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: Rajeev Kumar **Mobile No: 9341062431**

Roll Number: b20124 **Branch: cs**

PART - A

1 a.

		Prediction Outcome
True Label	71	14
	5	145

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
LabelTrue	78	7
	47	103

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
LabelTrue	75	10
	8	142

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
LabelTrue	72	13
	18	132

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	**91.915**
4	**77.021**
8	**92.340**
16	**86.809**

Inferences:

1. The highest classification accuracy is obtained with Q =8
2. Increasing the value of Q decreases the prediction accuracy first and then starts increasing and then decreases..
3. This happens because adding nodes with less weight causes the model to overfit on training data hence the accuracy decreases.

4. If the classification accuracy increases with the increase in value of Q the number of diagonal elements in the confusion matrix increase.
5. As the classification accuracy increases with the increase in value of Q the number of off- diagonal elements decrease.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	83.9
2.	KNN on normalized data	98.5
3.	Bayes using unimodal Gaussian density	94.345
4.	Bayes using GMM	92.340

Inferences:

1. the classifiers with the highest accuracy is KNN on normalized data and lowest accuracy is of KNN.
2. the classifiers in ascending order of classification accuracy. KNN < Bayes using GMM < Bayes using unimodal Gaussian density < KNN on normalized data
3. the reason for increase in classification accuracy after data normalization is because it changes the value of data in one column to a fixed scale

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 attribute with the highest correlation coefficient was used for predicting the target attribute Rings as it represent high dependency.
2. the best fit line doesn't fit the training data perfectly.
3. the best fit line doesn't fit the training data perfectly as it's oversimplified.
4. High bias and low variance trade-off for the best fit line.

b.

The prediction accuracy on training data is 2.528.

c.

the prediction accuracy on testing data is 2.468.

Inferences:

1. Amongst training and testing accuracy, training accuracy is high.

d.

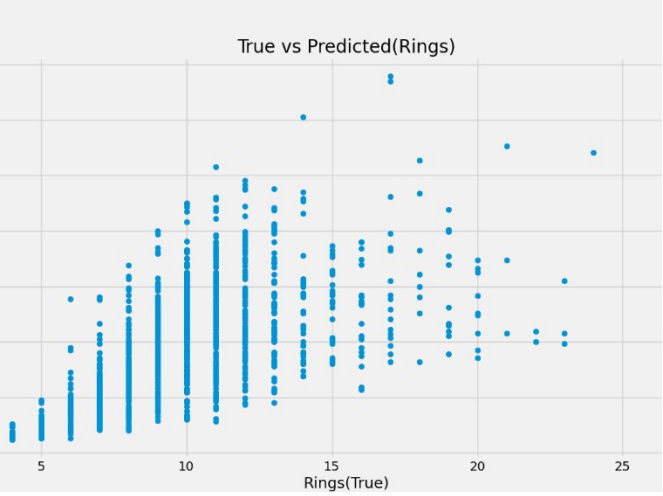

Figure 6 Univariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, the predicted temperature is inaccurate.
2. Because the spread if actual rings is 2-23 while that of predicted is 6-20

3

a.

the prediction accuracy on training data is 2.216.

b.

the prediction accuracy on testing data is 2.219.

Inferences:

Amongst training and testing accuracy, testing data has slightly high accuracy.

c.

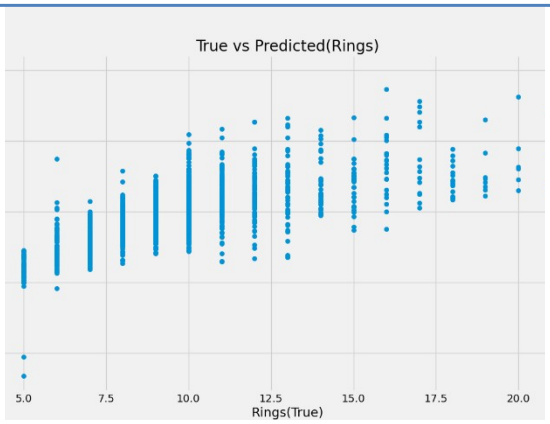

Figure 7 Multivariate linear regression model: Scatter plot of predicted rings from linear regression model vs. actual rings on test data

Inferences:

1. Based upon the spread of the points the predicted number of ring is high.
2. The spread of Actual Rings is 5-23 and that of Predicted Rings is 4.8-22.

4

a.

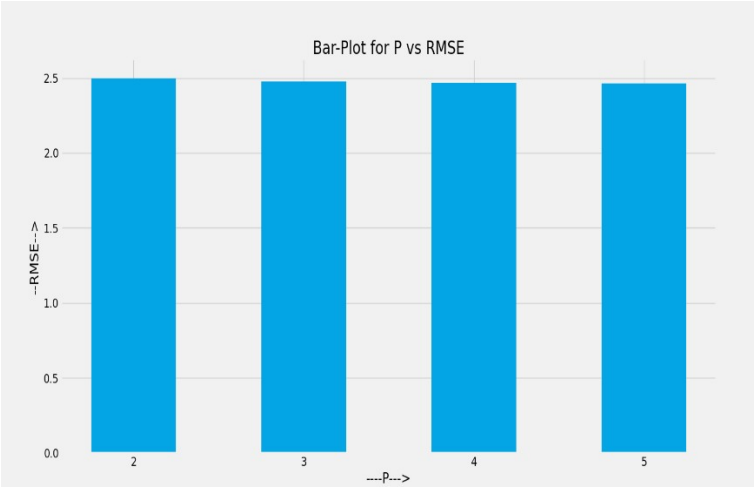

Figure 8 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. Decreases from p =2 to 3 and then become gradual.
3. As the degree increases the curve fits the data more better so RMSE decreases.
4. From the RMSE value, 5 degree curve will approximate the data best.
5. bias decreases and variance increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

b.

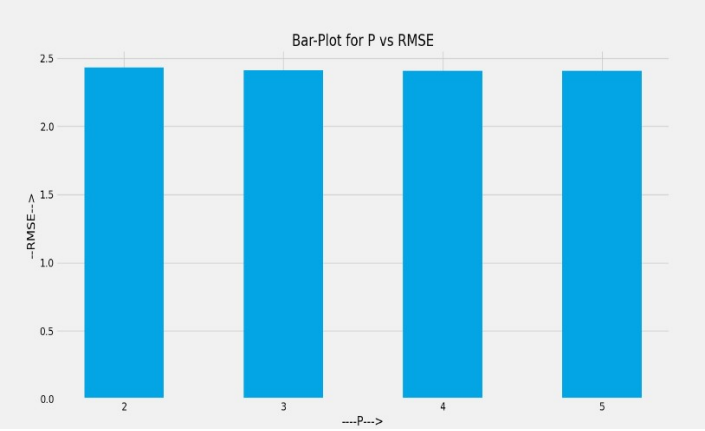

Figure 9 Univariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. Infer whether RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. after a certain p-value the decrease becomes gradual.
3. As the degree increases the curve fits the data more better so RMSE decreases.
4. From the RMSE value, 4 degree curve will approximate the data best.
5. bias decreases and variance increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

c.

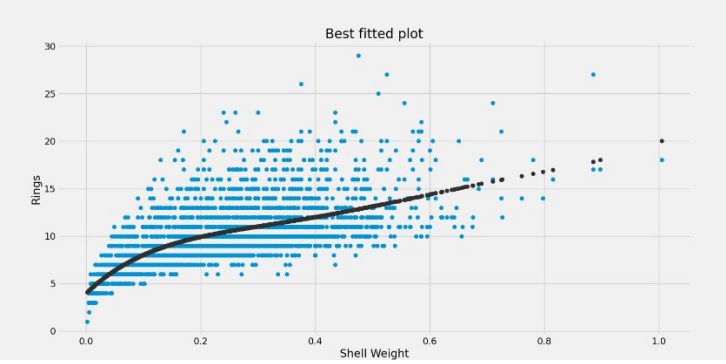

Figure 10 Univariate non-linear regression model: Rings vs. chosen attribute(replace) best fit curve using best fit model on the training data

Inferences:

1. the p-value corresponding to the best fit model is p=4
2. bias decreases and variance increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

d.

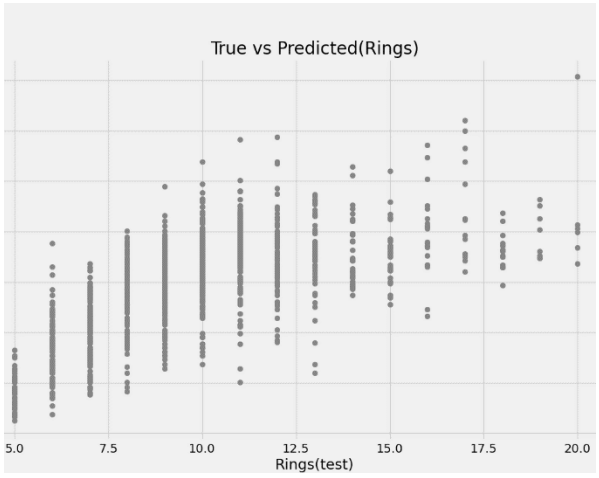

Figure 11 Univariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, infer how accurate the predicted temperature is more likely accurate.
2. State the reason for Inference 1.
3. The accuracy for Univariate non-linear is the highest closely followed by Multivariate Linear model and least is for univariate linear model.
4. In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high.

4a.

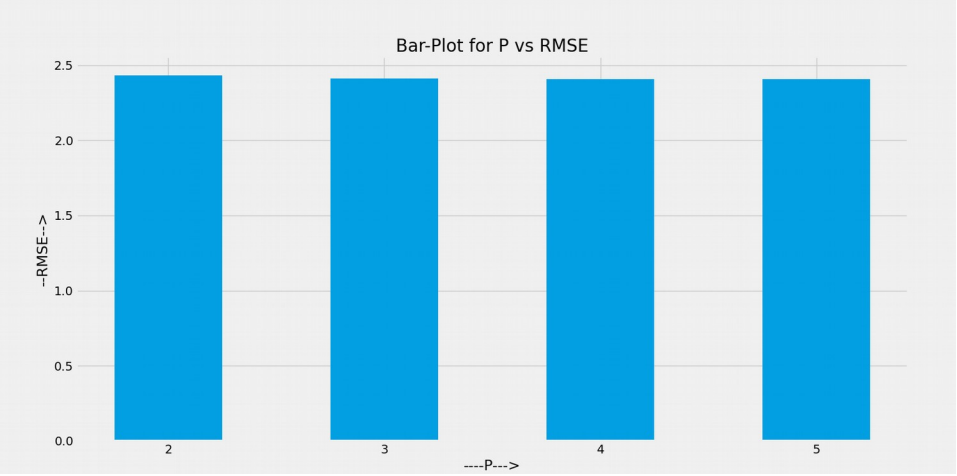

Figure 12 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. Infer whether RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. Decreases from p=2 to 3 then become gradual.
3. From the RMSE value degree curve p=5 will approximate the data best.
4. Bias decreases and variance increases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).

b.

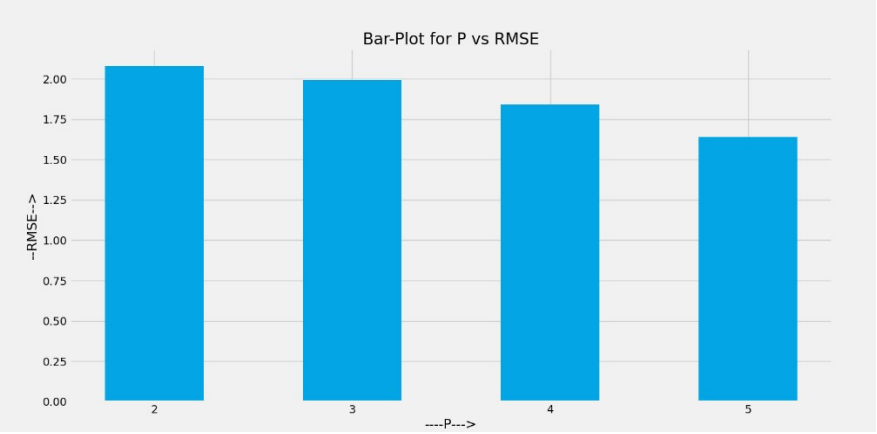

Figure 13 Multivariate non-linear regression model: RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. Infer whether RMSE value decreases with respect to the increase in the degree of the polynomial (p = 2, 3, 4, 5).
2. Decrease is uniform till p=2 to 3 after decrease is more.
3. As we increased the degree of polynomial our model became overfitted.
4. From the RMSE value p= 5 degree curve will approximate the data best.

c.

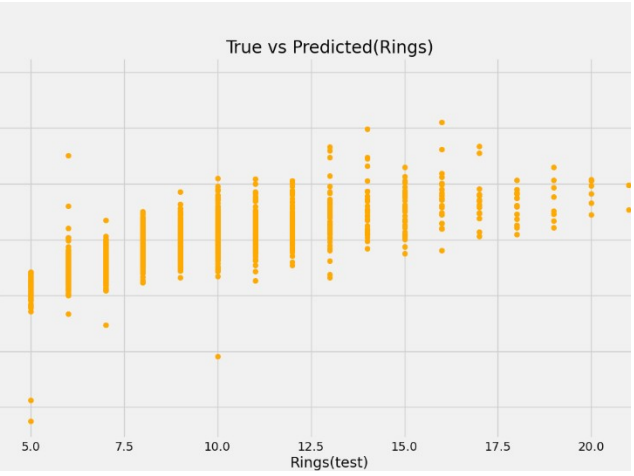

Figure 14 Multivariate non-linear regression model: Scatter plot of predicted rings vs. actual rings on test data

Inferences:

1. Based upon the spread of the points, infer how accurate the predicted temperature is more likely accurate.
2. The spread of actual rings is 3-23 and that of predicted rings is also 3-22.

3. The multivariate non-linear regression model has the highest accuracy followed by univariate non-linear model and the accuracy of multivariate linear is less than that of univariate non-linear model but more than univariate linear regression model.
4. In linear regression models bias is high, variance is low and in non-linear regression models bias is low, variance is high.