Machine Learning Applications in Wine Production Quality and Satisfaction

CSCI 5622: Machine Learning

Author: Samuel Kwon

Department of Computer Science
University of Colorado Boulder
12/04/2022

Contents

1 Introduction

The varieties of red and white wine are increasing every year. The increased production and sales of wine make it harder for consumers to purchase good quality wine. Furthermore, from a process standpoint, it is important to find key processes and chemical parameters that produce the best wine to provide the best wine sales to customers. In this paper, machine learning techniques will be applied to wine-related datasets to explore the relationship between learning techniques and their potential to model wine production.

Wine production starts with the harvest of grapes. Different grape varieties grow in various regions. Depending on the variety, the grape can influence the chemical composition, taste of the final product, and overall consumer satisfaction. The most popular and well-known grape varieties are Cabernet Sauvignon, Shiraz, Pinot Noir, Sauvignon Blanc, and Merlot. Not only can the grape variety influence the final product, but the production process plays a role in the quality of the wine. The production process and the steps taken to prepare the grapes play a role in the wine's chemical composition, which changes the aroma and taste of the wine. For example, some producers might decide to keep the skin of the grape on while others decide to take it off. The decision to keep or remove the skin of the grape can impact the concentration of tannins in the grape, which will influence the texture aspect when drinking the wine. Depending on the goal of the final product, one variety or a combination of multiple varieties will be harvested. The combination of multiple varieties is called blended wines, which tend to have their unique chemical composition and profile. After the preprocessing of the grapes, the next step of wine production is fermentation. Fermentation is performed using yeast. Red wine production will include the skin during fermentation and storage. White wines tend to use grapes without the skin, although some practices use the skin contact method. The fermentation process is divided into a two-step process where the first step is yeast fermentation, and the second is malic acid fermentation. During alcohol production and fermentation, wines develop various congeners, influencing the taste and flavor profile. Congeners are organic compounds that influence the overall flavor profile and quality of the wine. The last step in the production process is aging, where the product is left inside a wooden, glass, clay, or steel container. Just like how whiskey barrels give a wooden, oaky flavor and scent to the final product, aging the wine also changes the alcoholic content and concentration of congener compounds. Wines can also go through additional clarification processes giving the final product.

Some past works explore the applications of machine learning techniques to predict wine quality. Yogesh Gupta applies linear regression, neural networks, and support vector machines on wine chemical composition datasets to perform selection and prediction. Linear regression was utilized to examine the relationship between the labeled data and prediction variables. Then, an Artificial Neural Network (ANN) with three layers and support vector machines is used to predict the dependent variables. Gupta utilizes red wine and white wine datasets with 12 chemical characteristics. The quality variable is the label data, and the 11 other features are predictors. Using ANN and support vector machines, Gupta concluded that the quality of the wine could be accurately represented if correct features were included in the model. For example, when running linear regression, the author found that citric acid, chlorides, and total sulfur dioxides play a statistically insignificant role in the model. As a result, during model preparation, optimizing the number and predictors is important. Besides the work performed by Yogesh Gupta, K.R Dahal et al. expand the techniques to include ridge regression and gradient boosting regression. The wine chemical compound dataset was also utilized in this paper. Comparisons of models indicated that the gradient boosting algorithm performed the best compared to ANN, ridge regression, and support vector machines.

Current research emphasizes the applications of supervised learning and deep learning methods to generate a model. To further extend current research, it would be important to apply unsupervised learning and text data methods. Exploring the models using text data would give insight into public sentiment towards wine and how the people or media perceive wine. This paper will apply unsupervised, supervised, and neural networks to various wine-related datasets. Text data gained from news outlets will be processed to explore associations related to wine. The project seeks to explore the properties and satisfaction associated with some red wines, the relationships between chemical composition and consumer satisfaction, and information about wine available to the public through news media outlets.

Unsupervised learning will be implemented first to explore chemical composition and quality. In some of the techniques used in this paper, the dataset will be discretized into good, average, and bad-quality wine. The prediction is that wine can generally be divided into qualities. In some countries, wine is regulated, and an independent institution oversees wine production. As a result, some wines might reflect that system in the quality ratings.

2 Data Cleaning and Visualization

2.1 Analyses

Four datasets have been collected and cleaned so far. The first dataset, called the Wine Enthusiast dataset, has qualitative and quantitative variables on consumer reviews. The second dataset is the Chemical Analysis dataset which only has quantitative data on chemical compounds in wine. The third dataset is Portugal Wine data, where the set contains values of chemical compounds and consumer satisfaction. The fourth dataset is News API data showing current news articles related to wine.

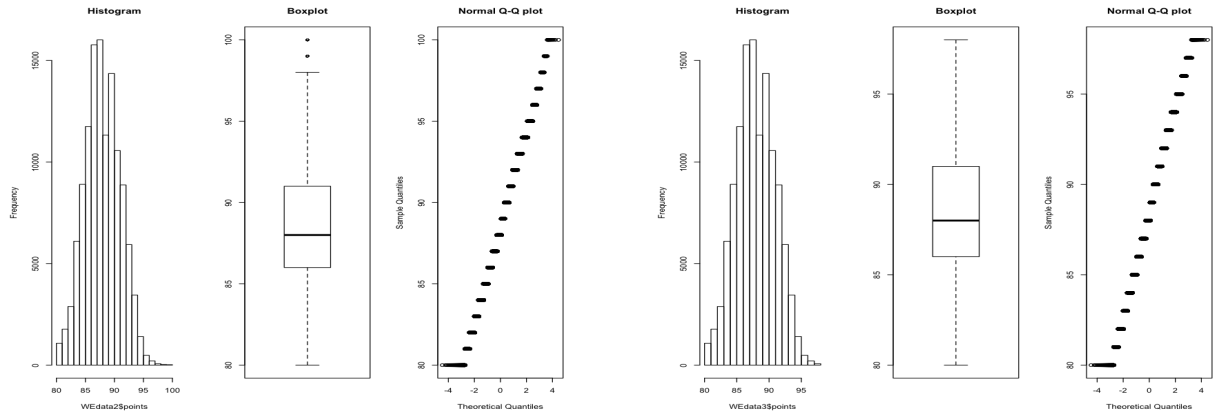
Wine Enthusiast Dataset

This dataset was obtained through Kaggle, where a user posted web scrapping data from the Wine Enthusiast. The Wine Enthusiast examines independent wines, and the following data values are provided: country, description, points, price, province, title, and variety. The dataset has seven columns and 129971 rows which is a relatively large dataset.

Figure 1 shows that the price and point vectors are incomplete and have some empty values. It is hard to replace wine price data points as they vary drastically and have no standard in determining price. The range of the price dataset is large, meaning that replacing it with mean or median would impact the interpretation and proper representation of the dataset. It is important to note that the dataset has 129953 total rows of data, and only 8996 are incomplete. As a result, the whole row with missing price and points data values was removed from the dataset. This dataset is a combination of qualitative and quantitative values. The price and point data set were examined individually using a histogram, boxplot, and QQ plot. The points histogram, boxplot, and QQ plot are presented in figure 2a. Note that the dataset presents some outliers seen in the boxplot and QQ plot. Because the dataset is relatively large and has some extreme values, the interquartile range method was utilized to remove outliers. Figure 2b is the cleaned points dataset. Notice that the boxplot and QQ plot show that the outliers were removed. The price dataset was also cleaned using the same method. It is important to note that price has extreme outliers since wine prices can vary drastically. Figure 3a and 3b show the uncleaned and cleaned dataset. The cleaned dataset has some outliers shown in the boxplot and QQ plot. The interquartile range method could not remove all the outliers; however, those remaining outliers will be kept as noise for the model. Lastly, the country, description, province, title, and variety data were cleaned by converting the capital words to lowercase letters for text data usage. Figure 4 represents a sample of the cleaned dataset. The final dataset has around 113730 rows, which show a reduction from the earlier dataset.

country	description	points	price	province	title	variety
Italy	Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.	87	NA	Sicily & Sardinia	Nicosia 2013 Vulkà Bianco (Etna)	White Blend
Portugal	This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.	87	15	Douro	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red
US	Tart and snappy, the flavors of lime flesh and rind dominate. Some green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented.	87	14	Oregon	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris
US	Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish.	87	13	Michigan	St. Julian 2013 Reserve Late Harvest Riesling (Lake Michigan Shore)	Riesling
US	Much like the regular bottling from 2012, this comes across as rather rough and tannic, with rustic, earthy, herbal characteristics. Nonetheless, if you think of it as a pleasantly unfussy country wine, it's a good companion to a hearty winter stew.	87	65	Oregon	Sweet Cheeks 2012 Vintner's Reserve Wild Child Block Pinot Noir (Willamette Valley)	Pinot Noir
Spain	Blackberry and raspberry aromas show a typical Navarran whiff of green herbs and, in this case, horseradish. In the mouth, this is fairly full bodied, with tomatoey acidity. Spicy, herbal flavors complement dark plum fruit, while the finish is fresh but grabby.	87	15	Northern Spain	Tandem 2011 Ars In Vitro Tempranillo-Merlot (Navarra)	Tempranillo-Merlot

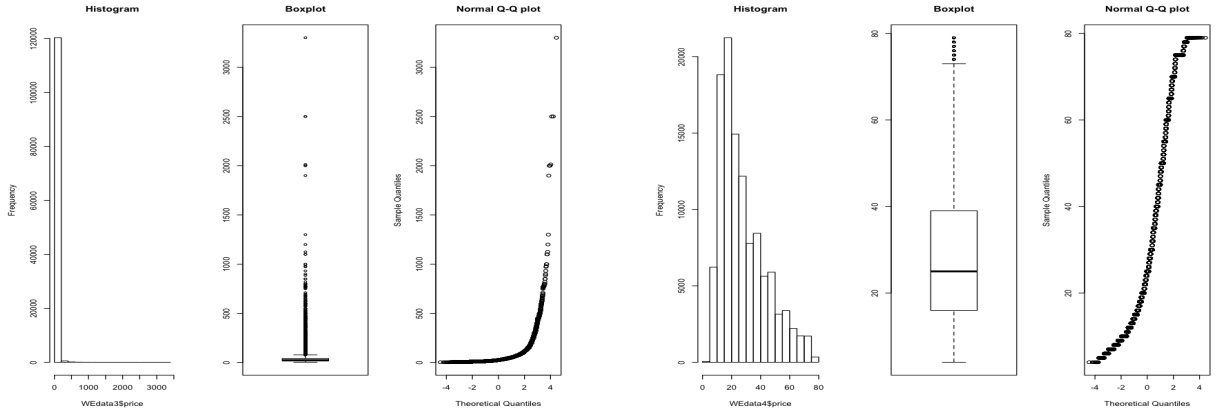
Figure 1: Wine Enthusiast Uncleaned Dataset



(a) Uncleaned Points Dataset

(b) Cleaned Points Dataset

Figure 2: Points Dataset



(a) Uncleaned Price Dataset (b) Cleaned Price Dataset

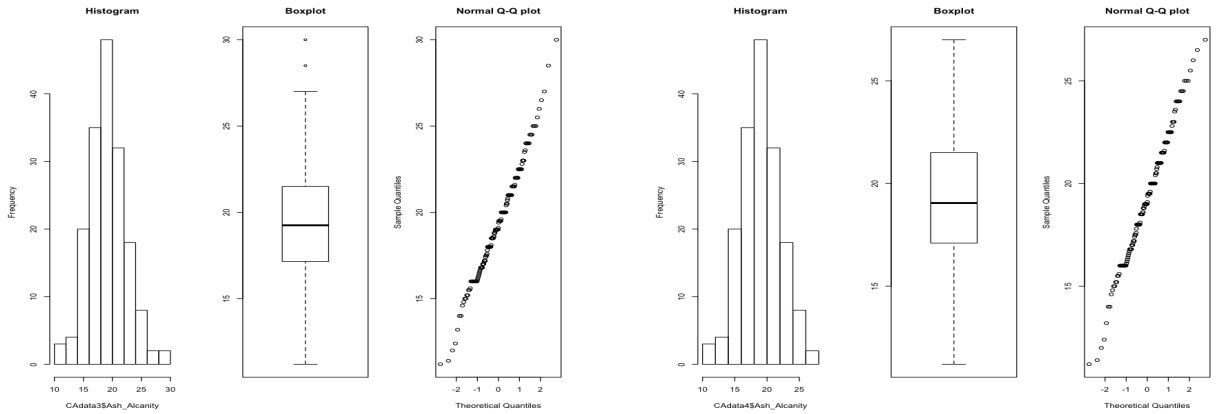
Figure 3: Price Dataset

	country	description	points	price	province	title	variety
2	portugal	this is ripe d fruity, a wine that is smooth while still structured. firm tnins are filled out with juicy red berry fruits d freshened with acidity. it's already drinkable, although it will certainly be better from 2016.	87	15	douro	quinta dos avidagos 2011 avidagos red (douro)	portuguese red
3	us	tart d snappy, the flavors of lime flesh d rind dominate. some green pineapple pokes through, with crisp acidity underscoring the flavors. the wine was all stainless-steel fermented.	87	14	oregon	rainstorm 2013 pinot gris (willamette valley)	pinot gris
4	us	pineapple rind, lemon pith d orge blossom start off the aromas. the palate is a bit more opulent, with notes of honey-drizzled guava d mgo giving way to a slightly astringent, semidry finish.	87	13	michigan	st. julian 2013 reserve late harvest riesling (lake michigan shore)	riesling
5	us	much like the regular bottling from 2012, this comes across as rather rough d tnic, with rustic, earthy, herbal characteristics. nonetheless, if you think of it as a pleastly unfussy country wine, it's a good compion to a hearty winter stew.	87	65	oregon	sweet cheeks 2012 vintner's reserve wild child block pinot noir (willamette valley)	pinot noir
6	spain	blackberry d raspberry aromas show a typical navarr whiff of green herbs d, in this case, horseradish. in the mouth, this is fairly full bodied, with tomatoey acidity, spicy, herbal flavors complement dark plum fruit, while the finish is fresh but grabby.	87	15	northern spain	tandem 2011 ars in vitro tempranillo-merlot (navarra)	tempranillo-merlot
7	italy	here's a bright, informal red that opens with aromas of cdied berry, white pepper d savory herb that carry over to the palate. it's balced with fresh acidity d soft tnins.	87	16	sicily & sardinia	terre di giurfo 2013 belsito frappato (vittoria)	frappato

Figure 4: Wine Enthusiast Cleaned Dataset

Chemical Analysis Dataset

The Chemical Analysis dataset is from the UCI Machine Learning repository. This dataset results from a chemical analysis of wines grown in Italy. Around 13 column variables are present, and they are Alcohol, Malic Acid, Ash, Ash Alcanity, Magnesium, Total Phenols, Flavanoids, Nonflavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280, and Proline. The dataset has 178 rows, and no values are missing from the set. Since the dataset is quantitative, the sets were examined, and outliers were removed using the interquartile range method because the distributions vary between the variables. The histogram, boxplot, and QQ plots were used to determine outliers that should be removed. Some variables, such as alcohol, total phenols, flavonoids, OD280, and nonflavonoid phenols did not have outliers. In most cases, the IQR method did not remove all of the outliers, but kept some upper and lower-bound outliers. Since the number of outliers was minimal, they were kept in the set as noise. Figure 5 is a sample of cleaned and uncleaned Ash Alcanity data. Please refer to the references section to see the cleaned version of each variable.

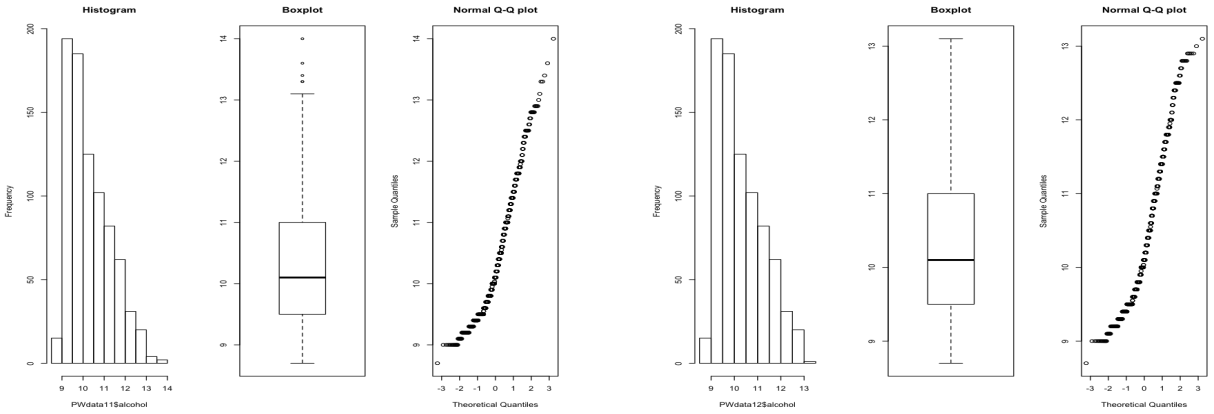


(a) Ash Alcanity Uncleaned Dataset (b) Ash Alcanity Cleaned Dataset

Figure 5: Ash Alcanity Dataset

Portugal Wine Dataset

The Portugal Wine dataset is from the UCI Machine Learning repository. The dataset is related to red and white variants of the Portuguese Vinho Verde wine. Around 13 variables are present: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and quality. The other variables are considered when creating the discrete variable, quality. The Portugal Wine dataset is relatively large, with 1143 rows, and only contains quantitative data. All of the variables are complete and have no empty values. As a result, outliers were examined for cleaning the data. When cleaning this dataset, the free sulfur dioxide variable was removed from the set since total sulfur dioxide is present. The total sulfur dioxide is a combination of free sulfur dioxide and bounded sulfur dioxide. The same procedure was applied to examine outliers. Histograms, boxplots, and QQ plots were generated to examine outliers. After, the interquartile range method was utilized to eliminate outliers from the dataset since the distribution also varies. Figure 6 represents the uncleaned and cleaned graphs of the variables. Like in the other datasets, some outliers remained in the set as noise. Please review the references section for the dataset variables' uncleaned and cleaned plots. The cleaned dataset has around 804 rows which is slightly lower than the original uncleaned dataset.



(a) Alcohol Uncleaned Dataset

(b) Alcohol Cleaned Dataset

Figure 6: Alcohol Dataset

News API Text Dataset

API data was gained through News API regarding current events on red wine. The dataset presented with 9 column variables. Those are X, source, author, title, description, url, urlToImage, publishedAt, and content. Based on examining the dataset, variables X, source, url, urlToImage, publishedAt, author, and content were removed from the dataset because they all contained insignificant information. The content variable had unreadable symbols, which was the basis for removing it from the dataset. Personal identifiers such as author were removed. The remaining variables are the title and description. News API dataset contains around 300 rows of data which is relatively large. Since the text data is from news resources, the assumption was made that no misspelled words were contained in the set. The dataset contained capital letters. As a result, all text data was converted to lowercase letters. Figure 7 is a comparison of uncleaned and cleaned data.

X	source	author	title	description	url	
0	{'id': None, 'name': 'Lifehacker.com'}	Claire Lower	Your BLT Needs This Stealth Bacon	A BLT is a pretty perfect sandwich. I don't thing I need to explain why the combination of crispy, salty bacon, crunchy lettuce, creamy mayo, and juicy tomatoes taste good together. Each component has a part to play, and they play it oh-so-well, but I think t...	https://lifehacker.com/your-blt-needs-this-stealth-bacon-1849448203	media/image/upload/c_fill,f_auto,fl_progressive,g_c

(a) Uncleaned API Dataset

title	description
your blt needs this stealth bacon	a blt is a pretty perfect sandwich. i don't thing i need to explain why the combination of crispy, salty bacon, crunchy lettuce, creamy mayo, and juicy tomatoes taste good together. each component has a part to play, and they play it oh-so-well, but i think t...
8 italian villages you may have never heard of	travelers tend to head to the big name destinations like venice, florence, rome and even naples when visiting italy. but there are many equally, if not more, beautiful villages in the european country that are largely unknown, even among some italians.
opinion: new yorkers sent an election message. and it wasn't about inflation	it's becomingly increasingly clear that when abortion is on the ballot, abortion rights win. that's one lesson from this week's special election in a swing district of new york's hudson valley, in which pro-choice democrat pat ryan beat republican marc molina...
best places to go for fall: a dozen cool spots around the world	from fall classics such as bavaria to vibrant cities like buenos aires, here are a dozen places around the world that would make an excellent autumn getaway.
pioneering japanese woman photojournalist dies, aged 107 - reuters	tsuneko sasamoto, a pioneering photographer considered to be japan's first female photojournalist, who swore by a glass of red wine each night as one of the keys to good health, has died at the age of 107, domestic media said on monday.
five underrated french wines – and the overrated to avoid | fiona beckett on drinks	france makes some of the world's best wines, as well as some of the most over-rated. but you can you still nab a bargain if you shop wiselyif you've just come back from a holiday in france, you may well be basking in the glow of all things french, particulari...

(b) Cleaned API Dataset

Figure 7: APIDataset

2.2 Results

Through machine learning, information about consumer satisfaction, the chemical makeup of good quality wine, and words associated with good wine can be explored. The Wine Enthusiast dataset can bring insight into the relationships between wine ratings and descriptions. Certain words associated with high-rated wines can help characterize good quality wine and wine that consumers should seek. The Chemical Analysis dataset will allow the exploration of chemical compounds in wines and how they vary from bottle to bottle. A simple correlation plot between the variables in Chemical Analysis was performed. Figure 8 represents the variable correlation plot. The plot shows negative and positive correlations, which can help identify relationships between the compounds created during the production process. This information can be used to learn about the wine production process and optimization of congeners. The Portugal Wine dataset will help characterize what chemical compounds cause consumers to give high satisfaction ratings. Since this dataset provides the chemical composition of Portuguese wine with quality, the question of what chemical composition makes a good quality wine can be explored. Lastly, the API data will list words associated with wine in current news. Based on this data, the public attitude towards wine can be analyzed.

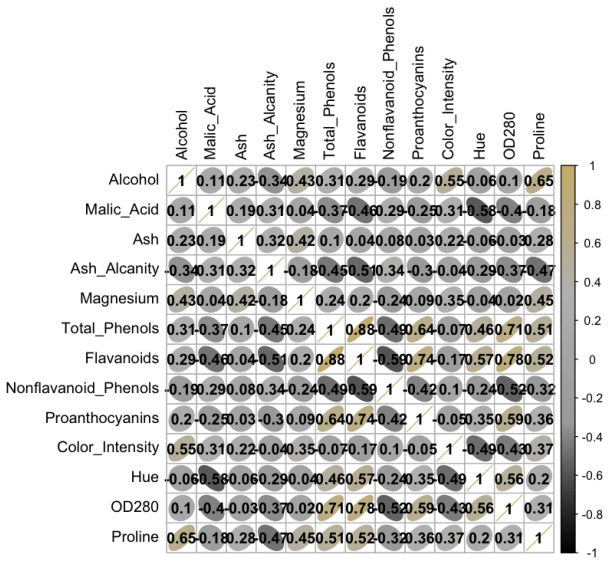


Figure 8: Wine Enthusiast Cleaned Dataset

3 Unsupervised Learning: Clustering and Association Rule Mining

3.1 Analyses

3.1.1 Clustering

Data clustering includes three different methods: K Means Clustering, Density-Based Spatial Clustering, and Hierarchical Clustering. Clustering methods will be performed on the Portugal Wine and Chemical Analysis datasets. The Portugal Wine dataset includes 11 variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, density, pH, sulfates, alcohol, and quality. The quality variable is discrete, where consumers rate the wine from 1-10. After data cleaning, the mean of the quality variable was 5.631, maximum of 7 and a minimum of 4. In order to determine the clustering pattern of this dataset, the discrete variable was converted to labeled data. Wines with quality greater than or equal to 6 were determined as "good" wine, the quality between ranges of 6 and 4 was determined as "average" wine, and the quality of less than or equal to 4 was determined as "bad" wine. After converting the quality variable into labeled data, it was removed to perform Clustering. The Chemical Analysis dataset has 13 variables: Alcohol, Malic Acid, Ash, Ash Alcanity, Magnesium, Total Phenols, Flavanoids, Nonflavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280, and Proline. It is important to note that this dataset does not include labeled data; therefore, no additional steps were taken to prepare the dataset for Clustering.

K Means clustering, Density-Based Spatial Cluster, and Hierarchical Clustering are unsupervised methods that require unlabeled data; therefore, preparing the dataset before clustering is important. Clustering is used

to discover groups or clusters in the dataset. The user can determine variables and vectors in the unlabeled dataset if clusters are identified. K Means clustering is a partitioning clustering method where clusters are associated with a centroid. In every iteration, clusters are assigned to the closest centroid, making up the number of clusters, K. The goal of this method is to minimize the sum of distances with respect to the centroid. When implementing this in Python, users will plot the dataset and determine the number of clusters. It is important to note that K Means clustering requires the user to know the number of clusters.

Density Based clustering is a clustering method that partitions the dataset into clusters based on the density of regions. The dataset is converted into highly dense and low dense regions. Two variables, eps and minpts, are identified by the user. Eps defines the neighborhood around the individual data point. If two data points have a distance lower than the eps, they are considered to be in the same neighborhood. Next, minpts is the minimum number of data points in the neighborhood. In Density Based clustering, clusters can be divided into core and border points. The core point is where the data points are heavily clustered. The border point is the point that surrounds the core point. Additional noise and outlier points are also shown away from the border point. Density Based clustering does not require the user to know the number of clusters which is helpful when the dataset values are similar.

Hierarchical clustering can be divided into two types: Agglomerative and Divisive. Agglomerative clustering starts with a data point as an individual cluster, and then after each step, the closest pair of data points merge into the cluster. Divisive clustering is where all data points are considered as one cluster, and then after each step, the cluster is split into individual clusters based on distance. Unlike other clustering methods, Hierarchical clustering can be represented as a dendrogram where the distance between points is evaluated and represented in each cluster. Based on the dendrogram, users can identify the overall distance between each cluster and the number of clusters present. In Hierarchical clustering, linkages such as single, complete, and average can be employed in addition to a distance metric. Single linkage returns the minimum distance between two data points in two independent clusters. Complete linkage returns the maximum distance between each data point. Average linkage returns the average distance between pairs of data points in individual clusters.

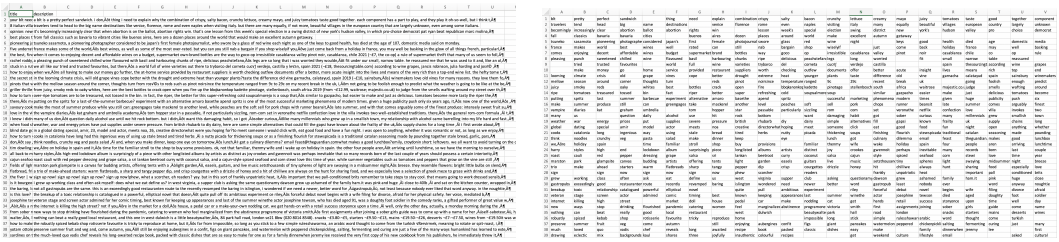
In all three clustering methods mentioned above, a distance metric is utilized to differentiate data points. Distance metrics are important in clustering because they identify clusters' similarity and dissimilarity. Three distance metrics will be employed in this model: Euclidean, Cosine, and Manhattan. The Euclidean distance is also identified as the Pythagorean Theorem's distance metric, which calculates the longest distance of a right triangle. Manhattan distance, also called "city block," calculates the distance between two points in a block-type manner. If we are given a right triangle, the distance is calculated by the sum of traveling along the x and y axes. The Cosine distance measures the angle between two vectors. Generally, it is favorable to use Manhattan distance if the dataset vectors are characterized as a grid. Euclidean distance is the default metric; however, the dimensionality of the data increases, which should be considered when deciding on a distance metric. Lastly, Cosine distance is utilized to better identify similarities between two data points; however, it also leads to higher dimensionality of the dataset.

The Portugal Wine dataset is expected to see three different clusters since the labeled data can be divided into three categories. However, the clustering behavior for Chemical Analysis is hard to determine since chemical concentrations can drastically vary depending on the bottle.

3.1.2 Association Rule Mining

Association Rule Mining (ARM) is an unsupervised learning method that utilizes transactional text data. The ARM procedure will observe frequently occurring patterns and associations in the dataset to form rules. Rules can be formatted into an if-then statement where dependencies between items in the dataset are identified. The most famous example is the Market Basket example. A rule for the Market Basket would be that if a consumer purchases milk and bread, then the probability of purchasing coke is high. In forming these rules, the Association Rule Mining algorithm utilizes three parameters: Support, Confidence, and Lift. Support indicates how frequently the rule appears in the database. Confidence is how often the rules are found to be true. Lift is the ratio of confidence in the rule over the expected confidence of the rule. Generally, a lift of greater than one is desired as it indicates positive associations between the elements. A confidence value greater than 0.8 is desired since it indicates that the rule is valid and frequently occurs in the dataset.

This model will utilize the API data gained from News API. It is important to note that the API data gathered includes chunks of descriptions; therefore, it is important to find a way to split the sentences into individual words while filtering out stop words. Please refer to the R code generated to find a way to transfer chunks of code to transactional data. The descriptions were processed so that it was all in lowercase, numerics were removed, punctuations were removed, and stop words were removed. Figure 9 is a comparison of precleaned data with cleaned transactional data. Even after running the code and generating transactional data, some stop phrases were kept in the dataset. Since we are using news data, the phrase "continue reading" was kept in the dataset. Those types of phrases were removed to avoid inaccurate rules. In addition, stop words with unknown symbols were removed from the transactional dataset to improve readability.



(a) (b)

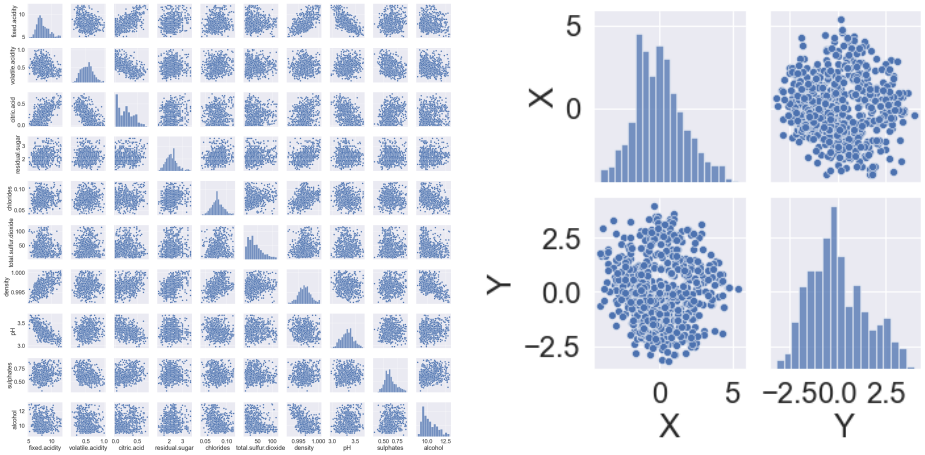
Figure 9: News API Data to Transactional Data

3.2 Results

3.2.1 Clustering

Portugal Wine Dataset

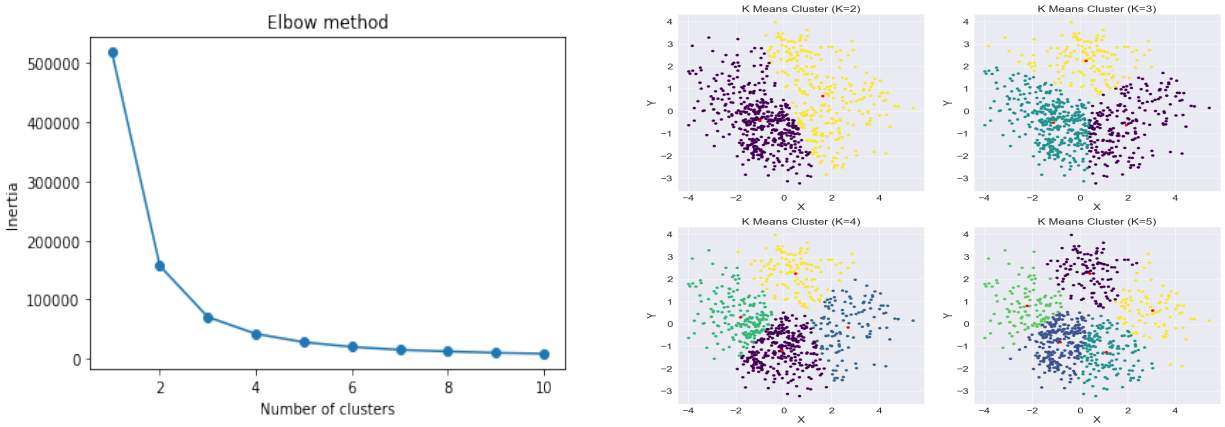
The Portugal Wine dataset is a high-dimension dataset with 11 variables. The dataset was prepared for clustering by scaling. Principle Component Analysis with a dimension of two was employed to reduce the dimensions, making it easier to visualize and perform clustering. The original dataset was explored using a pairwise plot. Figure 10a represents the pairwise plot of the original dataset. Most pairwise relationships show heavy clustering in the center and some randomness in the dataset. Figure 10b shows a normal distribution of datapoint in the center after performing PCA and scaling.



(a) (b)

Figure 10: Portugal Wine Dataset Initial Exploration

K Means clustering will be performed using K=2, K=3, K=4, and K=5. In order to determine the optimized K value, the Elbow method will be employed to compare the inertia at varying K values.The K Means clustering models will also be compared using the Silhouette Score. Comparisons of the K Means cluster with varying K values are shown in figure 11b. It is important to note that the Euclidean distance was utilized. In the figure, the red dot indicates the centroid of each cluster. The Silhouette score indicates how well the dataset clustered with respect to distance. This score ranges between -1 and 1, where -1 through 0 represent poor clustering. Any Silhouette value close to 1 represents good clustering. In this comparison model, the Silhouette score using K=2 is 0.380, K=3 is 0.397, K=4 is 0.382, and K=5 is 0.371. Comparing these four K Means clustering models, the model with K=3 is the better model since the Silhouette score is closer to one. The Elbow plot backs up this fact. Figure 11a shows the inertia at different numbers of clusters. The optimized number of clusters is determined where non-linearity starts to occur. K values from one to three show decreasing linearity. However, after three clusters, the graph starts to display nonlinear behavior. As a result, the elbow method indicates that the optimized number of clusters is three.

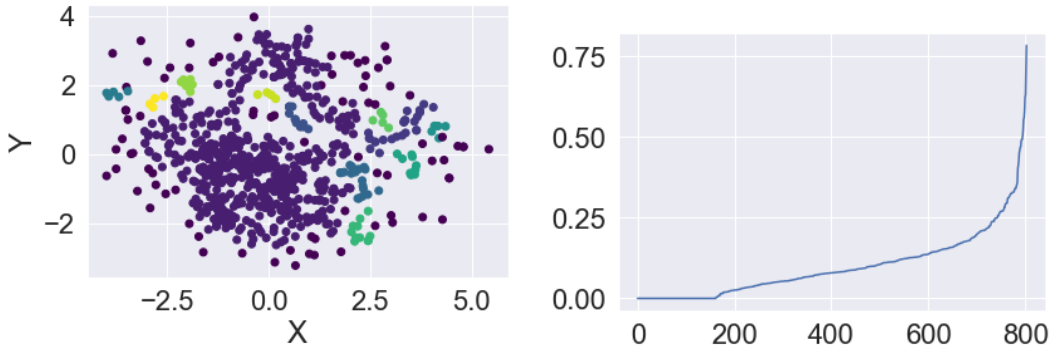


(a) Portugal Wine Elbow Method

(b) Portugal Wine K Means Plot

Figure 11: Portugal Wine K Means Optimization

Density Based Clustering does not require the user to input a number of clusters, but optimizing the eps and minpts values is important. The min pts value, in this case, will be four since the dimension of the dataset is two. The eps value was optimized by plotting the nearest neighbor curve, represented in figure 12b. When determining the optimized eps value, it is important to identify the maximum curvature value. In this case, the optimized eps value for the Portugal Wine dataset is 0.3. Figure 12a shows the result of Density-Based Clustering. The purple represents the core point where the dataset is heavily dense. The colors outside the core point could be the boundary point and noise from the dataset.

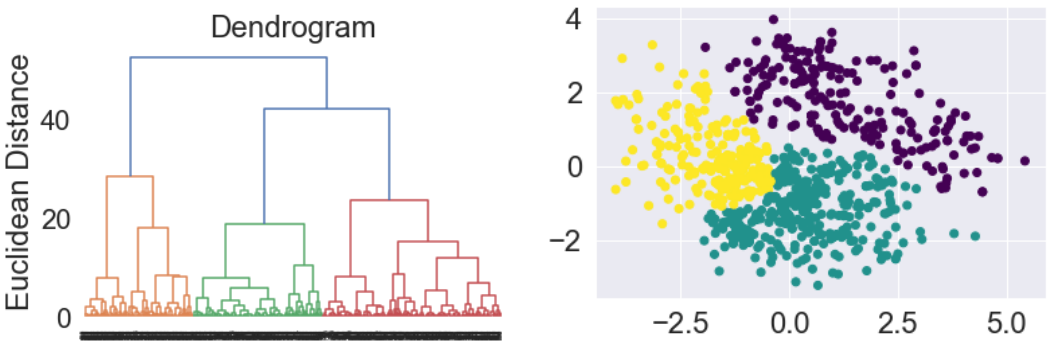


(a)

(b)

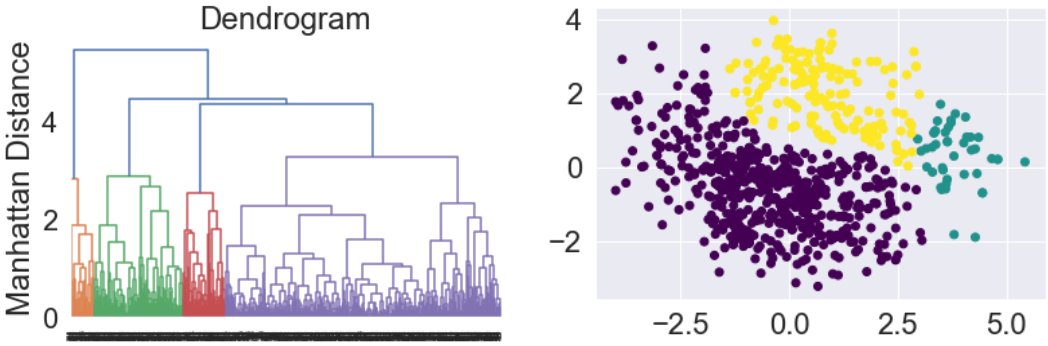
Figure 12: Portugal Wine Density Based Clustering

Hierarchical clustering was performed with three different distance metrics and Agglomerative clustering. Based on the Elbow method, the optimized number of clusters is three; therefore, three clusters were utilized with Agglomerative clustering. The Euclidean Agglomerative clustering and dendrogram are represented in figure 6. When running this model, the ward linkage method was utilized. It is important to note that the default linkage setting is ward. Based on the Euclidean dendrogram, the longest Euclidean distance starts after 30, indicating that three clusters are ideal for this distance metric. The second Hierarchical model utilizes the Manhattan distance. As seen in figure 14a, the dendrogram is skewed to the left. Unlike the Euclidean distance, the Manhattan distance is shorter near the top, making it harder to differentiate the optimum number of clusters. The Agglomerative cluster in figure 14b shows heavy clustering in the purple region, which is not seen in K Means clustering. The third Hierarchical model utilizes the Cosine distance. As seen in figure 15a, the dendrogram shows that the distance is long after Cosine distance of 0.5. Thus, the optimum number of clusters is also three. The Agglomerative cluster using Cosine distance is similar to Euclidean distance and K means because the clusters are equally divided into three clusters.



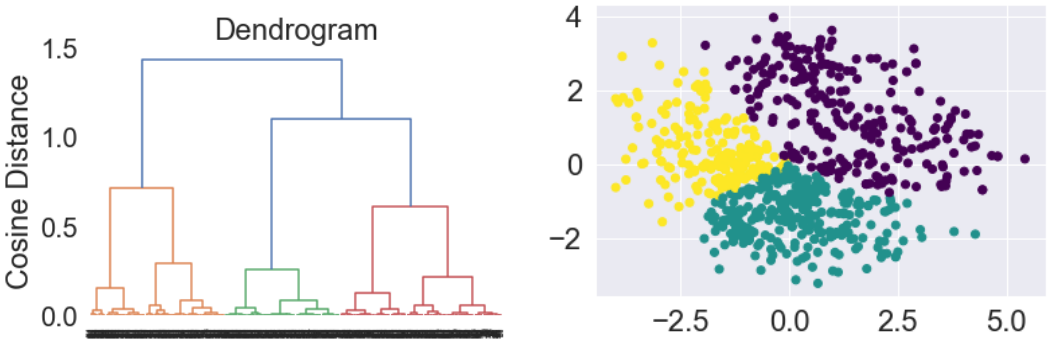
(a) (b)

Figure 13: Portugal Wine Dataset: Euclidean Distance



(a) (b)

Figure 14: Portugal Wine Dataset: Manhattan Distance

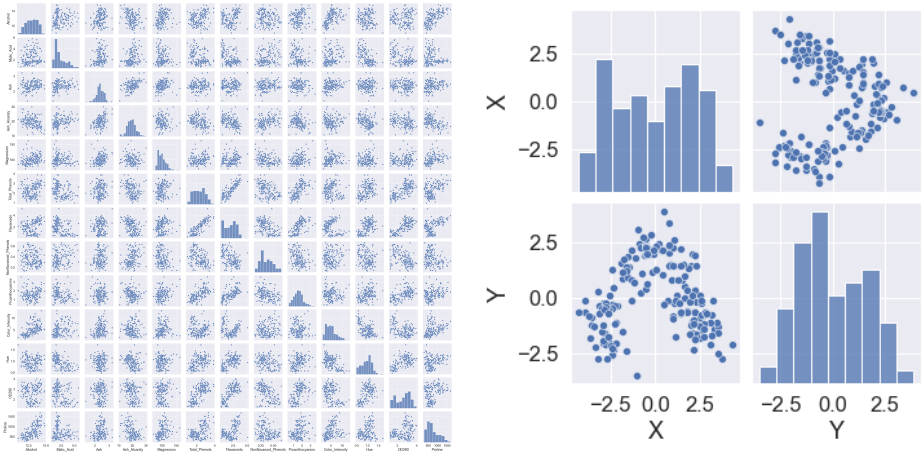


(a) (b)

Figure 15: Portugal Wine Dataset: Cosine Distance

Chemical Analysis Dataset

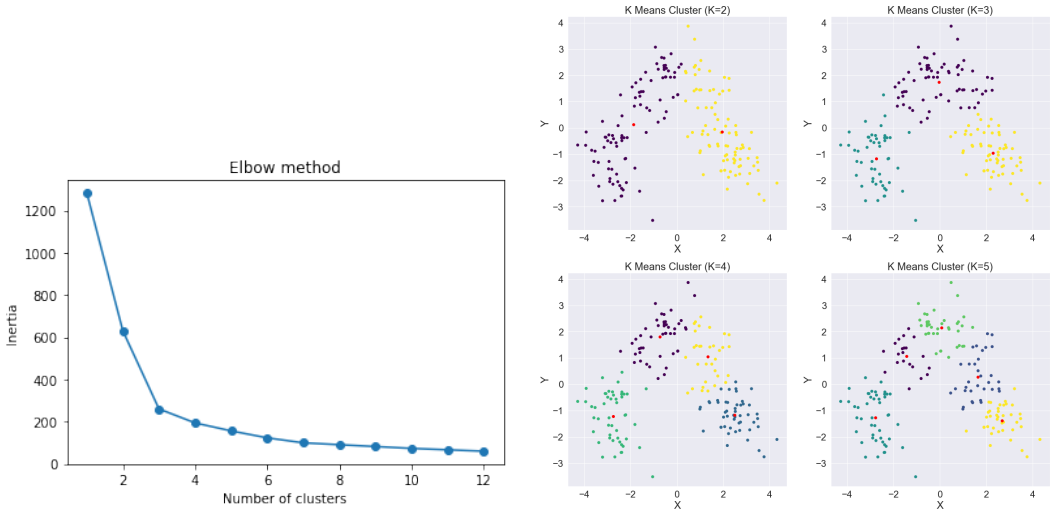
The Chemical Analysis dataset is a high-dimension dataset with 13 variables. The dataset was also prepared for clustering by scaling. Principle Component Analysis with a dimension of two was employed to reduce the dimensions, making it easier to visualize and perform clustering. The original dataset was explored using a pairwise plot. Figure 16b shows a pairwise plot of the transformed dataset. There is a pattern in this plot because it shows some curvature. K Means clustering will be performed using K=2, K=3, K=4, and K=5. The



(a) (b)

Figure 16: Chemical Analysis Dataset Initial Exploration

Elbow method will be employed to compare the inertia at varying K values. The K Means clustering models will also be compared using the Silhouette Score. Comparisons of the K Means cluster with varying K values are shown in figure 17b. In the figure, the red dot indicates the centroid of each cluster. The Silhouette score indicates how well the dataset clustered with respect to distance. In this comparison model, the Silhouette score using K=2 is 0.465, K=3 is 0.560, K=4 is 0.491, and K=5 is 0.441. Comparing these four K Means clustering models, K=3 is the better model since the Silhouette score is closer to one. Figure 17a shows the inertia at different numbers of clusters. Similar to the Silhouette score, the Elbow method also indicates that the optimized number of clusters is three.

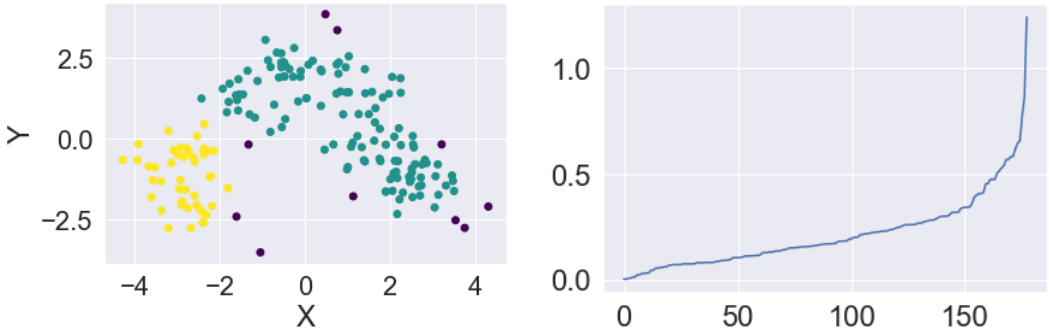


(a) (b)

Figure 17: Chemical Analysis: Elbow Method and K Means Clustering

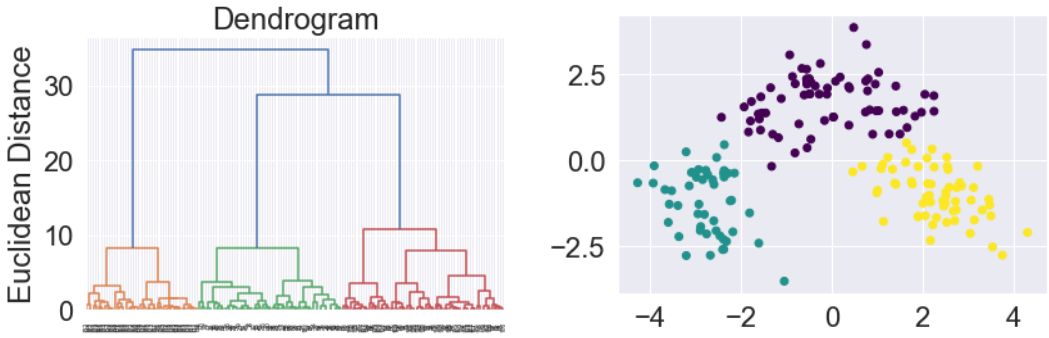
The min pts value for Density-Based clustering will be four since the dimension of the dataset is two. The eps value was optimized by plotting the nearest neighbor curve, represented in figure 18b. In this case, the optimized eps value for the Chemical Analysis dataset is 0.6. Figure 18a shows the result of Density-Based clustering.

The blue represents the core point where the dataset is heavily dense. The yellow represents another cluster in the dataset. The purple can represent boundary point and noise. Hierarchical Clustering was performed with three different distance metrics and Agglomerative clustering. Based on the Elbow method, the optimized number of clusters is three; therefore, three clusters were utilized with Agglomerative clustering. The Euclidean Agglomerative clustering and dendrogram are represented in figure 19. Based on the Euclidean dendrogram, the longest Euclidean distance starts after 10, indicating that three clusters are ideal for this distance metric. The second Hierarchical model utilizes the Manhattan distance. The Agglomerative cluster in figure 20 shows equal clustering similar to the K Means model. The third Hierarchical model utilizes the Cosine distance. As seen in figure 21, the dendrogram shows that the distance becomes longer after Cosine distance of 0.5. Thus, the optimum number of clusters is also three. The Agglomerative clusters show similar results to the K Means model in all three Hierarchical models. There is equal density in all three clusters.



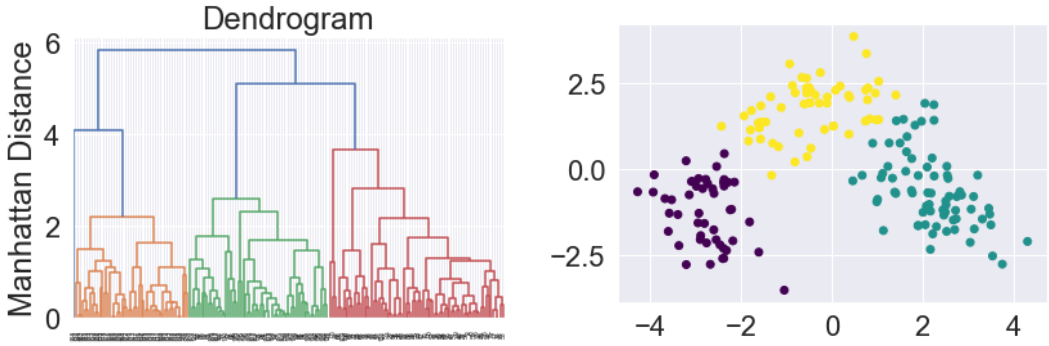
(a) (b)

Figure 18: Chemical Analysis Dataset Density Based Clustering



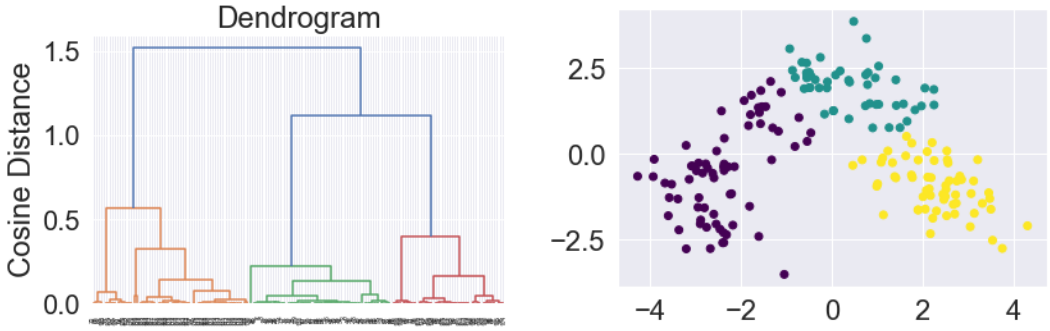
(a) (b)

Figure 19: Chemical Analysis Dataset: Euclidean Distance



(a) (b)

Figure 20: Chemical Analysis Dataset: Manhattan Distance



(a) (b)

Figure 21: Chemical Analysis Dataset: Cosine Distance

Based on all three clustering methods, it is clear that the Chemical Analysis dataset can be divided into three clusters. When comparing the Hierarchical clustering method, it should be noted that the Manhattan distance metric does not perform well compared to other metrics.

3.2.2 Association Rule Mining

Association Rule Mining was performed using the Apriori algorithm in R. The support was set to 0.01, and the confidence was 0.7. Anything with a confidence value greater than 70% was included in the rules. No rules were found when changing the support variable to greater than 0.02; therefore, the support was set to 0.01. Figure 15 and 16 represents the top 20 support, confidence, and lift rules. When examining the lift report, the greatest lift is the rule pinot with noir and noir with pinot. This is an obvious association since pinot noir is a type of wine.

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{started}	=>	{now}	0.01742160	1.00	0.01742160	20.500000	5
[2]	{york}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[3]	{avoid}	=>	{want}	0.01045296	1.00	0.01045296	28.700000	3
[4]	{dragon}	=>	{house}	0.01045296	1.00	0.01045296	47.833333	3
[5]	{varietals}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[6]	{cookbook}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[7]	{grapes}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[8]	{white}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[9]	{tasting}	=>	{wine}	0.01045296	0.75	0.01393728	5.817568	3
[10]	{impressive}	=>	{make}	0.01045296	1.00	0.01045296	16.882353	3
[11]	{pasta}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[12]	{travelers}	=>	{many}	0.01045296	1.00	0.01045296	19.133333	3
[13]	{noir}	=>	{pinot}	0.01045296	1.00	0.01045296	71.750000	3
[14]	{pinot}	=>	{noir}	0.01045296	0.75	0.01393728	71.750000	3
[15]	{fresh}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[16]	{late}	=>	{summer}	0.01045296	0.75	0.01393728	14.350000	3
[17]	{lunch}	=>	{dinner}	0.01045296	1.00	0.01045296	41.000000	3
[18]	{meal}	=>	{dinner}	0.01045296	0.75	0.01393728	30.750000	3
[19]	{started, wine}	=>	{now}	0.01045296	1.00	0.01045296	20.500000	3
[20]	{now, wine}	=>	{started}	0.01045296	1.00	0.01045296	57.400000	3

(a)

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{noir}	=>	{pinot}	0.01045296	1.00	0.01045296	71.750000	3
[2]	{pinot}	=>	{noir}	0.01045296	0.75	0.01393728	71.750000	3
[3]	{now, wine}	=>	{started}	0.01045296	1.00	0.01045296	57.400000	3
[4]	{dragon}	=>	{house}	0.01045296	1.00	0.01045296	47.833333	3
[5]	{lunch}	=>	{dinner}	0.01045296	1.00	0.01045296	41.000000	3
[6]	{pasta}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[7]	{fresh}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[8]	{meal}	=>	{dinner}	0.01045296	0.75	0.01393728	30.750000	3
[9]	{avoid}	=>	{want}	0.01045296	1.00	0.01045296	28.700000	3
[10]	{started}	=>	{now}	0.01742160	1.00	0.01742160	20.500000	5
[11]	{started, wine}	=>	{now}	0.01045296	1.00	0.01045296	20.500000	3
[12]	{travelers}	=>	{many}	0.01045296	1.00	0.01045296	19.133333	3
[13]	{impressive}	=>	{make}	0.01045296	1.00	0.01045296	16.882353	3
[14]	{late}	=>	{summer}	0.01045296	0.75	0.01393728	14.350000	3
[15]	{york}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[16]	{varietals}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[17]	{cookbook}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[18]	{grapes}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[19]	{white}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[20]	{tasting}	=>	{wine}	0.01045296	0.75	0.01393728	5.817568	3

(b)

Figure 22: Support(a) and Lift (b)

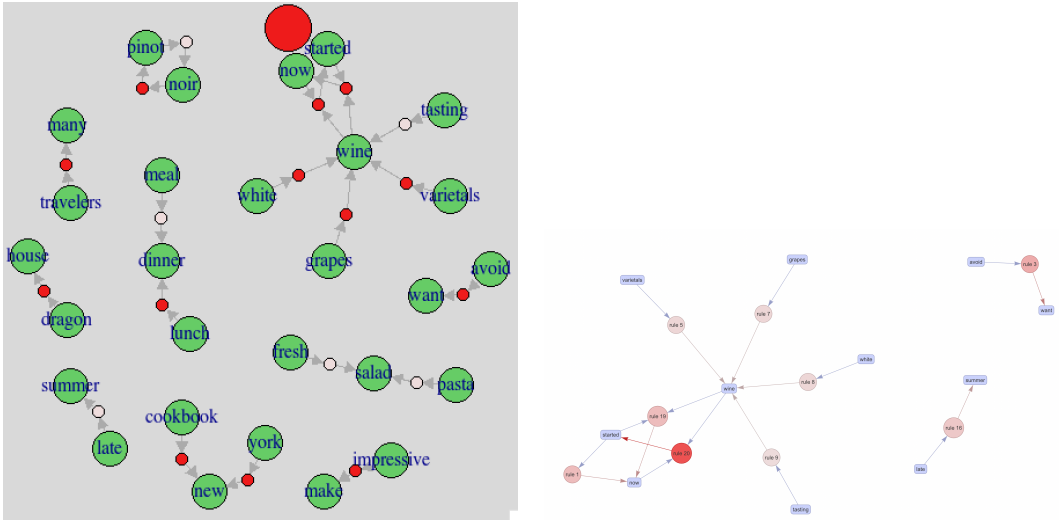
	lhs		rhs	support	confidence	coverage	lift	count
[1]	{york}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[2]	{avoid}	=>	{want}	0.01045296	1.00	0.01045296	28.700000	3
[3]	{dragon}	=>	{house}	0.01045296	1.00	0.01045296	47.833333	3
[4]	{varietals}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[5]	{cookbook}	=>	{new}	0.01045296	1.00	0.01045296	7.756757	3
[6]	{grapes}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[7]	{white}	=>	{wine}	0.01045296	1.00	0.01045296	7.756757	3
[8]	{impressive}	=>	{make}	0.01045296	1.00	0.01045296	16.882353	3
[9]	{travelers}	=>	{many}	0.01045296	1.00	0.01045296	19.133333	3
[10]	{noir}	=>	{pinot}	0.01045296	1.00	0.01045296	71.750000	3
[11]	{lunch}	=>	{dinner}	0.01045296	1.00	0.01045296	41.000000	3
[12]	{started}	=>	{now}	0.01742160	1.00	0.01742160	20.500000	5
[13]	{started, wine}	=>	{now}	0.01045296	1.00	0.01045296	20.500000	3
[14]	{now, wine}	=>	{started}	0.01045296	1.00	0.01045296	57.400000	3
[15]	{tasting}	=>	{wine}	0.01045296	0.75	0.01393728	5.817568	3
[16]	{pasta}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[17]	{pinot}	=>	{noir}	0.01045296	0.75	0.01393728	71.750000	3
[18]	{fresh}	=>	{salad}	0.01045296	0.75	0.01393728	30.750000	3
[19]	{late}	=>	{summer}	0.01045296	0.75	0.01393728	14.350000	3
[20]	{meal}	=>	{dinner}	0.01045296	0.75	0.01393728	30.750000	3

(a)

Figure 23: Confidence

Furthermore, it makes sense that pinot noir is associated with new articles because it is one of the most famous and popular wine varieties in the United States. As the lift decrease, food associations are observed. Foods such as pasta and salad are popular associations. When examining the support, the greatest support association is new with york and started with now. The support is a list of rules with the most frequent occurrences. In

this case, the most popular associations in news articles are started with now and new with york. Based on this result, it would be safe to assume that New York was highly associated with wines from August through September. When exploring the confidence list, the greatest confidence is new with york and varietal with wines. Like the support list, New York is also high on the confidence list. The second highest association is varietals and wine. This is also expected since wine types can be divided into different varietals. The News API data gained from August through September show high associations between wine and travel, food, New York, cookbook, tasting, white wine, and red wine.



(a) (b)

Figure 24: Overall Network (a) and "Wine" Network (b)

Figure 24 represents the network graphs of the 20 association rules generated by apriori and the network graph of the word wine. It is clear that wine is associated with white, tasting, varietal, and grapes. Those are expected associations with wine. In addition to the obvious associations, interesting associations include travelers and many.

4 Supervised Learning: Decision Trees, Naive Bayes, SVM

4.1 Analyses

4.1.1 Decision Trees

Decision tree is a popular classification method because it is easy to read and analyze. In addition, there is some familiarity associated with the tree-like model. Decision trees will take the dataset and create the root, internal, and terminal nodes. Like an actual tree, the terminal nodes are called leaves, and the internal nodes are called branches. Splitting the tree happens recursively until the tree displays all of the information and no impurities are present. Generally, splitting can happen using two criterion based methods: GINI or entropy. Both measure the amount of information gained during each split. The GINI impurity index formula is represented below, where p_i is the probability of being classified into a distinct class.

$$GINI = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{1}$$

GINI measures the frequency of a feature being mislabelled when chosen randomly. The criterion-based methods often describes tree splitting as pure or impure. When the GINI index is zero, the node is pure, indicating that the node contains all of the elements in a single class. The best splitting using GINI occurs when the probability of each class is the same. Generally, a GINI around 0.5 indicates good tree splitting where distinct classes are present. Entropy is the second criterion based method. The entropy formula is represented below, where p_i is the probability of class i.

$$Entropy = \sum_{i=1}^{C} -p_i * log_2(p_i) \tag{2}$$

The entropy definition is similar to the Boltzmann entropy equation ($S = klog(W)$), where comparisons of the amount of microstates in a macrostate are made, giving information about disorder and randomness. When

splitting a tree using entropy, it is ideal to find the number of splits with less disorder. In order to evaluate the quality of each split, the information gain will quantify the amount of features present after the split. Information gain can be calculated by the difference in entropy or GINI before and after the split. The equation of information gain is presented below.

$$\text{Information Gain} = \text{Entropy(before)} - \sum_{j=1}^{K} \text{Entropy(j,after)} \tag{3}$$

$$\text{Information Gain} = \text{GINI(Before)} - \sum_{j=1}^{K} \text{GINI(j,after)} \tag{4}$$

After generating a tree, pruning the tree is important to improve readability and eliminate unnecessary splits. Examining the theoretical maximum depth of a tree is one way of preventing the overfitting of the dataset and pruning the tree.

This paper will employ decision trees on the Portugal wine dataset. A total of three different trees will be present. The first tree will be generated using R and optimizing the complexity parameter. The second tree will be generated using Python. Entropy will be the main splitting source. The third tree will also be generated using Python and GINI as the main method of splitting. In all three cases, optimization of the trees will be made to produce a more accurate tree.

4.1.2 Naive Bayes

The basis of the naive Bayes classification is the Bayes Theorem. Bayes theorem calculates the probability of event A occurring given information B. This classification method is applied to a dataset to predict the probability of the label type given information about the features. The utilization of Bayes theorem is called "naive" because two assumptions are made. First, the features are independent. Second, the features are equal. This assumption is naive because it is unrealistic to assume that none of the features have a relationship and that all the features contribute equally to the response. Naive Bayes is formulated below, where $X = x_1, x_2, x_3....x_n$ represents the features of the dataset or called predictors. This statement is similar to equation six .

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)} \tag{5}$$

$$P(y|x_1, x_2....x_n) = \frac{P(x_1|y)P(x_2|y)....P(x_n|y)P(y)}{P(x_1)P(x_2)....P(x_n)} \tag{6}$$

Naive Bayes can be divided into three categories: Multinomial, Bernoulli, and Gaussian. Multinomial is generally used on discrete data or assigning text data to classes. The assumption is that the dataset follows a multinomial distribution. Bernoulli is similar to multinomial naive Bayes; however, it only predicts binary classes. Gaussian Naive Bayes assumes a continuous normal distribution. The predictors in a Gaussian distribution are continuous variables and cannot be discrete. One important feature of Naive Bayes is Laplace smoothing. This technique is utilized to avoid a zero probability and improve the fitting. The naive Bayes with Laplace is written below. Note that C is the number of classes. When implementing Laplace smoothing, the probability will never be zero because of the $+1$ and C terms.

$$P(A_i|C) = \frac{N_{ic} + 1}{N_c + C} \tag{7}$$

This paper will employ naive Bayes using the Portugal wine dataset and wine enthusiast dataset.

4.1.3 Support Vector Machines

Support vector machines transform the dataset into a higher dimensional space, so the features are classifiable by a hyperplane. Different kernels, such as linear, gamma, sigmoidal, radial, or polynomial, can be used to separate the dataset. Hyperplanes will decide the boundaries of the data points that can be classified into different groups. In addition to the hyperplane, a maximum margin is optimized to make predictions with higher accuracy. In other words, the margin is an area that considers the possible error associated with future prediction. Two types of margins can be generated: soft margin hyperplanes or hard margin hyperplanes. Soft margin hyperplanes occur when the constraint on maximizing the margin is less strict. Datasets that have some mixing with two classifications can be managed with a soft margin hyperplane. Hard margin hyperplanes are the opposite, where distinct boundaries are chosen with the margin and hyperplane. The dataset has clear classification boundaries, and no mixing between groups is present. The main goal of support vector machines is maximizing the margin, which is a quadratic convex optimization problem.

This paper will employ support vector machines using the Portugal wine dataset and chemical analysis dataset.

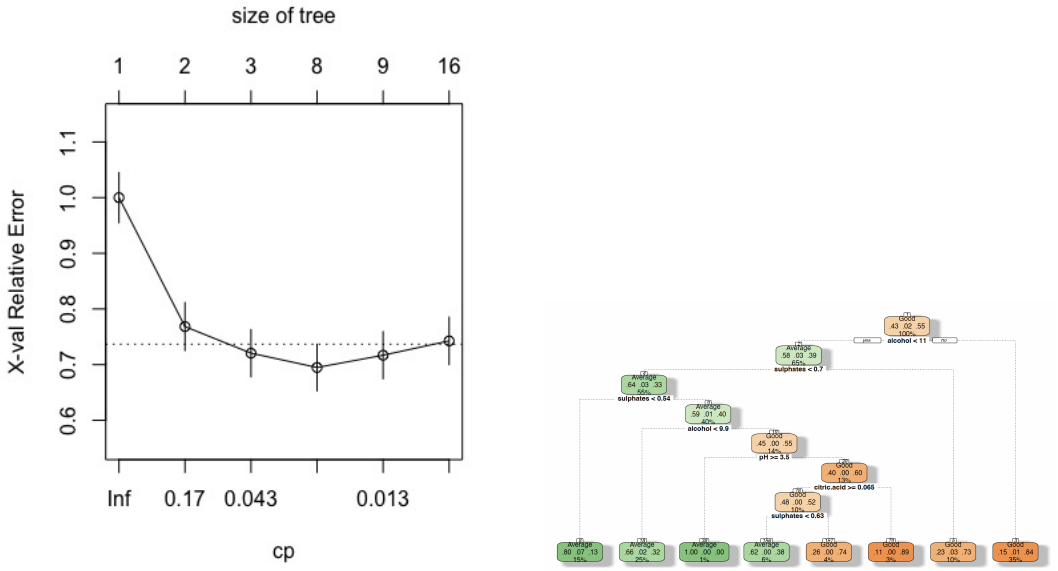
4.2 Results

4.2.1 Decision Trees

The Portugal wine dataset decision trees were generated using R and Python. The predictors of the dataset are the chemical properties that are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, density, pH, sulphates, and alcohol. The label data is going to be the quality. It is important to note that the label data is discretized into "bad", "average", and "good" quality wine. The dataset used in unsupervised learning methods was used in this case. The original decision trees produced using Python and R are relatively large. Therefore, images of those trees will be presented in the appendix section.

Supervised learning appendix section 1 represents an unpruned tree produced using R. The tree is relatively large, and it is hard to extract information. Furthermore, the label "bad" is not present in the tree. The lack of the label "bad" produces an unbalanced tree indicating that there could be some error during future predictions. Furthermore, the unbalanced tree could represent a lack of data. When summing the quality labels, there are 18 "Bad", 352 "Average", and 434 "Good" labels. Just examining these values, it is evident that there are more "Good" and "Average" data points. Thus, the tree will perform well in predicting future "Good" and "Average" wines. However, the lack of data points representing "Bad" wine provides limited training information for the decision tree algorithm. To mitigate this issue, the dataset can be sampled so that equal proportions of "bad", "good", and "average" labels are present. However, this is a hard task since the amount of "bad" data is very small, which could impact the overall learning process of the tree. As a result, the next time decision tree is performed on this dataset, it is important to set good boundary ranges for discretizing the labels.

Since the decision tree is relatively large, it is important to perform pruning to optimize the size. In R, the tree was pruned using the complexity parameter (CP).



(a) (b)

Figure 25: Complexity Parameter

Figure 25a represents the CP plot of the overall tree. When optimizing CP, choosing the CP value that produces the lowest error is key. The optimized CP value is 0.014706. After implementing the CP criterion, figure 25b shows the pruned tree. Comparing the original tree and the pruned tree, the size of the pruned tree is significantly smaller. A confusion matrix was calculated to evaluate the classification. The accuracy of the pruned tree is 63.16%, which is higher than the unpruned tree. Figure 26 is the confusion matrix.

Appendix section 3 and 5 represents the original GINI and entropy tree produced by Python. Similar to the results from R, the tree is large and has to be pruned. Pruning in Python was implemented by optimizing the GINI, entropy, and max depth. A for loop was implemented to test the accuracy associated with max depth ranges from one to thirty. The max depth with the highest accuracy was chosen to prune the tree. Figure 27 represents the GINI decision tree with a max depth of three. This significantly reduces the size of the tree with an accuracy of 73.9%. The accuracy of the GINI tree with a max depth of three is higher than the accuracy produced using the complexity parameter. Appendix section 6 represents the optimized entropy tree with a max depth of 27. The highest accuracy when testing the max depth of the entropy tree is 73.2%. The accuracy values produced by entropy and GINI are nearly the same. However, the tree produced by GINI requires a lower max depth which improves interpretation of the dataset.

	Average	Bad	Good
Average	64	4	32
Bad	0	0	0
Good	29	1	71

Figure 26: Complexity Parameter Optimization Confusion Matrix

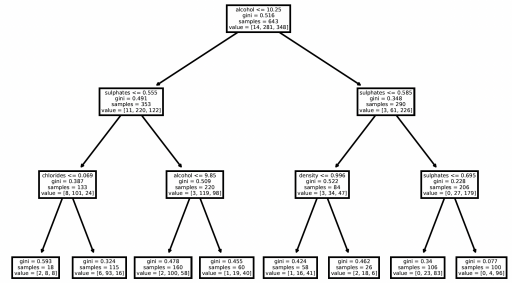


Figure 27: GINI Tree: Max Depth of Three

4.2.2 Naive Bayes

Naive Bayes was implemented on the Portugal wine dataset and the wine enthusiast dataset. It is important to note that the Portugal wine dataset is mixed data with some numerical and text data. The label in the Portugal wine dataset is quality, which is text data. The quality was discretized similarly to the dataset used for decision tree. Gaussian Naive Bayes was employed on the Portugal wine dataset. In order to prevent overfitting and a zero probability, Laplace smoothing was also included in the computation. The accuracy of this method was 64%. Figure 28 represents the confusion matrix of the Gaussian Naive Bayes. The same simulation was

	Average	Bad	Good
Average	65	5	37
Bad	2	0	1
Good	26	0	65

Figure 28: Confusion Matrix for Gaussian Naive Bayes (Portugal Wine Dataset)

run using Python. However, the quality column was converted to numeric values where 0 represents "bad", 1 represents "average", and 2 represents "good". The accuracy of the Gaussian Naive Bayes in Python is 67.7%, which is nearly the same as the accuracy produced using R. In order to further extend the applications of naive Bayes on this dataset, a multinomial naive Bayes was implemented in Python using the same dataset. The accuracy of this model is 59.6%, which is significantly lower than the accuracy produced using Gaussian Naive Bayes. As a result, when comparing the two naive Bayes methods, the Gaussian Naive Bayes method produced the highest accuracy rate. The Portugal wine dataset is relatively large; therefore, it is safe to assume a normal distribution.

Besides the Portugal wine dataset, multinomial naive Bayes was implemented on the wine enthusiast dataset. The wine enthusiast dataset is mixed data; however, most of it is text data. The dataset includes a description of the wine, country of origin, province of origin, price, grape varietal, and wine points. Text-based naive Bayes was run with the description as the predictor and points as the label. The first step in this process is to manage the text data. In this set, the description is presented in chunks of text. As a result, it is important to clean

the description by converting text into lowercase, removing numbers, stop words, punctuation, and any extra spaces. After this step is performed, the description is tokenized. The text-based multinomial naive Bayes will review the words produced in each description to predict the wine point. The accuracy of this model is around 43.77%. Since the accuracy is lower than 50%, multinomial naive Bayes is not a good way of classifying this dataset.

4.2.3 Support Vector Machines

Support vector machines were employed on the chemical analysis and Portugal wine datasets. Similarly to decision trees and naive Bayes, the label for the Portugal wine dataset is quality, which is represented in terms of discrete numerical values. Linear, radial, and the sigmoidal kernels were utilized on both datasets. The support vector machine never converged when running the linear kernel on the Portugal wine dataset. Furthermore, changing the cost variable from 10 to 20 didn't change the convergence pattern. Based on this result, it is safe to assume that the linear kernel is not adequate for classifying the Portugal wine dataset. The radial and sigmoidal kernels converged for the Portugal wine dataset using a cost parameter of 10. The accuracy using the radial kernel is 60.86%, and the sigmoidal kernel is 52.17%. Both kernels do not have a high accuracy rate. When the cost parameter of the radial kernel increased to 20, the accuracy of the model decreases slightly. Furthermore, when the sigmoidal kernel is used with the cost parameter at 20, the accuracy decreases to around 48.44%.

The chemical analysis dataset features are Malic Acid, Ash, Ash Alcanity, Magnesium, Total Phenols, Flavanoids, Nonflavanoid Phenols, Proanthocyanins, Color Intensity, Hue, OD280, and Proline. The label for this dataset is Alcohol. The label set was discretized where if the Alcohol level is greater than or equal to 13.4, it is classified as "High". If the alcohol level is less than 13.4, the wine is classified as "Average". Linear, sigmoidal, and radial kernels were utilized. The accuracy using the linear kernel with a cost parameter of one is 83.33%. As the cost parameter increases when using the linear kernel, the accuracy of the model decreases. The accuracy using the radial kernel with a cost parameter of 10 is 46.22%. When increasing the cost parameter to 20, the accuracy of the model increases to around 60%. The accuracy using the sigmoidal kernel with a cost parameter of 10 is 44.44%. When the cost parameter is increased to 20 for the sigmoidal kernel, the accuracy of the model jumps to around 50%. Considering this information, the linear kernel does a good job at classifying the chemical analysis dataset since it has the highest accuracy.

5 Artificial Neural Networks

5.1 Analysis

Neural networks generate an algorithm that imitates human neurons and the act of learning. Through trial and error, the neurons take in information to learn and predict labels. There are multiple varieties of neural network algorithms, such as Artificial Neural Networks, Convolutional Neural Networks, and Recurrent Neural Networks. All these algorithms seek to train the model through a training set and then predict new information provided. Neural network computation is based on a loss function. Generally, the MSE is utilized as the loss function, and the goal is to optimize the function. Neural networks will generate a weight and bias matrix representing each hidden later. Then, the gradient of the loss function is taken with respect to weight and bias. Depending on the epoch, the weight and bias are iteratively updated in the code. The final values are then utilized to compute the predicted values of the dataset. During the last stage, an activation function can transform the final prediction. Some examples of activation functions are sigmoidal, soft margin, logistic, linear, and ReLU. In this computation, backpropagation will be implemented to train the algorithm further. Backpropagation utilizes gradient descent or stochastic gradient descent to solve effectively for the weight and bias values. A Feed Forward Artificial Neural Network with one hidden layer will be applied to the Chemical Analysis dataset to predict the alcohol percentage in wine. Through this process, the goal is to visualize and compare the accuracy of the neural

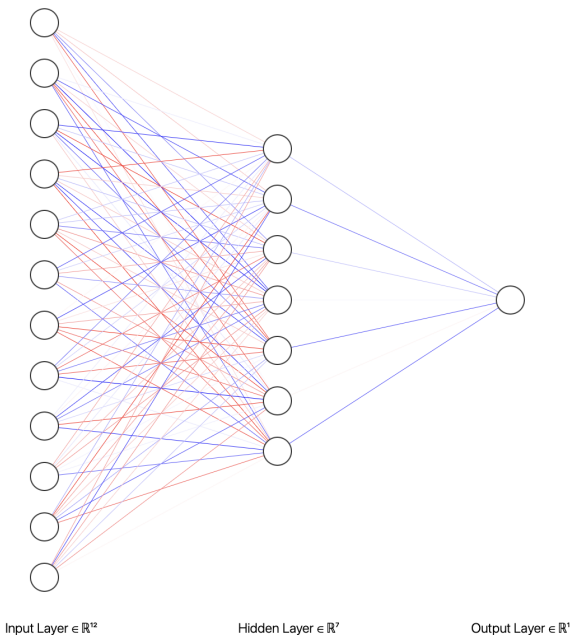


Figure 29: Wine Enthusiast Uncleaned Dataset

network model. The Chemical Analysis dataset includes 12 predictors: malic acid, ash, ash alcanity, magnesium, total phenols, flavonoids, non flavonoid phenols, proanthocyanins, color intensity, hue, OD280, and proline. The label data is going to be alcohol. Our neural network will have seven hidden units within the single hidden layer. Figure 29 shows the graphical representation of the neural network. The input layer represents the 12 predictors, the hidden layer shows seven units, and one output is predicted. It is important to note that a feed-forward neural network is direction sensitive and goes from left to right.

5.2 Results

The neural network was run with an epoch of 36. Figure 30 is the total loss vs. epochs plot. The plot indicates that as epochs increase, the total loss decreases. Any epoch value after 10 seems to decrease the total loss, which is the goal of neural networks. The accuracy of the neural network was calculated using a confusion matrix. The confusion matrix indicates that the neural network is 83.14% accurate. The neural network algorithm is relatively accurate and classifies the alcohol variable well. Figure 31 are snapshots of the last two epoch iterations.

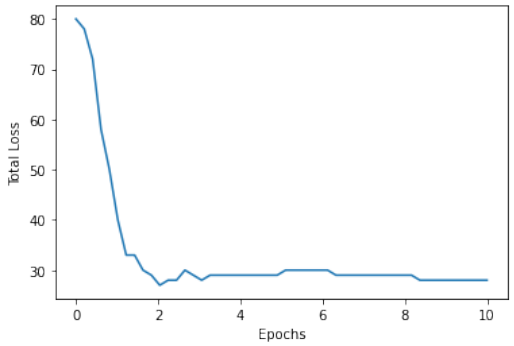
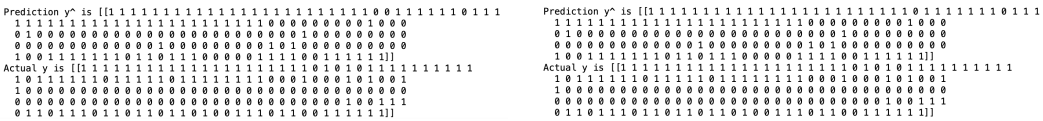


Figure 30: Total Loss and Epochs



(a) (b)

Figure 31: Neural Network Last Two Epoch Iterations

6 Conclusion

In this paper, unsupervised learning, supervised learning, and neural networks were explored and applied to wine-related datasets. Unsupervised learning, such as clustering and association rule mining, was implemented. Then, supervised learning such as naive Bayes, support vector machines, and decision trees were applied to labeled data. Lastly, a Feed Forward Artificial Neural Network (ANN) with a single hidden layer was utilized. It is important to note that all the collected datasets were processed and cleaned before these techniques were applied, which is key in improving the accuracy of the models generated.

Clustering was performed on the Portugal Wine and Chemical Analysis dataset. The first step was discretizing the quality variable and removing it since unsupervised learning requires unlabeled data. The quality was converted from numerical values to good, average, and bad wine. K Means, Hierarchical, and Density-based clustering were performed to compare the accuracy of each model. In all clustering models, three clusters were present in the Portugal Wine and Chemical Analysis datasets. These three different clusters could represent the "good", "bad", and "average" quality wines. It is important to note that the Chemical Analysis dataset only contains quantitative information about chemical compounds. Therefore, this clustering method highlighted that the dataset has three independent groups with similar chemical makeup. In clustering methods, distance is an important metric. Metrics such as Manhattan, Euclidean, and Cosine Similarity are methods of measuring distance and categorizing the data. In both datasets, the Euclidean distance performed consistently, and equal clustering was present. The results gained from the Portugal Wine and Chemical Analysis datasets revealed that wine quality depends on the chemical compound concentrations. Furthermore, it highlights the fact that there could be an optimum chemical combination and concentration that produces "good" wine.

The attitude of news outlets and the public toward wine was analyzed through Association Rule Mining. From August through September, the transactional data indicates that news outlets showed frequent relationships between wine and food, traveling, and varietals. Words such as varietal, grape, white, and tasting were frequently associated with wine, indicating the obvious associations with wine. Furthermore, cooking and food were commonly associated with wine, highlighting frequent associations between food and wine pairings. It is important to note that traveling and summer are themes from August through September, which could justify the frequent associations of "summer" words with wine. The Association Rule Mining results seem obvious since the associations discovered through this process are common relationships with wine.

Supervised learning methods such as support vector machines, naïve Bayes, and decisions trees were employed to explore the classification properties in the Portugal Wine, Chemical Analysis, and Wine Enthusiast datasets. When analyzing these Portugal Wine trees, the optimized GINI tree produced the best results. The size of the optimized GINI tree is relatively small, and the accuracy is greater than the trees produced when optimizing complexity parameters and entropy. The Portugal Wine dataset decision tree mainly consists of labels such as alcohol percent, sulfates, density, and chlorides. These features provide the most information for categorizing the dataset into "good", "average", and "bad" wine. When running the multinomial naive Bayes on the wine enthusiast dataset, the model's accuracy was lower than 50%, indicating that it would perform poorly during future classification. This result highlights that it might take much work to rank wine accurately by points given the description of the wine. Because taste is very biased, finding some relationship between wine description and points is challenging. When comparing the radial, sigmoidal, and linear kernels in support vector machines, the linear kernel performed the best for the chemical analysis dataset. This indicates that the chemical analysis dataset is linearly separable, and a linear relationship could represent the relationship between chemical composition and quality.

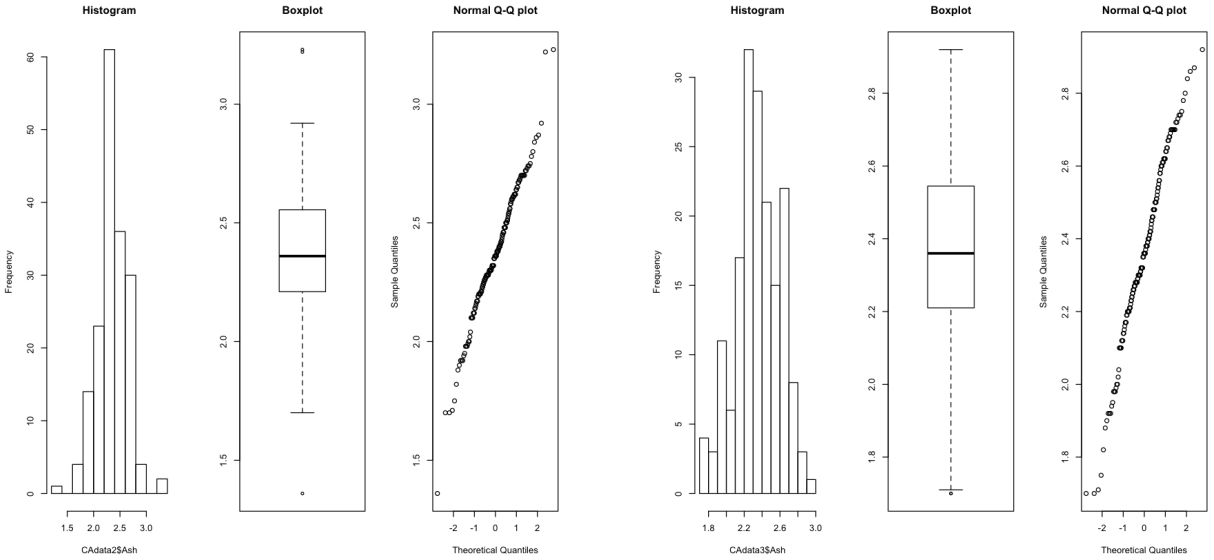
The goal of the neural network was to predict the alcohol level based on the chemical composition. Based on the high accuracy generated by the model, it is safe to state that the neural network does a good job of predicting the label data. Furthermore, the model can identify key input data patterns that produce high alcohol percentages. This result highlights that certain chemical combinations have a higher alcohol percentage. To further identify the reason for the high alcohol percentage, it would be important to investigate the production process of the wine.

By implementing machine learning techniques to wine-related datasets, it was possible to classify and generate models regarding chemical composition and quality. Furthermore, the wine-related associations in the news media indicated nothing unique. In other words, expected associations regarding wine are found in news media outlets. For future research, it would be important to explore white and red wine separately. Although this paper mainly focused on red wine, it would be interesting to see how dynamics would change when studying white wine. In the United States, wine quality is not regulated. However, many European countries, such as Italy, regulate wine quality by classifying them at different quality levels. Exploring and implementing learning methods on region-specific wine datasets would be another avenue to explore.

7 Appendix

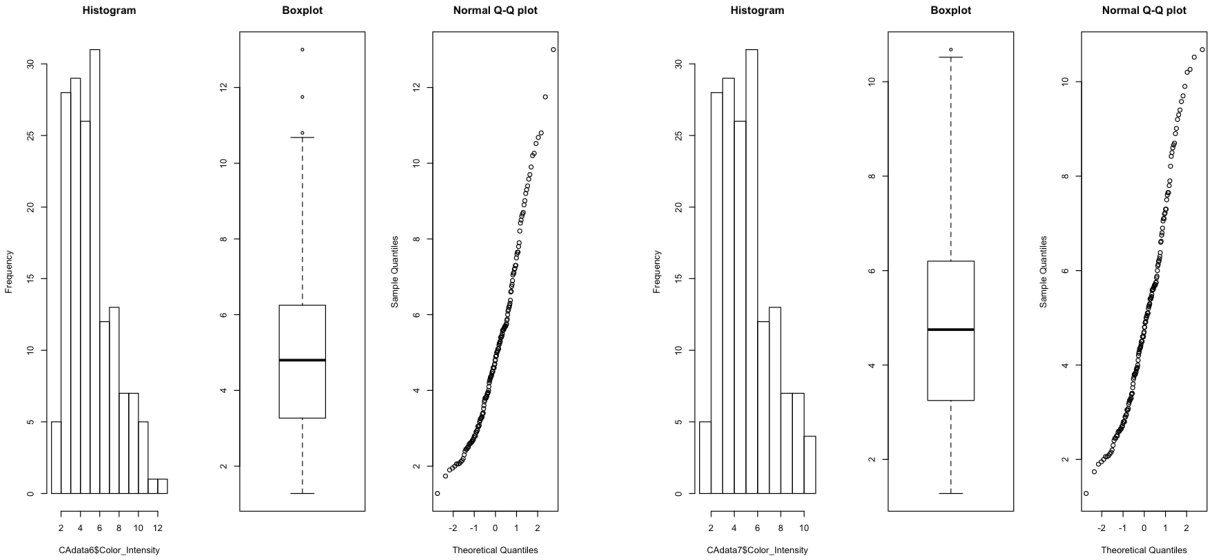
7.1 Data Cleaning and Visualization

Wine Enthusiast
https://www.kaggle.com/datasets/zynicide/wine-reviews
Chemical Analysis: Uncleaned and Cleaned Graphical Representation
https://archive.ics.uci.edu/ml/datasets/wine



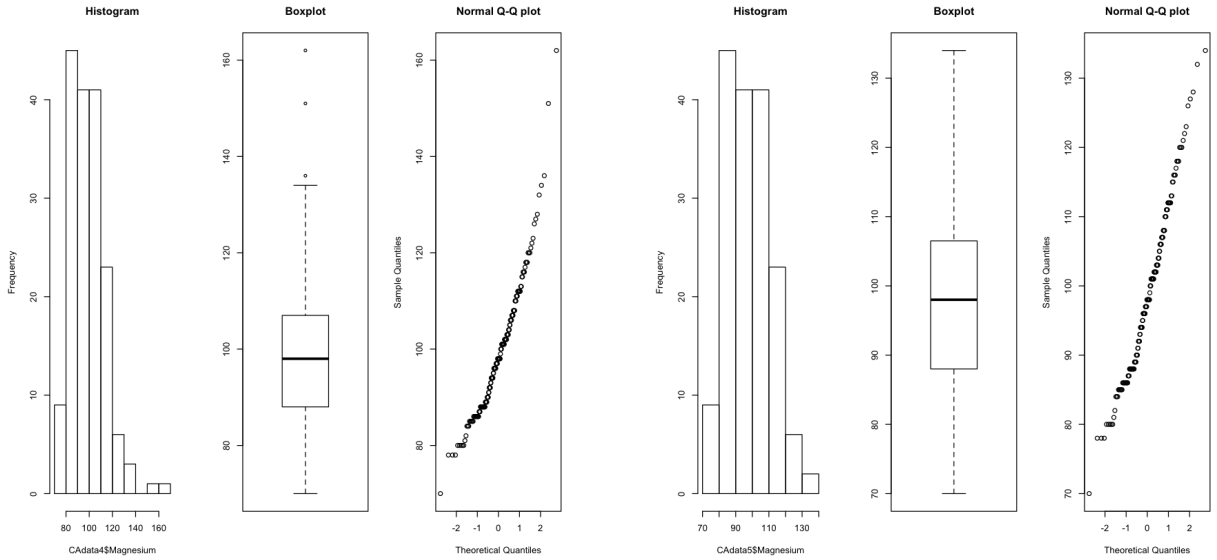
(a) Ash Uncleaned Dataset (b) Ash Cleaned Dataset

Figure 32: Ash Dataset



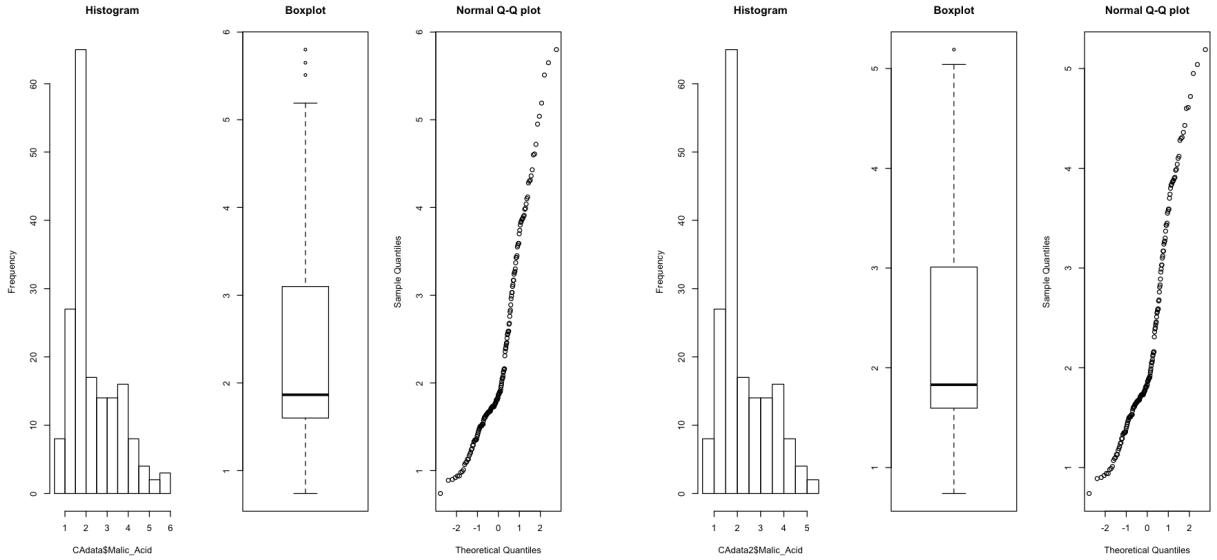
(a) Color Intensity Uncleaned Dataset (b) Color Intensity Cleaned Dataset

Figure 33: Color Intensity Dataset



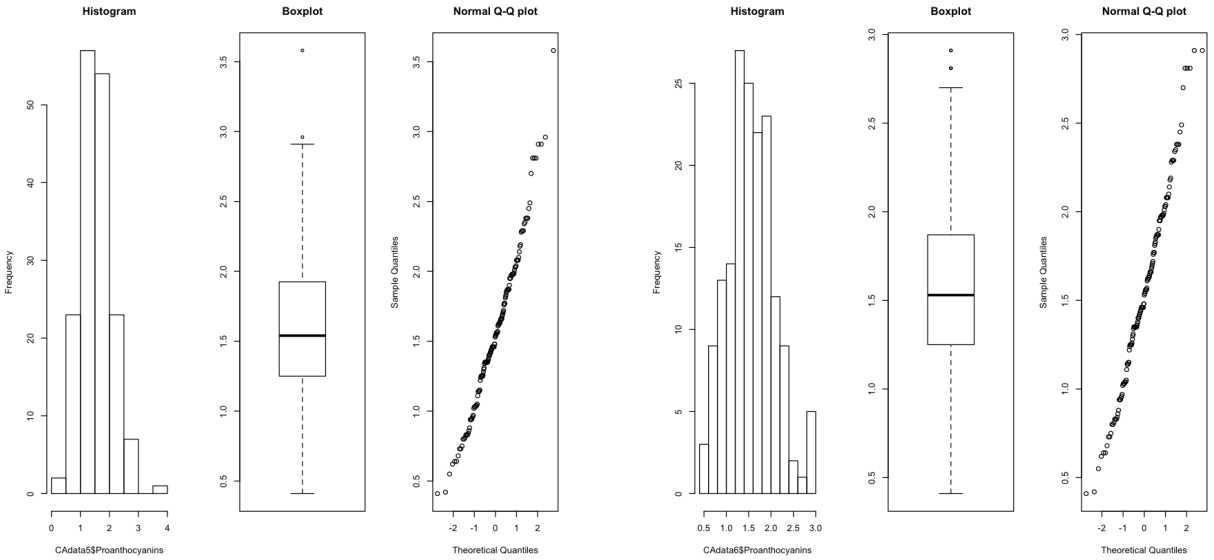
(a) Magnesium Uncleaned Dataset (b) Magnesium Cleaned Dataset

Figure 34: Magnesium Dataset



(a) Malic Acid Uncleaned Dataset (b) Malic Acid Cleaned Dataset

Figure 35: Malic Acid Dataset

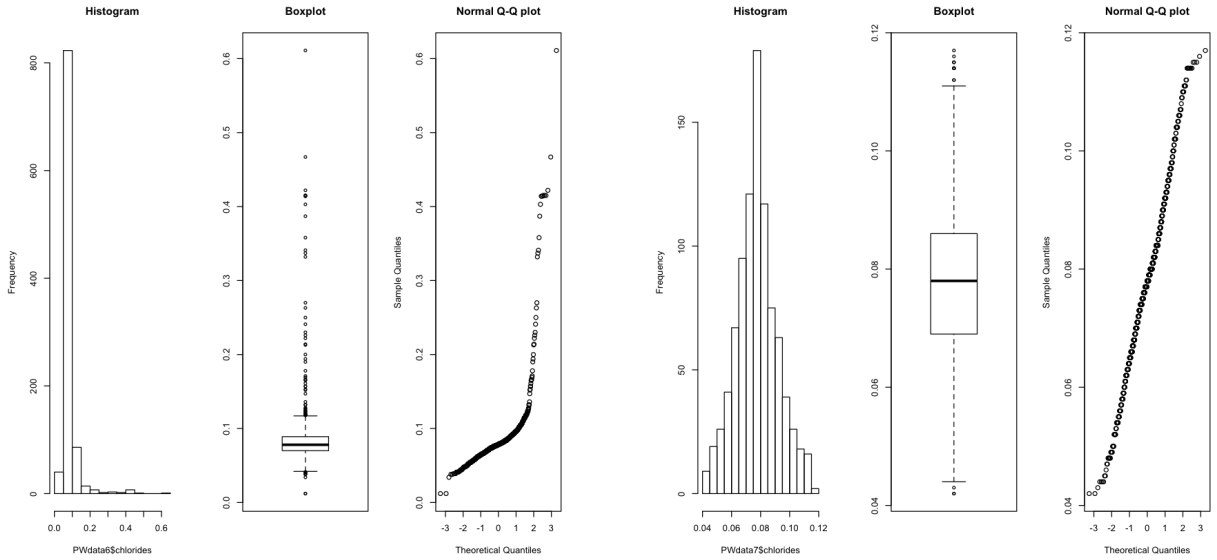


(a) Proanthocyanins Uncleaned Dataset

(b) Proanthocyanins Cleaned Dataset

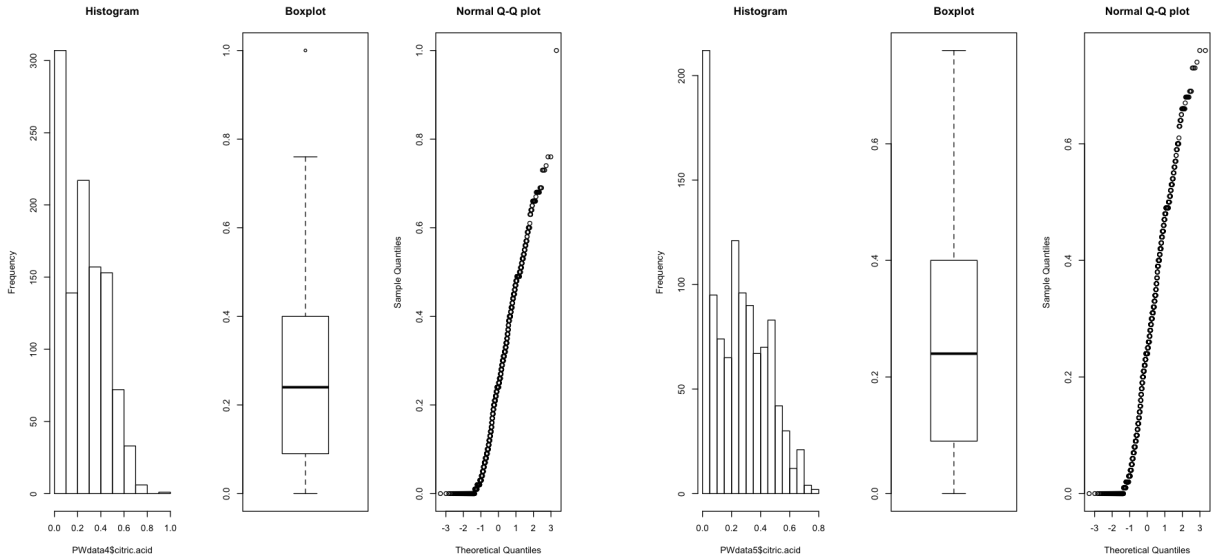
Figure 36: Proanthocyanins Dataset

Portugal Wine: Uncleaned and Cleaned Graphical Representation
https://archive.ics.uci.edu/ml/datasets/wine+quality



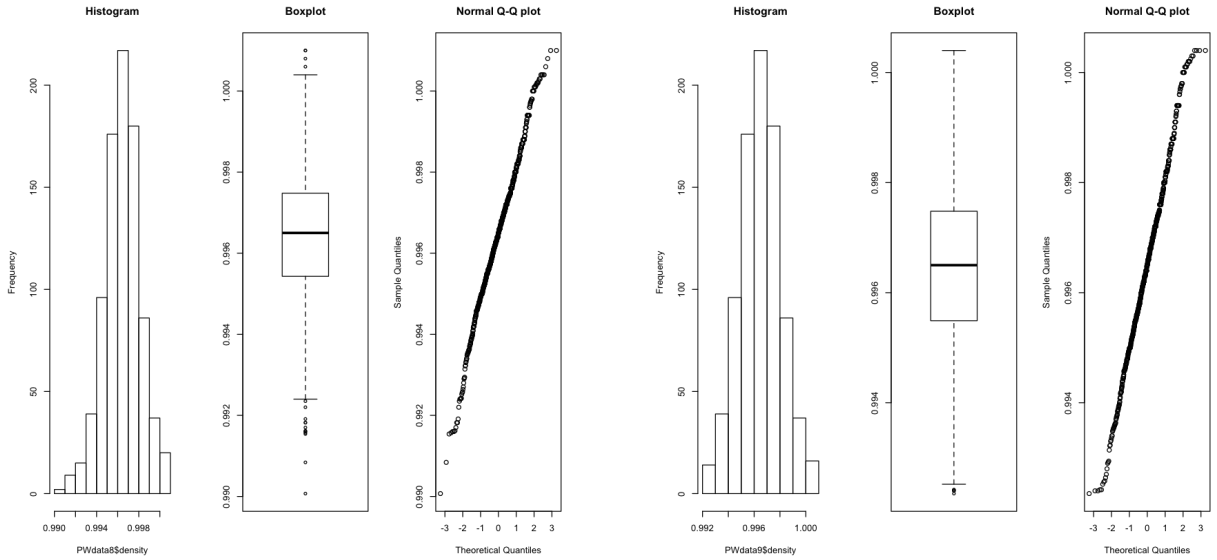
(a) Chlorides Uncleaned Dataset (b) Chlorides Cleaned Dataset

Figure 37: Chlorides Dataset



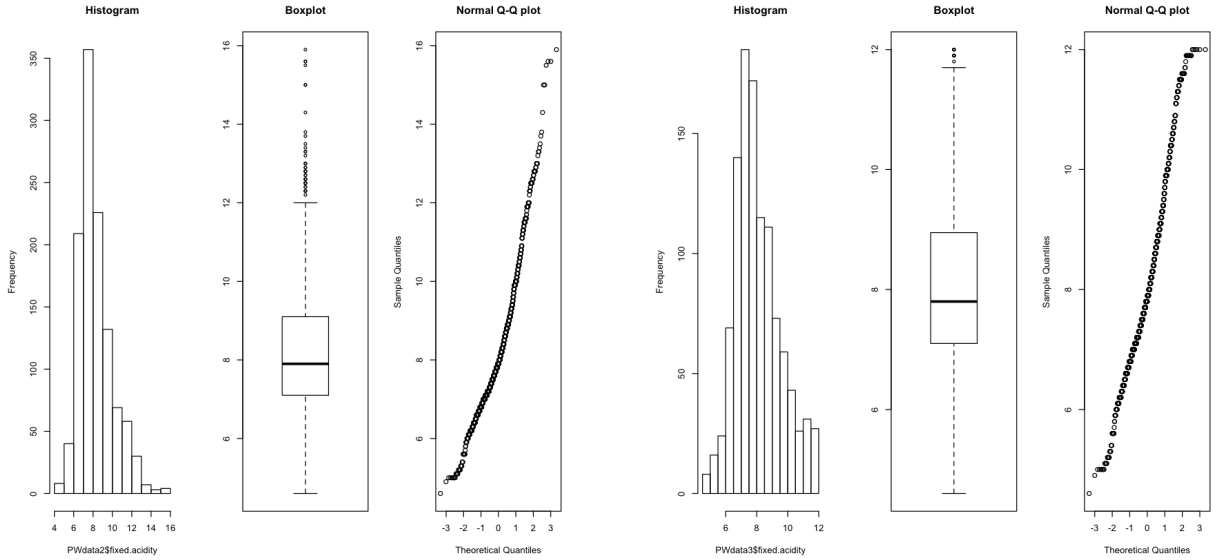
(a) Citric Acid Uncleaned Dataset (b) Citric Acid Cleaned Dataset

Figure 38: Citric Acid Dataset



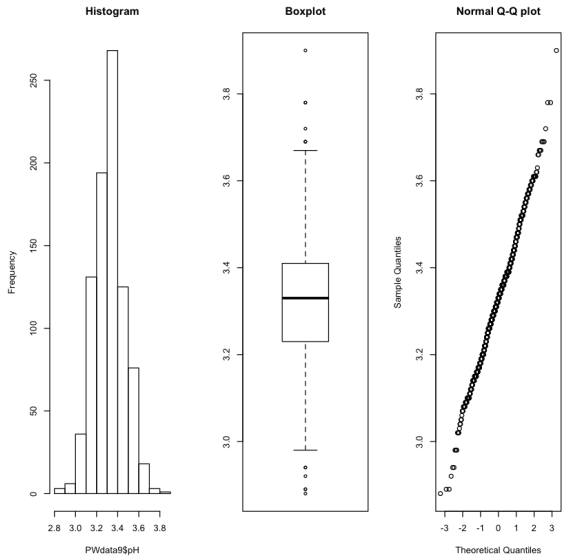
(a) Density Dataset (b) Density Dataset

Figure 39: Density Dataset

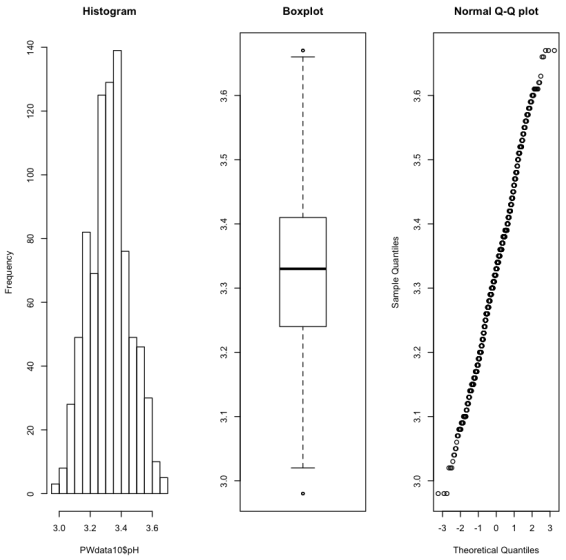


(a) Fixed Acidity Uncleaned Dataset (b) Fixed Acidity Cleaned Dataset

Figure 40: Fixed Acidity Dataset

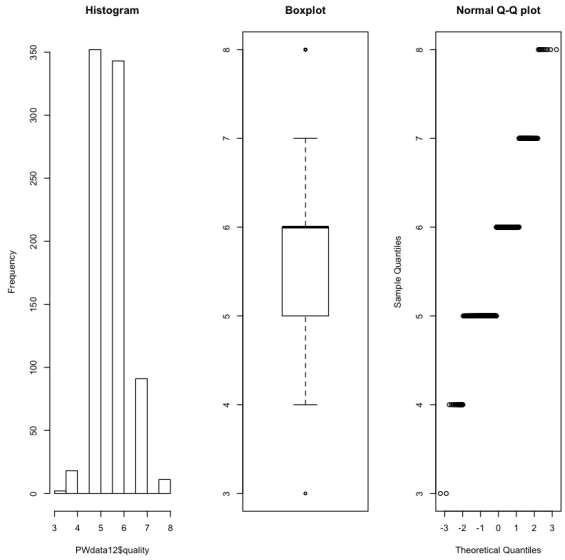


(a) pHUncleaned Dataset

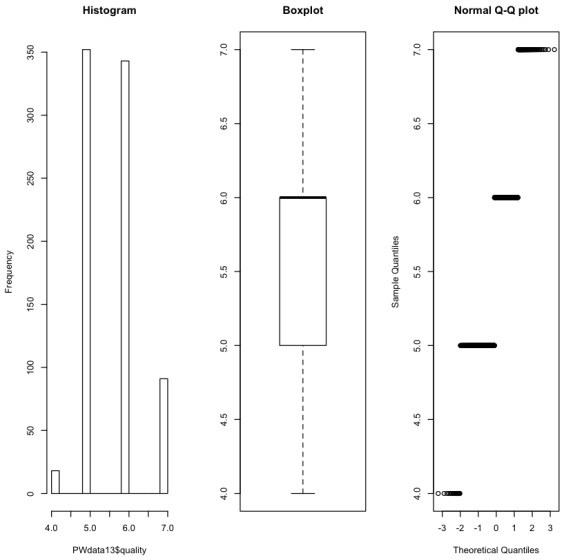


(b) pH Cleaned Dataset

Figure 41: pH Dataset

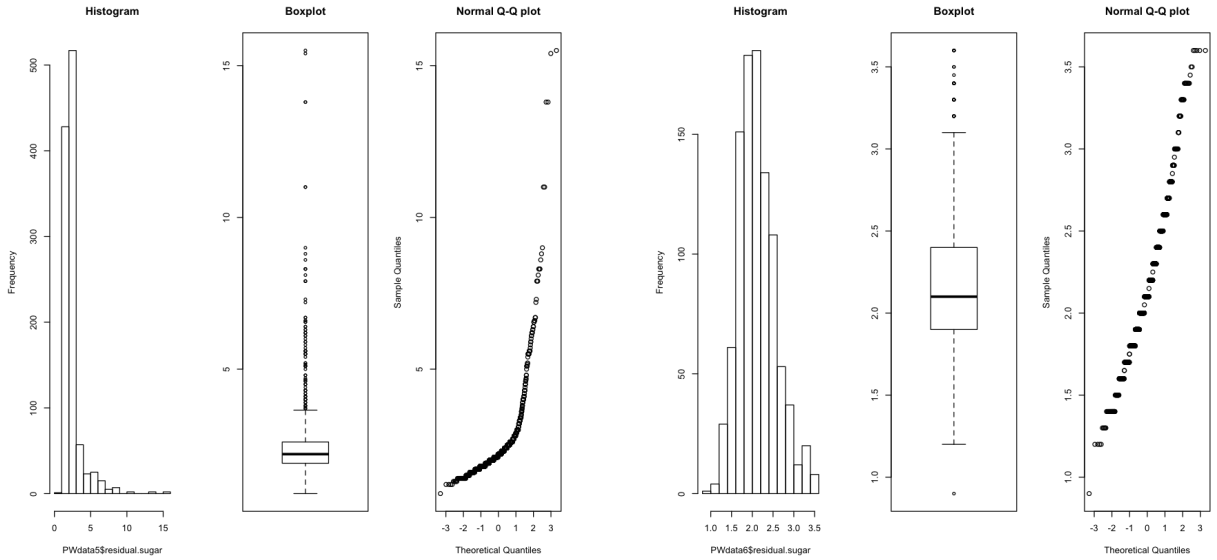


(a) Quality Uncleaned Dataset



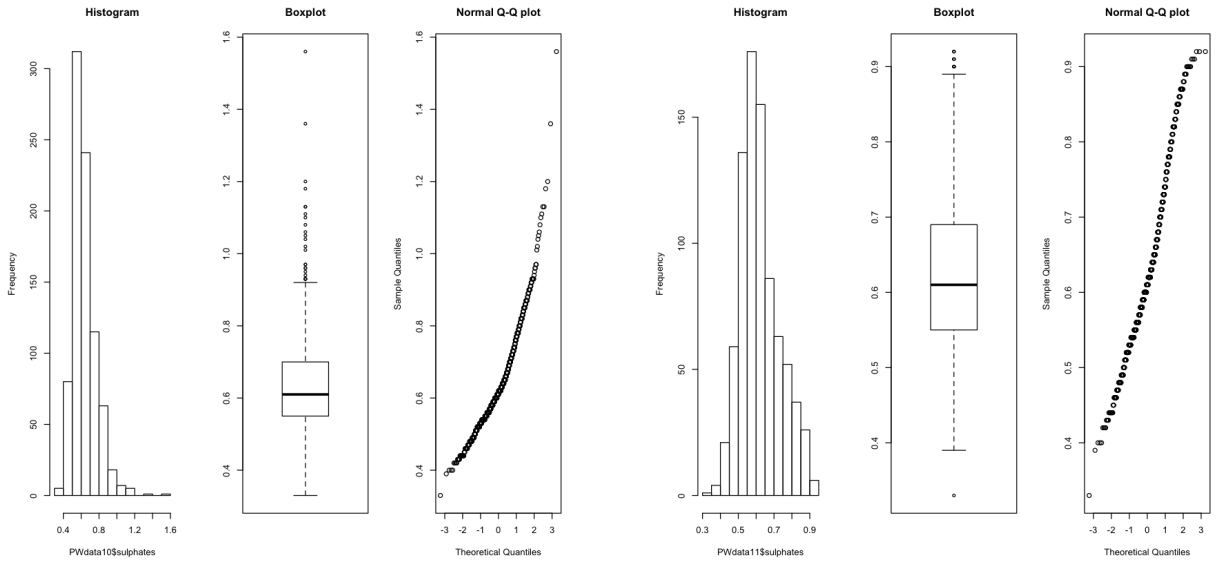
(b) Quality Cleaned Dataset

Figure 42: Quality Dataset



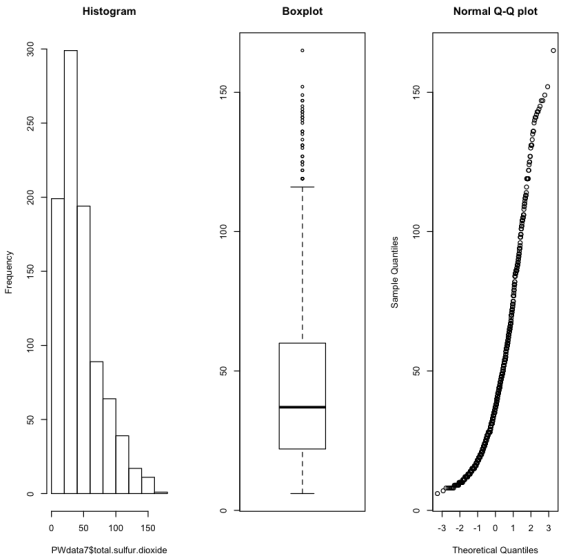
(a) Residual Sugars Uncleaned Dataset (b) Residual Sugars Cleaned Dataset

Figure 43: Residual Sugars Dataset

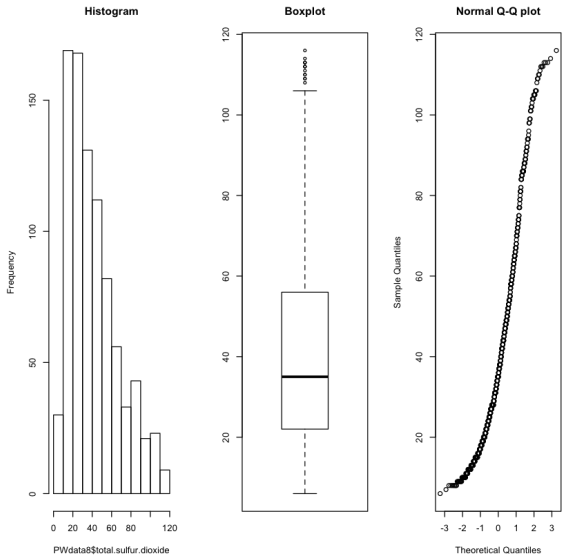


(a) Sulphates Uncleaned Dataset (b) Sulphates Cleaned Dataset

Figure 44: Sulphates Dataset

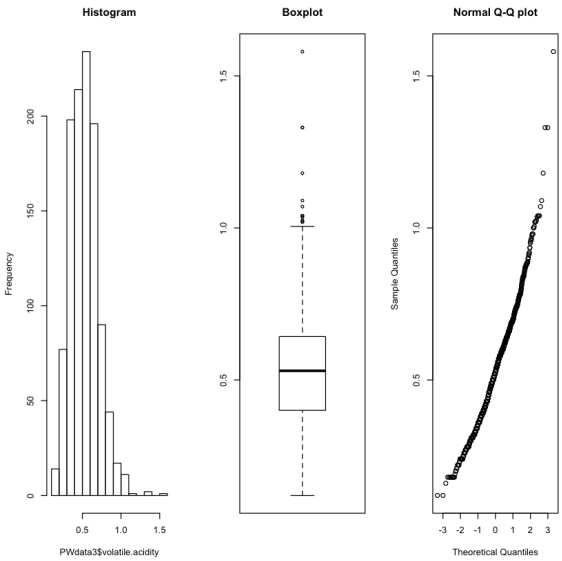


(a) Total Sulfur Dioxide Uncleaned Dataset

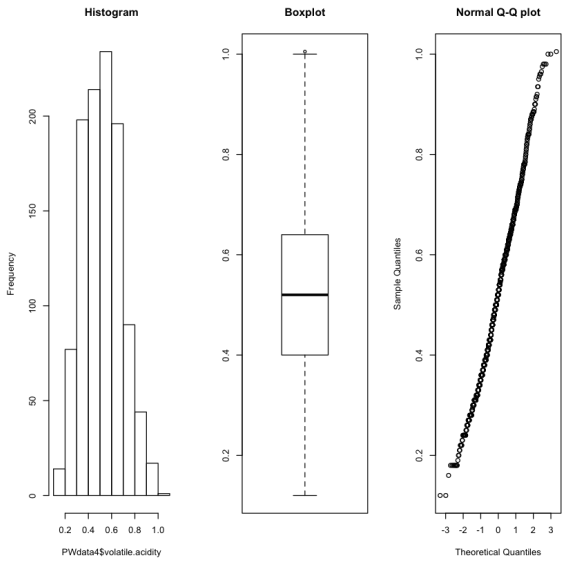


(b) Total Sulfur Dioxide Dataset

Figure 45: Total Sulfur Dioxide Dataset



(a) Volatile Acidity Uncleaned Dataset

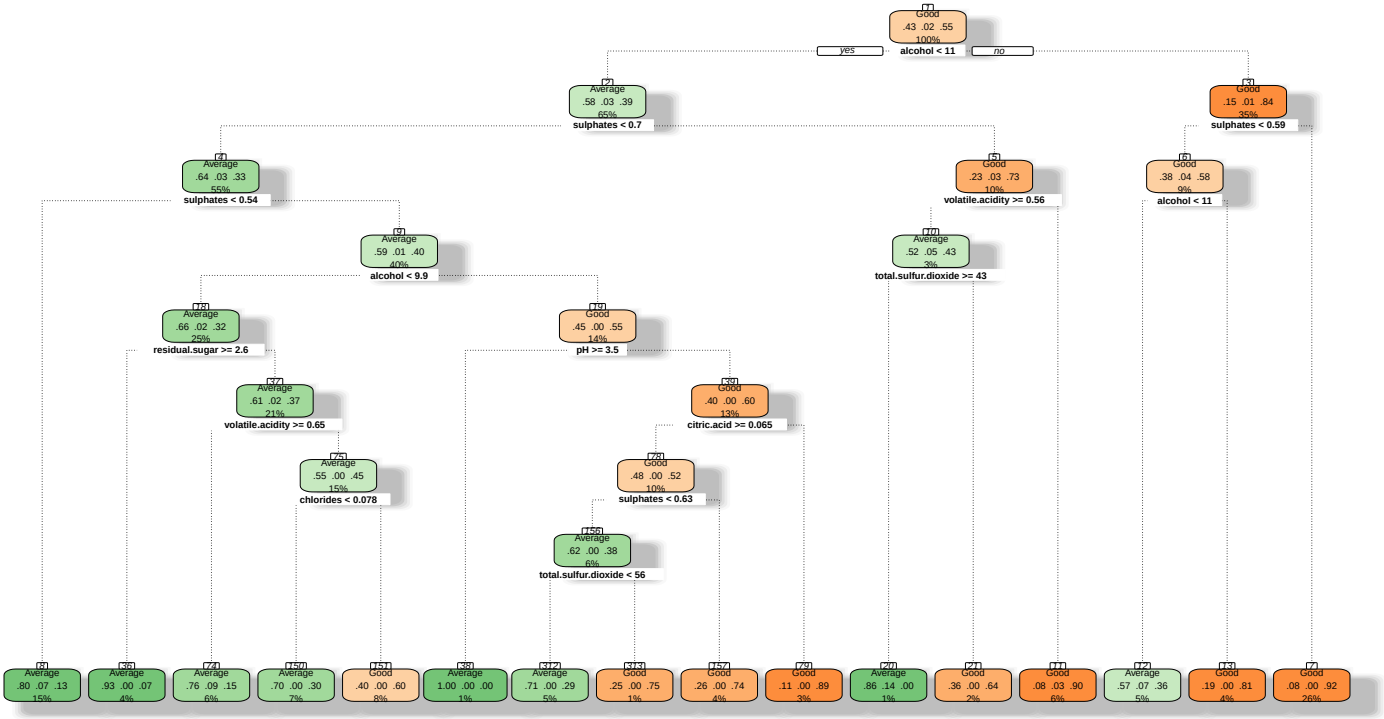


(b) Volatile Acidity Cleaned Dataset

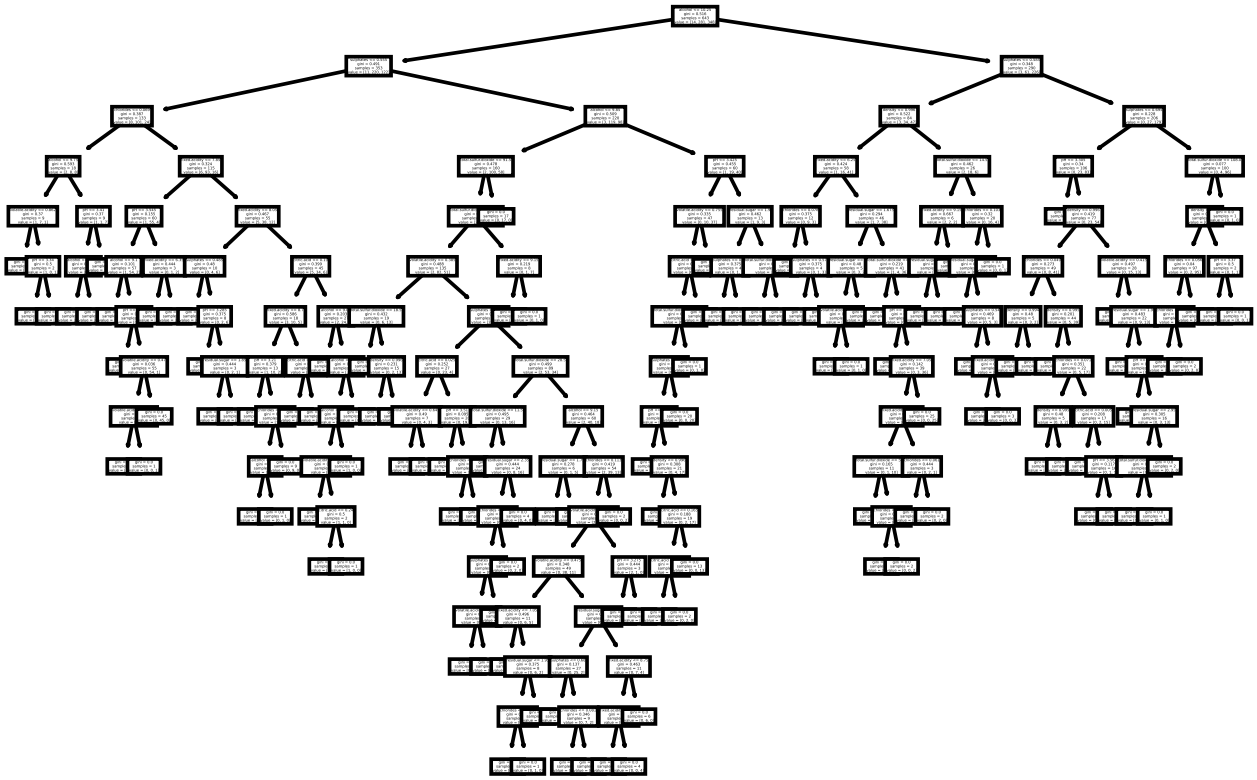
Figure 46: Volatile Acidity Dataset

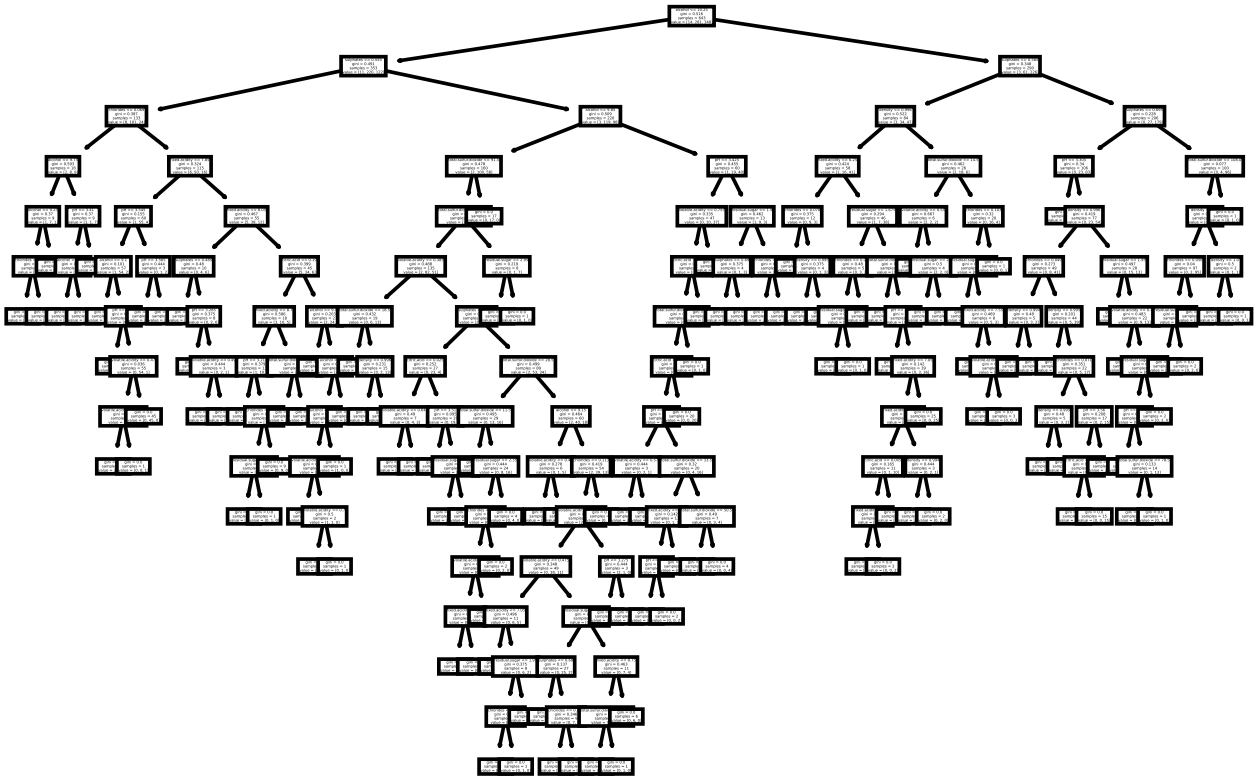
7.2 Supervised Learning

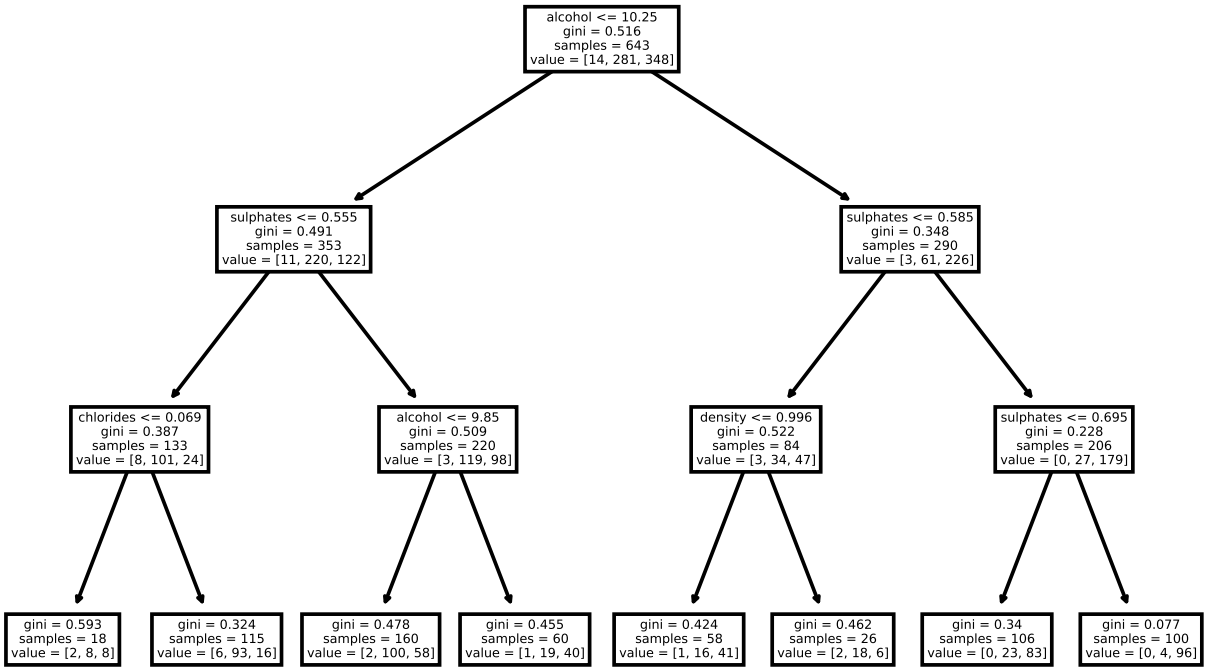
Reference Below for Decision Tree Plots.

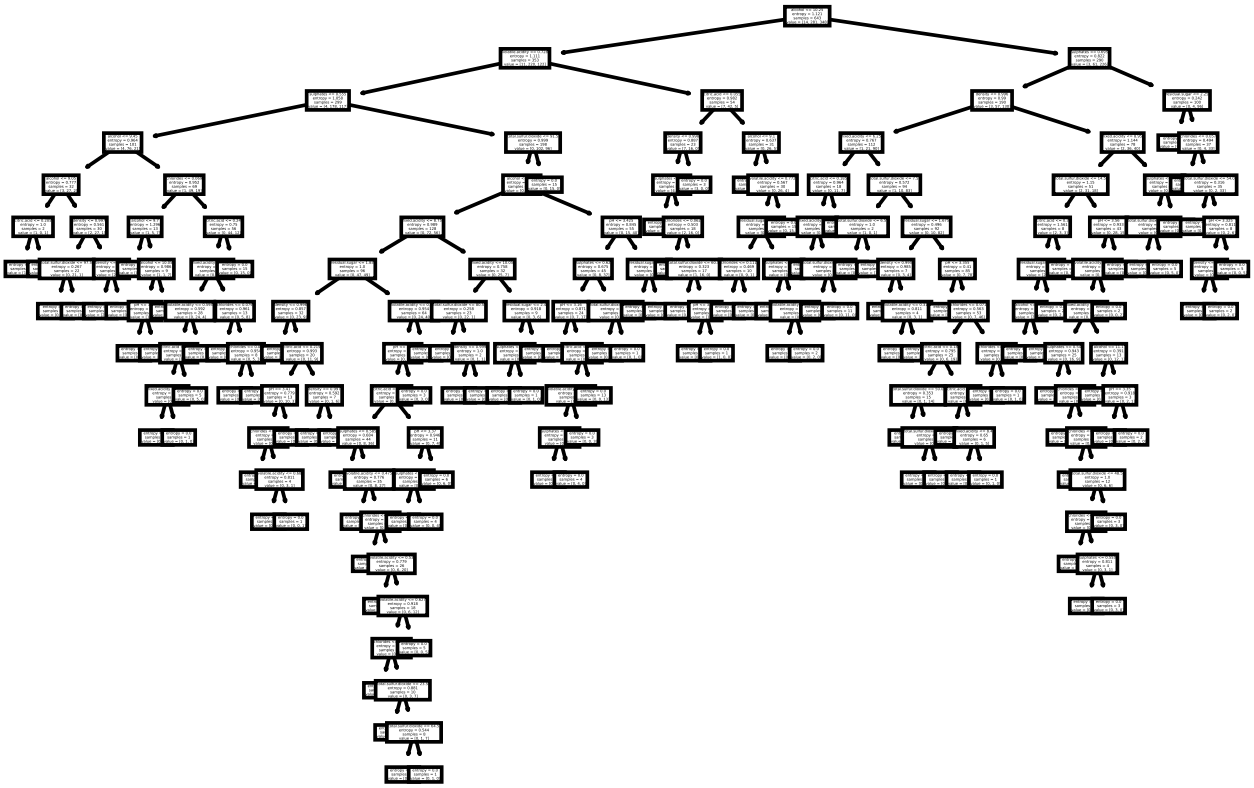


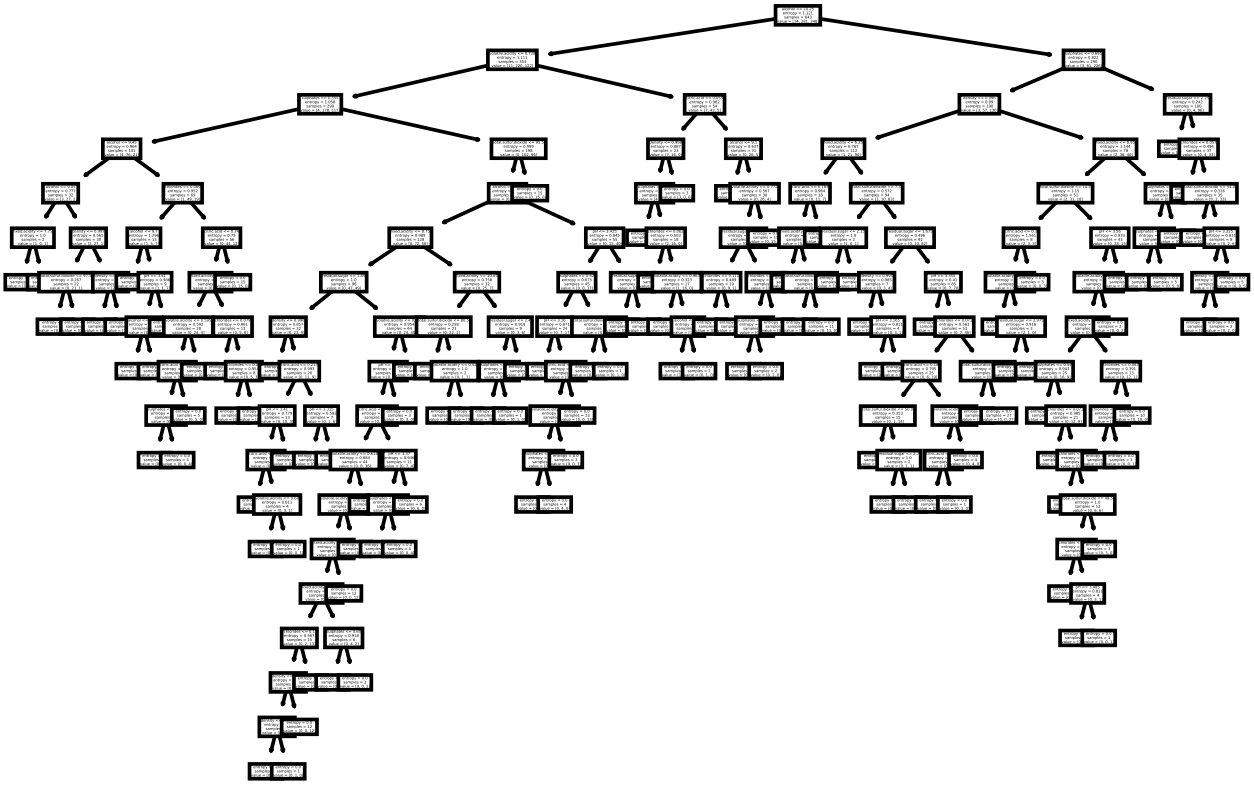
Rattle 2022−Oct−26 11:38:19 samuelkwon






alcohol <= 10.25
gini = 0.516
samples = 643
value = [14, 281, 348]
sulphates <= 0.555
gini = 0.491
samples = 353
value = [11, 220, 122]
sulphates <= 0.585
gini = 0.348
samples = 290
value = [3, 61, 226]
chlorides <= 0.069
gini = 0.387
samples = 133
value = [8, 101, 24]
alcohol <= 9.85
gini = 0.509
samples = 220
value = [3, 119, 98]
density <= 0.996
gini = 0.522
samples = 84
value = [3, 34, 47]
sulphates <= 0.695
gini = 0.228
samples = 206
value = [0, 27, 179]
gini = 0.593
samples = 18
value = [2, 8, 8]
gini = 0.324
samples = 115
value = [6, 93, 16]
gini = 0.478
samples = 160
value = [2, 100, 58]
gini = 0.455
samples = 60
value = [1, 19, 40]
gini = 0.424
samples = 58
value = [1, 16, 41]
gini = 0.462
samples = 26
value = [2, 18, 6]
gini = 0.34
samples = 106
value = [0, 23, 83]
gini = 0.077
samples = 100
value = [0, 4, 96]





7.3 Citation

[1] Yogesh Gupta, Selection of important features and predicting wine quality using machine learning techniques, Procedia Computer Science, Volume 125, 2018, Pages 305-312, ISSN 1877-0509, doi: 10.1016/j.procs.2017.12.041.
[2] Dahal, K. , Dahal, J. , Banjade, H. and Gaire, S. (2021) Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11, 278-289. doi: 10.4236/ojs.2021.112015.