Wine Quality Analysis Using Machine Learning Algorithms

Mahima, Ujjawal Gupta, Yatindra Patidar, Abhishek Agarwal and Kushall Pal Singh

Abstract Wines are being produced since thousands of years. But, it is a complex process to determine the relation between the subjective quality of a wine and its chemical composition. Industries use Product Quality Certification to promote their products and become concern for every individual who consumes any product. It is not possible to ensure quality with experts with such a huge demand of product as it will increase the cost. Wine-makers need a permanent solution to optimize the quality of their wine. This paper explores the space to easy out and make the whole process cost-effective and more trustworthy using machine learning. It allows to build a model with user interface which predicts the wine quality by selecting the important parameters of wine which play a significant role in determining the wines quality. Random forest algorithm is used in determining wines' quality whose correctness would further be escalated using KNN which makes our model dynamic. Output of this proposed model is used to determine the wines' quality on a scale of Good, Average or Bad. This proposed model can further be applied to several other products which need quality certification. Our prediction model provides ideal solution for the analysis of wine, which makes this whole process more efficient and cheaper with less human interaction.

Keywords KNN · CART algorithm · Random forest · Wine quality · Machine learning

Mahima (✉) · U. Gupta · Y. Patidar · A. Agarwal · K. P. Singh
SRM Institute of Science and Technology, NCR Delhi, Sonipat, India
e-mail: mahimasaini1997@gmail.com

U. Gupta
e-mail: audaciousujjawal@gmail.com

Y. Patidar
e-mail: yatindrapati@gmail.com

A. Agarwal
e-mail: a97agarwal@gmail.com

K. P. Singh
e-mail: kpal090@gmail.com

D. K. Sharma et al. (eds.), *Micro-Electronics and Telecommunication Engineering*, Lecture Notes in Networks and Systems 106,
https://doi.org/10.1007/978-981-15-2329-8_2

1 Introduction

Nowadays, demand of wine in market is growing day by day, in order to back up with the rise in demand by accepting new inventions. Along with producing wines, quality assurance certification is also a crucial issue for wine-makers. Currently, within wine industries, quality is estimated through physio-chemical data (e.g. PH levels) and sensory data (e.g. expert critics' involvement) [1]. Analytical data is provided through sensitivity analysis, i.e. the response measured when input variable is compared with its domain value. To produce distinct type and kind of wine producers, use diverse range of grapes and varieties of yeasts. To produce other than flavour, several other factors are also considered to improve quality of wine.

To analyse the quality of wine, a large dataset is taken which consists of huge variety of chemical and acidic aggregation of both red wine and white wine. By occupying smart business science techniques, we can discover the essential and exotic vision which could be productive for better-quality wine; this could be beneficial in economical, in financial and in business sectors in wine production companies [2]. Evolution of business strategies and values added were result shown by grade in excellence in refined model. Resulted produced product can be admiring if and only if it is produced with minimum cost along with maximum quality. The discovered model could be used independently either as wine quality prediction or as a replacement for human wine tasting appraisal by wine critics and could help in development of better wines by industries.

Section 2 is literature survey which describes our research. Section 3 is methodology which is going to tell about the whole process flow and data description which provides insights about the data. Section 4 is result analysis section which is going to tell about the advantages of using this method over the existing methods. Section 5 describes the conclusion and future scope of this model.

2 Literature Survey

2.1 Documentary Research

Linear regression is easy and simple to implement practically for making predictions in many fields. Using linear regression, the correlation between the attributes was determined. This helped in determining the important parameters with respect to quality [3]. After data analysis, it was found that alcohol shows maximum variation than other parameters. Higher the concentration of alcohol leads to better quality of wine and lowest density [4]. Two different machine learning techniques can be used to develop the prediction model, i.e. neural network and support vector machine. The two used is divided into two parts: red wine and white wine datasets. Both of them consist of 12 different physio-chemical characteristics [5].

While using KNN algorithm, it evaluates Strassen's matrix to calculate the maximum and minimum values of attributes that consist in dataset. K-nearest neighbour, random forests and support vector machines are evaluated on datasets. It shows precision to predict that wine quality can be improved to 90–92% from 75% [6]. How decision tree is formed from the dataset used and mean values are evaluated from 12 different attributes [7].

There are several machine learning algorithms which are analysed to distinguish the quality for both red wine and white wine such as k-nearest neighbour and random forests. The best fortunate to classify data should done using random forest algorithm, where the precision for prediction of good-quality wine is 96% and bad-quality wine is almost 100%, which give overall precisions around 96%. It also helps us to classify different parameters of wine with rating from 1 to 10 or good–bad. From the existing rating, 1–4 predicts bad quality, 5–6 gives average and 7–10 predicts good quality of wine [2].

2.2 Algorithm Analysis

It gives insights of the dependency of target variables on independent variables using machine learning techniques to determine the wine quality because it gives the best outcome for the assurance of quality of wine. The dependent variable is "quality rating", whereas other variables, i.e. alcohol, sulphur, etc., are assumed to be predictors or independent variables [6]. While hindering the effectiveness of the data model, various types of errors have occurred like over fitting, introduced from having too large of a training set and bias occur due to too small of a test set.

2.2.1 Random Forest

CART is a decision tree used for analysing both datasets (red and white wines). CART always generates binary decision trees, which consist of two branches, each for decision node. The tree grows by organizing data in each decision node, by splitting in all possible directions and selecting an optimal split. The decision tree supports a tool based on outcome that accesses a tree-like structure for making decisions and their desirable consequence, along with all the outcome chances, overall resource cost and efficiency. It is one of the ways for demonstration of algorithms which consist of provisional curb statement [4].

It is generally used for exploration operation, especially while analysing decision, helps to identify a strategy that reaches the desired goal and has proved to be an important tool in machine learning [7]. It is like a flowchart structure that consists of internal node which represents "Test" attribute (e.g. if we flip a coin, it comes out to be Heads or Tails), and every Branch represents conclusion for every tested data, as well as leaf node represents label of class (after computing, all parameters' decision is taken). The classification rule is represented from root to leaf.

Random forest is a method of classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [8].

Following are some of the features of random forest algorithm:

1. It runs efficiently on large databases.
2. It gives estimates of what variables are important in the classification.
3. It generates an internal unbiased estimate of generalization error as the forest building progresses occur. Random forest is similar to the decision tree method in that it builds trees—that is why known as "random forest" [7]. This is a learning method which creates a multitude of decision trees, and outputting the class that occurs most frequently among them and classify the output.

2.2.2 K-Nearest Neighbour

This classifier technique is depended on learning by analogy; this means a comparison between a test tuple with similar training tuples. The training tuples are described by n attributes. Each tuple corresponds a point in an n-dimensional space. All the training tuples are stocked in an n-dimensional pattern space. For an unknown tuple, a k-nearest neighbour classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. K training tuples are called as the k-nearest neighbours of the unknown tuple [2].

"Closeness" is a metric distance, likewise Euclidean distance between two points or tuples, say, $X_1 = x_{11},\ x_{12}, \ldots x_{1n}$ and $X_2 = x_{21},\ x_{22}, \ldots x_{2n}$ is:

$$\text{dist.}(X_1,\ X_2) = \sqrt{\sum_{i-1}^{n} (x_{1i} - x_{2i})^2}.$$

Standalone random forest algorithm is modelling the data with a RMSE of 0.6430 for white wine and 0.6322 for red wine. The proposed system is working in combination with KNN which reduces the RMSE of the above system. Quality formula is developed that connects random forests with KNN algorithm.

3 Methodology

This section gives insights the dependency of target variables on independent variables using machine learning techniques to determine the quality of wine because it gives the best outcome for the assurance of quality of wine [4]. The dependent variable is "quality rating," whereas other variables, i.e. alcohol, sulphur, etc., are assumed to be predictors or independent variables [4]. The analysis on these variables is done in two different ways:

1. Firstly, regression algorithms are used to depict the importance of each and every independent attributes or predictors.
2. Secondly, random forest and k-nearest neighbour techniques are used to evaluate the value of target variable, i.e. wine quality.

3.1 Model Overview

This architectural diagram shows the overall process flow along with the components of the system.

3.1.1 Dataset

Dataset is divided into test set and training set according to the splitting ratio. Training set then is taken to build a machine learning model which establishes the relation from the data of the dataset that issued to predict the quality of wine, i.e. output (Fig. 1).

3.1.2 Model Engineering

Machine learning model is using the random forest and k-nearest neighbour to build the prediction model. KNN is used dynamically with random forest. Output of random forest is further processed by KNN to predict output.

3.1.3 Training

Build model will undergo in training phase, which will train the model corresponding to the dataset provided with the help of algorithms used.

3.1.4 Testing

The final model output undergoes testing for its predicted output with the help of test set that was splatted prior building the model. If the tested output have the desired accuracy and shows it as the output, otherwise it undergoes in training.

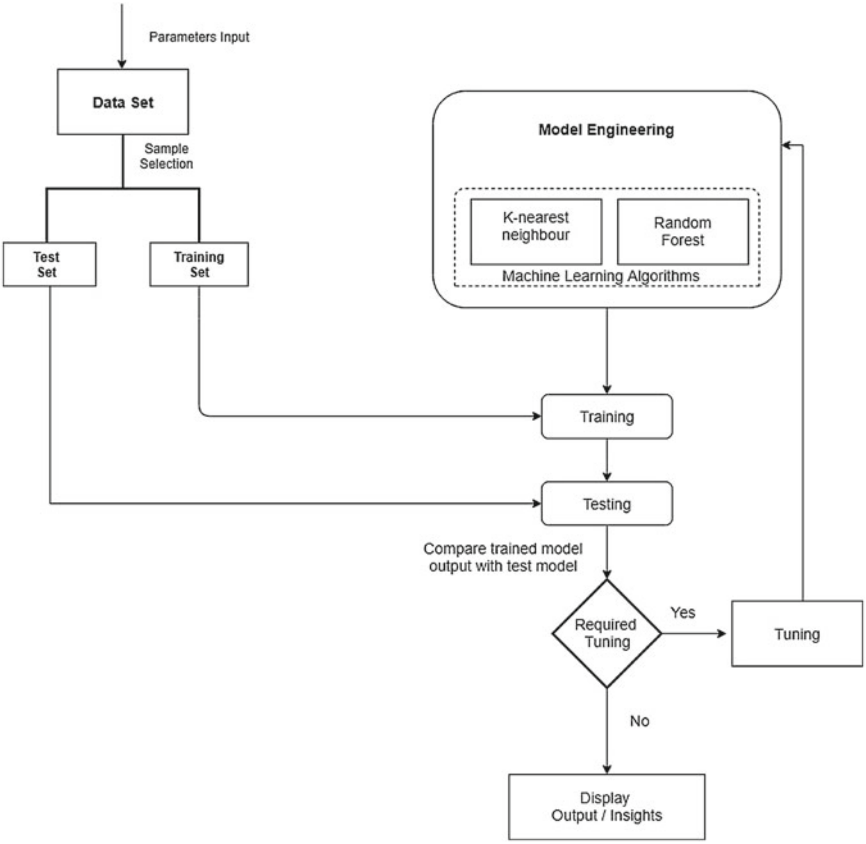

Fig. 1 Architectural diagram

3.2 Data Description

To analyse the quality of wine, a large dataset is taken from the research done by Paulo Cortez, in UCI Machine Learning Repository, contributed by University of Minho, Portugal [9]. This dataset consists of chemical information of 6499 types of Portugal wines, in which 4989 varieties are of white wines and 1650 varieties are of red wines. These datasets contain 1599 observations with 12 different feature variables/attributes such as alcohols, residual sugar, chloride, density, free sulphur dioxide, total sulphur dioxide and pH present in both red and white wines [9]. The quality of wine is analysed as follows:

$$\begin{aligned}\text{Quality} = \alpha 0 + \alpha 1\text{alcohol} + \alpha 2\text{volatile acidity} + \alpha 3\text{density} \\ + \alpha 4\text{chlorides} + \alpha 5\text{pH} + \in\end{aligned}$$

Conclusion based on the analysis of datasets is as follows:

1. The two most important features among all 12 attributes are sulphur dioxide (both free and total) and alcohol.
2. The most important factor to decide the quality of wine is alcohol; higher concentration of alcohol leads to better quality of wine and lower density of wine.
3. Sulphates are added by wine-makers to prevent spoilage and has positive correlation to wine quality.
4. Volatile acidity contributes to acidic tastes and has negative correlation to wine quality. Citric acid is added to give a freshness test and hence has a positive impact on wine quality.

4 Result and Analysis

After literature survey, we found that previously built models were using only single algorithm which gives root mean square error (RMSE) minimum of 0.6430 for white wine and 0.6322 for that of red wine by using random forest, and they were using all the attributes of the wine to build the model (Table 1).

Also, we found the research gap in which we come to know the fact that accuracy of the model can be further raised if we select only important features which play important role in determining wine quality; along this, KNN is used dynamically with random forest which minimizes the RMSE up to 0.541 for white wine and 0.584 for red wine of this model and is capable of predicting bad wine with 100% accuracy and good wine with 92% accuracy. KNN RMSE calculated the same to that of random forest and used along with it (Table 2).

Table 1 RMSE values for random forest

Algorithm	White wine RMSE	Red wine RMSE
Random forest	0.6430	0.6322

Table 2 RMSE values for random forest and KNN

Algorithm	White wine RMSE	Red wine RMSE
Random forest	0.541	0.584
K-nearest neighbour	0.541	0.584

5 Conclusion and Future Work

The classification tree provides the information that only 43% of the red wine tuples were classified in tree created using red wine dataset. Further, only 58% of white wine tuples were classified in tree created using white wine dataset. The quality value ranges from 4 to 7 which is classified in the decision tree. The limitation of this is that it does not classify extreme quality values, i.e. 0–3 and 7–10. Regression tree among another machine learning algorithms provides the best result with more accuracy. To make wine analyser model more dynamic, KNN algorithm is used through which we can predict quality of any produced wine. The quality value ranges from 4 to 7 which is classified in the decision tree.

Quality of wine is closer to the original value when we use only selected parameters to determine the quality which mainly influences the result. Wine manufacturers can use results to enhance the quality of wine by analysing the ranges in which different constituents should be for best-quality wine.

The parameters used in our dataset form a complex dimensional representation of each type of wine. But it can be possible that there are co-relations that cannot be visible immediately or need some calculation to be more specific and classified, e.g. PH and fixed acidity show a similar relation on quality, and hence, the datasets can be merged, to simplify the problem. As we know random forest tree is the best algorithm to analyse the datasets of wine and shows more accuracy then other algorithm, it can be used to improve the dataset and to reduce the number of dimensions formed while analysing wine quality.

References

1. Er Y (2016) The classification of white wine and red wine according to their physicochemical qualities. Int J Intell Syst Appl Eng 4(1):23–26
2. Executive Summary, Wine Process Monitoring, Wine Quality, Wine Safety, and Wine Complexity (2016) Wine analysis :from ‘Grape to Glass’ an analytical testing digest of the wine manufacturing process
3. Palmer J, Chen B (2018) Wine informatics : regression on the grade and price of wines through their sensory attributes
4. Tajini B, Paris OC (2017) BadrTajini—On campus Paris—DSTI 2017 47(4):547–553
5. Ghosh A (2018) Project report : red wine quality analysis final 3. An empirical red wine quality analysis of the Portuguese ‘Vinho Verde’ wine (2017, 2018)
6. A. Co “Final report,” Apr (2001)
7. Gupta Y (2018) Selection of important features and predicting wine quality using machinelearning techniques. Procedia Comput Sci 125:305–312
8. Predictive Model, Linear Regression, and RapidMiner Studio (2018) Building and evaluating a predictive model w/ linear regression in RapidMiner studio
9. Cortez P (2010) Wine quality dataset. https://archive.ics.uci.edu/ml/datasets/Wine+Quality