About our Dataset

The iris dataset is a simple and beginner-friendly dataset that contains information about the flower petal and sepal sizes. The dataset has 3 classes with 50 instances in each class, therefore, it contains 150 rows with only 4 columns.
RandomForestRegressorRandomForestRegressorRandomForestRegressorRandomForestRegressor

Our Target

Implement a machine learning classification or regression model on the dataset. Classification is the task of separating items into its corresponding class.

Importing all the important libray

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Now reading our .csv file using Pandas Library

In [2]:
```
df = pd.read_csv('iris.data' , header = None)
df
```

Out[2]:

	0	1	2	3	4
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

Giving the datdset its coloumns names

In [3]:
```
names = ['sepal length','sepal width','petal length','petal width' , 'class']
```

In [5]:
```
df.columns = names
df
```

Out[5]:

	sepal length	sepal width	petal length	petal width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

In [6]:
```
df['class'].unique()
```

Out[6]:
```
array(['Iris-setosa', 'Iris-versicolor', 'Iris-virginica'], dtype=object)
```

Checking for Null Values or missing Values in our Dataset

In [7]:
```
df.isna().sum()
```

Out[7]:
```
sepal length    0
sepal width     0
petal length    0
petal width     0
class           0
dtype: int64
```

Data Exploration

In [8]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   sepal length  150 non-null    float64
 1   sepal width   150 non-null    float64
 2   petal length  150 non-null    float64
 3   petal width   150 non-null    float64
 4   class         150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

In [9]:
```
df.describe()
```

Out[9]:

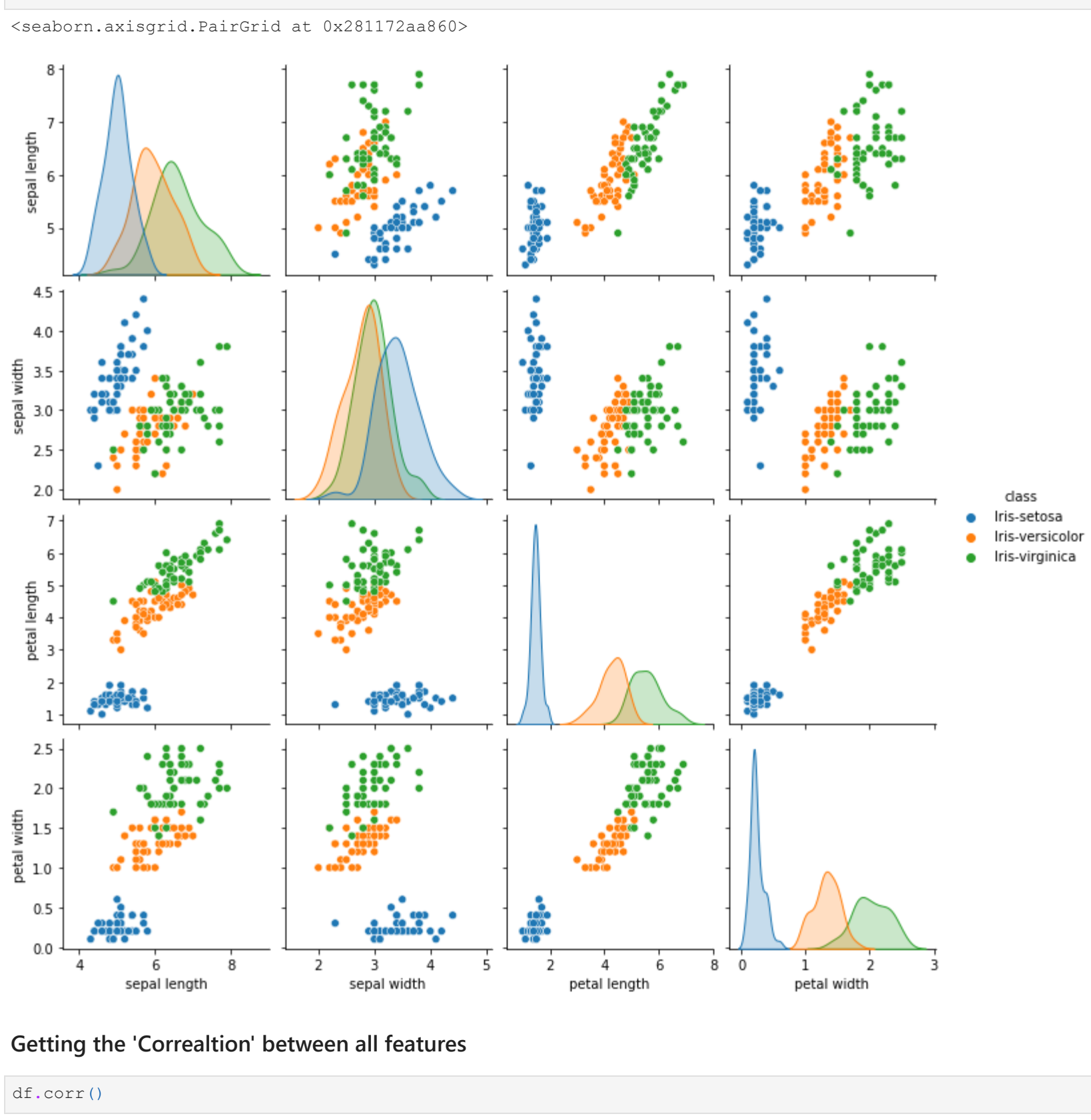
	sepal length	sepal width	petal length	petal width
count	150.000000	150.000000	150.000000	150.000000
mean	5.843333	3.054000	3.758667	1.198667
std	0.828066	0.433594	1.764420	0.763161
min	4.300000	2.000000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.350000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.400000	6.900000	2.500000

Plotting all the features against all other features.

We do this to get the main attributes that should be used for classifications

In [10]:
```
sns.pairplot(data = df ,hue ='class')
```

Out[10]:
```
<seaborn.axisgrid.PairGrid at 0x281172aa860>
```



Getting the 'Correaltion' between all features

In [11]:
```
df.corr()
```

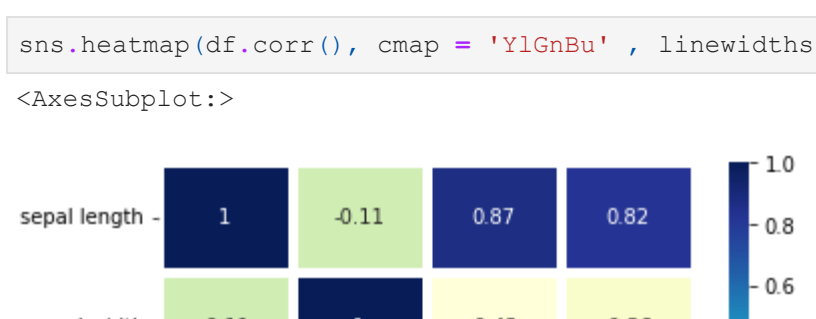
Out[11]:

	sepal length	sepal width	petal length	petal width
sepal length	1.000000	-0.109369	0.871754	0.817954
sepal width	-0.109369	1.000000	-0.420516	-0.356544
petal length	0.871754	-0.420516	1.000000	0.962757
petal width	0.817954	-0.356544	0.962757	1.000000

Ploting the Correlation using Heatmap

In [12]:
```
sns.heatmap(df.corr(), cmap = 'YlGnBu' , linewidths = 5 , annot= True)
```

Out[12]:
```
<AxesSubplot:>
```



In [32]:
```
df
```

Out[32]:

	sepal length	sepal width	petal length	petal width	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

Selecting our X and Y data

In [33]:
```
X = df.iloc[:,1:]
X
```

Out[33]:

	sepal width	petal length	petal width	class
0	3.5	1.4	0.2	Iris-setosa
1	3.0	1.4	0.2	Iris-setosa
2	3.2	1.3	0.2	Iris-setosa
3	3.1	1.5	0.2	Iris-setosa
4	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...
145	3.0	5.2	2.3	Iris-virginica
146	2.5	5.0	1.9	Iris-virginica
147	3.0	5.2	2.0	Iris-virginica
148	3.4	5.4	2.3	Iris-virginica
149	3.0	5.1	1.8	Iris-virginica

150 rows × 4 columns

In [45]:
```
y = df.iloc[:,0:1]
y
```

Out[45]:

	sepal length
0	5.1
1	4.9
2	4.7
3	4.6
4	5.0
...	...
145	6.7
146	6.3
147	6.5
148	6.2
149	5.9

150 rows × 1 columns

Transforming the Class or ['Class'] coloumn into numeric data.

We are using LabelEncoding for this

In [36]:
```
from sklearn.preprocessing import LabelEncoder

LE = LabelEncoder()
X['class'] = LE.fit_transform(X['class'])
X
```

Out[36]:

	sepal width	petal length	petal width	class
0	3.5	1.4	0.2	0
1	3.0	1.4	0.2	0
2	3.2	1.3	0.2	0
3	3.1	1.5	0.2	0
4	3.6	1.4	0.2	0
...	...	...	...	...
145	3.0	5.2	2.3	2
146	2.5	5.0	1.9	2
147	3.0	5.2	2.0	2
148	3.4	5.4	2.3	2
149	3.0	5.1	1.8	2

150 rows × 4 columns

In [37]:
```
X = X.values
```

Splitting our dataset into training and testing data

In [38]:
```
from sklearn.model_selection import train_test_split

x_train,x_test,y_train,y_test = train_test_split(X,y, test_size= 0.33, random_state= 101)
```

In [39]:
```
x_train[:5]
```

Out[39]:
```
array([[2.5, 3.9, 1.1, 1. ],
       [3.3, 1.7, 0.5, 0. ],
       [2.7, 5.1, 1.9, 2. ],
       [3.6, 1.4, 0.2, 0. ],
       [2.4, 3.3, 1. , 1. ]])
```

Fitting our data in the model

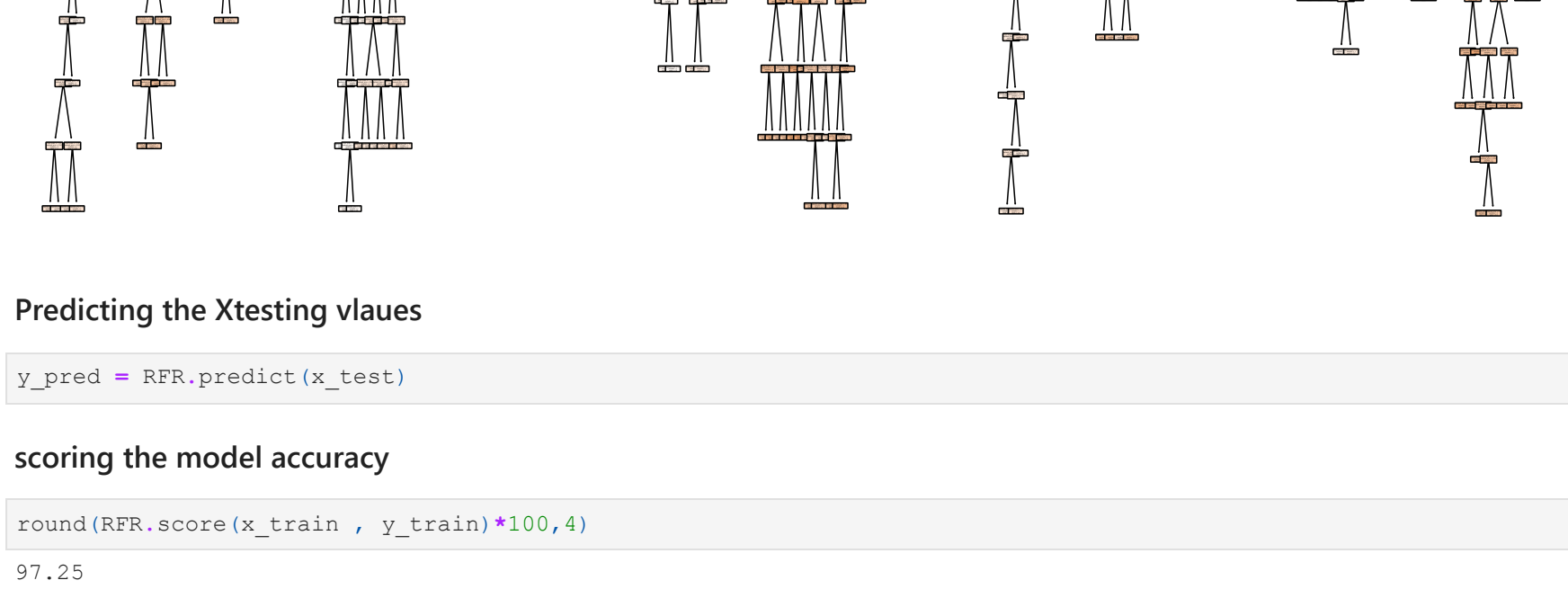
In [46]:
```
from sklearn.ensemble import RandomForestRegressor

RFR = RandomForestRegressor()
RFR.fit(x_train , y_train.values.ravel())
```

Out[46]:
```
RandomForestRegressor()
```

Plotting the RandomForestRegressor trees

In [47]:
```
from sklearn.tree import export_graphviz
import graphviz
```

In [48]:
```
from sklearn import tree

#fig = plt.figure(figsize=(10,8))
fig, axes = plt.subplots(nrows = 1,ncols = 5,figsize = (20,10) , dpi = 900)
a = 0
for i in range(0,100,20):

    tree.plot_tree(RFR.estimators_[i],    feature_names = names[:-1],class_names= df['class'].unique(),
                filled=True, impurity=True,rounded=True , ax = axes[a])

    axes[a].set_title('Estimators -: ' + str(i), fontsize = 11)
    a = a + 1



fig.savefig('RFR_5trees.png')
```



Predicting the Xtesting vlaues

In [49]:
```
y_pred = RFR.predict(x_test)
```

scoring the model accuracy

In [50]:
```
round(RFR.score(x_train , y_train)*100,4)
```

Out[50]:
```
97.25
```

In [62]:
```
y_test_values = y_test.values.ravel()
y_test_values
```

Out[62]:
```
array([5.5, 5.4, 5. , 7.2, 7. , 6.3, 6.2, 5.5, 6.3, 4.9, 6.5, 5.2, 5.4,
       7.7, 6.1, 6.4, 5.6, 5.7, 4.4, 6.1, 5.5, 5.3, 5.5, 6.7, 6.2, 5.6,
       5. , 5.8, 5. , 4.5, 6.5, 5.5, 6.5, 5.8, 6.8, 5.7, 6.7, 5.9, 5.6,
       7.7, 5.1, 4.6, 4.7, 6.7, 5.6, 4.3, 7.1, 5.8, 4.9, 5.1])
```

In [52]:
```
y_pred
```

Out[52]:
```
array([5.516     , 5.408     , 5.12425   , 6.615     , 6.273     ,
       5.987     , 5.894     , 5.98      , 5.765     , 4.603     ,
       6.285     , 5.515     , 4.95226667, 7.674     , 6.017     ,
       6.252     , 5.813     , 5.829     , 4.73      , 6.412     ,
       5.519     , 5.313     , 6.074     , 6.346     , 6.252     ,
       5.381     , 5.049     , 6.156     , 4.7085    , 4.601     ,
       6.713     , 5.915     , 6.335     , 5.498     , 6.592     ,
       6.085     , 6.366     , 5.664     , 5.783     , 7.292     ,
       5.247     , 5.365     , 4.7785    , 5.951     , 5.642     ,
       4.713     , 6.736     , 5.752     , 4.867     , 5.124     ])
```

In [63]:
```
print("Actual Value :-" , y_test_values[0])
print("Predicted  Value :-" , y_pred[0])
```

```
Actual Value :- 5.5
Predicted  Value :- 5.516
```