The Classification of Red Wine Quality According To Physiochemical Data

Elias Turk
Big Data Analytics and Management
Graduate Education Institute, BAU
Istanbul, Turkey
elias.turk@bahcesehir.edu.tr

***Abstract*:** The purpose of this study is to predict wine quality based on its physiochemical constituents. In this project , red wine quality dataset was taken from Kaggle but it was mainly uploaded to UCI Machine Learning Repository. This data set contains 1599 different red wine types with 11 features of physiochemical data such as alcohol, chlorides, density, total sulfur dioxide, free sulfur dioxide, residual sugar, and pH And 1 variable such as "quality" which is a score between 0 and 10. This task was approached using both regression and classification algorithms , but surely as it's more of classification problem, algorithms such as Logistic Regression and Random Forest with some feature selection showed more accuracy to predict Red Wine quality with 72 % and 79% respectively. On the other hand , Linear Regression was used but showed only an r-squared of 38%.Random Forest Classifier was the most successful classification Algorithm.

Keywords:*Classification , Regression , Random Forest , Linear Regression.*

1. Introduction

The red wine industry has always been on an exponential growth. Nowadays , companies are using product quality certifications to promote their products[1] . However this is not only a time-consuming process but it also requires assessment given by human experts which makes this process really expensive . The price of red wine always refers to how good it is , but it depends on an abstract concept of wine tasters , and their opinion is subject to high variability. Wine companies and the industry are researching new technologies for wine making and selling process in order to back up this growth [2].

2. Materials and Methods

2.1. Red Wine Data

The dataset of red wine quality is publicly available for research purposes on this website : https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009 but as we mentioned it was actually taken from https://archive.ics.uci.edu/ml/datasets/Wine+Quality .
The dataset contains 1599 instances with 11 features. It includes objective tests, independent variables (e.g. alcohol values) and the output is a dependent variable based on the average(median of at least 3 evaluations made by wine experts).Each wine taster graded the wine quality between 0 and 10 ranging from very bad to very good. The dataset is related to red variants of the Portuguese "Vinho Verde" wine.
The goal behind the dataset is to predict the rating that a wine taster will give to a wine sample using physiochemical features such as acidity, alcohol, sulphates composition.
Due to copyright rules , privacy and logistic issues , no data about grape type, wine brands(types) nor prices were listed in the dataset.

Table 1 presents the 11 independent different physiochemical features and data statistics of red wine quality.

Table 1. The physiochemical data statistics of red wine.

Attribute (units)	Min	Max	Mean	StDv
Fixed acidity (g(tartaric acid)/dm^3)	4.600	15.90	8.320	1.741
Volatile acidity (g(acetic acid)/dm^3)	0.120	1.580	0.528	0.179
Citric acid (g/dm^3)	0.000	1.000	0.271	0.195
Residual sugar (g/dm^3)	0.900	15.50	2.539	1.410
Chlorides (g(sodium chloride)/dm^3)	0.012	0.611	0.087	0.047
Free sulfur dioxide (mg/dm^3)	1.000	72.00	15.87	10.46
Total sulfur dioxide (mg/dm^3)	6.000	289.0	46.47	32.89
Density (g/cm^3)	0.990	1.004	0.997	0.002
pH	2.740	4.010	3.311	0.154
Sulphates(g(potassium sulphate)/dm^3)	0.330	2.000	0.658	0.170
Alcohol (%vol)	8.400	14.90	10.42	1.066

2.2. Algorithms :

2.2.1 Regression:

In order to predict quality score from 0 to 10 as mentioned above,we had the option to choose regression algorithms as this number is seen as a continuous variable but at the same type seen as categorical.

Linear Regression attempts to model the relationship between two variables by making a linear equation fit on the observed data. It has the equation of :

$$Y = \alpha + \beta X$$

The dataset is really small and we didn't do any prediction using linear regression , only printed the OLS (Least Squares) model .
We approached this problem by doing a manual feature selection by removing high p-values features.
Different interaction terms among features were also tried to see how the model got better with different interactions.

2.2.2 Classification:

Logistic regression is a model in statistics that uses the logistic function to model binary values of a dependent variable . Introducing a new variable that explains quality differently had to be implemented, I chose to give a rating of 0 and 1 for bad and good alcohol quality respectively , and by doing so , variable " quality " was dropped.
It then obtains odds ratio in presence of multiple independent variables to predict whether the rating is 0 or 1 (quality).
As the problem was clearly a classification task, logistic regression would logically perform better than linear regression however the implementation of other classification techniques was crucial to see which model algorithms performs better in our case [3].

Random Forest was another classification algorithm used in the analysis. It builds multiple decision trees and aggregate them to get a more accurate prediction, it nearly has the same hyperparameters as a decision tree however it adds more randomness to the model while growing the trees.
In order to do feature selection ,we had to specifically implement Random Forest classifier using a sequential forward selection.
This is one of the wrapper methods , which evaluates feature subsets in order to detect model performance between features.
Step forward selection is one wrapper method ,another is step backward selection however only step forward selection was used in this task [4].

3. Experimental Results

The 2 classification techniques and one regression technique was used in my study to classify the wine samples quality.
For the regression task ,a model was built using the Ordinary Least Squares (OLS) summary table so the dataset was not split because no predictions were made on the algorithm . Evaluating how the model performed was enough . Building the model with all features on the linear regression gave an r-squared of 36.1% and an adjusted R2 of 35.6% .
Because I have no experience in the field I just decided to remove the variables that displayed relatively high p-values.
Second model showed a relatively better ratio of adjusted R2 to R-squared but the model didn't really get any better.

The third time , I ran the model using interaction terms and the model got a higher r-squared but there was still some difference with the adjusted r-squared and that's basically caused by the small size of the dataset.
The regression task gave me a small insight that some features are really useless in predicting the quality of alcohol so I used to same technique for the classification tasks.

For the Logistic Regression , I created a Generalized Linear Regression model from the binomial family (Logistic Regression) using the added rating variable that we talked about earlier, and omitted high p-values variables.
With the remaining features of 5,we got an accuracy of 71.2% which showed an immense progress comparing it with linear regression .
I decided to proceed wit the 5 features training and testing datasets to fit the Random Forest because the rest of the variables were useless in both logistic and linear regression, keeping in mind that this could save not a lot of memory but also building and testing time .

Random Forest classifier with step forward selection showed us a result of 78% with 5 features which is also a lot better than Logistic Regression.
But that wasn't enough for me as I wanted to see how would the Random Forest perform with all features.

So I went back and created a copy of the original dataframe with the rating score and dropped quality and re-splitted it into training and testing and rebuilt the Random Forest Classifier algorithms with all features.
But this only showed an accuracy of 80% which is not worth the hassle because the trade off between the the accuracy given and the time spent to build and test the model isn't really good.
While the Sequential Forward Selection was iterating and displaying how the model was performing each time with

each new variable introduced I noticed that with 7 features it performed best.
I noticed that with 7 features the model gave 79.2% accuracy score which is kind of fair to me because it's the same with using ALL features and its accuracy is just slightly less .Our 5 feature option which gave 78.7% is my best option because its just 1% less accurate and saves a lot more time.

4. Insights on features

In the Exploratory Data Analysis stage I noticed some highly correlated features with the dependent variable, alcohol had the highest positive correlation with quality. Whenever alcohol levels increased the quality also increased.
Both sulphates and citric acid comes behind alcohol.

On the other hand, I noticed volatile acidity to be the highest correlated variable but negatively with quality. And after visualization I realized that whenever the volatile acidity decreased , red wine quality increased.

This illustration will prove what I've been looking for .

Fig 1. Different scores of SFS Random Forest Classifier Model during cross validations.

	feature_idx	cv_scores	avg_score	feature_names	ci_bound	std_dev	std_err
1	(4,)	[0.6785714285714286, 0.6964285714285714, 0.692...	0.702448	(alcohol,)	0.0380819	0.0237569	0.0137161
2	(1, 4)	[0.7142857142857143, 0.6964285714285714, 0.710...	0.721217	(citric_acid, alcohol)	0.0405183	0.0252768	0.0145936
3	(0, 1, 4)	[0.7535714285714286, 0.7535714285714286, 0.75,...	0.757834	(volatile_acidity, citric_acid, alcohol)	0.0153197	0.00955701	0.00551774
4	(0, 1, 3, 4)	[0.775, 0.7392857142857143, 0.7821428571428571...	0.774824	(volatile_acidity, citric_acid, sulphates, alc...	0.0367551	0.0229292	0.0132382
5	(0, 1, 2, 3, 4)	[0.7678571428571429, 0.7607142857142857, 0.771...	0.781989	(volatile_acidity, citric_acid, total_sulfur_d...	0.0429894	0.0268184	0.0154836

Alcohol alone gave a 70% score and got slightly higher when citric acid was added but jumped higher than 3% with volatile acidity.
Citric acid is the third contributor after VA and Alcohol in determining the quality.
Solving this task using classification algorithms definitely gave us better results than regression ones, and RF gave us better results than Logistic and Linear regression.

Let's check some of the results given in Table 2 conducted on the test dataset.

Table 2. Classification algorithms report

Algorithm	Accuracy	Precision	F1score	Recall
Logistic Regression	71%	76%	73%	70%
Random Forest Classifier	79%	81%	81%	81%

Accuracy is the ratio of the correctly predicted classifications to the total cases of the test dataset.

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{False Positives} + \text{False Negatives} + \text{True Negatives}}$$

It is well observed that Random Forest performed better in predicting better.

Precision is the total number of true positives over predicted positives.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

We can see that Random Forest also performed better with 81% precision while Logistic Regression made 76% which is also not bad.

In Random Forest, this is our confusion matrix :

```
array([[162,  51],
       [ 51, 216]], dtype=int64)
```

216 out of 266 were predicted correct so thats 81% precision .

Recall is the total number of true positives over actual positives.

Also gives us 81% .

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Precision and Recall are equals in RF this means false positives are equal to false negatives because recall and precision have the same denominator.

This may be a good thing if the dataset had an equal amount of good and bad quality wines.

In our case it certainly is , 744 wines are classified as bad and 855 are of good quality.

```
0    744
1    855
Name: rating, dtype: int64
```

5. Conclusions

For each classification and regression model used in our task , I analyzed the results and observed how they varied with different variables .
The study includes the analysis of classifiers on red wine dataset. The results are described in percentage of correctly classified classifications precision, recall, accuracy and F-score after applying both Logistic Regression and Random Forest.
Results from the analysis made us conclude that Random Forest Classifier algorithm performs better as compared to Logistic Regression.
I didn't standardize the data because its useless on tree based algorithms such as RF.
First of all I tried to see how the data fits on a Linear Regression model and the best modification on the model gave me results of 38% R2,and then I fit the same data on a Logistic Regression model and omitted the high p-vales in the GLM.After doing so I split the data into training and test(validation) .After validating the test data I got an accuracy of 71% .
Moreover, when I used Step Forward Feature Selection on a Random Forest, I got an accuracy of 79% then I tried to see how would the algorithm work with all of the 11 features only to realize that it gave us just 1% more accuracy which is really not worth the time to predict future data.
Random Forest was undoubtedly the best algorithm used among the other 2.However, despite the fact that the model with ALL features did slightly better than the one with 5 and 7 features, I still chose the model with 5 features because its accuracy is almost the same and saves a lot more time.

Acknowledgements

I am really grateful because I managed to complete my BDA-5001 project within the time given by my lecturer Serkan Ayvaz.
Honestly, it would've been impossible for me to complete it without Mr.Ayvaz's clear explanation about several data analytics topics that gave me the courage and enthusiasm to explore several other topics by myself. Emails and questions and clarifications were always answered spontaneously . Last but not least ,I would like to thank all my classmates and colleagues for making this course a pleasant experience this semester.

References

[1] S. Ebeler, "Linking Flavour Chemistry to Sensory Analysis of Wine," in *Flavor Chemistry*, Thirty Years of Progress, Kluwer Academic Publishers, 1999, pp. 409-422.

[2] P. Cortez, A. Cerderia, F. Almeida, T. Matos, and J. Reis, "Modelling wine preferences by data mining from physicochemical properties," *In Decision Support Systems, Elsevier*, 47 (4): 547-553. ISSN: 0167-9236.

[3] Ben-Assuli, O., Heart, T., Shlomo, N., & Klempfner, R. (2019). Bringing big data analytics closer to practice: A methodological explanation and demonstration of classification algorithms. *Health policy and technology*, *8*, 7-13.K. Elissa, "Title of paper if known," unpublished.

[4] M.Mayo . (2018) Step Forward Selection : A Practical Example in Python.Overviews.https://www.kdnuggets.com/2018/06/step-forward-feature-selection-python.html .

Fig 2 & Fig 3 . Displaying the Random Forest Classifier using Step Forward Feature Selection and their time stamps.

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concurrent workers.
[Parallel(n_jobs=-1)]: Done   3 out of   5 | elapsed:    2.5s remaining:    1.7s
[Parallel(n_jobs=-1)]: Done   5 out of   5 | elapsed:    2.6s finished

[2021-01-17 12:22:07] Features: 1/5 -- score: 0.7024481566820278[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concur
rent workers.
[Parallel(n_jobs=-1)]: Done   2 out of   4 | elapsed:    2.0s remaining:    2.0s
[Parallel(n_jobs=-1)]: Done   4 out of   4 | elapsed:    2.0s finished

[2021-01-17 12:22:09] Features: 2/5 -- score: 0.7212173579109062[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concur
rent workers.
[Parallel(n_jobs=-1)]: Done   3 out of   3 | elapsed:    1.6s finished

[2021-01-17 12:22:11] Features: 3/5 -- score: 0.7578341013824885[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concur
rent workers.
[Parallel(n_jobs=-1)]: Done   2 out of   2 | elapsed:    0.5s finished

[2021-01-17 12:22:12] Features: 4/5 -- score: 0.774823988735279[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concurr
ent workers.
[Parallel(n_jobs=-1)]: Done   1 out of   1 | elapsed:    0.5s finished

[2021-01-17 12:22:12] Features: 5/5 -- score: 0.781989247311828
```

```
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concurrent workers.
[Parallel(n_jobs=-1)]: Done   6 out of  11 | elapsed:    1.0s remaining:    0.8s
[Parallel(n_jobs=-1)]: Done  11 out of  11 | elapsed:    1.2s finished

[2021-01-17 12:22:17] Features: 1/11 -- score: 0.7024481566820278[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   5 out of  10 | elapsed:    1.1s remaining:    1.1s
[Parallel(n_jobs=-1)]: Done  10 out of  10 | elapsed:    1.2s finished

[2021-01-17 12:22:18] Features: 2/11 -- score: 0.7202924987199182[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   6 out of   9 | elapsed:    1.0s remaining:    0.5s
[Parallel(n_jobs=-1)]: Done   9 out of   9 | elapsed:    1.0s finished

[2021-01-17 12:22:19] Features: 3/11 -- score: 0.7703405017921147[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   5 out of   8 | elapsed:    1.1s remaining:    0.6s
[Parallel(n_jobs=-1)]: Done   8 out of   8 | elapsed:    1.2s finished

[2021-01-17 12:22:20] Features: 4/11 -- score: 0.7873591909882232[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   4 out of   7 | elapsed:    0.9s remaining:    0.7s
[Parallel(n_jobs=-1)]: Done   7 out of   7 | elapsed:    1.0s finished

[2021-01-17 12:22:22] Features: 5/11 -- score: 0.7936059907834101[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   3 out of   6 | elapsed:    0.8s remaining:    0.8s
[Parallel(n_jobs=-1)]: Done   6 out of   6 | elapsed:    0.8s finished

[2021-01-17 12:22:22] Features: 6/11 -- score: 0.7944988479262673[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   3 out of   5 | elapsed:    0.8s remaining:    0.5s
[Parallel(n_jobs=-1)]: Done   5 out of   5 | elapsed:    0.8s finished

[2021-01-17 12:22:23] Features: 7/11 -- score: 0.7998559907834102[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   2 out of   4 | elapsed:    0.7s remaining:    0.7s
[Parallel(n_jobs=-1)]: Done   4 out of   4 | elapsed:    0.7s finished

[2021-01-17 12:22:24] Features: 8/11 -- score: 0.7963069636456734[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   3 out of   3 | elapsed:    0.6s finished

[2021-01-17 12:22:25] Features: 9/11 -- score: 0.7909338197644649[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   2 out of   2 | elapsed:    0.6s finished

[2021-01-17 12:22:26] Features: 10/11 -- score: 0.796300563236047[Parallel(n_jobs=-1)]: Using backend LokyBackend with 12 concu
rrent workers.
[Parallel(n_jobs=-1)]: Done   1 out of   1 | elapsed:    0.5s finished

[2021-01-17 12:22:26] Features: 11/11 -- score: 0.7945116487455196
```