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: Priyanka kumari **Mobile No: 8328354314**

Roll Number: B20307 **Branch: ELECTRICAL ENGINEERING**

PART - A

1 a.

	Prediction Outcome	
True Label	95	13
	3	225

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	83	25
	4	224

Figure 3 Bayes GMM Confusion Matrix for Q = 8

	Prediction Outcome	
True Label	77	31
	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	**95.536**
4	**94.940**
8	**91.369**
16	**90.179**

Inferences:

1. The highest classification accuracy is obtained with Q = 2 which is equal to 95.536%
2. Increasing the value of Q decreases the prediction accuracy.

3. The perfect number of clusters for the above model must be around 2 and thus increasing the value of number of clusters away from this true value will decrease the classification accuracy of our model.
4. As the classification accuracy decreases with the increase in value of Q , the number of diagonal elements in the confusion matrix decreases.
5. The reason for decrease in diagonal elements is due to decrease in the number of true positive and negative values due to decrease in classification accuracy.
6. As the classification accuracy decreases with the increase in value of Q the number of off-diagonal elements increase.
7. The reason for increase in off-diagonal elements is the increase in false positive and negative values due to decrease in classification accuracy.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.60
2.	KNN on normalized data	1.00
3.	Bayes using unimodal Gaussian density	93.75
4.	Bayes using GMM	95.20

Inferences:

1. The classifier with the highest accuracy is ‘KNN on normalized data’ and lowest accuracy is ‘KNN’.
2. The classifiers in ascending order of classification accuracy. KNN < Bayes using GMM < Bayes using unimodal Gaussian Density < KNN on normalized data.
3. Since we assume independence between different attributes in Bayes Classifier , it becomes less effective than the KNN on normalized data for the given dataset with 24 attributes and it is more effective than simple KNN classifier without normalization because of outliers present in the dataset which decreases the efficiency of KNN method. Here Bayes using GMM is less effective than Bayes using unimodal Gaussian density because the data may be unimodal.

PART – B

1

a.

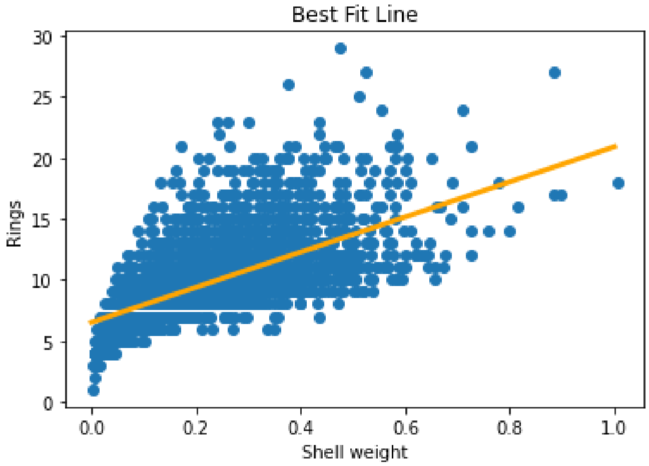

Figure 5: Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. The attribute with the highest correlation coefficient will be most efficient in predicting the target attribute Rings.
2. No, the best fit line does not fit the training data perfectly
3. Since the data is not perfectly linearly distributed thus the linear regression does not give a perfectly fitting line

b.

RMSE of training data = 2.528 .

c.

RMSE of testing data = 2.468 .

Inferences:

1. Test accuracy is slightly more than the training accuracy.
2. Since the RMSE for test data is 0.06 less than that of training data , it’s prediction accuracy is higher.

d.

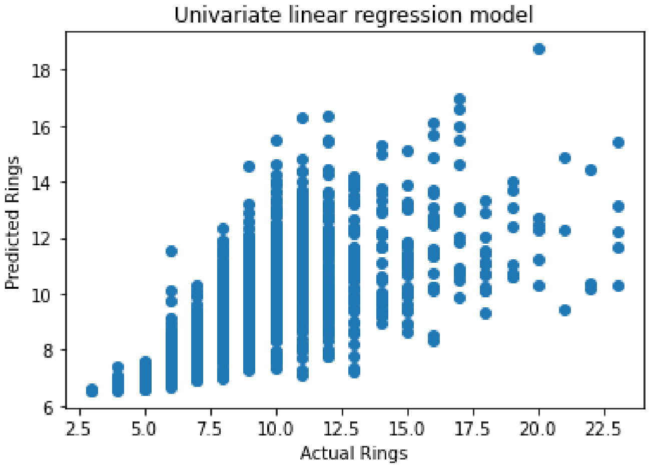

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, there seems to be a strong positive correlation between actual rings and predicted number of rings hence the prediction accuracy is quite good.
2. The RMSE for the predicted test dataset is very low and hence prediction accuracy is quite good.

2

a.

RMSE for training data = 2.216.

b.

RMSE for testing data = 2.205.

Inferences:

3. Testing accuracy is higher than training accuracy.
4. Since RMSE for testing data is less than training data thus testing accuracy is higher than training accuracy.

c.

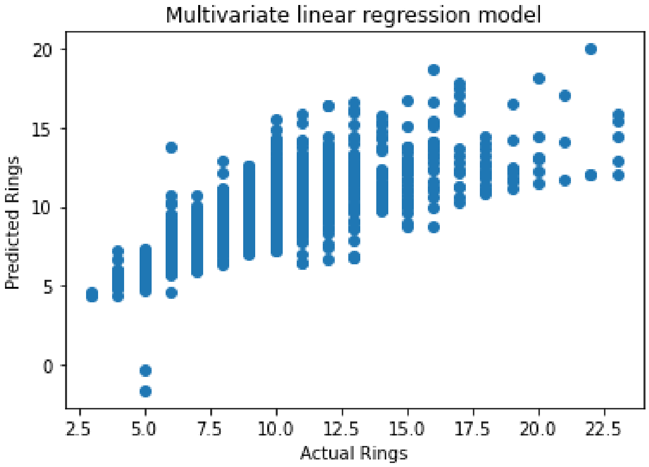

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, there seems to be strong positive correlation between actual and predicted rings ,thus the accuracy of our model is quite good.
2. The RMSE for test data prediction is very low and thus prediction accuracy is high.
3. The performance of multivariate linear is quite good than univariate linear regression model due to low RMSE.

3

a.

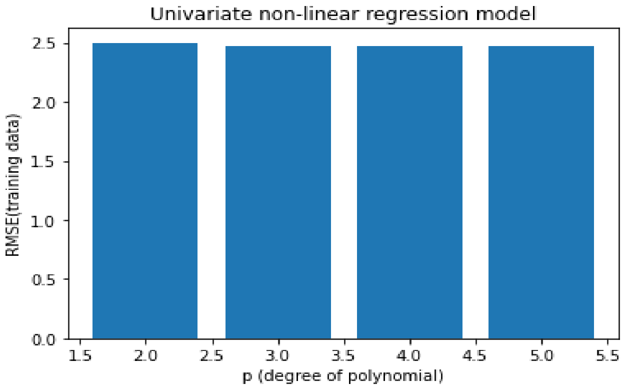

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. There is a gradual decrease after p = 4 as we can see the RMSE value is almost same for p =4 and p = 5
3. From the RMSE value, regression function polynomial of degree = 5 will approximate the data best.

b.

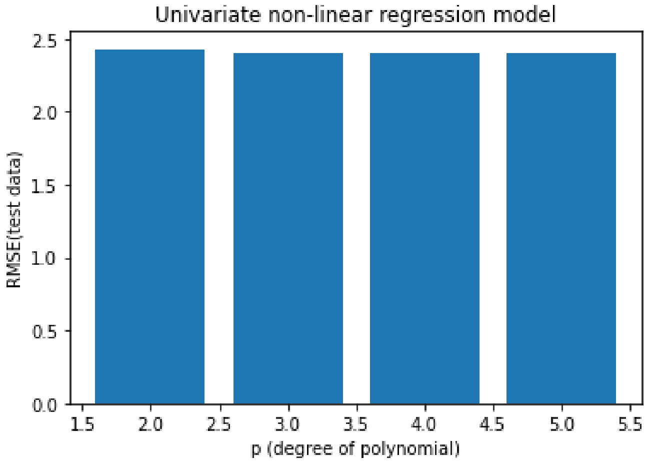

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. The decrease becomes gradual after p = 3.
3. The data can be perfectly predicted by the polynomial of degree 4.

c.

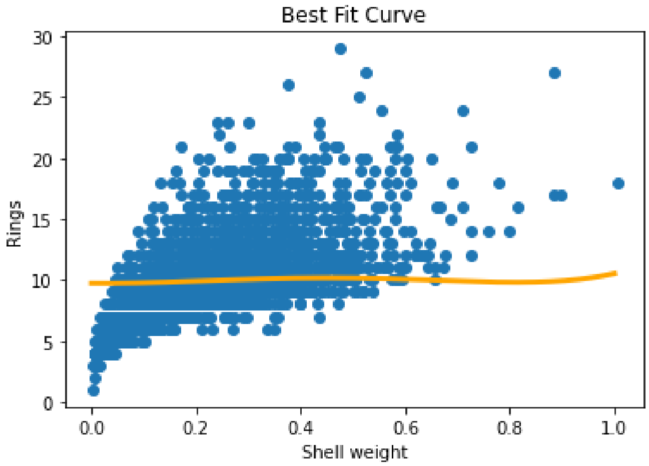

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. The p-value = 5 corresponds to the best fit model.
2. The RMSE value for p = 5 is the least and hence it is chosen to give the best fit model.

d.

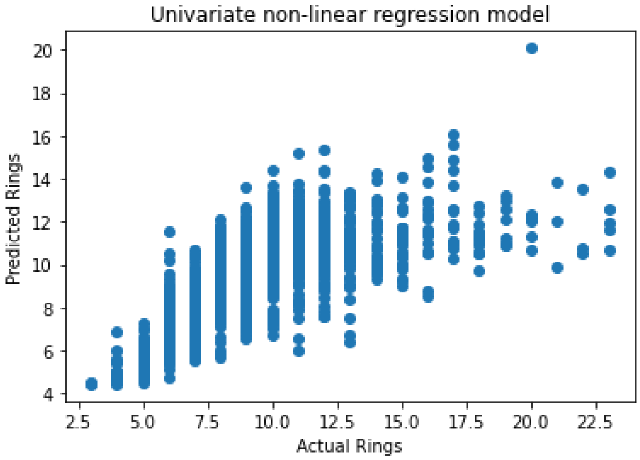

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted rings seem to be quite accurate.
2. The reason for Inference1 is low RMSE and high correlation between predicted and actual values.
3. Accuracy of prediction : univariate linear < univariate non-linear < multivariate linear regression model.
4. Inference 3 is made after looking at the RMSE values for all 3 models.

4

a.

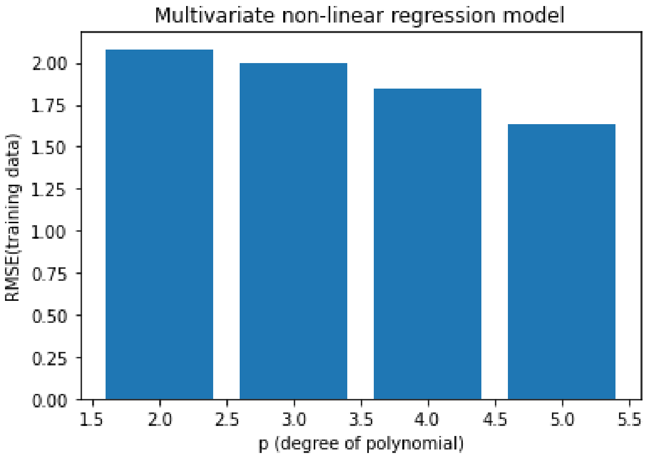

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. The decrease seems to be gradual for all p values.
3. The polynomial function of degree 5 will approximate the data best.

b.

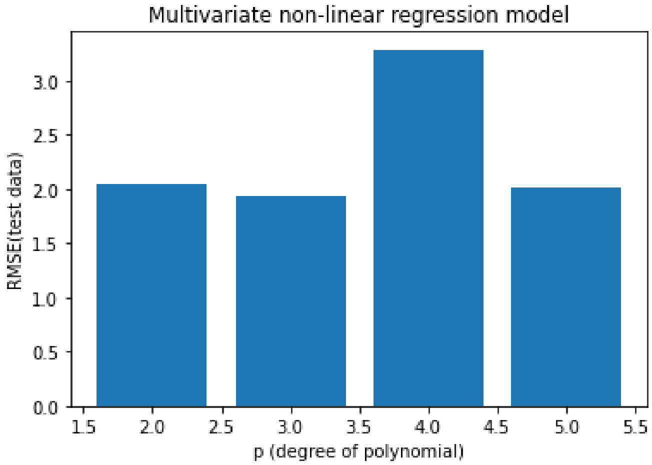

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases gradually first and then increases sharply for p = 4 and again shows sharp decrease for p = 5.
2. Non-uniform increase and decrease is observed as explained in inference 1.
3. From the RMSE value, 3 degree curve will approximate the data best.

c.

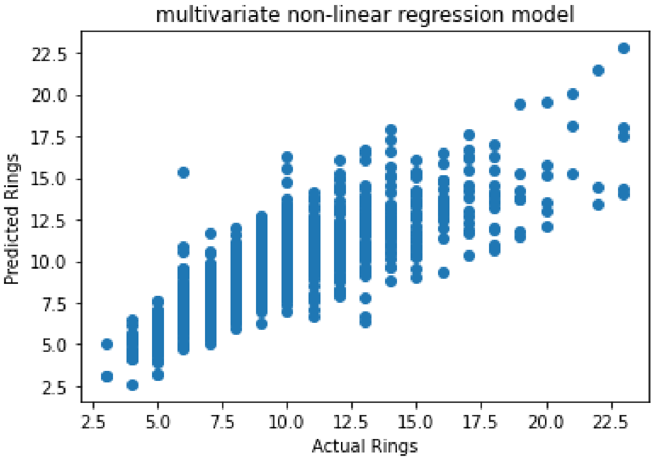

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, our model prediction seems to be quite good.
2. Due to strong positive correlation value between predicited and actual rings and low RMSE our prediction accuracy is high.
3. Univariate linear < univariate non-linear < multivariate linear < multivariate non-linear regression model is the order based upon the accuracy of predicted temperature value and spread of data points in Scatter Plot