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: Aryan Tiwari **Mobile No: 8982562898**

Roll Number: B20187 **Branch:** Electrical engineering

PART - A

1 a.

		Prediction Outcome
True Label	95	13
	2	226

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	90	18
	2	226

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	83	25
	3	225

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	**95.5**
4	**94.9**
8	**94.0**
16	**91.7**

Inferences:

1. The highest classification accuracy is obtained with Q =2.
2. Increasing Q decreases accuracy.
3. The data might not have more than 2 modes, hence the data might be bi-modal.
4. Diagonal elements decrease with decreasing accuracy.

5. They decrease as lesser true values are predicted.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	89.58
2.	KNN on normalized data	96.72
3.	Bayes using unimodal Gaussian density	95.83
4.	Bayes using GMM	95.5

Inferences:

1. Highest accuracy : KNN on normalized data
   Lowest accuracy : KNN
2. knn < Bayes using gmm < Cl Bayes using unimodal Gaussian density < KNN on normalised.

PART – B

1

a.

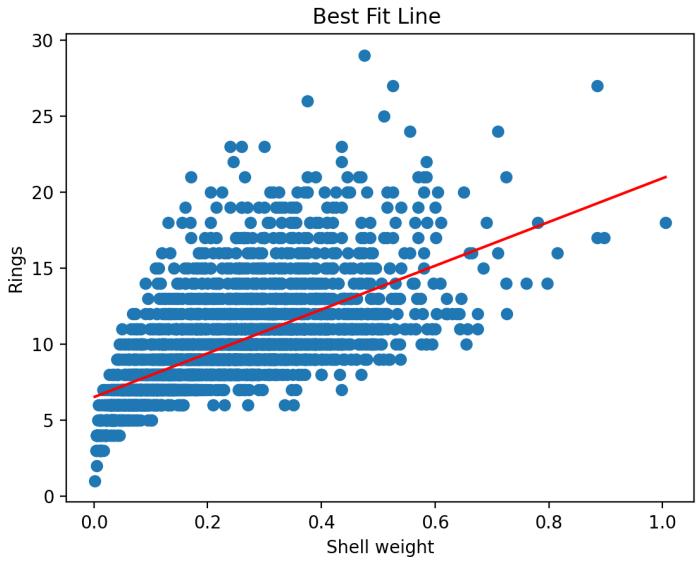

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. The highest correlation coeff attribute chosen so that the prediction is better.
2. The best line doesn't fit the data perfectly.
3. The line doesn't fit perfectly because its corr coeff is not 1.

b. Rmse on training data is 2.52

c. Rmse on testing data is 2.46

Inferences:

1. Prediction on training data is more accurate.
2. It is so because training data was used to create the regression model.

d.

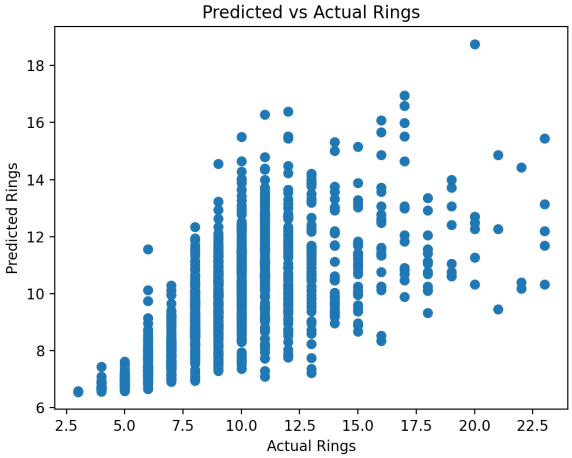

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The spread along a straight line with slope 1 is very much hence the accuracy is quite low.
2. The training attribute was not perfectly correlated to label attribute.

2

a.

Rmse on training data is 2.216

b.

Rmse on testing data is 2.219

Inferences:

3. Accuracy for testing is higher

c.

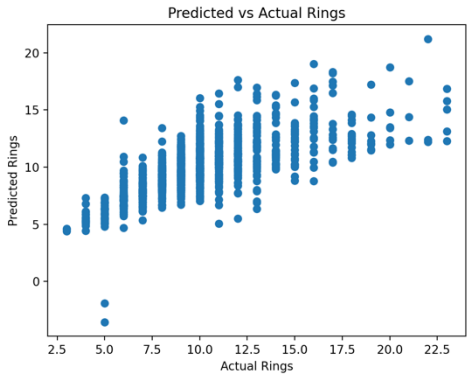

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. The predicted rings vary a lot from actual rings
2. The correlation coefficient is not perfect 1 for any attribute.
3. Multivariate training uses more data and hence predicts better.

3

a.

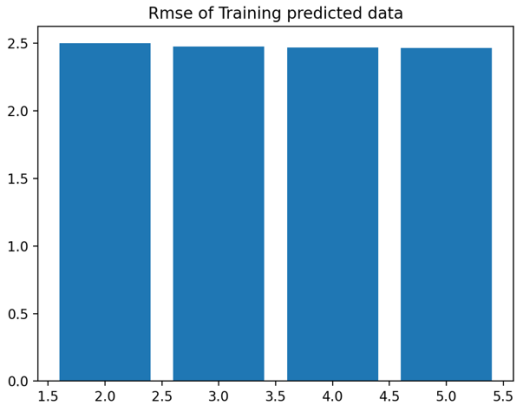

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases slightly with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. From the RMSE value, the 5-degree curve will approximate the data best.

b.

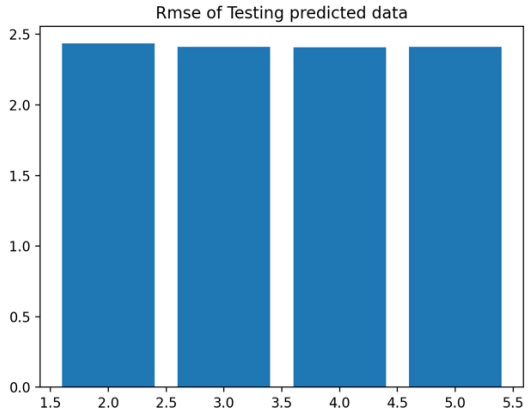

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases slightly and increase at the end with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. From the RMSE value, the 4-degree curve will approximate the data best.

c.

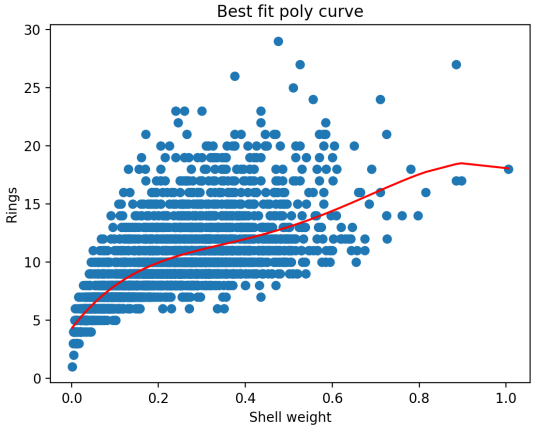

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. The p-value corresponding to the best fit model is 4.
2. Because its RMSE is least

d.

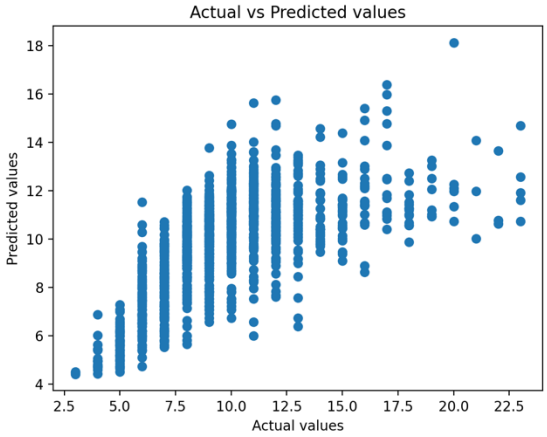

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. The predicted rings vary a lot from actual rings
2. The correlation coefficient is not perfect 1 for any attribute.
3. The linear multivariate classifier is best in this case

4

a.

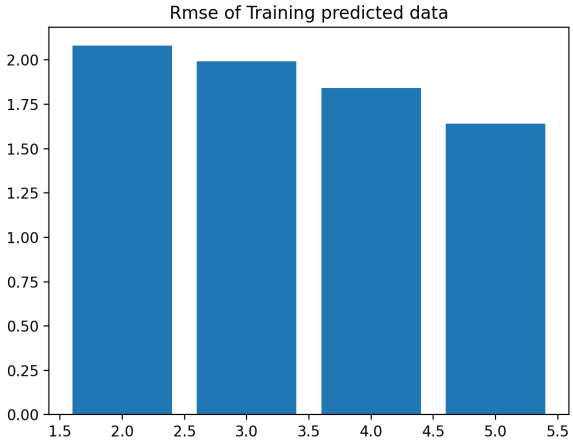

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. The decrease becomes more gradual with increasing p-values.
3. The curve with p=5 will best fit data.

b.

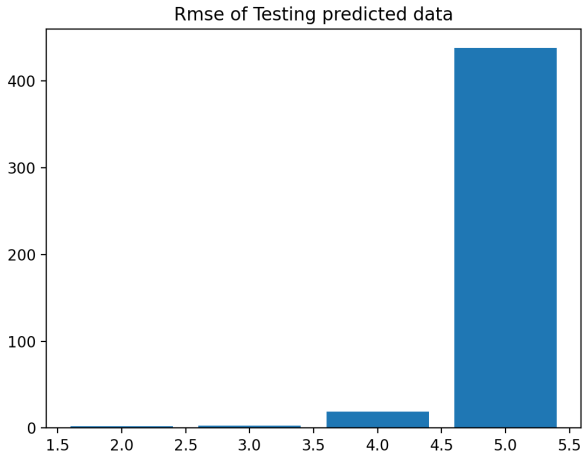

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. The increase in rmse is very upright.
3. The P=2 curve will best fit test data.

c.

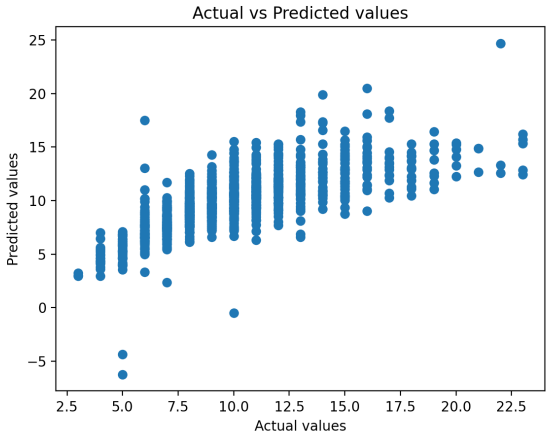

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. The prediction is quite inaccurate.