AN EFFECTIVE MACHINE LEARNING MODEL FOR WINE CLASSIFICATION

***Okikiola, F. M., Akinola, A. F., Ishola, P. E. and Onadokun, I .O.**

Department of Computer Technology, Yaba College of Technology, Yaba, Lagos State, Nigeria

***Corresponding author:** sade.mercy@yahoo.com

Manuscript received on the 8/10/22 and accepted on the 15/12/22

ABSTRACT

Wine companies spend a lot of money to test the quality of their wine because they need to buy specialized equipment and build elaborate winery labs to house it. Lab testing also takes a lot of time because it is so labour-intensive. Some people even go so far as to hire qualified taste consultants, which is an expensive alternative. In this instance, we created an effective machine learning model that can forecast wine quality based on some physicochemical traits, enabling the production and distribution of the highest quality wine. Using decision trees, random forests, and stochastic gradient descent. This was accomplished by using an industry-specific database for the wine-making industry. It was shown that the random forest strategy outperformed the other two strategies after training and testing on a set of dataa higher accuracy of 93%. This shows how wine companies can start saving money and making decisions that are much more informed.

Keywords: Decision trees, Random forests, stochastic gradient descent and Wine.

1. INTRODUCTION

Wine quality is a very crucial factor for the customer and the wine-making businesses. Wine is a widely consumed beverage nowadays, and companies use product quality certification to increase their marketability. Wine consumption has increased significantly over the past several years, not simply because it is enjoyable but also because of its heart-healthy qualities. Modern approaches and procedures are being used in various industries to boost output and enhance the effectiveness of the overall process. Both the price and the demand for these operations increase with time. Contrarily, while the chemicals required to manufacture wines are basically the same, they have a range of uses. We utilise these techniques to confirm since it is necessary to identify the kind of chemicals used. As a result, rating the quality of wine is a rather difficult task that requires outstanding knowledge, experience, and taste sensibilities (Maeve et al., 2022).

Testing for product quality used to be done at the conclusion of production. This process is time-consuming and expensive because it requires numerous human specialists to assess the product's quality and takes a lot of resources. Since every person has a different viewpoint on the test, it is challenging to evaluate the wine's quality based just on the opinions of others. Wine quality can be predicted using a variety of variables, however not all variables will lead to a more precise prediction. For the aim of quality control, it is essential to deduce wine quality from its chemical properties. Using the wine quality dataset and machine learning classification techniques such Stochastic Gradient Descent, Decision Tree, and Random Forest, the study determines the wine attributes necessary for a successful outcome.

In order to produce and distribute wine of the highest quality, an effective machine learning model that can forecast wine quality based on a few physicochemical criteria is provided in this work. We create a machine learning model to forecast wine quality, put it into practise, and assess the method that has been suggested. With the help of an ensemble machine learning algorithm, which uses classification machine learning techniques, it would be possible to assess the quality of various wines taken

from a common dataset. The user-friendly open system attribute that enables winemakers and buyers to quickly enter the physicochemical parameters of the wine and obtain a precise evaluation of the wine's predicted quality is not left out in this.

A wide variety of machine learning techniques are available for predicting wine quality. Stochastic Gradient Descent, Decision Trees, Logistic Regression, Random Forest, and Support Vector Machines are a few examples of machine learning techniques (SVM). The Wine Quality Dataset from Kaggle is only one of the many wine quality datasets that have been used in research on this subject. By removing and selecting different attributes from these databases, authors successfully come to a conclusion with their research. These studies are significant ones. In their study "A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality" published in 2021, Gupta and Vanmathi developed a model for predicting wine quality using a range of machine learning techniques, such as Support Vector Machines (SVM), Decision Trees, etc. Later, the authors used the multiple regression strategy to produce a final model.

The study "Prediction of Wine Quality Using Machine Learning Algorithms" by Dahal et al. (2021) demonstrated how statistical analysis may be used to identify the variables that have the greatest impact on wine quality before it is created. The effectiveness of the Ridge Regression (RR), Support Vector Machine (SVM), Gradient Boosting Regressor (GBR), and multi-layer Artificial Neural Network were evaluated by the authors to predict wine quality (ANN). There were many aspects of wine quality that were looked at. The analysis's findings showed that the GBR model outperformed all others, with MSE, R, and MAPE values of 0.3741, 0.6057, and 0.0873, respectively. In their article "Predictive modelling for wine authenticity using a machine learning technique" published in 2021, Da Costa et al. showed how to categorise wines from four different South American countries. The 83 samples that were gathered were tested for volatiles, semi-volatiles, and phenolic compounds. A classification decision-making system based on support vector machines (SVM) was created, along with correlation-based feature selection (CFS), random forest importance (RFI), which evaluates the relative importance of the input variables, and correlation-based feature selection (CFS). Using CFS, a subset of 190 chemical compounds' variables was selected. Thirteen compounds were selected as belonging to the group yet being unrelated to one another. The SVM was then used to classify these chemical elements, and they were subsequently placed in accordance with the RFI's importance ranking. According to the study, feature selection approaches and SVM were successfully combined to identify the most important compounds for classifying the wine samples. With a classification accuracy of 93.97%, the variable subset identified by the feature selection methods, which included catechin, gallic, octanoic acid, myricetin, caffeic, isobutanol, resveratrol, kaempferol, and ORAC.

Using data for red wine varietals of Portuguese Vinho Verde wine that were collected from the UCI machine learning collection, Sangodkar and Bapat (2021) "Wine Quality Prediction Using Machine Learning" was carried out. Wine industries were able to forecast wine quality using the data. Particularly, 30% of the data are used for testing, while 70% are used for training. Among the imported libraries were random, numpy, and pandas. SVM, Back Propagation Neural Network, Random Forest, and a 2*2 confusion matrix were used to extract PCA features, create the summary, and rate the dataset's quality. A number of performance indicators, such as specificity, precision, recall, F1-score for both models, and accuracy, were also determined for all three models: neural back propagation, random forest. In light of this, the outcomes were anticipated. This study utilising the UCI red wine dataset shows that random forest fits better than SVM model and BP neural network with accuracy of 80.9 percent.

In his paper "Red wine quality prediction with active learning" published by Tingwei

(2021), Tingwei gave an example of classification performed using active learning and a query. The data set for this study contained information on the quality of 1600 bottles of red wine as evaluated by famous wine tasters. By combining markers such as fixed acidity, chlorides, and density, the author built a feature matrix. The author then categorised those with a quality of at least 7 as good characteristics, those with less than 7, and took quality as outcome of the forecast. In order to improve prediction accuracy, the author picked the K-Nearest Neighbor technique for predictive modelling, and batch-mode sampling was rated as the best active learning algorithm.

A study employing a wine quality dataset obtained from the UCI machine learning library was conducted by Kumar et al. in 2020 for "Red Wine Quality Prediction Using Machine Learning Techniques." Various machine learning techniques were applied to the dataset in this study using the RStudio software. Using probabilities of 0.7 and 0.3, the data was split during use into a training set and a testing set. The results show that the accuracy of the Naive Bayes algorithm for the training set and testing set was 55.91 percent and 55.89 percent, respectively. The SVM method produced accuracy of 67.25 and 68.64 percent, respectively. The training set's high probability of division (0.7) led to the conclusion that the Support Vector Machine algorithm had the best performance followed by the Random Forest technique, and the Naive Bayes approach in third place. Wine Quality Prediction Using Machine Learning Algorithms, Pawar et al., 2019. With the aid of various machine learning techniques, such as Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, and Random Forest, the researchers focus on categorising wine quality. Exploratory data analysis was performed on the dataset, and they were able to employ random forest to achieve a maximum accuracy of 88 percent. Accuracy is 81% when using a stochastic gradient descent. SVC is 86% accurate, while logistic regression is 86% accurate.

The wine's quality and kind are assessed after the algorithm has been run before the result is decided. In their article titled "A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques," Aich et al. (2018) compared the performance metrics of various classifiers with different feature sets to predict the quality of different types of wine while accounting for various factors. The study developed a novel approach for measuring performance measures by considering different feature selection algorithms, such as Principal Component Analysis (PCA), Recursive Feature Elimination (RFE), and nonlinear decision tree-based classifiers. We found accuracy ranges of 94.51 percent to 97.79 percent for various feature sets using the Random Forest classifier. . Cortez et al. (2009) “A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques” described a revolutionary method that simultaneously selects variables and models for NN and SVM techniques. Sensitivity analysis, a computationally effective technique that assesses input relevance and directs the variable selection process, serves as the foundation for the variable selection process. The study also suggests a low-effort parsimony search technique to choose the optimal SVM kernel parameter. In some of the study, gradient boosting performed better, however if we can expand the training datasets, it is possible to make use of ANN's superior prediction capabilities. The SVM hyperplane was not adjusted therefore its accuracy remains lesser than other algorithms used

2. METHODOLOGY

Python was used to develop the suggested model in order to solve the machine learning issue. Through the usage of its frontend and backend development tools, Streamlit (a template for HTML, CSS, and JavaScript) will be fully responsible for designing the user interface for the online application. The database was likewise powered by MySQL. Heroku, a platform for deploying web

applications from the cloud, was used for the final deployment. To properly understand the various machine learning approaches and how they are employed, a basic literature review of prior studies will be undertaken after the first set of goals has been accomplished. Additionally, using the dataset of wines gathered from the UCI machine learning library, a wine quality prediction model is created (Cortez et al., 2009). It is anticipated that the collection will have examples from 6497 different wines. Eleven physicochemical characteristics, such as density (g/ml), pH, sulphates (g (potassium sulphate)/L), and alcohol. Fixed acidity (g (tartaric acid)/L), volatile acidity (g (acetic acid)/L), citric acid (g/L), residual sugar (g/L), chlorides (g (sodium chloride)/L), and free sulphur dioxide (mg/L), total sulphur dioxide (mg / (percent vol.) The wine quality is rated on a scale of 0 to 10. Quality should always come first with any product. The greatest attention and best techniques must be utilised to assess the product's quality. To gauge the quality of a wine, a mechanism for forecasting its quality based on its components and acidity levels should be created (Kavana et al., 2020).

2.1 Research Design

To effectively make the system user-friendly and valuable enough to encourage adherence to the system, we apply user-centered design. The following steps can be used to breakdown the iterative design process:

i. Create the system's first user interface for testing.
ii. Demonstrate the system to a range of potential users.
iii. Ask for comments and record any issues users may be having with the system.
iv. Modify the system to address the issues mentioned in step 3

Continue performing steps (ii - iv) until user complaints are brought down to a manageable level and a better system is developed

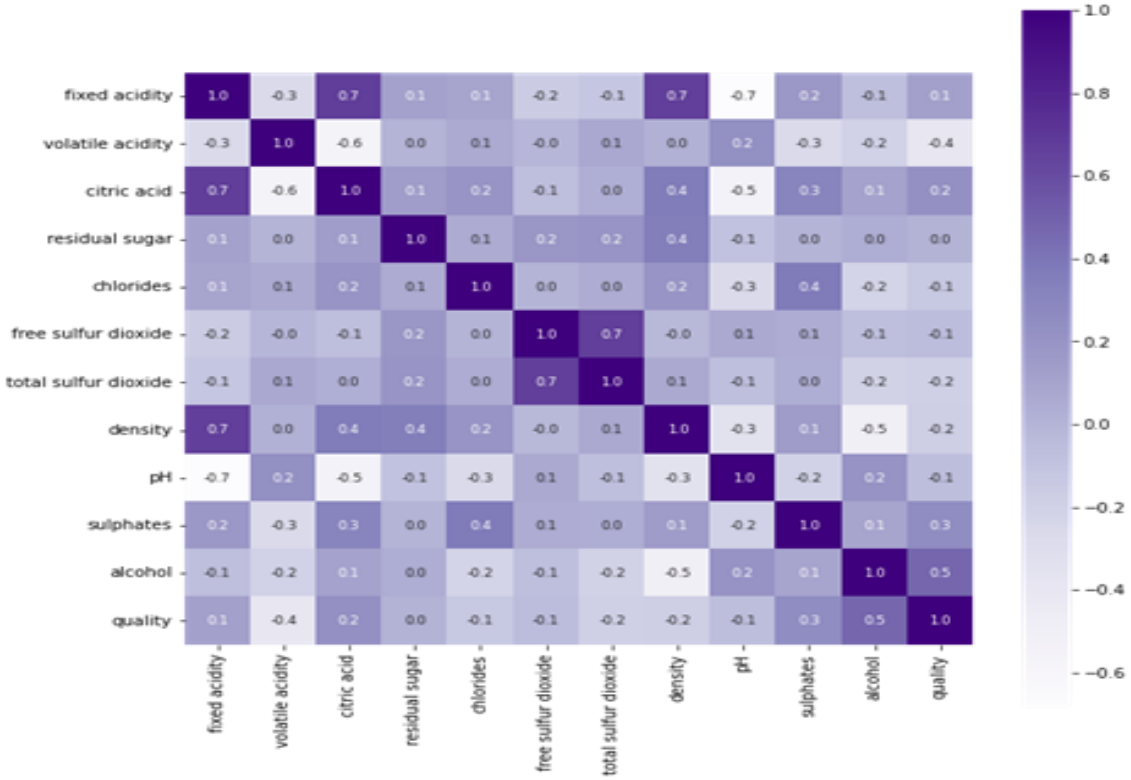
2.2 Data Collection

The first step was acquiring the dataset because we needed a wine data set with labelled chemical parameter values in order to train our machine learning model. The UCI machine learning repository offers free access to the wine quality dataset (Cortez et al., 2009). Two files—one for red wine and the other for white wine varieties of the Portuguese "Vinho Verde" wine—make up the dataset. The machine learning community makes use of its huge amount of datasets. There are 1599 cases in the red wine dataset and 4898 in the white wine dataset. Eleven input and one output characteristics are present in both files. The output variable is based on sensory data and scaled from 0 to 10, while the input attributes are based on physicochemical tests (0-very bad to 10-very good). Supervised learning is the sort of learning that takes place when you use labelled datasets to train a machine learning model, while unsupervised learning is the type of learning that takes place when you use an unlabelled data set. The next step was to seek for any missing data, but none were found. None of the values needed to be changed.

2.3 Data Analysis and Visualization

After receiving the great dataset, we performed some data analysis on it. The many properties of the value could also be connected in numerous ways. The association between citric acid concentration and wine quality would need to be established, for example, if we were to study the values for citric acid and content acidity. To do this, among other things, we would need to ascertain whether the wine's quality rises as the amount of citric acid does. A correlation matrix is used in this study to show how each input variable is related to the others. Simply described, a correlation matrix is a table that shows the correlation coefficients for various variables. The matrix shows each possible pairing of values in a table and their correlation. The tasks listed above were completed throughout the data analysis phase, and we also employed a number of visualisation techniques, including plots and graphs, to better understand the dataset at hand. A correlation matrix that shows the relationship between each input variable in this study is presented.

To put it simply, a correlation matrix is a table showing the correlation coefficients for various variables. The matrix shows the relationship between all possible value pairings in a table. The correlation matrix of the proposed model is displayed in Fig. 1



FigureI: Heat map of correlation for the proposed system

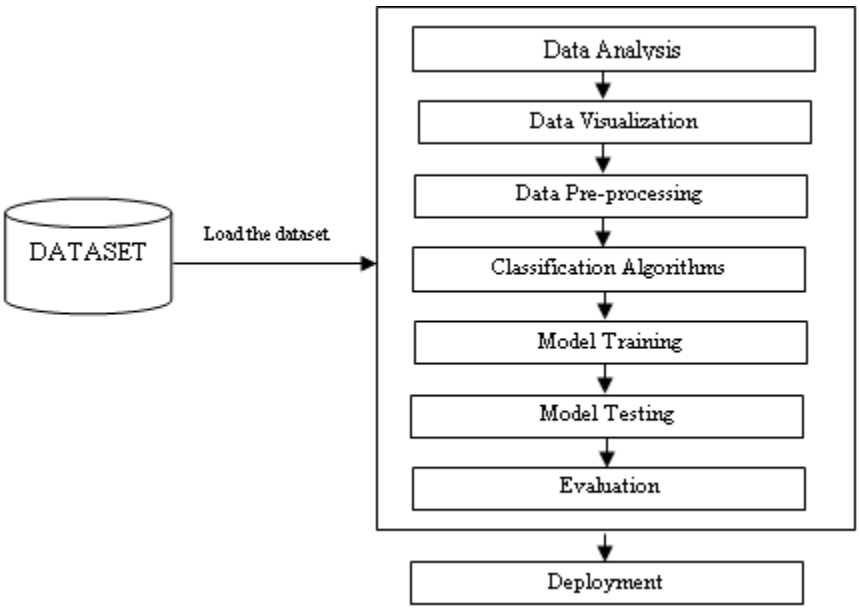
Data Preprocessing

Feature selection is the most popular method of preparing data before building a predictive model. To build the model, it selects the relevant features from the subset. According to the relevance of the features' weighted average, characteristics with an abnormally low weighting will be removed. This method will simplify the model, which will reduce training time and enhance model performance (Panday et al., 2018).We concentrate on feature selection because it affects how the study will be conducted. When assessing the performance of our model, accuracy, precision, recall, and f1 score are trustworthy indicators of the model's effectiveness. Wine quality is determined by physical and chemical tests as well as sensory evaluation. Physical and chemical tests have properties like density, alcohol content, and pH values. Sensory testing is a very time-consuming and expensive technique for human taste experts. We reviewed the literature and identified eleven physicochemical properties that practically all data sets used to evaluate the quality of wines have in common.Density, pH, sulphates, free sulphur dioxide, total sulphur dioxide, chlorides, citric acid, residual sugar, fixed acidity, volatile acidity, sulphates, and alcohol are the eleven physicochemical properties (Cortez et al., 2009). In this case, we separated the label from the data. In this case, the label is this quality column. We had to split all the data under this quality column since we will enter this data individually into our machine learning model; as a result, we put all the data in one variable under this quality column. We also used label binarization, which essentially means that quality values that are greater than or equal to 7 are replaced with 1 and quality values that are less than or equal to 6 are replaced with 0. Since numerical numbers are always better for processing, we decided to utilize the labels 1 and 0 for our data.

Proposed System Model

The suggested model, represented in Figure II, details the classification model employing Stochastic Gradient Descent

(SGD), Decision Tree (DT), and Random Forest algorithms (RF).



FigureII:Proposed model for the research

3. RESULTS

In this section, the process of classification is show in a step-wise manner. The actions were conducted to build the model and train the classification algorithm. Figures 3-18 show the instances of implementation of the process.

Step 1: Import dependencies

```
[1]  import numpy as np
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns
     from sklearn.model_selection import train_test_split
     from sklearn.tree import DecisionTreeClassifier
     from sklearn.linear_model import SGDClassifier
     from sklearn.ensemble import RandomForestClassifier
     from sklearn.metrics import classification_report
     from sklearn.metrics import accuracy_score
```

FigureIII: Code to import dependencies

Step 2: Import dataset to be used in ML

```
[2]  # loading the dataset to a Pandas DataFrame
     wine_dataset = pd.read_csv('/content/winequality-red.csv')
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

FigureIV: First 5 rows of the dataset

Step 3: Analyze the data

```
[ ]  # statistical measures of the dataset
     wine_dataset.describe()
```

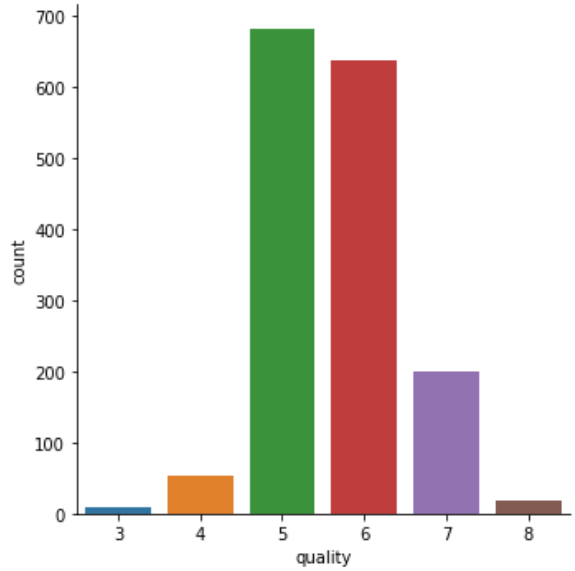
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

FigureV: Describing the dataset using statistical measures

Step 4: Visualize the data

```
# number of values for each quality
sns.catplot(x='quality', data = wine_dataset, kind = 'count')
```

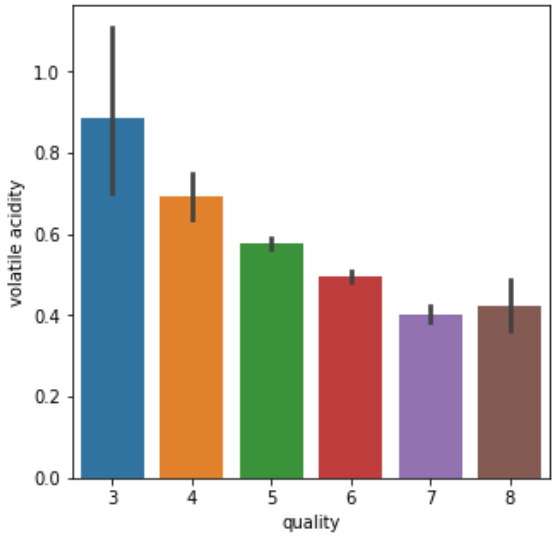
```
<seaborn.axisgrid.FacetGrid at 0x7f0c99bfe290>
```



FigureVI: Displaying values for each quality

```
# volatile acidity vs Quality
plot = plt.figure(figsize=(5,5))
sns.barplot(x='quality', y = 'volatile acidity', data = wine_dataset)
```

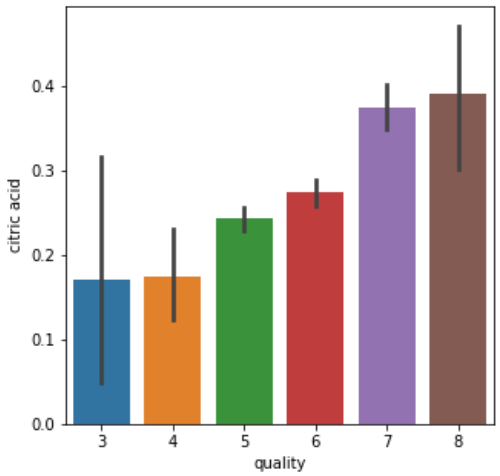
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0c99c2d1d0>
```



FigureVII:Comparing volatile acidity and quality

```
# citric acid vs Quality
plot = plt.figure(figsize=(5,5))
sns.barplot(x='quality', y = 'citric acid', data = wine_dataset)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f0c957cf510>
```



FigureVIII: Comparing citric acid and quality

Step 5: Perform data preprocessing (and label binarization)

Data Preprocessing

```
[ ] # separate the data and Label
    X = wine_dataset.drop('quality',axis=1)
```

```
[ ] print(X)
```

```
      fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  \
0               7.4             0.700         0.00             1.9      0.076
1               7.8             0.880         0.00             2.6      0.098
2               7.8             0.760         0.04             2.3      0.092
3              11.2             0.280         0.56             1.9      0.075
4               7.4             0.700         0.00             1.9      0.076
...             ...               ...          ...             ...        ...
1594            6.2             0.600         0.08             2.0      0.090
1595            5.9             0.550         0.10             2.2      0.062
1596            6.3             0.510         0.13             2.3      0.076
1597            5.9             0.645         0.12             2.0      0.075
1598            6.0             0.310         0.47             3.6      0.067

      free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  \
0                    11.0                  34.0  0.99780  3.51       0.56
1                    25.0                  67.0  0.99680  3.20       0.68
2                    15.0                  54.0  0.99700  3.26       0.65
3                    17.0                  60.0  0.99800  3.16       0.58
4                    11.0                  34.0  0.99780  3.51       0.56
...                   ...                   ...      ...   ...        ...
1594                 32.0                  44.0  0.99490  3.45       0.58
1595                 39.0                  51.0  0.99512  3.52       0.76
1596                 29.0                  40.0  0.99574  3.42       0.75
1597                 32.0                  44.0  0.99547  3.57       0.71
1598                 18.0                  42.0  0.99549  3.39       0.66

      alcohol
0         9.4
1         9.8
2         9.8
3         9.8
4         9.4
...       ...
1594     10.5
1595     11.2
1596     11.0
1597     10.2
1598     11.0

[1599 rows x 11 columns]
```

FigureIX: Code and output for data preprocessing

Label binarization

```
[ ] Y = wine_dataset['quality'].apply(lambda y_value: 1 if y_value>=7 else 0)
```

```
[ ] print(Y)
```

```
0       0
1       0
2       0
3       0
4       0
       ..
1594    0
1595    0
1596    0
1597    0
1598    0
Name: quality, Length: 1599, dtype: int64
```

FigureX:Code for label binarization or encoding

Step 6: Separate training and test sets of data

```
[ ] X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=3)

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

    (1599,) (1279,) (320,)
```

FigureXI: Train & Test split

Step 7: Train & test algorithms
In this stage, we will find the best of the three algorithms;

Stochastic Gradient Descent (SGD)

```
[ ] sgd = SGDClassifier()
    sgd.fit(X_train, Y_train)
    pred_sgd = sgd.predict(X_test)
    print(classification_report(Y_test, pred_sgd))
```

	precision	recall	f1-score	support
0	0.97	0.68	0.80	283
1	0.26	0.86	0.40	37
accuracy			0.70	320
macro avg	0.62	0.77	0.60	320
weighted avg	0.89	0.70	0.76	320

FigureXII:Code snippet for training SGD

Decision Tree (DT)

```
[ ] dt = DecisionTreeClassifier()
    dt.fit(X_train, Y_train)
    pred_dt = dt.predict(X_test)
    print(classification_report(Y_test, pred_dt))
```

	precision	recall	f1-score	support
0	0.95	0.94	0.94	283
1	0.55	0.59	0.57	37
accuracy			0.90	320
macro avg	0.75	0.77	0.76	320
weighted avg	0.90	0.90	0.90	320

FigureXIII**:** Code snippet for training DT

Random Forest (RF)

```
[ ] rf = RandomForestClassifier()
    rf.fit(X_train, Y_train)
    pred_rf = rf.predict(X_test)
    print(classification_report(Y_test, pred_rf))
```

	precision	recall	f1-score	support
0	0.95	0.98	0.96	283
1	0.78	0.57	0.66	37
accuracy			0.93	320
macro avg	0.86	0.77	0.81	320
weighted avg	0.93	0.93	0.93	320

FigureXIV: Code snippet for training RF

The model to employ in order to get the best results is Random Forest, which had the highest accuracy out of the aforementioned code snippets.

Step 8: Train the best model

```
[ ]  model = RandomForestClassifier()
```

```
[ ]  model.fit(X_train, Y_train)

     RandomForestClassifier()
```

FigureXV: Training the RF model

Step 9: Evaluate the model
After training the Random Forest model, which was the most suitable model with the highest accuracy, we evaluated the model to further test its accuracy on test data.

```
[ ]  # accuracy on test data
     X_test_prediction = model.predict(X_test)
     test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
```

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

     Accuracy :  0.928125
```

FigureXVI:Assessing the model's precision using test data

In comparison to other models, the accuracy of the Random Forest model was about 93%; Decision Tree's accuracy was 90%, and Stochastic Gradient Descent's accuracy was 70%. Consequently, the Random Forest model was the most suitable for the task.

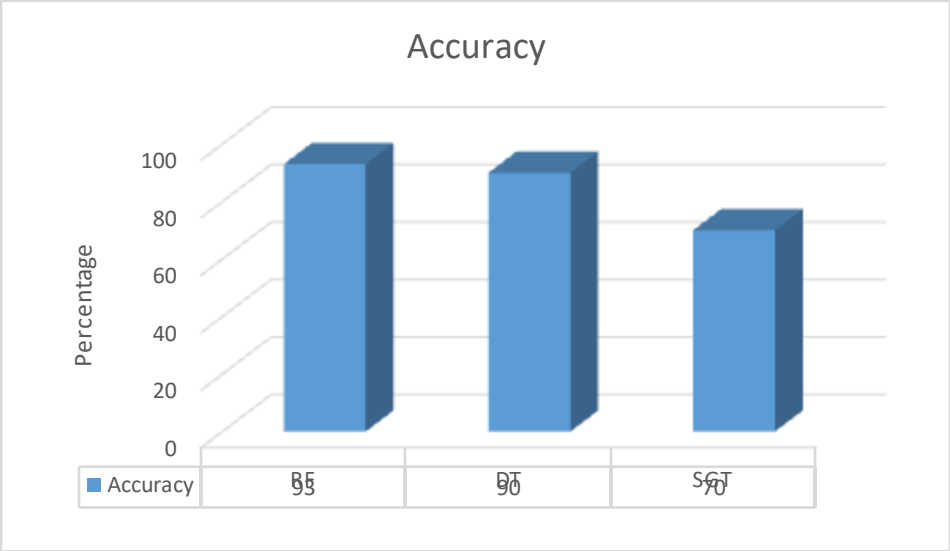


Figure XVII: Comparative evaluation of the machine learning techniques.

Step 10: Build a predictive system

```
[ ]  input_data = (7.5,0.5,0.36,6.1,0.071,17.0,102.0,0.9978,3.35,0.8,10.5)

     # changing the input data to a numpy array
     input_data_as_numpy_array = np.asarray(input_data)

     # reshape the data as we are predicting the label for only one instance
     input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

     prediction = model.predict(input_data_reshaped)
     print(prediction)

     if (prediction[0]==1):
       print('Good Wine')
     else:
       print('Bad Wine')

     [0]
     Bad Wine
```

FigureXVII: Predictive system for wine quality

4. DISCUSSION

To evaluate the effectiveness of different ML techniques, we employed three common machine learning algorithms to predict wine quality in the wine dataset: Stochastic Gradient Descent (SGD), Decision Tree Classifiers (DT), and Random Forest Classifiers (RF). This gives us the option to choose the best ML method for predicting wine quality based on the factors we've provided.The classification procedure was evaluated using measures such as training accuracy, precision, recall, f1-score, and confusion matrix.

To evaluate our result, we plotted a confusion matrix for the three algorithms used in this work using the Python library, matplotlib.

```
[[283   0]
 [ 35   2]]
<matplotlib.axes._subplots.AxesSubplot at 0x7f607e825490>
```

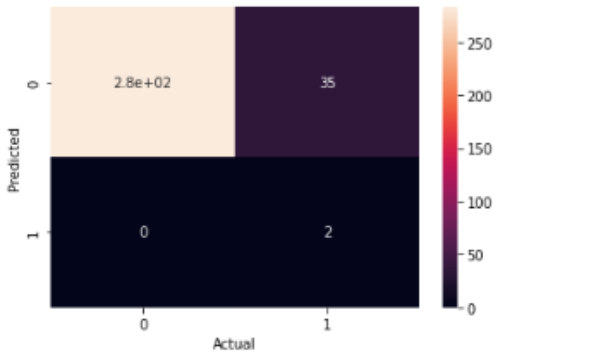
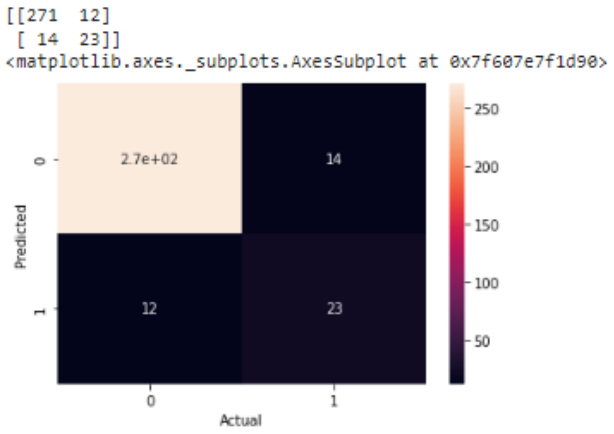
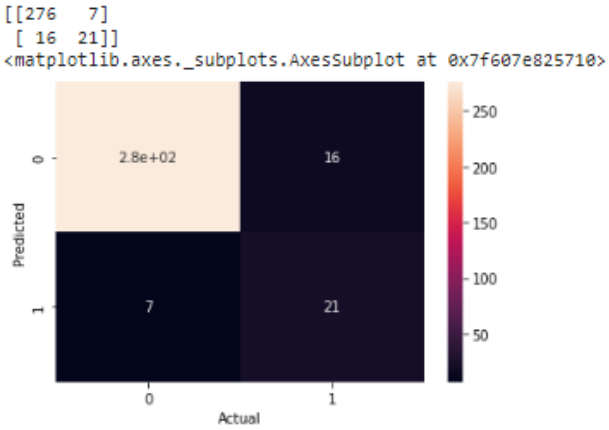


Figure XVIII:Confusion matrix for SGD



FigureXIX:Confusion matrix for DT



FigureXX:Confusion matrix for RF

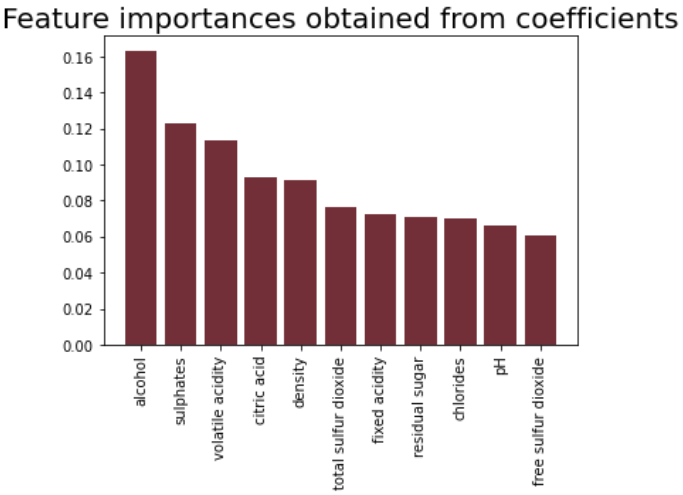
From the above, we can see three separate confusion matrixes for performance evaluation

Table I summarizes the findings from the three machine learning models that were utilized in this study

Table I: Summary of model evaluation

MODEL	PRECISION	SENSITIVITY	SPECIFICITY	ACCURACY (%)
Stochastic Gradient Descent	0.89	1.00	0.05	0.89
Decision Tree	0.95	0.96	0.62	0.92
Random Forest	0.95	0.98	0.57	0.93

Finally, all three machine learning algorithms created very accurate wine quality prediction models, with Random Forest being the most accurate. In order to better comprehend our outcome, we plotted a feature importance graph from the model with the best accuracy.



FigureXXI: Feature importance graph derived from Random Forest model

FigureXXI depicts the impact of RF factors in determining wine quality. When we plot the feature relevance of all features for our RF model, we observe that alcohol is the most essential feature for controlling wine quality, makes perfect sense because it affects the taste, texture, and structure of the wine itself, not just the sensations after drinking. The sulphates are the second most significant feature, and they are, by definition, partly connected to the first. The least essential attribute, according to the figure XXI, is free Sulphur(iv)oxide, a measure of the amount of SO2 (Sulphur(iv)oxide) used to prevent oxidation and microbiological development during the winemaking process.

5. CONCLUSION

The implementation of an effective machine learning model for classifying wine is done in this paper. This is purposed to help several wine companies lower the cost of testing during wine production and shorten the time required to conduct laboratory tests because most of the time, these tests are implemented at the end of production, or after the wine has been made and is prepared for sale. The firm then finds a way to test, either by hiring a consultant experienced in determining the quality of wine through testing, or by using a variety of testing tools that are highly expensive to buy and will also take a sizable amount of time for the company to employ. The RF, SGD and DT machine learning techniques were employed to train wine feature selection attributes. The density, pH, sulphates, free sulphur dioxide, total sulphur dioxide, chlorides, citric acid, residual sugar, fixed acidity, volatile acidity, sulphates, and alcohol are the physiochemical properties that are used. The RF had a higher accuracy of 93% than the other techniques at the conclusion of the technique evaluation. Our objective is for wine firms to use our trained model to examine how a small change to any component might affect the wine's qualityThey could swiftly decide which physicochemical characteristics of the wine were important and which ones might be ignored to reduce costs. They may also swiftly ascertain the precise quantity of each physicochemical component needed to raise the calibre of their wine and use it properly.

REFERENCES

Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T., & Sain, M. (2018). A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques.20th International Conference on Advanced Communication Technology (ICACT). https://doi.org/10.23919/icact.2018.8323673

Astray, G., Mejuto, J., Martínez-Martínez, V., Nevares, I., Alamo-Sanza, M., & Simal-Gandara, J. (2019). Prediction Models to Control Aging Time in Red Wine. *Molecules*, *24*(5), 826. https://doi.org/10.3390/molecules24050826.

Barbosa, C., Ramalhosa, E., Vasconcelos, I., Reis, M., & Mendes-Ferreira, A. (2022). Machine Learning Techniques Disclose the Combined Effect of Fermentation Conditions on Yeast Mixed-Culture Dynamics and Wine Quality. *Microorganisms*, *10*(1), 107. https://doi.org/10.3390/microorganisms10010107

Brand, J., Panzeri, V., &Buica, A. (2020). Wine Quality Drivers: A Case Study on South African Chenin Blanc and Pinotage Wines. *Foods*, *9*(6), 805. https://doi.org/10.3390/foods9060805

Copeland, M. (2021, July).*The Difference Between AI, Machine Learning, and Deep Learning*.https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modelling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4),547–553. https://doi.org/10.1016/j.dss.2009.05.016

Da Costa, N. L., Valentin, L. A., Castro, I. A., & Barbosa, R. M. (2021). Predictive modelling for wine authenticity using a machine learning approach. *Artificial Intelligence in Agriculture*, *5*, 157–162. https://doi.org/10.1016/j.aiia.2021.07.001

Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of Wine Quality Using Machine Learning Algorithms. *Open Journal of Statistics*, *11*(02), 278–289. https://doi.org/10.4236/ojs.2021.112015

Danner, L., Scientific, T. C., Ristic, R., & Johnson, T. (2016). Context and wine quality effects on consumers'mood,emotions, liking and willingness to pay for Australian Shiraz wines Context and wine quality effects on consumers ' mood , emotions , liking and willingness to pay for Australian Shiraz wines. . https://doi.org/10.1016/j.foodres.2016.08.006

Dornadula, V. N., & Geetha, S. (2019). Credit Card Fraud Detection using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 631–641. https://doi.org/10.1016/j.procs.2020.01.057

Eling, M., Nuessle, D., &Staubli, J. (2021). The impact of artificial intelligence along the insurance value chain and on the insurability of risks. *The Geneva Papers on Risk and Insurance - Issues and Practice*, *47*(2), 205–241. https://doi.org/10.1057/s41288-020-00201-7

Enginess. (2019, March). *Machine Learning: Everything You Need to Know*. Enginess Insights. https://www.enginess.io/insights/machine-learning-everything-you-need-to-know

Gupta, M., & Vanmathi, C. (2021). A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality. International Journal of Recent Technology and Engineering (IJRTE), 10(1), 314–319. https://doi.org/10.35940/ijrte.a5854.0510121

Kavana, Prasannavenkatesan, T., Achanta, S., Sufiyan, A., & Sushmitha. (2020). An Investigation of Wine Quality Testing using Machine Learning Techniques. *Journal of Xidian University*, *14*(4). https://doi.org/10.37896/jxu14.4/408

Kumar, S., Agrawal, K., & Mandan, N. (2020). Red Wine Quality Prediction Using Machine Learning Techniques. International Conference on Computer Communication and Informatics (ICCCI). https://doi.org/10.1109/iccci48352.2020.9104095

Lei, Y., Peng, Q., & Shen, Y. (2020). Deep Learning for Algorithmic Trading. *Proceedings of the* 6th International Conference on Computing and Artificial Intelligence. https://doi.org/10.1145/3404555.3404604

Li, H., Wang, H., Li, H., Goodman, S., van der Lee, P., Xu, Z., Fortunato, A., & Yang, P. (2018). The worlds of wine: Old, new and ancient. *Wine Economics and Policy*, *7*(2), 178–182. https://doi.org/10.1016/j.wep.2018.10.002

McMillan, R. (2022). State of the US Wine Industry 2022. Silicon Valley Bank Wine Division. https://www.svb.com/globalassets/trendsandinsights/reports/wine/svb-state-of-the-wine-industry-report-2022.pdf

Maeve Eunicia, Richie Skyszygfrid, Tiara Vitri, & Vicky Caren. (2022). Modeling Red Wine Quality Based on Physicochemical Tests: A Data Mining Approach. Formosa Journal of Multidisciplinary Research, 1(1), 89–110. https://doi.org/10.55927/fjmr.v1i1.414

Mattivi, F., Arapitsas, P., &Perenzoni, D. (2015). Influence of Storage Conditions on the Composition of Influence of Storage Conditions on the Composition of Red Wines. https://doi.org/10.1021/bk2015-1203.ch003

Paltrinieri, N., Comfort, L., & Reniers, G. (2019). Learning about risk: Machine learning for risk assessment. Safety Science, 118, 475–486. https://doi.org/10.1016/j.ssci.2019.06.001

Panday, D., Cordeiro De Amorim, R., & Lane, P. (2018). Feature weighting as a tool for unsupervised feature selection. *Information Processing Letters*, *129*, 44–52. https://doi.org/10.1016/j.ipl.2017.09.005

Pawar, D., Mahajan, A., &Bhoithe, S. (2019). Wine Quality Prediction using Machine Learning Algorithms. International Journal of Computer Applications Technology and Research, 8(9), 385–388. https://doi.org/10.7753/ijcatr0809.1010

Pyle, D., & José, C. S. (2019, February). An executive's guide to machine learning. McKinsey & Company.https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/an-executives-guide-to-machine-learning

Sangodkar, V. P., & Bapat, U. A. (2021). Wine Quality Prediction Using Machine Learning. International Journal for Research in Applied Science and Engineering Technology, 9(8), 1997–2001. https://doi.org/10.22214/ijraset.2021.37629

Sen, J. (Ed.). (2021). Machine Learning - Algorithms, Models and Applications. Artificial Intelligence. https://doi.org/10.5772/intechopen.94615

Sowmya, D., Sayyed, J., Ganavi, M., & Sankhya, N. (2019). Analysing Wine types and Quality using Machine Learning Techniques. International Journal of Engineering Applied Sciences and Technology, 4(3), 519–529. Https://Doi.Org/10.33564/Ijeast.2019.V04i03.080

Haptik,T (2022, February). 9 Best Chatbots in the Financial Services Industry. Haptik. https://www.haptik.ai/blog/best-chatbots-in-financial-industry

Thach, M. L. W. (2021, February). Overview of 2020 US Wine Market Stats and 10 Hot Wine Trends for 2021. https://lizthachmw.com/winestars/overview-of-2020-us-wine-market-stats-and-10-hot-wine-trends-for-2021/

Thach, M. L. W. (2022, April). U.S. Wine Market Sales Up 16.8% In 2021, Pointing Towards Hot Wine Trends In 2022. https://www.forbes.com/sites/lizthach/2022/04/25/us-wine-market-sales-up-168-in-2021-pointing-towards-hot-wine-trends-in-2022/?sh=6506e8f043d2

Tingwei, Z. (2021). Red wine quality prediction through active learning. Journal of Physics: Conference Series, 1966(1), 012021. https://doi.org/10.1088/1742-6596/1966/1/012021

Vilela, A. (2018). Lachancea thermotolerans, the Non-Saccharomyces Yeast that Reduces the Volatile Acidity of Wines. Fermentation, 4(3), 56. https://doi.org/10.3390/fermentation4030056

World Health Organization. (n.d.). Global status report on alcohol and health. https://www.who.int/substance_abuse/publications/global_alcohol_report/msbgsruprofiles.pdf

Yu, S., Chen, Y., & Zaidi, H. (2021). AVA: A Financial Service Chatbot Based on Deep Bidirectional Transformers. Frontiers in Applied Mathematics and Statistics, 7. https://doi.org/10.3389/fams.2021.604842