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: Yash Sharma **Mobile No: 8802131138**

Roll Number: B20241 **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	105	13
	7	212

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	88	30
	7	212

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	**89.021**

Inferences:

1. The highest classification accuracy is obtained with Q =4.
2. Increasing the value of Q increases the prediction accuracy only up to some specific value of Q (like here at Q=4) a decrease in prediction accuracy is observed when we increase Q further.

3. Increasing the value of Q increases the prediction accuracy as Q denotes the number of clusters and when number of clusters is assumed low the accuracy is low.
4. As the classification accuracy increases with the increase in value of Q, the number of diagonal elements increase as it represents the number of correct predictions
5. As the classification accuracy increases with the increase in value of Q thus, it makes our model predict more test samples correctly with their respective positive or negative class. Thus, we see increase in diagonal elements.
6. As the classification accuracy increases with the increase in value of K, the number of off-diagonal elements decreases. A opposite situation is observed when classification accuracy decreases i.e., after Q=4.
7. As we see the diagonal elements increases with increase in Q (up to 4) and thereby classification accuracy increases. Thus, proportionally off-diagonal elements decrease (as total test samples are fixed) and now our model makes more accurate and correct predictions (diagonal elements) thereby decreasing off-diagonal elements.
8. After Q=4, we see when Q is further increased the off-diagonal elements increases because now diagonal elements decrease as the prediction accuracy is now decreasing.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.614
2.	KNN on normalized data	97.033
3.	Bayes using unimodal Gaussian density	94.362
4.	Bayes using GMM	96.439

Inferences:

1. Classifiers with highest: KNN on normalized data and Lowest accuracy: KNN.
2. Ascending order of classification accuracy: KNN < Bayes using GMM < Bayes using unimodal Gaussian density < KNN on normalized data.
3. The reason behind the fact that K-NN performed on normalized data has higher accuracy than K-NN performed on normal data is as follow. As we know KNN method is based on finding the Euclidean distance between a given test tuple with all other input tuples. Euclidean distance calculated on non-normalized data does not give real distance between those two tuples as any attribute in data set with big range can caused a biased result while calculating the Euclidean distance, Thus, normalized data gives the accurate Euclidean distance thereby increasing the accuracy as seen.

4. The reason for the Bayes classifier having low accuracy than K-NN (on normalized data) is as follow. K-NN does better because of its inherent nature to optimize locally and make predictions based on maximum class of the k-nearest neighbors around it. But still we see Bayes Classifier still isn't very far from the accuracy 97.003 % thus it may not be regarded as a bad approach just based upon this data prediction results
5. Generally, Bayes Classifier is more accurate and beneficial on complex data sets as their K-NN method would fail due to its high time complexity and its inherent nature to optimize locally. Here, data set is not that much big and complex thus K-NN shows best accuracy among all other classifiers.
6. We may observe that GMM model also has improved prediction accuracy than the Bayes using unimodal Gaussian density possibly since the Gaussian Mixture Model approximates the probability density with a sum of Gaussians. Each gaussian is, in general, characterized by a full covariance matrix. Moreover, it does not assume the data to have only one cluster. But it is generally less accurate than KNN (on normalized data).

PART – B

1

a.

Figure 5 Univariate linear regression model: Rings vs. 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 to predict accurate values we need input which can represent the given data in best possible way or in other way shows a good correlation with change in output variable. This, way we choose attribute that has the highest influence on the given to be predicted attribute.
2. Approximately (excluding some data points) we may say that the fit line fit the training data seemingly good. But for some data points in train set, the line is not able to fit them.
3. The reason that the line is not able to **perfectly fit** the training data may be because the inputs and outputs are not linearly dependent that we are assuming while fitting this line. (They may have some other relation i.e., polynomial etc.)
4. Bias is low as the best fit line underfits the data, the model requires more complex function to fit the training data. Variance is low as the bias is high due to underfitting of data.

b.

The prediction accuracy on training data is 2.528 (RMSE Error).

c.

The prediction accuracy on test data is 2.468 (RMSE Error).

Inferences:

1. Amongst training and testing accuracy, the accuracy on test data is higher as it has low RMSE.
2. We observe that there is not so much accuracy difference between test data and train data. This may be since there may be no meaningful differences between the kind of data, we trained the model on and the testing data we are providing for evaluation.

d.

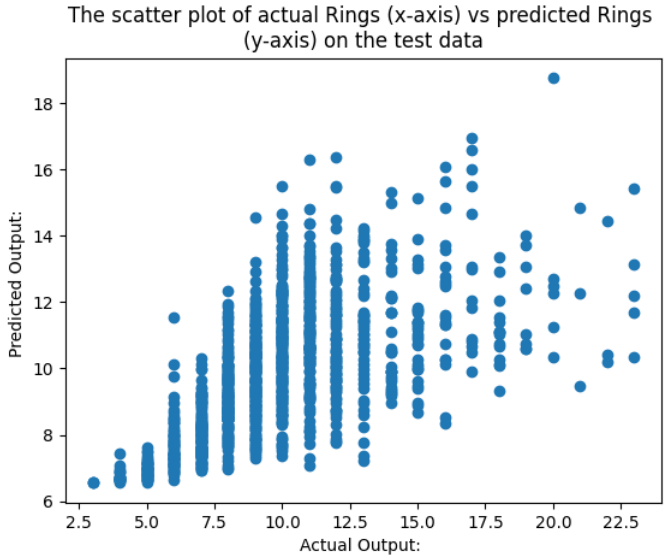

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The prediction accuracy of Rings is seemingly good.
2. The actual Rings is spread from 2.5 to 22.5 while the predicted Rings is more concentrated from 6 to 18 which shows that the prediction accuracy is seemingly good and accurate many times.
3. A good correlation value from scatter plot may be inferred which shows that our model is seemingly good at predicting correct number of Rings.

2

a.

The prediction accuracy on training data is 2.216 (RMSE Error).

b.

The prediction accuracy on training data is 2.219 (RMSE Error).

Inferences:

1. Amongst training and testing accuracy, the accuracy on training data is higher as it has low RMSE.
2. Training accuracy is higher because its RMSE is lower than testing accuracy, this is because the model is made on the training data so it will have less RMSE on training data.

3. We may also observe that there is not so much accuracy difference between test data and train data. This may be since there may be no meaningful differences between the kind of data, we trained the model on and the testing data we are providing for evaluation.

c.

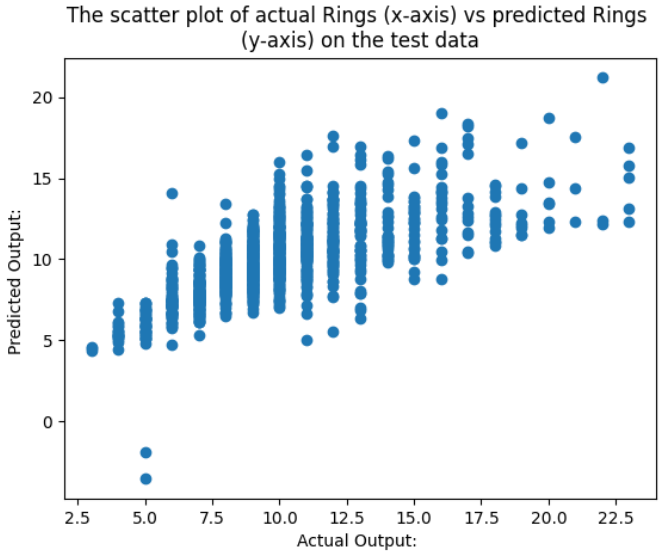

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The prediction accuracy of Rings is seemingly good.
2. The actual Rings is spread from 2.5 to 22.5 while the predicted Rings is more concentrated from 0 to 22 which shows that the prediction accuracy is seemingly good and accurate many times.
3. A good correlation value from scatter plot may be inferred which shows that our model is seemingly good at predicting correct number of Rings.
4. However, it is also visible that some negative values for Rings are also predicted that are not possible, thus our model is preferably not the best model to fit the given data but still it has some acceptable accuracy on test data also which can't be neglected.

3

a.

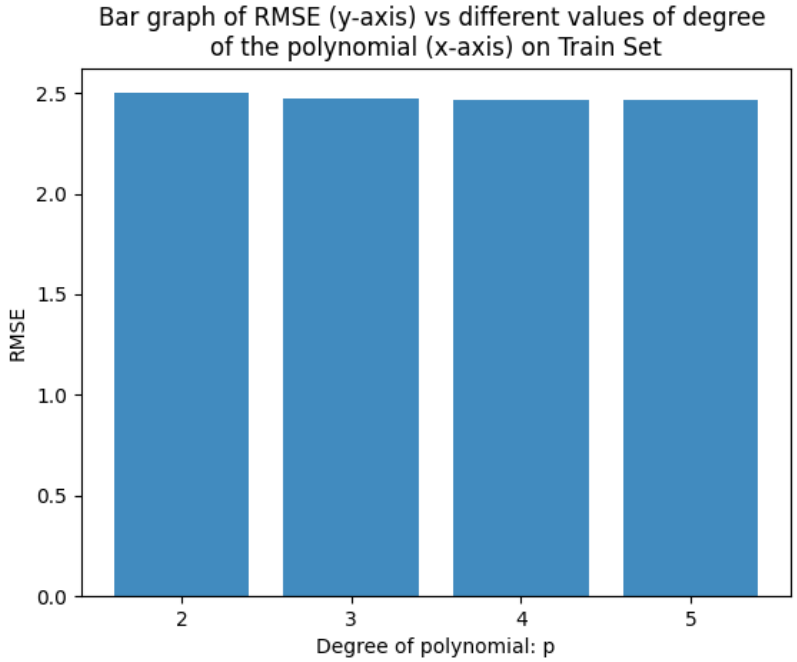

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. The RMSE decreases from p=2 to p=3 more compared to rest. From p=3 it decreases slightly or almost remains constant.
3. As the degree increases the curve fits the data better, so the RMSE decreases.
4. From the RMSE value, degree p= curve will approximate the data best.
5. As the degree increases the bias decreases and variance increases as the model starts becoming more complex and starts fitting the data better.

b.

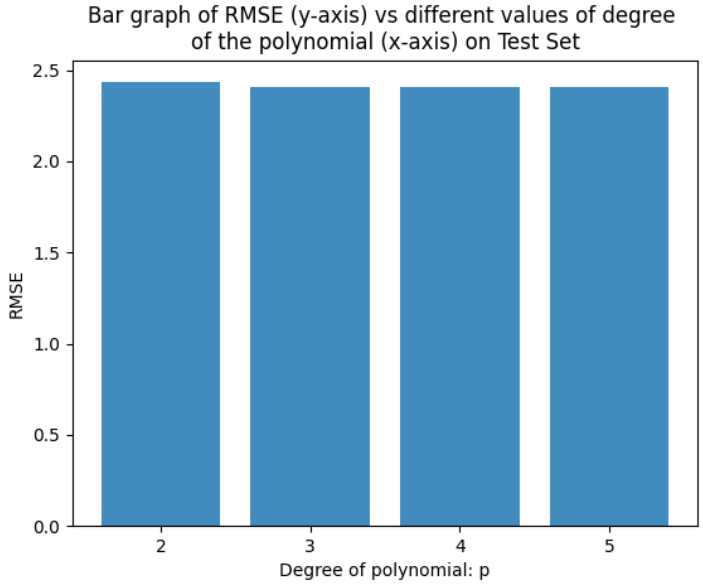

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. The RMSE decreases from p=2 to p=3 (though very slight) then it almost remains constant or decreases.
3. The RMSE decreases from p=2 to p=3 more compared to rest. From p=3 it decreases slightly or almost remains constant.
4. From the RMSE value, degree p=5 curve will approximate the data best.
5. As the degree increases the bias decreases and variance increases as the model starts becoming more complex and starts fitting the data better.

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 is 5 corresponding to best fit model.
2. p=5 is best fit model because it fits the data better as it is more complex and have higher variance.
3. As the degree increases the bias decreases and variance increases as the model starts becoming more complex and starts fitting the data better.

d.

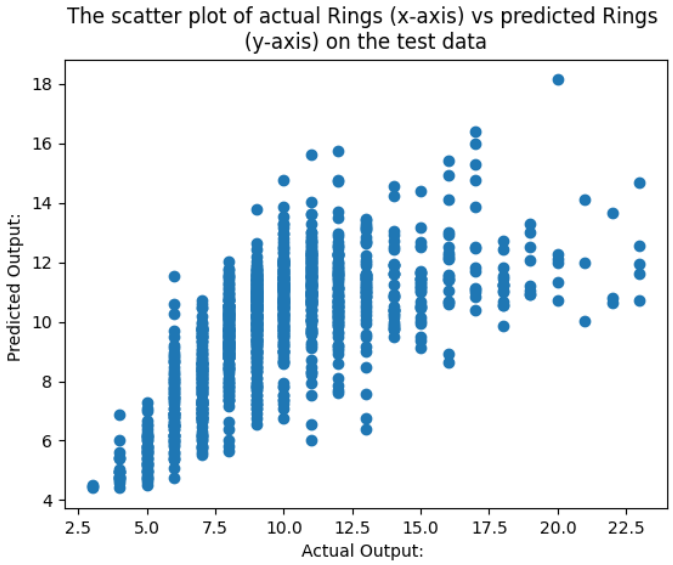

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. From the spread of points, we can see that accuracy of predicted Rings is quite good.
2. The actual Rings is spread between 2.5 and 22.5, similarly the predicted Rings is also spread between 4 to 19, thus we can say that the accuracy is seemingly good.
3. Prediction accuracy of non-linear univariate regression (2.408) is nearly same as that of univariate linear regression (2.468) as both have nearly same RMSE.
4. RMSE of multivariate nonlinear regression (2.185) lower than multivariate linear regression (2.185) and the spread of predicted value matches actual value better in nonlinear regression than linear, so we can say that nonlinear regression is better.
5. In linear regression bias is high and variance is low but in nonlinear regression variance is high and bias is low.

4

a.

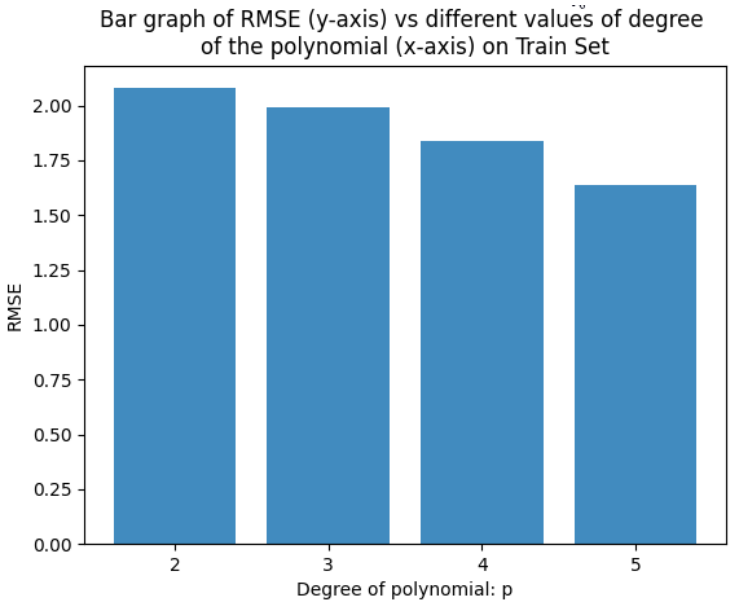

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. The RMSE decreases gradually from p=2 to p=5.
3. The RMSE decreases from p=4 to p=5 more compared to rest.
4. From the RMSE value, degree p=5 curve will approximate the data best.
5. As the degree increases the bias decreases and variance increases as the model starts becoming more complex and starts fitting the data better.

b.

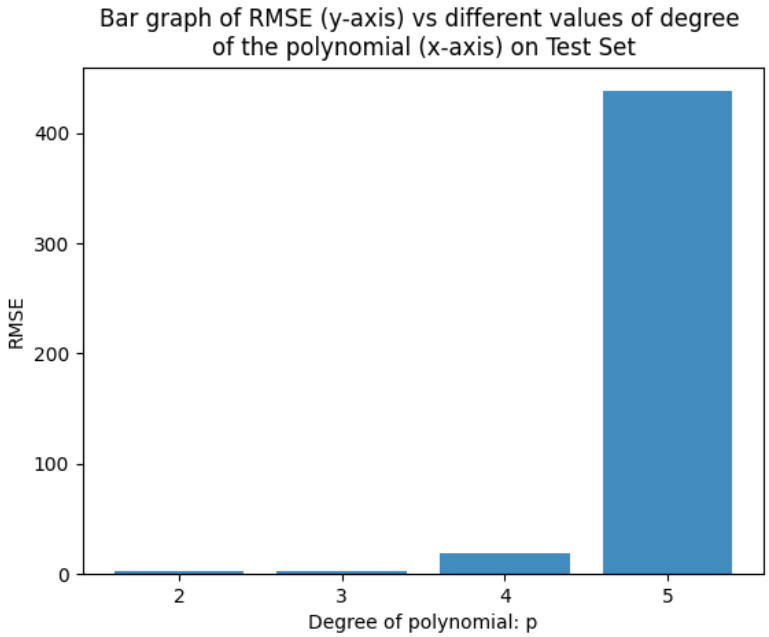

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value increases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. The RMSE value remains constant from p=2 to p=3 a sharp increase is observed from p=4 to p=5.
3. The RMSE increases from p=4 to p=5 more compared to rest. Before p=4 it increases slightly or almost remains constant.
4. From the RMSE value, degree p=2 curve will approximate the data best.
5. As the degree increases the bias decreases and variance increases as the model starts becoming more complex and starts fitting the data better.

c.

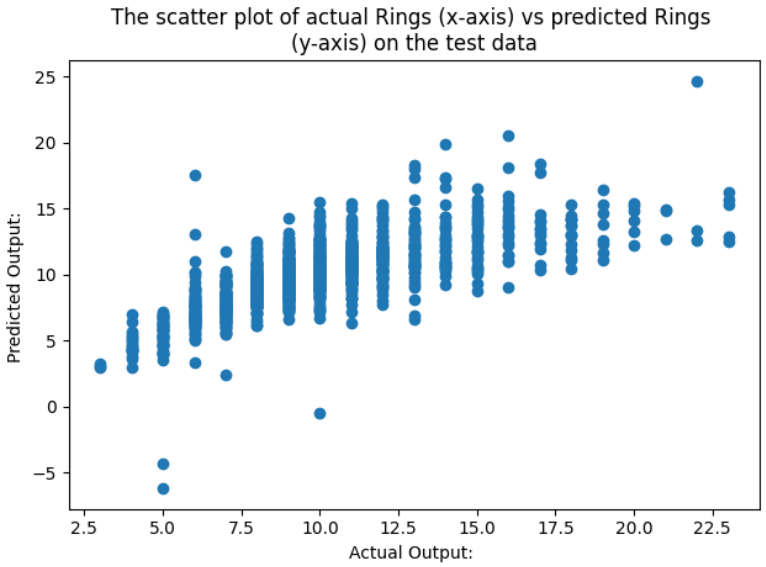

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. From the spread of points, we can see that accuracy of predicted Rings is quite good.
2. The actual Rings is spread between 2.5 and 22.5, similarly the predicted Rings is also spread between 3 to 20 (**except some outliers**), thus we can say that the accuracy is seemingly good.
3. Prediction accuracy of non-linear univariate regression (2.408) is nearly same as that of univariate linear regression (2.468) as both have nearly same RMSE.
4. RMSE of multivariate nonlinear regression (2.185) lower than multivariate linear regression (2.185) and the spread of predicted value matches actual value better in nonlinear regression than linear, so we can say that nonlinear regression is better.
5. In linear regression bias is high and variance is low but in nonlinear regression variance is high and bias is low.