Final Project – Rapid Miner Supplement

Random Forest and Logistic Regression Algorithm on UCI Wine Classification Dataset

Course: IS655 – Data Analytics for Information Systems

Professor: Dr. Lin Lin

Brian Rooney

Parvathy Neelakanta Sarma

Shivakalyan Soundarathiagarajan

Srishti Karakoti

**Table of Contents**

[***Process Overview***](#_heading=h.gjdgxs) ***3***

[**Description**](#_heading=h.30j0zll) **3**

[**Screenshot of process page**](#_heading=h.1fob9te) **4**

[**Discussion on Performance Measure**](#_heading=h.3znysh7) **4**

[***Random Forest Model***](#_heading=h.2et92p0) ***5***

[**Screenshot**](#_heading=h.tyjcwt) **5**

[**Parameters**](#_heading=h.3dy6vkm) **5**

[**Results**](#_heading=h.1t3h5sf) **5**

[Sample Decision Tree Output](#_heading=h.4d34og8) 5

[Confusion Matrix](#_heading=h.2s8eyo1) 6

[Performance Metrics](#_heading=h.17dp8vu) 6

[***Logistic Regression***](#_heading=h.3rdcrjn) ***7***

[**Screenshot**](#_heading=h.26in1rg) **7**

[**Parameters**](#_heading=h.lnxbz9) **7**

[**Results**](#_heading=h.35nkun2) **7**

[Coefficient Table](#_heading=h.1ksv4uv) 7

[Confusion Matrix](#_heading=h.44sinio) 7

[Performance Metrics](#_heading=h.2jxsxqh) 8

[***Comparison***](#_heading=h.z337ya) ***9***

# Process Overview

This project uses two Data Mining algorithms to classify wine into two discreet categories: either “good” or “bad”

Dataset Information

Name of dataset: Red Wine Quality Data Set

Source: UCI Machine Learning Repository

Input variables:

* fixed acidity
* volatile acidity
* citric acid
* residual sugar
* chlorides
* free sulfur dioxide
* total sulfur dioxide
* density
* pH
* sulphates
* alcohol

Output variable: quality (binary score of 0 (bad) or 1 (good))

Data Set Characteristics: Multivariate

Number of Observations: 1599

Number of Attributes/Variables: 12

Missing Values: N/A

Kaggle Link for the Notebook and Data: <https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

## Screenshot of process page

Diagram

Description automatically generated

## Discussion on Performance Measure

In this analysis we chose the best performing model based off of precision.

# Random Forest Model

## Screenshot

Diagram

Description automatically generated

## Parameters

Number of Trees: 300

Information Gain criterion: Gini Index

Maximal Depth: -1 (maximum depth)

## Results

### Sample Decision Tree Output

density > -0.719

| alcohol > 0.119

| | volatile acidity > 1.353

| | | alcohol > 0.401: Bad {Bad=19, Good=0}

| | | alcohol ≤ 0.401

| | | | total sulfur dioxide > -1.033

| | | | | total sulfur dioxide > -0.850

| | | | | | free sulfur dioxide > -0.848: Bad {Bad=1, Good=0}

| | | | | | free sulfur dioxide ≤ -0.848: Good {Bad=0, Good=1}

| | | | | total sulfur dioxide ≤ -0.850: Good {Bad=0, Good=5}

| | | | total sulfur dioxide ≤ -1.033: Bad {Bad=4, Good=0}

| | volatile acidity ≤ 1.353

| | | density > 3.115: Bad {Bad=5, Good=0}

| | | density ≤ 3.115

| | | | fixed acidity > -1.246

| | | | | pH > -2.080

| | | | | | sulphates > -0.137

| | | | | | | chlorides > 0.224

| | | | | | | | alcohol > 0.495

| | | | | | | | | free sulfur dioxide > -0.992

| | | | | | | | | | residual sugar > -0.276: Good {Bad=0, Good=30}

| | | | | | | | | | residual sugar ≤ -0.276

| | | | | | | | | | | volatile acidity > -0.546: Bad {Bad=1, Good=0}

| | | | | | | | | | | volatile acidity ≤ -0.546: Good {Bad=0, Good=5}

| | | | | | | | | free sulfur dioxide ≤ -0.992

| | | | | | | | | | residual sugar > 1.320: Bad {Bad=2, Good=0}

**NOTE**: This is a representative example of a single decision tree generated by the random forest model. The full decision tree took up roughly 15 pages and the graph was too large to fit into a screenshot.

### Confusion Matrix

Table

Description automatically generated

### Performance Metrics

accuracy: 81.86% +/- 4.44% | (micro average: 81.86%)

precision: 83.83% +/- 5.59% | (micro average: 83.51%)

recall: 82.34% +/- 4.45% | (micro average: 82.34%)

# Logistic Regression

## Screenshot

Diagram

Description automatically generated

## Parameters

Use Regularization: False

Standarize: True

Add intercept: True

Compute p-values: True

Remove colinear columns: True

## Results

### Coefficient Table

Table

Description automatically generated

### Confusion Matrix

Table

Description automatically generated

### Performance Metrics

accuracy: 74.36% +/- 2.56% | (micro average: 74.36%)

precision: 76.49% +/- 2.80% | (micro average: 76.39%)

recall: 75.31% +/- 4.25%. | (micro average: 75.32%)

# Comparison

We observed the key factors that determine and affect the quality of the red wine. Wine quality is ultimately a subjective measure. The ordered factor 'quality' was not very helpful and to overcome this, so we created another variable called 'rating'.

To make predictions of wine quality and any other if required, we trained two models i.e., sklearn's Logistic Regression model and Random Forest Classifier. The hyper parameters of both the models have been optimized and we are seeing performances very close in both the models.We decided to stick with 'Logistic Regression' if we had to make any more predictions,given the fact it had higher precision compared to the other model and relatively less hyper parameters to optimize (we are going with this decision based upon our observation - from results Python Lab).

The usage of this analysis will help to understand whether by modifying the variables, it is possible to increase the quality of the wine on the market. If you can control your variables, then you can predict the quality of your wine and obtain more profits.