Highlights

???Grape Quality and its Link to Regional Differences in the Australian Winegrowing Industry

Author

- ???
- ???
- ????
- ????

???Grape Quality and its Link to Regional Differences in the Australian Winegrowing Industry

Author[1,1,1]

Abstract

1. Introduction

The Australian wine-growing industry is a rich and diverse landscape that is separated into multiple Geographical Indicator Regions. Each region describing unique reputations, qualities and varietals of wine produced there. While a great deal has been done regarding individual regional properties and traits, there has been little statistical insight into broader regional comparisons; due to a lack of cross-regional and in-depth data sources (Keith Jones, 2002; Knight et al., 2019). In this study we use Classification Trees to compare regional differences and how these differences relate to sustainable practices.

A vineyard's region predetermines several physical parameters, such as: climate, geology and soil; making location a widely considered key determinant of grape yield and quality (Abbal et al., 2016; Agosta et al., 2012; Fraga et al., 2017). Through the use of classification trees this study aims to highlight the key differences in sustainable practices at a regional level and how these practices relate to the different grades of grape quality.

Preprint submitted to Elsevier *02/08/2023*

2. Methods

2.1. Data

Data used in this analysis were obtained from Sustainable Winegrowing Australia. Australia's national wine industry sustainability program, which aims to facilitate grape-growers and winemakers in demonstrating and improving their sustainability (SWA, 2022). Data recorded by the SWA is entered manually by winegrowers using a web based interface tool. A total of 6091 observations were collected from 2012 to 2022. Each observation contained 23 variables reflecting a vineyards account for the given year (see Table 2.1).

The data originally contained only two multiclass variables: year and region. Variables that measured the same metric from different sources (such as water collected from rivers versus water from dams) were converted into multiclass variables representing the source. The total amount used from these variables was retained as a separate variable. Occurrences of multiple sources were defined as separate classes. As harvest does not run by calendar year, years are in financial years. Region represents one of the 65 Geographical Indicator Regions (GI Region) used to describe different unique localised traits of vineyards across Australia (Halliday, 2009; Oliver et al., 2013; SOAR et al., 2008). Each region is explicitly defined under the Wine Australia Corporation Act of 1980 (Attorney-General's Department, 2010). Profit was also used as a binary variable, depicting whether a vineyard was profitable or not.

Table 1: Summary of variables used in the analysis. The recorded column indicate values that were either greater than zero or that were not missing.

Variable	**Units**	**Recorded**	**Number of Classes**
Water Used	Mega Litres	5846	
Diesel	Litres	5585	
Biodiesel	Litres	25	
LPG	Litres	958	
Herbicide Spray	Times per year	2026	
Year	Class	6091	10
Disease	Class	6091	2
Region	Class	6091	58
Solar	Kilowatt Hours	622	
Irrigation Type	Class	6091	20
Petrol	Litres	4309	
Slashing	Times per year	2290	
Yield	Tonnes	5935	
Irrigation Energy	Class	6091	16
Area Harvested	Hectares	6091	
Electricity	Kilowatt Hours	1015	
Insecticide Spray	Times per year	1092	
Fertiliser	Kilograms of Nitrogen	795	
Fungicide Spray	Times per year	2260	
Cover Crop	Class	6091	32
Water Type	Class	6091	39
Profit	AUD	853	
Operating Costs	AUD	853	

2.2. XGBoosted Trees

XGBoosted (eXtreme Gradient Boosting) trees were created using the XGBoost library (Chen and Guestrin, 2016) in the Python Programming language (G. van Rossum, 1995). They were chosen for this analysis as they provide both a high predictive performance and ability to effectively capture complex relationships. An XGBoosted tree was created for each variable to show how they interacted. Each tree included all but the economic variables (profit and operating cost), which were only included once as predicted variables.

Following Chen and Guestrin (Chen and Guestrin, 2016), XGboosted trees predict a value y_i from the input x_i. The method of prediction is achieved through a tree ensemble model, using K additive functions to predict the output.

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_K(x_i), f_K \in \mathcal{F}, \tag{1}$$

where each function f_K is a classification or regression tree, such that all functions are in the set of all decision trees $\mathcal{F}$, defined by $f(x) = \omega_{q(x)}(q : \mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T)$. Where, f_K corresponds to an independent tree structure q of ω weights. Each tree has T leaves, which contain a continuous score, represented by ω_i for the i-th leaf. The final prediction is determined by the sum of the score of the corresponding leaves, given by ω. The set of functions used by the tree is determined by minimising the regularised objective function, given by:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i^{t-1} + f_t(x_i)) + \sum_k \Omega(f_K). \tag{2}$$

The difference between the prediction and actual variable is a convex loss function l. To optimise l, the difference is calculated for the i-th instance at the t-th iteration. The function f_t is selected according to which value minimises (2). The model complexity is penalised by the function Ω, this acts to smooth weights in an attempt to prevent over fitting.

As predictions are made using additive tree functions, XGboosted trees can be used for classification and regression. Due to the mixture of continuous, binary and multiclass variables in this analysis, both classification and regression trees were created. The difference between the trees created for this analysis was the objective function used. XGBoosted regression trees were created for continuous variables, using the root-mean-square as the objective function. Binary class variables utilised the logistic loss function as the objective. And, Multiclass variable used the soft max function. All objective functions are defined within the SKlearn library (Buitinck et al., 2013), linked via an API to the XGBoost library (Chen and Guestrin, 2016).

Chen and Guestrin (Chen and Guestrin, 2016) further illustrate, using Taylor expansions, that for a fixed structure $q(x)$ the optimal weight ω_j^* for a leaf j can be derived. Furthermore, they show the loss reduction after the split is given by the function:

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda}\right] - \gamma, \tag{3}$$

with the tree structure defined using left I_L and right I_R instance sets of nodes, with $I = I_L \cup I_R$. Instead of enumerating all possible tree structures, a greedy algorithm iteratively adds branches to the tree minimising $\mathcal{L}_{split}$ in (3). The frequency of a variable's occurrence within a tree is directly attributed to the minimisation of the objective function (or loss) through the minimisation of $\mathcal{L}_{split}$.

The frequency of a variable appearing as a node within the ensemble was used as a measure of importance. This measure was chosen as it connected a variable to the minimisation of its associated objective function, translating the value into a simple count metric. Creating XGBoosted trees for each variable allowed the use of importance to show how strongly variables were associated with each other. The importance of predictor variables to economic variables was illustrated through the use of Sankey diagrams constructed using the Holoviews python library (Rudiger et al., 2020). Other variable's interconnectedness was demonstrated through the use of a chord diagram also created using Holoviews.

Each variable utilised 80% of the data to train the XGBoost ensemble, with 20% reserved for testing and validation. Testing was done through the iterative minimisation of the respective objective function for the variables type. For continuous variables 20% was used as testing data, minimising the root-mean-square function. The final model was validated using repeated k-fold cross validation for 10 folds, repeated 10 times. For binary and multiclass variables data was split into 80% training, 10% testing and 10% validation data. Due to class disparity in multiclass variables (most prominently in region) data was stratified into each subset at the same ratio of class occurrence. Validation was summarised through confusion matrices and their associated accuracy

The use of the XGBoost library incorporates regularisation techniques built into the software to mitigate over-fitting and enhance model generalisation. The further use of cross validated grid search functions allowed for the selection of better performing hyperparameters when selecting the final model. The performance measure for model selection was root-mean-square error for continuous variables. The receiver operator characteristic's area under the curve was used for category variables (Hanley and McNeil, 1982). Multiclass variables utilised the one verse one approach to minimise sensitivity to class disparity (Ferri et al., 2009; Hand and Till, 2001).

2.3. Classification and Regression Trees

Classification and Regression Trees were created for region, year, profit and operating cost. These models describe the partitions that are useful in predicting these variables; giving insight into the trees that make up the ensembles created by XGBoost. These trees were created using the rparts and caret packages (Kuhn, 2008; Terry Therneau and Beth Atkinson, 2022) in the R statistical programming language (R Core Team, 2021).

Classification trees were validated using K-fold cross validation. Each model was validated using 10 folds, utilising a random selection of different samples ten separate times to validate each of the classification trees. A summary confusion matrix was then constructed to show the class bias and overall accuracy of each tree.

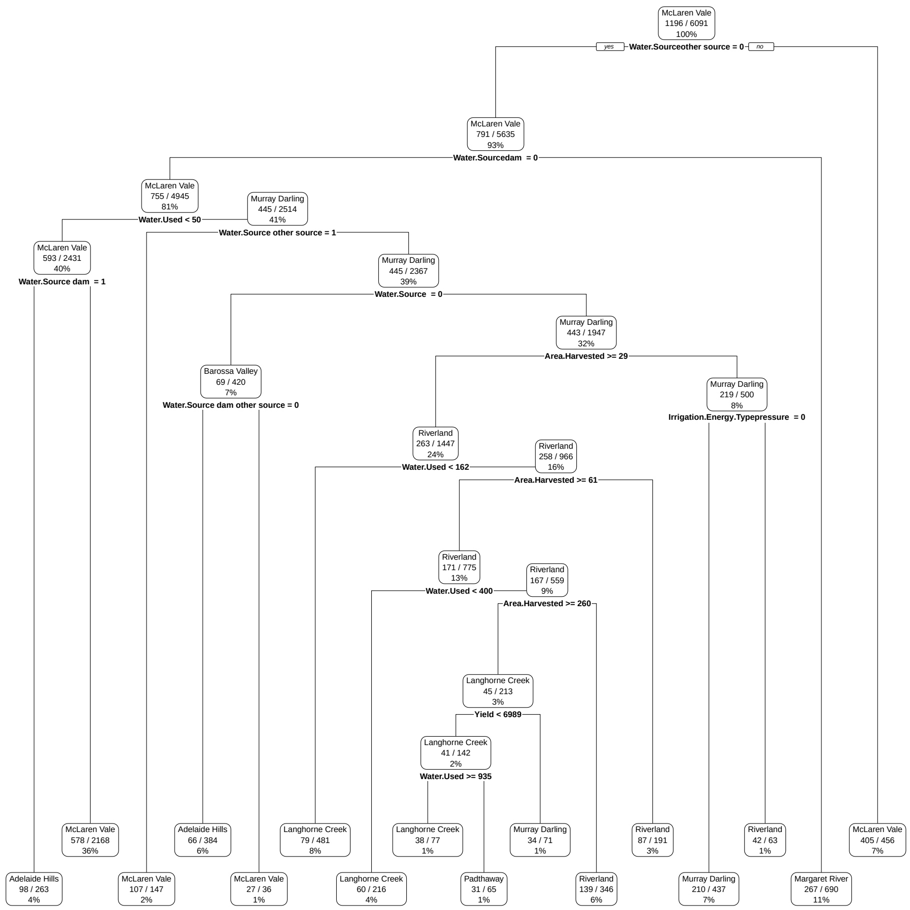

Figure 1: Decision tree predicting Region. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

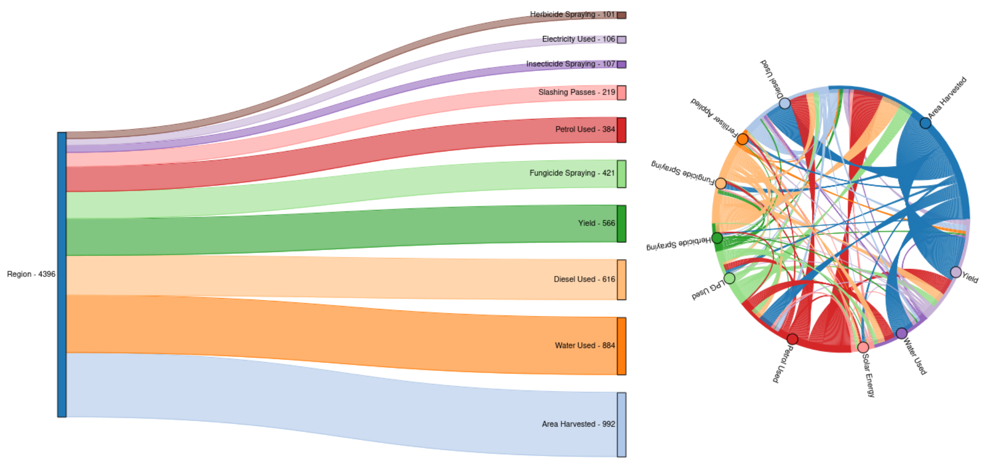

Figure 2: The left-hand side depicts the 10 most important variables in predicting Region using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

3. Results

3.1. Region

3.2. Year

3.3. Operating Costs

3.4. Profit

3.5. Validation

3.6. Model 1 GI Regions

The first Model was used to classify GI regions and resulted in an accuracy of 36.48% across 52 classes. The most prominent features used to classify regions were the types of water resources available (see Figure 1). Two regions, the Riverland and Coonawarra, were the most accurate classes being

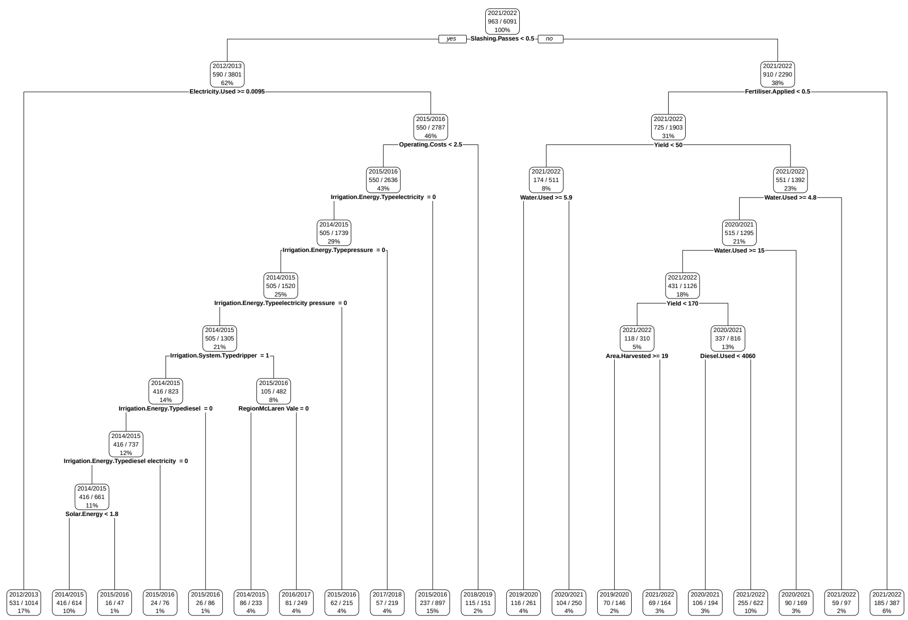

Figure 3: Decision tree predicting Year. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

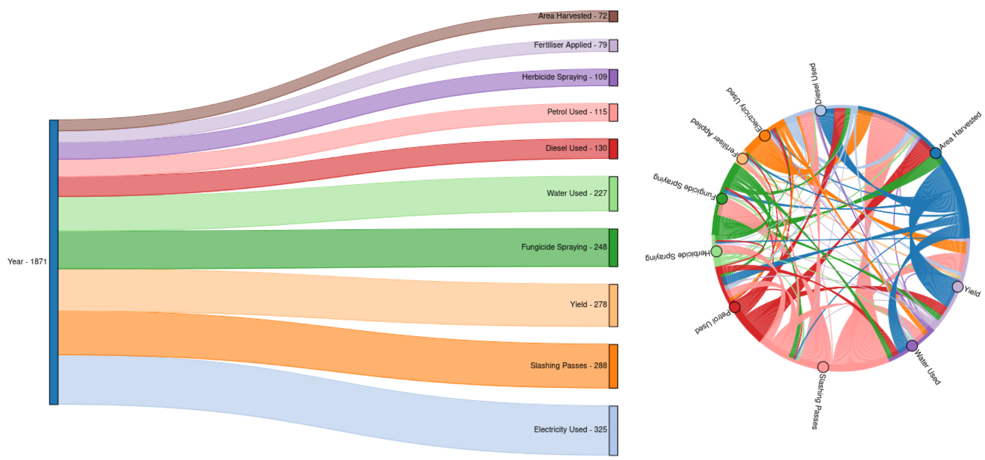

Figure 4: The left-hand side depicts the 10 most important variables in predicting Year using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

Table 2: Validation and training accuracies of each multiclass variable.

Variable	Validation	Training
cover crops	0.364086	0.396418
water type	0.742097	0.928905
profitable	0.705882	0.719737
irrigation type	0.841845	0.847554
giregion	0.505824	0.568242
irrigation energy	0.746293	0.836405
data year id	0.422003	0.518059

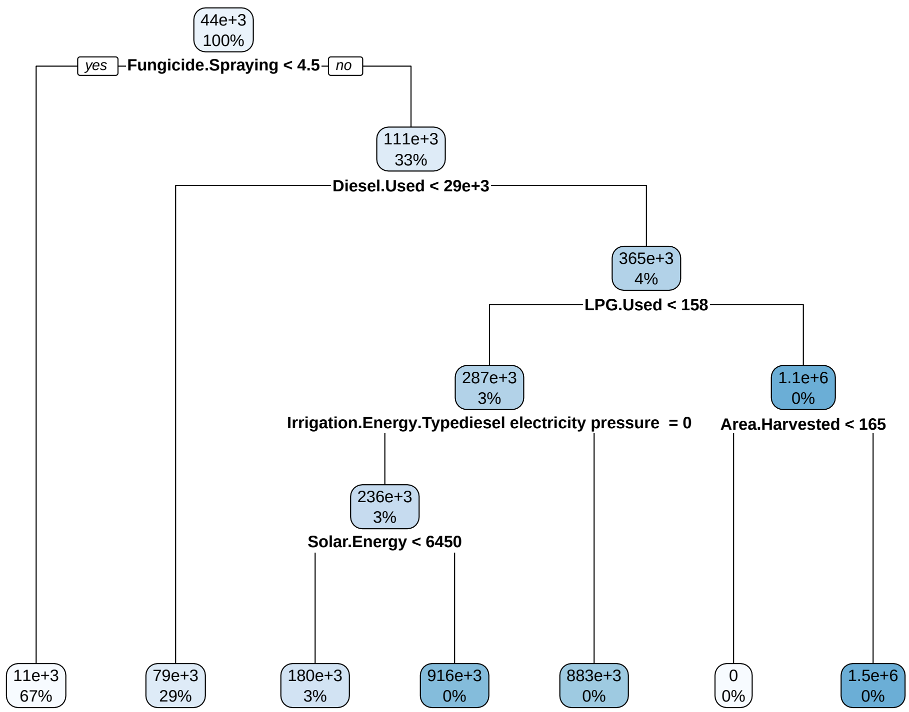

Figure 5: Decision tree predicting Operating Costs. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

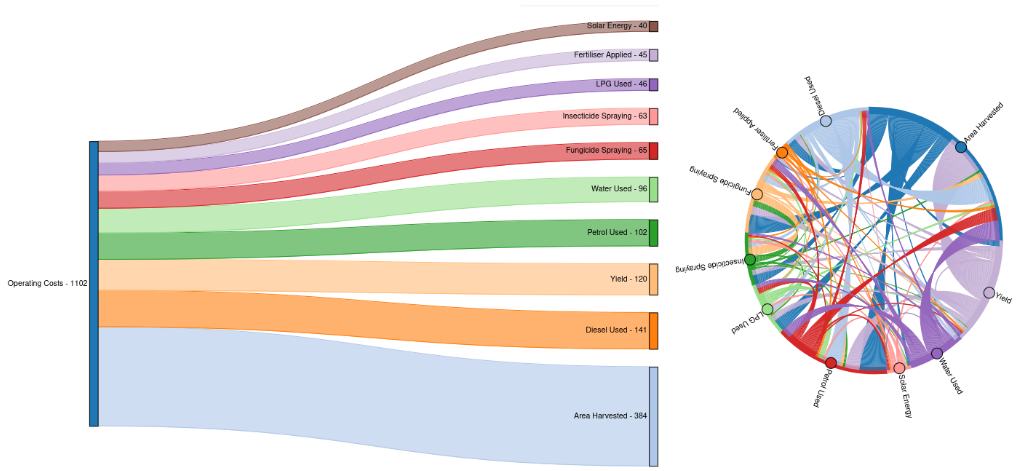

Figure 6: The left-hand side depicts the 10 most important variables in predicting Operating Costs using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

92.74% and 96.97% respectively. These regions differ greatly in practice and geophysical properties, with the Riverland being a dry warm inland region and Coonawarra being a cooler, wet coastal region. However, they are both similar in operational scales, with vineyards being relatively large compared with other regions. The differences in resources and practices between these regions are also significant, such as the Riverland utilising the river Murray as a water source. Many of the regions had significantly lower reporting rates, resulting much poorer classification performance. The regions with the most samples performed the best (see Table 1). Notably bordering regions were routinely grouped together and misclassified as the same region, for example the two closest regions to Coonawarra, Padthaway and Wrattonbulley, were misclassified as Coonawarra even though they had 147 and 137 samples

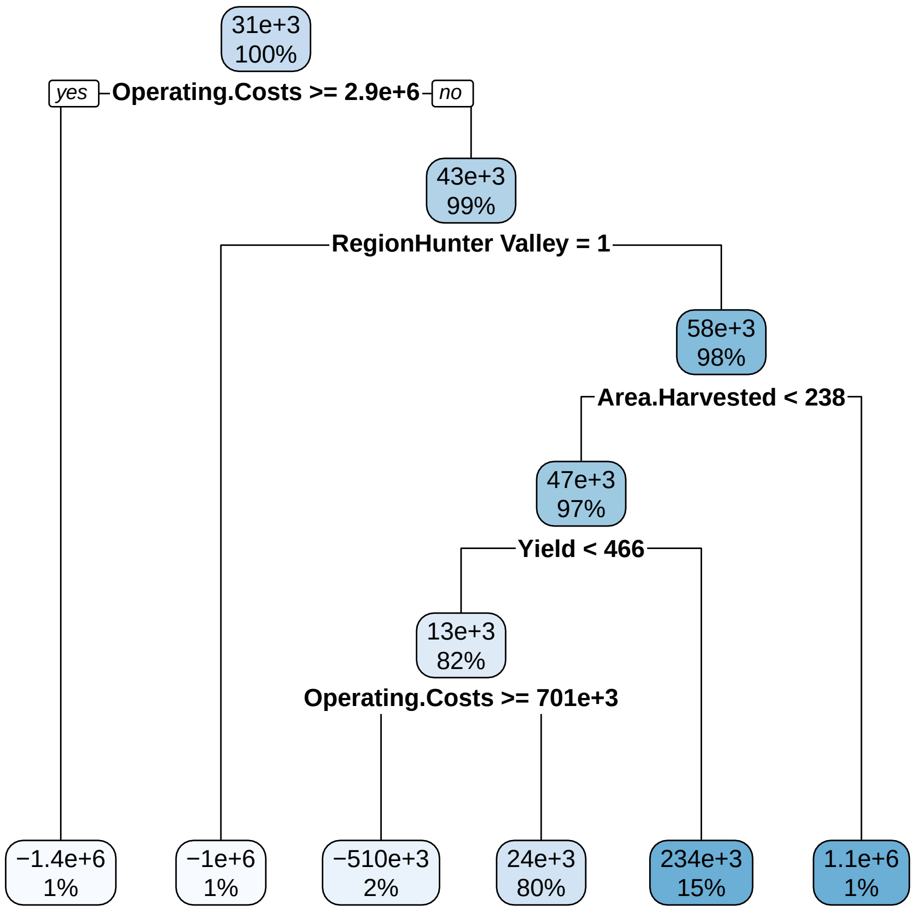

Figure 7: Decision tree predicting Profit. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

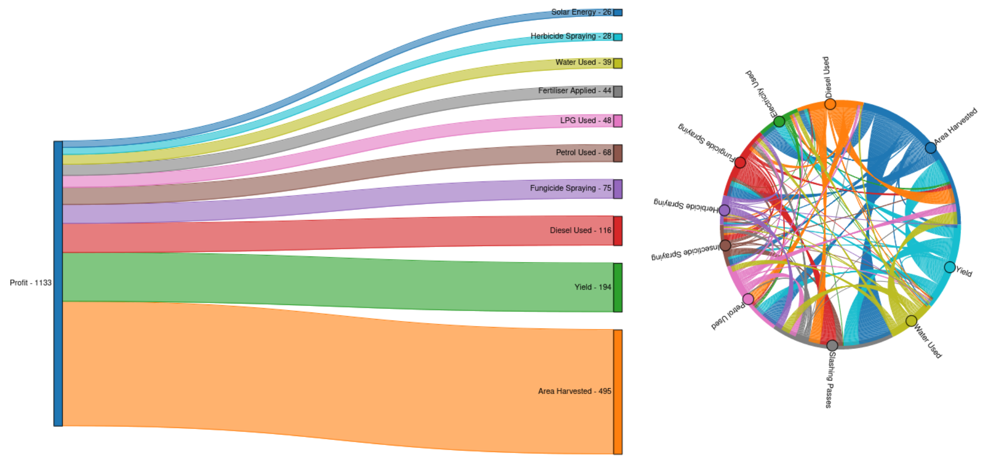

Figure 8: The left-hand side depicts the 10 most important variables in predicting Profit using XGBoosted trees as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

respectively. The same case was found for the Murray Darling, with 143 samples, it was misclassified as the Riverland. These misclassifications are likely due to the incredibly similar regional properties and close proximity these regions have with one another. Other misclassifications were most likely due to lower reporting rates with many regions being under represented.

3.7. Climate

Classifying the SWA climatic categorisation of the given regions had better performance than the GI Regions, with 41.66% being classified correctly. These categories were divided into 12 climatic classifications with 3 and 4 separate subsets for rainfall and temperature respectively. The decision tree behaved similarly and over classified climates with higher response rates. The results posed an interesting similarity with grape quality classifier, being influenced predominantly by water and area. The use of fungicide to separate regions that were 'Very dry' and 'Damp' can be considered as indicative of the different practices required due to climatic pressure; fungicides being more prominent in cooler regions with greater rainfall due to the higher risk of disease pressure (Reynolds, 2010). This could also potentially explain the use of contractor tractor use to discern differences in grape quality, where the lack of contractor use to prevent disease could have led to lowered quality of grapes.

3.7.1. Rainfall

The rainfall decision tree showed a greater use of fungicides sprays to discern between damp and very Dry as shown in Figure 4; with the accuracy improving to 62% but was unable to effectively discern between dry and very dry regions (see Table 3).

3.7.2. Temperature

The classification of GI Regions by their temperatures (see Figure 5) showed similarities to the other trees, with a heavy reliance on the types of water resources used as dominant predictors. The use of contractors was again used to differentiate between warm and cool regions, likely being due to disease pressure. The temperature classification tree was only a minor improvement over the regional classification tree, with an accuracy of 49.26% as shown in the confusion matrix (see Table 4).

3.8. Model 3 Grape Quality

The classification of grape quality through its grade had an accuracy of 55.72% across 5 separate grades. There was a notable issue with the classification of B grade grapes when compared to A and C (see Table 2). The classification tree itself shows similarities to that of classifying regions in Model 1, with the type of water resource used being a prominent determiner. Although not surprising the number of contractor tractor passes is new deciding factor due disease and pests reducing the potential quality of a crop. The prevalence of contractor use is greater in regions such as the Barossa Valley and the McLaren Vale, this could be due to the difference in operational scales, with larger sites being more likely to have ownership of their own equipment for weeding and spraying due to the cost benefit.

4. Discussion

The difference between grape quality is most notable between warm inland regions and coastal regions such as the Riverland and Coonawarra, respectively. Grape quality is only described by a singular variable within this study, however in reality it is driven by market demand and subject to complex forces such as international market pressure, fire, pests and disease (Wine Australia, 2019, 2020, 2021, 2022; Winemakers' Federation of Australia, 2015, 2016, 2017, 2018) The decision trees were able to offer some insights into the factors that influence grape quality and regional contrasts that contribute to different qualities. The most prominent being what readily available resources of each region were, particular the types of water available. Heavy water consumption is often linked to the mass production of grapes, where lower quality grapes are targeted in a quantity over quality strategy. These types of business decisions are unfortunately obfuscated by lack of in-depth data regarding vineyard business plans. Notably the literature shows that there are many complex decisions to be made on the ground depending on many compounding factors that influence both quality and yield (Abad et al., 2021; Cortez et al., 2009; Hall et al., 2011; I. Goodwin, et al., 2009; Kasimati et al., 2022; Oliver et al., 2013; Srivastava and Sadistap, 2018)

. There are also further differences when comparing winegrowers to other agricultural industries as they are vertically integrated within the wine industry, tying them to secondary and tertiary industries, such as wine production, packaging, transport and sales. This results in unique issues, where on-the-ground choices are influenced by other wine industry's decisions, such as the use of sustainable practices in vineyards to sell in overseas markets; notably these interactions are further complicated by some winegrowers being totally integrated into wine companies, while others are not (Knight et al., 2019). It is incredibly difficult to attribute external business decisions to produced grape quality but it is important to acknowledge that some growers are contracted to produce grapes of a particular grade; it is difficult to know whether another consumer may have graded the grape quality differently paying more or less for the same grapes given the opportunity to purchase them. It is difficult to untangle the contributing factors to the success of winegrowers and the quality of grapes produced without further specifics of choices made through out a season (Leilei He et al., 2022).

5. Conclusion

The type and availability of water resources were a major contributing factor when classifying grape quality and region. This was seen in the two most accurately classified regions, Coonawarra and the Riverland, with the

Riverland predominantly utilising river water. Furthermore, the study highlighted the influence of water use, fungicide application, and contractor use in differentiating grape quality, climate and region respectively. These models provide insight into the complex dynamics between regional characteristics, sustainable practices, and grape quality in the Australian winegrowing industry. It is important to acknowledge that grape quality is subject to external influences such as market demands and prior established business arrangements. Further in-depth data and understanding are necessary to fully grasp the nuances of decision-making and the interplay of factors impacting grape quality.

References

Abad, J., Hermoso de Mendoza, I., Marín, D., Orcaray, L., Santesteban, L.G., 2021. Cover crops in viticulture. A systematic review (1):
Implications on soil characteristics and biodiversity in vineyard. OENO One 55, 295–312. doi:10.20870/oeno-one.2021.55.1.3599.

Abbal, P., Sablayrolles, J.M., Matzner-Lober, É., Boursiquot, J.M., Baudrit, C., Carbonneau, A., 2016. Decision Support System for Vine Growers Based on a Bayesian Network. Journal of agricultural, biological, and environmental statistics 21, 131–151. doi:10.1007/s13253-015-0233-2.

Agosta, E., Canziani, P., Cavagnaro, M., 2012. Regional climate variability impacts on the annual grape yield in Mendoza, Argentina. Journal of Applied Meteorology and Climatology 51, 993–1009.

Attorney-General's Department, 2010. Wine Australia Corporation Act 1980.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G., 2013. API design for machine learning software: Experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, pp. 108–122.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA. pp. 785–794. doi:10.1145/2939672.2939785.

Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Using data mining for wine quality assessment, in: Discovery Science: 12th International Conference, DS 2009, Porto, Portugal, October 3-5, 2009 12, Springer. pp. 66–79.

Ferri, C., Hernández-Orallo, J., Modroiu, R., 2009. An experimental comparison of performance measures for classification. Pattern Recognition Letters 30, 27–38. doi:10.1016/j.patrec.2008.08.010.

Fraga, H., Costa, R., Santos, J.A., 2017. Multivariate clustering of viticultural terroirs in the Douro winemaking region. Ciência Téc. Vitiv. 32, 142–153.

G. van Rossum, 1995. Python tutorial, Technical Report CS-R9526. Centrum voor Wiskunde en Informatica (CWI),.

Hall, A., Lamb, D.W., Holzapfel, B.P., Louis, J.P., 2011. Within-season temporal variation in correlations between vineyard canopy and winegrape composition and yield. Precision Agriculture 12, 103–117.

Halliday, J.C.J.C., 2009. Australian Wine Encyclopedia. Hardie Grant Books, VIC.

Hand, D.J., Till, R.J., 2001. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45, 171–186. doi:10.1023/A:1010920819831.

Hanley, J.A., McNeil, B.J., 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143, 29–36.

I. Goodwin,, L. McClymont,, D. Lanyon, A. Zerihun, J. Hornbuckle, M. Gibberd, D. Mowat, D. Smith, M. Barnes, R. Correll, 2009. Managing soil and water to target quality and reduce environmental impact.

Kasimati, A., Espejo-García, B., Darra, N., Fountas, S., 2022. Predicting Grape Sugar Content under Quality Attributes Using Normalized Difference Vegetation Index Data and Automated Machine Learning. Sensors 22. doi:10.3390/s22093249.

Keith Jones, 2002. Australian Wine Industry Environment Strategy.

Knight, H., Megicks, P., Agarwal, S., Leenders, M., 2019. Firm resources and the development of environmental sustainability among small and mediumsized enterprises: Evidence from the Australian wine industry. Business Strategy and the Environment 28, 25–39. doi:10.1002/bse.2178.

Kuhn, M., 2008. Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Articles 28, 1–26. doi:10.18637/jss.v028.i05.

Leilei He, Wentai Fang, Guanao Zhao, Zhenchao Wu, Longsheng Fu, Rui Li, Yaqoob Majeed, Jaspreet Dhupia, 2022. Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods 195.

Oliver, D., Bramley, R., Riches, D., Porter, I., Edwards, J., 2013. Review: Soil physical and chemical properties as indicators of soil quality in Australian viticulture. Australian Journal of Grape and Wine Research 19, 129–139. doi:10.1111/ajgw.12016.

R Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Reynolds, A.G., 2010. Managing Wine Quality : Viticulture and Wine Quality. Woodhead Publishing Series in Food Science, Technology and Nutrition ; v.1., Elsevier Science, Cambridge.

Rudiger, P., Stevens, J.L., Bednar, J.A., Nijholt, B., Andrew, B, C., Randelhoff, A., Mease, J., Tenner, V., maxalbert, Kaiser, M., ea42gh, Samuels, J., stonebig, LB, F., Tolmie, A., Stephan, D., Lowe, S., Bampton, J., henriqueribeiro, Lustig, I., Signell, J., Bois, J., Talirz, L., Barth, L., Liquet, M., Rachum, R., Langer, Y., arabidopsis, kbowen, 2020. Holoviz/holoviews: Version 1.13.3. Zenodo. doi:10.5281/zenodo.3904606.

SOAR, C., SADRAS, V., PETRIE, P., 2008. Climate drivers of red wine quality in four contrasting Australian wine regions. Australian journal of grape and wine research 14, 78–90. doi:10.1111/j.1755-0238.2008.00011.x.

Srivastava, S., Sadistap, S., 2018. Non-destructive sensing methods for quality assessment of on-tree fruits: A review. Journal of Food Measurement and Characterization 12, 497–526.

SWA, S.W.A., 2022. Sustainable Wingrowing Australia. https://sustainablewinegrowing.com.au/case-studies/.

Terry Therneau, Beth Atkinson, 2022. Rpart: Recursive Partitioning and Regression Trees.

Wine Australia, 2019. National Vintage Report 2019 .

Wine Australia, 2020. National Vintage Report 2020 .

Wine Australia, 2021. National Vintage Report 2021 .

Wine Australia, 2022. National Vintage Report 2022 .

Winemakers' Federation of Australia, 2015. National Vintage Report 2015 .

Winemakers' Federation of Australia, 2016. National Vintage Report 2016 .

Winemakers' Federation of Australia, 2017. National Vintage Report 2017 .

Winemakers' Federation of Australia, 2018. National Vintage Report 2018 .