iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VIII **Month of publication:** August 2021

DOI: https://doi.org/10.22214/ijraset.2021.37629

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

Wine Quality Prediction Using Machine Learning

Vrusha P. Sangodkar[1], Umesh A. Bapat[2]
[1]*PG Student Computer Science and Engineering, Department of Computer, Goa College of Engineering, Farmagudi, Goa University*
[2]*Associate Professor, Department of Computer, Goa College of Engineering, Farmagudi, Goa University*

Abstract: Nowadays people are living a luxurious lifestyle, wine has become a part of one's culture. consumption of wine is very common throughout the world so its quality is very important. hence its important to analyse wine quality quality of the wines are usually checked by humans through tasting but it has other physicochemical attributes which affects the taste but the process is slow hence machine learning methods can be used for the same. dataset is taken and feature selection is done using pca feature selection and then accuracy is find using SVM, backpropagation neural network and Random forest algorithm to find which model fits best and gives greater accuracy.
Keywords: Data Extraction, PCA, SVM,BP neural network, Randomforest

I. INTRODUCTION

Data mining is the trail toward discovering new examples to separate the quality information from the immense storehouse.It incorporates various varieties of measurements, machine learning and arrangement of databases. The fundamental target is to isolate significant information from a tremendous database.[1]. And find patterns within the data.it has a large sized data so knowledge extraction hence its complex. Machine learning is a subfield of artificial intelligence (AI). The goal of machine learning generally is to understand the structure of data and fit that data into models that will be understood and utilized by people.[2].machine learning focuses on the data which can be utilized to learn themselves and later predict the output the process begins by taking examples and training the system for each example it train themselves and used to make future predictions.

Two of the foremost widely adopted machine learning methods are supervised learning which trains algorithms with examples of input and output data that's labeled by humans, and unsupervised learning which provides the algorithm with no labeled data in order to allow it to find out structure within its input data[3]

II. LITERATURE

Paulo Cortez discussed that wine was once viewed as luxury is now is consumes by wider range portugal being the highest exporting country of its *vinho verde* wine.due it its increasing growth wine industries are investing more in testing and selling techniques.quality verification is part of certification process and can be improved in wine making.they have done case study on regression methods taking dataset with output ranging from 0-10.they have used sensitive analysis for extracting knowledge.they used svm and NN model and since svm is less effective to outliers it gives higher accuracy,[4]

Sunny Kumar ,Kanika Agrawal ,Nelshan Mandan in their paper states that quality checking is important as consumption is greater they have suggested machine learning models for the same.they have take random forest,support vector machine and naive bayes and accuracy is calculated using f-score,precision,recall and found out that support vector machine fits best with highest accuracy of 67% and them random forest and naive bayes[1]

Satyabrata Aich , Ahmed Abdulhakim Al-Absi , Kueh Lee Hui , John Tark Lee and Mangal Sain proposed a paper they have considered different feature selection algorithm such as Principal Component Analysis (PCA) as well as Recursive Feature Elimination approach (RFE) approach for feature selection and nonlinear decision tree based classifiers for analyzing the performance metrics and found different accuracies using random forest model [5]

Àngela Nebot , Francisco Mugical and Antoni Escobet in their paper states that wine classification is task since its least understood by humans sense they proposed In this research we propose to use hybrid fuzzy logic techniques to predict human wine test preferences based on physicochemical properties from wine analyse[6]

S. Kallithraka ", L.S. Arvanitoyannis h, P. Kefalas, A. El-Zajouli, E. Soufleros E. Psarra they used pca on factors to classified 33 geek wines according to geographic origin.[7]

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VIII Aug 2021- Available at www.ijraset.com

A. Research Methodology

Wine data was collected from UCI machine learning repository[8].and in this experiment red wine data is taken in consideration there were all together 1599 samples and 12 variables which are listed above which contains 11 physicochemical properties and output.data had 1-10 output range which was converted to good wine or bad wine ,below 5 was considered as bad wine and above 5 was considered to be good.

1) *Performance Measure:* Performance measures are calculated to measure effectiveness and efficiency of the method.and to check which model effectively fits the data
2) *Confusion Matrix:* Confusion matrix is mainly used for evaluating performance in classification model it is N*N matrix where N is number of target classes it has true positive(TP), true negative(TN),false positive(FP),false negative(FN)

a) Accuracy: it is calculated as true positive plus true negative divided by sum of all.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

b) Precision:- precision is calculated as true positive divided by true positive plus false positive.

$$Precision = \frac{TP}{TP + FP}$$

c) Recall :- Recall is calculated as true positive divided by true positive plus false negative

$$Recall = \frac{TP}{TP + FN}$$

d) F1-score: is calculated as recall multiplies by precision divided by recall plus precision into 2.

$$\text{F1-Score} = 2 * \frac{recall * Precision}{recall + precision}$$

III. DATASET

A. Wine Attributes and Properties

Different properties we consider are :

1) *Fixed Acidity:* Acidity in wine balances sweetness and bitterness in wine.Reducing acids significantly might lead to wines tasting flat Volatile acidity:is a measure of the low molecular weight (or steam distillable) fatty acids in wine . The volatile acidity is expressed in *g(acetic acid)/dm3* in the dataset
2) *Citric Acid:* Is very important to give wine its freshness.its also added to prevent ferric hazes. It's usually expressed in *g/dm3* in the dataset.
3) *Residual Sugar:* It refers to the natural sugar from fruit that remains even after the fermentation process is stopped it gives the wine sweetness. It's usually expressed in *g/dm3* in the dataset.
4) *Chlorides:* This is a large contributor to saltiness in wine. It's usually expressed in *g(sodium chloride)/dm3* in the dataset.
5) *Free Sulfur Dioxide:* This is the part of the sulphur that is free after the remaining part is bound to other chemicals.. This variable is expressed in *mg/dm3* in the dataset.
6) *Total Sulfur Dioxide:* This is the sum total of the bound and the free sulfur dioxide (*SO2*). Here, it's expressed in *mg/dm3*..
7) *Density:* It is generally used as a measure of the conversion of sugar to alcohol. Here, it's expressed in *g/cm3.*
8) *pH:* potential of hydrogen, this is a numeric scale to find if the wine is acidic or basic. Most of the wines have a pH between 2.9 and 3.9 hence they are acidic.
9) *Sulphates:* These are mineral salts containing sulfur They are an important part of the winemaking and are considered essential. Here, it's expressed in *g(potassium sulphate)/dm3* in the dataset.
10) *Alcohol:* Alcohol is formed during the fermentation process as a result of yeast converting sugar. The alcohol percentage is different for every wine. It's usually measured in % vol or alcohol by volume (ABV).
11) *Quality:* wine quality is 0 or 1. 0 is bad and 1 is good

IV. PROPOSED SYSTEM

Dataset is extracted from UCI machine learning repository for red wine variants of the Portuguese "Vinho Verde" wine.it contains 1599 samples and 12 features quality being one of them. The data is to predict the quality of wine which can be further used by wine industries. first quality is changed 1-10 to “good” or”bad” below 5 is bad and above 5 is good. The data is split into 70% and 30%, 70% is for training and 30% for testing. Libraries like numpy, pandas, random is imported. PCA feature extraction is applied and summary is calculated using SVM and Back Propagation neural network, random forest .and 2*2 confusion matrix is generated on dataset observation and quality is calculated.

Further various performance measures are calculated like specificity,precision,recall,F1-score for both the models and accuracy is calculated for all three model Support vector machine, Back Propagation neural network and random forest .The results are predicted based on this. This research on UCI’s red wine dataset have found that Random forest fits better then svm model and BP neural network with accuracy 80.9%.”fig 1” shows the whole dataset with all features. ”Fig 2” shows the accuracy calculated for SVM Model and all the performance measure calculations for SVM model where -1 is bad and 1 is good, ”Fig 3” shows accuracy and performance measure calculations for BP Neural Network.: ”Fig 4” shows accuracy and performance measure calculations for random forest. ”Fig 5” shows the plotting of quality parameter of test data before classification and “Fig 6” shows the quality parameter of test data after classification.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	0
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	0
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	0
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	1
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	0
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	1
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	1
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	0
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	1

1599 rows × 12 columns

Fig 1.Dataset

```
############ PRINT SVM INFO ################
C: 1
max_iter: 10000
epsilon: 0.01
kernel_type: linear

Classification report :
              precision    recall  f1-score   support

          -1       0.65      0.78      0.71       148
           1       0.77      0.64      0.70       172

    accuracy                           0.70       320
   macro avg       0.71      0.71      0.70       320
weighted avg       0.71      0.70      0.70       320


Accuracy: 0.703125

F1-Score: 0.6984126984126984
```

Fig 2..Accuracy and Performance Measure for SVM

```
Classification report :
              precision    recall  f1-score   support

           0       0.72      0.64      0.68       148
           1       0.72      0.78      0.75       172

   micro avg       0.72      0.72      0.72       320
   macro avg       0.72      0.71      0.71       320
weighted avg       0.72      0.72      0.72       320
 samples avg       0.72      0.72      0.72       320


Accuracy:  71.875 %

F1-Score: 0.71875
```

Fig 3. Accuracy and Performance Measure for BP Neural Network

```
Classification report :
              precision    recall  f1-score   support

           0       0.79      0.79      0.79       146
           1       0.82      0.83      0.83       174

    accuracy                           0.81       320
   macro avg       0.81      0.81      0.81       320
weighted avg       0.81      0.81      0.81       320


Accuracy:  80.9375 %

F1-Score: 0.809375
```

Fig 4. Accuracy and Performance Measure for Random forest

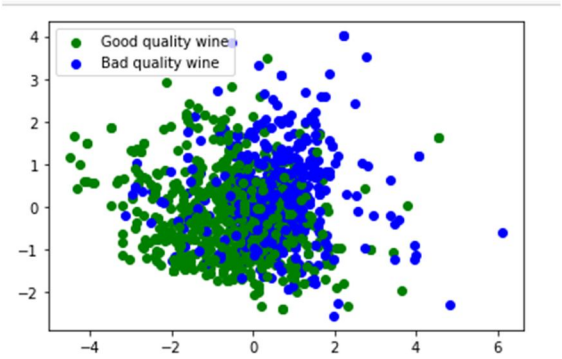


Fig 5.Test Data Before Classification

Fig 6.Test Data After Classification

V. RESULT

Nowadays wine is very important and quality testing is important since consumption is high and bad quality can affect one's health.so research dataset contains wine information which is used to detect quality of wine. two machine learning algorithm is written in jupyter notebook and accuracy is calculated and find out which model fits best for the dataset. dataset is separated into training set and testing set in ratio 70-30%.the result shows that SVM gives 70% Accuracy and BP neural network gives 71% Accuracy and Random Forest gives 80.9% Accuracy. if the SVM hyperplane is adjusted more than accuracy of SVM can be higher than BP neural network. Table 1 shows performance measure for SVM for both the set bad and good. and Table 2.shows performance measure for BP neural network. Table 3.shows performance measure for Random Forest.

Table 1

	Wine quality = bad	Wine quality= good
precision	0.65	0.77
recall	0.78	0.64
F1-score	0.71	0.70
support	148	172

Accuracy = 70%

Table 2

	Wine quality = bad	Wine quality= good
precision	0.72	0.72
recall	0.64	0.78
F1-score	0.68	0.75
support	148	172

Accuracy = 71.87%

Table 3

	Wine quality=bad	Wine quality=good
precision	0.79	0.82
recall	0.79	0.83
F1-score	0.79	0.83
support	146	174

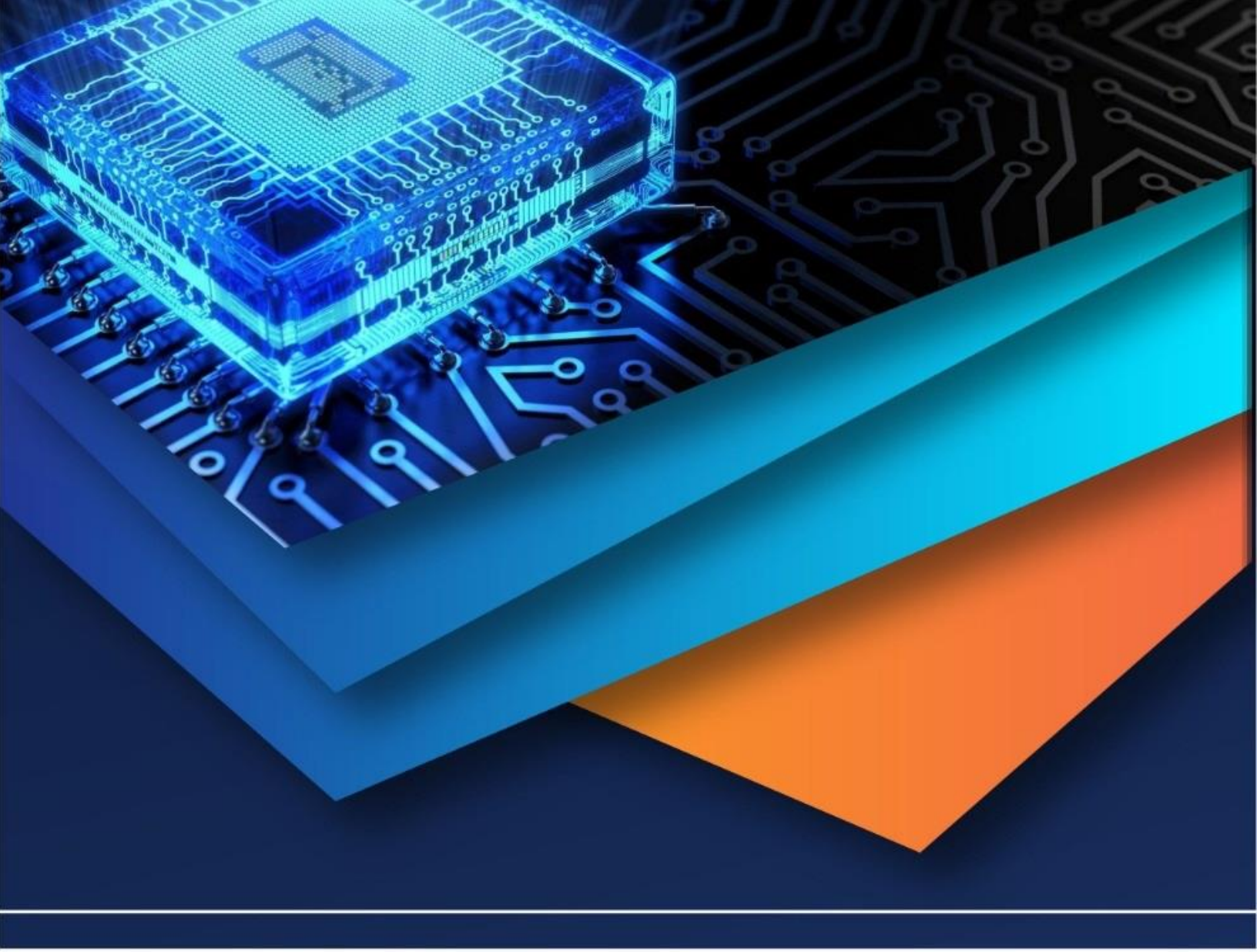
accuracy-80%

VI. CONCLUSION

Wine quality check can be used by many wineries in order to check quality of their wine ,and since machine learning is a promising technology this modes can be used on their dataset to train their dataset and predict the future wine as good or bad and can be accepted and discarded based on the output.on this data For this portuguese wine dataset according to this research Random forest fits best with accuracy 80.9% and back propagation neural network gives accuracy of 71% and then support vector machine gives accuracy of 70% and also performance measures like precision,recall,specificity,F1-score is calculated.

REFERENCES

[1] Sunny Kumar,Kanika Agrawal,Nelson Mandan "Red Wine Quality Prediction Using Machine Learning Techniques" in 2020 International Conference on Computer Communication and Informatics
[2] Akanksha Trivedi, Ruchi Sehrawat "Wine Quality Detection through Machine Learning Algorithms"in 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)
[3] https://www.digitalocean.com/community/tutorials/an-introduction-to-machine-learning.
[4] P. Cortez, A. Cerderia, F. Almeida, T. Matos, and J. Reis, "Modelling wine preferences by data mining from physicochemical properties," In Decision Support Systems, Elsevier, 47 (4): 547-553. ISSN: 0167-9236.
[5] "Satyabrata Aich, Ahmed Abdulhakim Al-Absi , Kueh Lee Hui , John Tark Lee and Mangal SainA" Classification Approach with Different Feature Sets to Predict the Quality of Different Types of Wine using Machine Learning Techniques
[6] "Àngela Nebot , Francisco Mugical and Antoni Escobet"Modeling Wine Preferences from Physicochemical Properties using Fuzzy Techniques
[7] S. Kallithraka , L.S. Arvanitoyannis h, P. Kefalas, A. El-Zajouli, E. Soufleros E. Psarra"Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin"
[8] UCI Machine Learning Repository, Wine quality data set, [Online].Available: https://archive.ics.uci.edu/ml/datasets/Wine+Quality.
[9] Sowmya D Sayyed ,Ganavi M Sankhya N Nayak "Analyzing wine types and quality using machine learning techniques"
[10] Gongzhu Hu, Tan Xi, Faraz Mohammed and Huaikou Miao"Classification of Wine Quality with Imbalanced Data"



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY