Out[1]: The raw code for this IPython notebook is by default hidden for easier reading. To toggle on/off the raw code, click here.

Exercice 2

Prérequis

- Scipy
- Numpy
- Scikit learn
- Pandas

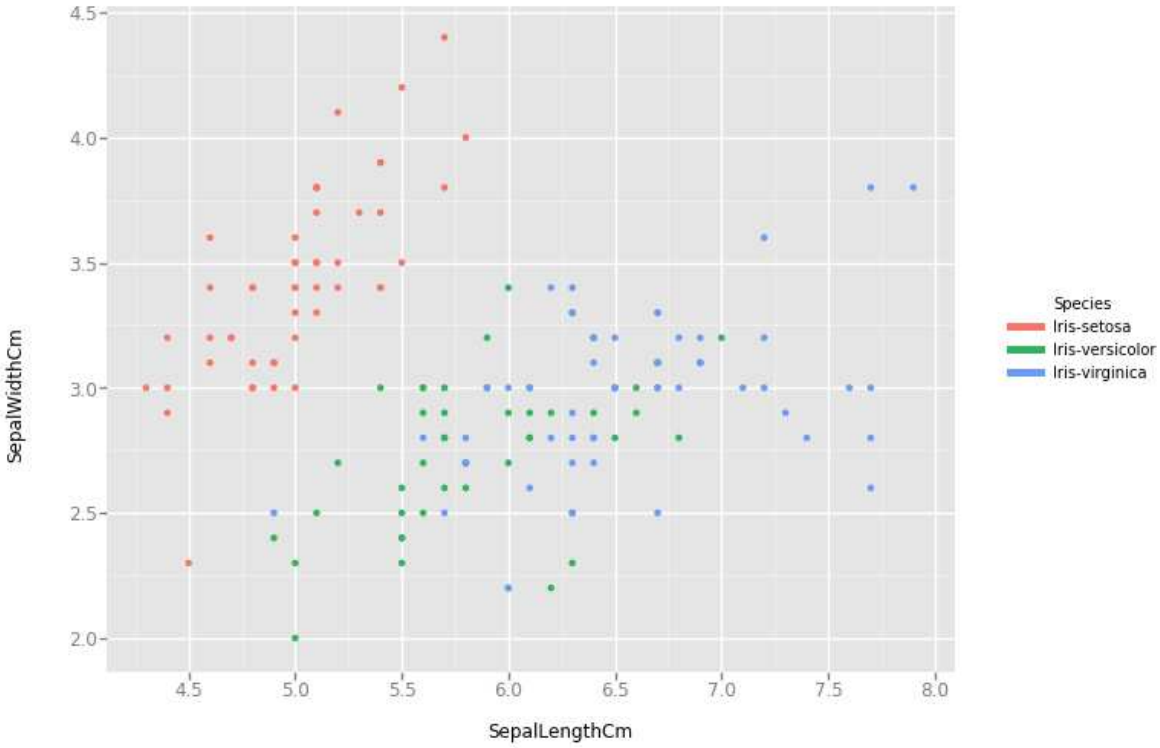
Problème

Dans ce premier exercice, analyser les données sur les fleurs. Cela consiste à essayer de prédire (détecter) le n'espèce d'une fleur en fonction de ses différentes caractéristiques. Nous allons dans un premier temps utiliser les arbres de décision. Nous allons utiliser le package arbre de décision de scikit-learn, la documentation se trouve à cet endroit (https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html). Analysez les paramètres du module et essayer de comprendre ce que chaque paramètre fait.

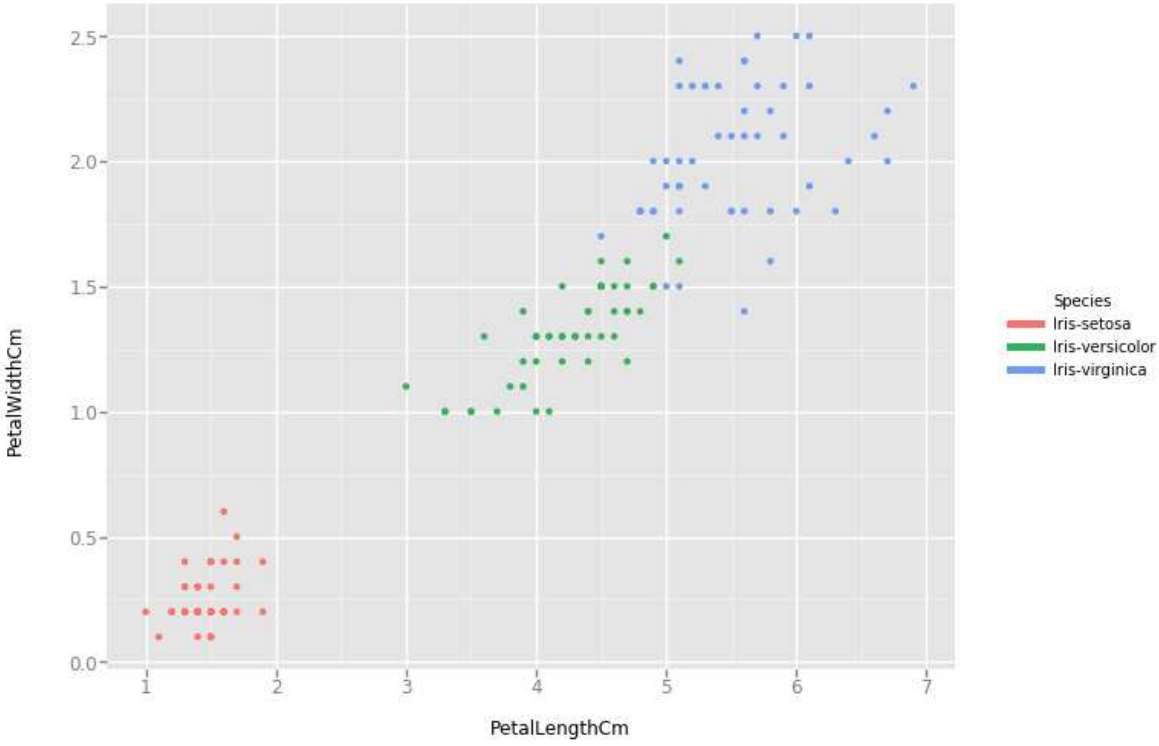
- Commencez par charger les données sur un dataframe et analyser leur contenu.
- Essayer de visualiser ces données en utilisant différents graphes.
- Choisissez une variable explicative, et à l'aide la librairie DecisionTreeClassifier essayer de construire un modèle capable de prédire l'espèce de la fleur à l'aide de cette variable.
- Graphviz est un outil de construction et de visualisation de graphe. La librairie tree permet d'exporter un arbre de décision en un graphe de type graphviz à l'aide la fonction export_graphviz (https://scikit-learn.org/stable/modules/generated/sklearn.tree.export_graphviz.html#sklearn.tree.export_graphviz). Exportez le graphe résultant de votre analyse, et visualisez-le.
- A quelle moment fallait il s'arrêter de splitter ? Pourquoi ? Refaire la classification en configurant celà.
- Refaire la démarche sur d'autres variables.
- Quelle est la variable qui explique le mieux l'espèce ?
- Peut-on arriver aux mêmes conclusions à priori (en se limitant à l'étape 1 et 2 de cet exercice) ?
- Refaire la classification en prenant en compte d'autres variables.

Out[2]:

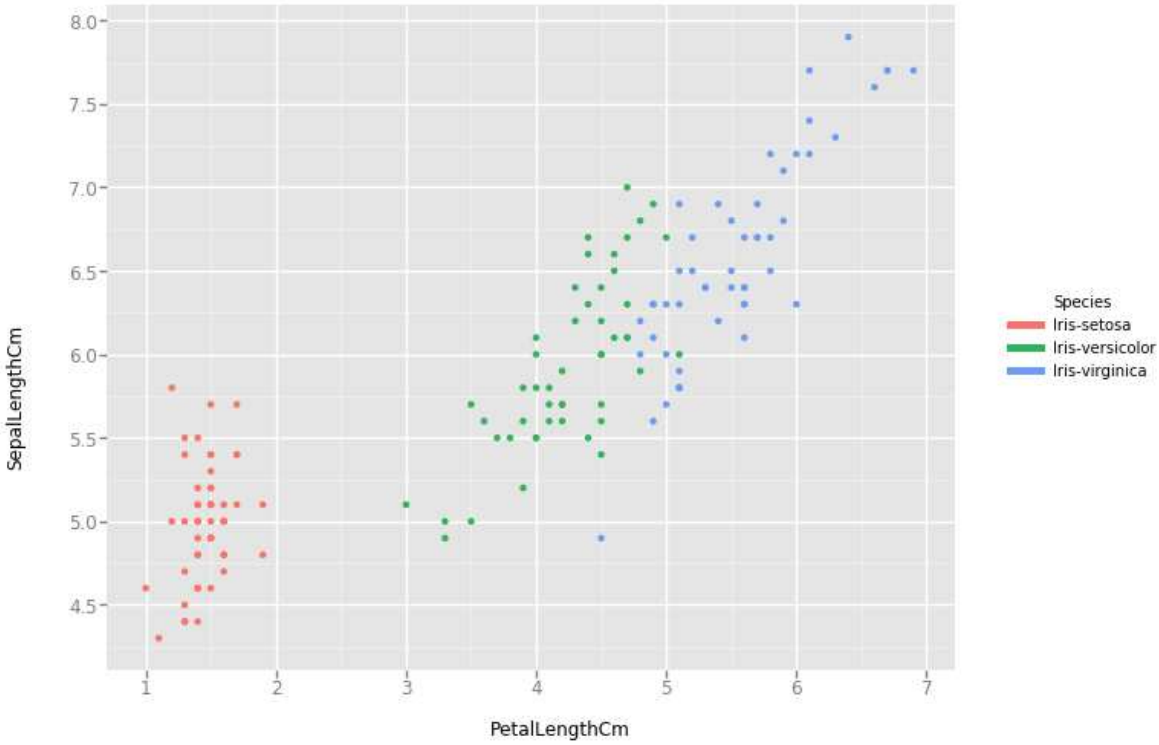
	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa



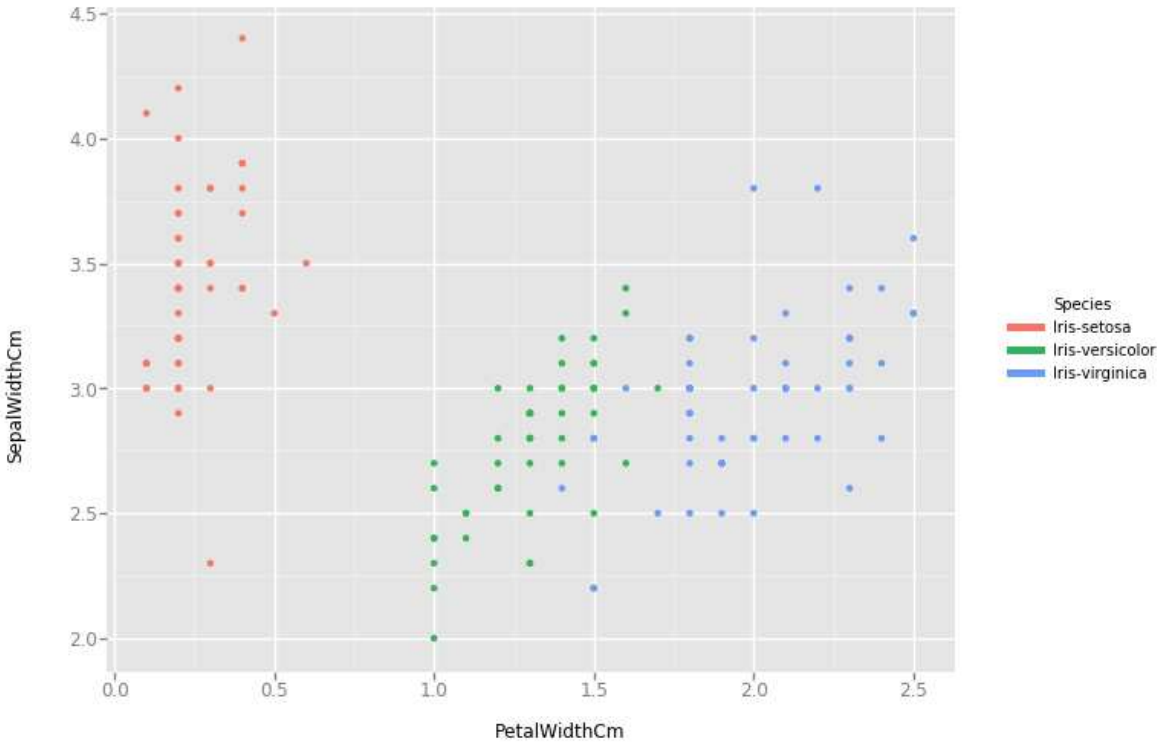

```
Out[20]: <ggplot: (10362281)>
```




```
Out[19]: <ggplot: (-9223372036844432124)>
```



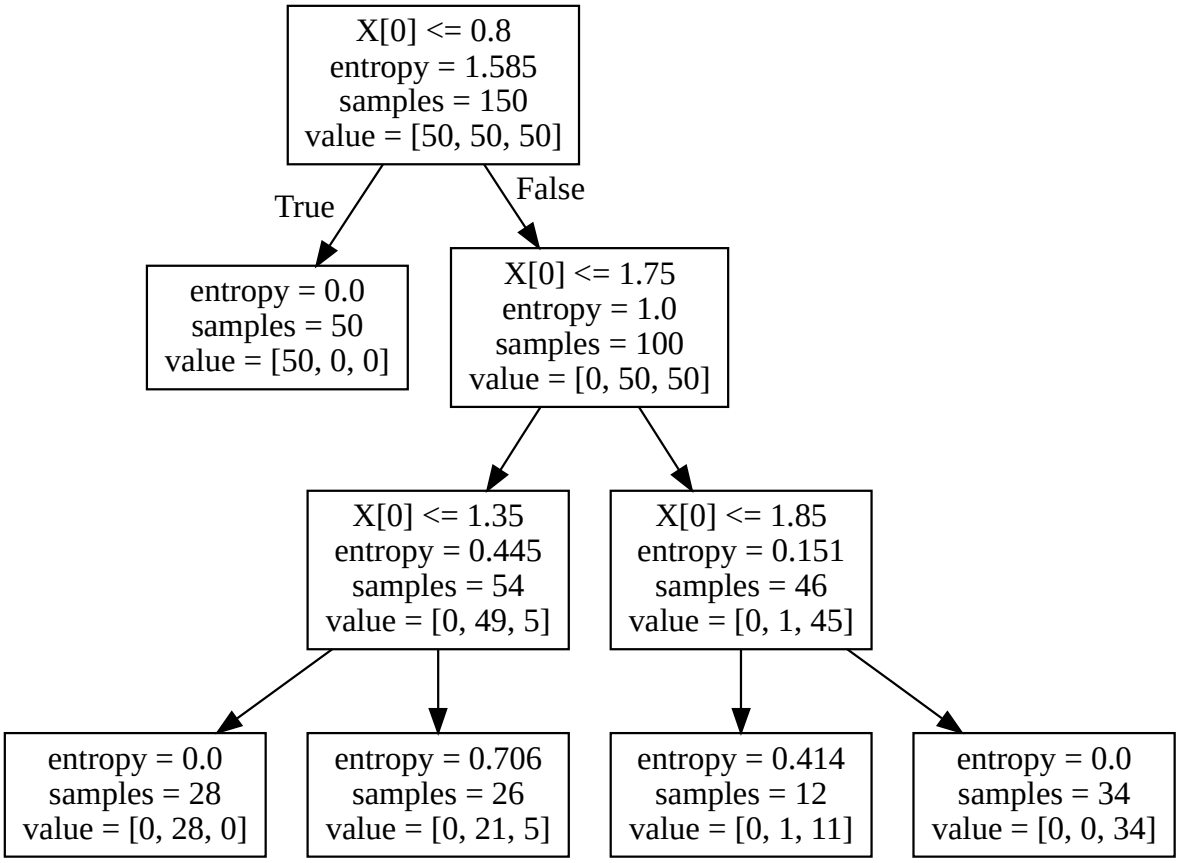

```
Out[22]: <ggplot: (-9223372036844195938)>
```



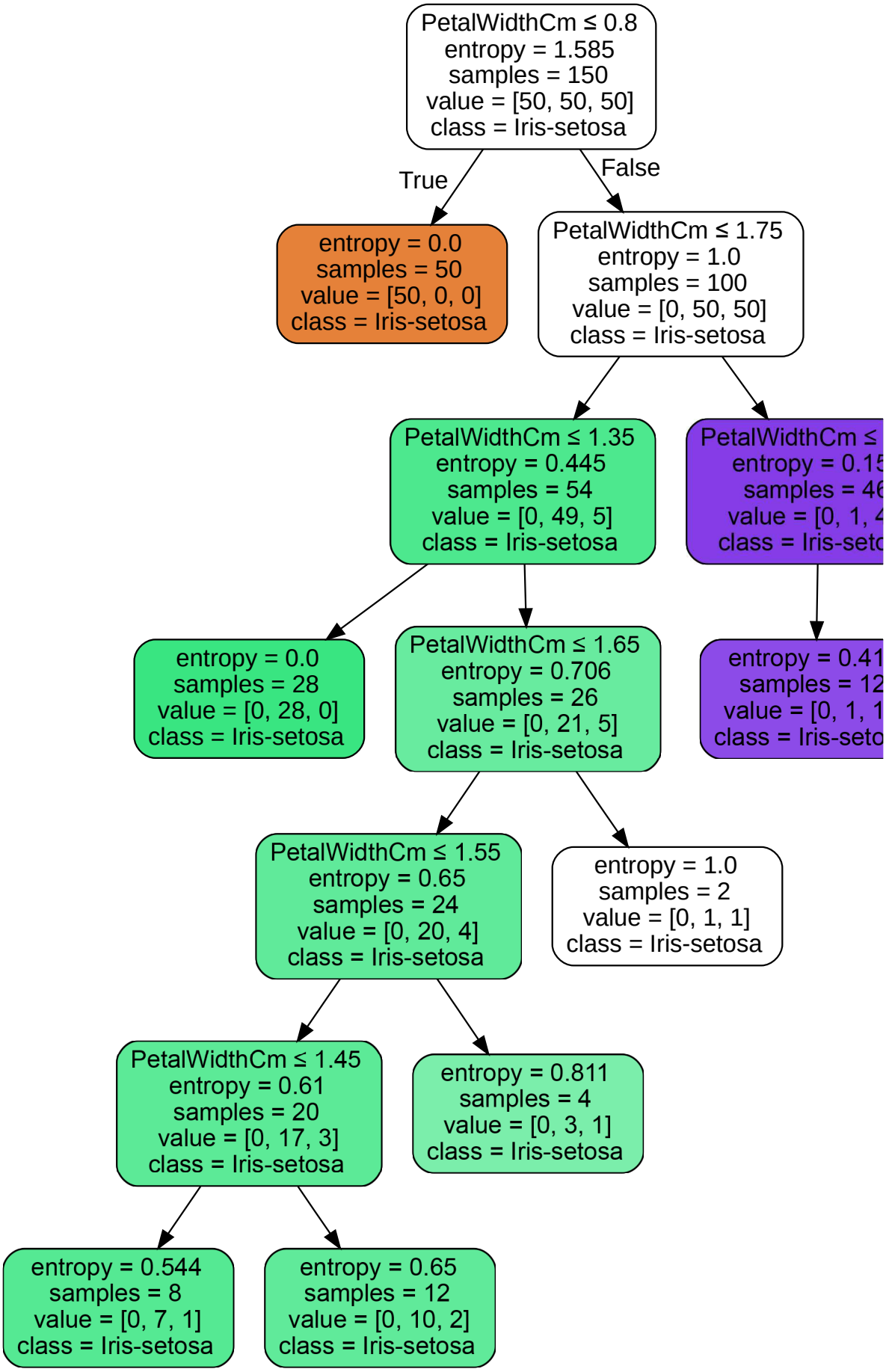

```
Out[24]: <ggplot: (10585408)>
```

```
Out[3]: DecisionTreeClassifier(class_weight=None, criterion='entropy', max_depth=3,
                               max_features=None, max_leaf_nodes=None,
                               min_impurity_decrease=0.0, min_impurity_split=None,
                               min_samples_leaf=1, min_samples_split=2,
                               min_weight_fraction_leaf=0.0, presort=False, random_state=None,
                               splitter='best')
```

Out[4]:


X[0] <= 0.8
entropy = 1.585
samples = 150
value = [50, 50, 50]
True
False
entropy = 0.0
samples = 50
value = [50, 0, 0]
X[0] <= 1.75
entropy = 1.0
samples = 100
value = [0, 50, 50]
X[0] <= 1.35
entropy = 0.445
samples = 54
value = [0, 49, 5]
X[0] <= 1.85
entropy = 0.151
samples = 46
value = [0, 1, 45]
entropy = 0.0
samples = 28
value = [0, 28, 0]
entropy = 0.706
samples = 26
value = [0, 21, 5]
entropy = 0.414
samples = 12
value = [0, 1, 11]
entropy = 0.0
samples = 34
value = [0, 0, 34]

Out[38]:


PetalWidthCm ≤ 0.8
entropy = 1.585
samples = 150
value = [50, 50, 50]
class = Iris-setosa
True
False
entropy = 0.0
samples = 50
value = [50, 0, 0]
class = Iris-setosa
PetalWidthCm ≤ 1.75
entropy = 1.0
samples = 100
value = [0, 50, 50]
class = Iris-setosa
PetalWidthCm ≤ 1.35
entropy = 0.445
samples = 54
value = [0, 49, 5]
class = Iris-setosa
PetalWidthCm ≤
entropy = 0.15
samples = 46
value = [0, 1, 4
class = Iris-seto
entropy = 0.0
samples = 28
value = [0, 28, 0]
class = Iris-setosa
PetalWidthCm ≤ 1.65
entropy = 0.706
samples = 26
value = [0, 21, 5]
class = Iris-setosa
entropy = 0.41
samples = 12
value = [0, 1, 1
class = Iris-seto
PetalWidthCm ≤ 1.55
entropy = 0.65
samples = 24
value = [0, 20, 4]
class = Iris-setosa
entropy = 1.0
samples = 2
value = [0, 1, 1]
class = Iris-setosa
PetalWidthCm ≤ 1.45
entropy = 0.61
samples = 20
value = [0, 17, 3]
class = Iris-setosa
entropy = 0.811
samples = 4
value = [0, 3, 1]
class = Iris-setosa
entropy = 0.544
samples = 8
value = [0, 7, 1]
class = Iris-setosa
entropy = 0.65
samples = 12
value = [0, 10, 2]
class = Iris-setosa