

```
In [3]: import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, GridSearchCV
from sklearn.preprocessing import MinMaxScaler
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error, r2_score, accuracy_score
```

```
In [5]: data = pd.read_csv("data/sustainability_dataset2.csv")
```

```
In [7]: print("\n--- Dataset Overview ---")
print("Shape:", data.shape)
print("Columns:", data.columns)
print("\nData Types:\n", data.dtypes)
print("\nMissing Values:\n", data.isnull().sum())
print("\nDescriptive Statistics:\n", data.describe())
```

--- Dataset Overview ---

Shape: (1500, 7)

Columns: Index(['Product Name', 'Material Type', 'Carbon Footprint (kg CO<sub>2</sub>)',  
'Energy Consumption (kWh)', 'Brand Policy (Eco-Certified)',  
'Number of Certifications', 'Sustainability Score'],  
dtype='object')

Data Types:

Product Name	object
Material Type	int64
Carbon Footprint (kg CO <sub>2</sub> )	float64
Energy Consumption (kWh)	float64
Brand Policy (Eco-Certified)	int64
Number of Certifications	int64
Sustainability Score	float64
dtype:	object

Missing Values:

Product Name	0
Material Type	0
Carbon Footprint (kg CO <sub>2</sub> )	0
Energy Consumption (kWh)	0
Brand Policy (Eco-Certified)	0
Number of Certifications	0
Sustainability Score	0
dtype:	int64

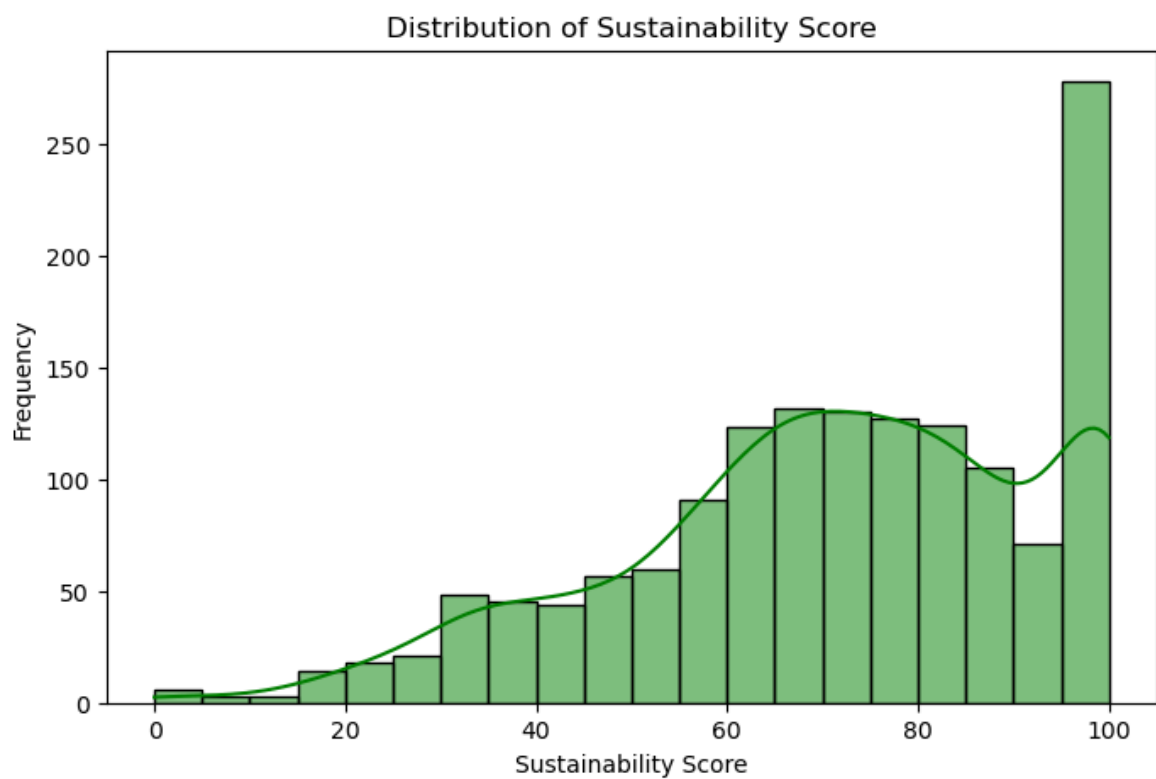
Descriptive Statistics:

	Material Type	Carbon Footprint (kg CO <sub>2</sub> )	Energy Consumption (kWh)
\			
count	1500.000000	1500.000000	1500.000000
mean	2.142000	44.987216	151.237163
std	0.945039	28.731402	74.660456
min	1.000000	10.092640	48.730088
25%	1.000000	23.941632	91.005148
50%	2.000000	40.294500	136.960072
75%	3.000000	61.429049	203.103788
max	4.000000	373.142395	697.941023

	Brand Policy (Eco-Certified)	Number of Certifications	\
count	1500.000000	1500.000000	
mean	0.509333	2.304667	
std	0.500080	1.618721	
min	0.000000	0.000000	
25%	0.000000	1.000000	
50%	1.000000	2.000000	
75%	1.000000	4.000000	
max	1.000000	5.000000	

	Sustainability Score
count	1500.000000
mean	71.372955
std	21.868172
min	0.000000
25%	58.155674
50%	73.190102
75%	88.414978
max	100.000000

```
In [9]: plt.figure(figsize=(8, 5))
sns.histplot(data["Sustainability Score"], bins=20, kde=True, color="green")
plt.title("Distribution of Sustainability Score")
plt.xlabel("Sustainability Score")
plt.ylabel("Frequency")
plt.show()
```



```
In [11]: X = data.drop(columns=["Product Name", "Sustainability Score"])
y = data["Sustainability Score"]
```

```
In [13]: scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

```
In [15]: X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

```
In [17]: # hyper parameter
param_grid = {
    "n_estimators": [50, 100, 200],
    "max_depth": [None, 10, 20, 30],
    "min_samples_split": [2, 5, 10],
    "min_samples_leaf": [1, 2, 4],
    "max_features": ["auto", "sqrt", "log2"]
}
```

```
In [19]: rf = RandomForestRegressor(random_state=42)
grid_search = GridSearchCV(rf, param_grid, scoring='r2', cv=5, n_jobs=-1)
grid_search.fit(X_train, y_train)
import warnings
warnings.filterwarnings('ignore')
```

4/27

5/27

6/27

7/27

8/27



9/27

10/27

11/27

12/27

13/27

4/27

15/27

16/27



17/27

18/27

9/27

20/27

21/27

```
[CV] END max_depth=30, max_features=auto, min_samples_leaf=1, min_samples_
split=10, n_estimators=100; total time= 0.0s
```

```

/opt/anaconda3/lib/python3.11/site-packages/sklearn/model_selection/_validation.py:528: FitFailedWarning:
540 fits failed out of a total of 1620.
The score on these train-test partitions for these parameters will be set
to nan.
If these failures are not expected, you can try to debug them by setting e
rror_score='raise'.

```

Below are more details about the failures:

```

-----
265 fits failed with the following error:
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.11/site-packages/sklearn/model_selectio
n/_validation.py", line 866, in _fit_and_score
    estimator.fit(X_train, y_train, **fit_params)
  File "/opt/anaconda3/lib/python3.11/site-packages/sklearn/base.py", line
1382, in wrapper
    estimator._validate_params()
  File "/opt/anaconda3/lib/python3.11/site-packages/sklearn/base.py", line
436, in _validate_params
    validate_parameter_constraints(
  File "/opt/anaconda3/lib/python3.11/site-packages/sklearn/utils/_param_v
alidation.py", line 98, in validate_parameter_constraints
    raise InvalidParameterError(
sklearn.utils._param_validation.InvalidParameterError: The 'max_features'
parameter of RandomForestRegressor must be an int in the range [1, inf), a
float in the range (0.0, 1.0], a str among {'log2', 'sqrt'} or None. Got
'auto' instead.

```

```

-----
275 fits failed with the following error:
Traceback (most recent call last):
  File "/opt/anaconda3/lib/python3.11/site-packages/sklearn/model_selectio
n/_validation.py", line 866, in _fit_and_score
    estimator.fit(X_train, y_train, **fit_params)
  File "/opt/anaconda3/lib/python3.11/site-packages/sklearn/base.py", line
1382, in wrapper
    estimator._validate_params()
  File "/opt/anaconda3/lib/python3.11/site-packages/sklearn/base.py", line
436, in _validate_params
    validate_parameter_constraints(
  File "/opt/anaconda3/lib/python3.11/site-packages/sklearn/utils/_param_v
alidation.py", line 98, in validate_parameter_constraints
    raise InvalidParameterError(
sklearn.utils._param_validation.InvalidParameterError: The 'max_features'
parameter of RandomForestRegressor must be an int in the range [1, inf), a
float in the range (0.0, 1.0], a str among {'sqrt', 'log2'} or None. Got
'auto' instead.

```

```

    warnings.warn(some_fits_failed_message, FitFailedWarning)
/opt/anaconda3/lib/python3.11/site-packages/sklearn/model_selection/_search
h.py:1108: UserWarning: One or more of the test scores are non-finite: [
nan          nan          nan          nan          nan          nan
      nan          nan          nan          nan          nan          nan
      nan          nan          nan          nan          nan          nan
      nan          nan          nan  0.90458065  0.90422603  0.9042407
0.90154555  0.90109919  0.90123471  0.89411844  0.89432123  0.89605091

```

```

0.89746687 0.89857306 0.89914058 0.89677983 0.89776524 0.89922844
0.88998722 0.8917746 0.89274801 0.88318668 0.8841201 0.88481313
0.88318668 0.8841201 0.88481313 0.88334807 0.88316757 0.8825256
0.90458065 0.90422603 0.9042407 0.90154555 0.90109919 0.90123471
0.89411844 0.89432123 0.89605091 0.89746687 0.89857306 0.89914058
0.89677983 0.89776524 0.89922844 0.88998722 0.8917746 0.89274801
0.88318668 0.8841201 0.88481313 0.88318668 0.8841201 0.88481313
0.88334807 0.88316757 0.8825256 nan nan nan
nan nan nan nan nan nan
nan nan nan nan nan nan
nan nan nan nan nan nan
0.90140684 0.90211433 0.90256605 0.89834074 0.89939992 0.90093736
0.89415359 0.89646797 0.89573559 0.89470218 0.89839301 0.8992903
0.89714071 0.89756731 0.89862595 0.89236609 0.89337565 0.89340479
0.88096668 0.88264196 0.88372937 0.88096668 0.88264196 0.88372937
0.88416862 0.88349838 0.88211967 0.90140684 0.90211433 0.90256605
0.89834074 0.89939992 0.90093736 0.89415359 0.89646797 0.89573559
0.89470218 0.89839301 0.8992903 0.89714071 0.89756731 0.89862595
0.89236609 0.89337565 0.89340479 0.88096668 0.88264196 0.88372937
0.88096668 0.88264196 0.88372937 0.88416862 0.88349838 0.88211967
nan nan nan nan nan nan
nan nan nan nan nan nan
nan nan nan nan nan nan
nan nan nan nan nan nan
0.90154555 0.90109919 0.901236 0.89411844 0.89432123 0.89605091
0.89746687 0.89857306 0.89914058 0.89677983 0.89776524 0.89922844
0.88998722 0.8917746 0.89274801 0.88318668 0.8841201 0.88481313
0.88318668 0.8841201 0.88481313 0.88334807 0.88316757 0.8825256
0.90458065 0.90424931 0.90424649 0.90154555 0.90109919 0.901236
0.89411844 0.89432123 0.89605091 0.89746687 0.89857306 0.89914058
0.89677983 0.89776524 0.89922844 0.88998722 0.8917746 0.89274801
0.88318668 0.8841201 0.88481313 0.88318668 0.8841201 0.88481313
0.88334807 0.88316757 0.8825256 nan nan nan
nan nan nan nan nan nan
nan nan nan nan nan nan
nan nan nan nan nan nan
0.90458065 0.90422603 0.9042407 0.90154555 0.90109919 0.90123471
0.89411844 0.89432123 0.89605091 0.89746687 0.89857306 0.89914058
0.89677983 0.89776524 0.89922844 0.88998722 0.8917746 0.89274801
0.88318668 0.8841201 0.88481313 0.88318668 0.8841201 0.88481313
0.88334807 0.88316757 0.8825256 0.90458065 0.90422603 0.9042407
0.90154555 0.90109919 0.90123471 0.89411844 0.89432123 0.89605091
0.89746687 0.89857306 0.89914058 0.89677983 0.89776524 0.89922844
0.88998722 0.8917746 0.89274801 0.88318668 0.8841201 0.88481313
0.88318668 0.8841201 0.88481313 0.88334807 0.88316757 0.8825256 ]
warnings.warn(

```

```
In [20]: print("Best Hyperparameters:", grid_search.best_params_)
best_rf = grid_search.best_estimator_
```

Best Hyperparameters: {'max\_depth': None, 'max\_features': 'sqrt', 'min\_samples\_leaf': 1, 'min\_samples\_split': 2, 'n\_estimators': 50}

```
In [21]: # Step 6: Model Evaluation
y_pred = best_rf.predict(X_test)
mse = mean_squared_error(y_test, y_pred)
```

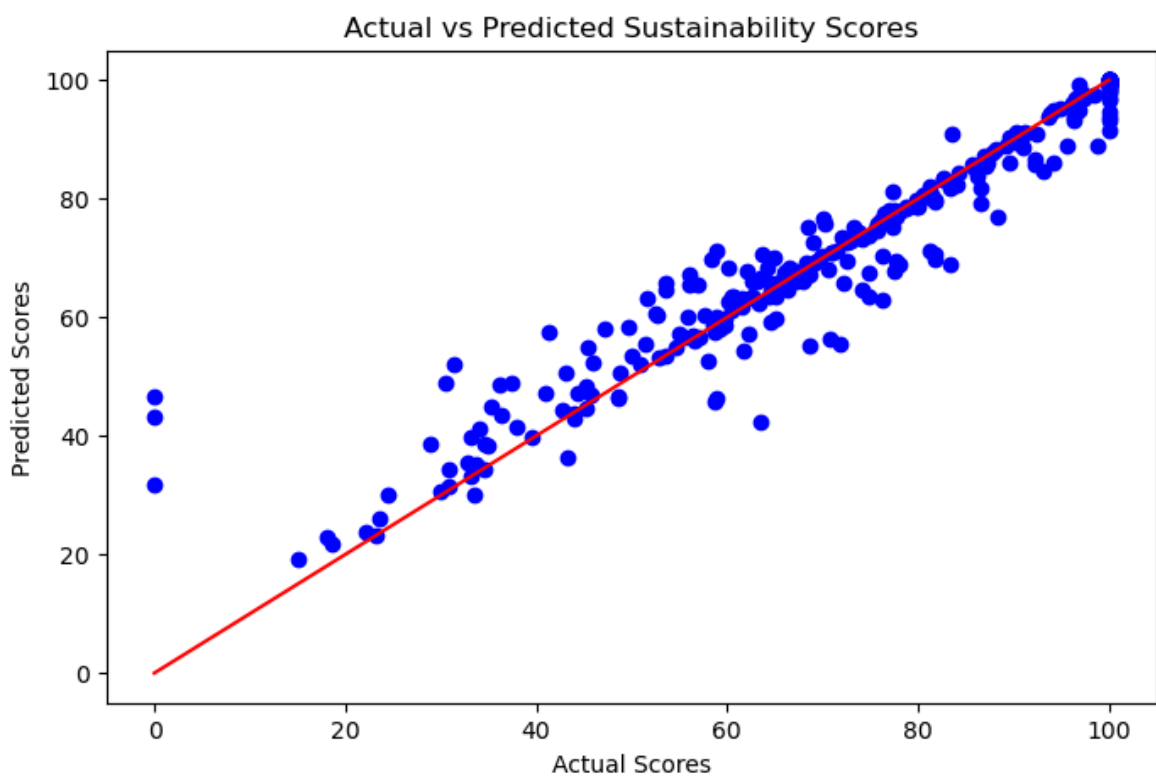


```
r2 = r2_score(y_test, y_pred)
accuracy = best_rf.score(X_test, y_test) * 100
```

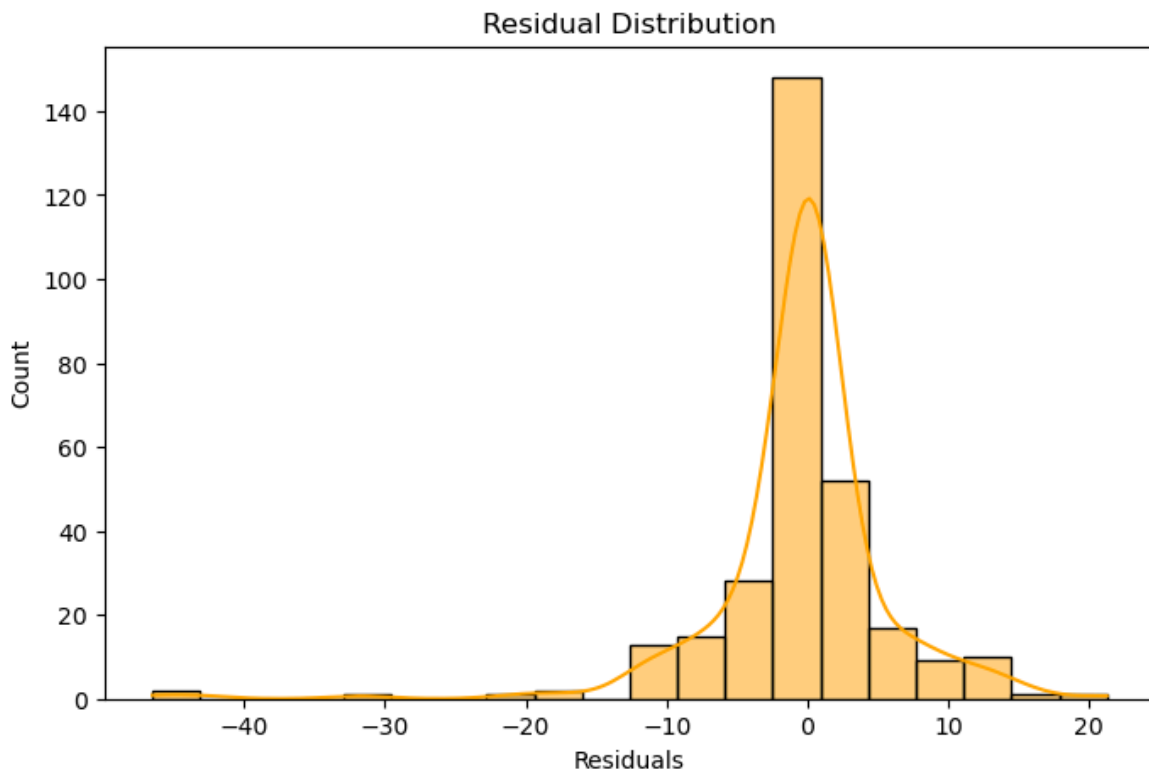
```
In [22]: print("\n--- Model Performance ---")
print("Mean Squared Error (MSE):", mse)
print("R2 Score:", r2)
print("Accuracy:", accuracy, "%")
```

```
--- Model Performance ---
Mean Squared Error (MSE): 42.832928455178845
R2 Score: 0.9166630627596599
Accuracy: 91.666306275966 %
```

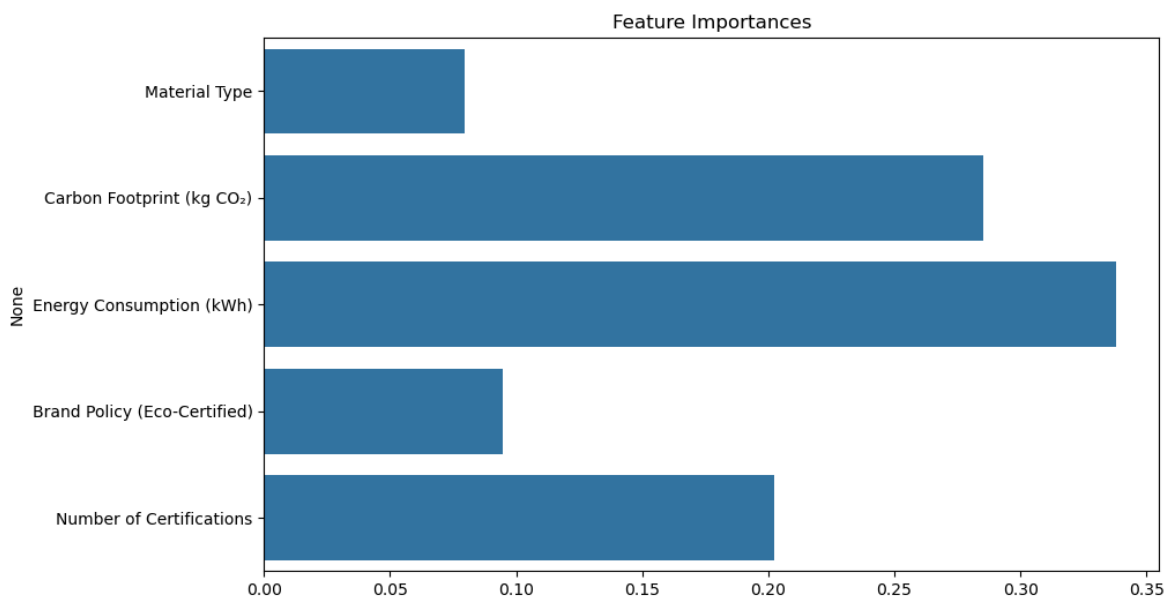
```
In [23]: plt.figure(figsize=(8, 5))
plt.scatter(y_test, y_pred, color="blue")
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], color="r")
plt.title("Actual vs Predicted Sustainability Scores")
plt.xlabel("Actual Scores")
plt.ylabel("Predicted Scores")
plt.show()
```



```
In [24]: residuals = y_test - y_pred
plt.figure(figsize=(8, 5))
sns.histplot(residuals, bins=20, kde=True, color="orange")
plt.title("Residual Distribution")
plt.xlabel("Residuals")
plt.show()
```



```
In [25]: feature_importances = best_rf.feature_importances_  
features = X.columns  
plt.figure(figsize=(10, 6))  
sns.barplot(x=feature_importances, y=features)  
plt.title("Feature Importances")  
plt.show()
```



```
In [26]: import joblib  
joblib.dump(best_rf, "sustainability_rf_model.joblib")
```

```
Out[26]: ['sustainability_rf_model.joblib']
```

```
In [27]: # Example hard-coded input values (replace with actual values)  
input_values = np.array([[10, 20, 30, 0.5, 15]]) # Replace with actual f  
  
# Scale the input values using the same scaler  
input_values_scaled = scaler.transform(input_values)
```

```
# Predict the sustainability score  
predicted_score = best_rf.predict(input_values_scaled)  
  
print("Predicted Sustainability Score:", predicted_score[0])
```

Predicted Sustainability Score: 91.51675662420931

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: