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 stability, 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 a scope to do. To learn how well these properties help in quality assessment procedure, linear discriminant 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, disease, 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 characteristics 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
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 statistically formulate and yield good result, continous improvements are making this field of great worth.

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

![Different Types of Models](http://www.ijesrt.com)

**Discriminative models**:
- Linear Regression
- Logistic Regression
- Support Vector Machines
- Neural 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 Discriminant 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 flavour, 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 statistics 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].
Posterior Probability

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

$\pi_k$ 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:-

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

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

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

$$\log \frac{\Pr(G = k | X = x)}{\Pr(G = l | X = x)} = \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l}$$

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

$$+ x^T \Sigma^{-1} (\mu_k - \mu_l)$$  \hspace{1cm} \text{eq}(4)$$
So we estimate $\hat{\pi}_k, \hat{\mu}_k, \hat{\Sigma}$

\begin{align*}
\hat{\pi}_k &= \frac{N_k}{N}, N_k \text{ is the number of Class } k \text{ data} \\
\hat{\mu}_k &= \frac{\sum_{g_i=k} x_i}{N_k}; \\
\hat{\Sigma} &= \sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T / (N - K). \\
\delta_k(x) &= x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k.
\end{align*}

\textit{LDA rule:} \quad g_k(x) = \max_i \{\delta_i(x)\}

\textit{Decision boundary} \quad \{x \mid \delta_k(x) = \delta_i(x)\}

**MATERIALS AND METHODS**

Data Retrieval \rightarrow Data Analysis & Class Division \rightarrow Parameter Evaluation \rightarrow Classification \rightarrow Result Evaluation

\textit{Data Retrieval:} Data about wines was taken from online website [7]. The data retrieved was all about white wine which is mainly fashioned in a 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 independently tasted by a three autonomous tasters. Final taste was assigned by considering median of three rankings. Ranking was basically a depiction 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].
Data Analysis and Class division: 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 classes 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 ofstudy. Following graph represent the statistics of wine ranking which were median of ranking as given by tasters.

![Figure 4: Rank Data by wine Tasters](image)

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

<table>
<thead>
<tr>
<th>Classes</th>
<th>Rank Division</th>
<th>No of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class low</td>
<td>Rank(1-5)</td>
<td>633</td>
</tr>
<tr>
<td>Class Medium</td>
<td>Rank (6)</td>
<td>944</td>
</tr>
<tr>
<td>Class High</td>
<td>Rank(7-10)</td>
<td>761</td>
</tr>
</tbody>
</table>

**Parameter Evaluation**: Out of twelve features which are basically chemical physical properties of wine, two specific features were chosen 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 discriminant analysis, which is a generative approach. The classification procedure enables one to classify an 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, precision and recall.

**RESULTS AND DISCUSSION**

Data was divided into training data and test data according to the validation applied. The validation factor was chosen 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 chosen 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.
Table 2. Division of Samples into Classes

<table>
<thead>
<tr>
<th></th>
<th>Total Samples</th>
<th>Training Data</th>
<th>Test Data</th>
<th>No of Folds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total No of Samples</td>
<td>2037</td>
<td>1576 (76.75%)</td>
<td>1052 (66.75%)</td>
<td>3</td>
</tr>
<tr>
<td>Class Low and Medium</td>
<td>1405</td>
<td>937 (66.69%)</td>
<td>468 (33.31%)</td>
<td>3</td>
</tr>
<tr>
<td>Class Medium and High</td>
<td>1093</td>
<td>729 (66.39%)</td>
<td>364 (33.61%)</td>
<td>3</td>
</tr>
</tbody>
</table>

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.

![Confusion Matrix](image)

Figure 7: Confusion Matrix

Performance Measures for the Models

- **Accuracy**
  It is defined as portion of correctly classified values compared to the ground truth. Formula for its mathematical computation is given below:
  \[
  \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}
  \]

- **Error rate:**
  It depicts the misclassified elements in comparison to the ground truth. The mathematical formulation is given below:
  \[
  E = \frac{FN + FP}{TP + TN + FP + FN}
  \]

- **Precision**
  It is also known as positive predictive value. It is fraction retrieved instances that are relevant [8].
  \[
  P = \frac{TP}{TP + FP}
  \]

- **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.
  \[
  R = \frac{TP}{TP + FN}
  \]
Figure 8 LDA with class 0 and 1

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

<table>
<thead>
<tr>
<th>Sample Classification</th>
<th>Predicted low</th>
<th>Predicted Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>446</td>
<td>186</td>
</tr>
<tr>
<td>Actual Medium</td>
<td>381</td>
<td>563</td>
</tr>
</tbody>
</table>

Figure 9 LDA with class 1 and 2
Table 1.9.2 Sample Classification for Class low and Medium (1405 Samples)

<table>
<thead>
<tr>
<th>Sample Classification</th>
<th>Predicted low</th>
<th>Predicted Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Medium</td>
<td>602</td>
<td>342</td>
</tr>
<tr>
<td>Actual High</td>
<td>148</td>
<td>313</td>
</tr>
</tbody>
</table>

Figure 10 LDA with class 0 and 2

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

<table>
<thead>
<tr>
<th>Sample Classification</th>
<th>Predicted low</th>
<th>Predicted High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual low</td>
<td>377</td>
<td>255</td>
</tr>
<tr>
<td>Actual High</td>
<td>211</td>
<td>250</td>
</tr>
</tbody>
</table>

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

Table 1.9.3 Performance Measures for Linear Discriminant Model for different classes

<table>
<thead>
<tr>
<th>Classification</th>
<th>Accuracy</th>
<th>Error Rate</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>For class low and Medium</td>
<td>64.02%</td>
<td>35.97%</td>
<td>53.92%</td>
<td>70.539%</td>
</tr>
<tr>
<td>For Class Medium and high</td>
<td>65.1245%</td>
<td>34.44%</td>
<td>80.266%</td>
<td>63.77%</td>
</tr>
<tr>
<td>For Class High and low</td>
<td>57.36%</td>
<td>42.63%</td>
<td>64.11%</td>
<td>59.65%</td>
</tr>
<tr>
<td>Average</td>
<td>62.17%</td>
<td>37.68%</td>
<td>66.09%</td>
<td>64.65%</td>
</tr>
</tbody>
</table>
Figure 11: Graphical Representation of Performance Measures

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

CONCLUSION

It is seen that PH and sugar have quite an influence on taste of wines. They can be used as a criteria to classify wines 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 specially my aunt and grandmom for their immense love and true support at every point.

REFERENCES


