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: Sachin Mahawar **Mobile No:** 9166843951

Roll Number: B20129 **Branch:** CSE

PART - A

1 a.

		Prediction Outcome
True Label	106	12
	5	214

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	111	7
	4	215

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	109	9
	8	211

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	91	27
	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	**94.95**
4	**96.73**
8	**94.95**
16	**91.39**

Inferences:

1. The highest classification accuracy is obtained with Q = 4.
2. Increasing the value of Q increases the prediction accuracy first and then starts decreasing.
3. It occurs because adding nodes with lower weight overfit the model on training data thus resulting in lower accuracy.

4. As the classification accuracy increases with the increase in value of Q the number of diagonal elements in the confusion matrix increase.
5. As accuracy improves, more correct predictions and fewer incorrect predictions are made, resulting in higher true positive and true negative frequencies.
6. As the classification accuracy increases with the increase in value of Q the number of off-diagonal elements in the confusion matrix decreases.
7. Increased accuracy means fewer incorrect predictions and fewer false positives and negatives, lowering the frequency of false positives and negatives.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	**89.614 %**
2.	KNN on normalized data	**97.329 %**
3.	Bayes using unimodal Gaussian density	**94.362 %**
4.	Bayes using GMM	**96.735%**

Inferences:

1. KNN on normalized data have highest and KNN have the lowest accuracy.
2. The classifiers in ascending order of classification accuracy : KNN < Bayes using unimodal Gaussian density < Bayes using GMM < KNN on normalized data.
3. As KNN is based on Euclidean distance so by normalizing data we are less prone to error that's why normalized data has more accuracy than nonnormalized data. And bayes is very simple classifier but given data is real world complex its accuracy is less than KNN But Multimodal Bayes performs better as we are now using multiple clusters which increases the relative accuracy.

PART – B

1

a.

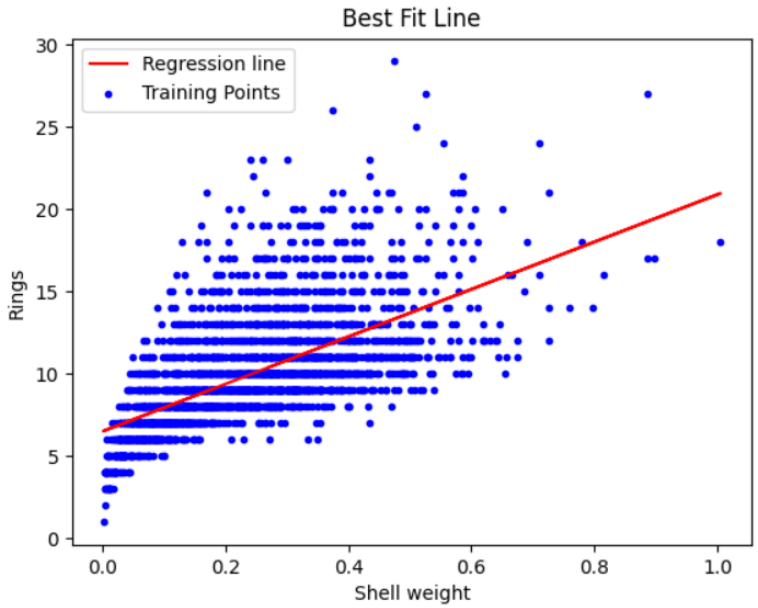

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. Because the target attribute is anticipated to be more dependent on the property with the highest correlation coefficient, so attribute with the highest correlation coefficient is used to forecast the target attribute Rings.
2. No, the best fit line does not fit the training data perfectly
3. Because it is simple fit line for this data, it needs complex curve to perfectly fit the data.
4. The bias is high and variance is low.

b.

The prediction accuracy on training data is 2.5278.

c.

The prediction accuracy on test data is 2.4679.

Inferences:

1. Training data accuracy is higher
2. Because we have trained the model with training data set that's why it has higher accuracy on training data.

d.

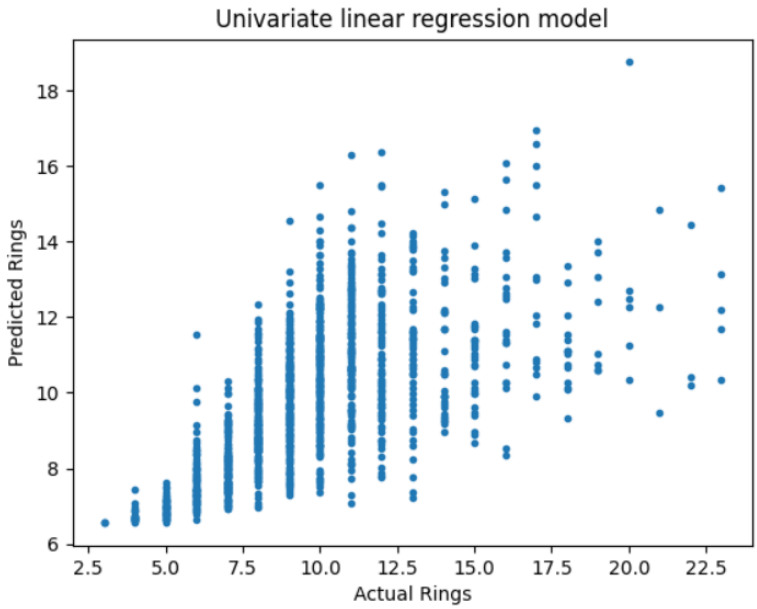

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based on spread of points, predicted number of rings is not very accurate.
2. Because the spread of data is very high that's why the prediction is not that accurate.

2

a.

The prediction accuracy on training data is 2.2161.

b.

The prediction accuracy on testing data is 2.2192.

Inferences:

1. Both the training and testing data have almost same accuracy.
2. Because we have trained the model with training data set that's why it has higher or same accuracy on training data.

c.

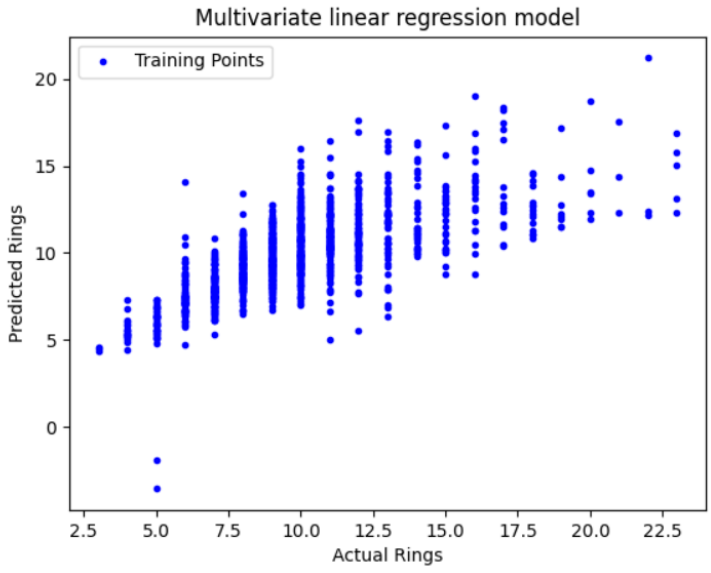

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based on spread of points, predicted number of rings is accurate.
2. Because the spread of data is quite low that's why it has higher accuracy or prediction is accurate.

3. Multivariate linear regression performs better than univariate linear regression because in it we have more information available of dependent variables.

3

a.

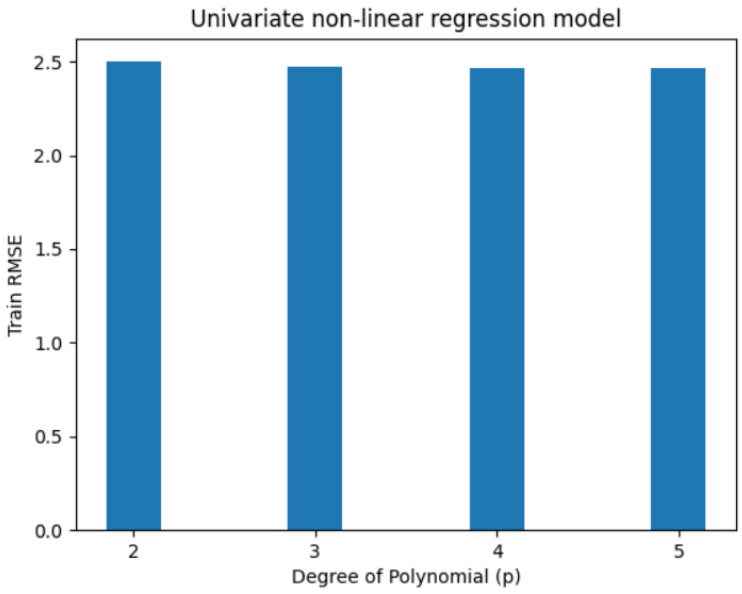

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with increase in degree of polynomial (p).
2. Decrement is more for p=2 to p=3 but after this it is more gradual.
3. With increase in degree of polynomial best fit curve fits the data better thus RMSE decreases.
4. From the RMSE value, curve for p=5 with approximate the data best.
5. Bias decreases and variance increases with increase in the degree of polynomial (p).

b.

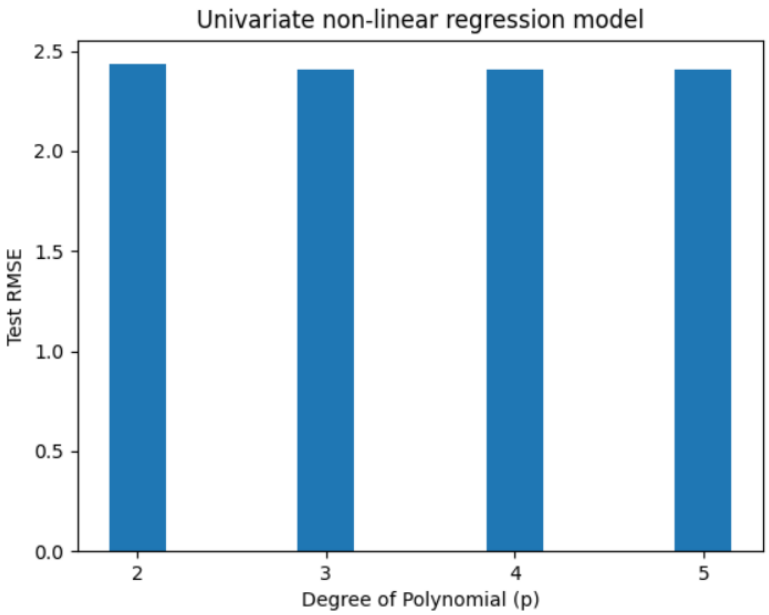

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with increase in degree of polynomial p=2,3,4 then increases for p=5.
2. Decrement is more for p=2 to p=3 but after this it is more gradual.
3. With increase in degree of polynomial best fit curve fits the data better thus RMSE decreases.
4. From the RMSE value, curve for p=4 will approximate the data best.
5. Bias decreases and variance increases with increase in the degree of polynomial (p).

c.

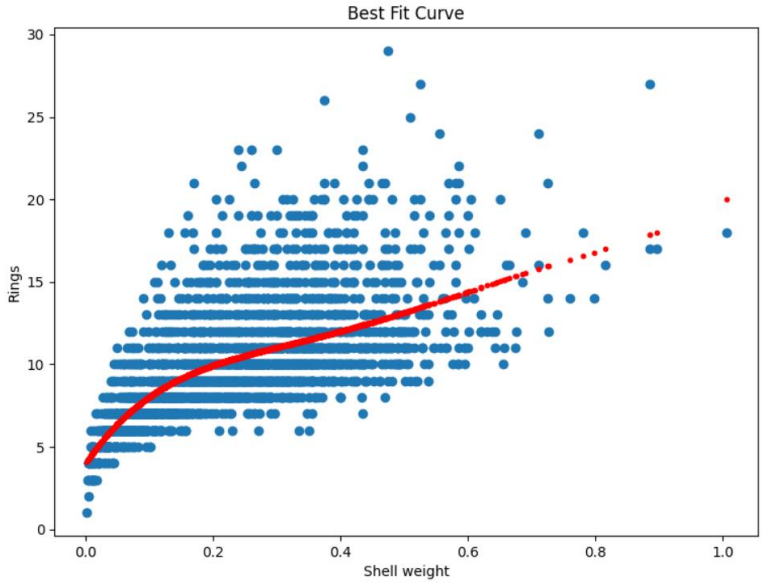

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. p-value corresponding to the best fit model is 5.
2. Because for p=5, it has high accuracy thus fitting data more accurately.
3. Bias decreases and variance increases with increase in the degree of polynomial (p).

d.

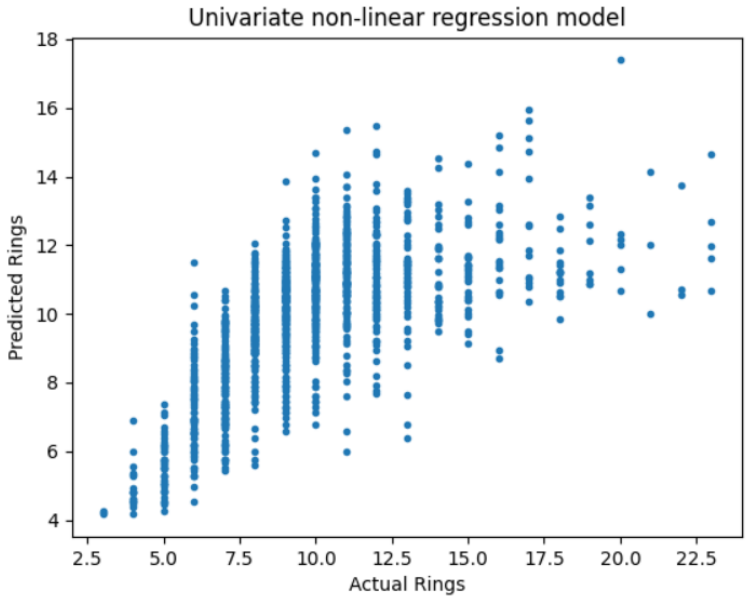

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based on spread of points, predicted number of rings is quite accurate.
2. Because the spread of data is quite low that's why it has higher accuracy or prediction is accurate.
3. Accuracy is as follows : Univariate non-linear > Multivariate Linear model > univariate linear model.
4. Because the RMSE values are lower for non-linear regression model than linear models.
5. In linear regression models bias is high, variance is low but in non-linear regression models bias is low, variance is high.

4

a.

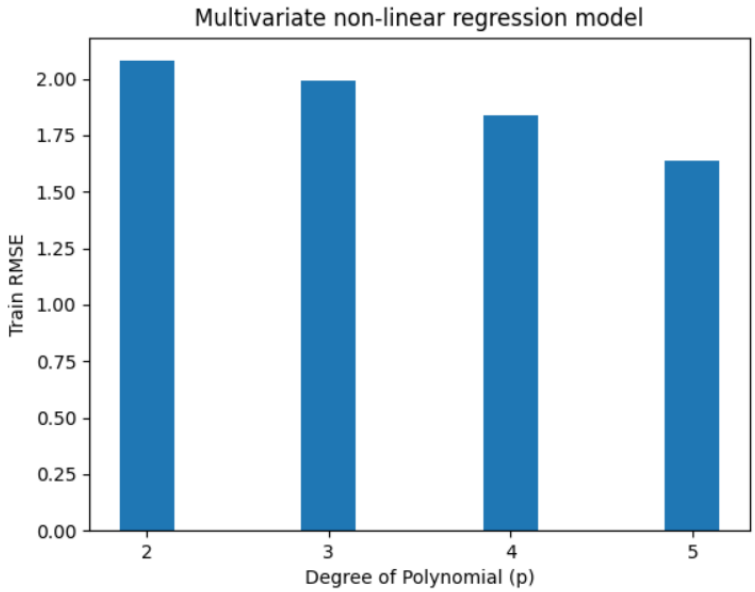

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with increase in degree of polynomial (p).
2. Decrement is gradual for p=2 to p=4 but after this it is significant.
3. With increase in degree of polynomial best fit curve fits the data better thus RMSE decreases.
4. From the RMSE value, curve for p=5 will approximate the data best.
5. Bias decreases and variance increase with increase in the degree of polynomial (p).

b.

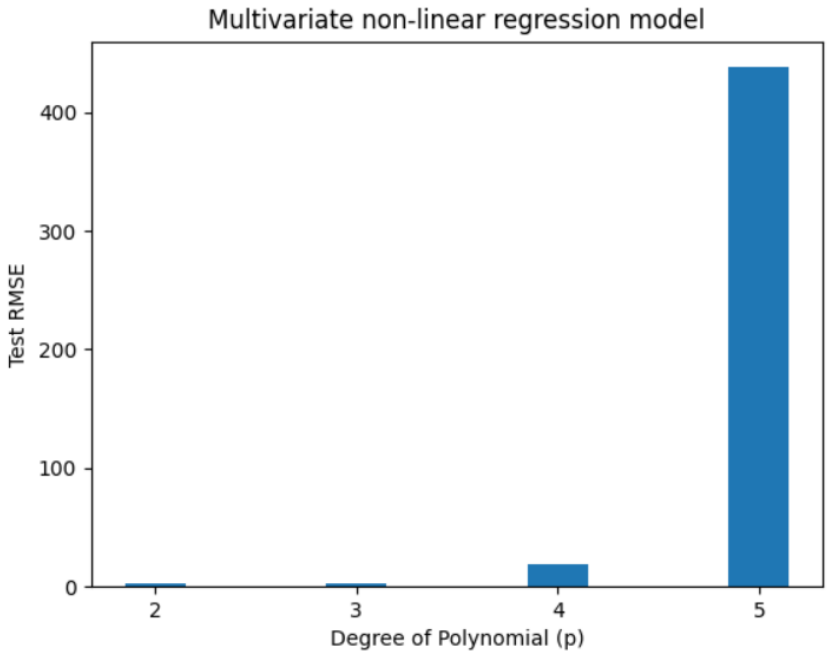

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value increases with increase in degree of polynomial (p) from p=1 to 5.
2. Decrement is gradual for p=1 to p=4 but after this it is much significant.
3. With increase in degree of polynomial our model is getting overfitted.
4. From the RMSE value, curve for p=1 will approximate the data best.
5. Variance increases with increase in the degree of polynomial (p) but bias decreases for p=1 to p=3 but after this bias increases significantly.

c.

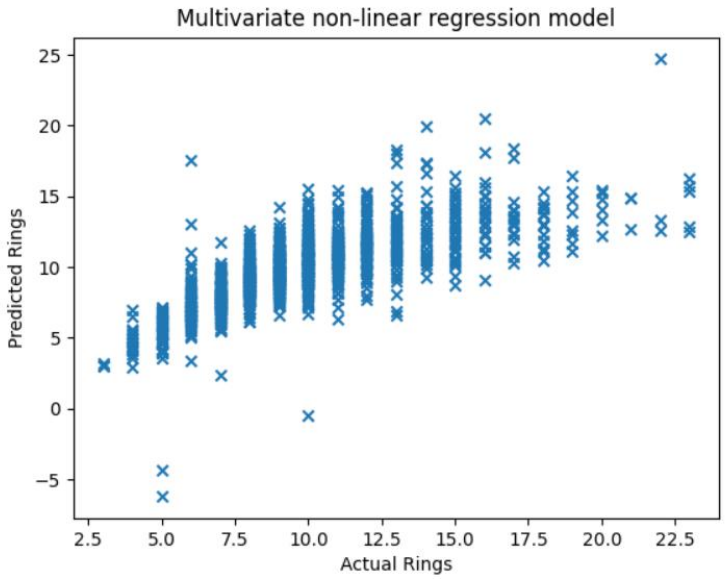

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based on spread of points, predicted number of rings is quite accurate.
2. Because the spread of data is quite low that's why it has higher accuracy or prediction is accurate.
3. Accuracy is as follows: Multivariate non-linear > Univariate non-linear > Multivariate Linear model > univariate linear model.
4. Because the RMSE values are lower for non-linear regression model than linear models and thus multivariate models are better than univariate.
5. In linear regression models bias is high, variance is low but in non-linear regression models bias is low, variance is high.