# Classifying the Quality of Wine: An Ethical Case Study

COSC 4931 - Project 1

David Helminiak
Marquette University
Milwaukee, Wisconsin
david.helminiak@marquette.edu

## ABSTRACT

This work sought to analyze a dataset describing subjective quality assessments of Vinho Verde wine in relation to typical physiochemical tests. This analysis was performed using basic statistical methods and a decision tree algorithm using the Gini impurity index implemented using the Python 3 coding language and without the use of external machine learning packages. This analysis was supplemented by and conducted with a primary interest on the ethical implications imparted by the biases introduced through the original creator(s) of the dataset, those from the analysis itself, and the possible ramifications of conclusions drawn from the results. This overview is to be used as a singular and random case study into how ethical considerations may be being performed within the professional data science community. Basic ethical considerations were made using an existing checklist: deon, produced by DrivenData Labs, which acted as a guideline for the discussion of the possible implications for internal and external regulation within the field of data science.

## CCS CONCEPTS

• **Social and professional topics → Codes of ethics**; • **Computing methodologies** → *Supervised learning by classification*;

## KEYWORDS

Ethics, Case Study, Machine Learning, Python, CART, Classifiers, Decision Trees, Wine Science, Sensory Preferences

**ACM Reference Format:**
David Helminiak. 2018. Classifying the Quality of Wine: An Ethical Case Study: COSC 4931 - Project 1.

## 1 GITHUB

Project code may be found at:
https://github.com/Yatagarasu50469/Helminiak-project1

## 2 INTRODUCTION

A possible definition within the Merriam-Webster Dictionary describes science as, "a system of knowledge covering general truths or the operation of general laws especially as obtained and tested through scientific method" [2] which in turn is defined as the "principles and procedures for the systematic pursuit of knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses" [3]. Each of these definitions limit data science to facts, but lack any sense of ethical responsibility that should be accompanied with their execution. This is a fact that other sciences and engineering fields have long sought to rectify through codes of conduct, as seen with the NSPE (National Society of Professional Engineers), AAPC (American Academy of Professional Coders) and the IEEE (Institute of Electrical and Electronics Engineers) Codes of Ethics, or through organizations such as the Order of the Engineer, which take oaths to follow professional and ethical standards. This also becomes legally enforced with systems including the PE (Professional Engineers) license, that give a level of personal attribution for the performance and failure of designs, particularly pertaining to their public effects. Data science is unique among the sciences as being particularly far-reaching through the sudden machine learning revolution. The field requires relatively minimal levels of understanding for individuals and organizations to develop critical societal infrastructure given the availability of sophisticated packages and services like TensorFlow, Deep Cognition and Google Cloud Machine Learning. This rapid implementation of machine learning technology means there most likely exists a large number of systems and research projects produced without any form, or minimal ethical oversight.

One website for finding datasets to field test machine learning algorithms is the University of California Irvine's (UCI) Machine Learning Repository [9]. A random dataset, the multivariate Wine Quality Data Set [8], was found within the repository and selected for this paper's set of analyses. The dataset will be analyzed through basic statistical analysis, a decision tree formulation code using Gini impurity, and then subsequently examined in light of the deon ethics checklist [4]. The results of this meta-examination will then be used to form possible social and ethical implications both implied and absent from the checklist, as well as impending outcomes.

## 3 DATA INTRODUCTION & EXAMINATION

As a commodity, specific wine varieties can have a large following, being strongly regulated internally by the industry to both certify the end products' quality and adherence to categorical standards. While the actual certification process varies according to the type of wine and their desirous qualities, most perform a regular set of physiochemical tests along with a qualitative evaluation. The multivariate dataset selected for this study: "Wine Quality" was donated to the UCI machine learning repository in October 2009 by Cortez et al. who published a research paper based on their

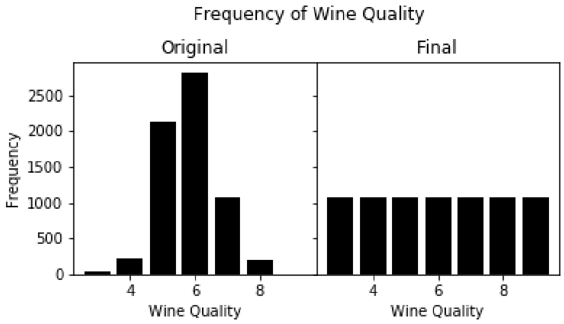

**Figure 1: Frequency distribution for qualitative assessments of the original and balanced (through oversampling) Wine Quality Data Set**

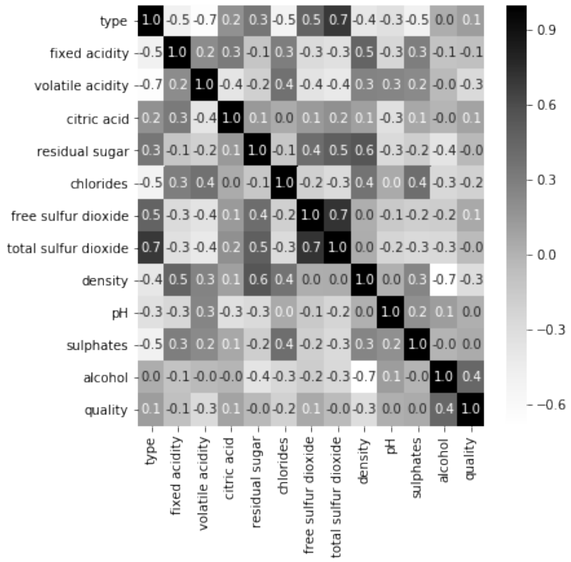

**Figure 2: Correlation between variable values from the Wine Quality Data Set**

own analysis: "Using Data Mining for Wine Quality Assessment," from the University of Minho in Portugal [7]. The set comprises the results of 11 typical physiochemical tests, its type (red, or white wine) and a qualitative assessment conducted on 1599 red and 4898 white Portuguese Vinho Verde wine samples (6497 samples in total), rated from a poor score of 0 to an excellent score of 10. Vinho Verde, characterized by "vibrant freshness, elegance, lightness and aromatic and flavorful expression, especially its fruity and floral notes" has exclusive production in northwestern Portugal using indigenous grape varieties [1].

The set (large, relative to similar studies) was formed to relate the typical physiochemical markers in wines to the subjective expert quality evaluation. Portuguese law dictates that the sensory analysis must be performed by humans, which, given that the measure is inherently subjective, means that in all probability there exists a constant churn in the value assessment data and a fluctuating error rate. Given that the relationships between chemical composition and perception of taste is not presently understood comprehensively, regression and machine learning techniques are a good solution to forming a model. Such a model may be used to generate an expected quality rating that may then inform professional evaluations, price point, and the overall study of wine: oenology.

Statistics for the physiochemical and sensory test results may be seen in Table 1 and the distribution of quality assessments in the left side of Figure 1. The qualitative value is stated in the dataset's documentation to have been averaged from at least 3 evaluations produced by experts [8]. While the set was listed to contain only real values, some variables were observed to be he type of wine was listed as a string "red" or "white". More vital data was absent from the documentation and was only found in Cortez et al. [7]. The data was collected between 2004 and 2007, being tested by the official certification entity, the Viticulture Commission of the Vinho Verde Region (CVRVV). The database includes distinct wine samples in each row and includes the most frequent physiochemical tests. Further, the paper also indicates that the qualitative assessments were made with blind taste tests.

Following Wickham's principles of tidy data [17], the dataset was seen to already have only a single variable name forming each column, keeping a single observation per row, and in total, forming a single observational unit. Binary values were kept on the far left, with fluctuating values in the middle, and the classifier on the far right, which eases the process of understanding the data structure. There were multiple types of values in the structure, with the "type" column containing strings to distinguish "red" and "white" wines. These were replaced with real values to match with the other data types, using 1 for "white" wine and 0 for "red" wine. 0.58% of the dataset samples contained at least 1 missing variable value, with a maximum missing data for a single variable being 0.15%. The distribution of missing values with respect to each variable and their correlation with the desired classifier may be seen in Table 1. The alcohol content and chlorides had the most significant correlation to the wine quality, but each would still lack too much information for classifying the quality on their own. The variables can be seen to have significant inter-variable correlations in Figure 2, while Table 2 shows the absolute maximum correlation values found for each variable with respect to another. The strongest and most expected correlation between different variables was found to be between the free and total sulfur dioxide levels, with the second strongest being between total sulfur dioxide levels and wine type.

## 4 EXISTING RESEARCH

The dataset selected was created alongside a paper from the University of Minho, by Cortez et al. [7], examining data mining approaches to predict wine quality through the use of chemical analytical tests typically performed on wines at the time of their certification. This work compared the accuracy of multiple regression (MR), neural network (NN), and support vector machine (SVM) approaches, in order to create a support system that could inform expert evaluations of wine samples. While the MR model gives a relatively simplistic analysis, the NN and SVM approaches offer non-linear learning capabilities. This variety of models was used

**Table 1: Statistics for the Wine Quality Data Set physiochemical tests [7]**

| Variable | Units | Maximum | Minimum | Mean | #Missing | %Missing | Classifier Correlation |
|---|---|---|---|---|---|---|---|
| type | g(tartaric acid)/dm^3 | 0.00000 | 1.00000 | 0.753520 | 0 | 0.000000 | 0.119323 |
| fixed acidity | g(acetic acid)/dm^3 | 3.80000 | 15.90000 | 7.217755 | 10 | 0.153917 | -0.077031 |
| volatile acidity | g/dm^3 | 0.08000 | 1.58000 | 0.339589 | 8 | 0.123134 | -0.265953 |
| citric acid | g/dm^3 | 0.00000 | 1.66000 | 0.318758 | 3 | 0.046175 | 0.085706 |
| residual sugar | g(sodium chloride)/dm^3 | 0.60000 | 65.80000 | 5.4439585 | 2 | 0.030783 | -0.036825 |
| chlorides | g(sodium chloride)/dm^3 | 0.00900 | 0.61100 | 0.056056 | 2 | 0.030783 | -0.200886 |
| free sulfur dioxide | mg/dm^3 | 1.00000 | 289.00000 | 30.516865 | 0 | 0.000000 | 0.055463 |
| total sulfer dioxide | mg/dm^3 | 6.00000 | 440.00000 | 115.694492 | 0 | 0.000000 | -0.041385 |
| density | g/cm^3 | 0.98711 | 1.03898 | 0.994698 | 0 | 0.000000 | -0.305858 |
| pH | | 2.72000 | 4.01000 | 3.218332 | 9 | 0.138525 | 0.019366 |
| sulphates | g(potassium sulphate)/dm^3 | 0.22000 | 2.00000 | 0.531150 | 4 | 0.061567 | 0.038729 |
| alcohol | % vol. | 8.00000 | 14.90000 | 10.492825 | 0 | 0.000000 | 0.444319 |
| quality | 1-10 (poor to excellent) | 3.00000 | 9.00000 | 5.818505 | 0 | 0.000000 | 1.000000 |

**Table 2: Maximum absolute correlation values found between different variables in the Wine Quality dataset**

| Variable | Abs. Max Correlation | Cross Variable |
|---|---|---|
| type | 0.700521 | total sulfur dioxide |
| fixed acidity | 0.488552 | type |
| volatile acidity | 0.653374 | type |
| citric acid | 0.377512 | volatile acidity |
| residual sugar | 0.551494 | density |
| chlorides | 0.512705 | type |
| free sulfur dioxide | 0.721476 | total sulfur dioxide |
| total sulfur dioxide | 0.721476 | free sulfur dioxide |
| density | 0.687432 | alcohol |
| pH | 0.328474 | type |
| sulphates | 0.486715 | type |
| alcohol | 0.687432 | density |
| quality | 0.444637 | alcohol |

in order to account for the trends of simplistic models, such as MR, failing to comprehensively represent the data, while those that are highly complex, including NN and SVM models, tend to overfit the training data. Using 2/3 of the data for training and the remainder for validation, a sensitivity analysis was performed after the training phase to discard irrelevant inputs. The unbalanced class distribution was preserved, with variables selected as inputs based on their Median Absolute Deviation (MAD) value. The data was standardized to have a 0 mean and 1 standard deviation. The neural network was generated using the Comprehensive R Archive Network (CRAN) RMiner nnet package, set to use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm and end training after 2 iterations without improvement, or else when only 1 input was available. The support vector machine was based on a sequential minimal optimization implementation created using the CRAN kernlab package, provided through A Library for Support Vector Machines (LIBSVM) and using a gaussian kernel. The variable selection found that only 1-2 of the variables were not highly relevant to the quality class. The evaluations of the final models were conducted with a 5-fold cross-validation. With an acceptable error tolerance of 0.5 points, the percent accuracy values were 51.7 +/- 0.1 for the multi-regression, 52.6 +/- 0.3 for the neural network, and 64.3 +/- 0.4 for the support vector machine, improving to 84.3 +/- 0.1, 84.7 +/- 0.1, 86.8 +/- 0.2 respectively with an increased tolerance of 1 point. The SVM was the most expensive computationally, but was found to provide the best accuracy, with the NN slightly outperforming the MR model. Their results were shown to match with oenological theory: increases of alcohol content mapped to higher quality wines, volatile acidity content has a negative impact, sulfates correlate positively, etc. The authors note that a future model generation and comparison could be performed using ranking data mining algorithms, specifically regression trees.

The same dataset was used by Appalasamy et al. [13] as a follow-up in 2012 with Naïve Bayes (NB) and a more successful ID3 decision tree model, training and testing using a 10-fold cross-validation method. Computation times were improved using the initial measure of possible information gain to remove entire chemical tests that had minimal correlation with the desired output classifier: wine quality. White wine had accuracy of 60% for the ID3 tree and 58.8% with the NB model while red wine achieved 52.3% and 50.5% respectively. Given these low accuracy ratings, it was concluded that while the tree allows for more interpretability and therefore more active promotion of quality during the production process, the poor accuracy of both models implies that neither are suitably comparable to the SVM generated by Cortez et al. [7]. Most recently in 2018, Gupta [11] used the set to explore qualitative prediction using linear regression (LR), Multi-Layer Perceptron (MLP) NN, and SVM models. The study found that wine quality was seen to have a limited dependence on the predictors used for the set and that both the NN and SVM models became more accurate when using less of the physiochemical test results.

A similar end objective often used in existing research has been the assessment of wine authenticity, or regional provenance through chemical analysis and subsequent machine learning and/or statistical modeling. The 2013 publication by Gómez-Meire et al. [10], uses gas chromatography and the identified wine aromas of 42 samples from 6 regions to find a model to distinguish wine authenticity through 5 different classifiers: SVM, Random Forests (RF), MLP NN, k-Nearest Neighbor (KNN) and NB with particular emphasis on providing the chemical testing methodologies. It was found that the SVM was the worst for performing the analysis, with perfect classification accuracy obtained by the RF model when using all the available information, but the MLP provided the most accurate results when given limited information. In contrast, using near infrared spectroscopy and least-squares support vector machines Yu et al. [18] correctly classified rice wine, according to 3 age categories from 147 samples, reaching 96% to 100% accuracy in 2008. An earlier 2006 spectroscopy study by Moreno et al. [12] was able to sort 54 commercial wines by their mineral content with 90% to 95% accuracy. Later Azcarte et al. [5] in 2015 also used spectroscopy results to form LDA models to classify 57 white wines from 4 regions with 96% accuracy.

2017 research by Portinale et al. [14] performed an assessment on qualifying the authenticity of 9 wine types using 146 samples to show that Supervised Bayesian Network (BN), SVM, and MLP NN models could provide classification without specialized chemical analysis through progressive removal of chemical predictors. Using 40 predictors, the accuracy levels were 87%, 94%, and 92% respectively, while decreasing to 13 features, resulted in 79%, 82%, and 89% respective accuracies. A Linear Gaussian Bayesian Network (LGBN) was also utilized to generate 14,600 synthetic samples to overcome the lack of available data, resulting in respective accuracies of 92%, 90%, and 94% with the use of all 40 features.

A more comprehensive overview of prior authenticity studies by Versari et al. [16] was published in 2013. For classification of authenticity using minerals, Discriminant Analysis (DA), Analysis of Variance (ANOVA), Classification and Decision Tree (CART), Linear Discriminant Analysis (LDA), Soft Independent Modeling of Class Analogy (SIMCA), Artificial Neural Network (ANN), Canonical Variate Analysis (CVA), Cluster Analysis (CA), Unequal Dispersed Classes (UNEQ), KNN, and Stepwise Linear Discriminant Analysis (SLDA) models have all previously been used, achieving accuracy ratings between 74% to 100%. This summary also covers studies that classify the authenticity of wines using phenolic compounds, additionally using Hierarchical Cluster Analysis (HCA), Principal Component Analysis (PCA), Stepwise Discriminant Analysis (SDA), Quadratic Discriminant Analysis (QDA), Partial Least Square-Discriminant Analysis (PLS-DA), ANOVA-Least Significant Difference (LSD), and Canonical Discriminant Analysis (CDA) models obtaining between 68% to 100% accuracies.

Two unique studies include the 2003 paper by Riu Jr. et al. [15] as well as the Corcoran et al. [6] 1994 publication. Combining ultra-thin film taste sensors with neural networks the former research showed how 900 samples from 6 red wines could be classified with 99-100% accuracy using MLP NN models, picking out the correct vintage, variety, and producer. The second, used a genetic algorithm (GA) on 178 samples from three different grape species (cultivars) to evolve a structure of classification rules, which achieved a maximum 100% classification rate and average of 98.3%.

## 5 DATA ANALYSIS

Given the success of SVMs and NNs within previous works, the general failure of decision trees, and as the presence of a labeled classifier within the dataset allowed for a supervised learning model; a CART decision tree, using a Gini impurity measurement, was chosen for implementation. This decision was also formed by the requirement that the code be produced from scratch without external machine learning packages, using either the Python, or R programming languages. The model has significant room for improvement, seen in the Appalasamy et al. [13] research to only reach a maximum accuracy of 60% for white wine and 52.3% for red wine. However, it offers potential in the development of a high interpretability model. The Gini impurity provides a measure of relative inequality between the original set and the information gained from splitting the data. On its own, the metric does not provide information as to how well the model represents the data, but it may be used to determine how well a hyperplane partitions a population.

$$\mathbb{G}(D) = \sum_{i=1}^{k} Prob(c_i|D)(1 - Prob(c_i|D)) = 1 - \sum_{i=1}^{k} Pob(c_i|D)^2 \quad (1)$$

This value is calculated through Equation 1, where D is the dataset, i is the characteristic index, c, the current characteristic, and k, the maximum number of characteristics. If there is only a single category, then the Gini value is 1, if there are 2 categories then the Gini impurity is 0.5. Comparing the Gini impurity before and after bisecting the data, with a hyperplane for a given attribute, provides a measure of information gain. The CART tree therefore is formed by examining the dataset's starting population, in a starting node, to determine which split, over all of the variables, provides the greatest information gain; executing that split and repeating over the resulting populations, the starting node's left and right children (progressing from top-to-bottom), until the termination requirement is met. The dataset is always randomized when loaded to ensure randomness in the training and testing sets.

For this tree, the termination criteria for splitting populations was initially set as the Gini impurity of the set's classifier being reduced to 1% of its original value, or when the nodal population size was decreased to a given size. The code allowed for arbitrary specification of this size, or automatically assigned it as 1/12 (12 being the number of defining characteristics) of the dataset's starting size. An example tree using these criteria and 0.1% of the data for training can be seen in Figure 3. Though the code can generate these visual models for any given percentage of the data used for training, they are too large to include in this report.

Given the unbalance in the class distribution, it was known that the model would have a higher accuracy rating for classes that appeared frequently. To compensate, options to use oversampling, or synthetic value generation from normal and truncated normal distributions were created; The oversampling method draws randomly selected samples from the training set in order to build a balanced distribution of classes. This can be seen to function as desired in the right side of Figure 1. However, the results leaves little variation for the values in the final training data for classes that were under-represented in the starting set. Using the normal distribution option, each class' mean and variance from the starting training set can be used to create distributions, from which values are randomly taken to synthetically manufacture balanced sample sets. This should allow the program to generate values slightly outside the originally under-represented class' data purview. The truncated normal distribution does the same synthetic generation

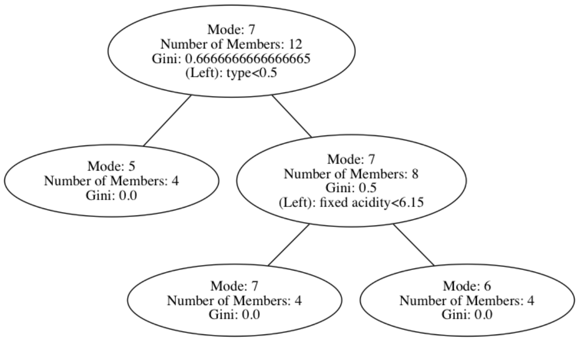

**Figure 3: Example CART decision tree model trained using 1% of the Wine Quality Data Set balanced using oversampling**

process as the normal, but cuts off sample generation according to the original training set classes' minimum and maximum values. Doing so should increase the program's fit of the training data, but could be proven to produce an overfit model, preventing correct classification of values not encountered during training.

Binary classifiers, such as the wine 'type' were also anticipated in the code design, making sure that the number of samples generated for each binary classifier were always equal. Should a classifier be unique to a single binary classifier, it will still have an equal number of samples produced as the other classifiers.

Initial experimentation with training sets with 0.1% and 1% of the sample population were able to reach up to 48% accuracy. However, this is most likely due to the unbalanced set, where only more frequent values were used to form the balanced training set. Using a typical 80% training, 20% testing split, resulted in 15.3% accuracy. In this instance, the Gini impurity at the end nodes, was noted to be extremely high. The code was therefore modified to terminate when there were only a single class remaining. While this produces trees containing 100s to 1000s of nodes, the results may still be considered highly interpretable, as a human could follow the chain of decisions by which any singular decision was made. Also, the training set had been set to be taken from a balanced version of the whole population, to eliminate the imbalance of accuracy ratings seen between low and high training/testing population allocation. The most computationally intensive section, simulation of the data split points, was constructed for independent parallel pipeline processing execution, being run on a dual-socket 32-thread 2.6 GHz Hackintosh workstation. Variations in the training set size were set to be 0.1%, 1%, 10%, and 80% of the dataset's length, with the remainder kept for use in testing. The program remains very computationally expensive, with an early version of the 80% training set evaluation, taking 23 hours to complete. To further accelerate the computation, variables that had correlation values with the wine quality less than the mean of these values were removed from the training dataset. However, the 80% population sizes still require several hours to compute and were unable to be evaluated in this report given time limitations.

The testing results that could be computed are seen in Table 3 and may be seen to show that the model failed to produce any suitable classification tree for the intended task. The truncated sampling technique gave the best accuracy of 28.489%. However, both the synthetic sample generation models had lower accuracy results for their 10% training sets than either the 0.1%, or 1% values. In contrast, the oversampling technique had a consistent upward accuracy trend, suggesting that it may be a more valid method for future experimentation.

**Table 3: Accuracy tests for CART decision tree with test options listed as 0 for No and 1 for Yes: Over?: oversampling performed; Synth?: synthetic generation performed; Norm?: normal distribution sampling; Trunc?: truncated normal distribution sampling; %Train: percent of the dataset used for training; %Accuracy: percent accuracy found using the testing data**

| Over? | Synth? | Norm? | Trunc? | %Train | %Accuracy |
|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0.1 | 13.722 |
| 1 | 0 | 0 | 0 | 1 | 21.238 |
| 1 | 0 | 0 | 0 | 10 | 25.563 |
| 1 | 0 | 0 | 0 | 80 | NA |
| 0 | 1 | 1 | 0 | 0.1 | 14.047 |
| 0 | 1 | 1 | 0 | 1 | 23.972 |
| 0 | 1 | 1 | 0 | 10 | 22.83 |
| 0 | 1 | 1 | 0 | 80 | NA |
| 0 | 1 | 1 | 1 | 0.1 | 10.717 |
| 0 | 1 | 1 | 1 | 1 | 28.489 |
| 0 | 1 | 1 | 1 | 10 | 24.256 |
| 0 | 1 | 1 | 1 | 80 | NA |

## 6 ETHICAL CHECKLIST

The deon ethics checklist [4] for data scientists was used to evaluate the ethical ramifications of this case study pertaining to data collection, storage, analysis modeling, and system deployment. Researching beyond the dataset and its immediately provided documentation, into the indirectly associated Cortez et al. [7] paper, revealed critical information regarding many criteria considered within the checklist, particularly for data collection. In order to provide insight into the importance of this information, the set and its immediate documentation will be considered on their own merits, prior to consideration with the information provided by the Cortez et al. [7] research.

During the data collection process an unknown number of human subject(s) were used to generate at least 3, but not necessarily the same number, of qualitative analyses for each of the wine samples. These subject(s) are not identified through any personally identifiable information and no information was provided with the dataset, or subsequent analyses as to whether there was a mechanism where consent could be granted, or denied. Given that the intent behind the dataset was to generate a program by which professional tasters could be potentially replaced by chemical analysis and the study's resulting algorithm, it may be considered that the tasters could have deliberately sabotaged their results in order to protect their job security. Further, not having informed them would also constitute grounds for an ethical violation.

The experts' actual testing methodology, specific tasting criteria, or professional qualifications were not provided by the dataset's direct documentation. It could not be confirmed if the tasting was performed as a blind, or double-blind test and therefore it was unknown if any financial incentives were provided to these individuals that may have biased the results. This lack of information on the testers also calls into question what their innate bias was for wines. For example, it is possible that some of them specialized in red, or white wines, leading to the over-criticism of certain samples. Whether each wine underwent the 3-part evaluation process multiple times remained unclear, leading to the possibility that singular wines were used more often to pad out the data. Other pertinent information was not provided as to whether these test samples were of the same vintage, grape composition, cask, grown in the same soil composition, or even evaluated under identical test conditions by the same people, such as the temperature that the wine was served, or how the palette was cleansed between tests. The possible injected bias from the data collection became further compounded by a lack of references to the actual chemical testing procedures, or third-party laboratory that enacted the evaluations. This called into question the accuracy of the results presented, along with questions as to why random data points are absent. Values may have been found at these points and removed as being statistically deviant from the general trends, but neither this possibility, or the failure to collect a sample was evidenced by the documentation.

The data collection process completely succeeded in eliminating any personal identifying information from the end dataset, with the only names mentioned within the accompanying documentation being those who utilized the end dataset for published research. Wine producers, brands, grapes, and price were all censored in consideration of privacy and logistic concerns, leaving only test results.

Looking in the Cortez et al. paper [7] renders the majority of these concerns mute, as it indicates that the tasters and the physiochemical tests were conducted through the wine's certifying authority, CVRVV. This means that consent and professional levels of certification are significantly more likely to have been obtained for the dataset's creation. The dataset can also be found, through the Cortez et al. work [7], to have been collected over several years and that each row is a unique wine. There still remains no indication as to how, or from whom consent was obtained to make use of the data. The actual number of expert tastings for each sample remain unknown, but given that the tasters most likely had credentials suitable for the assessment, this becomes significantly less problematic. The same evaluation is true of the accuracy for the physiochemical tests, as knowing who conducted them makes it possible to contact them directly to determine and cross-validate their results.

Regarding storage, the dataset has no indicated plan to protect and secure the data, as it was donated to the UCI Machine Learning Repository and is publicly accessible without any internal mechanisms for access controls The repository does not list how it might be secured, while the site does transit the information through a secure hypertext protocol. Hashes to check successful transference of data are neither provided by the repository, nor through a relevant research paper. The repository lists the number of times that the page has been visited, so while it cannot be confirmed, it may have a simple IP logging system implemented. A scan using OWASP ZAP (Open Web Application Security Project Zed Attack Proxy) shows a single significant potential risk vulnerability that could result in the use of clickjacking techniques. However, this risk does not present an immediate exploitation that could manipulate the dataset as downloaded by the user. Mechanisms are not present for an individual to request for their personal information to be removed. However, this criterion was judged to not be applicable given the measures taken to remove any personal identifiable information from the data and its documentation. The data has no plans, or schedule to be deleted, but was donated for other people to perform analyses on the set. Again, given the steps taken for privacy prior to the set's publication this does not raise any undue concerns, though the UCI repository does not have any immediate option for requesting that a dataset be removed from the repository, except through direct contact with the website host.

The dataset has been documented and shown to provide unbalanced classes and absent data points. Given the relatively minimal proportion of points missing data of 0.58% of 6497 samples these samples were eliminated from this study's analysis. The unbalanced classes were addressed through oversampling and synthetic value generation based on the original set using the variance, mean, maximum, and minimum values for each class. Data variables were found to be sufficiently clear, with descriptions and units provided in the accompanying documentation. Given logistic constraints for this study neither the relevant stakeholders for the dataset, nor subject matter experts could be contacted to address possibilities of finding information relevant to expressed concerns, nor for discussion of the implications resultant from this study.

Visualizations and summary statistics were generated to honestly represent the data provided. Without information in the dataset containing personally identifiable information, no censorship was needed for ethical visualization of the dataset. The actual code used to generate the analysis presented in this report has been publicly posted, with a link provided at the start of this report, making the results highly reproducible and open to critique.

Both information on the actual physiochemical tests used prevents assurance that the model does not rely on values that unfairly discriminate and the possibility of generating/evaluating relevant error rates. The data was optimized to remove data points that lacked information and provide input types that are interpretable by the code. The actual values processed from the dataset could be considered to have been altered through the synthetic sampling methods, however, they are, as has been indicated, to still be representative as their derivation was performed from the original set. Additional metrics, though not used, were considered to augment the models presented by this work. Particularly, it was noted that a direct comparisons with the codes used by previous works could have shown optimizations and patterns not yet considered. Another metric that has been pointed to was that as the decision tree was generated using the Gini impurity, the model provides immediate justification for how a conclusion was reached by the model. General explanations of the shortcomings and limitations of the models generated have been indicated for readers of this work.

A plan has not been established for the potential harm of users affected by the results of this work, or for possible drift in the oenology field. Since the code and dataset that generated the models made herein are publicly available, they are easily reproduced and modified to reflect any updates, modifications, or entirely new versions of the dataset. Should users be harmed by the model results, the code base is available for further evaluation and audits. Given that the modeling and analysis techniques here are regularly used within the statistical and machine learning communities there are no plans to prevent, or monitor for, unintended uses and abuse of the model.

## 7 IMPLICATIONS

The literature review conducted for this study found many existing classifier models that have high to perfect accuracy levels and compared them with alternative modeling techniques. The most significant deviations in model performances between studies appeared to be due to the variation of obtained variable types (sensory evaluations, chemical testing, chromatography, and spectroscopy), sample sizes (ranging from starting sets of 10s to 1,000s), and objective classification goals (authenticity, regional, type, and quality). Although there are aggregate, overview evaluations of the effectiveness of machine learning usage for wine classifications, such as by Versari et al. [16], none have generated an idealized, singular formula that can objectively and accurately state the quality of any singular wine type. Models that have high interpretability, such as the ID3 CART tree, simply did not have high enough accuracy to be considered for general purpose use. Those that had sufficient accuracy, most generally SVMs and NNs, were only good for very specific applications and datasets. They were not generalized for broader conception of how, or why, any singular decision was made. These observations imply that a final, idealized classification tree from this study would also lack sufficient accuracy. However, given that SVMs and NNs have been shown to exist that can produce high accuracy, it is implied there exists a line of reasoning by which these classifiers can be predicted in an interpretable manner.

Ideally, it was intended that this study could have replicated and built on the results of Cortez et al. [7] and Appalasamy et al. [13] generating multiple models to determine an optimum strategy for classification. However, this study was not successful in producing a useful classifier model. While this study eliminated variables that possessed lower correlation with the output classifier, to handle the computational expense encountered by the code used, ideally all variables should be included for the analysis. An alternative computation method that could bypass this issue entirely for this dataset would be GPU acceleration. This is particularly desirous as while some variables might not appear in the whole population to be significantly variant, they may be statistically important when splitting a sub-population. Implementing this would require significant modification to the current code structure, since Graphical Processing Units (GPUs) typically only allow the use of float32 and float64 datatypes in their execution space. Also, while the dataset is randomized each time the algorithm is run to produce variation in the testing and training sets, implementing a k-fold cross-validation method, or at least some averaging of multiple runs should be added to help mitigate variations in test results. Another important factor not optimized in this study was the use of Gini impurity to split nodal populations. It is suspected that the poor accuracy of the decision tree models, both in this and other studies is due to the hard line hyperplane separation, which prevents overlap between correlation results. This might be compensated for by combining hyperplanes for multiple variables that yield high information gain to split the data according to the interactions and correlations between variables, rather than between single variables and the output classifier.

However, given a final, optimized version of this study's decision tree it would be interesting and desirous to attempt to make an idealized NN that could be combined with the tree to determine the splitting points for each nodal population. This NN would take in a set of data, deciding which samples belonged to a singular class, splitting the remainder into another node. The resulting two sections could then be analyzed to determine which variables were statistically the most diverse and therefore responsible for the decision made by the NN. This could produce a hybrid machine learning classifier containing both high accuracy and interpretable results. Should this prove infeasible, the reverse procedure might be possible, taking high accuracy NN and SVM models, querying them with synthetic data, and then reverse engineering the most significant classifier values. An entirely alternative direction for this work, not examined by the existing works discovered, would be the evaluation of wine consumers' preferences and the market in general in order to form a new qualitative measurement to inform which chemical proportions yield the most profitable results.

With regards to the ethical checklist there was limited information regarding the possible mitigation of biases involved within the dataset and its direct documentation, save for averaging quality evaluations for each given sample. While the data collection process did take into account privacy concerns, only providing the end results, this action could have prevented attribution and evaluation of the biases that were possibly injected by them. Prior to finding Cortez et al. [7], the results found for their, this, and any other research findings derived from the set provided could have been considered invalid, given the lack of experimental procedure and information regarding test accuracy. Particularly problematic was the lack of evidence for certification on the part of the wine evaluators, meaning that any results could not be trusted from this study until it could be cross-validated by another researcher(s), with evidence given to support that it was performed independently from the original sources. Of course, this would mean that the researcher(s) would need to know what sources to avoid, made difficult by them not being provided! While these concerns were ultimately mitigated by finding and examining the Cortez et al. [7] paper, it highlights a key danger with regards to the use of data without provenance. The deon ethical checklist [4] does not provide any mandate that sufficient attribution is provided for results to be replicated, or validated by independent researchers. Even if such a criterion would be bounded by a requirement for consent by the parties involved, it remains critical for the study to be considered in line with the scientific method.

This study's analysis is unlikely to produce social, or ethical issues particularly due to the large number and variety of shortcomings within the generated model. More generally, the study's selected model placed the primary value on its interpretability rather than its accuracy. The public exhibition of code allows for a high degree of auditability, as well as availability for anyone to test for and then call attention to concept drift. Given this public deployment, usage cannot be observed for unintended uses, or abuse of the model, nor planned was it for.

## 8 CONCLUSION

This study was intended to develop a model by which physiochemical test results from a relatively large dataset could be mapped to a subjective expert evaluation. The maximum accuracy obtained during testing was approximately 48%, largely influenced by inefficient sampling and balancing of the selected training subset. More typical accuracy was found to be between 20% and 30% for training populations between 0.1% and 10%, with the 80% training population sizes unable to be run in the study's time frame. The oversampling method for class balancing was seen to be more promising than the synthetic normal and truncated normal sample generation, given its consistent increase in accuracy with respect to training population size. Future work intends to finalize analysis of this study's classifier tree model and possibly implement GPU acceleration for reasonable computation times. Should the model fail with larger training population sizes to produce higher accuracy levels than previous studies, then integration with more complex machine learning techniques will be considered to maximize the model's usefulness, while preserving its interpretability.

The study further examined the social and ethical implications of the analysis performed and the dataset used. It was identified that prioritizing privacy and logistical concerns over the ability to cross-validate the data collection methodology could produce results unacceptable for practical usage, given that they have no evidence to disprove any allegations of biased inputs and/or evaluations. The deon ethics checklist [4] was able to point directly towards the possible biases involved within the data, but its encouragement of the obfuscation of personally identifying information may detrimentally affect the potential for studies to be wholly replicated and cross-validated as required by the scientific method. Unless more focus becomes placed on the combination of generally considered requisite ideals of the scientific method in with ethical considerations, data science will continue to fail in the production of meaningful, accurate results.

## ACKNOWLEDGMENTS

The author would like to thank Cortez et al. for the public donation of the Wine Quality Data Set to the UCI Machine Learning Repository and the repository administrators for maintaining the website for public consumption.

## REFERENCES

[1] [n. d.]. About Vinho Verde. Retrieved October 12, 2018 from http://www.vinhoverde.pt/en/about-vinho-verde.

[2] [n. d.]. science. Retrieved October 16, 2018 from https://www.merriam-webster.com/dictionary/science.

[3] [n. d.]. scientific method. Retrieved October 16, 2018 from https://www.merriam-webster.com/dictionary/scientific%20method.

[4] 2018. An ethics checklist for data scientists. Retrieved October 12, 2018 from http://deon.drivendata.org/#data-science-ethics-checklist.

[5] Silvana M. Azcarate, Luis D. Martinez, Marianela Savio, José M. Camiña, and Raúl A. Gil. 2015. Classification of monovarietal Argentinean white wines by their elemental profile. *Food Control* 57 (nov 2015), 268–274. https://doi.org/10.1016/j.foodcont.2015.04.025

[6] A.L. Corcoran and S. Sen. [n. d.]. Using real-valued genetic algorithms to evolve rule sets for classification. In *Proceedings of the First IEEE Conference on Evolutionary Computation. IEEE World Congress on Computational Intelligence*. IEEE. https://doi.org/10.1109/icec.1994.350030

[7] Paulo Cortez, Juliana Teixeira, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Using Data Mining for Wine Quality Assessment. In *Discovery Science*, João Gama, Vítor Santos Costa, Alípio Mário Jorge, and Pavel B. Brazdil (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 66–79.

[8] Paulo Cortez, Juliana Teixeira, António Cerdeira, Fernando Almeida, Telmo Matos, and Jos Reis. 2009. Wine Quality Data Set. Retrieved October 12, 2018 from https://archive.ics.uci.edu/ml/datasets/Wine+Quality.

[9] Dua Dheeru and Efi Karra Taniskidou. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[10] S. Gómez-Meire, C. Campos, E. Falqué, F. Díaz, and F. Fdez-Riverola. 2014. Assuring the authenticity of northwest Spain white wine varieties using machine learning techniques. *Food Research International* 60 (jun 2014), 230–240. https://doi.org/10.1016/j.foodres.2013.09.032

[11] Yogesh Gupta. 2018. Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science* 125 (2018), 305–312. https://doi.org/10.1016/j.procs.2017.12.041

[12] Isabel Moreno, Dailos González-Weller, Valerio Guiterrez, Marino Marino, Ana Cameán, A. González, and Arturo Hardisson. 2007. Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks. *Talanta* 72, 1 (apr 2007), 263–268. https://doi.org/10.1016/j.talanta.2006.10.029

[13] A. Mustapha, P. Appalasamy, N.D. Rizal, F. Johari, and A.F. Mansor. 2012. Classification-based Data Mining Approach for Quality Control in Wine Production. *Journal of Applied Sciences* 12, 6 (jun 2012), 598–601. https://doi.org/10.3923/jas.2012.598.601

[14] Luigi Portinale, Giorgio Leonardi, Marco Arlorio, Jean Daniel CoÃŕsson, Fabiano Travaglia, and Monica Locatelli. 2017. Authenticity assessment and protection of high-quality Nebbiolo-based Italian wines through machine learning. *Chemometrics and Intelligent Laboratory Systems* 171 (dec 2017), 182–197. https://doi.org/10.1016/j.chemolab.2017.10.012

[15] Antonio Riul, Humberto C. de Sousa, Roger R. Malmegrim, David S. dos Santos, André C.P.L.F. Carvalho, Fernando J. Fonseca, Osvaldo N. Oliveira, and Luiz H.C. Mattoso. 2004. Wine classification by taste sensors made from ultra-thin films and using neural networks. *Sensors and Actuators B: Chemical* 98, 1 (mar 2004), 77–82. https://doi.org/10.1016/j.snb.2003.09.025

[16] Andrea Versari, V. Felipe Laurie, Arianna Ricci, Luca Laghi, and Giuseppina P. Parpinello. 2014. Progress in authentication, typification and traceability of grapes and wines by chemometric approaches. *Food Research International* 60 (jun 2014), 2–18. https://doi.org/10.1016/j.foodres.2014.02.007

[17] Hadley Wickham. 2014. Tidy Data. *Journal of Statistical Software, Articles* 59, 10 (2014), 1–23. https://doi.org/10.18637/jss.v059.i10

[18] Haiyan Yu, Hongjian Lin, Huirong Xu, Yibin Ying, Bobin Li, and Xingxiang Pan. 2008. Prediction of Enological Parameters and Discrimination of Rice Wine Age Using Least-Squares Support Vector Machines and Near Infrared Spectroscopy. *Journal of Agricultural and Food Chemistry* 56, 2 (jan 2008), 307–313. https://doi.org/10.1021/jf0725575