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: Nikhil **Mobile No: 8949463760**

Roll Number: B20219 **Branch: CSE**

PART - A

1 a.

	Prediction Outcome	
True Label	106	12
	4	215

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	103	15
	5	214

Figure 3 Bayes GMM Confusion Matrix for Q = 8

	Prediction Outcome	
True Label	90	28
	1	218

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	**95.252**
4	**96.439**
8	**94.065**
16	**91.395**

Inferences:

1. The highest classification accuracy is obtained with Q = 4.
2. The classification accuracy increases with increase in Q at first but then starts decreasing.
3. This happens because adding nodes with less weights causes the model to overfit the training data hence the accuracy decreases.

4. The number of diagonal elements in the confusion matrix increase till Q = 4 but then start decreasing.
5. The accuracy is related to the number of diagonal elements in the confusion matrix which increases till Q = 4 but starts to decrease after Q = 4 that's why number of diagonal elements also increase/decrease with increase/decrease in accuracy.
6. The number of off-diagonal elements in the confusion matrix decrease till Q = 4 but then start increasing.
7. As the number of off-diagonal elements are total number of elements minus the number of diagonal elements so when the number of diagonal elements increase, the number of off diagonal elements decrease and vice-versa.

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	96.736

Inferences:

1. The K-NN model on normalized data shows the highest accuracy while K-NN model on actual data shows the lowest accuracy.

2. Accuracy of KNN on normalized data > Bayes using GMM > Bayes using unimodal Gaussian density > KNN model on actual data.

3. For KNN, since we are classifying based on Euclidean distance, greater accuracy will be there if distances with respect to all the attributes are considered equally significant. Equal significance can only be achieved if the spread of data in all attributes is same. Same spread can be achieved by scaling the spread in all the attributes to a common spread, and min-max normalization does this job. That's why we see a greater accuracy for normalized KNN than KNN on actual data.

4. The accuracy of Bayes Classifier is 94.362% which is less than the normalized K-NN model accuracy. This is because bayes classifier assumes the attributes to be independent of each other. But they might not be that independent, and this is indeed the case here as we are getting lower accuracy.

5. Accuracy of KNN model on actual data is less than Bayes Classifier because it finds Euclidian distance without normalizing. Greater accuracy will be there if distances with respect to all the attributes are considered equally significant. But this is not the case in the given dataset. Some attributes lie between 0 and 1 while others range to values in millions. This leads to huge fall in accuracy for KNN model on actual data

6. The accuracy of bayes classifier using GMM is higher than the bayes classifier using unimodal gaussian density because the unimodal bayes classifier assumes the distribution to be unimodal while bayes using GMM examines the accuracies considering every multimodal distribution and takes that number of modes in consideration for which accuracy is the highest. The bayes using GMM uses more information available to increase the accuracy. But the accuracy for GMM using bayes is still lower than KNN on normalized data because any bayes classifier assumes the distribution to be gaussian and, it might not be gaussian. KNN takes care of it since it does not make any assumption about the distribution.

PART – B

1

a.

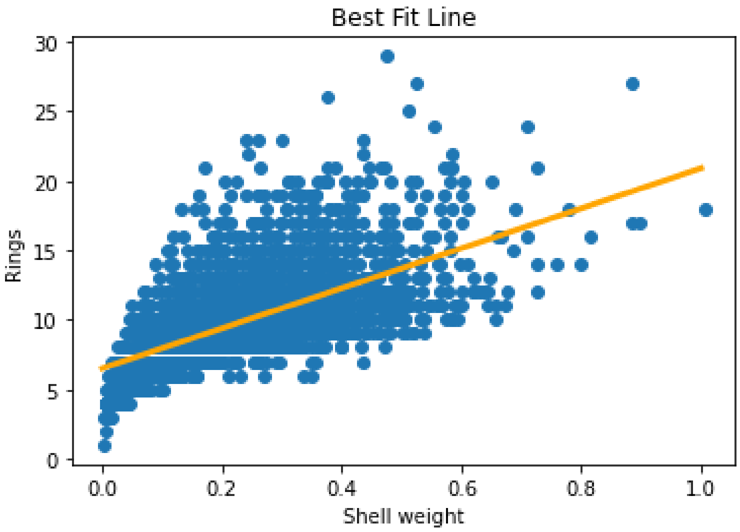

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. The target attribute is more likely to be more dependent on attribute having highest correlation coefficient with itself. That’s why the attribute with the highest correlation coefficient with the target attribute is used for prediction.
2. The best fit line does not fit the data perfectly, because it is oversimplified for the data, a more complex curve is needed to fit the data.
3. The bias is high, and variance is low.

b.

The RMSE for training data is 2.528

c.

The RMSE for testing data is 2.468

Inferences:

1. Accuracy is higher for training dataset.
2. This is because we have trained the model on the training data so it will give higher accuracy for the data samples resembling to the training data.

d.

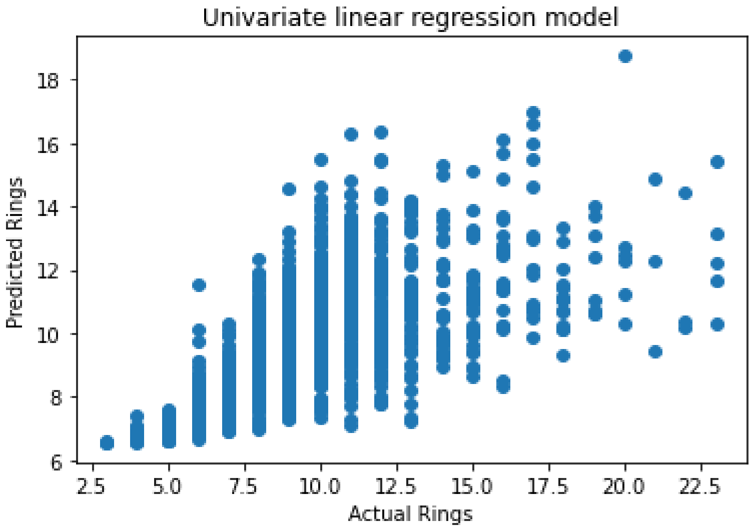

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of points, the accuracy is not so high.
2. This is because we did not use all the information available to us: we just used one attribute for prediction, not all the attributes.

2

a.

The RMSE for training data is 2.216

b.

The RMSE for testing data is 2.219

Inferences:

1. Amongst training and testing accuracy, testing accuracy is almost same as training accuracy.
2. This is because the testing dataset is almost of the same distribution as the training dataset in this case.

c.

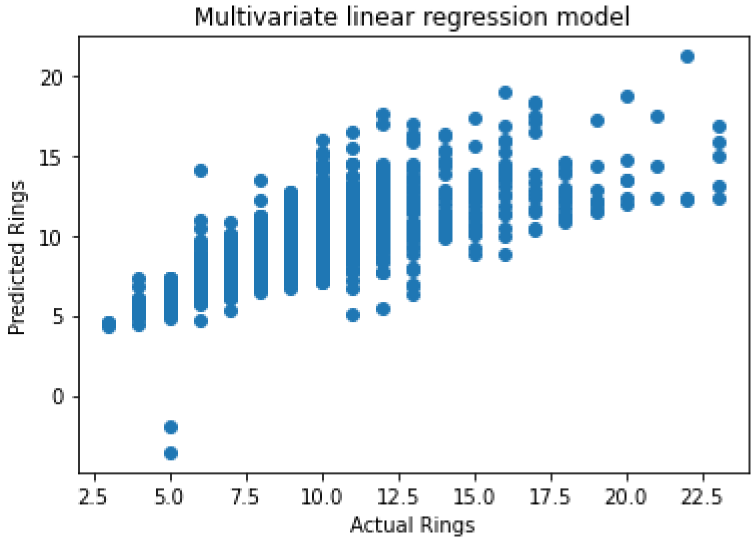

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based on the spread of points, the predicted number of rings is highly accurate.
2. The unimodal regression model gives less accuracy as compared to the multivariate regression model.
3. This is because the multivariate model uses all the information available to us in an efficient way: we used all the attributes for the prediction not just one single attribute.

3

a.

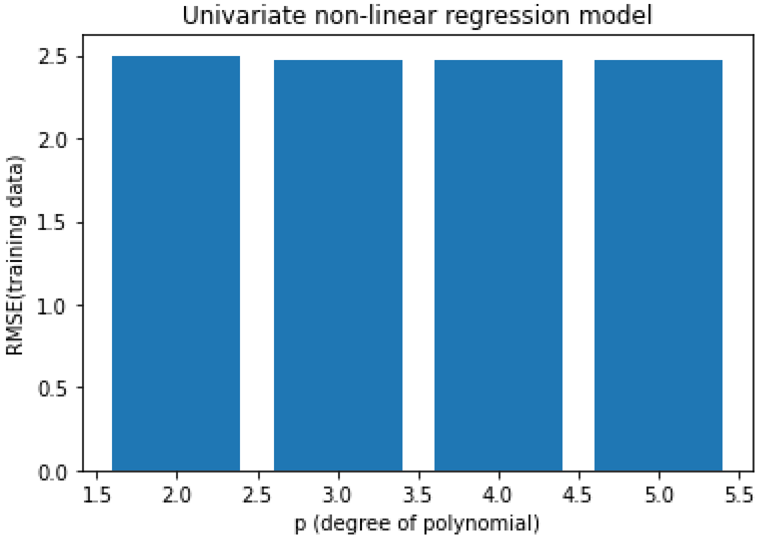

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE values slightly decrease with respect to the increase in the degree of the polynomial.
2. The decrease is more from 2 to 3 and the gradual, as compared to other transitions.
3. As the degree increases the curve fits the data better so RMSE decreases.
4. From the RMSE value, p=5 curve will approximate the data best. 5. As the degree increases, the bias decreases and variance increases.

b.

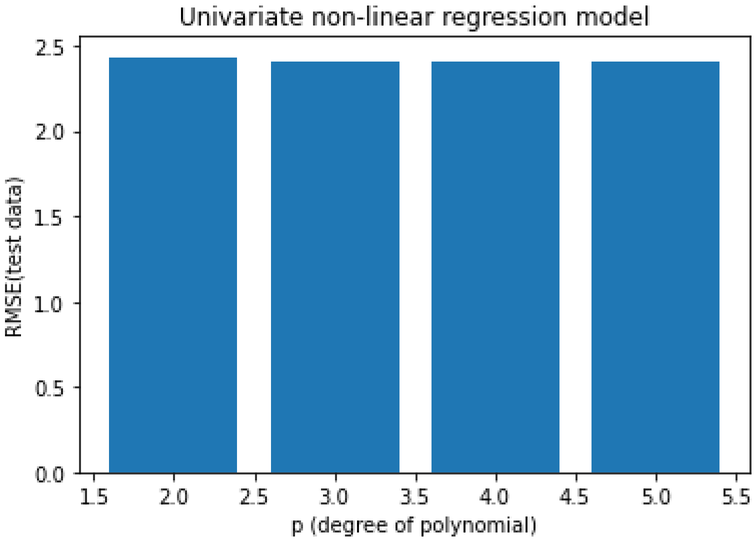

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE values slightly decrease with respect to the increase in the degree of the polynomial.
2. The decrease is more from 2 to 3 and the gradual, as compared to other transitions.
3. As the degree increases the curve fits the data better so RMSE decreases.
4. From the RMSE value, p=5 curve will approximate the data best. 5. As the degree increases, the bias decreases and variance increases.

c.

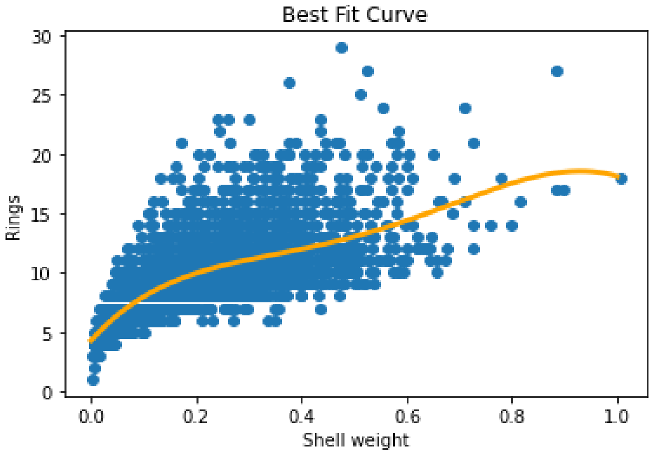

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. The p-value corresponding to the best fit model is 4.
2. This is because it neither overfits nor underfits the data so it the 'perfect' value.
3. The bias decreases and variance increases with increasing value of p.

d.

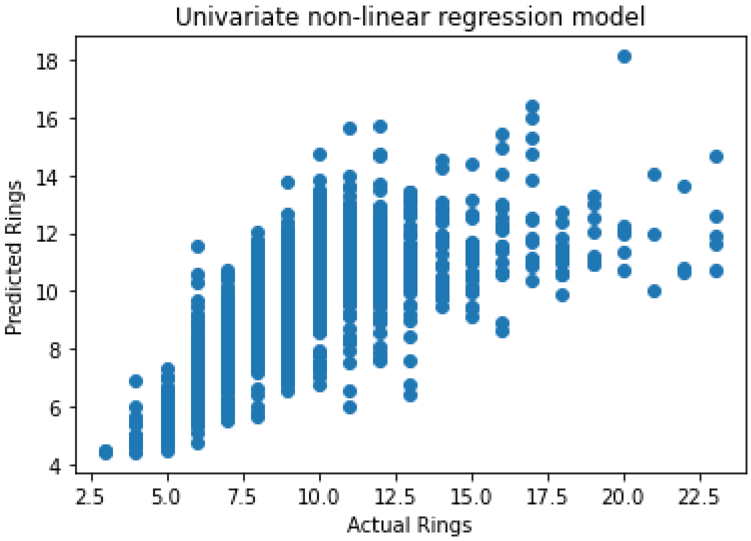

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted number of rings is quite accurate.
2. This is because we calculated the 'perfect' curve the fits the data, not just some assumed linear curve.
3. Both univariate non-linear model and multivariate linear model give high and almost same accuracy followed by univariate linear model which does not give so good accuracy
4. The non-linear univariate model uses a complex curve to best fit the data not just linear curve, and the linear multivariate model uses most of the information available to us for the prediction: it uses all the attributes for prediction. But since it is still linear, the accuracy is not higher than the non-linear univariate model. The univariate linear model uses little information for prediction: just one attribute

and uses only a linear curve to fit the data which not a very good approximation in most cases, and this is why it has low accuracy.

5. In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high.

4

a.

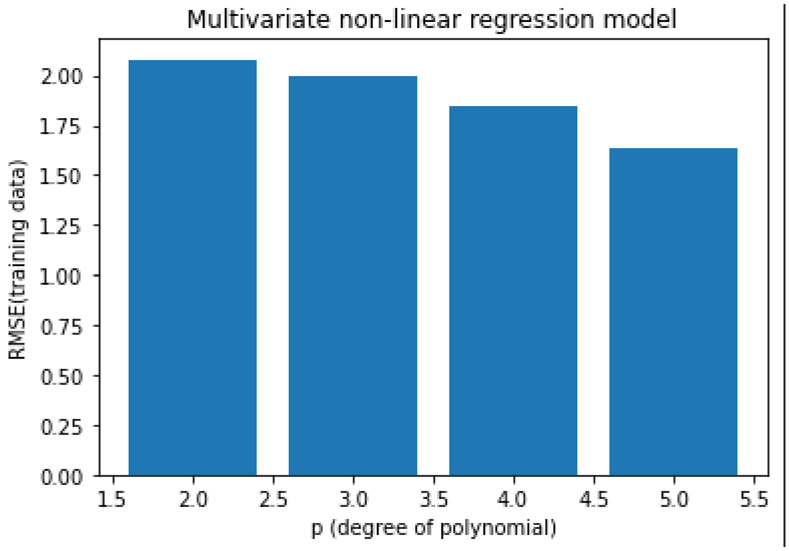

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE decreases with increase in degree of polynomials.
2. The decrease in RMSE is somewhat uniform but the rate of decrement gradually increases from degree 2 to degree 5.
3. As the degree increases the curve fits the data better hence increase in accuracy or decrease in RMSE.
4. From the RMSE values we see that since p=5 has so less RMSE value, it should be used for better accuracy.
5. The bias decreases and variance increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

b.

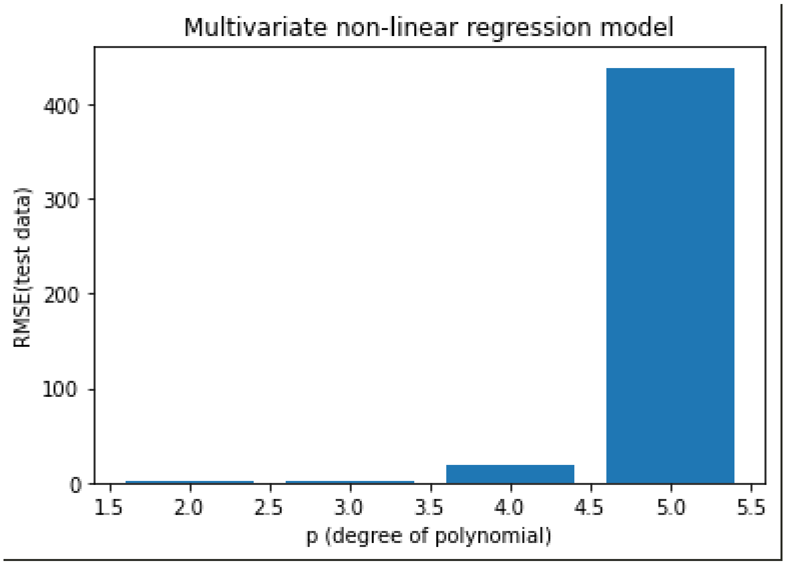

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. The RMSE values increase with increase in degree of the polynomial.
2. The increase in RMSE values is not so high till p = 3 but it is significant at p = 4 and it is extremely high for p = 5.
3. The increase is RMSE is a result of overfitting the data by the polynomial curves.
4. Since we see the RMSE for p = 2 to be least, it is the best degree of polynomial for predicting the data.
5. The bias gradually decreases till p=3 and then suddenly increases after p=3 and the variance increases as the model becomes more complex with increasing degree of polynomial.

c.

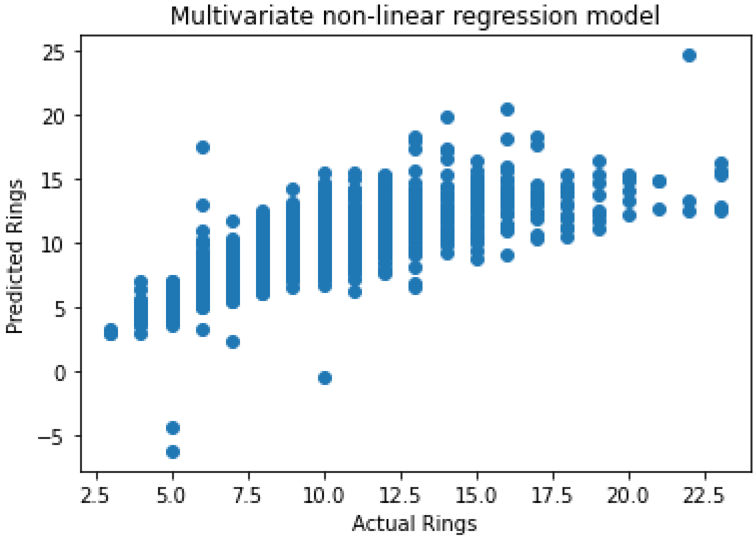

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted number of rings is quite accurate.
2. The high accuracy is because we used most information available to us and at the same time, also used the best fit curve for prediction.
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. The high accuracy for multivariate non-linear model is because we made very less number of assumptions while predicting: we did not assume that attribute of the data is independent of other attributes (unlike univariate regression models) nor did we assume that the linear fit curve is the best fit for the data (unlike linear regression models): we used most of the information available to us by using information in all the other attributes and at the same time, also found and used the best fit curve for predicting the data.
5. In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high**.**