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: Vikas Dangi **Mobile No: 9406661661**

Roll Number: B20238 **Branch:EE**

PART - A

1 a.

	Prediction Outcome	
True Label	117	1
	28	191

Figure 1 Bayes GMM Confusion Matrix for Q = 2

	Prediction Outcome	
True Label	117	1
	23	196

Figure 2 Bayes GMM Confusion Matrix for Q = 4

	Prediction Outcome	
True Label	112	6
	20	199

Figure 3 Bayes GMM Confusion Matrix for Q = 8

	Prediction Outcome	
True Label	97	21
	8	211

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	**91.4**
4	**92.9**
8	**92.3**
16	**91.4**

Inferences:

1. The highest classification accuracy is obtained with Q =4.
2. When Q is increased, accuracy improves at initially, but after Q=4, accuracy drops.
3. It may be due to the fact that our data has around 4 clusters which are more prominent and thus provide us better estimation when we use around 4 clusters in multimodal.

4. The number of diagonal elements in the confusion matrix increase and decrease with the increase and decrease of the accuracy respectively.
5. The diagonal elements are a measure of the elements that were accurately predicted, hence better accuracy equals higher value.
6. The number of off-diagonal elements in the confusion matrix increase and decrease with the decrease and increase of the accuracy respectively
7. The off-diagonal elements are a measure of the items that were mistakenly predicted, hence more accuracy equals lesser value.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	0.896
2.	KNN on normalized data	0.970
3.	Bayes using unimodal Gaussian density	0.946
4.	Bayes using GMM	0.929

Inferences:

1. KNN on the normalized data has the highest accuracy and KNN classifier has the lowest accuracy.
2. The classifiers in ascending order of classification accuracy are KNN < Bayes using GMM < Bayes using unimodal Gaussian Density < KNN on normalized data.
3. The reasons for KNN being lowest is that it is not normalized so the attributes corresponding to the higher value gets a huge edge over the others. Also, in the other classifiers it is mostly depending on the type of data and its distribution. If our data has only one cluster and we are forcing it in multiple clusters then it will definitely affect the accuracy. Also whether data is in normal distribution or not also plays a major role.

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

PART – B

1

a.

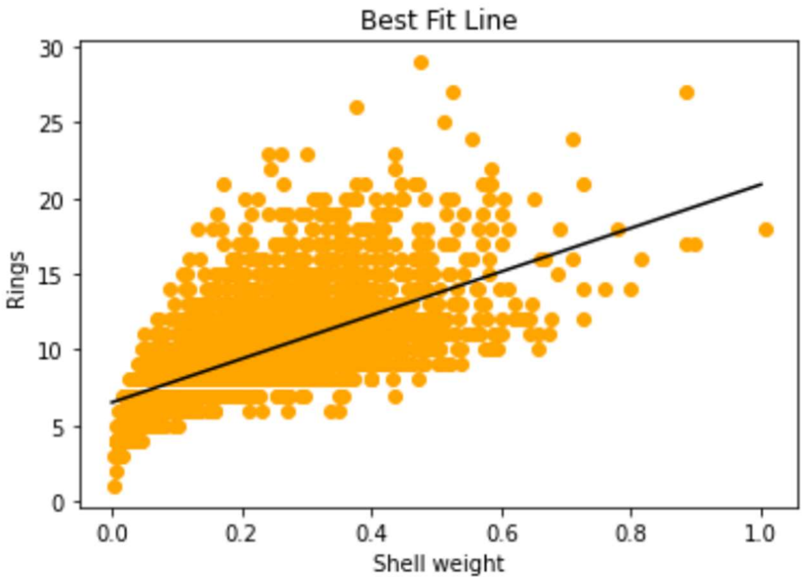

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. Because the attribute with the highest correlation coefficient will have the most impact on the target attribute, choosing it will result in superior prediction results.
2. NO, it doesn’t.
3. Because only one attribute was used as an input variable for the regression model. Infer upon bias and variance trade-off for the best fit line.

b.

The prediction accuracy on the training data is 2.528

c.

The prediction accuracy on the test data is 2.468

Inferences:

1. The accuracy of the test data is higher due to low RMSE.
2. This is the polar opposite of what one would predict, and upon more investigation, the cause appeared to be the test-train data splitting ratio.

d.

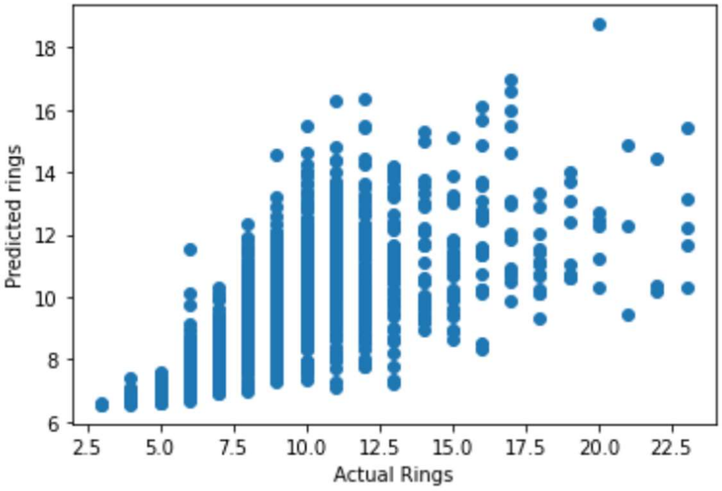

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The actual rings should have been discrete integers, but the predicted rings are spread continuously over the y-axis.
2. We have used only one attribute so the modelling is not perfect.

2

a.

The prediction accuracy on the training data is 2.216

b.

The prediction accuracy on the test data is 2.219

Inferences:

3. The training data has lower RMSE, hence higher accuracy.
4. We have built the model onto the training data it has already seen the data which it is testing hence higher accuracy.

c.

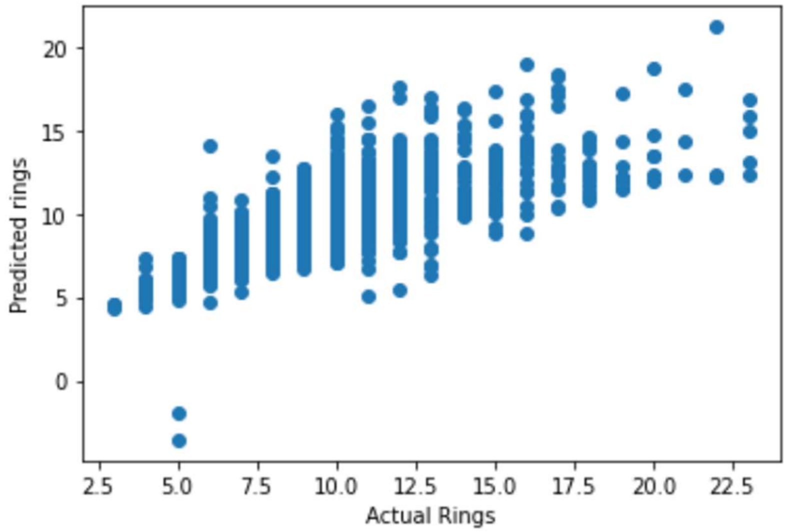

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The actual rings should have been discrete integers, but the predicted rings are spread continuously over the y-axis
2. Because it predicts the result using several inputs (all qualities) rather than a single input, multivariate linear regression performs better.

3

a.

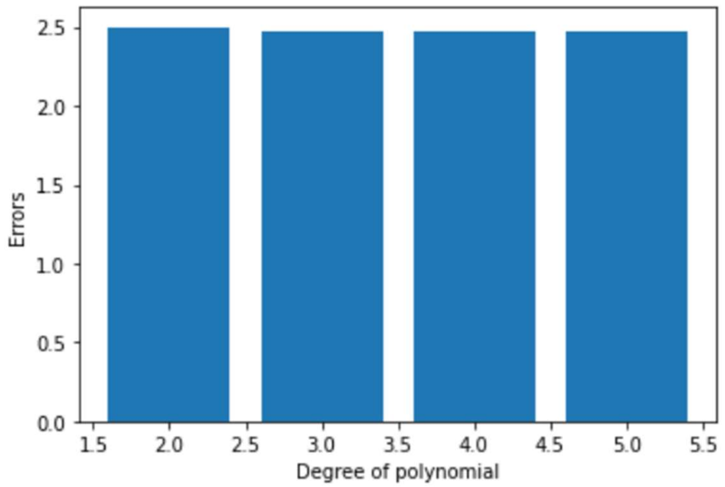

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. The RMSE value drops as the degree of the polynomial increases.
2. Is the decrease being uniform and gradual after a certain p-value.
3. As we increase p from 1, the accuracy improves initially as the model fitting more exactly captures the training data, but at a certain p, the model begins to overfit the data, resulting in better training accuracy but poorer testing accuracy.
4. The data will be best approximated by the 5th degree curve, which has the lowest test RMSE and thus the highest test accuracy.
5. As the model complexity grows, the bias diminishes, resulting in improved training data accuracy.

b.

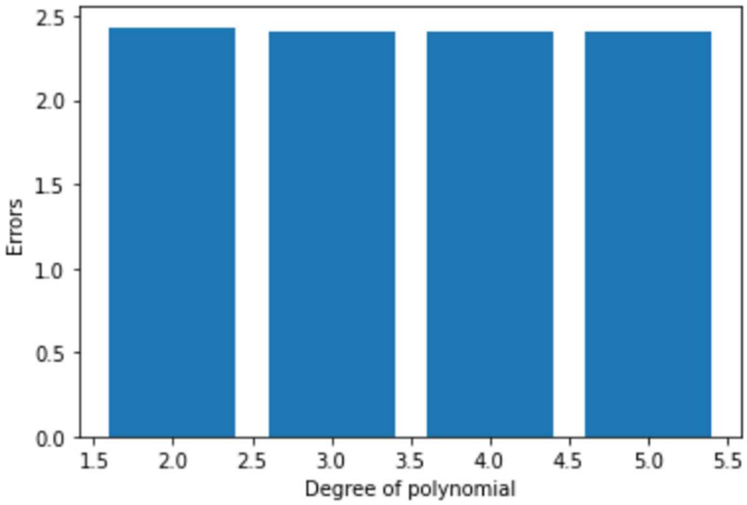

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. The RMSE value drops as the degree of the polynomial increases.
2. Is the decrease being uniform and gradual after a certain p-value.
3. As we increase p from 1, the accuracy improves initially as the model fitting more exactly captures the training data, but at a certain p, the model begins to overfit the data, resulting in better training accuracy but poorer testing accuracy.
4. The data will be best approximated by the 4th degree curve, which has the lowest test RMSE and thus the highest test accuracy.

c.

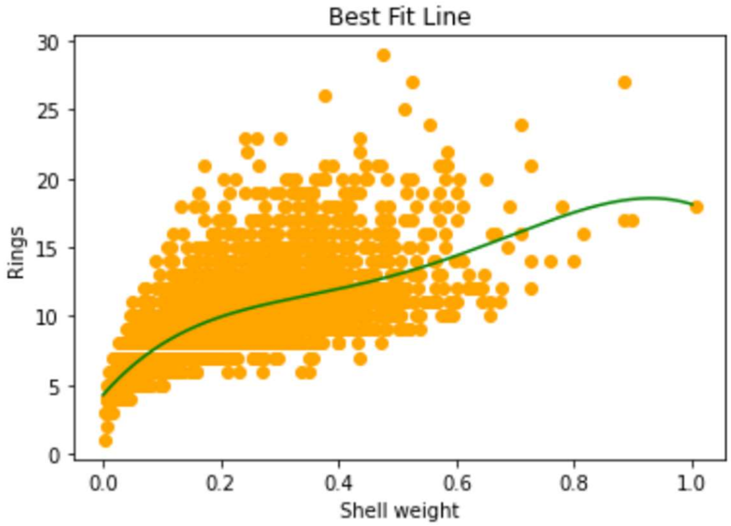

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. The p-value corresponding to the best fit model is 5.
2. Because a higher p value suggests that a higher order polynomial is utilized for regression, the training data is better fitted.
3. As the model complexity grows, the bias diminishes, resulting in improved training data accuracy.

d.

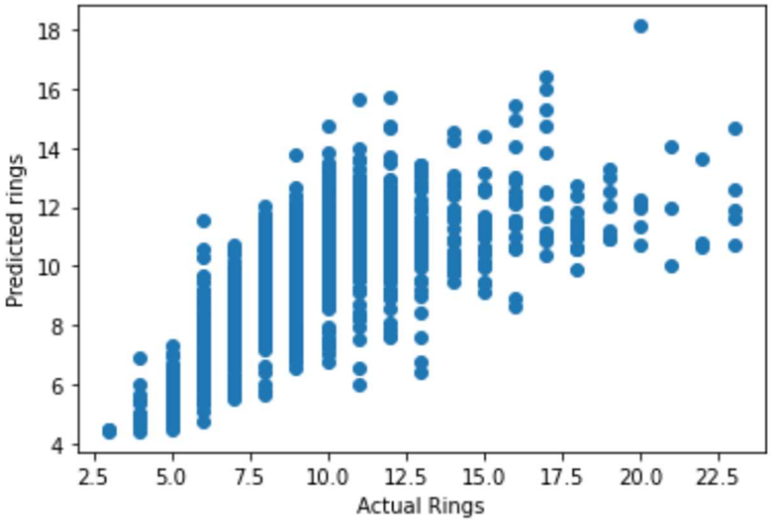

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. The actual rings should have been discrete integers, but the predicted rings are spread continuously over the y-axis
2. Although no model is perfect it still seems better than the other two models.
3. As the multivariate model takes in account of all the qualities it is better than the unimodal and the polynomial modelling takes care of the non-linearity (if present) in the data it is better in most of the cases than single variate ones.
4. As the model complexity grows, the bias diminishes, resulting in improved training data accuracy.

4

a.

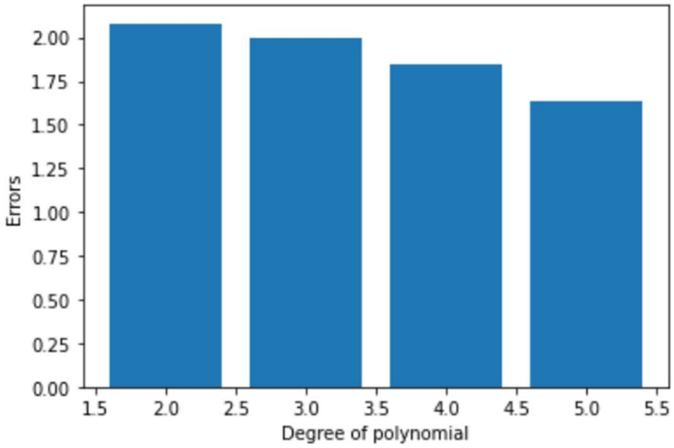

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. The decrease is uniform and slow.
3. As we increase p from 1, the accuracy improves initially as the model fitting more exactly captures the training data, but at a certain p, the model begins to overfit the data, resulting in better training accuracy but poorer testing accuracy.
4. The data will be best approximated by the 5th degree curve, which has the lowest test RMSE and thus the highest test accuracy.
5. As the model complexity grows, the bias diminishes, resulting in improved training data accuracy.

b.

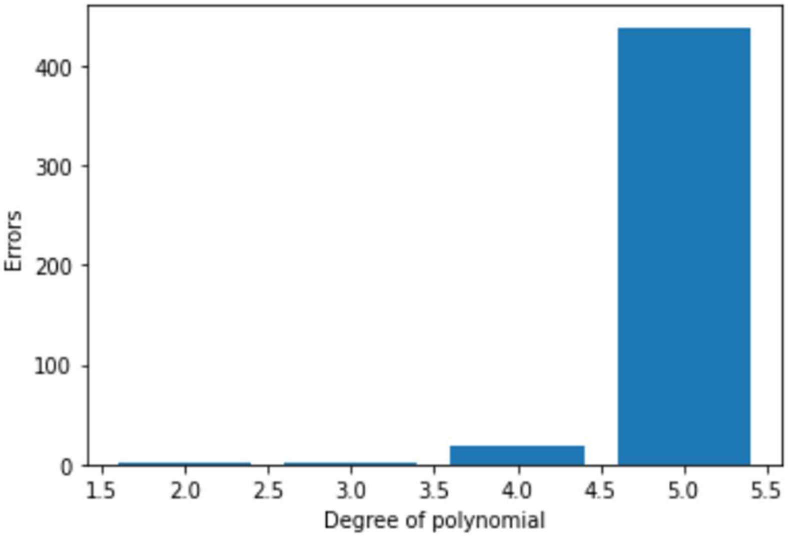

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. The RMSE value increases with increase in p-value
2. The increase is very steep at the end.
3. As we increase p from 1, the accuracy improves initially as the model fitting more exactly captures the training data, but at a certain p, the model begins to overfit the data, resulting in better training accuracy but poorer testing accuracy. Here it is the case of outfitting due to the increase in number of dimensions.
4. The curve with 2nd degree is the most suitable.
5. As the model complexity grows, the bias diminishes, resulting in improved training data accuracy.

c.

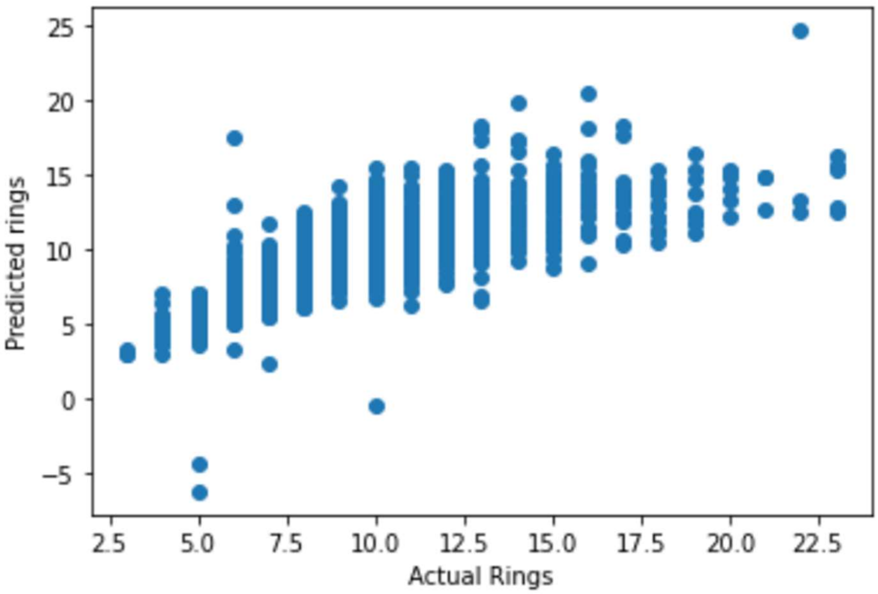

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. As there are continuous lines it is not totally accurate but it is spread in a small area which means it is still better than the most of the others.
2. The multivariate polynomial regression predicts the output variable using all the input attributes, and using non-linear regression which performs better than a simple linear regression.
3. Multivariate polynomial regression surpasses all three, with multivariate linear regression coming in second and univariate regressions following closely behind. It has the least spread.

Thank You