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: ISHAAN GUPTA **Mobile No: 9179242114**

Roll Number: B20292 **Branch:** MECHANICAL ENGINEERING

PART - A

1 a.

	Prediction Outcome	
True Label	95	13
	2	226

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	85	23
	3	225

Figure 3 Bayes GMM Confusion Matrix for Q = 8

	Prediction Outcome	
True Label	79	29
	2	226

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	**95.535**
4	**94.940**
8	**92.261**
16	**90.773**

Inferences:

1. The highest classification accuracy is obtained with Q =2.
2. Increasing the value of Q decreases the prediction accuracy.
3. Higher values of Q can make the data separated.

4. As the classification accuracy increases/decreases with the increase in value of Q infer does the number of diagonal elements in the confusion matrix increase/decrease.
5. State the reason for the increase/decrease in diagonal elements.
6. As the classification accuracy increases/decreases with the increase in value of Q infer does the number of off-diagonal elements increase/decrease.
7. State the reason for increase/decrease in off-diagonal elements.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.583
2.	KNN on normalized data	96.726
3.	Bayes using unimodal Gaussian density	95.833
4.	Bayes using GMM	95.535

Inferences:

1. The classifier with the highest accuracy is KNN on normalized data and lowest accuracy is KNN.
2. Arrange the classifiers in ascending order of classification accuracy. KNN < Bayes using GMM <Bayes using unimodal Gaussian density < KNN on normalized data.

PART – B

1

a.

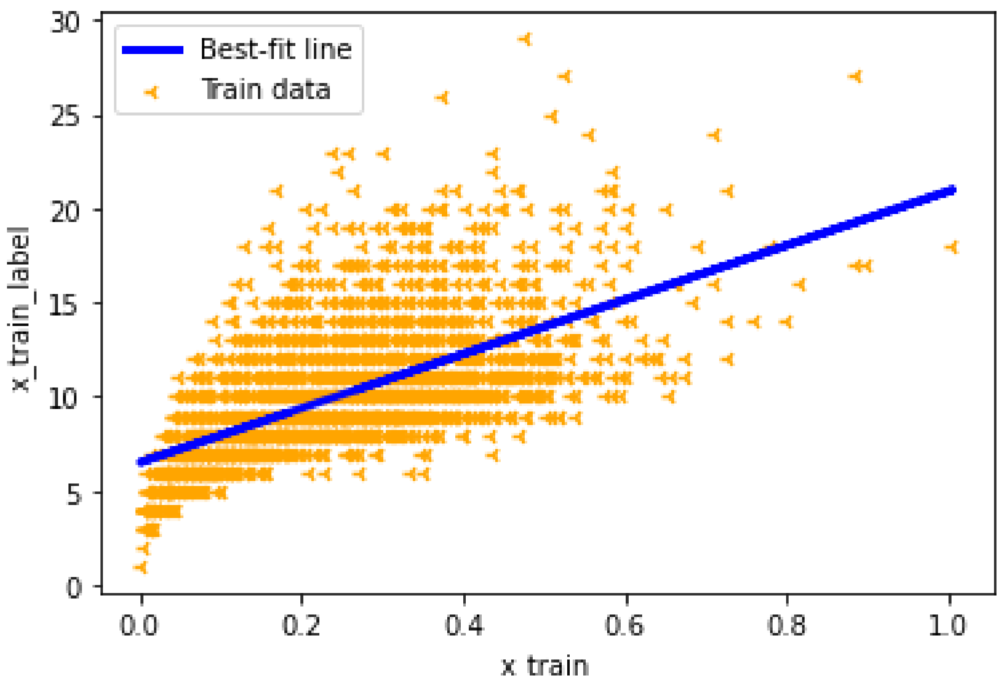


Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. The attribute with the highest correlation coefficient was used for predicting the target attribute Rings. Because it highly depend on that.
2. Does the best fit line fit the training data perfectly?
3. If not, why?
4. Infer upon bias and variance trade-off for the best fit line.

b.

The prediction accuracy on the training data using root mean squared error: 74.681 %

c.

The prediction accuracy on the testing data using root mean squared error: 74.859 %

Inferences:

1. Testing accuracy is higher than the training accuracy.

d.

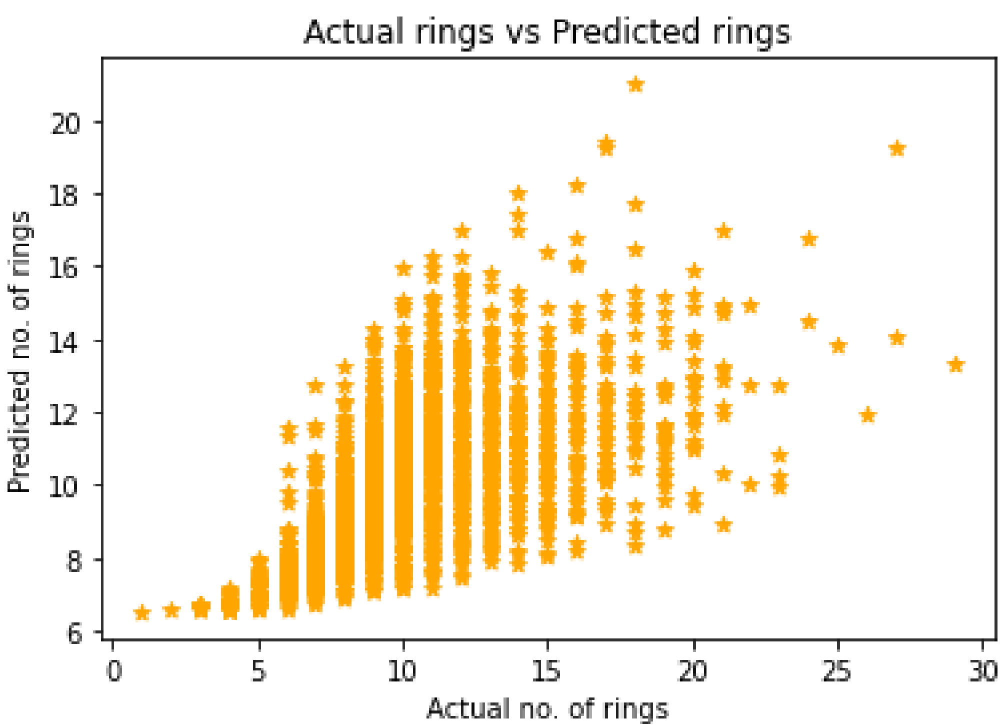


Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, infer how accurate the predicted temperature is?

a.

The prediction accuracy on the training data using root mean squared error:: 77.802 %

b.

The prediction accuracy on the testing data using root mean squared error:: 77.393 %

Inferences:

2. Training accuracy is higher.

c.

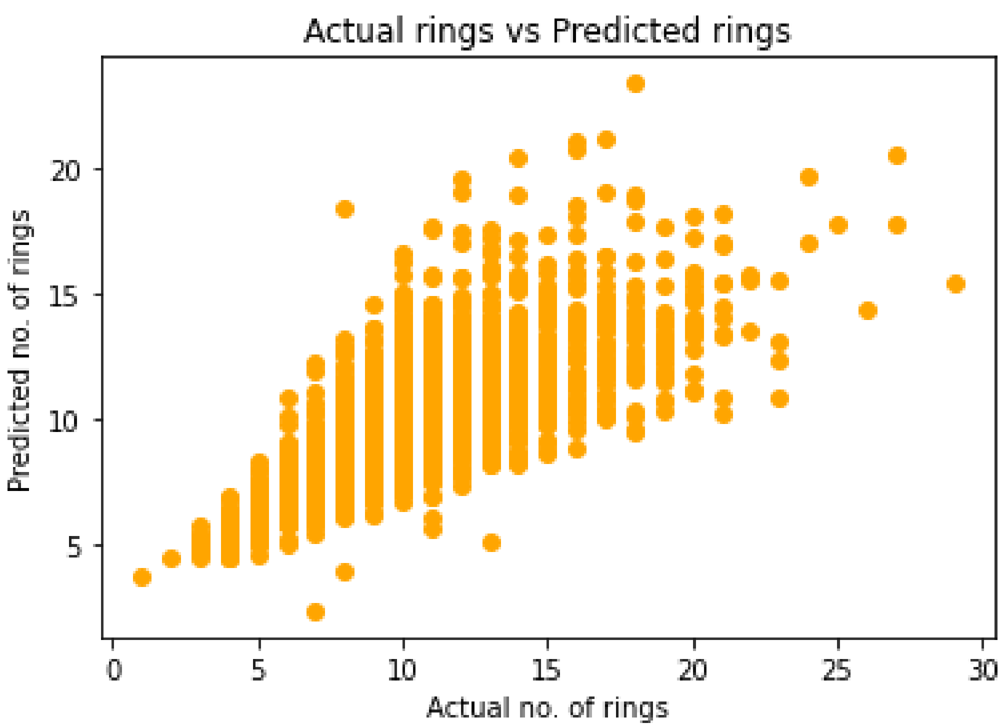


Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, infer how accurate the predicted temperature is?
2. State the reason for Inference 1.
3. Compare and contrast the performance of univariate linear with multivariate linear regression.

2

a.

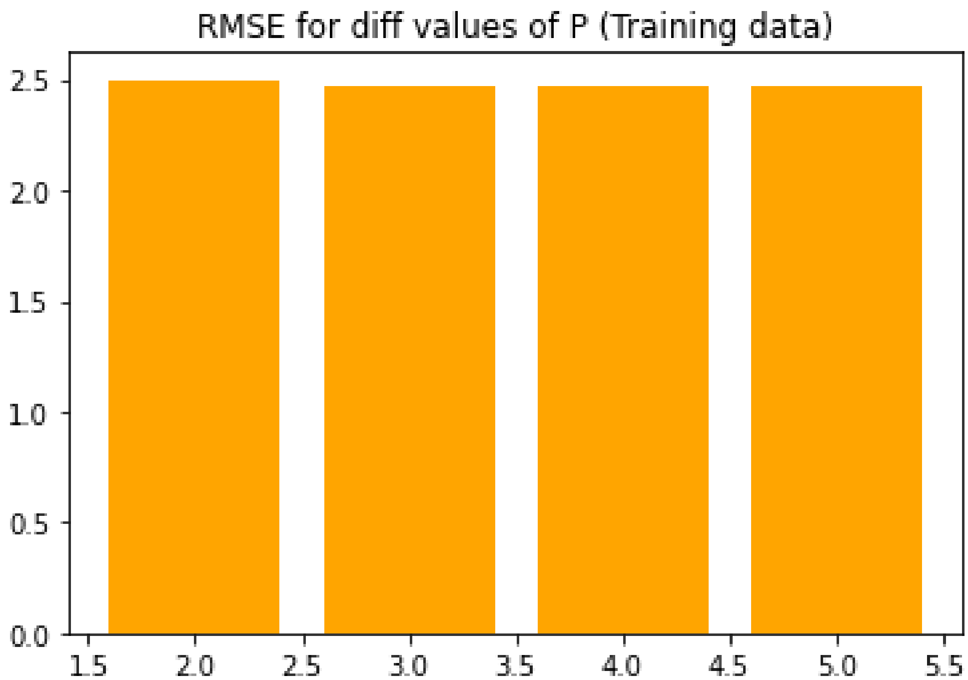


Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. Is the increase/decrease uniform or after a certain p-value the increase/decrease becomes gradual?
3. State the reason for Inference 1 and 2.
4. From the RMSE value, infer which degree curve will approximate the data best.
5. Infer based upon bias and variance trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

b.

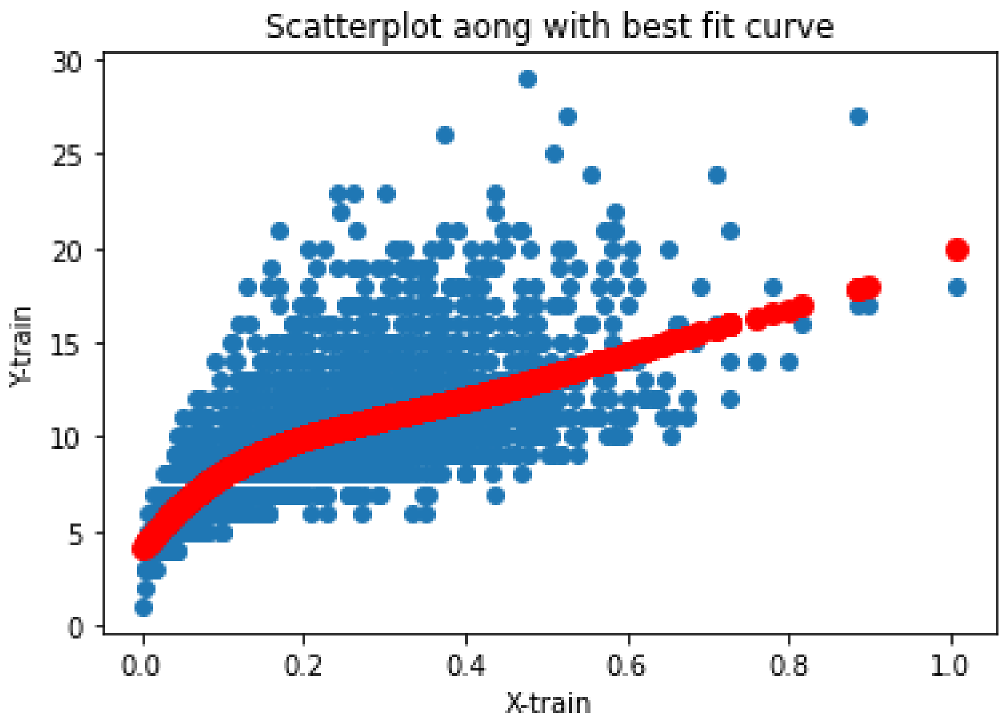


Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. Infer whether RMSE value decreases/ increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. Is the increase/decrease uniform or after a certain p-value the increase/decrease becomes gradual.
3. State the reason for Inference 1 and 2.

4. From the RMSE value, infer which degree curve will approximate the data best.
5. Infer based upon bias and variance trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

c.

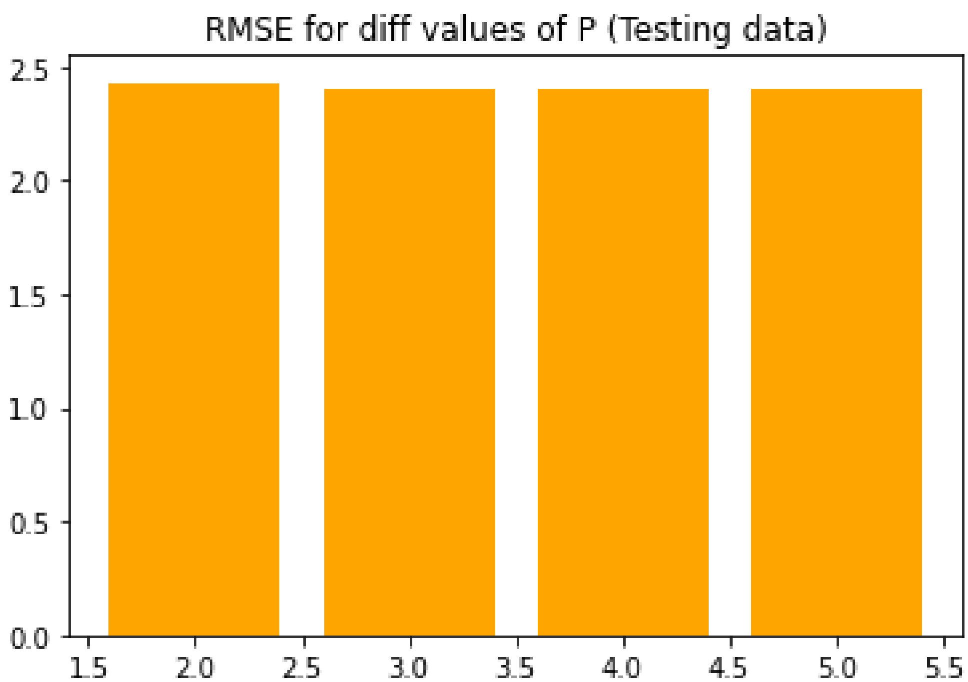


Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. State the p-value corresponding to the best fit model.
2. State the reason behind inference 1.
3. Infer based upon bias and variance trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

d.

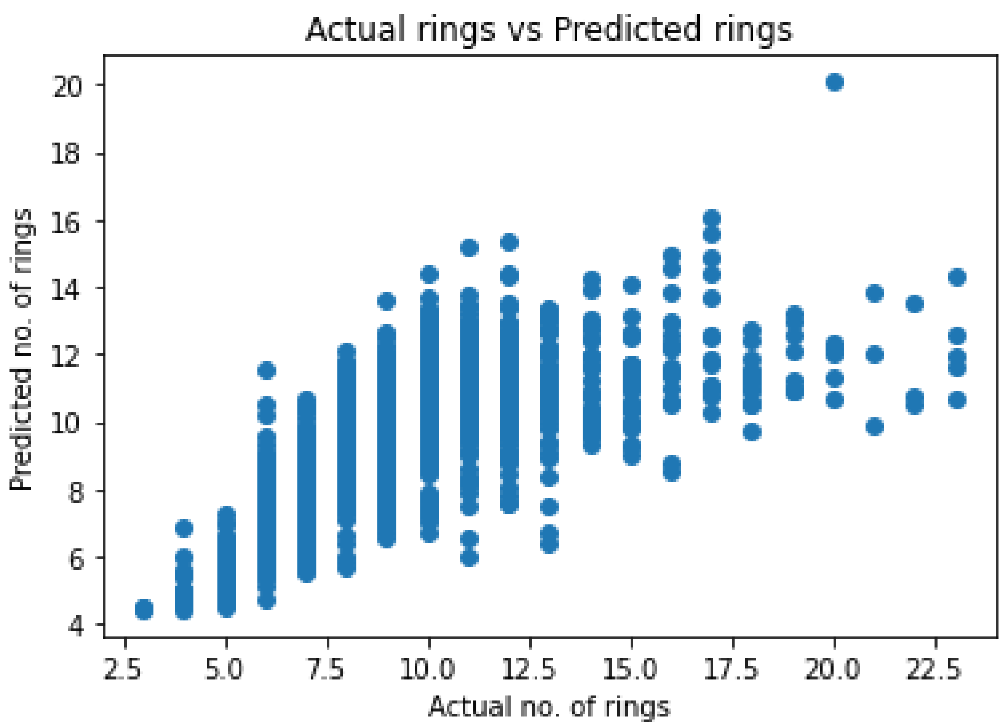


Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, infer how accurate the predicted temperature is?
2. State the reason for Inference 1.
3. Compare and contrast univariate linear, multivariate linear and non-linear regression model based upon the accuracy of predicted temperature value and spread of data points in Scatter Plot
4. State the reason for Inference 3.
5. Inference based upon bias and variance trade-off between linear and non-linear regression models.

Note: The above scatter plot is for illustration purposes only. Replace it with scatter plot obtained by you.

3

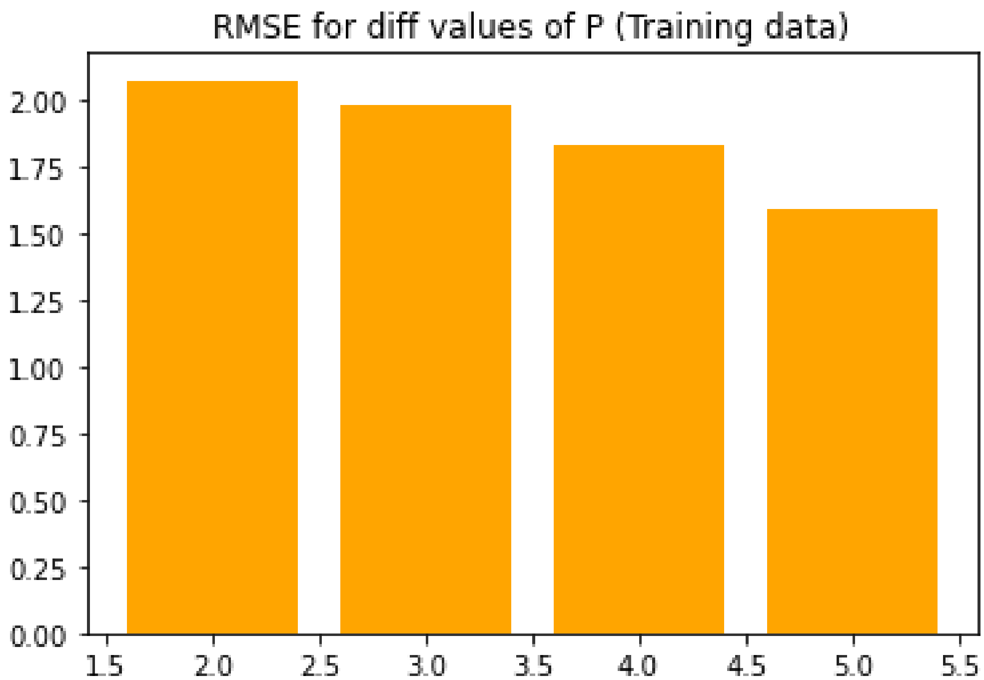
a. 

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. Infer whether RMSE value decreases/ increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. Is the increase/decrease uniform or after a certain p-value the increase/decrease becomes gradual?
3. State the reason for Inference 1 and 2.
4. From the RMSE value, infer which degree curve will approximate the data best.
5. Infer based upon bias and variance trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

b.

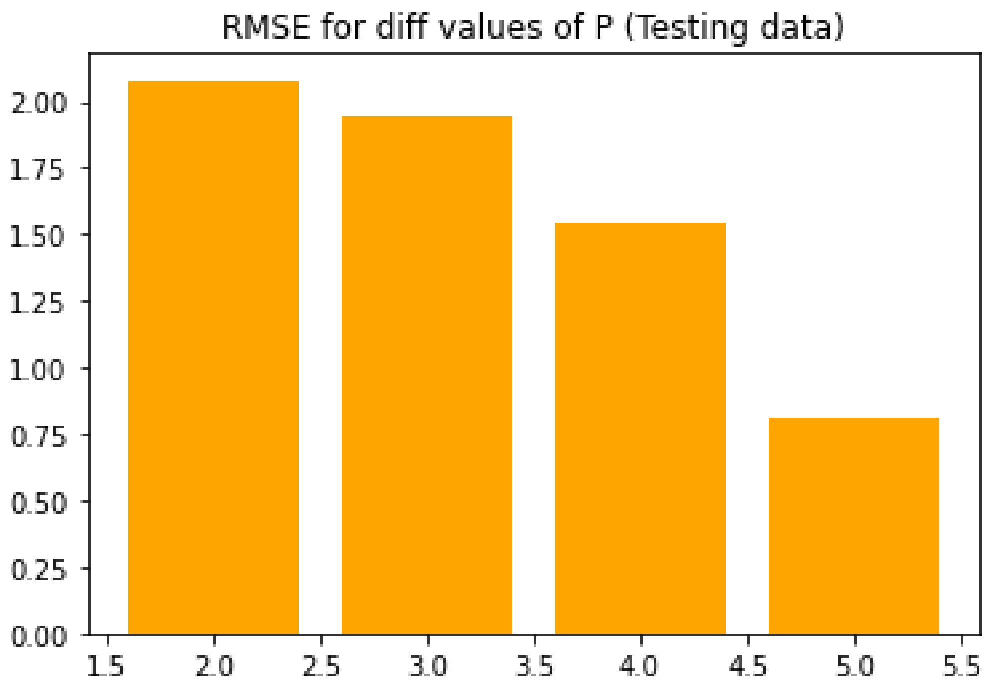


Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. Infer whether RMSE value decreases/ increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. Is the increase/decrease uniform or after a certain p-value the increase/decrease becomes gradual.
3. State the reason for Inference 1 and 2.
4. From the RMSE value, infer which degree curve will approximate the data best.
5. Infer based upon bias and variance trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

c.

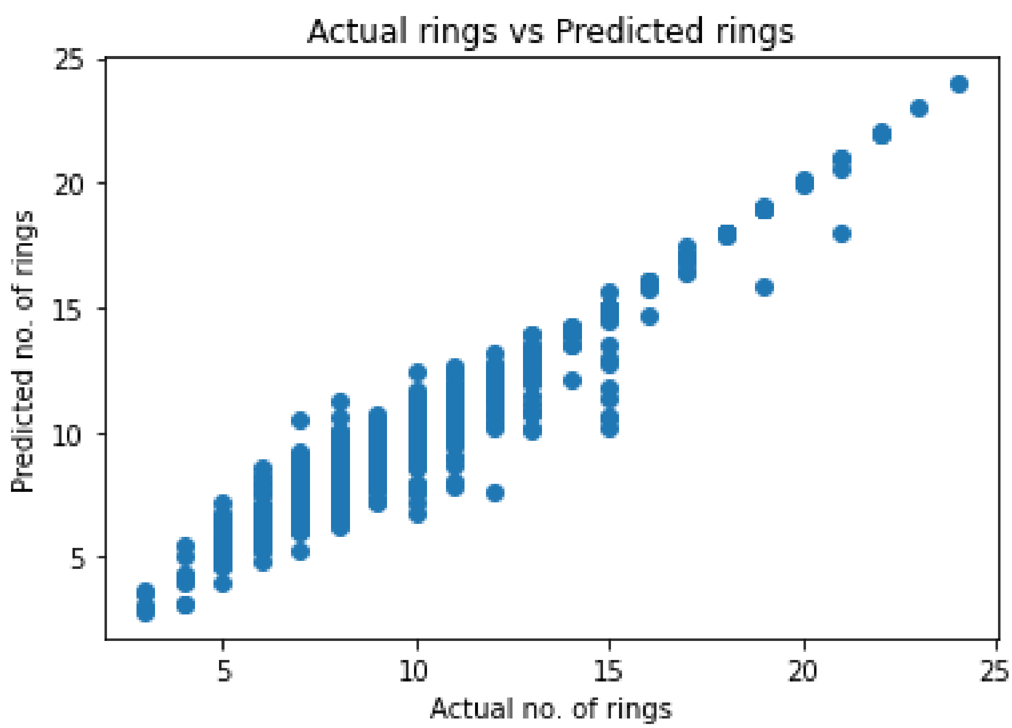


Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, infer how accurate the predicted temperature is?
2. State the reason for Inference 1.

3. Compare and contrast univariate linear, multivariate linear, univariate non-linear and multivariate non-linear regression model based upon the accuracy of predicted temperature value and spread of data points in Scatter Plot
4. State the reason for Inference 3.
5. Inference based upon bias and variance trade-off between linear and non-linear regression models.