IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – V
Data classification using Bayes Classifier with Gaussian Mixture Model (GMM); Regression using Simple Linear Regression and Polynomial Curve Fitting

Student's Name: Aryan Garg **Mobile No: +91-8219383122**

Roll Number: B19153 **Branch: EE**

PART - A

1 a.

		Prediction Outcome
True Label	27	24
	11	40

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome
True Label	41	10
	18	33

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome
True Label	24	27
	33	18

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome
True Label	25	26
	46	5

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	65.686
4	72.549
8	23.529
16	~0

Inferences:

1. The highest classification accuracy is obtained with Q = 4.
2. Accuracy increases till K = 4 then decreases exponentially.
3. Increasing till 4 components is viable and fits the data well but more components than that lead to overfitting of data.
4. Diagonal elements increase as accuracy increases (till K = 4) due to the same reason as in point 3.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	75.490
2.	KNN on normalized data	72.549
3.	Bayes using unimodal Gaussian density	78.431
4.	Bayes using GMM	72.549

Inferences:

1. Bayes classifier using unimodal density produced the best result while Bayes using GMM and KNN on normalized data performed poorly.
2. KNN on normalized data < Bayes using GMM < KNN < Bayes using unimodal density
3. Clearly, Bayes on unimodal density does better than KNN models because it considers all of the data while KNN considers certain number of neighbors. Normalizing data leads to a slight loss which is how we justify the loss in accuracy (trade-off for time complexity). Bayes using GMM is also performing poorly due to overfitting on a small dataset.

PART – B

1

a.

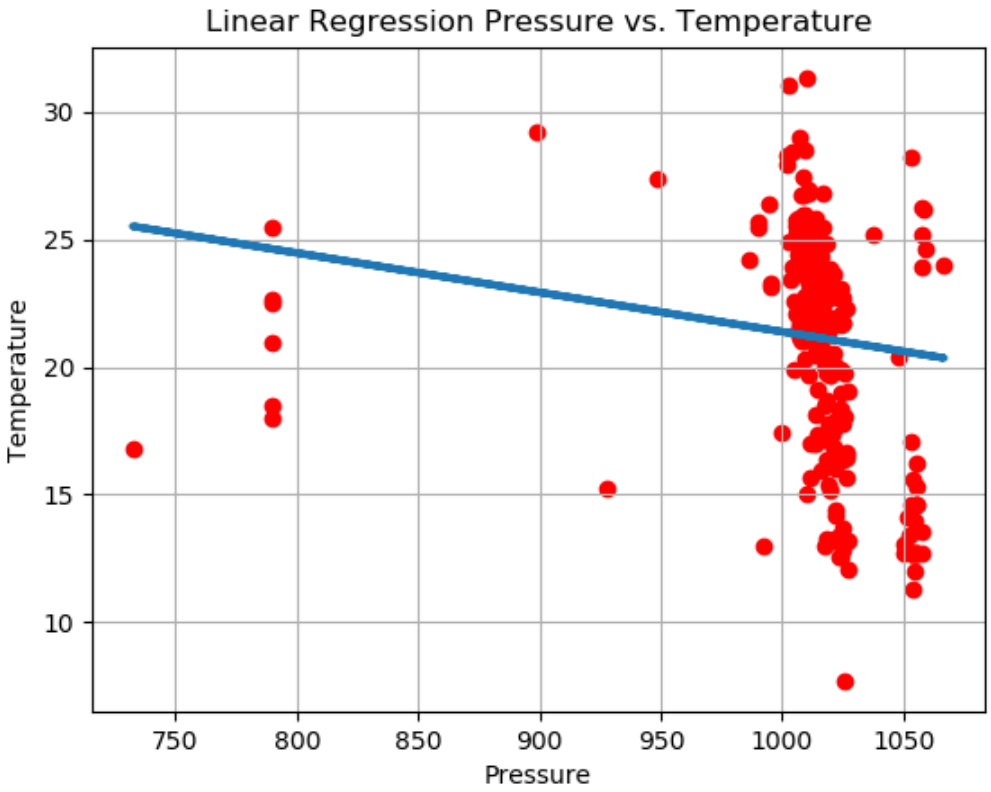

Figure 5 Pressure vs. temperature best fit line on the training data

Inferences:

1. The best fit line does not fit the training data perfectly
2. Clustering of data around a region and some outliers on far off places in the hyperspace produce poor results for linear regression as is evident in figure 5.
3. An ideal model is one which has low variance and low bias. If the variance is high and low bias, the model is prone to overfitting and if the bias is high and low variance, to underfitting. This simple linear regression model clearly has low variance and a high bias for certain values. It doesn't capture the underlying patterns of our dataset.

b.

Prediction accuracy of train data: 8.388 (Prediction accuracy = 1.96 * RMSE)

c.

Prediction accuracy of test data: 8.402

Inferences:

1. Training accuracy is higher as its prediction accuracy is lower.
2. The mere fact that the model was built considering this data is the reason behind better accuracy on that data.

d.

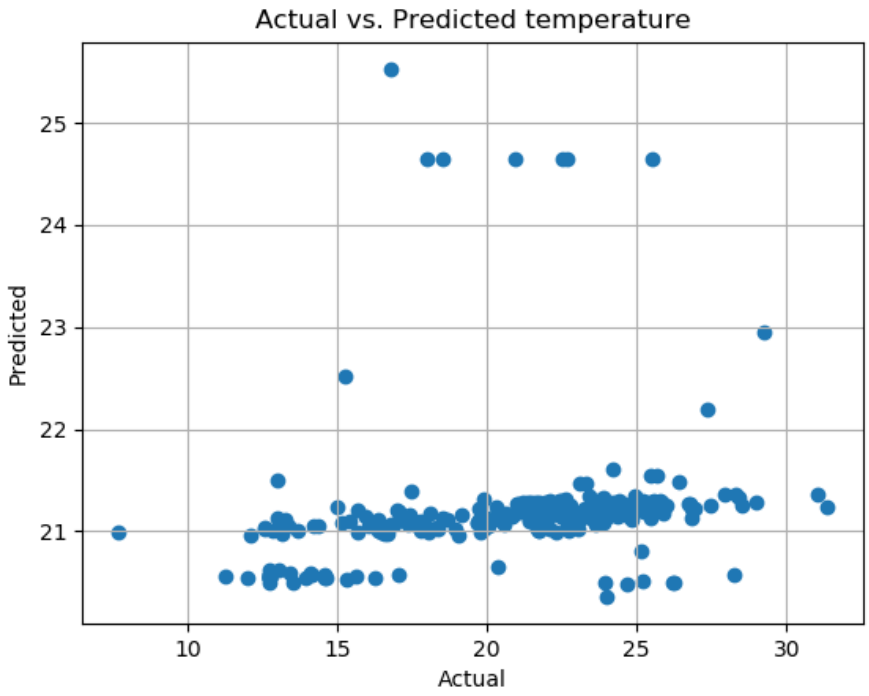

Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. The model predicts the temperature at a higher value than it actually is.
2. Due to the model's high bias, the model has underfitted the data and gives us a poor estimation.

2

a.

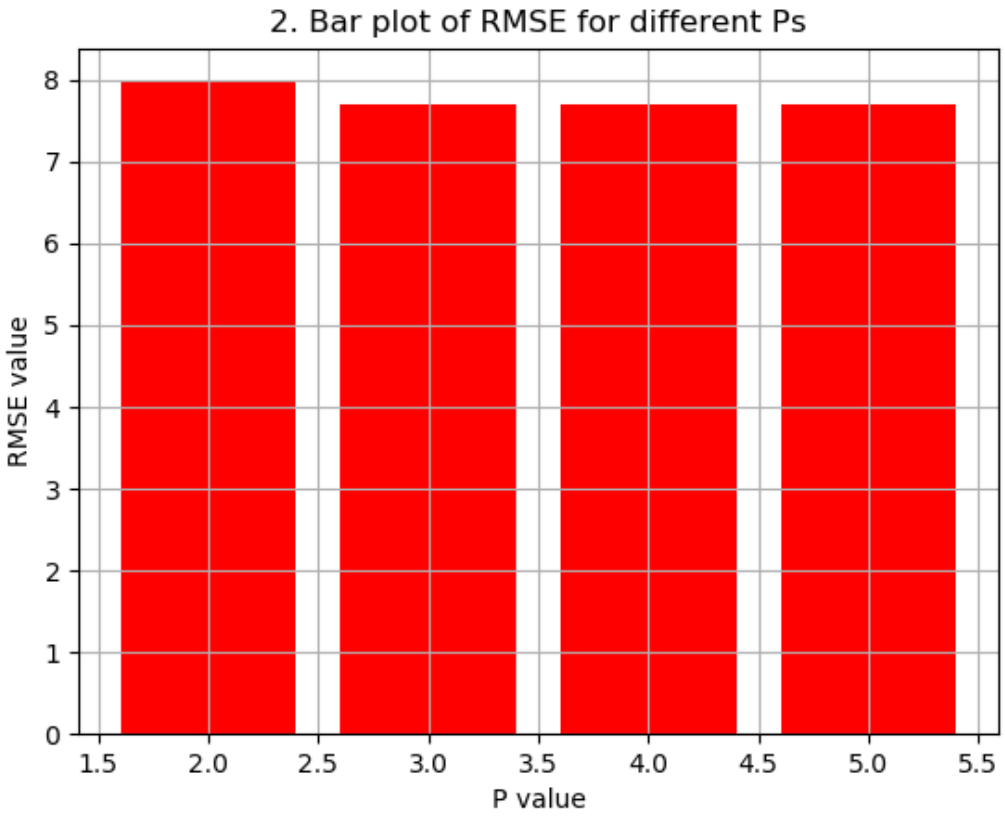

Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE decreases till a certain degree then it almost becomes constant/gradually declines.
2. A polynomial of degree = 3 is probably best to fit this data and further increasing the degree of the polynomial is just going to overfit the training data on noise.
3. As we increase the degree of the polynomial, we increase the variance on the data but reduce the overall bias. An optimal trade-off is reached by a third degree polynomial.

b.

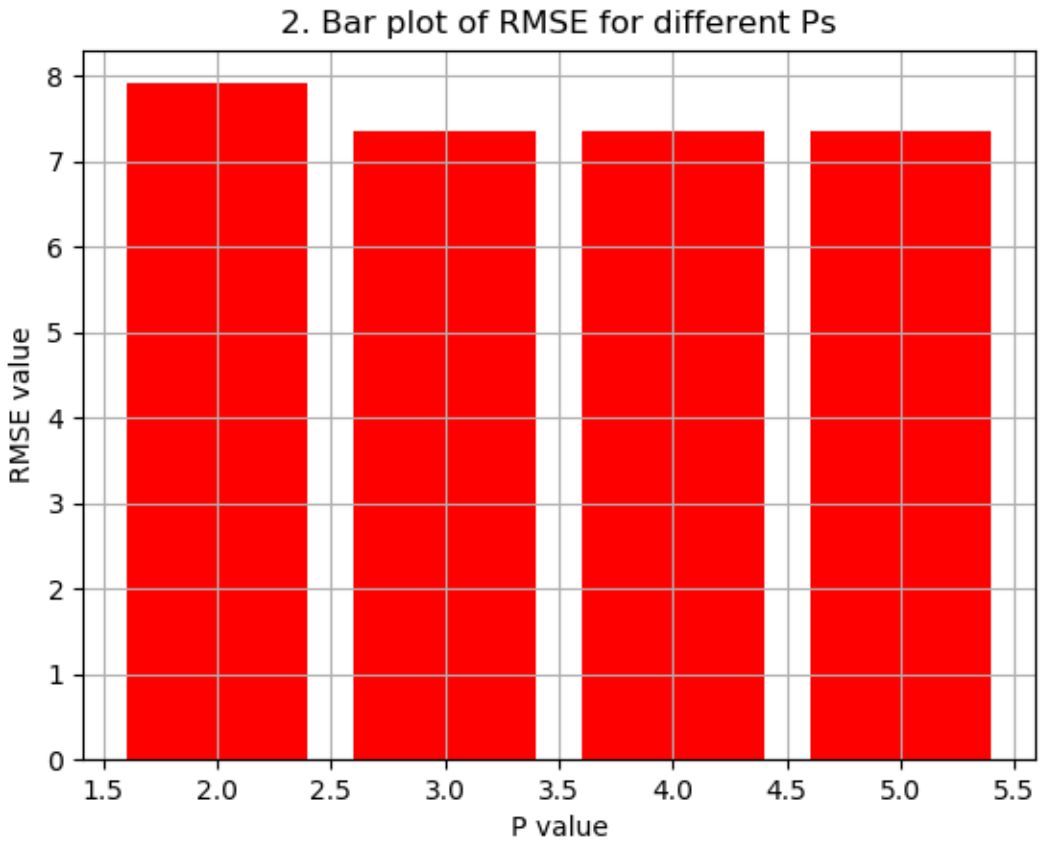

Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE decreases as we increase the degree of the polynomial.
2. Initially, RMSE drops drastically then gradually declines.
3. Polynomials are innately bound to fit real world data but also run the risk of overfitting data if their degree is unchecked.
4. Degree = 5 gives the lowest error on test data and might be a viable option to start with.

c.

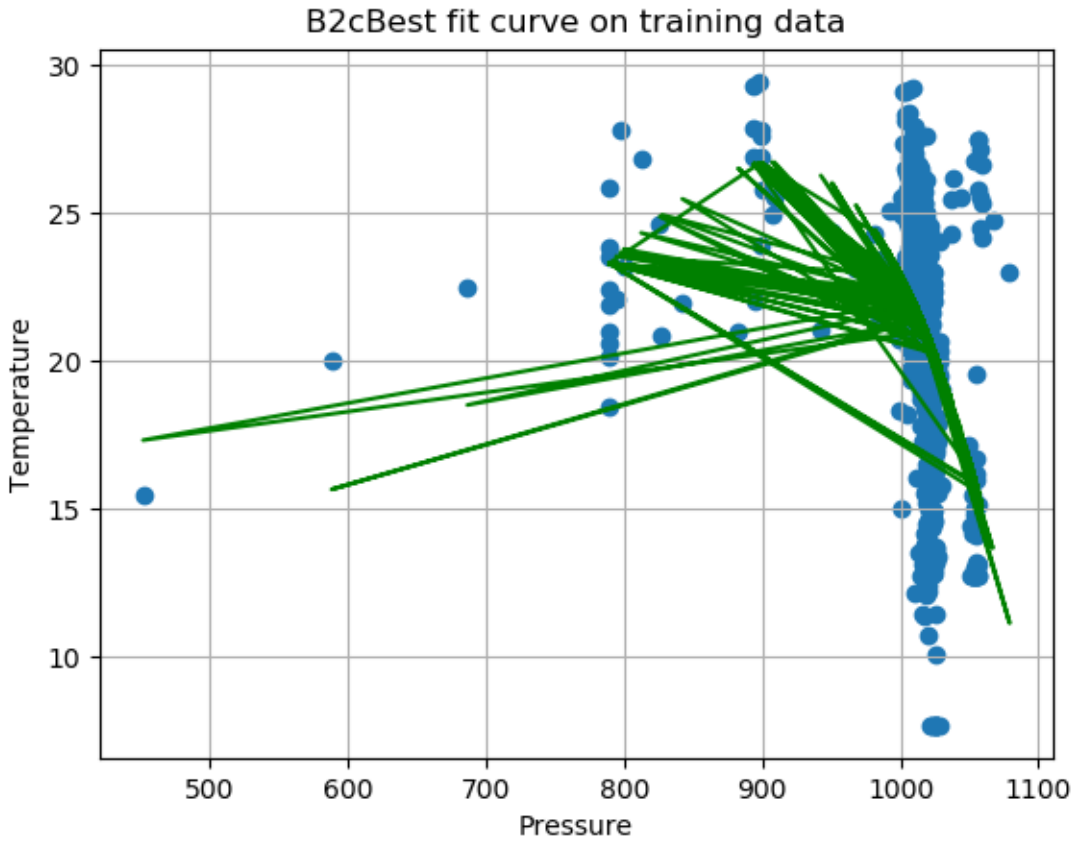

Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

Inferences:

1. The p-value is 5.
2. As we plotted the bar charts of the RMSE values on test data, we have taken this value to proceed with.
3. Taking p = 5 has higher variance and lower bias than a more stable choice of taking p = 3 or 4, but we prefer this because of the highly clustered nature of the data which is evidently non-linear.

d.

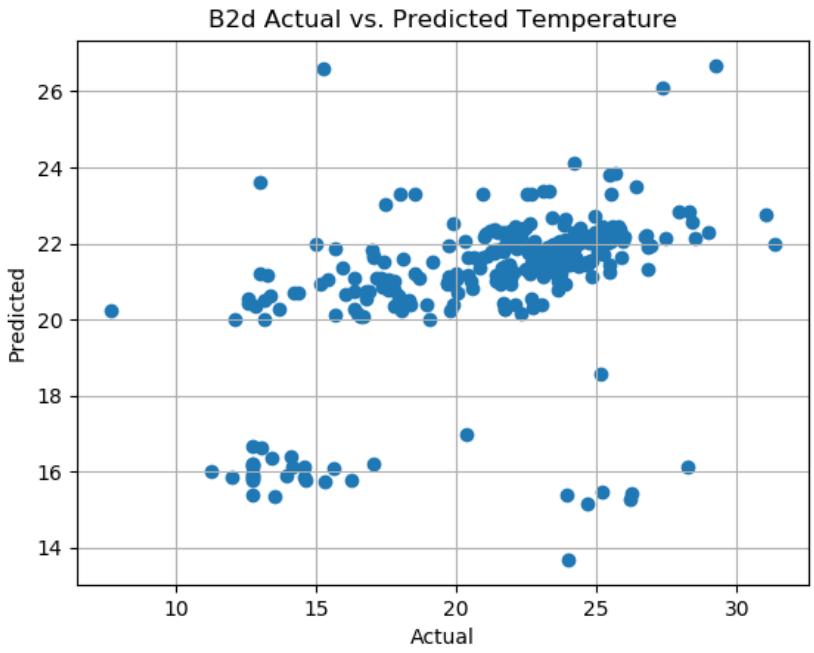

Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. The non-linear regressor closes in on the gap between reality and prediction.
2. Polynomial fitting of the data suits this dataset better as compared to linear modelling.
3. Clearly, this model has produced more accurate results as compared to a linear regression mode.
4. By going for a polynomial fitting, we automatically increase variance at the root-level thus radicalizing the same algorithm. This higher variance was much needed for the dataset at hand and was clearly non-linear on simple scatter-plot inspections.