

Red Wine Classification Using SVM and RBF Kernel

Kevin Silvanus Hutabarat ^{1*}, Rosalia Arum Kumalasanti¹

¹*Department of Informatics, Faculty of Science and Technology, Sanata Dharma University, Yogyakarta, Indonesia*
Corresponding Author: kevinilvanus@gmail.com

(Received 27-10-2023; Revised 27-06-2024; Accepted 01-08-2024)

Abstract

Today's cultural diversity has influenced lifestyle, especially at the time of certain events. Food and drink are important in the event. Quality food is a key ingredient in a person's eating. Red wine is one of the most popular beverages in the West because of its cold climate but today Red wine has become a popular drink not only among Western countries. The love of red wine should also be balanced with his knowledge of the quality of the beverage because it has various variations. The duration of fermentation and the materials used will give different quality products. Nowadays technology is present to provide solutions to these problems by using the SVM algorithm on Machine Learning. The approach is carried out by performing several experiments to obtain optimal evaluation results. The study has achieved accuracy of 90.93%, precision of 72.5% and recall of 61.70%.

Keywords: SVM, machine learning, accuracy, precision, recall

1 Introduction

The recent cultural diversity has influenced people's lifestyles, especially in celebrating certain events. The event usually has a companion dish which is the main attraction in enlivening the event. Often classy dishes and drinks become a specific value at the event. One drink that is often served at certain events is "anggur merah" or red wine. Red Wine is an alcoholic drink made from red grapes which undergoes a fermentation process. The fermentation process here is the process of dipping the grape skins and seeds into the squeezed fruit juice and letting it sit for some time [1]. Red Wine is a drink that is commonly consumed in western countries, because the climate conditions are very cold so this drink is quite popular there.



In fact, Red Wine is already quite popular apart from in western countries, but not all Red Wine fans know the quality of this drink. The length of time during the fermentation process will provide different variations in alcohol levels and this will affect the quality of each Red Wine. Until now, Red Wine has various variants and this drink still has many fans. The number of variants, the various quality produced. Red Wine Producers can of course easily determine the quality of the Red Wine to be produced, but consumers must also know that choosing a Red Wine variant based on quality will also be a very important thing to know before buying.

In determining the quality of Red Wine, the current technology is here to provide convenience in determining quality. It's well known that producing wine is an age-old craft that necessitates in depth understanding of the circumstances and ingredients that a wine may contain. Since wine consumption has increased across the board throughout the epidemic, wineries should look into less expensive ways to improve wine quality [2].

One method that is quite popular today is Machine Learning. Machine Learning is a sub-field of artificial intelligence which is still frequently used today. Machine Learning enables software applications to be more accurate in predicting outcomes without being explicitly programmed. One of the Machine Learning algorithms that is often used for classification is SVM. SVM (Support Vector Machine) is one of the Machine Learning algorithms used for classification and regression. The primary objective of SVM in classification tasks is to find the best line or hyperplane that can separate two different data classes as best as possible. The SVM was used by a team of researchers to identify the quality of red wine with an accuracy of 87.5% [1]. Other research has also been done by comparing the Naïve Bayes algorithm with the SVM in classifying wine. These studies yield significant results with fairly good accuracy [3].

The SVM on this study will be used to classify the quality of red wine by setting the correct parameters. It is hoped that SVM can provide optimal accuracy results in classifying red wine by quality, so that red wine lovers can use it in balancing quality before deciding to buy.

SVM is an algorithm to help classify data. Classification is a process involving the creation of models that can be used to predict a class or category of an object based on a set of attributes or features given. The aim is to generalize patterns from training data that

are known as class labels on new data that does not have labels [4] In the process of classification, there are several concepts that need to be understood:

Data Training: Data sets used to train classification models. Each example in this data set has an attribute or feature that is associated with a known class label.

Testing: Once the classification model is built, the next step is to test the model using test data that has never been used before. Test data to measure model performance in predicting classes of unknown objects. The evaluation results of accuracy, precision, and recall are used to evaluate the performance of the classification model.

2 Material and Methods

Red Wine Data Set

Based on experiments, it is noticed that the majority of the chemical components used in the manufacture of wine are the same for different wines, and that the influence or concentration of each grade of chemical composition varies depending on the kind of wine. The objective of this case study was to forecast a wine's quality based on feature sets that were provided as inputs and output rating scale ranging from 0 to 10. With regard to red wine, the quality is measured on a scale with values of [3,4,5,6,7,8]. The quality gets better as the scale value increases (3 = lowest and 8 = highest) [2]

Preprocessing

Preprocessing is a set of steps that must be taken before applying data analysis methods or building models. The goal is to prepare the data in accordance with the requirements of the algorithm to be used, as well as maximize the quality of the analysis to be performed [5].

Kernel Function

The kernel on SVM is one of the key concepts that enables SVM to deal with nonlinear problems effectively. The kernel functions on the SVM perform the transformation of data from the original feature space to the higher feature space. There are several common types of kernels that are often used in SVM, such as [4]:

- **Linear Kernel:** is the simplest kernel and only performs linear mapping of data into the same feature space.
- **Polynomial Kernel:** This kernel carries data into a higher feature space using a polynomic function.
- **Radial Base Function (RBF) Kernel:** This kernel is very commonly used to mapp data into an unlimited feature space using the Gauss function. Gamma parameters need to be set to control how sharply the function of the kernel decreases.
- **Sigmoid Kernel:** This kernel also performs nonlinear transformations and is used in some special cases.
- **Custom Kernels:** In addition to the above-mentioned kernels, there are also specific custom kernels to customize classification/regression tasks.

Support Vector Machine

SVM is a powerful and well-known classification method that operates on the principles of optimization theory and utilizes kernel function. When there are fewer training samples available and the image is represented in more spectral bands, SVM has been shown to be more effective classification tool [6] [7]. SVM is a technique for making predictions, both in the case of regression and classification. The SVM technique is used to obtain an optimal hyperplane to separate observations that have different target variable values [8]. Hyperplane visualization can be seen in Fig. 1 where there are two classes separated by hyperplane lines. The best separator between the two classes can be found by measuring the margin and looking for the maximum point. Margin is the distance between the hyperplane and the closest pattern of each class. The nearest pattern is called a support vector.

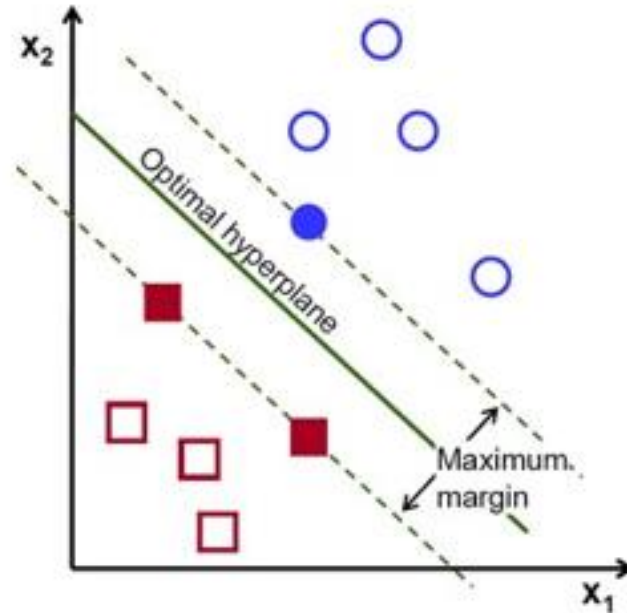


Figure 1. Visualization of discrimination [4].

Hyperplane can be searched using the following equation:

$$W \cdot X_i + b = 0 \tag{1}$$

Where:

W: Weight of vector

X_i : Value of attribute

B: Bias

In this equation there are two classes: positive class and negative class. There are patterns that are members of these two classes. Positive and negative classes can be obtained from the following equations:

$$W \cdot X_i + b \geq +1 \text{ if } y_i = +1 \tag{2}$$

$$W \cdot X_i + b \leq -1, \text{ if } y_i = -1 \tag{3}$$

In obtaining the best hyperplane is by maximizing the margin or distance between two sets of objects of each class.

Confusion Matrix

The confusion matrix is very useful for analyzing the quality of the classifier in recognizing the tuples of existing classes. In the evaluation process of classification there are four possibilities that occur from the classification process. The result of binary classification on a dataset can be represented in Table 1 below.

Table 1. Binary classification on a dataset.

Prediction	Actual		
	Class	Positive	Negative
	Positive	True Positive	False Negative
Negative	False Positive	True Negative	

There are several common formulas used to calculate the performance of classification, namely:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{4}$$

$$Precision = \frac{True\ Positive}{True\ Positive+False\ Positive} \tag{5}$$

$$Recall = \frac{True\ Positive}{True\ Positive+False\ Negative} \tag{6}$$

This study goes through several stages and is the steps that will be taken to obtain optimal results. The results can be seen in Fig. 2 below.

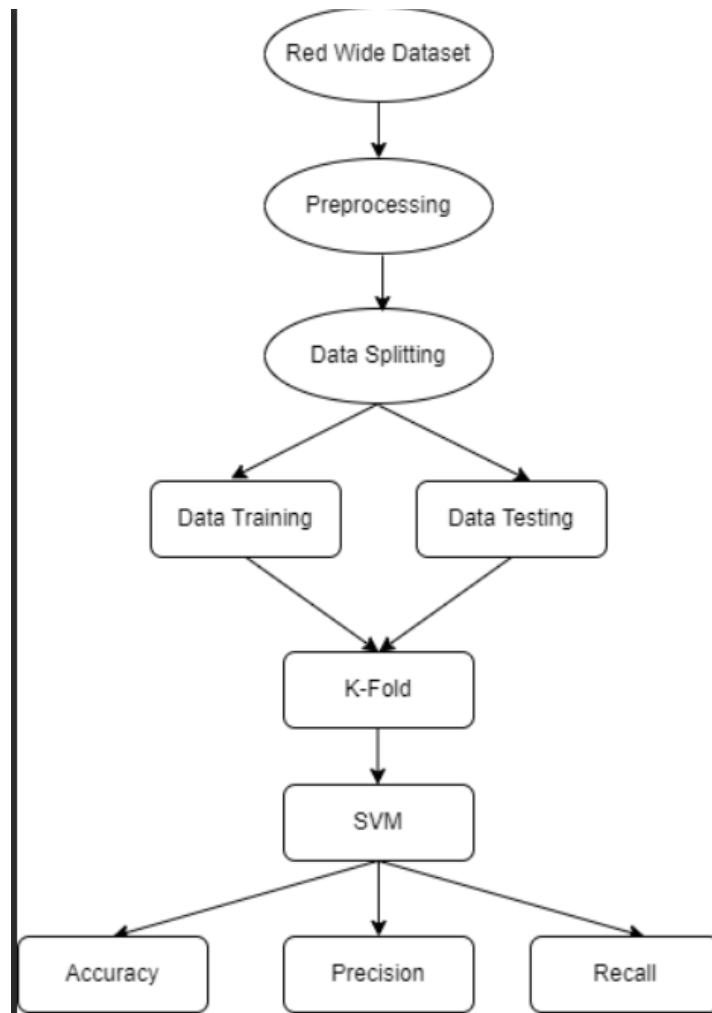


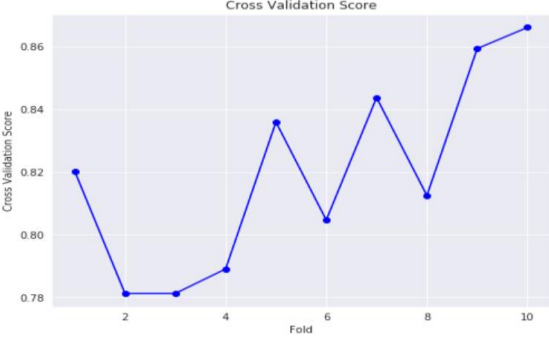
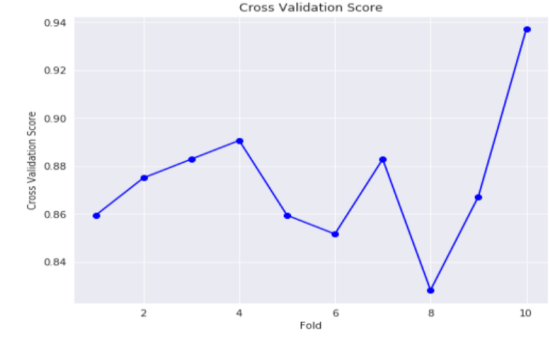
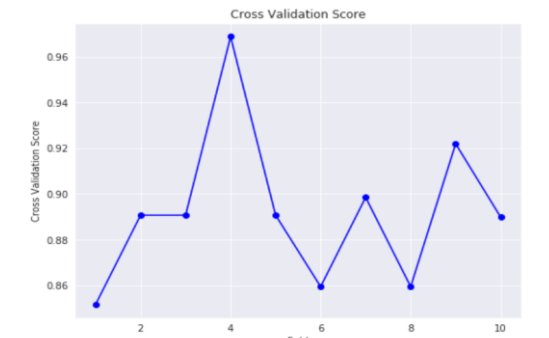
Figure 2. Flow of Research

3 Results and Discussions

The study used a data set of Red Wine with 12 attributes namely solid acid, volatile acid, citric acid, residual sugar, chloride, free sulfur, total sulfur dioxide, density, pH, sulfate, alcohol and quality. The quality limit for red wines that have good quality is with the label 2-6, while the bad quality is 7-8 on the label.

SVM offers several kernels to use, so in this study it is necessary to compare several nuclei to get optimal results. The sigmoid, polynomial and RBF kernels have been made and can be seen in Table 2 below.

Table 2. SVM’s Kernel

<p style="text-align: center;">Kernel Sigmoid</p>  <table border="1" data-bbox="347 376 898 712"> <caption>Kernel Sigmoid Cross Validation Score Data</caption> <thead> <tr> <th>Fold</th> <th>Cross Validation Score</th> </tr> </thead> <tbody> <tr><td>1</td><td>0.82</td></tr> <tr><td>2</td><td>0.78</td></tr> <tr><td>3</td><td>0.78</td></tr> <tr><td>4</td><td>0.79</td></tr> <tr><td>5</td><td>0.835</td></tr> <tr><td>6</td><td>0.805</td></tr> <tr><td>7</td><td>0.845</td></tr> <tr><td>8</td><td>0.815</td></tr> <tr><td>9</td><td>0.855</td></tr> <tr><td>10</td><td>0.86</td></tr> </tbody> </table>	Fold	Cross Validation Score	1	0.82	2	0.78	3	0.78	4	0.79	5	0.835	6	0.805	7	0.845	8	0.815	9	0.855	10	0.86	<p>Accuracy: 85.3 Precision: 50.0 Recall :59.57</p>
Fold	Cross Validation Score																						
1	0.82																						
2	0.78																						
3	0.78																						
4	0.79																						
5	0.835																						
6	0.805																						
7	0.845																						
8	0.815																						
9	0.855																						
10	0.86																						
<p style="text-align: center;">Kernel Polynomial</p>  <table border="1" data-bbox="347 795 898 1131"> <caption>Kernel Polynomial Cross Validation Score Data</caption> <thead> <tr> <th>Fold</th> <th>Cross Validation Score</th> </tr> </thead> <tbody> <tr><td>1</td><td>0.86</td></tr> <tr><td>2</td><td>0.875</td></tr> <tr><td>3</td><td>0.885</td></tr> <tr><td>4</td><td>0.89</td></tr> <tr><td>5</td><td>0.86</td></tr> <tr><td>6</td><td>0.85</td></tr> <tr><td>7</td><td>0.885</td></tr> <tr><td>8</td><td>0.83</td></tr> <tr><td>9</td><td>0.87</td></tr> <tr><td>10</td><td>0.935</td></tr> </tbody> </table>	Fold	Cross Validation Score	1	0.86	2	0.875	3	0.885	4	0.89	5	0.86	6	0.85	7	0.885	8	0.83	9	0.87	10	0.935	<p>Accuracy : 90.93 Precision : 72.5 Recall :61.7</p>
Fold	Cross Validation Score																						
1	0.86																						
2	0.875																						
3	0.885																						
4	0.89																						
5	0.86																						
6	0.85																						
7	0.885																						
8	0.83																						
9	0.87																						
10	0.935																						
<p style="text-align: center;">Kernel RBF</p>  <table border="1" data-bbox="347 1214 898 1550"> <caption>Kernel RBF Cross Validation Score Data</caption> <thead> <tr> <th>Fold</th> <th>Cross Validation Score</th> </tr> </thead> <tbody> <tr><td>1</td><td>0.85</td></tr> <tr><td>2</td><td>0.89</td></tr> <tr><td>3</td><td>0.89</td></tr> <tr><td>4</td><td>0.965</td></tr> <tr><td>5</td><td>0.89</td></tr> <tr><td>6</td><td>0.86</td></tr> <tr><td>7</td><td>0.895</td></tr> <tr><td>8</td><td>0.86</td></tr> <tr><td>9</td><td>0.92</td></tr> <tr><td>10</td><td>0.89</td></tr> </tbody> </table>	Fold	Cross Validation Score	1	0.85	2	0.89	3	0.89	4	0.965	5	0.89	6	0.86	7	0.895	8	0.86	9	0.92	10	0.89	<p>Accuracy :90.93 Precision : 72.5 Recall : 61.70</p>
Fold	Cross Validation Score																						
1	0.85																						
2	0.89																						
3	0.89																						
4	0.965																						
5	0.89																						
6	0.86																						
7	0.895																						
8	0.86																						
9	0.92																						
10	0.89																						

This research also carries out several approaches by processing cost parameters to obtain optimal results. The cost experiment can be seen in Fig. 3 below.

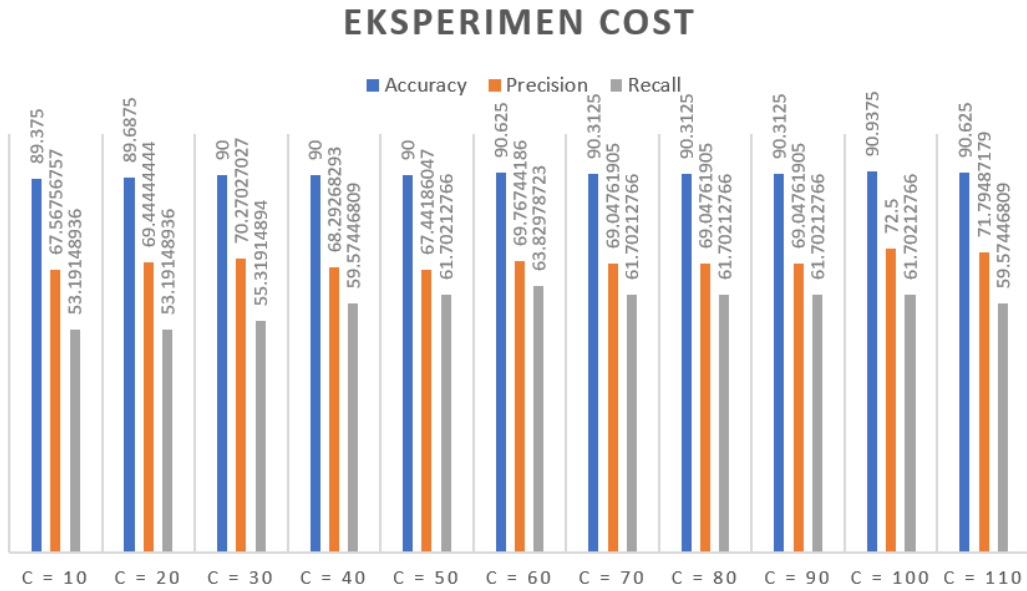


Figure 3. Experimental cost graph

After performing the fitting classifier then the next step is to determine the most optimal K-Fold Cross Validation by performing several K values experiments. K-fold cross validation is a model validation technique commonly used in machine learning. Experiments to obtain the best K value have been carried out ranging from K=1 to K=11 and in can K=10 the most optimal. The experiment can be seen in Fig. 4 below.

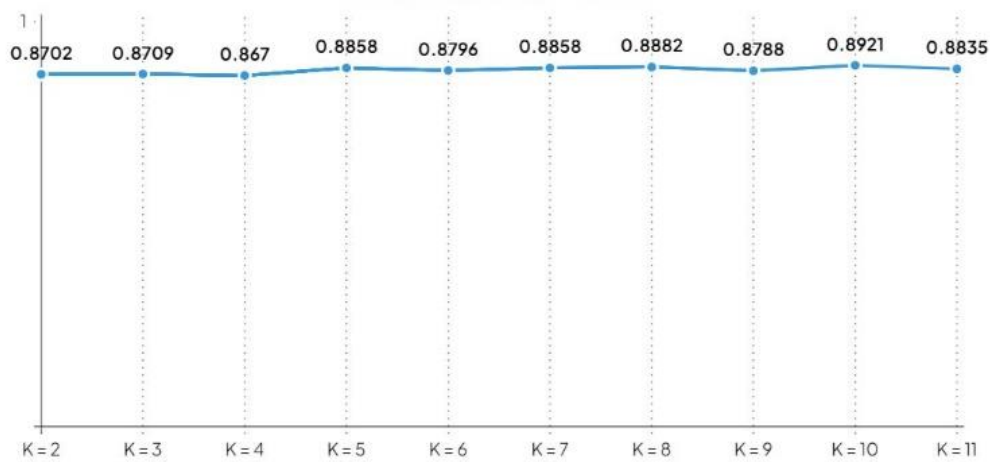


Figure 4. K-fold experiment

4 Conclusions

This research has been carried out using the SVM approach and some parameters in it. The most important parameters are cost, tolerance, gamma and degree. The Radial Basic Function (RBF) kernel gets the most optimal outcomes compared to other kernels. The evaluation results showed that optimum accuracy was achieved at 90.93%, precision at 72.5% and recall at 61.70%.

Acknowledgements

The work is supported by the Faculty of Science Technology, University Sanata Dharma.

References

- [1] M. R. Kushalatha, P. Rachana, P. Sameep and S. B. Shreesha, "Machine Learning Approach for Attribute Identification and Quality Prediction of Red Wine," *An International Open Access*, vol. 9, no. 4, pp. 4906-4910, 2021.
- [2] A. C. Benjamin, "Wine Quality Classification Using Machine Learning Algorithms," *International Journal of Computer Applications Technology and Research*, vol. 11, no. 6, pp. 241-246, 2022.
- [3] M. Aslam, "Wine Quality Prediction By using Machine Learning Algorithms," *Global Scientific*, vol. 10, no. 12, pp. 631-649, 2022.
- [4] G. Bonaccorso, *Machine Learning Algorithms*, Brimingham: Packt, 2018.
- [5] C. M. Andreas and S. Guide, *Machine Learning With Python*, USA: O'reilly, 2018.
- [6] B. N. Ramya, S. Adithi and R. Kruthika, "Study on Red Wine Quality Detection Using Machine Learning," *International Journal of Research Publication and Reviews*, vol. 4, no. 12, pp. 1579-1587, 2023.
- [7] A. G. Bhavya, "Wine Quality Prediction Using Different Machine Learning Techniques," *International Journal of Science Engineering and Technology*, vol. 8, no. 4, pp. 1-6, 2020.
- [8] D. Haroon, *Python Machine Learning Case Studies*, Karachi, Pakistan: Apress, 2017.