–Classification of Wine Quality–

Ozan Can ALPER

April 2022

Contents

1 Introduction

Wine is one of the popular beverages as its consumption is very popular among generations and its quality is time-dependent, and generally the older wine tastes better. One of the growing research areas in engineering is machine learning. The quality of wine includes different properties such as alcohol content, pH value, and density. In this study, a classifier is trained to detect wine quality over different characteristics with an approach using a deep Multilayer Perceptron (MLP) neural network. It is shown how the number of different hidden layers and the number of neurons affect the success of the classifier. Apart from this, the effect of the data augmentation method on the system is shown.

2 Data set

Since the purpose is to classify wine quality, a dataset containing each quality value is needed. When the data set used is examined, it is seen that the quality values contain the set 3,4,5,6,7,8. In addition, there are 11 features to determine the quality value in the data set. These features are named as follows;

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

	fixed acidity	volatile acidity	citric acid	...	alcohol	quality
0	7.4	0.700	0.00	...	9.4	5
1	7.8	0.880	0.00	...	9.8	5
2	7.8	0.760	0.04	...	9.8	5
3	11.2	0.280	0.56	...	9.8	6
4	7.4	0.700	0.00	...	9.4	5
...	...	...	...	...	...	...
1138	6.3	0.510	0.13	...	11.0	6
1139	6.8	0.620	0.08	...	9.5	6
1140	6.2	0.600	0.08	...	10.5	5
1141	5.9	0.550	0.10	...	11.2	6
1142	5.9	0.645	0.12	...	10.2	5

Figure 1: Data frame

3 Data Preprocessing

Data preprocessing is the process of preparing raw data and fitting it into a machine learning model. It is the first and most important step when building a machine learning model. When creating a machine learning project, it is not always possible to come across clean and formatted data. And when doing any operation with data, it is imperative to clean it and put it in a formatted way. For this, data must be preprocessed. When the data set we use is examined, it is seen that the values of the features do not have the same scale. This is a factor that can negatively affect the network we will train. For this reason, the Standard Scaler process was applied to the values of the features.

3.1 Standard Scaler

The variables are translated into a distribution with a mean of 0 and a standard deviation of 1. It can be found by subtracting the corresponding column mean from all the data in the data set and dividing it by the column standard deviation. Thus, it is ensured that all observation units in the data set take values between -1 and 1.

$$z = \frac{x - u}{s} \tag{1}$$

4 Hyperparameter Tuning

Hyperparameter tuning captures a snapshot of the current performance of a model, and compares this snapshot with others taken previously. In any machine learning algorithm, hyperparameters need to be initialized before a model starts the training. Fine-tuning the model hyperparameters maximizes the performance of the model on a validation set. On the other hand, the values of model parameters are derived via training the data. Model parameters refer to the weights and coefficients, which are derived from the data by the algorithm. GridSearchCV was used in this study. Therefore, it is also important to understand the Cross validation method. Cross validation

(CV) is a statistical method used to estimate the accuracy of machine learning models. Assurance is needed regarding the accuracy of the prediction performance of the model. To evaluate the performance of a machine learning model, some unseen data are needed for the test. Based on the model's performance on unseen data, we can determine whether the model is underfitting, overfitting, or well-generalized. Cross-validation is considered a very helpful technique to test how effective a machine learning model is when the data in hand are limited Elgeldawi et al. [2021].

4.1 Grid Search

The most intuitive traditional approach for performing hyperparameter optimization is perhaps Grid Search Shekar and Dagnew [2019]. It generates a Cartesian product of all possible combinations of hyperparameters. Grid Search trains the machine learning algorithm for all combinations of hyperparameters; this process should be guided by a performance metric, typically measured using the "cross-validation" technique on the training set. This validation technique ensures that the trained model obtains most of the patterns from the dataset. Grid Search is obviously the most straightforward hyperparameter tuning method. With this technique, we simply build a grid with each possible combination of all the hyperparameter values provided, calculating the score of each model, in order to evaluate it, and then selecting the model that gives the best resultsElgeldawi et al. [2021]. To perform Grid Search, one selects a finite set of reasonable values for each hyperparameter; the Grid Search algorithm then trains the model with each combination of the hyperparameters in the Cartesian product. The performance of each combination is evaluated on a held-out validation set or through internal cross-validation on the training set. Finally, the Grid Search algorithm outputs the settings that achieve the highest performance in the validation procedure. The best set of hyperparameter values chosen in the Grid Search is then used in the actual model. Grid Search guarantees the detection of the best hyperparameters. However, one of its drawbacks is that it suffers severely when it comes to rapid convergence and dimensionalityLorenzo et al. [2017].

5 Data Augmentation

The best way to make a machine learning model more successful is to train it on more data. However, there is a limit to the amount of data owned. One way to get around this problem is to create synthetic data and add it to the data set. However, while performing this process, it should be done by paying attention to what kind of data set is workingGoodfellow et al. [2016].Data augmentation is also used in imbalanced classificationRamyachitra and Manikandan [2014]. One of the problems with imbalanced classification is that there are too few samples in the minority class for a classifier to be successfully trained. Thus, successful results cannot be obtained when the classifier is used for classification. One way to solve this problem is to use data augmentation techniques. Different data augmentation techniques have been suggested in the literature. In this study, SMOTE technique is used among the techniques in the literature.

5.1 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE is a statistical technique used to increase the number of data in a data set in a balanced wayChawla et al. [2002]. However, there is an important difference between the techniques described in the classical data augmentation section and the SMOTE technique. When the classical techniques are examined, it is seen that new images are obtained by applying operations such as rotating,

zooming, changing the brightness. However, in the SMOTE technique, the attributes of the images are determined first. Then, the determined features are associated with the neighboring features. As a result of this association, new data is produced. Different from other methods is that the work space is a feature fieldYavaş et al..

6 Methods

This section describes the methods used during the study.

6.1 Machine Learning and Deep Learning

Interactions in constantly used social media accounts, search engines and traces left behind when searching, movements made with bank accounts, blogs, mails, sensors in technological devices, biomedical data, images, music are seen as elements that support the concept of big data. Moreover, it seems that these data sources will continue to increase. This deluge of data calls for automated methods of data analysis, which is what machine learning provides. In particular, machine learning is defined as a method that can automatically detect patterns in data using mathematical and statistical methods, makes predictions about the future patterns with these inferences, or to perform other kinds of decision making under uncertainty. Hence, this method becomes more significant day by dayMurphy [2012].

Deep Learning is a subfield of machine learning that deals with algorithms inspired from the structure and function of the brain called artificial neural networks. This method allows computers to learn from experiences and define each concept by considering its hierarchical relationship with other concepts. In this approach, which collects information from experience, people do not need to specify all the information computers need for machine learning. The hierarchy of concepts enables the computer to learn complex concepts by building them from simpler ones. Thus, a deep structure is formed with many layers. To understand deep learning well, it is necessary to have a solid understanding of the fundamentals of machine learningGoodfellow et al. [2016].

6.2 Artificial Neural Network

Artificial neural networks are computational networks that try to simulate the neuronal networks in the biological nervous system in a large way. It is the foundation of artificial intelligence (AI) and solves problems that prove impossible or difficult by human or statistical standards. ANNs have self-learning capabilities that enable them to produce better results as more data is available. ANNs use backpropagation algorithms, to perfect their output results Graupe [2013].

6.3 Classification

The concept of classification is to distribute data among various classes defined on a data set. Classification algorithms learn this way of classification by applying certain algorithms to the training data, and then they try to correctly classify the test data for which the class label is not certain.

6.4 Multilayer Perceptron

Multi-layer Perceptron (MLP) is a supervised learning algorithm that learns a function $f(\cdot) : R^m \rightarrow R^o$ by training on a data set, where m is the number of dimensions for input and is the number of dimensions for output. Given a set of features $X = x_1, x_2, x_3, ..., x_m$ and a target , it can learn a non-linear function approximator for either classification or regression. It is different from logistic regression, in that between the input and the output layer, there can be one or more non-linear layers, called hidden layers. Figure 1 shows a one hidden layer MLP with scalar output.

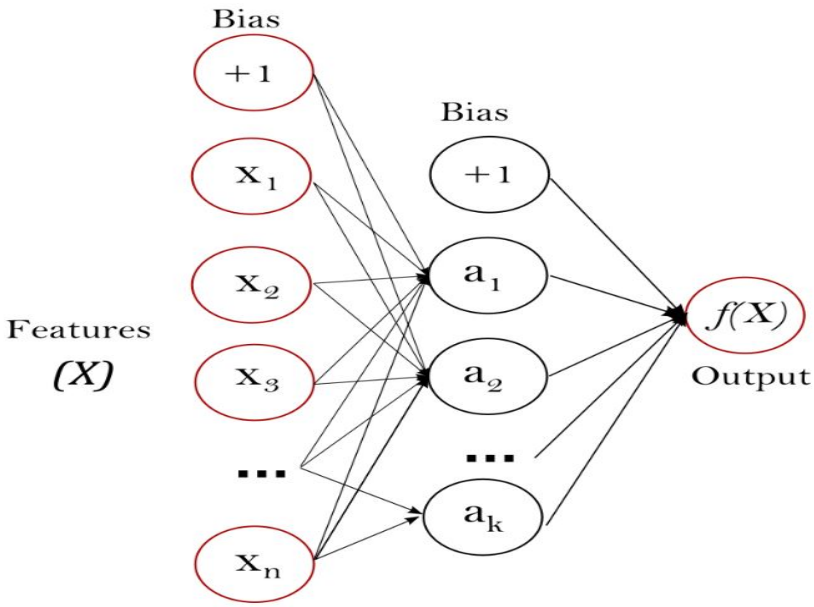

Figure 2: One hidden layer MLP

The leftmost layer, known as the input layer, consists of a set of neurons $x_i = x_1, x_2, ..., x_m$ representing the input features. Each neuron in the hidden layer transforms the values from the previous layer with a weighted linear summation $w_1x_1 + w_2x_2 + ... + w_mx_m$ followed by a non-linear activation function $g(\cdot) : R \rightarrow R$ - like the hyperbolic tan function. The output layer receives the values from the last hidden layer and transforms them into output values.

7 Evaluation Metrics

Some evaluation metrics are used to evaluate the performance of trained classifiers. In this section, these evaluation metrics are introducedHossin and Sulaiman [2015]. The meaning of the expressions used in the explanations are as follows;

- TP = True Positive
- TN = True Negative

- FP = False Positive
- FN = False Negative

7.1 Accuracy

The accuracy metric measures the ratio of correct predictions number over the total number of samples. Accuracy formula can be represented as Equation (2).

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{2}$$

7.2 Precision

The precision metric measure the ratio of the positive patterns that are correctly predicted over the total predicted patterns in a positive class. Precision formula can be represented as Equation (3).

$$\frac{TP}{TP + FP} \tag{3}$$

7.3 Recall

The recall (also known as sensitivity) metric is used to measure the fraction of positive patterns that are correctly classified. Recall formula can be represented as Equation (4).

$$\frac{TP}{TP + TN} \tag{4}$$

7.4 F1-Score

F1-Score metric represents the harmonic mean between recall and precision values. For imbalanced dataset, this metric provides a more robust decision about the model. F1-Score formula can be represented as Equation (5).

$$2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

7.5 Confusion Matrix

The confusion matrix summarizes the classifier's classification performance based on some test data. A confusion matrix has two-dimensions, one dimension is indexed by the actual class, the other is indexed by the class that the classifier predictsSammut and Webb [2011].

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

Table 1: Confusion Matrixcon [2021]

8 Applications & Results

First of all, the MLP model is trained without using any data processing techniques. As a result of this training process, the accuracy rate was found to be 54%. Then, the StandardScaler process was applied to the data with different value ranges, and the attribute values were passed to the same value range. The accuracy rate of the MLP trained on these data was found to be 58%. Confusion matrix, Loss curve and classification report are shown in Figure 3, Figure 4 and Table 2, respectively.

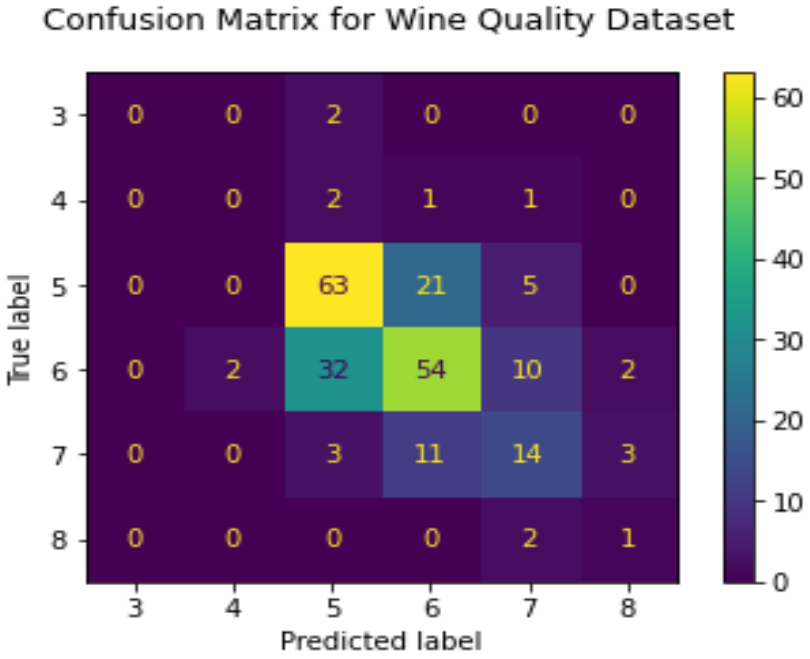

Figure 3: Confusion Matrix for Imbalanced Dataset without Hyperparameter Tuning

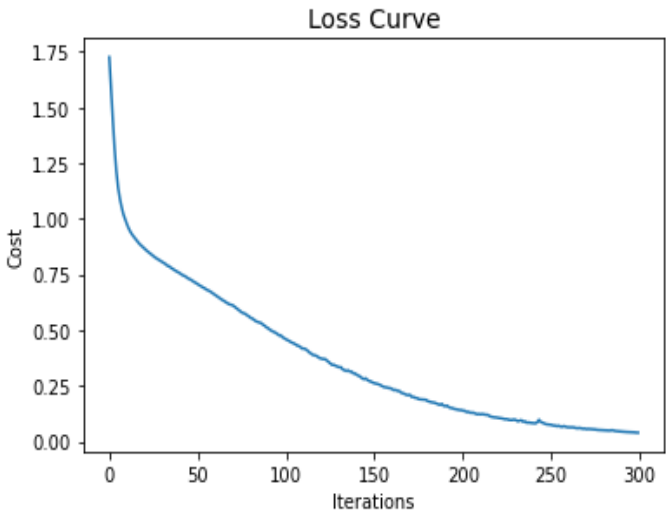

Figure 4: Loss Curve for Imbalanced Dataset without Hyperparameter Tuning

Classification Report				
	Precision	Recall	F1-Score	Support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	4
5	0.62	0.71	0.66	89
6	0.62	0.54	0.58	100
7	0.44	0.45	0.44	31
8	0.17	0.33	0.22	3
Accuracy			0.58	229
Macro avg	0.31	0.34	0.32	229
Weighted avg	0.57	0.58	0.57	229

Table 2: Classification Report for Imbalanced Dataset without Hyperparameter Tuning

The most successful hyperparameters were found using GridSearchCV. The parameters found are as follows.

- 'activation': 'relu'
- 'alpha': 0.0001
- 'hidden layer sizes': (150, 100, 50)
- 'learning rate': 'invscaling'
- 'max iter': 150
- 'solver': 'adam'

Then, the model was retrained with these parameters and it was seen that the accuracy increased to 62%. Confusion matrix, Loss curve and classification report are shown in Figure 5, Figure 6 and Table 3, respectively.

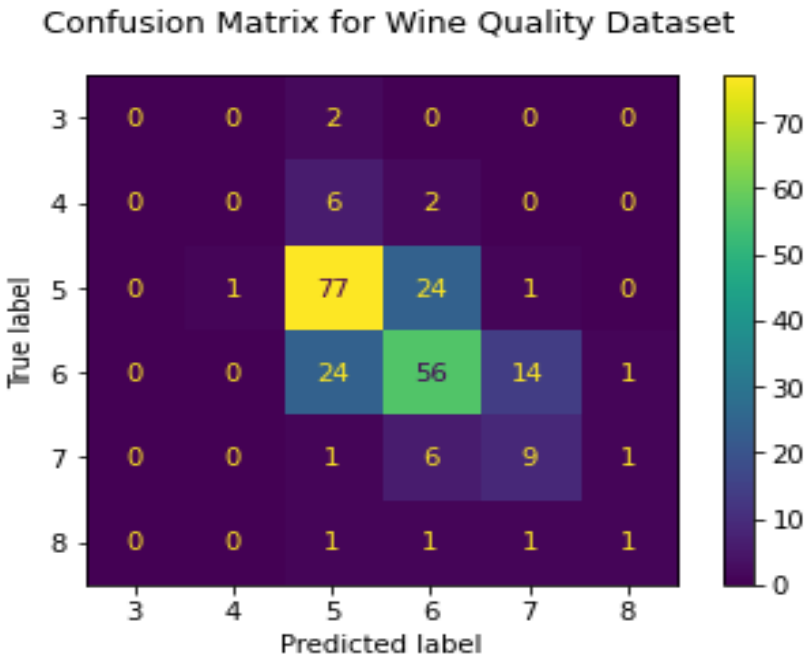

Figure 5: Confusion Matrix for Imbalanced Dataset with Hyperparameter Tuning

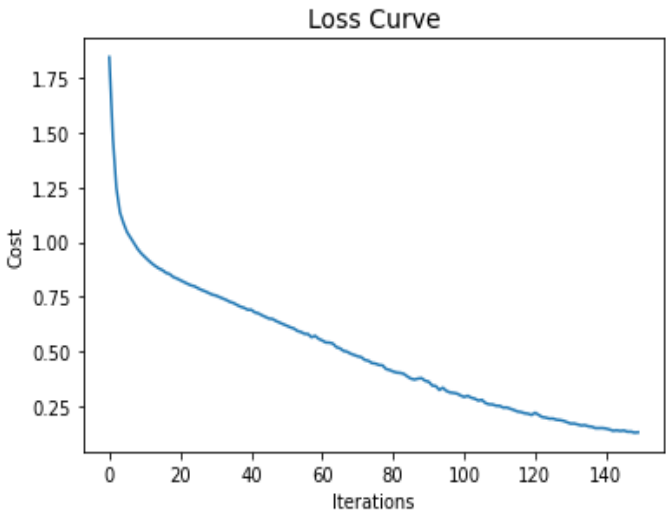

Figure 6: Loss Curve for Imbalanced Dataset with Hyperparameter Tuning

Classification Report				
	Precision	Recall	F1-Score	Support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	8
5	0.69	0.75	0.72	103
6	0.63	0.59	0.61	95
7	0.36	0.53	0.43	17
8	0.33	0.25	0.29	4
Accuracy			0.62	229
Macro avg	0.34	0.35	0.34	229
Weighted avg	0.61	0.62	0.61	229

Table 3: Classification Report for Imbalanced Dataset with Hyperparameter Tuning

When the data set seen in Table 4 is examined, it is seen that the data numbers of the classes are not equal to each other. This situation is called as unbalanced data set in the literature. The unbalanced data set is a factor that prevents the trained model from being successful. Because there are very few examples of successful training of a classifier in the minority class. One way to solve this problem is to use data augmentation techniques. In this study, the SMOTE technique, which is one of the data augmentation techniques in the literature, was used. The imbalance was eliminated by setting the data count to 500 for all classes.

Dataset		
Quality value	Imbalanced Dataset	Balanced Dataset
3	6	500
4	33	500
5	483	500
6	462	500
7	143	500
8	16	500
Total:	1143	3000

Table 4: Summary of Dataset

The MLP model was trained using the optimal hyperparameters with the dataset obtained after the SMOTE process. After this training, the accuracy value was found to be 89%. Confusion matrix, Loss curve and classification report are shown in Figure 7, Figure 8 and Table 5, respectively.

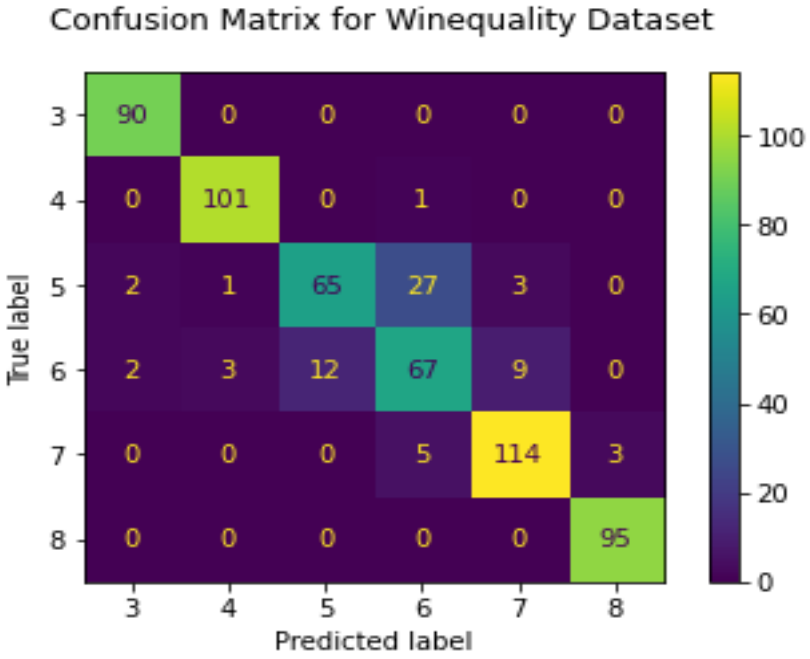

Figure 7: Confusion Matrix for Each Class has 500 sample

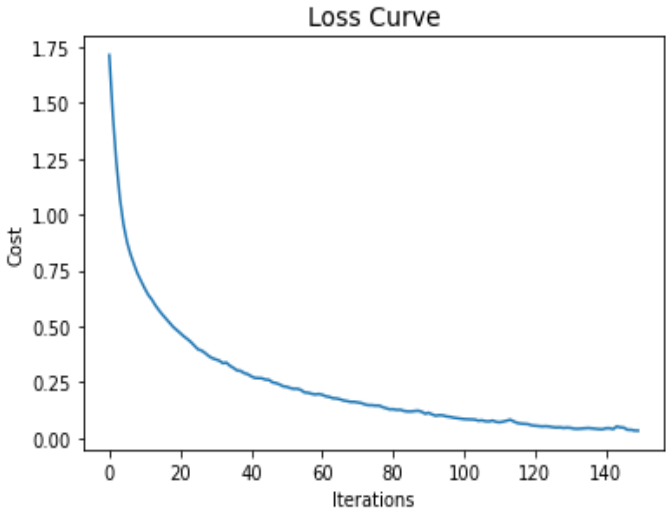

Figure 8: Loss Curve for Each Class has 500 sample

Classification Report				
	Precision	Recall	F1-Score	Support
3	0.96	1.00	0.98	90
4	0.96	0.99	0.98	102
5	0.84	0.66	0.74	98
6	0.67	0.72	0.69	93
7	0.90	0.93	0.92	122
8	0.97	1.00	0.98	95
Accuracy			0.89	600
Macro avg	0.88	0.88	0.88	600
Weighted avg	0.89	0.89	0.88	600

Table 5: Classification Report for Each Class has 500 sample

Then, using SMOTE, the number of data for all classes was set to 1000 and the MLP model was trained again.

Dataset		
Quality value	Imbalanced Dataset	Balanced Dataset
3	6	1000
4	33	1000
5	483	1000
6	462	1000
7	143	1000
8	16	1000
Total:	1143	6000

Table 6: Summary of Dataset

After this training, the accuracy value was found to be 95%. Confusion matrix, Loss curve and classification report are shown in Figure 9, Figure 10 and Table 7, respectively.

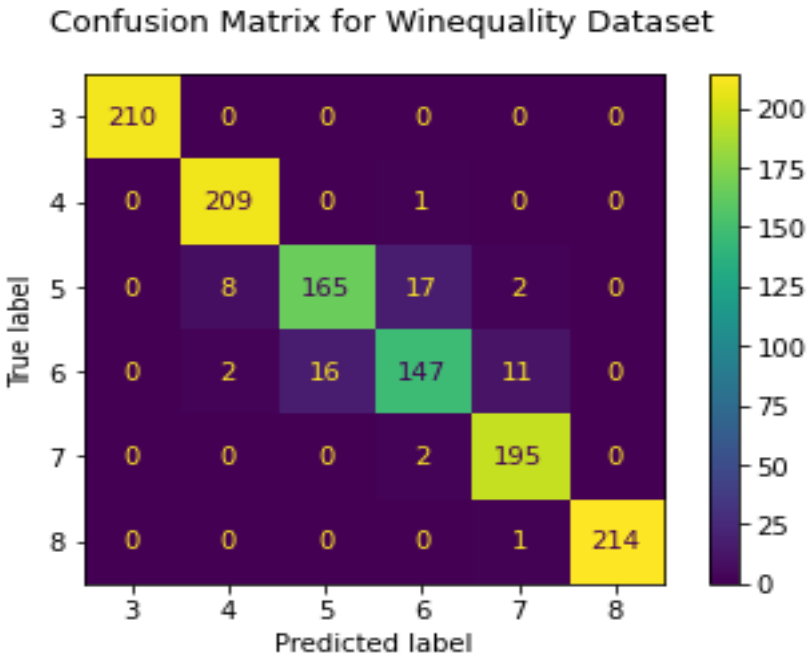

Figure 9: Confusion Matrix for Each Class has 1000 sample

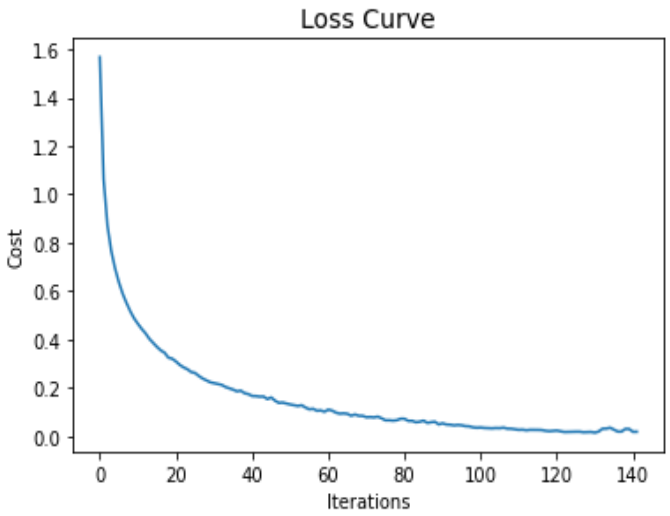

Figure 10: Loss Curve for Each Class has 1000 sample

Classification Report				
	Precision	Recall	F1-Score	Support
3	1.00	1.00	1.00	210
4	0.95	1.00	0.97	210
5	0.91	0.86	0.88	192
6	0.88	0.84	0.86	176
7	0.93	0.99	0.96	197
8	1.00	1.00	1.00	215
Accuracy			0.95	1200
Macro avg	0.95	0.95	0.95	1200
Weighted avg	0.95	0.95	0.95	1200

Table 7: Classification Report for Each Class has 1000 sample

9 Conclusion

In this study, the determination of wine quality was carried out depending on the wine characteristics. Since the study was carried out using Multi-layer Perceptron, processing techniques were used like StandardScaler on the data set. It was seen that this standardization process had a positive effect on the success rate. In addition, it was observed that hyperparameters affect model success. In order to make this inference, the GridSearchCV method was applied to the data set and optimal hyperparameters were found. As a result of this optimization process, it was seen that the success was positively affected. Considering that the data set used has an uneven distribution, the model cannot be expected to work successfully enough. For this reason, the data set was expanded using SMOTE, which is a data augmentation technique. It was observed that the contribution of syn-

thetic data to the success of the system was significantly higher. As a result, the wine classification problem was examined from various perspectives and a successful classifier was trained.

References

Enas Elgeldawi, Awny Sayed, Ahmed R Galal, and Alaa M Zaki. Hyperparameter tuning for machine learning algorithms used for arabic sentiment analysis. In *Informatics*, volume 8, page 79. Multidisciplinary Digital Publishing Institute, 2021.

BH Shekar and Guesh Dagnew. Grid search-based hyperparameter tuning and classification of microarray cancer data. In *2019 second international conference on advanced computational and communication paradigms (ICACCP)*, pages 1–8. IEEE, 2019.

Pablo Ribalta Lorenzo, Jakub Nalepa, Michal Kawulok, Luciano Sanchez Ramos, and José Ranilla Pastor. Particle swarm optimization for hyper-parameter selection in deep neural networks. In *Proceedings of the genetic and evolutionary computation conference*, pages 481–488, 2017.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

D Ramyachitra and P Manikandan. Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, 5(4):1–29, 2014.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Mustafa Yavaş, Aysun Güran, and Mitat Uysal. Covid-19 veri kümesinin smote tabanlı örnekleme yöntemi uygulanarak sınıflandırılması. *Avrupa Bilim ve Teknoloji Dergisi*, pages 258–264.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Daniel Graupe. *Principles of artificial neural networks*, volume 7. World Scientific, 2013.

Mohammad Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

Claude Sammut and Geoffrey I Webb. *Encyclopedia of machine learning*. 2011.

Confusion matrix — Wikipedia, the free encyclopedia, 2021.