ECE 418: Machine Learning Project Report

Prof. Udaya Sankar V

Team Members:
Yeshwanth sai Gokarakonda (AP19110030008, Mechanical)
Raja Venkata Pramod (AP19110030017, Mechanical)
Prajwal Katakam (AP19110010286, CSE-E)
Devendra Sai (AP19110010279, CSE-E)
Sai Krishnam Raju (AP19110010267, CSE-E)
Surya (AP19110010299, CSE-E)

Introduction:

Quality testing is one attribute which the companies are now using in order to promote their products.

This project aims at predicting wine quality based on the features available in the dataset, the prediction can be done by using different machine learning algorithms but for our project we are using KNN Algorithm, SVM Algorithm, Logistic Regression, Decision Tree Algorithm and Random Forest Algorithm.

Based on the outcomes of these algorithms, we are comparing the accuracy and concluding which algorithm has the best overall accuracy.

Problem Statement:

This project predicts whether the given wine is of good quality or bad using different algorithms which are stated above.

Explanation:

Firstly we need to import all the required libraries, then we need to import and read the data from the CSV file.

We need to check whether the data is imported successfully or not by printing the head. We also need to check if there are any null values or missing values and if there are, we need to replace them by the mean of that particular column values.

Since this is a classification project, we need to add a new column and tell whether the quality of wine is good or bad by comparing it with the quality index.

Then we go ahead and test our feature importance. It basically gives the score of importance of each feature that we have. In this project we have 12 features and hence it prints 12 scores.

Next, we need to split the data into training set and testing set. Training set is used to train our model and testing is used to test the predictions of our model, based on what it has learned, then we can see the accuracy of the model. Here we used a test size of 0.3 which means that 70% of the dataset is used for training and the remaining 30% for testing purposes.

We have used a few models to test and see which model has the best accuracy in predicting the correct values. First one is Logistic regression, we have imported the Logistic regression module from scikit, then we need to fit the training data into the model and later predict the good quality for the testing set.

Second we have done using KNN, Here the root mean squared value was 3 that is the value of K is 3. Based on this we have fitted the training data into the model and predicted whether the wine was of good quality or bad quality.

In the similar way we fitted the data into different models using SVM, Decision trees, Random Forests and predicted whether it was of good quality or bad quality and then calculated their respective accuracies.

After executing all of these, we got accuracies as shown in our result section. Based on these outputs we can say that the Random Forests model predicts data with the most accuracy, with an accuracy rate of 0.89375.

Our Approach & Code:

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

from warnings import filterwarnings
filterwarnings(action='ignore')

wine = pd.read_csv("/content/winequality-red.csv")
print("Successfully Imported Data!")
wine.head()

print(wine.shape) #returns the dimensions of an array
wine.describe(include='all') #describes the dataset wine
print(wine.isna().sum()) #detecting missing values in the dataset

wine.groupby('quality').mean()

# Create 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']

# See proportion of good vs bad wines
wine['goodquality'].value_counts()

X # attributes which are responsible for classification

print(Y)

from sklearn.linear_model import LogisticRegression
model = LogisticRegression()

from sklearn.ensemble import ExtraTreesClassifier
classifiern = ExtraTreesClassifier()
classifiern.fit(X,Y)
score = classifiern.feature_importances_
print(score)

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)

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
print("Accuracy Score:",accuracy_score(Y_test,Y_pred))

confusion_mat = confusion_matrix(Y_test,Y_pred)
print(confusion_mat)

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
print("Accuracy Score:",accuracy_score(Y_test,y_pred))

SVM:
from sklearn.svm import SVC
model = SVC()
model.fit(X_train,Y_train)
pred_y = model.predict(X_test)

from sklearn.metrics import accuracy_score
print("Accuracy Score:",accuracy_score(Y_test,pred_y))

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
print("Accuracy Score:",accuracy_score(Y_test,y_pred))

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
print("Accuracy Score:",accuracy_score(Y_test,y_pred2))
```

Code Result:

Algorithms:	KNN	SVM	Decision Tree	Logistic Regression	Random Forest

Accuracy Score:	0.8729166666666667	0.86875	0.8645833333333334	0.8729166666666667	0.89375

Literature Survey:

Selection of important features and predicting wine quality using machine learning techniques - Yogesh Gupta

This paper explores the usage of machine learning techniques such as linear regression, neural network and support vector machine for product quality in two ways. Firstly, determine the dependency of the target variable on independent variables and secondly, predicting the value of the target variable. In this paper, linear regression is used to determine the dependency of the target variable on independent variables. On the basis of computed dependency, important variables are selected that make a significant impact on dependent variables. Further, neural networks and support vector machines are used to predict the values of dependent variables. All the experiments are performed on Red Wine and White Wine datasets. This paper proves that the better prediction can be made if selected features (variables) are being considered rather than considering all the features

A Study and Analysis of Machine Learning Techniques in Predicting Wine Quality - Mohit Gupta, Vanmathi C

This chapter demonstrates the usage of various machine learning techniques in predicting the quality of wine and results are validated through various quantitative metrics. Moreover the contribution of various independent variables facilitating the final outcome is precisely portrayed

Wine Quality Prediction using Machine Learning Algorithms - Devika Pawar, Aakanksha Mahajan, Sachin Bhoithe

An automatic predictive system can be integrated into a decision support system, helping the speed and quality of the performance. Furthermore, a feature selection process can help to analyze the impact of the analytical tests. If it is concluded that several input variables are highly relevant to predict the wine quality, since in the production process some variables can be controlled, this information can be used to improve the wine quality. Classification models used here are:
1) Random Forest
2) Stochastic Gradient Descent
3) SVC
4) Logistic Regression

Conclusion:

This project helped us get practical hands-on knowledge on everything which has been said in the class.
It has helped us understand and implement the algorithms more clearly.
We would like to thank Prof. Udaya Sankar Sir for giving us this opportunity.