Classification of Physicochemical Properties of Wine by Machine Learning Models

---Based on a dataset from a Portuguese region of wine production

Executive Summary

Owing to technological limitation, human experts of wine knowledge are the only eligible measurements for commenting and evaluating wine taste quality since ancient time. With the development of physicochemical measurement instruments, scientists are able to capture diverse wine attributes from the sample drops, which triggers them to rethink about the connections among these extracted features and the benchmarks inherited from ancestors. In the following paragraphs, different machine learning-based methods are applied on trying to understand the relationship between taste quality of wine and its corresponding data, background by a Portuguese wine speciality. Since this type of wine speciality are made by two distinctive grapes, the measured data are significantly different from each other, so the question of taste quality class will occur after the classification of wine types, which generates some calculated errors varied from model's performance. Fortunately, random forest model received the highest accuracy rate in both type and quality classification at 78%, showing that almost three quarters of unmarked wine would be known their taste before removing the corks, which is beneficial for wine manufactory to check the acceptance rate of prepare-to-launch products or make quality prediction in advertisement. Furthermore, it is interesting that the alcohol degree has become the most vital feature on determining the sensory taste of wine, which might be a good recommendation of wine industry to optimise their new products.

Introduction

To enjoy a bottle of tasty wine anytime and anywhere has been always concerned by its regular consumers. It is mainly because of the grape type that determines the wine taste quality as its nature controls how it is influenced by the environmental factors during growth, like temperature and light exposure (Fracassetti, D. et al., 2021). A good-quality wine hence is rarely to be massively produced. Wine tasting thus has warranted as a subcultural interest of high-class families that assign it with symbolic meaning of extravagance and curiosity (Teil, 2021). Several sociologists noticed that elite consumers and winemakers preferred to create unique wine language on measuring how their wine taste is completely distinguishing from the others, to emphasise social stratification through their 'supernormal knowledge' of taste sensitivity (Schwarz, 2013; Maguire, 2018). While business scholars debated that taste is an elusive perception that mingles the perceptions of extrinsic attributes out from product itself, such as brand advertising, promotional skills, and a predominated price (Lee et al., 2006; Charters & Pettigrew, 2007). Yet it is a weak robustness on

building a statistical system for classifying wine quality through consumer cognitive scoring. As their beliefs about one evaluable attribute affect the others, the scoring system will be biased. A systemic taste-preference bias happened when wine is labelled as organic product that fictional elements (e.g., health and vitality) are more added into the taste to make it better from some consumer's perspectives (Sörqvist et al., 2013). Reviewed the different understandings about organic products, the literatures on wine's halo effect remain ambiguous evaluations about its taste quality perception. Stolz and Schmid (2008) underlined that organic wine might be associated with inferior taste by abandoning complicated processing techniques, but still being welcomed by clients who embrace its notion of health life and environmentally friendly treatment.

Considered the variance under diverse comprehensions about wine language and its corresponding commercial concepts, are there any repeatably quantitative measurements about wine quality? Some recent studies about the characteristics of taste-activating compounds in wine taste have been carried out (Yu et al., 2015), discovering that the organic acids, residual sugars, chlorides, and sulphides are among the main compounds in the grapes. The more properties and concentrations of these components will heavily effect on wine's sensible flavours and intensity of taste characteristics. Different types of grapes contain different concentrations of these compounds. Physicochemical tools are commonly tested to explore the relationship between tasting-activating compounds and sensory traits of wine. For instance, to progress the wine certification for market launching, physicochemical laboratories have to test a variety of wine attributes, including the determination of chemical components, beverage density, alcohol content and PH assessments, which generates a series of float data, combined with the categorial benchmark modified by human experts (Ebeler, 1999). Because numerous characteristic information of wine has been stored into the dataset in terms of classes and numbers, it is possible for several advanced classification algorithms to attempt the construction of specific associations between these explanatory information and the perceived taste.

Cortez et al. (2009) collected and collated an integrated dataset of the wine production in the northwest region of Portugal from 2004 to 2007, which mainly concentrates a special product--- Vinho Verde, a medium alcohol wine made by both red and white grapes, particularly appreciated for its freshness and sweetness. Multiple physicochemical parameters and a discrete metric of quality evaluation were packaged inside.

Therefore, it is necessary to initially distinguish the wine type made by different fruit sources since they generated utterly distinctive data in explanatory parameters and then build models on identifying the quality class of each wine type.

Explanatory Data Analysis

Probably respected the privacy and logistic issues, Cortez et al. did not demonstrate any business-related information, like brand names or price of this speciality. On

contrast, their dataset included the raw material information into red and white grapes, displaying as eleven physicochemical attributes as explanatory parameters (e.g., acidity, sugar contents, chlorides, sulphur dioxide, concentration, alcohol degrees, and pH values). The only quantitative feature is the wine's sensory quality (abbr. quality), as output result. The entire dataset contains 6,497 samples, illustrated by eleven continuous variables and two discrete variables (Fig.1).

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	type
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	2	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	2	0
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	2	0
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	2	0
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	2	0

Fig.1 A summary of the Dataset (first five rows)

Because the entire digital information was summarised for more than a decade ago, it is customary for technical miss input and occasionally missing data, which might misinterpret the exploratory variables if the mistakes have occupied a large proportion of the dataset (Dong & Joanne, 2013). Fortunately, none of data marked as Not a Number thereof the whole dataset is meaningful nominally. In addition, the data of each sample are assumed to be generated randomly independent under unknown distributions because few theoretical cues show how they are correlated, which applicably fits the hypothesis of normal distribution used in social science. To check the outlier's influence thus becomes crucial to validate these variables has whether been tendentiously recorded. According to Maddala's definition about outliers under the normal distribution (1992), it refers to some data points are notably different from the rest observations away from the sample mean over three standard deviations. More outliers aggregate that a stronger skewness will shape as heavy top-or-tail effects. Thereof, the proportion of outlier's numbers on its measured parameter is firstly measured by setting the threshold limits at 0.01, four out of eleven variables are discovered of having heavy heads, illustrating their smaller values are probably over their variable averages (Fig.2-6).

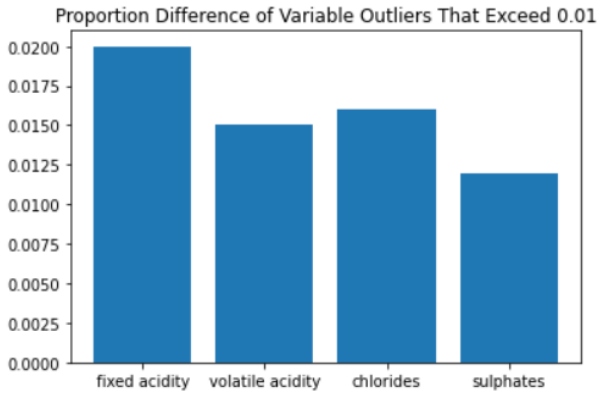

Fig.2 Proportion Difference of Variable's Outliers That Exceed 0.01

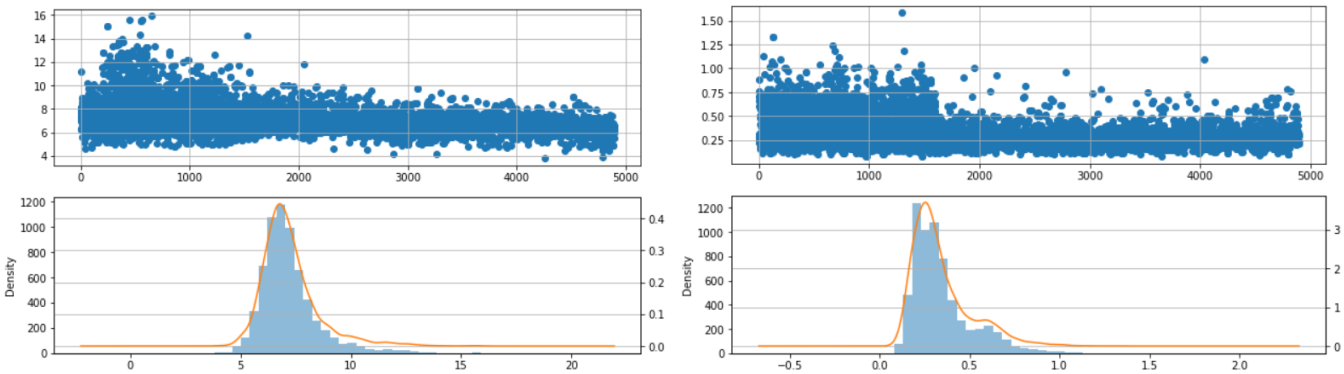

Fig.3-4 Scatter Plots and Distribution Plots of Fixed Acidity and Volatile Acidity (from left to right)

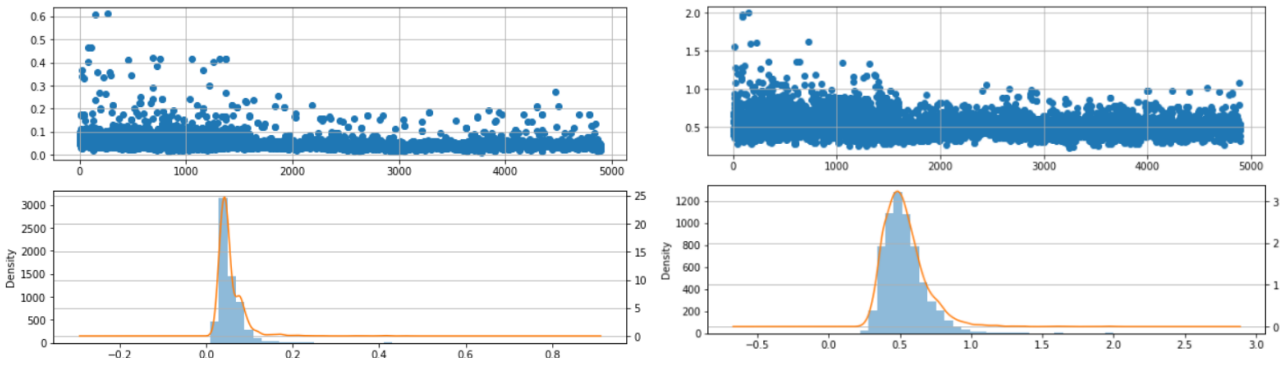

Fig.5-6 Scatter Plots and Distribution Plots of Chlorides and Sulphates (from left to right)

Cortez's research team used to classify their wine quality into ten different hierarchies, which is redundantly complicated for this experiment that has being shrunk into five categorical classes: 'Very Bad', 'Slightly Bad', 'Normal', 'Slightly Good', and 'Very Good', which is similar as 'Red Grapes' and 'White Grapes' in the binomial classification of wine types.

Ahead of model selecting, it is important to split the pre-processed dataset into training and test sets. A small-sized test set should be isolated and validated independently, otherwise its information will contaminate the training set by miss weighting, leading poor interpretation of models. Based on conventionally empirical rules, 80% of original data is allocated into training set for better training than usual.

The last procedure before classification modelling is to normalise the chosen explanatory features because the large absolute values inside them will empower themselves huge weights when comparing to the other smaller values in other columns, even these small values are the comparably dominators within their groups. To balance the intra-group weighting, a same range scaling technique should be applied to each variable' column through transforming their values within a fixed boundary---MinMax Scaler is appropriate to be introduced that restricts the whole observational data into [0,1] through punishing the outliers for their extremes. It counts the multiple relationships among the absolute difference of individual values, the difference to the top, and the scaled range from minimum to maximum, displaying as formula $MinMaxScaler = \frac{X - Xmin}{Xmax - Xmin}$. It helps researcher to protect the robustness of the other seven features that own small values in small standard deviations.

Model Construction and Performance Analysis

Because wine quality classification belongs to multiple element classification that scarcely be separated by linearity-based model, non-linearity-based models, like classification tree structures and neural network are recommended since no requirements about Gaussian assumption, though they have different emphasises on choice interpretation and hidden factor analysis. In this article, Decision Tree (includes

prune tree experiment), Bagging Tree, and Random Forest will be elaborated, while Artificial Neural Network will be a rivalry for comparing trees' model performance on classification correctness.

All experimented models are served for the entire dataset, so classification accuracy, the most valuable indicator, will check the interpretation on entire dataset throughout the training data-generated models from different patterns. To alleviate the chaos of simple drawing, each accuracy will be pre-processed by ten times of cross-validation (CV). If the accuracy results are in a similar level, the accuracy rates of test validation will secondarily be considered.

In order to make the two classification experiments coherently, the classification model for the raw material type will be initially constructed, which will be inherited by the wine quality model after comparing their performance.

Decision-Tree Model

In classification decision tree, the available result choices can be explained as sequential branches that assigning class labels to numerous hyper-rectangles generated by conditional features in an intuitive way that all of them have not been distorted by only applying normal weights during the pre-processing sections (James et al., 2013). A single tree separates the original data into multiple sub-rectangles through testing their purity that whether the conditions in the sub-zones equal to the conditions of target variables. If the balance has not been achieved, this process will recurse until no more space are split out.

To build an easy-to-understand model in this case, Gini Index, a measure to demonstrate the frequency at each node (splitting point) of data has been randomly mislabelled, is applied as the testing method of purity. It is interesting that both train and test datasets climb quicky at the starting point and eventually swing around at a certain value of accuracy when depth of split increases (Fig.7), which implies the maximum point of climbing acceleration is a 'sweet point' that overfitting will happen when it exceeds. It is because that a complex tree has been constructed through absorbing excessive information for split making, which results in good suitability in training set, but weak generalisation in test set (Mithrakumar, 2019). Therefore, the sixth split is selected as a threshold as test set's accuracy reaches the maximum.

Although the depth of maximum splits has been controlled, the number of leaf (conditional choices) attached to the tree structure are still too many for evaluating which feature is the most significant (Fig.8). A best solution is to prune the weakest nodes for reducing the calculation cost spend on insignificant leaves. After the cost complexity optimisation, only chlorides dominate the feature importance, and the test dataset accuracy maintains at 91.3%, while the CV accuracy of the entire dataset arises to 91.7%, meaning that over 91% of grapes are correctly classified.

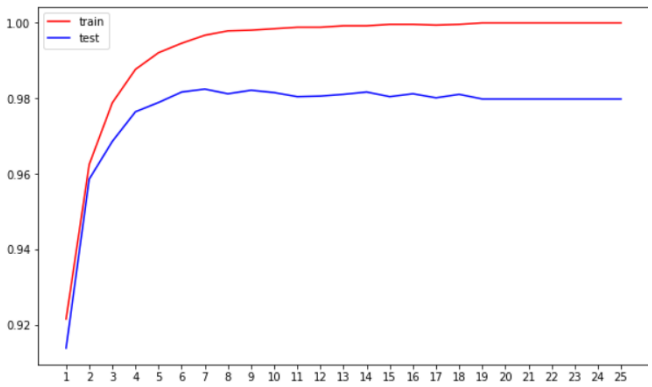

Fig.7 Relationship between the Classification Accuracy and the Depth of Splits in Both Train and Test Dataset

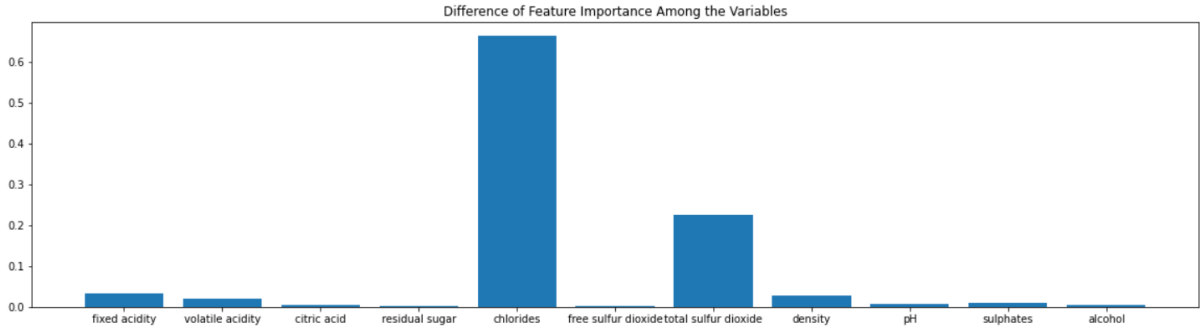

Fig.8 Feature Importance under the Single Decision-Tree Model without Pruning

Bagging-Tree Model

Although the pruned decision-tree model receives a high accuracy score, it does not optimise the value difference when solving some close-to-outlier numbers, that probably generates high variances among the features. To compress the predictors' flexibility into a small range, it is recommended to create some random-generated subsets from the training datasets for substituting the only one mother tree employed in previous method. Following that each subset dataset trains their own decision tree and finally get a series of distinctive models. Using the average of all predictions from these models, the new-generated aggregation is more robust than the normal weighting on classification (Nagpal, 2017). In this experiment, the accuracy rate of the test dataset under the fitted model flows severely from 56% to 61% as the number of estimated sub-tree increases, while the whole dataset's accuracy ratio only maintains at 9% (Fig.9), which might because of the inadequate sample size that highly correlated features have overfitted the algorithmic structure.

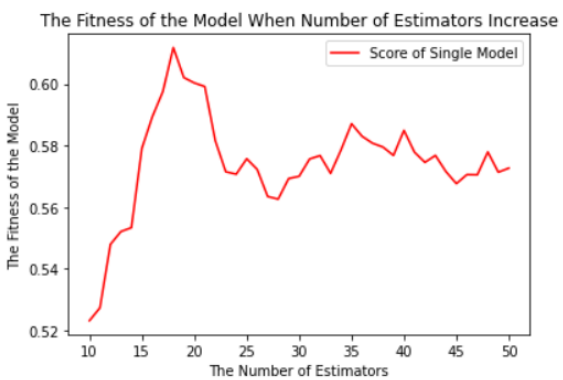

Fig.9 The Relationship between Accuracy Rate and Numbers of Estimators in Test Dataset

Random Forest Model

Although bagging-tree method declines the feature variance through resampling, those highly correlated predictors are still troubles for shrinking their variances down comparing with unrelated quantities. Therefore, the bootstrap concept should be also referred on the feature selection that only 1/3 of the original features will be randomly reselected from an empirical perspective, to reserve comparably independent candidates in each split dataset (James et al., 2013). In this case, four feature parameters will be stochastically drew out from eleven dimensions that not only significant predictor is valued but also the other predictors can highlight themselves inside the model. In the default condition of 100 sub-trees, 93.3% and 99.06% are received as the accuracy metrics of test set and the whole dataset individually, while chlorides and sulphur dioxide again become as the dominations of variables, as coincident to the results of decision-tree model (Fig.10).

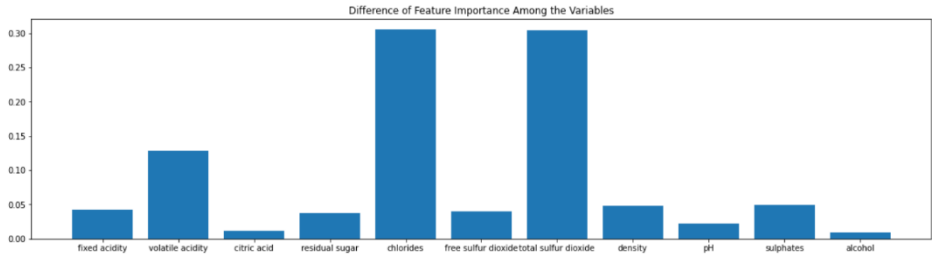

Fig.10 Feature Importance under the Random Forest

Artificial Neutral Network Model (ANN)

Artificial Neutral Network is a deep-learning-based classification algorithm simulated from human neuron system. A small amount of shallow neurons accept the original input data and then transfer them to a large number of hidden neurons for weight assignment. The purity and effectiveness of information will be distinctively labelled by these deep processors during a collaborative work, resulting in classifiable conclusions (Sharma, 2017). The secret behind the different weighting is 'ReLU' activation that automatically enlarges the significant choices by positive gradient while punishing the insignificant alternatives at zero weight. However, it is a challenging question when the number of hidden layers increase, the neural framework gains deeper explanation about dataset itself at the cost of overfitting hazards. Krishnan (2021) underlines the rules of thumb that will ensure the ANN model works properly: the number of layers should be appropriate by taking 2/3 of the number of input layers, which is 4 in this case. Then drop 25% of unrepresentable data after layer modelling to keep to the robustness. Moreover, how to determine the exact number of neural units in each layer also follows an experimental principle that only numbers built from the power of 2 are capable, which is 32 after comparing the accuracy performance through several alternative units (Fig.11). In all, the ANN model achieved 94.3% and 75.3% in test dataset and whole dataset accuracy respectively.

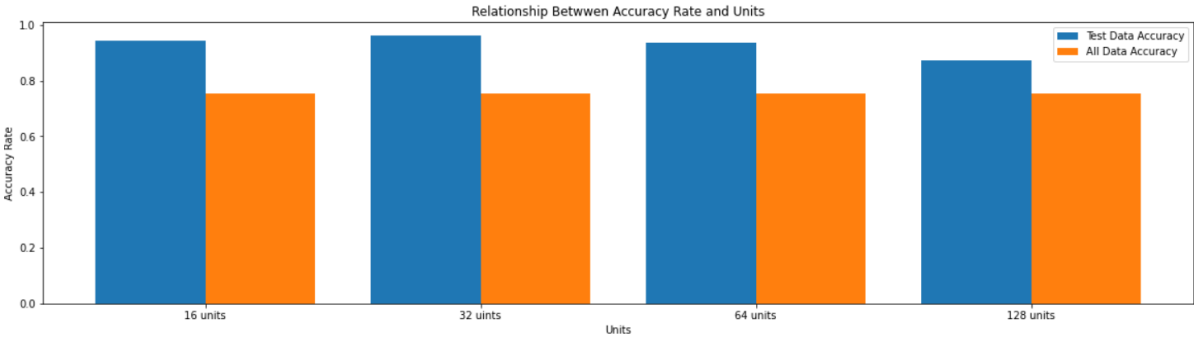

Fig.11 Relationship Between Accuracy Rate and Unit Numbers

Considering the different accuracy performance generated by four machine learning methods, those classified data from the random forest model are selected for the further classification of wine taste quality, and both pruned decision-tree method and random forest method are reserved since their predictive performance.

Further Discussion

Inherited the model structures, the predictive accuracy of the five-class wine quality received 76.5% and 78.0% in pruned decision tree and random forest separately, which means more three quarter of wine quality labels are properly identified, while the detailed information is formed into two confusion matrices (Fig.12-13).

Owing to the uneven distribution across the five-class quality labels, Class ‘Normal’ and ‘Slightly Good’ occupy around 98% of the entire labels, which triggers severe bias on the other light-weighted features that Class ‘Slightly Good and ‘Very Good’ receive almost zero predictions. However, the random forest method provides diverse predictions for each class, except for Class ‘Very Bad’ since no data are captured inside (Fig.14). To better comparing the performance of two classifiers, some measurements, like precision and recall could be employed for evaluating the frequency of true positive cases and relevant cases occurred in these samples. Selected the highest frequent label, ‘Normal’, is scored 14.4% in precision and 100% in recall through applying the decision model, while 82.1% in precision and 83.1% in recall after modelling on random forest model. Two comparably similar scores calculated from the rear method demonstrate there is a large possibility to reach true relevant classifications out from the other prediction choices, which also become the main reason of choosing random forest as the exclusive selection of the model.

Furthermore, it is interesting the feature importance indicator has found that alcohol degree is the most influential feature on wine taste quality, rather than beverage density and wine acidity, which might be useful for wine manufactories to review their production formulations (Fig.15).

CF_DT	Very Bad	Slightly Bad	Normal	Slightly Good	Very Good
Very Bad	0	0	0	0	0
Slightly Bad	53	985	261	1	0
Normal	0	0	45	0	0
Slightly Good	0	4	6	2	0
Very Good	0	0	0	0	0
CF_RF	**Very Bad**	**Slightly Bad**	**Normal**	**Slightly Good**	**Very Good**
Very Bad	0	12	105	54	0
Slightly Bad	0	0	0	0	0
Normal	0	41	809	122	1
Slightly Good	0	0	71	85	0
Very Good	0	0	0	0	0

Fig.12-13 The Confusion Matrices Generated by Pruned Decision Tree and Random Forest

Numbers	Very Bad	Slightly Bad	Normal	Slightly Good	Very Good
	0	246	9948	3816	20

Fig.14 Counts of Five-Class Labels of Wine Taste

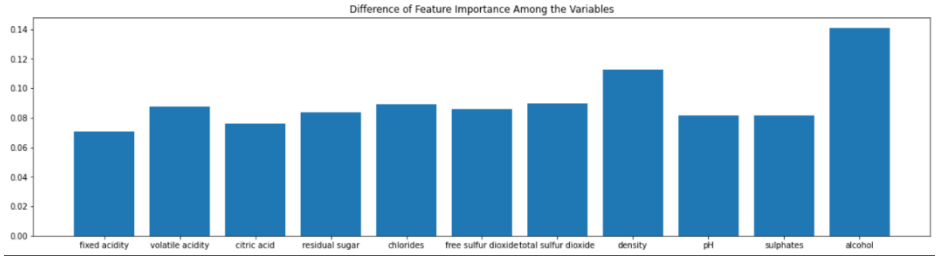

Fig.15 Feature Importance among Features

Limitation and Further Work

Since the wine quality classification is at the sub-level of the raw material classification, the 1% misidentified data generated by the random forest model will be discarded, which increases the data deviation of the lower-level model sampling. Cross class errors can be solved by applying more powerful upper-level classifiers, such as finding the optimal number of trees in the random forest algorithm through cyclic sequences to improve the model's accuracy, rather than selecting the default value as mentioned above. Furthermore, it might be a superior solution by applying Boosting method compared to the random forest framework because it builds trees according to the previous grown trees, rather than growing independently, which might be sufficient for some small-sized subsets of labels when their tree numbers are limited. In addition, Boosting is a self-learning method that generates data sequentially, so its accuracy may be higher than parallel-based random forest model.

References

Charters, S. & Pettigrew, S. (2007) 'The dimensions of wine quality', *Food Quality and Preference,* 18(7), pp. 997-1007.

Dong, Y. & Joanne, C.(2013) 'Principled missing data methods for researchers', *Springerplus,* Volume 2, p. 222.

Ebeler, S.(1999) Flavor Chemistry — Thirty Years of Progress, Kluwer Academic Publishers. In: *Linking flavour chemistry to sensory analysis of wine.* s.l.:s.n., pp. 409-422.

Fracassetti, D. et al. (2021) 'Light-struck taste in white wine: Reaction mechanisms, preventive strategies and future perspectives to preserve wine quality', *Trends in Food Science & Technology,* Volume 112, pp. 547-558.

James, G. et al. (2013) *An Introduction to Statistical Learning.* New York: Springer.

Krishnan, S. (2021) *How to determine the number of layers and neurons in the hidden layer?* Available at: https://medium.com/geekculture/introduction-to-neural-network-2f8b8221fbd3 (Accessed: 10 Feburary 2021)

Lee, L., Frederick, S. & Ariely, D, (2006) 'Try it, you'll like it: the influence of expectation, consumption, and revelation on preferences for beer', *Psychological Science,* 17(12), pp. 1054-1058.

Maddala, G. S. (1992) Outliers. In: *Introduction to Econometrics.* New York: MacMillan, p. 89.

Maguire, S. (2018) 'The taste for the particular: A logic of discernment in an age of omnivorousness', *Journal of Consumer Culture,* Volume 18, pp. 3-20.

Mithrakumar, M. (2019) *How to tune a Decision Tree?* Available at: https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680#:~:text=The%20theoretical%20maximum%20depth%20a,one%20big%20reason%20being%20overfitting. (Accessed: 9 February 2021)

Nagpal, A. (2017) *Decision Tree Ensembles- Bagging and Boosting.* Available at: https://towardsdatascience.com/decision-tree-ensembles-bagging-and-boosting-266a8ba60fd9 (Accessed: 9 Feburary 2021)

Schwarz, O. (2013) ;Bending Forward, One Step Backward: On the Sociology of Tasting Techniques', *Cultural Sociology,* 7(4), p. 415–430.

Sharma, S. (2017) *Artificial Neural Network (ANN) in Machine Learning.* Available at: https://www.datasciencecentral.com/artificial-neural-network-ann-in-machine-learning/ (Accessed: 10 Feburary 2021)

Sörqvist, P. et al. (2013) 'Who needs cream and sugar when there is Eco-Labeling? Taste and willingness to pay for 'Eco-Friendly' coffee', *PLoS One,* 8(12), p. e80719.

Stolz, H. and Schmid, O. (2008) '*Consumer attitudes and expectations of organic wine',* Modena, 16th IFOAM Organic World Congress.

Teil, G. (2021) 'Amateurs' Exploration of Wine: A Pragmatic Study of Taste', *Theory, Culture & Society,* 38(5), p. 137–157.

Yu, H.Y. et al. (2015) 'Characterization of Chinese rice wine taste attributes using liquid chromatographic analysis, sensory evaluation, and an electronic tongue', *Journal of Chromatography B,* Volume 997, pp. 129-135.