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: Ankit Pal Singh **Mobile No: 9149024234**

Roll Number: B20181 **Branch:EE**

PART - A

1 a.

		Prediction Outcome
True Label	108	3
	9	216

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	106	1
	11	218

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	98	2
	19	217

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	97	1
	20	218

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	**96.428**
4	**96.428**
8	**93.75**
16	**93.75**

Inferences:

1. The highest classification accuracy is obtained with Q =. 4
2. As Q increases the number of clusters increases so precision increases upto a limit and then eventually decreases.

3. Since by increasing Q , the number of cluster increases , and more cluster means more similar data samples to be clustestered with the cluster domain but increasing the cluster to a high limit will reduce this thing as more number of clusters are created.
4. Yes, the number of diagonal elements also increases with increasing precision.
5. Since as Q increases upto 4 the model becomes more generalized and accurate so that it provides better predictions for TP and TN .
6. As classification precision increases, the number of off-diagonal items decreases.
7. With an increased value of Q (up to a cutoff value) our model becomes more efficient and more accurate and therefore the number of false positive and false negative results is reduced..

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	88.392
2.	KNN on normalized data	97.023
3.	Bayes using unimodal Gaussian density	96.428
4.	Bayes using GMM	96.428

Inferences:

1. Highest accuracy – KNN(with normalized data) and lowest accuracy --- KNN (without normalized data).
2. KNN < Bayes (unimodal) < Bayes (GMM) < KNN (normalized).
3. By doing normalization the data falls within the specified ranges and this helps in better predictions for the test data. But since the data is not completely Gaussian so KNN is more effective with normalization.

PART – B

1

a.

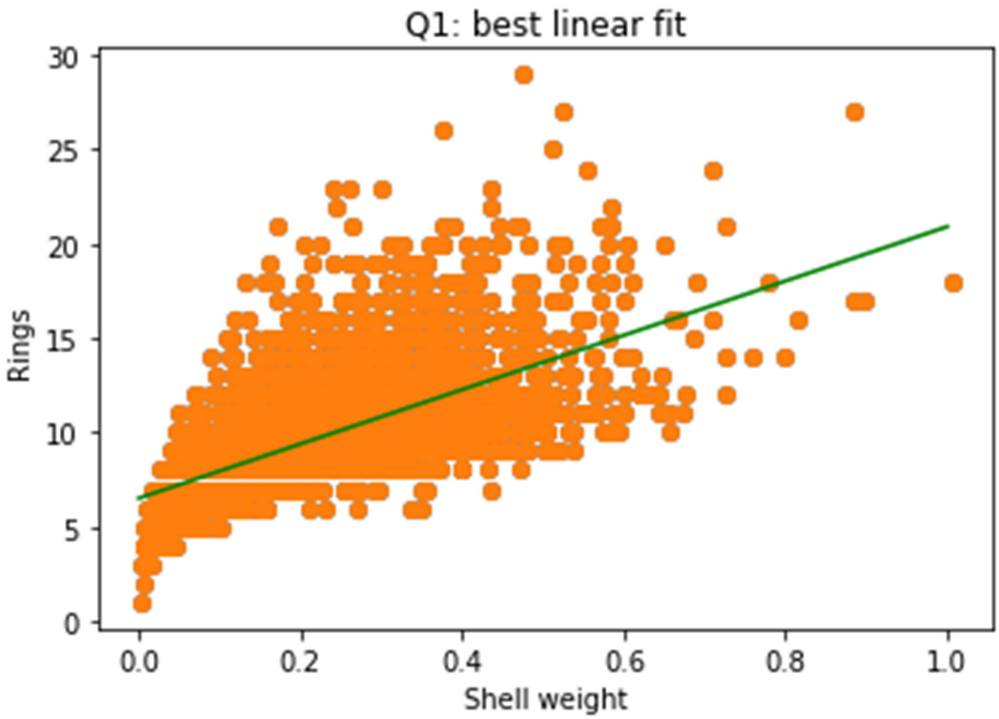


Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. As the target attribute rings are more dependent on the attribute with the highest correlation coefficient , so linear regression very efficient.
2. No , the line of best fit does not perfectly match the training data.
3. Since the assumption is that the attributes are linearly related, but donot have perfectly linear relation ship.
4. The variance is higher than the bias, so suggesting a complex model fit for low error.

b.

Accuracy of prediction of training data is 74.681 %.

c.

Accuracy of prediction of testing data is 74.859 %

Inferences:

1. The testing accuracy is higher than training accuracy.
2. Since the testing examples are less in size so this is the reason of increasing accuracy.

d.

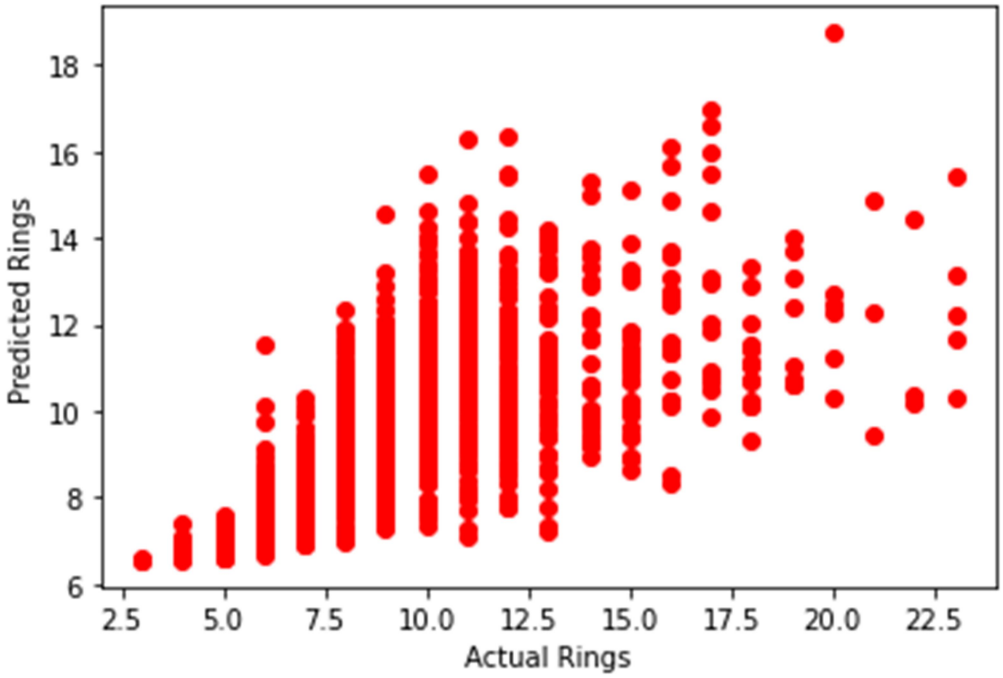


Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The predicted temperature is fairly accurate.
2. The data density is high in mid and follows a pattern, i.e., when actual no. increases, predicted no. also increases by fairly the same ratio.

2

Inferences:

1. Training is a little bit higher
2. It may be because there are more examples in train to get the model accustomed but still the model may not be perfectly trained for the samples and thus testing accuracy reduces.

c.

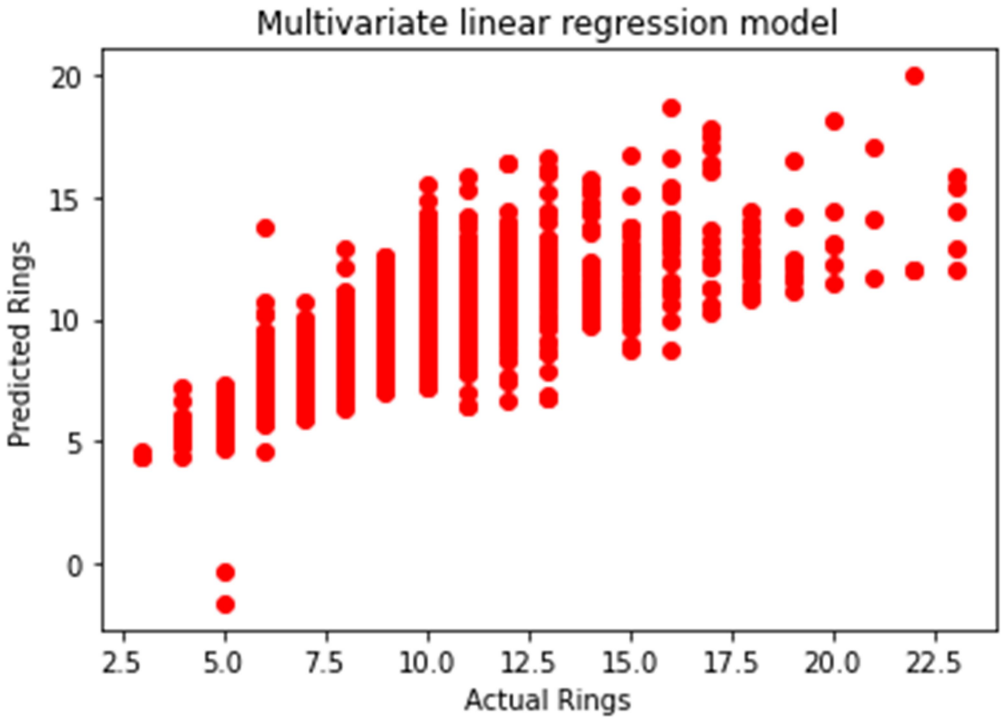
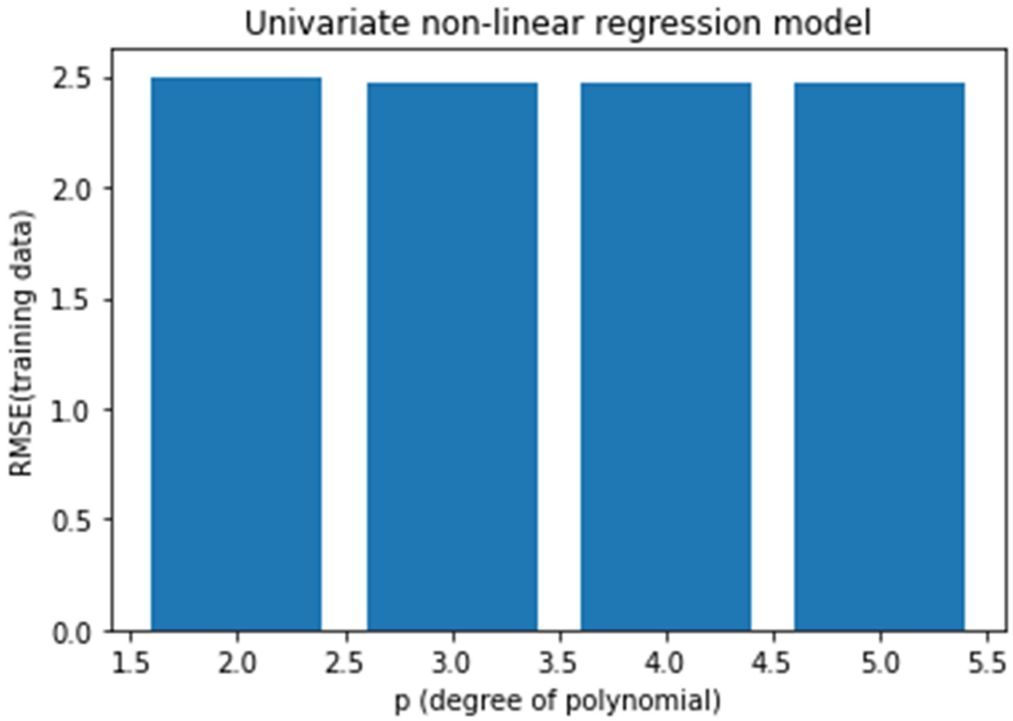


Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The temperature predicted is quietly accurate.
2. The data is dense in the middle and follows a clear pattern; both increase in the same proportion.
3. Multivariate regression is better than univariate linear regression because it takes training data from various attributes and efficiently trains the model to predict our attribute.

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 decreases as the degree increases.

2. After p = 3 , it is gradual.
3. The complexity of dependency equation increses , when the degree is increasing.
4. P = 5
5. Biasness in the data increases a bit as we increase degree and variance decreases just a little bit with increasing p

b.

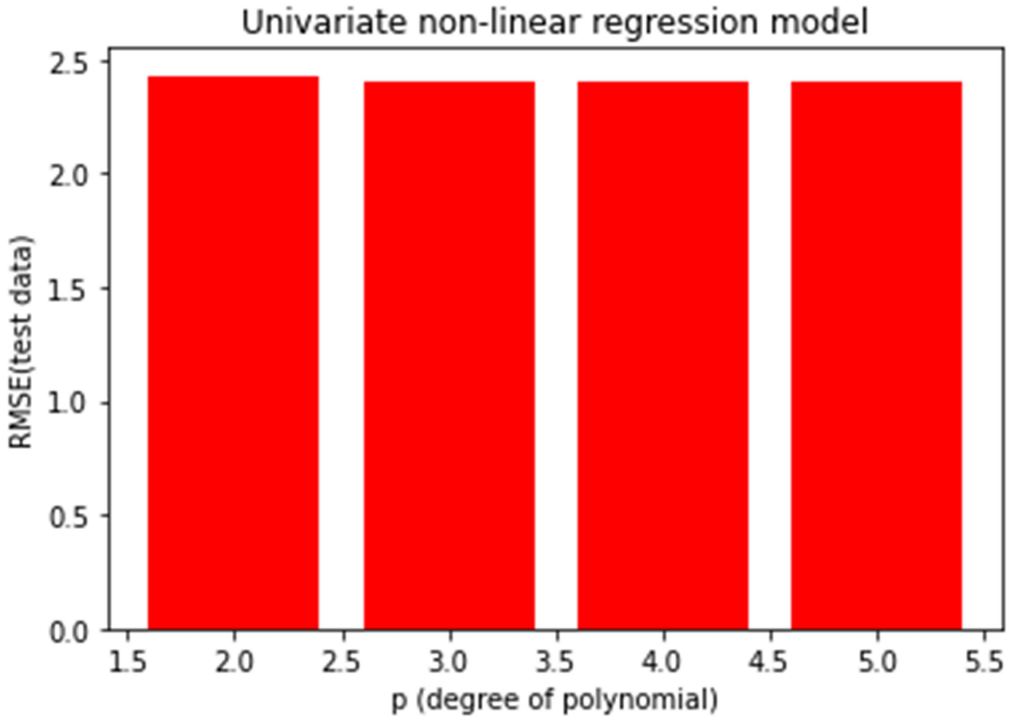
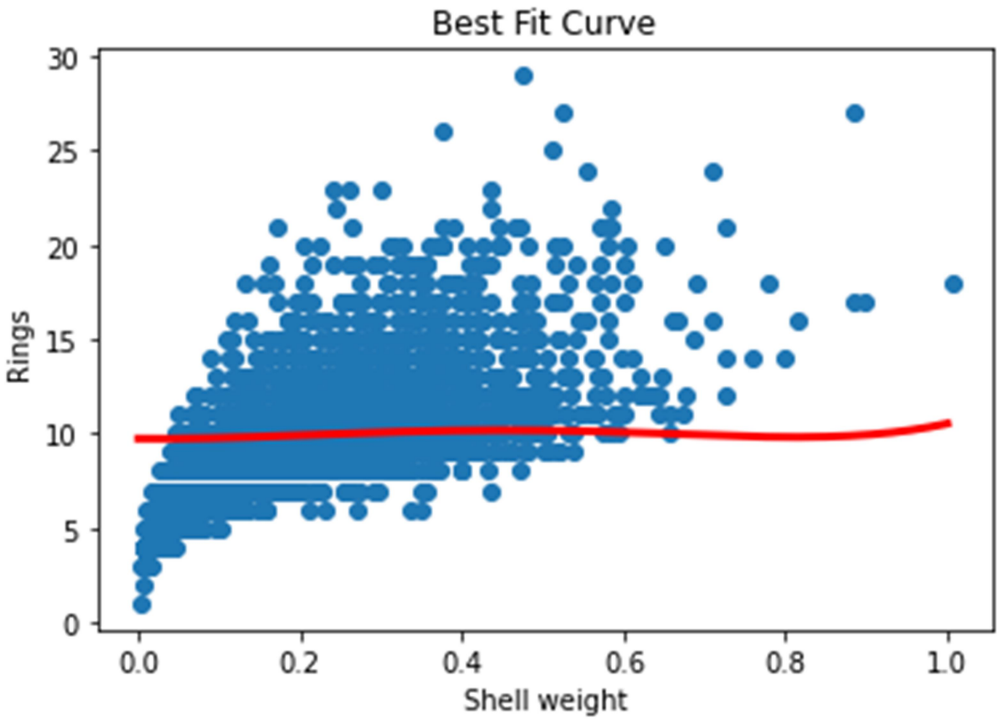


Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. As the value of P increases the Rmse increases.
2. After p = 3 it is gradual.
3. By increasing degree, the complexity of dependency equation increases, and it fits the data in a better way.
4. P = 5
5. Biasness in the data increases as we increase degree and variance decreases just a bit with increasing p



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 corresponding to the best fit model is p = 5
2. Increasing degree makes the equation of predicted case bit complex and better expresses the proportions of each attribute here

3. Bias is bit low, and variance is also low near p=4 so this may be the best degree point for expressing data.

d.

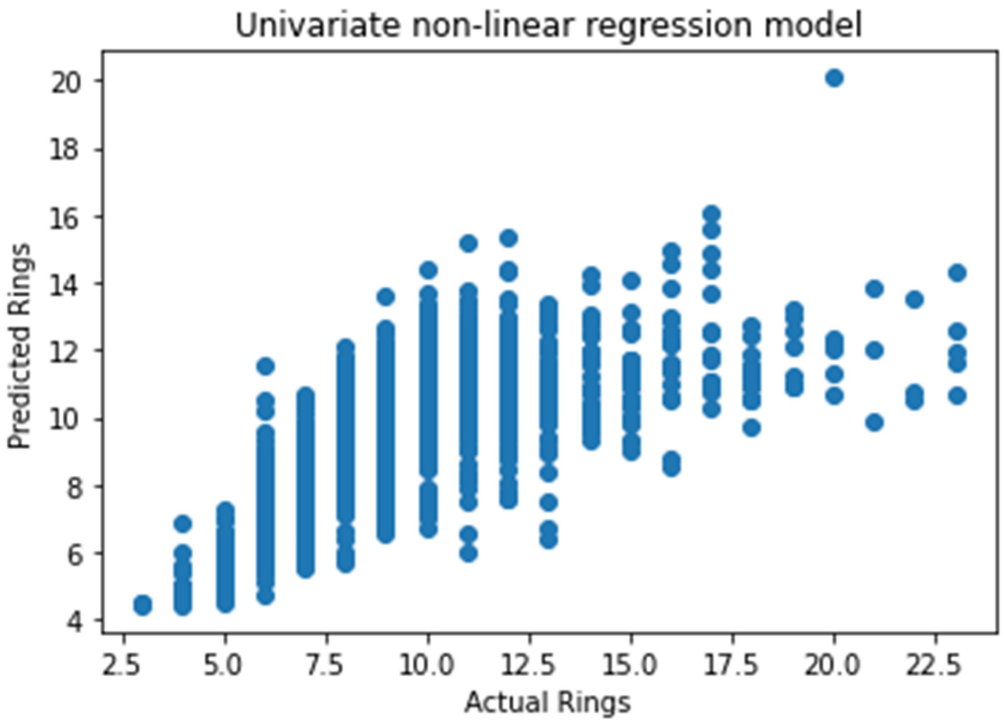


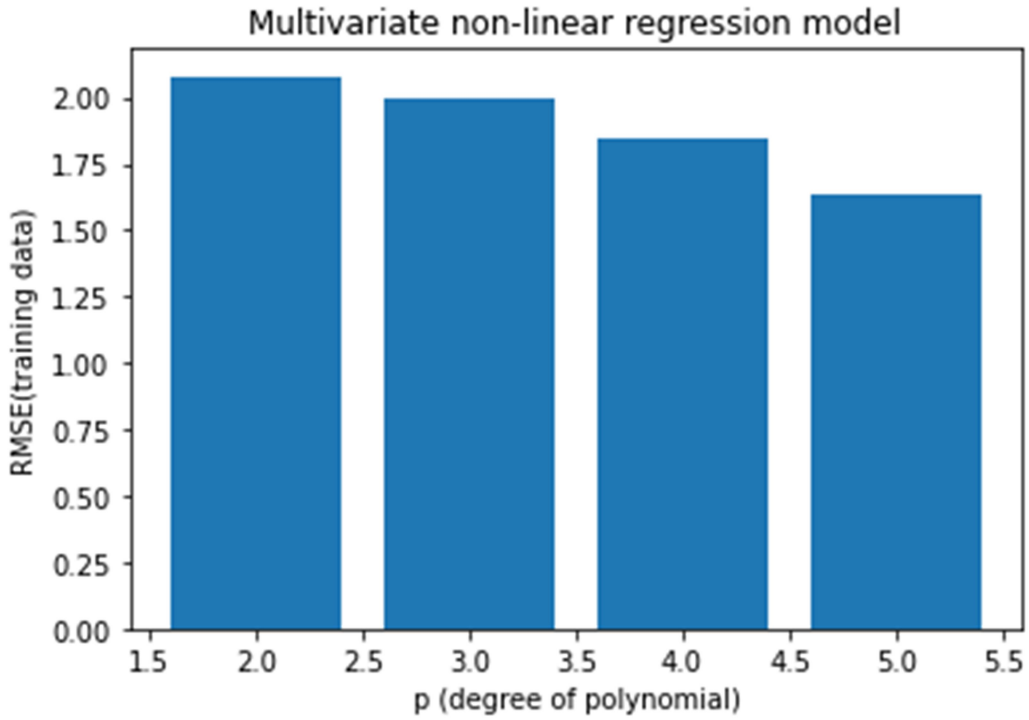
Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. The predicted temperature is quite accurate.

2. The plot follows a pattern, that is, when the real number increases, the predicted number increases. it also increases in roughly the same proportion
3. The margin (mainly outliers) is smaller for non-linear than for linear. In linear multivariate it is better for prediction.
4. The data depends on several attributes and is not linearly connected to them
   In linear regression, the compensation is high, but here it is low, suggesting a better fit and a trained model.

4

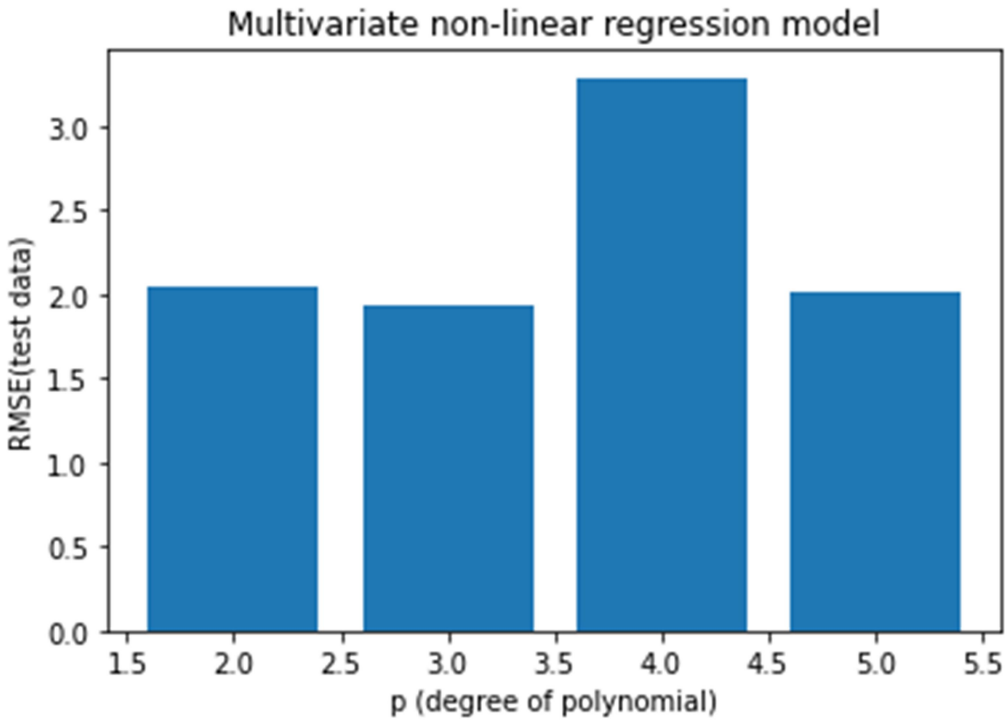


a.

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 decreases with increasing p
2. Decrease is uniform.
3. By increasing degree, the complexity of dependency equation with multiple attributes increases, and it trains data in a better way..
4. P=5. Variance reduces with increasing p but bias increases.



5. **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 decreases with increasing p
2. The decrease is gradual from p = 3.
3. By increasing degree, the complexity of dependency equation with multiple attributes increases, and so accuracy also increases.
4. P = 5.
   Variance reduces with increase in degree and so bias increases, but it looks like p=4 is a good point for reducing this trade-off.

c.

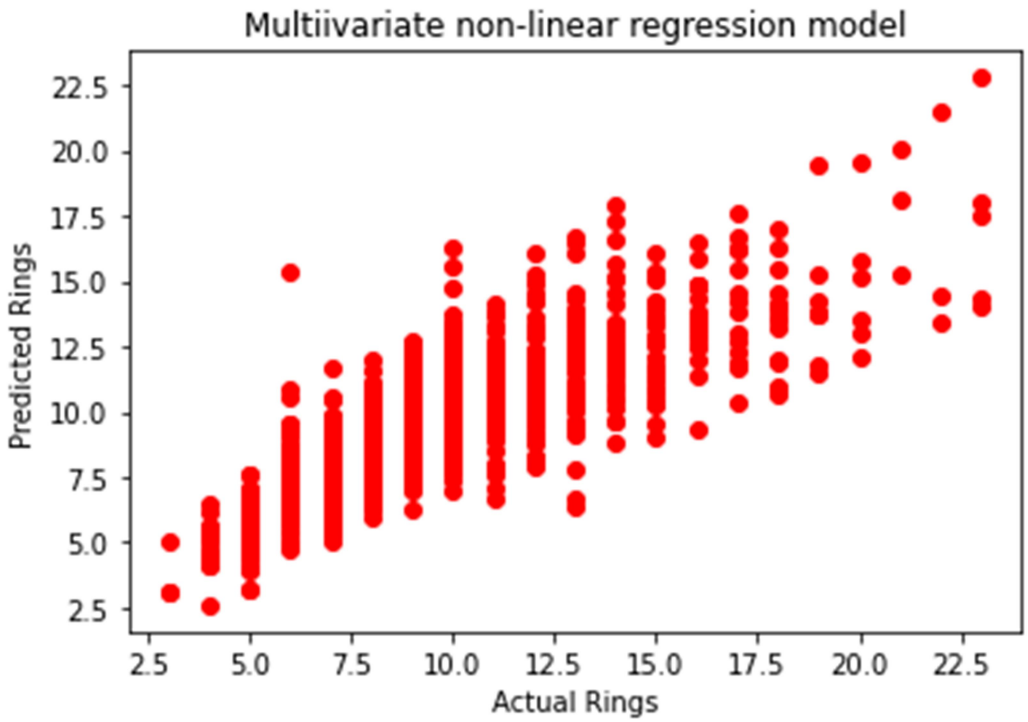


Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. The predicted temperature is fairly accurate
2. The data are less random and more evenly distributed along the origin
3. Multivariate is better than univariate regression and non-linear is good compared to linear regression
4. Nonlinear multivariate offers the best precision because it takes dependencies on multiple attributes and predicts data in its true nature. Linear regression provides comparatively less precision because the data is not linearly related to other attribute
5. The variance is low for nonlinear multivariate data and the scatterplot shows that the offset for the bias variance is small when we use our change in linear to nonlinear model, suggesting better predictive quality for the test cases