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: Abhay Vijayvargiya **Mobile No: 6377967485**

Roll Number: B20176 **Branch:DSE**

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
	4	215

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	97	21
	5	214

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	84	34
	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	**0.95**
4	**0.97**
8	**0.92**
16	**0.90**

Inferences:

1. The highest classification accuracy is obtained with Q = 4.
2. Classification accuracy first increases and then decreases.

3. Higher value of Q will increase the accuracy upto a certain value as there will be more no. of clusters per class so but after increasing Q after a certain point will decrease accuracy as there will be overfitting of data.
4. Diagonal elements represent the data samples correctly predicted hence these will increase with increase in accuracy.
5. No. of Off-Diagonal elements will decrease with increase in classification accuracy.
6. It happens because the total no. of off-diagonal elements represents the data samples incorrectly predicted.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	0.896
2.	KNN on normalized data	0.973
3.	Bayes using unimodal Gaussian density	0.943
4.	Bayes using GMM	0.967

Inferences:

1. Highest Accuracy = KNN on normalized data, Lowest Accuracy = KNN.
2. KNN on normalized data > Bayes using GMM > Bayes using unimodal density > KNN.
3. Bayes classifier using GMM will give more better results than Bayes classifier using unimodal density because there it does not assumes the data to be unimodal and forms no. of clusters for a class by observing class distribution. Moreover, KNN on normalized data will perform better than all other because it evaluates the Euclidian distance with every data point and compares with all the datapoints and assigns a class using a closest distance members.

PART – B

1

a.

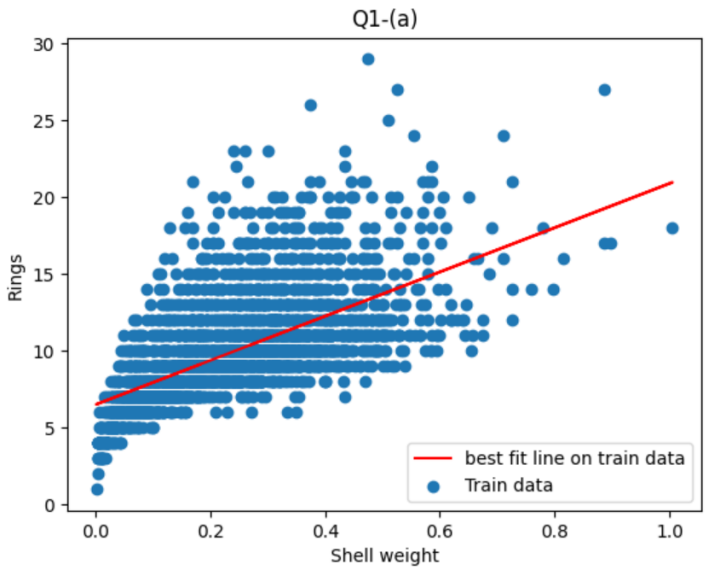

Figure 5 Univariate linear regression model: Rings vs. the chosen attribute name (replace) best fit line on the training data

Inferences:

1. Attribute - Shell weight is highly related with Rings. Highly related attributes will give more accurate linear regression line.
2. No, the regression line does not fit the data properly.
3. Because there are various attributes on which target attributes depends but we take only that attribute which have a comparatively high correlation coefficient.
4. The bias is high and the variance is low for the best fit line.

b.

The prediction accuracy on training data using RMSE error is 2.528.

c.

The prediction accuracy on testing data using RMSE is 9.803.

Inferences:

1. Among test and train data accuracy for train data is higher (lesser RMSE).
2. Since we are predicting the exact data that we used to train the model hence accuracy is higher for train data.

d.

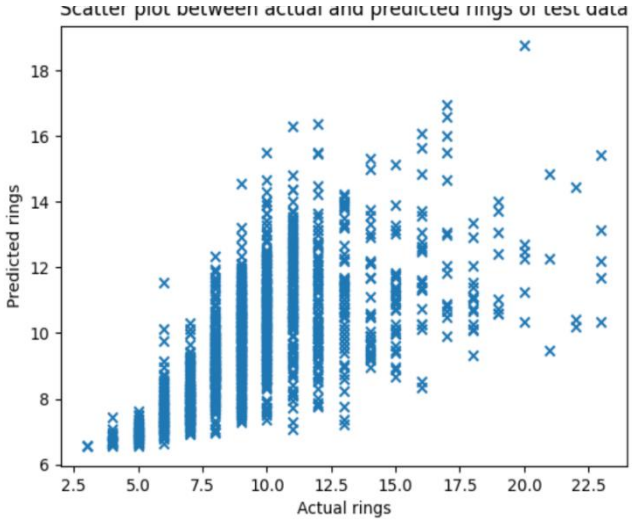

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based on above plot the prediction is not very accurate.
2. Because the spread of actual rings is 2-23 and that of predicted is 6-20.

2

a.

The prediction accuracy on training data using RMSE is 2.216.

b.

Report the prediction accuracy on testing data using RMSE is 2.219.

Inferences:

3. Testing data accuracy is very slightly higher than training data.
4. It seems that our model is significantly accurate since the test and train accuracy is almost equal .

c.

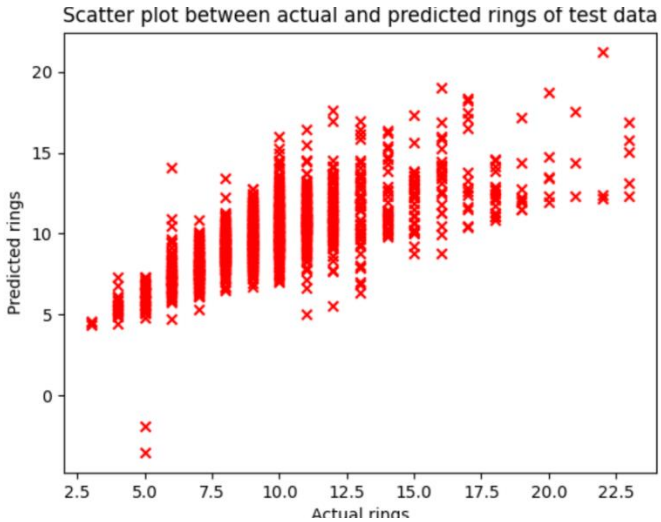

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of data points, the predicated rings are more accurate than previous one
2. Because the spread of the actual data is 2-23 and that of predicted is 4.8 to 21.
3. Multivariate linear regression seems to be more accurate than the univariate linear regression because it consider all the attributes in training model.

3

a.

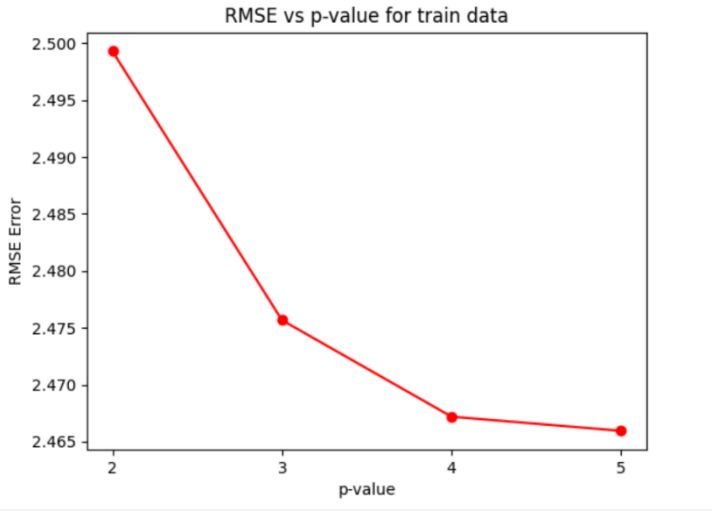

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE values decreases with respect to the increase in the degree of the polynomial.
2. The decrease is more from 2 to 3 and then gradual.
3. As the degree increases the curve fits the data better so RMSE decreases.
4. Degree 5 will fit the data more accurate.
5. As the degree increases, the bias decreases and variance increases.

b.

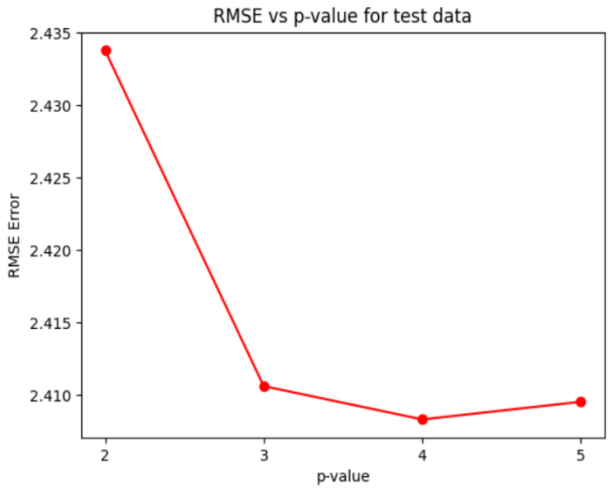

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial.
2. The decrease is more from 2 to 3 and after that its gradual.
3. As the degree increases the curve fits the data better so RMSE decreases.
4. Degree = 4 will fit the curve better than other p values.
5. As the degree increases, the bias decreases and variance increases.

c.

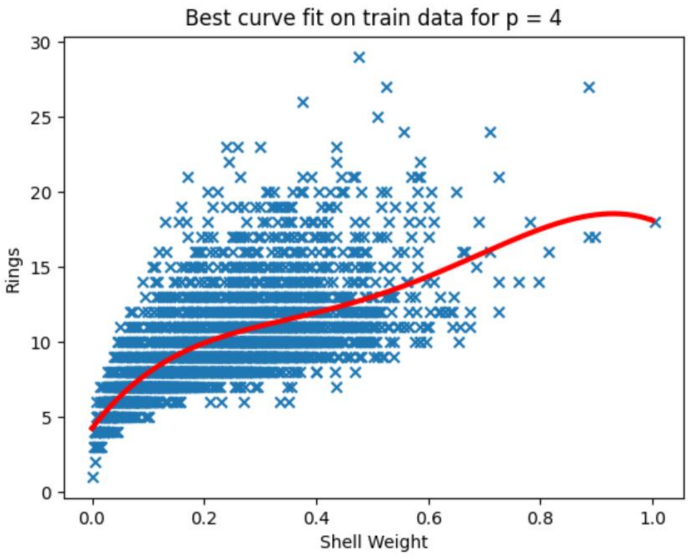

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. P = 4.
2. The RMSE value for p = 4 is lowest among all other so it will fit more accurately.
3. The bias decreases and variance increase with increasing value of p.

d.

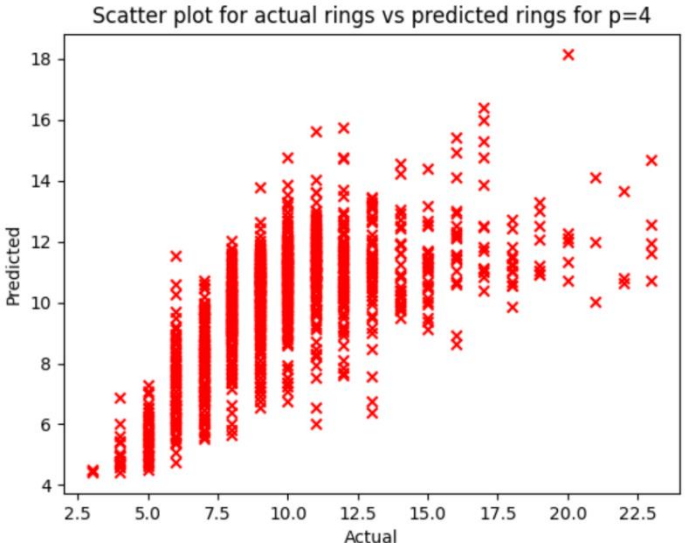

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted temperature is quite accurate.
2. The spread of actual rings is 3-23 while that of predicted rings is 4-20.
3. The accuracy for Univariate non-linear is the highest closely followed by Multivariate Linear model and least is for univariate linear model.
4. Polynomial fitting is better than linear fitting since it takes account of various ups and down in data.
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.

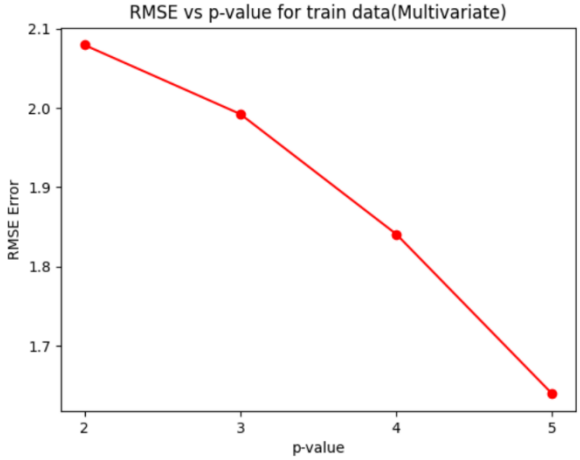

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial.
2. The decrease is more or less uniform but after p=4 the decrease is more.
3. As the degree increases the curve fits the data better so RMSE decreases.
4. From the RMSE value, 5-degree curve will fit the data best among all other values.
5. The bias decreases and variance increase with respect to the increase in the degree of the polynomial.

b.

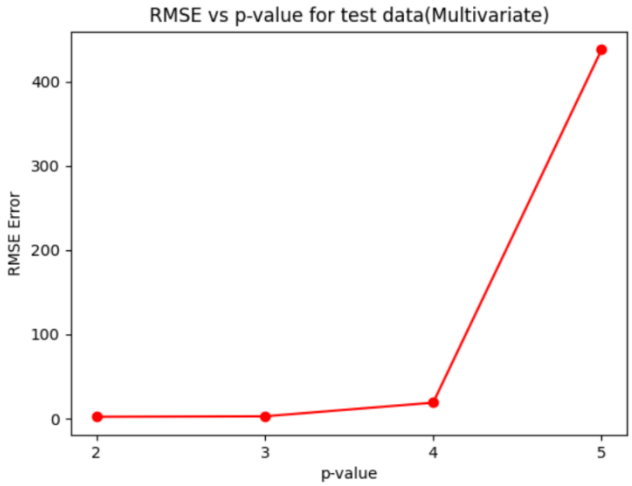

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial and starts increasing after p=3.
2. The decrease is uniform till p=3 but after p=3 the increase is drastic.
3. As we increase the degree of polynomial our model will become overfit.
4. 2-degree curve will fit best.
5. The bias gradually decreases till p=3 and then suddenly increases after p=3 and the variance increase as the model becomes more complex with increasing degree of polynomial.

c.

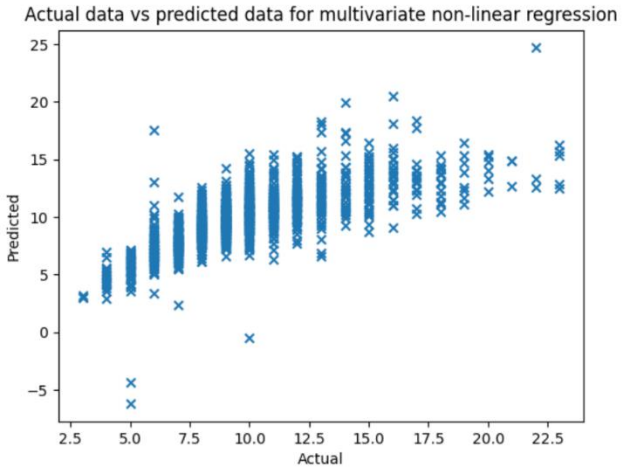

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted data points are very accurate
2. The spread for both actual and predicted data is almost similar i.e., 2.5-23.
3. The multivariate non-linear regression model has the highest accuracy followed by univariate nonlinear model and the accuracy of multivariate linear is less than that of univariate non-linear model but more than univariate linear regression model
4. Multivariate non-linear regression will give highest accuracy since it fits the curve taking account of all the attributes.
5. n linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high.