Data Science Project
Sub-Project – V
Data classification using Bayes Classifier with Gaussian Mixture Model (GMM); Regression using Simple Linear Regression and Polynomial Curve Fitting

Student's Name: Anmol Bishnoi **Mobile No: 7042845211**

Roll Number: B19069 **Branch: CSE**

PART - A

1 a.

	Prediction Outcome	
True Label	706	19
	42	9

Figure 1 Bayes GMM Confusion Matrix for Q = 2

	Prediction Outcome	
True Label	682	43
	42	9

Figure 2 Bayes GMM Confusion Matrix for Q = 4

	Prediction Outcome	
True Label	719	6
	47	4

Figure 3 Bayes GMM Confusion Matrix for Q = 8

	Prediction Outcome	
True Label	716	9
	48	3

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	**92.14**
4	**89.05**
8	**93.17**
16	**92.65**

Inferences:

1. The highest classification accuracy is obtained with Q = 8
2. The Q value first decreases then increases and then decreases again.
3. This is because, realistic data is hardly ever perfectly n-modal. Hence, to accommodate the deviations the data shows despite having say n modes, adding more modes with less weights takes in those

deviations. However, overdoing this is not only a waste of computer time and effort, but also it may make the model overfitted on the train data, and hence decrease the test accuracy As the classification accuracy increases with the increase in value of K, the number of diagonal elements increases for true positive and true negative.

4. When the classification accuracy increases, the number of diagonal elements in the Confusion matrix increases and when the classification accuracy decreases, the number of diagonal elements in the Confusion matrix decreases.
5. When the classification accuracy increases, the number of off-diagonal elements in the Confusion matrix decreases and when the classification accuracy decreases, the number of off-diagonal elements in the Confusion matrix increases.
6. Off-Diagonal elements decrease when Q increases because increasing Q increases accuracy which is calculated on the basis of diagonal elements, so if diagonal elements increase then off diagonal will decrease.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	92.40
2.	KNN on normalized data	92.40
3.	Bayes using unimodal Gaussian density	88.92
4.	Bayes using GMM	93.17

Inferences:

1. Highest Accuracy => KNN on Normalised Data
   Lowest Accuracy => Bayes using unimodal Gaussian Density
2. Bayes Unimodal < KNN = KNN (normalized) < Bayes using GMM
3. Unimodal Bayes assumes perfectly Normal distribution of data, which is not the case here and hence gives the lowest accuracy.
4. KNN assumes nothing and finds out classes by optimizing locally and hence gives very accurate results.
5. Multimodal Bayes performs better as we are now using multiple clusters which allow for relatively higher accuracy.

PART – B

1

a.

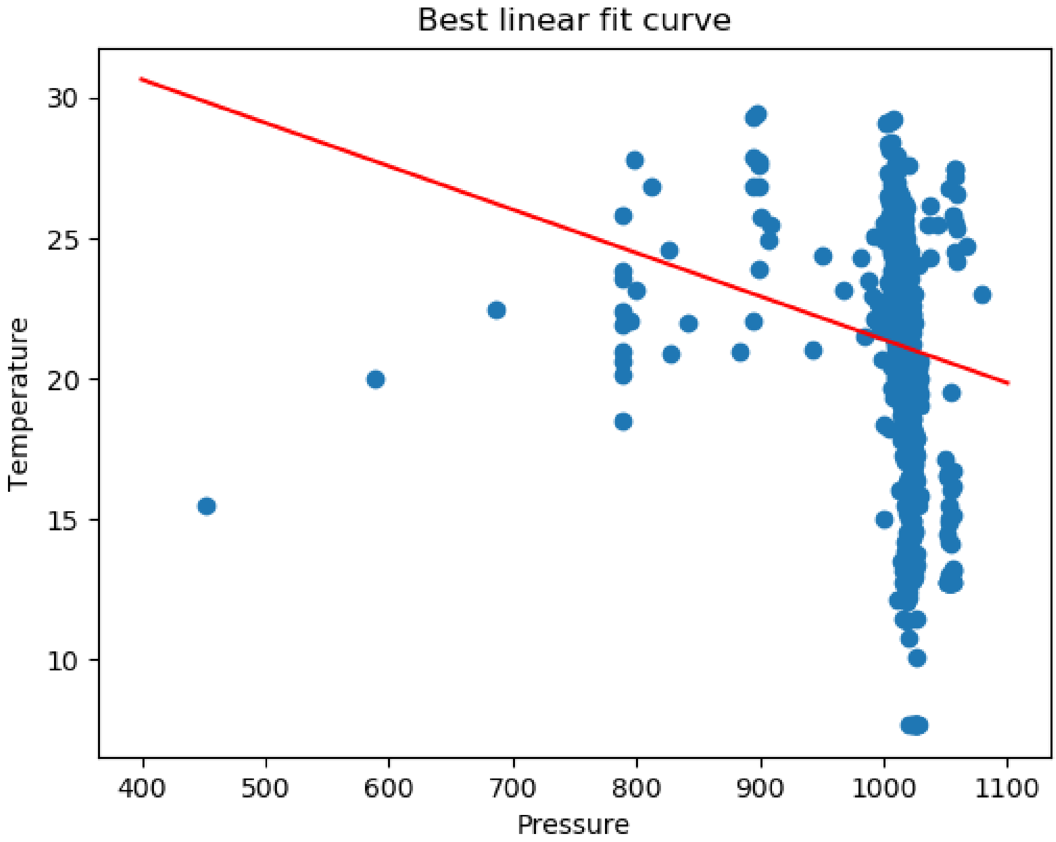

Figure 5 Pressure vs. temperature best fit line on the training data

Inferences:

1. No, the best fit line does not fit the training data perfectly.
2. It does not fit the training data perfectly as it is oversimplified for the data, a more complex function is required to fit the data.
3. Bias is high as the best fit line underfits the data, the model requires more complex functions to fit the training data. Variance is also low due to underfitting.

b.

RMSE for Training Data Set on Linear Regression Model: 4.280

c.

RMSE for Training Data Set on Linear Regression Model: 4.287

Inferences:

1. The accuracy on the training data is higher as the rmse value for it is lower.
2. Because we trained the data on the training set, therefore it is being predicted a little better.

d.

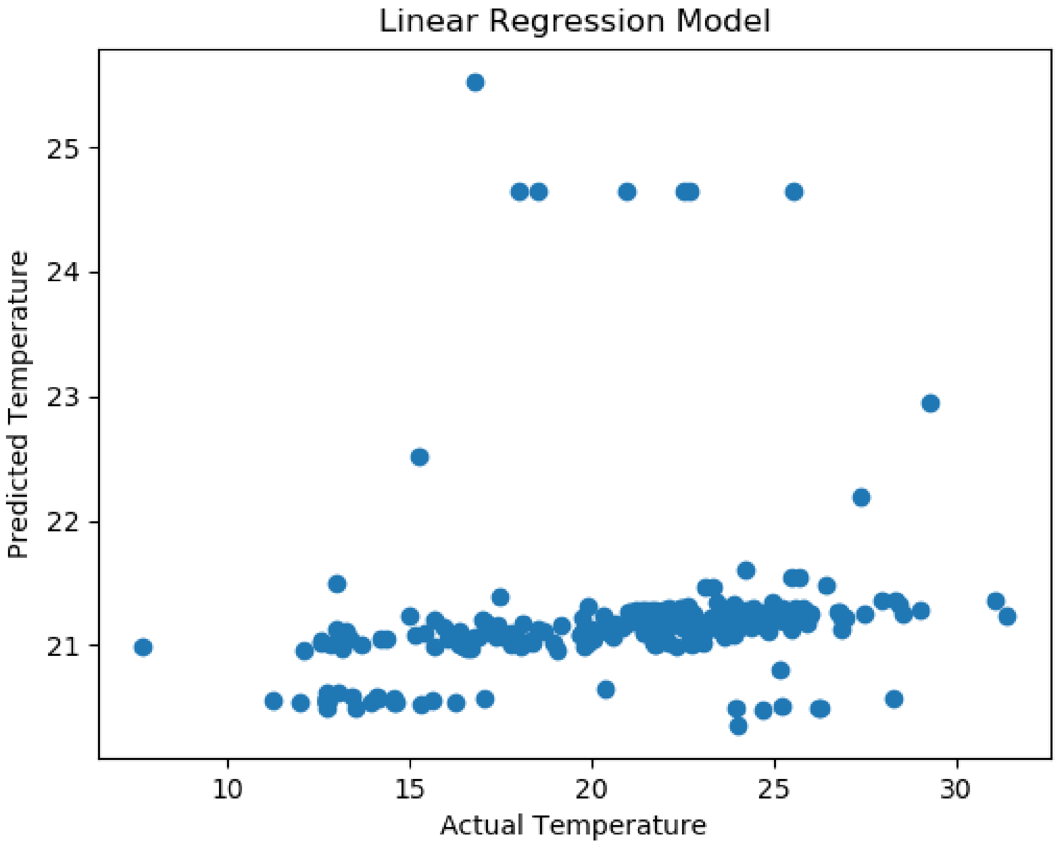

Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. The prediction accuracy is not very high.
2. The actual temperature is spread from 10 to 30 but the predicted temperature is more concentrated from 20 to 23 which shows that the prediction accuracy is not high.

2

a.

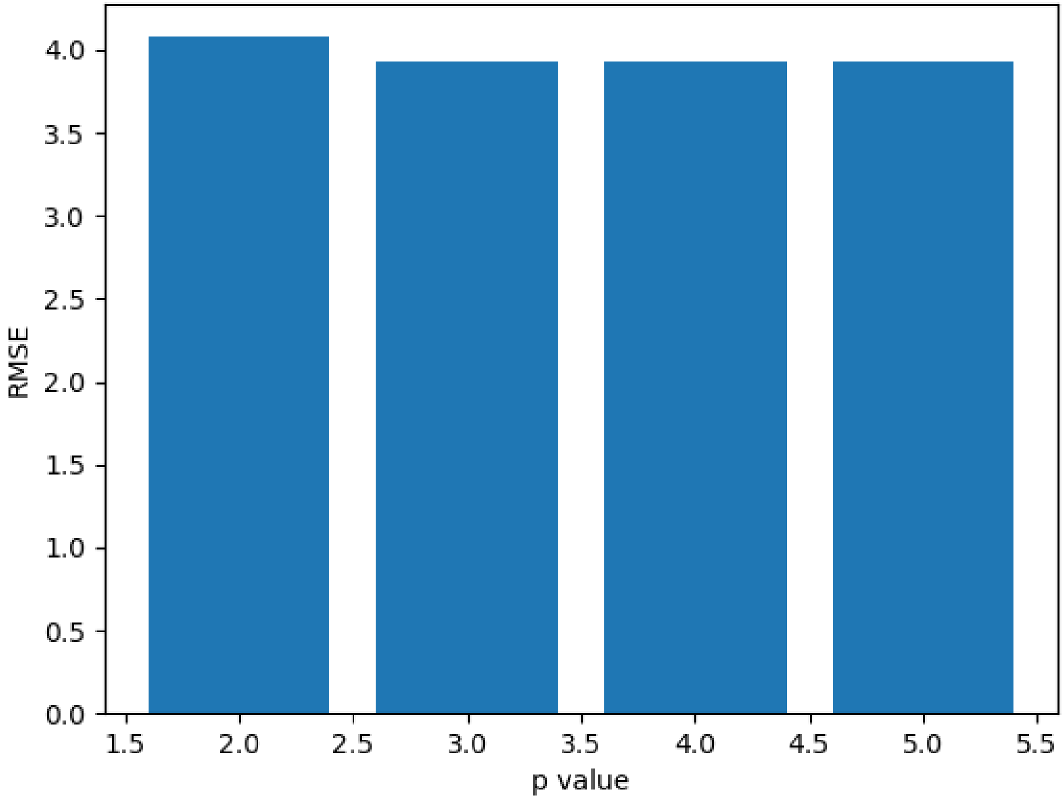

Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

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 more better, so the RMSE decreases.
4. From the RMSE value, degree p = 4 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.

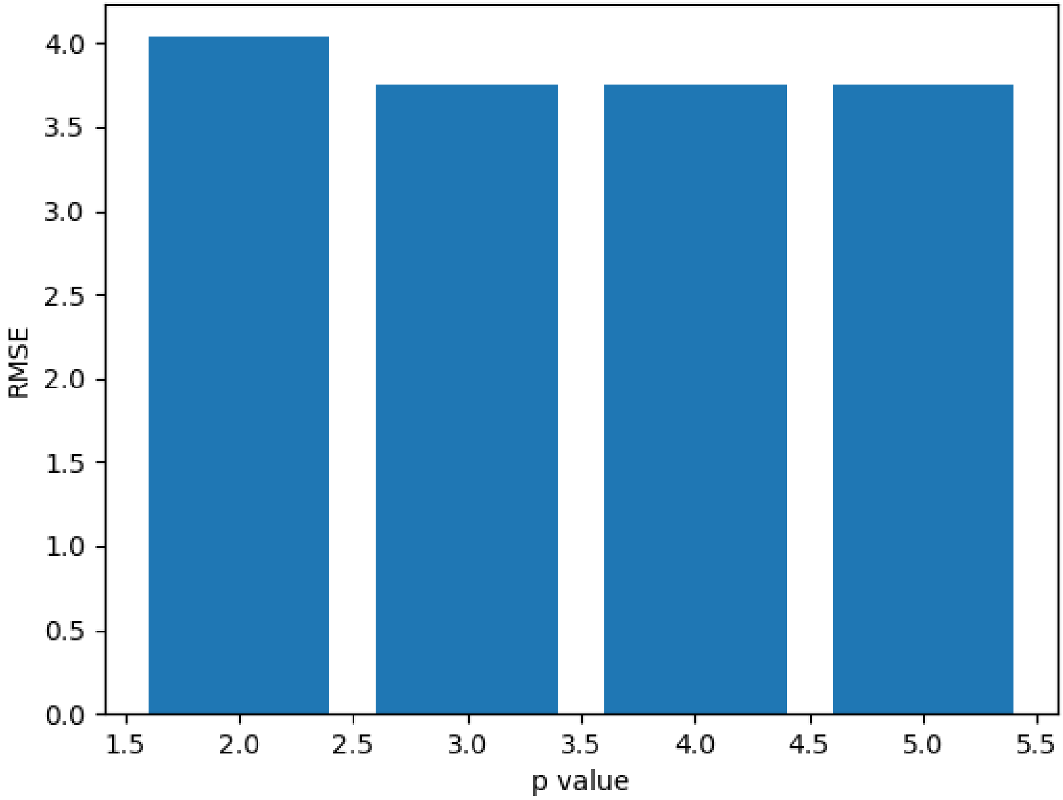

Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

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 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.

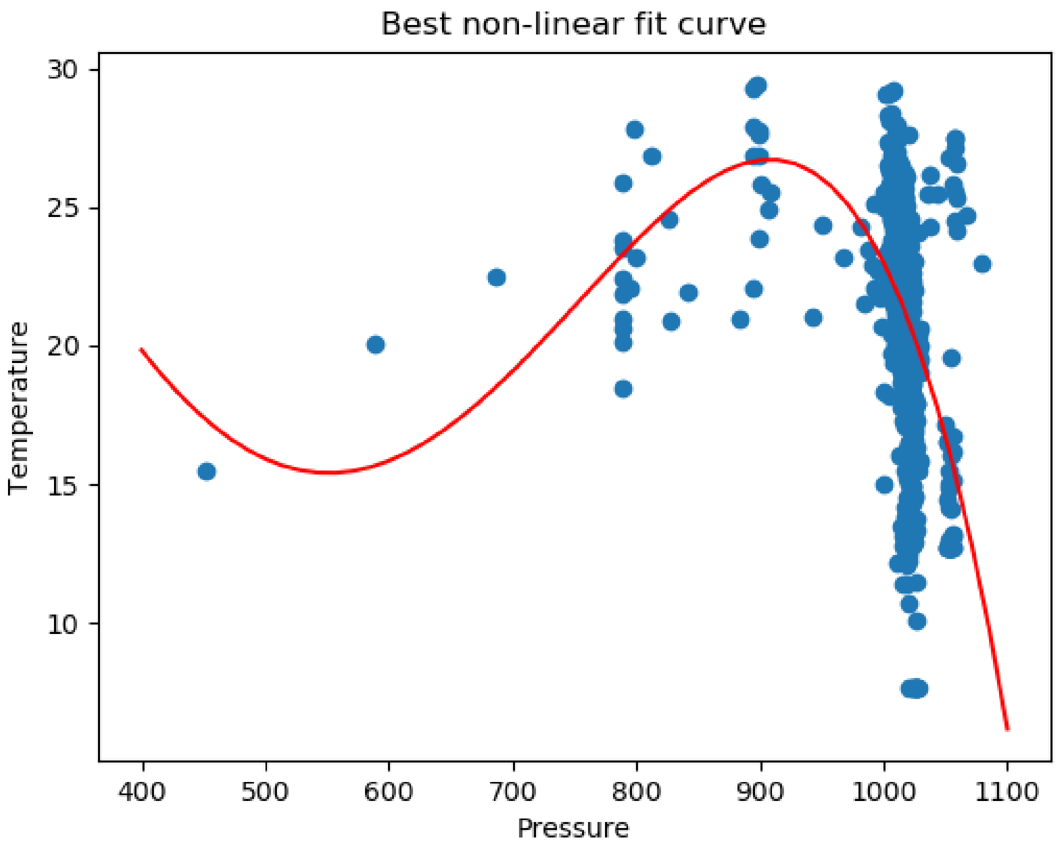

Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

Inferences:

1. p-value is 5 corresponding to the 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.

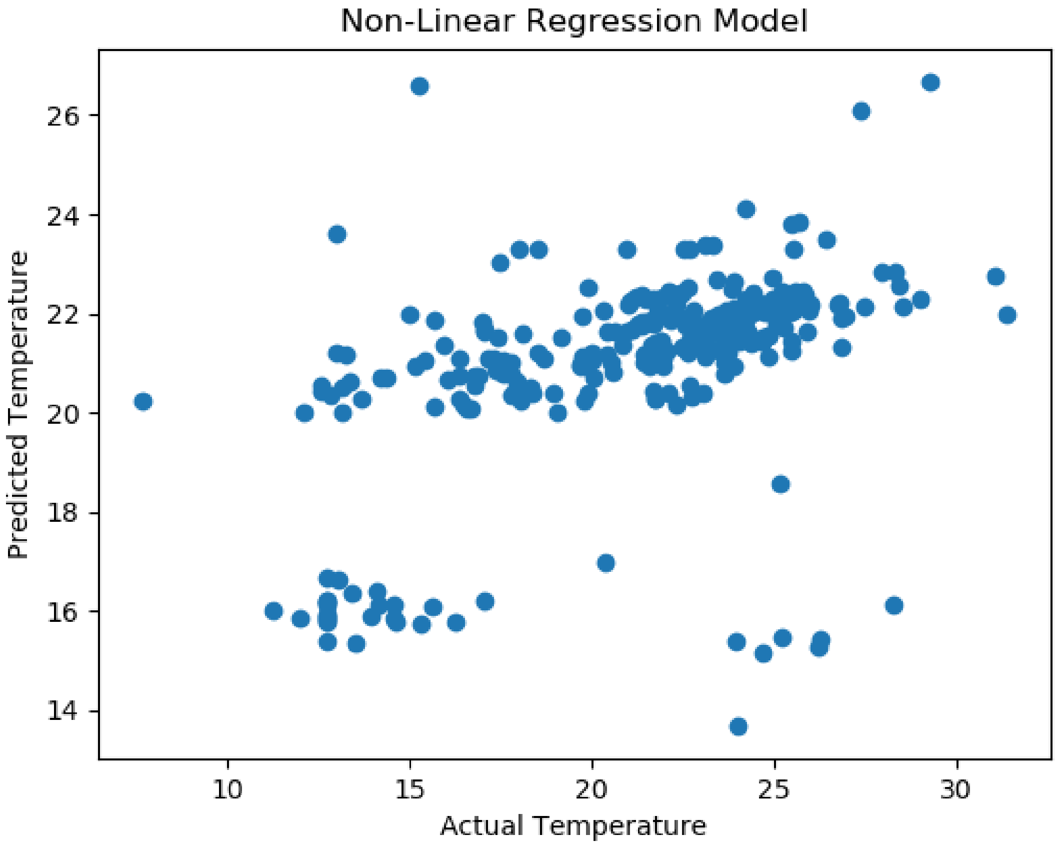

Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. From the spread of points we can see that accuracy of predicted temperature is quite good.

2. The actual temperature is spread between 10 and 30, similarly the predicted temperature is also spread between 10 to 30, thus we can say that the accuracy is good.

3. Prediction accuracy of non linear is better as the rmse is lower for it, also from the spread of data we can see that the non linear regression is better than linear regression.

4. Rmse of non linear regression is lower than linear and the spread of predicted value matches actual value better in nonlinear regression than linear, so we can say that non linear 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.