Surabhi Singh

The Sparks Foundation

Decision Tree Classification

For the given 'Iris' dataset, create the Decision Tree classifier and visulize it graphically

Python libraries being used:

A:Data Analysis

In [1]:

```
import pandas as pd
import numpy as np
```

B: Data Visualization

In [2]:

```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:

```
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn import tree
from sklearn import metrics
```

Data Extraction

Importing data from our Task File

In [4]:

```
from sklearn import datasets
```

In [6]:

```
iris_data = datasets.load_iris()
iris = pd.DataFrame(iris_data['data'],columns=['sepal_lengthCm', 'sepal_widthCm', 'petal
_lengthCm', 'petal_widthCm'])
```

In [7]:

```
iris['Species']=iris_data['target']
iris['Species']=iris['Species'].apply(lambda x: iris_data['target_names'][x])
iris.head()
```

Out[7]:

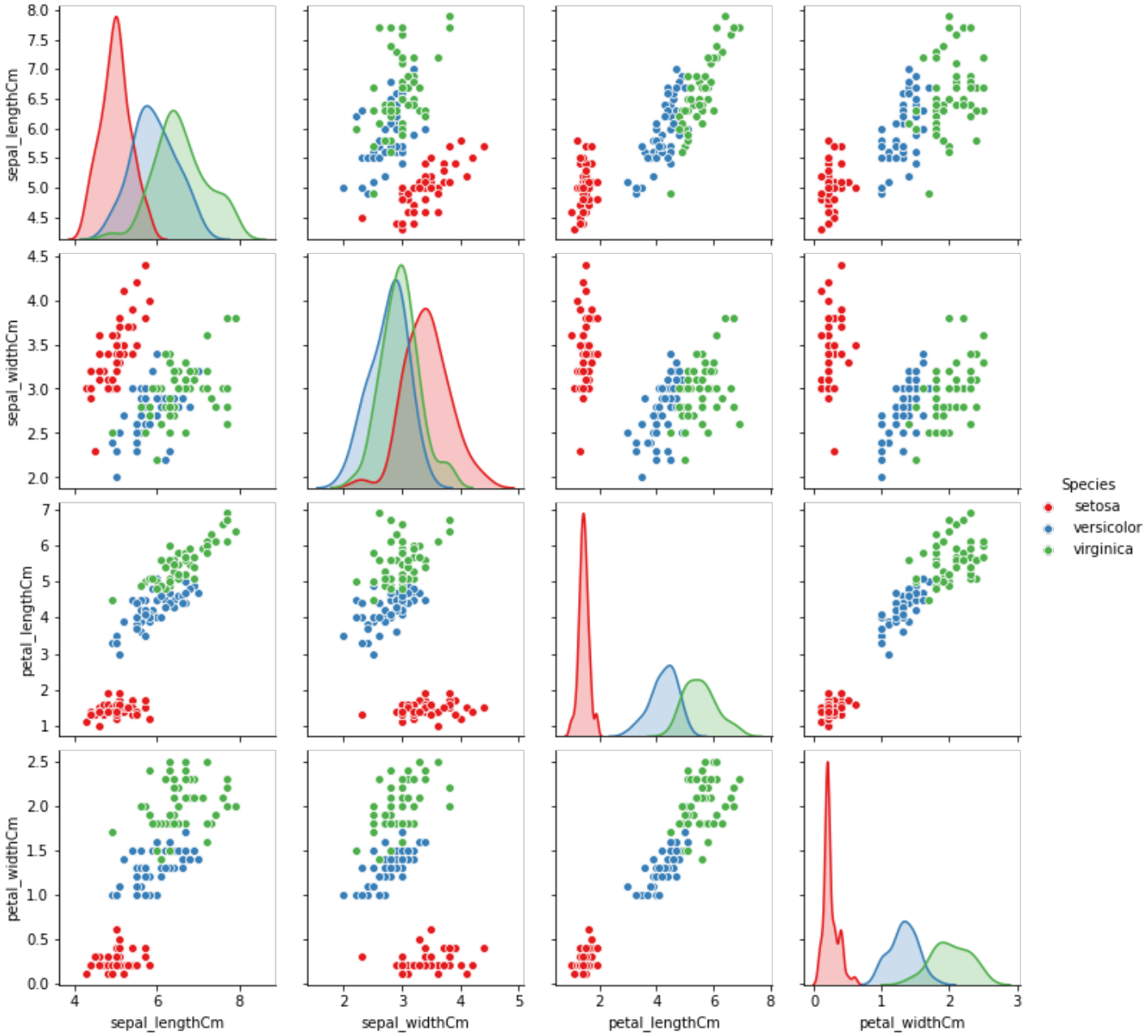
	sepal_lengthCm	sepal_widthCm	petal_lengthCm	petal_widthCm	Species
0	5.1	3.5	1.4	0.2	setosa

	sepal_lengthCm	sepal_widthCm	petal_lengthCm	petal_widthCm	Species
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

Data Visualization

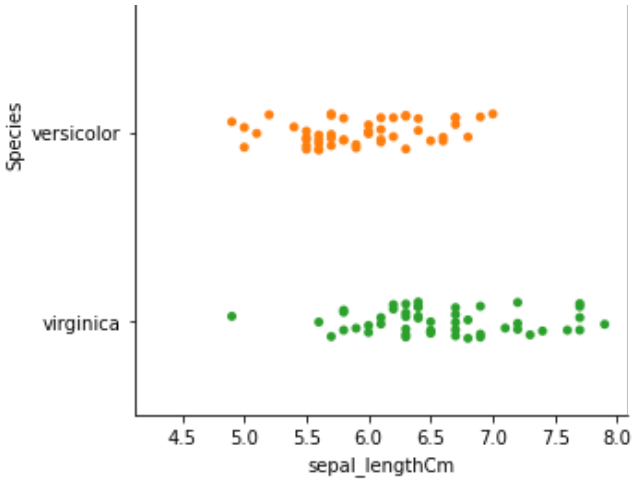
In [8]:

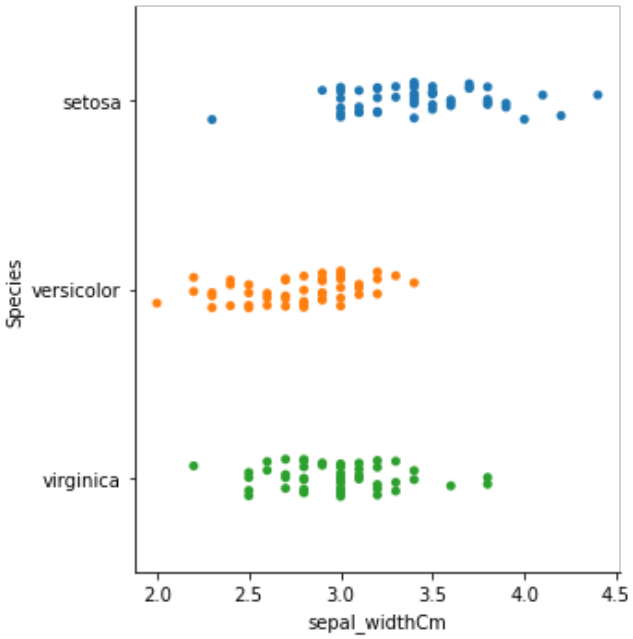
```
sns.pairplot(iris, hue = 'Species', palette="Set1")
plt.show()
```

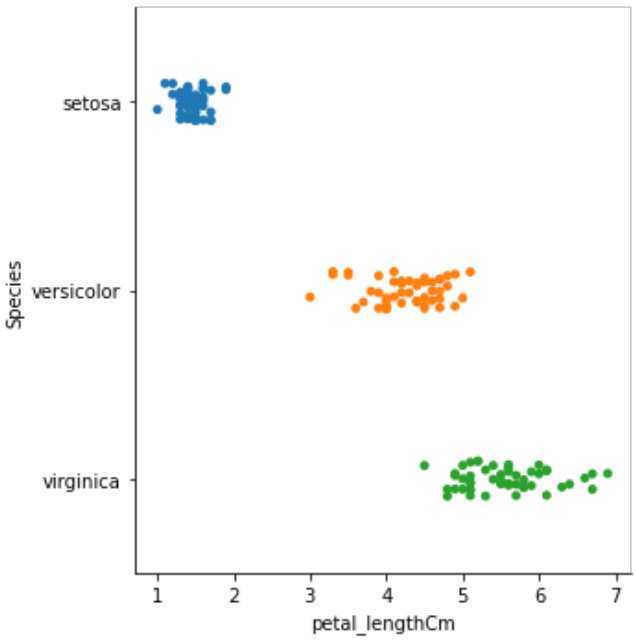


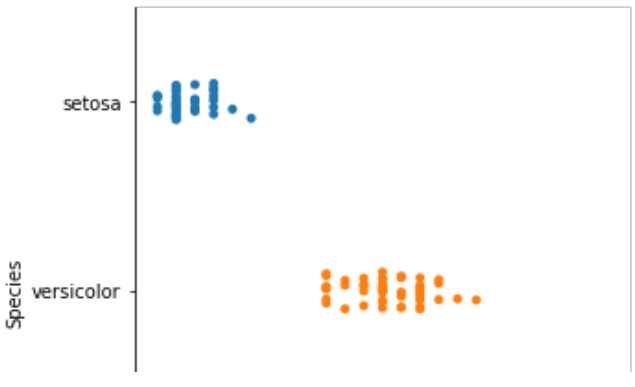
In [10]:

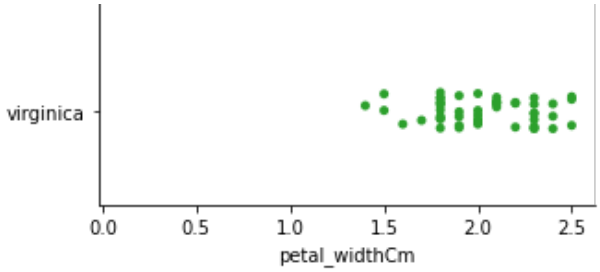
```
d = ["sepal_lengthCm", "sepal_widthCm", "petal_lengthCm", "petal_widthCm"]
for d in d:
    sns.catplot(data=iris, x=d, y="Species")
    plt.show()
```


Species
versicolor
virginica
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
sepal_lengthCm


Species
setosa
versicolor
virginica
2.0
2.5
3.0
3.5
4.0
4.5
sepal_widthCm


Species
setosa
versicolor
virginica
1
2
3
4
5
6
7
petal_lengthCm


Species
setosa
versicolor



Here, iris-setosa species is clearly separable from the other species

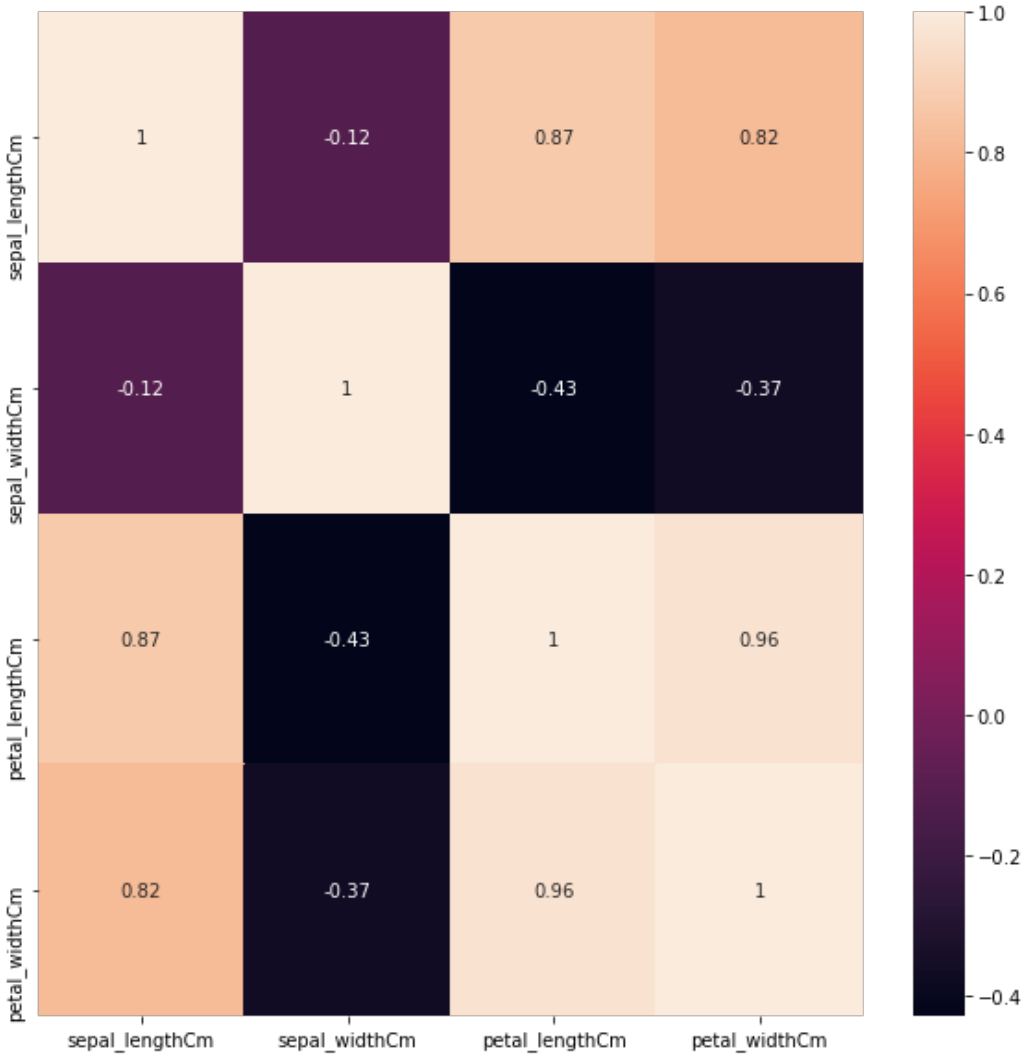
Correlation Plot

In [11]:

```
plt.figure(figsize=(10,10))
sns.heatmap(iris.corr(),annot=True)
plt.plot()
```

Out[11]:

[]



Here, Petal length is highly related to petal width and Sepal lenth is not related to sepal width

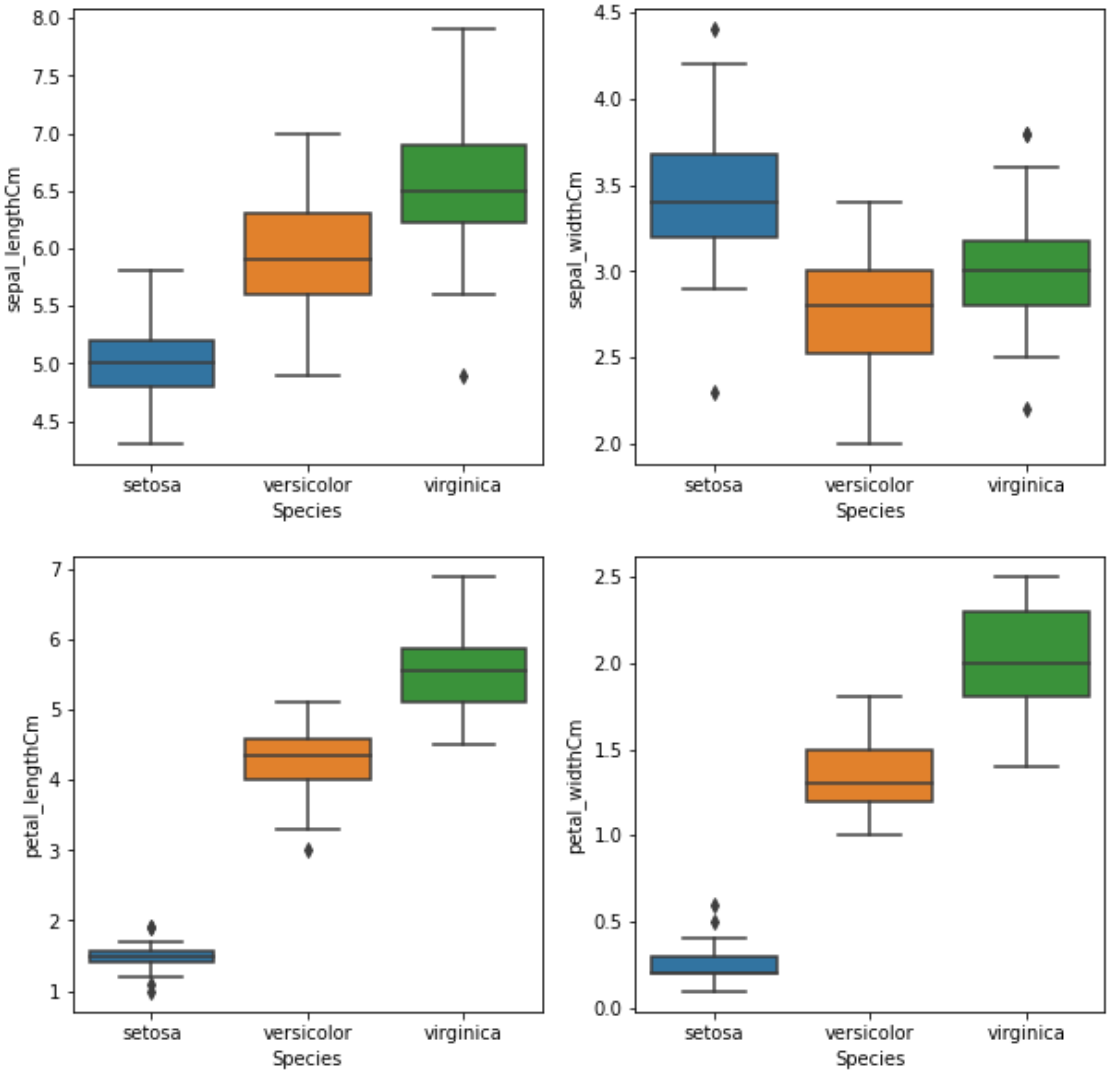
Categorical Distribution of Species using BoxPlot

In [12]:

```
plt.figure(figsize=(10,10))
plt.subplot(2,2,1)
sns.boxplot(data=iris, x="Species",y="sepal_lengthCm")
plt.subplot(2,2,2)
sns.boxplot(data=iris, x="Species",y="sepal_widthCm")
plt.subplot(2,2,3)
sns.boxplot(data=iris, x="Species",y="petal_lengthCm")
plt.subplot(2,2,4)
sns.boxplot(data=iris, x="Species",y="petal_widthCm")
```

Out[12]:

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f8407933a10>
```



Data Modelling

Splitting data into Train and Test sets

In [13]:

```
train,test = train_test_split(iris,test_size=0.3)
```

Training

In [14]:

```
train_X = train[["sepal_lengthCm", "sepal_widthCm", "petal_lengthCm", "petal_widthCm"]]
train_y = train.Species
```

Testing

In [16]:

```
test_X=test[["sepal_lengthCm", "sepal_widthCm", "petal_lengthCm", "petal_widthCm"]]
test_y=test.Species
```

In [17]:

```
dectree = DecisionTreeClassifier(random_state=12)
model = dectree.fit(train_X, train_y)
```

In [18]:

```
dectree.predict(test_X)
```

Out[18]:

```
array(['versicolor', 'virginica', 'versicolor', 'virginica', 'versicolor',
       'setosa', 'virginica', 'setosa', 'versicolor', 'virginica',
       'virginica', 'virginica', 'virginica', 'virginica', 'setosa',
       'versicolor', 'virginica', 'setosa', 'virginica', 'virginica',
       'setosa', 'versicolor', 'virginica', 'setosa', 'versicolor',
       'setosa', 'virginica', 'setosa', 'virginica', 'setosa',
       'virginica', 'versicolor', 'setosa', 'setosa', 'virginica',
       'setosa', 'versicolor', 'setosa', 'setosa', 'virginica', 'setosa',
       'setosa', 'virginica', 'versicolor', 'setosa'], dtype=object)
```

Evaluating Accuracy

In [19]:

```
dectree.score(test_X, test_y)
```

Out[19]:

```
0.9777777777777777
```

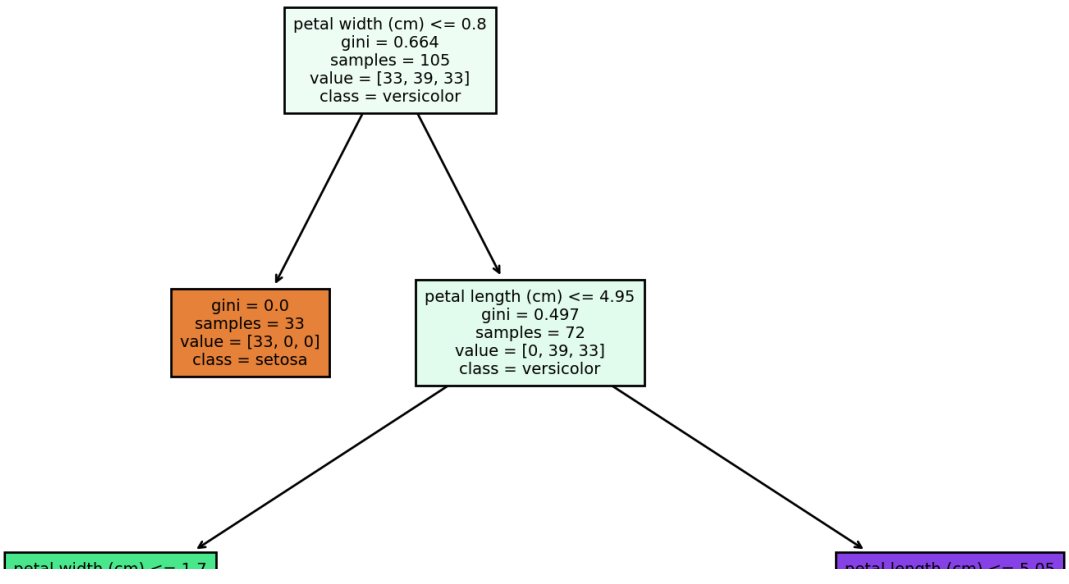
In [20]:

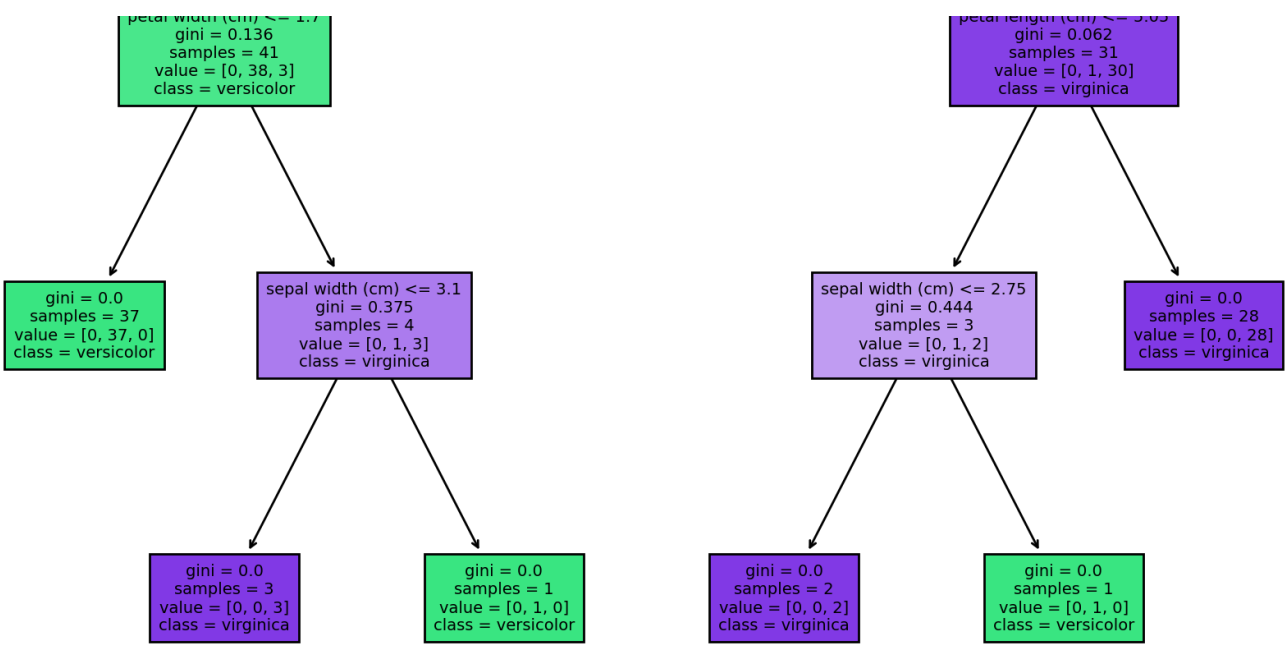
```
y_pred = dectree.predict(test_X)
print('Accuracy Score:', metrics.accuracy_score(test_y, y_pred))
```

```
Accuracy Score: 0.9777777777777777
```

In [21]:

```
fig, axes = plt.subplots(nrows = 1,ncols =1,figsize = (11,11), dpi=200)
_ = tree.plot_tree(dectree,feature_names=iris_data.feature_names,class_names=iris_data.ta
rget_names,filled=True,fontsize=7)
```




petal width (cm) <= 1.7
gini = 0.136
samples = 41
value = [0, 38, 3]
class = versicolor
petal length (cm) <= 5.05
gini = 0.062
samples = 31
value = [0, 1, 30]
class = virginica
gini = 0.0
samples = 37
value = [0, 37, 0]
class = versicolor
sepal width (cm) <= 3.1
gini = 0.375
samples = 4
value = [0, 1, 3]
class = virginica
sepal width (cm) <= 2.75
gini = 0.444
samples = 3
value = [0, 1, 2]
class = virginica
gini = 0.0
samples = 28
value = [0, 0, 28]
class = virginica
gini = 0.0
samples = 3
value = [0, 0, 3]
class = virginica
gini = 0.0
samples = 1
value = [0, 1, 0]
class = versicolor
gini = 0.0
samples = 2
value = [0, 0, 2]
class = virginica
gini = 0.0
samples = 1
value = [0, 1, 0]
class = versicolor

In [ ]: