**Rajarambapu Institute Of Technology, Rajaramnagar**



**Department of Computer Science and Engineering**

**Project Report**

|  |  |
| --- | --- |
| **Area of the Project** | Machine Learning |
| **Title of the project** | Wine Quality Prediction Model |
| **Project Guide Name** | Prof. V. Lokare. |
| **Group Number** | G5 (Div. A) |

**Members**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No.** | **Roll No.** | **Name** | **Email** | **Sign** |
| 1 | 1903024 | Aniruddha Vaze | 1903024@ritindia.edu |  |
| 2 | 1903027 | Aryan Jadhav | 1903027@ritinida.edu |  |
| 3 | 1903046 | Digvijay Gore | 1903046@ritindia.edu |  |
| 4 | 1903055 | Pruthvi Belgaonkar | 1903055@ritindia.edu |  |

|  |  |  |
| --- | --- | --- |
| **Prof. V. Lokare** |  | **Dr. N. V. Dharwadkar** |
| **Project Guide** |  | **Head of department** |

# 

# **Introduction**

The quality of the wine is a very important part for the consumers as well as the manufacturing industries. Industries are increasing their sales using product quality certification. Nowadays, all over the world wine is a regularly used beverage and the industries are using the certification of product quality to increases their value in the market. Previously, testing of product quality will be done at the end of the production, this is time taking process and it requires a lot of resources such as the need for various human experts for the assessment of product quality which makes this process very expensive. Every human has their own opinion about the test, so identifying the quality of the wine based on humans experts it is a challenging task.

There are several features to predict the wine quality but the entire features will not be relevant for better prediction. The research aims to what wine features are important to get the promising result by implementing the machine learning classification algorithms such as Support Vector Machine (SVM), Naïve Bayes (NB), and Artificial Neural Network (), using the wine quality dataset. The wine quality dataset is publically available on the UCI machine learning repository (Cortez et al., 2009). The dataset has two files red wine and white wine variants of the Portuguese “Vinho Verde” wine. It contains a large collection of datasets that have been used for the machine learning community. The red wine dataset contains 1599 instances and the white wine dataset contains 4898 instances. Both files contain 11 input features and 1 output feature. Input features are based on the physicochemical tests and output variable based on sensory data is scaled in 11 quality classes from 0 to 10 (0-very bad to 10-very good).

Feature selection is the popular data preprocessing step for generally (Wolf and Shashua, 2005). To build the model it selects the subset of relevant features. According to the weighted of the relevance of the features, and with relatively low weighting features will be removed. This process will simplify the model and reduce the training time, and increase the performance of the model (Panday et al., 2018). We pay attention to feature selection is also the study direction. To evaluate our model, accuracy, precision, recall, and f1 score are good indicators to evaluate the performance of the model.

The report is divided into 7 sections, including this one. In Section 2 we discuss the background and related work. In Section 3 we formulate our research question and hypothesis. Section 4 describes the methodologies. Section 5 discusses the experimental design. In Section 6 results and discussion of the whole work. In Section 7 we discuss the conclusions and future work.

# **Background**

A wide range of machine learning algorithms is available for the learning process. This section describes the classification algorithms used in wine quality prediction and related work.

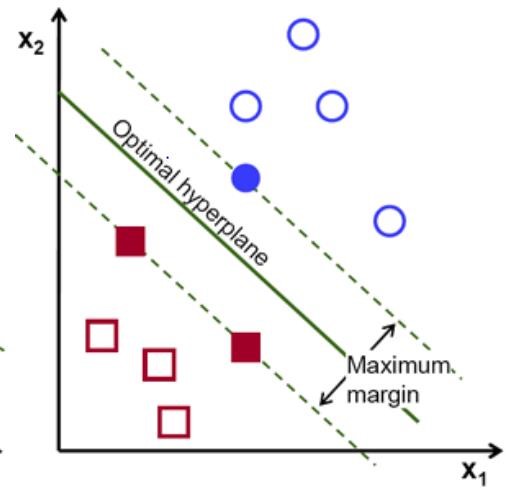
## 

## **Classification algorithm**

### **Support Vector Machine**

The support vector machine (SVM) is the most popular and most widely used machine learning algorithm. It is a supervised learning model that can perform classification and regression tasks. However, it is primarily used for classification problems in machine learning (Gandhi, 2018).

The SVM algorithm aims to create the best line or decision boundary that can separate n-dimensional space into classes. So we can put the new data points easily in the correct groups. This best decision boundary is called a hyperplane.



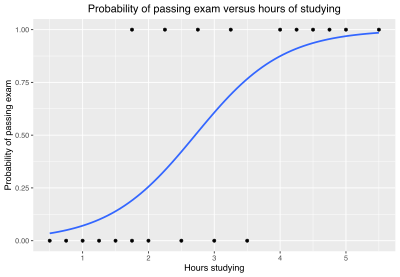
The support vector machine selects the extreme data points that helping to create the hyperplane. In Figure 1, two different groups are classified by using the decision boundary or hyperplane:

The SVM model is used for both non-linear and linear data. It uses a nonlinear mapping to convert the main preparing information into a higher measurement. The model searches for the linear optimum splitting hyperplane in this new measurement. A hyperplane can split the data into two classes with an appropriate nonlinear mapping to suitably high measurements and for the finding, this hyperplane SVM uses the support vectors and edges (J. Han et al., 2012). The SVM model is a representation of the models as a point in space, the different classes are isolated by the gap to mapped with the aim that instances are wide as would be careful. The model can perform out a nonlinear form of classification (Kumar et al., 2020).

### **Logistic Regression**

### 

In [statistics](https://en.wikipedia.org/wiki/Statistics), the logistic model (or logit model) is a [statistical model](https://en.wikipedia.org/wiki/Statistical_model) that models the probability of an event taking place by having the [log-odds](https://en.wikipedia.org/wiki/Log-odds) for the event be a linear combination of one or more independent variables. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (the coefficients in the linear combination). Formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1", while the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling, the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit, from **log**istic un**it**, hence the alternative names.



The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification (it is not a classifier), though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other; this is a common way to make a binary classifier.

Analogous linear models for binary variables with a different sigmoid function instead of the logistic function (to convert the linear combination to a probability) can also be used, most notably the probit model.

### **Random Forest Classifier**

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

Decision trees are a popular method for various machine learning tasks. Tree learning "come[s] closest to meeting the requirements for serving as an off-the-shelf procedure for data mining", "because it is invariant under scaling and various other transformations of feature values, is robust to inclusion of irrelevant features, and produces inspectable models. However, they are seldom accurate".

In particular, trees that are grown very deep tend to learn highly irregular patterns: they overfit their training sets, i.e. have [low bias, but very high variance](https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff). Random forests are a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of reducing the variance. This comes at the expense of a small increase in the bias and some loss of interpretability, but generally greatly boosts the performance in the final model.



## **Problem Statement**

Based on the articles reported in section 2.2, the significance of each feature for the wine quality prediction is not yet quantified. And in terms of performance, the current accuracy is about 67.25%. Thus, in this thesis, we considered two aspects of the problems mentioned above. The first one is the study of the importance of the features for the prediction of wine quality. The secondly, performance of the prediction model can be improved using a neural network with other ordinary classifiers used by the articles cited above.

The following research question and hypothesis are formulated.

1. What wine features are important to get a promising result?

The researchers have used a neural network for the regression task but for the classification task neural network was never used. Hypothetically, the current prediction model that has been obtained by researchers will be improved by using the neural network.

To address the research question the following objectives are formulated :

* To balance the dataset.
* To analyze the impact of the features.
* To optimize the classification models through hyperparameter tuning.
* To model and evaluate the approaches.

# 

# **Method and Approach**

## **Data Description**

The red wine and white wine datasets have been used in this paper which is obtained from the UCI machine learning repository it contains a large collection of datasets that have been used for the machine learning community. The dataset contains two excel files, related to red wine and white wine variants of the Portuguese “Vinho Verde” wine (Cortez et al., 2009). The red wine dataset contains 1599 instances and the white wine dataset contains 4898 instances. Both datasets have 11 input variables (based on physicochemical tests): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulfates, alcohol, and 1 output variable (based on sensory data): quality. Sensory data is scaled in 11 quality classes from 0 to 10 (0-very bad to 10-very good). Below Table 1 description of the attributes.

*Table 1: Attribute description*

|  |  |
| --- | --- |
| **Attributes Description** | |
| **Fixed acidity** | Fixed acids, numeric from 3.8 to 15.9 |
| **Volatile acidity** | Volatile acids, numeric from 0.1 to 1.6 |
| **Citric acid** | Citric acids, numeric from 0.0 to 1.7 |
| **Residual sugar** | residual sugar, numeric from 0.6 to 65.8 |
| **Chlorides** | Chloride, numeric from 0.01 to 0.61 |
| **Free sulfur dioxide** | Free sulfur dioxide, numeric: from 1 to 289 |
| **Total sulfur dioxide** | Total sulfur dioxide, numeric: from 6 to 440 |
| **Density** | Density, numeric: from 0.987 to 1.039 |
| **pH** | pH, numeric: from 2.7 to 4.0 |
| **Sulfates** | Sulfates, numeric: from 0.2 to 2.0 |
| **Alcohol** | Alcohol, numeric: from 8.0 to 14.9 |
| **Quality** | Quality, numeric: from 0 to 10, the output target |

## **Feature** **selection**

Feature selection is the method of selection of the best subset of features that will be used for classification (Fauzi et al., 2017). Most of the feature selection method is divided into a filter and wrapper, the filter uses the public features work individually from the learning algorithm and the wrapper evaluates the features and chooses attributes based on the estimation of the accuracy by using a search algorithm and specific learning model (Onan and Korukoğlu, 2017).

In this study, for a better understanding of the features and to examines the correlation between the features. The Pearson correlation coefficient is calculated for each feature in Table 1, this shows the pairwise person correlation coefficient P, which is calculated by using the below formula (Dastmard, 2013).

cov (X, Y)

𝑃𝑥,𝑦 =

𝜎𝑋, σY

Where the 𝜎 is the standard deviation of the features X and Y and cov is the covariance. The range of the correlation coefficient from -1 to 1. Point 1 value implies linear equation is describes the correlation between X and Y strong positive, which is all data points are lying on a line for Y increases as X increases. Where point -1 value indicates that strong negative correlations between data points. All data points lie on a line in which Y decreases as X increases. And point 0 indicates that there is an absence of correlation between the points (Dastmard, 2013).

## **Evaluation**

The performance measurement is calculated and evaluate the techniques to detect the effectiveness and efficiency of the model. There are four ways to check the predictions are correct or incorrect:

* True Positive: Number of samples that are predicted to be positive which are truly positive.
* False Positive: Number of samples that are predicted to be positive which are truly negative.
* False Negative: Number of samples that are predicted to be negative which are truly positive.
* True Negative: Number of samples that are predicted to be negative which are truly negative.

Below listed techniques, we use for the evaluation of the model.

Accuracy – Accuracy is defined as the ratio of correctly predicted observation to the total observation. The accuracy can be calculated easily by dividing the number of correct predictions by the total number of predictions.

True Positive + True Negative

Accuracy =

True Positive + False Positive + False Negative + True Negative

Negative + True Negative

1. Precision – Precision is defined as the ratio of correctly predicted positive observations to the total predicted positive observations.

True Positive

Precision =

True Positive + False Positive

1. Recall – Recall is defined as the ratio of correctly predicted positive observations to all observations in the actual class. The recall is also known as the True Positive rate calculated as,

True Positive

Recall =

True Positive + False Negative

1. F1 Score – F1 score is the weighted average of precision and recall. The f1 score is used to measure the test accuracy of the model. F1 score is calculated by multiplying the recall and precision is divided by the recall and precision, and the result is calculated by multiplying two.

Recall ∗ Precision

F1 score = 2 ∗

Recall + Precision

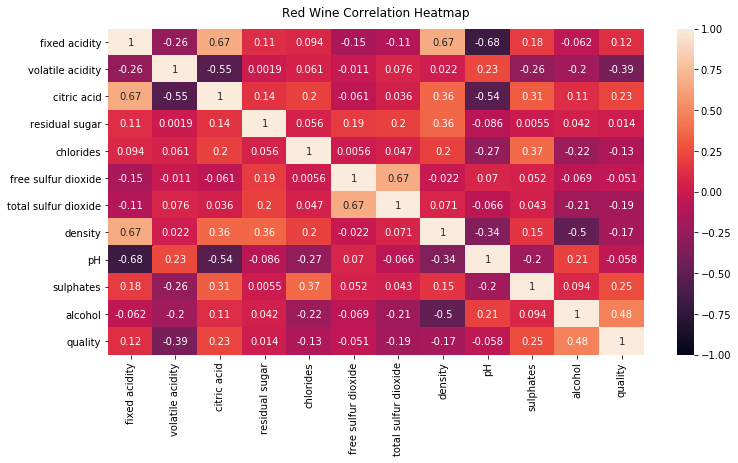
Accuracy is the most widely used evaluation metric for most traditional applications. But the accuracy rate is not suitable for evaluating imbalanced data sets, because many experts have observed that for extremely skewed class distributions, the recall rate for minority classes is typically 0, which means that no classification rules are generated for the minority class. Using the terminology in information retrieval, the precision and recall of the minority categories are much lower than the majority class. Accuracy gives more weight to the majority class than to the minority class, this makes it challenging for the classifier to implement well in the minority class.

For this purpose, additional metrics are coming into widespread usage (Guo et al., 2008).

The F1 score is the popular evaluation matric for the imbalanced class problem (Estabrooks and Japkowicz, 2001). F1 score combines two matrices: precision and recall. Precision state how accurate the model was predicting a certain class and recall state that the opposite of the regrate misplaced instances which are misclassified. Since the multiple classes have multiple F1 scores.

## **Feature Selection**

For a better understanding of the features and to examines the correlation between the features. We use the Pearson coefficient correlation matrices to calculate the correlation between the features.



*Figure 5: correlation matrices red wine*

From Figure 5 red wine correlation matrix we ranked the features according to the high correlation values to the quality class such as freatures are 'alcohol', 'volatile acidity', 'sulphates', 'citric acid', 'total sulfur dioxide', 'density', 'chlorides', 'fixed acidity', 'pH', 'free sulfur dioxide', 'residual sugar'.

## **Data Standardization**

Scikit-learn is a python module, it integrates the newest machine learning algorithm for supervised and unsupervised problems (Pedregosa et al., 2011).

The data standardization technique can scale the features among 0 and 1, it will be useful for learning the model, by applying it to all the numeric features and then separating data by standard derivation (Pedregosa et al., 2011). So, we use this technique to standardize the data. The formula of standardization is :

zi = xi – u

σ

σ is the standard derivation, xi is each value, and u is the mean value of the array x.

## **Data Separation**

The scikit-learn library is splitting the data into a training and testing set. So we split the dataset test size is equal to 0.2. The train test split method randomly splits the sample data into the testing set and the training set, so this will avoid the unseen division of the sample data.

## **Model and Evaluation**

For the implementation of the model, we used machine learning algorithms such as support vector machine (SVM), naïve Bayes (NB), and artificial neural network (). To adobe algorithms, we use the scikit-learn python machine learning libraries (scikit-learn, 2021).

The evaluation results were achieved from each implementation of the classification algorithm calculated. As mentioned in the Evaluation sub-section.

## **Model Results**

The importance of the features are identified and from both dataset's first 10 features were selected and the last feature was excluded, above red wine performance analysis Figure 7 and white wine performance analysis Figure 8 shows that the performance in terms of accuracy.

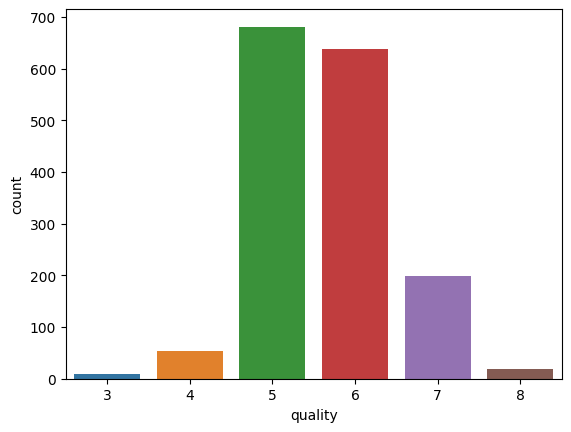
Firstly, these selected features were implemented on the unbalanced classes, Figure 3 shows the unbalanced classes and the performance of the prediction model, in terms of accuracy, precision, recall, and F1 score is examined, as expressed in Table 4 red wine and Table 5 white wine. Then these selected features were implemented on the balanced class, Figure 4 shows that the balancing of each class and the performance of the prediction model, in terms of accuracy, precision, recall, and f1 score is examined, as expressed in Table 6 red wine and Table 7 white wine.

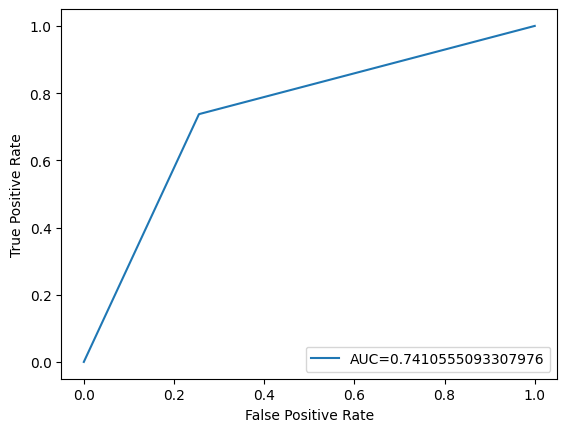
## **Observation and Results:-**

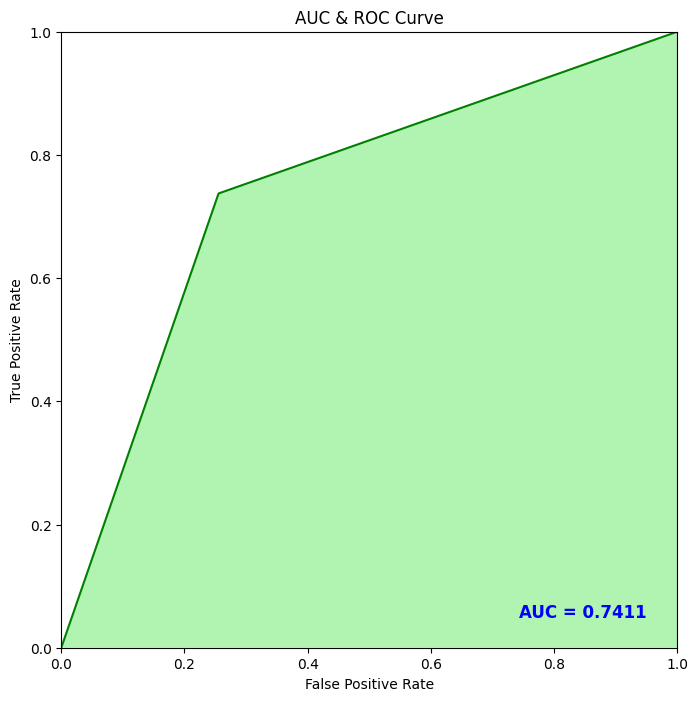
## 

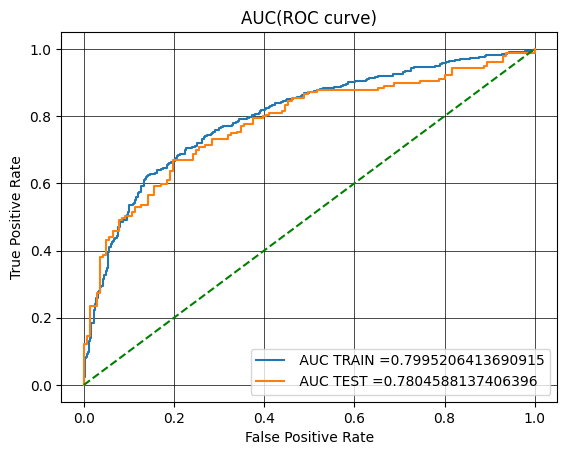
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ML Model | Precision | Recall | F1-score | Support |
| Logistic Regression | 0.69 | 0.74 | 0.72 | 141 |
| Support Vector Machine RBF | 0.77 | 0.77 | 0.82 | 320 |
| Random Forest | 0.82 | 0.80 | 0.81 | 179 |

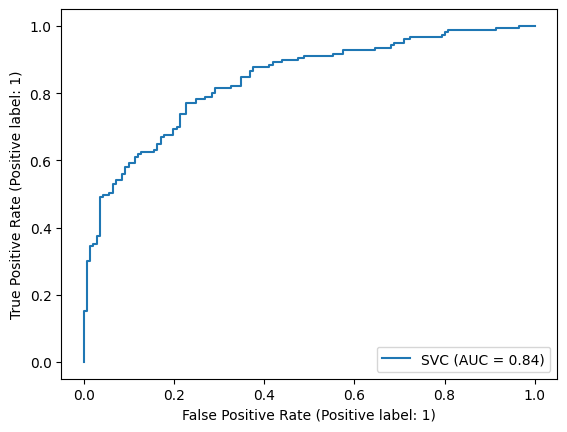
**Graphical Analysis :-**

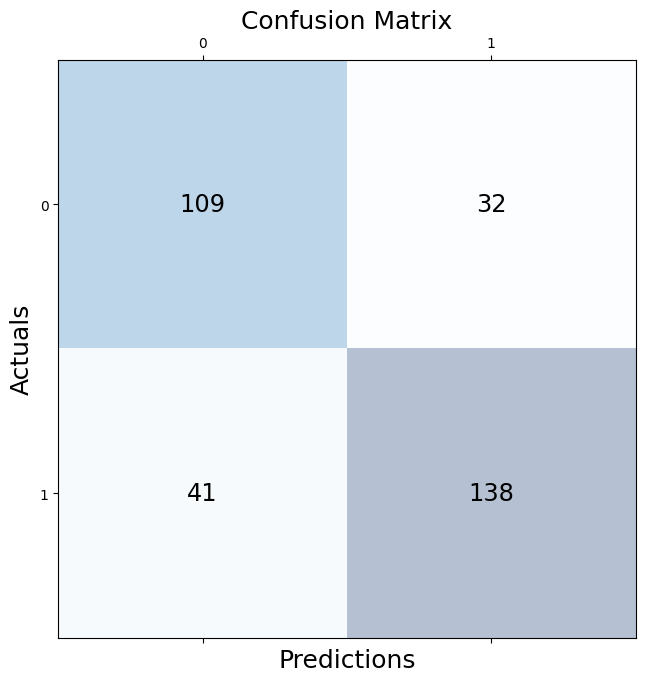


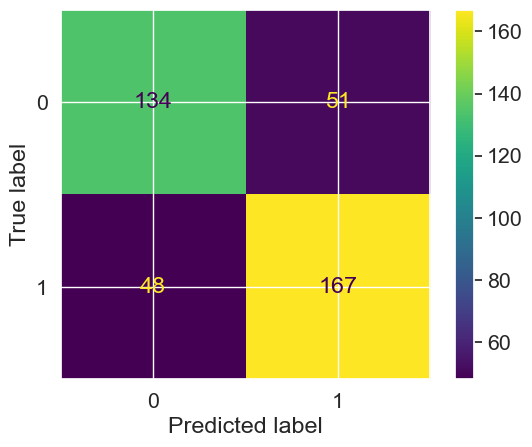












# **Conclusions**

This report uses the two types of wine dataset red and white, of Portuguese “Vinho Verde” wine to predict the quality of the wine based on the physicochemical properties.

First, we used oversampling to balance the dataset in the data preprocessing stage to optimize the performance of the model. Then we look for features that can provide better prediction results. For this, we used Pearson coefficient correlation matrices and ranked the features according to the high correlation among the features. After applying the sampling datasets which is balancing dataset the performance of the model is improved. In general, removing irrelevant features of the datasets improved the performance of the classification model. To conclude that the minority classes of a dataset will not get a good representation on a classifier and representation for each class can be solved by oversampling and undersampling to balance the representation classes over datasets.

The accuracy of the support vector machine (SVM) algorithm is

83.52% from the red wine and 86.86% from the white wine, the naïve Bayes (NB) algorithm is 46.33% from the red wine and 46.68% from the white wine, and the artificial neural network () is 85.16% from the red wine and 88.28% accuracy from the white wine. Among these three machine learning algorithms, we achieved the best accuracy result from the artificial neural network () on both red and white wine datasets. Therefore, in the classification algorithms by selecting the appropriate features and balancing the data can improve the performance of the model.

## **Future Work**

In the future, to improve the accuracy of the classifier, it is clear that the algorithm or the data must be adjusted. We recommend feature engineering, using potential relationships between wine quality, or applying the boosting algorithm on the more accurate method.

In addition, by applying the other performance measurement and other machine learning algorithms for the better comparison on results. This study will help the manufacturing industries to predict the quality of the different types of wines based on certain features, and also it will be helpful for them to make a good product.

## **LocalHost URL**

http://localhost:8501

## 

# **References**

Chawla, N.V., 2005. Data Mining for Imbalanced Datasets: An Overview, in: Maimon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook. Springer US, Boston, MA, pp. 853–867. https://doi.org/10.1007/0-387-25465-X\_40

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J., 2009. Modeling wine preferences by data mining from physicochemical properties.

Decis. Support Syst. 47, 547–553.

https://doi.org/10.1016/j.dss.2009.05.016

Dahal, K., Dahal, J., Banjade, H., Gaire, S., 2021. Prediction of Wine Quality Using Machine Learning Algorithms. Open J. Stat. 11,

278–289. https://doi.org/10.4236/ojs.2021.112015

Dastmard, B., 2013. A statistical analysis of the connection between test results and field claims for ECUs in vehicles.

Drummond, C., Holte, R.C., 2003. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling. pp. 1–8.

Er, Y., Atasoy, A., 2016. The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. Int. J. Intell. Syst.

Appl. Eng. 4, 23–26. https://doi.org/10.18201/ijisae.265954 Er, Y., Atasoy, Ayten, 2016. The Classification of White Wine and Red Wine According to Their Physicochemical Qualities. Int. J. Intell.

Syst. Appl. Eng. 23–26. https://doi.org/10.18201/ijisae.265954

Estabrooks, A., Japkowicz, N., 2001. A Mixture-of-Experts Framework for Learning from Imbalanced Data Sets, in: Hoffm, F., Hand, D.J., Adams, N., Fisher, D., Guimaraes, G. (Eds.), Advances in

Intelligent Data Analysis, Lecture Notes in Computer Science.

Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 34–43.

https://doi.org/10.1007/3-540-44816-0\_4

Fauzi, M., Arifin, A.Z., Gosaria, S., Prabowo, I.S., 2017. Indonesian News

Classification Using Naïve Bayes and Two-Phase Feature Selection Model. Indones. J. Electr. Eng. Comput. Sci. 8, 610–615.

https://doi.org/10.11591/ijeecs.v8.i3.pp610-615

Gandhi, R., 2018. Support Vector Machine — Introduction to Machine Learning Algorithms [WWW Document]. Medium. URL https://towardsdatascience.com/support-vector-machineintroduction-to-machine-learning-algorithms-934a444fca47 (accessed 6.6.21).