ISSN: 2277-9655
Impact Factor:4.116

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

CLASSIFICATION OF WINES BASED ON QUALITY ASSESSMENT USING ITS CHEMICAL PROPERTIES WITH APPLICATION OF DISCRIMINANT ANALYSIS

Arshpreet Kaur*, Raman Chadha, Kumar Shashvat
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.54941

ABSTRACT

To know the quality of wine is utmost important for the producers,consumers and most people involved in this industry. Better is transparency in any procedure more appropriateness is added to it; similar is case with wine quality assessment procedure. Wine is characterized by its taste ,odor, flavor, aroma, mouthgood feel and after taste sensation it leaves with the taster. It is many a times perceived that costlier wines taste better , but it is just a mere perception and not in every respect true. The problem of quality assessment of wine through its taste is of consideration , as the procedure is complex as a whole; and various factors such as pricing, age of taster, color etc. do affect it making it inappropriate to be considered as true standard. On secondary basis the sensation of taste differs from person to person on the basis of sensitivity to a substance, origin of person and his genes. Number of taste buds one has also affect the taste sensation thus, a particular standard can not be set. Chemical properties of wine offer more stablity, and certain properties one of them being PH, which is responsible for the acidity in wine, and other such as sweetness on bases of study are found to collectively affect the taste of wine and hence, can be incorporated to predict the quality. Classifiaction techniques in machine learning provide ba scope to do. To learn how well these properties help in quality assessment procedure, linear discriminanat analysis has been applied on data set of wines produced in a exacting area of Portugal.

KEYWORDS:Classification, Discriminant Analysis, Wine Quality, PH, Residual Sugar

INTRODUCTION

Wine tasting is complex procedure. Taste is a chemical sagacity professed by specific receptor cells that make up taste buds[1][2]. Taste in general is influenced by many factors such as age of taster[10], pricing[3], previous experiences in context to taste, geographic origin, price, reputation, color,desease,adaptation etc [4]. Different research have shown that individuals appreciate the same wine more when they think that it is more expensive (Brochet, 2001; Plassmann et al., 2008)[5]. Human wine tasting is complicated procedure involving lots of steps and errors. It involves different steps after which the results are collaborated which are, In glass aroma of the wine ,"in mouth" sensations and finish (aftertaste); these properties afterwards these cooperatively establish the following properties of a wine: complexity and character, potential (suitability for aging or drinking) and possible faults[12]. Though to ensure impartial judgment of a wine, blind tasting procedure is also followed. Blind tasting might involve serving the wine from a black wine glass to mask the color of the wine. But it is believed that though it is considered to be one of better techniques it also has certain set of drawbacks such as ; it is considered that sighted reactions can divulge fascinating characterstics of the wine that will expected to evade us under conditions of blind tasting.

With intend of modern technology and machine learning capabilities; we intend to have a more upper base approach for classification of wine on basis of its quality. A wine can be marked as good , bad and average using its chemical properties which have a huge say in taste of wine through computerized methods. Though these are still evolving to reach high level of accuracy but with newer models and better formulations these models can be more reliable and

ISSN: 2277-9655
Impact Factor:4.116

accurate in near future. Many machine learning techniques have evolved up which are able to provide good classification and categorization when presented with certain set of feature. They stasticaly formulate and yield good result, continous improvements are making this field of great worth.

Different types of approaches which are available for classification are:-

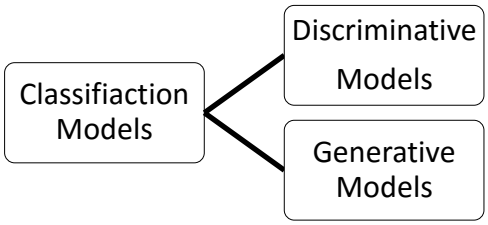

Figure 1Different Types of Models

Discriminative models :

- Linear Regression
- Logistic Regression
- Support Vector Machines
- Nueral Networks

Generative Models:

- Naïve Bayes
- Linear Discriminant Analysis
- Gaussian Mixture Models

Previously, using the same data various discriminative and data mining based approaches have been applied as in [12] and [14]. The intension of this work was to apply a generative model; Linear Discriminanat Analysis and evaluate its performance based on certain matrices.

The taste of wine is affected by various chemical properties it exibit. PH is one such important parameter which affect the overall taste of wine to a large extend; it not only plays an important role in overall flavor but gives its affect to every aspect of taste of a wine. Importance of PH is so because with right amount of acidity it is possible to easily bolt the flvour, aroma and give a pleasant color to wine. All these qualities have a important role to play in adding a feel good sensation to ones mouth. A perfect balance of acid level is utmost important to create a wine of good quality as low acid levels cause a wine to lack body and appropriate test. One of more significant parameter is sugar level in wine, Sugar is heart in wine making process and also plays an important role in giving a taste to wine. Sugar is not added to wine but comes from the natural sugars found in wine grapes that include fructose and glucose Perfect blend of acidity mixed with right amount of sugar are good to tickle your taste buds.

Linear Discriminant Analysis

It is a generative approach used in statstics and pattern recognition. It works with data that is previously classified into groups to derive rules for classifying new (and as yet unclassified) individuals on the basis of their observed variable values. This approach computes the sample mean of each class. Then it computes the sample covariance by first subtracting the sample mean of each class from the observations of that class, and taking the empirical covariance matrix of the result.The below figure represent discriminant analysis procedure[11].

ISSN: 2277-9655
Impact Factor:4.116

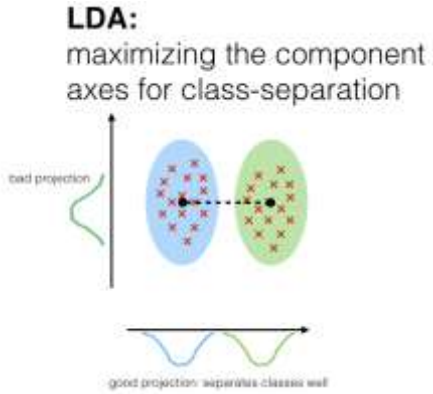

Figure 2 Linear Discriminant Analysis

Posterior Probability

$$\Pr(G = k \mid X = x) = \frac{f_k(x)\pi_k}{\sum_{l=1}^{K} f_l(x)\pi_l} \qquad \text{eq(1)}$$

π_k is the prior probability for class k
$f_k(x)$ is class conditional density or likelihood density

Multivariate Gaussian Density is given by the following equation:-

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^{-1}(x-\mu_k)\right) \qquad \text{eq(2)}$$

- The linear log-odds function above implies that the class k and l is linear in x; in p dimension a hyperplane.

Linear Discriminant Function

$$\delta_k(x) = x^T \Sigma^{-1}\mu_k - \frac{1}{2}\mu_k \Sigma^{-1}\mu_k + \log \pi_k \qquad \text{eq(3)}$$

Comparing two classes k and l , assume $\Sigma_k = \Sigma, \forall k$

$$\begin{aligned} \log \frac{\Pr(G = k \mid X = x)}{\Pr(G = l \mid X = x)} &= \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \\ &= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)\Sigma^{-1}(\mu_k - \mu_l) \\ &\quad + x^T \Sigma^{-1}(\mu_k - \mu_l) \end{aligned} \qquad \text{eq(4)}$$

ISSN: 2277-9655
Impact Factor:4.116

So we estimate $\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}$

$$\hat{\pi}_k = \frac{N_k}{N}, N_k \text{ is the number of Class k data} \quad \text{eq (5)}$$

$$\hat{\mu}_k = \sum_{g_i=k} x_i / N_k; \quad \text{eq(6)}$$

$$\hat{\Sigma} = \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K). \quad \text{eq(7)}$$

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k \Sigma^{-1} \mu_k + \log \pi_k$$

$$LDA \text{ rule}: \ g_k(x) = \text{var max}_l \{\delta_l(x)\}$$

$$Decision \text{ boundary } \{ x \mid \delta_k(x) = \delta_l(x) \}$$

MATERIALS AND METHODS

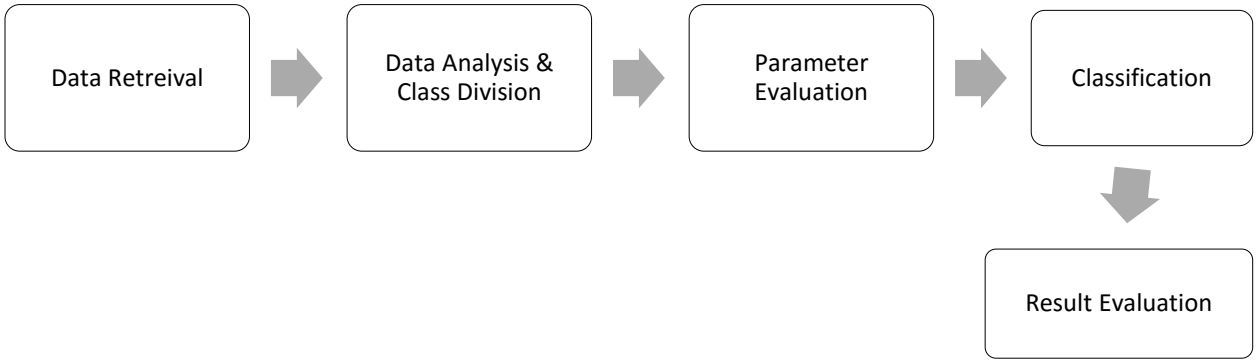

Figure 3 Methodology of Work

Data Retreival: Data about wines was taken from onlie website [7]. The data retrieved was all about white wine which is mainly fashioned ina particular area of Portugal . Their were different chemical properties on which analysis was conducted. To label quality human help was sought by data collectors; each wine was independenly tasted by a three autonomous tasters. Final taste was assigned by considering median of three rankings. Ranking was basically a depicture of quality of wine in context to the mouth good feel as felt by the tasters. The rank assigned was a number between 1-10 ; amongst which 1 was for worst taste and 10 for the best as explained in[7]and analysis procedure as given in [13].

ISSN: 2277-9655
Impact Factor:4.116

Data Analysis and Class divisoin: Data was analysed and quality rankings were studied. It was seen that ratings provided by wine tasters were mostly ranking of 3,4,5,6,7, 8 and 9 were used to represent the quality. Three clases were used to divide wine on bases of its quality ranking as done previously [7]. The three classes were Class Low which depicts poor quality wine, class Average which depict medium quality wine and class high which depict superior quality of wine. The statsicts of data were such that wine with ranking less than or equal to 5 were considered to be of poor quality. Rank 6 was taken to be of average quality and wine with more than or equal to rank 7 was taken to be a high quality wine as in consideration of study. Following graph represent the statistics of wine ranking which were median of ranking as given by tasters.

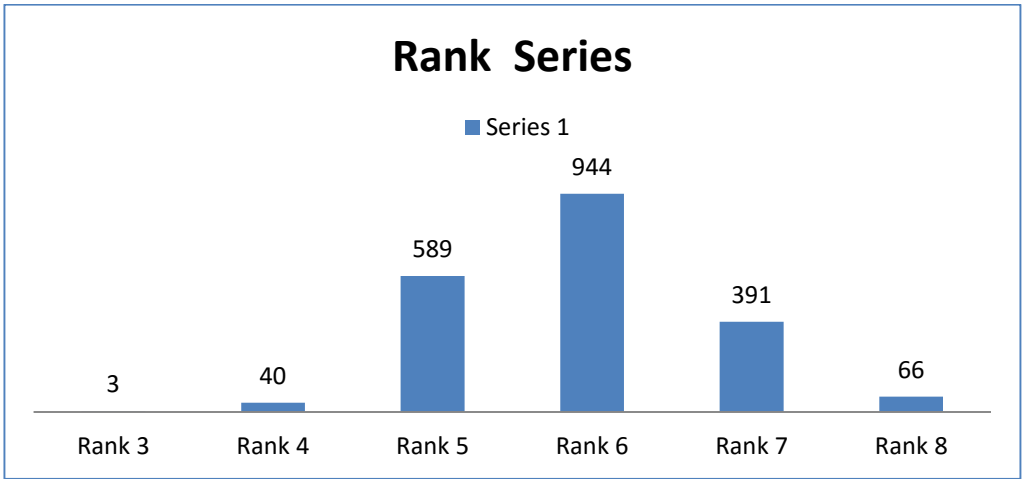

Figure 4: Rank Data by wine Tasters

Considering the above statsicts of data, classes were divided such as wine with ranking less than or equal to 5 were considered to be of poor quality. Rank 6 was taken to be of average quality and wine with more than or equal to rank 7 was taken to be a high quality wine as in consideration of study.Following graph represent the statistics of wine ranking.

Figure 5 Division of Samples

ISSN: 2277-9655
Impact Factor:4.116

Following are the details of number of samples that fell into each class:

Table 1. Division of Samples into Classes

Classes	Rank Division	No of Samples
Class low	Rank(1-5)	633
Class Medium	Rank (6)	944
Class High	Rank(7-10)	761

Parameter Evaluation : Out of tweleve features which are basically chemical physical properties of wine, two specific features were choosen to evaluate which were Ph and Residual sugar. Firstly, because PH of sample greatly influences the taste. Residual sugar reason being that Ph and residual sugar together affect the taste of wine.

Classification : The wines divided into categories were classified using discriminanat analysis, which is a generative approach. The classification procedure enables one to classify a unknown sample into a particular category using its chemical properties.

Result Evaluation : Results were evaluated, on basis of performance parameters such as accuracy, error rate, precesion and recall.

RESULTS AND DISCUSSION

Data was divided into training data and test data according to the validation applied. The validation factor was choosen in accordance to divide the data such that 70% data was taken as training data and 30% data approximately. as test data in k folds . The validation procedure allows data to be choosen randomly k number of times. Where 'k' is number of folds and division is also in according to value of k. This allows training data to be shuffled k times thus allowing a better performance evaluation.

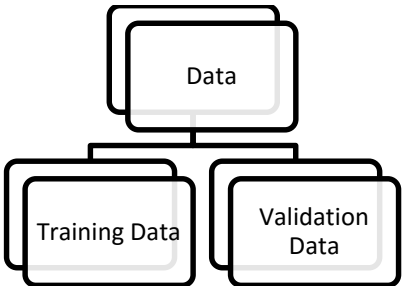

Figure 6 Data Division forClassifiacation

ISSN: 2277-9655
Impact Factor:4.116

Table 2. Division of Samples into Classes

	Total Samples	Training Data	Test Data	No of Folds
Total No of Samples	2037			
Class low and Medium	1576	1052 (66.75%)	524 (33.25%)	3
Class Medium and High	1405	937 (66.69%)	468 (33.31%)	3
Class High and Low	1093	729 (66.39%)	364 (33.61%)	3

Performance Parameter:

Confusion Matrix

The confusion matrix is also known to be as error matrix. It helps in visualization and calculation of performance of a algorithm over a set of given data. It is representation of how well a particular algorithm performed over given data. Diagonal elements depict number of elements which were predicted correctly; while off diagonal elements show the wrongly predicted elements. It is a method through which classification results and ground truth is compared and contrasted.

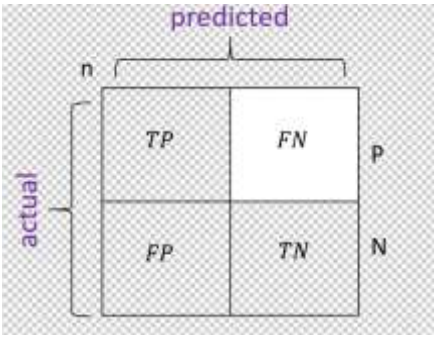

Figure 7: Confusion Matrix

Performance Measures for the Models

- **Accuracy**
  It is defined as portion of correctly classified values compared to the grounf thruth.Formula for its mathematical computation is given below:
  $$\text{Accuracy} = \frac{\text{Tp} + \text{Tn}}{\text{Tp} + \text{Tn} + \text{Fp} + \text{Fn}}$$
- Error rate:

It depicts the misclassified elements in comparison to the ground truth. The mathematical formulation is given below:

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

- **Precision**
  It is also known as positive predictive value. It is fraction retrieved instances that are relevant [8].

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

- **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.
  $$\text{R} = \frac{\text{Tp}}{\text{Tp} + \text{Fn}}$$

ISSN: 2277-9655
Impact Factor:4.116

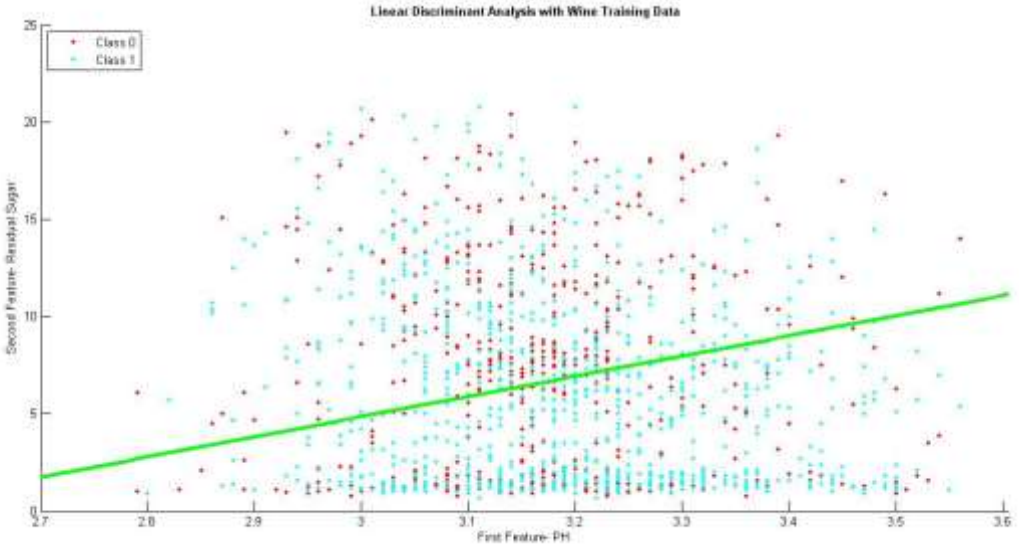

Figure 8 LDA with class 0 and 1

Table 1. Sample Classification for Class low and Medium(1576 Samples)

Sample Classification	Predicted low	Predicted Medium
Actual low	446	186
Actual Medium	381	563

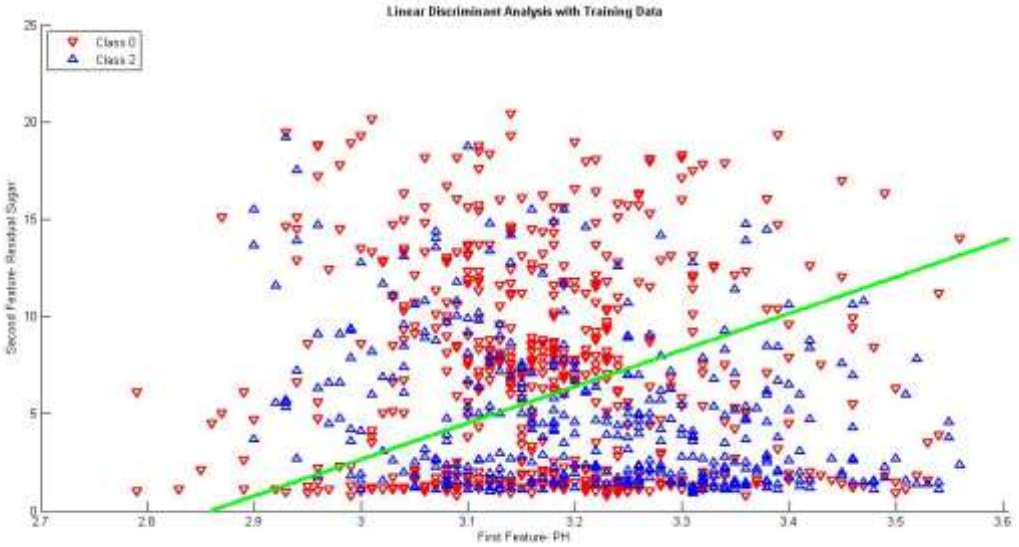

Figure 9 LDA with class 1 and 2

ISSN: 2277-9655
Impact Factor:4.116

Table 1.9.2 Sample Classification for Class low and Medium (1405 Smples)

Sample Classification	Predicted low	Predicted Medium
Actual Medium	602	342
Actual High	148	313

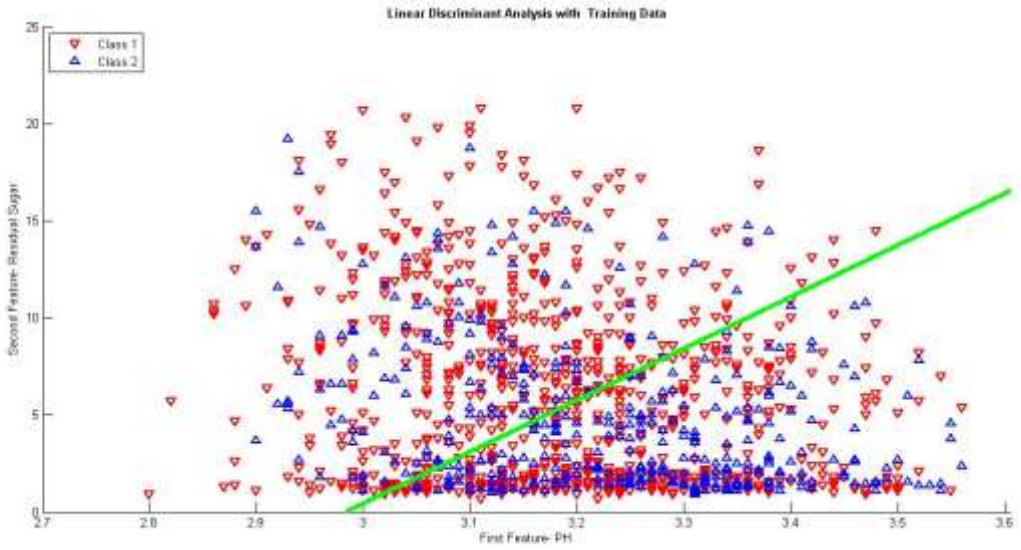

Figure 10 LDA with class 0 and 2

Table 1.9.2 Sample Classification for Class low and Medium(1093 Samples)

Sample Classification	Predicted low	Predicted High
Actual low	377	255
Actual High	211	250

The table below describes performance evaluation of discriminant analysis over classes.

Table1 9.3 Performance Measures for Linear Discriminant Model for different classes

Classification	Accuracy	Error Rate	Precision	Recall
For class low and Medium	64.02%	35.97%	53.92%	70.539%
For Class Medium and high	65.1245 %	34.44%	80.266%	63.77%
For Class High and low	57.36%	42.63%	64.11%	59.65%
Average	62.17%	37.68%	66.09%	64.65%

ISSN: 2277-9655
Impact Factor:4.116

Figure 11 : Graphical Representation of Performance Measures

The graphical represent how the algorithm performs on evaluation and categorization of new sample data.

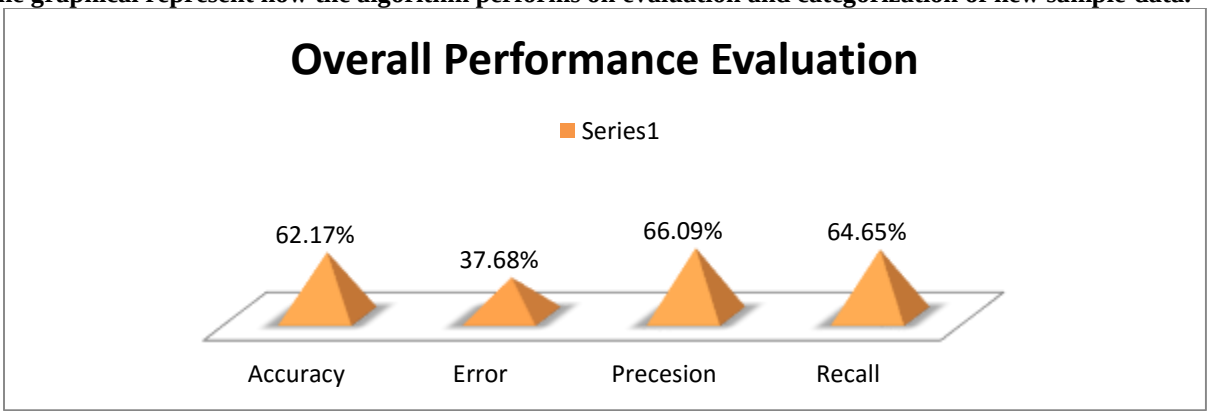

Figure 12 Graphical Representation of overall Performance

CONCLUSION

It is seen that PH and sugar have quite a influence on taste of wines. They can be used as a criteria to classify wines and and as a prediction base , though there is scope of improvement; as data can be better analysed and more generative can be applied to get better results. Though linear discriminant analysis has shown better results from the previously implemented models which were discriminative approaches.

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. Most importantly, I thank my parents, family specialy my aunt and grandmom for their immense love and true support at every point.

REFERENCES

[1] Anderson, E.N. *Everyone Eats: Understanding Food and Culture*. New York and London: New York University Press, 2005.
[2] Taste Preception, 25-04-16,http://health.howstuffworks.com/mental-health/human-nature/perception/taste
[3] Personal, M., Archive, R., Pashchenko, S., & Porapakkarm, P. (2012). Mp r a, (41193). doi:10.5897/JAERD12.088
[4] Services, S. (1900). Factors Influencing Taste Perception, 1–2.
[5] Brochet, F (2001) Chemical Object Representation In The Field of Consciousness. Working paper, General Oenology Laboratory, France.
[6] Linear Discriminant Analysis, 28-04-16, http://sebastianraschka.com/Articles/2014_python_lda.html
[7] Wine Data,24-05-16, https://onlinecourses.science.psu.edu/stat857/node/229
[8] Precesion and Recaall,21-05-16,https://en.wikipedia.org/wiki/Precision_and_recall
[9] Landon, S and C E Smith (1997) The Use of Quality and Reputation Indicators by Consumers: The Case of Bordeaux Wine. Journal of Consumer Policy 20: 289-323.
[10] Cohen T, Gitman L. Oral complaints and taste perception in the aged. J Gerontol.1959;14:294-298.
[11] Linear Discriminant Analysis, 28-04-16, http://sebastianraschka.com/Articles/2014_python_lda.html
[12] Wine Tasting,20-04-16, https://en.wikipedia.org/wiki/Wine_tasting
[13] Teixeira, J. (n.d.). Using Data Mining for Wine Quality Assessment
[14] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In *Decision Support Systems*, Elsevier, 47(4):547-553, 2009.