Wine Quality Prediction Using Data Mining

Shruthi P
Department of CSE
ATME College of Engineering
Mysuru, INDIA
shruthip.111292@gmail.com

***Abstract*— Certifying the quality of food product is the major concern of the country. The citizens of the country are recommended to use only quality assured products. The same thing need to be applied for the wine industry also. The quality of wine need to be assessed and it should be classified into different category based on the quality assessment. Data mining is the right approach to achieve this as it extracts the useful information by analyzing the data set. In this paper, the samples of different wines with their attributes required for quality assurance is collected and different data mining classification algorithms- Naive Bayes, Simple Logistic, KStar, JRip, J48 are applied on it. The wine will be classified into three main categories and the accuracy of the algorithms are compared.**

Keywords— data mining, classification, attributes

I. INTRODUCTION

The quality of wine is very important since consuming low quality wine will adversely effect the human health. The wine industry is huge sector where they will cheat people in quality of wine. The manual approach to identify the quality of wine and labeling them to different quality levels is time consuming and not accurate. There is a need of system which can classify the product to different quality level by considering its attributes like alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, Hue, OD280/OD315 of diluted wines, proline. By analysing the amount of these attributes present in the wine, it can be labelled into different quality levels. The best approach to this problem is the use of data mining classification algorithms.

TABLE I: LIST OF ATTRIBUTES

Sl No	Attribute Name	Value
1	alcohol	Real Number
2	malic acid	Real Number
3	ash	Real Number
4	alcalinity of ash	Real Number
5	magnesium	Real Number
6	total phenols	Real Number
7	flavanoids	Real Number
8	nonflavanoid phenols	Real Number
9	proanthocyanins	Real Number
10	color intensity	Real Number
11	Hue	Real Number
12	OD280/OD315 of diluted wines	Real Number
13	proline	Real Number
14	Class	1, 2, 3

In this paper the data set of 178 samples of wine with 13 attributes and the respective classes of quality (1, 2, 3) are considered. The data set is loaded in to weka 3.9. The data set is divided with the split ratio of 80%, where 80% will be considered as training set and 20% is considered as test set. Different classification algorithms such as Naive Bayes, Simple Logistic, KStar, JRip, J48 are applied on the training set to classify the test set. The accuracy of these algorithms are compared. The above Table I shows the different attribute and its value.

In the above attribute list, 13 attributes has a value as real numbers but the class attribute has a value restricted to 1, 2, 3. All test set is classified to the class 1, 2, 3, denoting 1 as the excellent quality, 2 as the average quality and 3 as the low quality wine.

II. LITRATURE REVIEW

Different machine learning algorithms such as linear regression, neural networks and support vector machine is used to predict the quality of wine in two phases- determining the dependency of target variable on independent variable and predicting the value of target variable. All experiments are performed on red and white wine and it is concluded that the better prediction is achieved by considering selected features instead of considering all features of wine[1]. The quality of good wine depends on scientific, chemical and technical factors. Principal Component Analysis (PCA) as well as Recursive Feature Elimination approach (RFE) was used for identifying the features and nonlinear decision tree based classifiers was used for analyzing the performance metrics. It was noticed that by using Random Forest classifier with different feature sets, the accuracies ranges from 94.51% to 97.79%. This results will help the wine experts to know the important factors to consider while selecting the good quality wine[2]. As the occurrence of events in the data set obtained from the Minho region in Portugal was imbalanced with about 93%, Synthetic Minority Over- Sampling Technique (SMOTE) was used to overcome this. The balanced data was used to model a classifier that categorize a wine into three categories as high quality, normal quality, and poor quality. Decision tree, adaptive boosting (AdaBoost), and random forest was used for prediction. The results shows that the random forest technique is more accurate with minimum errors. The Logistic regression and BP neural network and SVM classification algorithms are used to identify the modeling analysis of wine quality[4]. The database consisting of 32 wine characteristics applied to 180 wine samples with wine quality labels assigned by a wine expert is considered for classification. The good classifiers are identified and the optimal subset of features to maximize the performance of the best classifier and also minimize the overall cost of the measurements are searched. The result shows the best performing subset of tests for a given threshold cost[5].

III. METHODOLOGY

178 samples of wine with their attributes and class is loaded into weka 3.9 and the following classification algorithms are applies and accuracy is compared.

A. Naive Bayes Classifier

In data mining, naive Bayes classifiers belongs to the family of simple "probabilistic classifiers". This classifier works by applying Bayes theorem with strong independent assumptions between attributes.

Naive Bayes Classifier is applied on 178 samples with the split ration of 80%, which means 80% of samples are used as training set and 20% of samples are used as test set. The bellow Figure 1 shows the result inferred.

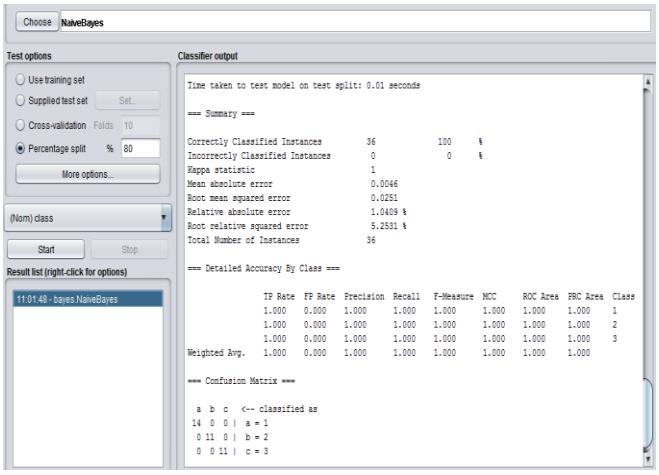

Figure 1: Naive Bayes Classifier

As shown in the above Figure 1, the percentage of split is 80%. In 178 samples, 142 (80% of 178) is used as traning set and 36 samples (20% of 178) is used as test set. The classifier is applied to classify 36 test set into respective classes. Correctly classified instances is 36, which shows 100% of accuracy and incorrectly classified instances is 0, which shows 0% of inaccuracy.

B. Simple Logistic Classifier

Simple logistic classifier is used where the response variable is categorical. The idea behind Logistic Regression is to identify a relationship between features and probability of particular outcome.

Simple Logistic Classifier is applied on 178 samples with the split ration of 80%, which means 80% of samples are used as training set and 20% of samples are used as test set. The bellow Figure 2 shows the result inferred.

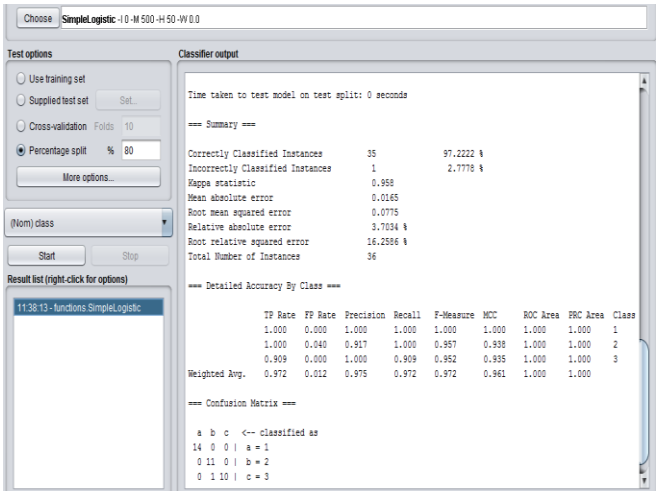

Figure 2: Simple Logistic Classifer

As shown in the above Figure 2, the percentage of split is 80%. In 178 samples, 142 (80% of 178) is used as traning set and 36 samples (20% of 178) is used as test set. The classifier is applied to classify 36 test set into respective classes. Correctly classified instances is 35, which shows 97.22% of accuracy and incorrectly classified instances is 1, which shows 2.77% of inaccuracy.

C. KStar Classifier

KStar is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function.

KStar Classifier is applied on 178 samples with the split ration of 80%, which means 80% of samples are used as training set and 20% of samples are used as test set. The bellow Figure 3 shows the result inferred.

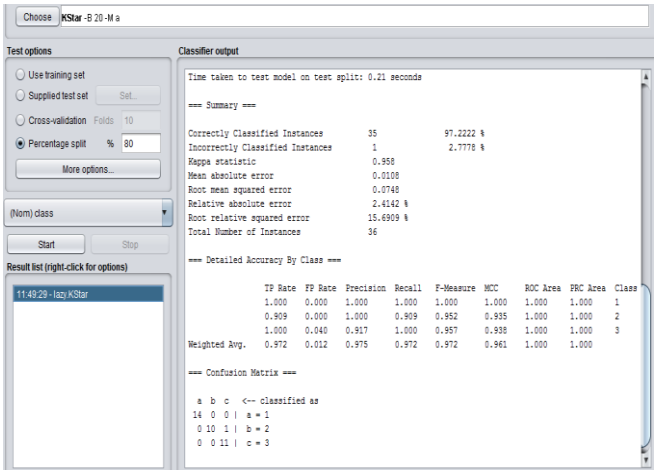

Figure 3: KStar Classifer

As shown in the above Figure 3, the percentage of split is 80%. In 178 samples, 142 (80% of 178) is used as traning set and 36 samples (20% of 178) is used as test set. The classifier is applied to classify 36 test set into respective classes. Correctly classified instances is 35, which shows 97.22% of accuracy and incorrectly classified instances is 1, which shows 2.77% of inaccuracy.

D. JRip Classifier

This classifier implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP.

JRip Classifier is applied on 178 samples with the split ration of 80%, which means 80% of samples are used as training set and 20% of samples are used as test set. The bellow Figure 4 shows the result inferred.

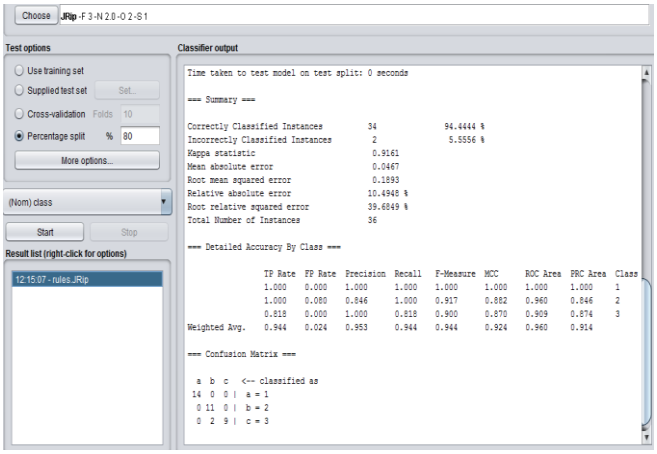

Figure 4: JRip Classifer

As shown in the above Figure 4, the percentage of split is 80%. In 178 samples, 142 (80% of 178) is used as traning set and 36 samples (20% of 178) is used as test set. The classifier is applied to classify 36 test set into respective classes. Correctly classified instances is 34, which shows 94.44% of accuracy and incorrectly classified instances is 2, which shows 5.55% of inaccuracy.

E. J48 Classifier

J48 classifier is a decission tree based classifier. Decision tree J48 is the implementation of algorithm ID3 (Iteractive Dichotomiser) developed by the WEKA project team.

J48 Classifier is applied on 178 samples with the split ration of 80%, which means 80% of samples are used as training set and 20% of samples are used as test set. The bellow Figure 5 shows the result inferred.

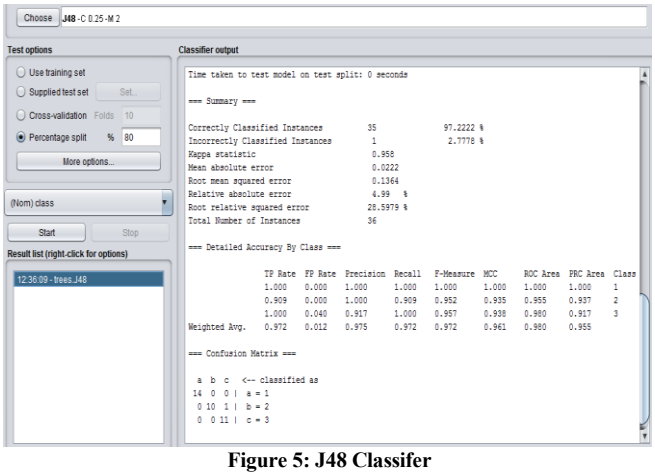

Figure 5: J48 Classifer

As shown in the above Figure 5, the percentage of split is 80%. In 178 samples, 142 (80% of 178) is used as traning set and 36 samples (20% of 178) is used as test set. The classifier is applied to classify 36 test set into respective classes. Correctly classified instances is 35, which shows 97.77% of accuracy and incorrectly classified instances is 1, which shows 2.77% of inaccuracy.

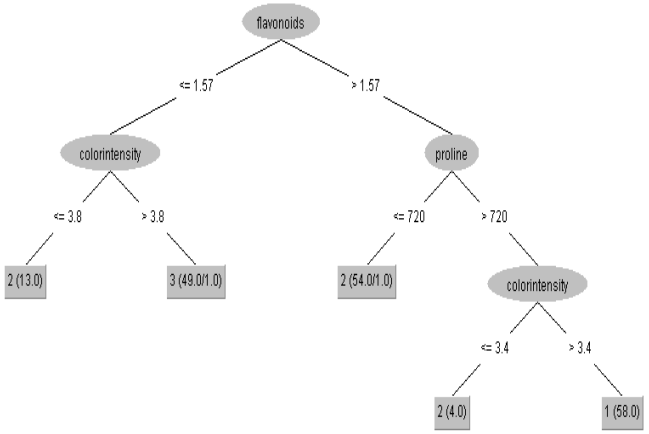

Figure 6: Tree Visualization of J48 Classifer

The above Figure 6 shows the tree structure bulit on executing J48 algorithm for classification. By traversing this tree the test samples will be classified into different classes.

IV. Result and Discussion

As discussed in methodology chapter, 178 samples of wine with their attributes and classes are loaded in weka 3.9. The data set is divided with the split ratio of 80% where 80% of data set is used as training set and 20% of data set is used as the test set. Naive Bayes, Simple Logistic, KStar, JRip, J48 classification algorithms are applied on the training set to classify the test set. The accuracy of these algorithms are shown in below Table II.

TABLE II: ACCURACY COMPARISON

Algorithm	Accuracy in %	Inaccuracy in %
Naive Bayes	100	0
Simple Logistic	97.22	2.77
KStar	97.22	2.77
JRip	94.44	5.55
J48	97.22	2.77

With this discussion, the accuracy of all five classification algorithms is very high and it can be used to classify the wines belonging to different quality. Among all the five algorithm it can be noted that Naive Bayes Classification algorithm has highest accuracy of 100%. It has classified all instances correctly.

The Naive Bayes classifier is the simplest and fastest classification algorithm. It handles continuous and discrete values to make probabilistic predictions. It is highly scalable and insensitive to irrelevant features. It predicts accurately with small training set also. With all the above advantages of Naive Bayes algorithm, it can predict more accurately. But if the training set becomes big, the better classifier can be chosen by repeatedly applying the data set to different classifiers using cross validation.

V. CONCLUSION

Different classification algorithms (Naive Bayes, Simple Logistic, KStar, JRip, J48) are applied on same data set of 178 wine samples, all the five algorithm's efficiency is good but Naive Bayes is more accurate classifier among all. These algorithms can be used to classify the wine to the respective quality levels. It helps the consumers of wine and reduces the number of fraud in wine industry. It also helps the quality labeling companies of government to issue quality certificate. It reduces the errors compared to manual quality assurance. In future cost of wine can also be predicted based on the quality assessment done using data mining techniques. Other classification and prediction algorithms can also be applied and studied in future.

REFERENCES

[1] Yogesh Gupta, "Selection of Important Features and Predicting Wine Quality using Machine Learning Techniques," Article in Procedia Computer Science January 2018, pp 305-312

[2] Satyabrata Aich, Ahmed Abdulhakim Al-Absi, Kueh Lee Hui, John Tark Lee and Mangal Sain, "A Classification Approach with Different Feature Sets to Predict the Quality of Different Types of Wine using Machine Learning Techniques", International Conference on Advanced Communications Technology (ICACT), pp 139-143

[3] Gongzhu Hu, Tan Xi, Faraz Mohammed, Huaikou Miao, "Classification of Wine Quality with Imbalanced Data", IEEE 2016, pp 1712-1717

[4] Zhang Lingfeng, Feng Feng, Huang Heng, "Wine quality identification based on data mining research", IEEE 2017, pp 358-361

[5] Răzvan Andonie, Anne M. Johansen, Amy L. Mumma, Holly C. Pinkart, Szilárd Vajda, "Cost efficient prediction of Cabernet Sauvignon wine quality", IEEE 2017

[6] Neelamadhab Padhy, Dr. Pragnyaban Mishra, Rasmita Panigrahi, "The Survey of Data Mining Applications And Feature Scope", IJCSEIT, June 2012, pp 43-58

[7] Amiya Kumar Sahu, "The Criticism of Data Mining Applications and Methodologies", IJARCS, 2016, pp 52-55

[8] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011", Elsevier, 2012

[9] Shikha Agrawal, Jitendra Agrawal, " Survey on Anomaly Detection using Data Mining Techniques", Elsevier, 2015, pp 708-713

[10] Mrs. Bharati M. Ramageri, "Data Mining Techniques and Applications", Indian Journal of Computer Science and Engineering

[11] Sukhdev Singh Ghuman, "A Review of Data Mining Techniques", IJCSMC, 2014, pp 1401-1406

[12] Mansi Gera, Shivani Goel, "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity", International Journal of Computer Applications, 2015, pp 22-29

[13] Ashish Kumar Dogra, TanujWala, "A Review Paper on Data Mining Techniques and Algorithms", IJARCET, 2015

[14] Abhijit Raorane, R.V.Kulkarni, "Data Mining Techniques: A Source for Consumer Behavior Analysis", IJDMS, 2011

[15] Hais Han Tian, Qiaohong Pang, "Data Mining Application for Upgrading Quality of Wine Production", IEEE, 2010, pp 109-111

[16] Subana Shanmuganathan and Philip Sallis, Ajit Narayanan, "Data Mining Techniques for Modelling Seasonal Climate Effects on Grapevine Yield and Wine Quality", IEEE, 2010, pp 84-89

[17] Angela Zinnai, Francesca Venturi, Gianpaolo Andrich, "Cold maceration in production of high quality wine", IEEE, 2006