Data_Analysis

March 15, 2022

Data Fields and Types

```
      attrname attrtype factororresponse
0  SepalLength  numeric           factor
1   SepalWidth  numeric           factor
2  PetalLength  numeric           factor
3   PetalWidth  numeric           factor
4    IrisClass  nominal         response
```

attribute_datas None

Getting a Response Column

response_attributes: IrisClass

Getting a Non-Factor Column

non_factor_attributesname: []

Getting a Factor Column

factor_attributesname: ['SepalLength' 'SepalWidth' 'PetalLength' 'PetalWidth']

Getting a Nominal Type Column

nominal_attributes: ['IrisClass']

Getting a Numeric Type Column

numeric_attributes: ['SepalLength' 'SepalWidth' 'PetalLength' 'PetalWidth']

Getting a Binary Type Column

binary_attributes: []

Sample Data

	SepalLength	SepalWidth	PetalLength	PetalWidth	IrisClass
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

data head None

Data Columns

```
data columns:  Index(['SepalLength', 'SepalWidth', 'PetalLength', 'PetalWidth',
'IrisClass'], dtype='object')
```

Summary Statistics

Data Describe

	count	mean	std	min	25%	50%	75%	max
SepalLength	150.0	5.843333	0.828066	4.3	5.1	5.80	6.4	7.9
SepalWidth	150.0	3.054000	0.433594	2.0	2.8	3.00	3.3	4.4
PetalLength	150.0	3.758667	1.764420	1.0	1.6	4.35	5.1	6.9
PetalWidth	150.0	1.198667	0.763161	0.1	0.3	1.30	1.8	2.5

data.describe None

Data Count and Values

```
data count:
 SepalLength    150
SepalWidth     150
PetalLength    150
PetalWidth     150
IrisClass      150
dtype: int64
```

Data Shape

data shape (150, 5)

Displaying Counted Unique Values of Nominal Attributes

Column > IrisClass < unique values

['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']

Column > IrisClass < unique values count:3

```
Index(['IrisClass'], dtype='object')
data head:
```

	SepalLength	SepalWidth	PetalLength	PetalWidth	IrisClass
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

None

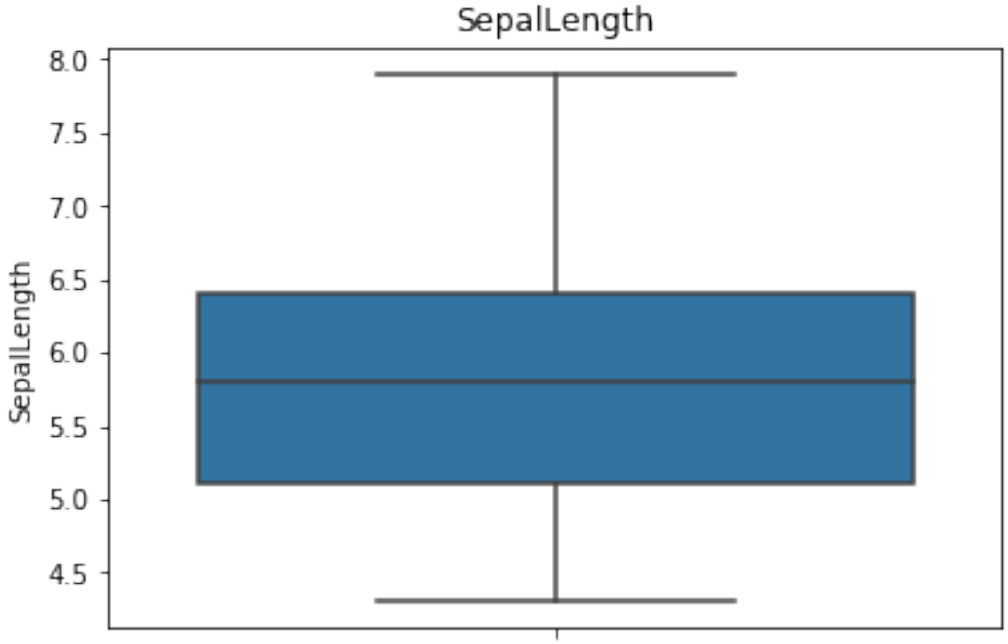
Columns to be Dropped

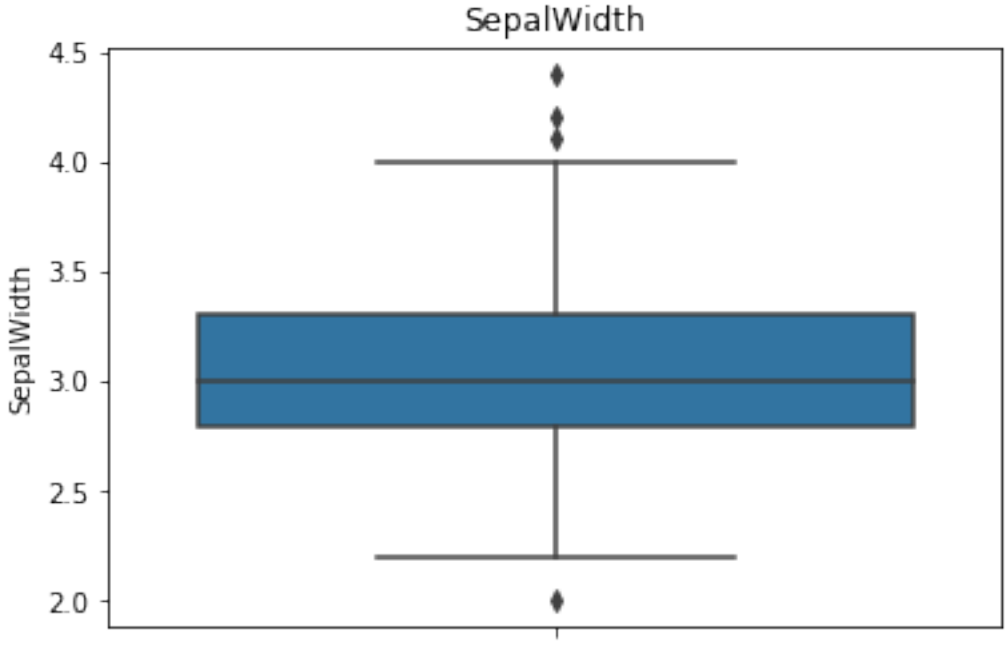
dropping_columns : []

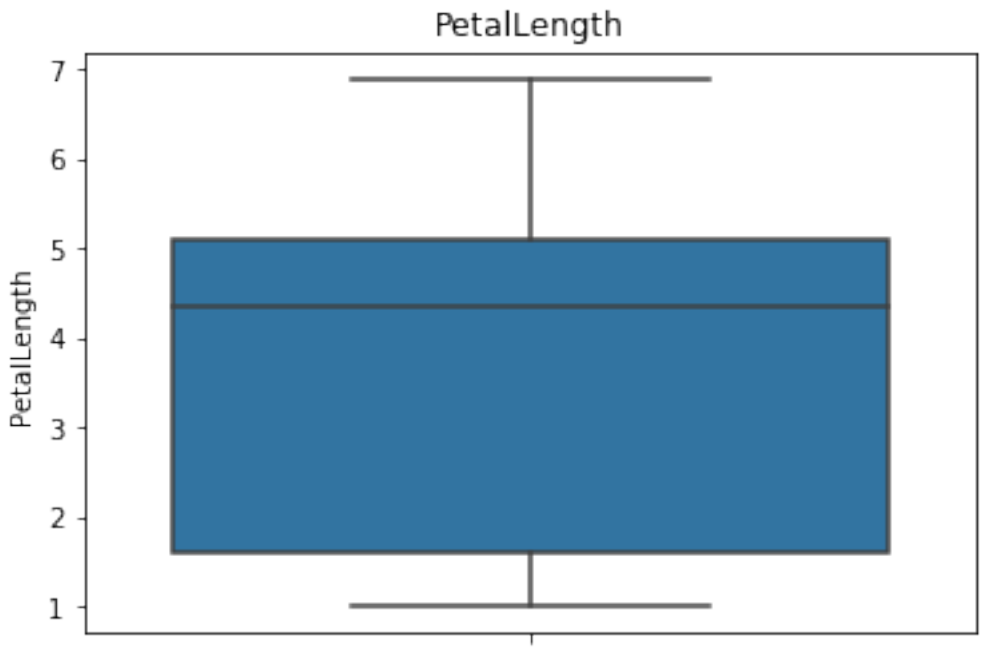
processing_columns : ['SepalLength', 'SepalWidth', 'PetalLength', 'PetalWidth']

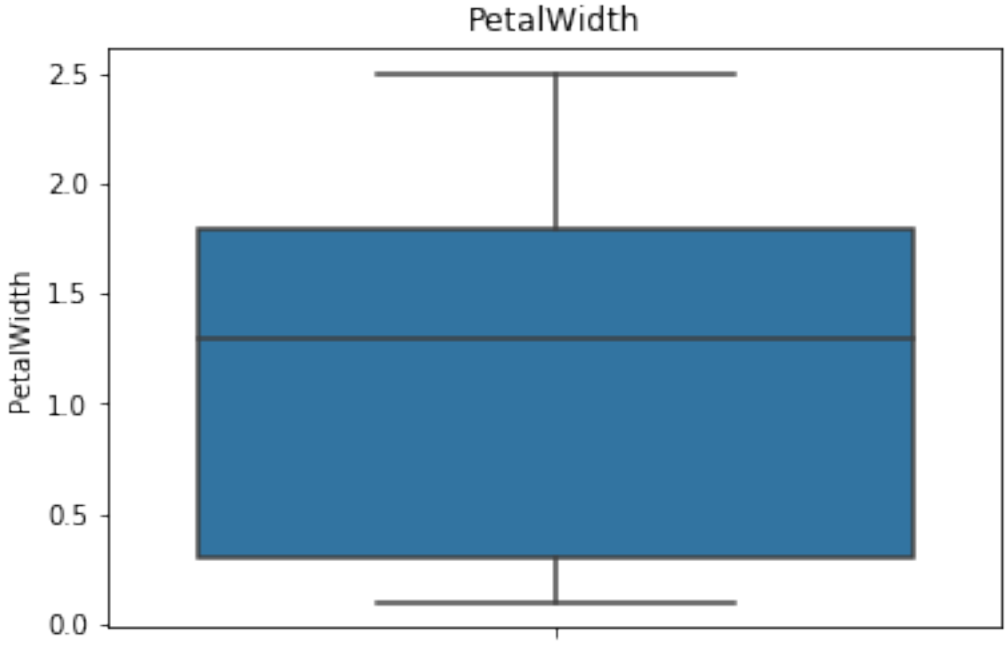
Univariate Analysis

Box Plots for Numeric Attributes

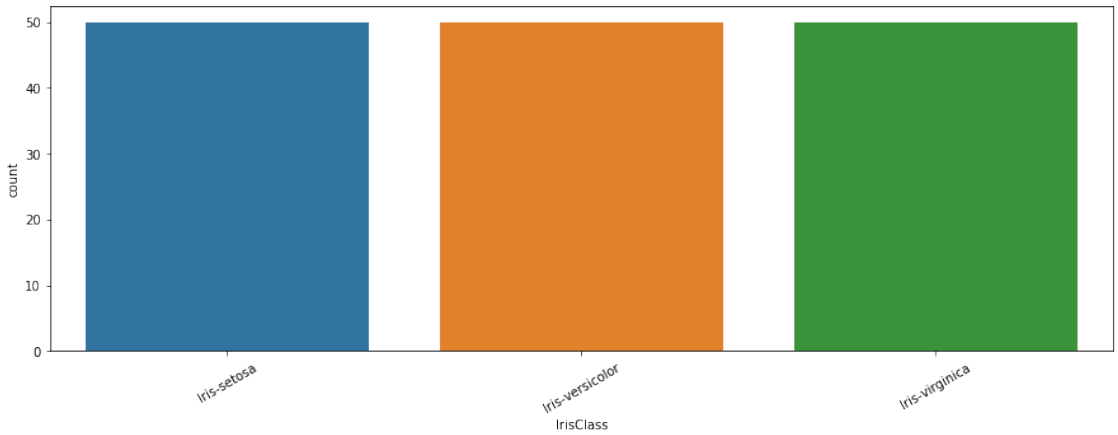








Frequency Charts for Nominal(Categorical) Attributes



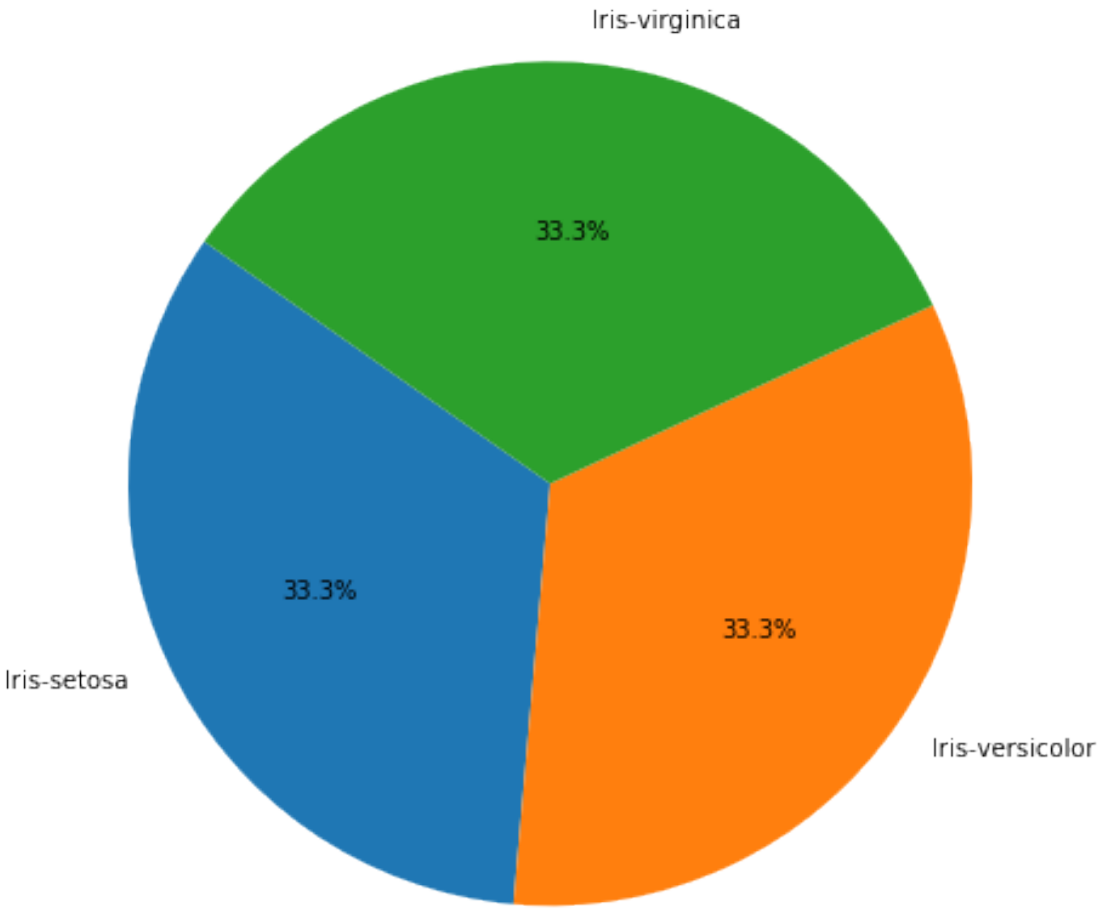

```
IrisClass :
Iris-setosa        50
Iris-versicolor    50
Iris-virginica     50
Name: IrisClass, dtype: int64
A number of Class : 3
```

Displaying Binary Variables

Pie Charts for Nominal(Categorical) Attributes

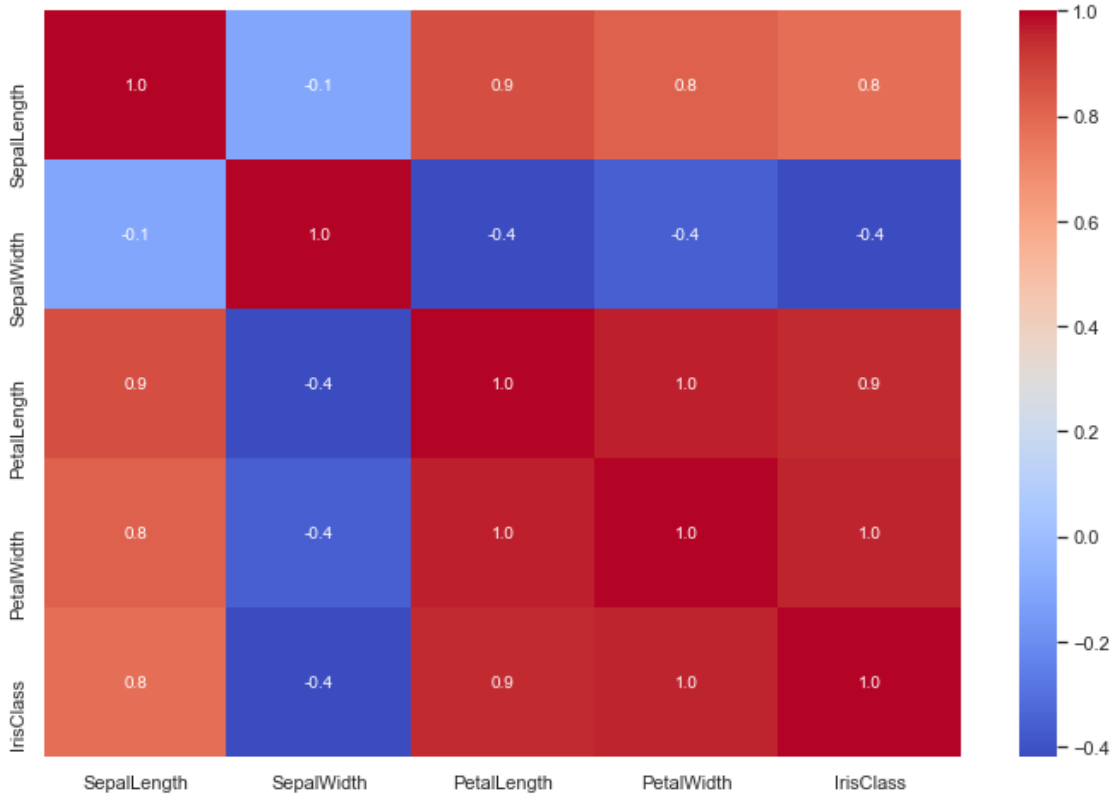
Display Nominal Features Value Counts on pie graph

Displaying IrisClass in pie graph.




```
Correlation Between Numeric Column

numeric_values_for_correlation:  ['SepalLength' 'SepalWidth' 'PetalLength'
'PetalWidth']
```



Matrix that involves correlation values between numeric columns

	SepalLength	SepalWidth	PetalLength	PetalWidth	IrisClass
SepalLength	1.000000	-0.109369	0.871754	0.817954	0.782561
SepalWidth	-0.109369	1.000000	-0.420516	-0.356544	-0.419446
PetalLength	0.871754	-0.420516	1.000000	0.962757	0.949043
PetalWidth	0.817954	-0.356544	0.962757	1.000000	0.956464
IrisClass	0.782561	-0.419446	0.949043	0.956464	1.000000

corr table: None

Analysis for Values of Nominal Attributes

Nominal Features: IrisClass - Mean Of Numeric Features (First 10)

	IrisClass	SepalLength	SepalWidth	PetalLength	PetalWidth
0	0	5.006	3.418	1.464	0.244
1	1	5.936	2.770	4.260	1.326
2	2	6.588	2.974	5.552	2.026

None

Nominal Features: IrisClass - Count Of Numeric Features (First 10)

	IrisClass	SepalLength	SepalWidth	PetalLength	PetalWidth
0	0	50	50	50	50

```
1          1          50          50          50          50
2          2          50          50          50          50
```

```
None
```

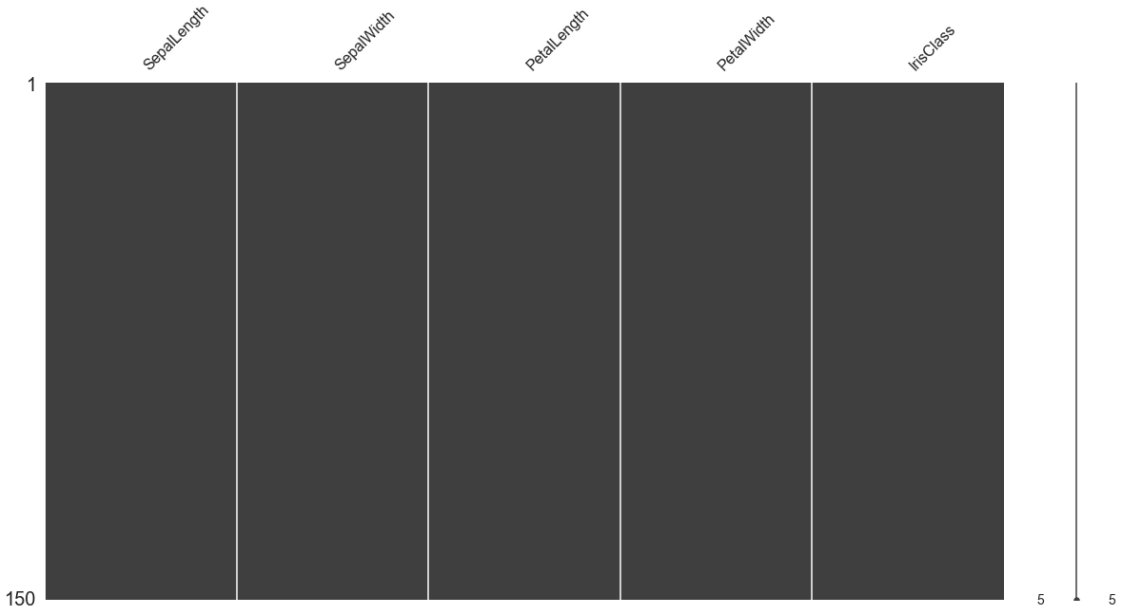
Frequency Charts for Nominal(Categorical) Attributes

Analysis of Missing(Null) Values

Check null values and processing on data

```
data null sum each attribute:
 SepalLength    0
SepalWidth     0
PetalLength    0
PetalWidth     0
IrisClass      0
dtype: int64
```

```
data null sum :  0
```




```
             Null Values Sum  % Value
SepalLength                0      0.0
SepalWidth                 0      0.0
PetalLength                0      0.0
PetalWidth                 0      0.0
IrisClass                  0      0.0
```

```
data shape :  (150, 5)
```

```
data null sum:
 SepalLength    0
SepalWidth     0
PetalLength    0
PetalWidth     0
IrisClass      0
dtype: int64
```

Encoding Nominal Attributes

```
data columns Index(['SepalLength', 'SepalWidth', 'PetalLength', 'PetalWidth',
'IrisClass'], dtype='object')
```

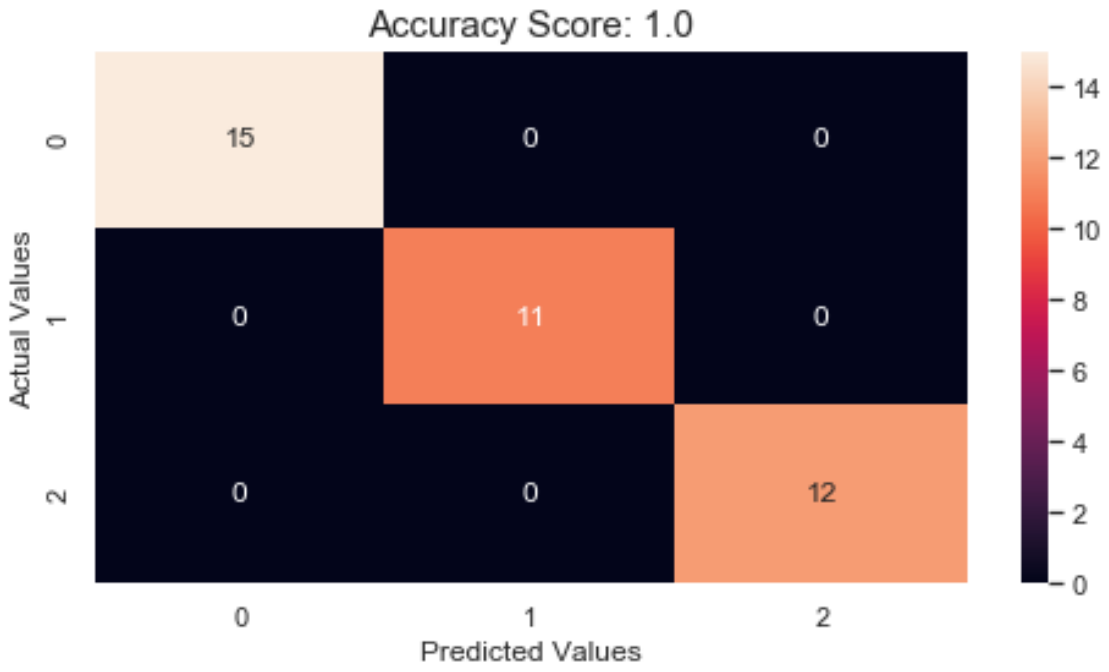
```
data columns after dropping Index(['SepalLength', 'SepalWidth', 'PetalLength',
'PetalWidth', 'IrisClass'], dtype='object')
```

```
Splitting our data as test and train
```

	SepalLength	SepalWidth	PetalLength	PetalWidth
0	0.222222	0.625000	0.067797	0.041667
1	0.166667	0.416667	0.067797	0.041667
2	0.111111	0.500000	0.050847	0.041667
3	0.083333	0.458333	0.084746	0.041667
4	0.194444	0.666667	0.067797	0.041667

```
train data head table:  None
```

```
Decision Tree Classification
decision_tree_regression_model DecisionTreeClassifier(ccp_alpha=0.0,
class_weight=None, criterion='gini',
                       max_depth=None, 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='deprecated',
                       random_state=None, splitter='best')
[[15  0  0]
 [ 0 11  0]
 [ 0  0 12]]
accuracy_score 1.0
```



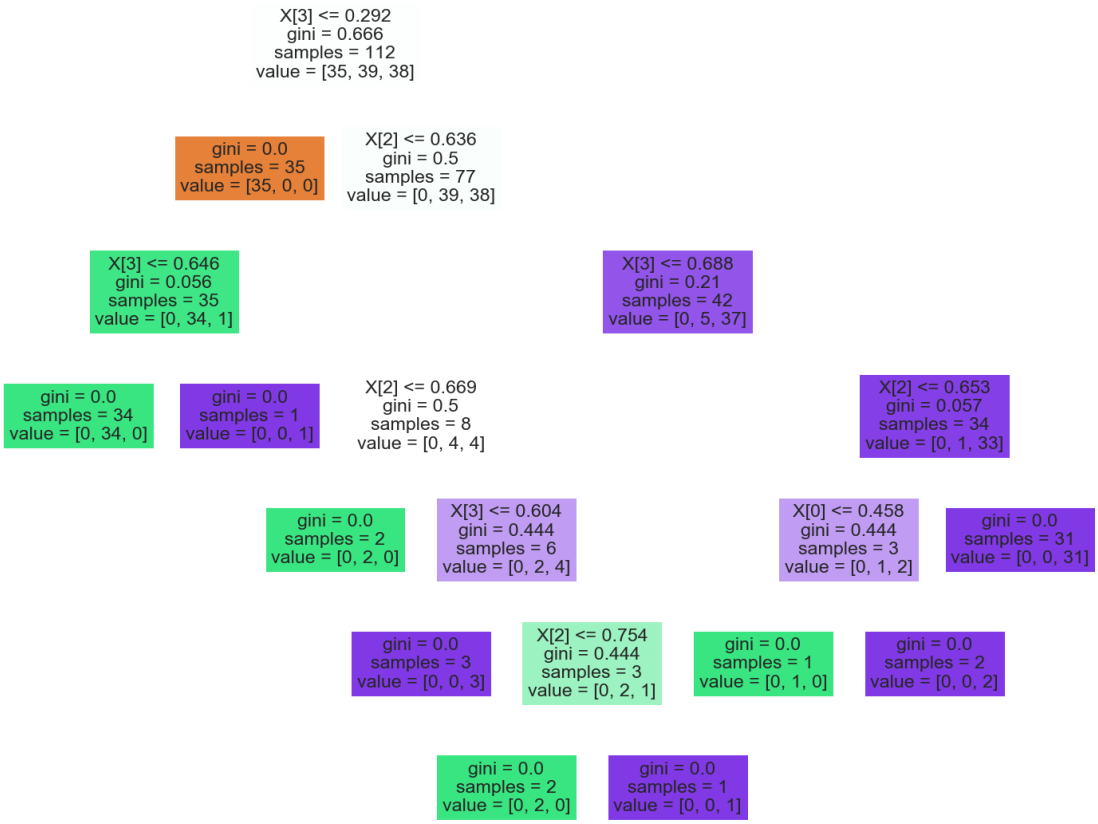
	precision	recall	f1-score	support
0	1.00	1.00	1.00	15
1	1.00	1.00	1.00	11
2	1.00	1.00	1.00	12
accuracy			1.00	38
macro avg	1.00	1.00	1.00	38
weighted avg	1.00	1.00	1.00	38

```
Fitting 10 folds for each of 30 candidates, totalling 300 fits

[Parallel(n_jobs=-1)]: Using backend LokyBackend with 8 concurrent workers.
[Parallel(n_jobs=-1)]: Done  25 tasks      | elapsed:    3.9s
[Parallel(n_jobs=-1)]: Done 300 out of 300 | elapsed:    4.6s finished

model_cv_name:  best_score : 0.9371212121212119
model_cv_name:  best_params : {'max_depth': 3, 'min_samples_split': 10}
model_cv_name:  best_estimator : DecisionTreeClassifier(ccp_alpha=0.0,
class_weight=None, criterion='gini',
                       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=10,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
accuracy_score 1.0
```

Decision tree rules



0