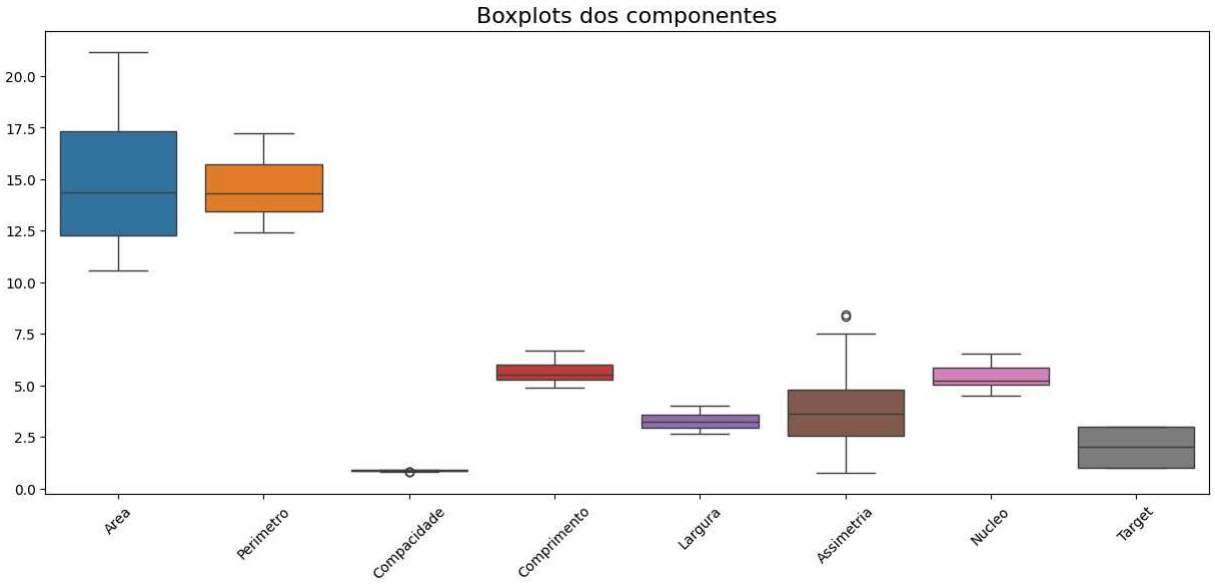
```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Read the file
file_path = 'seeds_dataset.csv'
data = pd.read_csv(file_path)

# 2. Show the first lines
print("Step 2 First 5 rows of the dataset:")
print(data.head())

# 3. Calculate descriptive statistics
print("Step 3 Statisticas:")
stats = data.describe()
median = data.median()
print("\nEstatísticas Descritivas:")
print(stats)
print("\nMediana de cada valor:")
print(median)

# 4. Visualize the distribution of features
# Histograms
print("Step 4 - Histogramas:")
data.hist(bins=15, figsize=(15, 10), color='skyblue', edgecolor='black')
plt.suptitle('Histogramas dos componentes', fontsize=16)
plt.show()

# Boxplots
print("Step 4 - Boxplots:")
plt.figure(figsize=(15, 6))
sns.boxplot(data=data)
plt.title('Boxplots dos componentes', fontsize=16)
plt.xticks(rotation=45)

plt.show()

# 5. Scatter plots to identify possible relationships
print("Step 5 - Possíveis relacionamentos:")
sns.pairplot(data, diag_kind='kde', hue='Target', markers=["o", "s", "D"])
plt.suptitle('Gráfico de dispersão dos componentes', fontsize=16)
plt.show()

# 6. Identify and handle missing values
print("Step 6 - Identificando valores faltantes:")
missing_values = data.isnull().sum()
print("\nValores faltantes em cada componente:")
print(missing_values)

if missing_values.any():
    print("\nPreenchendo os valores faltantes com a média, caso necessário")
    data.fillna(data.mean(), inplace=True)
```

```python
# 7. Assess the need for scaling and normalization
print("Step 7 - Normalização - se necessário:")
print("\nVerificando o intervalo para dimensionamento:")
print(data.max() - data.min())

# Applying normalization (Min-Max Scaling)
normalized_data = (data - data.min()) / (data.max() - data.min())

# Applying standardization (Z-score Scaling)
standardized_data = (data - data.mean()) / data.std()

# Display normalized and standardized samples
print("\nAmostra do dado normalizado:")
print(normalized_data.head())
print("\nAmostra do dado padronizado:")
print(standardized_data.head())
```

```
Step 2 First 5 rows of the dataset:
    Area  Perimetro  Compacidade  Comprimento  Largura  Assimetria  Nucleo  \
0  15.26      14.84       0.8710        5.763    3.312       2.221   5.220
1  14.88      14.57       0.8811        5.554    3.333       1.018   4.956
2  14.29      14.09       0.9050        5.291    3.337       2.699   4.825
3  13.84      13.94       0.8955        5.324    3.379       2.259   4.805
4  16.14      14.99       0.9034        5.658    3.562       1.355   5.175

   Target
0       1
1       1
2       1
3       1
4       1
Step 3 Statisticas:

Estatísticas Descritivas:
             Area   Perimetro  Compacidade  Comprimento     Largura  \
count  210.000000  210.000000   210.000000   210.000000  210.000000
mean    14.847524   14.559286     0.870999     5.628533    3.258605
std      2.909699    1.305959     0.023629     0.443063    0.377714
min     10.590000   12.410000     0.808100     4.899000    2.630000
25%     12.270000   13.450000     0.856900     5.262250    2.944000
50%     14.355000   14.320000     0.873450     5.523500    3.237000
75%     17.305000   15.715000     0.887775     5.979750    3.561750
max     21.180000   17.250000     0.918300     6.675000    4.033000

       Assimetria      Nucleo      Target
count  210.000000  210.000000  210.000000
mean     3.700201    5.408071    2.000000
std      1.503557    0.491480    0.818448
min      0.765100    4.519000    1.000000
25%      2.561500    5.045000    1.000000
50%      3.599000    5.223000    2.000000
75%      4.768750    5.877000    3.000000
max      8.456000    6.550000    3.000000

Mediana de cada valor:
Area           14.35500
Perimetro      14.32000
Compacidade     0.87345
Comprimento     5.52350
Largura         3.23700
Assimetria      3.59900
Nucleo          5.22300
Target          2.00000
dtype: float64
Step 4 - Histogramas:
```



Step 4 - Boxplots:



Step 5 - Possíveis relacionamentos:

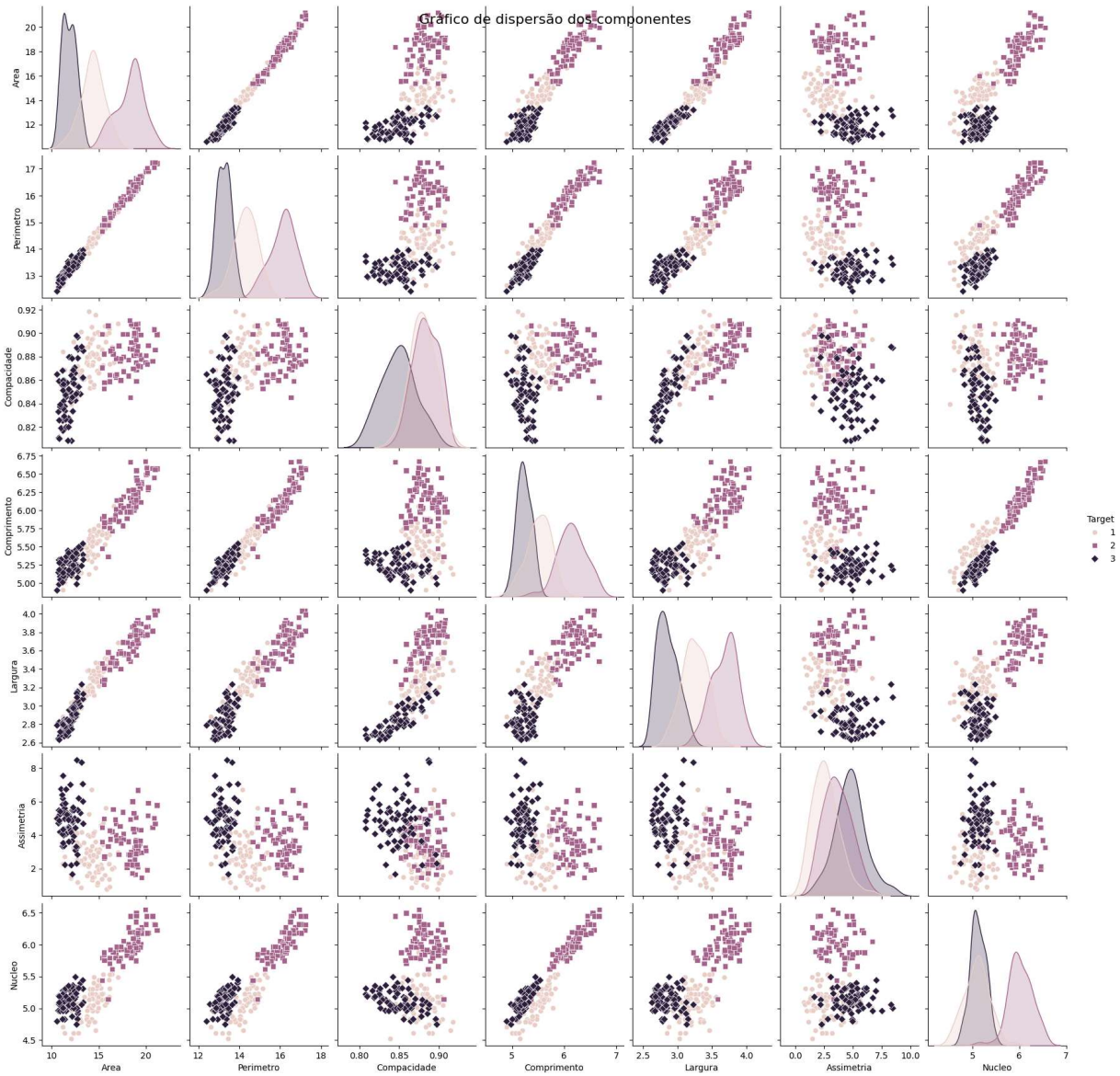


Gráfico de dispersão dos componentes
Area
Perimetro
Compacidade
Comprimento
Largura
Assimetria
Nucleo
Target
1
2
3

```
Step 6 - Identificando valores faltantes:

Valores faltantes em cada componente:
Area           0
Perimetro      0
Compacidade    0
Comprimento    0
Largura        0
Assimetria     0
Nucleo         0
Target         0
dtype: int64
Step 7 - Normalização - se necessário:

Verificando o intervalo para dimensionamento:
Area           10.5900
Perimetro       4.8400
Compacidade     0.1102
Comprimento     1.7760
Largura         1.4030
Assimetria      7.6909
Nucleo          2.0310
Target          2.0000
dtype: float64

Amostra do dado normalizado:
       Area  Perimetro  Compacidade  Comprimento   Largura  Assimetria  \
0  0.440982   0.502066     0.570780     0.486486  0.486101    0.189302
1  0.405099   0.446281     0.662432     0.368806  0.501069    0.032883
2  0.349386   0.347107     0.879310     0.220721  0.503920    0.251453
3  0.306893   0.316116     0.793103     0.239302  0.533856    0.194243
4  0.524079   0.533058     0.864791     0.427365  0.664291    0.076701

     Nucleo  Target
0  0.345150     0.0
1  0.215165     0.0
2  0.150665     0.0
3  0.140817     0.0
4  0.322994     0.0

Amostra do dado padronizado:
       Area  Perimetro  Compacidade  Comprimento   Largura  Assimetria  \
0  0.141759   0.214949     0.000060     0.303493  0.141364   -0.983801
1  0.011161   0.008204     0.427494    -0.168223  0.196962   -1.783904
2 -0.191609  -0.359342     1.438945    -0.761817  0.207552   -0.665888
3 -0.346264  -0.474200     1.036904    -0.687336  0.318747   -0.958528
4  0.444196   0.329807     1.371233     0.066507  0.803240   -1.559768

     Nucleo    Target
0 -0.382663 -1.221825
1 -0.919816 -1.221825
2 -1.186357 -1.221825
3 -1.227051 -1.221825
4 -0.474223 -1.221825
```

In [ ]: