# IC 272: DATA SCIENCE - III
# LAB ASSIGNMENT – V
## Data classification using Bayes classifier with Gaussian mixture model (GMM); regression using linear regression and polynomial curve fitting

**Student's Name: Mayank Bansal** **Mobile No: +919636993445**

**Roll Number: B20156** **Branch:CSE**

## PART - A

**1 a.**

| | Prediction Outcome | |
|---|---|---|
| True Label | 95 | 13 |
| | 3 | 225 |

**Figure 1 Bayes GMM Confusion Matrix for Q = 2**

| | Prediction Outcome | |
|---|---|---|
| True Label | 95 | 13 |
| | 4 | 224 |

**Figure 2 Bayes GMM Confusion Matrix for Q = 4**

| | Prediction Outcome | |
|---|---|---|
| True Label | 87 | 21 |
| | 4 | 224 |

**Figure 3 Bayes GMM Confusion Matrix for Q = 8**

| | **Prediction Outcome** | |
|---|---|---|
| **True Label** | 81 | 27 |
| | 1 | 227 |

**Figure 4 Bayes GMM Confusion Matrix for Q = 16**

**b.**

**Table 1 Bayes GMM Classification Accuracy for Q = 2, 4, 8 & 16**

| **Q** | **Classification Accuracy (in %)** |
|---|---|
| 2 | **0.952** |
| 4 | **0.949** |
| 8 | **0.926** |
| 16 | **0.917** |

**Inferences:**

1. The highest classification accuracy is obtained with Q =2.
2. Prediction accuracy decreases on a regular basis by increasing Q.
3. Adding nodes with less weights can cause problems for model.
4. Number of diagonal elements increase with increase in accuracy
5. Diagonal elements correspond to correct predictions. Hence increase in correct predictions leads to increase in accuracy.
6. Number of off diagonal elements increase with increase in value of Q.
7. Increase in value of Q decrease the value of accuracy, hence leading to an increase in off diagonal elements.

**2**

**Table 2 Comparison between Classifiers based upon Classification Accuracy**

| S. No. | Classifier | Accuracy (in %) |
|---|---|---|
| 1. | KNN | 0.896 |
| 2. | KNN on normalized data | 1 |
| 3. | Bayes using unimodal Gaussian density | 0.9375 |
| 4. | Bayes using GMM | 0.952 |

**Inferences:**

1. KNN Normalized and KNN have the highest and lowest accuracy respectively.
2. KNN< Unimodal Gaussian Bayes Model < Multimodal Gaussian Bayes Model< KNN Normalized
3. KNN Normalized explores the neighborhood and considers all the attributes equally leading to best accuracy. GMM Bayes Model tries to make multiple clusters to judge the pattern which is better than Unimodal Bayes Model which tries to accommodate all elements in one single cluster. KNN has least accuracy because attributes in this model can overpower each other.

## PART – B

**1 a.**

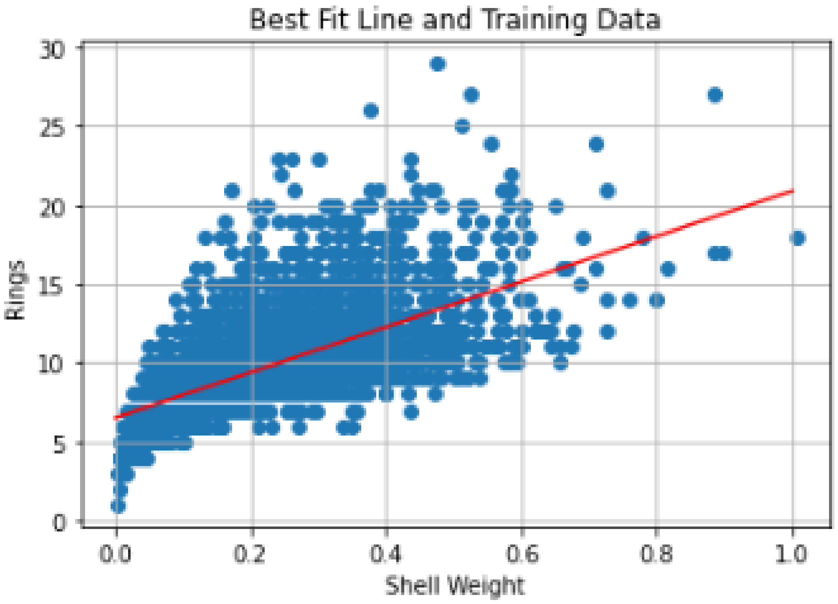

**Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data**

**Inferences:**

1. Shell weight was used in this case because Shell weight has the most say on the value of Rings. Both attributes are dependent on each other very much.
2. No, the best line doesn't fit the data properly.
3. Because the line can't recognize the curve in the data spread.
4. Bias is high and variance is low.

**b.**

Prediction accuracy on training data=2.528.

**c.**

prediction accuracy on testing data=2.468.

**Inferences:**

1. Training data has higher accuracy.
2. Because the model is made on training data, it will be biased in favor of training data.

**d.**

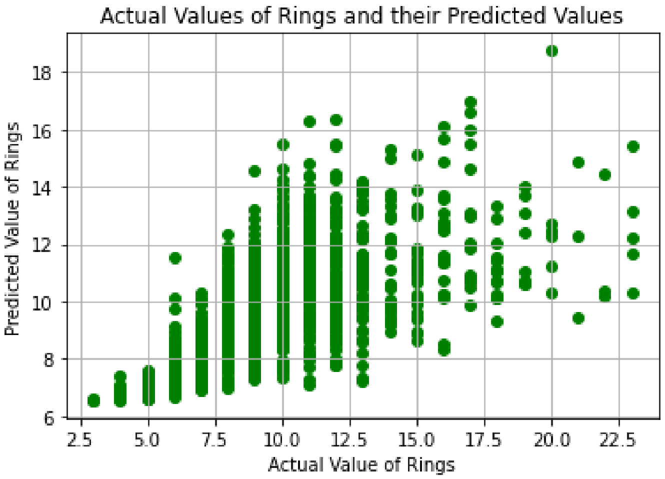

**Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data**

**Inferences:**

1. It is not very accurate.
2. Range of the actual data and predicted data is different as seen from the graph.

**2**

**a.**

Prediction accuracy on training data=2.216.

**b.**

Prediction accuracy on testing data=2.219.

**Inferences:**

3. Accuracy of testing data is on higher side but it just has a difference of 0.003 which is very less. Hence both can be considered same.
4. The Multivariate model considers whole data frame as the input making the model very good. Hence RMSE is less and almost same for both train and test data.

**c.**

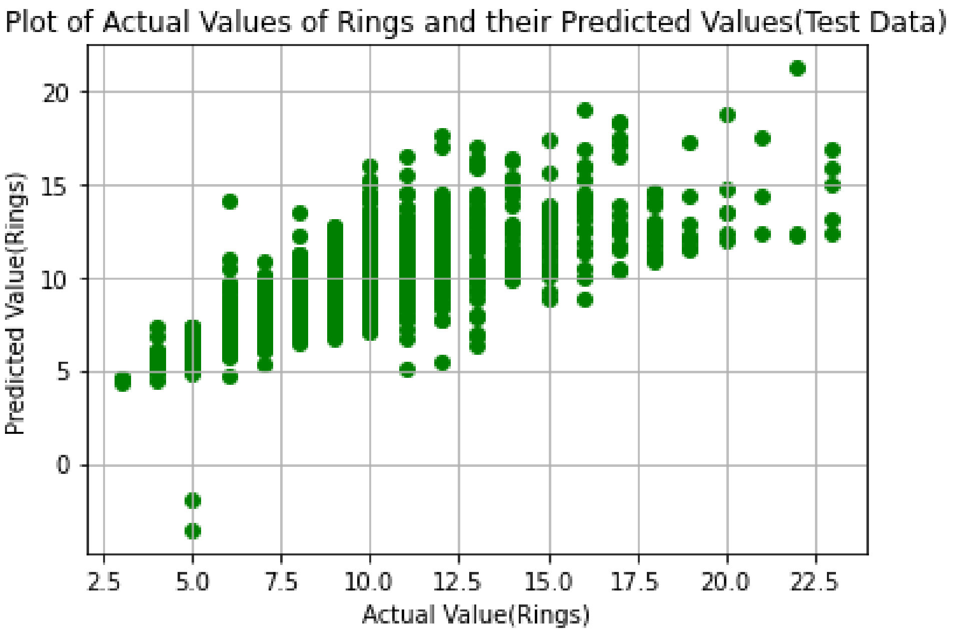

**Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data**

**Inferences:**

1. Based on the spread of data, the data is moderately accurate.
2. It is moderately accurate because the spread is inclined at an almost 45 degree angle.
3. Multivariate linear regression is more accurate than univariate regression because RMSE is less in case of Multivariate regression.

**3**

**a.**

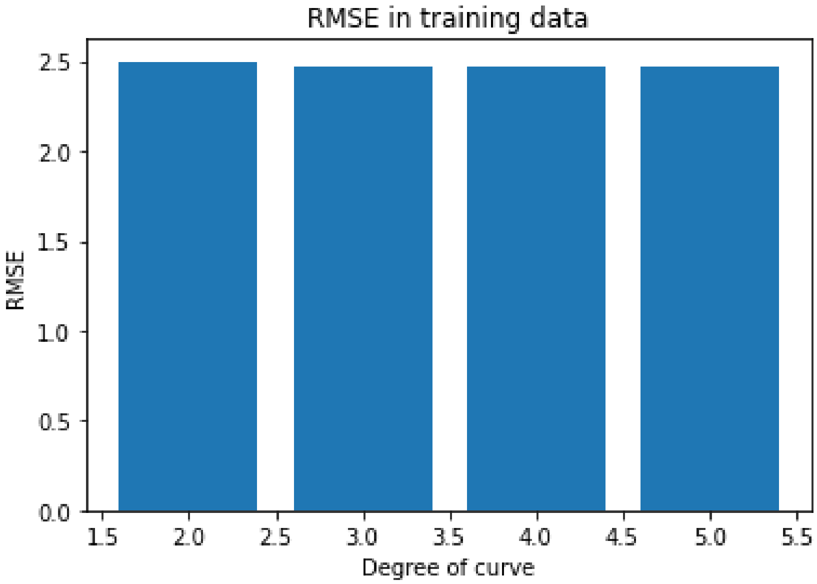

**Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data**

**Inferences:**

1. RMSE value decreases with increase in degree.
2. The decrease is so less that it can be ignored eventually.
3. As the degree increases the curve fits the data better so RMSE decreases.
4. Degree = 5 will approximate the data best
5. As the degree increases, the bias decreases and variance increases.

**b.**

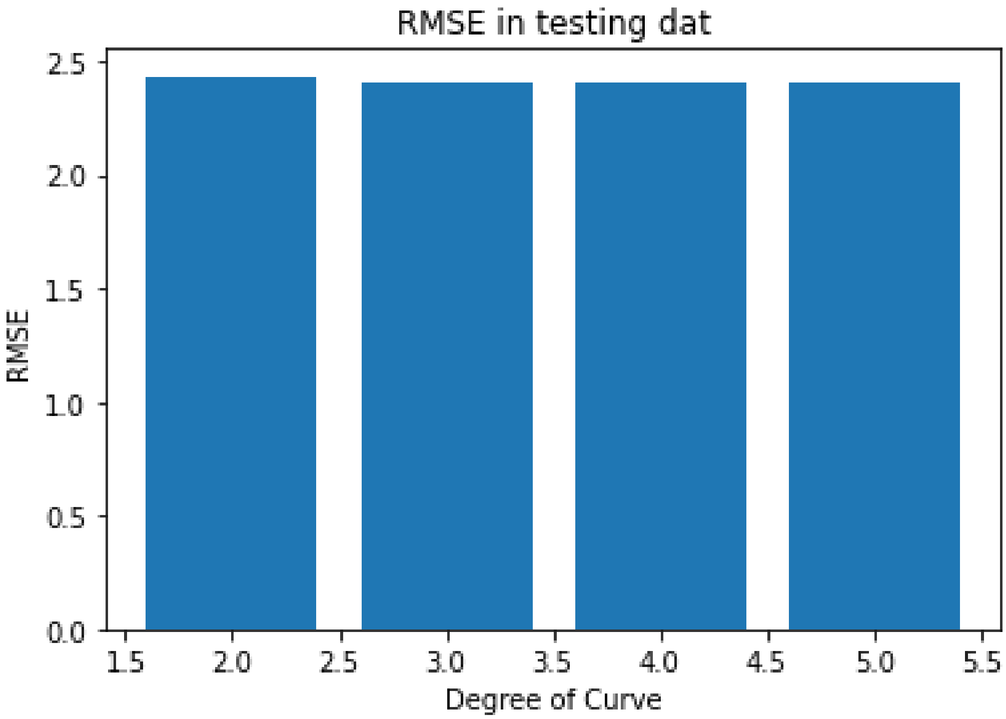

**Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data**

**Inferences:**

1. RMSE decrease with increase in value of degree.
2. The decrease is so less that it can be ignored eventually.
3. As the degree increases the curve fits the data better so RMSE decreases.
4. Degree = 4 will approximate the data best.
5. As the degree increases, the bias decreases and variance increases.

**C.**

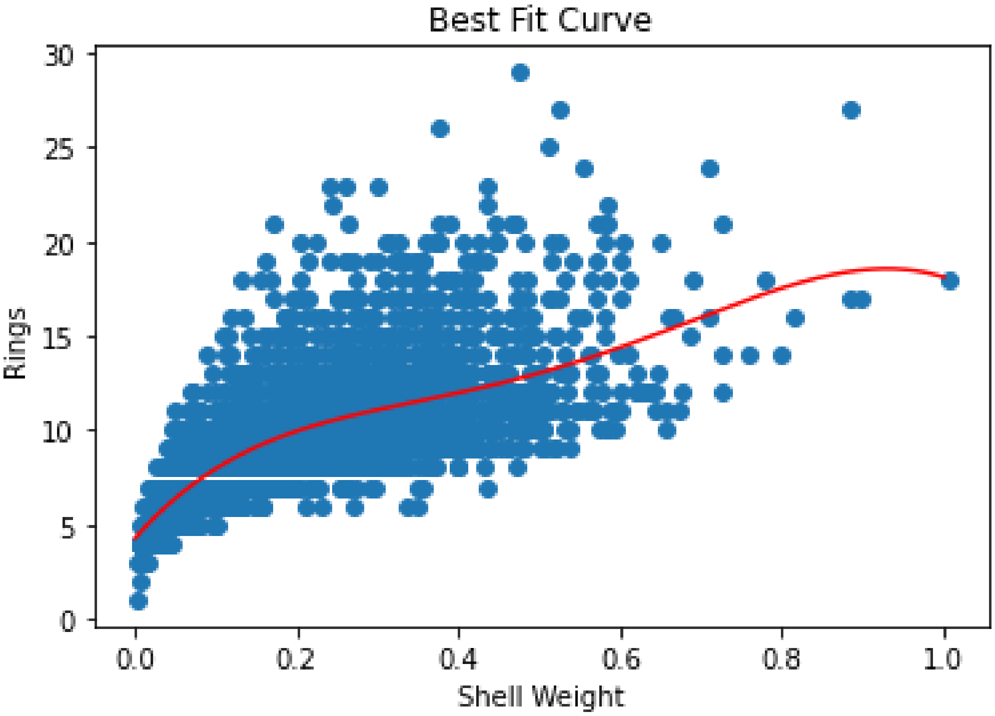

**Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data**

**Inferences:**

1. p-value= 4 corresponding to the best fit model.
2. It can predict data more accurately.
3. The bias decreases and variance increases with increasing value of p.

**d.**

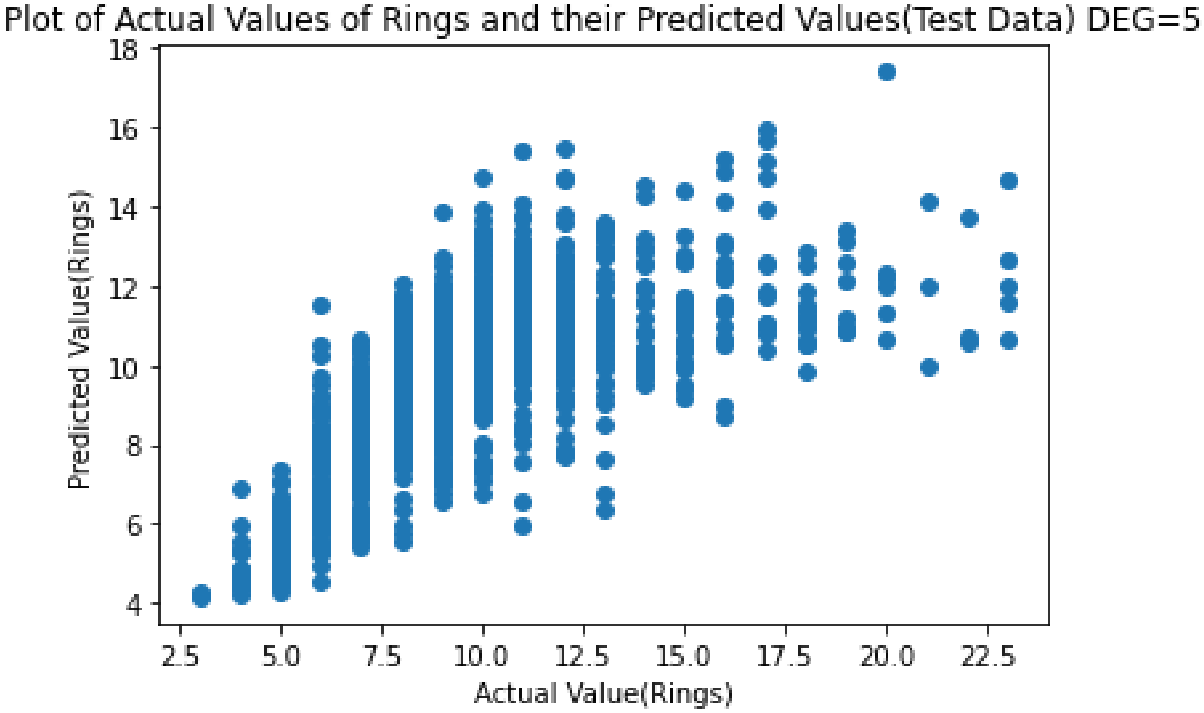

**Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data**

**Inferences:**

1. The predicted data is moderately accurate.
2. The Range is different of actual and predicted data as seen from graphs.
3. Accuracies follow the given order: univariate linear < multivariate linear <non-linear regression model
4. RMSE value follow the reverse order hence less RMSE means more accuracy.
   In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high.

**4 a.**

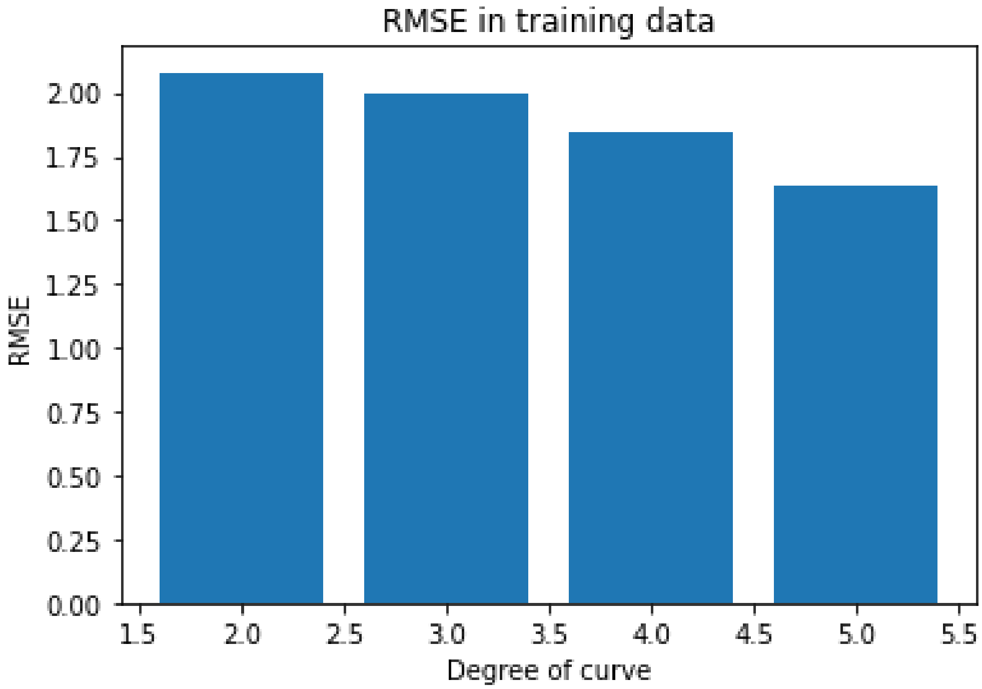

**Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data**

**Inferences:**

1. RMSE value decreases with increase in degree.
2. The decrease is almost same for all degrees except from deg=4 to 5 the decrease is more.
3. Increase in degree makes the curve to accommodate the data more accurately.
4. Degree = 5 will approximate the data best.
5. The bias decreases and variance increases with respect to the increase in the degree

**b.**

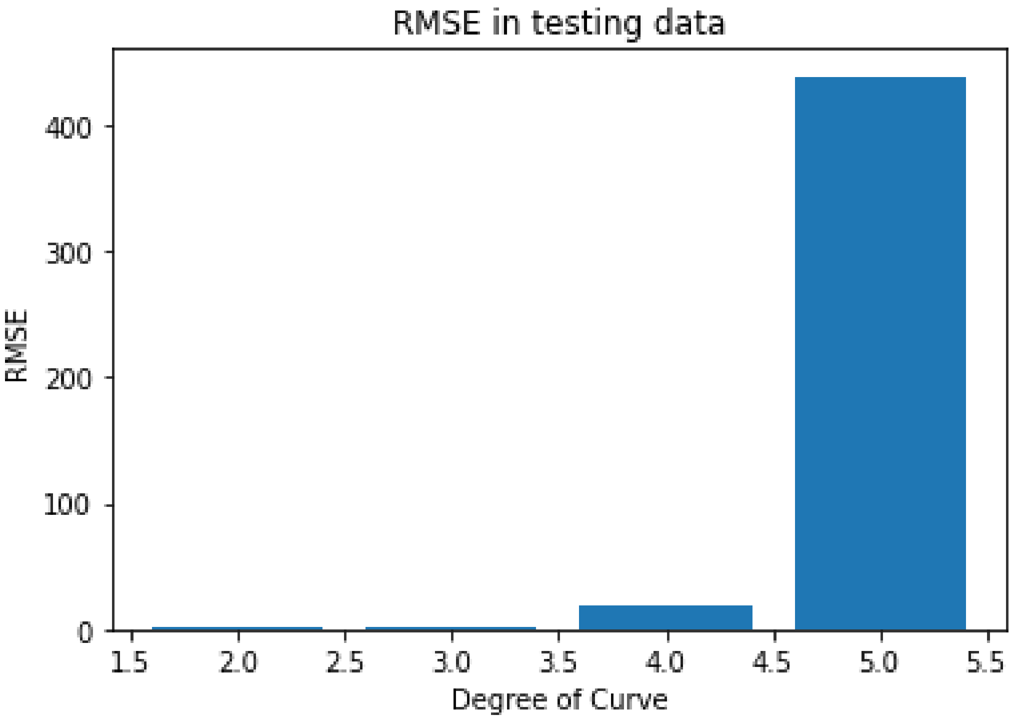

**Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data**

**Inferences:**

1. RMSE value increases with increase in degree.
2. The increase is non uniform but increase after degree=4 is sudden.
3. Increase in degree leads makes our model imperfect because our model makes our overfitted.
4. Degree = 2 will approximate the data best.
   The bias gradually decreases 6ll p=3 and then suddenly increases after p=3 and the variance increases as the model becomes more complex with increasing degree of polynomial.

**c.**

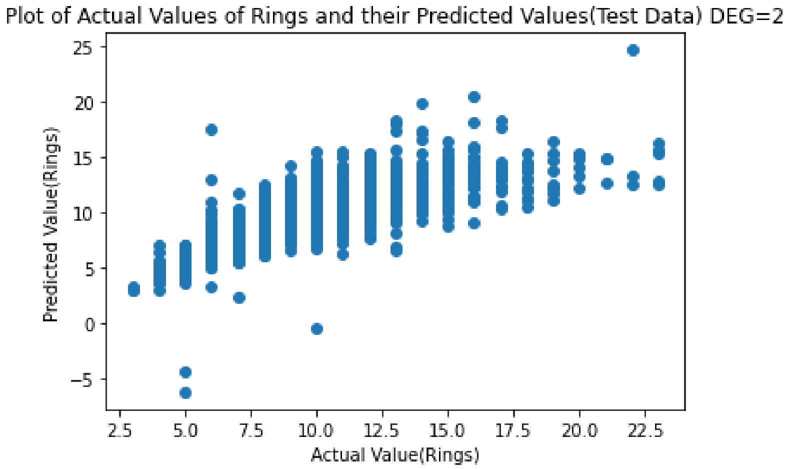

**Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data**

**Inferences:**

1. Prediction is accurate.
2. Range is same and the spread is inclined at 45 degrees.
3. Accuracies follow the given order: univariate linear < multivariate linear < univariate non-linear regression model< multivariate nonlinear regression model.
4. RMSE values follow the reverse order.
5. In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high