iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** I **Month of publication:** January 2024

DOI: https://doi.org/10.22214/ijraset.2024.58014

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: 7.538
Volume 12 Issue I Jan 2024- Available at www.ijraset.com

Wine Quality Prediction using Machine Learning

Shagun Davessar[1], Shashank Shekhar Tiwary[2]
[1]*Department of Computer Science and Engineering (AI ML), Manipal University Jaipur, India*
[2]*Department of Information Technology (IT), Manipal University Jaipur, India*

Abstract: ***According to the most recent statistics, India is a developing wine market in Asia's New World countries. Although the Indian wine business is in its infancy in terms of area, predictions, and marketing of wines, eighty percent of consumption is concentrated in cities. In today's world, it is very crucial to find out the quality of the wine we are using as it can affect our health, badly. [8] This model aims to predict the quality of different types of wines, based on parameters like percentage of acidity, sulphates, chlorides, sugar, pH used. This model demonstrates, how statistically it can be used to identify the components that mainly control the wine before it is produced in the country. This will also assist winemakers regulate quality, which will benefit both, the country's economy, and people's health.***
Keywords: Wine prediction, Machine learning, Random Forest Classifier.

I. INTRODUCTION

Wine industry is worth over $ 340 billion globally. According to latest stats, wine industry is expected to grow until it reaches a value of over $456 billion in 2028, which would be a CAGR of 4.3% from 2021 to 2028. Nowadays, with rising world's population consumption of wine is too increasing and becoming common. From a teenager to an Oldman, everybody is consuming it either for fun or occasionally. So, it is important to know the quality of wine people intaking as it can affect their health badly in future. [8]
In our model, we can predict the quality of the wine in an easier and more accurate way. We have used Machine Learning for the prediction, using Random Forest Classifier Model which comes under supervised learning.

A. Machine Learning Algorithms

1) *Random Forest:* It is a Machine Learning algorithm which comes under Supervised Learning and is used for Classification as well as Regression. It contains several decisions trees and takes average for a better predictive accuracy of the specific dataset. This model prevents overfitting.

We can say, more is the number of trees, higher is the accuracy.
We are using this model because-

- It predicts higher accuracy as the output.
- Ignores overfitting.
- Lesser training time.
- Can handle larger dataset.

Mathematically, [1]

$$normfi_i = \frac{fi_i}{\sum_{j \in all\ features} fi_j}$$

Fig. 1. Random forest model [1]

Where, normfi sub(i)= the normalized importance of feature I; fi(i), fi(j)= sum of importance of feature i and j respectively.

2) *Support Vector Machine (SVM):* It is a binary classifier which assumes that the data in problem statement contains two possible target values.

In this prediction, we didn't used SVM because:

- Large training time.
- New and more features, hence more complexity.
- Not suitable for larger datasets.

3) Naïve Bayes: It is a classifier algorithm which is based upon mathematic's Bayes Theorem.
Bayes Theorem explains the probability of occurrence of an event related to some condition. [10]
Mathematically, [1]

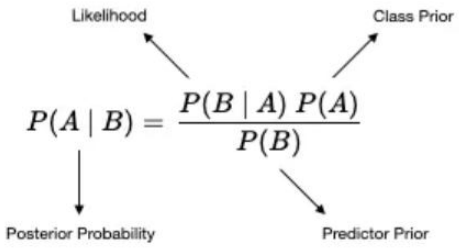


Fig. 2. Bayes Theorem [1]

We did not apply Naïve Bayes Classifier here because: [5]

- This algorithm cannot learn the relationship between features as it takes an assumption that all features are independent of each other.
- Often used for text classification problem statements.

4) Logistic Regression- It is an algorithm which comes under Supervised learning. It predicts the output as either Yes or No, 0 or 1, True or False. It is used for classification problems.

We did not apply this model because [14]

- Overfitting situation arises.
- Nonlinear problems cannot be solved
- Difficult to capture complex relationships.

Mathematically, [15]

$$\log\left[\frac{p}{1-p}\right] = \beta_o + \beta(\text{Age})$$

Fig. 3. Logistic Regression

5) K-nearest Neighbors (KNN)- It is simplest algorithm which comes under Supervised learning. It is mostly used for classification. The main point to be noted is, it doesn't learn from training set suddenly, instead stores dataset and make action when it is at classification stage.

We did not apply this model because [5]

- High computation cost.
- Difficult to calculate value of K (which is also mandatory)
- Cannot handle missing values.

Mathematically,

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Fig. 4. KNN Model's Euclidean distance

B. Related Work

Chen et al put forward an approach using the human savory reviews [3] to predict wine quality. They got an accuracy of 73.79-82.58% to predict wine quality using Hierarchical clustering approach and association rule algorithm.

Thakkar et al. to rank the attributes primarily they used analytical hierarchy process followed by machine learning classifiers such as random forest and they found accuracy of 70.33% using the same model.

These are some researchers who really worked on Predicting quality of wine using different Machine Learning models in the past. (Refer Table below)

II. METHODOLOGY

Accuracy	Literature Review in Wine Quality Prediction (Related work)		
	Author	*Year*	*Model used*
64.37%	Rohan Kothawade	2019	ANN
70.33%	Thakkar et al	2019	Hierarchical clustering
67.25%	Kumar et al	2020	SVM
88%	Mayur Bdole [9]	2022	Random Forest

Fig. 4. Correlation matrix

The workflow of the research has been done into different small parts, which are as follows:

1) *Data Collection:* The dataset is randomly taken from Kaggle [2]. It contains 13 different columns and 1143 rows and inserted on google colab. [11] For the data exploration part, importing all the necessary libraries like, Matplotlib, NumPy, Pandas, Seaborn. Data columns like, fixed acidity, citric acid, chlorides, pH, alcohol, sulphates, etc.
2) *Data Preprocessing:* There are quite chances of missing values inside the dataset which creates inconsistency. To cure this issue, data processes to achieve better and consistent results. inconsistency. To cure this issue, data processes to achieve better and consistent results.
3) *Data Splitting and Training:* Data gets split into two sets- Training and Testing datasets. Overall, approximately 70% is considered as training data and rest 30% of as testing data.
4) *Choosing a model:* Selecting one model such as, Random Forest, Logistic Regression, Support Vector Machine, etc.
5) *Model evaluation/Prediction*: By prediction the resultant accuracy, it is 95.10%

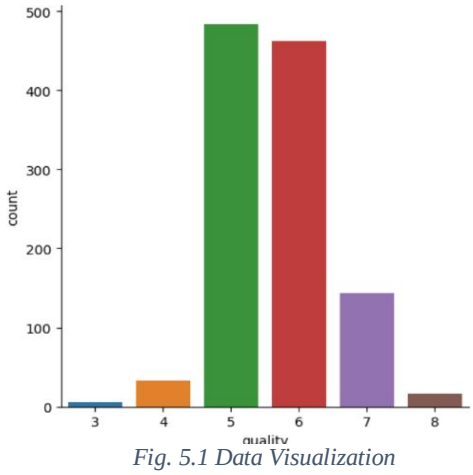


Fig. 5.1 Data Visualization

Fig. 8. Training and Testing data

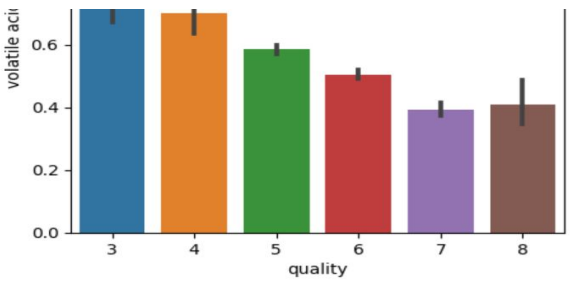


Fig. 5.2 Data Visualization

```
[ ]  print('Accuracy:',test_data_accuracy)

Accuracy: 0.951048951048951
```

Figure 7. Accuracy

```
[ ] X_train,X_test,Y_train,Y_test=train_test_split(X,Y,test_size=0.25,random_state=6) #0.2 means 20% of data is test data

print(Y.shape,Y_train.shape,Y_test.shape)

(1143,) (857,) (286,)
```

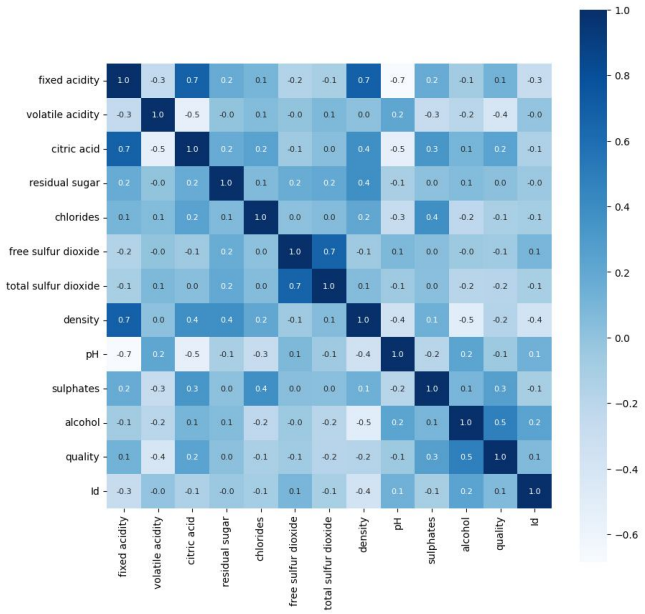


Fig. 6. Correlation (heatmap)

III. RESULT

The analysis of our applied machine learning model shows the accuracy achieved of 95.10% which is highest using Random Forest Classifier model as till now.

IV. FUTURE SCOPE

In this project, Quality of wine is predicted accurately using Random Forest Classifier. In future, this prediction system can be enhanced for a better result with better accuracy by using bigger datasets or using different ML algorithms. This model has doors opened for further improvement. During manufacturing of wines, organization can provide more minute details about the product so that more accurate result can be achieved.

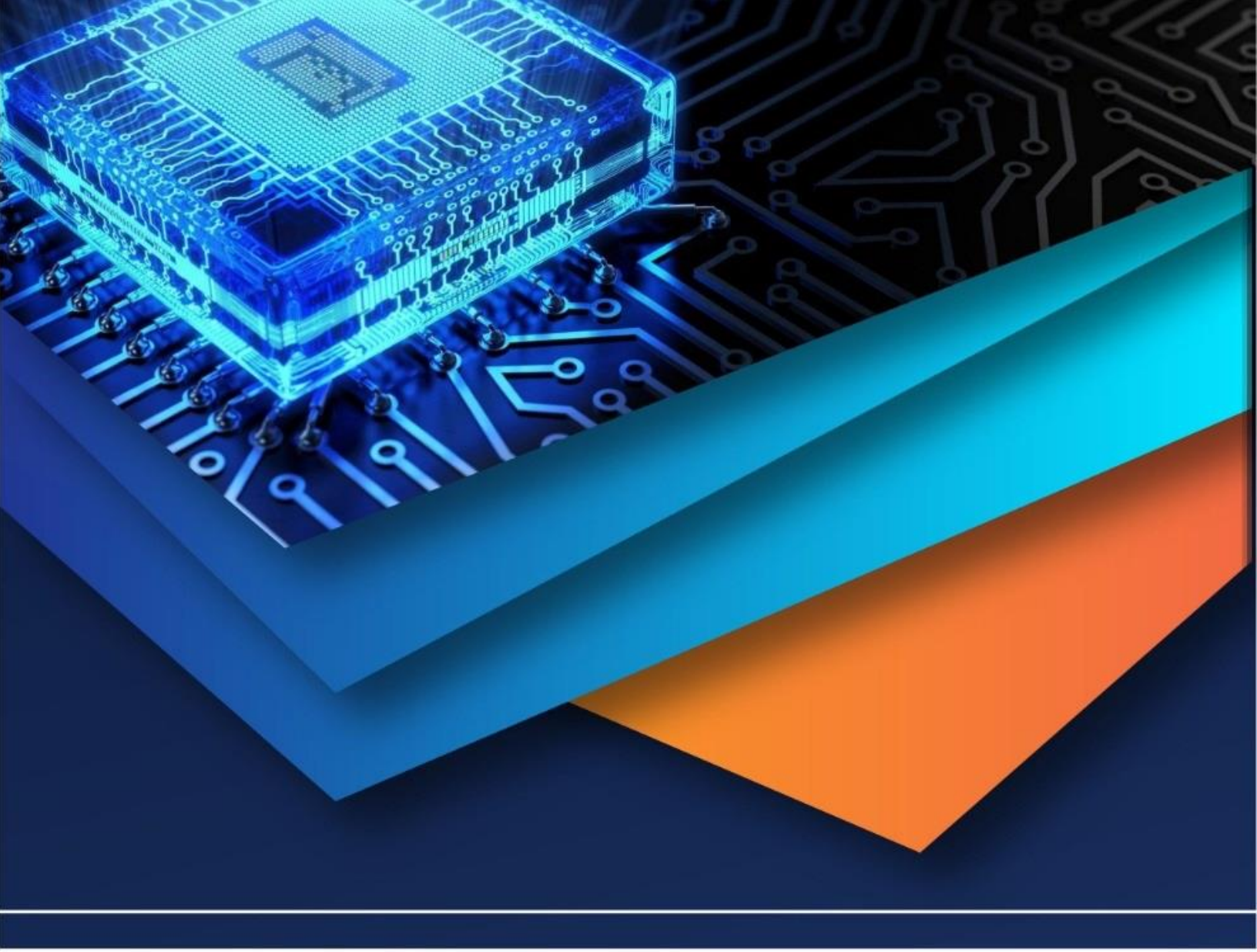
V. ACKNOWLEDGMENTS

Words cannot express my gratitude to my professors for his invaluable patience and feedback. I would be remiss in not mentioning my family, especially my parents. Their belief in me has kept my spirits and motivation high during this process.

REFERENCES

[1] www.towardsdatascience.com
[2] www.kaggle.com
[3] www.grin.com
[4] Dahal, Keshab & Dahal, Jiba & Banjade, Huta & Gaire, Santosh. (2021). Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics. 11. 278-289. 10.4236/ojs.2021.112015
[5] www.javatpoint.com
[6] Avinash Sanjay Gawale, 'Wine quality prediction using Hybrid Modeling and Machine Learning'- https://norma.ncirl.ie/6124/1/avinashsanjaygawale.pdf

[7] www.analyticsvidhya.com
[8] https://www.zippia.com/advice/wine-industry-statistics/
[9] https://www.analyticsvidhya.com/blog/2021/04/wine-quality-prediction-using-machine-learning/
[10] www.byjus.com
[11] www.colab.research.google.com
[12] www.scikit-learn.org
[13] www.researchgate.net
[14] www.iq.opengenus.org
[15] www.analyticsvidhya.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY