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: Kuldeep Jain Dugar **Branch:**

Roll Number: B20112 **CSE**

Mobile No: 8986388665

PART - A

1 a.

		Prediction Outcome
True Label	96	12
	4	224

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	97	11
	6	222

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	83	25
	2	226

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	79	29
	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	**95.238**
4	**94.94**
8	**91.964**
16	**91.071**

Inferences:

1. The highest classification accuracy is obtained with Q = 2
2. Increasing the value of Q first increases then start decreasing the prediction accuracy.

3. Increasing the value of Q decreases the prediction accuracy because adding the nodes with less weight cause the model to overfit on training data
4. As the classification accuracy increases with the increase in value of Q the number of diagonal elements in the confusion matrix increase.
5. the reason for the increase in diagonal elements is because accuracy increase result in increase in no. of correct prediction
6. As the classification accuracy increases with the increase in value of Q infer does the number of off-diagonal elements decrease.
7. the reason for decrease in off-diagonal elements is due to the decrease in no. of wrong prediction.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.6
2.	KNN on normalized data	97
3.	Bayes using unimodal Gaussian density	93.9
4.	Bayes using GMM	95.238

Inferences:

1. KNN on normalised data have the highest accuracy and KNN have lowest accuracy .
2. KNN < Bayes using unimodal Gaussian Density < Bayes using GMM < KNN on normalised data.
3. KNN on normalised data is better than knn because Euclidean distance is used so normalized data works better. Also in the above example which involves just 2 clusters, KNN will give more accurate predictons than Bayes. Multimodal Bayes performs better as we are now using multiple clusters which increases the relative accuracy.

PART – B

1

a.

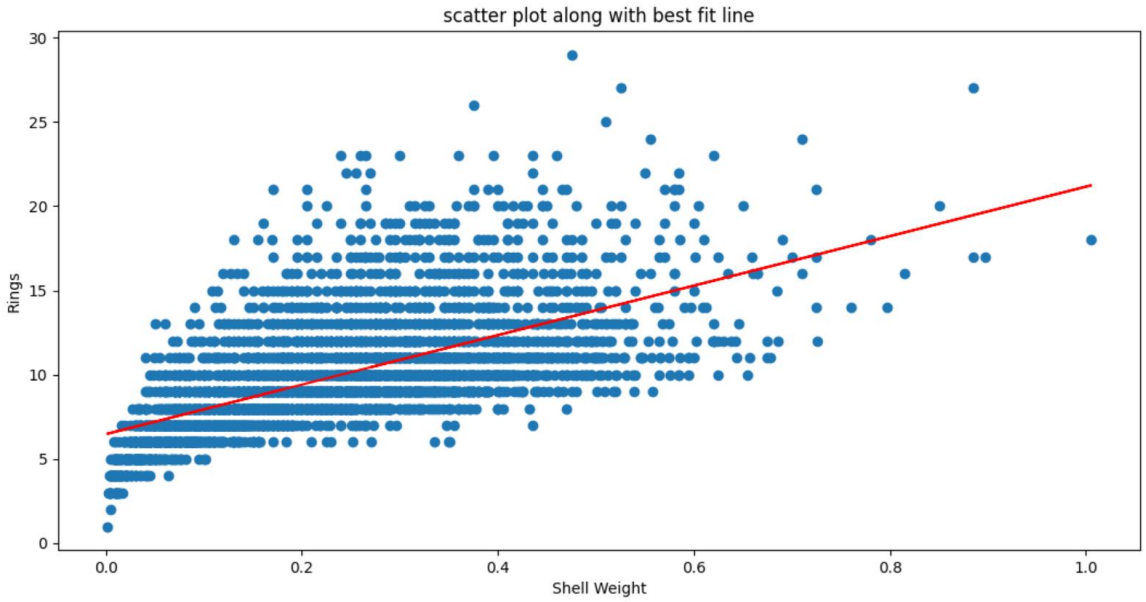


Figure 5 Univariate linear regression model: Rings vs. Shell weight best fit line on the training data

Inferences:

1. The attribute with the highest correlation coefficient was used for predicting the target attribute Rings. Because target attribute will be most dependent on it.
2. Does the best fit line fit the training data perfectly?- NO
3. Because a straight line can not be fit as the distribution is complex.
4. The bias is high and variance is low for the best fit line.

b.

Report the prediction accuracy on training data.- 2.527

c.

Report the prediction accuracy on testing data.- 2.467

Inferences:

1. Training Accuracy is higher

2. Because model is based on training data.

d.

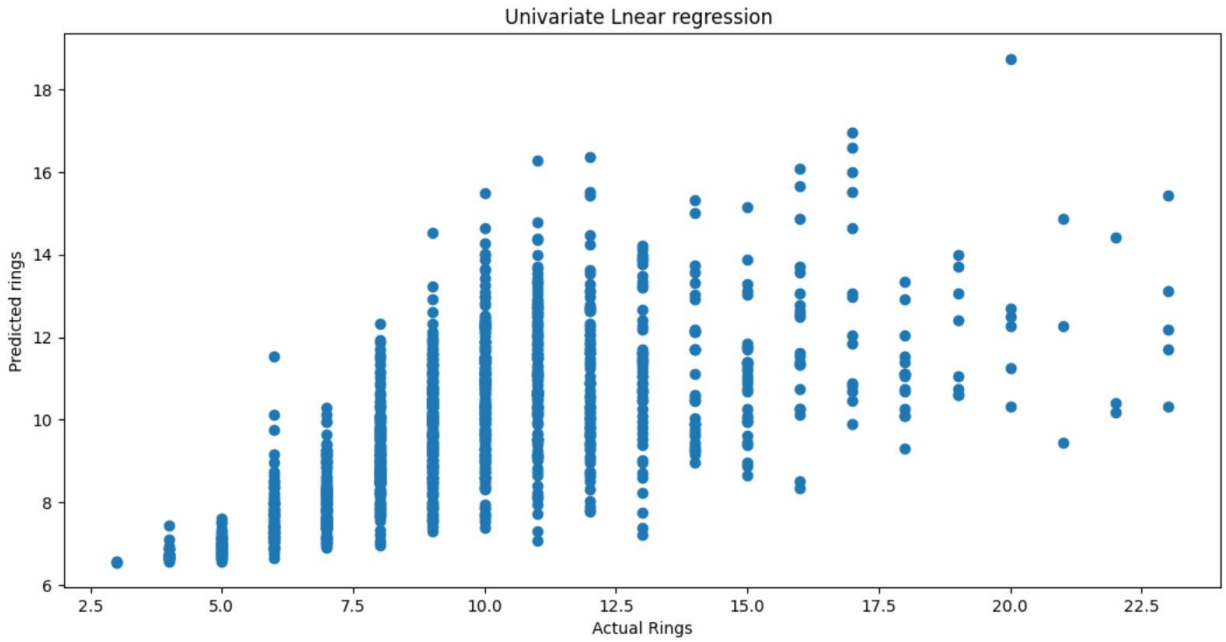


Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points predicted temperature is not very accurate.
2. Range of actual data and predicted data us varying.

2

a.

Report the prediction accuracy on training data.-**2.216**

b.

Report the prediction accuracy on testing data.- **2.219**

Inferences:

3. Amongst training and testing accuracy – both are almost equal
4. The data has fit with the modal more appropriately and prediction is better

c.

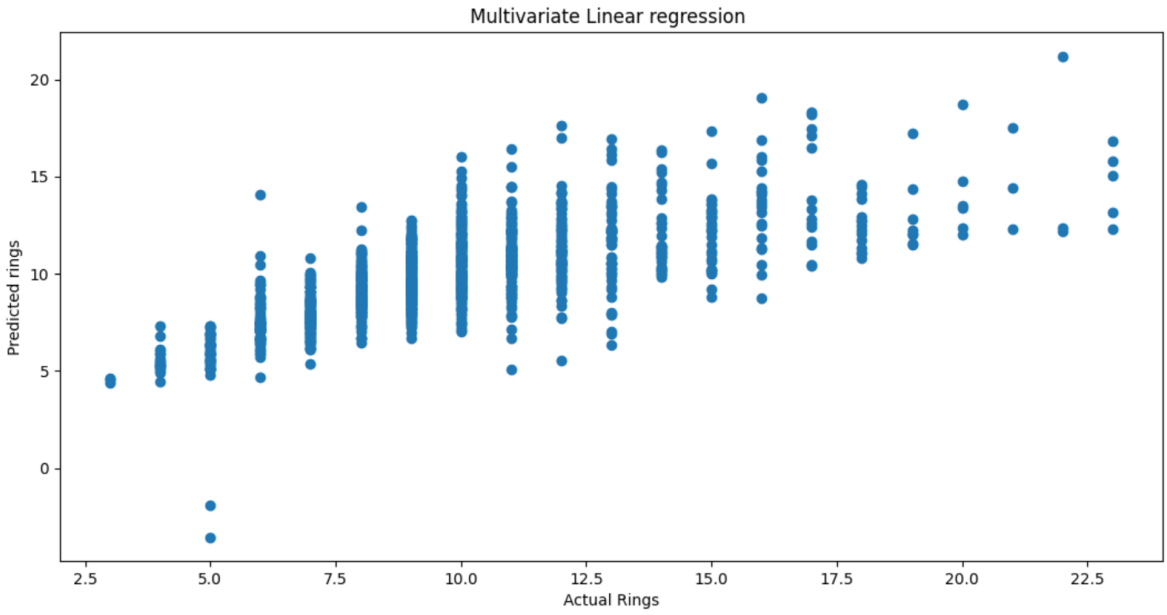
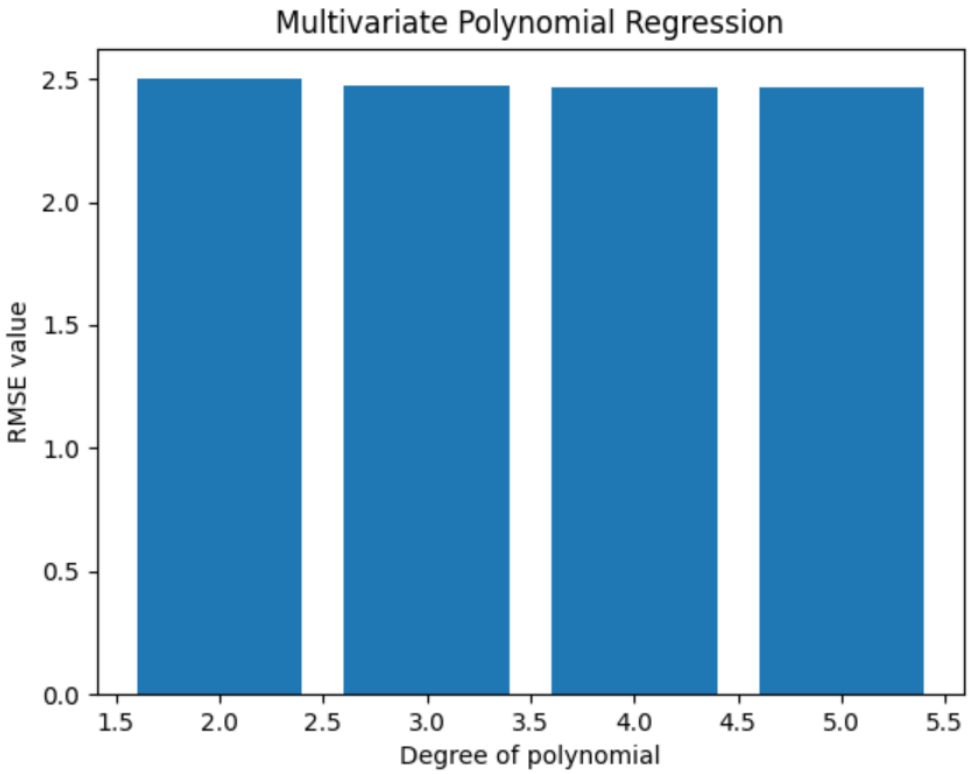


Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points the prediction seems to be little accurate
2. It has better model prediction method
3. Compare and contrast the performance of univariate linear with multivariate 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 values decreases with respect to the increase in the degree of the polynomial
2. The decrease is more from 2 to 3 and the gradual.
3. As the degree increases the curve fits the data more better so RMSE decreases.
4. From the RMSE value, p=5 curve will approximate the data best.
5. As the degree increases, the bias decreases and variance increases.

b.

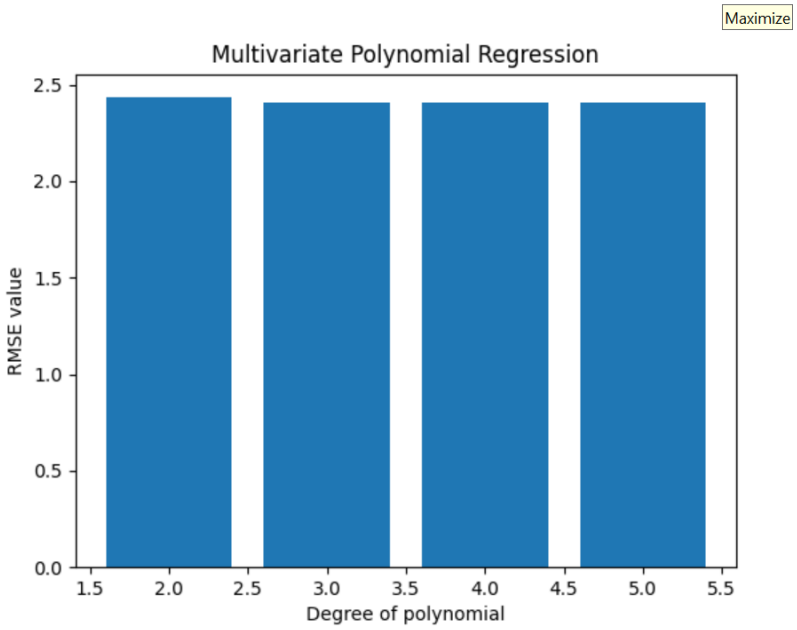
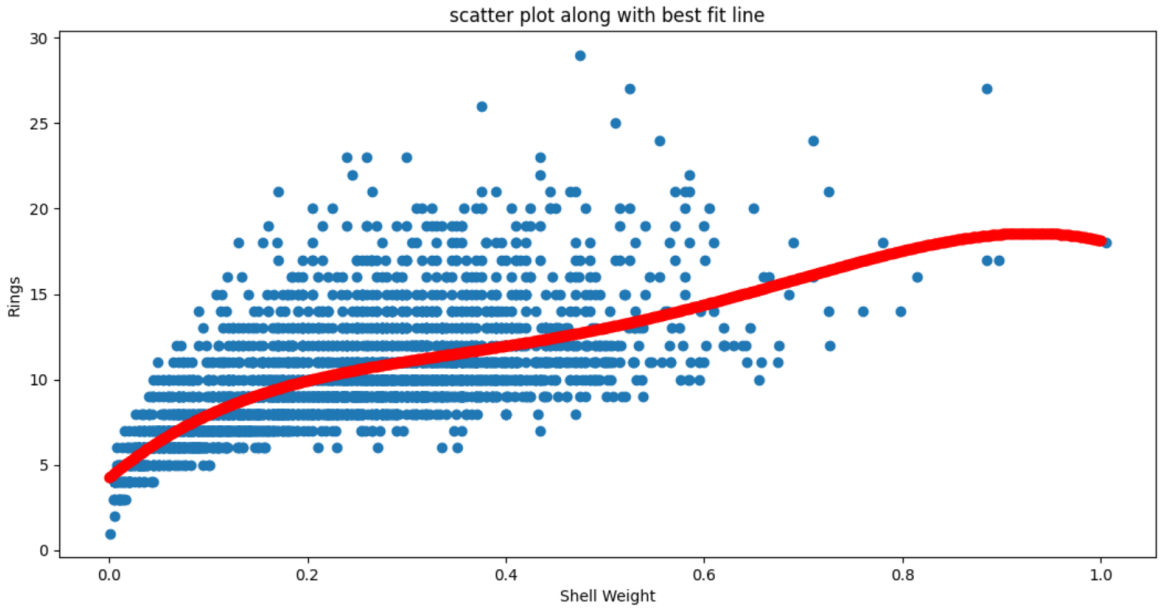


Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE values decreases with respect to the increase in the degree of the polynomial
2. The decrease is more from 2 to 3 and the gradual.
3. As the degree increases the curve fits the data in better way so RMSE decreases.
4. From the RMSE value, p=4 curve 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. State the p-value corresponding to the best fit model.= 4
2. Because it fits the data more and has more variance.

3. The bias decreases and variance increases with increasing value of p.

d.

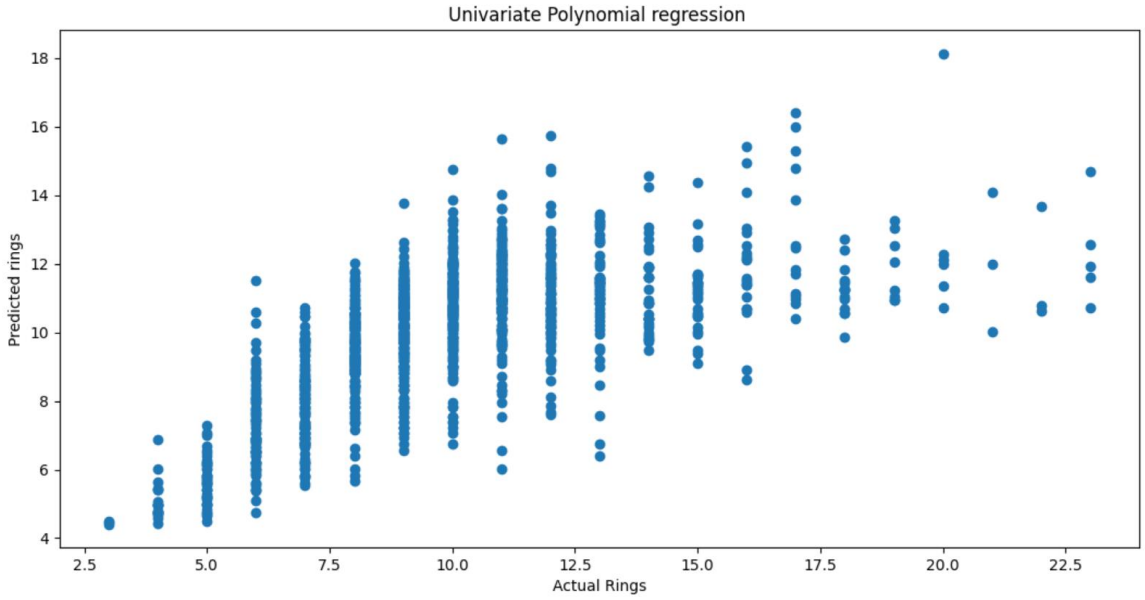


Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points the prediction is significantly accurate
2. The difference in the range is not high enough.
3. The accuracy for Univariate non-linear is the highest closely followed by Mul6variate Linear model and least is for univariate linear model.
4. RMSE values for non-linear regression model is lower than that of linear models hence 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.

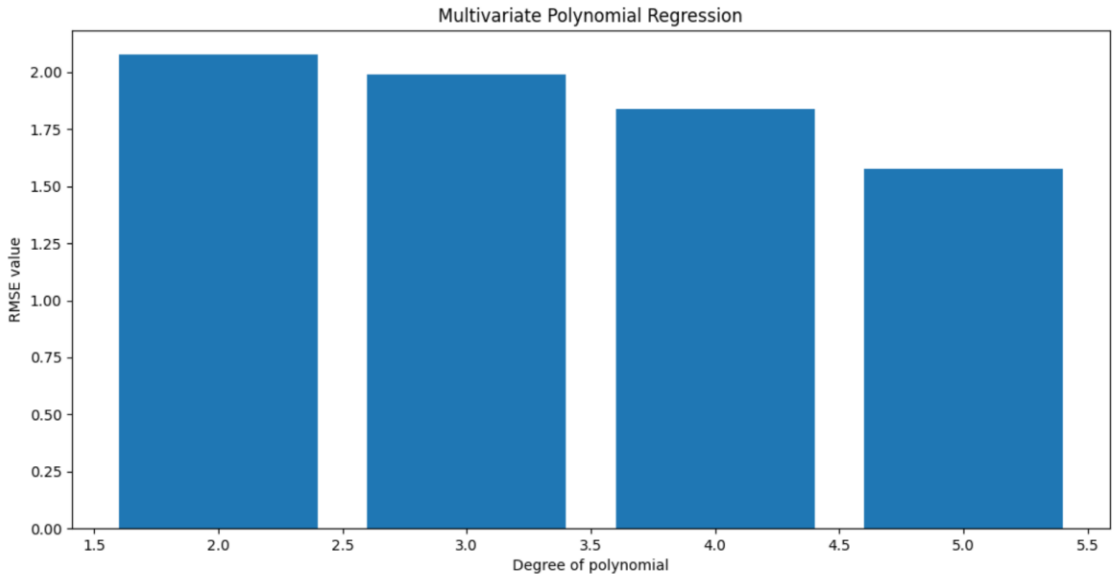
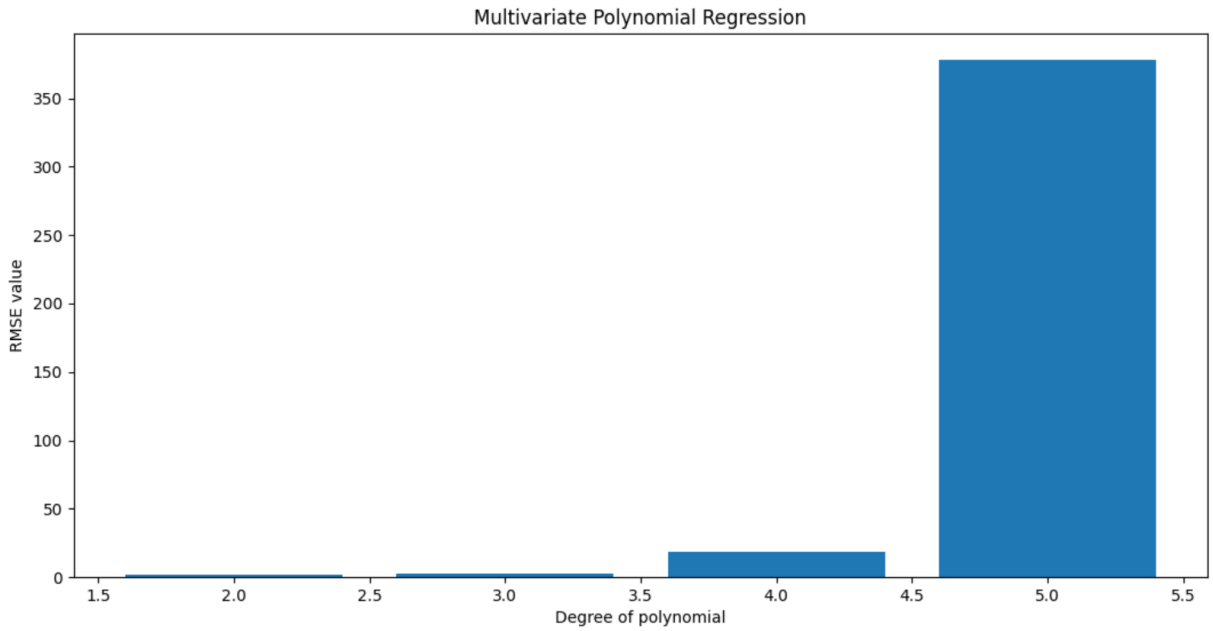


Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial .
2. The decrease is almost uniform but at p=4 the decrease is more.
3. As the degree increases the curve fits the data better so RMSE decreases.
4. From the RMSE value, p=5 degree curve will approximate the data best.
5. The bias decreases and variance increases with respect to the increase in the degree of the polynomial



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. Infer whether RMSE value decreases with respect to the increase in the degree of the polynomial and starts increasing when p=3.
2. The decrease is uniform when p=3 but when p=3 the increase is much more.
3. As we increased the degree of polynomial our model became overfitted.
4. From the RMSE value, p=2 curve will approximate the data best.

5. The bias gradually decreases 6ll p=3 and then suddenly increases ader p=3 and the variance increases as the model becomes more complex with increasing degree of polynomial.

c.

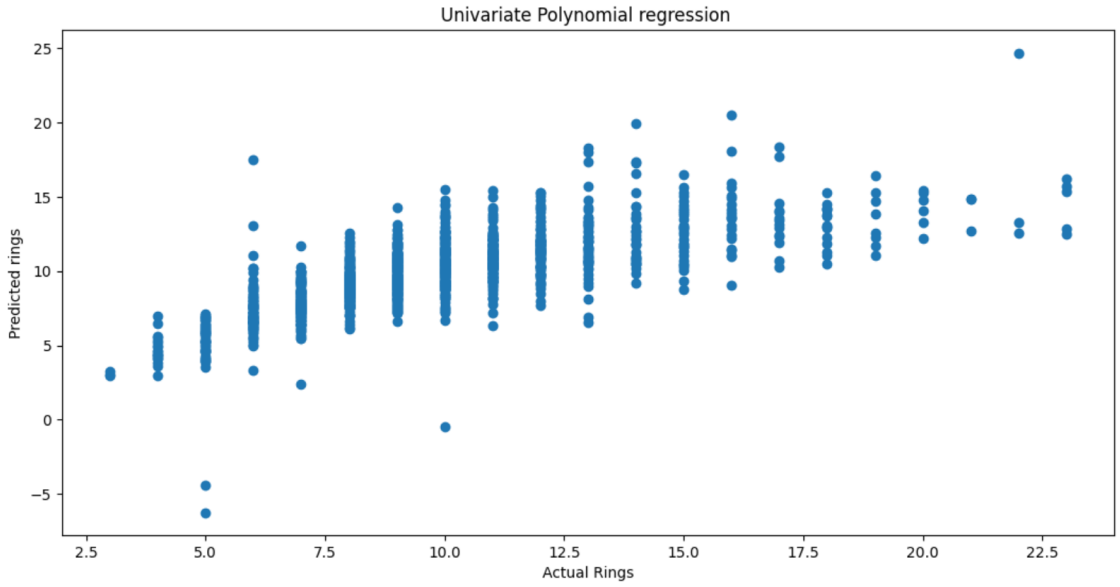


Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points it is accurate
2. Because model is fitted appropriately
3. Multvariate non-linear regression model has the highest accuracy followed by univariate non-linear model and the accuracy of mul6variate linear is less than that of univariate non-linear model but more than univariate linear regression model.
4. Due to the rmse values.
5. Inference based upon bias and variance trade-off between linear and non-linear regression models.