WINE QUALITY ANALYSIS

MACHINE INTELLIGENCE (UE20CS302)

MINI PROJECT

ANUBUTHI K (PES1UG20CS065)

ATTILI SUBHA VIDISHA (PES1UG20CS091)

DEVANG SARAOGI (PES1UF20CS122)

ABSTRACT

The aim of this project is to predict human wine taste preferences that are based on easily available analytical tests at the certification step. We expect to get an accuracy score of ~ 85%. Predicted value could be used for designing new types of wine, defining pricing policy or supporting decision making in advisory systems. Our goal was to build an artificial neural network from scratch using only numpy and pandas and compare its accuracy with various other python library functions (KNN, SVC,LOGISTIC REGRESSION , DESCISION TREE , NAIVE BAYES AND SO ON)and also we used tensorflow to create and run a deep neural network for classifined the quality of wine.

Assumptions:

The **null hypothesis (H0)** is that none of the variance in the quality ranking is explained by physicochemical properties. The **alternate hypothesis (H1)** is that physicochemical properties contribution to the variance in the quality ranking and make a wine 'best' or vice verse 'bad'.

The data

We will use a real data set related to red Vinho Verde wine samples, from the north of Portugal. This dataset is available from the UCI machine learning repository, https://archive.ics.uci.edu/ml/datasets/wine+quality and also on kaggle https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009. This dataset can be viewed as classification or regression tasks.

input variables (based on physicochemical tests):

1. fixed acidity;
2. volatile acidity;
3. citric acid;
4. residual sugar;
5. chlorides;
6. free sulfur dioxide;

7. total sulfur dioxide;
8. density;
9. pH;
10. sulphates;
11. alcohol.

Output variable (based on sensory data):

1. quality (score between 0 and 10).

Execution details

The project requires libraries like tensorflow , skleran , pandas and numpy to be imported before execution. We begain the execution by running some priliminary analysis on the data by performing exploratory data analysis by checking for null values , correlation among variables , density of variables . We measure feature importance using random forest regressor , from which we find that density, residual sugars and free sulphar dioxide are least important features in determining quality of wine. We define **best quality** as 1 when the quality of wine is greater then 6 else as 0 . This gives us a binary class for classfication . We then split the data into train ands test datasets in ratio 70:30.

The next step is to run the various classification models :

1. LOGISTIC REGRESSION
2. KNN
3. SVC
4. DECISION TREE
5. GAUSSIAN NB
6. RANDOM FOREST
7. XG BOOST
8. DEEP NEURAL NETWORK (KERES)

The last part of the project is the implemenatation of ANN from scratch using only numpy and pandas The accuracies from all the

Conclusion

The various models give us accuracies ~85% . The best accuracy is obtained from Random Forest model (90.37%). The ANN model built from scartch gives and accuracy of (86%).