SMU Data Science Review

Volume 7 | Number 1 Article 8

Comparison of Sampling Methods for Predicting Wine Quality Based on Physicochemical Properties

Robert Burigo
Southern Methodist University, rburigo@mail.smu.edu

Scott Frazier
Southern Methodist University, scottf@mail.smu.edu

Eli Kravez
Southern Methodist University, ekravez@mail.smu.edu

Nibhrat Lohia
Southern Methodist University, nlohia@mail.smu.edu

Comparison of Sampling Methods for Predicting Wine Quality Based on Physicochemical Properties

Robert Burigo[1], Scott Frazier[1], Eli Kravez, Nibhrat Lohia[2]
[1] Master of Science in Data Science, Southern Methodist University,
Dallas, TX 75275 USA

Abstract. Using the physicochemical properties of wine to predict quality has been done in numerous studies. Given the nature of these properties, the data is inherently skewed. Previous works have focused on handful of sampling techniques to balance the data. This research compares multiple sampling techniques in predicting the target with limited data. For this purpose, an ensemble model is used to evaluate the different techniques. There was no evidence found in this research to conclude that there are specific oversampling methods that improve random forest classifier for a multi-class problem.

1 Introduction

Wine is one of the oldest and most popular drinks in the world. With that comes the recurring question, how do people pick what wine to drink? Some are picked based on taste, some price, others by the label. But is this really the best way to judge the quality of the wine? This research will use a wine quality data set, that takes both red and white wines of Portuguese origin, to attempt to use the physicochemical features to determine wine quality.

Before going into predicting wine quality or classifying wine quality. It must first be discussed what the term "quality" means. For some, quality might be the cross-section between taste and price. For others, it might mean what does not give them a headache in the morning. Maybe quality can be determined by the perceived value of the wine due to the status symbol that high quality wine might be.

According to Hopfer et al. (2014), the term quality, when in the context of wine, has a strong correlation to aroma and flavor descriptions. Both consumers and wine experts use hedonic descriptors when discussing what quality of wine means (Hopfer et al., 2014). Hedonic here refers to the pleasant or unpleasant sensations when drinking wine. It can be implied that the quality score given to wine is created by experts to appeal to the consumer segments. While this research will focus mostly on the physiochemical features of wine it must be understood that the quality score given by experts is a combination of those features in addition to the more hedonic descriptions (Hopfer et al., 2014).

While the score given to represent the quality of wine is sometimes subjective, one thing that is usually agreed upon is that a big component of wine quality is grape quality. This means how the grape was grown, where it was grown, the types of seeds, the timing of the grow, and many other chemical features of the grape depending on

the type of wine. This is not something that the research will aim to separate (grape and overall wine quality) but rather just provide the context that some quality scores value the quality of the grape more than others who value the process more.

This research will aim its focus on Portuguese wine, both white and red. While many think of popular wine regions in France, Italy, or the United States, the Portugal wine industry has been gaining traction. According to Caldas et al., there have been an increasing number of Portuguese wines submitted to two of the most popular wine rating publications, Robert Parker, and Wine Spectator (Caldas et al., 2013). The main takeaway after reviewing the Portuguese wine is that red wines tend to have higher quality scores than white wines. In particular, the Douro region wines tend to be scored amongst the best (Caldas et al., 2013). This will be considered in the research when classifying both the red and white wines from Portugal.

This research will focus on the actual physicochemical features of Portuguese wine. First, physicochemical features are ones that describe the physical and chemical makeup of something, in this case wine. This includes properties like alcohol level, pH level, citric acid, etc. To get a better understanding of the features that will be used to generate a model, the researchers got some insight from wine expert Brook Ray. According to B. Ray (personal communication, October 31, 2022), there were features that Ray felt would not provide much use in the model like citric acid or volatile acid. Two of the more intriguing features would include residual sugars, as this is a feature that is desired in some wines and not others, and chlorides because there is discussion in the wine community whether proximity to the sea impacts the making of wine (Ray, 2022). A main point that will have to be considered in this analysis is that wine cannot be judged all on the same scale. This means that the quality of some styles of wine may vary greatly compared to other styles. Even the different regions within Portugal might value quality differently. So, while this research hopes to provide a more blanket answer to the classification of wine quality, that context is considered in the analysis.

In previous studies using these features with Portuguese wine, oversampling techniques were used with random forest to classify wines into poor, normal, or high-quality wines. According to Hu et al., this technique provided the desired results and had the least amount of error. The problem with this model is that with 93% of the data belonging to one class, the sampling and model methods must be chosen with much more scrutiny (Hu et al., 2016). There were not any under sampling studies that were done, almost all used some form of oversampling and usually SMOTE. Another problem looked at was both white wine and red wine being evaluated together. This research will aim to provide additional insight and information regarding imbalanced Portuguese wine data by using various sampling techniques and treating white and red wine as separate tasks.

While wine is popular around the world people still generally rely on reviews and suggestions from others to determine the quality. It would be beneficial if there was a way to determine wine quality based on physicochemical features and not have to rely on word of mouth. Wine from all regions will tend to skew towards the most popular types in that region. In which case, predicting the quality of wine will always result in an imbalanced data set. After reviewing what quality means in relation to wine, the features that will be used in this analysis, the experts' initial thoughts, and issues stemming from previous analyses, there is a clear direction as to what the next step is

regarding wine quality prediction. This research aims to compare random sampling methods to determine the optimal method to use when predicting wine quality.

2 Literature Review

2.1 History of wine prediction models

According to Gupta et al., efforts to predict the quality of wines through a means of classification models proved to be effective, but there was still room for improvement in the regards for the wine manufacturers to improve their quality of wine during the production process (Gupta et al., 2021). These researchers make no mention of the significance of the class imbalance between the white and red wine datasets involved. A research paper written by Beh et al., did a review of the original wine quality study conducted by Cortez et al., and these researchers recommend that future researchers should attempt to pursue an approach involving non-symmetrical correspondence analysis (Beh et al., 2012) (Cortez et al., 2009). These researchers also do not touch on the prevalence of the data imbalance or how to address this issue (Beh et al., 2012) (Cortez et al., 2009).

2.2 Physicochemical Features

When discussing the actual makeup of the properties of wine, there are a few key physicochemical features that play an important role in the resulting quality of the wine. These have all been mentioned in this paper, as well as previous studies conducted involving the analysis of wine quality, but it is worth diving into the actual effect each of these features has.

One group of researchers created a trained model to give customers the ability to predict a wine's quality based on the physicochemical features (Mor et al., 2022). The metrics the researchers built this model from include the physicochemical features which will also be present in this study, and according to these researchers, fixed acids (also known as non-volatile acids), include tartaric, malic, citric, and succinic acids (Mor et al., 2022). These acids do a good job of resisting the effects of other acids present within the wine, and as a result do not evaporate as easily. Fixed acids will vary in strength depending on which acids are present overall. For acids such as tartaric, the wines see an elevated level of impact in the overall taste and color profile of the wines (Mor et al., 2022).

On the other end of the spectrum, volatile acids such as acetic acid (which happens to be the primary volatile acid that winemakers are concerned with). The reason this acid is of great importance is because acetic acid is responsible for the level of sourness in the wines; so too much acetic acid will lead to a wine which will approximate that of vinegar in both taste and scent (Mor et al., 2022). If a volatile acid is present in wine, it is considered a flaw. Citric acid should also be kept to a miniscule level; and this is because once citric acid has been included, it is typically an indicator

that a wine has been acidified in a cheap manner (Ray, 2022). Residual sugar, according to Horváth et al., is a byproduct resulting from the original elevated levels of sugar in the grapes/wine itself after the fermentation process has been completed. Once a wine has reached a residual sugar level of more than 35g/L to 45 g/L, it can be classified as a sweeter wine (Horváth et al., 2020). It will depend on the wine, but some wines result in a more desirable taste when residual sugar levels are higher. This is also contingent on the way it interacts with the acids present (Ray, 2022).

When dealing with chloride within wines, it is essentially the amount of salt within the composition of the wine. According to Sonegheti et al., the amount of salt, in this case chloride, is related to something called terroir and the type of the grape involved in making the wine (Sonegheti et al., 2015). Free sulfur dioxide and total sulfur dioxide are known as sulfites. These characteristics of the wines are utilized to preserve and maintain the integrity and taste of the wine (Sonegheti et al., 2015) Sulfur dioxide does occur naturally in wines, but it can also be added to the mixture for the wine by the winemaker. It is also worth mentioning this is something which is regulated by law depending on the area in which the wine is being created (Ray, 2022).

The density of the wine is a mixture of a few elements. These include the mixture of the alcohol content, as well as the presence and amount of sugar, glycerol, and other miscellaneous elements. Some wines are meant to be denser than others; and wines such as Vinho Verde are not meant to be dense (Ray, 2022). The pH level is defined as a measurement of a liquid's acidity or basicity on a scale ranging from 0 to 14. When putting wine on the pH scale, wines on average will receive a score typically ranging from 3 to 5, making most wines more acidic in nature (Sonegheti et al., 2015).

Sulfites can relate to the free sulfur dioxide as well as the total sulfur dioxide mentioned earlier in this section, but in this case, it is a wine additive. These can be present in the wine naturally without adding any SO2 though. In this instance, sulfites are being used to act as an antioxidant. Alcohol is the percentage of alcohol present in each container of wine. There are no strict guidelines on how much alcohol is to be present in each container of wine, but they do typically contain anywhere from around 5% to 25% alcohol by volume (ABV) depending on the type of wine (Sonegheti et al., 2015).

2.3 Under Sampling Method

One of the most common techniques to resolve issues with imbalanced datasets is random majority under-sampling (RUS), in this process, samples of the majority class are removed at random from the dataset. However, the main drawback of this approach is that some potentially useful information contained in these deleted examples is lost due to this process (Lin et al., 2017). To overcome this limitation, they propose to use clustering technique instead of RUS. Rather than deleting random data points they propose to cluster majority class (group similar data samples into the same cluster). The number of clusters in the majority class is set to be equal to the number of data points in the minority class (Lin et al., 2017). The authors used two different strategies to represent the class. The first consisted of using k-means to cluster data and to use centroids to represent these clusters. The second was to use k-means but instead of selecting the centroid, which is often a new data point created by the algorithm, they suggest using nearest neighbors and finding the nearest real data point to the centroid

(Lin et al., 2017). The machine learning design which was proposed follows the following steps: 1. Split data into training and test. 2. Select majority data from the training set and apply clustering-based under-sampling. After this step, it will produce balanced training data. 3. In this step certain methods such as classification algorithms will be used on the balanced data. 4. Use test set to evaluate the results (Lin et al., 2017).

In another study, Arefeen et al. focused on using Neural Network-based under sampling techniques to address the described above RUS problem of losing information when deleting sab sample of the majority class. Authors propose two recently composed algorithms that are using neural network-based methodology to remove samples of the majority class that exist in the surrounding area of the samples of the minority class (Arefeen et al., 2022). Thereby under sampling the majority class to resolve the imbalance issue. Based on the paper, there are two main drawbacks of any imbalanced data. First, the existence of a nearly overlapped region of two different classes. It is difficult for a model to learn the minority data in such an environment (Arefeen et al., 2022). There is also the existence of a high imbalance ratio between the two classes. The main strategy of these two algorithms is to perform under sampling using Neural networks by removing majority class data that are more like minority class data (Arefeen et al., 2022). It chooses the number of majority samples which is exactly equal to the number of the minority samples. And by this process achieving data which is balanced.

2.4 Over Sampling Method

Another common solution to imbalanced data sets is oversampling. Oversampling is the process of adding additional records on to the dataset in order to balance out the classes to help with the model. The Synthetic Minority Oversampling Technique (SMOTE) method is considered the unofficial standard as it related to balancing classes (Fernandez et al., 2018). The SMOTE algorithm aims to rebalance the original data set by creating synthetic data. Fernandez et al. describe the process as a construction of new data points between samples of the different minority classes that are within a defined area (Fernandez et al., 2018). They explain the foundation of this novel technique which was to create new samples of the minority class. SMOTE preprocessing technique became an industry standard for the data science community in dealing with imbalanced classification. Since inception, many add-ons and other methods have been implemented to improve its performance under different data circumstances. To give an example of how the process of SMOTE works, Fernandez et al. explain that first the total oversampling amount is set up, called N. Then, from the training data, the minority instance is selected using a nearest neighbors of 5 (K, by default although other values can be selected), N of these K instances are then chosen to compute the new instances that will be added to the data set via interpolation (Fernandez et al., 2018).

The next solution to solving the issue of an imbalanced data set is random oversampling (ROS). Liu et al. describe the process of ROS as randomly selecting existing minority class data points to duplicate and balance out the data (Liu et al., 2018). While it seems, simple there is some evidence to support that ROS performs on par with other methodologies. Liu et al. conclude that ROS performs on par when it comes to text classifications (Liu et al., 2018). Although, when faced with a problem

that uses nearest neighbors, there is evidence to support that SMOTE would outperform ROS (Liu et al., 2018).

2.5 Multiclass Imbalanced data

All the approaches discussed thus far are for binary classification problems. The data set discussed below is not binary, but a multi-class problem. Janicka et al. discuss in their paper that the current approach to dealing with multiple imbalanced classes is mainly based on translating multi-class problems into special binary subtasks (Janicka et al., 2019). One of the common approaches for the multi class problem is to use one-versus-one (OVO) or one-versus-all (OVA). The problem with this approach as described by the authors is that the information about the decision boundaries between different classes as well as information about data distributions of these classes is lost (Janicka et al., 2019). As in the original problem one class may influence several neighboring classes. As well as this process of creating two classes for multiple class problems ignores the mutual relations between classes that are different for majority and minority classes and increases the complexity of the learning task. The authors propose a new hybrid algorithm called Similarity Oversampling and Under sampling Preprocessing (SOUP). In SOUP, all majority classes undergo an under-sampling process, and all minority classes undergo an oversampling process to the number of elements being the mean of the count of the largest minority and the smallest majority class (Janicka et al., 2019). Implementing SOUP data processing before OVO or OVA significantly improves results of classification algorithms.

Yao et al. introduce another novel algorithm for data level pre-processing for the multi-class imbalanced data. They called this approach Evolutionary Mahalanobis Distance Oversampling (EMDO). EMDO uses a set of ellipsoids to approximate the decision regions of the minority class Synthetic minority samples are generated based on Mahalanobis distance within every ellipsoid (Yao et al., 2021). And the number of synthetic minority samples generated by EMDO in every ellipsoid is determent by the density of the minority samples in every ellipsoid. EMDO was evaluated against other multi-class imbalanced data learning algorithms on different number of data sets. EMDO was demonstrated to outperform competing oversampling approaches in simulation (Yao et al., 2021).

As seen from the above, multi-class imbalanced classification is difficult. Lango et al. discuss in their paper ‘What makes multi-class imbalance problem difficult’ results of multiple experiments with different data sources to provide an answer to this problem (Lango et al., 2022). Especially multi-minority, multi-majority classification problems. In this paper authors describe the experimental study they conducted to test impact of various multi-class imbalanced difficulty factors on performance of the 3 classifiers: CART tree,4 k-Nearest Neighbours and bagging ensemble. It was noticed that the imbalance between classes is often accompanied by additional data difficulty factors, such as rare sub-concepts in the minority classes, overlapping regions between classes, or rare minority examples located inside the majority class region (Lango et al., 2022). It was determined that it is the combination

of various data difficulty factors that makes learning from class imbalanced data challenging (Lango et al., 2022).

Jingjun et al. provide great insight on the most up-to-date multi-class classification models that are being used today (Jingjun et al., 2018). After giving context, they then created a new model, which they call DECOC, which basically combines the improved ECOC model with methods used in ensemble models (Jingjun et al., 2018). This research will be helpful to determine which model will be most beneficial when comparing sampling methods to make sure what this study consists of is up to the industry standard.

3 Data

3.1 Data Ingestion and Preparation

The data being utilized will consist of two datasets from the UCI Machine Learning Repository, with one of them holding the data for red wines and the other for white wines. It is also worth mentioning that according to the owners of the data, the wine data available here is dealing with variants of a Portuguese wine known as "Vinho Verde". No specific information such as grape types, brand, or sale price was available with the data given to us. The size of the data is as follows: The red wine dataset contains 12 variables with 1,598 rows of data not including the header; and the white wine dataset contains 12 columns with 4,898 rows of data. Of these 12 attributes which will be used throughout this study, 11 of them are predictor (independent) variables, with the 1 remaining outcome (dependent) variable. There are no missing values in either data set. Each row in the data set is of a different wine that was collected in 2009.

1. Fields in datasets
    a. Fixed acidity
    b. Volatile acidity
    c. Citric acid
    d. Residual sugar
    e. Chlorides
    f. Free sulfur dioxide
    g. Total sulfur dioxide
    h. Density
    i. pH
    j. Sulphates
    k. Alcohol
    l. Quality (score between 0 and 10)

The target variable is the quality score of the wine. This score for both the white wine and the red wine is between 0 and 10. In order to develop classification models with an imbalanced data set, the white wine data was grouped into “low”, “medium” and “high” buckets based on the quality score. Wine of low quality was scored between 0 and 5.5. Wine of medium quality was scored between 5.5 and 6. Lastly, wine of high quality was scored greater than 6. For red wine, the buckets were just split into “low” and “high”. Low quality was anything between 0 and 6. High

quality was 7 and above. These buckets were determined in order to get sufficient data in each bucket before undergoing any sampling techniques. Without sufficient data, it was determined that any synthetic data created would result in a poor experiment overall. After bucketing the data, the distribution of the target groups broke out for each wine as demonstrated in Table 1. Due to the disparity in scores and feature importance for white wine and red wine, the data needed to be split up as opposed to combining the two types of wine.

Table 1. Target Group Distribution

Target Group	% of White Wine Data	% of Red Wine Data
Low	33.48%	86.43%
Medium	44.88%	N/A
High	21.64%	13.57%

3.2 Exploratory Data Analysis

As explained in section 3.1, data is highly unbalanced for the red wine data source. The smallest number of samples are for wines with a quality score of 3 and it occurs in only a total of 10 examples. The largest number of samples is for wine with a quality score of 5 and it occurs in a total of 681 samples as shown in Figure 1. This imbalanced breakdown made bucketing into target groups difficult. In order to set up an experiment to compare sampling techniques with red wine data, there needed to be a split in groups between those with a quality score of 0 to 6 and those with a quality score of 7 and above. It was determined that wine with a quality score of 5 and 6 would be put in the “low” group and wine with a quality score of 7 would be put in the “high” group.

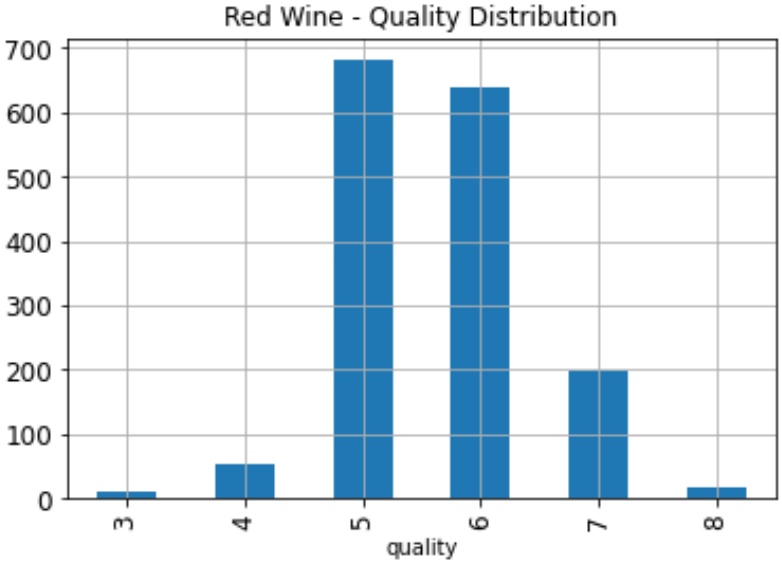

Figure 1. Red Wine Quality Distribution

To get a look at what features might be useful in a model to classify quality, a chart was created to show the correlation between each variable and the target variable. Of note there appears to be some positive correlation between alcohol and red wine quality. There also appears to be a negative correlation between volatile acidity and wine quality as shown in Figure 2. For clarity, alcohol is the concentration of alcohol in the wine and volatile acidity is the measure of the wine's gaseous acids. This shows that as wine becomes less acidic in flavor and concentration, the quality score would be higher on average. The quality score also is higher as the concentration of alcohol rises although it's thought that this will plateau. With two features showing a relationship with the target variable, a check of correlation between features is also needed. If such a relationship exists, then there would have to be a discussion of which feature to drop from the model. In this case, there are no highly correlated features (over 0.8 correlation) for the red wine data set, as shown in Figure 3, so no features will be dropped.

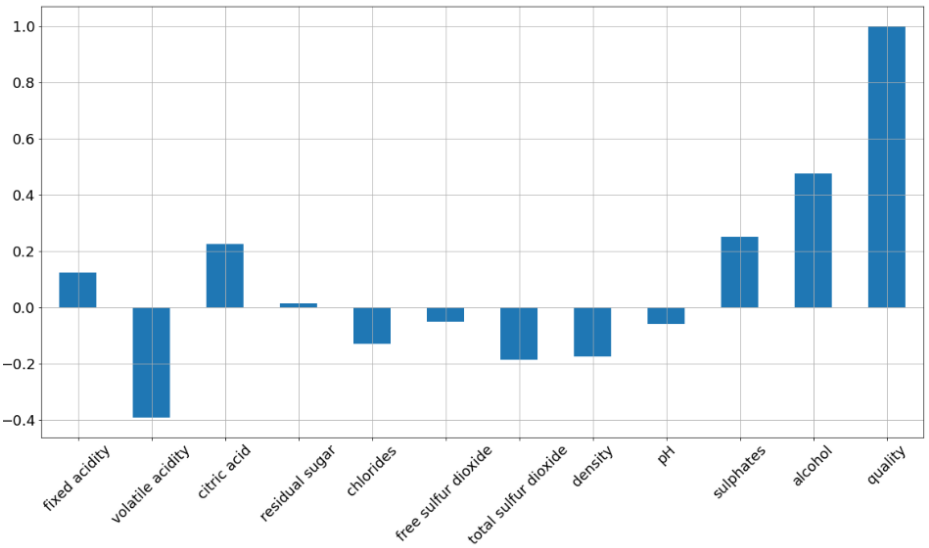

Figure 2. Red Wine Correlation with Target Variable (*quality*)

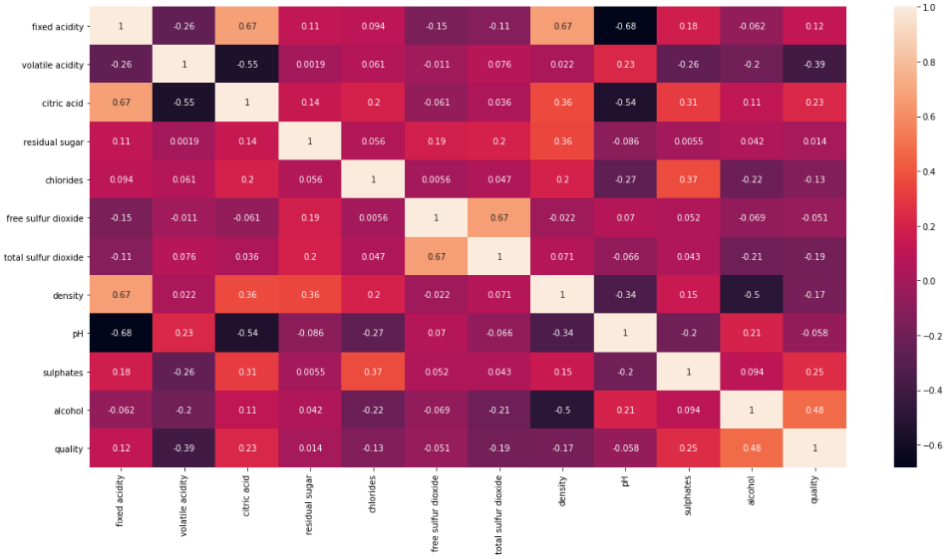

Figure 3. Red Wine Variable Correlation Matrix

Next, in the exploration of the white wine data set, a highly unbalanced data set is once again revealed. The smallest number of samples for a particular quality score was 5 samples with a score of 9. On the other hand, there were 2,198 samples with a quality score of 6. The white wine data lends itself to be split up in buckets a

little easier due to roughly similar number of samples below 6 and above 6 as seen in Figure 4.

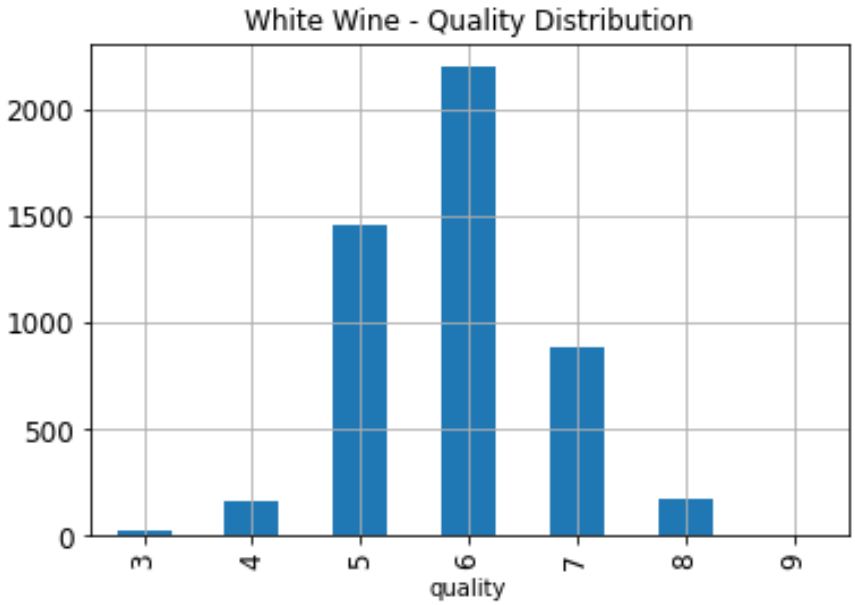

Figure 4. White Wine Quality Distribution

Another correlation check was done for white wine, first between the features and the target variable (Figure 5) and then between each of the features as shown in Figure 6. In a similar fashion to red wine, white wine also showed a positive correlation between alcohol and wine quality. There are some negative correlations that start to pop up like density or chlorides, but that relationship is not quite as strong as alcohol. Once again, there is no correlation between the features so nothing needs to be dropped before the models are created, seen in Figure 6

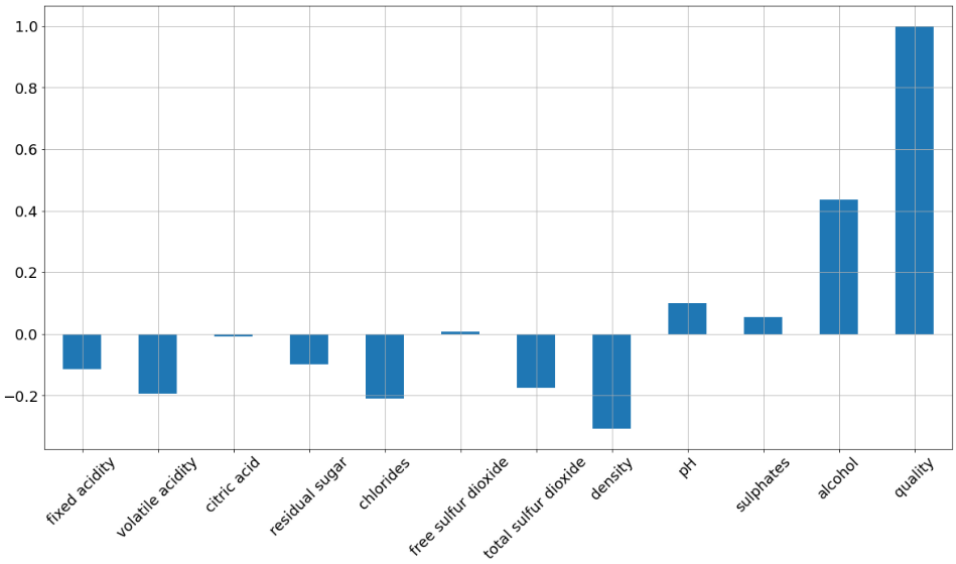

Figure 5. White Wine Correlation with Target Variable (*quality*)

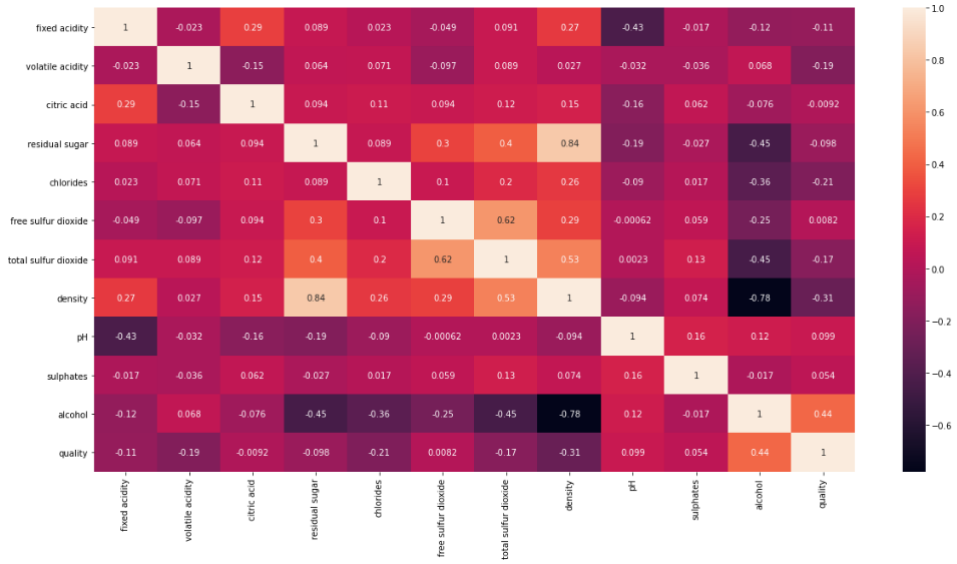

Figure 6. White Wine Variable Correlation Matrix

4 Methods

This wine quality data set does not have an even balance amongst the classes. This presents a problem when creating a final model and one that must be solved. This research will explore the idea of oversampling with different variations of SMOTE or Random Oversampling. Under sampling was explored but the distribution of quality of wine scores did not meet the requirements to do under sampling or neural network based under sampling. There are several approaches to deal with imbalanced data. The decision was made to focus on a data-level pre-processing approach. In this approach the skewed data distribution will be rebalanced by adding new minority class samples. The main advantage of this approach is that they are independent of the classifiers. The goal is to provide a comparison of the different techniques based on performance within the context of wine quality.

While these topics have been discussed in previous work, this research aims to bridge the gap between the two. There is a disconnect between learning machine learning models in a controlled setting and applying it to real world data. The topic of imbalanced data is usually discussed but not in the depth that is needed. Other researchers have started trying to predict the quality of wine with either XGBoost models or Random Forest models. There have also been papers that have gone more in depth with one sampling method. While this has been relatively successful there is still a gap that has been uncovered where the methods chosen to deal with imbalanced wine quality data have not been explored in the detail needed for an optimal model. While there will be quite a few sampling techniques with results that will be shared, this section will focus in more detail on a select few.

4.1 Baseline Random Forest

To evaluate the effectiveness of the different sampling methods, there must first be an established baseline. For this research, a random forest model will be sed. A random forest model is an ensemble classification model that uses multiple uncorrelated decision tree models. The decision tree individually determines which classification to set the sample data as and then as a group, the classification with the most results will be assigned to that sample. When building this model, the actual data with no sampling is being used, F1 score will be used as the metric to determine the effectiveness of the model. The reason for using F1 score is that it is a metric that considers both false positives and false negative and thus gives a wholistic view of the model. The formula for F1 score is shown in Figure 7 below.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Figure 7. F1 Score Calculation Formula

This baseline model will be done for both the multi-class problem for white wine and the binary class problem for red wine.

4.2 Variations of SMOTE

The first type of over sampling method that this research wanted to explore was SMOTE. As referenced in section 2.4, the SMOTE algorithm aims to rebalance the original data set by creating synthetic data, shown in Figure 9 (De Dios Santos, 2019). Fernandez et al. describe the process as a construction of new data points between samples of the different minority classes that are within a defined area (Fernandez et al., 2018). In Figure 8, the original data is shown as well as the rebalanced white wine data using SMOTE.

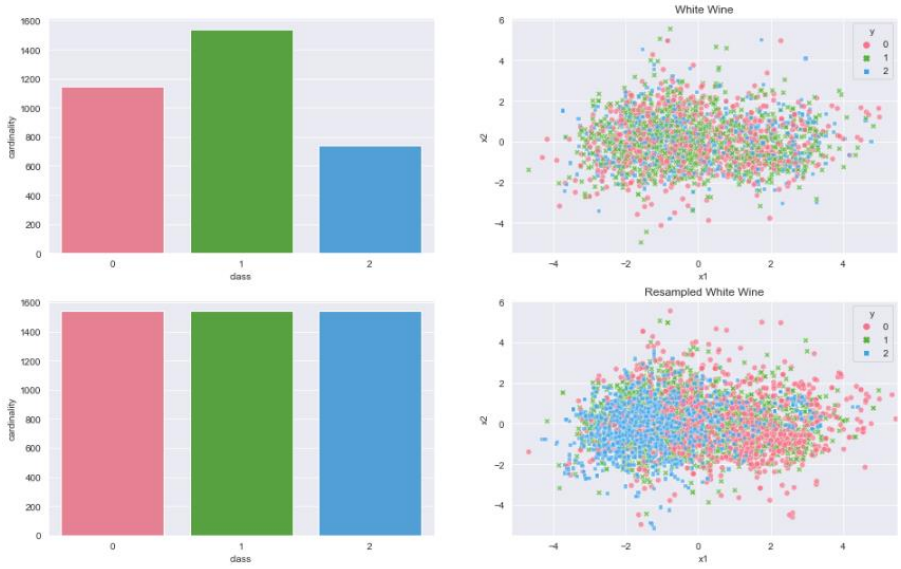

Figure 8. White Wine Resampling - SMOTE

SMOTE appears to have created some better separation in the groups with high quality (score group = 2) than the original data but still not as great as what was hoped. This led the group to investigate more hyper focused versions of SMOTE like SVMSMOTE, SMOTEN, and KMeansSMOTE among others. SMOTEN is done after the group is split into classification groups. SVMSMOTE works to establish borders amongst the data to aid in classification, and KMeansSMOTE runs a K-Means model before the oversampling is done. These are just some examples that were used but are some of the more common ones in the industry. Once all these SMOTE variations were run, the data was then used to create a random forest classifier. From this mode, F1 scores were extracted and compared to the baseline random forest classifier to evaluate if SMOTE or other variations were able to improve on the baseline model. This was done for both red and white wine.

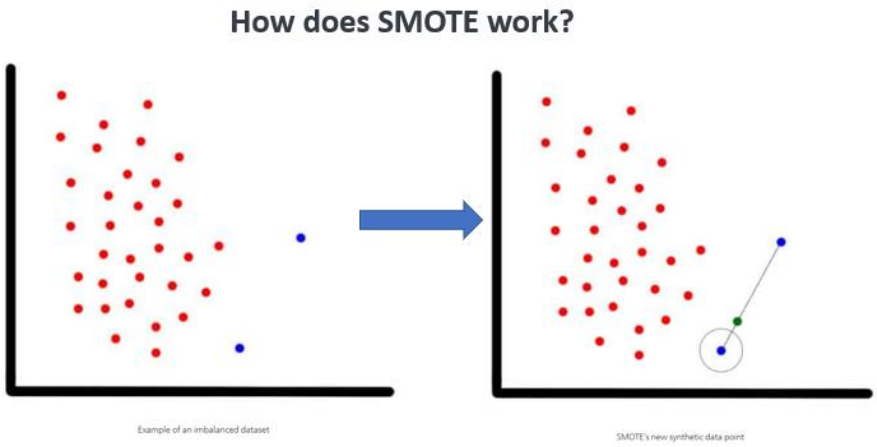

Figure 9. SMOTE Visualization

4.2 Random Over Sampling

While SMOTE methods create new synthetic data in attempts to output a more balanced data set, random over sampling randomly selects existing minority class data points and duplicates them in order to balance out the data (Liu et al., 2018). In the case of the white wine data set, the medium data set (swine quality score = 2) was the majority class. This means with the random over sampling model, low score data and high score data were randomly selected and duplicated to equal the majority class, seen in Figure 10. For the red wine data, the high score class was the minority class, and it was severely imbalanced. So, in this case, the over sampling technique took random samples of high-quality wines and duplicated them to even out the red wine data set, as shown in Figure 11.

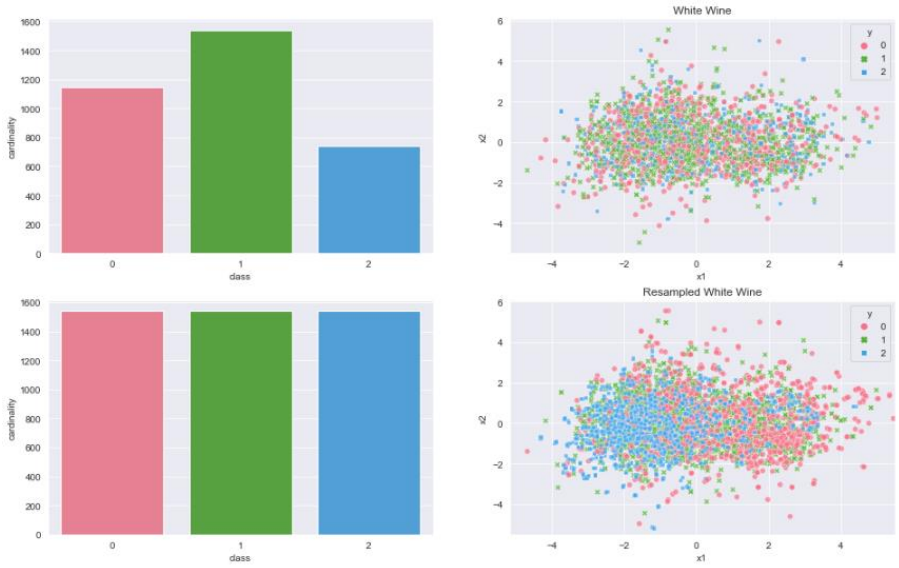

Figure 10. White Wine Resampling – Random Oversampling

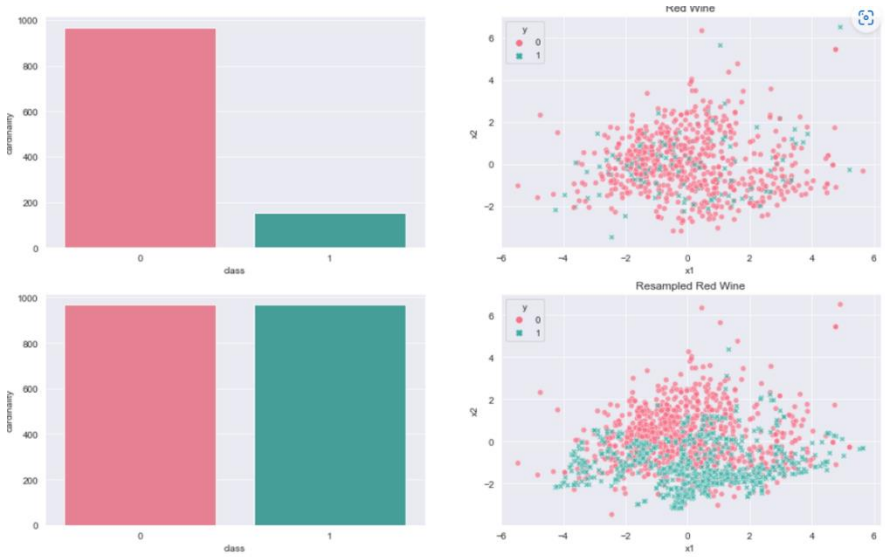

Figure 11. Red Wine Resampling – Random Oversampling

For the white wine there appears to be separation between high-quality wines and then low and medium quality wines. The separation between low and medium does not appear to be that strong and it is expected this is where the model might have some difficulty. For red wine, there does appear to be good separation between low-

and high-quality wines, it is expected that the results of these models will be similar as this is a duplication of data

5 Results

5.1 Baseline Model Results

For both the white wine and red wine data sets, a baseline model was created. For white wine, the F1 score of the random forest classifier was .70. For red wine, the F1 score for the random forest classifier was .90. As suspected, the binary model for the red wine performed a little bit better at limiting false positives and false negatives. There does not leave a lot of room for improvement for red wine. White wine has more room for improvement but due to the problem being multi-class it will be interesting to see how the sampling techniques impact the overall model. For both baseline models, it was important to look at feature importance to gain more understanding of which features were more prominent in the model. There was not much difference between white and red wine regarding feature importance. The four that stand out above the rest for being more prominent are alcohol, volatile acidity, density, and free sulfur dioxide. Citric acid, sulfates, and fixed acidity stand out as the three that are not as prominent shown in Figure 12.

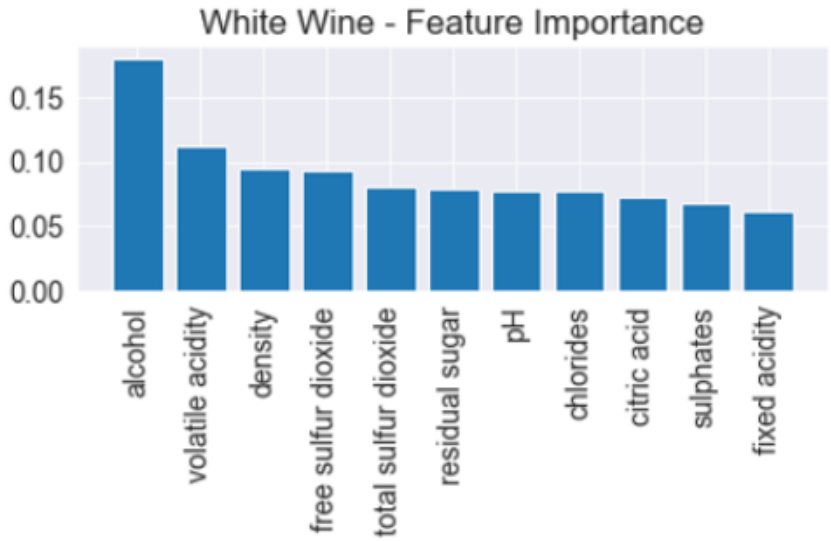

Figure 12. White Wine Feature Importance shows some of the top features (*shown on left*) when predicting quality include alcohol, volatile acidity, and density.

5.2 White Wine Model Results

White wine is the multi-class data set where the base line random forest classifier had an F1 score of .70. Ten different over sampling techniques were used to compare to the baseline. Of those ten, only two met the baseline F1 score (SMOTE, GlobalCS). This is shown in Table 2 below.

Table 2. White Wine Model Performance (*shown below*). SMOTE and GlobalCS were able to obtain a similar F1 score to our baseline, while the lowest scoring technique was the SOUP (Similarity Oversampling Undersampling Preprocessing) technique.

Oversampling Algorithm	F1 Score
Base Case (No oversampling)	0.70
SMOTE	0.70
SVMSMOTE	0.69
KMeansSMOTE	0.69
SMOTEN	0.69
SMOTETomek	0.69
Random Oversampling	0.68
ADASYN	0.68
MDO	0.68
GlobalCS	0.70
SOUP	0.67

SMOTE has been discussed but GlobalCS has not. GlobalCS uses all classes and then duplicates the samples in each class equally until a balance is reached. While it is concerning that there was no real improvement in F1 score for any of the over sampling models, it does give some clarity regarding the white wine data set.

5.3 Red Wine Model Results

With red wine, there was more optimism due to the problem being binary. Although, with the baseline F1 score at .90, there was not as much room to improve upon the model. For red wine, there were five random forest models that were analyzed in addition to the base line model. Of those five, three of them matched the baseline F1 score of .90, while one of the models (SMOTEN) improved the baseline model with an F1 score of .91 as shown in Table 3.

Table 3. Red Wine Model Performance (*shown below*). SMOTEN results in highest F1 score with 0.90, while MDO (Mahalanobis Distance-based Oversampling) has the lowest F1 score with 0.88.

Oversampling Algorithm	F1 Score
Base Case (No oversampling)	0.90
SMOTEN	0.91
SMOTETomek	0.90
Random Oversampling	0.90
MDO	0.88
GlobalCS	0.90

SMOTEN is essentially SMOTE but with the resample only being used on categorical features. When there was specific resampling for categorical features the model had more success in reducing false classifications. MDO was the only model that had decreased performance compared to the baseline. MDO uses the covariance of the minority class in order to create new synthetic data. The red wine data is encouraging in that when predicting wine, it might be useful to only have two classes. With imbalanced multiclass data there might be too many variables in the process to make sampling worth it.

5.4 Metrics

The primary metric that we employed in this research study was F1-score. The reason for this decision is because of the data we are working with, we found that both false positives and false negatives are undesirable errors, and as such F1-score will take the harmonic mean of both FN (precision) and FP (recall) heavy models. Should there be an instance in which multiple sampling methods are used but they have similar results, we will choose the machine learning algorithm and sampling method we deem to be the best option based on certain factors such as the resulting group separation between the data classes.

6 Discussion

Different algorithms to solve unbalanced data for binary classification problems were developed and are very common in use. multi-class imbalance data classification problems are much more complicated and the research in this field is ongoing. Most of the packages which were used in this paper are not part of the common libraries like sklearn. It requires a significant investment to discover implementations of algorithms for multi-class problems as well as the correct ways to use these packages.

Working with multi-class unbalanced data requires domain knowledge. Using domain knowledge and combining classes to reduce the number of unbalanced classes can be a good approach to improve performance of the classification algorithms.

One of main lessons learned was to make sure to use oversampling algorithms only on the train data. Using these algorithms on all the data before splitting into the train and test would produce incorrect estimations of accuracy due to the bias. This is a very important point, as by using oversampling new data points will be created which are very similar to existing data. If the train/test split occurs after oversampling applied, there will be same data points in both train and test. Therefore, classification algorithms will learn and be tested on the same data. Further on, it will create a biased machine learning model which might not be performing well on unseen data.

6.2 Ethics

The ethics revolving around academic research studies will always be an important aspect of conducting research. Regarding this study, there does not appear to be any major concerns for a breach of ethics. We would like to make a note that we are only attempting to utilize the feature set within our data to predict a wine's quality, and not attempting to promote the consumption of alcohol to anyone based on the quality scores given.

6.3 Future Research

We would love to also take the results of this research and potentially apply it to other types of drinks besides wine. With some tweaks to the data/models, we could see this being used with other drinks that contain a similar physicochemical makeup. A tweak that might be necessary would be not using random forest classifier and to switch to something a little more simplistic. After the research, using the binary red wine data, a simple naïve bayes model was run and recall was checked compared to a baseline. In Table 4, it is clear there is improvement in recall when using sampling techniques on a simpler model.

Table 4. Red Wine Naïve Bayes Model Performance

Oversampling Algorithm	Recall
Base Case (No oversampling)	0.72
SMOTE	0.87
SVMSMOTE	0.81
SMOTEN	0.80
SMOTEENN	0.89
SMOTETomek	0.87
Random Oversampling	0.84
MDO	0.75
GlobalCS	0.84
SOUP	0.86

SMOTE, SMOTEENN, and SMOTETomek all had a 0.15 improvement in recall over the base. It would be very interesting to dive into these numbers more and possibly rerun the original analysis with a different model picked to analyze. It is worth mentioning that this study focuses on Portugues wine; which in many cases deals with wines that are made from a blend of multiple grapes, whereas in many other parts of the world, wines are made from just one grape. It is unclear how this affects the overall result in terms of the quality score of a wine, but it is something that should be investigated further if the required data is available to the researchers.

7 Conclusion

This research explored the idea of using oversampling and under sampling as ways to improve industry standard models using imbalanced data to classify wine by its quality. Based on the results, it was not proven that oversampling methods on imbalanced data would improve random forest classifiers when put into a multi-class problem. There was some evidence to suggest that over sampling, specifically SMOTEN, could improve a random forest classifier when faced with a binary case of predicting wine by its quality.

In addition, this research showed that the use of synthetic data runs the risk of decreasing model performance specifically when using ensemble models. There was also preliminary research showing that using a simpler model, such as naïve bayes, may be the route to take when predicating wine quality with sampling.

Acknowledgments
Jacquie Cheun. – Capstone Professor
Brook E. Ray – CSW, WSET II Sommelier and Wine Educator

References

0. Arefeen, M., Nimi, S., Rahman, M. (2022). Neural Network-Based Undersampling Techniques. IEEE transactions on systems, man, and cybernetics. Systems, 2022, Vol.52 (2), p.1111-1120

1. Caldas, J., & Rebelo, J. (2013). Portuguese wine ratings: An old product a new assessment. Wine Economics and Policy, 2(2), 102-110. 10.1016/j.wep.2013.11.004

2. Coli, M. S., Rangel, A. G. P., Souza, E. S., Oliveira, M. F., & Chiaradia, A. C. N. (2015, March). Chloride concentration in red wines: Influence of terroir and grape type. Food Science and Technology.

3. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4),.

4. Dal Pozzolo, A., Caelen, O., Bontempi, G. (2015). When is Undersampling Effective in Unbalanced Classification Tasks?. In: Appice, A., Rodrigues, P., Santos Costa, V., Soares, C., Gama, J., Jorge, A. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2015. Lecture Notes in Computer Science(), vol 9284. Springer, Cham.

5. Fernandez, A., Garcia S., Herrera, F., Chawla, N. (2018) SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. The Journal of artificial intelligence research, 2018, Vol.61, p.863-905

6. G. Hu, T. Xi, F. Mohammed and H. Miao, "Classification of wine quality with imbalanced data," 2016 IEEE International Conference on Industrial Technology (ICIT), 2016, pp. 1712-1217, doi: 10.1109/ICIT.2016.7475021.

7. Gupta, M., & C, V. (2021). A study and analysis of machine learning techniques in predicting wine quality. *International Journal of Recent Technology and Engineering (IJRTE)*, *10*(1), 314–319.

8. Hopfer, H., & Heymann, H. (2014). Judging wine quality: Do we need experts, consumers or trained panelists? Food Quality and Preference, 32, 221-233. 10.1016/j.foodqual.2013.10.004

9. Horváth, B. O., Sárdy, D. N., Kellner, N., & Magyar, I. (2020). Effects of High Sugar Content on Fermentation Dynamics and Some Metabolites of Wine-Related Yeast Species Saccharomyces cerevisiae, S. uvarum and Starmerella bacillaris. Food technology and biotechnology, 58(1), 76–83.

10. Janicka, M., Lango, M., Stefanowski, J. (2019) . Using Information on Class Interrelations to Improve Classification of Multiclass Imbalanced Data: A New Resampling Algorithm. International journal of applied mathematics and computer science, 2019, Vol.29 (4), p.769-781

11. J. Beh, E., & I. Holdsworth, C. (2012). A visual evaluation of a classification method for investigating the physicochemical properties of Portuguese wine. *Current Analytical Chemistry*, *8*(2).

12. Jingjun, B., Chongsheng, Z. (2018). An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme. Knowledge-based systems, 2018, Vol.158, p.81-93

13. Lango, M., Stefanowski, J. (2022) What makes multi-class imbalanced problems difficult? An experimental study. Expert systems with applications. 2022, Vol.199, p.116962

14. Lin, W., Tsai, C., Hu, Y., Jhang, J. (2017). Clustering-based undersampling in class-imbalanced data. Information sciences, 2017, Vol.409-410, p.17-26

15. Liu, A., Ghosh, J., & Martin, C. (2007, June). Generative Oversampling for Mining Imbalanced Datasets. In DMIN (pp. 66-72).

16. Mor, N. S. (2022). Wine quality and type prediction from physicochemical properties using neural networks for Machine Learning: A Free Software for winemakers and customers. Wine Quality and Type Prediction from Physicochemical Properties Using Neural Networks for Machine Learning: A Free Software for Winemakers and Customers.

17. Tanha, J., Abdi, Y., Samadi, N. et al. Boosting methods for multi-class imbalanced data classification: an experimental review. J Big Data 7, 70 (2020).

18. Yao, L., Lin, T. (2021). Evolutionary Mahalanobis Distance-Based Oversampling for Multi-Class Imbalanced Data Classification. Sensors (Basel, Switzerland), 2021, Vol.21 (19), p.6616

19. De Dios Santos, Juan. “Handling Imbalanced Datasets with SMOTE in Python” Kite, August 21. 2019, https://www.kite.com/blog/python/smote-python-imbalanced-learn-for-oversampling/. How does SMOTE work.