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: Madhur Jajoo **Mobile No: 7597389137**

Roll Number: B20211 **Branch:DSE**

PART - A

1 a.

		Prediction Outcome
True Label	106	12
	6	213

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	111	7
	5	214

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	105	13
	4	215

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	90	28
	2	217

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	**94.659**
4	**96.439**
8	**95.252**
16	**91.098**

Inferences:

1. The highest classification accuracy is obtained with Q =4.
2. Initially the value of accuracy increases with increase in value of q but later on it decreases with increase in value of q.
3. As we increase the value of q the complexity of the model increases resulting in decreased accuracy.

4. As the accuracy decreases the number of diagonal elements in the confusion matrix decrease.
5. Diagonal elements represents the number of times the model predicted correct, as the accuracy decreases i.e. our model is predicting wrong therefore the diagonal elements decreases.
6. As the accuracy decreases the number of off diagonal elements in the confusion matrix increase.
7. Off Diagonal elements represents the number of times the model predicted wrong, as the accuracy decreases i.e. our model is predicting wrong therefore the off diagonal elements increases.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	86.614
2.	KNN on normalized data	97.329
3.	Bayes using unimodal Gaussian density	94.362
4.	Bayes using GMM	96.439

Inferences:

1. KNN have the lowest accuracy while KNN on normalized data have highest accuracy.
2. KNN < Bayes using unimodal Gaussian density < Bayes using GMM< KNN on normalized data.
3. KNN performs better when data is normalized because, the attributes on a bigger scale can no longer overpower and influence the results in their favor. In the above example which involves just 2 clusters, KNN will give more accurate predictions than Bayes. Multimodal Bayes performs better as we are now using multiple clusters which increases the relative accuracy.

PART – B

1

a.

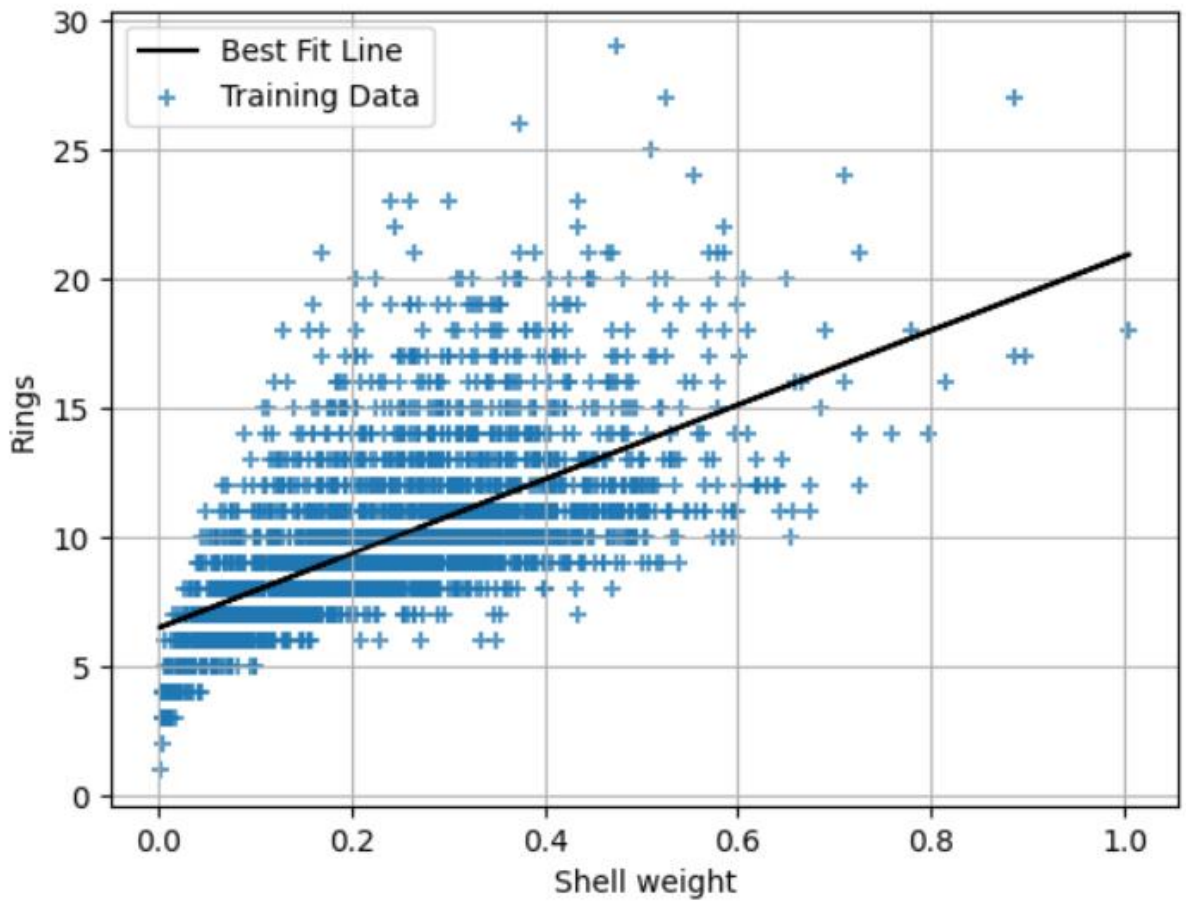

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. In univariate linear regression we use the attribute with high correlation because, because of high correlation is easier to predict the target values.
2. No, the best fit line doesn't fit perfectly.
3. It doesn't fit perfectly because it is oversimplified, a more complex curve is needed.
4. Bias is high while the variance is low for the best fit line.

b.

The prediction accuracy on training data is 2.528.

c.

The prediction accuracy on testing data is 2.466.

Inferences:

1. Training accuracy is higher
2. Training accuracy is higher because it is on the same data we trained model.

d.

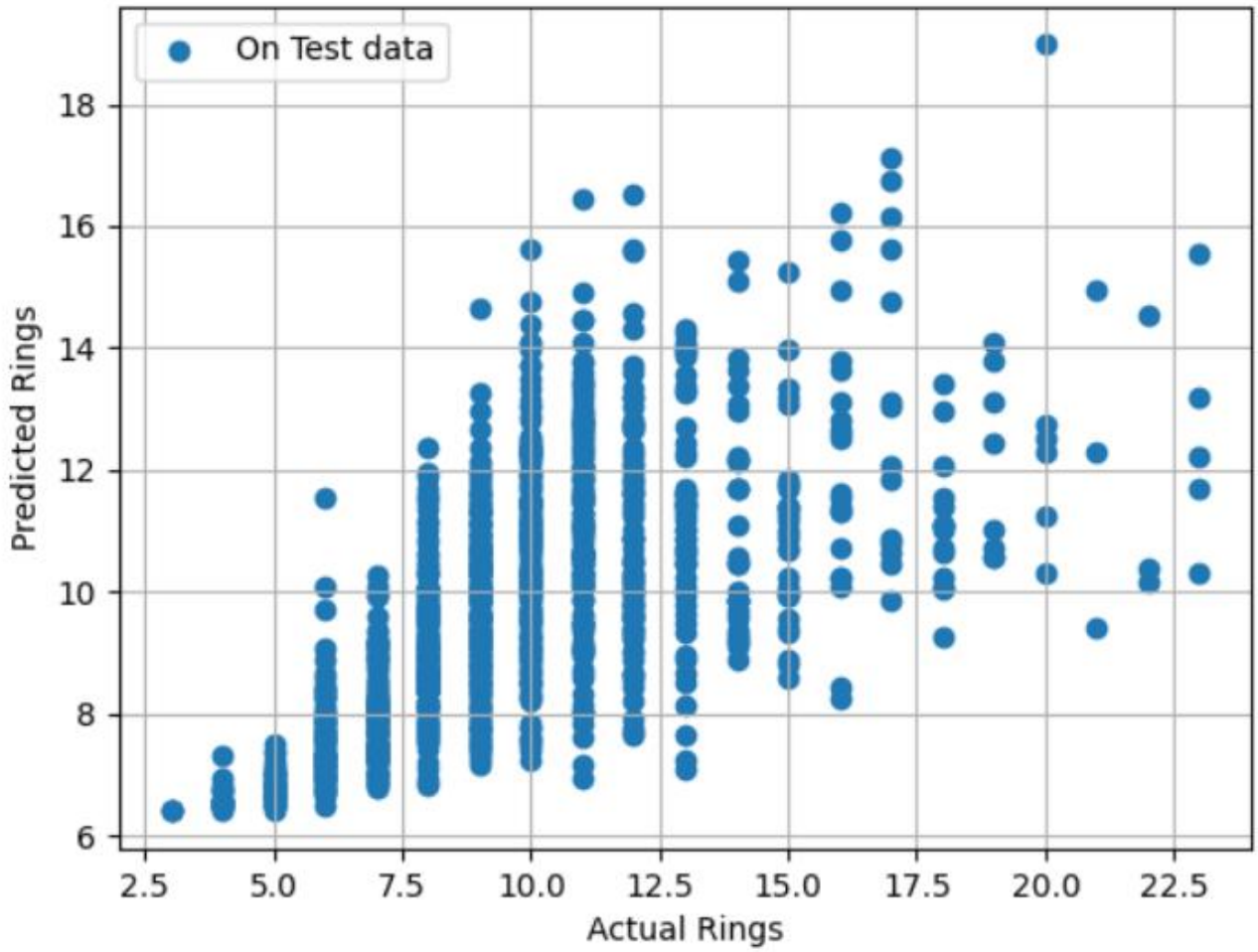

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, we didn't predicted the no. of rings accurately.
2. Because the original spread of rings is 2-23 while ours is 6-20.

2

a.

The prediction accuracy on training data is 2.216.

b.

The prediction accuracy on testing data is 2.205.

Inferences:

3. Amongst training and testing accuracies both are almost same.
4. The model is good.

c.

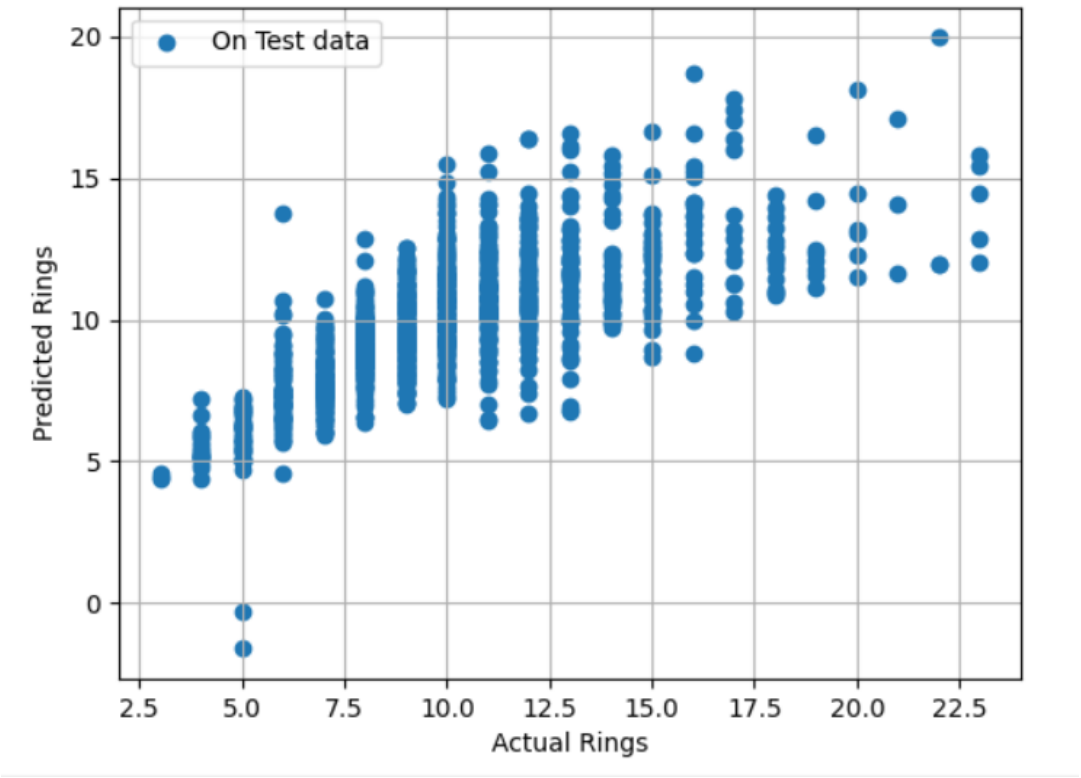

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, no. of Rings is predicted high.
2. Because the original spread of rings is 5-23 while ours is 4-22.
3. Multivariate linear regression performed better than Univariate linear regression.

3

a.

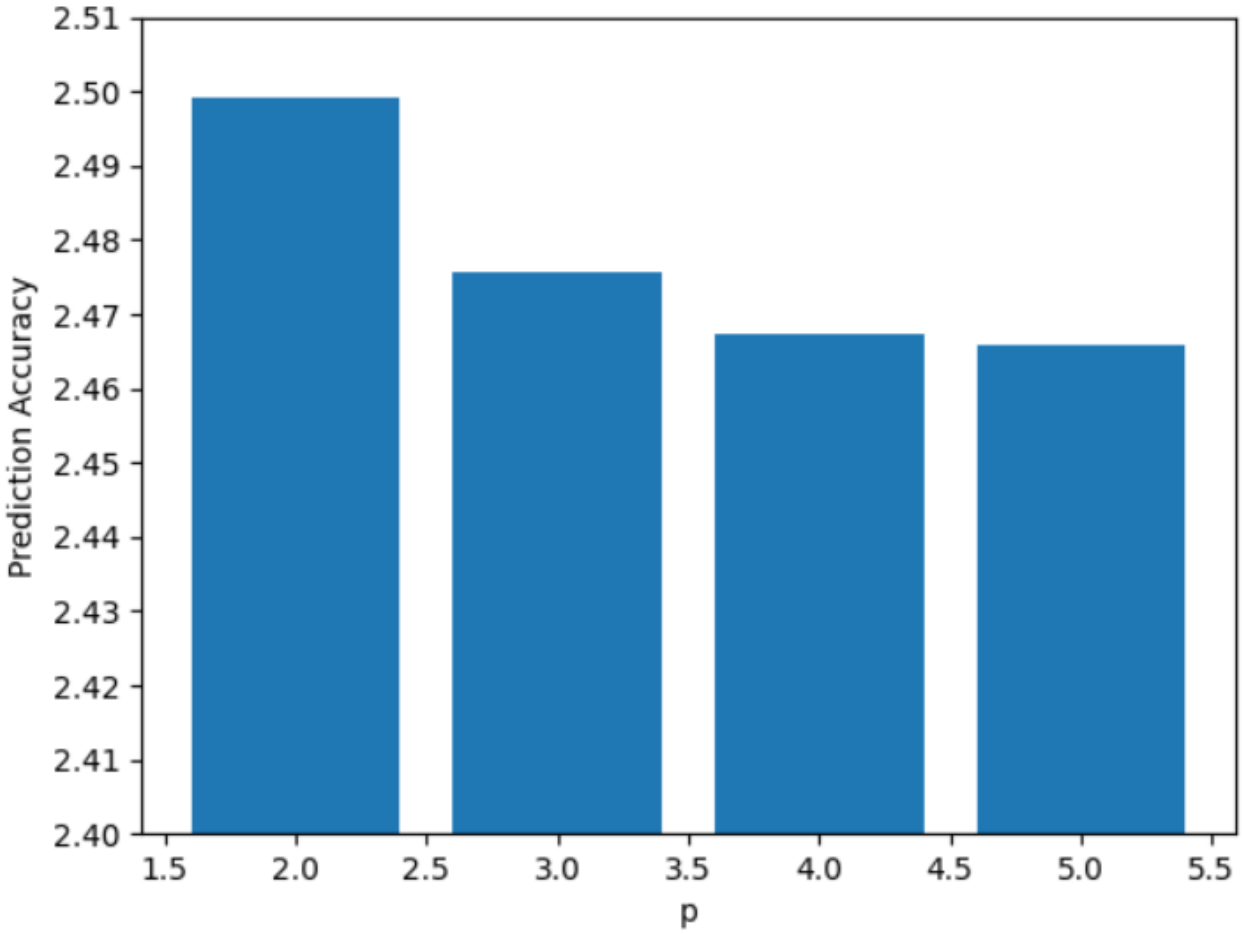

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases as the degree of the polynomial increases.
2. The decrease is more in degrees 2 to 3 and further the decreasing nature is gradual.
3. The more the degree the good the curve fits the data, as a result of this the RMSE decreases

4. From the RMSE value, p = 5 degree curve will approximate the data best.
5. The bias decreases and the variance increases as increase in degree happens.

b.

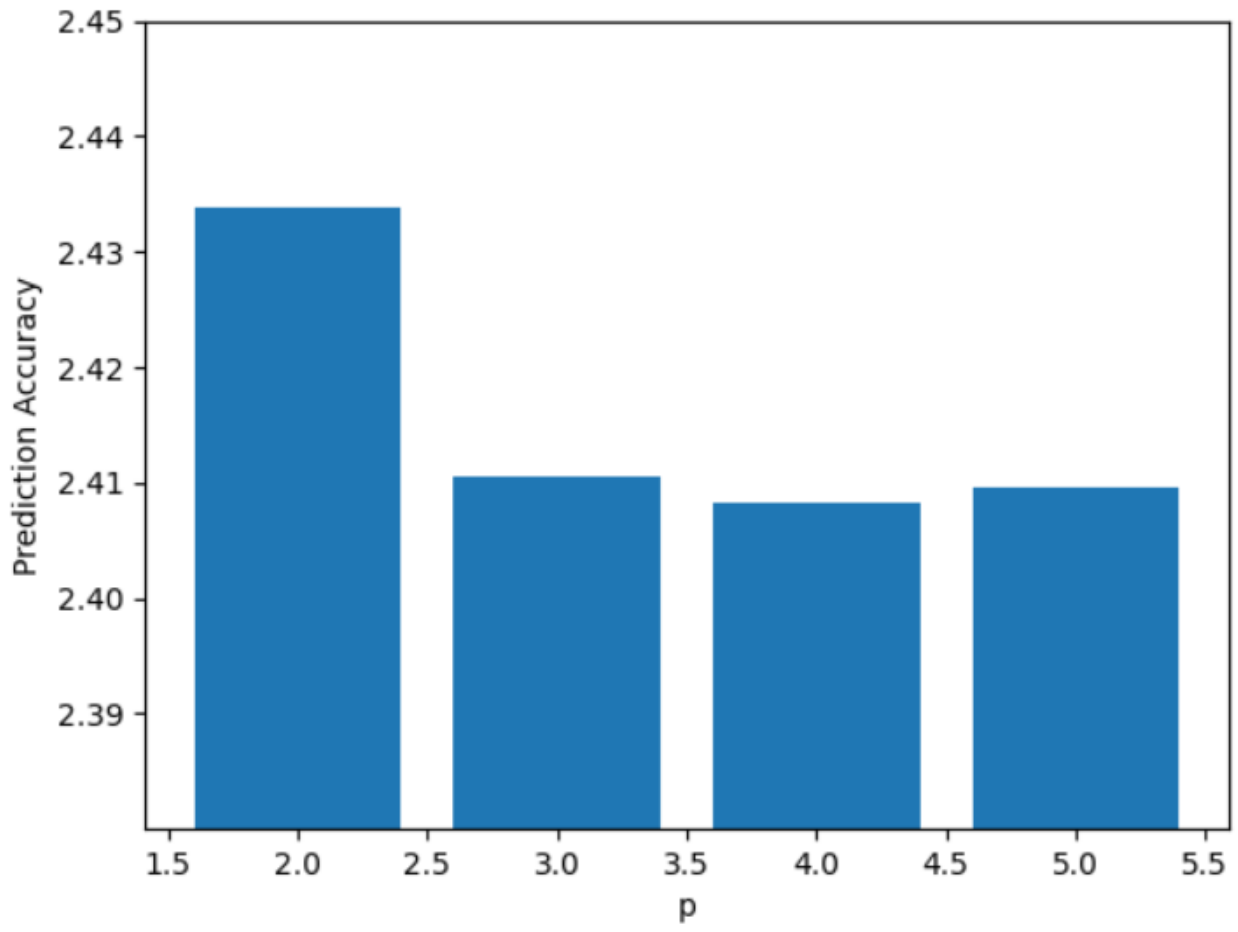

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases as the degree of the polynomial increases.
2. The decrease is more in degrees 2 to 3 and further the decreasing nature is gradual.
3. The more the degree the good the curve fits the data, as a result of this the RMSE decreases
4. From the RMSE value, p = 4 degrees curve will approximate the data best.
5. The bias decreases and the variance increases as increase in degree happens.

c.

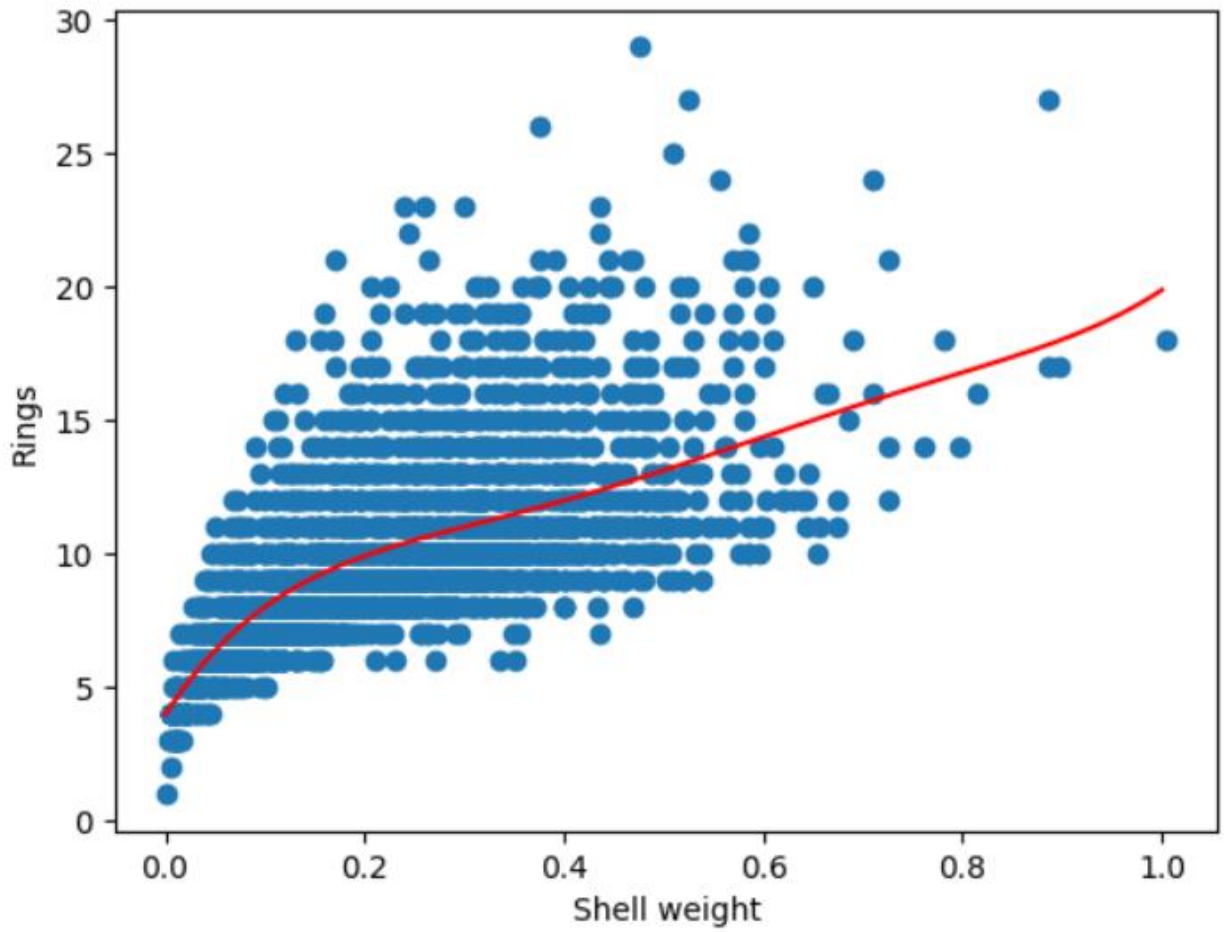

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. The p-value corresponding to the best fit model is p=4
2. Because it has more variance and fits the data more as compared to others.
3. The bias decreases and the variance increases as increase in degree happens.

d.

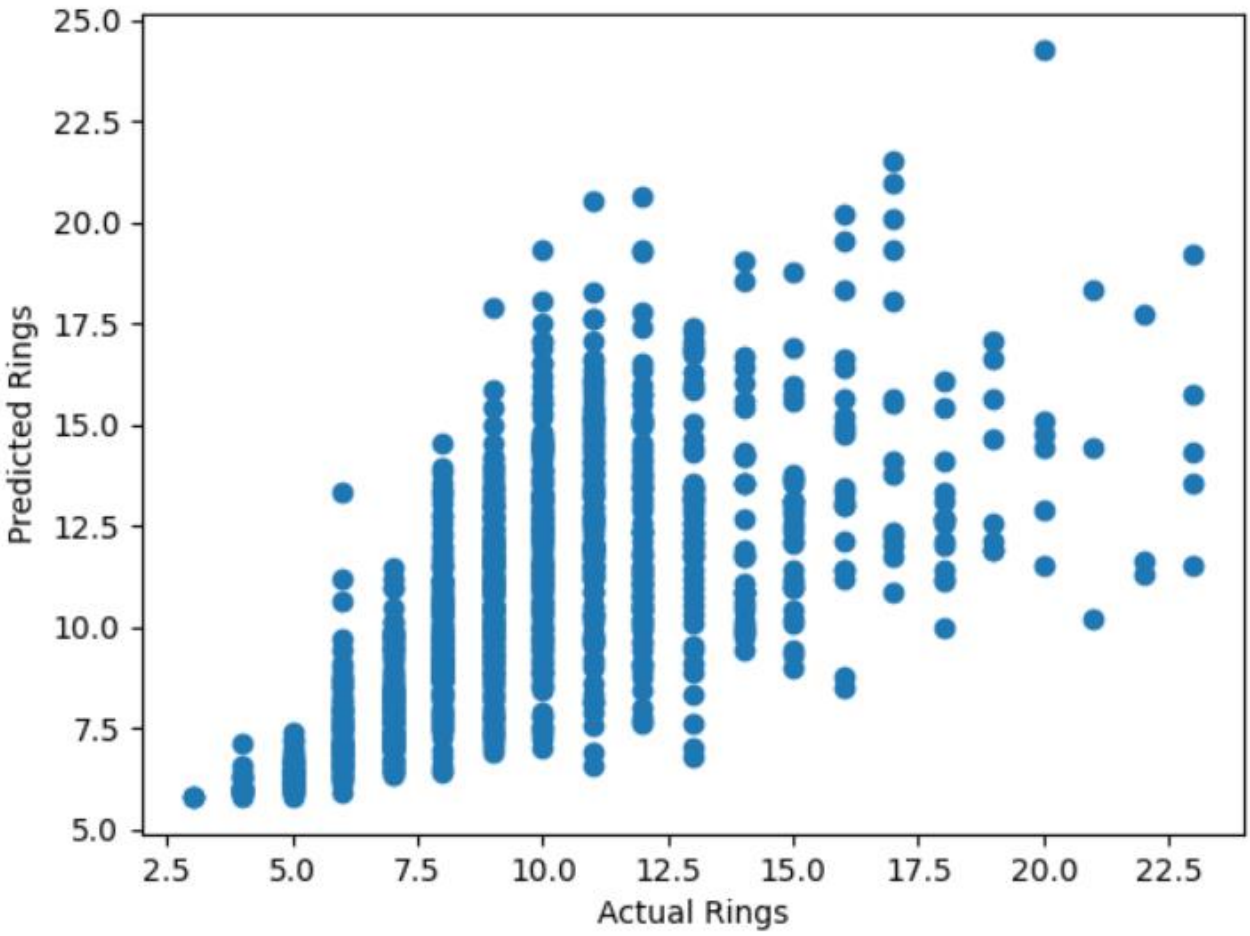

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted number of rings is almost accurate.
2. The spread of actual rings is 3-23 and that of predicted is 4-20.
3. The accuracy for Univariate non-linear is highest, multivariate linear is also close to the highest i.e. is of univariate non-linear and is least for Univariate linear regression model.
4. RMSE values of non-linear regression is lower than that of linear model therefore it is better.

5. In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high.

4

a.

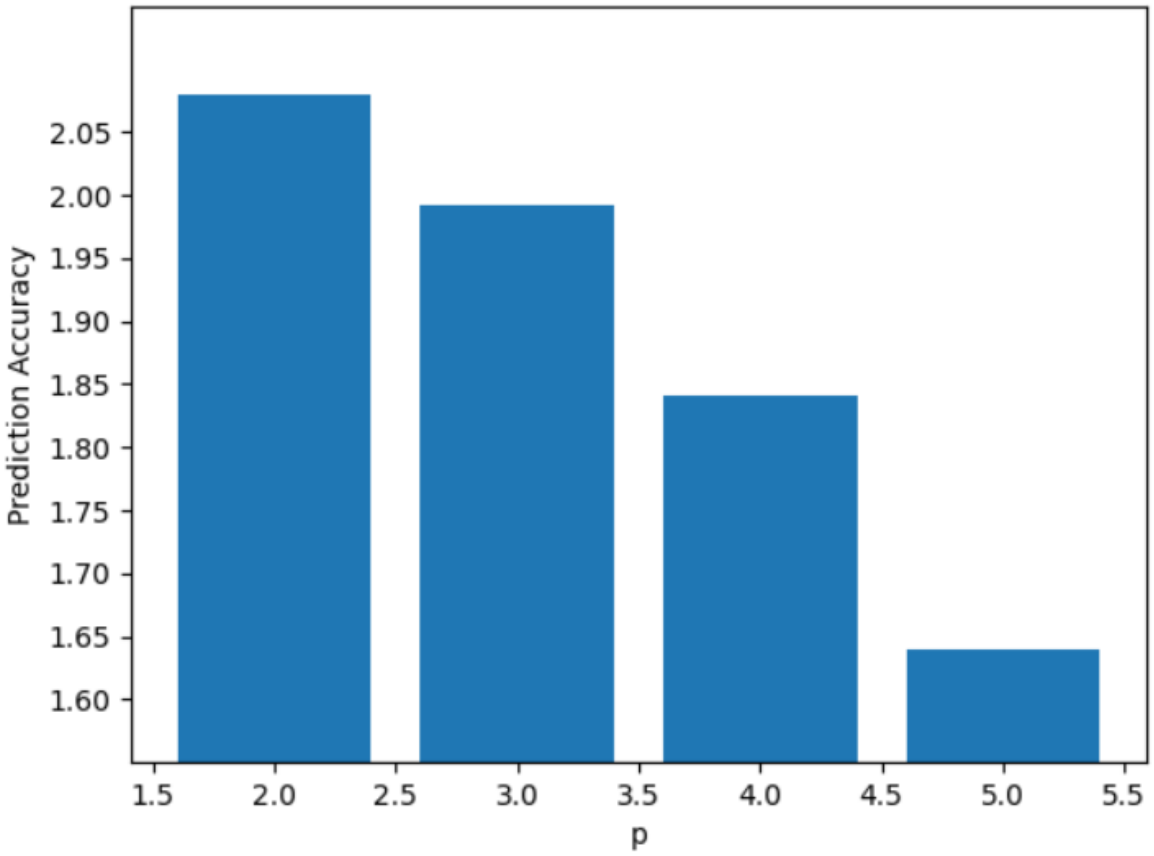

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases as the degree of the polynomial increases.
2. The decrease is more after degree 4 but before that the decreasing nature is gradual.
3. The more the degree the good the curve fits the data, as a result of this the RMSE decreases
4. From the RMSE value, p = 5 degrees curve will approximate the data best.

5. The bias decreases and the variance increases as increase in degree happens.

b.

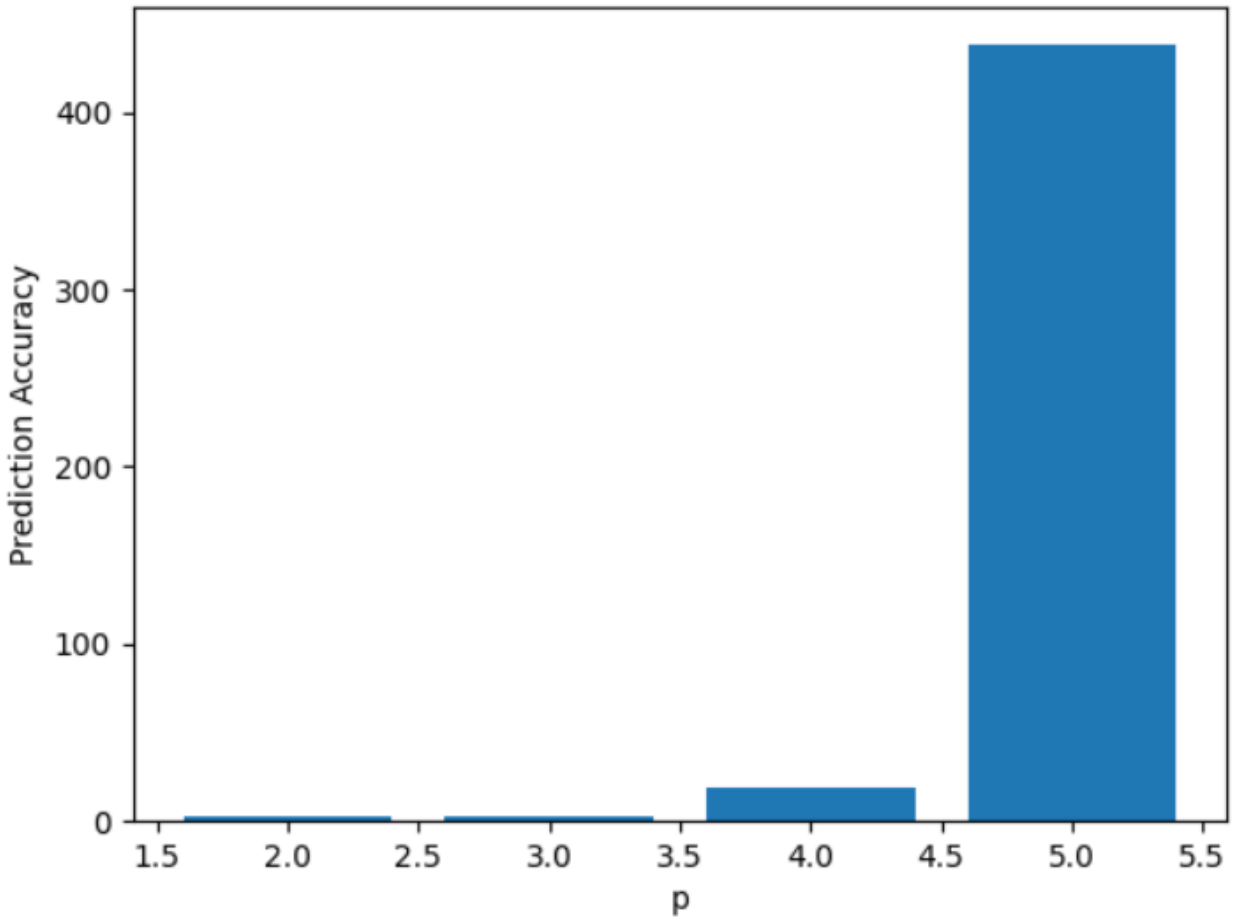

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial but it increases after p=3
2. Decrease is uniform till p=3 but after that it increases.
3. As we increase the value of degree our model is overfitted.
4. From the RMSE value, p = 2 degrees curve will approximate the data best.

5. Bias decreases till p=3 and then suddenly increases after p=3 and variance increases as models becomes complex as degree increases.

c.

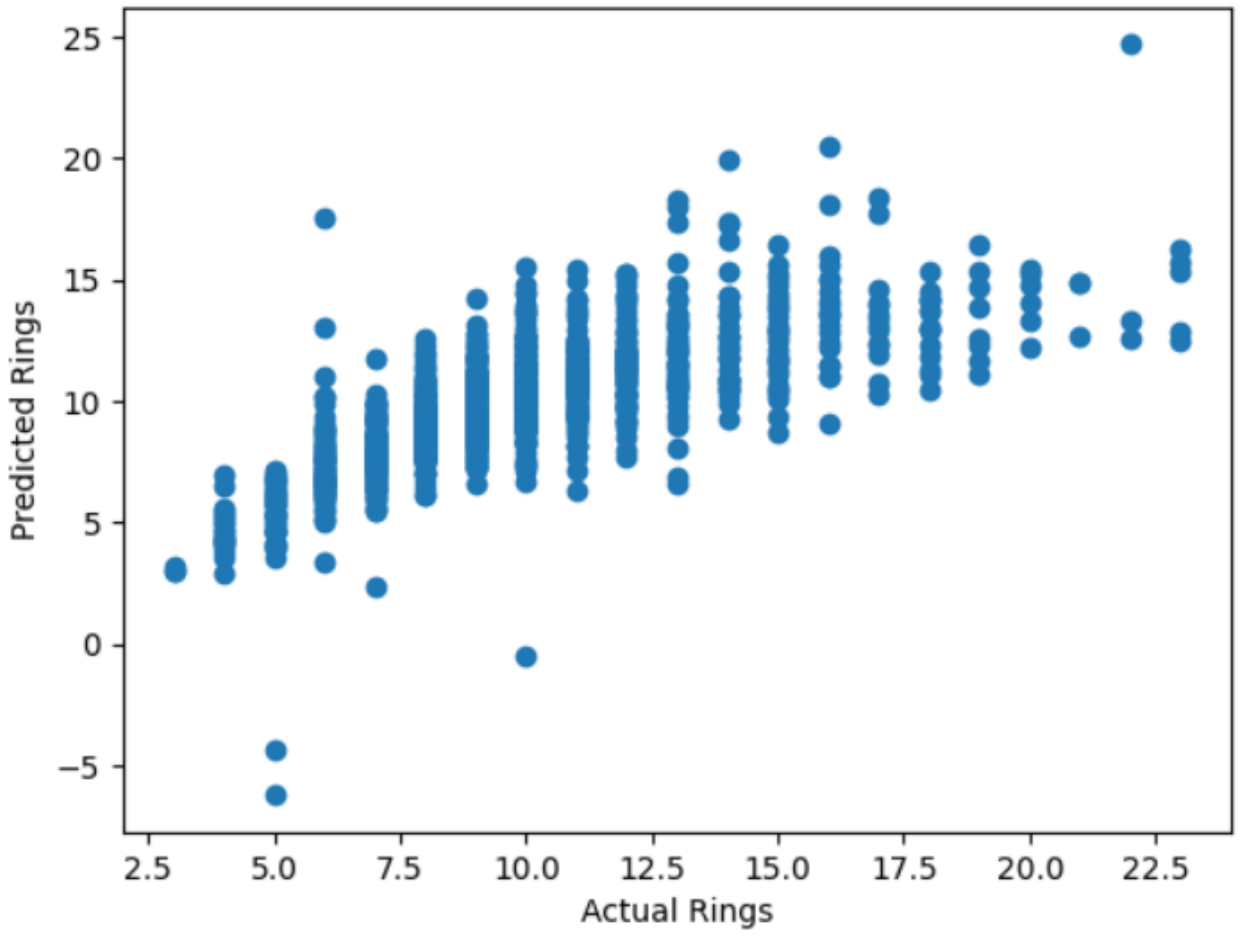

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted number of rings is almost accurate.
2. The spread of actual rings is 3-23 and that of predicted is 3-22.

3. The accuracy for Multivariate non-linear is highest, followed by univariate non-linear, and the accuracy of multivariate linear is less than that of univariate non-linear but more than that of univariate linear regression model.
4. RMSE values of non-linear regression is lower than that of linear model therefore it is better.
5. In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high