DWDM ASSIGNMENT – 8

21BCE7371
RADHA KRISHNA GARG

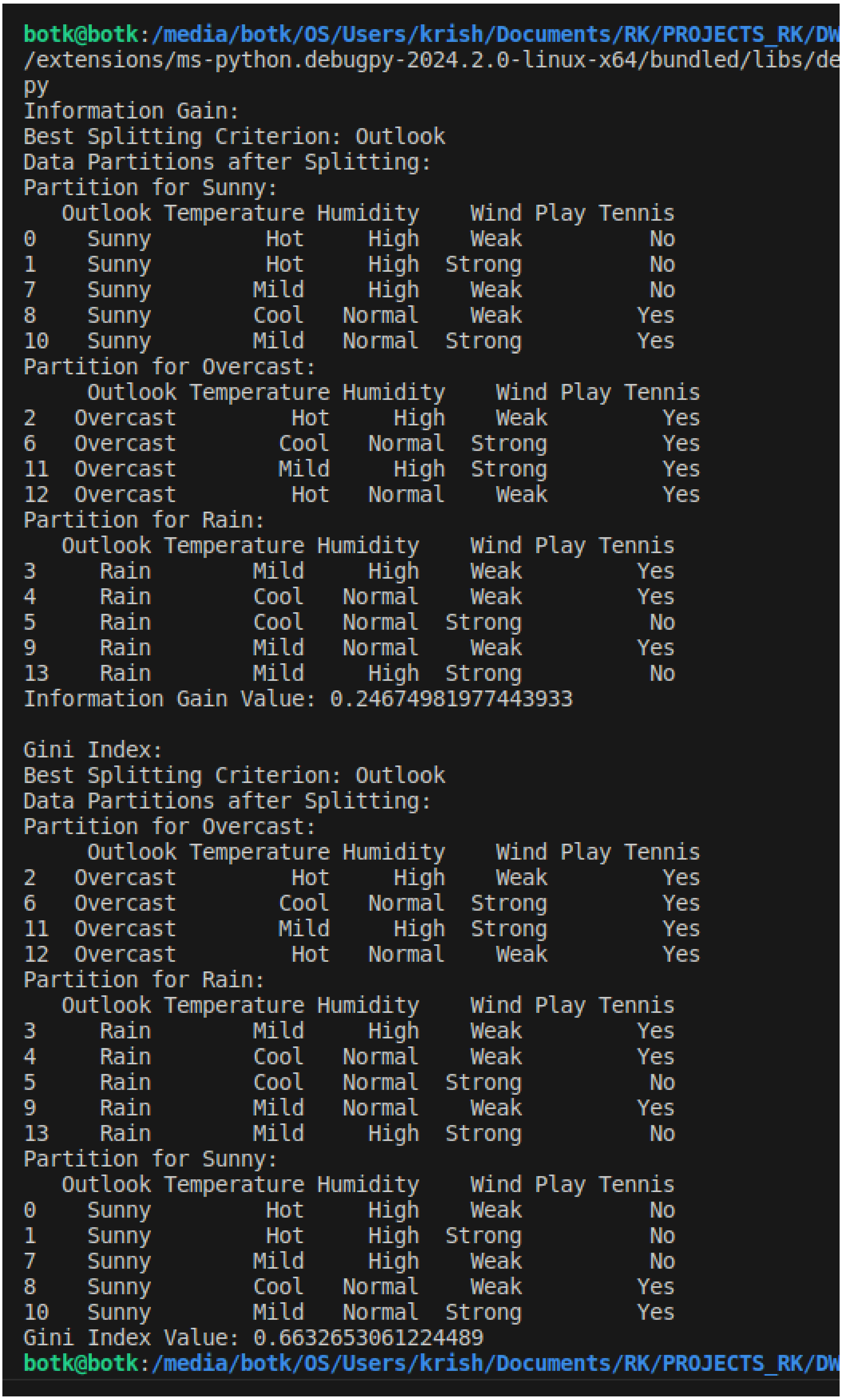
CODE

FOR tennis.csv

```
# Modify the script to use the provided dataset

# Importing necessary libraries
import pandas as pd
import numpy as np

# Function to calculate entropy
def entropy(labels):
"""Calculate the entropy of a list of labels."""
unique_labels, counts = np.unique(labels, return_counts=True)
probabilities = counts / len(labels)
entropy_value = -np.sum(probabilities * np.log2(probabilities))
return entropy_value

# Function to calculate information gain
def information_gain(data, split_attribute_name, target_name):
"""Calculate the information gain for a given split attribute."""
total_entropy = entropy(data[target_name])
values, counts = np.unique(data[split_attribute_name], return_counts=True)
weighted_entropy = np.sum([(counts[i] / np.sum(counts)) *
entropy(data.where(data[split_attribute_name] == values[i]).dropna()
[target_name])
for i in range(len(values))])
information_gain_value = total_entropy - weighted_entropy
return information_gain_value

# Function to calculate Gini index
def gini_index(labels):
"""Calculate the Gini index of a list of labels."""
unique_labels, counts = np.unique(labels, return_counts=True)
probabilities = counts / len(labels)
gini_index_value = 1 - np.sum(probabilities**2)
return gini_index_value

# Function to find the best splitting criterion
def find_best_split(data, target_name, measure):
"""Find the best splitting criterion based on the specified measure."""
best_measure_value = 0
best_split_attribute = None
partitions = None
```

```
for column in data.columns[:-1]:
if measure == 'Information Gain':
current_measure_value = information_gain(data, column, target_name)
elif measure == 'Gini Index':
current_measure_value = gini_index(data[column])

if current_measure_value > best_measure_value:
best_measure_value = current_measure_value
best_split_attribute = column

if measure == 'Information Gain':
partitions = {value: data[data[best_split_attribute] == value] for value in
data[best_split_attribute].unique()}
elif measure == 'Gini Index':
partitions = {value: data[data[best_split_attribute] == value] for value in
np.unique(data[best_split_attribute])}

return best_split_attribute, partitions, best_measure_value

# Load the tennis dataset
tennis_data = pd.DataFrame({
'Outlook': ['Sunny', 'Sunny', 'Overcast', 'Rain', 'Rain', 'Rain',
'Overcast', 'Sunny', 'Sunny', 'Rain', 'Sunny', 'Overcast', 'Overcast',
'Rain'],
'Temperature': ['Hot', 'Hot', 'Hot', 'Mild', 'Cool', 'Cool', 'Cool', 'Mild',
'Cool', 'Mild', 'Mild', 'Mild', 'Hot', 'Mild'],
'Humidity': ['High', 'High', 'High', 'High', 'Normal', 'Normal', 'Normal',
'High', 'Normal', 'Normal', 'Normal', 'High', 'Normal', 'High'],
'Wind': ['Weak', 'Strong', 'Weak', 'Weak', 'Weak', 'Strong', 'Strong',
'Weak', 'Weak', 'Weak', 'Strong', 'Strong', 'Weak', 'Strong'],
'Play Tennis': ['No', 'No', 'Yes', 'Yes', 'Yes', 'No', 'Yes', 'No', 'Yes',
'Yes', 'Yes', 'Yes', 'Yes', 'No']
})

# Example usage with tennis.csv
print("Information Gain:")
split_criteria, data_partitions, measure_value =
find_best_split(tennis_data, 'Play Tennis', 'Information Gain')
print("Best Splitting Criterion:", split_criteria)
print("Data Partitions after Splitting:")
for value, partition in data_partitions.items():
print("Partition for {}: \n{}".format(value, partition))
print("Information Gain Value:", measure_value)

print("\nGini Index:")
split_criteria, data_partitions, measure_value =
find_best_split(tennis_data, 'Play Tennis', 'Gini Index')
print("Best Splitting Criterion:", split_criteria)
print("Data Partitions after Splitting:")
for value, partition in data_partitions.items():
print("Partition for {}: \n{}".format(value, partition))
print("Gini Index Value:", measure_value)
```

OUTPUT



```
botk@botk:/media/botk/OS/Users/krish/Documents/RK/PROJECTS_RK/DW
/extensions/ms-python.debugpy-2024.2.0-linux-x64/bundled/libs/de
py
Information Gain:
Best Splitting Criterion: Outlook
Data Partitions after Splitting:
Partition for Sunny:
   Outlook Temperature Humidity    Wind Play Tennis
0    Sunny         Hot     High    Weak          No
1    Sunny         Hot     High  Strong          No
7    Sunny        Mild     High    Weak          No
8    Sunny        Cool   Normal    Weak         Yes
10   Sunny        Mild   Normal  Strong         Yes
Partition for Overcast:
     Outlook Temperature Humidity    Wind Play Tennis
2   Overcast         Hot     High    Weak         Yes
6   Overcast        Cool   Normal  Strong         Yes
11  Overcast        Mild     High  Strong         Yes
12  Overcast         Hot   Normal    Weak         Yes
Partition for Rain:
   Outlook Temperature Humidity    Wind Play Tennis
3     Rain        Mild     High    Weak         Yes
4     Rain        Cool   Normal    Weak         Yes
5     Rain        Cool   Normal  Strong          No
9     Rain        Mild   Normal    Weak         Yes
13    Rain        Mild     High  Strong          No
Information Gain Value: 0.24674981977443933

Gini Index:
Best Splitting Criterion: Outlook
Data Partitions after Splitting:
Partition for Overcast:
     Outlook Temperature Humidity    Wind Play Tennis
2   Overcast         Hot     High    Weak         Yes
6   Overcast        Cool   Normal  Strong         Yes
11  Overcast        Mild     High  Strong         Yes
12  Overcast         Hot   Normal    Weak         Yes
Partition for Rain:
   Outlook Temperature Humidity    Wind Play Tennis
3     Rain        Mild     High    Weak         Yes
4     Rain        Cool   Normal    Weak         Yes
5     Rain        Cool   Normal  Strong          No
9     Rain        Mild   Normal    Weak         Yes
13    Rain        Mild     High  Strong          No
Partition for Sunny:
   Outlook Temperature Humidity    Wind Play Tennis
0    Sunny         Hot     High    Weak          No
1    Sunny         Hot     High  Strong          No
7    Sunny        Mild     High    Weak          No
8    Sunny        Cool   Normal    Weak         Yes
10   Sunny        Mild   Normal  Strong         Yes
Gini Index Value: 0.6632653061224489
botk@botk:/media/botk/OS/Users/krish/Documents/RK/PROJECTS_RK/DW
```

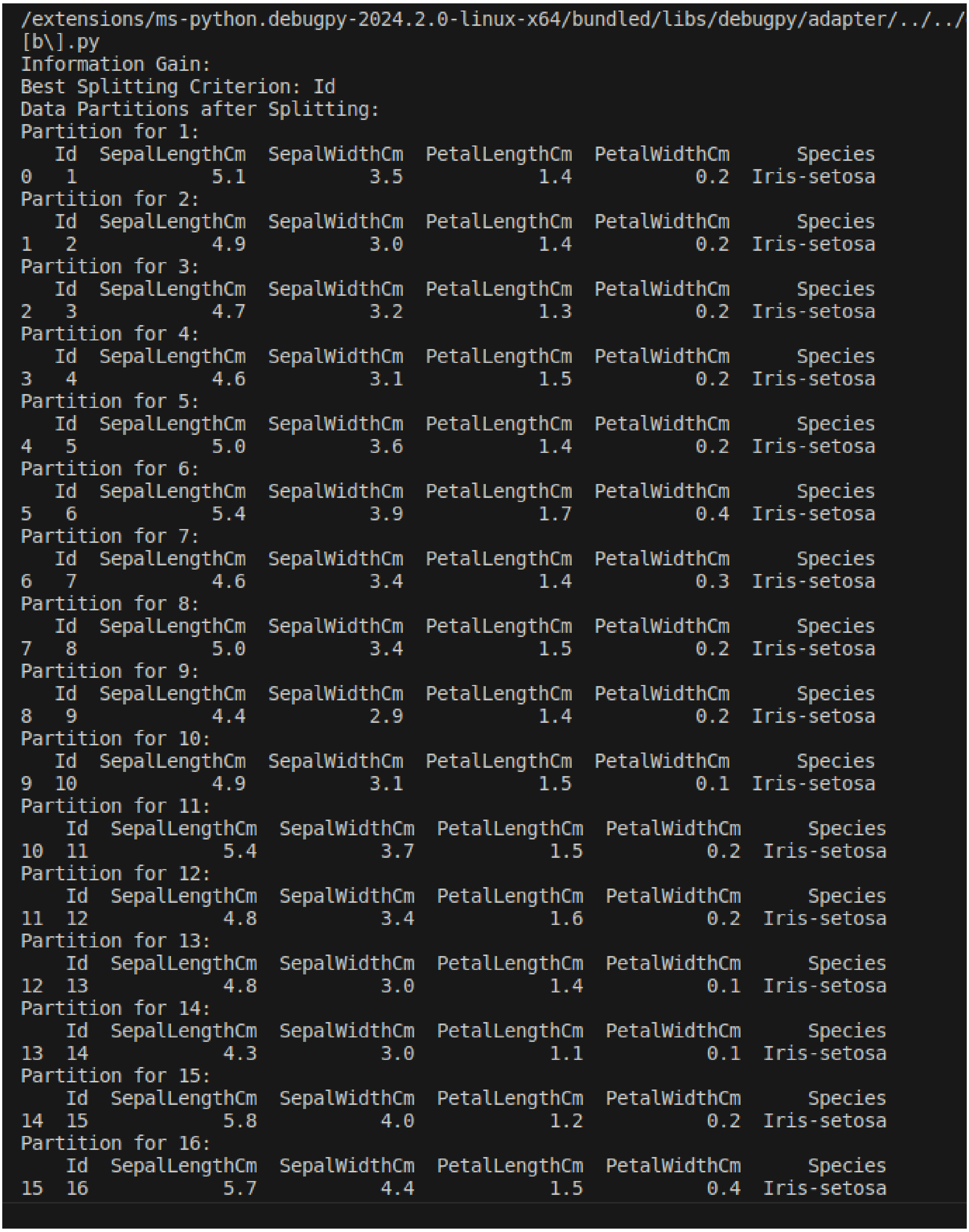
FOR iris.csv

CODE

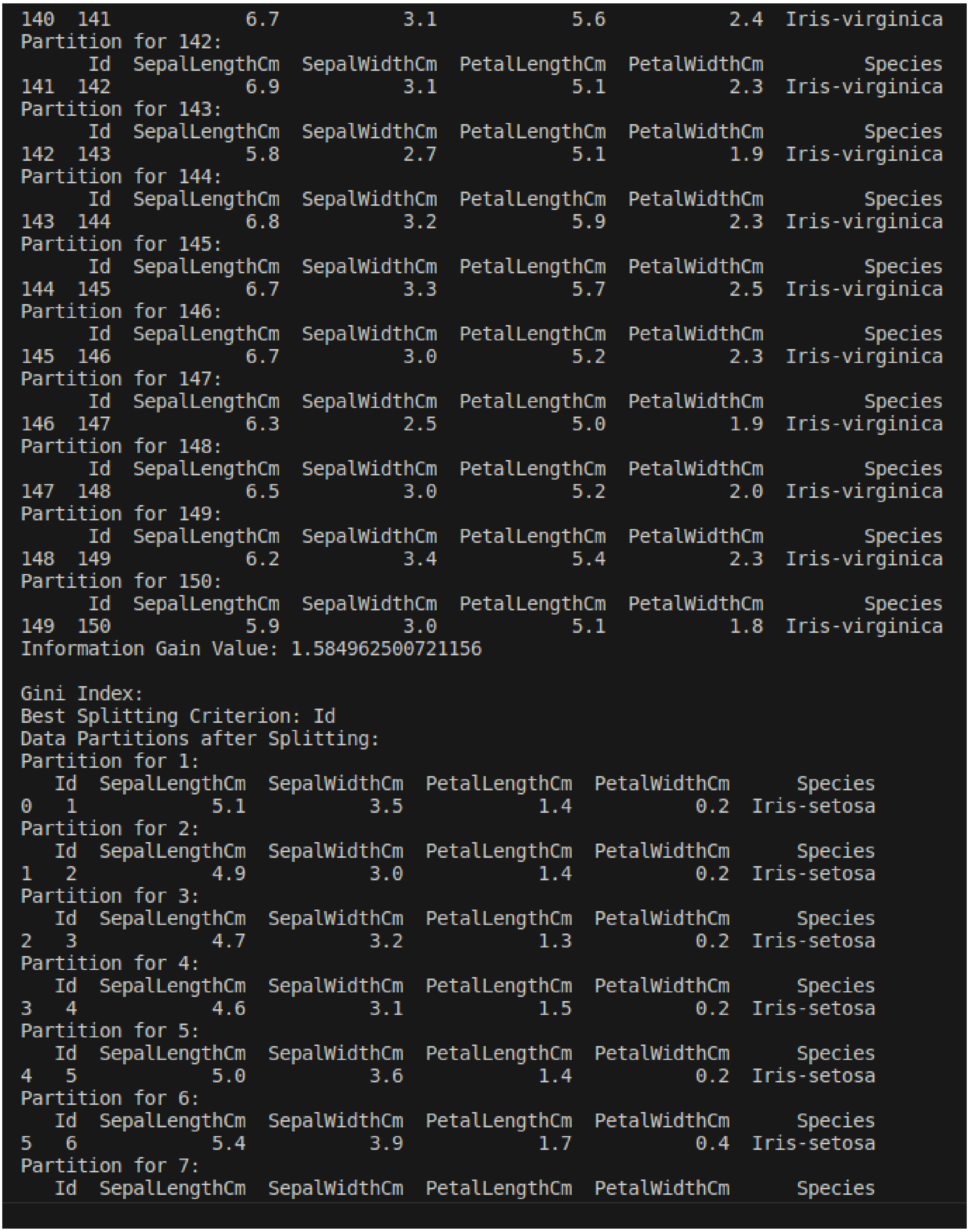
```
# Importing necessary libraries
import pandas as pd
import numpy as np

# Function to calculate entropy
def entropy(labels):
"""Calculate the entropy of a list of labels."""
unique_labels, counts = np.unique(labels, return_counts=True)
probabilities = counts / len(labels)
entropy_value = -np.sum(probabilities * np.log2(probabilities))
return entropy_value

# Function to calculate information gain
def information_gain(data, split_attribute_name, target_name):
"""Calculate the information gain for a given split attribute."""
total_entropy = entropy(data[target_name])
values, counts = np.unique(data[split_attribute_name], return_counts=True)
weighted_entropy = np.sum([(counts[i] / np.sum(counts)) *
entropy(data.where(data[split_attribute_name] == values[i]).dropna()
[target_name])
for i in range(len(values))])
information_gain_value = total_entropy - weighted_entropy
return information_gain_value

# Function to calculate Gini index
def gini_index(labels):
"""Calculate the Gini index of a list of labels."""
unique_labels, counts = np.unique(labels, return_counts=True)
probabilities = counts / len(labels)
gini_index_value = 1 - np.sum(probabilities**2)
return gini_index_value

# Function to find the best splitting criterion
def find_best_split(data, target_name, measure):
"""Find the best splitting criterion based on the specified measure."""
best_measure_value = 0
best_split_attribute = None
partitions = None

for column in data.columns[:-1]:
if measure == 'Information Gain':
current_measure_value = information_gain(data, column, target_name)
elif measure == 'Gini Index':
current_measure_value = gini_index(data[column])
```

```
if current_measure_value > best_measure_value:
best_measure_value = current_measure_value
best_split_attribute = column

if measure == 'Information Gain':




partitions = {value: data[data[best_split_attribute] == value] for value in
data[best_split_attribute].unique()}
elif measure == 'Gini Index':
partitions = {value: data[data[best_split_attribute] == value] for value in
np.unique(data[best_split_attribute])}

return best_split_attribute, partitions, best_measure_value

# Load the iris dataset
iris_data = pd.read_csv('iris.csv')

# Example usage with iris.csv
print("Information Gain:")
split_criteria, data_partitions, measure_value = find_best_split(iris_data,
'Species', 'Information Gain')
print("Best Splitting Criterion:", split_criteria)
print("Data Partitions after Splitting:")
for value, partition in data_partitions.items():
print("Partition for {}: \n{}".format(value, partition))
print("Information Gain Value:", measure_value)

print("\nGini Index:")
split_criteria, data_partitions, measure_value = find_best_split(iris_data,
'Species', 'Gini Index')
print("Best Splitting Criterion:", split_criteria)
print("Data Partitions after Splitting:")
for value, partition in data_partitions.items():
print("Partition for {}: \n{}".format(value, partition))
print("Gini Index Value:", measure_value)
```

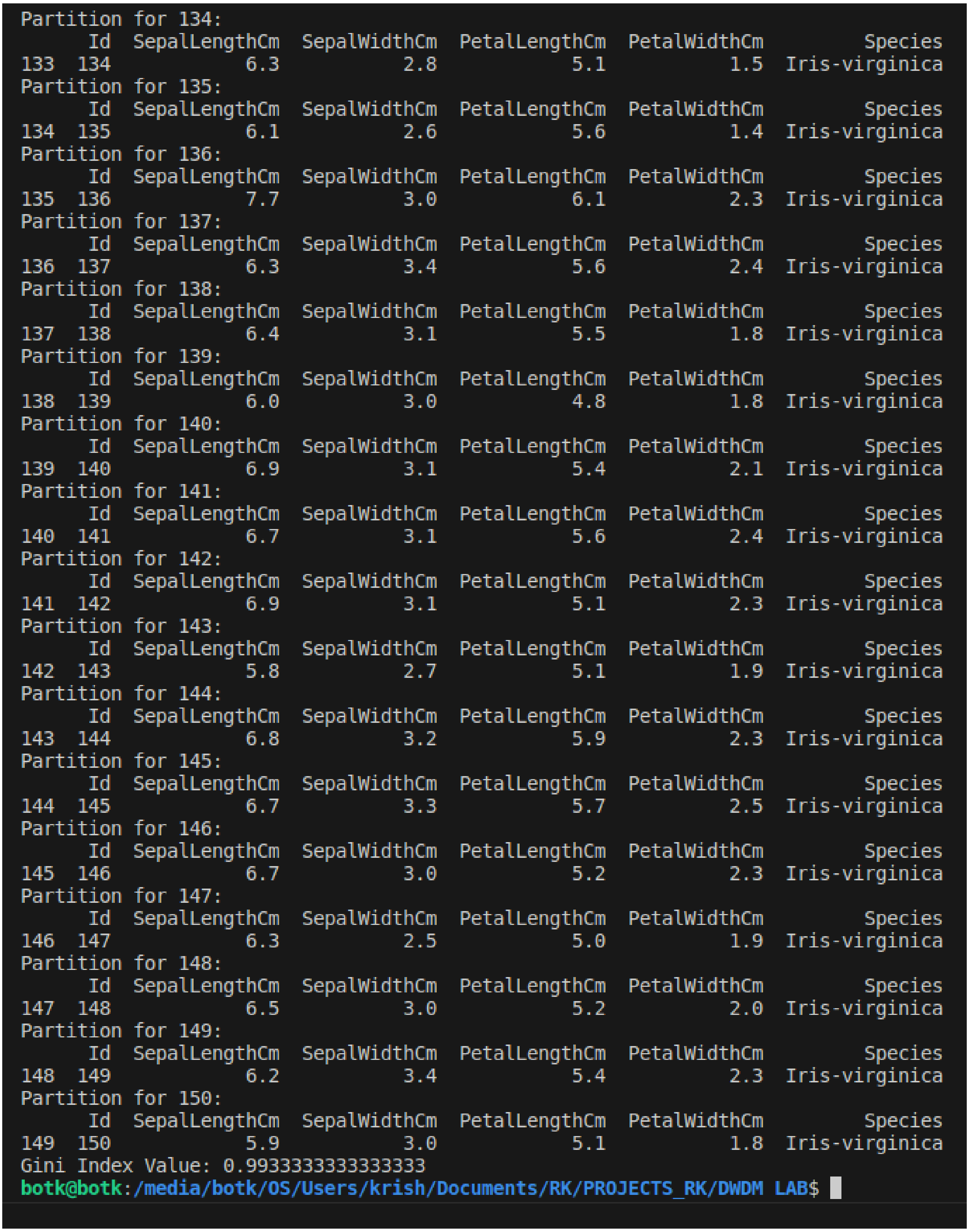
OUTPUT



```
/extensions/ms-python.debugpy-2024.2.0-linux-x64/bundled/libs/debugpy/adapter/../../
[b\].py
Information Gain:
Best Splitting Criterion: Id
Data Partitions after Splitting:
Partition for 1:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
0   1            5.1           3.5            1.4           0.2  Iris-setosa
Partition for 2:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
1   2            4.9           3.0            1.4           0.2  Iris-setosa
Partition for 3:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
2   3            4.7           3.2            1.3           0.2  Iris-setosa
Partition for 4:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
3   4            4.6           3.1            1.5           0.2  Iris-setosa
Partition for 5:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
4   5            5.0           3.6            1.4           0.2  Iris-setosa
Partition for 6:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
5   6            5.4           3.9            1.7           0.4  Iris-setosa
Partition for 7:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
6   7            4.6           3.4            1.4           0.3  Iris-setosa
Partition for 8:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
7   8            5.0           3.4            1.5           0.2  Iris-setosa
Partition for 9:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
8   9            4.4           2.9            1.4           0.2  Iris-setosa
Partition for 10:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
9  10            4.9           3.1            1.5           0.1  Iris-setosa
Partition for 11:
    Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
10  11            5.4           3.7            1.5           0.2  Iris-setosa
Partition for 12:
    Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
11  12            4.8           3.4            1.6           0.2  Iris-setosa
Partition for 13:
    Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
12  13            4.8           3.0            1.4           0.1  Iris-setosa
Partition for 14:
    Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
13  14            4.3           3.0            1.1           0.1  Iris-setosa
Partition for 15:
    Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
14  15            5.8           4.0            1.2           0.2  Iris-setosa
Partition for 16:
    Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
15  16            5.7           4.4            1.5           0.4  Iris-setosa
```



```
140  141            6.7           3.1            5.6           2.4  Iris-virginica
Partition for 142:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
141  142            6.9           3.1            5.1           2.3  Iris-virginica
Partition for 143:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
142  143            5.8           2.7            5.1           1.9  Iris-virginica
Partition for 144:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
143  144            6.8           3.2            5.9           2.3  Iris-virginica
Partition for 145:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
144  145            6.7           3.3            5.7           2.5  Iris-virginica
Partition for 146:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
145  146            6.7           3.0            5.2           2.3  Iris-virginica
Partition for 147:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
146  147            6.3           2.5            5.0           1.9  Iris-virginica
Partition for 148:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
147  148            6.5           3.0            5.2           2.0  Iris-virginica
Partition for 149:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
148  149            6.2           3.4            5.4           2.3  Iris-virginica
Partition for 150:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
149  150            5.9           3.0            5.1           1.8  Iris-virginica
Information Gain Value: 1.584962500721156

Gini Index:
Best Splitting Criterion: Id
Data Partitions after Splitting:
Partition for 1:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
0   1            5.1           3.5            1.4           0.2  Iris-setosa
Partition for 2:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
1   2            4.9           3.0            1.4           0.2  Iris-setosa
Partition for 3:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
2   3            4.7           3.2            1.3           0.2  Iris-setosa
Partition for 4:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
3   4            4.6           3.1            1.5           0.2  Iris-setosa
Partition for 5:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
4   5            5.0           3.6            1.4           0.2  Iris-setosa
Partition for 6:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
5   6            5.4           3.9            1.7           0.4  Iris-setosa
Partition for 7:
   Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm      Species
```



```
Partition for 134:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
133  134            6.3           2.8            5.1           1.5  Iris-virginica
Partition for 135:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
134  135            6.1           2.6            5.6           1.4  Iris-virginica
Partition for 136:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
135  136            7.7           3.0            6.1           2.3  Iris-virginica
Partition for 137:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
136  137            6.3           3.4            5.6           2.4  Iris-virginica
Partition for 138:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
137  138            6.4           3.1            5.5           1.8  Iris-virginica
Partition for 139:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
138  139            6.0           3.0            4.8           1.8  Iris-virginica
Partition for 140:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
139  140            6.9           3.1            5.4           2.1  Iris-virginica
Partition for 141:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
140  141            6.7           3.1            5.6           2.4  Iris-virginica
Partition for 142:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
141  142            6.9           3.1            5.1           2.3  Iris-virginica
Partition for 143:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
142  143            5.8           2.7            5.1           1.9  Iris-virginica
Partition for 144:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
143  144            6.8           3.2            5.9           2.3  Iris-virginica
Partition for 145:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
144  145            6.7           3.3            5.7           2.5  Iris-virginica
Partition for 146:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
145  146            6.7           3.0            5.2           2.3  Iris-virginica
Partition for 147:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
146  147            6.3           2.5            5.0           1.9  Iris-virginica
Partition for 148:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
147  148            6.5           3.0            5.2           2.0  Iris-virginica
Partition for 149:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
148  149            6.2           3.4            5.4           2.3  Iris-virginica
Partition for 150:
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm         Species
149  150            5.9           3.0            5.1           1.8  Iris-virginica
Gini Index Value: 0.9933333333333333
botk@botk:/media/botk/OS/Users/krish/Documents/RK/PROJECTS_RK/DWDM LAB$ 
```