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: Sarthak Saptami Kumar Jha **Mobile No: 8825319259**

Roll Number: B20317 **Branch: CSE**

PART - A

1 a.

		Prediction Outcome
True Label	108	0
	0	208

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	107	1
	0	208

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	107	1
	0	208

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	107	1
	0	208

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	**100**
4	**99.70**
8	**99.70**
16	**99.70**

Inferences:

1. The highest classification accuracy is obtained with Q = 2.
2. There is a decrease in value of accuracy with increase in Q.
3. The reason can be due to lack of sufficient data, which leads to decrease in accuracy
4. Number of elements decrease.

5. Number of diagonal elements decrease due to reduction of accuracy.
6. The number of off-diagonal elements increase slightly.
7. Due to fall in accuracy with increase in Q.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.58
2.	KNN on normalized data	97.02
3.	Bayes using unimodal Gaussian density	86.31
4.	Bayes using GMM	100.00

Inferences:

1. Bayes using GMM has highest accuracy and Bayes using unimodal Gaussian density has lowest accuracy.
2. Bayes using GMM > KNN on normalized data > KNN > Bayes using unimodal Gaussian density.
3. Bayes using GMM has highest accuracy because of better classification methods than KNN.

PART – B

1

a.

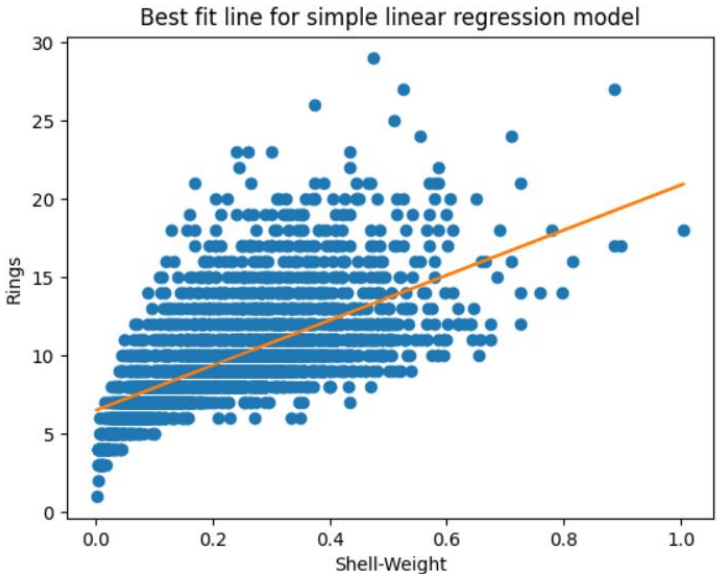

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. Shell Weight has the highest correlation with rings.
2. The best fit line does not fit the line quite well.
3. Because the it is a straight line and does not capture all details.
4. There is bias and variance trade-off because of lack of polynomial curve in the model.

b.

Prediction accuracy on training data = 2.52

c.

Prediction accuracy on testing data = 2.46

Inferences:

1. Training data has higher RMSE values than Testing data.
2. The reason can be due to the amount of data in the model.

d.

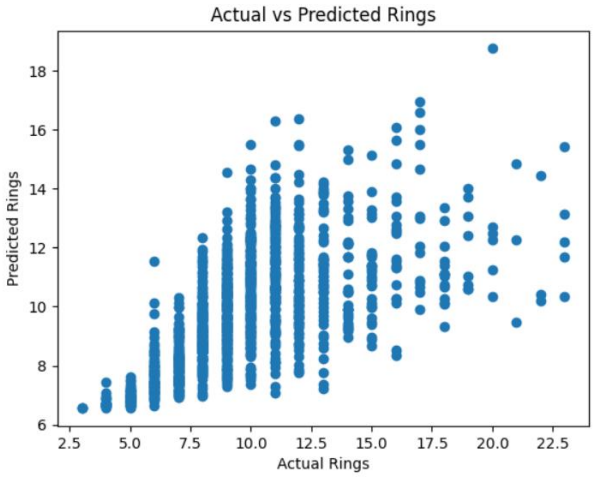

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. We can see that for lower range of values, the prediction is accurate and becomes less accurate with larger values.
2. There are some errors in prediction due to the linear nature of the prediction model.

2

a.

Prediction accuracy on training data = 2.2161

b.

Prediction accuracy on testing data = 2.2192

Inferences:

3. Testing data has higher RMSE than training data.
4. This is because with more data, the training data will have lower error compared to testing data.

c.

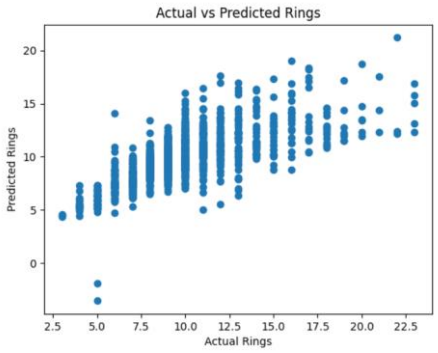

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The prediction is quite accurate.
2. The prediction is accurate as we can see that much of the data is closely correlated to the prediction.
3. There is better performance as we have included more data for the model.

3

a.

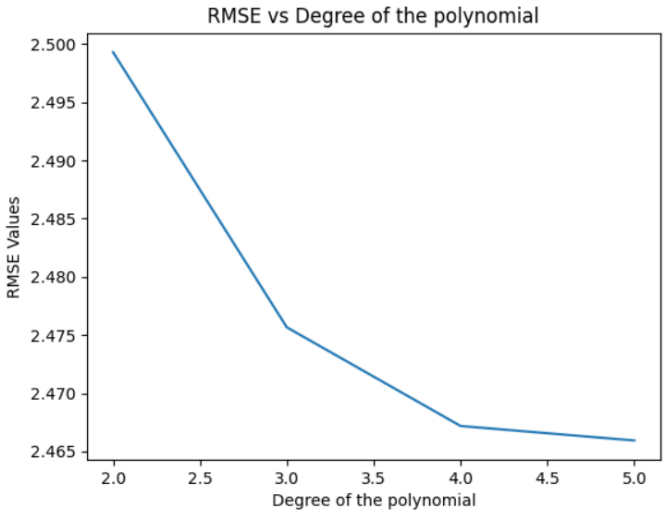

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value has a general downward trend.
2. The decrease of becomes more gradual after p = 4.
3. With increase in value of p, the curve fits better, but with more increase in value of p, there is over-fitting.
4. p = 5 fits the curve the best.
5. Lower p values have high bias and low variance, while higher p values have low bias and higher variance.

b.

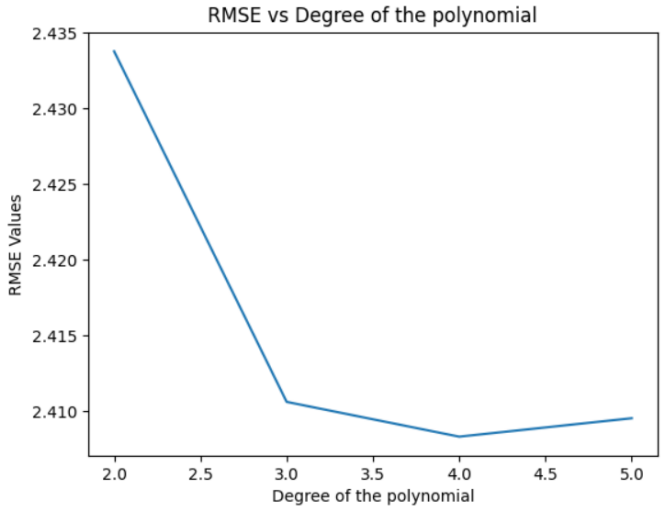

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

6. RMSE value has a general downward trend.
7. The decrease of RMSE after p = 3 is gradual.
8. With increase in value of p, the curve fits better, but with more increase in value of p, there is over-fitting.
9. p = 4 fits the curve the best.
10. Lower p values have high bias and low variance, while higher p values have low bias and higher variance.

c.

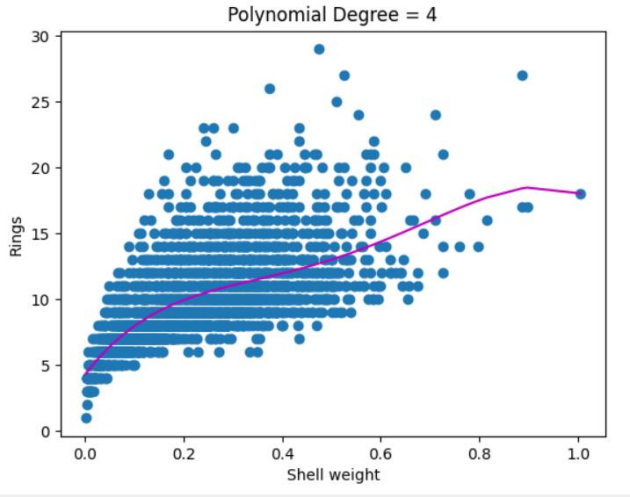

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. P = 4
2. For the value of p there is lowest error, and so it has the most adequate curve fitting.
3. Lower p values have high bias and low variance, while higher p values have low bias and higher variance.

d.

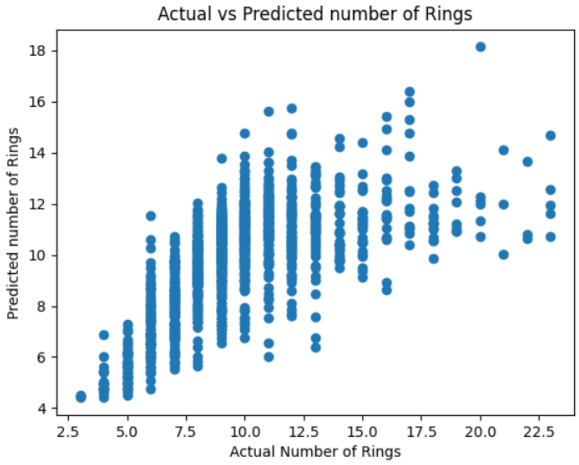

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. The predicted number of rings is quite accurate.
2. We have a strong correlation between the prediction data and the actual data thus, there is better prediction result.
3. Non-linear regression model is better than multivariate model which in turn is better than univariate model.
4. This is due to non-linear regression model having better curve fitting than the uni and multivariate models.
5. Lower p values have high bias and low variance, while higher p values have low bias and higher variance.

4

a.

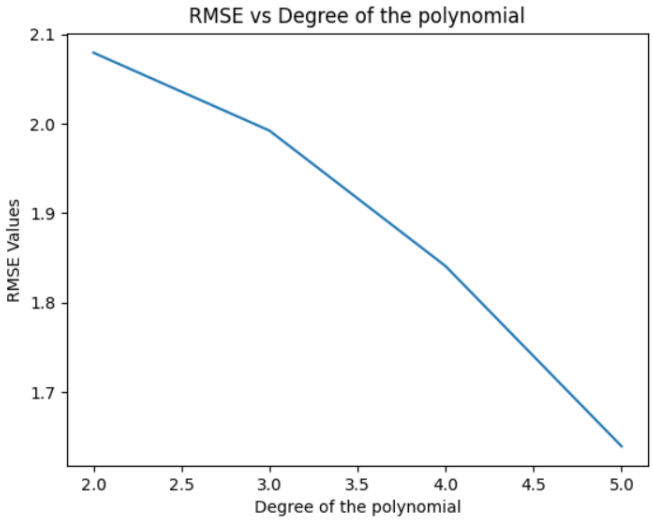

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. There is a more rapid decrease in value of RMSE with increase in p.
2. The trend of decrease is more rapid with increase in p.
3. With more data, taken into account, there is a steeper decline in error values.
4. P = 5 will fit the curve best,
5. Lower p values have high bias and low variance, while higher p values have low bias and higher variance.

b.

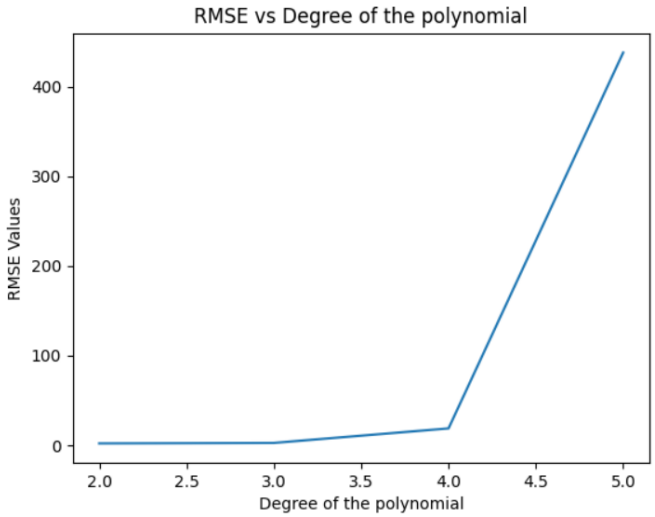

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. There is a general upward trend in RMSE values with increase in value of p.
2. After value of p=4, there is a rapid increase in the error values.
3. When we add new data to the model, the error increase for higher values of p due to over-fitting of the curve.
4. P = 2, will have the best fitting curve
5. Lower p values have high bias and low variance, while higher p values have low bias and higher variance.

c.

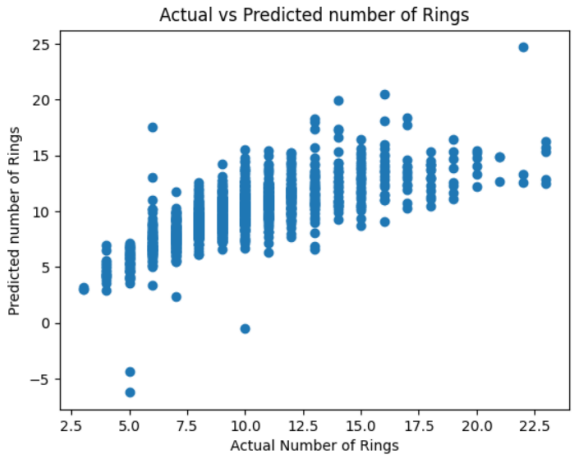

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. The prediction of data is accurate.
2. The prediction and actual data are quite correlated and thus are good predictions
3. Multivariate non-linear regression model is better than univariate non-linear due to more data being considered for the model. The non-linear models are comparatively better than univariate linear, multivariate linear models.
4. Non-linear regression curves fit the data much better than linear models, and thus have better prediction data.
5. Linear regression models have higher bias and variance than non-linear regression models.