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: Arpit Singh **Mobile No: 6265104315**

Roll Number: B20084 **Branch: CSE**

PART - A

1 a.

	Prediction Outcome	
True Label	106	12
	6	213

Figure 1 Bayes GMM Confusion Matrix for Q = 2

	Prediction Outcome	
True Label	111	7
	5	214

Figure 2 Bayes GMM Confusion Matrix for Q = 4

	Prediction Outcome	
True Label	111	7
	7	212

Figure 3 Bayes GMM Confusion Matrix for Q = 8

	Prediction Outcome	
True Label	93	25
	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	**94.658**
4	**96.439**
8	**95.845**
16	**91.691**

Inferences:

1. The highest classification accuracy is obtained with Q =4.
2. increasing the value of Q increases the prediction accuracy initially but later it decreases as we increase the value of Q.
3. Reason for increasing the value of Q increases then decreases the prediction accuracy is complexity of the model when Q becomes comparatively large it decreases the prediction accuracy.
4. As the accuracy decreases the number of diagonal elements decreases.
5. Diagonal matrix represents the number of times a model has correctly predicted the value. The model is predicting wrongly as the accuracy decreases.
6. As the classification accuracy decreases with the increase in value of Q the number of off-diagonal elements increases.
7. Off-diagonal matrix represents the number of times a model has wrongly predicted the value decrease in accuracy means the model is predicting wrongly.

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	96.439

Inferences:

1. KNN and KNN on normalised data are the classifiers with the lowest and highest accuracy.
2. KNN< Bayes using unimodal Gaussian density< Bayes using GMM< KNN on normalized data.
3. KNN performs better when data is normalized because data is more equally distributed. That is the reason for lowest accuracy of KNN without normalized data as attribute with wider spread is influencing the result. Multimodal bayes performs better as we are using multiple clusters which increases the relative accuracy.

PART – B

1

a.

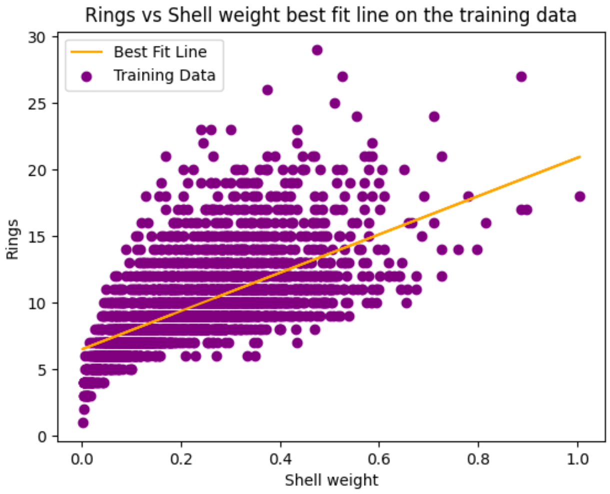

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute shell weight best fit line on the training data

Inferences:

1. The attribute with the highest correlation coefficient was used for predicting the target attribute Rings because it is easier to predict the target value if correlation is high.
2. No, the best fit line doesn't fit the training data perfectly.
3. It doesn't fit perfectly because it is oversimplified.
4. Bias is high and variance is low , trade-off for the best fit line.

b.

The prediction accuracy on training data is 2.527.

c.

The prediction accuracy on testing data is 2.466.

Inferences:

1. Training accuracy is higher.
2. Training accuracy is higher because it is on the same data we trained the model.

d.

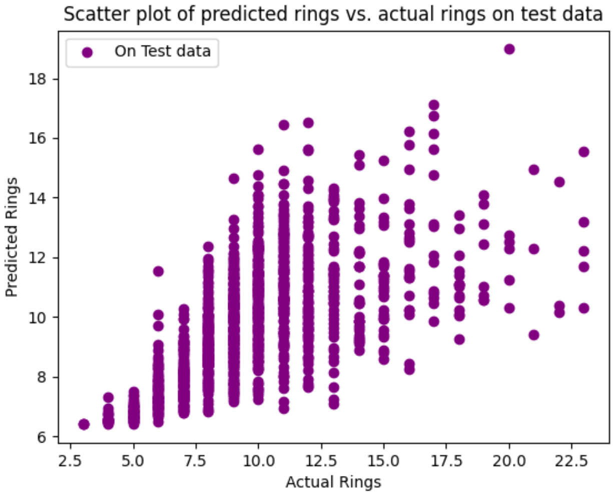

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, we didn’t predict the number of rings accurately.
2. We didn’t predict the number of rings accurately because the original spread of rings is 2 to 23 while ours is 6 to 20.

2

a.

The prediction accuracy on training data is 2.216.

b.

The prediction accuracy on testing data is 2.205.

Inferences:

3. The training data has slightly higher accuracy.

4. The difference between the accuracy of training and testing data is very small which shows that this model is good.

c.

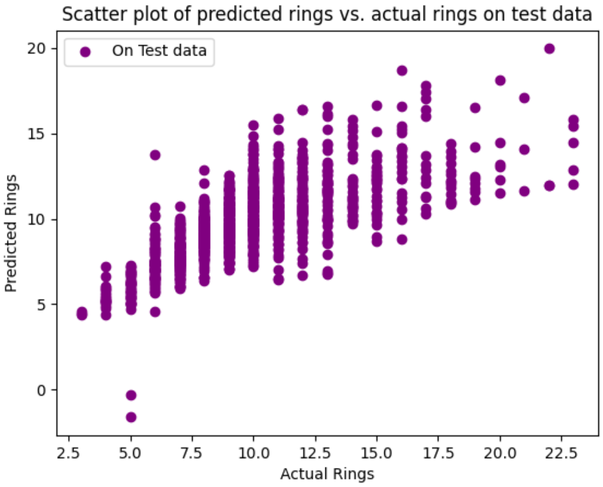

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, Number of rings is predicted to be higher than the original.
2. Reason for the higher prediction is because the original spread of rings is 3 to 23 while ours is -3 to 20.
3. The performance of multivariate linear regression is better than univariate linear regression.

3

a.

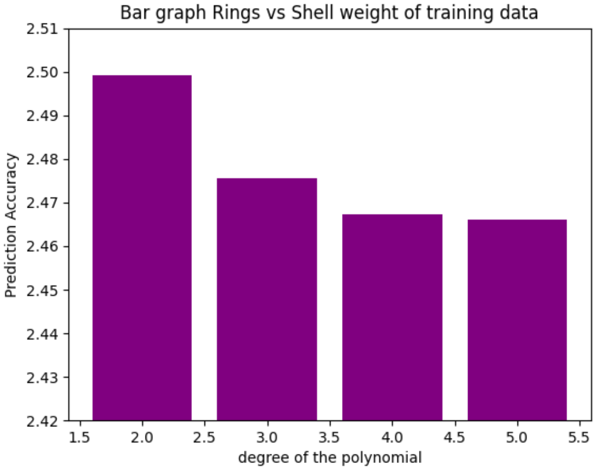

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases as the degree of the polynomial (p = 2, 3, 4, 5) increases.
2. Decrease is more for p = 2,3 but uniform after that.
3. The more the degree, the better fitting of the data.
4. From the RMSE value, the degree p=5 curve will approximate the data best.
5. Bias decreases and variance increases, trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

b.

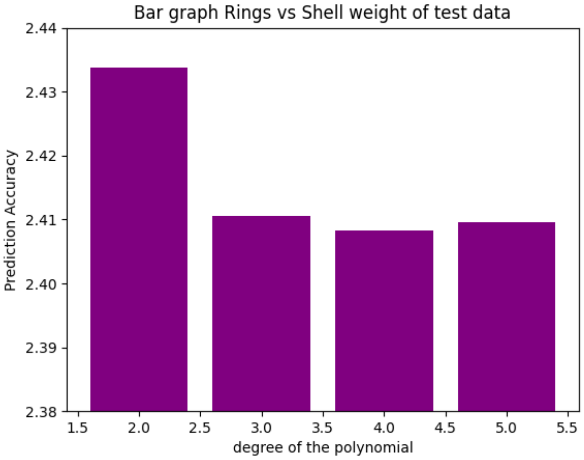

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases as the degree of the polynomial (p = 2, 3, 4, 5) increases.
2. Decrease is more for p = 2,3 but uniform after that.
3. The more the degree, the better fitting of the data.
4. From the RMSE value, the degree p=4 curve will approximate the data best.
5. Bias decreases and variance increases, trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

c.

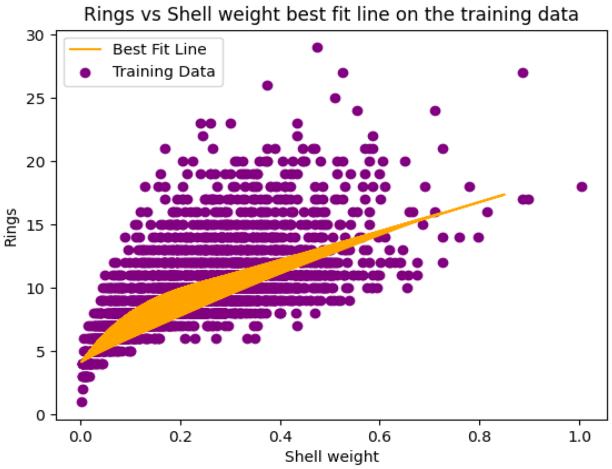

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute shell weight best fit curve using best fit model on the training data

Inferences:

1. The p-value corresponding to the best fit model is 4.
2. The p-value corresponding to the best fit model is 4 because it has more variance.
3. Bias decreases and variance increases, trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

d.

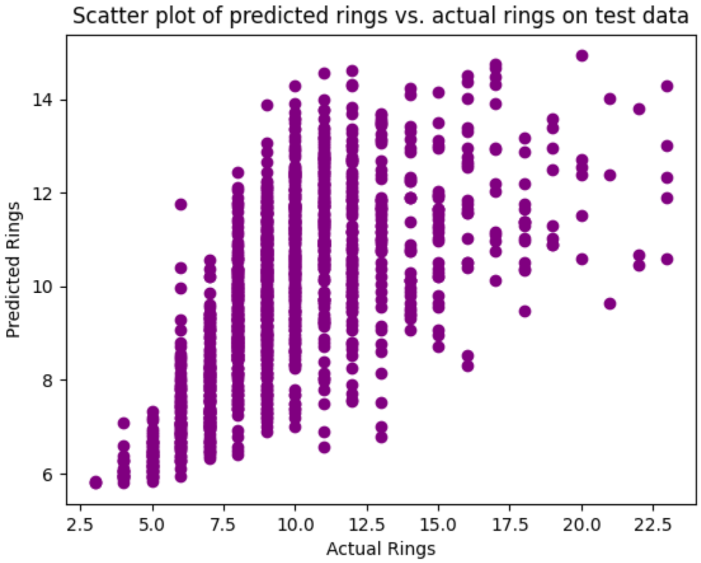

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, The predicted data is almost accurate.
2. The spread of actual rings is 2 to 23 and predicted is 5 to 21.
3. The accuracy of the univariate non-linear model is highest then multivariate linear then univariate linear regression model.
4. RMSE value of non-linear regression is lower than linear and multivariate is better than univariate regression.
5. Bias is higher and lower and variance is lower and higher trade-off between linear and non-linear regression models respectively.

4

a.

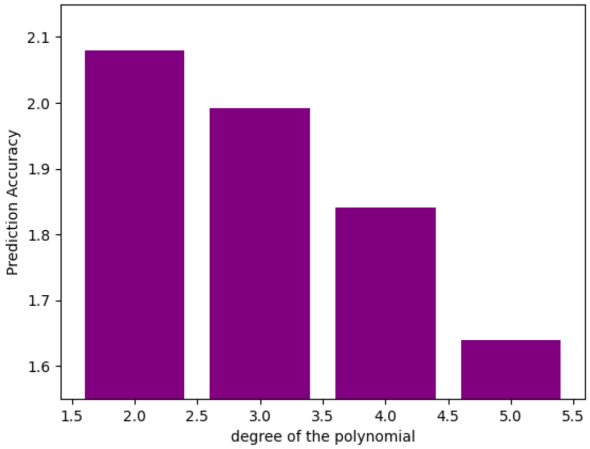

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases as the degree of the polynomial (p = 2, 3, 4, 5) increases.
2. The decrease is more after 4 before that decrease is gradual.
3. More degree of polynomial better the fitting.
4. From the RMSE value, the degree p=5 curve will approximate the data best
5. Bias decreases and variance increases, trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

b.

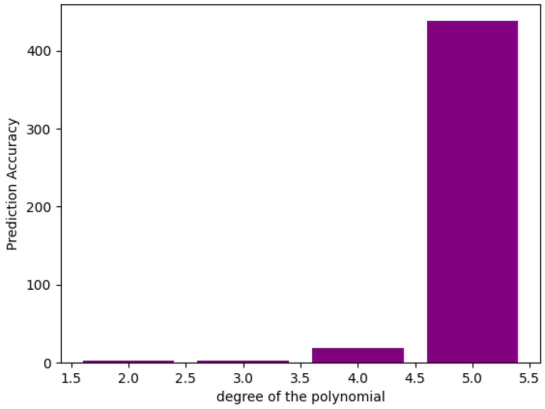

Figure 13 Multivariate non-linear regression model: RMSE vs different values of degree of polynomial (p=2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases as the degree of the polynomial (p = 2, 3, 4, 5) increases but here it is increasing after p=3.
2. Decrease is uniform till p=3 after that it increases.
3. Our model is becoming overfitted.
4. From the RMSE value, the degree p=2 curve will approximate the data best
5. Bias decreases till p=3 after that it increases rapidly and variance increases, trade-off with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

c.

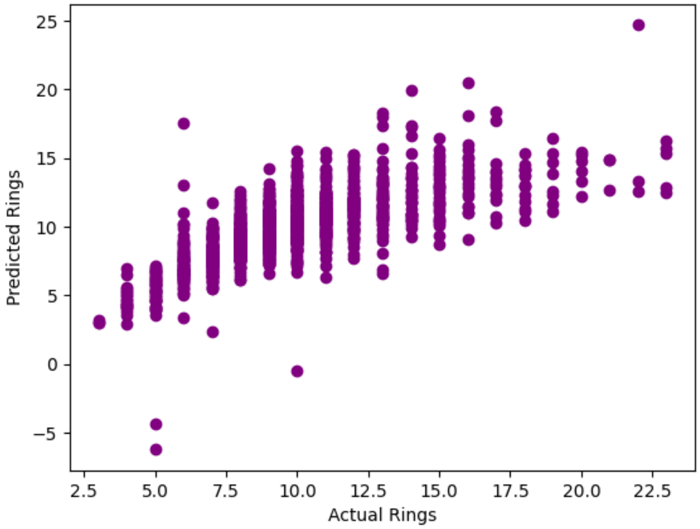

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, The predicted data is almost accurate.
2. The spread of actual rings is 2 to 23 and predicted is 3 to 22.
3. The accuracy of multivariate non-linear is highest then univariate non-linear then multivariate linear then univariate linear regression model.
4. RMSE value of non-linear regression is lower than linear and multivariate is better than univariate regression.
5. Bias is higher and lower and variance is lower and higher trade-off between linear and non-linear regression models respectively.