Application of SMOTE Data Augmentation on Imbalanced Wine Quality Dataset Regression

1st Kevin Wijaya
School of Computer Science
Binus University
Jakarta, Indonesia
kevin.wijaya018@binus.ac.id

2nd Abyaan Syauqi Muhammad
School of Computer Science
Binus University
Bogor, Indonesia
abyaan.muhammad@binus.ac.id

3rd Jordan Rubin Kurnia
School of Computer Science
Binus University
Bandung, Indonesia
jordan.kurnia@binus.ac.id

4th Albert
School of Computer Science
Binus University
Jakarta, Indonesia
albert037@binus.ac.id

5th Muhammad Abdullah Faqih
School of Computer Science
Binus University
Jakarta, Indonesia
muhammad.faqih002@binus.ac.id

6th Ivan Halim Parmonangan
School of Computer Science
Binus University
Jakarta, Indonesia
ivan.parmonangan@binus.edu

***Abstract*—Wine judges and testers are notoriously inconsistent at judging a wine's quality. In an effort to make a more objective and consistent prediction of wine quality, we endeavored on making a machine learning regression model based on physiochemical indicators in wine to predict wine quality. This research specifically explores SMOTE as a data augmenting method in order to solve the poor regression model performance in previous researches because of an imbalanced dataset. We found that SMOTE significantly increase all model performance in this particular task. The best model, that we found in this research is a Random Forest model with SMOTE that produced a score of 0.166 MSE, 0.959 R^2-Score, 0.043 MAPE, and 0.407 RMSE.**

***Index Terms*—SMOTE, Wine, Regression, Random Forest, TabNet, XGBoost**

I. Introduction

Wine quality prediction has been an elusive and subjective craft since its beginning. It has been known that wine testers are notoriously unreliable at predicting wine quality blindly [7]. It is not uncommon to see wines win gold in one competition, yet fail to win medals in others. In fact, according to research by Hodgson (2008), only 10% of wine judges managed to replicate their scores within a single medal group. Perceived wine quality can also be impacted by other external factors such as the types of glassware used [11]. This makes determining wine quality via human testing extremely subjective.

Knowing these facts, many people have tried to find an objective way to predict wine quality. One of the ways that people have tried is by training machine learning models on wine datasets to act as an objective agent in determining wine quality. Most of these researches have mainly focused on classification tasks and suffer from unequal data distribution on both ends of the spectrum because most wines are rated average. This has resulted in a severe bottleneck in the model's output quality and its inability to accurately discern both bad and good wines, rating almost all wines as average [6]. This phenomenon happened because, in the presence of an imbalanced dataset, machine learning models tend to prioritize reducing the error rates for the majority classes at the expense of the minority classes to increase their performance [14].

In this research, we decided to explore the use of data augmentation techniques [9] in a regression model on a wine quality dataset to improve the quality of the machine learning models in this task and fill in the existing research gap.

II. Literature Review

There have been several pieces of research done that are being used as the backbone for our research endeavor on this topic. Here are some of them:

Research by Dahal et al. (2021) [6] indicated that Gradient Boosting Regressor (GBR) performs the best compared to Ridge Regression (RR), SVM, and multi-layer ANN with a score of 0.3741 MSE, 0.6057 R^2-Score, and 0.0873 MAPE for regression task in red wine dataset.

A study by Kumar et al. (2020) [10] discovers that the due to imbalanced red wine dataset that they used, it resulted in severely decreased accuracy at the margins of their prediction result.

A previous study by Hu et al. (2016) [8] has indicated that SMOTE can provide substantial improvement in the classification of the imbalanced white wine dataset. They have also found that Random Forest provides the best result among the models they tested.

In a study by Shaw et al. (2020) [12], they also found that in wine classification Random Forest managed to outperform SVM and Multilayer Perceptron scoring 0.8196 accuracy. However, some research on other datasets [15] suggests that XGBoost actually outperforms Random Forest on tabular datasets.

Considering the facts discovered by previous pieces of research, we decided to test SMOTE as a data augmentation on a regression task as it had not been done before and seems to produce promising results in the classification task of a similar dataset. Aside from that, we wanted to compare the results on the current state-of-the-art machine learning models for tabular datasets to achieve maximum results, so we use Random Forest and XGBoost as a point of comparison. Aside from that, no research seems to have explored newer deep learning models made specifically for tabular datasets such as TabNet for any wine quality dataset. So we decided to explore this avenue and observe the impact of SMOTE on deep learning models and compare them to more traditional machine learning models.

III. Methodology

In order to prepare and train our machine learning model we needed to do several things.

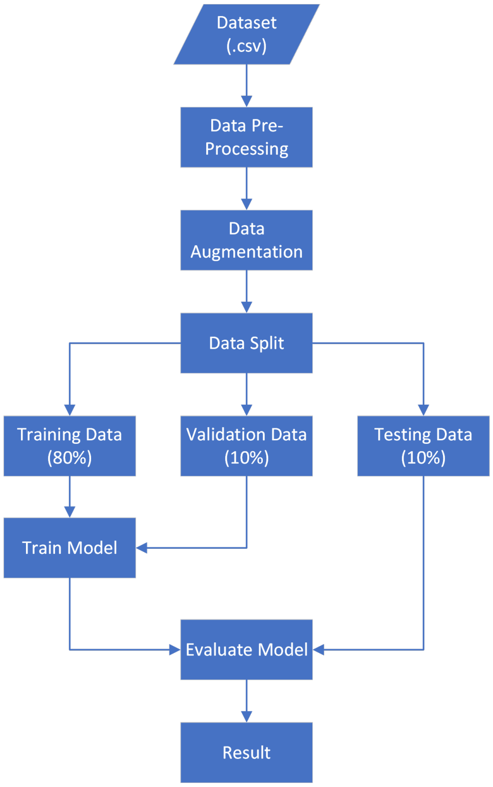

Fig. 1. Project Workflow

The workflow diagram on Fig. 1 represents our workflow step-by-step in making this research. In general, we first collect our dataset, then process and clean-up our data. After that, we do model selection, training, and evaluation.

A. Dataset

The dataset used in this research is a dataset from a research by P. Cortez (2009) [5]. The dataset is obtained from a wine variant from Portugal called *vinho verde*. It contained 6497 data with a distribution of 25% white wine and 75% red wine. There are 12 independent variable (wine type, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol) to predict the quality (dependent variable) of the wine. This dataset or some variant of it has been the dataset used for most wine quality machine learning research done since its creation.

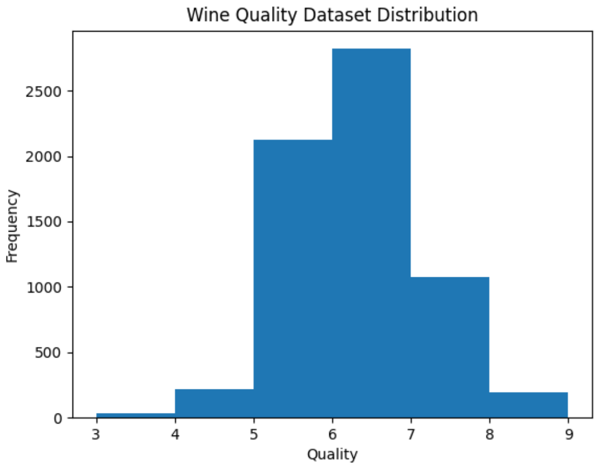

Fig. 2. Dataset Distribution Based On Wine Quality

TABLE I
Dataset Example

Type	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Free Sulfur Dioxide	Total Sulfur Dioxide	Chlorides	Density	pH	Sulphates	Alcohol	Quality
white	7	0.27	0.36	20.7	45	170	0.045	1.001	3	0.45	8.8	6
white	6.3	0.3	0.34	1.6	14	132	0.049	0.994	3.3	0.49	9.5	6
white	8.1	0.28	0.4	6.9	30	97	0.05	0.9951	3.26	0.44	10.1	6
white	7.2	0.23	0.32	8.5	47	186	0.058	0.9956	3.19	0.4	9.9	6
white	7.2	0.23	0.32	8.5	47	186	0.058	0.9956	3.19	0.4	9.9	6

From Fig. 2, we can see that the dataset distribution is heavily skewed to the center with most wine being scored 5, 6, or 7 while leaving the other score with a very minimal portion of the dataset. The impact of this distribution and how we address it will be discussed further in the data augmentation segment.

B. Data Pre-Processing

First, we filter out incomplete data in the dataset. After filtering, we found that only 6463 data are usable for experimentation. Next, we encode the wine type variable into an integer of 1 for white wine and 0 for red wine. Next, we use MinMaxScaler to scale the independent variables to normalize the data and improve the model outcome.

C. Data Augmentation (SMOTE)

As we can see from Fig. 2, the dataset distribution is heavily skewed to the center with most wine being scored 5, 6, or 7, while leaving the other score with a very minimal portion of the dataset. This causes the inaccuracy of prediction on either end of the quality score spectrum from previous research and in our opinion, contributes t1o the stifling of the model's accuracy in several other research. In order to solve this, we implemented the Synthetic Minority Oversampling Technique (SMOTE) data augmentation/oversampling technique [3]. SMOTE will create new data with independent variable values between the data already available. We decided to use SMOTE with 4 k_neighbors as a parameter because that is the maximum amount of neighbors that we can use due to our limited dataset. This may cause a lack of variation in our augmented dataset which could inflate the scores in the metrics that these models are tested at, whilst underperforming in real-world applications. After utilizing SMOTE, we ended up with an equalized dataset of 19740 distributed equally among all quality scores.

D. Random Forest

Random Forest is an ensemble machine learning model based on decision trees [2]. The advantage of Random Forest compared to traditional decision trees is its resistance to overfitting and can handle large datasets with high dimensionality. We decided to use Random Forest in this task because we saw that in other research about wine quality classification, Random Forest has in general performed the best among other machine learning models. In view of this, we'd like to try applying Random Forest models to the regression task of a similar dataset. In this experiment, we decided to use the default parameters of Random Forest.

E. XGBoost

XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting machine learning algorithm [4]. We decided to use XGBoost because it has been established in several pieces of research that XGBoost is one of the best machine-learning algorithms for tabular data [13]. Besides that, the previous research on wine quality regression found that the standard Gradient Boosting Algorithm performs the best among the models they tested for wine quality regression [6]. In this experiment, we decided to use the default parameters of XGBoost.

F. TabNet

TabNet is a deep learning model explicitly designed for tabular data [1]. It is a variant of Deep Neural Network (DNN) made by Google that can do tabular data classification and regression tasks better than other deep learning models. We choose this model because we feel that using a deep learning model for tabular data is under-explored, and specifically for wine quality prediction, TabNet has never been used yet. We also want to see how deep-learning models would respond to a more balanced dataset by data augmentation as compared to machine-learning models. In this experiment, we decided to run the TabNet up to 200 epochs.

G. Evaluation Metrics

We partitioned the dataset into 80% training set, 10% validation set, and 10% test set. We use four standard evaluation metrics to examine the regression models in this research. Those are Mean Squared Error (MSE), R^2-Score, Mean Absolute Percentage Error (MAPE), and Root Mean Squared Error (RMSE).

IV. Results and Discussion

This research is carried out using Google Colab as an IDE and processing platform. Here are the results of our experiment on Random Forest, XGBoost, and TabNet before and after using SMOTE.

TABLE II
Experiment Result Before SMOTE

Algorithms	**Metrics**			
	MSE	***R^2-Score***	***MAPE***	***RMSE***
Random Forest	0.385	0.491	0.081	0.62
XGBoost	0.0402	0.468	0.084	0.634
TabNet	0.468	0.382	0.095	0.684

TABLE III
Experiment Result After SMOTE

Algorithms	**Metrics**			
	MSE	***R^2-Score***	***MAPE***	***RMSE***
Random Forest	0.166	0.959	0.043	0.407
XGBoost	0.236	0.942	0.061	0.485
TabNet	0.285	0.93	0.068	0.534

From table II and III, we conclude that SMOTE does improve the performance of machine learning models substantially. Even the worst model (TabNet) managed to outperform the best model (Random Forest) after applying SMOTE. In this experiment, the best-performing model is the Random Forest model after SMOTE with a score of 0.166 MSE, 0.959 R^2-Score, 0.043 MAPE, and 0.407 RMSE.

This very good result is attributed to the balanced dataset acquired after the application of SMOTE. However, this data augmentation may come at a cost. Due to the small amount of data points that we have, SMOTE's algorithm has to do a lot of extrapolating from the existing data. In fact, we do have to use a lower k-neighbors of 4 due to this exact limitation. This may cause SMOTE to generate very similar data that may not capture a lot of the characteristics and variation possible in a certain data point. Thus, the resulting model might not perform as well in real-world scenarios or different datasets.

V. Conclusion

From our research, we can conclude that SMOTE can substantially increase the quality of machine learning models in predicting wine quality from imbalanced data to the point that the worst-performing model after SMOTE (TabNet) managed to outperform the best model before SMOTE (Random Forest) with a score of 0.285 MSE vs 0.385 MSE. However, this model might be unable to perform as well in a different dataset due to the small amount of data points from which SMOTE generated its data. The best-performing model in this research is the Random Forest after SMOTE model with a score of 0.166 MSE, 0.959 R^2-Score, 0.043 MAPE, and 0.407 RMSE.

In the future, we would like to see more data augmentation/oversampling techniques being explored for this dataset. We also want to have a more complete wine dataset with wine from multiple countries to capture more diverse variations in regional preferences and characteristics. Finally, the result in this paper can still be further amplified by tweaking the parameters of each model, particularly on Random Forest and XGBoost as we only used the default parameter for this experiment and mainly focused on SMOTE application.

References

[1] Sercan Ö Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6679–6687, 2021.

[2] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[5] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. Smart Business Networks: Concepts and Empirical Evidence.

[6] K R Dahal, J N Dahal, H Banjade, and S Gaire. Prediction of wine quality using machine learning algorithms. *Open J. Stat.*, 11(02):278–289, 2021.

[7] Robert T. Hodgson. An examination of judge reliability at a major u.s. wine competition. *Journal of Wine Economics*, 3(2):105–113, 2008.

[8] Gongzhu Hu, Tan Xi, Faraz Mohammed, and Huaikou Miao. Classification of wine quality with imbalanced data. In *2016 IEEE International Conference on Industrial Technology (ICIT)*, pages 1712–1217, 2016.

[9] Pradeep Kumar, Roheet Bhatnagar, Kuntal Gaur, and Anurag Bhatnagar. Classification of imbalanced data:review of methods and applications. *IOP Conference Series: Materials Science and Engineering*, 1099(1):012077, mar 2021.

[10] Sunny Kumar, Kanika Agrawal, and Nelshan Mandan. Red wine quality prediction using machine learning techniques. In *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6, 2020.

[11] Wendy V. Parr, Claire Grose, Duncan Hedderley, Marcela Medel Maraboli, Oliver Masters, Leandro Dias Araujo, and Dominique Valentin. Perception of quality and complexity in wine and their links to varietal typicality: An investigation involving pinot noir wine and professional tasters. *Food Research International*, 137:109423, 2020.

[12] Bipul Shaw, Ankur Kumar Suman, and Biswarup Chakraborty. Wine quality analysis using machine learning. In *Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018*, pages 239–247. Springer, 2020.

[13] Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

[14] Fadi Thabtah, Suhel Hammoud, Firuz Kamalov, and Amanda Gonsalves. Data imbalance in classification: Experimental evaluation. *Information Sciences*, 513:429–441, 2020.

[15] Mehdi Zamani Joharestani, Chunxiang Cao, Xiliang Ni, Barjeece Bashir, and Somayeh Talebiesfandarani. Pm2. 5 prediction based on random forest, xgboost, and deep learning using multisource remote sensing data. *Atmosphere*, 10(7):373, 2019.