IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

PERFORMANCE INVESTIGATION OF GENERATIVE MODELS FOR CLASSIFICATION OF ALCOHOLS

Kumar Shashvat [*], Raman Chadha, Arshpreet Kaur
M.Tech Student,CGC Technical Campus, Jhanjeri, Mohali
Professor, Head(CSE),CGC Technical Campus,Jhanjeri, Mohali
M.Tech Student,CGC Technical Campus, Jhanjeri, Mohali

DOI: 10.5281/zenodo.54675

ABSTRACT

Classification is the process related to categorization, the process in which ideas and objects are understood. It helps in clear identification of species for classification of various chemical compounds like Alcohol, Wine various discriminative approaches have been used .Discriminative methods offer good predictive performance and have been widely used in many applications but are unable to make efficient use of the unlabelled information. In such scenarios generative approaches have better applicability, as they are able to knob problems, such as in scenarios where variability in the range of possible input vectors is enormous. Generative models are integrated in machine learning for either modeling data directly or as a transitional step to form an uncertain probability density function. In this paper the generative models like Linear Discriminant Analysis and Naive Bayes have been used for classification of the alcohols. Linear Discriminant Analysis is a method used in data classification, pattern recognition and machine learning to discover a linear combination of features that characterizes or divides two or more classes of objects or procedures. The Naive Bayes algorithm is a classification algorithm base on Bayes rule and a set of conditional independence supposition. Naive Bayes classifiers are highly scalable, requiring a number of constraints linear in the number of variables (features/predictors) in a learning predicament. The main advantages of using the generative models are usually a Generative Models make stronger assumptions about the data, specifically, about the distribution of predictors given the response variables. The experimental results have been evaluated in the form of the performance measures i.e. are accuracy, precision and recall. The experimental results have proven that the overall performance of the Linear Discriminanat Analysis was better in comparison to the Naive Bayes Classifier on alcohol dataset.

Keywords: Classification, Generative Models, Naive Bayes, Linear Discriminant Analysis

INTRODUCTION

1.1 Classification

Classification is a common process related to categorization, the process in which ideas and objects are recognized, differentiated, and understood. Classification has essential part to play especially in assisting in the search process. By classifying things into different sections it enables us to retrieve things or information that we needed to look for, without the peril of too much time consuming in retrieving that particular things or information. When we want to search about particular possessions in school library, for example we want to look for any information about 'bear', we automatically will look for animal section to find these information. It is only because we know that bear is classified as an animal and that the provisions animal is a broad generalization of classification of bear. Animal can also be divided into subgroup such as herbivore, carnivore, mammal, reptile, and omnivore[3]. This is an example of animal hierarchical classification. We can rupture this classification into a more detail information like, carnivore there is lion, tiger, bear, puma and so on.

1.1 Need and Importance of Classification

- It helps in the clear identification of species by scientists.
- It also helps in the general study, observation and the organization of all rigorous conservation efforts to preserve the different species that exist in our biodiversity.

- It is a very important way of differentiating and recognizing different types of organisms, making important scientific and biological predictions as regards organisms of the same type, classifying how the different types of organisms relate with one another and providing specific names for each of the organisms.

1.2 Generative Models

The generative models are used in the machine learning for either modeling the data directly or as the intermediate step to forming a conditional probability density function. A conditional distribution can be formed from a generative model through Bayes rule [2]. Generative models contrast with discriminative models, in that a generative model is a full probabilistic model of all variables, whereas a discriminative model provides a model only for the target variable(s) conditional on the observed variables. Thus a generative model can be used, for example, to simulate (i.e. generate) values of any variable in the model, whereas a discriminative model allows only sampling of the target variables conditional on the observed quantities.

1.2.1 Types of Generative Model

- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naive Bayes
- Gaussian Discriminant Analysis
- Gaussian Mixture Model
- Hidden Markov Model

1.2.2 Applications of Generative Models

- **Operations research:-** It is an interdisciplinary branch of applied mathematics and formal science that uses methods such as mathematical modeling, statistics, and algorithms to arrive at optimal or near optimal solutions to complex problems [1].
- **Population ecology:-** It is a sub-field of ecology that deals with the dynamics of species populations and how these populations interact with the environment [1].
- **Psychometric:-** It is the theory and technique of educational and psychological measurement of knowledge, abilities, attitudes, and personality traits.
- **Quality control:-** It reviews the factors involved in manufacturing and production; it can make use of statistical sampling of product items to aid decisions in process control or in accepting deliveries.

METHODOLOGY

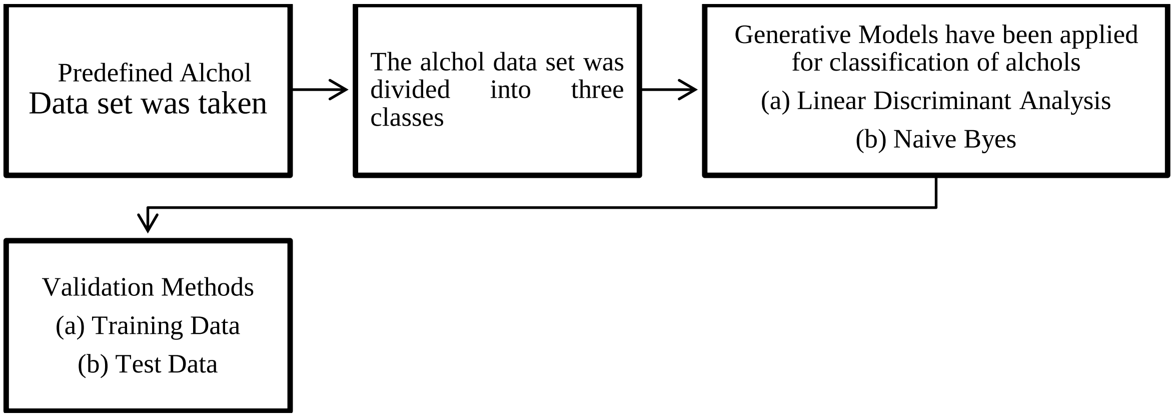
Predefined Alchol Data set was taken
The alchol data set was divided into three classes
Generative Models have been applied for classification of alchols
(a) Linear Discriminant Analysis
(b) Naive Byes
Validation Methods
(a) Training Data
(b) Test Data

Figure 1 Methodlogy

- **Predefined alcohol data set was taken**

The predefined alcohol data set was taken from UCI machine Learning[4].

- **The alcohol dataset was taken divided into three classes**

There are total of 76 samples of alcohol data set which was divided into three classes i.e. Kirsch, Mirab, and Piore

Table 1.1 Classification of Alcohols

Kirsch (0)	Mirab (1)	Piore (2)
19	28	29

The Features for these classes are shown in table below:

Table 1.2 Features of Alcohols

Feature 1	MEOH
Feature 2	ACET
Feature 3	BU1
Feature 4	MEPR
Feature 5	ACAL

Various generative models have been applied for classification

We have implemented the two generative models
(a) Linear Discriminant Analysis
(b) Naive Bayes

(a) Linear Discriminant Analysis

Linear Discriminant Analysis is a method used in, pattern recognition, Data Classification and machine learning to find a linear combination of features that exemplify or separates two or more classes of objects or events [2]. LDA is also closely related to principal component analysis (PCA) and factor analysis in that they both look for linear combinations of variables which best explain the data. LDA explicitly attempts to model the difference between the classes of data. PCA on the other hand does not take into account any disparity in class, and factor analysis builds the feature combinations based on differences rather than similarities.

Mathematical formulation for Linear Discriminnat Analysis

The Linear Discriminant Analysis (LDA) can be derived from simple probabilistic models which model the class conditional distribution of the data $P(x|y = k)$ for each class K. Predictions can then be obtained by using Bayes' rule:

$$p(y = k|x) = \frac{p(x|y = k)\, P(y=k)}{p(x)} \qquad \text{eq (1)}$$

To use this model as a classifier, we just need to estimate from the training data the class priors $P(y = k)$ (by the proportion of instances of class k), the class means μ_k (by the empirical sample class means) and the covariance matrices

$$p(x|y = k) = \frac{1}{(2\pi)^{n/} |\Sigma_k|^{1/2}} \exp(- \frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \qquad \text{eq (2)}$$

- **Score Function for Linear Discriminant Analysis**

The score function for Linear Discriminant Analysis

$$S(\beta) = \frac{\beta^T\mu_1 - \beta^T\mu_2}{\beta^T C\beta}$$

(b) Naive Baye's

Bayesian classifiers use Bayes theorem, which says

$$p(cj \mid d) = \frac{p(d \mid cj)\, p(cj)}{p(d)} \qquad \text{eq. (3)}$$

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}} \qquad \text{eq. (4)}$$

• $p(cj|d)$ = probability of instance d being in class cj, this is what we are trying to compute[2].
• $p(d|cj)$ = probability of generating instance d given class cj, We can imagine that being in class cj , causes you to have feature d with some probability.
• $p(cj)$ = probability of occurrence of class cj, This is just how frequent the class cj , is in our database.

• $p(d)$ = probability of instance d occurring this can actually be ignored, since it is the same for all classes.

Let μ_c be the mean of the values in x associated with class c, and let σ_c^2 be the variance of the values in x associated with class c [1]. Then, the probability distribution of $\mathcal{V}$ given a class c,p(x = v|c), can be computed by plugging v into the equation for a Normal distribution parameterized by μ_c and σ_c^2. That is,

$$P(x = v|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e - \frac{(v-\mu c)^2}{2\sigma_c^2} \qquad \text{eq (5)}$$

- **Validation Methods**

Method validation is the process used to confirm that the analytical procedure employed for a specific test is suitable for its intended use.

The validation methods which are used on the data set are as follows:

- Training Data (70% of the data set was used for training).
- Test Data (30% of the data set was used for test).

RESULTS AND DISCUSSION

Results

After obtaining the data the generative models have been applied i.e. Linear Discriminant Analysis and Naive Bayes

3.1 Results with Linear Discriminant Analysis

The Figure 1.1 shows the classification of class 0(Kirsch) and 1(Mirab) with respect to the feature first (MEOH) and second feature (ACET).

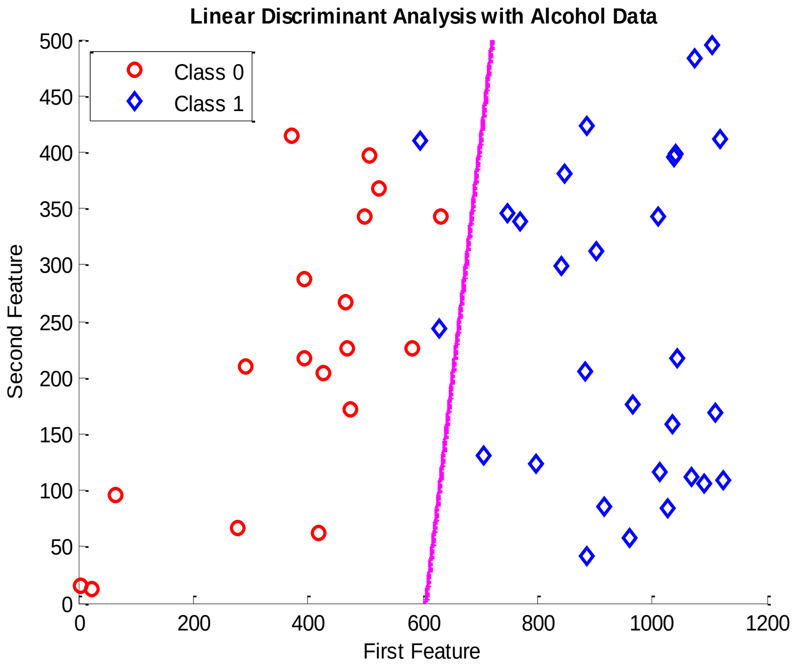

Figure 1.1 LDA with class 0 and 1

The Figure 1.2 shows the classification of class 0 (Kirsch) and 2 (Piore) with respect to the first and second feature.

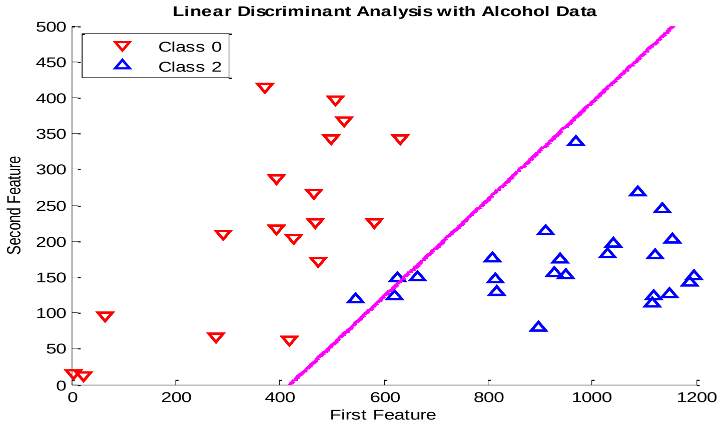

Figure 1.2 LDA with Class 0 and 2

The Figure 1.3 shows the Classification between class 1 (Mirab) and class 2 (Poire) with respect to the First and Second feature

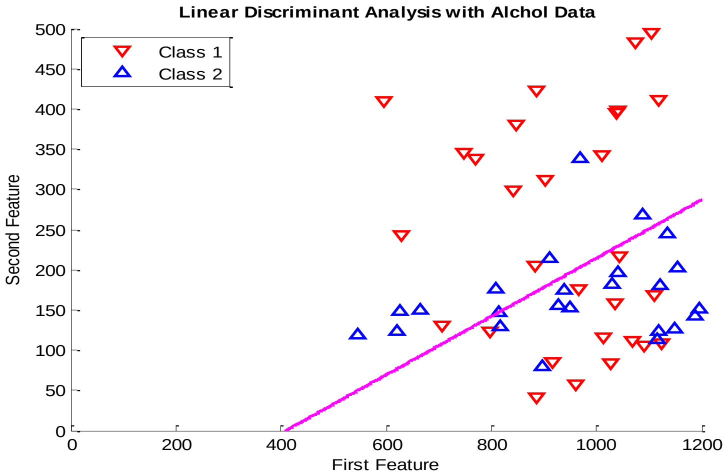

Figure 1.3 LDA with class 1 and 2

3.2.1 Confusion Matrix

The confusion matrix is a table that is used to describe the performance of the classification model or classifier. The confusion matrix of the model is may be calculated as shown in the table:

Table 1.3Confusion Matrix

	Predicted Class(N)	Predicted Class (M)
Actual (N)	True Positive	False Negative
Actual (M)	False Positive	True Negative

3.2.2 Confusion Matrix for Linear Discriminant Analysis

Table 1.4 Sample Classification for class Kirsch and Mirab

Sample classification for class Kirsch and Mirab	Predicted Kirsch	Predicted Mirab
Actual Kirsch	18	0
Actual Mirab	2	27

Table 1.5 Classification for class Kirsch and Poire

Sample classification for class Kirsch and Poire	Predicted Kirsch	Predicted Poire
Actual Kirsch	18	0
Actual Poire	2	28

Table 1.6 Sample Classification for class Mirab and Poire

Sample classification for class Mirab and Poire	Predicted Mirab	Predicted Poire
Actual Mirab	16	13
Actual Poire	18	34

3.2.3 Performance Measures for the Models

- **Accuracy**

The accuracy is may be define as the closeness of a measured value to standard or known value. The accuracy is define as the

$$\text{Accuracy} = \frac{\text{Tp} + \text{Tn}}{\text{Tp} + \text{Tn} + \text{Fp} + \text{Fn}}$$

- **Precision**

The precision refers to the closeness of two or more measurements to each other. The precision (P) is defined as the number of true positives (Tp) over the number of true positives plus number of false positives (Fp) i.e.

$$\text{P} = \frac{\text{Tp}}{\text{Tp}+\text{Fp}}$$

- **Recall**

The recall refers to evaluate the classifier output quality. The recall (R) is defined as the number of true positives (Tp) over the number of true positives plus number of false negatives (Fn) i.e.

$$\text{R} = \frac{\text{Tp}}{\text{Tp} + \text{Fn}}$$

3.2.4 Performance Measures for the Linear Discriminant Analysis

Table 1.7 Performance Measures for Models

Classification	Accuracy	Precision	Recall
For class Kirsch and Mirab	95.7%	90%	100%
For Class Kirsch and Poire	95.8%	90%	100%
For Class Mirab and Poire	70.4%	47%	55.1%

3.3 Results with Naive Bayes Classifier

The Figure 1.4 shows the classification of class 0(Kirsch) and 1(Mirab) with respect to the feature first (MEOH) and second feature (ACET).

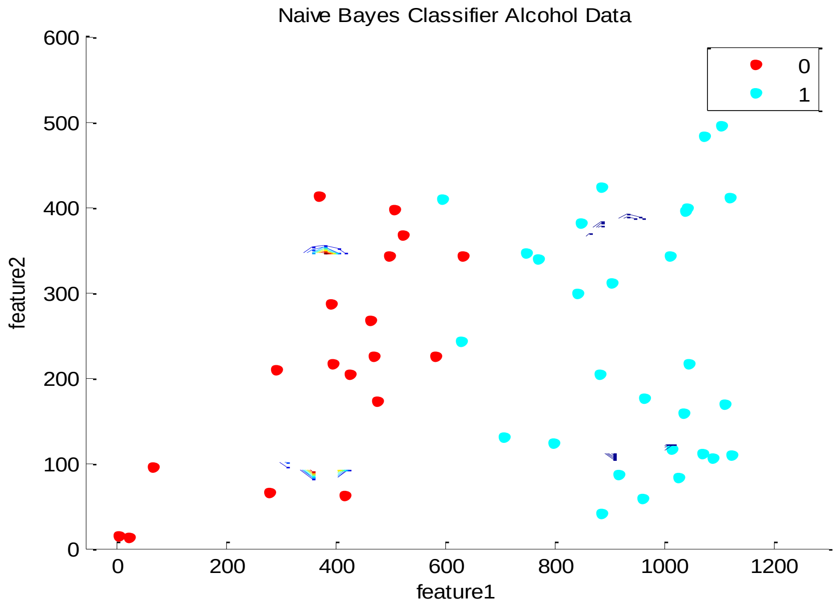

Figure 1.4 Naive Bayes classifier with class 0 and 1

The Figure 1.5 shows the classification of class 0(Kirsch) and 2 (Poire) with respect to the feature first (MEOH) and econd feature (ACET).

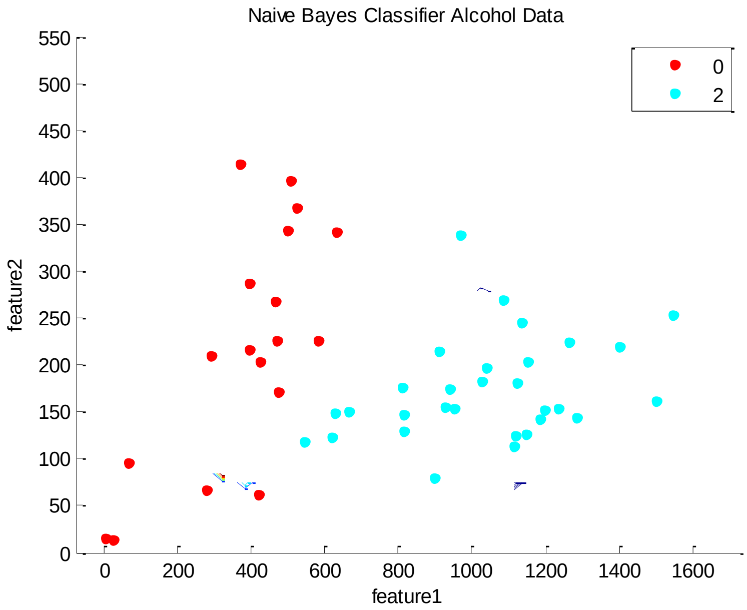

Figure 1.5 Naive Bayes classifier with class 0 and 2

The Figure 1.6 shows the classification of class 1 (Kirsch) and 2 (Poire) with respect to the feature first (MEOH) and second feature (ACET).

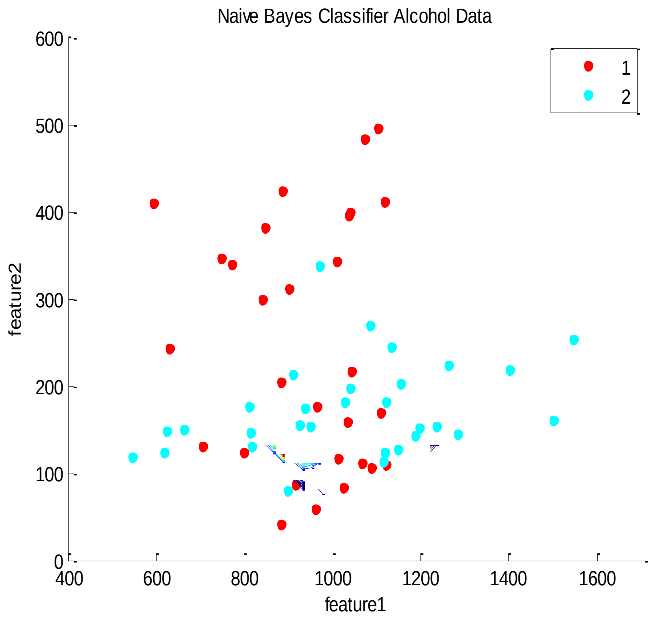

Figure 1.6 Naive Bayes classifier with class 0 and 2

3.3.1 Confusion Matrix for Naive Bayes Classifier

Table1.8 Sample Classification for class Krisch and Mirab(Naïve Bayes)

Sample Classification	Predicted Kirsch	Predicted Mirab
Actual Kirsch	5	1

Table 1.9 Sample Classification for class Mirab and Poire (Naive Bayes)

Sample Classification	Predicted Mirab	Predicted Poire
Actual Kirsch	4	1
Actual Poire	0	9

Table 1.9.1 Sample Classification for Class Krisch and Poire(Naïve Bayes)

Sample Classification	Predicted Kirsch	Predicted Mirab
Actual Mirab	4	8
Actual Poire	1	8

3.2.2 Performance Measures for the Linear Discriminant Analysis

Table 1 9.2 Performance Measures for Models

Classification	Accuracy	Precision	Recall
For class Kirsch and Mirab	85.7%	83.3%	83.3%
For Class Kirsch and Poire	92.8%	100%	80%

For Class Mirab and Poire	70.5%	80%	50%

3.2.3 Overall Performance of the Linear Discriminant Analysis and Naive Bayes

Table 1.9.3 Performance Measures for Models

Performance Measures	Linear Discriminant Analysis	Naive Bayes
Accuracy	87.3%	83%
Precision	75.6%	87.7%
Recall	85%	71.1%

3.2.4 Comparison of Linear Discriminant Analysis and Naive Bayes

The Figure 1.7 shows the comparison of Linear Discriminant Analysis and Naive Bayes

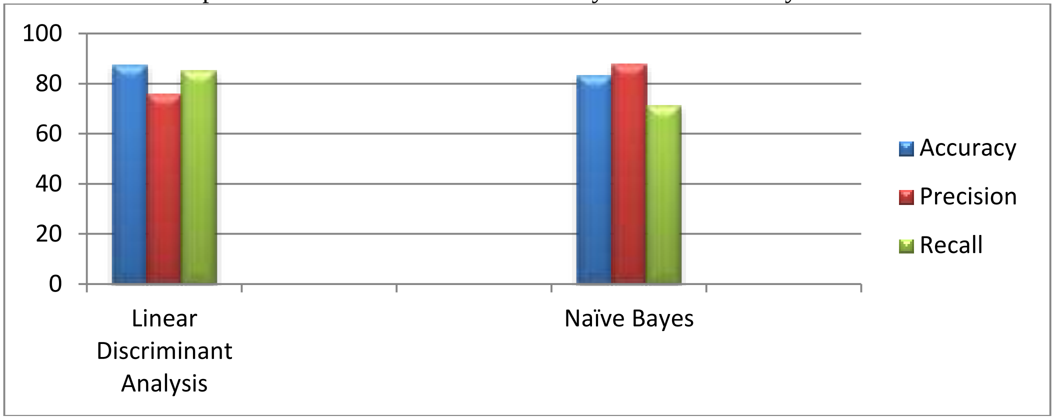

Figure 1.7 comparison of Linear Discriminant Analysis and Naive Bayes

CONCLUSION

Previously most of the classification work has been implemented with discriminative approaches. In this work two models generative models have been implemented i.e. Linear Discriminant Analysis and Naive Bayes. The results have been evaluated based upon the performance measures i.e. are Accuracy, Precision, and Recall. As the accuracy of Linear Discriminant Analysis is better as compared to Naive Bayes so Linear Discriminant Analysis shows better results on Alcohol dataset.

ACKNOWLEDGEMENTS

I wish to express my gratitude to all those individuals who have contributed their ideas, time and energy in this work. It is my privilege to thank Dr GD Bansal Director CGC Technical Campus. Lastly I wish gratitude to my family and friends for their support and encouragement.

REFERENCES

[1] Ranzato, M., Susskind, J., Mnih, V., & Hinton, G. (2011). On deep generative models with applications to recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2857–2864. http://doi.org/10.1109/CVPR.2011.5995710.

[2] Ng, A. (2008). CS229 Lecture notes 2 - Generative Learning algorithms, (0), 1–14.

[3] Science, C., & Engineering, S. (2014). Research on Data Mining Classification. International Journal of Advanced Research in Computer Science and Software Engineering, 4(4), 329–332.

[4] Citation, A., Donate, P., Contact, D. S., Learning, M., Systems, I., By, S., … Trajectories, G. P. S. (2016). View all Data Sets Welcome to the UC Irvine Machine Learning Repository.