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: Bharat Kumar **Mobile No: 6299862503**

Roll Number: B20090 **Branch:**CSE

PART - A

1 a.

		Prediction Outcome
True Label	101	17
	6	213

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	110	8
	8	211

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	112	6
	8	211

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	104	14
	3	216

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	**93.17**
4	**95.25**
8	**95.84**
16	**94.95**

Inferences:

1. The highest classification accuracy is obtained with Q =.8.
2. First accuracy increases then decreases.
3. After q=8 there is over-fitting.

4. No. of diagonal elements increases with increase in accuracy.

5. No. of off-diagonal elements decreases with increase in accuracy.

6. Off-diagonal elements decreases because they are wrongly classified datapoints.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.61
2.	KNN on normalized data	97.33
3.	Bayes using unimodal Gaussian density	94.35
4.	Bayes using GMM	95.84

Inferences:

1. Highest accuracy: KNN on normalized data
2. Arrange the classifiers in ascending order of classification accuracy. Classifier 2 < Classifier 4 < Classifier 3< Classifier 1.

PART – B

1

a.

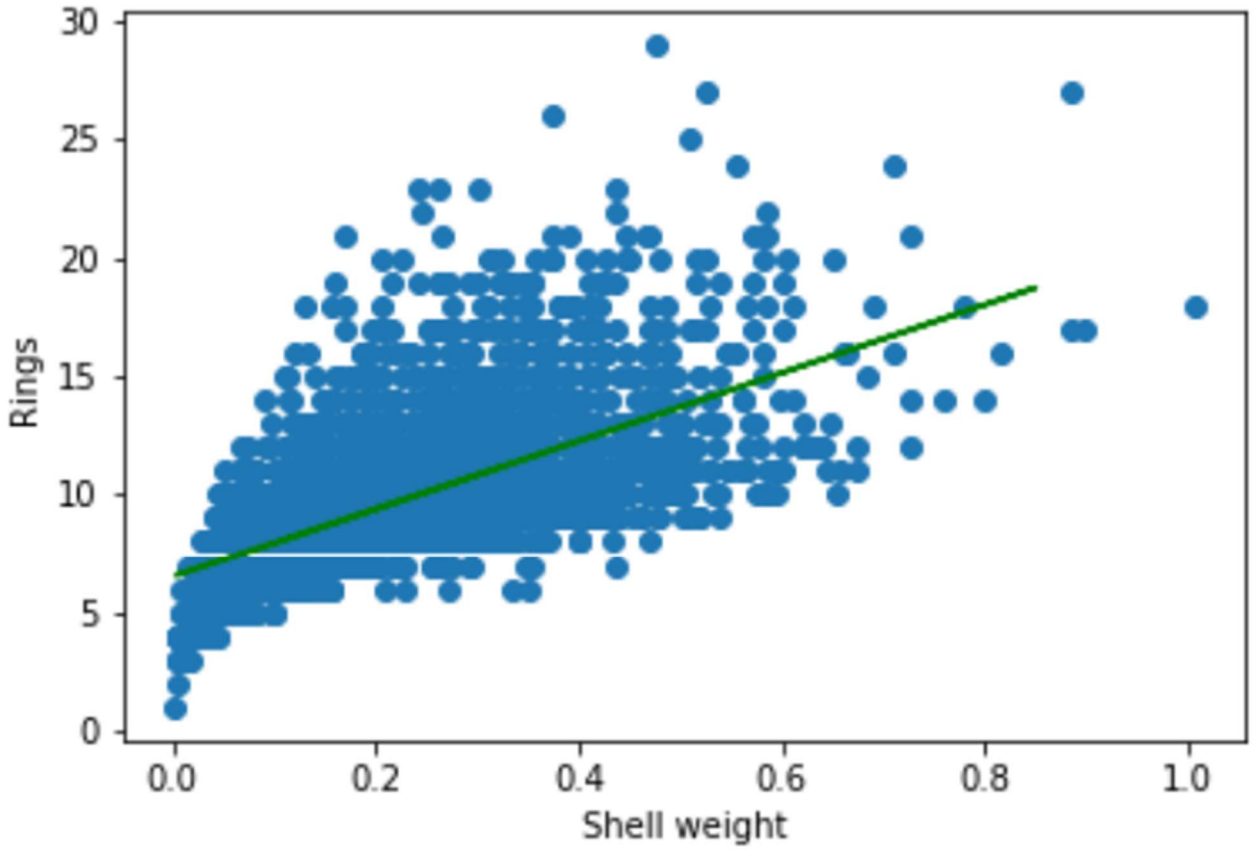

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. The attribute with the highest correlation coefficient was used for predicting the target attribute Rings because they will have the best relation compared to others.
2. The best fit line does not fit the training data perfectly. because correlation is moderate but not very high.
3. Because correlation is moderate but not very high.
4. High bias implies less accuracy.

b.

RMSE on training data is 2.527

c.

RMSE on testing data 2.467

Inferences:

1. Amongst training and testing accuracy, training accuracy is higher.
2. Because the model was trained using training data. So, the best fit line is more suitable for training data than any other random data.

d.

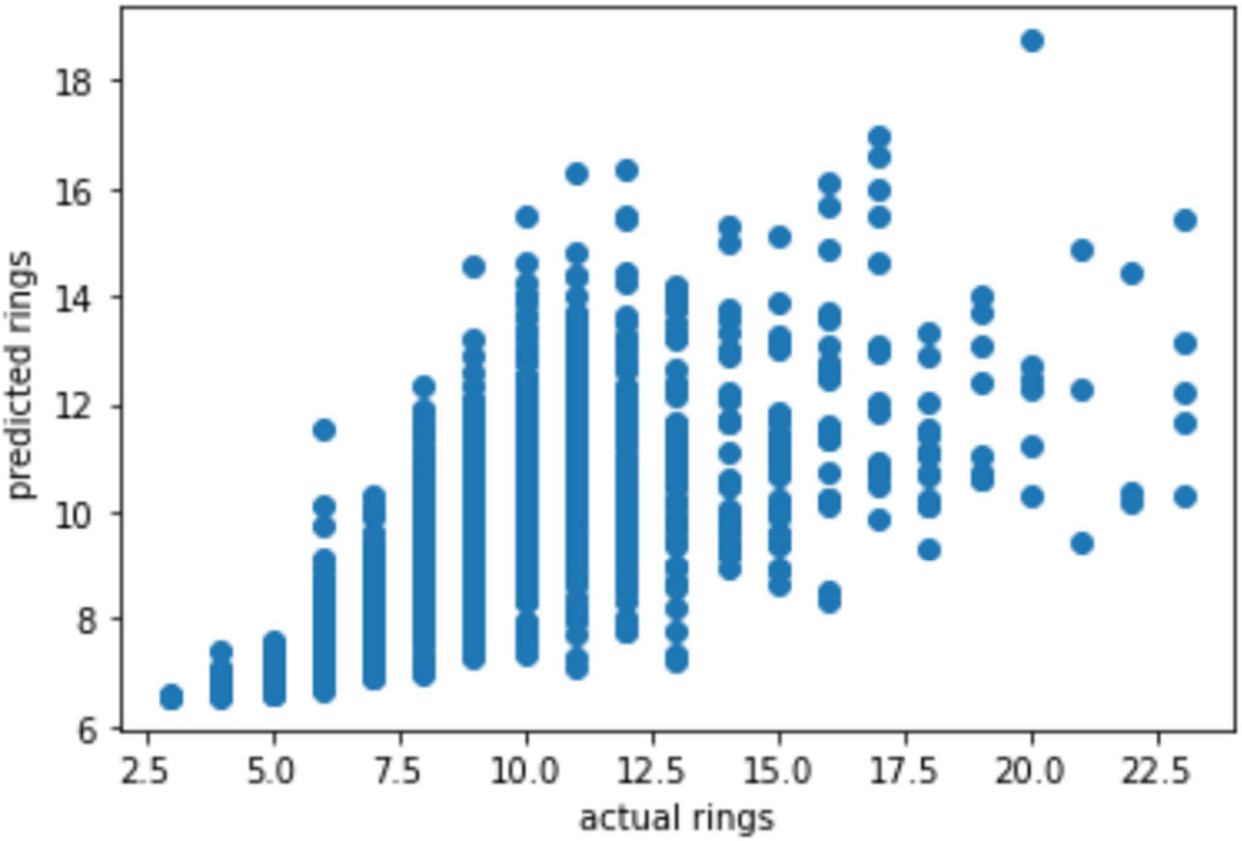

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, it is not perfect as variance is high.
2. High variance implies prediction varies over wide spread which makes it imperfect.

2

a.

```
RMSE on training data 2.216
```

b.

```
RMSE on testing data is 2.219
```

Inferences:

1.) Here both are almost same.

c.

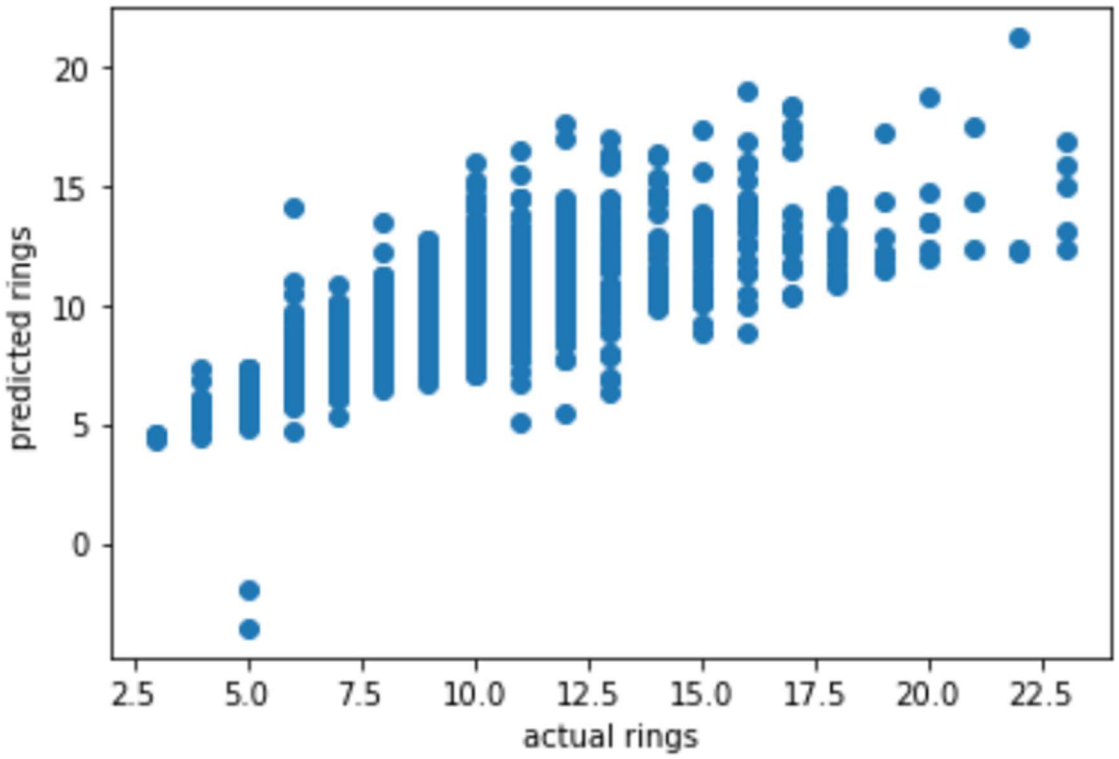

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. It is better than the previous predictor.
2. The reason behind this is that now more features are included.
3. In this case multivariate is giving better results than univariate.

3

a.

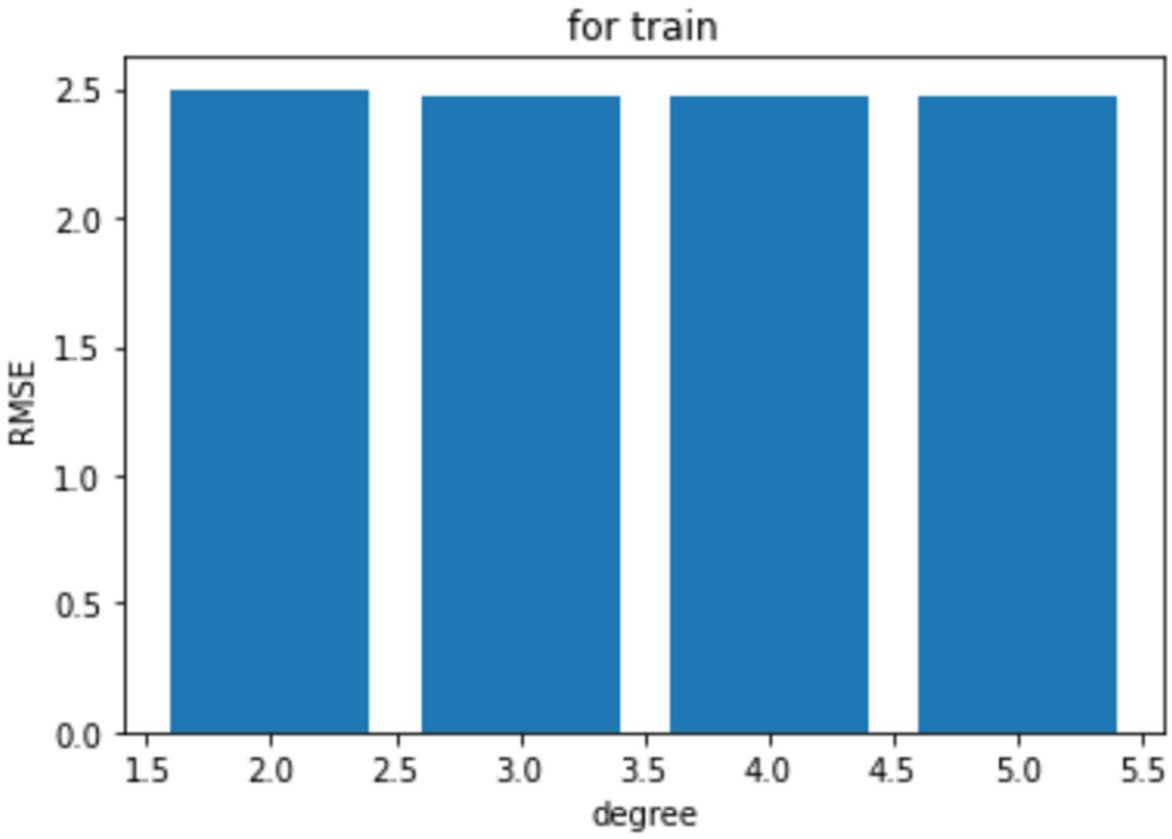

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. Here , RMSE decrease with increase in p.
2. After 4 it decreases.

3. Increasing the degree causes better coverage of data, but if we increase it too much then that will cause overfitting and RMSE will increase.
4. 4^{th} degree curve is best fitted here.

b.

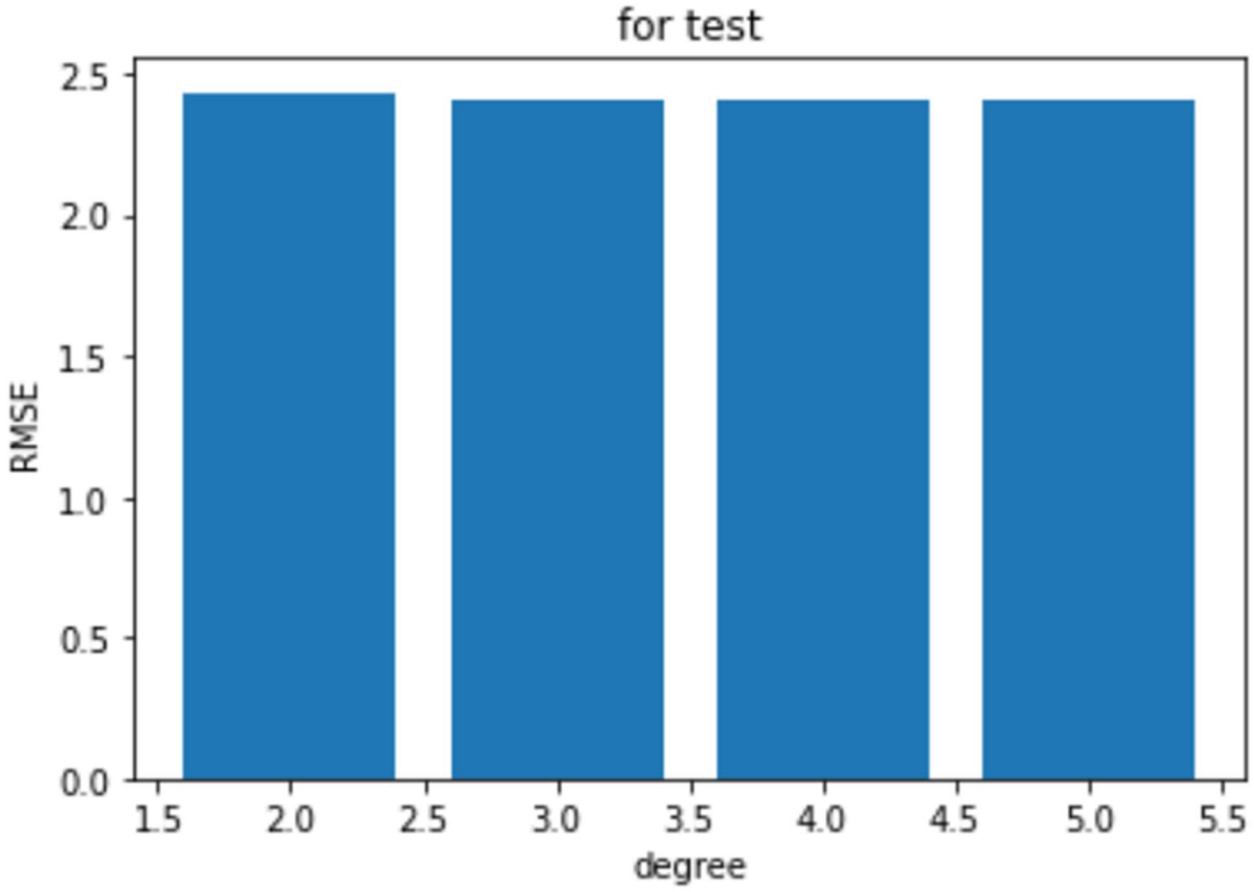

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. Here , RMSE decrease with increase in p.

2. After 4 it decreases.
3. Increasing the degree causes better coverage of data, but if we increase it too much then that will cause overfitting and RMSE will increase.
4. 4^{th} degree curve is best fitted here.

c.

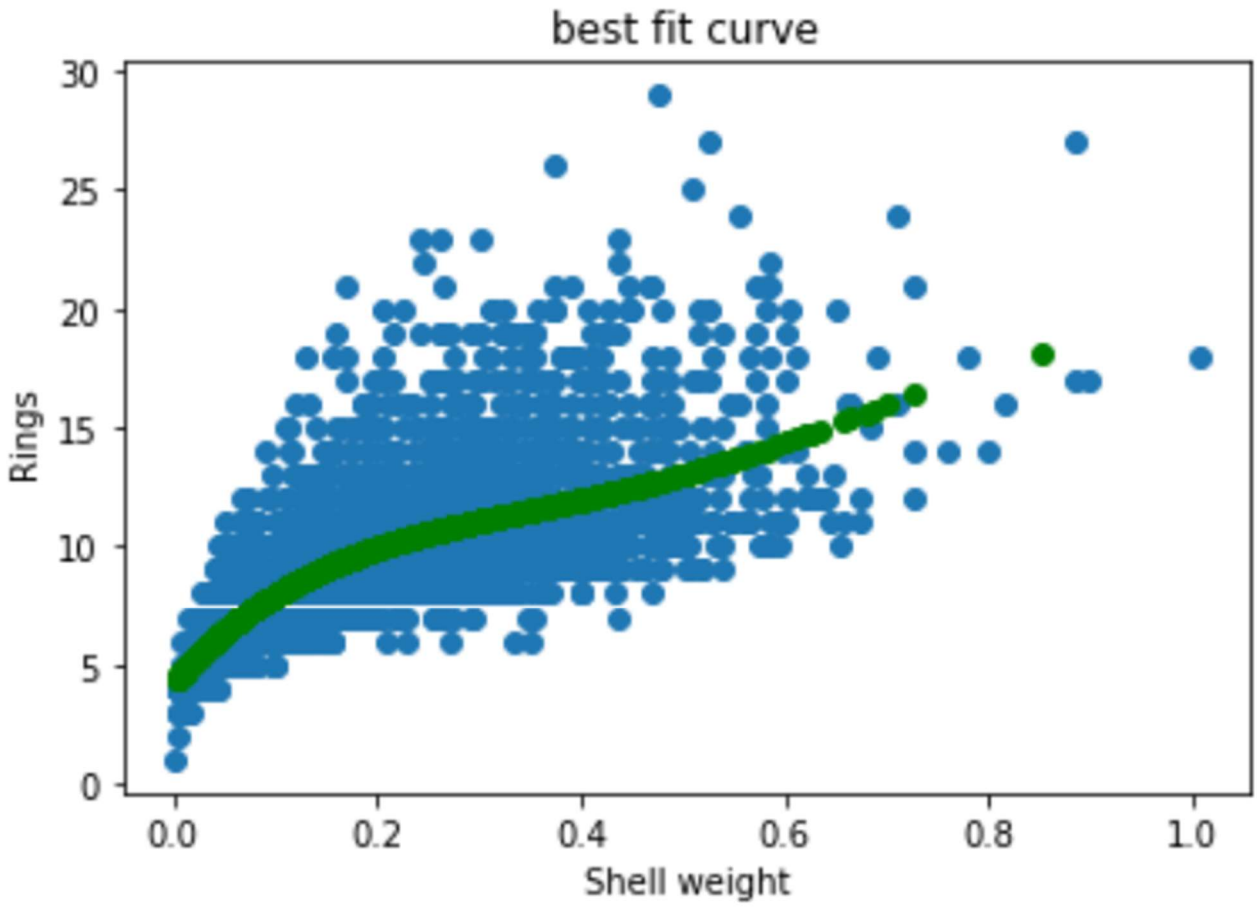

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. The model is best fitted for p=4.

2. The RMSE decreases till 4 and then increases due to overfitting.

d.

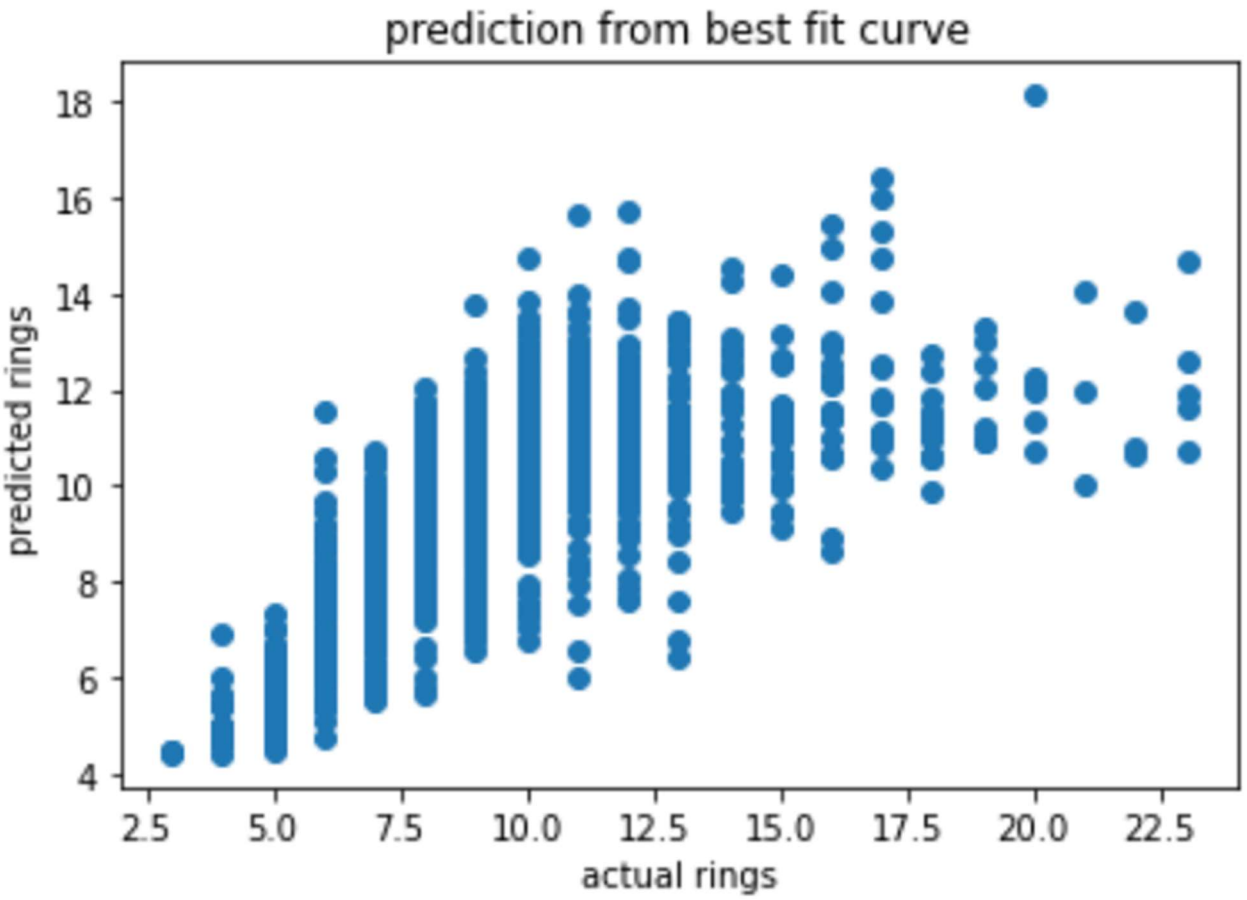

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. It is similar to the first classifier.
2. We are taking only one feature this time also.
3. Still multi variate is better than univariate.

4

a.

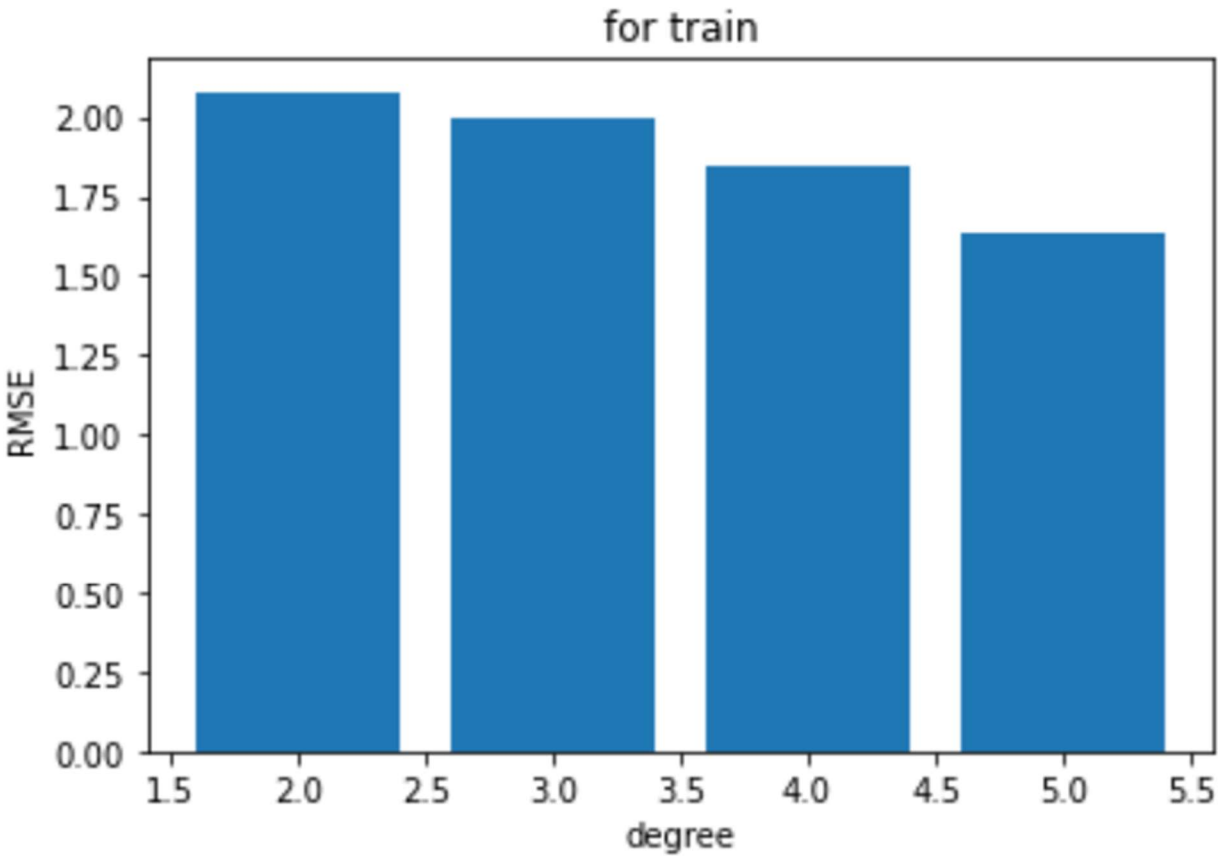

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE decreases for increase in p.
2. It decreases till p=5.
3. The overfitting might occur for higher degrees and then it will increase.

b.

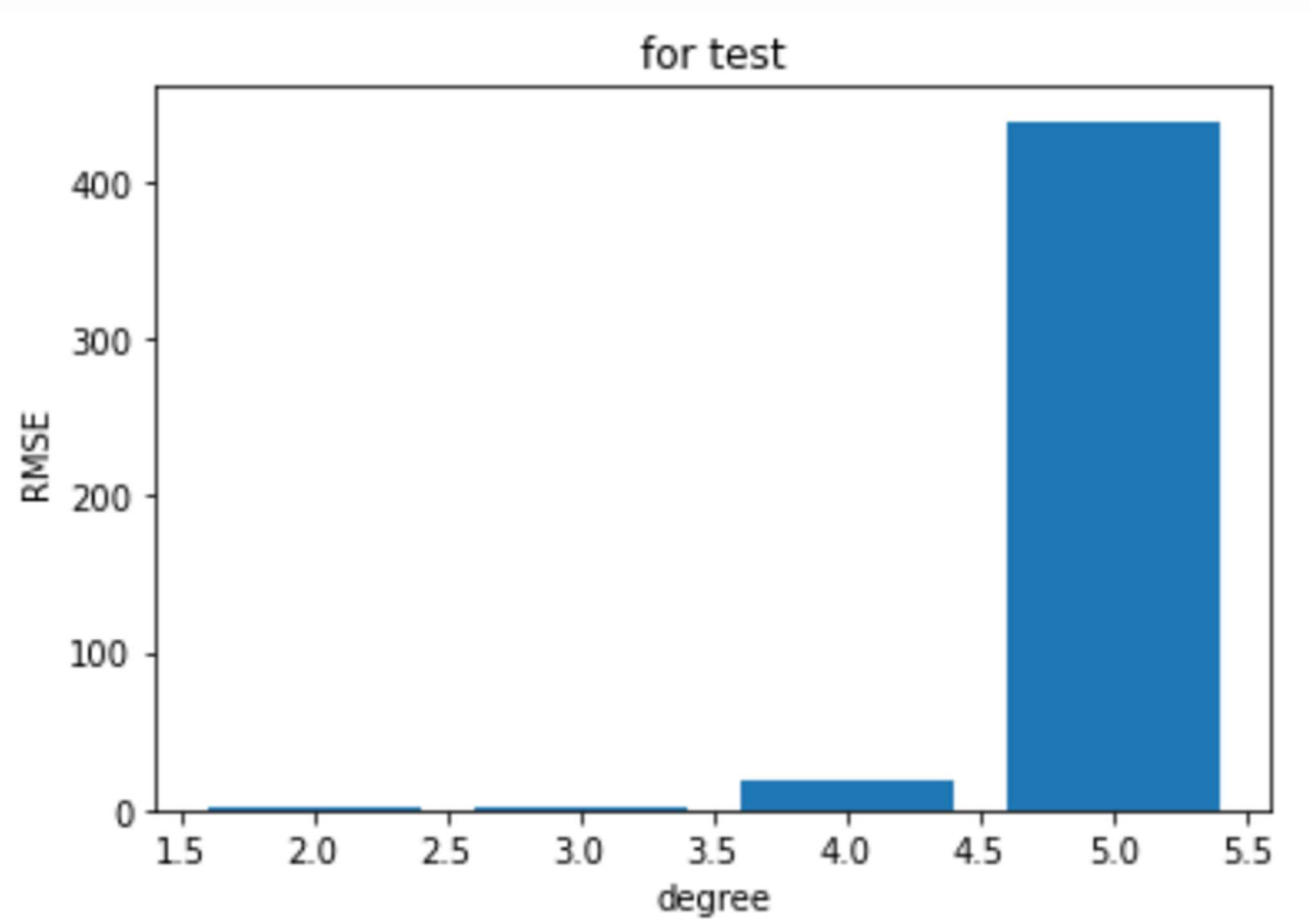

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE increases till p=5.
2. There is no decrease in between.
3. The overfitting started occurring after p=5.
4. From the RMSE value, 2nd degree curve will approximate the data best.

c.

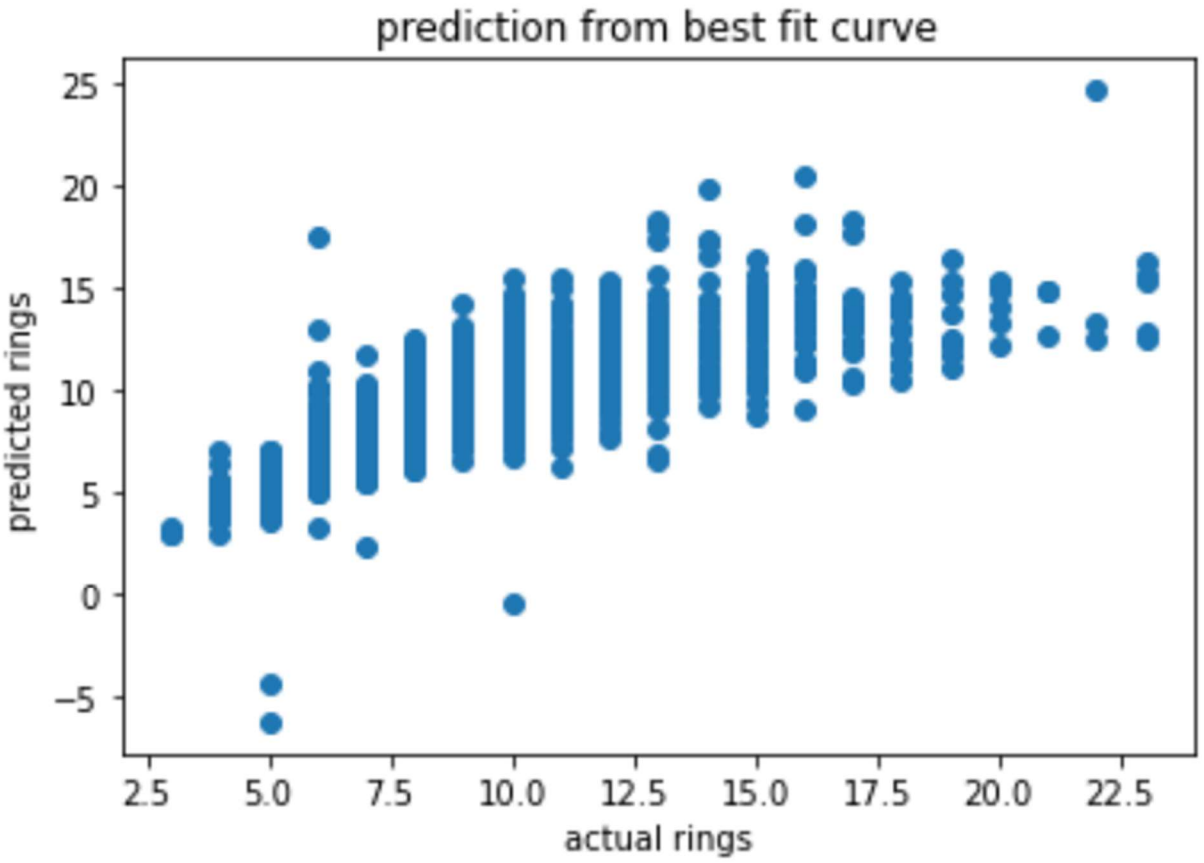

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. This prediction looks better than all the previous predictors.
2. Here, we are taking multivariate non-linear predictor.
3. Univariate is less accurate than multivariate.
4. Non-linear is more accurate than linear.