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: Gajraj Singh Chouhan **Mobile No: +91-9351159849**

Roll Number: B19130 **Branch: DSE**

PART - A

1 a.

		Prediction Outcome	
True Label		**Negative Class**	**Positive Class**
	Negative Class	396	20
	Positive Class	329	31

Figure 1 Bayes GMM Confusion Matrix for Q = 2

		Prediction Outcome	
True Label		**Negative Class**	**Positive Class**
	Negative Class	705	49
	Positive Class	20	2

Figure 2 Bayes GMM Confusion Matrix for Q = 4

		Prediction Outcome	
True Label		**Negative Class**	**Positive Class**
	Negative Class	687	40
	Positive Class	38	11

Figure 3 Bayes GMM Confusion Matrix for Q = 8

		Prediction Outcome	
True Label		**Negative Class**	**Positive Class**
	Negative Class	721	51
	Positive Class	4	0

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	**55.025**
4	**91.108**
8	**89.948**
16	**92.912**

Inferences:

1. The highest classification accuracy is obtained with Q = 16.
2. Increasing the value of Q increases the prediction accuracy.
3. With increase of Q, the gaussian distributions are better fit on the data, decreasing the error and so increasing the accuracy.
4. As the classification accuracy increases with the increase in value of Q, the number of diagonal elements in Confusion matrix increase.
5. As accuracy is proportional to the correct predictions which are represented by diagonal elements, hence increasing accuracy means increasing diagonal elements.
6. As the classification accuracy increases with the increase in value of Q, the number of off-diagonal elements decrease.
7. Since number of off diagonal elements equal total elements - on diagonal elements, as on diagonal elements increase off diagonal elements decrease.

2

Table 2 Comparison between Classifiers based upon Classification Accuracy

S. No.	Classifier	Accuracy (in %)
1.	KNN	93.170
2.	KNN on normalized data	92.894
3.	Bayes using unimodal Gaussian density	87.757
4.	Bayes using GMM	92.912

Inferences:

1. Highest - KNN, lowest - Bayes with Unimodal Gaussian Density.
2. Bayes < KNN on normalized < Bayes using GMM < KNN.
3. KNN can fit the data better than Bayes with unimodal Gaussian Density.

<u>PART – B</u>

1 a.

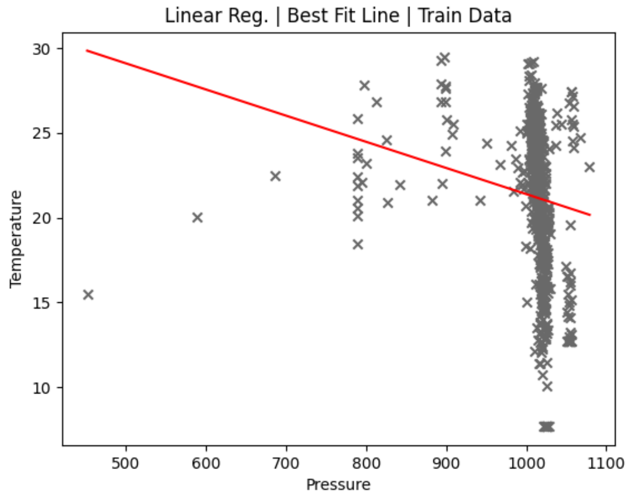

Figure 5 Pressure vs. temperature best fit line on the training data

Inferences:

1. Best fit line doesn't fit it perfectly.
2. There is very large number of bias in the best fit line, the amount of error is also large due to slope of the line not matching the spread of data.
3. There is large number bias as this doesn't fit the distribution of training data, as seen by RMSE the variance is also high.

b. Value of RMSE for training data is 4.279.

c. Value of RMSE for testing data is 4.286.

Inferences:

1. The accuracy is higher for the testing data.
2. We choose RMSE as a way to check the accuracy of the data, and if the error is more the accuracy will be less.

d.

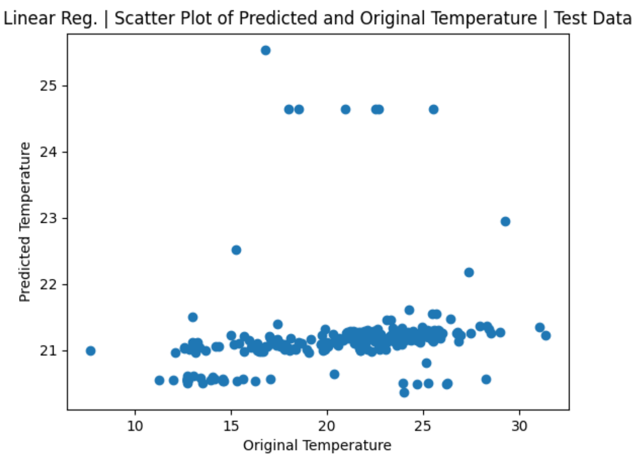

Figure 6 Scatter plot of predicted temperature from linear regression model vs. actual temperature on test data

Inferences:

1. The prediction isn't much accurate, as even for low original temperature we are getting high prediction e.g. for <10 original temp. we are getting ~21C in prediction.
2. Due to high spread of prediction temperature and mostly constant prediction (around ~21), the accuracy has become low.

2 a.

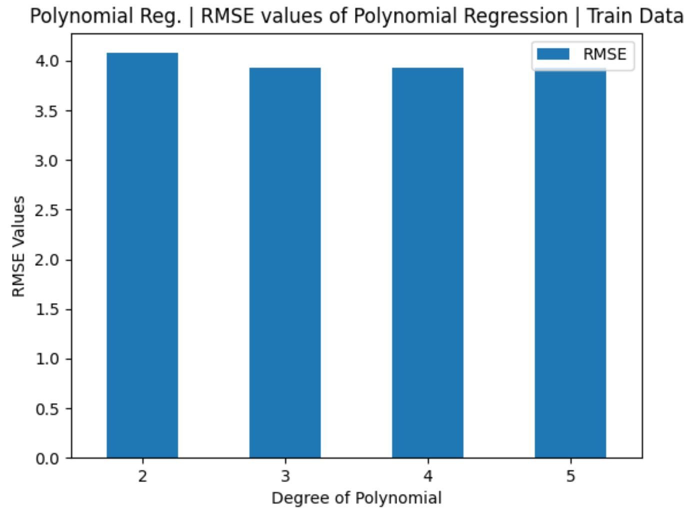

Figure 7 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the training data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. After a p=4 the decrease becomes gradual.
3. As the Degree of Polynomial increases, the RMSE is decreasing from 4 to ~3.9, after p=4 the difference between consecutive RMSE value is less than ~0.001 so we can decrease becomes gradual.
4. From the RMSE value, p=4 will approximate the data best.
5. The polynomial regression has lower bias as it fits the training data better, as degree of polynomial increases the bias will decrease. In this case the variance is decreasing due to polynomial regression.

b.

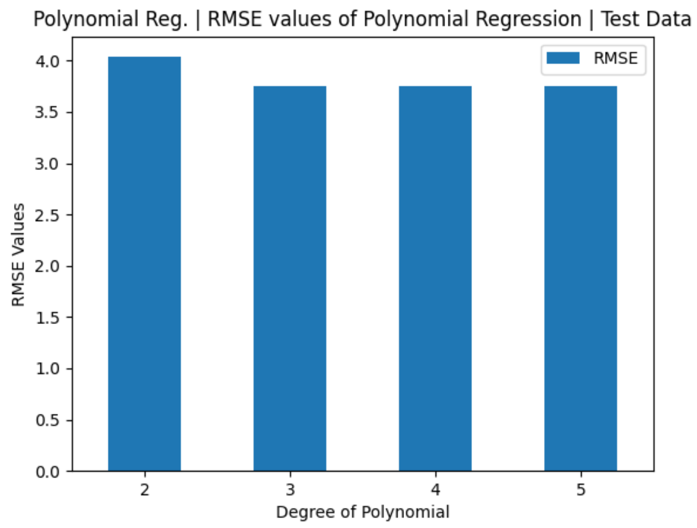

Figure 8 RMSE vs. different values of degree of polynomial (p = 2, 3, 4, 5) on the test data

Inferences:

1. RMSE value decreases with respect to increase in degree of polynomial (p = 2, 3, 4, 5).
2. After a p=3 the decrease becomes gradual.
3. As the Degree of Polynomial increases, the RMSE is decreasing from 4 to ~3.74, after p=4 the difference between consecutive RMSE value is less than ~0.001 so we can decrease becomes gradual.
4. From the RMSE value, p=5 will approximate the data best.

c.

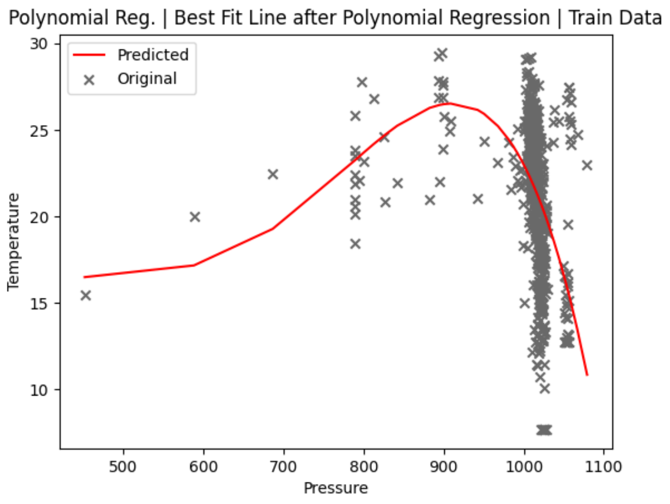

Figure 9 Pressure vs. temperature best fit curve using best fit model on the training data

Inferences:

1. Degree of polynomial with best fit line for training data = 4
2. Due to minimum value of RMSE we choose best fit line with degree = 4.

d.

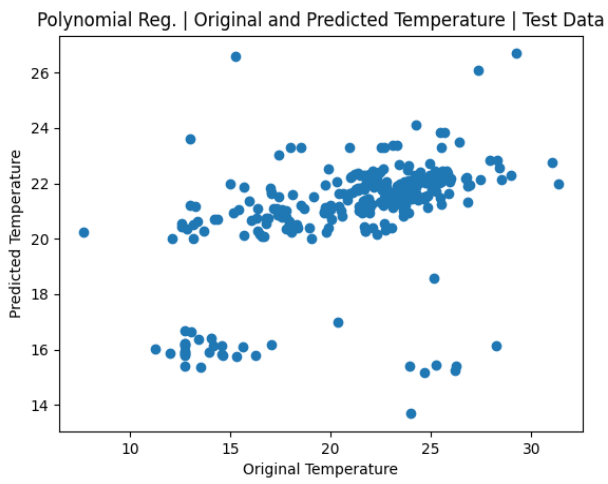

Figure 10 Scatter plot of predicted temperature from non- linear regression model vs. actual temperature on test data

Inferences:

1. Unlike linear regression model, this is more accurate as the variance is lower.
2. We can tell variance is lower as we are getting more accurate result for lower x axis.
3. The spread, density of data is high in linear regression model while it's lower in non linear regression.