

1 An analysis of underlying relationships between factors
2 related to operating costs and revenue in Australian
3 vineyards.

4 Author^{1,1,1}

5 **Abstract**

6 The Australian wine industry is a major contributor to Australia's agri-
7 cultural sector and economy. As global market demands change and new
8 pressures on the industry present themselves, a more sustainable approach is
9 needed. Through a nationwide data set, collected over ten years, we link key
10 variables in determining vineyard operational costs and revenue through the
11 use of XGBoost. We use a measure of relative importance to show the inter-
12 related nature of these variables and the comparative influence they have on
13 one another. We present these connections through the use of Sankey and
14 Chord diagrams to show the important predictors of revenue and operating
15 costs and their strong interrelatedness. Furthermore, we connect these vari-
16 ables to different wine regions, highlighting the complex influence of location
17 on the use of different resources. The study provides valuable insights into
18 the multifaceted dynamics governing operational costs and revenue, illustrat-
19 ing how factors such as water and fuel use impacts operational costs and how
20 different seasonal events affect these operations.

21 1. Introduction

22 Historically strong demands for Australian wine have helped to create a
23 thriving industry. However, recent pressures brought on by a loss of tourism
24 and labour due to the COVID-19 pandemic, the global freight crisis, war
25 in Europe, tariffs and rising inflation have negatively affected the industry’s
26 outlook (Wine Australia, 2021; Australia, 2021a). The 2021-2022 financial
27 year alone saw a decline of 19% in exports solely due to tariffs (Wine Aus-
28 tralia, 2022). A greater understanding of the different underlying conditions
29 leading to improved performance in agricultural productivity and sustainabil-
30 ity at scale is key to making data-informed decisions to increase a nation’s
31 agricultural sustainability (OECD, 2019). Specifically within the Australian
32 wine and vine industry, there is a need to further understand the driving
33 relationships between resource use and economic output, which can help to
34 determine more cost effective and efficient methods and develop benchmarks
35 with local growers (Luke Mancini, 2020).

36 An unprecedented amount of data regarding the Australian winegrowing
37 industry has been collected through the Sustainable Winegrowing Australia
38 program, offering the potential for new insights into the driving economic
39 forces of the Australian wine industry. A major part of the potential for
40 insight within this dataset comes from the incorporation of operating costs
41 and grape revenue from grape sales within the data. In this paper, we use
42 data to study economic outcomes and their statistical relationships to vine-
43 yards’ utilisation of the resources. We further compare the relationships
44 between different resources to address the extensive collinearity found within
45 the data (Chen and Guestrin, 2016). We adopt a popular, relatively new

46 machine learning method, XGBoost, for this analysis because it is able to
47 overcome multicollinearity as well as highlight the level of importance that
48 predictor variables have on response variables.

49 **2. Methods**

50 *2.1. Data*

51 Data used in this analysis were obtained from Sustainable Winegrowing
52 Australia (SWA), Australia’s national wine industry sustainability program.
53 SWA aims to support grape growers and winemakers in demonstrating and
54 improving their sustainability (SWA, 2022). Data recorded by SWA are
55 entered manually by winegrowers using a web based interface tool. A total
56 of 6049 observations were collected from 2012/2013 to 2021/2022 financial
57 years, with each observation comprising 23 variables reflecting a vineyard’s
58 state for the given year (see Table 1).

59 The data originally contained only two multiclass variables: year and
60 region. For this case study, related binary variables, such as the use of river
61 water and the use of dam water, were combined to create multiclass variables
62 such as water source/type. Further details about these variables, their classes
63 and their frequency is available in the Appendix.

64 The variable Region represented one of the 65 Geographical Indicator
65 Regions (GI Region) used to describe unique localised traits of vineyards
66 across Australia (Halliday, 2009; Oliver et al., 2013; SOAR et al., 2008).
67 Each region is explicitly defined under the Wine Australia Corporation Act
68 of 1980 (Attorney-General’s Department, 2010).

Table 1: Summary of variables used in the analysis. The recorded column indicate the number of values that were either greater than zero or that were not missing (see Appendix for more information).

Variable	Units	Number of Classes	No. Records
Water Used	Mega Litres		5846
Diesel	Litres		5585
Biodiesel	Litres		25
LPG	Litres		958
Herbicide Spray	No. Times per year		2026
Year	Class	10	6049
Disease	Class	2	6049
Region	Class	58	6049
Solar	Kilowatt Hours		622
Irrigation Type	Class	20	6049
Petrol	Litres		4309
Slashing	No. Times per year		2290
Yield	Tonnes		5935
Irrigation Energy	Class	16	6049
Area Harvested	Hectares		6049
Electricity	Kilowatt Hours		1014
Insecticide Spray	No. Times per year		1092
Fertiliser	KGs of Nitrogen		795
Fungicide Spray	Times per year		2260
Cover Crop	Class	32	6049
Water Type/Source	Class	39	6049
Grape Revenue	AUD		853
Operating Costs	AUD		853

69 2.2. *XGBoost*

70 XGBoost (eXtreme Gradient Boosting), described in more detail below
71 (and further in the Appendix), were created using the XGBoost library (Chen
72 and Guestrin, 2016) in the Python Programming language (G. van Rossum,
73 1995). XGBoost is a type of machine learning method that constructs and
74 ensemble of decision trees to predict or estimate an output variable (the re-
75 sponse) based on a number of input variables. The ensemble, can be used
76 to classify classes or predict a continuous response, depending on the nature
77 of the output variable. They were chosen for this analysis as the data con-
78 tained a mixture of class and continuous variables. Moreover, XGBoost is
79 unaffected by multicollinearity, and offer high predictive performance for a
80 wide variety of purposes, and are capable of identifying and ranking variables
81 and interactions in order of relative importance (Chen and Guestrin, 2016).

82 Four sets of analyses were conducted. In the first set, two XGBoost mod-
83 els were developed, with operational cost and grape revenue as the response
84 variables. The analysis of operational cost and revenue included all variables
85 in Table 1. The second set of analyses focused on identifying relationships
86 between the input variables themselves, creating XGBoost models for each
87 other variable so that every variable would have a measure of its relative
88 importance to every other variable (see Section 2.3). Together these mod-
89 els were used to measure the interrelationships of the ten most important
90 variables in determining operational cost and grape revenue using variable
91 importance. These measures of relative importance were used to illustrate
92 the highly interrelated nature of variables within vineyards. The interaction
93 between variables was depicted through the use of Sankey and Chord dia-

grams; with variable importance measures being used to show the strength of connection between the respective predictor variables and the response (see section 2.3).

The third analysis was an XGBoost tree with Region as the response variable. The difference for this model was that relative variable importance for each variable would be measured for the overall importance in determining region, as opposed to a variable's connection to each region specifically. The fourth analysis focussed on profit (the difference between revenue and operational costs) and year, however these results were not included due to low average loss values and model stability (see Appendix).

XGBoost is an ensemble method that combines multiple decision trees together to create a more accurate predictive model. The gradient boosting aspect of the ensemble is the use of a loss function to create new decision trees that add to the ensemble, improving its predictive power. The loss function is optimised iteratively to improve upon prior trees. The loss function can be any convex function, allowing gradient descent to traverse the loss space until no substantive improvements can be made via traversal. Because the loss function is only required to be convex, both classifiers and regressors can be used. Regularisation methods can also be incorporated to help prevent over fitting. This makes XGBoost incredibly versatile and accurate, whilst still being interpretable compared to other machine learning methods.

2.3. Variable Importance

XGBoost creates a large number of decision trees in the ensemble, it is hard to directly interpret the model and the derived intricate relationship between the variables. Variable importance can be measured in multiple

ways, in this paper we used the frequency of a variable appearing as a node within the ensemble as a measure of its importance. This measure can be interpreted as how often a variable was the optimal choice in reducing the loss function of the ensemble. Multiclass variables are given an importance score for each individual class; for example, in the first set of analyses each specific region will have its own importance score, as will Year, Irrigation Type, etc (see Table 1).

The Sankey and Chord diagrams were constructed using the Holoviews python library (Rudiger et al., 2020). Both Chord and Sankey diagrams illustrated variable importance through the size of the bands between two variables. The number at the end of a connection in a Sankey diagram indicates a variable’s importance (the number of times it appeared within the ensemble). Sankey and Chord diagrams are presented together; with Sankey diagrams showing the connection of a variable to its ten most important predictor variables and chord diagrams showing the interconnectedness of the ten most prominent variables within its associated Sankey diagram. Chord diagrams formed circles, with variables being connected through their relative importance.

2.4. Validation

The predictive accuracy of each tree was assessed through a validation process. For each model, a sample of 80% of the data was used for training the model and the remaining 20% was used for testing and validation. Categorical data were stratified to conserve the same proportion of class occurrences between the training, testing and validation data. The models were validated using 10 repetitions of the sampling process (10-fold cross

validation). R^2 scores were used to determine the best regression models during validation. For analyses with continuous responses R^2 was used instead of RMSE to allow the comparison of models with different units to each other when considering how well each model extrapolated to further data. For binary and multiclass variables, validation was summarised through the accuracy, the proportion of true negatives and positives.

As with most machine learning methods, a key component of the XGBoost model setup was the tuning of hyperparameters. The XGBoost library incorporates regularisation techniques built into the software to mitigate over-fitting and enhance model generalisation. This allowed us to utilise cross validated grid search functions when selecting for better performing hyperparameters. 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).

3. Results

The below sections present each of the analyses conducted within this study. This includes the three analyses for Revenue, Operational Cost and Region, with the fourth and final analysis on profit and yield presented in the appendix.

3.1. Revenue

The predictive performance of the XGBoost model for revenue performed similarly to operating cost, for achieving an R^2 of 0.77 (with a standard

168 deviation of 0.15).

169 The most important predictors of revenue were fuel use (petrol 307 and
 170 diesel 144), yield (285), size (216) and water use (199). The values attached
 171 to each variable indicate the relative importance of the variable (number
 172 of times selected in the tree ensemble, see Section 3.1). Overall regions
 173 contributed to 234 nodes in the ensemble making them collectively the third
 174 most important variable. The chord diagram (see Figure 3.1) illustrates that
 175 vineyard area is also of high relative importance to other variables, especially
 176 slashing. The overall importance of Area to other variables is evident by its
 177 larger circumference within the chord diagram.

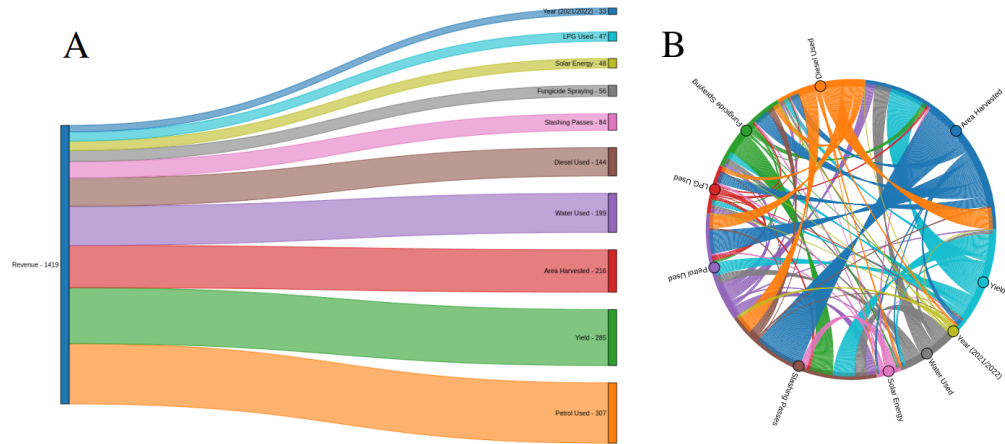


Figure 1: The left-hand side depicts the 10 most important variables in predicting revenue using XGBoost as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

178 3.2. Operating Costs

179 Compared to revenue, the predictive performance of XGBoost model for
180 operating cost was slightly better, with an R^2 of 0.80 (with a standard deviation of 0.10). Similar to revenue, the most important predictors of operating
181 cost were fuel, water, area and yield (see figure 2). A surprising difference was
182 the change in relative importance of activities involving tractor passes where
183 the use of fungicide was more important for operational costs, compared to
184 revenue, where slashing was more important (see Figure 3). The variables
185 that feed into these decisions are also very different with diesel having the
186 highest relative importance to slashing, and area having the greatest relative
187 importance to the need for fungicide.
188

189 Again, Region played a determining factor overall, contributing to 334
190 nodes within the ensemble making it the most important variable when considering all regions together. It was surprising that electricity, slashing and
191 spraying passes were not more prominent in operating costs due to the intrinsic nature as an agricultural expense.
192
193

194 3.3. Region

195 Region was a highly informative variable based on measures of importance
196 for both operating cost and revenue. As noted above, Region was the third
197 most important variable for determining revenue. The Barossa Valley region
198 and Tasmania were the two most important regions in relation to revenue;
199 these two regions are considered to be some of the highest revenue per hectare
200 regions in Australia (Wine Australia, 2022). These two regions are also
201 relative opposites in winegrowing climates with the Barossa having a warm

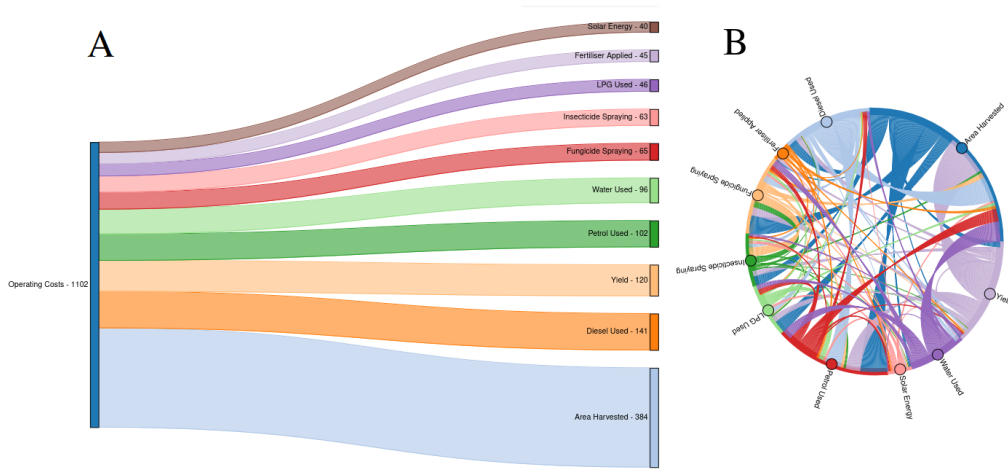


Figure 2: The left-hand side, A, depicts the 10 most relative important variables in predicting Operating Costs using XGBoost as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable’s importance. The right-hand side, B, depicts the importance of the 10 variables in Sankey diagram relative to one another.

202 and dry climate focussing on Shiraz grapes and Tasmania having a cool wet
 203 climate that favours Pinot/Chardonnay (Wine Australia, 2022).

204 As also noted above, Region was also a key determinant of operating
 205 costs. Tasmania had the highest relative importance, followed by the Ade-
 206 laide Hills. In contrast, the regions of the highest relative importance were
 207 warmer and drier, such as the Barossa. The higher relative importance of
 208 fungicide spraying is the likely due to fungal pressure being greater in cooler
 209 wetter regions variables than in drier regions.

210 The XGBoost ensemble for Region achieved an accuracy of 56.82% (and
 211 50.58% validation accuracy). The difference in accuracy compared to the
 212 other models is in part due to the large number of classes (58 regions). The

213 ensemble had an emphasis on area, water, fuel and yield as determining
 214 factors (see Figure (3)).

215 A number of regions had lower reporting rates, resulting in much poorer
 216 classification performance. The regions with the most samples performed
 217 the best likely due to the disparity in sample sizes. Bordering regions were
 218 routinely grouped together and misclassified as the same region. When scru-
 219 tinising each class explicitly, the two areas that effected the most from this
 220 were the Limestone Coast (cool coastal areas in South Australia) and the
 221 warmer inland regions along the Murray Darling.

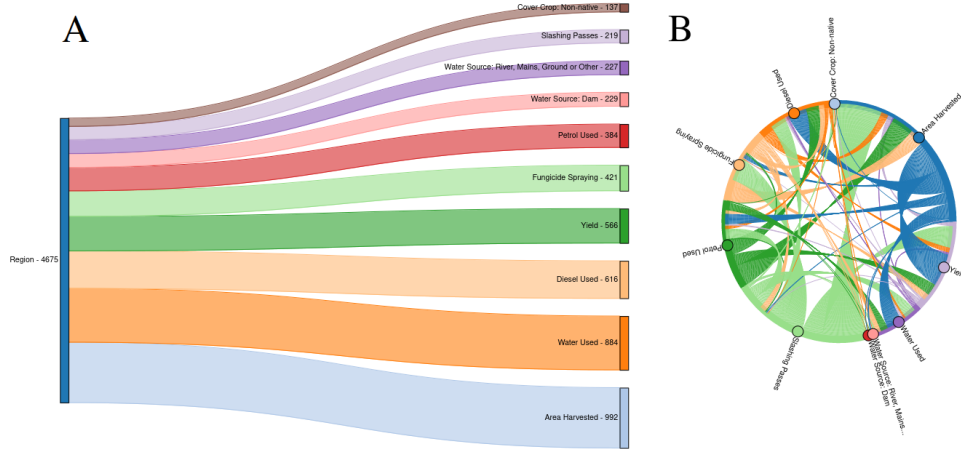


Figure 3: The left-hand side, A, depicts the 10 most relative important variables in predicting Region using XGBoost as a measure of node occurrence, using a Sankey diagram. The number at the end of each band in the diagram is that variable’s importance. The right-hand side, B, depicts the importance of the 10 variables in Sankey diagram relative to one another.

222 4. Discussion

223 This study explored the relationships between vineyard resource use, op-
224 erations and geographical properties to revenue and operating costs. The
225 analysis was based on a large national study of 6049 samples collected over
226 ten years. Three main findings were identified. First, the most important
227 predictors of revenue and operating costs were fuel, yield and area. Secondly,
228 area and fuel were highly interrelated to other variables (see Figure 2 and
229 Figure 3.1A). Finally, the relative importance of predictor variables for Re-
230 gion, differed from Revenue and Operating costs, with Water Use being more
231 prominent than Yield. Region was also more prominent than illustrated in
232 the Sankey diagrams due to the relative importance for operating cost and
233 revenue being calculated for individual regions and not all regions together.
234 In its entirety Region was the third most important predictor of revenue
235 and the most important predictor for operating costs, relative to the other
236 variables consideration in the analyses.

237 Several physical parameters such as climate, geography and soil are pre-
238 determined by a vineyard’s location, making it a widely considered key de-
239 terminant of grape yield and quality (Abbal et al., 2016; Agosta et al., 2012;
240 Fraga et al., 2017). The association between yield and region is demonstrated
241 by yield appearing as the fourth most important variable when determining
242 region (see Figure 3).

243 Warmer regions are known to be beneficial in hastening the ripening
244 process of winegrapes (Webb et al., 2011). Warmer regions are also associ-
245 ated with lower quality grapes, caused largely due to this hastened ripening
246 (Botting et al., 1996). It is likely that the combination of larger vineyards

247 with higher water use is a determining factor in classifying regions which
248 favour larger production of grapes; reflected through region using water use
249 so prominently in the XGBoost ensemble. The link to water resources in
250 defining regions is also an important consideration, as vineyards can leverage
251 higher irrigation rates if water resources are available. A further considera-
252 tion in the link between revenue and region is that grape prices are set at a
253 regional level by buyers (Wine Australia, 2022). It is also important to con-
254 sider that some regions carry particular fame regarding the quality of their
255 produce such as Tasmania, the Hunter Valley and Barossa Valley (Halliday,
256 2009). This classification can be contrasted with other warmer regions of
257 higher rainfall that use the warmer climate to concentrate their grapes, in-
258 creasing the flavour profile (Goodwin I, Jerie P, 1992; MG McCarthy et al.,
259 1986).

260 In part, yield is sometimes restricted simply through access to water
261 resources. Regions are likely to have varying access to different water sources,
262 such as those along the River Murray being able to utilise river water for
263 crops, unlike most coastal regions which may be drawing from surface or
264 underground water sources. Similarly, the connection between region and
265 fuel use is likely an indicator of the level of infrastructure within the region
266 because vineyards in regions without pressurised water will need to use more
267 fuel or electricity to pressurise their irrigation systems.

268 Operational costs showed similar importance across fuel, water and trac-
269 tor use. The dominating factor of area likely played a large part in deter-
270 mining how costly a tractor pass would be, or in defining the ratio of water
271 applied to the amount of vines. The relative importance was high for area

272 but much lower in general across the other variables, which could indicate the
273 need to be specific when attempting to determine the cause of a operational
274 cost. Although these analyses attempted to capture the complexity between
275 how variables interacted when determining operational costs (see Figure 2),
276 in reality these relationships are likely even more complicated. An example
277 of how interrelated operational costs can be, is the optimisation of tractor
278 passes to achieve multiple goals in a pass, being shown to reduce energy use
279 in vineyards, decreasing running costs, as well as reducing soil compaction
280 (Capello et al., 2019).

281 When determining revenue, similar variables were used to operational
282 cost; with region also being of high variable importance relative to other
283 variables (when considering all regions together in importance). It is difficult
284 to extrapolate the specific influence of location on a vineyard’s outcomes due
285 to the broad and varying definition of a region. Utilising the Geographical
286 Indicator regions defined by Wine Australia (Australia, 2021b) is a limitation
287 in one way, as it is too broad to fully capture a vineyards location and how
288 that influences variables at a more granular level. However, as buyers set
289 prices at regional levels, it is still important to consider this factor.

290 Decisions made on the ground have far-reaching effects and are difficult
291 to completely capture. A larger number of tractor passes used as a preven-
292 tative measure for occurrences such as disease may incur higher operational
293 costs but could be critical in preventing long term losses. Although the
294 models demonstrated a good predictive fit (via large R^2 values), the ability
295 to predict operational costs is limited by the variables incorporated in the
296 analysis. Other factors such as erosion and soil health are also influenced by

297 tractor use and would contribute to these operational costs but are difficult
298 to measure and were not available as part of the data (Capello et al., 2019,
299 2020). Reductions in fuel, water and tractor use are obvious methods to
300 reduce operational costs but not necessarily achievable decisions. Without
301 fully capturing more granular activities for example the specific reasons for
302 fuel use, it is difficult to determine what decisions specifically influence the
303 operational costs.

304 The reasoning for any particular decision can be widely varying. More
305 sophisticated models, specifically those that utilise expert opinion, may also
306 help to capture and address the decision-making process. An example is the
307 optimisation of fungicide sprays using Bayesian models that forecast disease
308 risk (Lu et al., 2020).

309 Separately, revenue and operating cost did have a greater predictability
310 than their counterpart profit (see Appendix). The disparity in accuracy be-
311 tween profit and other economic outcomes is reflective of the complexity in
312 trying to address challenges such as climate change, disease and changing
313 market demands (Wine Australia, 2020, 2021, 2022). The difference between
314 turning a profit or loss is dependent on predictable factors unforecasted fac-
315 tors, farming practice and farmers’ decisions. The difference between vine-
316 yards that make profit and those that do not could be a multitude of factors
317 including differences in farming practices not captured within this study.

318 5. Conclusion

319 This study has provided valuable insights into the multifaceted dynamics
320 governing operational costs and revenue in vineyards. The impact of dif-

ferent regions highlighted the complex interrelatedness of variables within a vineyard. We relate how factors such as water and fuel intersect to impact operational costs and how different seasonal events affect these operations; as well as the significance of context-specific decision-making. While this investigation utilised a broad regional classification, the potential benefits of adopting a more nuanced approach and incorporating expert knowledge have been highlighted. Further work could pursue causal models and the creation of decision support systems. It is difficult to untangle the predictive and correlative nature of a variable compared to the causal reasons. By delving deeper into the complex interplay of variables, further advancements can be made in optimising vineyard management strategies for lowering operational costs, increasing revenue and enhancing sustainability.

References

- 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.

343 Australia, W., 2021a. Australian Wine: Production, Sales and Inventory
344 2019–20.

345 Australia, W., 2021b. Wine Australia-Open Data.

346 Botting, D., Dry, P., Iland, P., 1996. Canopy architecture-implications for
347 Shiraz grown in a hot, arid climate .

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

354 Capello, G., Biddoccu, M., Cavallo, E., 2020. Permanent cover for soil and
355 water conservation in mechanized vineyards: A study case in Piedmont,
356 NW Italy 15.

357 Capello, G., Biddoccu, M., Ferraris, S., Cavallo, E., 2019. Effects of Tractor
358 Passes on Hydrological and Soil Erosion Processes in Tilled and Grassed
359 Vineyards. *Water* 11. doi:10.3390/w11102118.

360 Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System,
361 in: *Proceedings of the 22nd ACM SIGKDD International Conference on*
362 *Knowledge Discovery and Data Mining*, ACM, New York, NY, USA. pp.
363 785–794. doi:10.1145/2939672.2939785.

364 Ferri, C., Hernández-Orallo, J., Modroi, R., 2009. An experimental com-
 365 parison of performance measures for classification. *Pattern Recognition*
 366 *Letters* 30, 27–38. doi:10.1016/j.patrec.2008.08.010.

367 Fraga, H., Costa, R., Santos, J.A., 2017. Multivariate clustering of viticul-
 368 tural terroirs in the Douro winemaking region. *Ciência Téc. Vitiv.* 32,
 369 142–153.

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

372 Goodwin I, Jerie P, 1992. Regulated deficit irrigation: Concept to prac-
 373 tice. *Advances in vineyard irrigation. Australian and New Zealand Wine*
 374 *Industry Journal* 7.

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

377 Hand, D.J., Till, R.J., 2001. A Simple Generalisation of the Area Under the
 378 ROC Curve for Multiple Class Classification Problems. *Machine Learning*
 379 45, 171–186. doi:10.1023/A:1010920819831.

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

382 Lu, W., Newlands, N.K., Carisse, O., Atkinson, D.E., Cannon, A.J., 2020.
 383 Disease Risk Forecasting with Bayesian Learning Networks: Application
 384 to Grape Powdery Mildew (*Erysiphe necator*) in Vineyards. *Agronomy*
 385 (Basel) 10, 622. doi:10.3390/agronomy10050622.

386 Luke Mancini, 2020. Understanding the Australian Wine Industry: A growers
 387 guide to the background and participants of the wine grape industry.

388 MG McCarthy, RM Cirami, DG Furkaliev, 1986. The effect of crop load and
 389 vegetative growth control on wine quality. .

390 OECD, 2019. Innovation, Productivity and Sustainability in Food and Agri-
 391 culture.

392 Oliver, D., Bramley, R., Riches, D., Porter, I., Edwards, J., 2013. Review:
 393 Soil physical and chemical properties as indicators of soil quality in Aus-
 394 tralian viticulture. Australian Journal of Grape and Wine Research 19,
 395 129–139. doi:10.1111/ajgw.12016.

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

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

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

407 Webb, L.B., Whetton, P.H., Barlow, E.W.R., 2011. Observed trends in

Table A.2: Summary statistics of continuous variables used in XGBoost models.

	count	mean	std	min	0.25	0.5	0.75	max
Vineyard Solar	622	22916.89	104808	1	1170.75	5500	14866.25	2300000
Biodiesel	25	6635.932	11768.832104	1	200	500	10000	37216
Fungicide Spray	2260	7.724801	3.279794	1	6	7	9	68
LPG	958	327.831399	861.538804	1	40	95.835	240	11950
Petrol	4309	825.276809	1556.621119	1	135	306.66	903	38568
Insecticide Spray	1092	1.707189	1.316042	0	1	1	2	12
Water Used	5846	7301838	558206600	0.0007	13.2655	43	146.875	42680000000
Fertiliser	795	91149.89	483913.4	1	560	4759.5	45148.5	11358000
Diesel	5585	11677.070183	24380.588742	0.1267	1240	3850	12500	591000
Yield	5935	772.902449	2175.113895	0.03	68	192.3	601.8795	72305
Herbicide Spray	2026	2.646199	2.598899	0	2	2	3	103
Slashing	2290	3.311485	1.826788	1	2	3	4	26
Electricity	1014	58223.07	177626.3	0.019	2160	9637	36498.25	3000000
Area Harvested	6049	66.52604	133.4525	2.220446E-16	10.13	24.5	66.8	2436.15
Grape Revenue	875	377972	606286.8	1	76000	172964	386747	5700000
Operating Costs	853	314187.1	511522.6	1	57315	140000	327408	4482828

408 winegrape maturity in Australia. Global change biology 17, 2707–2719.
409 doi:10.1111/j.1365-2486.2011.02434.x.

410 Wine Australia, 2020. National Vintage Report 2020 .

411 Wine Australia, 2021. National Vintage Report 2021 .

412 Wine Australia, 2022. National Vintage Report 2022 .

413 **Appendix A. Continuous variables**

414 Table A.2 below shows the ranges of each of the continuous variables:

415 **Appendix B. Categorical Variables**

416 The tables below describe each possible class a multiclass variable could
417 have taken and the frequency that it occurred.

418 *Appendix B.1. Water Source Types*

419 Table B.3 below shows the different class types for water sources used by
420 vineyards and their frequency of occurrences.

Table B.3: Frequency and class types of water types used
by vineyards.

Water types	frequency
river water	1578
groundwater	1433
surface water dam	617
recycled water from other source	386
groundwater and surface water dam	256
not listed	235
mains water	170
river water and groundwater	147
groundwater and recycled water from	145
other source	
other water	101
river water and surface water dam	92

Continued on next page

Table B.3 – continued from previous page

Water types	frequency
groundwater and water applied for frost control	90
groundwater and mains water	76
river water and groundwater and surface water dam	70
recycled water from other source and mains water	63
groundwater and recycled water from other source and mains water	60
river water and mains water	57
surface water dam and mains water	56
groundwater and other water	33
river water and groundwater and mains water	30
groundwater and surface water dam and recycled water from other source	27
river water and water applied for frost control	27
groundwater and surface water dam and mains water	22
surface water dam and recycled water from other source	21
Continued on next page	

Table B.3 – continued from previous page

Water types	frequency
river water and recycled water from other source	19
river water and other water	19
river water and surface water dam and mains water	18
river water and groundwater and sur- face water dam and mains water	18
mains water and other water	16
groundwater and surface water dam and water applied for frost control	12
surface water dam and other water	12
groundwater and recycled water from other source and other water	11
groundwater and surface water dam and recycled water from other source and mains water	8
recycled water from other source and mains water and other water	8
river water and recycled water from other source and mains water	8
river water and surface water dam and recycled water from other source	8
Continued on next page	

Table B.3 – continued from previous page

Water types	frequency
surface water dam and mains water and other water	7
recycled water from other source and other water	7
river water and groundwater and recy- cled water from other source	6
groundwater and mains water and other water	5
groundwater and surface water dam and other water	5
groundwater and surface water dam and mains water and other water	5
river water and groundwater and re- cycled water from other source and mains water	5
river water and groundwater and wa- ter applied for frost control	5
river water and surface water dam and water applied for frost control	4
surface water dam and water applied for frost control	4

Continued on next page

Table B.3 – continued from previous page

Water types	frequency
river water and groundwater and sur- face water dam and recycled water from other source and mains water and other water	4
river water and groundwater and recy- cled water from other source and other water	3
groundwater and surface water dam and recycled water from other source and water applied for frost control	3
river water and groundwater and sur- face water dam and recycled water from other source	3
river water and recycled water from other source and other water	3
surface water dam and recycled water from other source and mains water	2
river water and recycled water from other source and mains water and wa- ter applied for frost control	2

Continued on next page

Table B.3 – continued from previous page

Water types	frequency
groundwater and surface water dam	2
and recycled water from other source	
and mains water and other water	
river water and groundwater and	2
mains water and other water	
river water and groundwater and sur-	2
face water dam and other water	
river water and surface water dam and	2
other water	
river water and mains water and water	2
applied for frost control	
river water and groundwater and sur-	2
face water dam and recycled water	
from other source and mains water	
river water and mains water and other	2
water	
river water and surface water dam and	2
mains water and other water	
river water and groundwater and	1
mains water and water applied for	
frost control	

Continued on next page

Table B.3 – continued from previous page

Water types	frequency
surface water dam and other water and water applied for frost control	1
water applied for frost control	1
groundwater and other water and wa- ter applied for frost control	1
other water and water applied for frost control	1
groundwater and surface water dam and recycled water from other source and other water and water applied for frost control	1
mains water and water applied for frost control	1
groundwater and surface water dam and recycled water from other source and other water	1
groundwater and mains water and wa- ter applied for frost control	1
river water and groundwater and sur- face water dam and mains water and other water	1

Continued on next page

Table B.3 – continued from previous page

Water types	frequency
river water and surface water dam and	1
recycled water from other source and	
mains water	

422 *Appendix B.2. Cover Crop Types*

423 Table B.4 below shows the different cover crop types used together and
424 their frequency.

Table B.4: Frequency and class types of cover crop types
used by vineyards.

Cover crop types	frequency
Cover crop types	frequency
permanent cover crop volunteer sward	1822
permanent cover crop non native	936
permanent cover crop native	490
annual cover crop	479
groundwater and surface water dam	406
annual cover crop and permanent cover crop volunteer sward	309
bare soil	225
permanent cover crop non native and permanent cover crop volunteer sward	214
annual cover crop and permanent cover crop non native	169
bare soil and permanent cover crop volunteer sward	129
Continued on next page	

Table B.4 – continued from previous page

Cover crop types	frequency
bare soil and permanent cover crop non native	115
annual cover crop and permanent cover crop non native and permanent cover crop volunteer sward	101
bare soil and annual cover crop	93
permanent cover crop native and per- manent cover crop volunteer sward	80
bare soil and permanent cover crop na- tive	78
annual cover crop and permanent cover crop native	78
permanent cover crop native and per- manent cover crop non native	68
permanent cover crop native and per- manent cover crop non native and per- manent cover crop volunteer sward	44
annual cover crop and permanent cover crop native and permanent cover crop non native and permanent cover crop volunteer sward	44

Continued on next page

Table B.4 – continued from previous page

Cover crop types	frequency
bare soil and annual cover crop and permanent cover crop volunteer sward	33
bare soil and permanent cover crop non native and permanent cover crop volunteer sward	26
annual cover crop and permanent cover crop native and permanent cover crop volunteer sward	17
bare soil and annual cover crop and permanent cover crop native	15
annual cover crop and permanent cover crop native and permanent cover crop non native	15
bare soil and annual cover crop and permanent cover crop non native	13
bare soil and annual cover crop and permanent cover crop native and per- manent cover crop non native and per- manent cover crop volunteer sward	12
bare soil and annual cover crop and permanent cover crop non native and permanent cover crop volunteer sward	11
Continued on next page	

Table B.4 – continued from previous page

Cover crop types	frequency
bare soil and annual cover crop and permanent cover crop native and per- manent cover crop non native	8
bare soil and permanent cover crop na- tive and permanent cover crop non na- tive	7
bare soil and permanent cover crop na- tive and permanent cover crop volun- teer sward	6
bare soil and permanent cover crop na- tive and permanent cover crop non na- tive and permanent cover crop volun- teer sward	4
bare soil and annual cover crop and permanent cover crop native and per- manent cover crop volunteer sward and	2

426 *Appendix B.3. Irrigation Types*

427 Below in Table B.5 are the frequency and different irrigation types.

Table B.5: Frequency and class types of irrigation types used by vineyards.

Irrigation types	frequency
Irrigation type	frequency
dripper	4800
dripper and non irrigated	342
Not listed	319
dripper and overhead sprinkler	201
dripper and undervine sprinkler	91
non irrigated	65
undervine sprinkler	53
dripper and flood	53
overhead sprinkler	46
dripper and overhead sprinkler and undervine sprinkler	28
overhead sprinkler and undervine sprinkler	12
dripper and non irrigated and overhead sprinkler	11
flood and undervine sprinkler	10
Continued on next page	

Table B.5 – continued from previous page

Irrigation types	frequency
dripper and flood and undervine sprinkler	7
dripper and flood and non irrigated and overhead sprinkler and undervine sprinkler	3
dripper and flood and overhead sprinkler	3
non irrigated and undervine sprinkler	2
dripper and flood and non irrigated	1
dripper and non irrigated and overhead sprinkler and undervine sprinkler	1
flood and	1

429 *Appendix B.4. Irrigation Energy Type*

430 Below, Table B.6 shows the different types of energy used to power vine-
 431 yards and their frequency.

Table B.6: Frequency and class types of irrigation energy types used by vineyards.

Irrigation Energy types	frequency
Irrigation energy type	frequency
electricity	2162
not listed	2053
pressure	586
electricity and pressure	396
diesel	254
diesel and electricity	227
electricity and solar	96
diesel and electricity and pressure	90
diesel and pressure	74
solar	50
electricity and pressure and solar	23
diesel and electricity and solar	14
diesel and electricity and pressure and solar	10
pressure and solar	9
Continued on next page	

Table B.6 – continued from previous page

Irrigation Energy types	frequency
diesel and solar	4
diesel and pressure and solar and	1

433 *Appendix B.5. Year*

434 Below in Table B.7 is the list of years and the number of sample collected
 435 in each.

Table B.7: Frequency and class types of year

Year	frequency
Year	frequency
2021/2022	954
2020/2021	860
2019/2020	599
2012/2013	590
2013/2014	549
2015/2016	548
2014/2015	505
2017/2018	493
2016/2017	485
2018/2019	466

436

437 *Appendix B.6. Region*

438 Below in Table B.8 are the number of collected samples for each region.

Table B.8: Frequency and class types of regions.

Regions	frequency
giregion	frequency
McLaren Vale	1195
Barossa Valley	584
Murray Darling	521
Riverland	472
Adelaide Hills	454
Langhorne Creek	347
Margaret River	344
Coonawarra	284
Padthaway	202
Wrattonbully	195
Clare Valley	149
Yarra Valley	122
Eden Valley	92
Tasmania	89
Swan Hill	83
Grampians	73
Orange	72

Continued on next page

Table B.8 – continued from previous page

Regions	frequency
Hunter Valley	70
Bendigo	53
Great Southern	51
Rutherglen	41
Robe	36
Tumbarumba	35
Mornington Peninsula	32
King Valley	32
Southern Fleurieu	30
Heathcote	29
Adelaide Plains	25
Currency Creek	24
	23
Henty	22
Canberra District	21
Southern Flinders Ranges	20
Upper Goulburn	20
Mudgee	20
Mount Benson	20
Other	19
Riverina	18
Alpine Valleys	15
Continued on next page	

Table B.8 – continued from previous page

Regions	frequency
Barossa Zone	14
Pemberton	12
Mount Gambier	11
Blackwood Valley	10
Kangaroo Island	10
Big Rivers Zone Other	9
Geographe	7
Cowra	6
Gundagai	5
Strathbogie Ranges	5
Glenrowan	4
Geelong	4
Swan District	4
Goulburn Valley	3
Beechworth	3
Southern Highlands	3
Macedon Ranges	2
Pyrenees	2
Sunbury	1

440 Appendix C. XGBoost

441 Following Chen and Guestrin (Chen and Guestrin, 2016), XGBoost pre-
 442 dicted a value y_i from the input x_i . The method of prediction is achieved
 443 through a tree ensemble model, using K additive functions to predict the
 444 output. Each of f_k functions is a classification or regression tree, such that
 445 all functions are in the set of all decision trees, given by \mathcal{F} , is defined by
 446 $f(x) = \omega_{q(x)}(q : \mathbb{R}^m \rightarrow T, \omega \in \mathbb{R}^T)$. Where each function corresponds to an
 447 independent tree structure q of ω weights. Each tree has T leaves, which
 448 contain a continuous score, represented by ω_i for the i -th leaf. The final
 449 prediction is determined by the sum of the score of the corresponding leaves,
 450 given by:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^K f_k(x_i), f_k \in \mathcal{F}, \quad (\text{C.1})$$

451 The set of functions, \mathcal{F} , used by the tree is determined by minimising a
 452 regularised objective function, \mathcal{L} given by:

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

453 , where

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (\text{C.3})$$

454 As predictions are made using additive tree functions, XGboost can be
 455 used for classification or regression. The difference between a prediction,
 456 $\phi(x_i)$, and actual variable, $f_k(x_i)$, is a differentiable convex loss function l .
 457 These properties of l allow the function to be versatile in which objective
 458 we choose to optimise for, which is also important in being able to process

both continuous and categorical variables. To optimise l , the difference is calculated for the i -th instance at the t -th iteration.

Appendix C.1. Loss functions

The functions included as parameters in equation C.2 mean that traditional optimisation methods for Euclidean space cannot be used. Chen and Guestrin (Chen and Guestrin, 2016) illustrate, using Taylor expansions, that for a fixed structure $q(x)$ the optimal weight ω_j^* for a leaf j can be derived. Importantly a loss function can be used to fit a model iteratively to data. For this analysis several loss functions were used, as variables took the form of continuous, binary and multi-class data. The loss function for making a split within the tree structure is given by:

$$\mathcal{L}_{split} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma. \quad (\text{C.4})$$

The tree structure being 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 (C.4). The frequency of a variable's occurrence within a tree is directly attributed to the minimisation of the loss function through the minimisation of \mathcal{L}_{split} .

The loss functions used for this analysis were the root-mean-square function for continuous variables, the logistic loss function for binary class variables, and the soft max function for Multiclass variables. All objective functions are defined within the SKlearn library (Buitinck et al., 2013), which was utilised via an API to the XGBoost library (Chen and Guestrin, 2016).

481 *Appendix C.2. Year*

482 The classification tree and XGBoost performed similarly for classifying
483 year with 35.20% (6.28% standard deviation) and 51.81% (42.20% validation
484 accuracy) respectively. Electricity and the type of irrigation were highly
485 influential within the classification tree. Similarly, electricity was the most
486 frequently occurring node in the XGBoost ensemble. Other variables such
487 as slashing passes, and fungicide and herbicide spraying were more prevalent
488 than in the classification tree. Weed and disease outbreaks are likely an
489 influential factor when classifying different years, making the decisions to
490 spray and slash unique factors that differ year to year. Climatic differences
491 between years are likely tied to the influence of yield and water use.

492 Over half of the interrelated importance of the predictor variables is domi-
493 nated by area harvested, yield and slashing passes. Although all the predictor
494 variables are highly connected, their relative importance is not as prominent
495 as the three major variables. It is of particular note of the relative importance
496 of slashing passes to area, fuel and yield; as these are not directly related ac-
497 tivities. The connection between the number of slashing and spraying passes
498 is that those who do a set number of spraying or slashing passes tended to
499 do that many passes for all slashing and spraying activities.

500 *Appendix C.3. Profit*

501 Predictions of profit performed poorly compared to operating cost and
502 revenue with an average R^2 of 0.2535 and standard deviation of 0.3126. With
503 the large standard deviation being indicative of how unstable the models
504 created were.

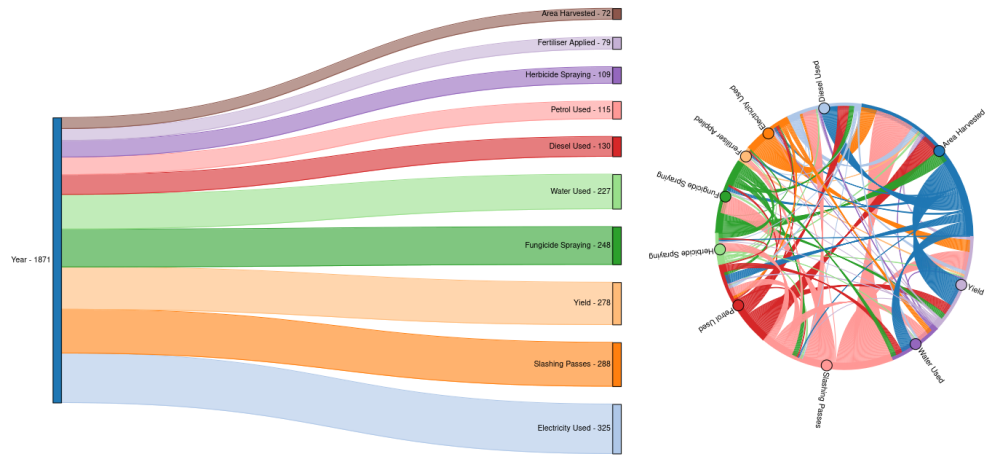


Figure C.5: The left-hand side depicts the 10 most relative important variables in predicting Year using XGBoost as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.

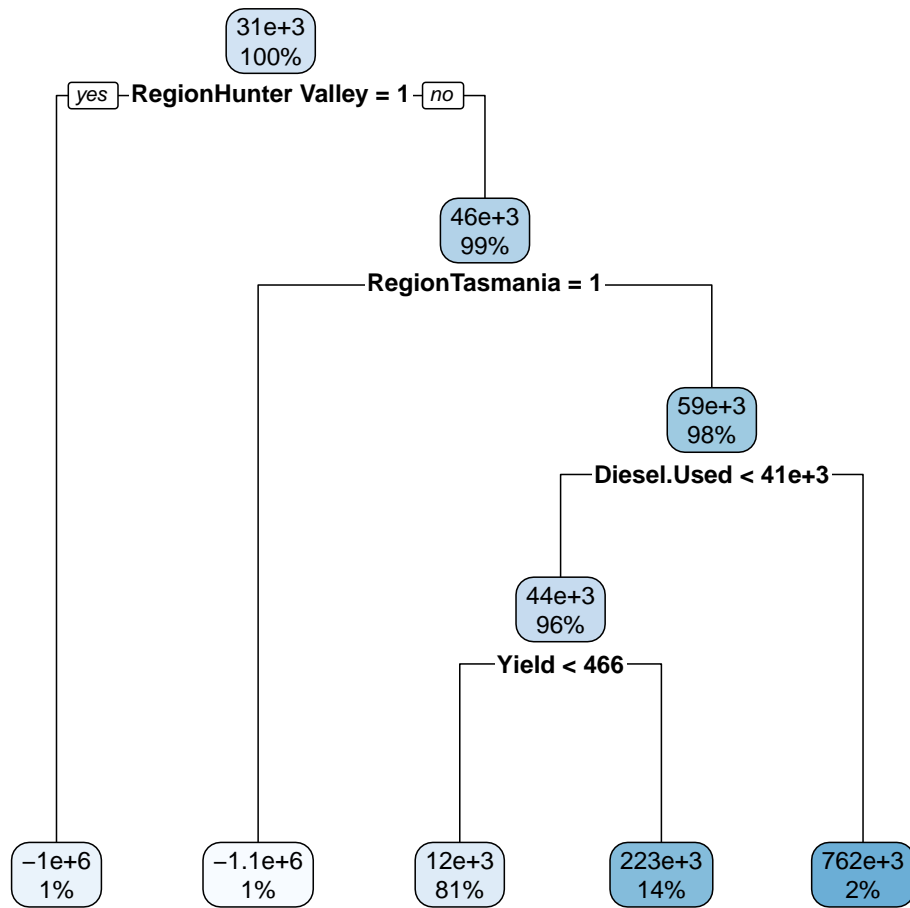


Figure C.6: Decision tree predicting revenue. Each node indicates the class predicted, and the proportion of elements agreeing with nodes partitioning, with the left direction indicating a yes to the nodes rule.

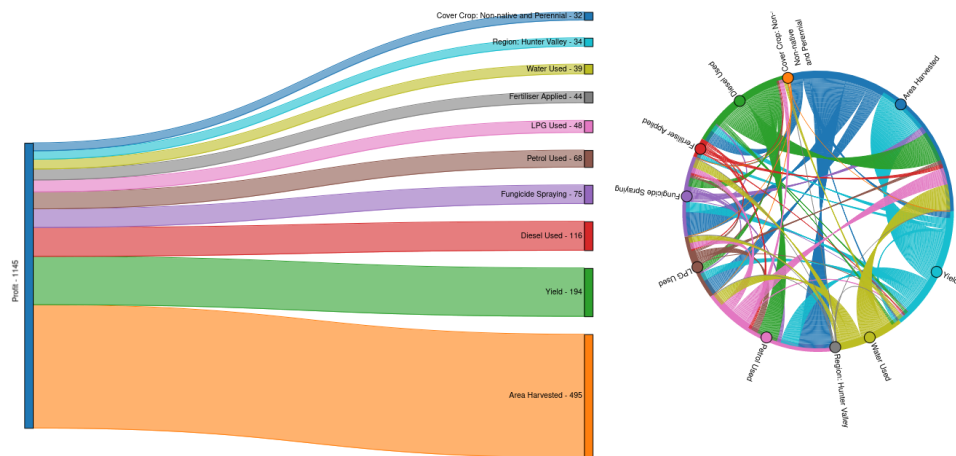


Figure C.7: The left-hand side depicts the 10 most relative important variables in predicting revenue using XGBoost as a measure of node occurrence, using a Sankey diagram. The right-hand side depicts the interrelated importance of the ten predictor variables using a chord diagram.