Portuguese Vinho Verde
Psychochemical Data Analysis

M^{a} Almeida 97136

Nuno Rosinha 34290

Sebastião Almeida 97115

Sebastião MC 97108

Abstract

Quality assessment is an important issue in the wine industry. Wine quality is currently assessed mainly by physicochemical (e.g. alcohol levels) and sensory (e.g. expert human studies) tests, the later having a certain component of subjectivity related to human factor. In this work we adopted a statistic approach to predict wine quality, based on easily available analytical tests. A dataset is considered with white and red *vinho verde* samples from the Minho region of Portugal and wine quality is modeled under a Suport Vector Machine, Decision Trees and Random Forests. Unsupervised clustering methods are also applied to the dataset, for comparison with the original partition into groups by quality.

1 Introduction

There is no doubt that wine has an important role in the Portuguese culture. In fact, we are known for having some of the best wines in the world. Considering that, it is in our best interest to know what defines its quality and what distinguishes between different types of wine. Considering this, we have chosen to study wine in its different specific characteristics. To do so we will work with a dataset regarding two different types of *vinho verde*, red and white, described through 12 features. We will then perform a classification task that will classify wine according to its quality and do a clustering analysis that will distinguish the different wines. We believe that the results of this project may be of great interest, not only as to know what features will define a great wine, but also to learn more about the differences in the different types of *vinho verde.*

2 Methods

2.1 Feature Description

The first step into any data analysis should be the one of knowing the data we will be working with. Firstly, we are dealing with two subsets, one for the white wine and another for the red wine, with 4898 and 1599 observations respectively, defined by 12 features. We are aware that the difference in observations may influence our analysis in the sense that our classifier will most likely predict the data for white wine with more accuracy. On the other hand, regarding the twelve features mentioned, these concern specific aspects that define wine (Table 1) and are all numerical, except for "Quality" which will be considered as categorical and ordinal. In addition, both datasets have no missing values.

Table 1: Features, and corresponding description, considered for the red and white wine datasets.

Features	**Description**
Fixed Acidity (g(tartaric acid) / dm^3)	One of the main constituents of wine. Total acidity can be divided into "Fixed" and "Volatile".
Volatile Acidity (g(acetic acid) / dm^3)	One of the main constituents of wine. Total acidity can be divided into "Fixed" and "Volatile".
Citric Acid (g/dm^3)	Added in order to increase its acidity, type of fixed acidity.
Residual Sugar (g/dm^3)	Amount of sugar in the wine.
Chlorides (g(sodium chloride)/dm^3)	Amount of salt in the wine.
Free Sulfur Dioxide (mg/ dm^3)	Sulfur dioxide is used as an antioxidant and preservative. The "Free" one is not related to compounds already present in wine.
Total Sulfur Dioxide (mg/ dm^3)	Sum of both types of sulfur dioxide (the free and the one naturally present in other compounds of the wine).
Density (g/cm3)	Determined by the amount of alcohol, sugar, amongst others.
pH	Measures the degree of acidity compared to alkalinity.
Sulphates(g(potassium sulphate)/dm^3)	Type of ""sulfur"" that is added to preserve the wine. Part of it goes into free sulfur dioxide.
Alcohol(% vol)	Amount of alcohol in the wine.
Quality(0-10)	Three experts evaluated each wine and its median was saved.

2.1.1 Univariate Analysis

The next step was to perform a brief univariate analysis on each feature. Firstly, regarding the red wine, we noticed that the variables have very different ranges of values for most of the parameters (Table 2). Moreover, the variances also differ a lot for each feature. The fact that certain characteristics vary a lot from wine to wine while others are relatively similar is probably primarily a consequence of the different scales of measure used, but it may also reflect the presence of outliers. In addition, this is a good indicator that we will most likely need to consider normalizing the variables, as such different values can negatively influence the performance of certain classifiers. On the other hand, we also looked into the histograms of each feature and concluded that only the variables “Density” and “pH” follow an approximate normal distribution. Besides, by analysing the variable “Quality”, we saw that it is highly unbalanced, having more than 80% of the observation in the middle classes (5 and 6). That is, there are few very bad or very good wines. This scenario can present itself as a problem on the future analysis as, on the one hand, the classifiers will struggle to predict with accuracy the classes with less observations and, on the other hand, the clustering algorithms will most likely not replicate this partition at all. We then made a similar analysis for the white wine which leaded to same conclusions, that is, the variables have very different values between them, and the only one following an approximate normal distribution is "pH". In addition, the variable "Quality" is once more very unbalanced.

Table 2: Summary Statistics for the red wine - Call: Summary(winequalityred)

	Fixed Ac.	Volatile Ac.	Citric Acid	Residual Sugar	Chlorides
Mean	8.320	0.530	0.271	2.539	0.088
Median	7.900	0.527	0.260	2.200	0.079
Var.	3.031	0.032	0.038	1.988	0.002

	Free S.D.	Total S.D.	Density	pH	Sulphates	Alcohol	Quality
Mean	15.870	46.470	0.997	3.311	0.658	10,420	5.636
Median	14.000	38.000	0.997	3.310	0.620	10.200	6.000
Var.	109.400	1082.000	0.000	0.024	0.029	1.135	0.652

The final step on our univariate analysis was the one of comparing the two datasets. Firstly, we can see that the main differences occur in the features “Free Sulfur Dioxide” and “Total Sulfur Dioxide”, having lower values on the red wines; in “Residual Sugar”, being white wines sweeter than red wines; and, finally, in “Fixed Acidity” and “Volatile Acidity”, being red wines more acidic that white wines. Contrarily, certain features present very similar behaviours between the two datasets, those are “Density”, “pH”, “Alcohol” and “Quality”. This allows us to conclude that the type of *vinho verde*, red or white, isn’t a factor that will affect the quality of the wine. Given this information, from our point of view, the datasets present some differences between them and, because of that, we opted do the analysis on the datasets separately, choosing the classifier and the clustering method that leads to the best results in each case.

2.2 Principal Component Analysis

The first step on our Principal Components Analysis (PCA) was the one of choosing between robust or classical PCA and with or without normalized data. We opted to go with robust PCA, given that we are dealing with real data and this is a more secure approach, that is, the chances of existing outliers in the data are considerable. Regarding the normalized data, as the information loss could be considerable by doing so, we opted to test for both methods and chose the one whose Principal Components would have a better interpretation, by looking at their loadings. For both datasets this happened with the normalized data. To chose the number of principal components to retain we based on three criteria, those were the proportion of variance explained, the number of principal components above the average standard deviation and the analysis of the scree plot. That being said, we opted to keep 5 principal components for both datasets.

Table 3: PCA on the red wine- Call: PcaCov(x = data[, 1:11], scale = TRUE)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.773	1.475	1.33	1.053	0.891	0.85	0.757	0.57
Proportion of Variance	0.286	0.198	0.161	0.101	0.072	0.066	0.052	0.03
Cumulative Proportion	0.286	0.484	0.645	0.745	0.818	0.883	0.935	0.965

Table 4: PCA on the white wine- Call: PcaCov(x = data[, 1:11], scale = TRUE)

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.887	1.246	1.114	1.057	0.967	0.857	0.833	0.745
Proportion of Variance	0.324	0.141	0.113	0.102	0.085	0.067	0.063	0.050
Cumulative Proportion	0.324	0.465	0.577	0.679	0.764	0.831	0.894	0.944

2.3 Outlier identification and removal proposal

One of our main goals with this work is to try to understand if it is possible to train classifiers on the wine datasets that yield a good performance. Although some of the most common Machine Learning algorithms are robust enough to not be very sensitive to outliers, some of them are very sensitive to these observations. For that reason, it may be very important, for some of the methods, to use a version of the dataset where, for each class, the observations that deviate too much from the behaviour of the others have been removed. This does not necessarily mean that the removed points, which we refer to as "outliers", contain wrong measures, poorly conserved or poorly processed wines, although that can be the case for some of them, particularly the most extreme outliers. To identify these outliers, we utilized a method based on Robust Principal Component Analysis. The estimation of the robust PCA's, projection of the data, calculation of the Score Distance and Orthogonal Distance thresholds and classification of the data as outlier or not was done separately for each subset of each dataset containing all the observations of one class for one type of wine. A cutting value of 0.9999 was used for determining the thresholds, since we wanted to be conservative with the removal of points from the data. The information about the points removed for each class can be found on Table 5.

We acknowledge that the removal of these points comes with the removal of some information, and interesting observations might have been discarded, but our hope is that this cost is offset by the gain in performance of the classifiers, which would translate to a better real-world application.

Table 5: Number of outliers removed and proportion relative to the total number of observations.

	Quality 3	Quality 4	Quality 5	Quality 6	Quality 7	Quality 8
White Wine	3 (15%)	12 (7.36%)	56 (3.84%)	91 (4.14%)	15 (1.7%)	8 (4.57%)
Red Wine	n.a.	5 (9.43%)	48 (7.05%)	31 (4.86%)	7 (3.52%)	2 (11.1%)

2.4 Feature Selection

Considering that certain features could be irrelevant, and that to many features can make the models overfit, we opted to do feature selection. Moreover, taking into consideration that the variable "Quality" is categorical, we considered that the best way to do so should be through the method Minimum Redundancy Maximum Relevance (MRMR). In addition, we also considered better to perform this analysis on the dataset without the outliers, as this observations can influence the mutual information between the variables. As there is no rule of thumb into how many features we should keep, we opted to exclude the three most irrelevant variables on a first approach and, if the results present considerable improvements further on, we will be removing more features. According to the MRMR procedure the most irrelevant features were "Fixed Acidity", "Citric Acid" and "Free Sulfur Dioxide" for the red wine, and "Density", "Total Sulfur Dioxide" and "Residual Sugar" for the white wine. For both datasets the most relevant features were "Alcohol" and "Volatile Acidity".

2.5 Dataset Transformations

Considering the analysis performed previously we built several datasets. The first thing we approached was the imbalance of the classes in “Quality”, taking this into consideration, we opted to remove the most extreme classes - classes 3 and 8 on the red wine dataset and classes 3, 8 and 9 on the white wine dataset. This translated to a removal of 28 observations for the red wine dataset and 200 observations for the white wine dataset (approximately 1% and 4% of observations, respectively). In addition, we also grouped the classes 4 and 5, being the final three classes "5", "6" and "7". "Quality" has now 1620 observations for class 5, 2198 for class 6 and 1055 for class 7 for the white wine dataset and 734 observations for class 5, 638 for class 6 and 217 for class 7, for the red wine dataset.

On the other hand, we also built a dataset without the outliers, for both red and white wine. Furthermore, considering the feature selection procedure mentioned previously, we also made a new dataset for each wine without the three most irrelevant features. Finally, considering the Principal Components Analysis, we built a dataset using the new components as the new variables. These datasets have the 5 chosen Principal Components as their variables, which are a linear combination of the previous features.

2.6 Supervised Learning

2.6.1 Support Vector Machine

The first classifier we tested was a Support Vector Machine (SVM), due to the fact that is an algorithm that we have worked with and which has shown good performance in classification tasks. This is a consequence of, amongst others, the fact that it uses the "Kernel Trick" allowing it to project the data into a higher dimensional space, while keeping the computations in the lower dimensional space.

In the present scenario we opted to test for different types of kernel as we weren’t sure which would lead to the best result. In addition, we had to define a set of hyperparameters related to each kernel - the degree on the polynomial kernel, the Gamma value and the cost. To do so we opted to, once more, test for different values and chose the ones that would lead to the best result. Furthermore, to train and test this classifier we chose to do 10 fold cross validation having 75% of observations for training and the remaining for testing. We opted for this procedure as we believe that it will lead to more robust results. The measure of evaluation considered was accuracy. Moreover, we also opted to normalize the data for this classifier as it is good practice to normalize the features when using a Support Vector Machine, the normalisation was made through standardisation. The results obtained for each kernel, on the different datasets, can be seen on Tables 7 and 8. In addition, it’s relevant to underline that these results, except for the ones highlighted on the table, consider cost equal to 1, the degree of the polynomial kernel equal to 3 and the gamma value equal to 1 divided by the dimension of the dataset.

By analysing Tables 6 and 7 we can see that the best results happened for the Radial Kernel in both datasets, achieving a maximum accuracy for the dataset with three classes.

This was around 0.666 and 0.679 for the red and white wine, respectively. However, considering that the second best results happened on the original dataset and these didn't present a major change in accuracy, more specifically a difference of 0.026 for the red wine and 0.031 for the white wine, we considered the best option to be keeping the original dataset, as it is a less complex approach and keeps the complete dataset. Regarding the remaining datasets, the results didn't go as expected, as neither has presented considerable different results.

Table 6: Accuracy obtained using the SVM classifier on the red wine.
'Gamma = 1 Cost = 6 "Cost = 9 "'Gamma = 0.1, Cost = 6 "" Gamma = 0.1

Dataset/ Kernel	Linear	Polynomial	Radial	Sigmoid
Original	0.569	0.581	0.640'	0.513
PCA	0.573"	0.473	0.492	0.438
3 Classes	0.618	0.626	0.666"'	0.537
Outliers	0.592	0.631	0.634	0.507
Feature Selection	0.573	0.574	0.631""	0.505

Table 7: Accuracy obtained using the SVM classifier on the white wine.
'Gamma = 1, Cost = 6 " Gamma = 1, Cost = 11 "' Gamma = 1, Cost = 7

Dataset/ Kernel	Linear	Polynomial	Radial	Sigmoid
Original	0.522	0.538	0.648'	0.453
PCA	0.449	0.449	0.447	0.407
3 Classes	0.547	0.560	0.679"	0.494
Outliers	0.526	0.533	0.643'	0.445
Feature Selection	0.515	0.519	0.629"'	0.452

2.6.2 Decision Trees & Random Forest

The second classifier we decided to use was a Decision Tree. This choice was made because they have a relatively straightforward representation, which results in a classification model that is quite easy to interpret, they are non-parametric, i.e., the algorithm does not make any assumptions about the underlying distribution of the data; and they can be used for both regression and classification problems. Considering we were using decision trees, we decided to also test one of the most common Ensemble methods using them, Random Forests, since Ensemble methods usually improve results, are robust to outliers and preserve the good characteristics of the base methods used. We decided to create both a decision tree classifier and a random forest classifier. For both models we used 10-Fold Cross Validation.

We recorded the accuracy score for each of the datasets and classifier and presented it in Table 8. We decided to inspect the tree for the original datasets and for the datasets that yielded the highest accuracy scores. Since we used 10-fold cross validation, meaning that we grew 10 different trees for each dataset, we recorded the features that were used most of the times for node splitting, which we can interpret as the set of most important features to predict the wine classification. For the original white wine dataset, we verified that "Alcohol" was always used for root split and that both "Volatile acidity" and "Free Sulfur Dioxide" were used to split the subsequent nodes. For the original red wine dataset, "Alcohol" was also always used to split the root node and, "Sulphates", "Total Sulfur

Dioxide" and "Volatile Acidity" were the features used in the nodes that followed. We can interpret this as a confirmation of our 2.5 section conclusion on what are the most relevant features to classify wine quality in both white and red datasets.

Table 8: Accuracy obtained using the Decision Tree and Random Forest for white and red wine.

Wine Type	White		Red	
Dataset	Decision Tree	Random Forest	Decision Tree	Random Forest
Original	0.526	0.697	0.561	0.707
PCA	0.442	0.621	0.489	0.628
3 Classes	0.557	0.736	0.620	0.737
Outliers	0.514	0.700	0.571	0.716
Feature Selection	0.526	0.702	0.566	0.708

2.7 Unsupervised Learning

2.7.1 Clustering

For the second part of this assignment we did a cluster analysis applying Hierarchical K-Means and K-Medoids. The choice of the algorithms was due to the fact that, in case of Hierarchical K-Means, we were interested in doing K-Means but aware that is highly sensitive to the choice of the initial centroids. Furthermore, considering K-Medoids, we chose this algorithm as it has some advantages when compared to K-Means, namely that is more robust to noise and outliers.

Firstly, starting with Hierarchical K-Means, we still had to give the algorithm a certain number of clusters. That being said, we opted to perform Hierarchical Clustering separately, using as distance metrics Single, Complete, Average and Ward, and then, by looking at the dendrograms, extract the optimal number of clusters. We also crossed this result with the one obtained by the average silhouette width, that is, we chose the number of clusters that would maximize this value. Considering this analysis we concluded that we should keep 2 clusters and that the best metric to be used on the Hierarchical step should be "Ward", for all datasets and for red and white wine. Regarding the number of clusters to be considered on the K-Medoids algorithm we once more looked at the average silhouette width which led to the same result of keeping two clusters. The next step was to compare the results obtained for the different methods on each dataset (Tables 9 and 10). Finally, it's worth mentioning that we opted to analyse the silhouette average width as it is a coefficient familiar to us and that allows for a good comparison between the two clustering methods.

By analysing Tables 9 and 10 we can see that the best results happened after applying feature selection for both wines. Moreover, we can see that the Hierarchical K-Means algorithm presents better results on the red wine dataset, and relatively similar results on the white wine dataset. Taking this into consideration we opted to do a deeper analysis on the K-Means algorithm performed on the dataset after feature selection as it seems, in our opinion, the best algorithm and dataset for this scenario.

Table 9: Clustering results for the Hierarchical K-Means and K-Medoids algorithms on the red wine and white wine, based on the average silhouette width.

Wine type	White		Red	
Datset	Hierarchical K-Means	K-Medoids	Hierarchical K-Means	K-Medoids
Original	0.506	0.506	0.603	0.579
Outliers	0.507	0.508	0.595	0.568
Feature Selection	0.578	0.563	0.647	0.621

Table 10: Hierarchical K-Means results for the white and red wine.

Wine type	White		Red	
	Cluster 1	Cluster 2	Cluster 1	Cluster 2
Size	1784	3114	1192	407
BSS	865372.4		1170116	
WSS	307029.0	255222.8	251802.4	312411.5
BSS/TSS	0.606		0.675	

In order to better understand the clusters created we looked at specific details in each cluster, namely the summary statistics for the features, the size, the within sum of squares, amongst others (Table 10). That being said, we concluded that the clusters have considerable differences in size and, by looking at the Between Sum of Squares (BSS) and Within Sum of Squares (WSS) we can also say that the clustering is slightly better for the red wine dataset. Both clusters have a similar constitution when considering of the quality of the wines they contain, so we tried to understand what variables were being used the most to discriminate between clusters, by analysing the summary statistics, and we concluded that all the variables have very similar values in both clusters, except for "Total Sulfur Dioxide" and "Free Sulfur Dioxide" on the red and white wine datasets, respectively (Figure 1).

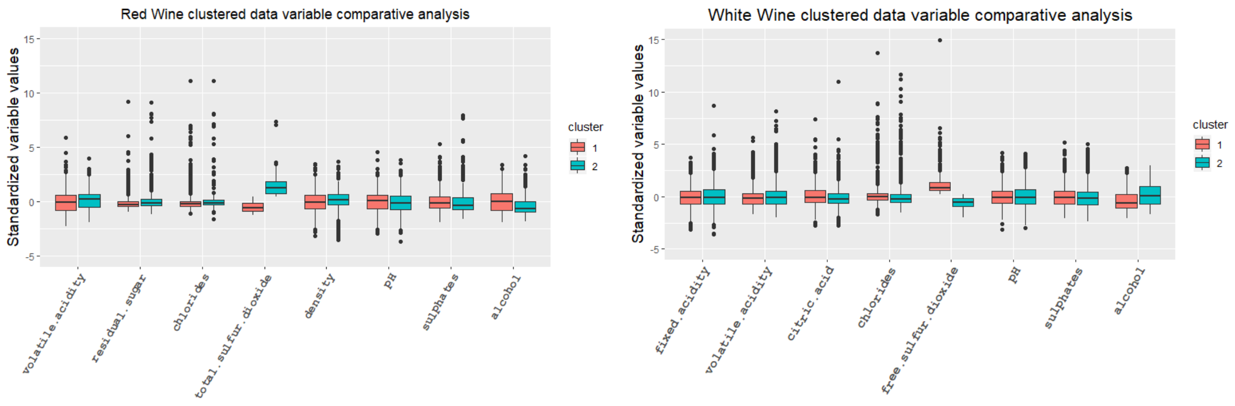

Figure 1: Behaviour of the features in the different clusters on the white wine dataset on the left and red wine dataset on the right.

With these results we can say that the amount of sulfur in the wine is the main reason for differentiating between the types of wine. In addition, by analysing the variable "Quality" summary statistics we can see that it is not influencing the cluster formation, as its values are similar in both clusters.

Finally, we would like to mention that the clustering algorithms were also made with the standardised datasets, however, the results were considerably worse when we standardised hence why we didn't refer them. Moreover, considering that the dataset after applying feature selection led to the best results we also tested for a dataset removing the 5 most irrelevant features, instead of 3, but the results didn't improve. Finally, the variable "Quality" wasn't considered while performing the clustering analysis, as it would biase the analysis.

2.8 Applying the Clustering Results

The final step on this study was the one of applying the clustering results obtained using the Hierarchical K-Means algorithm to the classifiers mentioned previously. Firstly, we should keep in mind that the clustering algorithms didn't differentiate the types of wine based on their quality, but in their amount of sulfur dioxide. That being said, the classifiers predictions will be completely different from the supervised learning performed previously. Moreover, we opted to apply the Suppor Vector Machine to the white wine dataset and the Random Forest to the red wine dataset, as they were the classifiers that got the best results for each dataset. Firstly, we started by splitting our dataset into two parts, with 70% of the observations in one part and the remaining on the other part. Then, we ran the Hierarchical K-Means algorithm separately for each part and extracted the clusters. Next, we used the biggest dataset to train the classifier and then tested it using the smaller dataset. The results, in case of the SVM for the white wine, translated to an accuracy of approximately 0.978. Regarding the Random Forest the accuracy was 0.987 for the red wine. These values are a good indicator that the cluster partition was well performed.

3 Conclusions

The different analysis presented have given us several insights on both wines datasets. Firstly, we were able to predict with an accuracy surrounding the 60% what will be the wine's quality, given the features stated. Moreover, considering the decision tree interpretation we can also say that the main features influencing its quality are "Alcohol" and "Volatile Acidity". We consider this to be a very interesting and relevant result for the wine producer. In addition, given the fact that changing the datasets on the classifiers didn't produce a considerable change on the overall accuracy we can conclude that the problem on the data doesn't concern redundant variables, outliers or the imbalance of classes. In fact, in our opinion, the main problem on the classification task is on the fact that "Quality" is ordinal, and the classification methods presented don't consider this component. On the other hand, regarding the clustering analysis, we can conclude that the discrimination between the different types of wine isn't related with their quality, but with the amount of sulfur in the wine. This is, once more, an interesting result as it shows us that the remaining variables don't have an impact when distinguishing the wines. The first one if we are interested in knowing what will define the quality of the wine and the last one if we want to know what will distinguish between the different types of wine, given a set of features. However, considering the problem at hand, we believe it's more relevant to do a further study on what will determine the quality of a wine. Finally, in order to improve this analysis we think it would be interesting to look at it from a regression point of view, as this would consider the label “Quality” as having an implicit order. Moreover, another key aspect should be increasing the number of very bad and very good wines, as it would enable the classifiers to predict those classes better.

References

[1] UC Davis, Waterhouse Lab. What is Wine. [Online]. Available: https://waterhouse.ucdavis.edu/tags/what-wine. [Accessed: 23-Dec-2019].

[2] R. A. Johnson and D. W. Wichern. Applied Multivariate Statistical Analysis, Sixth Edition, Prentice-Hall, New Jersey, 2007.

[3] A. J. Izenman. Modern Multivariate Statistical Techniques: Regression, Classication, and Manifold Learning. Springer, 2008.

[4] N. Jay, S. Papillon-Cavanagh, C. Olsen, G. Bontempi, and B. Haibe-Kains. "mRMRe: an R package for parallelized mRMR ensemble feature selection", Bioinformatics and Computational Biology Laboratory, Institut de recherches cliniques de Montreal, Canada and Machine Learning Group, Universite Libre de Bruxelles, Belgium, 2020, 7 pp.

[5] Carnegie Mellon University, School of Computer Science. Cross Validation. [Online]. Available: https://www.cs.cmu.edu/~schneide/tut5/node42.html. [Acessed: 28-Dec-2019]

[6] E. Osuna Edgar, Freund Robert and Girosi Federico. "Support Vector Machines: Training and Applications", Massachusetts Institute of Technology Artificial Intelligence Laboratory and Center for Biological and Computational Learning Department of Brain and Cognitive Sciences. March, 1997.

[7] Chang Chih-Chung, Hsu Chih-Wei and Lin Chih-Jen. "A Practical Guide to Support Vector Classification", National Taiwan University, Department of Computer Sciences. Taiwan, 2016.

[8] R Package Documentation, hkmeans: Hierarchical k-means clustering. [Online]. Available: https://rdrr.io/cran/factoextra/man/hkmeans.html?fbclid=IwAR19ObhXutYAWmG7HSO380K5ql3kzzdUXrfJkvtYkmKPj1FZQPZdurKXDmk. [Accessed: 28-Dec-2019].

[9] Cultural Modeling for Behavior Analysis and Prediction. [Online]. Available: https\\www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm?fbclid=IwAR2Y3LbdlLlL_ZWdy1Q655QUdaFuVs3wqBudPAr34qol0kIBLBHymHfm1ZQ. [Accessed: 03-Jan-2020].

[10] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.