Wine quality prediction and type classification.

Vaibhav Somani
Arizona State University
Tempe, 85281, Arizona, USA
vsomani3@asu.edu

ABSTRACT

With growing consumer awareness, companies and corporations are required to label their products to ensure the consumer of the product quality, which in the food industry holds vital significance as the companies need to provide labels with accurate nutrition facts and the ingredients used. Therefore, quality management has become an integral part of every company. In this project, I will be evaluating Knn, Random Forest, and linear svm on 3 data sets, one with all features, 2nd with feature selected using linear regression (to determine the dependency of target variables on independent variable) as proposed by Y.Gupta[1] and 3rd with PCA reduction technique. Predictions of the quality of wine have been documented using physicochemical properties previously utilizing various techniques. But can I utilize these properties of wine to predict the type of wine(ie, red or white)? Given that I have the same attributes provided for both, the red and the white wine, I will be attempting to classify wine into red or white.

ACM Reference Format:
Vaibhav Somani. 2022. Wine quality prediction and type classification.. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Over the years, Quality Analysis has been a manual process, requiring human expertise. But with the rapid increase in production, repeatable and automated quality analysis methods are now required. With the advancement in Machine Learning and data mining techniques, and with an increasing dataset of records of properties on which the production and final product depend, I can utilize different models and techniques to understand the data and make predictions on the quality of the product. Moreover, Quality management not only allows the company to acquire certifications, and assure the consumers(especially with regards to adulteration), but also provides the producer with an opportunity to define parameters to estimate and control the quality of the final product even before the production starts. With its long aging process, wine production can be a beneficiary of such a set of parameters, since the producer can estimate the quality of the final product in advance, allowing them to make crucial decisions in the initial stages, saving years of work and investment. The long aging process also adds to the complexity of wine cost, as white wine requires way less time, thus providing an opportunity for fraudulent behaviour such as adultration of wine, and counterfeiting white wine as red to increase profit margins. Wine, in particular, is interesting to analyze due to its physicochemical properties. The UCL Machine Learning Repository dataset for wine(I will be using this data set for creating our framework) provides us with 11 such physicochemical properties/parameters: fixed acidity (g/dm3), volatile acidity (g/dm3), total sulfur dioxide (mg/dm3), chlorides (g/dm3), pH level, free sulfur dioxide (mg/dm3), density (g/cm3), residual sugar (g/dm3), citric acid (g/dm3), sulfates (g/dm3), and alcohol (vol by percentage). Using these, I can keep a track of the quality of the wine from the start to the end of production. This also allows us to record the initial measurements and the process to reproduce the same quality every time. Before I start working on the data, I need to preprocess the dataset due to variable amplitude for different parameters, thus requiring standardization of data. I will also need to take care of irrelevant, redundant, and insignificant predictors while preprocessing data. Further, I will be making a 3:1 data split for training and testing. After preprocessing, I will analyze Y.Gupta's[1] feature selection using linear regression to determine the target variable's dependency on independent variables. Using these best predictors, I will create a new dataset. Another dataset will be created using the PCA dimensionality reduction technique. All further work will be done on each of the 3 data sets. I will contrast k nearest neighbor, tree-based classifier, naive Bayes, SVM, and Random Forest, as to determine the best fit models for the dataset in hand, and finally test an ensemble of the best performing models. I will be creating a separate dataset, with equal records of red and white wine(since I have fewer records for red wine) with an extra attribute type(0 for red, 1 for white). I will be using this data set to predict the type of wine The main contribution of the proposed approaches are - e -

- The paper will provide a comprehensive understanding of when and where to use the feature selection techniques and dimensionality reduction. The paper will also delve into understanding the difference between feature selection and dimensionality reduction.
- To classify the wine into two different types(Red and White) based on the 11 physicochemical properties and thus determine the correlation between the 2 types.

The paper is organized as follows - Section 2: Related work - an overview of the papers referred and the related work previously performed on the wine quality analysis and classification. Section 3: Methodology - This section describes the methods and techniques used for the analysis. The section will be divided into the following subsections: 3.1- Understanding dataset and Preprocessing - This section intends to provide a holistic view of the dataset used, and the preprocessing of the data. 3.2: Models - Understanding and implementation of the models(Stated in the introduction). Section

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

4: Results and evaluation - This section discusses the results of the preprocessing(and attribute selection), the different models used, and the bias in the dataset. Section 5: Conclusion and discussion about future prospects and improvement in the implementation and framework obtained.

2 RELATED WORK

Given the luxurious status of wine, the classification and prediction of wine had been around for a good time. Sun[2] classified wine based on geographic information. Grape maturity level and chemicals were used by Vlassides[3], and similarly, other properties have been used to classify wine. By far these classifications have dealt with the problem of a very less number of records to work with. The UCI data set for wine quality with 11 physicochemical properties provide us with a data set with 5000+ records, which I will be utilized in this project. I will be referring to 3 papers, utilizing Y.Gupta's[1] work to understand and use feature selection, and the [4] and [5] to access our results by comparing the different models used by [4] and [5]

- Y.Gupta[1] Uses linear regression to determine the dependency of wine quality on independent variables. After extracting the most important features, a new data set is created, over which Neural Network and Support Vector Machine have been implemented. Details regarding the selection of features will be discussed further in the methods section. According to the paper, SVM performed best and concluded that feature selection helped improve the performance.
- Paulo Cortez[4] Uses SVM and Neural NRetworks. I will refer to this paper to evaluate our work with respect to results achieved by [4]
- Terry Hui-Ye Chiu[5] provides an interesting methodology for using GA-based hybrid models. They first encode a set of classifiers and hyperparameters into a chromosome. The fitness functions including the accuracy and macro-F1 score are employed to evaluate the goodness of every chromosome. After the evolution process, the appropriate hybrid model and the hyperparameters are used for wine quality prediction. I will be contrasting this work with the ensemble method I will be creating to test how an advanced genetic-based algorithm works compared to an ensemble.

3 METHODOLOGY

3.1 Pre-processing

First, I will be following the provided steps for both, the red and the white wine dataset individually-

1 . Understanding and preprocessing the wine dataset.
2 . Creating a new dataset utilizing Y.Gupta's[1] feature selection method.
3 . Analyzing PCA to select the number of components, and thus transform the dataset into a new dataset.

Second, I will be creating a new dataset with equal records of red and white wine. I will be using the same process described above on this data set, but I will need to redo feature selection to test if the quality of the dataset could help classify the wine into red and white.

3.1.1 Red Wine. At the first glance at the description of the red wine dataset, I notice the amplitude difference between different properties. Thus standardization is required.

Table 1: Data sample - Red wine

index	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
0	7.4	0.7	0.0	1.9	0.076
1	7.8	0.88	0.0	2.6	0.098
2	7.8	0.76	0.04	2.3	0.092
3	11.2	0.28	0.56	1.9	0.075
4	7.4	0.7	0.0	1.9	0.076

Table 2: Data Description - Red wine

index	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
count	1599.0	1599.0	1599.0	1599.0	1599.0
mean	8.0	1.0	0.0	3.0	0.0
std	2.0	0.0	0.0	1.0	0.0
min	5.0	0.0	0.0	1.0	0.0
25%	7.0	0.0	0.0	2.0	0.0
50%	8.0	1.0	0.0	2.0	0.0
75%	9.0	1.0	0.0	3.0	0.0
max	16.0	2.0	1.0	16.0	1.0

I also notice that I only have qualities between 3 and 8 -

Quality	distribution
5	681
6	638
7	199
4	53
8	18
3	10

It is worth noting that our dataset contains duplicate values but they cannot be disregarded as data collected from vine-wards can be the same if the wine process used by them is the same or related.

Now let's have a look at the correlation heatmap-

I see that alcohol has the highest positive correlation with wine quality, followed by various other variables such as acidity, sulfates, density chlorides. I notice that Fixed acidity has a high correlation with citric acid and density. These features were disregarded by Y. Gupta's [1] feature selection as well.

Next, let's look at the distribution of the data -

I see that the distribution of the attributes "density" and "pH" are quite normally distributed, while "alcohol" seems to be positively skewed.

Our dataset does not contain any null values, thus I do not need to handle null values.

Now that our dataset has been standardized and quality converted to 3 classes, I will go forward and create a new data set attributes selected by utilizing Y.Gupta's [1] feature selection.

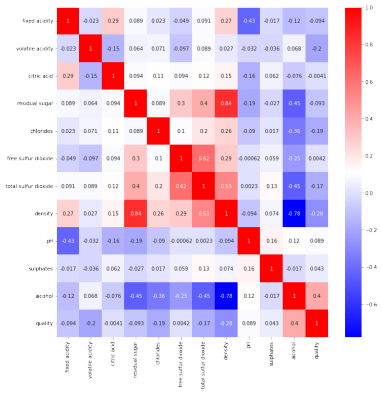

Figure 1: Correlation heatmap

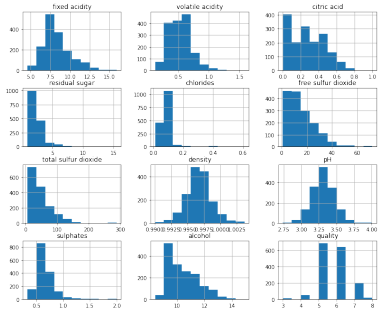

Figure 2: Distribution of the dataset

Next, I use PCA to reduce the dimensions of the data set and create a new dataset with the transformed dimensions.

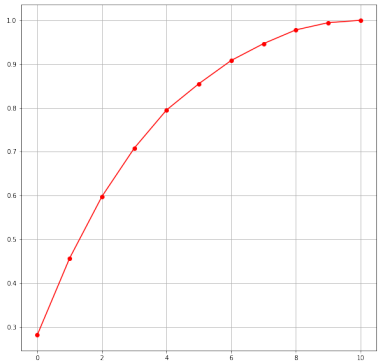

Figure 3: PCA Explained variance ration

I see that 6 principal components attribute to 90 percent of the variation in the data. I will transform our data using 6 principal components.

Now let's look at the white wine dataset as I will be utilizing it for Wine Type prediction.

I have sampled the same number of white wine records as red, and a new data set is created. Let's look at the data distribution, heatmap, and PCA's.

Table 3: Data sample - white wine

index	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
0	7.0	0.27	0.36	20.7	0.045
1	6.3	0.3	0.34	1.6	0.049
2	8.1	0.28	0.4	6.9	0.05
3	7.2	0.23	0.32	8.5	0.058
4	7.2	0.23	0.32	8.5	0.058

Table 4: Data Discription - white wine

index	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
count	4898.0	4898.0	4898.0	4898.0	4898.0
mean	7.0	0.0	0.0	6.0	0.0
std	1.0	0.0	0.0	5.0	0.0
min	4.0	0.0	0.0	1.0	0.0
25%	6.0	0.0	0.0	2.0	0.0
50%	7.0	0.0	0.0	5.0	0.0
75%	7.0	0.0	0.0	10.0	0.0
max	14.0	1.0	2.0	66.0	0.0

Table 5: Data Description - Both wine Combined

index	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides
count	6497.0	6497.0	6497.0	6497.0	6497.0
mean	7.0	0.0	0.0	5.0	0.0
std	1.0	0.0	0.0	5.0	0.0
min	4.0	0.0	0.0	1.0	0.0
25%	6.0	0.0	0.0	2.0	0.0
50%	7.0	0.0	0.0	3.0	0.0
75%	8.0	0.0	0.0	8.0	0.0
max	16.0	2.0	2.0	66.0	1.0

Figure 4: Correlation heatmap

From the regression report, 'fixed acidity','volatile acidity', 'residual sugar', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', and 'alcohol', are extracted.
For PCA reduction, we will be reducing to 6 dimensions.

3.1.2 PCA. Principal component analysis is a technique for analyzing datasets containing a high number of features, increasing the interpretability of data while preserving the maximum amount

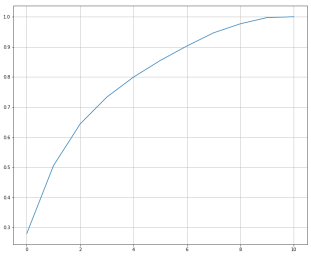

Figure 5: PCA

OLS Regression Results

Dep. Variable:	type	R-squared (uncentered):	0.279
Model:	OLS	Adj. R-squared (uncentered):	0.277
Method:	Least Squares	F-statistic:	227.7
Date:	Fri, 09 Dec 2022	Prob (F-statistic):	0.00
Time:	19:26:49	Log-Likelihood:	-8158.1
No. Observations:	6497	AIC:	1.634e+04
Df Residuals:	6486	BIC:	1.641e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
fixed acidity	0.1306	0.023	5.594	0.000	0.085	0.176
volatile acidity	-0.0908	0.015	-6.169	0.000	-0.120	-0.062
citric acid	0.0198	0.013	1.482	0.138	-0.006	0.046
residual sugar	0.3803	0.028	13.413	0.000	0.325	0.436
chlorides	-0.0231	0.013	-1.713	0.087	-0.049	0.003
free sulfur dioxide	-0.0978	0.015	-6.345	0.000	-0.128	-0.068
total sulfur dioxide	0.2531	0.018	14.004	0.000	0.218	0.289
density	-0.5889	0.042	-13.996	0.000	-0.671	-0.506
pH	0.0605	0.017	3.602	0.000	0.028	0.093
sulphates	-0.0205	0.013	-1.566	0.117	-0.046	0.005
alcohol	-0.1996	0.023	-8.654	0.000	-0.245	-0.154

Omnibus:	1783.565	Durbin-Watson:	0.095
Prob(Omnibus):	0.000	Jarque-Bera (JB):	592.617
Skew:	-0.544	Prob(JB):	2.86e-129
Kurtosis:	1.997	Cond. No.	9.61

Figure 6: Regression report

of information, and enabling the visualization of multidimensional data.
Equation used -
$X = \{x_n\}_{n=1}^{N}$

3.1.3 Linear Regression using OLS. Linear Regression using OLS estimates coefficients of linear regression equations thus describing the relationship between one or more independent quantitative variables and a dependent variable (simple or multiple linear regression).
. Equation used -
. $Y = 0 + j = 1..pjXj+$

3.1.4 Initial Hypothesis. Looking at the data from both datasets, I have attributes that differ a lot in range, which could help us determine the type of wine. One interesting finding is that PCA on the overall data suggests that reducing dimensionality to 6 should cover more than 90 percent. The correlation between various variables could justify this. I predict that Feature selection will work better for the red wine data set for quality analysis as compared to PCA since there is a correlation between variables, but it is not high. For predicting the type of wine, I predict that we will not see much accuracy difference in feature selection, PCA, and no feature reduction or selection as the data from red and white wine suggest that there are features in red and white wine data sets that differ in range. The value of Adjusted R2 is 0.3561 which shows a 35.61 percent dependency of quality on all predictors as a whole. This thus indicates feature selection will be helpful in increasing accuracy.

3.2 Models

3.2.1 SVM. The support vector machine (SVM) is a supervised machine learning model for solving a classification problem. The central concept of SVM is to utilize the kernel function to find the hyperplane that can separate instances into categories. It utilizes a nonlinear mapping to change the primary preparing information into a higher estimation. It scans for the linear optimal isolating hyperplane in this new estimation. A hyperplane can isolate information from two classes, with a reasonable nonlinear mapping to adequately high estimation. An SVM model is a portrayal of the models as points in space, mapped with the goal that instances of the different classes are isolated by a gap that is as wide as would be prudent. SVM can play out a nonlinear type of classification.

3.2.2 K nearest neighbor. The k-nearest-neighbours (kNN) classifies a data record by using its k-nearest neighbors, where these k-nearest neighbors are collected, forming a neighborhood. The majority vote among data records in the neighborhood is used to determine the classification for the target, with or without taking distance-based weighting into account. However, in order to use kNN, we must first select an appropriate value for k, and the classification's success is heavily dependent on this number. In certain ways, the kNN approach is influenced by k.

3.2.3 Random Forest. Random forest is a supervised learning algorithm, and several studies have shown that using RF can provide good prediction accuracy. In general, the RF algorithm creates different decision trees using randomly sampled instances. Then, in the prediction phase, based on the prediction results of the trees, a voting technique is used to determine the best solution. Due to using multiple decision trees for prediction, the advantage of RF over other methods is that it can reduce overfitting.

3.2.4 Gridsearch Hyper-Parameter tuning. We will be using gridsearchcv from sklearn library for tuning our parameter. It provides an exhaustive search over specified parameter values for an estimator. Since random forest and knn requires parameters to be tuned, we need to first find the set of best parameters.
These are the parameters used selected for knn and random forest based of gridSearch for red wine quality analysis -

knn - all attributes - 'metric': 'cosine', 'n$_neighbors'$: 36,$'$ $weights'$:$'$ $distance'$
$knn-FeatureSelection-'metric'$:$'$ $l1'$,$'$ $n_neighbors'$: 35,$'$ $weights'$:$'$ $distance'$
$knn-PCAreduction-'metric'$:$'$ $l1'$,$'$ $n_neighbors'$: 37,$'$ $weights'$:$'$ $distance'$

$RF-allattributes-'criterion'$:$'$ $gini'$,$'$ $max_features'$:$'$ $sqrt'$,$'$ $n_estimators'$: 75
$RF-FeatureSelection-'criterion'$:$'$ $entropy'$,$'$ $max_features'$:$'$ $log2'$,$'$ $n_estimato$
$RF-PCAreduction-'criterion'$:$'$ $entropy'$,$'$ $max_features'$:$'$ $log2'$,$'$ $n_estimators'$

4 RESULTS AND EVALUATION

4.1 Red Wine Quality Analysis

We received the following accuracy for our models-

Table 6: Accuracy Scoreboard

Algos		Accuracy
SVM	All attributes	0.64
	Feature Selection	0.64
	PCA Reduction	0.6
Knn	All attributes	0.64
	Feature Selection	0.67
	PCA Reduction	0.66
Random Forest	All attributes	0.66
	Feature Selection	0.68
	PCA Reduction	0.67

Table 7: Accuracy Scoreboard

Algo		Accuracy
SVM	All attributes	0.74
	Feature Selection	0.74
	PCA Reduction	0.74
Knn	All attributes	0.74
	Feature Selection	0.73
	PCA Reduction	0.73
Random forest	All attributes	0.71
	Feature Selection	0.71
	PCA Reduction	0.71

Random Forest with feature selection gives the most accuracy with Feature selection, while PCA reduction on SVM performed the worst.

Random Forest on average performed the best. PCA reduction does not show any boost as compared to all features, which gives less accuracy for SVM. This can be explained due to the correlations between variables being around 30 percent to 70 percent, which indicates that PCA might be effective for some algorithms. Feature selection improved the quality of all the methods implemented. Using p-value and regression to select features does help boost the scores.

4.1.1 PCA vs Feature Selection. The wine dataset provides us with a really good opportunity to understand the nuances between Dimensionality reduction and Feature Selection. Following are the important differences noted from the experiment-

- PCA reduction performs much better on the datasets with a high magnitude of predictors and has a high correlation between the predictors as we can save computational time by converting highly correlated predictors into a single dimension. In the wine dataset, we only have 11 predictors, and none of them have more than 0.7 correlation, thus reducing the dimensions might save us computational time, but will not provide us with a much-improved result.
- When we have a low percentage of dependency of quality on all predictors as a whole (indicated by adjusted R squared value in regression results), it is better to perform feature selection as we have one or more predictors which are not good. Feature selection can also help us understand which predictors are the most important, as to be able to get a more in-depth understanding of the product, which in our case is the red wine.

4.2 Wine Type Analysis

We received the following accuracy for our models-

We were able to get 74 percent accuracy, thus these could potentially help find a counterfeit wine from the original.

Interestingly, feature selection and PCA reduction had no effect on the accuracy. This could be due to the difference in the range of individual attributes between red and white wine which helps us get approx 75 percent accuracy, but due to not so high correlation between the attributes, failed to provide any reasonable reduction or feature extraction.

5 CONCLUSION AND DISCUSSION

Estimating the final quality of the product even before production starts plays a crucial role in deciding the processes to opt for and the measures to take. This in turn is beneficial for the producer to produce the same quality product every time and thus be able to provide the right information to the end consumer. As consumer awareness grows, we tend to require labels on almost everything, special if the category is food. But one problem that still stays afloat is adulteration and counterfeit products. I attempted to classify wines based on their physio-chemical properties, as to be able to know if a physicochemical test over a product can help find adulteration and counterfeit products. I received an accuracy of 74 percent, which indicates that 2 products, with the same physio-chemical properties, can be tested against each other to know if the product is in fact what was promised to the consumer. Further, the same approach can be improved using more complex models and can be extended to other food products to help reduce adulterated and counterfeit products.

6 REFERENCES

[1] Modeling wine preferences by data mining from physicochemical properties - Y.Gupta - DOI - https://doi.org/10.1016/j.dss.2009.05.016
[2]Sun, LX., Danzer, K. Thiel, G. Classification of wine samples by means of artificial neural networks and discrimination analytical methods. Fresenius J Anal Chem 359, 143–149 (1997). https://doi.org/10.1007/s00216
[3] Using historical data for bioprocess optimization: modeling wine characteristics using artificial neural networks and archived process information - Vlassides, Ferrier and Block - DOI - 10.1002/1097-0290(20010405)73:1<55::aid-bit1036>3.0.co;2-5
[4]Modeling wine preferences by data mining from physicochemical properties - Paulo Cortez,António Cerdeira,Fernando Almeida,Telmo Matos,José Reis - DOI - https://doi.org/10.1016/j.dss.2009.05.016
[5] A Generalized Wine Quality Prediction Framework by Evolutionary Algorithms - Terry Hui-Ye Chiu, Chienwen Wu, Chun-Hao Chen - https://doi.org/10.9781/ijimai.2021.04.006
//TODO