ESTIMATION OF WINE QUALITY USING CHEMICAL ANALYSIS DATA

S. JACINDHA
Department of Information Technology
Sri Siva Subramaniya Nadar College of Engineering
Chennai 603110,Tamil Nadu,India
Email:-sjacindha428@gmail.com

T.HYMADEVI
Department of Computer Science and Engineering
Dr.MGR Educational and Research Institute
Chennai 6000095,Tamil Nadu,India
Email:-hymadevi1234@gmail.com

Y.SAIKUMAR
Department of Computer Science and Engineering
Dr.MGR Educational and Research Institute
Chennai 6000095,Tamil Nadu,India
Email:-saikumaryatirajula2@gmail.com

A.NANDHINI
Department of Computer Science and Engineering
Dr.MGR Educational and Research Institute
Chennai 6000095,Tamil Nadu,India
Email:-nandhu3999@gmail.com

Abstract— The price of wine has been determined from sensory points(quality) and the vintage. Thus finding the quality of the wine has become a tedious job. Each type of wine has its own composition. Hence our idea is to use the type, alcohol, density, volatile acidity, free sulfur dioxide, total sulfur dioxide, sulphates and other chemical factors to determine the quality and hence, the price of the wine. We use random forest, logical regression and linear regression.Thereby helping the customer choose the type of wine by his financial status and making a profit for the wine keepers without sacrificing the quality.

Keywords—*Wine,Logistic regression, Linear regression, Random forest, Price, Quality, Chemical Factors*

INTRODUCTION

Quality management is important when it comes to business.Today, all types of industries are improving by adopting new technologies and applying them in all areas. These technologies help to enhance the production and making the whole production process smooth. But, still there are different areas, which demands human expertise such as product quality assurance. Nowadays, it becomes an expensive process as the demand of a product grows over time. Therefore, this project aims to use a machine learning algorithm for product quality assurance. These techniques performs quality assurance process with the help of available characteristics of product and automate the process by minimizing human interference. The work also identifies the important features to predict the values of dependent variables.

The wine quality determines the price of the wine. The quality is measured from every aspect such as the vintage, variety, type, place , the grape used, and or the various chemical analysis. Each type of wine has its own composition and taste.

There are about five types of wine: Red wine, White wine, Rose wine, Sparkling wine and Fortified wine.Since there are various varieties and blends under each type, the quality estimation has become difficult. In this report, we want to derive a new method of estimating quality using chemical analysis.

OBJECTIVE

- To visualize and understand the given database by using graphs, diagrams and images.
- To build a model which can estimate the quality of wine which varies by its composition and chemical factors .
- To predict the price of wine based on the sensory points
- To create an application that can predict the price of the wine.

PROBLEM STATEMENT

Wine is categorised based on quality and price is set accordingly allowing the customer to purchase wine of his choice. Quality in the sense, the ingredients used while manufacturing of wine. However, each type of wine has its own uniqueness. This classification is also used to set the prices accordingly so that the firm enjoys the profit. By prediction of the quality of wine every individual can purchase it as per their financial status. Since there is no, such a system to predict the quality of wine.

RELATED WORKS

Ian Xiao has done chemical analysis on Red wine and White wine and differentiated the high quality and low quality based on attributes like fixed acidity, volatile acidity,residual sugar,chlorides, alcohol, Sulphur dioxide, density, pH and sulphates and sensory data.

Helene Hopfer(.et.al) has done study on, the sensory, volatile and elemental profiles of 27 Californian Cabernet Sauvignon wines were correlated to the quality proxies (i) points awarded during a wine competition, (ii) wine expert liking scores, (iii) retail bottle price, (iv) vintage, and (v) wine region**.**

Limitations:

No single compound or sensory descriptor is able to fully describe all aspects of wine quality.
Different quality determining techniques for different types of wines.

TOOLS AND MODULES USED

The following tools were used to build this mode:

1. Python (3.5.7)
2. Sklearn to built the models: Random forest, Linear regression, Logistic regression.
3. Matplotlib and seaborn to plot graphs for visualisations
4. tkinter for UI
5. Google Colabs

DATA COLLECTION

We use two datasets for our project:

1. Wine chemical analysis:Wine Quality by Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis (CVRVV) @ 2009 (Fig.1)

```
(6497, 13)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6497 entries, 0 to 6496
Data columns (total 13 columns):
type                  6497 non-null object
fixed acidity         6487 non-null float64
volatile acidity      6489 non-null float64
citric acid           6494 non-null float64
residual sugar        6495 non-null float64
chlorides             6495 non-null float64
free sulfur dioxide   6497 non-null float64
total sulfur dioxide  6497 non-null float64
density               6497 non-null float64
pH                    6488 non-null float64
sulphates             6493 non-null float64
alcohol               6497 non-null float64
quality               6497 non-null int64
dtypes: float64(11), int64(1), object(1)
memory usage: 659.9+ KB
```

Fig1. Information on wine chemical dataset

- fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- volatile acidity:the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- citric acid: found in small quantities, citric acid can add 'freshness' and flavor to wines
- residual sugar: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
- chlorides: the amount of salt in the wine
- free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine
- total sulfur dioxide: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
- density: the density of water is close to that of water depending on the percent alcohol and sugar content
- pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
- sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
- alcohol: the percent alcohol content of the wine
- quality: output variable (based on sensory data, score between 0 and 10)

2. Wine reviews: The data was scraped from WineEnthusiast during the week of June 15th, 2017.

```
(108959, 11)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 108959 entries, 0 to 108958
Data columns (total 11 columns):
Unnamed: 0     108959 non-null int64
country        108956 non-null object
description    108959 non-null object
designation    76734 non-null object
points         108959 non-null int64
price          98126 non-null float64
province       108956 non-null object
region_1       91134 non-null object
region_2       43341 non-null object
variety        108959 non-null object
winery         108959 non-null object
dtypes: float64(1), int64(2), object(8)
memory usage: 9.1+ MB
```

Fig 2. Information on wine reviews dataset

- country: The country that the wine is from

description

- designation: The vineyard within the winery where the grapes that made the wine are from
- points:The number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
- price: The cost for a bottle of the wine
- province: The province or state that the wine is from
- region_1: The wine growing area in a province or state (ie Napa)
- region_2: Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank
- taster_name
- taster_twitter_handle
- title: The title of the wine review, which often contains the vintage if you're interested in extracting that feature
- variety: The type of grapes used to make the wine (ie Pinot Noir)
- winery: The winery that made the wine

PROPOSED SYSTEM

Fig 3. shows our proposed system's flow chart. Model 1 uses Linear regression to predict the price by training the wine review dataset. Model 2 first uses Random forest to find out the important features in both white and red wine dataset obtained from chemical analysis data. Now we run the Logistic Regression to find the quality which is scaled to the factor of 100. By running model 1 again, we can predict the price for the estimated quality from chemical analysis.

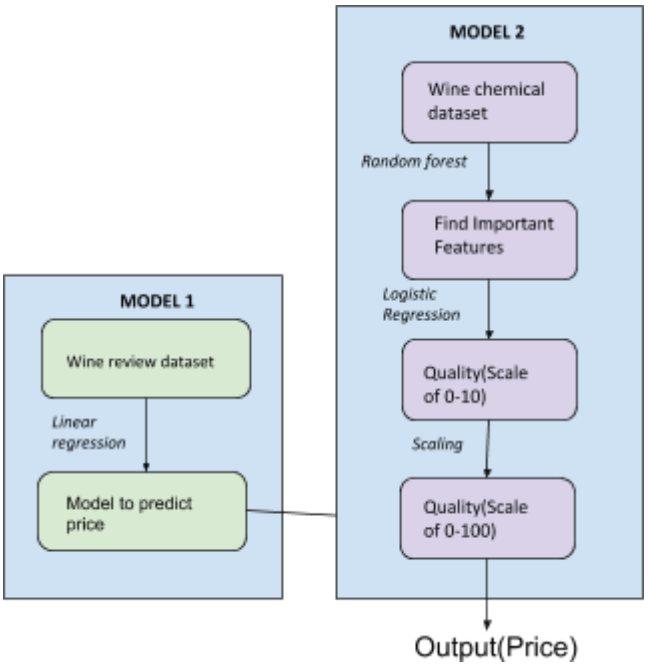


Fig 3. Our proposed system

LIST OF MODULES

1. Analysis of wine chemical dataset
2. Analysis of wine reviews datasets
3. Predicting the price of the wine based on sensory points
4. Estimation of quality based on chemical properties
5. Determining the price for the estimated quality
6. Comparing the results

I. Analysis of wine chemical datasets

The wine chemical dataset is first loaded by using pandas. Then the white and red wine datasets are divided into two separate frames.

Percentage of white wine: 75.38864091118978 %
Percentage of red wine: 24.611359088810218 %

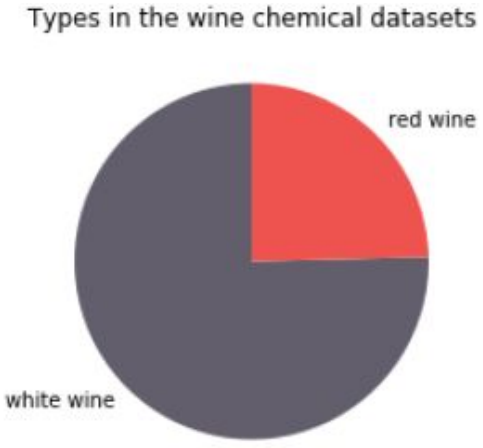


Fig 4. Percent of red and white wine in the wine chemical dataset

Fig 5. Red wine analysis: Minimum and maximum value for each feature under each quality

quality: 3

Name	min	max
fixed acidity	6.7	11.6
volatile acidity	0.44	1.58
citric acid	0.0	0.66
residual sugar	1.2	5.7
chlorides	0.061	0.267
free sulfur dioxide	3.0	34.0
total sulfur dioxide	9.0	49.0
density	0.99471	1.0008
pH	3.16	3.63
sulphates	0.4	0.86
alcohol	8.4	11.0

quality: 4

Name	min	max
fixed acidity	4.6	12.5
volatile acidity	0.23	1.13
citric acid	0.0	1.0
residual sugar	1.3	12.9
chlorides	0.045	0.61
free sulfur dioxide	3.0	41.0
total sulfur dioxide	7.0	119.0
density	0.9934	1.001
pH	2.74	3.9
sulphates	0.33	2.0
alcohol	9.0	13.1

quality: 5

Name	min	max
fixed acidity	5.0	15.9
volatile acidity	0.18	1.33
citric acid	0.0	0.79
residual sugar	1.2	15.5
chlorides	0.039	0.611
free sulfur dioxide	3.0	68.0
total sulfur dioxide	6.0	155.0
density	0.99256	1.00315
pH	2.88	3.74
sulphates	0.37	1.98
alcohol	8.5	14.9

quality: 6

Name	min	max
fixed acidity	4.7	14.3
volatile acidity	0.16	1.04
citric acid	0.0	0.78
residual sugar	0.9	15.4
chlorides	0.034	0.415
free sulfur dioxide	1.0	72.0
total sulfur dioxide	6.0	165.0
density	0.9900700000000001	1.00369
pH	2.86	4.01
sulphates	0.4	1.95
alcohol	8.4	14.0

quality: 7

Name	min	max
fixed acidity	4.9	15.6
volatile acidity	0.12	0.915
citric acid	0.0	0.76
residual sugar	1.2	8.9
chlorides	0.012	0.358
free sulfur dioxide	3.0	54.0
total sulfur dioxide	7.0	289.0
density	0.99064	1.0032
pH	2.92	3.78
sulphates	0.39	1.36
alcohol	9.2	14.0

quality: 8

Name	min	max
fixed acidity	5.0	12.6
volatile acidity	0.26	0.85
citric acid	0.03	0.72
residual sugar	1.4	6.4
chlorides	0.044000000000000004	0.086
free sulfur dioxide	3.0	42.0
total sulfur dioxide	12.0	88.0
density	0.9908	0.9988
pH	2.88	3.72
sulphates	0.63	1.1
alcohol	9.8	14.0

Fig 6. White wine analysis: Minimum and maximum value for each feature under each quality

quality: 3

Name	min	max
fixed acidity	4.2	11.8
volatile acidity	0.17	0.64
citric acid	0.21	0.47
residual sugar	0.7	16.2
chlorides	0.022000000000000002	0.244
free sulfur dioxide	5.0	289.0
total sulfur dioxide	19.0	440.0
density	0.9911	1.0001
pH	2.87	3.55
sulphates	0.28	0.74
alcohol	8.0	12.6

quality: 4

Name	min	max
fixed acidity	4.8	10.2
volatile acidity	0.11	1.1
citric acid	0.0	0.88
residual sugar	0.7	17.55
chlorides	0.013000000000000001	0.29
free sulfur dioxide	3.0	138.5
total sulfur dioxide	10.0	272.0
density	0.9892	1.0004
pH	2.83	3.72
sulphates	0.25	0.87
alcohol	8.4	13.5

quality: 5

Name	min	max
fixed acidity	4.5	10.3
volatile acidity	0.1	0.905
citric acid	0.0	1.0
residual sugar	0.6	23.5
chlorides	0.009000000000000001	0.3460000000000003
free sulfur dioxide	2.0	131.0
total sulfur dioxide	9.0	344.0
density	0.9872200000000001	1.00241
pH	2.79	3.79
sulphates	0.27	0.88
alcohol	8.0	13.6

quality: 6

Name	min	max
fixed acidity	3.8	14.2
volatile acidity	0.08	0.965
citric acid	0.0	1.66
residual sugar	0.7	65.8
chlorides	0.015	0.255
free sulfur dioxide	3.0	112.0
total sulfur dioxide	18.0	294.0
density	0.9875799999999999	1.03898
pH	2.72	3.81
sulphates	0.23	1.06
alcohol	8.5	14.0

quality: 7

Name	min	max
fixed acidity	4.2	9.2
volatile acidity	0.08	0.76
citric acid	0.01	0.74
residual sugar	0.9	19.25
chlorides	0.012	0.135
free sulfur dioxide	5.0	108.0
total sulfur dioxide	34.0	229.0
density	0.98711	1.0004
pH	2.84	3.82
sulphates	0.22	1.08
alcohol	8.6	14.2

quality: 8

Name	min	max
fixed acidity	3.9	8.2
volatile acidity	0.12	0.66
citric acid	0.04	0.74
residual sugar	0.8	14.8
chlorides	0.013999999999999999	0.121
free sulfur dioxide	6.0	105.0
total sulfur dioxide	59.0	212.5
density	0.98713	1.0006
pH	2.94	3.59
sulphates	0.25	0.95
alcohol	8.5	14.0

The number of wines in each quality point of both red and white wine are shown in fig 7.

Wines in each quality

Quality	white wine	red wine
1	0	0
2	0	0
3	20	10
4	163	53
5	1457	681
6	2198	638
7	880	199
8	175	18
9	5	0
10	0	0

Fig 7. Number of wine in each type

Since the quality points mentioned here is in the scale of 0-10, we will change it into scale of 100.

NewValue = (((OldValue - OldMin) * (NewMax - NewMin)) / (OldMax - OldMin)) + NewMin -(1)

II. Analysis of wine review datasets

The wine reviews are in two separate csv files. There is no specific column mentioned whether the wine is white or red wine, but the variety is given.

```
Common and unique varieties in both csv
are:['Syrah-Malbec','Grenache-Carignan',
'Primitivo','Silvaner-Traminer', Touriga
Nacional-CabernetSauvignon',....,
'Sylvaner', 'Malbec-Cabernet Franc']
Number of varieties : 578
```

From these variety, we select only the white and red wine varieties.

```
Number of white wine varieties
considered: 51
Number of red wine varieties
considered: 35
```

We divide the white and red dataset and, concat the data from both the csv.

III. Predicting the price of the wine based on sensory points

From the wine reviews data, we create a model for price against points(sensory points). We use Linear Regression to create the model.

```
from sklearn.linear_model import
LinearRegression
import numpy as np
model_price_white = LinearRegression()
model_price_white.fit(x_train,y_train)
y_pred = model_price_white.predict(x_test)
r =metrics.mean_squared_error(y_test,y_pred)
print("MEAN SQUARE ERROR: ", r)
print("ROOT MEAN SQUARE ERROR: ",np.sqrt(r))
```

```
MEAN SQUARE ERROR:  1002.4361745421751
ROOT MEAN SQUARE ERROR:  31.661272471936044
```

(a)

```
MEAN SQUARE ERROR:  1000.1440841486901
ROOT MEAN SQUARE ERROR:  31.62505469005058
```

(b)

Fig 8. Linear Regression: (a) white wine (b) red wine

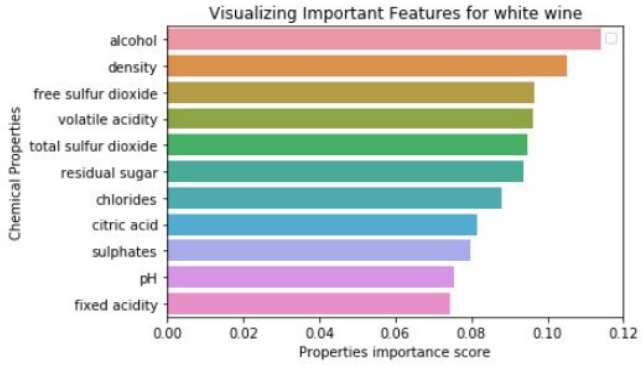
This model is used to predict prices(unknown) for the wine chemical dataset.

IV. Estimation of quality based on chemical properties

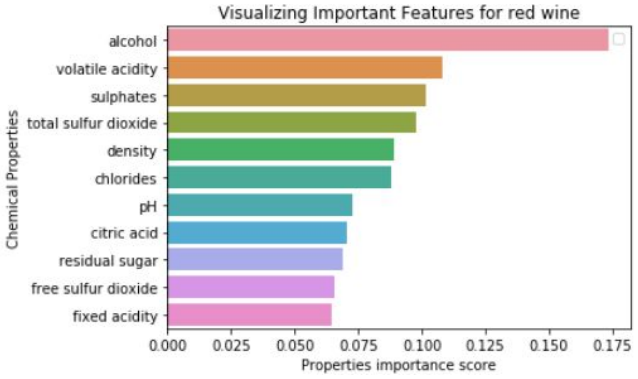
The random forest algorithm is used to find the important chemical properties that determine the quality of the wine.For this we take the wine chemical anlysis data.

```
from sklearn.ensemble import
RandomForestClassifier

m = RandomForestClassifier(n_estimators=10)
m.fit(x_train,y_train)
y_pred = m.predict(x_test)
print("PRECISION RECALL: ",
metrics.recall_score(y_test,y_pred,average='
weighted'))
```



(a)



(b)

Fig 9.Random forest: (a) white wine (b) red wine

From the results, we take the first five important features in both white wine and red wine. Coincidently, four features: alcohol, density, volatile acidity and total sulfur dioxide are common in both the cases. The other important feature in white wine, free sulfur dioxide does not much affect in the case of red wine. The same goes to the sulphates in red wine. Thus we can add them to the important features to estimate the quality.

We use Logistic regression to predict the quality.

```
feature = ['alcohol','density','volatile
acidity','free sulfur dioxide','total sulfur
dioxide','sulphates']
model = LogisticRegression()
model_white.fit(x,y)
```

The quality is predicted in the scale of 0-10. We use equation 1 to convert it into 0-100 scale.

V. Determining the price for the estimated quality

The price is calculated by running the model created in III step. We use the same Linear regression trained with the points and price of wine review dataset.

We can also use other dataset with points and price trained model to determine the price of the wine.

EXPERIMENTAL RESULTS

The price calculated from the III step and V step is compared. It is shown below that there is profit in white wine and loss in red wine. But, overall there is a profit.

```
Predicted price from sensory points
Total price for white wine:  175740.74
Total price for red wine:  82357.55
Total price :  258098.29
Predicted price from chemical analysis
Total predicted price for white wine:
208020.78
Total predicted price for red wine:
77284.04
Total predicted price :  285304.82
```

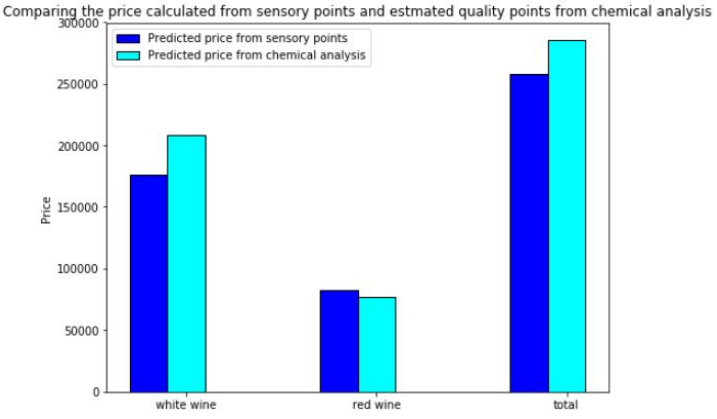


Fig 10. comparing the price with the price determined for the estimated quality

The profit percent and loss percent is calculated and shown:

```
Profit gained for white wine:
  12.506880594068939 %
```

```
Loss gained for red wine:
  1.9657291780859656 %
Total Profit gained:  10.54115141598298 %
```

quality	points	price	predicted quality	predicted points	predicted price
6	90.0	37.802585	5	88.0	28.354767
6	90.0	37.802585	5	88.0	28.354767
6	90.0	37.802585	6	92.0	47.250403
6	90.0	37.802585	6	92.0	47.250403
6	90.0	37.802585	6	92.0	47.250403

Fig 11.From the white wine

quality	points	price	predicted quality	predicted points	predicted price
5	88.0	38.060375	5	88.0	38.060375
5	88.0	38.060375	5	88.0	38.060375
5	88.0	38.060375	5	88.0	38.060375
6	92.0	59.200014	5	88.0	38.060375
5	88.0	38.060375	5	88.0	38.060375

Fig 12.From the red wine

Fig. 11 and 12 will show the final table after applying various models. Fig 13 is the application we created that uses our proposed model for finding the quality and price from chemical properties.

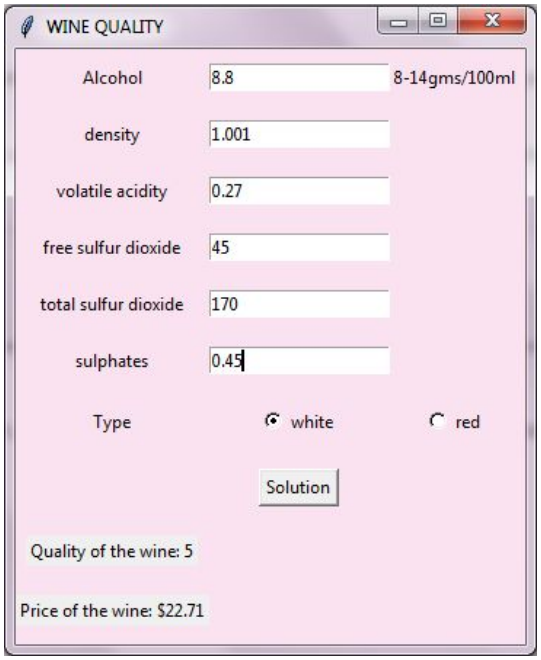


Fig 13. UI of the application

CONCLUSION

Sensory points are awarded by tasting the wine. This is not standard and will vary among many tasters. Hence estimating its quality from the sensory points is not accurate.

Our method estimates the quality from the chemical properties, which even the computer can calculate.The price can be changed according to the winery, if they can train the computer with their own quality vs price datasets.

REFERENCES

1. Red and White Wine Analysis, by Ian Xiao on June 27, 2015
2. Correlating Wine Quality Indicators to Chemical and Sensory by Helene Hopfer, Jenny Nelson, Susan E. Ebeler and Hildegarde Heymann on Molecules 2015, 20
3. How to Choose a Good Wine.
4. List of white wine varieties: http://frenchscout.com/white-wine-varietals
5. List of red wine varieties: http://frenchscout.com/red-wine-varietals

Datasets:

1. Wine chemical dataset downloaded from kaggle provided by Raj Kumar
2. Wine Review downloaded from kaggle provided by zackthout