ISSN: 2249-7196
IJMRR/ May 2023/ Volume 13/Issue 2/Article No-1/01-9
DrK.SivanagiReddy/ International Journal of Management Research & Review

WINE QUALITY PREDICATION AND CLASSIFYING USING MECHNIE LEARNING

DrK.SivanagiReddy[1],Niharika.K[2], Venusri.M[3], Sravika.N[4]

[1]*Professor, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India*

[234] *B. Tech Student, Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India*

[1234]*Department of Electronics and Communication Engineering, Sridevi Women's Engineering College, Hyderabad, India*

[1]*sivanagireddykalli@gmail.com*
[2]*niharikagoud245@gmail.com*
[3]*mukkuvenu@gmail.com*
[4]*sravikanuka01@gmail.com*

ABSTRACT: ***The majority of industries base the promotion of their goods on the product certifications for quality. The classic method of while evaluating the quality of a product takes time, wine quality first analysed by data scientist then execution of wine take place. It has become faster and more efficient with the development of machine learning algorithms. In this paper, we used jupyter as platform for writing the code and python for writing the code. Random Forest, MLR, SVC are the different algorithms used for classifier then best algorithm for predication. Some of the machine learning techniques to assess the quality of wine based on the attributes of wine that depends on quality. Here we required of 12 attributes for predication of alcohol. Then, Logistic regression and Random forest, SVC classifier are performed individually on data to predict the test data values. Random forest (RF) classifier with accuracy 90% while MLP has 55% accuracy rate and SVC has 48%.***

Key words - Wine Quality, RANDOM FOREST, MLP, SVC

I.INTRODUCTION

The prediction of wine in wine making industries where do so long, so that the wine quality can be known and create of wine take place after the prediction is done.

By the year's the prediction of wine quality is done in different method [1], [2], [3]. Using different algorithms and process for prediction of wine quality coming across different algorithm using different language for prediction. But accuracy of algorithm is less in past and prediction was not accurated. Here we classify the data into different algorithm to get correct accurate value where it is good or normal or bad.

We took a set of data from kaggle and split the data into train and test and pre-process the raw data for all the algorithms. We used jupyter as code platform and we detected the best algorithm and then quality of wine will be shown in percentage with very bad or good or normal at output.

In section (1) we will explain the proposed system and section (2) Literature survey, in section (3) proposed system. We present the implementation in section (4) we present the conclusion.

II. LITERATURESURVEY

For development of this project, various researches conducted by many researchers.

In this report [1] "Wine Quality Detection through Machine Learning Algorithms". Machine learning techniques are utilized to analyze those attributes. Firstly, data pre-processing takes place i.e. making data appropriate for the models that are built for prediction. Defining independent and dependent variables, missing data handling, feature scaling and data splitting is done to improve the data standard. Then, Logistic regression and Random forest classifier are performed individually on data to predict the test data values. Random forest (RF) classifier outperforms logistic regression (LR) with accuracy 84% while LR has 53% accuracy rate.

[2] Research on Red Wine Quality Based on Data Visualization. Due to the rapid development of modern society, red wine has gradually become popular. Research on the quality of red wine has turn into an important topic. Red wine contains more than 600 kinds of ingredients, in terms of alcohol, minerals, tannic acid, citric acid, chloride and other substances. This paper analyses 12 factors affecting red wine quality in the data set and studies the influence of each ingredient on the quality of red wine through data mining algorithm. Based on the data visualization of Python processing, classical visualization tools such as histogram, heat map, boxplot and Pearson correlation coefficient algorithm are used for data mining. The histogram is adopted for univariate analysis and the heat map composed of Pearson coefficient is used for multivariate analysis.

[3] Navajas , Eva Campo , Angela Sutan , Jordi Ballester, "Perception of wine quality according to extrinsic cues: The case of Burgundy wine consumers". Show that it exists an important trade-off in quality perception among different extrinsic cues such as origin, denomination of origin (1er Cru vs vin de pays), label aesthetic (classical vs modern), bottling (estate vs cooperative bottled), the presence of awards as well as different cues commonly linked to tradition such as "special curve" or being produced by independent winemakers or being perceived as a wine with a potential for ageing.

[4] "Variety seeking behaviour in the wine domain: A consumers segmentation using big data". Panel Francesco Caracciolo, Marilena Furno, Mario D'Amico, Giovanbattista Califano, Giuseppe Di Vita. The largest group is "switchers," which includes consumers showing a relatively higher wine diversity than brand diversity. Estimates reveal the "habitual" group, that lives in the southern Italy and consumes wine less frequently than all other groups. The "loyal" group includes the youngest consumers with an above average income, who reside in the northern regions. Finally, the "variety seekers" are older, have the highest incomes, and live in the central regions. This grouping provides insights into the effects of brand and wine typology on consumers' choices.

[5] Alex A. Freita "A Survey of Evolutionary Algorithms for Data Mining and Knowledge Discovery". This discusses the use of evolutionary algorithms, particularly genetic algorithms and genetic programming, in data mining and knowledge discovery. We focus on the data mining task of classification. In addition, we discuss some preprocessing and post processing steps of the knowledge discovery process, focusing on attribute selection and pruning of an ensemble of classifiers. We show how the requirements of data mining and knowledge discovery influence the design of evolutionary algorithms. In particular, we discuss how individual representation, genetic operators and fitness functions have to be adapted for extracting high-level knowledge from data.

[6] "Judging wine quality: Do we need experts, consumers or trained panelists?". Helene Hopfer, Hildegarde Heymann. wine experts and consumers evaluated 27 Californian Cabernet Sauvignon wines with varying quality scores. Descriptive Analysis revealed several aroma and flavor descriptors driving quality scores. For all consumer segments as well as the wine experts, hedonic liking was shown to highly correlate to perceived quality, but for some consumers liking and perceived quality was not at all correlated to the quality scores of the wines. Wine experts were able to find significant differences in liking and quality, but did not agree completely with the assigned quality scores from the wine judgment. Wine experts also used a combination of both descriptive and hedonic terms when describing a high quality wine, indicating that they are better at communicating and describing what they like.

III. WINE QUALITY USING ML ALGORITHMS

A. Preparation of dataset:

So here we used a dataset for Kaggle this dataset is viewed as classifier or regression tasks. The raw dataset will first pre-processed. We took 10 attributes for prediction of wine quality in data we also have quality column we separate the quality value make 11 attributes as X-axis label and quality as Y-axis label. If we get a random data value of each attributes it will

compare the previous data and if it has it will take the values from previous and if we don't have then it will tell good or bad or normal also show the accuracy of all algorithms in graphical representataion. F1-score, prediction, recall of algorithms.

1)Fixed acidity
2)Volatile acidity
3) Citric acid
4)Residual sugar
5)Chlorides
6)Free sulphur dioxide
7)Total sulphur dioxide
8)Density
9)pH
10)sulphates
11)Alcohols
12)quality

B. Data Processing:

Data processing which means the data after analysis we got a raw data to our ML model so we need pre-process so that it become compatible with our ML model so once we process data there is another important step called as test split so in the particular step we will split our original dataset into training & test data .The reasons for this is we will train our ML model using this training data & we won't show this test data to our ML model for training so this disk data is used to evaluate our model to find how good our model is performing. So, this the reasons for splitting the data into training & test data. Training data is used to train the model to learn from the data and the test data is used to evaluate or test the model.

C. Architecture

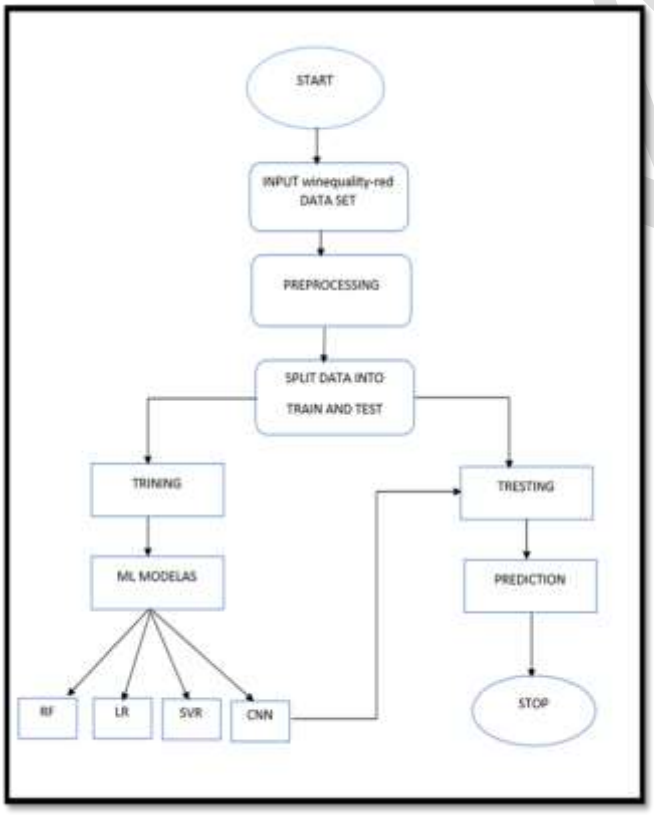

Figure: Architecture of wine predication.

3.1 Models Classification:

Here for classification of data set we four model.
They are

A. Random Forest

A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that improves the accuracy of machine learning algorithms.

B. Multiple Logistic Regression

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable
(target) is categorical.

C.SVC

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The objective of the SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane became a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

D.CNN

The convolutional layers are the key component of a CNN, where filters are applied to the input image to extract features such as edges, textures, and shapes. The output of the convolutional layers is then passed through pooling layers, which are used to down-sample the feature maps, reducing the spatial dimensions while retaining the most important information. The output of the pooling layers is then passed through one or more fully connected layers, which are used to make a prediction or classify the image. It's used for data storing, because it will in feature selection & extraction of different values and also separates the unique values quality.

IV. IMPLEMENTATION

Here we take 1599 with 12 attributes data set from kaggle[17] website for classification which is raw dataset. Then we pre-process the data set and then data split into test and train dataset. Here we see that after classification of all algorithms, RF shows more accuracy then all algorithms.so, we use RF for our predication. In order to categorise the wine's quality as excellent or very bad because it wasn't distributed regularly. A sample of wine is used to forecast the wine's quality. Labels were provided based on whether the wine was of GOOD or BAD quality according to how the data was organised. The techniques are covered. Additionally, shown for performances report by confusion matrix. Accuracy, Precision, Recall, F1-Score are used to assess how well the approaches work.

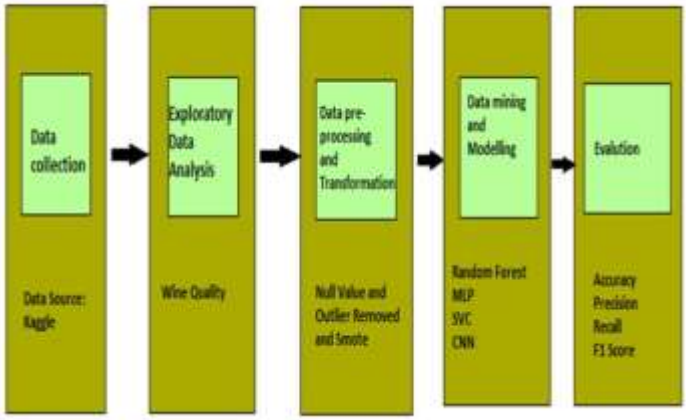

Figure: Databases that support knowledge discovery

A. Classification with Random Forest

Classification	Random forest
Precision	86.82
Recall	83.71
F1 score	85.24

Table 1: Classification Report of Random Forest

As it can be seen in Table 1, the accuracy of the Random Forest was 90.00 %. Precision, recall, and f1_score is, respectively, 86.82%,83.71%,85.24%.

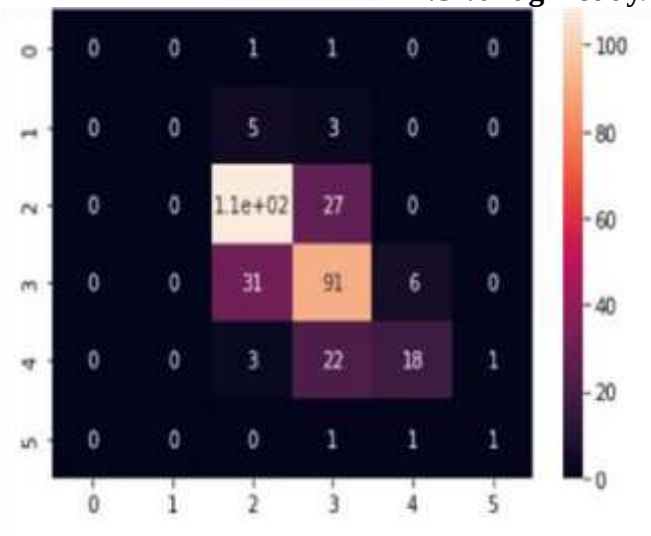

Figure: Confusion Matrix of Random Forest

B. Classification with Multiple Logistic Regression

Classification	MLP
Precision	55.00
Recall	57.00
F1 Score	26.00

Table 2: Classification of MLP

As it can be seen in Figure, the accuracy of the Random Forest was 57.00 %. Precision, recall, and f1_score is, respectively, 55.00, 57.00, 26.00.

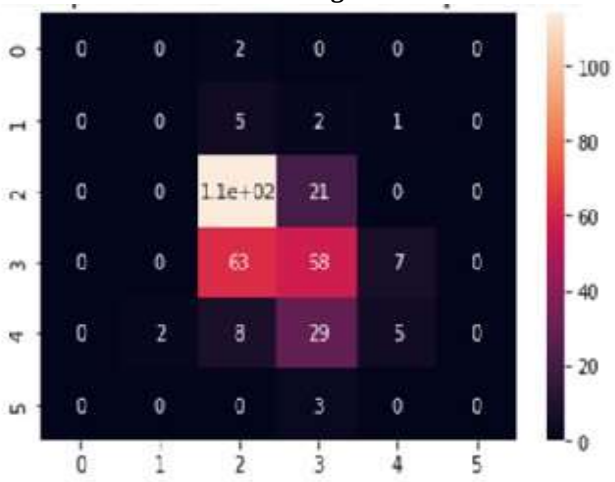

Figure: Confusion Matrix of Random Forest

C. Classification with SVC

Classification	SVC
Precision	44.00
Recall	39.00
F1 Score	47.00

Table 3: Classification of SVC

As it can be seen in Figure, the accuracy of the Random Forest was 48%. Precision, recall, and f1_ score is, respectively, 44.00, 39.00, 47.00.

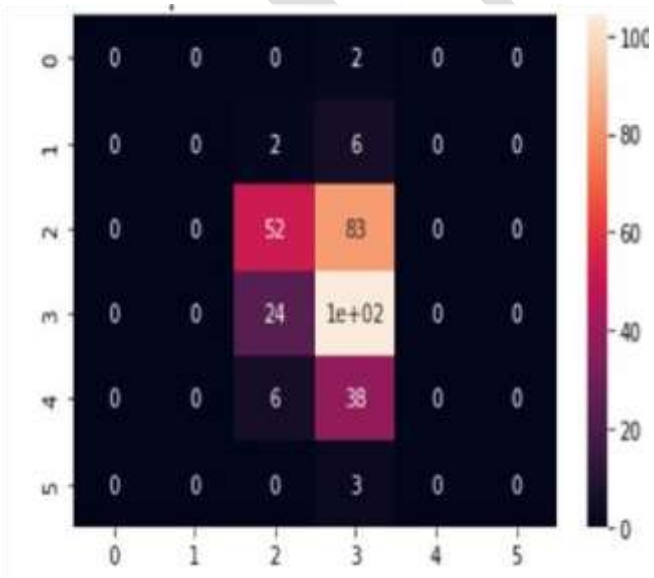

Figure: Confusion Matrix of SVC

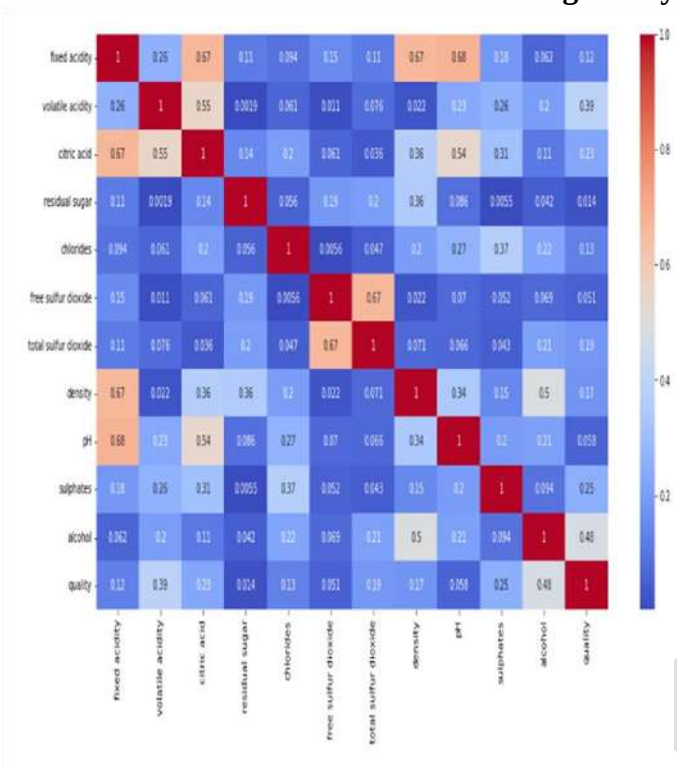

Figure: Correlation matrix of features

Similarly, from wine correlation matrix we ranked the features according to the high correlation values to the quality class such as features are 'alcohol', 'density', 'chlorides', 'volatile acidity', 'total sulphur dioxide', 'fixed acidity', 'pH', 'residual sugar', 'sulphates', 'citric acid', 'free sulphur dioxide'. Here we see all the attributes are correlations at sum of point .

V. CONCLUSION

A wine quality can be classified as “GOOD” or “BAD” depending upon a number of variables. In order to estimate the quality, this article concentrated on these parameters and employed several machine learning algorithms. In specification, we created ad examined two classification algorithms that were targeted at the identical Red Wine dataset but used various method to forecast the dependent variables. It was shown that Random Forest produced prediction that were superior to those produced by Logistic regression. The accuracy rate indicated by the RF model is 90%, compared to LR’s accuracy rate of 55% and SVC’s accuracy rates 48%. Random forest is used of predication; it’s having more accuracy so we chose RF as classifier model.

S.NO	Models	Accuracy
1	Random Forest Classifier	90.59400
2	MLP Classifier	55.625000
3	SVC	48.750000

4	CNN	66.430002

Table 4: Accuracy of all algorithms

FUTURE SCOPE

In the future, Improving the accuracy of the models classifier, it is clear that the algorithms like MLP, SVC has less accuracy. We recommend feature engineering, using potential relationships between wine quality, or applying the boosting algorithms on the more accurate method. In additional, by applying the other performance measurement and other machine learning algorithms for the better comparison on results. This study will help the manufacturing industries to predict the quality of the different types of wines based on certain features, and also it will be helpful for to make good product of wine.

REFERENCES

[1] Eva Campo, Angela Satan, Jordi Baluster**, "**Perception of wine quality according to extrinsic cues: The case of Burgundy wine consumers**".**https://www.sciencedirect.c om/science/article/abs/p ii/S0950329312001085.

[2] "Variety seeking behaviour in the wine domain: A consumer segmentation usingbigdata" panelFrancesco Caracciolo,Marilena Fu rno, Mario D'Amico, Giovanbattista Califano,G iuseppe DiVita. https://www. sciencedirect.com/science/article/abs/p ii/S0950329321003633

[3] Alex A. Freitas "A Survey of Evolutionary Algorithms forDataMiningandKnowledgeDiscovery". https://link.springer.com/chapter/10.1007/978-3642-18965-4_33

[4] "Judging wine quality: Do we need experts, consumers or trained panellists?". Helene Hopfer, Hildegarde Heymann. https://www.researchgate.net/publication/25913699.Judgewine_quality_Do_we_need_experts_consumers_or_trained_panel ists

[5] "A Survey of Web Usage Mining Techniques" Path Suthar, Prof. Bhavesh Oza https://www.ijcsit.com/docs/Volume%206/vol6issu e06/ijcsit2015060644.pdf

[6] "Comparison of a logistic regression and Naïve Bayes classifier in landslide susceptibility assessments: The influence of models complexity and training dataset size"
DOI:10.1016/j.catena.2016.06.004

[7] A survey on data mining classification algorithms
S. Umadevi; K. S. Jeen Marseline

[8] M. Mattheakis, P. Protopapas, K. Rader. "Generalized Linear Models: Logistic Regression and Beyond". github.io /2020-CS109A/a sections/asection3/ Asec3 _2020_notes_ GLM.pdf

[9] Guangzhou Hu, Tan Xi, Faraz Mohammed, Huaikou Miao. (2016). Classification of wine quality with imbalanced data. 2016 IEEE International Conference on Industrial Technology (ICIT). doi:10.1109/icit.2016.7475021.

[10] Jambhulkar and Baporikar. (2015) "Review on Prediction of Heart Disease Using Data Mining Technique with Wireless Sensor Network". International Journal of Computer Science and Applications 8 (1) 55-59.

[11] Sun, Danzer and Thiel. (1997) "Classification of wine samples by means of artificial neural networks and discrimination analytical methods". Fresenius Journal of Analytical Chemistry359 (2) 143–149.

[12] Agricultural and Food Chemistry 56 307–313. Beltran, Duarte-Mermoud, Soto Vicencio, Salah and Bustos. (2008) "Chilean Wine Classification Using Volatile Organic Compounds Data Obtained With a Instrumentation and Measurement 57 2421-2436.
[13]https://www.kaggle.com/datasets/uciml/redwine-quality-cortez-et-al-2009
[14] D. Smith, R. Margolskee, Making sense of set, Scientific American, Paulo Cortez, Antonio, Fernando.Introduction ToMachineLearningalgorithms. https://towardsdatascience.com/supportmachine-learning introduction-to-machinelearningalgorithms-934a444fca47(accessed 6.1).
[15] Dahal, K., Dahal, J., Banjade, Gaire, S.2021.Prediction of Wine Quality Using Machine Learning States used 11, 278289.https://doi.org/10.4236/ojs.2021.112015
[16] Wine Quality Detection through Machine Learning Algorithms
[17]https://www.kaggle.com/datasets/uciml/redwine- quality-cortez-et-al-2009