Table of contents

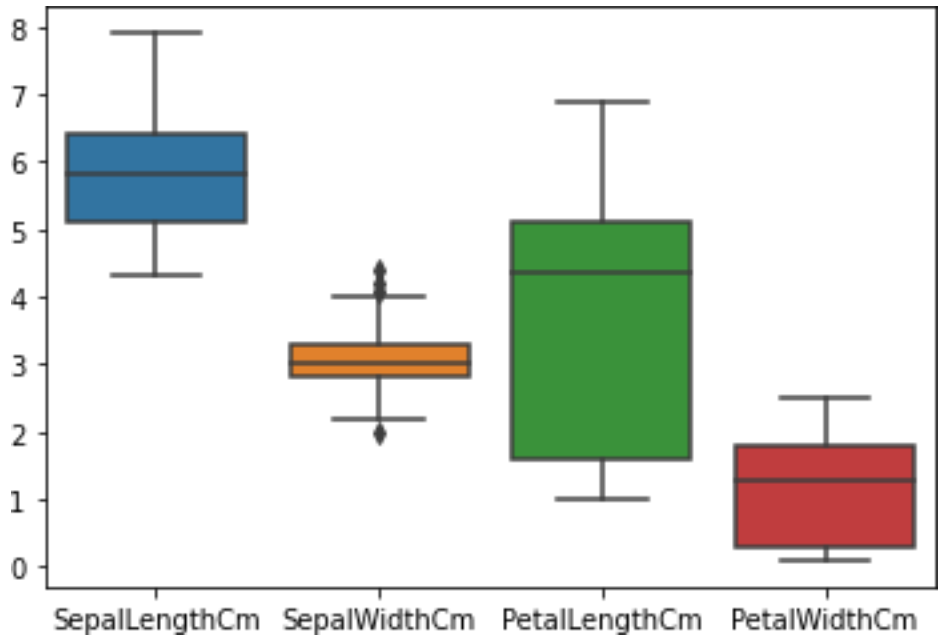
DESCRIPTION:

The Iris Data has 150 instances and 5 attributes with 4 Independent(Sepal length, Sepal Width, Petal length and Petal width) and one Dependent Variable (Species).

The numerical values of all 4 independent variables have the same scale (centimeters) and similar ranges between 0 and 8 centimeters. There are three duplicate observations so after removing the duplicates dataset size decreased to 147 instances. There are no null values found. Replacing a few Outliers with median instead of removing them as sample size if very small.

The Dependent Variable has three unique values with each has 50 instances, After removing duplicates, we have one class not more than 4% of other class. so usually

depends on dataset when one class is 30% to 60% greater than other is considered class imbalance.



	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	147.000000	147.000000	147.000000	147.000000
mean	5.856463	3.037415	3.780272	1.208844
std	0.829100	0.392779	1.759111	0.757874
min	4.300000	2.200000	1.000000	0.100000
25%	5.100000	2.800000	1.600000	0.300000
50%	5.800000	3.000000	4.400000	1.300000
75%	6.400000	3.300000	5.100000	1.800000
max	7.900000	4.000000	6.900000	2.500000

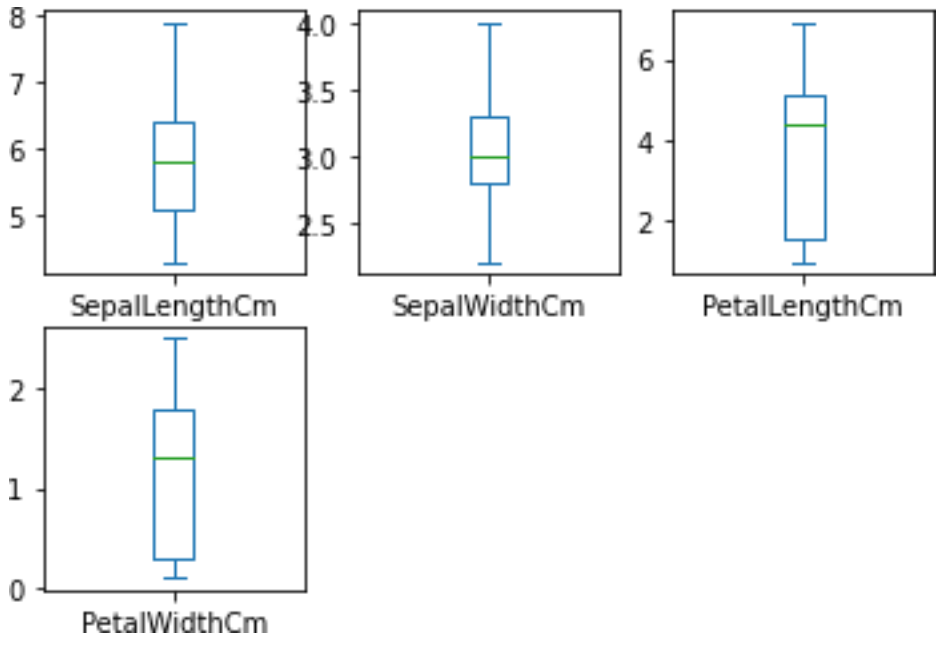
Number of Classes inside Species

Species	
Iris-setosa	48
Iris-versicolor	50
Iris-virginica	49

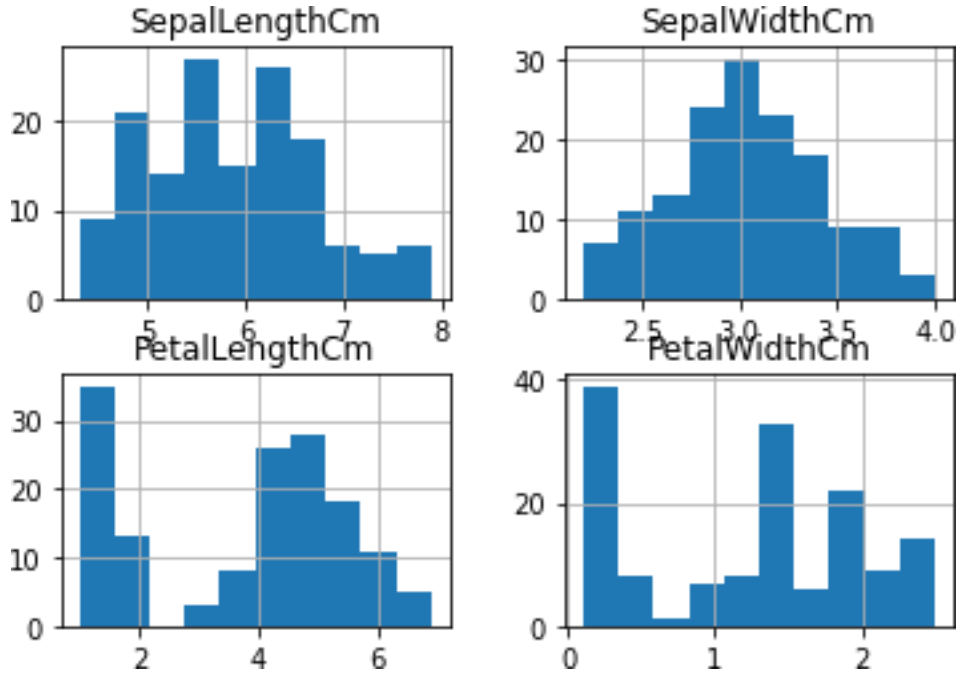
DATA VISUALIZATION

Histogram shows two of the input variables have a Gaussian distribution. After replacing outliers with Median there are no outliers visible in Box plot.

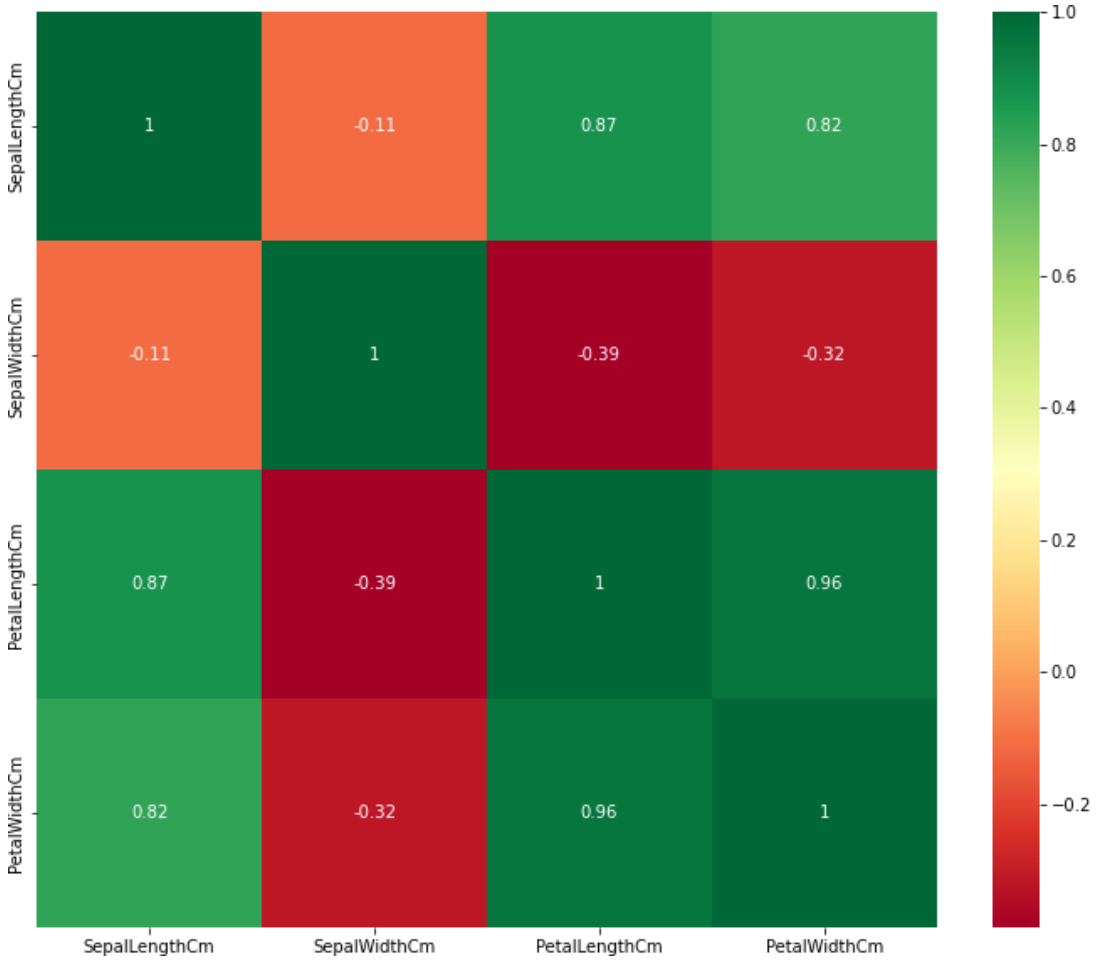
BOX PLOT



HISTOGRAM

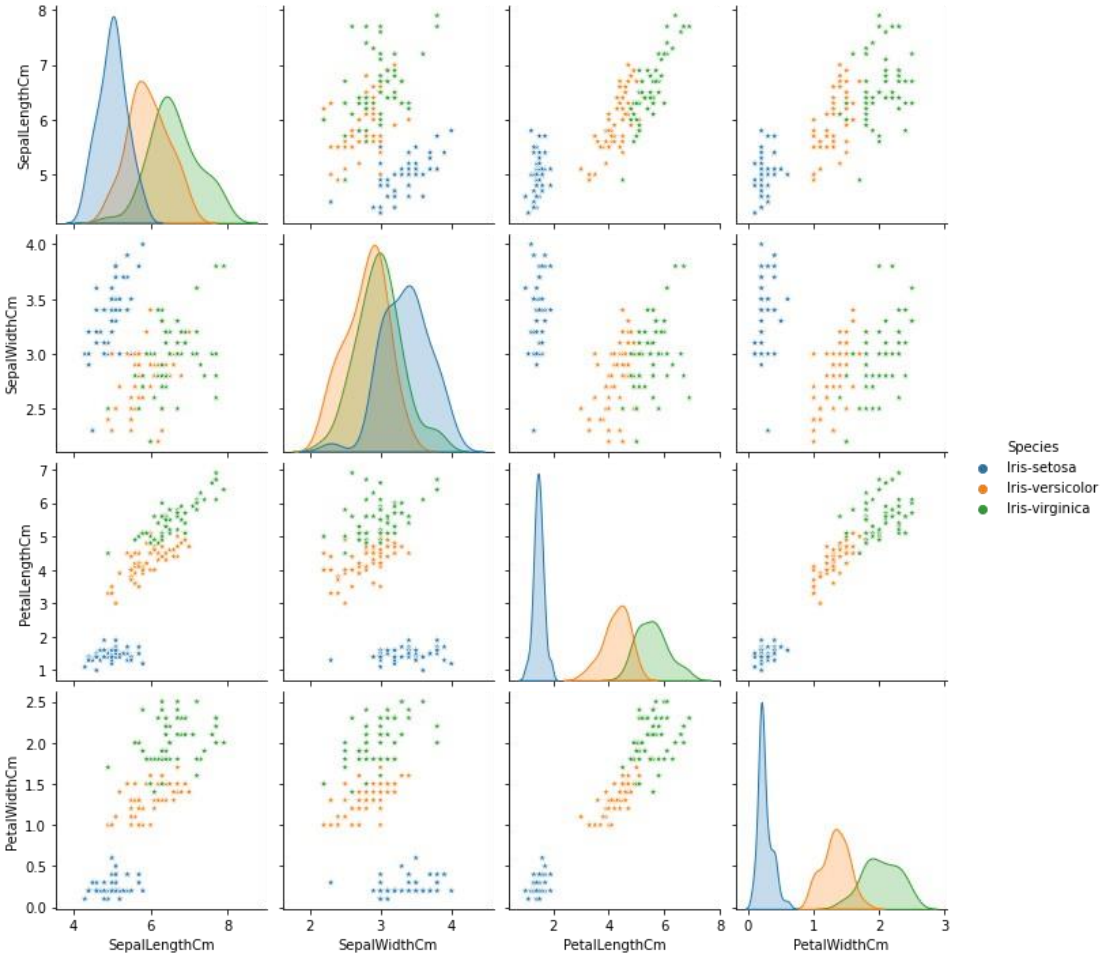


HEATMAP to see Correlations



SCATTER PLOT for data variations

The relation between pairs of inputs of an iris-setosa (in pink) is particularly unique in relation to those of the other two species. There is some cross-over in the pairwise connections of the other two species, iris-versicolor (brown) and iris-virginica (green).



Splitting the loaded dataset into two, 80% of which we will use to train our models and20% that we will hold back as a validation dataset

---

SUPPORT VECTOR MACHINES:

Accuracy Score with Validation Data: 0.9666666666666667

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	8
Iris-versicolor	0.90	1.00	0.95	9
Iris-virginica	1.00	0.92	0.96	13
accuracy			0.97	30
macro avg	0.97	0.97	0.97	30
weighted avg	0.97	0.97	0.97	30

RANDOM FOREST:

Accuracy Score with Validation Data: 0.9333333333333333

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	8
Iris-versicolor	0.89	0.89	0.89	9
Iris-virginica	0.92	0.92	0.92	13
accuracy			0.93	30
macro avg	0.94	0.94	0.94	30
weighted avg	0.93	0.93	0.93	30

KNN CLASSIFIER:

Accuracy Score with Validation Data: 0.9333333333333333

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	8
Iris-versicolor	0.89	0.89	0.89	9

Iris-virginica	0.92	0.92	0.92	13
accuracy			0.93	30
macro avg	0.94	0.94	0.94	30
weighted avg	0.93	0.93	0.93	30

LOGISTIC REGRESSION:

When Tried with Two Solvers Newton-cg and Liblinear clearly Liblinear has better Cross Valdiation Score with 0.9575 since our dataset is Pretty Small with size of just 150 Observations.

Mean Cross Validation Score using solver newton-cg : 0.95

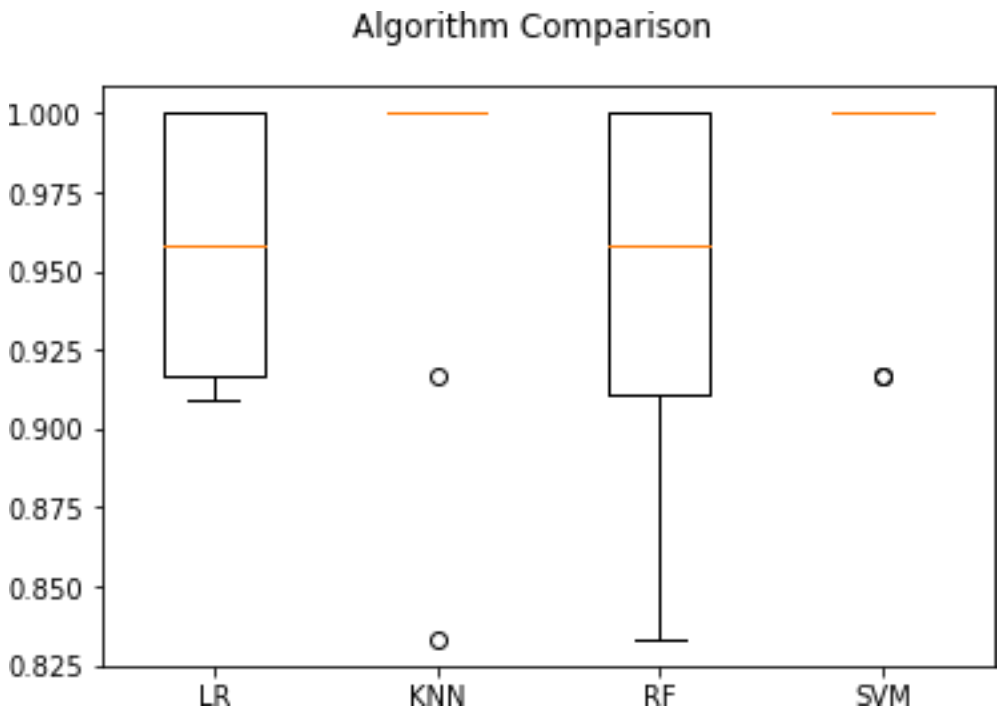
Mean Cross Validation Score using solver liblinear : 0.9575757575757574
Accuracy Score with Validation Data: 0.9666666666666667[[8 0 0]
[0 8 1]
[0 0 13]]

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	8
Iris-versicolor	1.00	0.89	0.94	9
Iris-virginica	0.93	1.00	0.96	13
accuracy			0.97	30
macro avg	0.98	0.96	0.97	30
weighted avg	0.97	0.97	0.97	30

STRATIFIED K_FOLD for all MODELS:

Model with Mean Cross validation score and Standard Deviation

LR: 0.957576 (0.042478)
KNN: 0.975000 (0.053359)
RF: 0.940909 (0.065415)
SVM: 0.983333 (0.033333)



Model with Highest Accuracy : SVM Accuracy Score of SVM: 0.9666666666666667

	precision	recall	f1-score	support
Iris-setosa	1.00	1.00	1.00	8
Iris-versicolor	0.90	1.00	0.95	9
Iris-virginica	1.00	0.92	0.96	13
accuracy			0.97	30
macro avg	0.97	0.97	0.97	30
weighted avg	0.97	0.97	0.97	30

CONCLUSION:

Clearly the F1 score of Iris-Setosa is perfect 1 for all four models but other only SVM hasbetter F1 score for other two Classes. No Matter where we used GridSearch CV or just Stratified K-fold the output of Support Vector Machine is Same.

As we can see below time to run models with Gridsearch CV is more than the time taken forall other Models with Stratified Kfold. Here our results are pretty similar no matter whether it is Stratified or Gridsearch so process which take shorter time to run and executewill be ideal for Production.

Time to Run SVM with Grid Search CV : 0:00:00.252399
Time to Run Random Forest with Grid Search CV : 0:00:42.708611
Time to Run KNN with Grid Search CV : 0:00:00.158763
Time to Run Logistic Regression with Grid Search CV : 0:00:00.098761
Time to Run Stratified Kfold with all Models : 0:00:00.767912