**Analytical Report**

**Prepared By:**

Vemuri Varun Sri Sai

[vemuriv2233@gmail.com](mailto:vemuriv2233@gmail.com)

Feb 28,2022

**Overview:**

This report covers linear regression and logistic regression using discriminative methods. We apply these models to the Iris data set and how summarize the accuracies of these models for different features and also for the entire dataset. In this report, we analyze how some features like sepal length, sepal width, petal length, petal width differs for different species. And also, we also study how L2 regularization effects the stochastic gradient regressor (Linear regression) and how the learning rate effect the stochastic classifier (logistic regression).

**Scope:**

This work focuses on creating 12 models of Stochastic regressor and 6 models of Stochastic classifier and study their accuracies (Mean square error for linear regression and log loss and accuracy score for logistic regression).

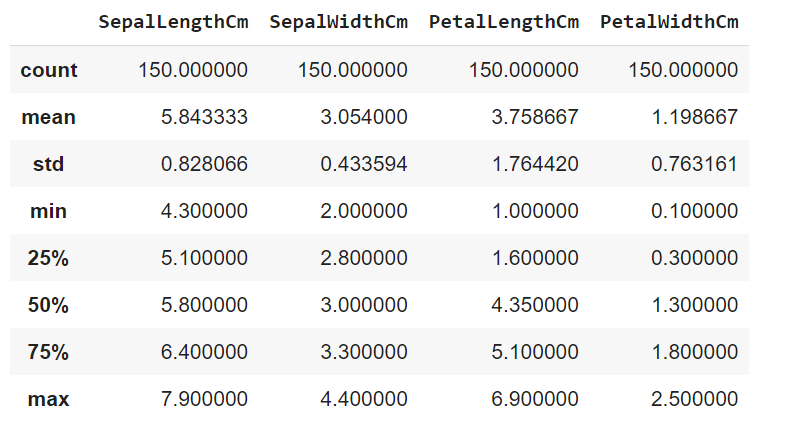
**Methodology:**

This report uses the Iris flower dataset organized by Ronald Fisher in 1936. We first split the iris data set into train and test data sets selecting 10% of dataset randomly using Stratified train-test, ensuring even split of the class. Then we apply the stochastic regressor and stochastic classifier to the data sets.

**Analysis:**

**EDA:**

Before splitting the data:

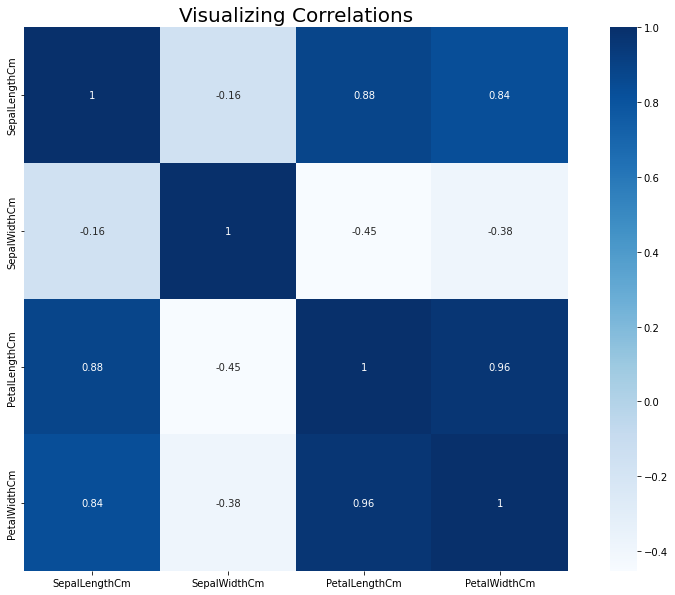


Then we split the data using stratified train-test split.

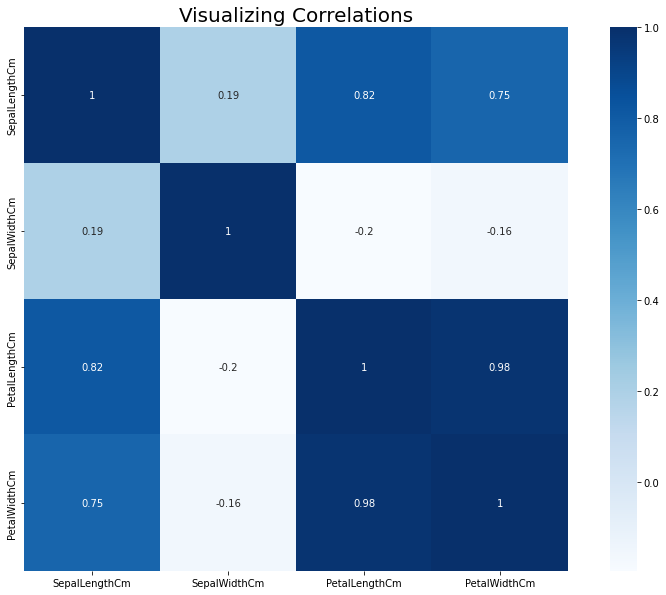
**Heat map:**

We use a heat map to describe the correlation between the features.

Train dataset:



Test dataset:

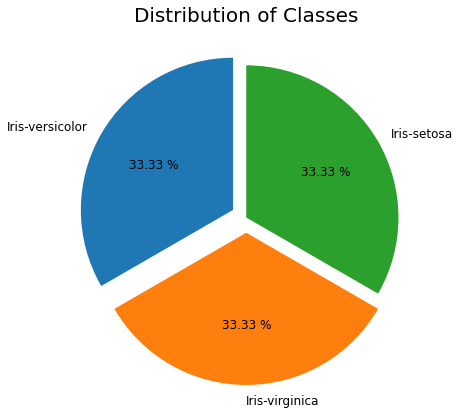


**Observations**:

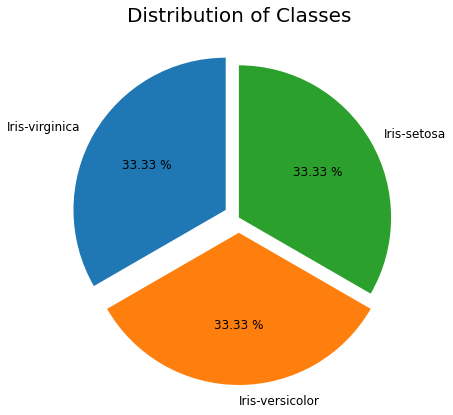
* In the train data set petal length and petal width are highly correlated and sepal length and petal length are next to them.
* In the test data set petal length and petal width are highly correlated.

**Pie Chart:**

Train dataset:



Train dataset:

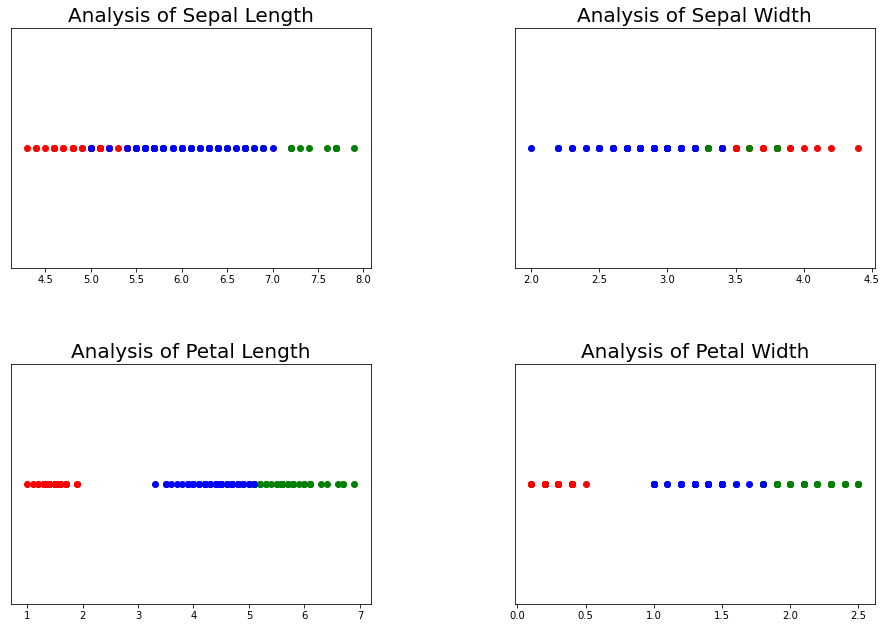


**Conclusions:**

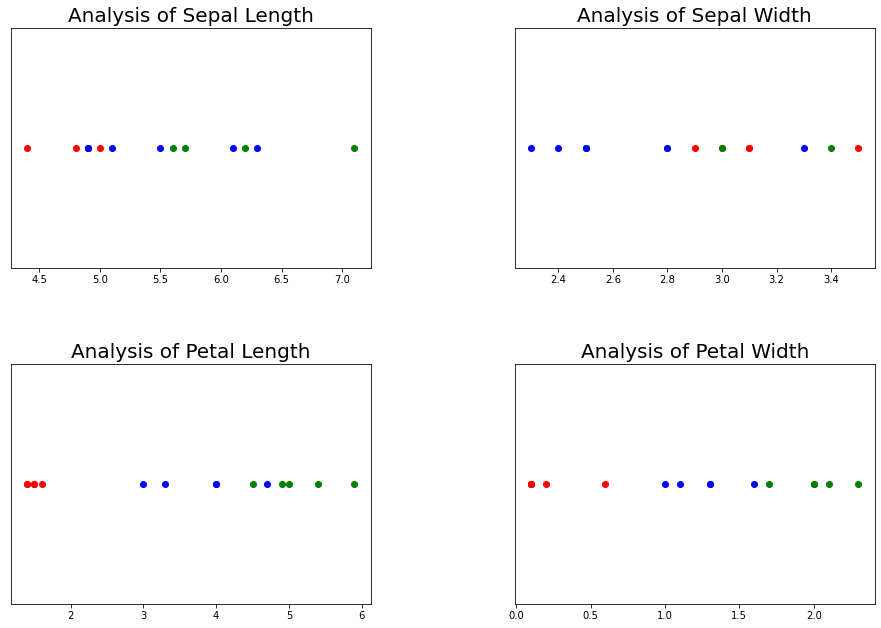
* Pie charts show that the data points are equally distributed for each class in both train and test data sets.

**Univariate Analysis:**

Train dataset:

****

Test dataset:



**Observation:**

Train dataset

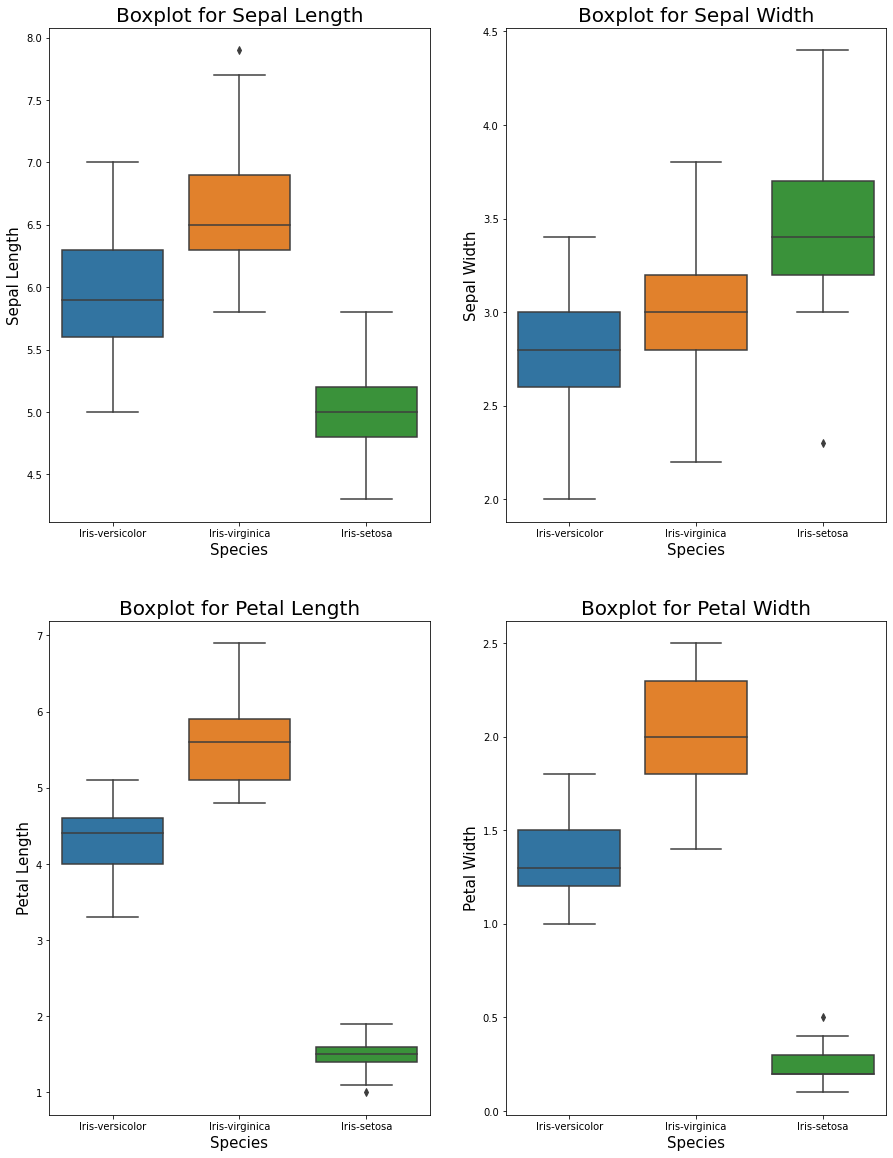
* SepalLengthCm: Iris-setosa has the smallest sepal length. We can see two outliers (red dots) in the graph. Iris-versicolor has a sepal length of about 5 cm to 7 cm. Iris-virginica has the largest sepal length (above 7 cm)
* SepalWidthCm: Iris-versicolor has the smallest sepal width. The distinction between setosa and virginia is not so prominent in the range of 3.5 cm to 4 cm. However, for a sepal width equal to or greater than 4 cm, all the flowers belong to the Iris-setosa species.
* PetalLengthCm: Iris-setosa has the smallest petal length. The length of the petals does not exceed 2 cm. For Iris-versicolor the petal length is in the range of 3 cm to 5 cm. Iris-virginica has the largest petal length (5 cm or greater).
* PetalWidthCm: Iris-setosa has the smallest petal width. Iris-versicolor has a petal length from 1 cm to slightly less than 2 cm. Iris-virginica has a petal width that is approximately greater than 1.8 cm.

Test dataset

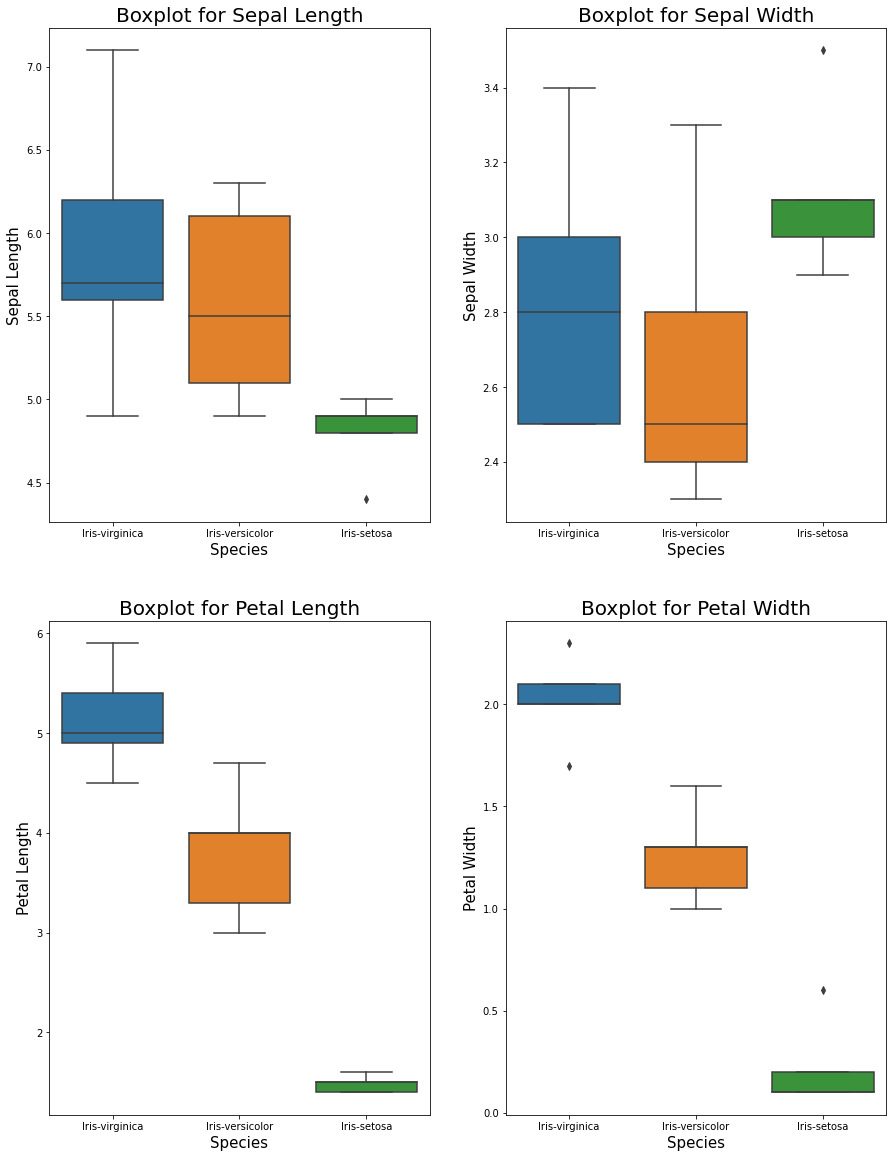
* SepalLengthCm: Iris-setosa has the smallest sepal length. We can see one outlier (red dot) in the graph. Iris-versicolor has a sepal length of about 4.5 cm to 6.5 cm. Iris-virginica has the largest sepal length (above 7 cm)
* SepalWidthCm: Iris-versicolor has the smallest sepal width. The distinction between setosa and virginica is prominent in the range of 3 cm to 3.5 cm.
* PetalLengthCm: Iris-setosa has the smallest petal length. The length of the petals do not exceed 2 cm. For Iris-versicolor the petal length is in the range of 3 cm to 4.5 cm. Iris-virginica has the largest petal length (5 cm or greater).
* PetalWidthCm: Iris-setosa has the smallest petal width. Iris-versicolor has a petal length from 1 cm to less than 2 cm. Iris-virginica has a petal width that is approximately greater than 1.8 cm.

**Box** **plot:**

Train dataset



Test dataset



**Observations:**

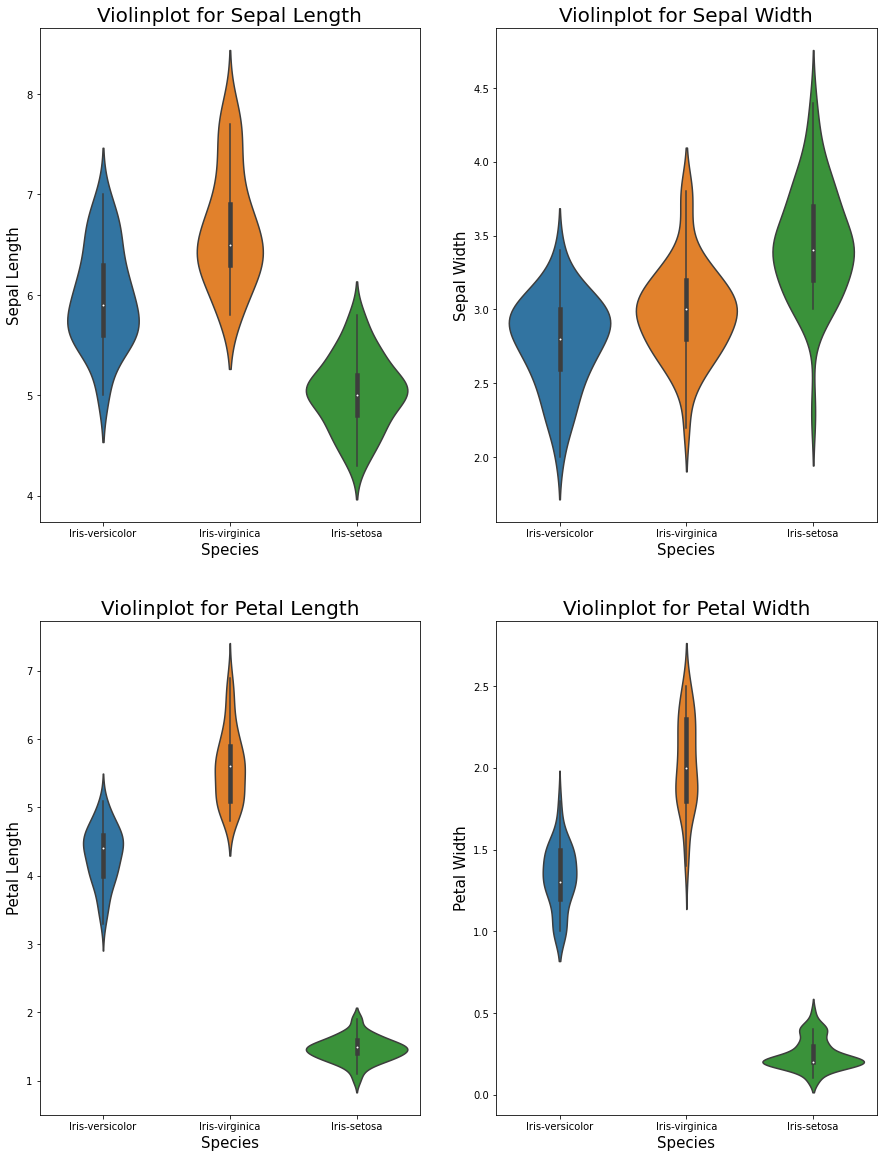
Train dataset

* SepalLengthCm: The median value for sepal length is the least for Iris-setosa and the maximum for Iris-virginica. We have an outlier for the Iris-setosa
* SepalWidthCm: The median value for sepal width is the least for Iris-versicolor and the largest for Iris-setosa. We have an outlier for Iris-setosa less than 2.5.
* PetalLengthCm: Visually it is very evident that Iris-setosa has the least petal length. Even though there is an outlier less than 2 cm. Iris-versicolor has the second-largest median petal length, whereas Iris-virginica has the largest median petal length. We have an outlier in Iris-setosa.
* PetalWidthCm: Similar to the petal lengths of the species, we can see that Iris-setosa has the lowest median petal width, whereas Iris-virginica has the largest median petal width. We have an outlier in Iris-setosa greater than 0.5.

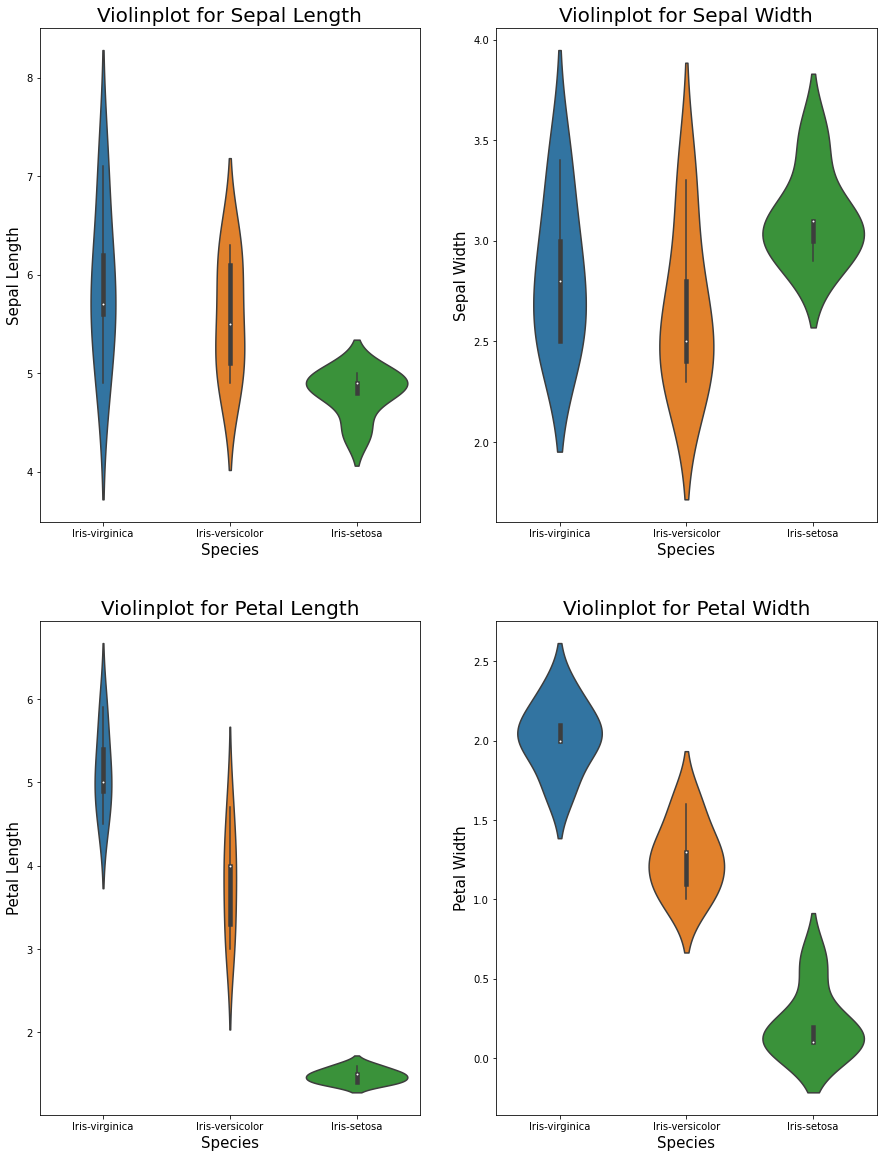
Test dataset

* SepalLengthCm: The median value for sepal length is the least for Iris-setosa and the maximum for Iris-virginica. We have an outlier for the Iris-setosa
* SepalWidthCm: The median value for sepal width is the least for Iris-versicolor and the largest for Iris-setosa. We have an outlier for Iris-setosa greater than 3.4.
* PetalLengthCm: Visually it is very evident that Iris-setosa has the least petal length. Iris-versicolor has the second-largest median petal length, whereas Iris-virginica has the largest median petal length.
* PetalWidthCm: Similar to the petal lengths of the species, we can see that Iris-setosa has the lowest median petal width, whereas Iris-virginica has the largest median petal width. We have an outlier in Iris-setosa greater than 0.5. We have two outliers for Iris-virginica.

**Violin Plot:**

Train dataset

Test dataset



**Observations:**

Train dataset

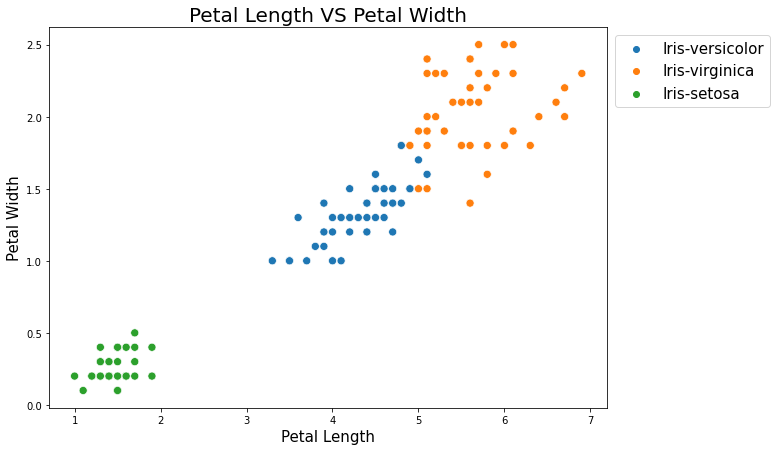
* SepalLengthCm: There is a high probability that Iris-setosa will have a sepal length of 5 cm.
* SepalWidthCm: For Iris-versicolor and Iris-virginica there is a high probability that they will have approximately the same sepal width. Thus identification of the species based on this feature only might not yield good results.
* PetalLengthCm: Most of the petal lengths for Iris-setosa are about 1.5 cm.
* PetalWidthCm: There is a high probability that the petal width for Iris-setosa species would be approximately 0.25 cm. The violin plots for Iris-versicolor and Iris-virginica are not as broad as that of Iris-setosa.

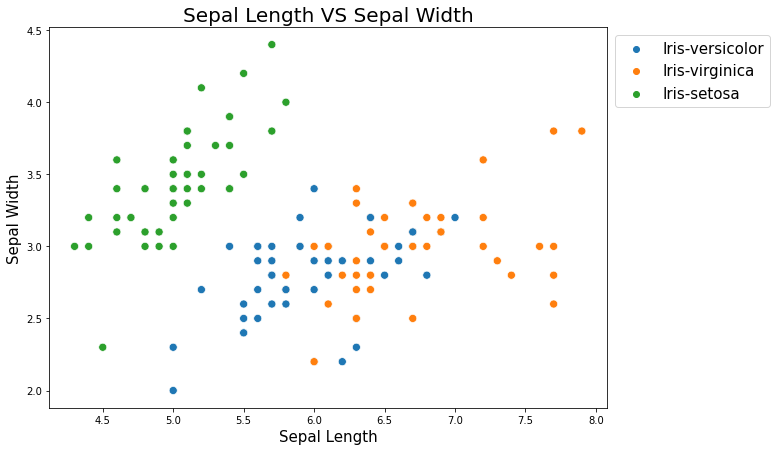
Test dataset

* SepalLengthCm: There is a high probability that Iris-setosa will have a sepal length of 4.8 cm.
* SepalWidthCm: For Iris-versicolor and Iris-virginica there is a high probability that they will have approximately the same sepal width. Thus identification of the species based on this feature only might not yield good results. There is a high probability that Iris-setosa will have sepal width of 3 cm.
* PetalLengthCm: Most of the petal lengths for Iris-setosa are about 1.5 cm.
* PetalWidthCm: There is a high probability that the petal width for Iris-setosa species would be approximately 0.25 cm. The violin plots for Iris-versicolor and Iris-virginica are not as broad as that of Iris-setosa.

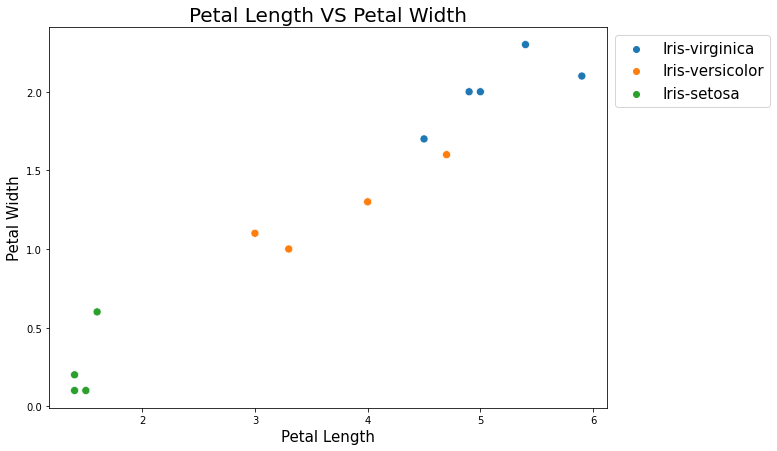
**Bivariate Analysis:**

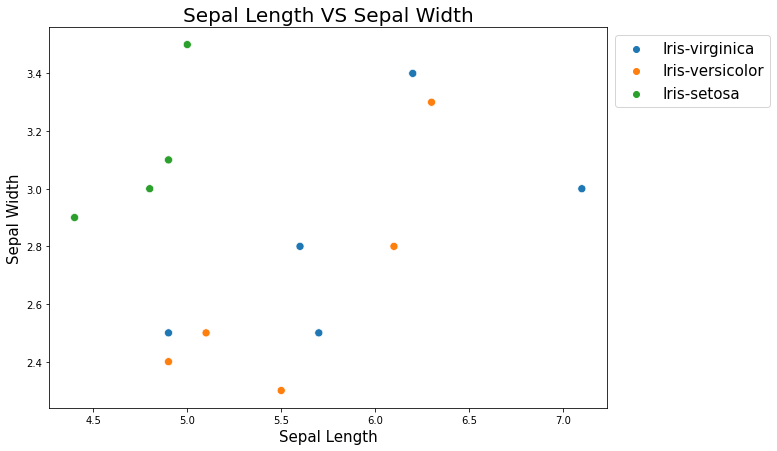
Train dataset





Test dataset

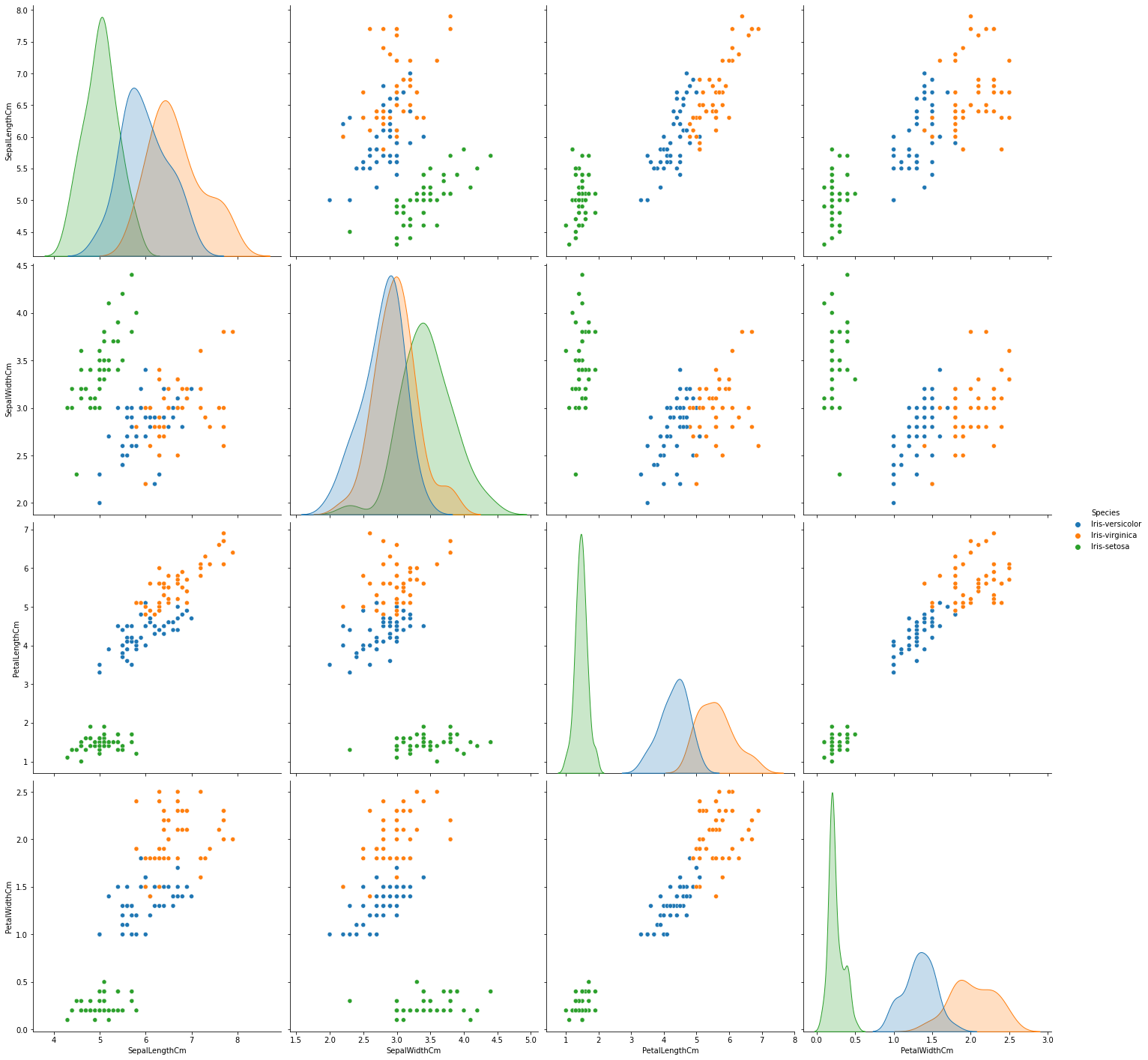


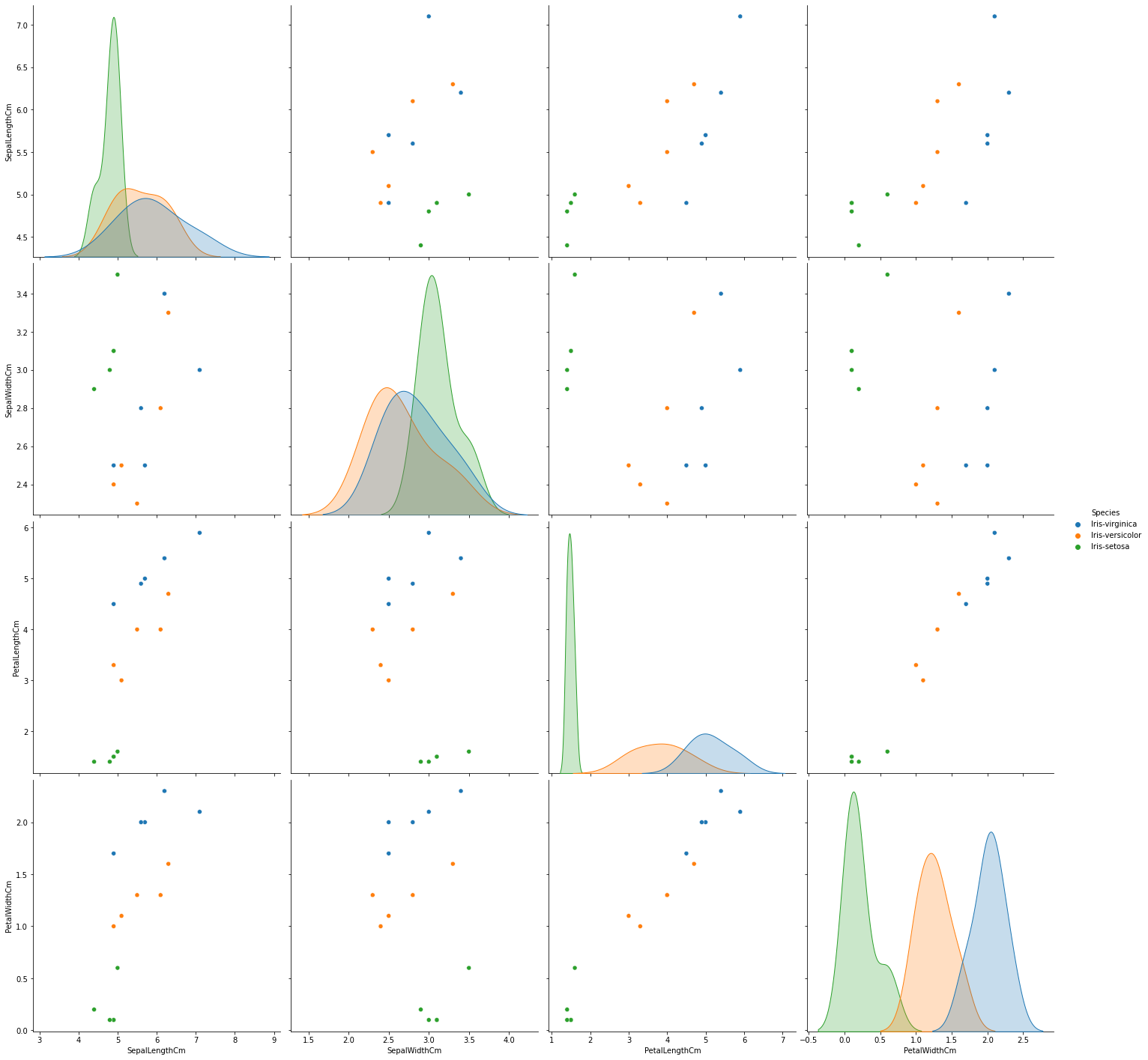


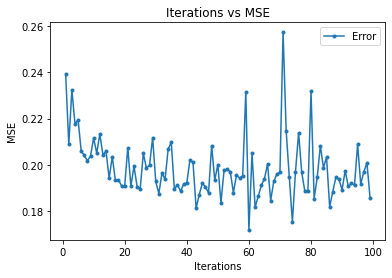
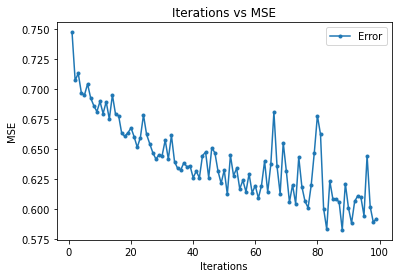
**Observations:**

* We can see that the petal length and petal width is the least for Iris-setosa. The petal width and length for Iris-versicolor lies in an intermediate range, between that of setosa and virginica. Iris-virginica has the largest petal length and width. A few outliers exist in the case of both versicolor and virginica.
* We can observe that Iris-setosa has a relatively lower sepal length (as compared to versicolor or virginica). Iris-setosa has a large sepal width. It is hard to distinguish between versicolor and virginica based on sepal length and width.

**Multivariate Analysis:**

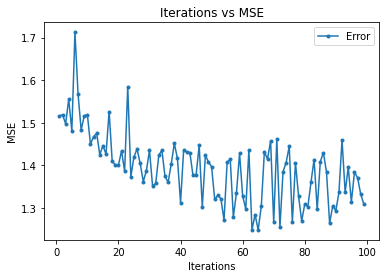
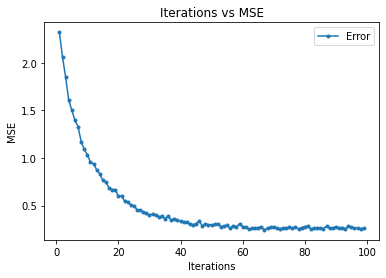
Train dataset

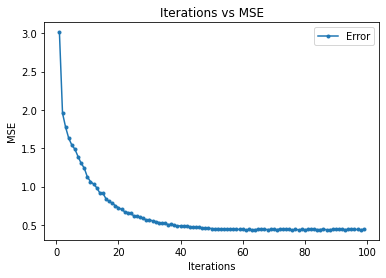
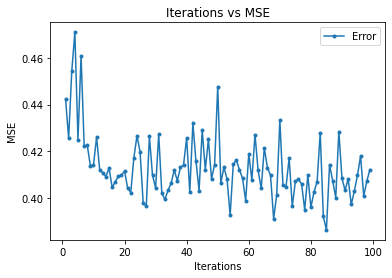
Test dataset

**Linear Regression (Stochastic gradient regressor):**

**Sepal width and Sepal Length**

**Sepal length and Sepal Width**



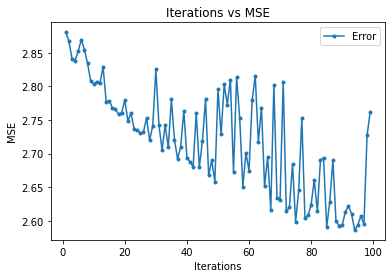
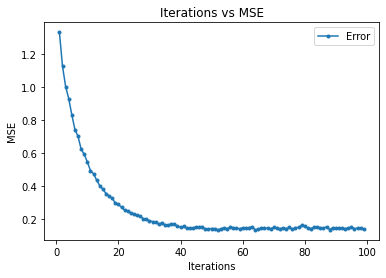


**Petal length and Sepal length**

**Sepal length and Petal length**

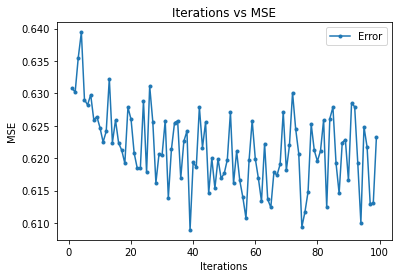
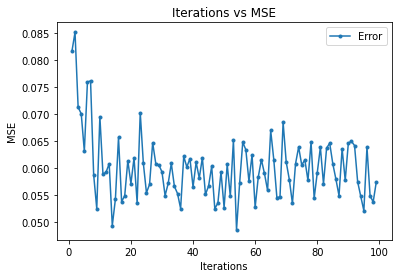
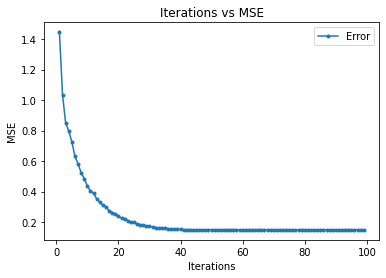
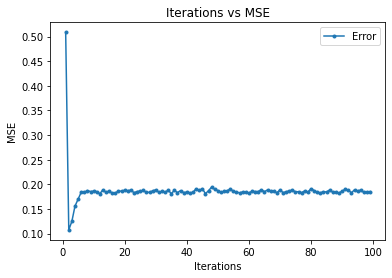
**Sepal length and Petal width**

**Petal width and Sepal length**



**Sepal width and Petal length**

**Petal Length and Sepal width**



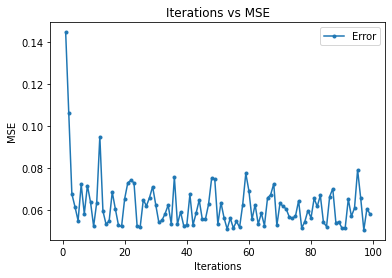
**Petal width and Petal length**

**Petal length and Petal width**

**Petal width and Sepal width**

**Sepal width and Petal width**

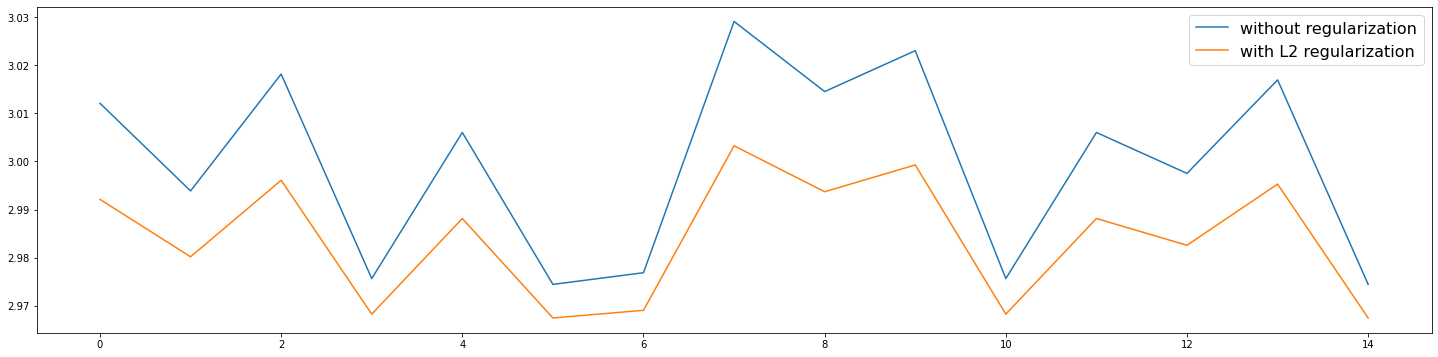
Main Model (Predicting species with the help of features)



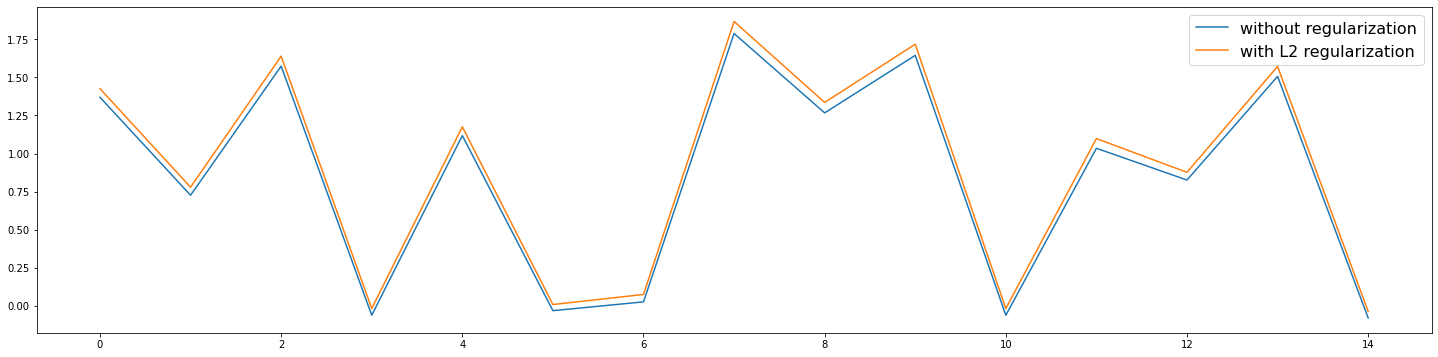
* Out of 12 models model 11 i.e., predicting the petal width using petal length have less mean square error of 0.04681 for 53 iterations (steps).
* For the main model mean square error is 0.05059 for 96 iterations (steps).

**Effect of L2 regularization on Stochastic gradient regressor:**

I selected model 8 for comparison and the model with iteration 55 is

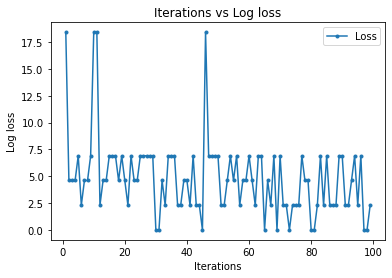
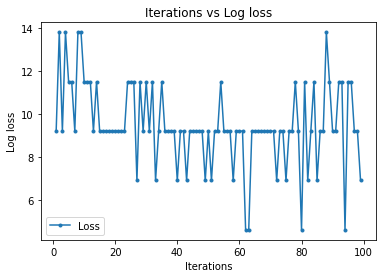
identical one we apply L2 regularization for that model

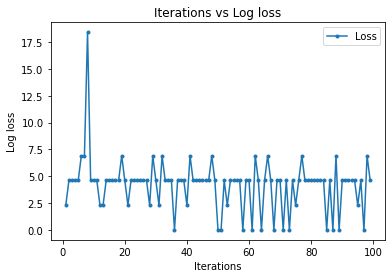
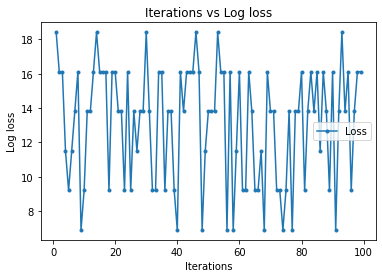
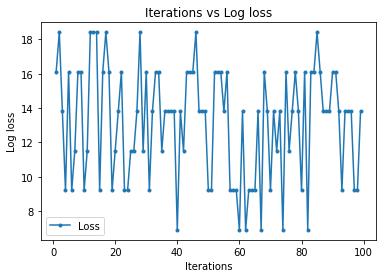
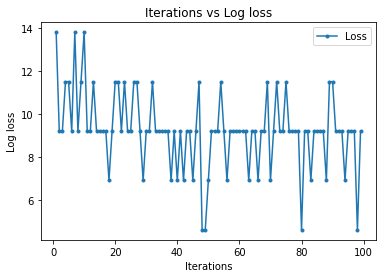
|  |  |  |
| --- | --- | --- |
| **Model** | **Slope** | **Intercept** |
| Without L2 Regularization | 0.01214738 | 2.95742216 |
| With L2 Regularization | 0.00795337 | 2.95631389 |

**Main model:**

**Conclusion:**

* Without L2 regularization stochastic regressor coefficient is 0.01214738 and intercept is 2.95742216.
* With L2 regularization stochastic regressor coefficient is 0.00795337.
* For main model without L2 regularization coefficients are [-0.03605329 -0.12314695 0.33905776 0.21356018] and intercept is -0.03464323
* For main model without L2 regularization coefficients are [-0.02803606 -0.12170901 0.34016686 0.21995688] and intercept is -0.03664186.
* Because of L2 regularization the mean square error is decreasing.

**Logistic Regression (Stochastic gradient classifier):**



**Model 6**

**Model 5**

**Model 4**

**Model 2**

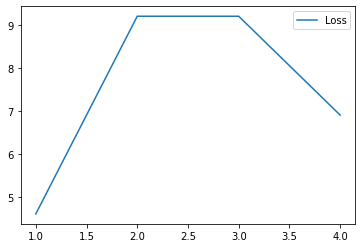
**Model 3**

**Model 1**

**Observations:**

* Model 6 has less log loss of 4.605170185988091 for 47 iterations.

**Effect of Learning rate on model convergence:**

****

* After training the model with learning rates α= 0.0001,0.001,0.01 and 0.1 we can clearly state that the loss for the model having α=0.0001 is less than other α values
* For less α values it takes more time to converge and for high α values it takes less time to converge.

**Conclusions:**

Linear regression

* Out of 12 models model 11 i.e., predicting the petal width using petal length have less mean square error of 0.04681 for 53 iterations (steps).
* For the main model mean square error is 0.05059 for 96 iterations (steps).
* Without L2 regularization stochastic regressor coefficient is 0.01214738 and intercept is 2.95742216.
* With L2 regularization stochastic regressor coefficient is 0.00795337.
* For main model without L2 regularization coefficients are [-0.03605329 -0.12314695 0.33905776 0.21356018] and intercept is -0.03464323
* For main model without L2 regularization coefficients are [-0.02803606 -0.12170901 0.34016686 0.21995688] and intercept is -0.03664186.
* Because of L2 regularization the mean square error is decreasing.

Logistic regression

* Model 6 has less log loss of 4.605170185988091 for 47 iterations.
* After training the model with learning rates α= 0.0001,0.001,0.01 and 0.1 we can clearly state that the loss for the model having α=0.0001 is less than other α values
* For less α values it takes more time to converge and for high α values it takes less time to converge.