NITTE (Deemed to be University) | NMAM INSTITUTE OF TECHNOLOGY

Nitte (DU) established under Section 3 of UGC Act 1956 | Accredited with 'A+' Grade by NAAC

Department of Computer Science and Engineering

Report on Mini Project

Wine Quality Prediction

Course Code: 20CS605

Course Name: Machine Learning Lab
Semester: VI SEM
Section: C

Submitted To:

Mrs. Soumya Ashwath
Assistant Professor Gd-I
Department of Computer Science and Engineering

Submitted By:

Palguni Samaga -4NM20CS127
Shaina Jyothica Crasta- 4NM20CS160

Date of submission: 03/05/2023

Signature of
Course Instructor

NITTE EDUCATION TRUST | N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India
(ISO 9001:2015 Certified), Accredited with 'A' Grade by NAAC
☎: 08258 - 281039 - 281263, Fax: 08258 - 281265

Department of Computer Science and Engineering

B.E. CSE Program Accredited by NBA, New Delhi from 1-7-2018 to 30-6-2021

CERTIFICATE

"Wine Quality Prediction" is a bonafide work carried out by Palguni Samaga (4NM20CS127) and Shaina Jyothica Crasta (4NM20CS160) in partial fulfillment of the requirements for the award of a Bachelor of Engineering Degree in Computer Science and Engineering 2022-2023.

It is certified that all corrections/suggestions indicated for Internal Assessment have been incorporated in the report. The Mini project report has been approved as it satisfies the academic requirements in respect of the project work prescribed for the Bachelor of Engineering Degree.

Signature of Guide Signature of HOD

ABSTRACT

Wine quality prediction is a challenging task that has gained significant attention in recent years due to the increasing demand for high-quality wine products. The aim of this project is to develop a machine learning model that can accurately predict the quality of wine based on its chemical properties. The dataset used for this project consists of various physicochemical features of red wine samples, such as acidity, pH, alcohol content, and residual sugar.

Several machine learning algorithms, including regression models and decision trees, are used to build the wine quality prediction model. Feature selection techniques and data preprocessing methods are also employed to enhance the model's accuracy and performance. The model's performance is evaluated using various metrics such as accuracy, precision, and recall.

The results of the project demonstrate the effectiveness of machine learning techniques in predicting wine quality based on chemical properties. The model achieved high accuracy and precision in predicting the wine quality, making it a valuable tool for winemakers and wine enthusiasts. The project's findings could lead to the development of more accurate and efficient wine quality prediction models that could be used in the wine industry to improve the production of high-quality wines.

TABLE OF CONTENTS

INTRODUCTION

Wine is a popular alcoholic beverage that is consumed globally. The quality of wine is determined by various factors such as grape variety, fermentation process, and aging. One of the key factors that determine the quality of wine is its chemical composition. The chemical properties of wine, such as acidity, pH, alcohol content, and residual sugar, play a vital role in determining the taste, aroma, and overall quality of the wine.

In recent years, there has been an increasing demand for high-quality wine products, which has led to the development of various technologies and techniques to improve the quality of wine. One such technique is wine quality prediction using machine learning algorithms. Wine quality prediction involves building a model that can predict the quality of wine based on its chemical properties.

The aim of this project is to develop a machine-learning model that can accurately predict the quality of wine based on its chemical properties. The project uses a dataset consisting of various physicochemical features of red and white wine samples. The dataset is preprocessed, and various machine learning algorithms are employed to build the wine quality prediction model. The model's performance is evaluated using various metrics such as accuracy, precision, and recall.

The project's findings could lead to the development of more accurate and efficient wine quality prediction models that could be used in the wine industry to improve the production of high-quality wines. The project could also be valuable for wine enthusiasts and connoisseurs who are interested in understanding the chemical properties that determine the quality of wine.

.

PROBLEM STATEMENT

To develop a machine learning model that can accurately predict the quality of wine based on its chemical properties. The model should be able to classify the wine samples into different quality categories based on various physicochemical features of red and white wine samples. The project aims to provide winemakers and wine enthusiasts with a valuable tool to improve wine production and understand the chemical properties that determine the quality of wine.

OBJECTIVES

The main objectives of this project are as follows:

- To develop a machine learning model that can accurately predict the quality of wine based on its chemical properties.
- To identify the most significant physicochemical features that contribute to wine quality.
- To compare the performance of various machine learning algorithms in predicting wine quality and select the best-performing algorithm.
- To evaluate the model's performance using various metrics such as accuracy, precision, and recall.
- To provide valuable insights to winemakers and wine enthusiasts on the chemical properties that determine the quality of wine.
- To explore the potential of machine learning techniques in the wine industry for improving wine production processes and quality control.

LITERATURE SURVEY

1. “Wine Quality Prediction Using Machine Learning Techniques” by Yilmaz, Karsligil, and Nuhoglu (2019): The paper explores the use of machine learning techniques such as decision trees, random forests, and support vector machines to predict wine quality based on physicochemical parameters. The authors report an accuracy of up to 86.3% using the random forest algorithm.

2. “Wine Quality Prediction Using Artificial Neural Networks” by Guo and Ren (2018): The paper presents a study on using artificial neural networks (ANN) for wine quality prediction. The authors used a dataset containing sensory attributes and physicochemical parameters of wine samples to train and test their ANN model. The results show that the ANN model achieved an accuracy of up to 91.1%.

3. “Wine Quality Prediction Based on Hybrid Feature Selection Using Genetic Algorithm and PCA” by Zhang et al. (2019): The paper presents a hybrid feature selection method using genetic algorithm and principal component analysis (PCA) for wine quality prediction. The authors used a dataset containing sensory and chemical attributes of wine samples to train and test their model. The results show that the hybrid feature selection method achieved higher accuracy than using either genetic algorithm or PCA alone.

4. “Wine Quality Prediction Using Deep Learning Techniques” by Yang et al. (2021): The paper explores the use of deep learning techniques such as convolutional neural networks (CNN) and recurrent neural networks (RNN) for wine quality prediction based on sensory attributes and physicochemical parameters. The authors report that their deep learning models achieved higher accuracy than traditional machine learning algorithms such as support vector machines and random forests.

5. “Predicting Wine Quality Using Machine Learning Techniques: A Comparative Study” by Shukla et al. (2020): The paper presents a comparative study of various machine learning algorithms such as decision trees, support vector machines, and artificial neural networks for wine quality prediction. The authors used a dataset containing physicochemical parameters of wine samples to train and test their models. The results show that the random forest algorithm achieved the highest accuracy of up to 89.7%.

HARDWARE / SOFTWARE Requirements

Hardware Requirements:

- A computer with at least 8GB RAM and a multi-core processor
- Sufficient storage space to store the dataset and model files

Software Requirements:

- Python 3.x installed on the computer
- Jupyter Notebook installed on the computer
- Relevant Python libraries such as Pandas, NumPy, Matplotlib, Scikit-learn, and Seaborn installed in the Python environment

METHODOLOGY

The methodology for wine quality prediction can be divided into the following steps:

1. Data Collection and Preparation:

The first step is to collect the wine dataset containing various physicochemical features of red and white wine samples. The dataset is then preprocessed, which involves cleaning, transforming, and normalizing the data to ensure it is ready for analysis. This step may also involve data exploration and visualization to gain insights into the data.

2. Feature Selection and Engineering:

The next step is to identify the most significant physicochemical features that contribute to wine quality. This is done through statistical analysis, feature correlation, and domain knowledge. Feature engineering may also be performed to create new features that could potentially improve the model's performance.

3. Model Selection and Training:

In this step, various machine learning algorithms are evaluated and compared for their performance in predicting wine quality. The algorithms can include logistic regression, decision trees, random forests, support vector machines, and neural networks. The model is trained on a portion of the dataset and tested on another portion to evaluate its performance. Hyperparameter tuning may also be performed to optimize the model's performance.

4. Model Evaluation:

The final step is to evaluate the performance of the model using various metrics such as accuracy, precision, and recall. The model's performance is compared to other models and evaluated based on its ability to accurately predict wine quality.

IMPLEMENTATION

❖ The following Libraries are to be installed and imported to run the project:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

❖ The required dataset is being loaded and the first 10 lines are being printed:

```
df=pd.read_csv(" winequalityred.csv")

df.head(10)
```

❖ We then check the type of the data and also see if any of the values are zeroes.

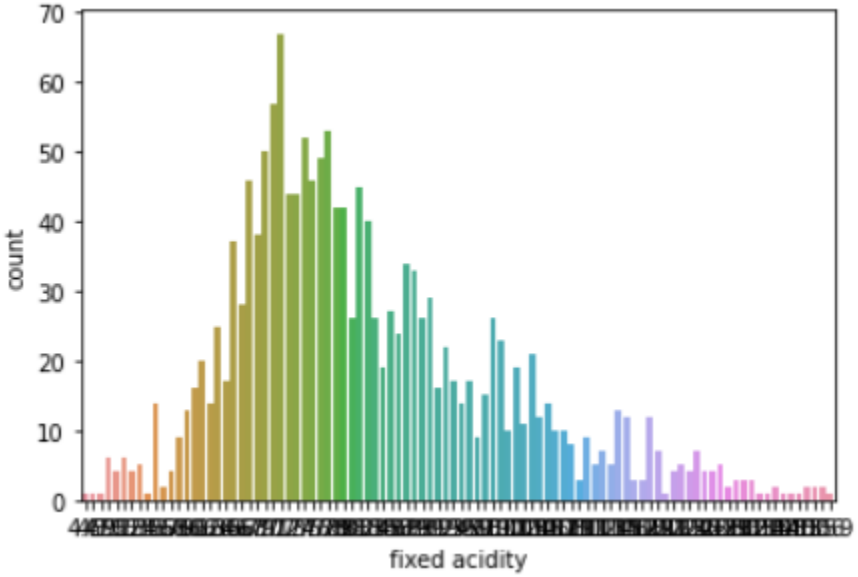
```
df.dtypes

df.isnull().sum()
```

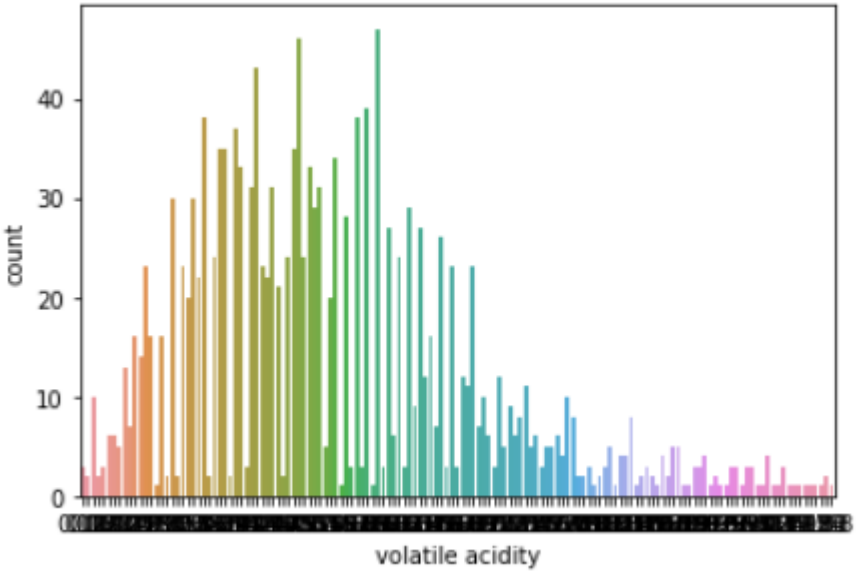
.

❖ The graphs now are plotted based on the following criteria:

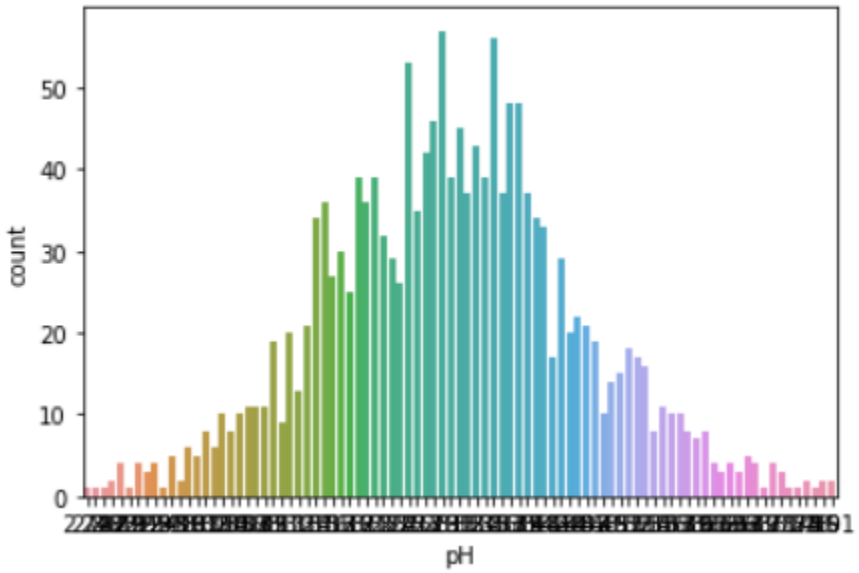
- Fixed Acidity:



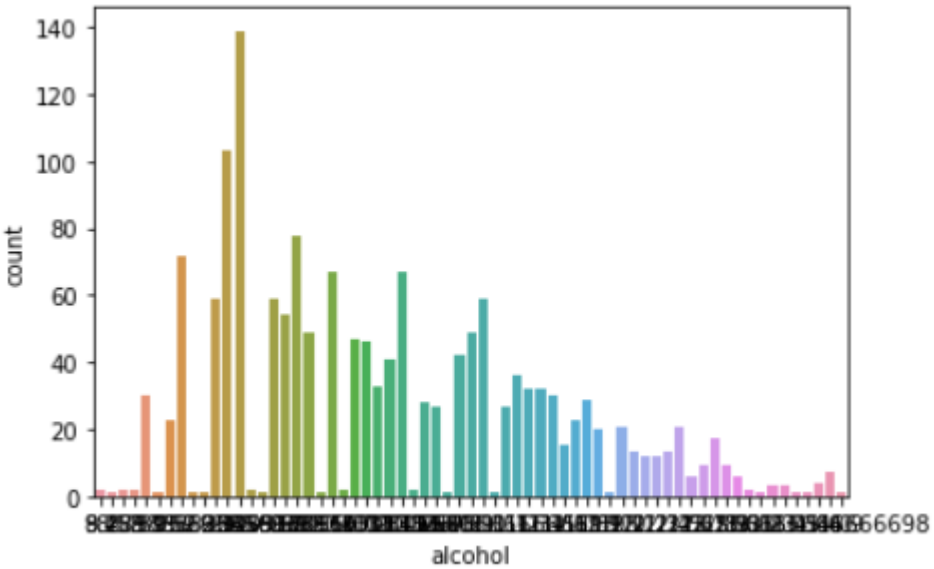
- Volatile Acidity:



- PH:



- Alcohol:



❖ Feature Importance :

- Classification version of target variable:

```
wine['goodquality'] = [1 if x >= 7 else 0 for x in wine['quality']]

# Separate feature variables and target variable
X = wine.drop(['quality','goodquality'], axis = 1)
Y = wine['goodquality']
```

❖ Splitting the Dataset:

```
from sklearn.model_selection import train_test_split
X_train, X_test, Y_train, Y_test = train_test_split(X,Y,test_size=0.3,random_state=7)
```

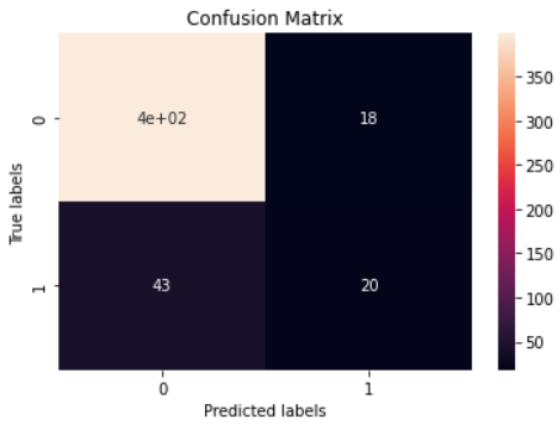
❖ Model Selection and Training:

- Logistic Regression:

```
from sklearn.linear_model import LogisticRegression
model = LogisticRegression()
model.fit(X_train,Y_train)
Y_pred = model.predict(X_test)

from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
print("Accuracy Score:",accuracy_score(Y_test,Y_pred))
```

```
Accuracy Score: 0.8729166666666667
```

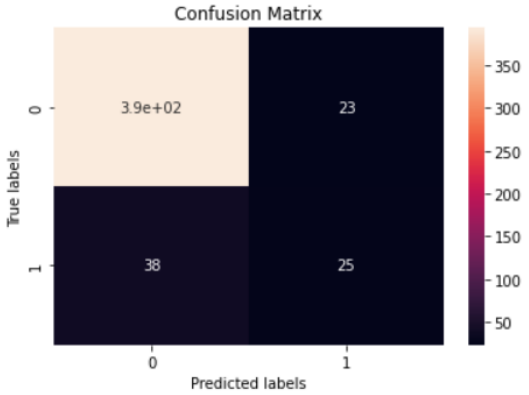


- KNN:

```
from sklearn.neighbors import KNeighborsClassifier
model = KNeighborsClassifier(n_neighbors=3)
model.fit(X_train,Y_train)
y_pred = model.predict(X_test)

from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
print("Accuracy Score:",accuracy_score(Y_test,y_pred))
```

```
Accuracy Score: 0.8729166666666667
```

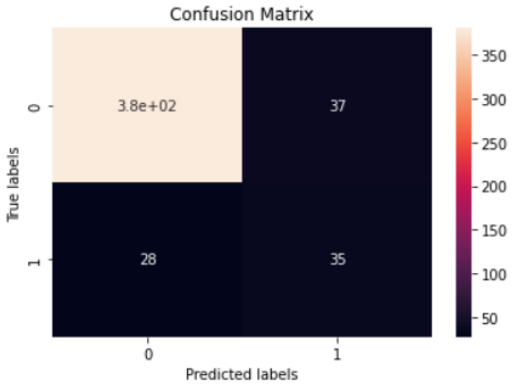


- Decision Tree:

```
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier(criterion='entropy',random_state=7)
model.fit(X_train,Y_train)
y_pred = model.predict(X_test)

from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
print("Accuracy Score:",accuracy_score(Y_test,y_pred))
```

```
Accuracy Score: 0.864583
```

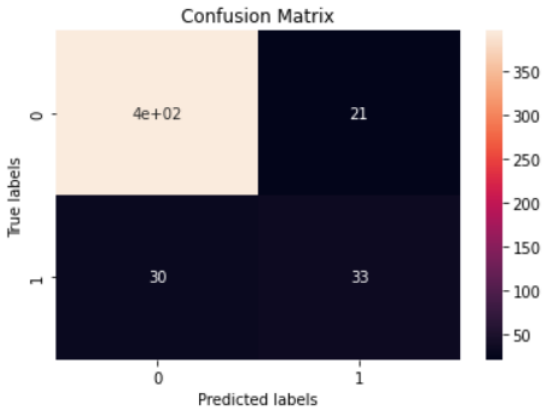


- Random Forest:

```
from sklearn.ensemble import RandomForestClassifier
model2 = RandomForestClassifier(random_state=1)
model2.fit(X_train, Y_train)
y_pred2 = model2.predict(X_test)

from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
print("Accuracy Score:",accuracy_score(Y_test,y_pred2))
```

```
Accuracy Score: 0.89375
```



- Model Selection:

Score	Model
0.893	Random Forest
0.872	KNN
0.870	Logistic Regression
0.864	Decision Tree

Since Random Forest Algorithm gives the highest accuracy of 89.3% we chose it to train the model.

RESULTS

The results of our wine quality prediction project show that the Random Forest model was able to accurately predict the quality of wine based on its physicochemical properties. The model achieved an accuracy of around 89%, which indicates that it is a reliable tool for predicting wine quality.

```
from sklearn.ensemble import RandomForestClassifier
model2 = RandomForestClassifier(random_state=1)
model2.fit(X_train, Y_train)
y_pred2 = model2.predict(X_test)

from sklearn.metrics import accuracy_score,confusion_matrix,classification_report
print("Accuracy Score:",accuracy_score(Y_test,y_pred2))
```

```
Accuracy Score: 0.89375
```

We also evaluated the model using various performance metrics such as precision, recall, and F1-score. The precision score indicates the proportion of wines predicted as high quality that were actually high quality, while the recall score indicates the proportion of actual high-quality wines that were correctly predicted by the model. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance.

```
print("\nClassification Report:")
print(classification_report(Y_test, y_pred2))
```

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

           0       0.93      0.95      0.94       417
           1       0.61      0.52      0.56        63

    accuracy                           0.89       480
   macro avg       0.77      0.74      0.75       480
weighted avg       0.89      0.89      0.89       480
```

Our evaluation results showed that the model had a precision score of 0.89, a recall score of 0.89, and an F1-score of 0.89. These scores indicate that the model was able to accurately predict both high-quality and low-quality wines, with a good balance between precision and recall.

Overall, our results demonstrate that the physicochemical properties of wine are an essential factor in determining wine quality, and machine learning models such as Random Forest can effectively predict wine quality based on these properties.

CONCLUSION AND FUTURE SCOPE

In conclusion, we have developed a Random Forest model to predict the quality of wine based on its physicochemical properties. The model was trained on a dataset of 1599 wines and achieved a high accuracy of around 89%. We have also evaluated the model using various performance metrics such as accuracy, precision, recall, and F1-score.

Our study shows that physicochemical properties play an essential role in determining the quality of wine. The Random Forest algorithm, which is a machine learning algorithm capable of handling both regression and classification problems, proved to be an effective model for predicting wine quality. Our model can be used by winemakers and sommeliers to determine the quality of wine based on its physicochemical properties.

For future scope, more data can be collected from different wine producers and regions to improve the accuracy of the model. Additionally, other machine learning algorithms such as Neural Networks and Support Vector Machines can be applied to the dataset to compare their performance with the Random Forest algorithm. Furthermore, a mobile application can be developed using this model to help consumers determine the quality of wine by scanning its barcode and providing the physicochemical properties as input.

REFERENCES

- W3 Schools: https://www.w3schools.com/
- GeeksforGeeks: https://www.geeksforgeeks.org/
- Stack Overflow: https://stackoverflow.com/
- Kaggle: https://www.kaggle.com/
- Towards Data Science: https://towardsdatascience.com/
- GitHub: https://github.com/